A Multi-Domain Process Design and Improvement Framework

by
Robert Nicol

B.S., Mechanical Engineering
University of Houston, 1993

S.M., Chemical Engineering
S.M., Management
Massachusetts Institute of Technology, 2001

Submitted to the Engineering Systems Division in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Engineering Systems
at the
Massachusetts Institute of Technology
February 2010
[JUNE 2010]

Signature of Author:______________________________
Engineering Systems Division
February 17, 2010

Certified by:______________________________
Christopher L. Magee
Professor of the Practice, Mechanical Engineering and Engineering Systems

Certified by:______________________________
Deborah J. Nightingale
Professor of the Practice, Aeronautics and Astronautics and Engineering Systems

Certified by:______________________________
Andrew B. Onderdonk
Professor of Pathology, Harvard Medical School

Accepted by:______________________________
Nancy G. Levenson
Professor of Aeronautics and Astronautics
Chair, Engineering Systems Division Education Committee

A Multi-Domain Process Design and Improvement Framework

by

Robert Nicol

Submitted to the Engineering Systems Division on February 17, 2010 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Engineering Systems

Abstract

Processes in manufacturing, services, and healthcare are complex socio-technical systems composed of intricately sequenced activities supported by elements drawn from multiple domains. While many of these processes offer high performance, their complexity can make their design, improvement, troubleshooting, and change difficult due to the many possible and unforeseen interactions between elements. This thesis develops a design methodology and multi-domain network model for complex process design, change management, process improvement, and troubleshooting. As part of the methodology a feasibility analysis method based on solving the minimum cost flow problem for a network of process alternatives is presented to identify feasible processes subject to stakeholder requirements and constraints including performance, flexibility, modularity, and other system properties. A model based on Multi-Domain Matrix (MDM) concepts is developed specifically for process analysis called the Multi-Domain Process Matrix model (MDPM) to enumerate and analyze the interactions between process elements such that process performance under change and troubleshooting scenarios can be improved. The graph theory basis of the MDPM model enables its analysis using a proposed set of metrics derived from communications, social, and process network literature.

As a demonstration of the use of the methodology, a complex DNA sequencing based surveillance process for Methicillin resistant Staphylococcus aureus (MRSA) in the US healthcare system is designed and a prototype implemented. Rapid advances in DNA-based technologies have greatly expanded the range of processes available to the clinical microbiology laboratory, however, integrating these new processes into a comprehensive surveillance system presents significant challenges. Many of these new technologies are still in early stages of development, require multidisciplinary teams to support them, and must undergo significant optimization presenting significant barriers to their rapid adoption despite the pressing need to understand and control antibiotic resistance. Data from the prototype MRSA surveillance process show significant variation at the DNA level between patient cases, providing evidence for the urgent need for a DNA sequencing based microbial surveillance process as part of clinical microbiology efforts in the US healthcare system. However, results of applying the process design methodology and MDPM model analysis indicate significant work remains to reduce complexity, further improve key technology elements, gain acceptance, develop key organizational infrastructure, and redesign the process to efficiently absorb the rapid technology change expected in DNA sequencing. The MDPM model is used to develop a roadmap of specific multi-domain projects addressing these issues to accelerate the deployment of a national DNA sequencing based surveillance system.

Thesis Supervisors:

Christopher L. Magee
Professor of the Practice, Mechanical Engineering and Engineering Systems

Deborah J. Nightingale
Professor of the Practice, Aeronautics and Astronautics and Engineering Systems

Acknowledgements

I wish to thank my advisors Professors Deborah Nightingale and Chris Magee for all their support, encouragement, and advice throughout my years at MIT and the research journey in this thesis. I also wish to thank Professor Andrew Onderdonk for his invaluable support of my research and incredible patience through all the trials in the research. Numerous ESD, MIT, Harvard, and Broad Institute members contributed to my learning, research, and encouragement and I would like to thank the community and the network it represents. Finally, I wish to thank my parents most of all for all their encouragement and support.

Table of Contents

List of Figures

List of Tables

List of Equations

Acronyms and Abbreviations

ABCs	:	Active Bacterial Core
AMA	:	American Medical Association
AMA	:	American Medical Association
AoN	:	Activity on node
ASM	:	American Society for Microbiology
CA-MRSA	:	Community Associated MRSA
CDC	:	Centers for Disease Control
CLIOS	:	Complex Large Interconnected Open Socio-technical
COIM	:	CLIOS-OPM Integrated Method
CPT	:	Current Procedural Terminology
DMM	:	Domain Mapping Matrices
DOE	:	Design of Experiments
DSM	:	Design Structure Matrix
DVM	:	Domain Version Matrices
EIP	:	CDC Emerging Infections Program Network
EPC	:	Event-driven process chains
ESM	:	Engineering Systems Matrix
HCPCS	:	Healthcare Common Procedure Coding System
HCUP	:	Health Care Cost & Utilization Project databases
HHS	:	Health and Human Services
ICD-9-CM	:	International Classification of Diseases – rev. 9 – clinical modification
IDEF	:	Integration DEFinition family of languages
IDEF0	:	Integration DEFinition function modeline
IDSA	:	Infectious Diseases Society of America
LIMS	:	Laboratory Information Management System
MCNF	:	Minimum cost network flow
MDM	:	Multiple-Domain Matrices
MDPM	:	Multi-Domain Process Matrix model

MIC	:	Minimum Inhibitory Concentration
MLST	:	Multi-locus sequence typing
MRSA	:	Methicillin Resistant Staphylococcus aureus
MSSA	:	Methicillin Sensitive Staphylococcus aureus
NCBI	:	National Center for Biological Information
NHSN	:	National Healthcare Safety Network (),
NIH	:	National Institutes of Health
NIS	:	Nationwide Inpatient Sample (HHS Cost Database)
NNIS	:	National Nosocomial Infections Surveillance System
OPM	:	Object Process Methodology
PCR	:	Polymerase Chain Reaction
PFGE	:	Pulse-Field Gel Electrophoresis (microbial typing)
PVL	:	Panton Valentine leukocidin
qPCR	:	quantitative PCR
SCCmec	:	Staphylococcus Chromosomal Cassette mec
SNP	:	Single Nucleotide Polymorphism
UML	:	Unified Modeling Language ()
VRE	:	Vancomycin Resistant Enterococcus
VRSA	:	Vancomycin Resistant Staphylococcus aureus

1 Introduction

Nearly every human activity can be described as a process. Everything from cooking dinner to launching a space shuttle to processing a loan can be broken down into a series of steps and instructions for performing them. While we think of process models in modern terms representing them as flowcharts and other sophisticated ways, human beings developed the essential machinery to represent processes several millennia ago with the invention of language. This enabled the codification and transmission of work instructions such that valuable discoveries and experience could be transmitted without each individual having to rediscover knowledge. This allowed for knowledge to build upon itself and for improvement to take place. The verbal instructions for how to build a stone ax, find food, and build shelter were the very first process models, useful abstractions of reality. The ability to build processes gave humans an enormous advantage as hunting, farming, construction, and social organization all became codified through "work instructions" of various kinds including formal written ones, experiential (apprenticeship) ones, and verbal ones. From that initial breakthrough of language humans have built ever more complex processes to the point that modern processes defy comprehension by any single human being, instead existing as distributed knowledge sets across multiple organizations and individuals. Our ability to build processes as complex as the Apollo moon missions, air traffic control, oil refining, heart transplant surgery, and modern supply chains necessarily requires the distribution and modularization of knowledge. And around these processes humans have built equally complex systems to support them. Our society relies on these complex processes and infrastructure for everything from communications to transportation to critical systems essential for its very survival, such as healthcare. It is therefore essential that we understand and control the complexity in them. That is the objective of process complexity research in Engineering Systems and the analysis tools described in this thesis are a contribution to this endeavor.

1.1 Motivation

The motivation for this thesis is two-fold:

1. To develop new tools and understanding relative to complex processes
2. To make progress in understanding antibiotic resistance, an important real world problem area

These two motivations are made explicit here because this thesis was developed around both as will be evident in reading this thesis. The first motivation stems from the author's extensive involvement in process design from professional experience in the petrochemical industry and high throughput DNA sequencing. The second motivation comes from the author's belief, research experience, and growing evidence that antibiotic resistance is a grave threat to human society, and must be addressed using systems approaches given the magnitude, complexity, and multi-disciplinary nature of the problem. The two motivations are connected where processes must be built to understand and manage antibiotic resistance. The intersection of complex process design and antibiotic resistance research is the focus of this thesis.

Bacteria are the oldest life form on planet earth, first appearing over 4 billion years ago, they have adapted to every environmental condition imaginable and changed the earth itself. Evolution over millennia has equipped bacteria to exploit every possible niche and adapt as a group to a rapid change in environmental conditions. An estimated 10^7 to 10^9 bacterial species are thought to exist on the planet as estimated by various surveys including (Curtis, Sloan et al. 2002), of which we have catalogued a minuscule 7,031 and fully sequenced the complete genomes of approximately 600 species as of 2008 as estimated by (Achtman and Wagner 2008). The vast majority of life on the planet is bacterial. By comparison human beings are very recent arrivals to the planetary eco-system, only appearing about 200,000 years ago. Our primate ancestors evolved in symbiosis with bacteria, and every human, composed of 10^{14} human cells, co-exists with 10^{15} bacteria in their body. The vast majority of these are beneficial, helping digest food, produce needed enzymes, and even fight infection from pathogenic bacteria. Only a tiny fraction of the bacteria on the planet are human pathogens, but this small group has caused enormous suffering over most of human history. It is only very recently, in the last 60 years, that bacterial infection has been controlled through modern antibiotics. But this advantage may be short lived. Bacteria have been fighting antibiotic warfare for millennia, and within all their species have evolved defenses against almost every chemical agent possible. Already many of the antibiotics in our medical arsenal are ineffective against new strains of pathogenic bacteria. The struggle to combat bacterial infection will be an ongoing one, not a permanent victory. This thesis addresses a crucial element in this struggle, the need for molecular level monitoring of pathogenic bacteria in hospital clinical microbiology processes.

The healthcare system within which these processes are performed is perhaps the most complex enterprise on the planet intertwined with nearly every infrastructure system from transportation to power, continuously incorporating new technologies, and impacting every person throughout their lives. The benefits of this system in the US have been enormous, increasing life expectancy from 47.3 years in 1900 to 77.7 years in 2008[1]. An important portion of this gain can be attributed to the control of infectious disease through vaccinations and antibiotics, which in turn facilitated the development of increasingly complex surgical interventions including organ transplants, vascular repair, and assistive devices. And as the interventions have grown more complex, so has the system supporting them in terms of the number of stakeholders, technologies, reimbursement methods, and organizational architecture. This growing system complexity makes the introduction of new processes difficult because of the many elements that must be addressed from the fundamental effectiveness of the technology to the reimbursement mechanism to the organizational location. A key objective of this thesis is to develop a methodology to model and analyze process complexity and feasibility in healthcare systems to identify key bottlenecks and speed the implementation of critically needed processes such as pathogenic bacteria surveillance.

[1] Source: CDC Statistics, N. C. f. H. (2009). Health, United States, 2008.

1.2 Research Opportunity

This thesis focuses on the analysis of healthcare processes from a multi-disciplinary Engineering Systems perspective including social science, management, and engineering dimensions. While disciplines such as chemical engineering have developed significant knowledge bases in physical process design, relatively little attention has been paid to "messy" socio-technical processes such as those found in healthcare. Yet, these processes are the basis for our modern health care system, and if it is to be improved, these processes must be improved. The work presented here combines three main research areas within engineering systems: first, the critically important health care area of clinical microbiology services delivery and its associated technology, second, enterprise architecture and tools for stakeholder identification and alignment, and third, network and complexity science. The research opportunity is in developing an integrated process analysis methodology that can be used to visualize the true extent of a process within a system and analyze its true complexity beyond the traditional analysis of physical workflow. This methodology is a first step to the optimization and rational design of socio-technical processes from a systems perspective in the engineering tradition.

1.2.1 Engineering Systems and Processes

As a new field of study, engineering systems has to date focused primarily on understanding large-scale infrastructure systems such as air traffic control, power distribution, petrochemical infrastructure, and manufacturing. However the processes embedded within these systems have received less attention, and their understanding is important as they can influence and even define their systems. Healthcare in particular is an extremely large and growing fraction of the economy, and it is largely composed of processes from routine checkups to vaccinations to heart transplants. Understanding the interaction between these processes and the healthcare system is critical to the "macro" perspective engineering systems can provide. Processes are in some sense the building blocks of healthcare delivery, since they are what is billed for, what is performed, what is offered at each hospital, what is reimbursed, what the staff is organized around, and what is trained for. One important goal of the research in this thesis is to expand the knowledge around processes from an engineering systems perspective.

1.2.2 Big P Processes

Much like "big M" manufacturing, which expands the traditional view of manufacturing away from the shop floor to include design, supply chain, finance, and other areas of the enterprise, processes need to undergo a similar broadening. And so "big P" processes expand away from the simple physical flowchart view to include stakeholders, information flows, alternative paths, alternative designs, system infrastructure, and other system elements. Including these other dimensions shows the true extent of a process, which cannot be captured by simply describing the physical process steps. The way the organization is structured can have a very significant impact on the structure of the process, such as poorly communicating stakeholders leading to fragmented or suboptimal processes. Conversely highly technically complex processes likely require substantial stakeholder and organizational communication to manage their complexity. The need for integration and coordination of multiple deep knowledge bases is critical to the management of processes within complex systems. The need for substantial improvement in modern processes, such as the significant cost performance now demanded from healthcare in particular, places an even stronger emphasis on ensuring complex processes are also able to change and adapt quickly despite their complexity. Understanding all of the elements associated with modern processes and their interactions is an important first step to managing change and enabling improvement.

1.3 Thesis

This thesis evolved from an initial desire to apply network theory to study processes in the healthcare system. A number of methodology alternatives were explored to provide a meaningful engineering systems insight to the vast complexity contained in the healthcare system. In particular, an approach to mine the national Health Care Cost & Utilization Project (HCUP) databases which contain reported cost data for a representative sampling of 20% of US hospitals for evidence of "learning curve" behavior following the adoption of a new technology and whether the variation on learning curves could be explained in terms of system variables such as stakeholder alignment, resource utilization, and process control. However, billing variation and presumed "gaming" of prices made the data difficult to correlate with any process improvement behavior in addition to the large complexity (and variability) of technologies and their implementation. In mining this data some interesting results did appear in terms of which procedures (processes) were associated with the highest hospitalization cost (irrespective of learning curve behavior) in the Nationwide Inpatient Sample (NIS, a subset of the HCUP databases). The top 20 categories mainly consist of well-known diseases such as heart disease, births, diabetes, pulmonary disease, and complications of aging. However, two related categories stood out as not belonging with the others and these were all associated with presumably preventable microbial infectious disease treatable with antibiotics. These cost categories in the HCUP data were: Pneumonia and Septicemia highlighted in yellow in Figure 1 along with the other top 10 categories.

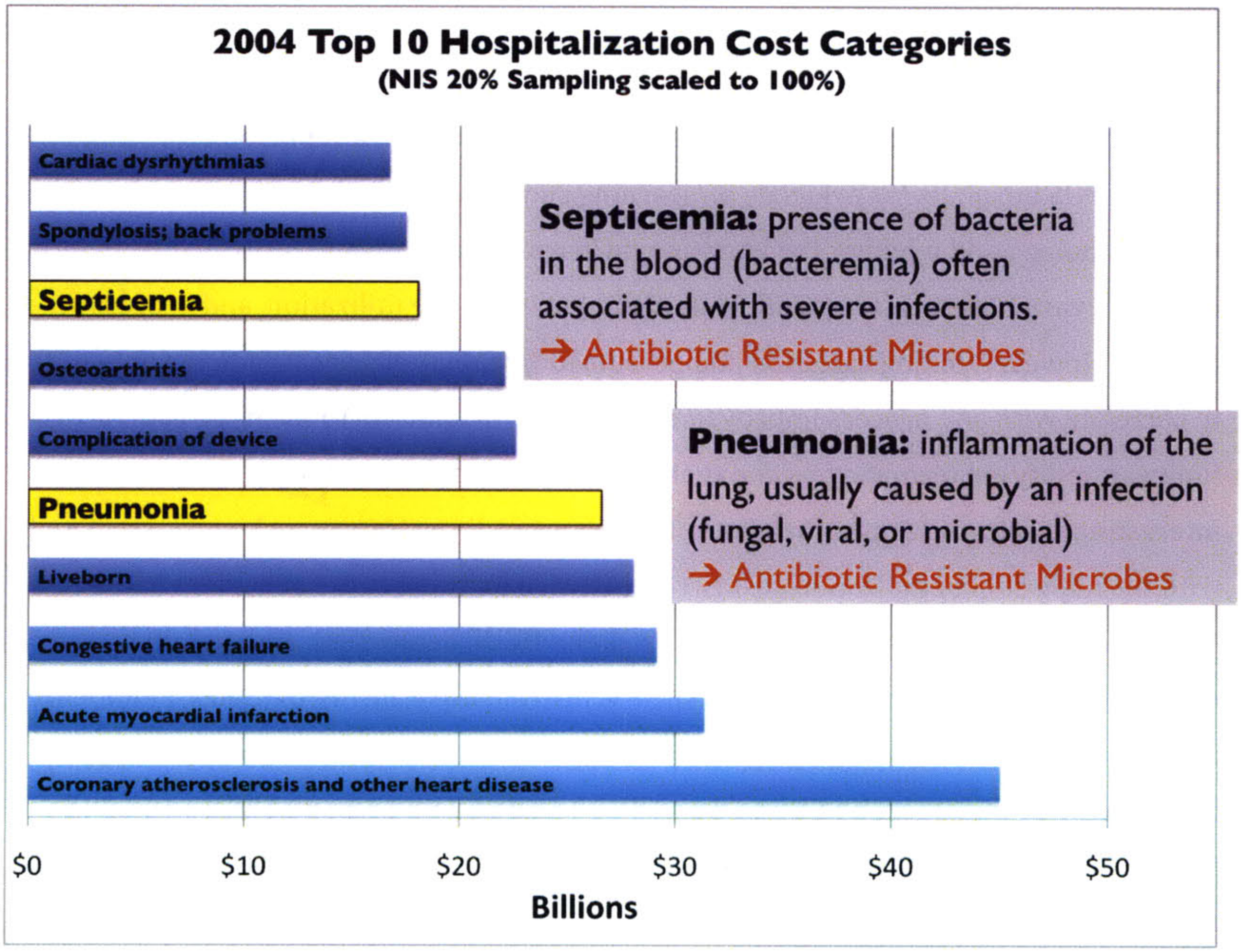


Figure 1 - Top 10 Hospitalization Cost Categories (NIS Data 2005)

On further investigation of these categories there were enormously high costs per patient for the septicemia category, which is defined[2] as:

> **Septicemia**: the presence of bacteria in the blood (bacteremia) and is often associated with severe disease.
>
> - Septicemia is a serious condition that requires a hospital stay. You may be admitted to an intensive care unit (ICU).
> - Fluids and medicines are given by an IV to maintain the blood pressure.
> - Oxygen will be given. Antibiotics are used to treat the infection.

[2] National Library of Medicine definition

- Plasma or other blood products may be given to correct any clotting abnormalities.

Based on this definition, the high costs seemed especially surprising given that antibiotics are usually highly effective. The HCUP databases could also be mined to produce a histogram of all associated procedure codes (i.e. what procedures were performed on patients with Septicemia) and the results were also surprising as shown in Figure 2 suggesting that these patients were being admitted and treated with a wide range of procedures not necessarily associated with antibiotic therapy. Or was it possible the antibiotic therapy was not working and therefore all these other procedures were necessary?

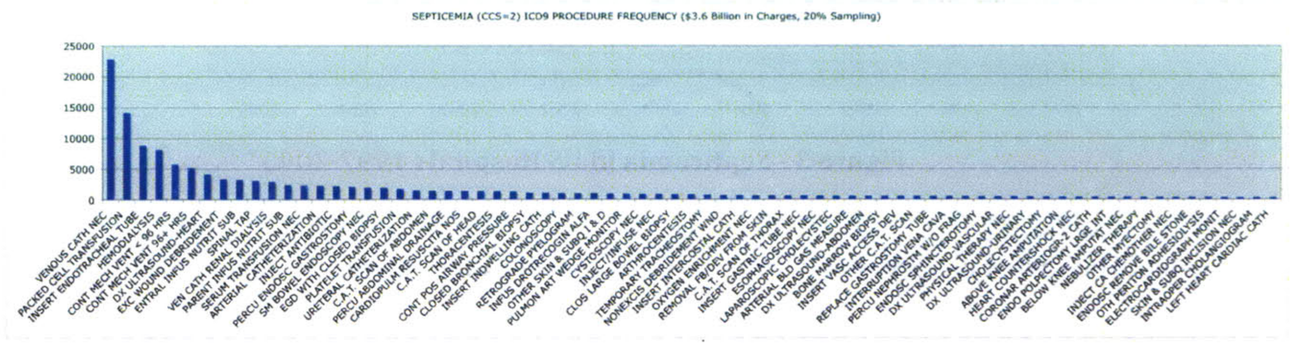


Figure 2 - Septicemia Associated ICDS Procedure Frequencies

Further mining the HCUP databases, a time series was constructed for the Septicemia category from 1997-2005 in terms of number of discharges, mean charges, and the "National Bill" extrapolating the 20% sampling of the NIS database to 100%. The trends were deeply concerning and pointed to some deeper and rapidly emerging system problem as shown in Figure 3.

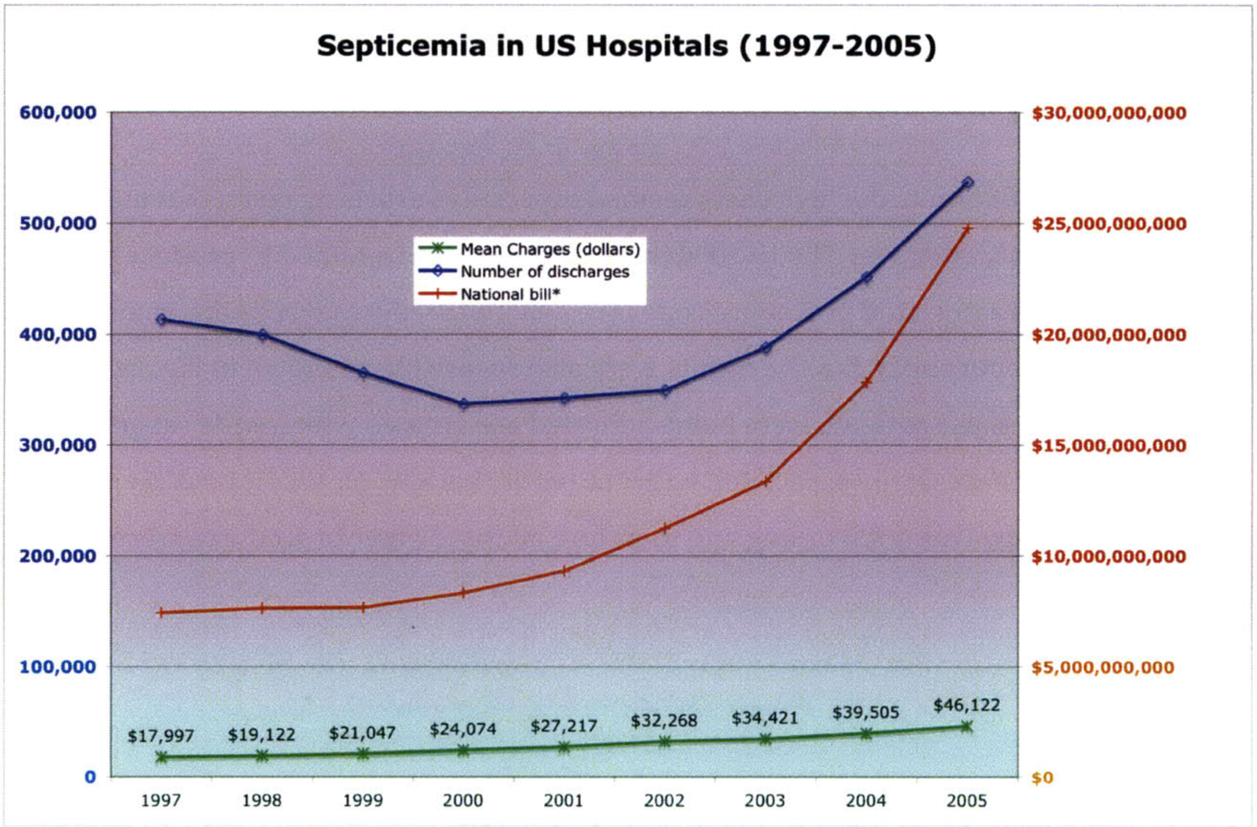


Figure 3 - Septicemia in US Hospitals 1997-2005

Beyond the financial costs, the human costs of this problem were also deeply troubling as not only was the cost of treatment rising, but the outcomes were not improving, rather the contrary, as mortality associated with Septicemia increased from 14.2% to 18.3% over the 1997-2005 time period and not necessarily because of older or sicker patients as the average age of these patients remained nearly the same as did the mean length of stay shown in Figure 4.

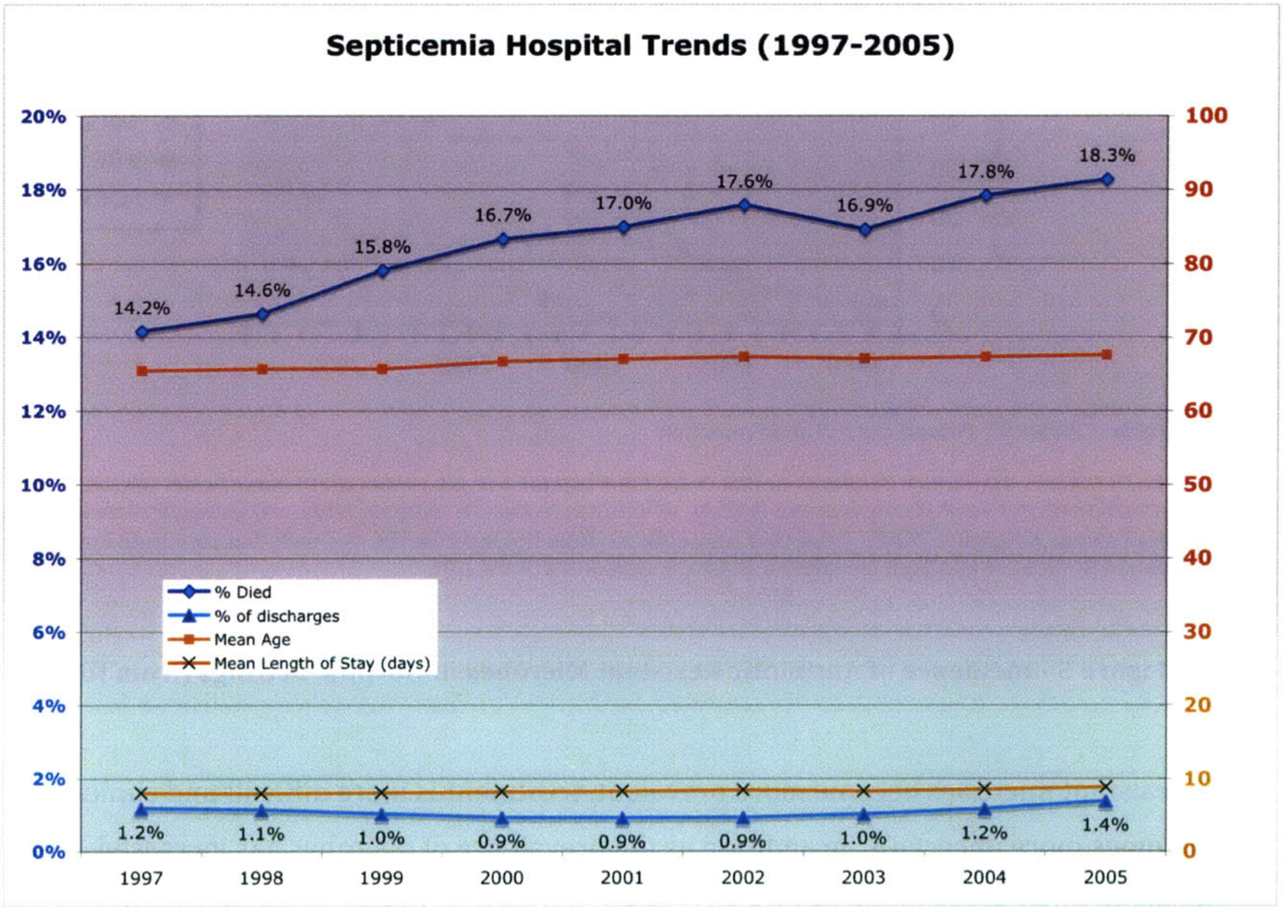
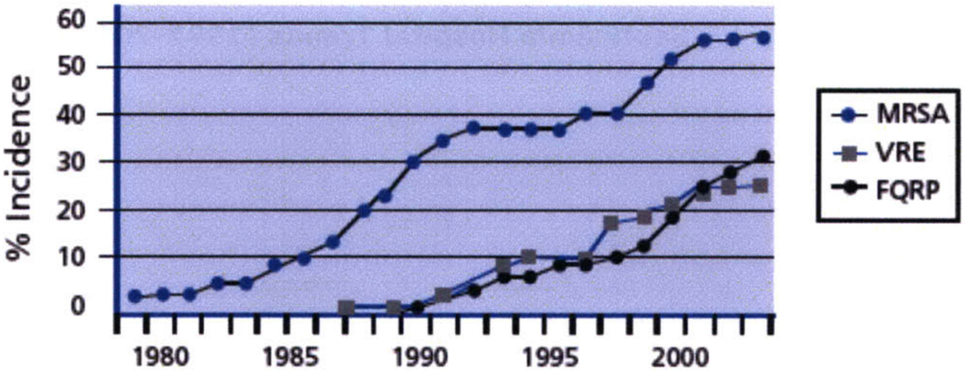


Figure 4 - Septicemia Hospital Trends (1997-2005)

The author's involvement with DNA sequencing provided opportunities to ask practicing Infectious Disease Physicians about the cause of these trends and the answer was that rapidly increasing rates of microbial resistance were rendering antibiotics ineffective and the spread of these resistant microbes provided further opportunity to generate resistance as "last line of defense" antibiotics such as Vancomycin previously used only sparingly had to be used in much larger numbers of cases. The rapid increase of resistant microbes can be seen in Figure 5 from the Centers for Disease Control (CDC).



Source: Centers for Disease Control and Prevention

This chart shows the increase in rates of resistance for three bacteria that are of concern to public health officials: methicillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-resistant enterococci (VRE), and fluoroquinolone-resistant *Pseudomonas aeruginosa* (FQRP). These data were collected from hospital intensive care units that participate in the National Nosocomial Infections Surveillance System, a component of the CDC.

Figure 5 - Incidence of Antibiotic Resistant Microbes in Hospital Settings (from IDSA)

As resistant microbes become more prevalent, treatment is more difficult and clinical outcomes more uncertain. In addition as the prevalence of microbes has increased, their impact in other areas of the healthcare system has also increased, so for example complications associated with routine surgery can also increase if resistant bacteria cause infection as part of the procedure. Of the resistant organisms, Methicillin Resistant Staphylococcus aureus (MRSA) is the most famous with constant headlines about its spread in both hospitals and in the community including schools and nursing homes. The impact of MRSA alone is enormous and a study by (Noskin, Rubin et al. 2007) using many of the same methods presented here including analysis of the NIS database estimated the economic burden of MRSA per year: $14.5 billion in 2003. The first comprehensive nationwide study was performed by the CDC as reported in (Klevens, Morrison et al. 2007) to estimate the US burden of MRSA infections in 2005 estimated 94,360 patients had an invasive MRSA infection (including Septicemia) of which 18,650 died from MRSA infections. These estimates are for invasive infections only, such as Septicemia, which are relative small compared to non-invasive infections, which could be an order of magnitude larger as reported by Klevens et al.

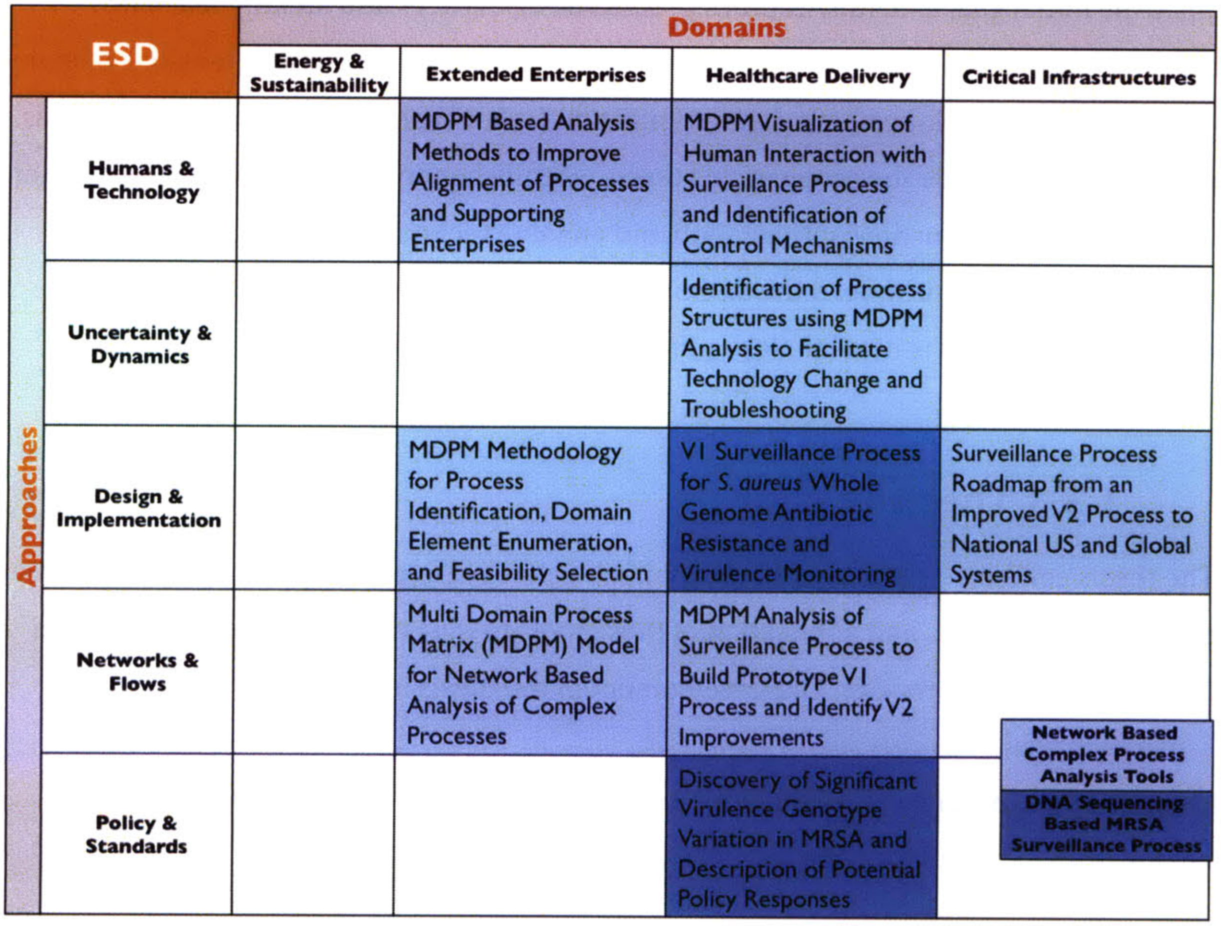
Of great interest in the Klevens et al. study was the attempt at molecular characterization of the MRSA isolates from the surveillance sites, which used pulse-field gel electrophoresis typing (PFGE) where the genome of a microbe is cut into a few (100-1000) fragments using restriction enzymes able to cut at certain unique sites to produce a "fingerprint" for the particular strain. This relatively crude method can distinguish gross differences between strains, but not any gene level mutations or changes. But even PFGE showed significant genome variation among the strains collected in the Klevens et al. CDC study despite all the strains being phenotypically identical (i.e. all were resistant to antibiotics). Other studies have shown that even small differences in gene sequence can lead to significant virulence (ability to infect and survive) beyond the necessary antibiotic resistance, such as (Diep and Otto 2008) who describe virulence mechanisms of Community Associated MRSA (CA-MRSA) and their encoding genes, many of which would not be visible using standard PFGE methods including variation in the Panton Valentine leukocidin (PVL) encoding genes which are thought to be key virulence determinants. DNA sequencing can provide the needed resolution to see these differences down to the nucleotide level, but it is a very new and complex technology and although costs are rapidly decreasing is still expensive and cumbersome for an MRSA surveillance application. The concept developed in this thesis is to adapt existing processes to create a prototype DNA based surveillance process that satisfies the various stakeholder constraints including cost, ease of use, integration into existing workflows, and most importantly the ability to track the virulence and other changes in the rapidly growing MRSA threat. This idea formed the basis for this thesis and established a real world process to use in developing and testing the process modeling approach. This idea was built upon the initial research done on learning curves in healthcare processes, which although it represented nearly a year of work is only described in this section to provide background for the current thesis and to describe the significant exploration needed to tackle "systems" problems in healthcare.

1.3.1 Approach

The research approach for this thesis was to study deeply an existing process in healthcare, specifically surveillance of Methicillin Resistant Staphylococcus aureus (MRSA), an increasingly common and dangerous microbe in hospital settings. The first step was to observe current diagnostic processes in a hospital clinical microbiology laboratory to provide a reference data set for the development of a process representation and analysis methodology. The next step was to use this methodology to help develop and demonstrate a new clinically important process in DNA sequencing based MRSA surveillance. This demonstration provided a data set describing the true process complexity when implementing a novel process in healthcare including the effect of other process elements beyond the physical steps such as additional layers of information and organization, which contain a large amount of previously unmapped interactions necessary in the initial stages of process development that contribute to high complexity. This higher complexity is not found in more mature processes as most of this "scaffolding" has been removed as a dominant process design emerges and alternatives are culled out, design spaces refined, and optimizations performed. One of the key issues encountered in complex processes is the difficulty of change due to what some would argue is a fundamental level of irreducible complexity. However, it is possible that the high complexity is not irreducible but is in fact associated with poor process design to manage this complexity. While it is clear that deep knowledge bases (specialties) are required to build modern processes and their existence implies a certain level of irreducible complexity, one of the objectives of this thesis is to show that if complexity is addressed as any other design variable (such as yield or efficiency) that great improvements can be made to reduce complexity (and perhaps even eliminate some previously "irreducible" complexity through novel design that minimize the use of certain specialties). The methodology developed here is thus an effort to develop a set of tools to manage this process complexity. The value of the demonstration is that it is used to refine the process mapping and analysis methodology, as well as to guide simulations on representative process designs to suggest enterprise and process alternatives to better manage process complexity.

1.3.2 Scope and Limitations

This thesis is interdisciplinary in nature, attempting to provide an integrative framework for process design spanning engineering, management, and social sciences domains to properly reflect the actual complexity of processes in industry using MRSA surveillance as a specific example. Figure 6 shows the contributions of this thesis mapped onto the Engineering Systems Research Space, including its impact on a critical healthcare infrastructure, extended enterprises, and network methods. Because of the dual focus of the thesis on network based complex process analysis tools and DNA sequencing based MRSA surveillance, each category is highlighted in Figure 6.



ESD		Domains			
		Energy & Sustainability	Extended Enterprises	Healthcare Delivery	Critical Infrastructures
Approaches	Humans & Technology		MDPM Based Analysis Methods to Improve Alignment of Processes and Supporting Enterprises	MDPM Visualization of Human Interaction with Surveillance Process and Identification of Control Mechanisms	
	Uncertainty & Dynamics			Identification of Process Structures using MDPM Analysis to Facilitate Technology Change and Troubleshooting	
	Design & Implementation		MDPM Methodology for Process Identification, Domain Element Enumeration, and Feasibility Selection	V1 Surveillance Process for *S. aureus* Whole Genome Antibiotic Resistance and Virulence Monitoring	Surveillance Process Roadmap from an Improved V2 Process to National US and Global Systems
	Networks & Flows		Multi Domain Process Matrix (MDPM) Model for Network Based Analysis of Complex Processes	MDPM Analysis of Surveillance Process to Build Prototype V1 Process and Identify V2 Improvements	
	Policy & Standards			Discovery of Significant Virulence Genotype Variation in MRSA and Description of Potential Policy Responses	

Network Based Complex Process Analysis Tools

DNA Sequencing Based MRSA Surveillance Process

Figure 6 - Thesis Contributions Mapped onto ESD Research Space

A key objective of this thesis was to provide a network based analytical framework for "extended enterprise" process design encompassing not just the process steps, but also the organizations, information, and design alternatives that exist in real processes. This framework is specifically applied to the MRSA surveillance process also developed in this thesis and the value of the method is demonstrated in this case. Although the framework can likely be generalized, and specific future directions are described in chapter 7, the scope of this thesis is limited to the initial development of both the process analysis framework and the MRSA surveillance process, which are both prototypes. This thesis builds on established process design, network analysis, microbiology, and DNA sequencing methods but extends each of these through interdisciplinary connections and also develops new methods in each area. The thesis emphasizes finding "feasibility" over "optimization" since the initial goal is to fully capture process design spaces and identify feasible alternatives. Optimization can be performed at later stages and using extensions of many of the methods developed here, but optimization (other than to make processes work as prototypes) is outside the scope of this thesis. Significant enhancement and improvement is possible through further optimization and ongoing technology and process improvement, and it is indeed the author's hope that this will happen, but much work remains. This thesis is nevertheless a first step and an initial roadmap to get there.

1.3.3 Thesis Outline

The thesis combines elements of both main objectives in every chapter: new tools relative to complex processes and application of these tools to the critical healthcare need of an MRSA surveillance process. The development of the thesis is as follows:

Chapter 1: A general description of the research, along with the important motivation behind MRSA surveillance and complex processes.

Chapter 2: Literature review and background on the range of disciplines and topics addressed in this thesis including on the surveillance side: MRSA biology, Clinical Microbiology, Antibiotics, Antibiotic Resistance, Virulence. And on the process tools side: process models, process improvement methods, network models, network analysis, and network applications in processes.

Chapter 3: Describes the Multi Domain Process Model (MDPM) developed in this thesis to analyze complex processes and describes each of its elements in the organizational, information, and process domains. Chapter 3 also provides a listing of potential metrics that might be used in the analysis of process network models.

Chapter 4: Shows the potential uses of the MDPM model including metrics sets useful in the analysis of each domain as well as use of the MDPM model to describe and manage process improvement methodologies such as Lean.

Chapter 5: Applies the MDPM model to the real-world MRSA surveillance process design and implementation performed in this thesis. This chapter also describes the associated methodology to select feasible processes, identify model elements, describe critical process sub-networks, and demonstrates the use of network analysis tools to understand the surveillance process. This Chapter also highlights potential process configurations to improve performance under rapid technology change conditions or complex troubleshooting based on the MDPM model.

Chapter 6: Presents the biological results from the V1 Surveillance Process, including a comparison of the MRSA samples to references and identification of differences between samples. Recommendations for improvements in a V2 process are also made in terms of the MDPM model.

Chapter 7: Describes a roadmap from the V1 and V2 processes described in this thesis to further versions including a national and global surveillance system. Future work related to the MDPM model is also presented including the concept of process options.

2 Background and Related Work

Processes form a part of every Engineering System, from supply chain fulfillment to gasoline production in refineries to the manufacturing of commercial jet aircraft and everything in between. Understanding these processes is a critical part of understanding these complex socio-technical systems. Healthcare processes in particular have emerged as an important research area because of their direct impact on human wellbeing but also societal resources. A number of tools have been brought to bear on the understanding of processes and the systems that support them, and network theory has emerged as a potentially very useful analysis methodology to understand and manage the complexity of these processes. This chapter presents some background information on existing process analysis methodologies, network analysis methods, and the healthcare system to support the analyses presented in subsequent chapters.

2.1 Processes

At its most basic level, a process is a series of steps or tasks required to produce a product or deliver a service. In manufacturing this might mean the series of tasks required to build an aircraft or produce refined gasoline. In the service economy this might mean the series of steps needed to fulfill a customer order or process an insurance claim. In healthcare it can be the series of steps to correctly diagnose a bacterial infection or perform an x-ray. The majority of services provided by the US healthcare system can be described as processes, where a series of steps delivers a diagnosis, treatment, monitoring, insurance collection, vaccination, and many others. These simple examples obscure the real complexity of healthcare processes, where multiple stakeholders, multiple technical disciplines, complex feedback loops, information rather than physical products, and infrastructure items from other processes are required. Some of this complexity arises from the multiple disciplines needed to fully utilize the highly scientific and technologically advanced procedures now offering real improvements in health care but some complexity arises because our organization of health care has evolved in a way that often ignores the challenge of controlling complexity.

Traditional process definitions from chemical engineering focus largely on efficient physical transformation of raw materials into higher value chemicals and materials and the optimization of physical process steps. These processes rely on highly detailed specifications for the equipment involved and the particular operating values of the variables involved at each step. Typical chemical engineering process representations such as flow sheets are diagrams showing the final selections for process steps, control systems, operating parameters, input materials, and product yield. While, highly useful for the operation of these processes these diagrams do not show the many design alternatives, stakeholder preferences, regulations, design choices, design space explorations, designer's tacit knowledge bases, optimizations, explicit physical knowledge bases, and other information sources needed to build them. These additional elements are critical to the development of processes but most process representations and models do not show them focusing instead on "final" processes where the "messy" design and implementation stages are complete. The greater complexity of processes in these early stages is a potentially very significant barrier to adoption that is not commonly modeled although some effort exists within the chemical engineering community as described in (Nagl, Westfechtel et al. 2003), (Heller, Jäger et al. 2004), and (Sargent 2005) who state the importance of "big P" process encompassing organizations, information, and physical steps.

The business process reengineering efforts championed by Hammer (Hammer and Stanton 1999) and others in the 1990s brought many of the engineering process concepts to the business realm and expanded them to the enterprise and system view. The working definition used by these efforts for processes included the following from (Davenport 1993) "Process Innovation: reengineering work through information technology":

> *"a structured, measured set of activities designed to produce a specific output for a particular customer or market. It implies a strong emphasis on how work is done within an organization, in contrast to a product focus's emphasis on what. A process is thus a specific ordering of work*

> *activities across time and space, with a beginning and an end, and clearly defined inputs and outputs: a structure for action. ... Taking a process approach implies adopting the customer's point of view. Processes are the structure by which an organization does what is necessary to produce value for its customers."*

And from Michael Hammer (Hammer and Champy 2003) "Reengineering the corporation, A manifesto for business revolution",

> *"a collection of activities that takes one or more kinds of input and creates an output that is of value to the customer."*

These definitions imply the need to clearly define: process steps, relationships between steps, objectives of the process, stakeholders, organizations in which the process exists, and the policy environment in which the process is executed. Implicit in these process definitions is the need for a model and representation such that a process can be precisely described and analyzed. In addition to the chemical engineering representations, a number of methods for business process representation have been developed including flow charts, petri nets, event driven process chains, simulation including system dynamics, role activity diagramming, and information systems based techniques including IDEF and data flow diagramming. However there is no accepted standard for process representation and analysis, and each of the existing methods has strengths but also significant shortcomings especially in the areas of process design, change, and optimization. The lack of an accepted and comprehensive process modeling methodology makes it difficult to manage process design and evolution in areas with significant technology change, which is the case for many healthcare processes. This thesis focuses on the dynamic nature of processes where steps and their order may change due to technology advances, enterprise architecture optimization, regulatory environment, and most importantly in response to changing needs. Improving the dynamics of process evolution is critical to rapid adaptation and faster implementation times for improvements, but is often not explicitly the goal of process analysis.

As described in Chapter 1, this thesis also expands on the traditional view of processes as the transformation of materials into useful products or the series of steps associated with delivering a service in the business world. Big "P" processes are defined to encompass the physical steps of a process in traditional definitions but also the entire set of process related activities, environment (physical and regulatory), and infrastructure needed to support the process. In complex processes this set of process related elements is especially important because these elements may interact with the process in unintended or emergent ways not initially predicted by process designers. Similarly changes to one of these elements can propagate through the system in unforeseen ways impacting elements throughout the process that may not have previously been considered. The following sections describe current practice in complex process analysis as well as describing improvement methodologies critical to the ongoing operation (and change) of complex processes.

2.1.1 Definitions

The definition of process depends on the particular field it is applied to. The following is a brief listing of process definitions from various fields:

Process Type	Definition	Typical Model
Chemical Engineering Process	A set of operations that transform raw materials into desired products	Process Flow Diagram
Business Process	A set of activities or tasks that produce a specific service or product for a customer	Flowchart Process Map
Systems Engineering Process	A method for applying structured design techniques to system development	Project Plan, Process-Data Model
Thermodynamic Process	Transition of a thermodynamic system from an initial state to an allowable final state	Thermodynamic space diagram (P-V, h-s)
Mathematics: Stochastic Process	A variable whose time development is analyzable in terms of probability	Markov chain, Monte Carlo Simulation

Process Type	Definition	Typical Model
Computer Science Process	An instance of a computer program of one or more threads being sequentially executed	Process state diagram

Table 1 - Process Definitions

This thesis focuses primarily on the first three definitions of Table 1, as the complex process studied necessitates that we combine elements of each. The main emphasis of this thesis is in providing a process design methodology for clinical microbiology processes, and because of their similarity to other diagnostic processes in the healthcare system. The physical steps associated with these diagnostic (and surveillance) processes share much in common with chemical engineering process design as they can be thought of as a series of physical unit operations to produce data as a product such as a diagnosis or surveillance. Business processes take more consideration of the Enterprise and the tasks associated with delivering services beyond the physical steps in generating the data, since there are still a number of "non-engineering" steps between the delivery of the raw diagnostic or surveillance data and the processing of insurance forms, updating of medical records, patient consultations, etc. This set of "business" processes at the start and end of the "physical chemical engineering" are necessary to integrate the diagnostic process into the Enterprise, but can add significant complexity. Finally, there is the "Systems Design Process" where a new process (or system) is created according to a set of design rules that ensures the overall system is considered rather than individual elements. The design of the MRSA surveillance system and the design methodology developed can thus be thought of as a "system design process" which has as an outcome a "business process" that contains a "physical, chemical engineering - like process".

2.1.2 Processes in Engineering Systems

Contained within most engineering systems are processes and many engineering systems are built for the processes they support. For example, critical infrastructures such as airports and air traffic control support the process of air transportation from one location to another. And much effort has gone into streamlining that process including the "physical" process elements (or series of tasks) such as improved routing, navigational aids to simplify pilot workload, and more efficient aircraft able to directly connect almost any two airports on the globe. But the "business" processes have also been improved such as airline operations and airport demand management. Many of these improvements have come from "systems thinking" such as the application of system design methodologies or processes.

Many extended enterprises are also created to support processes such as the design and manufacturing of Boeing's new 787 which has a globally distributed supply-chain but more importantly also globally distributed its "design chain". The manufacturing process of every component of the 787 had to be integrated with the overall assembly process and Boeing suffered significant difficulties due to the lack of integration of these distributed processes with the final assembly process. Energy production and distribution are engineering systems that contain many chemical engineering processes such as the extraction of crude oil and refining into gasoline. However, there are many layers of "business processes" to manage this physical process such as demand forecasting, capital expense planning, and supply-chain management.

The processes within these engineering systems are different from products or enterprises as they are performed within enterprises and are dynamic. One of the goals of the Engineering Systems Division is to develop a body of knowledge around healthcare delivery systems issues, which is mainly the study of processes. Yet, many of the tools to study processes do not capture their systems nature, focusing strictly on the steps rather than the complex socio-technical enterprises they are performed in. The interaction between process and enterprise is a fundamental research area for Engineering Systems, and is especially important for the study of healthcare

2.1.3 Process Design and Improvement

Given an existing process an enterprise's ability to improve is critical to its long-term success in modern competitive markets. A number of methods for manufacturing process improvement have been described in the literature such as lean production methods derived from the Toyota Production System described by (Womack, Jones et al. 1990) in "The Machine that Changed the World". Womack et al. found that while lean production had certain structures to be followed these were mainly to enable learning and at the core of this methodology is a learning organization and one where the people on the frontlines have much greater accountability (and knowledge) for how to improve a process. Quoting from Womack et al.: "The truly lean plant has two key organizational features: it transfers the maximum number of tasks and responsibilities to those workers actually adding value to the car on the line, and it has in place a system for detecting defects that quickly traces every problem, once discovered, to its ultimate cause." This implies that vital process knowledge exists closest to the process activities but also that the organization must strive to ensure that the workers on the "frontline" have access to as much of the enterprise's knowledge network as possible.

So for example, a separate R&D team that hands off new designs to frontline production workers but is not continuously engaged with production is not lean as any problems that may occur due to the designs (and found in production by front-line workers) must be quickly resolved and this requires access to any specialized R&D knowledge or the R&D team itself. In addition the rapid response to problems and the deployment of resources to resolve them also requires that the lean enterprise have a highly efficient and highly connected network (both formal and informal) to support its processes. Process changes are also an important element of lean production and indeed the original genesis for the Toyota production system came from the need to rapidly switch production between various vehicle models at a single production plant. However, in complex processes change propagation can be a very difficult problem as the effects of a small and seemingly inconsequential change can have significant impacts on the rest of the process. This is in some sense a corollary to the need to quickly trace problems and find root causes in a “troubleshooting mode” but in a design change mode requires the same highly connected (networked) infrastructure needed for troubleshooting.

Lean production has proven to be an enormous competitive advantage to enterprises that have adopted this methodology as an "infrastructure" that enables them to keep improving and even adapt the methodology to their changing needs. That is to say simply adopting "just-in-time" or "kanban” ideas is not enough nor is it the basis of lean production but rather it is to recognize that there is always room for improvement, that improvement takes effort and is the responsibility of the entire organization, and that every activity must create value throughout the process. These ideas have been continuously refined and adapted to other industries than car manufacturing such as aerospace, healthcare and semiconductor manufacturing. These other industries are also highly complex and yet there are certain companies and organizations within them that consistently outperform others despite the enormous complexity of their industries and products. A recent book “Chasing the Rabbit” by (Spear 2008) describes the characteristics of these higher performing enterprises, and finds that improvements on some of the core ideas of lean production explain many of these performance differences. Spear describes 4 main capabilities for these high performing enterprises:

- **Capability 1**: Specifying Design to Capture Existing Knowledge and Building in Tests to Reveal Problems
- **Capability 2**: Swarming and Solving Problems to Build New Knowledge
- **Capability 3**: Sharing New Knowledge Throughout the Organization
- **Capability 4**: Leading by Developing Capabilities 1, 2, 3

As with Lean Production all of these capabilities require a significant “process infrastructure” in the form of highly connected organizations that can rapidly deploy resources where needed and learn from problems. This is especially important when dealing with complex processes that require multi-disciplinary efforts such as DNA sequencing, which can be depicted as a series of "silos" using Spear's ideas shown in Figure 7.

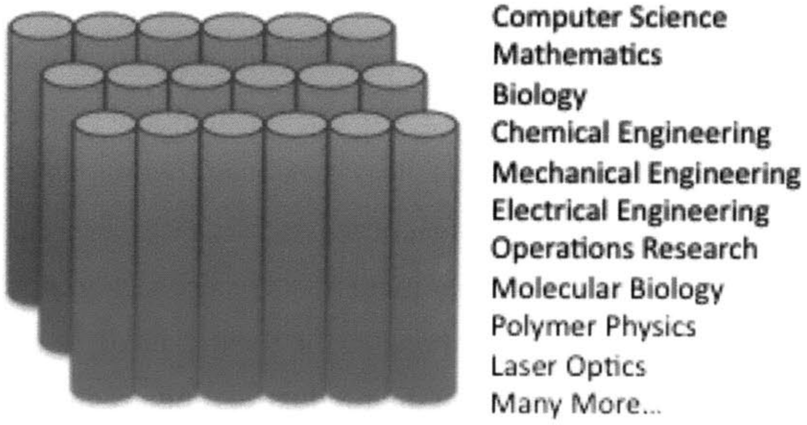


Figure 7 - Disciplines in DNA Sequencing

High throughput DNA sequencing originally developed to complete the human genome project requires a multitude of disciplines for the complex process to function. Just like many modern manufacturing processes, each of these "silos", as described by Spear, have grown deeper and deeper as each area specializes. The integration of all these silos requires connections ideally composed of very efficient networks where both the individuals and the information produced can be accessible. In order to facilitate this, integrated project teams are routinely used at one large-scale sequencing center, the Broad Institute, for new process development in order to facilitate the efficient networks needed. The integrated project teams are very similar to ones described in "The Machine that Changed the World" and serve a similar purpose: connecting the process, organization, and information domains such that communication can be efficient.

2.1.4 Process Models

Just as with process definitions, there are a large variety of process models corresponding to the different process types such as chemical, business, or design. This section presents an overview of the relevant process models considered in this thesis mainly aimed at physical, business, and design. The goal of all these process models is ultimately to produce an analytically useful representation of reality. No model however can fully represent every dimension of a complex process and the models described here are useful in their particular applications but no single model will provide all the answers a process designer or operator might need. The network-based process model developed in this thesis and described in Chapter 3 falls into the same category and while it provides a useful representation of the many information, organizational, and physical elements involved in complex processes, this model does not eliminate the need for other process models especially physical ones such as process flow diagrams. Some humility is necessary, as (Sterman 2002) writes in an introduction to his lecture accepting the Jay W. Forrester Prize:

> *"Most important, and most difficult to learn, systems thinking requires understanding that all models are wrong and humility about the limitations of our knowledge. Such humility is essential in creating an environment in which we can learn about the complex systems in which we are embedded and work effectively to create the world we truly desire."*

2.1.4.1 Flowcharts

Modern flowcharts can be traced back to the work of Frederick Taylor, Frank Gilbreth, and others who by 1911 held the first formal conference on "Scientific Management" as described in (Graham 2004). Gilbreth is mostly credited with the idea of a flowchart as part of the set of work analysis (and improvement) tools he and his wife, Lillian, developed in search of the "one best way". Gilbreth presented "Process Charts – First Steps in Finding the One Best Way" (Gilbreth and Gilbreth 1921), which described most of the elements of a modern flowchart at the 1921 ASME Annual meeting using the symbol set in Figure 8.

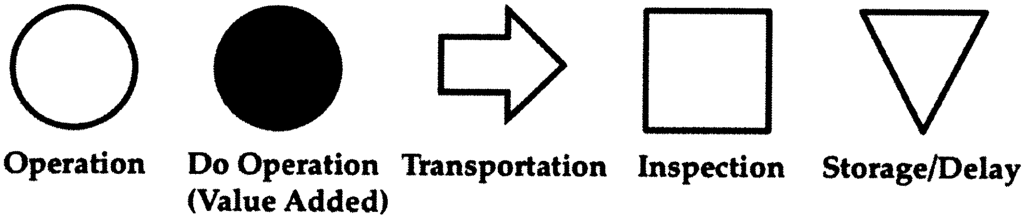


Figure 8 - Gilbreth's Basic Work Elements

The fundamental goal of flowcharts is to represent process activities, their precedence relationships, and properties such as duration with a standardized set of symbols such that the overall process can be analyzed, such as the "Do Operation" which emphasizes "value added" activities. This remains a useful tool in Industrial Engineering as it can quickly identify bottlenecks or non-value added activities and provides a process designer with an overall view of the relationships between process activities and their contribution to global process properties. But the emphasis on activities and the need to simplify the complexity of the process so it can be represented in a comprehensible diagram purposely omits other "views" as described by (Browning and Ramasesh 2007) such as the relationship networks between individuals, the information networks the work instructions for each activity were derived from, and the environment (policy or organizational) that the process is performed in. Indeed Gilbreth's symbol set contained no elements for information or people. Browning and Ramasesh's survey of process development models (including flowcharts and scheduling Gantt charts) finds that:

> "...models place a disproportionate emphasis on actions rather than interactions..." and "However, activity interactions, particularly the flow of deliverables (including information), give rise to a large portion of a PD process's behavior, such as iteration."

Browning and Ramesh also note that most process development models assume that all activities are known a priori, which is often not the case even for established processes (i.e. operating processes not in a development phase) where changes (and improvement) may lead to unforeseen interactions and uncertainty. In these cases it is likely that the elements flowcharts *do not* capture such as the organizational networks, information networks, and environment of the process facilitate the resolution of the problems by providing alternative paths (i.e. not explicitly defined links) between process elements such that they can be resolved (as in troubleshooting).

Manufacturing process flowchart ideas were later adapted to other fields including information processing, notably in the work of Herman Goldstine and John von Neumann (Goldstine and Von Neumann 1947) who used them to describe early computer programs on a successor to the ENIAC computer, the EDVAC. These computer science flow charts developed a different symbol set for each of the computer program "activities" including circles to represent start and end points, arrows to show "flow control" program elements, parallelograms for input/output, diamonds for conditional branches or decisions, and rectangles for processing steps. Many of these computer science focused flow chart symbols and enhancements formed the basis for modern flow chart techniques such as Unified Modeling Language (UML) which expands the modeling scope (and symbol set) to include not just computer programs but to model entire software intensive systems including "business processes" that are reliant on software. From a 2005 presentation on the UML standard by Satish Mishra, Humboldt University, the UML scope includes:

- Data Modeling concepts (Entity Relationship Diagrams)
- Business Modeling (work flow)
- Object Modeling
- Component Modeling

A criticism of these modern flow chart techniques is their significant complexity in attempting to model many different activities, processes, and elements, along with the correspondingly steep learning curve practitioners must go through to learn these methodologies. For example, the latest version (2.2) of the UML standard is composed of 14 different "diagrams" that represent "structural information", "behavior", and "aspects of interactions" as shown in Figure 9 to produce a master diagram.

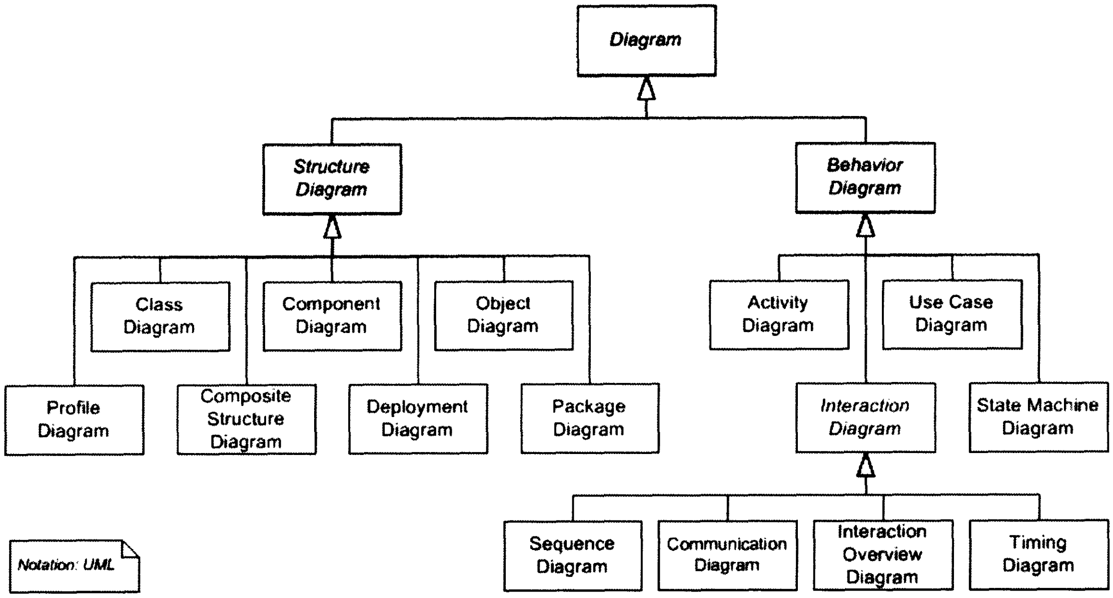


Figure 9 - UML Class Diagram

The closest of these diagrams to the original flow chart concept is perhaps the Activity diagram, which describes how objects in the UML diagram flow through a single process. Other diagrams may represent other interactions between the elements in the activity diagram such as information links, classification, etc., and the entire set of diagrams is intended to represent the system. For example Figure 10 shows an activity diagram for a customer interacting with an ATM machine.

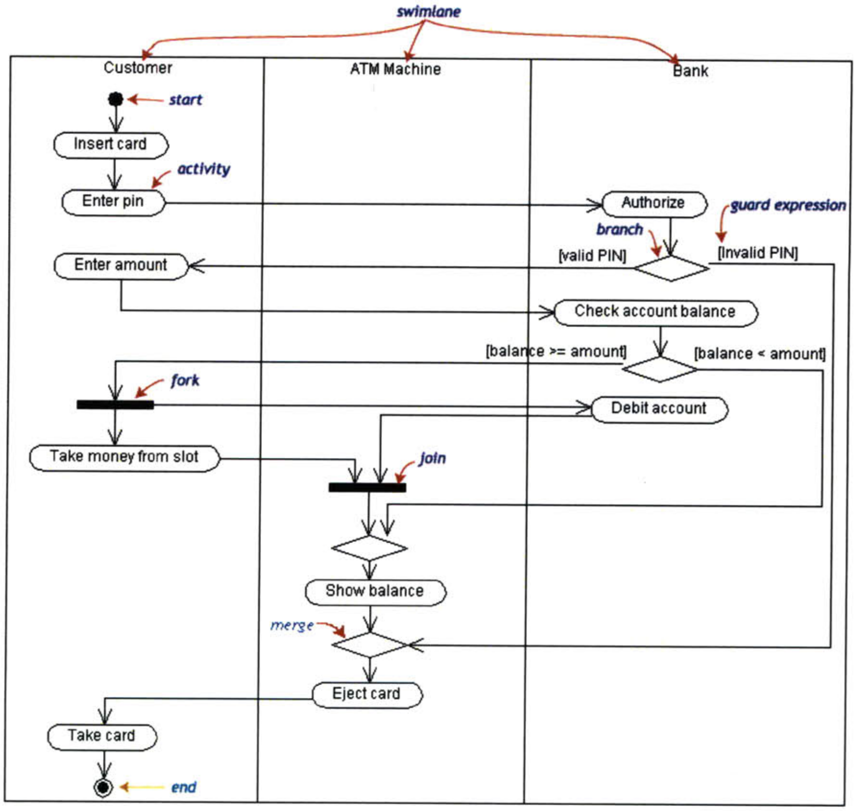


Figure 10 - UML Activity Diagram Example

A number of other business process and information systems modeling techniques exist, and a useful taxonomy was proposed by (Giaglis 2001) for the design of information systems including business process models. Giaglis classified models according to four dimensions:

- The **functional** perspective represents what process elements (activities) are being performed.
- The **behavioral** perspective represents when activities are performed (for example, sequencing) as well as aspects of how they are performed through feedback loops, iteration, decision-making conditions, entry and exit criteria, and so on.

- The **organizational** perspective represents where and by whom activities are performed, the physical communication mechanisms used to transfer entities, and the physical media and locations used to store entities.
- The **informational** perspective represents the informational entities (data) produced or manipulated by a process and their interrelationships.

The various modeling techniques were then described according to their "Depth" in each of these dimensions as shown in Table 2. It is apparent from the table that no single technique covers all the dimensions proposed by Giaglis, partly because a single all-encompassing modeling methodology would generate complex difficult to maintain and interpret models. Giaglis proposes using each type of model as appropriate according to the stage of process development from initial scoping to execution. Of particular interest in Table 2 are the functional (process activities themselves), the organizational (who does the activities), and the informational (the data needed to perform or produced by the activities), which this thesis focuses on. Highlighted in blue in Table 2 are only the models surveyed by Giaglis that cover each of the functional, organizational, and informational dimensions. Of the models surveyed by Giaglis only the highly complex UML model, data flow diagrams, system dynamics, discrete event simulation, and IDEF3 cover all 3 dimensions of functional, organizational, and informational. But they do so covering at least one of the dimensions in a limited fashion. Yet, the connections between these 3 dimensions are exceedingly important as they define the true "static" structure of the process in terms of all the elements needed to perform it. Associations between any 2 dimensions are not enough as these may be connected through the third in ways that will be invisible in any "2-dimensional" analysis.

	Modeling Perspective (Depth)			
BPM/ISM Techniques	Functional	Behavioral	Organizational	Informational
Flowchart	Yes	No	No	Limited
IDEF0	Yes	No	Limited	No
IDEF3	Limited	Limited	No	Limited
Petri nets	Yes	Yes	No	No
Discrete event simulation	Yes	Yes	Yes	Limited
System dynamics	Limited	Yes	Yes	Limited
Knowledge-based techniques	No	Yes	No	No
Role activity diagram	No	Limited	Yes	No
Data flow diagram	Yes	No	Limited	Yes
Entity-relationship diagram	No	No	No	Yes
State-transition diagram	No	Limited	No	Limited
IDEF1x	No	No	No	Yes
UML	Yes	Limited	Limited	Yes

Table 2 - Dimensions of Modeling Techniques (Adapted from Giaglis 2001)

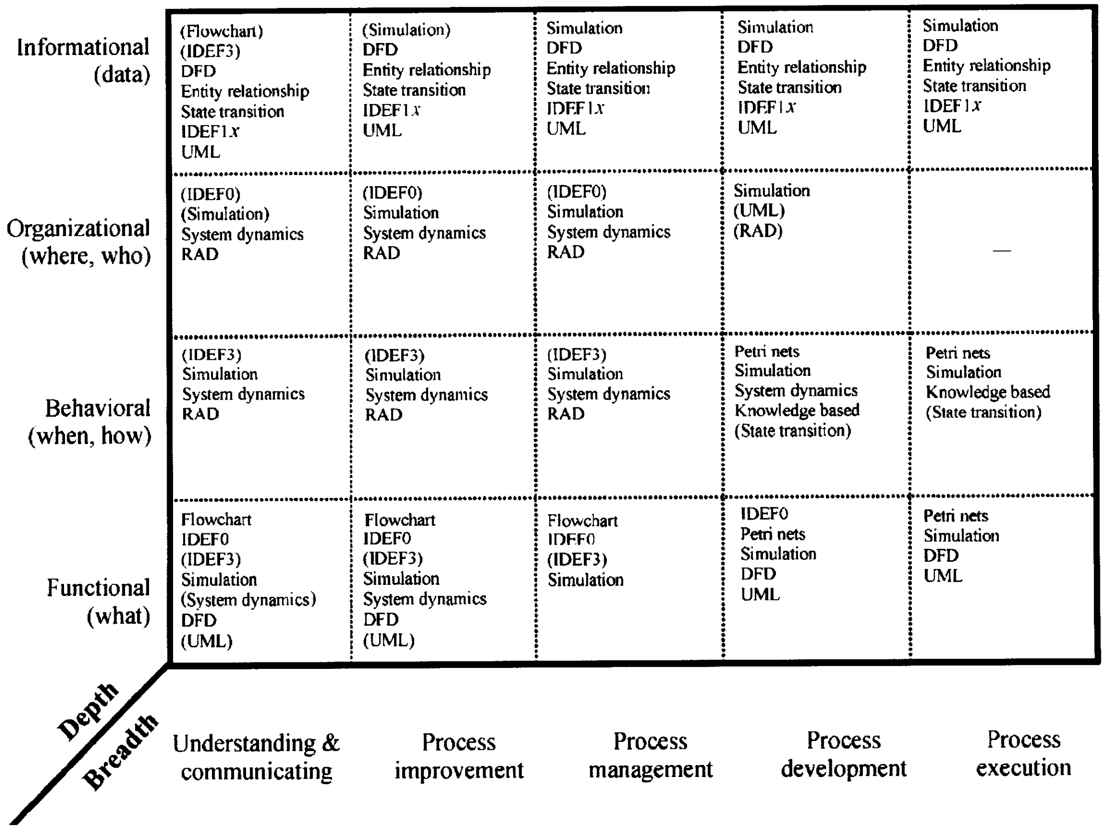
A description of the process models surveyed by Giaglis is given in Table 3 and many of these are derivations of flow charts with particular emphasis on data or other process elements.

BPM/ISM Techniques	Description
Flowchart	Graphical model to show the flow of work through a process and highlight key processing and decision points emphasizing the functional dimension
IDEF0	Functional modeling technique to describe decisions, actions, and activities of an organization with a single construct (ICOM) emphasizing the functional dimension
IDEF3	Complementary technique to IDEF0 to describe order and

BPM/ISM Techniques	Description
	sequences of events as flow models with a time dimension emphasizing the informational and behavioral dimensions
Petri nets	System modeling technique adapted to process consisting of a bipartite graph of concurrent states and transitions emphasizing the functional and behavioral dimensions
Discrete event simulation	A computer based model of a real system described as a chronological series of events that change the state of the system emphasizing functional and behavioral dimensions
System dynamics	System modeling technique to describe the causal relationships between system elements based on feedback and other control systems concepts emphasizing the behavioral dimension
Knowledge-based techniques	A modeling technique based on artificial intelligence methods where knowledge bases are built to represent system operation emphasizes behavioral dimension
Role activity diagram	Modeling notation describing roles as primary unit of analysis within organizations and showing the interactions between roles, emphasizes organizational dimension
Data flow diagram	Technique for depicting the flow of data among external entities, internal processing steps, and data storage elements in an enterprise emphasizes functional and informational dimensions
Entity-relationship diagram	Network based model describing the stored data layout of a system and the interrelationships in a manner independent of any processing, emphasizing informational dimension
State-transition diagram	Modeling technique derived from the analysis or real-time systems showing focusing on the time-related sequence of events, emphasizes behavioral and informational dimensions.
IDEF1x	Extension of the IDEF modeling family to analyze data structures in and out of computer systems to show relationships between entity classes, emphasizes informational dimension
UML	Multi-notation modeling technique providing multiple diagrams to address different system elements but which can be integrated into a single diagram. Covers all dimensions but complex.

Table 3 - Description of Information Systems Models (Adapted from Giaglis 2001)

When the "depth" along each of these dimensions is matched to particular applications (such as process improvement), which Giaglis calls "Breadth" the lack of a single modeling tool for all possible applications is apparent as shown in Figure 11 from (Giaglis 2001) where no single tool is listed in all (or even a majority of categories). The multi-purpose modeling tools such as UML or the Integration DEFinition (IDEF) family do cover several elements in that matrix, but at the expense of complexity. Based upon analysis of existing diagnostic processes, a major goal of this thesis was to develop a network based model that could tie together the informational, organizational, and functional dimensions, across each of process improvement, process management, and process development application dimensions. The Unified Modeling Language (UML) described in Chapter 3 is intended to fill these areas in the Giaglis taxonomy. Note that the IDEF family of models cover a large part of the same space the MDPM model, and indeed there are significant common elements as will be described in Chapter 3.



Depth / Breadth	Understanding & communicating	Process improvement	Process management	Process development	Process execution
Informational (data)	(Flowchart) (IDEF3) DFD Entity relationship State transition IDEF1*x* UML	(Simulation) DFD Entity relationship State transition IDEF1*x* UML	Simulation DFD Entity relationship State transition IDEF1*x* UML	Simulation DFD Entity relationship State transition IDEF1*x* UML	Simulation DFD Entity relationship State transition IDEF1*x* UML
Organizational (where, who)	(IDEF0) (Simulation) System dynamics RAD	(IDEF0) Simulation System dynamics RAD	(IDEF0) Simulation System dynamics RAD	Simulation (UML) (RAD)	—
Behavioral (when, how)	(IDEF3) Simulation System dynamics RAD	(IDEF3) Simulation System dynamics RAD	(IDEF3) Simulation System dynamics RAD	Petri nets Simulation System dynamics Knowledge based (State transition)	Petri nets Simulation Knowledge based (State transition)
Functional (what)	Flowchart IDEF0 (IDEF3) Simulation (System dynamics) DFD (UML)	Flowchart IDEF0 (IDEF3) Simulation System dynamics DFD (UML)	Flowchart IDEF0 (IDEF3) Simulation	IDEF0 Petri nets Simulation DFD UML	Petri nets Simulation DFD UML

Figure 11 - Taxonomy of Modeling Techniques (from Giaglis 2001)

Other techniques are also available, and although the MDPM model developed here has the much in common with IDEF, a brief description of these techniques is given in the following sections including ones not listed in the Giaglis taxonomy (which emphasized information systems) such as chemical engineering flow diagrams.

2.1.4.2 Process Flow Diagrams – Chemical Engineering

Unlike many of the Business Process Models discussed so far in this section, chemical engineering processes are bound by physical laws, which are integral to the process flow diagrams commonly used. These types of physically constrained models are especially relevant to this thesis, since unlike business process models or even models of product development tasks, the activities in chemical engineering flow diagrams must follow the physical reactions and processing steps and are not amenable to easy re-organization to improve task connectivity or sequence. The MRSA surveillance processes developed in this thesis is a physical process, and can be thought of as a series of bio-chemical reactions monitored by sensors which extract the DNA sequence information. The particular sequence of steps can be re-organized to a limited extent, but many steps are "fixed" in the sense they must be performed in particular order, and certain steps must be performed without exception (such as the lysis of MRSA cells). This is not to say that new technology cannot completely change the steps performed and perhaps even their sequence, but that within a technology process "sequence" there are certain physical constraints. An example of this might be from (Utterback 1996) which describes process innovation in plate glass manufacturing where a series of technological advances made the process successively simpler and more efficient as shown in Figure 12 with the series of processes (labeled as "steps") that were combined or eliminated through innovation.

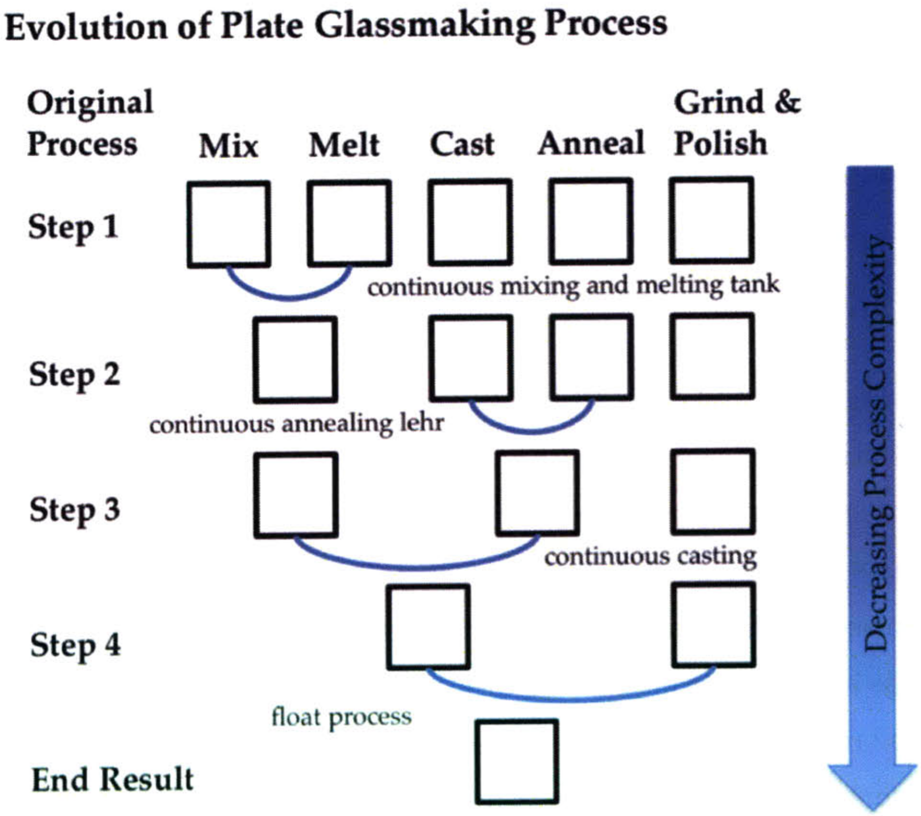

Evolution of Plate Glassmaking Process
Original Process
Mix
Melt
Cast
Anneal
Grind & Polish
Step 1
continuous mixing and melting tank
Step 2
continuous annealing lehr
Step 3
continuous casting
Step 4
float process
End Result
Decreasing Process Complexity

Figure 12 - Evolution of Plate Glassmaking Process (Adapted from Utterback 1996)

Pilkington, the glass manufacturer, made a massive and uncertain investment in developing the "Float" process which greatly simplified the number of steps involved although the development process was very complex in finding the right combination of variables and equipment to produce glass in this manner. This massive R&D effort required many additional resources beyond those needed to operate the eventual stable process in terms of personnel, information, and experimentation. And while the eventual process was simpler, there was a temporary increase in the "complexity" of the process, something akin to the "scaffolding" used during the construction of ships or buildings. However, at any given moment in between these major process innovations, the physical steps are "fixed" to a large extent and are not as easily changed as in the "business process" examples where the processing of a loan can be readily streamlined by moving personnel or arranging for a single representative to handle the overall processing of the loan as described in (Hammer and Champy 2003). Also because the steps in complex physical processes tend to be fixed, there is also a need for "fixed" organizations and procedures around each one that may require specialized knowledge and as a consequence can easily become "silos" resistant to change and not well integrated with other parts of the process. Ensuring that these silos are well connected and are able to understand the impact of their activities and changes on the rest of the process is a critical feature of "High-Velocity" organizations as described in (Spear 2008).

A significant benefit of the physical process centric view of process flow diagrams is the ability to perform significant optimization as the entire process can be modeled as a simultaneous set of equations subject to physical constraints and operational settings. A typical example taken from (Howell, Hanson et al. 2002) is given in Figure 13, a process flow diagram used in the optimization of a monoethylene glycol (MEG) regeneration system for use in a large offshore oil platform off Norway. MEG mitigates the formation of hydrates in oil recovery and is injected directly into the wellheads to be recovered on the platform, regenerated, and re-injected as the flow diagram shows. Extensive simulation and optimization of this process was performed prior to installation on the platform given the difficulty of post-installation redesign and the high capital cost associated with the facility. As reported by Howell et al, the process performed per the simulations once in operation, a tribute to nearly a century of advancement in chemical engineering modeling and our understanding of the physical laws that govern the chemistry and physics in each of the unit operations symbols shown in the MEG regeneration flow diagram.

But what of the personnel and the work instructions needed to operate the MEG regeneration process? Flow diagrams are not typically concerned with these other domains of a process other than to specify alarms in control systems, user inputs, and maintenance access valves and controls. Design exercises are always performed in offshore platform design to anticipate failure modes and many of these are often operational errors that the process design should anticipate and be robust to. However, the human operators and work instructions (i.e. operational rules) are often not explicitly described as part of a process flow diagram yet can have a very significant impact on the performance of a process. This is also true in a positive sense as even in physically difficult to change processes such as an offshore platform it is often possible to find ways to improve yields, quality, and reduce cost through control system adjustments, operational improvements, careful scheduling, and other means. These improvements happen through mechanisms not clearly specified in a process flow diagram, often through social networks, experimental observation, and planned experiments. A central part of the MDPM model in this thesis is to shine light on these "hidden networks" so critical to process change and improvement.

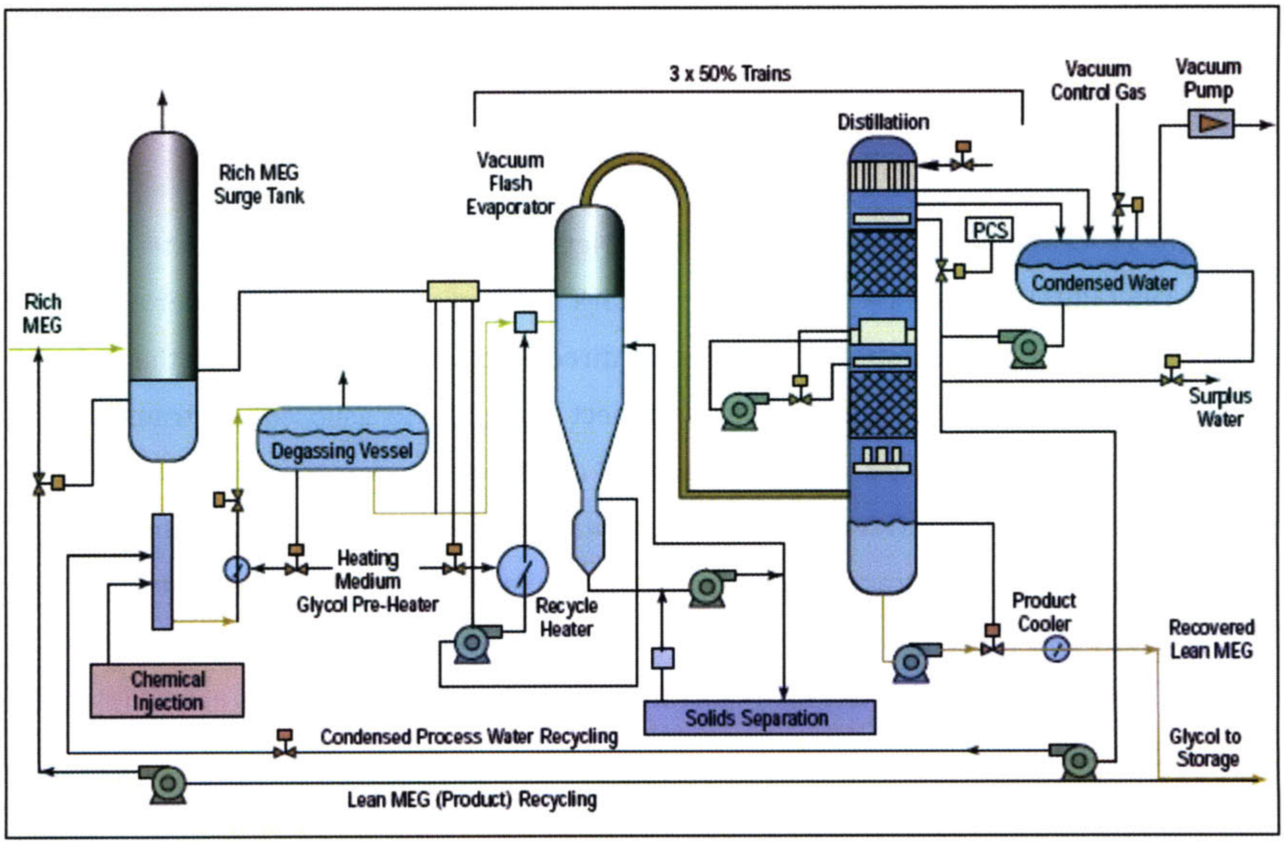


Figure 13 - Flow Diagram for the Asgard B Platform MEG Regeneration Process

Much like the plate glass making process evolution, but at an extraordinarily faster pace, measured in months, many of the DNA sequencing processes considered in this thesis for the MRSA surveillance system are improving in performance. However, because of the rapid pace of evolution much of the R&D "scaffolding" necessary to make major changes in process has remained in place making these processes enormously complex. And like the MEG regeneration process each evolutionary DNA sequencing process has a given physical configuration set by the biochemical steps needed that cannot immediately be reconfigured for greater simplicity. Successive generations and refinements can reduce complexity however, as in the plate glass example, but this requires ongoing effort, metrics to measure the complexity, and models to evaluate potential improvements. The proof of concept surveillance system developed in this thesis is the first step in an evolutionary chain that will eventually result in a much simpler, more efficient, and elegant process, just as the original plate glass process was the first step in that chain. The MDPM model developed in this thesis is a means to manage and accelerate this evolution using rational engineering tools.

2.1.4.3 Petri Nets

Of the process models described so far, Petri Nets developed by Carl Adam Petri as part of his 1962 PhD thesis are the closest to a network based process model (Van der Aalst 1998), and indeed are mathematically described as directed bipartite graphs where the nodes represent transitions or places and arcs connect each place to a subsequent transition. The elements of a Petri Net are shown in Figure 14 and the network is bipartite because it includes only places or transitions as nodes with arcs connecting places to transitions (but not places to places or transitions to transitions). Places are "passive" elements representing buffers, queues, physical locations (i.e. a warehouse), or communications mediums (a computer network). Transitions are active elements such as events (starting an operation, a traffic light change), transformations (a chemical reaction, updating a database), or transporting an object. Arcs are causal connections between places and events, which follow certain rules as shown in Figure 15. Finally, tokens represent the elements subject to change such as physical objects (a product or drug), information objects (a message), collections of objects (a warehouse with parts), an indicator of a state (process state), or an indicator of condition (order fulfillment). The simplicity of Petri Net elements and the directed bipartite structure allows for mathematical formalism and analysis of workflows constructed using them.

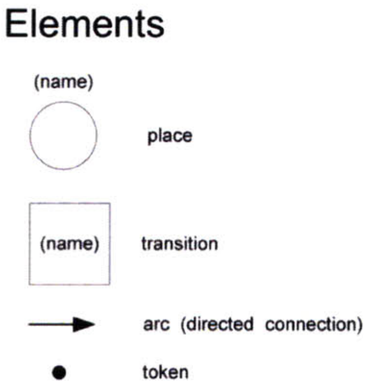
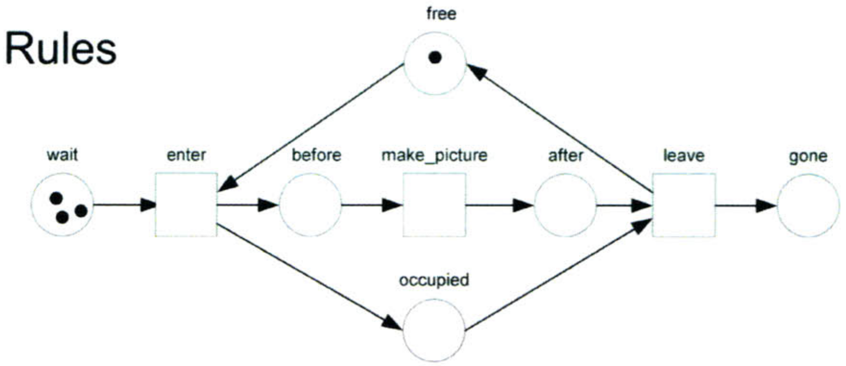


Figure 14 - Petri Net Elements (Adapted from van der Aalst)



- Connections are directed.
- No connections between two places or two transitions.
- Places may hold zero or more tokens.
- A transition is **enabled** if each of its input places contains at least one token.

Figure 15 - Petri Net Rules (Adapted from van der Aalst)

Business process modeling in particular has seen significant adoption of enhanced Petri Nets with some additional elements such as "color", "time", and "hierarchy" to represent more complex business processes that classic Petri Nets as described here cannot represent (Giaglis 2001). The mathematical framework of Petri Nets allows for the construction of many computer science constructs such as AND, OR, iteration, and nonsynchronous processing using tokens. And because Petri Nets are networks, a number of network algorithms can be used to test the integrity of workflows modeled using them, such as the "reachability" of the places (i.e. can a workflow be executed or are there parts that can't be performed because there are no arcs connecting them). A number of other network algorithms can be deployed, but the Petri Net emphasis on functional (activities) and behavioral (timing of activities) makes the use of Petri Nets to also describe organizational or informational aspects difficult. Nevertheless they are presented here in further detail because of their network nature and associated network metrics for their analysis with similarities to the analysis of the MDPM model described in Chapter 3.

2.1.4.4 Event Driven Process Chains

As part of the development of large-scale Enterprise Resource Planning systems such as SAP, a number of business process workflow methodologies emerged including ARIS (Architecture of Integrated Information Systems) developed by Prof. August-Wilhelm Scheer who also founded IDS Scheer, an information systems consulting company that popularized the method. ARIS models workflows using events, functions, interfaces, and connectors in what are called event-driven process chains (EPC). Events drive the process such that a temporal and logical representation of the business process can be made. Function elements capture the activities of a process. Event elements capture the pre-conditions and post-conditions of a function (an activity). An interface links multiple EPCs. And finally connectors can specify AND, OR, and XOR logic branches. For comparison, Figure 16 shows an example infection diagnosis and antibiotic treatment using both an EPC notation and a Petri Net notation. Because of their significant use in ERP systems, EPCs have received significant attention as business process models in part because they can also be analytically verified for consistency and errors using similar algorithms to software error checking as described in (Mendling 2009). In addition because EPCs are networks, many graph theory analysis algorithms can be brought to bear including density as a proxy for complexity as described in (Mendling 2006). A number of the MDPM model analysis metrics proposed in this thesis are based on validation metrics proposed by Mendling and others for EPC models. One significant limitation of EPCs and Petri Nets is the significant formalism required to build the models, as shown in Figure 16 with process activity sequences such as "diagnostic is requested" followed by "record diagnostic request". While highly useful to translate business processes into computer algorithms and workflows, the familiarity of most users, in healthcare for example, with such formal pseudo-computer languages is limited. The more familiar flowcharts or organizational charts being perhaps better alternatives to capture stakeholder inputs directly without the need to translate stakeholder knowledge directly into pseudo-code which likely requires the assistance of a "programmer" familiar with the language.

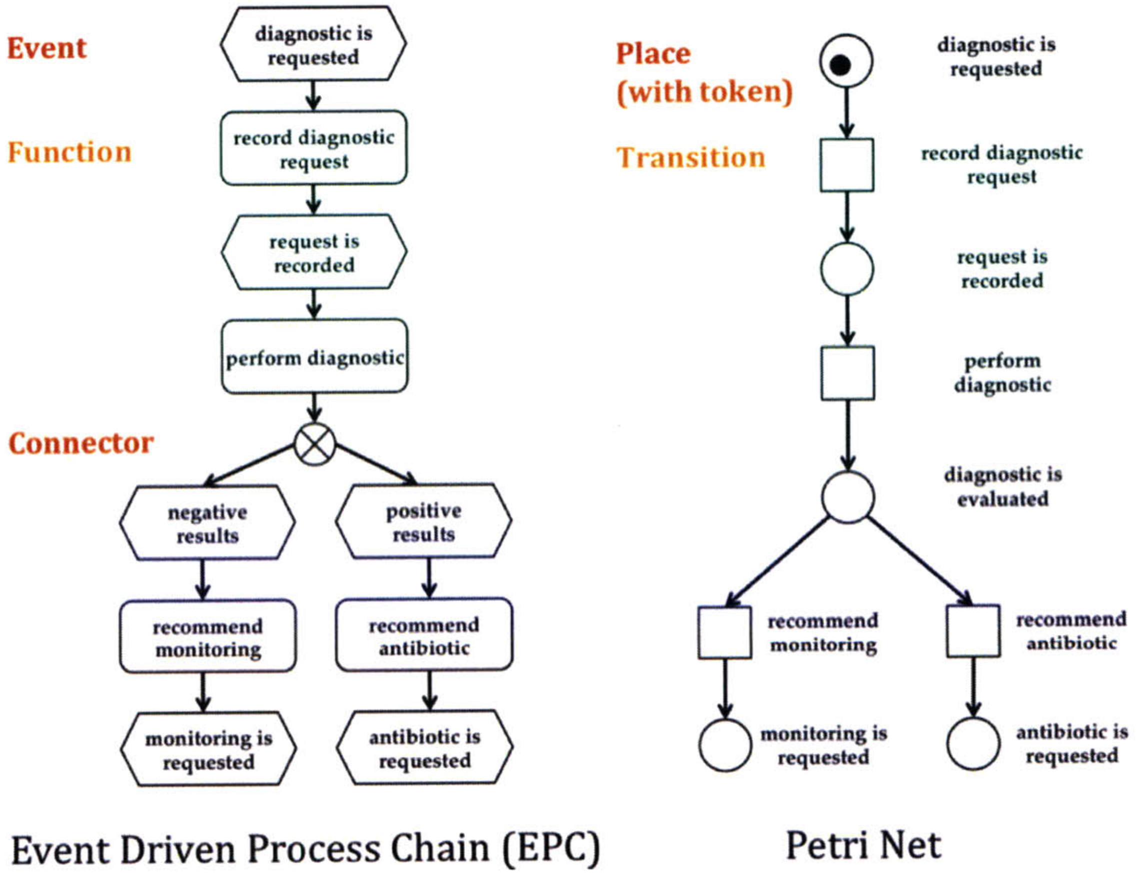


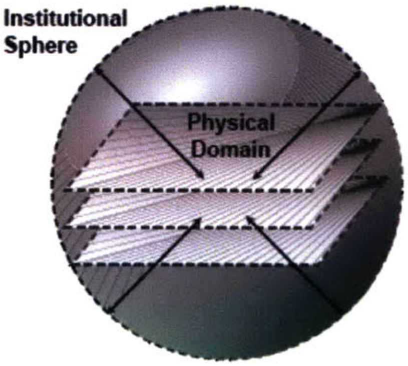
Figure 16 - Comparison of EPC and Petri Net Models (Adapted from Mendling 2008)

2.1.4.5 OPM, CLIOS, and COIM

The Object Process Methodology (OPM) developed by (Dori 2002) is a recent systems product design and systems engineering methodology allowing for the hierarchical decomposition of systems into objects whose states can be specified and processes. The relationships between objects and processes are specified by structural and procedural links that transform objects. Objects, processes, and links joining them are shown in an Object Process Diagram (OPD), which is used in OPM as a tool (rather than simply a diagram) to visualize the complexity of a system by "zooming in", "folding and unfolding", and "suppressing or expressing states". The ability to zoom in corresponds in some sense to layers, where the macro level view of a process may hide many of the details that zooming in to higher levels (layers) of detail might show. One advantage of OPM is that the system can be represented by one diagram (albeit with multiple "zoom" levels). However, OPM's focus on entities and process leaves out the organizational dimension (Osorio, Dori et al. 2009), where individuals may have important connections but which are perhaps not relevant to the process (and therefore not specified).

A general problem with many of the models described so far in terms of representing real processes is the need to formally represent process specific structures (that is only the elements needed to carry out the process in its "deterministic" or normal operational mode). However, it is under "non-standard" operations, such as during troubleshooting, attempts at improvement, and rapid process technology change that the "unspecified" connections become vitally important in the organizational and informational dimensions as it is not always possible to predict in a deterministic fashion the effects of changes or problems on complex processes. OPM can be useful in decomposition of systems (and processes) from high levels of abstraction to high levels of detail and this ability combined with another complementary modeling methodology that might address the limitations of OPM could yield a powerful hybrid methodology. The Complex Large Interconnected Open Socio-technical (CLIOS) methodology described in (Sussman, Sgouridis et al. 2005) is one such complementary methodology, which was joined with OPM into a hybrid methodology as described in (Osorio, Dori et al. 2009), called the CLIOS-OPM Integrated Method (COIM).

The CLIOS model separates a socio-technical system into a physical domain nested inside an "Institutional Sphere" allowing for layering of the physical domain into subsystems if necessary to provide more detail as shown in Figure 17 from (Dodder, McConnell et al. 2006).



Source: Dodder, Sussman et al. (2005)

Figure 17 - CLIOS System Representation

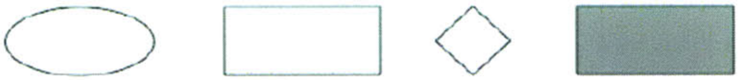
The CLIOS system modeling methodology is divided into 3 phases: system representation, design and evaluation, and implementation. The initial representation phase contains a useful methodology to describe a system's goals, elements, and connections between elements, and the initial steps of this methodology will be adapted for use in the MDPM methodology described in Chapter 5. The second CLIOS phase of design and evaluation also contains many useful elements for the development of the MDPM methodology although the DSM based MDM models and other network analysis tools are used in the MDPM methodology. The CLIOS representation phase has the following steps as described by Dodder et al.:

- *System Description*. The objective of this step is to describe the system, its major characteristics, goals, and the main issues at stake.
- *Identification of major subsystems of the physical domain and major actors of the institutional sphere.* In this step, we identify the major subsystems of the

CLIOS System, their nature, and relationships among them. An important aspect here is the definition of the Institutional Sphere and the identification of actors within this sphere.

- *Populating the physical domain and the institutional sphere.* In this step the functions and elements of each subsystem are described in greater detail. This is done by *nesting* the physical systems in the institutional sphere, *layering* the physical system into different subsystems and, if more detail is necessary, exploring some subsystems by *expanding* the analysis of some subsystems at finer granularity.

Dodder et al. defined four types of system components as shown in Figure 18 representing physical components, policy levers, common drivers, and an external factor



Examples	Physical Component	Policy Lever	Common Driver	External Factor
Electric Utility	Generator	Rights of Way Management	Reliability	National GDP
Internet Service Provider	Router	Telecommunications Act of 1996	Internet Protocol	Internet Backbone Provider

Source: prepared based on Dodder, Sussman et al. (2006)

Figure 18 - CLIOS Process Symbols (from Dodder et al.)

- *Describing components in the physical domain and organizations in the institutional sphere.* In this step we add detail and understanding by describing in detail the components on the physical domain and organizations in the institutional sphere. Detailed description results from gaining deep understanding of each component and organization's dynamics, behavior, relevance, critical factors for performance, and insights about their relationships.

- *Identifying links among components and organizations.* In this step, we identify the types of relationships between components, subsystems, and the various actors.

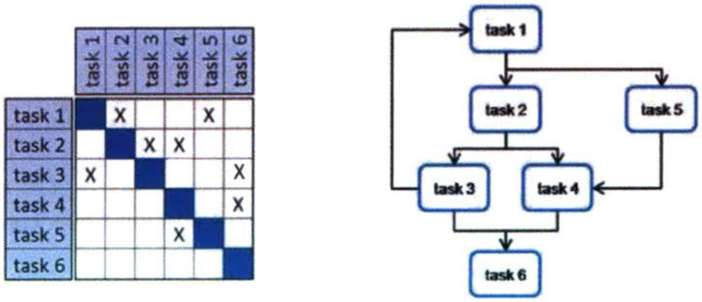
Three classes of links are defined by Dodder et al.:
(i) Class 1 links – links among elements of the physical system,
(ii) Class 2 links – links between the physical system and institutional sphere
(iii) Class 3 links – links among components of the institutional sphere.

- *Gaining insights about system behavior.* A major objective of CLIOS system representation is to understand its structure and behavior, at least to first order achieved by understanding its subsystems, components, relationships among them, and relationships with components of the IIS (internal institutional sphere) and the EIS (external institutional sphere).

A key limitation of the CLIOS method for process use is the lack of activity representation. With a focus on the elements of a system, CLIOS does not have an explicit task element for use in process description. However, CLIOS does have significant value in systematically enumerating all of the system goals, key physical elements, policy levers, major actors, and definition of the system environment. The MDPM model and the design methodology used in chapter 5 follows several of the steps in CLIOS to systematically identify, enumerate, classify, and link process elements but with a specific focus on process. The COIM model (Osorio, Dori et al. 2009) brings together elements of both CLIOS and OPM adding additional CLIOS elements to represent not just the physical element but also its function which could be modified to represent process activities as a potential process modeling technique. However, the central focus of COIM on systems as opposed to processes requires further development for this application and an expanded modeling notation to describe process activities specifically rather than components of a system.

2.1.4.6 DSM, DMM, MDM, and ESM

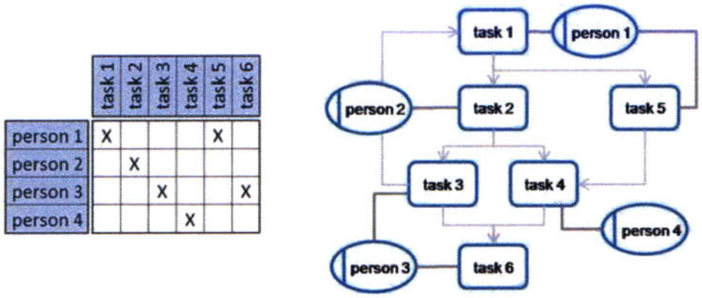
The Design Structure Matrix (DSM) is a systems analysis method based on matrix representation of graphs. DSMs can trace their history to early papers by (Steward 1965) and (Warfield 1973). The fundamental idea of a DSM is that design tasks (or subsystems in a product) can be enumerated along the rows and columns of a matrix, which corresponds to the adjacency matrix of a graph as shown in Figure 19.



	task 1	task 2	task 3	task 4	task 5	task 6
task 1	■	X			X	
task 2		■	X	X		
task 3	X		■			X
task 4				■		X
task 5				X	■	
task 6						■

Figure 19 - DSM Model (from DSMweb.org)

The DSM model can be extended to represent the mapping of two domains to each other such as people to tasks by forming a rectangular adjacency matrix that also represents a graph as shown in Figure 20. These models are called Domain Mapping Matrices (DMM) to distinguish them from the task or product element based DSM.



	task 1	task 2	task 3	task 4	task 5	task 6
person 1	X				X	
person 2		X				
person 3			X			X
person 4				X		

Figure 20 - DMM Model (from DSMweb.org)

Individual DSM and DMM models can be joined together into Multiple-Domain Matrices (MDM), which can show the relationships between various elements across an entire system. A representative MDM is shown in Figure 21 of a series of tasks and people with a resulting matrix sharing the interactions between all combinations of people-tasks, tasks-tasks, and people-people.

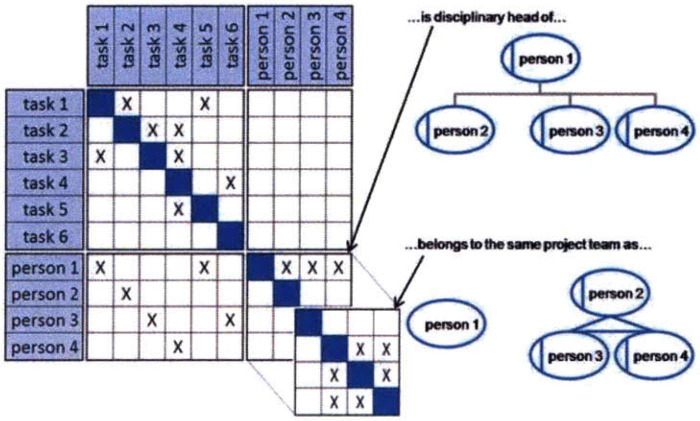


Figure 21 - MDM Model (from DSMweb.org)

Improvements to the DSM have most recently found valuable application in the design and development of complex products (Eppinger, Whitney et al. 1990) emphasizing the proper sequencing of tasks in product design processes. Algorithms can be applied to find more efficient task sequences in design processes such that the overall design cycle time can be reduced or the overall system knowledge can be managed. However, the DSM requires significant existing system knowledge (Dong 2002) and a mature product since any new tasks or new relationships between tasks cannot always be predicted a priori. This is especially problematic for the application to physical processes where for any given version of the process it may not be possible to re-sequence tasks and in complex new processes some of the tasks may not be fully specified. Also for troubleshooting or process improvement there may be in need to find (and rely upon) novel connections between process steps and the organization in order to find the root cause of a problem or model the impact of a change.

Multiple-Domain Matrices (MDM) however overcome some of the limitations of DSMs by mapping the "non-process specific" elements as additional domains such that the relationships between the process infrastructure and the "scaffolding" necessary during process development. These critically important connections between people, information, and the process represent a more realistic model of modern processes. In particular (Maurer and Lindemann 2008) describe the use of MDM models to address complexity in modern product design due to the growing number of necessary linkages across domains including technologies, components, process steps, and organizations. Maurer describes the application of MDM to represent an example of three different domains, which interact with themselves and with each other as shown in Figure 22. These domains might be product components, information, and people represented by the red circles, green triangles, and blue squares. There are linkages between elements of the product components for example represented by red circles and shown in the upper left corner of the MDM. However, there can also be cross-domain interactions such as product components with information as in the middle and right upper quadrants showing the interaction of product components (red circles) with people (green triangles) and information (blue squares). The overall matrix formed is an adjacency matrix representing the graph on the left of (figure), and the matrix can also be directed to show directionality in the linkages between elements and domains.

Although not found until much later in the execution of the research in this thesis, the MDM ideas presented by (Maurer 2007) have significant similarity to the MDPM model described in Chapter 3. The MDM ideas are seen to have significant potential to describe complex processes, which the MDPM model emphasizes. The ability to span multiple domains addresses many of the limitations of business process models described in the (Giaglis 2001) taxonomy since all of the domains (behavioral, organizational, functional, and informational) could be represented in an MDM as additional matrix elements. Despite these advantages there is limited literature on MDM modeling and application as this is a new and developing area of research with few "real world" examples to date in the MDM area, due in part to the significant complexity associated with the models.

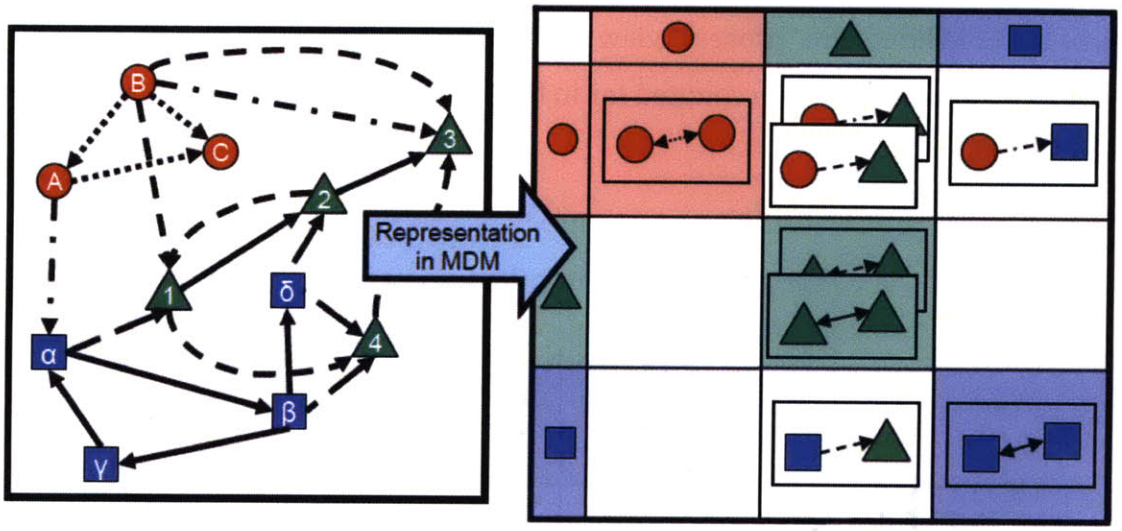
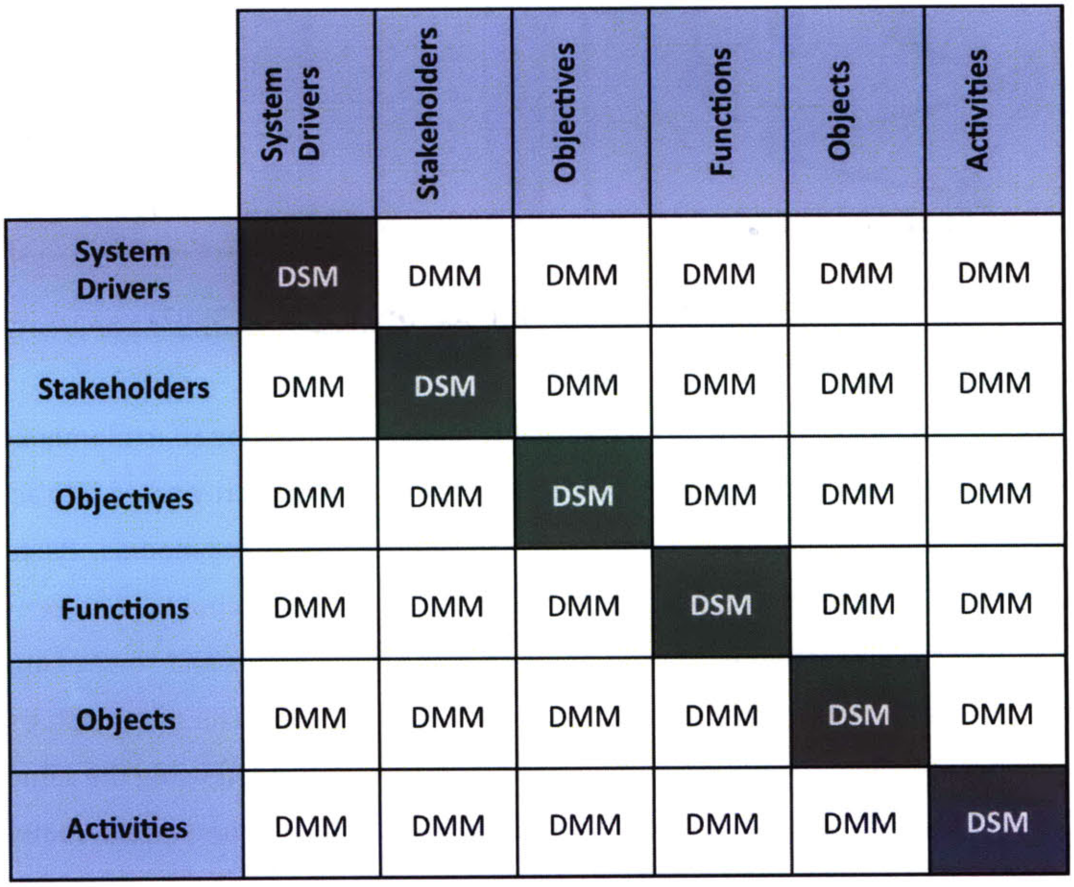


Figure 22 - MDM Graph and Matrix (from Maurer 2007)

Similarly, the Engineering Systems Matrix (ESM) model proposed by (Bartolomei 2007) is a multi-domain model with the same properties as MDMs that explicitly takes into account many of the system elements not traditionally considered by process models. Bartolomei describes an MDM structure that considers the system driver, stakeholders, objectives, functions, objects, and activities as domains each along the columns and rows of an MDM as shown in Figure 23 taken from (Bartolomei, Hastings et al. 2006). The relationships between elements in each of these domains can then be represented using the same tools as MDM analysis, such as the interaction of stakeholders with activities. Bartolomei's work also contains similarities to the MDPM model developed in chapter 3. The ESM model explicitly includes the system objectives (i.e. what a process must do), objects (physical infrastructure elements), and functions (things a system must do to fulfill its objectives), which are considered part of the initial methodology in specifying the MDPM model rather than parts of the model due to the significant additional complexity they would add. However, the ESM model does not address one critical dimension, the work instructions associated with processes, which are essential to every activity. Explicitly accounting for these protocols is essential since they are how design changes are propagated and must reflect the needs of stakeholders, objectives, and other domains just as any other of the elements described in the ESM. The MDPM model explicitly accounts for these work instructions and their interaction with the rest of the process. The product and system focus of the ESM limits the "process" view, but it is the author's view that these explicit work instructions should be accounted for in the "standard" Engineering Systems version of the ESM.



	System Drivers	Stakeholders	Objectives	Functions	Objects	Activities
System Drivers	DSM	DMM	DMM	DMM	DMM	DMM
Stakeholders	DMM	DSM	DMM	DMM	DMM	DMM
Objectives	DMM	DMM	DSM	DMM	DMM	DMM
Functions	DMM	DMM	DMM	DSM	DMM	DMM
Objects	DMM	DMM	DMM	DMM	DSM	DMM
Activities	DMM	DMM	DMM	DMM	DMM	DSM

Figure 23 - Engineering System Matrix (adapted from Bartolomei 2007)

Despite the potential limitations of DSM models to represent new processes as described in (Dong 2002) creative approaches have appeared to extend the utility of DSM (and by extension DMM and MDM) models to manage technology change. One such application is demonstrated in (Smaling 2005) who used DSM models to quantify the impact of "technology insertions" into existing systems. While the focus of these models is on product development, there is significant potential application on physical process design (i.e. MRSA diagnostic processes as opposed to product design processes). The approach taken by (Smaling 2005) is to augment a DSM by linking additional matrices of the same dimension as the original DSM to indicate a variety of technology architectures and types of changes such that a weighted sum of the changes corresponding to each of the architectures can be calculated which Smaling calls the "technology invasiveness" level. One limitation of this approach is that either all potential elements of all architectures must be listed in the original "baseline" DSM from which the technology invasiveness is derived, or changes must be limited to variations of an existing "mature" configuration. This might be a problem in technologies (and processes) that are rapidly changing and where design alternatives correspond to radically different process elements and sequences of these. Nevertheless the "technology invasiveness" methodology may have important applications in process design as developed in this thesis.

The literature surveyed on multi-domain models specifically applied to proceses finds some related research in the MDM models of (Maurer 2007) as described previously who apply DSM like tools to the mapping of multiple domains (and layers). However, Maurer's emphasis remains on design and products rather than "real" processes and there is no associated methodology for the enumeration of elements in each domain or the evaluation of feasible alternatives prior to mapping each of the domains. If all possible design alternatives and configurations are to be considered, the MDM models could be become inordinately complex or not meaningful as they are largely concerned with mapping other domains onto a single product or design process to identify interactions. Another related work using MDM like models is given by (Bradley and Yassine ; Bradley and Yassine 2008) who also use MDM models to map product development efforts attempting to link people, product elements, and manufacturing steps. Once again, the focus is on understanding the multi-domain impacts of existing processes without mention of a methodology to select feasible processes or how to enumerate and identify the necessary elements. The existence of these efforts is exciting as they already show alternative applications for MDPM-like models and could facilitate the wider adoption of these methods. In addition, although the MDPM model was developed independently, the similarities in the model, if not the methodology, provide some validation to the approach.

2.1.4.7 IDEF0

The IDEF family of system modeling languages traces its roots to the Integrated Computer Aided Manufacturing (ICAM) initiative of the United States Air Force and an initial contract to Doug Ross and his company SofTech to develop an industry wide model of how aerospace products are built (Feldmann 1998). This resulted in the Structured Analysis and Design Technique (SADT), a subset of which became the Air Force's ICAM language. The ICAM language in turn generated the Integrated DEFinition (IDEF) languages. A series of languages were developed from the initial ICAM set as shown in Table 4. The most relevant to this thesis are IDEF0 for functional (process) modeling, IDEF1 as a complement to IDEF0 for information modeling, and to an extent IDEF1x for data handling modeling. IDEF2 and IDEF3 are mainly intended for time-based modeling of systems and user interfaces (and user views) of the system. IDEF4 was intended as a design tool for software designers involved in object-oriented design. IDEF5 was intended for the capture of domain ontologies, which are used to capture objects in particular domains. Note that in Table 4 only languages IDEF0 to IDEF5 have completed (and accepted) standards while the majority of the other IDEF languages remain in various stages of development partly due to the very large and ambitious scope of the IDEF program.

Language	Purpose	Current Status
IDEF0	Function Modeling	Accepted Standard
IDEF1	Information Modeling	Accepted Standard
IDEF1X	Data Modeling	Accepted Standard
IDEF2	Simulation Model Design	Accepted Standard
IDEF3	Process Description Capture	Accepted Standard
IDEF4	Object-Oriented Design	Accepted Standard
IDEF5	Ontology Description Capture	In Development
IDEF6	Design Rationale Capture	In Development
IDEF8	User Interface Modeling	In Development

Language	Purpose	Current Status
IDEF9	Scenario-Driven IS Design	In Development
IDEF10	Implementation Architecture Modeling	In Development
IDEF11	Information Artifact Modeling	In Development
IDEF12	Organization Modeling	In Development
IDEF13	Three Schema Mapping Design	In Development
IDEF14	Computer and Communication Network Design	In Development

Table 4 - IDEF Modeling Language Family

IDEF0, developed originally to represent manufacturing processes, represents processes with a single generic process block that defines the activity, inputs, outputs, controls governing the activity, and mechanisms with which the activity is executed as shown in Figure 24. The "call" arrow represents the ability "drill down" into a more detailed expansion of the activity. This language is most relevant to this thesis due to its emphasis of process.

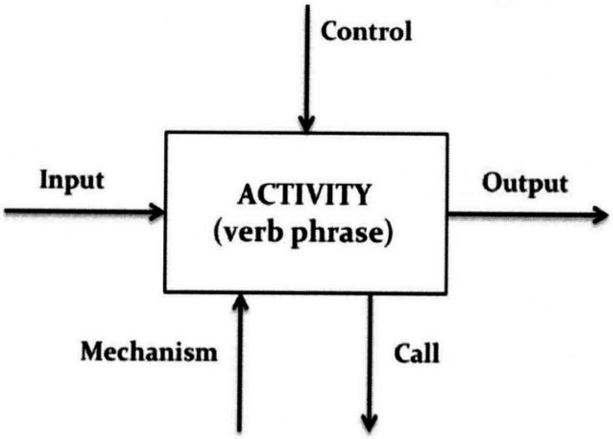


Figure 24 - IDEF0 Modeling Block

Shown in Figure 25 is an example IDEF0 diagram describing a design process. Note that while the activities are connected to each other as a graph, the "controls" and "mechanisms" have no such connection to each other and are shown as appearing on the diagram as needed such as the "designers" and the "drawing standards". This is a limitation of the IDEF0 language partly addressed by IDEF1 which can show the connections between information such as the "drawing standards" and "standard designs" in Figure 25, but this is not done so explicitly. The MDPM model developed in this thesis separates the "controls" and "mechanisms" into distinct domains and explicitly describes the linkages between all these elements in a manner similar to a Multiple-Domain Matrix (MDM) model.

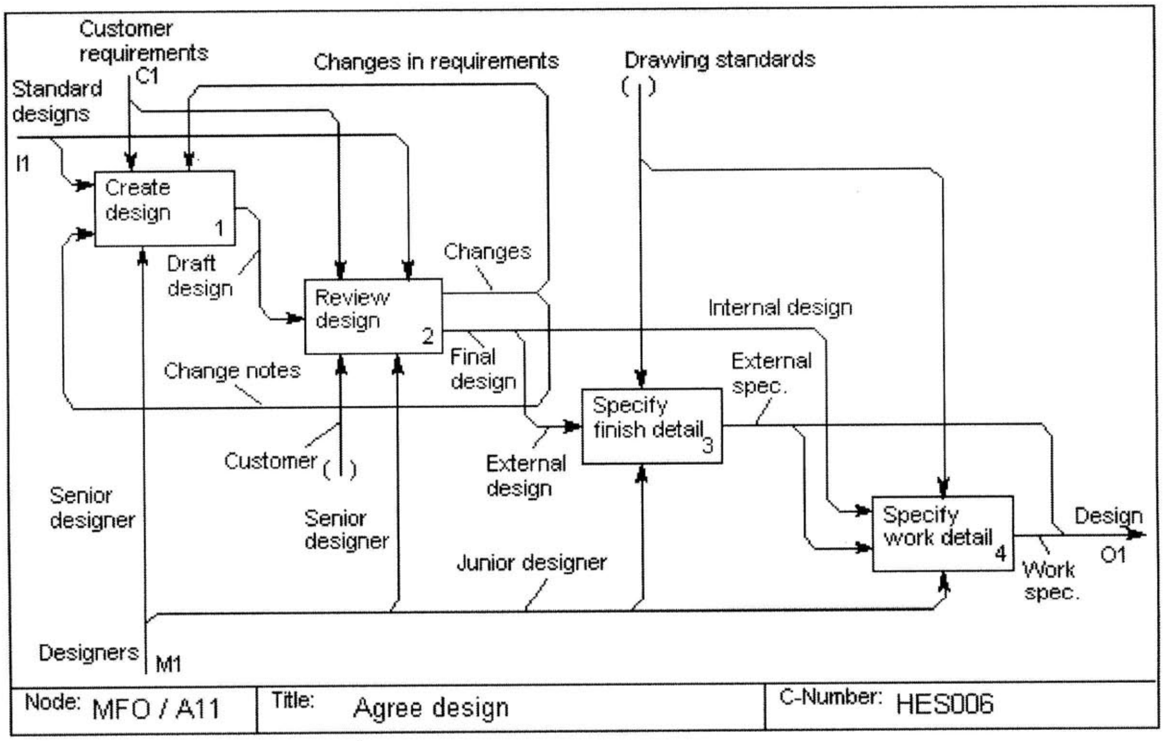


Figure 25 - Example IDEF0 Diagram (from syque.com)

The literature on IDEF0 tends to use the model as a complement to other analytical techniques and mainly to elicit system structure in terms of sets of activities. For example (Liu and Fang 2006) combine IDEF0 and Petri nets in developing a hybrid methodology called TTIPP to analyze processes as shown in Figure 26 where the IDEF0 model serves to define the places and transitions of the Petri Net which in turn allows for behavioral (performance) analysis. An IDEF0 example model developed by Liu and Fang is shown in Figure 27 for a model of emergency response.

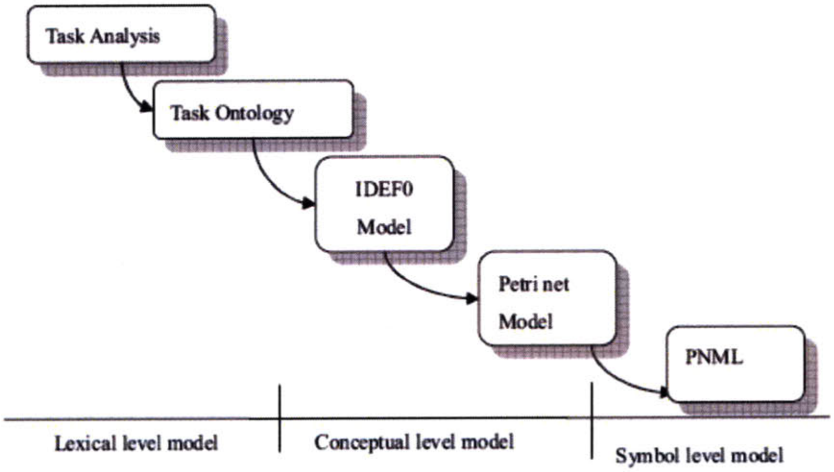


Figure 26 - TTIPP Methodology (from Liu and Fang 2006)

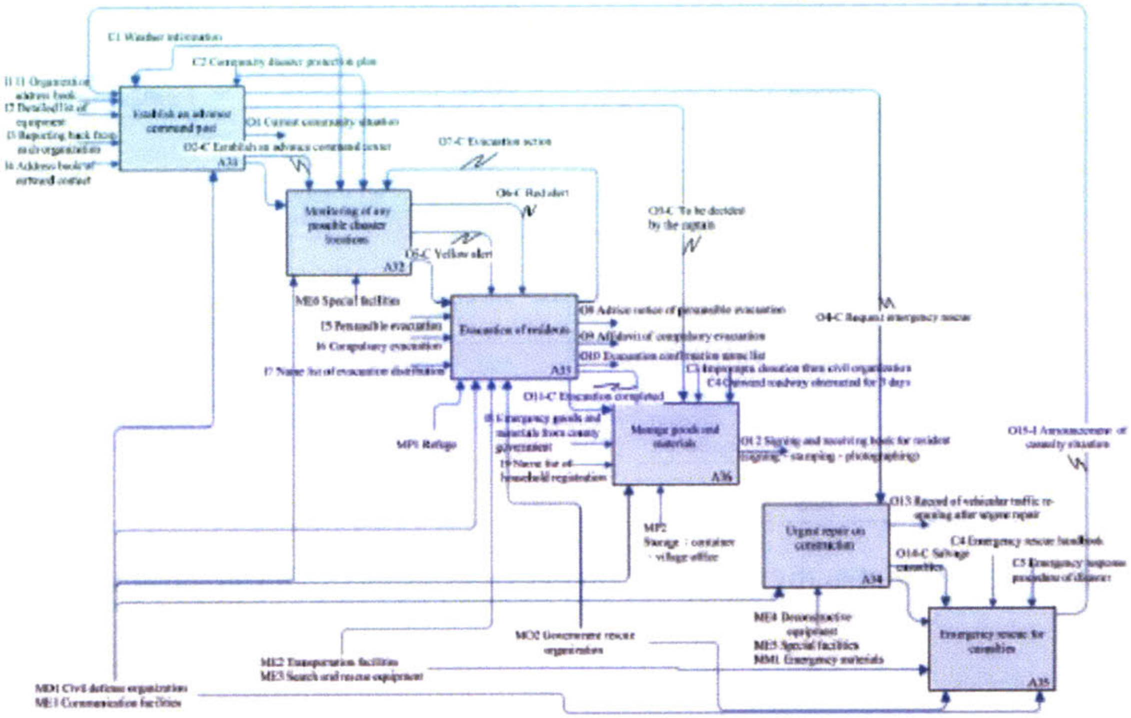

C1 Weather information
C2 Community disaster protection plan
Establish an advance command post
A31
O1 Current community situation
O2-C Establish an advance command center
O7-C Evacuation action
Monitoring of any possible disaster locations
A32
O6-C Red alert
O5-C Yellow alert
ME6 Special facilities
O3-C To be decided by the captain
Evacuation of residents
A33
O8 Advice notice of persuasible evacuation
O9 Affidavit of compulsory evacuation
O10 Evacuation confirmation name list
O4-C Request emergency rescue
I5 Persuasible evacuation
I6 Compulsory evacuation
I7 Name list of evacuation distribution
C4 Onward roadway obstructed for 3 days
O11-C Evacuation completed
MP1 Refuge
Manage goods and materials
A36
I9 Name list of household registration
O12 Signing and receiving book for resident (signing - stamping - photographing)
O15-I Announcement of casualty situation
MP2 Storage : container - village office
Urgent repair on construction
A34
O13 Record of vehicular traffic re-opening after urgent repair
C4 Emergency rescue handbook
O14-C Salvage casualties
C5 Emergency response procedure of disaster
ME4 Deconstructive equipment
ME5 Special facilities
MM1 Emergency materials
MO2 Government rescue organization
Emergency rescue for casualties
A35
ME2 Transportation facilities
ME3 Search and rescue equipment
MO1 Civil defense organization
ME1 Communication facilities

Figure 27 - IDEF0 Model of Emergency Response (from Liu and Fang 2006)

In healthcare IDEF0 models have been used to elicit system requirements from users prior to the development of a data model for a hospital information system as described in

(Staccini, Joubert et al. 2001) who modeled a hospital blood transfusion process. Staccini et al used a variation of IDEF0 to first categorize and enumerate all of the process elements in blood transfusion as shown in Figure 28 to force stakeholders to explicitly describe all of the process elements. Staccini et al also complement IDEF0 with a time-based description of the process as well as a data flow diagram as the IDEF0 diagram on its own did not provide sufficient detail.

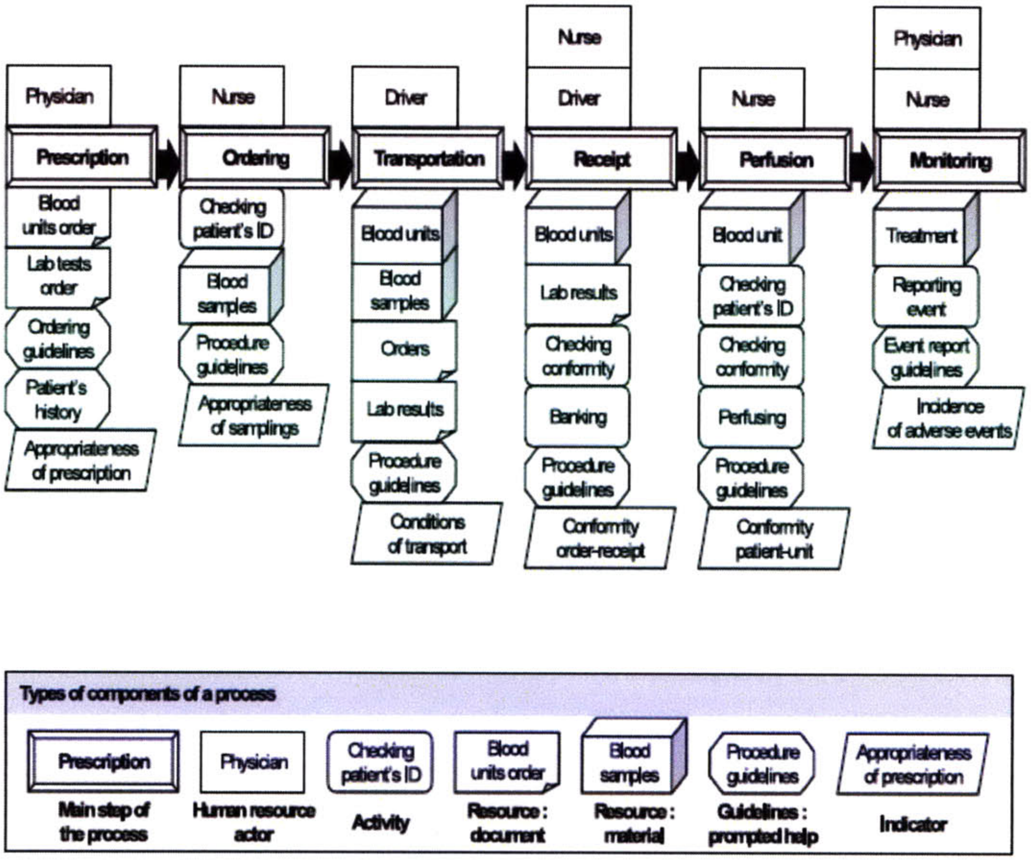


Figure 28 - Modified IDEF0 Model of Blood Transfusion (from Staccini et al 2001)

2.2 Network Analysis

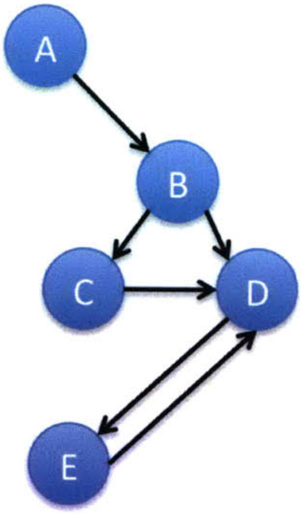
As described in the previous section networks form an integral part of the structure and analysis of some process models and an extensive literature spanning more than 250 years exists for networks and graph theory in general. Network analysis has recently entered a modern Renaissance being applied to fields as diverse as biology, organizations, marketing, and economics. But the roots of the field are quite old dating as far back as 1736 to Euler's solution of the bridges of Königsberg problem using graph theory ideas. This first period of network science, called the "Pre-Network period" from 1736 to 1966 as described in (Lewis 2009), saw the development of mathematics of graphs including formal descriptions of the mathematical objects used today such as vertices (nodes) and edges (links) with many of these fundamental concepts tracing back to Euler himself. An important but seemingly unrelated to networks discovery (Lewis 2009) occurred in 1927 with the publication of the first mathematical model (Anderson 1991; Kermack and McKendrick 1991) for the spread of infection in a population which although not in itself a network model relied on network ideas to explain the spread of infection along links connecting nodes and also formed the basis for models describing new product adoption where products (or information) spreads like an infection through a social network. Paul Erdos also made significant contributions to the mathematics graphs during this period including the analysis of random graphs and the Erdos-Renyi (ER) algorithm to generate random graphs.

The next major period of network science called the "Meso-Network period" from 1967 to 1998 (Lewis 2009) is marked by the significant application of graph theory to understand social structures such as Stanley Milgram's 1967 "6 degrees of separation" experiment (Milgram 1967) that showed social networks were not random but had certain "small world" properties that enabled large numbers of seemingly disconnected people to be connected with each other in less than six "hops". Much later (Watts and Strogatz 1998) would formally define small world networks and show that for small world networks the diameter (the longest number of hops between two nodes) increases as ln(n) while its size increases by $O(n)$ where n is the number of nodes. A related idea was proposed by (Granovetter 1973) who postulated that a few "weak ties" (long-distance connections between acquaintances) help tie together groups of "strong ties" (direct connections between family and friends) in social networks. Granovetter suggested that these "weak ties" help explain why it is possible to connect large sparse people networks in just a few hops as Milgram had found. Additional applications of network science to other fields occurred during this "Meso-Network period" including further work to relate infectious disease propagation models to new-product adoption network and technology diffusion models (Bass 1969; Norton and Bass 1987).

As described in (Lewis 2009) the current "Modern Period" of network science from 1998 to the present has seen further application to many other fields and a deeper understanding of the properties of networks including the analysis of both static properties (i.e. diameter) and dynamic properties (i.e. states and preferential attachment). A landmark publication in the "Modern Period" was (Watts and Strogatz 1998; Watts 1999; Watts 2003) demonstration of the universality and utility of "small world" networks and the Watts-Strogatz small world network generation algorithm. The discovery that biological networks such as the neural network of the nematode worm C. Elegans, the US electric power grid, and social networks all had small world properties further sparked interest in the "New" network science and a rediscovery of the graph theory in development for over 250 years since Euler. Other network types have been found such as "scale free networks" (Barabási, Dezsõ et al. 2003) which contain a few hubs of high degree (i.e. number of links) connecting many other nodes of much lower degree following a power law distribution. The small-world and scale-free network models produced an explosion of research and applications of network science in many fields but particularly relevant to this thesis in models of enterprises, communications networks, distribution of knowledge, and system complexity. The following sections describe selected network background and literature relevant to this thesis.

2.2.1 Properties

Graphs are defined mathematically as a 3-tuple: $G=(N,E,f)$, where N is a set of nodes, E is a set of Edges (Links), and f is a mapping function (matrix) that maps edges onto node pairs and can indicate directionality from one node to another. For example, Figure 29 shows a directed graph with nodes A-E and the corresponding adjacency matrix (the mapping function) listing all nodes and corresponding linkages between them as ones in the matrix. The location of matrix entries above or below the diagonal indicates directionality. All of the graph based process models (DSM, DMM, MDM, Petri Nets, EPCs) can be represented in this manner.



	A	B	C	D	E
A	0	1	0	0	0
B	0	0	1	1	0
C	0	0	0	1	0
D	0	0	0	0	1
E	0	0	0	1	0

Figure 29 - A Directed Graph and Corresponding Adjacency Matrix

The advantages of the matrix representation are that a number of metrics can easily and directly be computed in terms of graph properties. For example, the "in-degree" of a node (number of links "entering" the node) can be computed as the sum of that node's column entries and the "out degree" (links exiting that node) can be computed as the sum of that nodes row entries. A degree distribution for the network can then be calculated from individual node degrees. The diameter of a network can also be calculated as the longest path through the network (for the Figure 29 example path A-B-C-D-E with length 5). The shortest path length between every pair of nodes can similarly be found via search algorithms for more complex networks. Note that the network shown in Figure 29 is "weakly connected" which means that if we ignore the directional arrows (i.e. allow for things to go the "wrong way") then every node is connected to every other. But, if we follow the "rules of the road", node A cannot be reached from node E (or any other node). Directionality is especially important in network based process models, as activities must follow certain precedence and it is not always possible to go backwards in a process (i.e. many of the processing steps in a medical diagnostic are "irreversible" such as cell lysis). More sophisticated graph metrics can be calculated such as density of the network, which is the ratio of the actual number of links present in a network to the maximum possible density (i.e. the number of links to connect every node to every other node in a complete graph).

A number of additional and increasingly sophisticated metrics can be computed (Wasserman and Faust 1994; J. Carrington, Scott et al. 2005; Lewis 2009) but many of these are strictly associated with the type of network being studied to be meaningful. For example physical networks (i.e. a road network) require a different interpretation of path length (actual distance) than a social network (acquaintances in between) or a communications network (number of server hops). The relevant process network metrics used in the MDPM model are discussed in Chapter 4, where the network properties of the MDPM are exploited. Significant effort already exists to classify process networks according to sets of metrics for EPCs, Petri Nets, and others (Cardoso, Mendling et al. 2006; Gruhn and Laue 2006; Mendling 2009) and some of the metrics presented in Chapter 4 are based on such prior work. In addition there is a growing body of research on metrics to characterize the complexity of networks according to various measures (Costa, Rodrigues et al. 2005), although no single measure (or even set of measures) appears to be suitable for all network types and applications. Yet all of them are based on the fundamental concept that graphs can be represented as mathematical objects and matrices in particular as shown in Figure 29.

2.2.2 Organization Models

A fundamental need in models to represent all dimensions of complex processes is the representation of organizational elements in addition to the process activities. This is a particular strength of social network analysis, which traces its roots to Georg Simmel and other sociologists in the early 1900's interested in describing the relationships between individuals in groups such as classrooms, workgroups, and other interpersonal relationships. However, it took some time to develop analytical underpinnings for the sociological concepts and it wasn't until the 1960s that groups such as Harrison White, his students, and others (including Milgram and Granovetter) that modern analytical social network analysis was born. Among the areas of application for social network analysis are the diffusion of ideas "how quickly does an idea spread", collaboration patterns "who works with whom" as shown for scientific networks in, influence networks "who can influence who", and communication networks "who talks to who". All of these types of applications have significant importance in addressing the organizational dimension of complex processes as a key concept of this thesis is that any single dimension is not sufficient to model a real process. The relationships between individuals associated with a process are critically important to the "nonstandard" operation of processes such as troubleshooting, process improvement, and change since by definition these nonstandard events cannot be neatly specified a priori and require some "exploration" and "search" of necessary information and resources to resolve.

Social network analysis (J. Carrington, Scott et al. 2005) can represent these social elements (people and organizations) such that the relationships between elements can be specified but also key components identified and the efficiency of "searches" calculated. The elements of a social network are nodes, which represent "actors", and ties, which represent relationships between the actors. There may be multiple kinds of ties between actors such as work, social, and personal relationships but these can all be represented as graphs using adjacency matrices similar to Figure 29. From the network representation and adjacency matrix defining the relationships between nodes on number of useful measures can be calculated for individual actors in the social network as described in (Freeman 1979; Borgatti 2005) including:

- **Betweenness:** A measure of how important a node (actor) is to the communication paths in a network. Formally it is a measure of the number of shortest paths in a network that pass through the given node relative to all shortest paths. An actor with a high betweenness value is a part of more communications paths than others and is therefore more "important".

- **Closeness:** The degree an individual has shorter path lengths to reach other nodes and reflects the ability to get "information" from other actors.

Both of these measures might be particularly important to identify key individuals in processes to consult troubleshooting problems with, as they are more likely to either know information or know who might know. A number of other measures are available that define cliques in social networks, key players who can serve as hubs, and others as described in (Wasserman and Faust 1994). Some of these measures are used in Chapter 5 to identify key elements of the process that can help transfer knowledge because of their high betweenness centrality or elements that are highly disconnected and must have ties added (people communicating with them) to improve the communications performance and robustness of the network.

2.2.3 Information Models

Just as the organizational dimension is essential the informational dimension must be represented in a complete process model. Network science has also been applied to model information and knowledge networks including maps of citation networks, patent networks, communications networks, and nearly all computer networks. Some overlap exists with social networks in that these also contain knowledge and network analysis of firm collaborations as in (Owen-Smith and Powell 2004) to represent knowledge networks blur some of the distinctions. In this thesis information is treated as explicit knowledge meaning written policies, guidelines, work instructions, algorithms, and other strictly documented forms of information. This distinction is useful in process models because nearly every activity has an explicit knowledge component (i.e. a written protocol) and a tacit knowledge component (i.e. the operators experience). This explicit distinction also allows for the tracing of information "genealogies" where a particular work instruction might be derived from a local hospital guideline, which in turn might be derived from an industry-standard. When a change occurs such as a new industry standard, these changes must be propagated throughout the information network all the way down to the work instruction associated with a particular activity.

However, in complex processes the effect of a particular change as it propagates through this information network cannot always be predicted and is part of the reason to explicitly model information. An elegant description of network-based explicit information change analysis can be seen in (Giffin, de Weck et al. 2009) who mapped change requests during the design of a complex sensor system with the largest component of this change propagation network shown in Figure 30. In similar fashion the information domain of the MDPM model described in chapter 3 focuses on tracking work instructions associated with process activities and the associated information (guidelines, other protocols, and changes) these work instructions are derived from (and affected by).

Communications networks that transmit information have been modeled extensively in the literature, notably starting with (Shannon and Weaver 1948) and a line of literature developing information theory to apply to many other fields including biology. A key concept from the information theory literature is entropy, defined as the amount of information needed to know the state of a random variable, and can be used to quantify the amount of information contained in a message across a communication network in bits. These ideas have been extended to quantify the amount of information needed to search networks as described in (Rosvall, Gronlund et al. 2005), which is of particular importance to MDPM model developed here as most complex processes will require extensive searches during troubleshooting, improvement, and change conditions to find root causes of problems or find the impact of changes on the rest of the process.

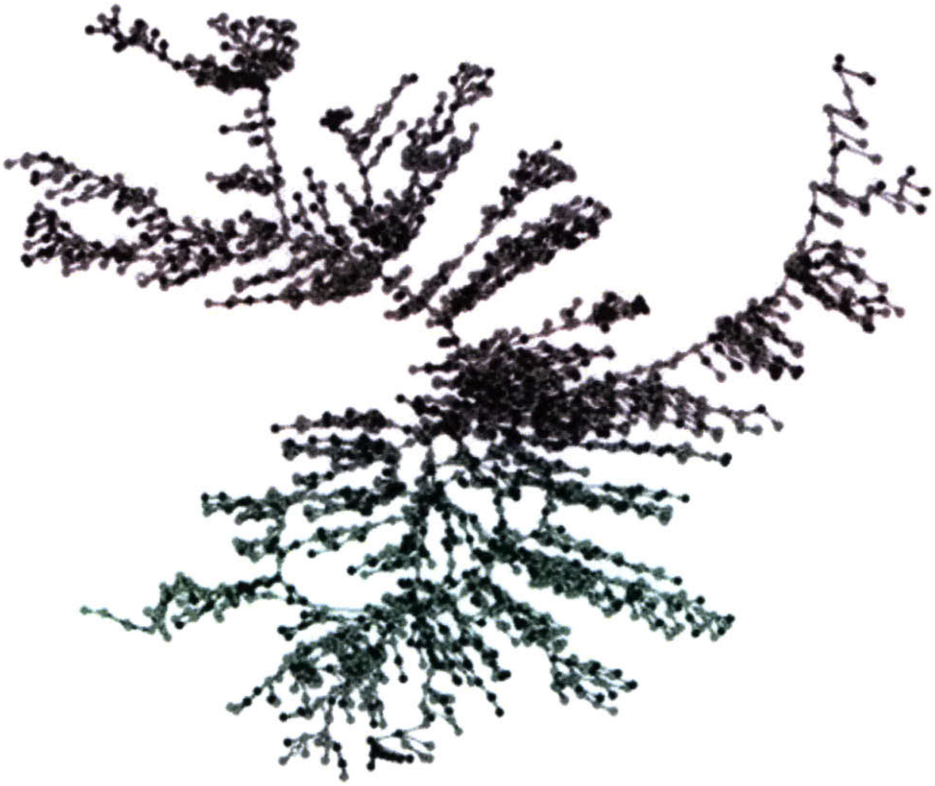


Figure 30 - Largest Engineering Change Request Network in a Complex System Design Program (from Giffin, de Weck 2009)

2.3 Healthcare as a Complex System

The US healthcare system is a vast collection of services, products, organizations, stakeholders, and regulatory agencies representing 16.2% of the US economy in 2007[3]. To constrain the scope of this thesis, the focus is on clinical microbiology, microbial surveillance, and DNA sequencing processes within the healthcare system and background on these is provided in the following sections. As described in Chapter 1, microbial surveillance has regained importance with rates of antibiotic resistant microbes including

[3] Source: National Health Expenditures – HHS 2008

MRSA and background is provided on the state of the antibiotic arsenal, which is nearly depleted, and MRSA such that the importance of the demonstration project described in chapters 5 and 6 can be understood.

2.3.1 Healthcare Processes

All of the diagnostic, therapeutic, and business processes in the healthcare system can be cast as processes within organizations subject to policy and system constraints. The healthcare system is in fact already organized around processes, although not explicitly described as such. In an effort to provide a means to classify disease and procedures to treat disease, catalogs were developed in the late 1800's. These evolved in the US into the "International Classification of Diseases, Adapted for Indexing of Hospital Records and Operation Classification" (ICDA), first published by the US Public Health Service in 1948. The current version used in the US is ICD-9-CM (clinical modification) published by the Department of Health and Human Services (HHS). This catalog contains codes to group and identify diseases, symptoms, and medical parameters, but also for purposes of this thesis codes to track health interventions (processes) performed by medical personnel. The ICD-9 is mainly aimed at disease diagnosis classification, however, and most insurance reimbursement for procedures is performed using the American Medical Associations' Current Procedural Terminology (CPT).

The CPT forms the basis for the Healthcare Common Procedure Coding System (HCPCS), which is the Federal standard for coding Health Care and includes everything from a routine blood pressure check to heart transplants, and are the means by which hospitals track, bill, and reimburse for procedures performed. HCPCS was first established in 1978 to standardize codes for insurance reimbursement of items and services primarily for Medicare and Medicaid. With the implementation of the Federal Health Insurance Portability and Accountability Act (HIPAA) in 1996 the use of HCPCS became mandatory. HCPCS is divided into 3 levels:

Level I - Comprised of AMA's CPT numeric coding structure, and further divided into 3 categories:

> **Category I** – Medical, Surgical, and Diagnostic Services: "consistent with contemporary medical practice and performed by many physicians in multiple locations"[4].
>
> **Category II** – Performance Measurement: Optional codes used to track certain measures "agreed upon as contributing to good patient care"
>
> **Category III** – Emerging Technology: Temporary codes used for new and emerging services and procedures.

Level II - Primarily non-physician services including ambulances and prosthetic devices. Alphanumeric.

Level III - Local (State Medicaid) codes for use in specific programs but use is now discontinued.

The majority of US insurance claims are contained within Level I and Category I of the HCPCS structure. The introduction of new technologies however can be lengthy and complex, and new codes start with a request for a temporary code in Level I Category III, where the AMA CPT Editorial Panel will review the request subject to the following requirements:

- A protocol for a study of procedures being performed
- Support from the specialties who would use the procedure
- Availability of U.S. peer-reviewed literature
- Descriptions of current United States trials outlining the efficacy of the procedure

Once sufficient data has been gathered, a new technology category III procedure can be reviewed for inclusion the "standard" category I codes, subject to the review of the Editorial Panel, and the following requirements:

[4] Source: AMA Website

- That the service/procedure has received approval from the Food and Drug Administration (FDA) for the specific use of devices or drugs;
- That the suggested procedure/service is a distinct service performed by many physicians/practitioners across the United States;
- That the clinical efficacy of the service/procedure is well established and documented in U.S. peer review literature;
- That the suggested service/procedure is neither a fragmentation of an existing procedure/service nor currently reportable by one or more existing codes; and
- That the suggested service/procedure is not requested as a means to report extraordinary circumstances related to the performance of a procedure/service already having a specific CPT code.

The cycle time for this process is long, and is not the end point for widespread adoption. Once a procedure has made it into category III, hospitals can bill insurers (and medicare) for these services. As an example of relevance in this thesis, CPT code 87641, DNA probe PCR based detection of MRSA, was only introduced into the standard CPT codes in 2007, and not all hospitals and insurers classify this procedure as "medically necessary" further delaying adoption. Partly due to the slow adoption the costs for this diagnostic remain high, $272 per one hospital system[5]. This compares to less than $5 per culture-based assay. To emphasize the long cycle times in adoption of new technology, the basic technology for the Polymerase Chain Reaction (PCR) was invented in 1984 as described in (Mullis and Faloona 1987), and the specific quantitative PCR (qPCR) technique used in this assay was invented in 1992 as described in (Higuchi, Dollinger et al. 1992). Applied Biosystems introduced the first commercial quantitative PCR system in 1996. While qPCR found numerous applications in other diagnostic areas, its use in MRSA diagnosis has lagged due to a number of factors:

1. Cost compared to culture based assays
2. Non-standard MRSA screening practices (not required in most hospitals)

[5] Source: UMC Health System, Texas: https://www.umchealthsystem.com/patientfinancialservices/pricing/

3. No perceived need (antibiotics working for most cases)
4. Physician training (how to interpret)
5. Laboratory training (how to perform and troubleshoot)
6. Workflow (speed of assay slowed by centralized laboratory handling, negating advantages)
7. Limited commercial availability (partly due to reimbursement restrictions)
8. Not phenotypic (detects resistance gene, but does not specify antibiotic to be used)
9. No accepted validation (no "authority" certified its validity, various studies, but AMA and others slow to certify)
10. Highly specific to particular genes and cannot detect changes in resistance or virulence (mutations or adaptations)

The cycle time for the qPCR assay can be represented by a series of phases from discovery to adoption as follows:

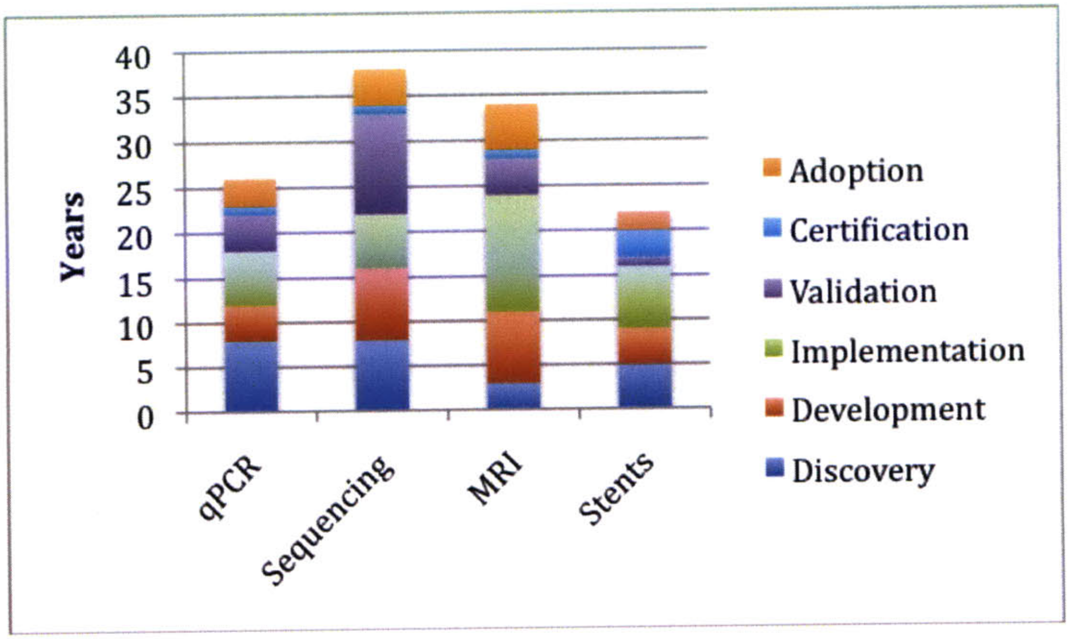


Figure 31 - Time to Widespread Adoption for Various Medical Technologies

Over 25 years elapsed from initial discovery of PCR to an approved reimbursable procedure. And while the exact start dates of each phase are somewhat arbitrary the main issue is the overall length of time from initial publication to a reimbursable procedure is exceedingly long. By contrast MRIs took longer, over 34 years as described in (Filler 2009) , but are also a more complex process with significant training, capital cost, facilities, and development requirements. DNA Sequencing of the type used in the Human Genome Project is plotted in Figure 31 and also had a lengthy maturation, and although it uses many of the same techniques as PCR, it is still not an "approved" procedure and the validation and adoption phases in Figure 31 are estimated. The lengthy development of DNA sequencing and high initial costs before new technologies appeared are described in numerous articles including (Hutchison 2007). These phases are representative of what most processes in healthcare must go through before they become commonplace best practices.

1. **Discovery:** Initial creation of a process and feasibility demonstration (discovery of PCR 1984 to application in qPCR 1992)
2. **Development:** Developing a prototype process (developing prototype qPCR 1992 to commercial product ABI 1996)
3. **Implementation:** Commercial robust process (product in the field and improvements 1996-2002)
4. **Validation:** Evaluation by multiple sites under real conditions (evaluation of qPCR for MRSA specifically 2002-2006)
5. **Certification:** AMA grants level I code for routine medical use (2007)
6. **Adoption:** Dissemination of the technology into routine medical practice (2007-2010?)

The adoption of qPCR into routine MRSA diagnosis is a particular example of the lengthy cycle-time for process adoption in healthcare due to the complexity of the processes and the system they are performed in. And by comparison to other processes it is not particularly complex. Sequencing has taken even longer from initial discovery and is still in an implementation phase. The work in this thesis is one of the first to demonstrate a robust process for using sequencing as an MRSA diagnostic, but likely it will still take several more years to pass through certification and adoption phases. Other industries have much faster process discovery to adoption cycle times, and have developed process focused design and development methodologies largely absent from healthcare. One of the objectives of this thesis is to show that the significant initial complexity of healthcare processes greatly slows their adoption, and it is only once the complexity of these processes has been reduced either through technology improvement or organizational changes that these processes can be adopted. Changes to the way processes are created, designed, implemented, and diffused should focus on ways to reduce this process complexity which many organizations in healthcare simply do not have the resources to deal with. One of the problems is that this complexity is largely hidden and only becomes apparent in initial implementations and it makes the barrier to entry significantly higher.

2.3.2 A Brief History of Microbiology and Antibiotics

One of the key processes in the healthcare system is the control of microbial infection, and for the last 60 years human beings have held temporary advantage in the fight against bacteria. But as discussed in chapter 1 this advantage is rapidly diminishing, and without significant effort to develop new antibiotics and perform much greater surveillance and research into the mechanisms of resistance we may lose this advantage entirely. Today we take for granted the ability to control infection during surgery, but this was not always so and returning to the terrifying situation before antibiotics would render our entire healthcare system inoperative. It is worth describing what the world was like before modern antibiotics as well as trace through the history of microbiology for some perspective, summarized from (Porter 2001; Kennedy 2004).

Microbiology begins in earnest with the development of the microscope in the Netherlands in the 1600s. Robert Hooke was the first to use the microscope as a tool and in 1665 published "Micrographia" in which he coined the term "cell". Further work followed with Von Leeuwenhoek and others who catalog the various cellular structures in nature and the human body. In the 1800s Germany took the lead in the development of advanced microscopes that further enabled the development of modern pathology, including catalogs of disease tissues for the education of the first generation of pathologists. The 1800s also sees the development of anesthesia and enabling the first internal complex operations, but also exposing the need for sterile conditions during surgery. Because of the poor understanding of bacterial infection in the mid 1800s many physicians resisted calls for antiseptic technique. In particular, Boston physician and Harvard anatomy professor, Oliver Wendell Holmes was one of the first to consider infection as caused by germs transmitted by attendants in hospital ward. He recommended washing with chlorinated water and changing clothes. However, other physicians disputed his theory and scoffed at the need for antiseptic technique.

At about the same time, Ignaz Semmelweiss at the Vienna General Hospital noted that the two maternity wards there had very different rates of infection. Ward 1, staffed by medical students, had a 29% mortality rate, whereas Ward 2, staff by midwifery pupils who did not attend autopsies had 3% mortality rates. Semmelweiss switched the medical students to Ward 2, and the mortality rate jumped following the medical students. In May 1847 he ordered handwashing with chlorinated water and mortality rates plummeted. His colleagues resisted these changes and in 1851 he resigned and moved to a Budapest hospital where he introduced chlorine disinfection and mortality rates in the maternity ward fell below 1%. But even there he could not convince his colleagues of the hazard of the microscopic organisms they could not see. Bacteriology was still in its infancy and it would not be until Pasteur that wide acceptance of antiseptic technique would occur. Yet despite modern knowledge of the causes of microbial infection, resistance to "process change" including changes to attitudes, work instructions, and technology remains pervasive even today where hospitals with significant patient isolation and transmission control (hand-washing etc.) programs have lower rates of MRSA infection (Harbarth 2006). Semmelweiss was fighting system change with only empirical evidence, but even today similar themes appear.

The next significant milestone in our understanding of microbiology came from the efforts of Joseph Lister, a surgeon in Glasgow, and Louis Pasteur a French chemist. Following surgeries Lister often encountered sepsis, which was believed to be caused by "miasma" or bad air. Lister noticed that when wounds healed cleanly results were good but anytime an infection developed it was likely to lead to sepsis and death. At about the same time John Snow, a leading physician of the time, established a clear connection between cholera outbreaks and sewage-contaminated water. In a famous and elegant epidemiological effort Snow was able to map cholera outbreaks on a London city map as shown in Figure 32 and track these outbreaks to local water pumps. In particular, Snow found 89 deaths associated with the Broad Street pump and advised the city to close the pump. Legend has it that when the city refused to close the pump, Snow stole the handle in the middle of the night and the epidemic in the neighborhood stopped. Later investigation showed that the Broad Street pump well was not deep enough and had been contaminated by an adjoining cesspool.

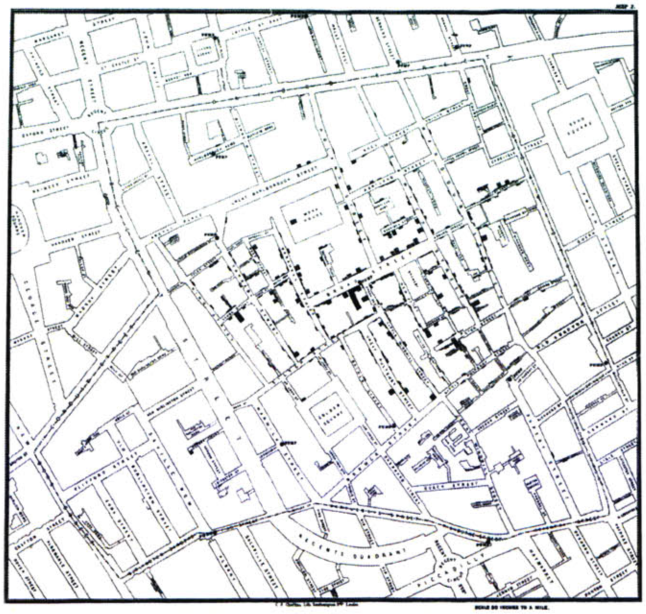


Figure 32 - John Snow's Map of Cholera Outbreaks in London

Snow became convinced that a living organism was causing cholera and not miasma. Following Snow's work the city of London greatly improved its water distribution and sewage systems greatly reducing cholera outbreaks. As part of the sanitary improvements, carbolic acid was found to have antiseptic properties and to greatly reduce the smell emanating from cesspools the city of London. Carbolic acid became widely used as a result, and with the sanitation improvements would today be thought of as a significant system change to critical infrastructures in response to microbial threats. Ironically, the rapid increase in resistant organisms not just in hospitals but also in food production may require similar changes in our food supply chains to manage growing resistance and the potential spread as described in (Soulsby 2007).

Lister in the meantime was looking for ways to reduce sepsis among his patients and became aware of the carbolic acid use. Lister began trials on patients with compound fractures where he would use lint soaked in carbolic acid linseed oil as a dressing. Lister expanded on this including the bathing of the wound in carbolic acid, spraying carbolic acid into the air during surgery, and the changing of dressings to ensure fresh carbolic acid covered the wound. As before with Semmelweiss, acceptance was not immediate and skeptics abounded amongst other surgeons. Some of this may have had to do with their reluctance to admit that they themselves could be the cause of high mortality rates in their wards with unclean hands and clothes. Continuing the theme of organizational resistance to process change, which must be addressed as part of any process change, especially today with far more complex but also harder to change processes. Other surgeons had adopted hand washing and clean operating rooms because of Florence Nightingale's efforts to improve antiseptic technique but refused to accept the "germ theory". Some German surgeons did adopt Lister's ideas however and in one Munich hospital the surgery mortality rates plummeted so much that the new "Listerism" rapidly won converts. Yet back in England, adoption was much slower. Lister and others had found ways to treat the source of infection but the cause remained unknown. John Bennett, professor of surgery and Edinburgh, said, "where are these little beasts? Show them to us and we shall believe in them. Has anyone seen them yet?" Today there is widespread belief that MRSA infections are caused by particularly well-adapted clones (i.e. all MRSAs of a particular outbreak are the same) and therefore the need for surveillance may not be as great. One of the key technical objectives of this thesis was to show significant variation even among MRSA samples that were thought to be phenotypically the same. Although we now know microbes cause disease we actually know very little about how they are able to adapt so quickly.

The cause of bacterial infection remained unknown and subject to much debate until the experiments of Louis Pasteur. Working on ways to improve fermentation Pasteur identified microorganisms for souring wine and showed that heating wine to 55 degrees centigrade killed the microorganisms and prevented the souring. Eventually, the "Pasteurization" principle was applied to commercial production of beer and milk. Pasteur followed this work with a demonstration that pebrine, a silkworm disease, was caused by protozoan and not "spontaneous generation". On February 19, 1878 he appeared before the French Academy of medicine to present the germ theory of disease and predicted that specific organisms would be found to produce specific diseases. In rapid succession Pasteur working with others including Robert Koch, isolated the organisms responsible for anthrax and rabies.

Koch continued the work begun by Pasteur and established the principles of modern bacteriology. The basic idea that bacteria must be identified and classified, then connected to the appropriate clinical disease became the basis of bacteriology. Koch developed many of the modern microbiology techniques including growing organisms on solid media such as agar, a Japanese seaweed extract that remains solid at body temperature. At about the same time Richard Petri invented a flat round dish to support these bacteriological studies. Koch and his lab went on to discover organisms responsible for tuberculosis, diphtheria, typhoid, pneumonia, gonorrhea, meningitis, leprosy, plague, tetanus, syphilis, whooping cough, Streptococcus, and Staphylococcus aureus, which would eventually evolve into today's MRSA.

Paul Ehrlich, a German pathologist joined Koch's Institute of infectious diseases and began searching for molecular mechanisms that might affect bacteria. In particular, Koch reasoned that since certain dyes could be used to differentiate human and bacterial cells, such as Gram's stain, so too might other chemicals affect microbes. He began to look for chemical agents that were toxic for bacteria but did not harm the host (which is a key constraint of antibiotic discovery). His work eventually led to a drug named "chloroquine" which provides effective prophylaxis against the malaria parasite. Ehrlich went on to tackle syphilis whose parasite had by then been identified along with a specific blood test in 1906. However, the only treatment at the time was Mercury in use since the 1500s. Ehrlich had been working on arsenical drugs to treat previous diseases and he decided to try variants of these going through over 600 until number 606 showed efficacy. Clinical trials of sorts began using "606" using volunteers and by 1910, over 10,000 syphilitics had been treated. The drug was commercially named "Salvarsan" and in 1914 an even more effective variant called Neo-Salvarsan was introduced although being arsenic-based remained toxic. Despite these advances it would be another 21 years until the first modern antibiotic was discovered.

The antibiotic properties of the Penicillium notatum mold had been observed before Fleming's discovery but Fleming was the first to fully investigate its properties. His observation of the growth inhibition of Staphylococcus and the presence of a Penicillium notatum mold colony was serendipitous since it required a particular sequence of temperatures to allow the mold to grow first for a number of days and then the Staphylococcus. The famous Petri dish left on a bench contained Staphylococcus on which mold spores from the lab below formed a Penicillium notatum colony. It also happened that London was suffering from a heat wave that would have prevented the Penicillium mold from growing, but the heat wave broke for 12 days allowing the mold to grow and then returned for two days allowing the Staphylococcus to grow. It was this particular sequence of events, a lab below Fleming's working on Penicillium notatum mold, Fleming forgetting to place the Petri dish in the incubator (which would have prevented Penicillium from growing), and the particular sequence of the London heat wave with 12 cool days and two hot ones that provided Fleming with his "discovery" of penicillin upon his return from vacation. Fleming then went about purifying and testing penicillin and found it effective against Staphylococcus, Streptococcus, Gonococcus, and Meningococcus. However, Fleming's purification methods meant his preparations quickly lost potency and Fleming put penicillin aside as nothing more than a topical agent similar to carbolic acid in an article summarizing his work. However, others read this article and asked Fleming for samples of his antibacterial mold. But it would be several years until Fleming's discovery was transformed into a mass-produced drug.

In the meantime Gerhard Domagk in 1935 reported the antibacterial effects of Prontosil a type of azo dye, which Paul Erlich had predicted might have antibacterial properties. In a poignant anecdote, Domagk's daughter became gravely ill with a Staphylococcus infection during his early studies of Prontosil (which was only marginally effective). In desperation he injected his daughter with Prontosil and she recovered. A derivative of this dye, where Prontosil is cleaved in half, became the first of the effective "sulfa" drugs, sulfanilamide. These worked by competing with active compounds within bacteria preventing these compounds from doing their work. Domagk was awarded the Nobel Prize in 1939 for this work although the Nazi government prevented him from receiving the prize at the time (he would later receive the Nobel in 1947). Work on Sulfa drugs continued during the war in other countries with the development of very effective drugs against pneumococcal infection (pneumonia).

George Dreyer, chairman of the school of pathology at Oxford had obtained a sample of Fleming's Penicillium culture after the 1929 paper to determine if penicillin was a bacteriophage on which he was an expert. He discontinued his research on finding penicillin was not a bacteriophage, but kept the culture nevertheless. A few years later Dreyer died and was replaced by Howard Florey who had earlier received reports of a successful penicillin broth treatment in his previous position as professor of pathology at Sheffield University. Florey assembled a team at Oxford enthusiastic about clinical research, and one of its members was Ernst Chain, a biochemist and fugitive from the Nazis. Through another remarkable series of coincidences and serendipity the Penicillium culture at Oxford caught the attention of Chain who suggested they investigate its antibacterial properties. Norman Heatley another team member found ways to measure the activity of penicillin and Chain was able to purify, crystallize, and stabilize it. The team then performed the critical experiment that had eluded many others before including Fleming. The team injected eight mice with lethal doses of streptococci, and then injected four of those with penicillin. The four mice that received the penicillin survived proving penicillin was an effective antibiotic. Further animal and then human trials (starting with the team themselves) followed during which the team identified effective dosage and necessary purity. However, large-scale production was still not possible when he was asked to treat a policeman dying of septicemia. They used what drug they had and the policeman improved, but they ran out before treatment was complete and the policeman died. Florey and the team focused on ways to increase production and subsequent trials were extremely successful. They sought American assistance in scaling up production providing details of their methods to the United States, which made significant improvements in penicillin production but incredibly refused to share these with the British as they were considered military secrets. Nevertheless, Britain was able to manufacture sufficient quantities of penicillin during the war using the Oxford team's methods. Fleming, Florey, and Chain received in 1945 Nobel Prize in medicine for their work on penicillin.

The success of penicillin inspired others to seek new antibiotics from other organisms in competition with bacteria including molds, fungi, and other bacteria. In 1944, Selman Waksman and his colleagues at Rutgers discovered a second antibiotic called streptomycin from Actinomyces bacteria. Waksman and his team discovered another species called Streptomyces griseus from which he isolated another antibiotic, streptomycin, which was especially important because it was effective against gram- negative bacteria and tuberculosis. Waksman was awarded the 1952 Nobel Prize in medicine for his discovery. In rapid succession Lederle laboratories announced Aureomycin in 1948 and Pfizer announced Terramycin in 1950, the first of a new class called tetracyclines. Parke-Davis developed chloramphenicol in 1949, which proved highly effective against typhoid fever. These antibiotics were seen as wonder drugs and used indiscriminately, yet very quickly the first cases of bacterial resistance were seen, and the first reported case of MRSA was given by (Barber 1961) in the United Kingdom. Sufficient modification potential existed in these classes of antibiotics to stay ahead of bacterial resistance but very few new classes were discovered following this golden age. The last of these "golden age" classes was discovered in 1962, the quinolones, and for 38 years no new classes were discovered (Walsh 2003). It was only until the oxazolidinones in 2000 and lipopeptides in 2003 that new types were introduced. And bacteria are already showing resistance to these new classes (Walsh 2003).

Why has there been a nearly 40-year gap in the discovery of new antibiotics? Some of the explanation is related to the economics of antibiotics which are not nearly as profitable as other drugs as described in (Christoffersen 2006) in addition to high effectiveness of "older" antibiotics until recent times. Now that there appears to be an incentive (growing markets due to growing resistance) there are still many regulatory hurdles that decrease the economic attractiveness of antibiotic R&D as reported in (Nathan 2004). But even without the economic and regulatory hurdles, the technical challenge of developing new antibiotics are enormous and highlight just how much of a gift the antibiotics discovered during the "golden age" were. Consider: an antibiotic must target only the pathogenic bacteria without damage to any human cells or other human flora (symbiotic bacteria). It must do so quickly and must be easily administered orally via stable compounds packaged in tablet form. Very few molecules can do this, partly because pathogenic bacteria are so similar to other useful bacteria but also carry many of the same cellular machinery as our own cells as described in (Walsh 2003). Enormous concerted effort will be required to develop new antibiotics, and an essential first step is to catalog all of the potential mutations and changes in the target microbes as part of a surveillance system, as there may be other gene targets in pathogenic microbes, such as virulence genes, that may allow for novel antibiotics as described in (Clatworthy, Pierson et al. 2007; Kristiansen, Hendricks et al. 2007; Cegelski, Marshall et al. 2008).

The devastating effects of bacterial infection as recently as 1941 have been forgotten, but the bacteria remain with us, continually working on new ways to defeat our weapons. The unimaginable return to the days of Florence Nightingale, Lister, and WWI field hospitals could return if we lose the battle against resistance. And we are starting to lose: the rates of healthcare associated bacterial infection have shown significant increases in the past 10 years, leading to increased costs and mortality as described in Chapter 1. But it is not just about developing new antibiotics, as significant process change will have to accompany any efforts to manage resistance. Changes in current practices of antibiotic overuse (Gonzales, Malone et al. 2001), improved hospital transmission control procedures, patient screening, and surveillance will be needed. The convoluted history of microbiology and antibiotics has many lessons to teach us, including the need for systems and processes to be able to change quickly and learn from observation. Even without antibiotics how many lives could have been saved by following Lister's and others ideas? Would we be in the precarious position we are now with few antibiotics left in the pipeline if we had actively "tested" the system to reveal resistance problems early on? The first patients with MRSA could have been isolated and the spread of this microbe controlled had there been a greater emphasis on testing and surveillance. Hindsight is of course very clear, but many of the problems encountered in microbiology and antibiotic development could have been lessened if more of the "process improvement" capabilities such as those described by (Spear 2005; Spear 2008) had been applied. This thesis provides some tools to apply these capabilities.

2.4 Clinical Microbiology

The focus of modern clinical microbiology labs is to support physicians in diagnosing and controlling patient infection. Patient samples are cultured, organisms isolated, and either tested for identity or effect of anti-microbial agents. Due in large part to the success of antibiotics, diagnostics are performed only when necessary and are not part of the routine healthcare of most individuals. Epidemiological surveillance is typically performed on an as needed basis, using whichever bacterial typing tools are available to the particular lab. Although some molecular based (i.e. reading a portion of the microbial genome) techniques are available, culture-based methods not much different from the Agar plates of Koch and Petri dishes dominate because they offer low cost and the ability to directly characterize the phenotype of the organism (i.e. which antibiotic to use which is the main concern of many clinicians). These culture-based methods essentially rely on the isolation of a microbe from a patient sample by a trained technician, followed by the growth of that isolate on petri dishes containing growth media but also a series of antibiotics in small disks on the surface of the petri dish. If the microbe does not grow around a particular disk, the microbe is susceptible to that antibiotic, but if they do and overrun it, they are resistant.

Cultures are also able to empirically determine the degree of resistance of a microbe by measuring the distance of growth inhibition around the disks, which should be proportional to the concentration gradient of the antibiotic, but these measurements require careful monitoring of growth conditions, measurement, and also take several days to complete. For the majority of the 20th century through today, hospital infection diagnostics have relied on these bacterial cultures. A sample is taken from a patient, incubated overnight, pathogenic bacteria are isolated, these bacteria are then incubated against antimicrobial agents, and the inhibition of growth is measured. Some automation has been introduced such as the Vitek system, and others, but the principles are largely the same and the cycle time is still measured in days, which can be problematic for critically ill patients such as those suffering from Septicemia. A sample non-automated workflow for the measurement of antibiotic resistance using the Bauer-Kirby disk diffusion method is shown in Figure 33 along with an image of the final product of this workflow, which is a petri dish with zones of inhibition around disks containing antibiotic and a ruler to measure the degree of inhibition as shown in Figure 34. These culture-based techniques, while tested and reliable, only provide phenotypic information and cannot identify any genotypic characteristics. A number of identification techniques are available to the clinical microbiology lab including methods based on cultures, restriction digests, Polymerase Chain Reactions (PCR), and partial sequencing (multi-locus sequence typing) but culture based methods dominate because of their low cost and integration with current practices (interpreting the other molecular-based methods is harder and not well established).

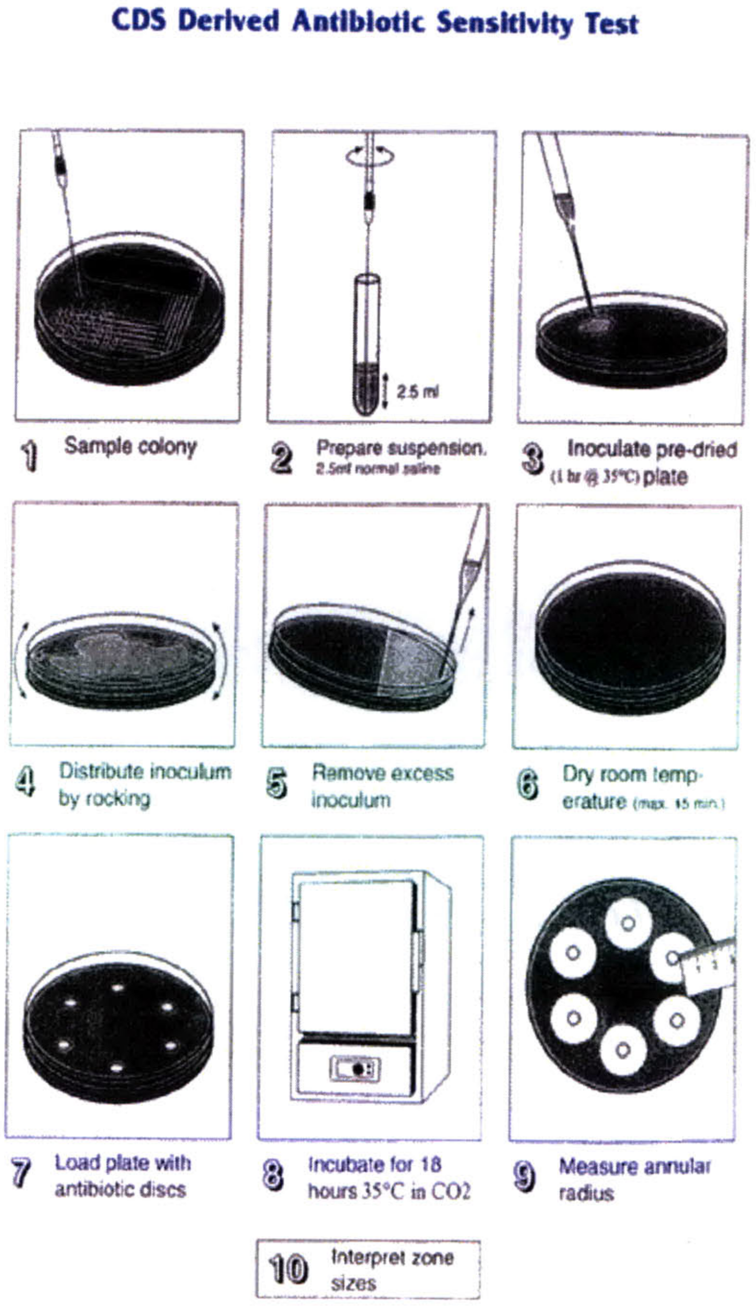

CDS Derived Antibiotic Sensitivity Test
2.5 ml
1 Sample colony
2 Prepare suspension. 2.5ml normal saline
3 Inoculate pre-dried (1 hr @ 35°C) plate
4 Distribute inoculum by rocking
5 Remove excess inoculum
6 Dry room temperature (max. 15 min.)
7 Load plate with antibiotic discs
8 Incubate for 18 hours 35°C in CO2
9 Measure annular radius
10 Interpret zone sizes

Figure 33 - Culture Based Antibiotic Sensitivity Workflow

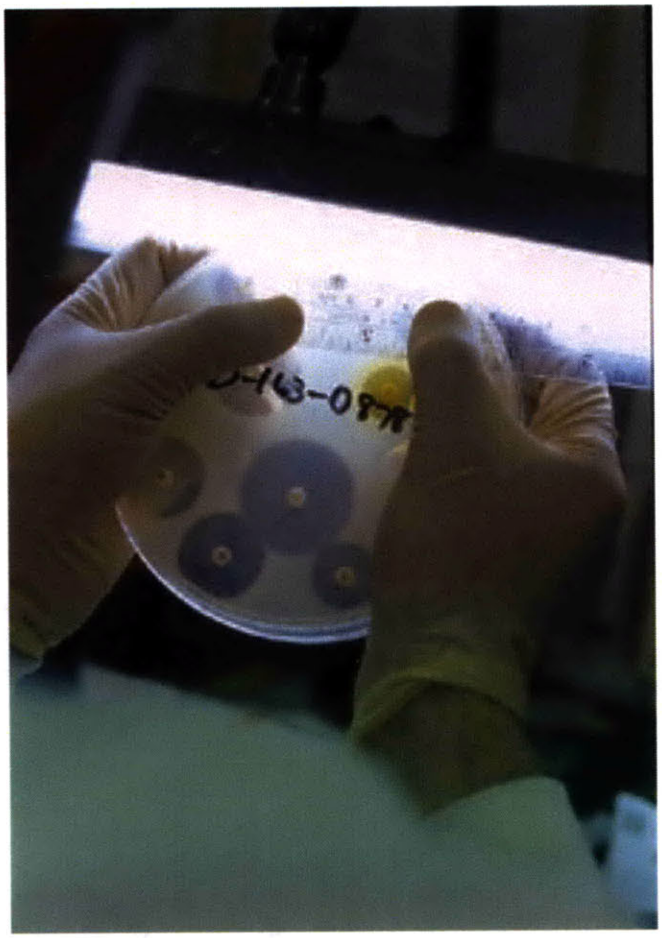


Figure 34 - Measurement of Antibiotic Disk Based Inhibitory Concentrations

This process was automated to a degree with the Vitek series of microbial diagnostic instruments originally developed for NASA by McDonnell Douglas during the space race, and eventually spun-off to Bio-Merieux in the 1980s. This instrument is largely the standard within hospital diagnostic laboratories, although traditional bacterial cultures are still performed using antibiotics disks on Petri dishes. The inhibition of growth around these antibiotic disks is then measured as the Minimum Inhibitory Concentration (MIC). This traditional procedure is labor intensive and subject to error due to the multi-step isolation of the bacteria required and the measurement error associated with manual measurement of the MIC. This traditional procedure also cannot identify or isolate organisms that may be nearing resistance but may are still inhibited by the antibiotic as shown by the different size regions in Figure 34. Nevertheless culture based methods have been the basis for recognizing the evolution of antibiotic-resistant bacteria starting in the 1970s and provide no information as to the genetic basis of resistance. More importantly, microbes, especially MRSA have a number of other adaptations beyond just resistance. Their ability to colonize skin and live longer in difficult environments allows them to spread more easily, and these virulence factors are thought to be critically important in the epidemiology of MRSA as discussed in (Chambers 2005; Diep and Otto 2008; Queck, Jameson-Lee et al. 2008). This has significant implications for clinical microbiology processes since only focusing on resistance (with culture based assays) will ignore any changes in virulence (genes encoding for improved adaptations to environment) that may allow certain strains of MRSA to spread faster, colonize patients, or simply adapt more quickly to antibiotic treatment. It is important to note however, that both genotype and phenotype of MRSA will be needed, at least until a comprehensive catalog of all possible mutations and changes exists (the so called "pan genome" describing all mutations in all strains of MRSA). Until then, techniques will have to coexist while this thesis emphasizes the need for complete whole genome sequencing as this information has previously not been available, many of the other existing technique can find application in patient screening and phenotype observations. The available phentoype and genotype methods in clinical microbiology are described in the following sections.

2.4.1 Microbial Phenotype

It is necessary to both characterize the genotype of the microbe, but also to match the genetic data with the phenotype of a microbe describing its physical properties: shape, cell wall, internal structures, metabolic pathways (nutrients, intermediates, and wastes), growth conditions (temperature, pH, time), and response to external agents (antibiotics, chemicals, and other bio-molecules). Microbes were first classified according to visible properties such as shape (rods – bacilli or spheres – cocci) and these properties are still useful classifications in a modern microbiology laboratory using similar microscopy techniques pioneered by Robert Koch over a century ago. Other physical classifications include various chemical stains that react with structures present in certain organisms only. The most famous of these staining methods is the Gram stain which groups organisms into gram-positives which turn violet from Methyl Violet 10B absorption and contain large amounts of peptidoglycan but no outer membrane, and gram-negatives which turn pink when exposed to the counter-stain safranin (which stains all cells) and have an outer membrane and less peptidoglycan. The majority of pathogenic bacteria are gram-negative, although there are notable exceptions: streptococcus and staphylococcus. These stains are clinically important as they provide fast assays not requiring significant preparation. Several other stains are available to modern clinical microbiology laboratories to assay patient samples, including some stains for MRSA (CHROMagar), which can provide relatively fast results (24h).

The next level of phenotype characterization involves metabolic pathways of the microbe including growth conditions, nutrients, and most importantly for clinical laboratories: antibiotic sensitivity. In routine clinical microbiology practice few tests are performed on metabolic pathways for clinical sample isolates, the focus being on antibiotic resistance, with most assays being the Bauer-Kirby disk diffusion type where microbial isolates are grown on solid media in the presence of disks containing antibiotics. This assay is usually read as a binary "yes/no" test to indicate the presence or absence of a microbe (i.e. MRSA) but can be used to estimate the degree of resistance to a particular antibiotic based on the diameter of the inhibition region around the antibiotic disks. Automated instruments are able to perform similar functions in liquid media to produce Minimum Inhibitory Concentration (MIC) curves, which can be useful in identifying both specific antibiotic regimens but also to identify borderline resistant organisms that may become threats. Using these methods a clinical lab can prepare an antibiogram showing the particular strain's resistance to various types of antibiotics.

Other phenotype assays are available to test various properties of microbes such as response to changes in nutrients, which activate secondary pathways, but because these are not clinically relevant they are not commonly performed. Implicit in all of these phenotype identification processes is the need to correctly isolate an organism and avoid contamination throughout the assay. The process of isolation also limits the ability to look for any other microbes that may be associated with the surveillance target, such as commensal bacteria, partially resistant microbes, and other pathogenic microbes. The primary clinical phenotype identification processes include:

- Antibiotic susceptibility – disk diffusion
- Antibiotic susceptibility – automated (Sensititre)
- Antibiogram – disk diffusion
- Antibiogram – VITEK
- Identification – disk diffusion
- Identification – automated liquid culture

2.4.2 Microbial Genotype

The availability of complete microbial genetic information for individual strains is relatively recent, dating from 1995 when the first bacterial genome, Haemophilus Influenzae, was fully sequenced (Fleischmann, Adams et al. 1995). Elements of bacterial genetics were well understood prior to that forming the basis for many biotech processes, but a complete genotype was only possible with the availability of fully sequenced genomes. Initial efforts to determine the genotype of microbes trace the evolution of molecular biology tools starting with restriction enzyme digests discovered in the 1970s on DNA extracted from cultured microbial isolates. These restriction digests cut microbial DNA at specific locations yielding a set of fragments read out on pulse field gel electrophoresis (PFGE) and are unique to particular species, although the resolution of this method is limited to large scale features on genomes and cannot distinguish nucleotide changes not affecting fragment length or genome size changes smaller than its resolution. This is particularly significant in distinguishing mutations in genes or recognizing the addition of genetic information from horizontal gene transfer (from another organism). In addition this method requires careful adherence to protocols as damage to the DNA in preparation can easily confound the results. The discovery of PCR in the mid 1980s enabled the identification of microbes based on PCR amplification of select regions in a microbial genome. These methods rely on either the amplification of a unique gene in a microbial genome, such as the mecA gene that confers methicillin resistance in MRSA, or the amplification of conserved but strain specific regions such as the 16s ribosomal RNA region. However, both of these PCR methods do not provide any information on other genes that may contribute to virulence or pathogenesis or provide accurate enough discrimination between strains.

An improvement on single locus PCR (i.e. mecA only) is multi-locus sequence typing (MLST) where as the name implies multiple regions in the genome are amplified via PCR to provide grater discrimination. The MLST approach is an intermediate to full genome sequencing largely driven by cost considerations as it is significantly cheaper only looking at a small fraction of the genome. Full genome sequencing was the first approach to provide complete genotype information, but was only possible with the advent of automated Sanger sequencing and capillary technology developed for the human genome project in the late 1990s when the sequencing of a single microbe could still cost nearly $1M. The rapid decrease in cost, faster than Moore's law, has enabled large scale sequencing using Sanger-based methods but also new highly efficient technologies including 454, Illumina, and others. The cost per microbial (3Mb) genome could approach less than $500 by the end of 2010. These whole genome sequencing approaches provide access to the entire genotype, eliminating many of the problems with partial genotype approaches such as PFGE, PCR, 16s, and MLST. The full sequence ensures mutations in all genes can be identified, especially ones associated with greater pathogenesis, identifies any genetic information transferred from other strains, species, or phages, and ensures that strains can be identified unambiguously. Nevertheless the rapid technology change in sequencing means many of the earlier approaches are in widespread use. Integration of new processes given an existing infrastructure is a problem and is a fundamental research objective of this thesis. The primary clinical genotype identification processes include:

- Restriction Digest PFGE
- PCR typing

The following are the available DNA sequencing methods for bacterial genotype:

- Sanger Sequencing (including MLST and 16s typing)
- 454 (recent)
- Illumina (recent)

Additional detail on the DNA sequencing methods is provided next:

Sanger Sequencing: The human genome project helped accelerate the development of sequencing technology to a rate surpassing the famed Moore's law in semiconductors. The current generation of sequencing technology began with Fred Sanger's invention of the chain termination sequencing method, for which he won his 2nd Nobel Prize in chemistry. This method relies on the construction of a template fragment with a specific primer sequence, followed by chain elongation and termination at a labeled nucleotide. This cycle is then repeated until labeled nucleotides exist at the end of chains matching every position on the template DNA. These fragments are then separated by size and the label of each chain terminator is read as the original DNA sequence. The industrialization of this technique and the supporting infrastructure around it, is largely what enabled the human genome project. Progressing from the manual early versions of this process, handled by high level scientists (Fred Sanger sequenced the phage f-X174 genome of 5,386 bases over a 2 year span from 1975 to 1977). The technique was gradually improved over the next 20 years, and by 1997 it was possible to imagine reading the entire 3Gb human genome with the development of automated fluorescent detection based capillary sequencing invented by Lee Hood, et. al. Yet, a limitation of this method is that the chain termination constrains the length of DNA that can be read in a single reaction to ~700 bases and reactions must be performed individually at significant expense.

454 Sequencing: Automated gel and capillary sequencing of fluorescent Sanger reaction products dominated the period from 1986 to 2006, but in 2005, 454 Life Sciences introduced the first of the next generation instruments. This technology relied on an entirely different technology, pyro-sequencing, which triggers emission of light through an enzyme cascade as each nucleotide is incorporated. Each burst of light is read out through a fiber optic plate as each type of nucleotide is introduced successively. A significant improvement of all the next generation technologies, including 454, is that the reactions occur in massively parallel 2-dimensional plates rather than "1-dimensional" gels or capillaries. The transition to 2-D means that efficiency increases are no longer a linear function of the number of capillaries or length of capillaries, but rather the square of the edge length of the surface. Doubling the dimensions of the plate quadruples the surface area and the amount of data produced. At its introduction, this technology approached the efficiency of the highly refined capillary methods, but its performance was not yet sufficient to displace the incumbent technology. Yet in the period 2005-2008 this technology improved sufficiently to represent an order of magnitude improvement over the existing technology following technology displacement curves similar to (Utterback 1996). Capillary technology vendors have tried to improve the performance of their systems to avoid displacement, but have proven unable to do so. Most capillary systems are being displaced and moved into lower performance applications in hospitals outside of high performance genome centers. 454 Workflow is greatly simplified over the traditional capillary sequencing workflow. Beads containing a single segment of the target DNA, but PCR amplified (all copies of DNA on the bead are the same) are deposited onto wells etched from a bundle of fiber optic fibers. The cycle time is also much shorter than traditional (capillary) sequencing methods (2 days vs. 2 weeks under optimum conditions for single samples). The plates on which the sequencing occurs are made from fusing a fiber optic bundle together, slicing it into plates, and then etching wells from the core of each fiber optic strand. DNA containing beads and enzyme beads are then deposited in each well. An enzymatic cascade then triggers light emission with the incorporation of each nucleotide on the DNA labeled beads. The 454 method has numerous advantages over traditional capillary sequencing including cost, speed, and quality of data. Because the 454 method is an entirely in-vitro method, none of the bacterial transformation steps leading to potential bias are required. This enables the 454 method to accurately represent DNA sequences that may not be present in traditional bacterial cloning and capillary sequencing methods.

The 454 method is also the most suitable of the next generation methods for in-hospital infection diagnosis, due to its speed, it is the only method that could generate data in the same day.

Illumina Sequencing: Following the introduction of the 454 process, another next generation company, Solexa introduced a revolutionary improvement to the Sanger fluorescent chemistry. Rather than separate fragments on gels or in capillaries, this technology produces clusters of identical DNA molecules on glass slides via PCR, then uses a variant of the Sanger chemistry to sequence each cluster. The input DNA is fragmented and adapters are added to bind the DNA onto glass slides and allow them to be amplified and then sequenced. This technology has significantly greater throughput per run than 454, and relies on well understood Sanger sequencing where each base receives a unique color (dye) making identification easier. However, the cluster amplification has some limitations in terms of the cluster density that can be achieved. Also, the cycle time to produce data is longer than 454, up to 12 days per run. However, the data output is much greater, upwards of 30Gb per run and will likely be higher in the near future.

2.5 DNA Sequencing Rate of Change

The historical cost data through 2009 and projection for 2010 of the cost of DNA sequencing using the best available technology in each year are shown in Figure 35 which plots the cost to produce 1 Megabase of DNA sequence. A Moore's law curve is plotted for reference starting with the cost per genome in 1999. A bacterial genome is on the order of (3-5Mb) suggesting that the cost to fully sequence a microbe to 30X depth (a standard level of redundancy needed for the assembly algorithms would have cost ~$1.5 Million in 2000 versus less than $30 as projected at the end of 2010. However, this is the cost for just the raw sequence without any of the additional sample acquisition, algorithm processing, assembly, and annotation, which can add significant additional cost. Nevertheless the cost of sequencing is rapidly approaching the cost of routine clinical microbiology procedures, suggesting a paradigm shift is possible where sequencing may rapidly displace more traditional approaches. This graph served as significant motivation for this thesis as it suggest the costs of whole genome sequencing appeared to rapidly be reaching the point that routine surveillance of microbes could be performed.

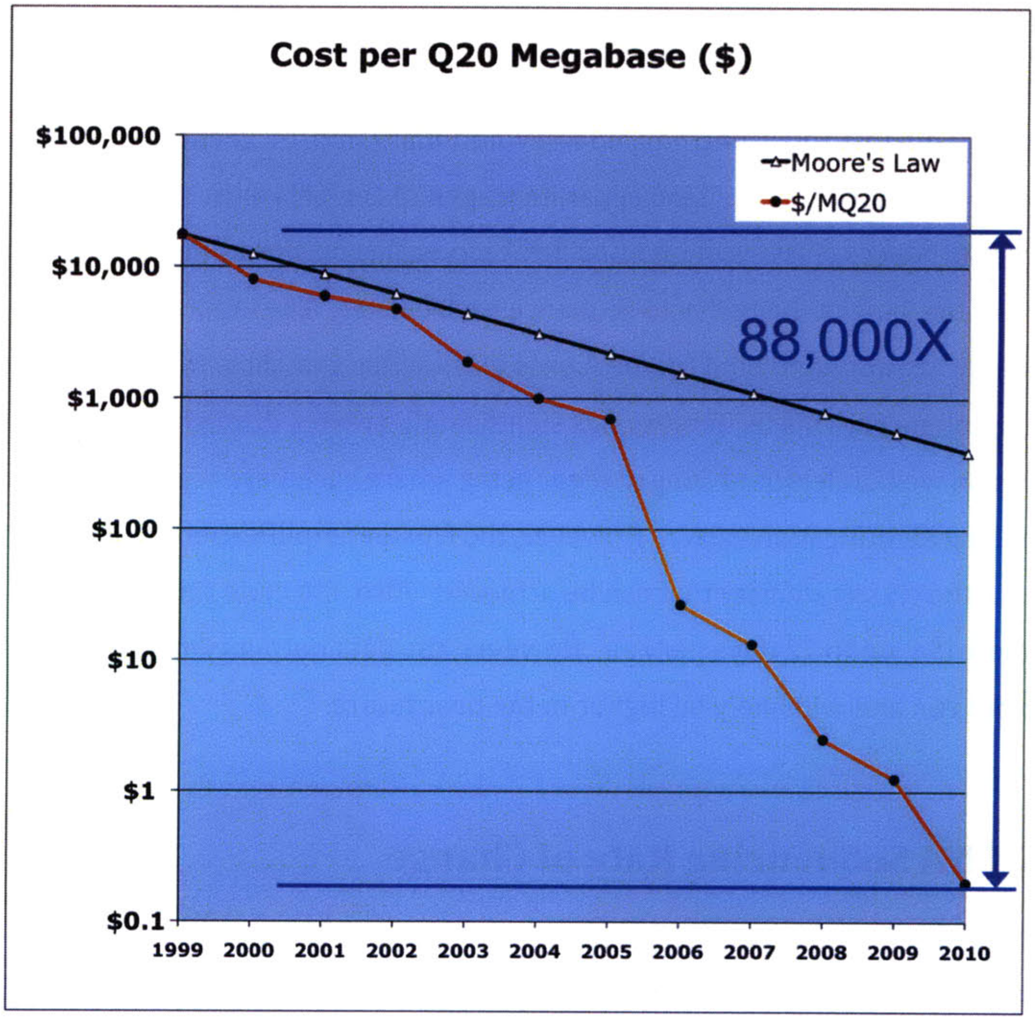


Figure 35 - Historical and Projected (2010) Cost per Megabase of DNA Sequence Compared to What Moore's Law Would Have Been Starting in 1999. There has been an 88,000 fold decrease since 1999.

2.6 MRSA Surveillance

Most of the MRSA surveillance efforts to date have focused on documenting cases, associating culture based (phenotype) data, or in certain efforts to characterize MRSA with PFGE or MLST methods (Klevens, Morrison et al. 2007) which have shown some gross differences, but have not had the resolution to find nucleotide level differences. Recent studies using a related DNA probe technology have shown significant differences although not with sufficient nucleotide level resolution (Kennedy, Otto et al. 2008). Fortunately, some of these efforts have yielded clone collections and samples that could be re-grown and used for sequencing efforts. The CDC surveillance efforts have focused on reporting infections associated with healthcare procedures as part of the National Nosocomial Infections Surveillance System (NNIS), which is a voluntary reporting system in ~300 hospitals in 42 states. NNIS focuses on high-risk units (intensive care) collecting data on infection rates and certain (voluntary) follow-up as to the organisms and degree of antibiotic resistance. However, no data on genotype is collected to identify differences in strains or other characteristics influencing changes in phenotype. NNIS is being replaced by an expanded system called the National Healthcare Safety Network (NHSN), which expands in scope to include electronic record submission but still focuses on infection rate and phenotype reporting to the exclusion of genotype as described in (Tokars, Richards et al. 2004). However, neither NNIS nor the new NHSN are surveillance systems of the type discussed in this thesis, as they do not have a sample collection or genotype collection procedures. The closest national CDC system is called the Active Bacterial Core surveillance (ABCs) part of the CDC Emerging Infections Program Network (EIP)[6]. The ABCs is a collaboration between CDC, state health departments, and universities in 10 states charged with determining the incidence of MRSA (and other select pathogens) and to "determine molecular epidemiologic patterns and microbiologic characteristics of public health relevance for isolates causing the above invasive infections". The ABCs do collect some MRSA isolates, as a sampling and these are sent to the Network on Antimicrobial Resistance in Staphylococcus aureus (NARSA), which distributes the isolates. The ABCs system produces data on the surveillance elements for MRSA described in Table 5:

CDC-ABCs MRSA Strain Characterization	**Notes**
Confirmatory identification of all isolates	Culture-based to determine if isolate grows in the presence of antibiotic.
Antimicrobial susceptibility testing,	Quantifies the degree of resistance via Bauer-

[6] Source: Centers for Disease Control, www.cdc.gov

selected isolates	Kirby tests (culture based)
Pulsed field gel electrophoresis and analysis using PulseNet methodology, selected isolates	PFGE based method with associated limitations in resolution and reproducibility
SCC*mec* cassette typing, selected isolates	PCR based assay looking for genes associated with methicillin resistance
Toxin testing for PVL and TSST-1, selected isolates	PCR based assay looking for gene encoding PVL

Table 5 - CDC MRSA Surveillance Elements

While valuable, the limited genotype information collected as part of the ABCs cannot distinguish variation within a strain but can identify gross differences between some strains (but not all). Yet, the techniques described in the ABCs require significant skill on the part of technicians, as PFGE is an intricate protocol despite its limitations suggesting that more complex processes such as DNA sequencing could be inserted into the ABCs workflow. The literature does not yet contain a reference describing whole genome based microbial surveillance, but some literature exists on "partial sequencing approaches" where a locus on the MRSA genome such as the S. aureus protein A gene (spa) might be chosen for sequencing and serve as a "fingerprint" for that particular strain as described in (Mellmann, Friedrich et al. 2006). However, all of the surveillance approaches to date suffer from a key flaw, from the lack of resolution of PFGE to the inability of PCR or "partial sequencing" approaches to find novel genes or mutations elsewhere on the genome present significant limitations for the use of this surveillance data to help develop new antibiotics or help understand MRSA biology. DNA sequencing is completely backwards compatible to all these approaches as the PFGE, PCR, and "partial sequencing" results can be generated on a computer from the complete genome. Because this produces a universal solution, whole genome DNA sequencing is the process pursued in this thesis, which could generate a "step change" in clinical microbiology.

2.7 MRSA

Methicillin resistant Staphylococcus aureus was first identified as a new strain in 1961 in the United Kingdom (Barber 1961). However, it did not appear in large numbers of cases in the US until the 1980s associated with intravenous drug users. The long time period between initial appearance and the growing epidemic suggests additional factors contributed to this, such as virulence genes among others. The genome of MRSA is relatively small, only 2.8Mb in size containing approximately 2,600 protein coding sequences (Lindsay and Holden 2004) with about 75% being highly conserved between isolates associated with metabolism and other housekeeping functions. The remainder and additional plasmids outside the core genome show much greater variability and also carry the majority of virulence genes identified to date including leukocidins such as PVL. The cell walls of MRSA are especially thick in some cases as part of its defense for antibiotics that bind cell wall proteins and peptides, which makes their lysis significantly more difficult than other species.

In terms of MRSA surveillance using DNA sequencing the literature is much more limited. Most of the approaches detail the sequencing of full genomes as a single research effort rather than focusing on building a process for surveillance. All of the reported fully sequenced MRSA genomes which number less than 20, are different from each other, and nearly every study using the less detailed but still useful PFGE or MLST techniques also show significant variation. The emphasis on building a process rather than producing a genome for research is also not found in the literature in exactly that focus.

The next chapter will describe in greater detail the matrix based MDPM model, and much of the specific case study of MRSA surveillance will be used to highlight its use.

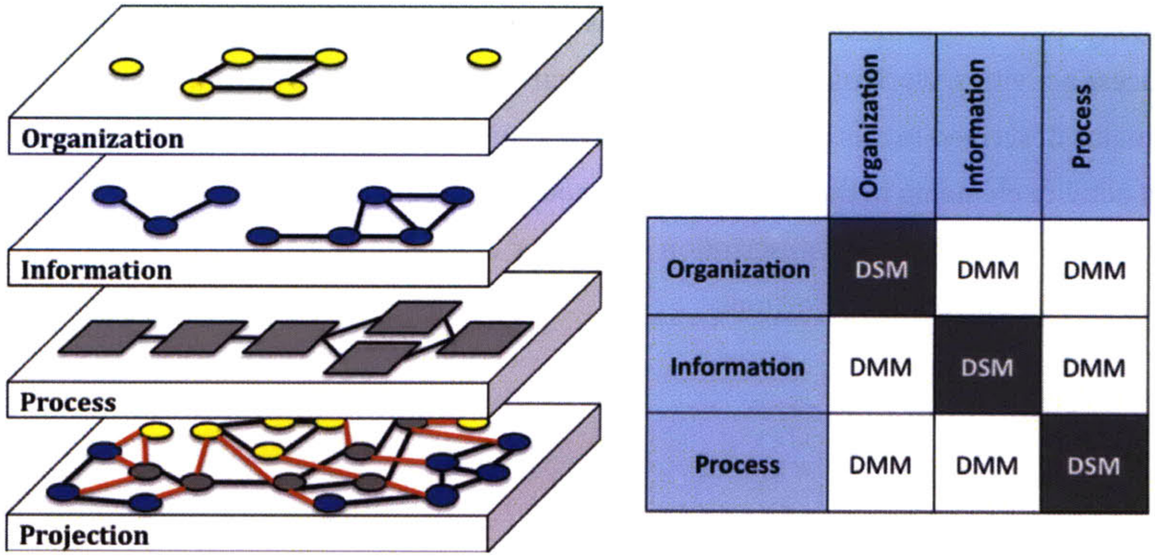
3 Multi-Domain Process Matrix Model (MDPM)

The goal of a process model is to represent reality in an analytically useful way according to the needs of the user recognizing that any model is an abstraction and simplification. A chemical engineer requires a process model that represents the physical flow and changes of reaction products as in a process flow diagram. A business consultant requires a process model that shows the interactions between tasks and individuals as in a swimlane process diagram. A software designer requires a process model that describes information flows from various modules in a computer program as in a control flow diagram. Each of these models is useful to their users, providing analytical value in a particular analysis dimension. However, none of these by themselves addresses the system level view needed to manage complex processes that these processes operate in. A process owner responsible for the physical, organizational, and information flow control of a process needs access to all of these domains and especially the interaction between them. The goal of the Multi-Domain Process Matrix model (MDPM) presented here is to provide a synthesis of these domains for system level process analysis, while preserving the ability to individually analyze each domain. This is accomplished by developing individual set representations for the activities (process steps), organization (stakeholders), and the necessary information (work instructions) then combining these onto a single analysis domain (matrix), which shows the connections between them. This model is fully described in this chapter.

3.1 Multi-Domain Process Matrix Model Description

The multi-domain process model used in this thesis expands on traditional process definitions and models such as flowcharts and other activity focused models described in chapter 2. The model separates processes into 4 distinct domains: information (work instructions), process (tasks), organizational (social network), and an integrative projection domain where all elements and interactions are described together. Each of these domains can be thought of as a graph with a corresponding matrix representation.

The goal of this decomposition is to capture the true range of activity, information, and organizational dimensions for the process. The integrative projection domain is the overall matrix describing the process, information, and organization domains and their interactions to represent the entire process in a single connected network that can be analyzed using many of the traditional network algorithms. This is possible because each of the other three domains in the model, process, information, and organizational have network representations. Processes have been modeled as networks in petri nets, workflow state diagrams, and in DSM methods as described in chapter 2. Information flows have been modeled as networks in software design, concept mapping, and communications networks as described in chapter 2 and especially in (Giffin, de Weck et al. 2009). Organizations have also been modeled as networks in sociology and the emerging field of social network analysis. These three domains can then be jointly analyzed in the overall matrix, which is a single "projection layer" that represents the complete process network as conceptually shown in the Figure 36. Each domain of the model is essentially a DSM, representing each of the domains (layers), and the overall model is a multi-domain matrix (MDM) model showing the links between elements in each of the domains.



	Organization	Information	Process
Organization	DSM	DMM	DMM
Information	DMM	DSM	DMM
Process	DMM	DMM	DSM

Figure 36 - MDPM Model

In the MDPM model, elements in each of the domains can have interactions with other elements in the same domain, or with elements in other domains. However, each domain can be analyzed independently to facilitate observation, analysis, and interpretation. The projection domain is an aggregation of each of the domains above it, but also of the interactions across domains shown conceptually in red. In MDM model terms, the projection domain is the entire matrix showing all connections simultaneously. The projection domain is "special" in that all the elements from each domain are treated as if they were the same, such that they can be analyzed as a single network. This abstraction is important in applying many of the network algorithms and metrics to quantify overall search, complexity, and connectedness of the process (and network). A key advantage of this approach is to show that individual elements, which may appear disconnected on a given domain, may in fact be connected to each other through elements in another domain. In complex modern processes many of the individuals performing the process are likely not directly connected to each other through direct social connections but through work instructions, information, and process steps. Likewise, the tacit knowledge of an individual operator may not readily flow into work instructions in the information domain or to process designers and other stakeholders appearing and the network representation as an isolated node. The projection domain is thus a useful abstraction that can represent the process activities, work instructions, and the various stakeholders all as networks. The domains described in this process model have a correspondence to the control, mechanism, and activity elements in the IDEF0 process modeling methodology as described in Figure 37 and the IDEF0 language description in Chapter 2. The individual domains are further described in the following sections.

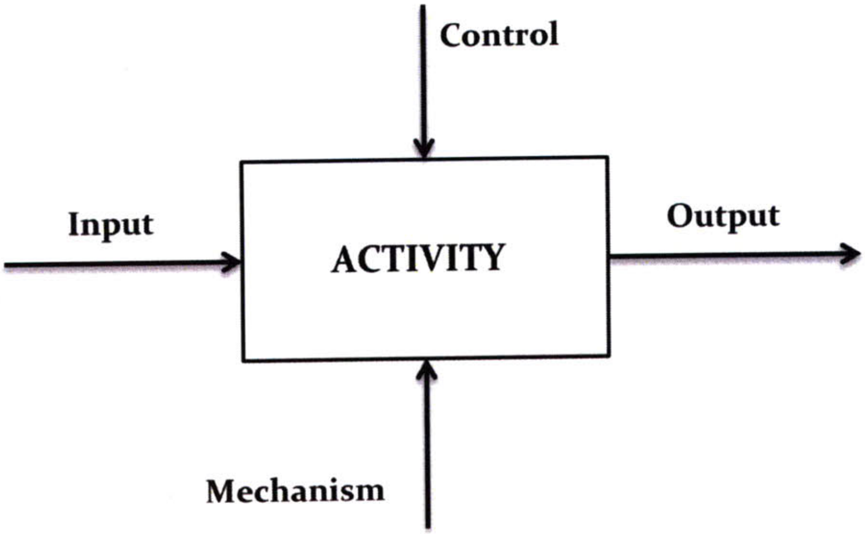


Figure 37 - IDEF0 Element

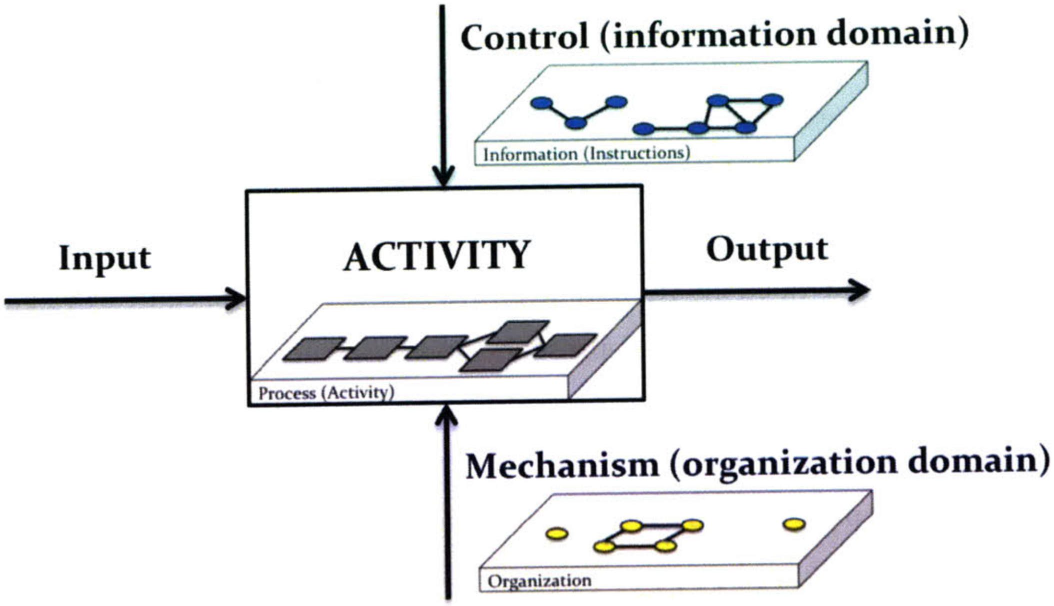

Control (information domain)
Information (Instructions)
Input
ACTIVITY
Output
Process (Activity)
Mechanism (organization domain)
Organization

Figure 38 - IDEF0 and MDPM Correspondence

The explicit modeling of the process work instructions associated with process activities is a key element of the MDPM model which does not yet have a corresponding category in the Engineering Systems Model currently under development by Bartolomei and others. Explicitly adding these work instructions and associated information will be a necessary addition to extend the application of the ESM model to process including manufacturing. The MDPM model can nevertheless be thought of as a subset of the ESM model and is completely compatible with it. Some ESM model elements not considered in this thesis to limit its scope, such as the object domain (physical elements of a system) could readily be added, and would bring additional model richness as well as provide additional troubleshooting and search capabilities to process stakeholders. The key differences in the MDPM modeling approach are an emphasis on activities as a distinct element and the connection of other domains around the activity elements. IDEF0 does not connect organization and information elements to each other. And other multi-domain approaches do not specifically emphasize processes.

3.2 Process Domain

As described in Chapter 2, a number of different process modeling methods exist. Their general goal is to represent the sequence of activities needed to perform a process, however their particular representations differ. The process domain is similar to traditional flowchart workflow representations, describing the process in terms of a series of activities with well-defined inputs, outputs, work instructions, and control systems. IDEF0 activity notation conveniently represents the process activities in this model, as only the activity, inputs, and outputs are described in the process domain. The IDEF0 control (information) and mechanism (organization) inputs are described in other domains. The particular process representation used as part of the model is "activity on node" or "AoN" where each node of a directed graph represents an activity from a traditional process flow diagram. The edges in the graph represent the inputs and outputs from each activity and can be mapped onto a separate layer if desired, where the mapping would be "activity on edge" and the nodes represent the inputs and outputs. To facilitate the integration and analysis with the information and organization domains, AoN provides a more useful representation. A notional process is shown below in flowchart form, showing the request of an antibiotic diagnostic. This process has only one branch point, and can be straightforwardly represented by the AoN directed graph shown in Figure 39.

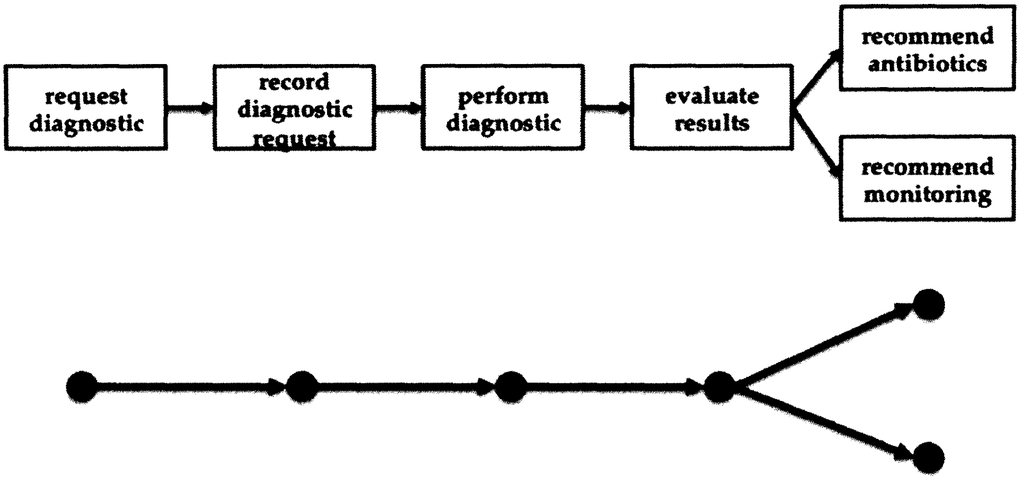


Figure 39 - Activity on Node Process Model

Real processes often have rework, where a particular activity may fail and action taken to perform it again. This is shown conceptually in the flow chart in Figure 40 adding a rework loop to the baseline diagnostic request process. This is represented on the AoN directed graph as a loop. Rework is not the only type of loop structure; others such as iteration during design refinement and optimization have similar structures. The additional graph structures associated with design and optimization is referred to in this thesis as "scaffolding" necessary during the development of processes and can add significant complexity. These ideas are discussed later in this chapter.

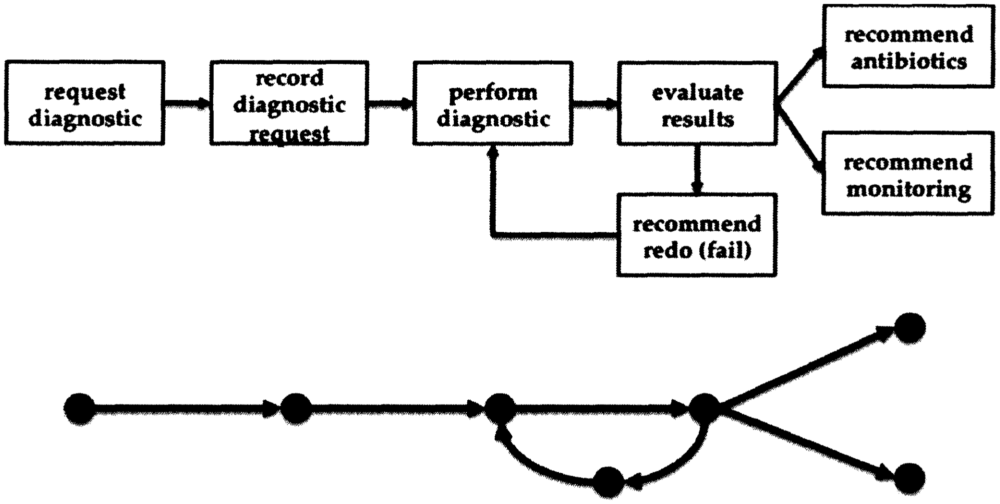


Figure 40 - Activity on Node Graph with Rework

Processes can also have significant operational choice where multiple steps are possible and traces must be made between them. For example, the conceptual process flowchart shown below adds three choices to the type of diagnostic to be performed. Implicit in this is the existence of a set of decision rules associated with the choice step, which in the MDPM model are explicitly represented in the information domain. Also implicit is the existence of an individual in the organization responsible for this decision, and in the MDPM model this is explicitly represented in the organization domain. The AoN directed graph is in Figure 41 and shows the additional graph structure associated with the three diagnostic choices.

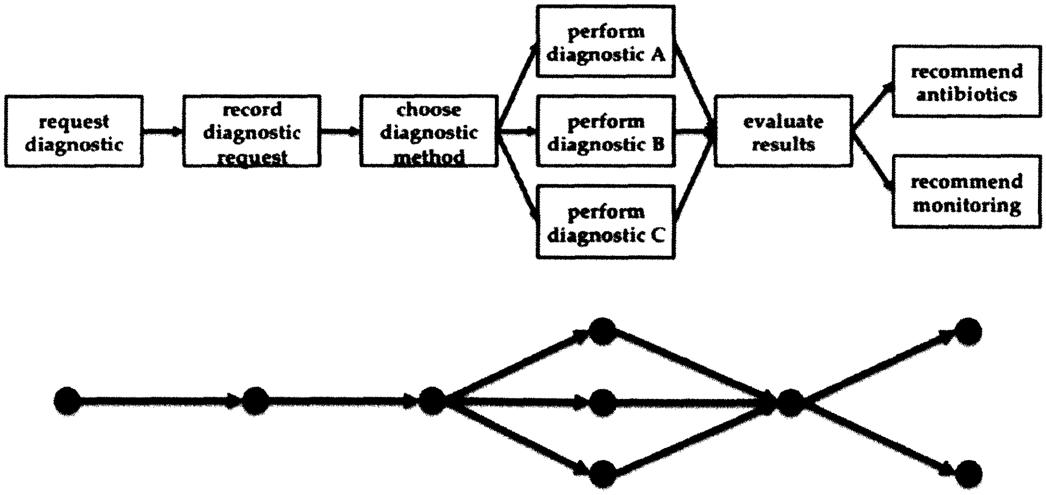


Figure 41 - Activity on Node Graph with Branching

Choice and rework are shown together in the following conceptual flowchart indicating the traditional representation. Certain workflow modeling techniques explicitly label decision points, such as connectors in EPC notations. For simplicity the use of these additional elements is avoided in this model, although branch points can be readily found in directed graphs. The branch points can be useful in assessing the "decision complexity" of a process, as a large number of branch points should correspond to high process complexity. The AoN graph representing the combined choice and rework process is shown in Figure 42.

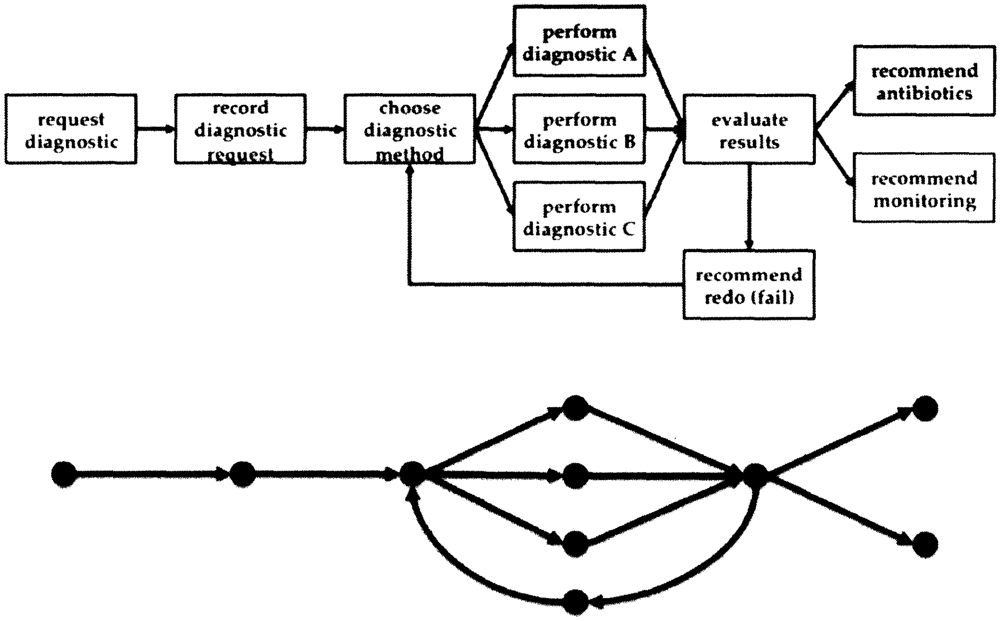


Figure 42 - Activity on Node Graph with Rework and Branching

Using the directed graph, all of the basic network properties can be calculated including node degree distribution, number of edges, average path length, diameter, and others. The basic network properties for the MDPM model will be described later in this chapter as well as the proposed complexity measures. Chapter 4 will describe in detail the application of the MDPM model to the real world process of DNA sequencing-based MRSA surveillance and a portion of that process is shown in Figure 43 in swim lane flow chart form. This particular flowchart was developed using a standard commercial tool, called iGrafx, for business process representation and modeling. The notation is the same as in the conceptual flowchart models shown in this section and the translation to a directed graph follows the same procedure of activity on node, AoN as will be shown in chapter 4.

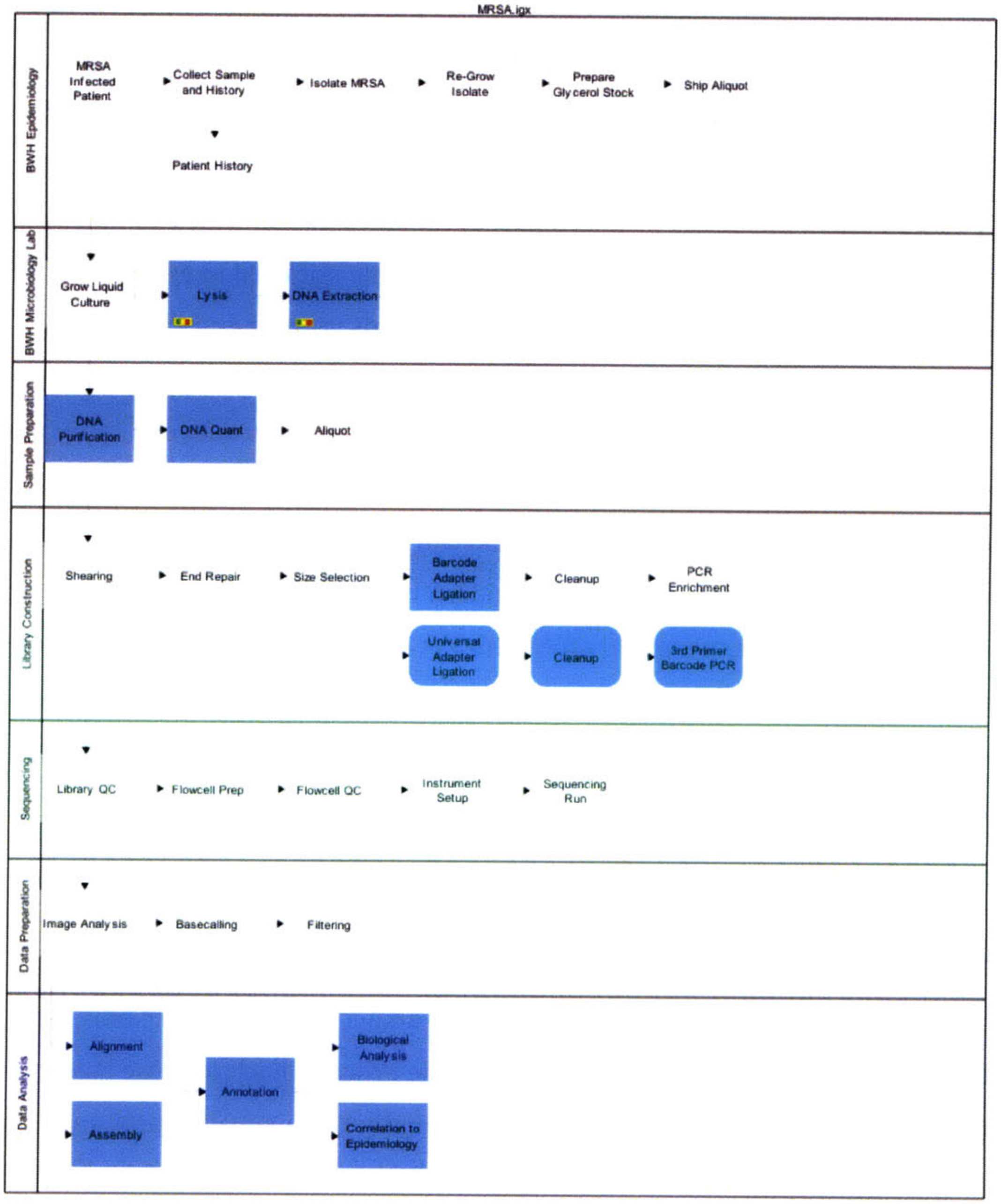


Figure 43 - Conceptual MRSA Surveillance Process in Flowchart Form

Some important considerations for the process layer relate to its scope, or "what to model?" In order to capture the true complexity of processes, it is important to enumerate all of the possible process paths within the system. So for example rarely used but feasible paths should be modeled since associated with these paths are a set of work instructions and individuals needed to "maintain" these paths. Indeed in quality control or safety, it is the least used paths that are most in need of definition and maintenance to ensure they will be available when needed. A manufacturing analogy is the need to maintain production lines for older products that may even be obsolete but still require parts for products still used by customers in the field. For microbial surveillance, the specific example is the need to maintain specialty or "reference" processes that while not used in everyday practice are occasionally used for difficult or "calibration" cases. This is another example of the "hidden" complexity of processes that basic flowcharts cannot capture in their effort to present a simplified version of a process. Similarly when there is not yet a "dominant" design and multiple viable alternatives exist it is important to show these as they represent the true complexity of the process. The existence of viable alternatives for steps in a process requires that someone in the organization whether it is users, designers, or the suppliers themselves maintain a corresponding set of work instructions and the processes themselves. In this model much of the complexity associated with the "scaffolding" early in the process lifecycle comes from the existence of these design alternatives and their supporting information and organizational networks.

3.3 Organization Domain

The organizational domain follows social network analysis models describing the interaction between the various stakeholders of the process. Each of these stakeholders can interact with multiple process elements, information domain elements, and with each other, such as communication between stakeholders, control of an individual process step, design responsibility for a process step, or as a source of information for one of the elements in the information domain. The organization domain is intended to represent individuals but may also represent organizations and other group entities. The goal of the organization domain is to show the rich and deep social network associated with every process but not normally shown in traditional process representations. The real social networks represented by this layer can have a very significant impact on the actual design, capabilities, and operational performance of a process and should therefore be explicitly shown as part of a process representation. In health care, organizational complexity can be particularly severe.

The diagram in Figure 44 again shows the conceptual example of a diagnostic request process but adds the individuals associated with each of the activities. Both the interactions with the activities and interactions between the individuals are shown. The IDEF0 notation uses the term "mechanism" to indicate the means by which the activity happens, and in this example the activity requires the individuals. It is possible to imagine that an additional domain might be included in the model to describe the infrastructure or equipment that represents the "mechanism" such as the “Objects” domain in the ESM, but that is not described here to limit the scope of the model. The interactions between the individuals and the process activities are shown here and will be carried over to the "projection" domain.

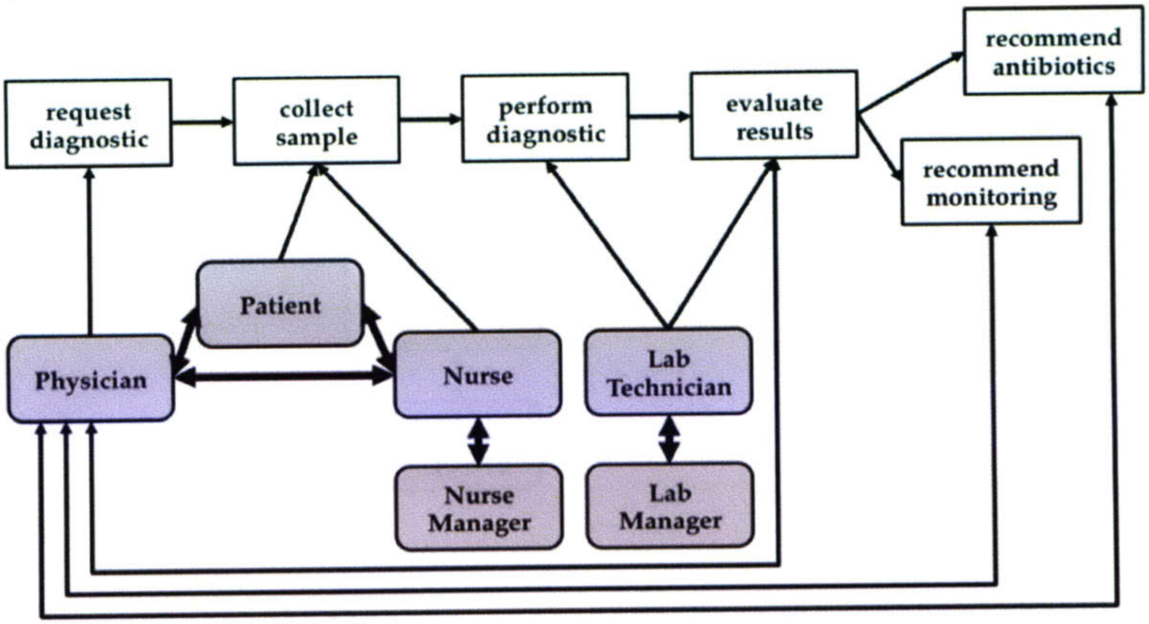


Figure 44 - Interaction of Process and Organization Domains

However, the interactions between individuals can be modeled separately in this layer as shown in Figure 45. These interactions can be modeled as a network where the nodes are the individuals and the edges represent communication between them. Note that not all the graph nodes are connected to each other, and although all of these individuals form part of a single process to lack of communication between them in this example could present problems in managing this process. For example, what if there was a significant quality issue with the diagnostic leading to a higher fraction of false negatives? Unless the technician reports the problem to the lab manager, who in turn has to take corrective action either by reporting the problem to the physician or explicitly adjusting the diagnostic confidence values the problem may go unnoticed. The physician may then recommend monitoring rather than antibiotic therapy with potentially serious consequences. This is an interesting observation as to the properties of social networks supporting processes: it's likely very important that all the individuals be connected and be reachable in relatively few steps. This becomes more important as the complexity of the process increases measured in either the number of activities or amount of information that needs to be processed. Small world networks may be characteristic of stable processes since changes and problems at various process activities impact all other downstream activities and the faster information can be exchanged the better.

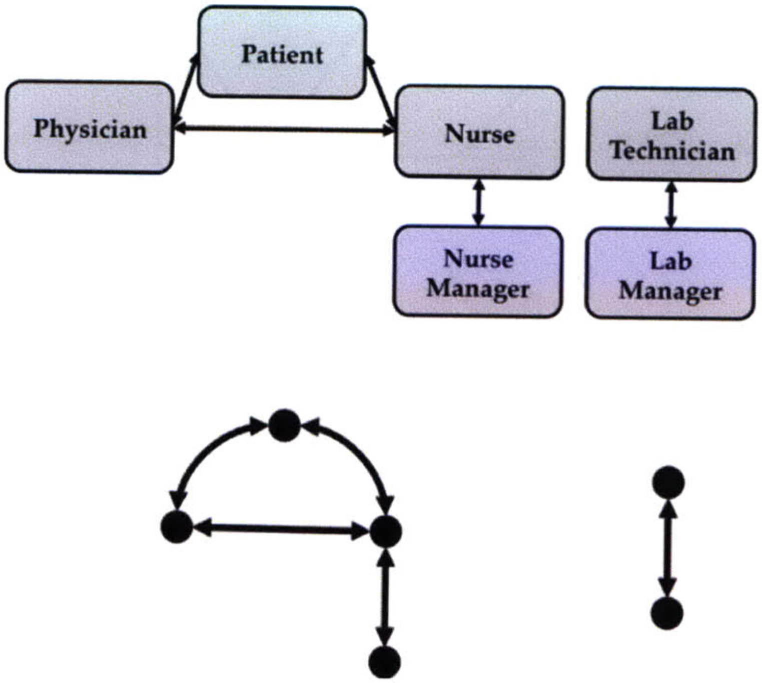


Figure 45 - Organization Domain as a Social Network Graph

Thus for example, a "process owner" could be designated to provide a connection between the lab technician and lab manager to the physician and nurse manager. This addition is shown in Figure 46, connecting the graph. And although a person has been added to connect the graph, an existing individual could just as easily have been designated to serve as the "process owner", such as the lab manager. This provides for a connected set that also means that every individual including the patient is reachable in the social network. These types of "design changes" may not be readily apparent from traditional flowchart process views that are critical to the implementation of change and troubleshooting in processes. While adding an individual such as the process owner to the social network will increase certain complexity measures such as number of nodes and edges it is likely to significantly reduce others such as connectivity, average path length, and diameter. The various complexity measures will be further discussed later in this chapter.

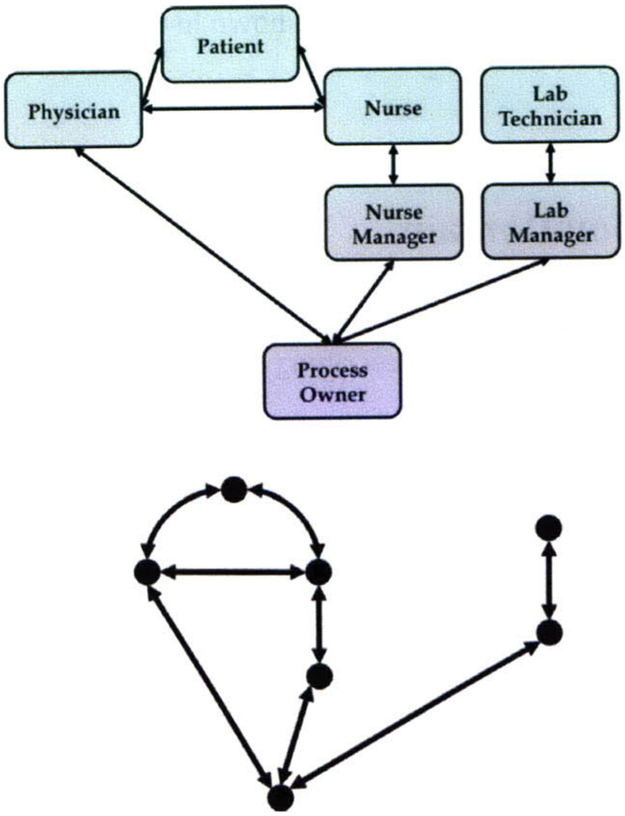


Figure 46 - Organization Domain with Addition of a Process Owner

This simple example does not show the additional organizational complexity found in real processes. However, one can imagine that the same framework can be used to represent the many additional stakeholders present in healthcare processes. For example, the organization domain for a healthcare diagnostic process should include representatives from hospital insurance reimbursement, senior hospital medical staff, hospital administrators, insurers, medical policy groups such as the AMA, representatives from the diagnostic manufacturers, government regulatory agencies such as the FDA, and others. The vast complexity of the social network associated with real healthcare processes is not often modeled but contributes directly to the performance and adaptability of these processes. These social networks are highly fragmented and distributed throughout lengthy processes (when viewed across their entire value and supply chains). Complexity reduction measures such as adding "process owners" as in the simple example to improve the search characteristics across the network by reducing the path length may prove valuable. A real example of this complexity is shown in the test case for the process model in this thesis.

3.4 Information Domain

The information domain describes all of the explicit knowledge needed to perform the process. It can also be interpreted as describing the data flow needed to manage the process in terms of all the work instructions, process data (i.e. metrics), design documents, operational observation logs, policy documents, regulations (ASME), physical models (i.e. steam tables), six sigma optimization results, vendor performance specifications, and any other sources of information needed. Unit operations in chemical engineering flow diagrams are information domain models of the physical properties of chemicals as they undergo certain transformations along with work instructions describing the particular set of parameters for the transformation. In a service business, the work instructions to process a loan can also be abstracted onto the information domain. The interactions between stakeholders of the organization domain can also be abstracted onto the information domain as is done for social network analysis describing individual stakeholders as nodes and various types of information exchange as different links between the nodes. The information domain can also be used to describe the complexity faced by process designers in considering the many alternatives available for each activity and the design effort associated with making choices amongst these alternatives to produce the final work instructions for activities in the process domain.

An important distinction of the information domain is that tacit knowledge, such as that held in a technician's head about he best way to hold a pipette (but not explicitly stated in the protocol) resides in the organization domain unless a specific mechanism for knowledge capture exists. So, for example if there is a formal process review where technician experience is to be captured into a design document, this design document resides in the information domain. Otherwise, every activity performed in the process domain is a combination of explicit information (such as a protocol) and tacit knowledge from a technician, physician, or other specialist whose entire tacit knowledge base cannot be easily transferred to explicit information. Every activity in the process domain thus requires a connection to at least one information domain item and one individual in the organization domain. This is the same as the IDEF0 definition where each activity requires both a control input and a mechanism input. Again, most traditional process flowchart representations hide the complexity associated with imperfect information, assuming that each activity (task) is fully specified and known as was described in chapter 2. In reality, the capture of this tacit information is critical to all process improvement methodologies such as lean and six Sigma, and in most process representations it is not explicitly identified or modeled.

The diagram in Figure 47 shows the now familiar process flowchart but adds the necessary work instructions associated with each of the process activities. Note that every activity must have a corresponding work instruction, but each work instruction can be derived from multiple explicit knowledge sources. A key goal of the information domain is to map the genealogy of the work instructions needed to perform the process. This is critically important in fast-changing technologically complex processes because work instructions should reflect changes in the policies, procedures, data, guidelines, and other knowledge sources they are derived from. For example a change in insurance reimbursement policy or a protocol change by a vendor must be reflected in the actual work instructions used to perform the process. The interaction between knowledge sources is also an important concept in the information domain. Thus in the example in Figure 47 hospital policy for diagnostic requests is influenced by insurer policies (what they will reimburse) and also American Medical Association guidelines as to standards of care (best practices). These guidelines and policies are "constraints" and instructions that are reflected in the actual treatment protocol work instructions. This correspondence between knowledge sources is critical to map since alignment of process steps subject to the constraints of policies and guidelines can be difficult if they are rapidly changing.

Another key concept in the information domain is the derivation of work instructions from vendor protocols and available literature evaluating a particular process step (activity). The goal is to capture the real process by which vendor instructions, such as equipment manuals, and available medical literature, such as validations, get synthesized into actual day-to-day work instructions. The synthesis process becomes even more important when there are numerous design alternatives (and changes) as conceptually shown by vendors A, B, and C. A single vendor was selected through some design process whose output is represented by the "Design Documents", from which the actual work instructions are derived. Once again in processes with rapidly changing technology, the explicit enumeration of the design alternatives from which the final day-to-day protocol was derived is critical to understanding the true process complexity. This is partly because until a "dominant design" emerges, the organization must still monitor and be aware of the performance of the available alternatives, as one of these could become the "dominant design". Of course these design alternatives could also be part of the process, as part of a decision branch in which case there would be direct work instructions associated with each one, and the design document would be associated with the decision branch. An important distinction should be made about the inherent process complexity and that brought on by having many design alternatives for each process step. Inherently complex processes will have difficulty incorporating rapid change, a phenomenon often seen in healthcare. Complex processes will also have difficulty fully optimizing their process steps as the "cost" of exploring all available alternatives increases with complexity. This problem gets worse as processes resistant to change because of their high complexity are faced with new design alternatives, sometimes coming from different fields or affecting only certain steps. The complexity of the overall process will now increase as the inherent process complexity is still there, but the new design alternatives must also be explored.

The question again arises as to whether processes, healthcare in particular, may have a certain level of irreducible complexity which cannot be engineered around and this is the fundamental problem facing many processes. Yet, before irreducible complexity can be accepted as an explanation, it should be measured to analyze the true extent of this complexity and whether it is in fact "irreducible" or whether its current configuration is merely not well designed to take into account complexity reduction. This thesis is an effort to provide tools to measure process complexity such that it can be assessed and managed like any other engineering variable. Like any engineering system, a process will be a set of compromises where greatly reducing the complexity may also significantly reduce the performance. Process designer will have to make these trade-offs, but they should do so with tools to allow them to analyze process complexity. Just as a Cessna 172 is much less complex than a Boeing 777 but it also has much lower performance, some processes are simpler but have much lower performance. Tradeoffs must be made and "complexity" should be "spent" wisely where it greatly improves performance rather than preventing improvement.

Many processes have also not been designed to accommodate change, just as many systems have not been designed in flexible and modular ways such that they can be modified in the future without becoming obsolete. The design of real options for systems is identical in concept to the real options that might be designed for processes where additional infrastructure in the form of personnel, steps, or other elements, might be built in to ensure processes remain flexible and adaptable. Often processes are not changed in their entirety, but in modular ways where the technology associated with one of their activities is changing rapidly and must be incorporated into the process. This is in many ways the example studied in this thesis where the insertion of DNA sequencing, a complex and rapidly changing technology is being inserted into standard clinical microbiology processes. And while complex, DNA sequencing processes are rapidly changing due to the enormous investment being made in R&D but also because no single "dominant design" has emerged to use Utterback's ideas.

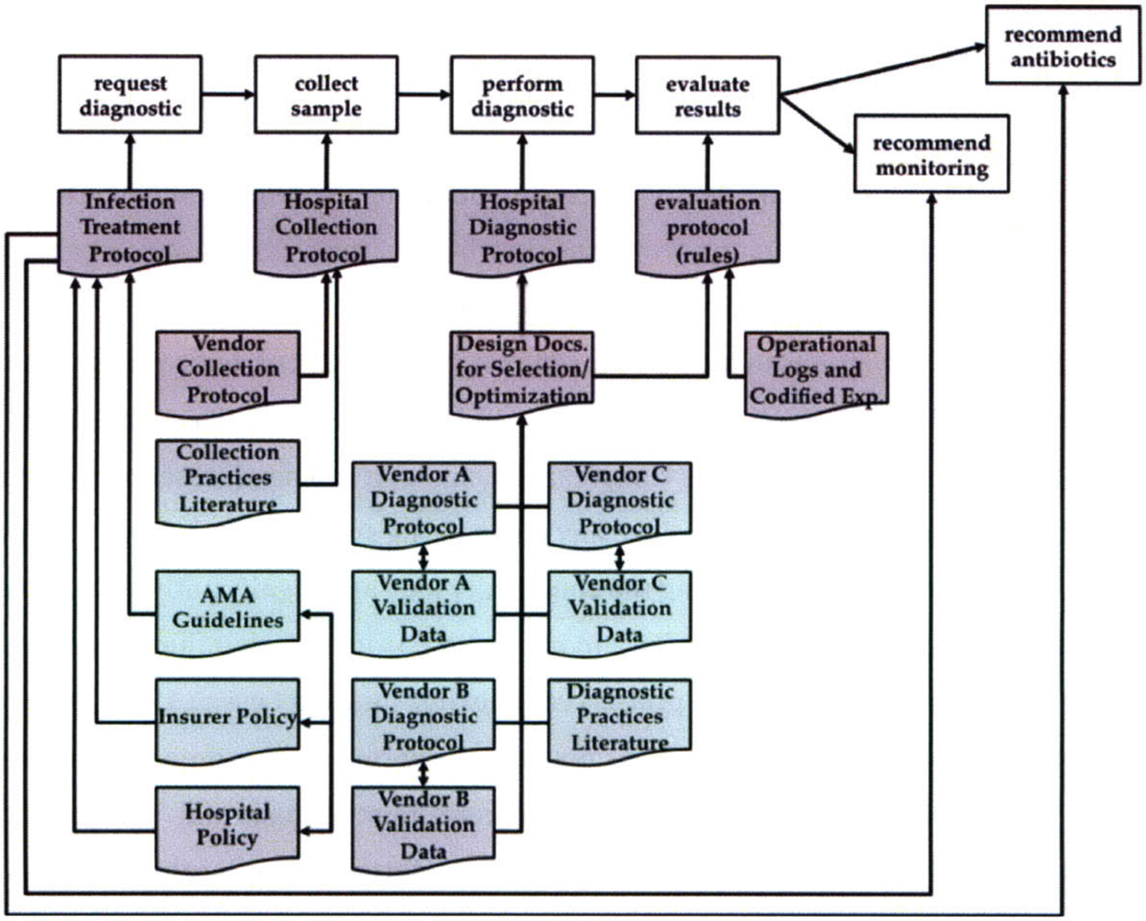
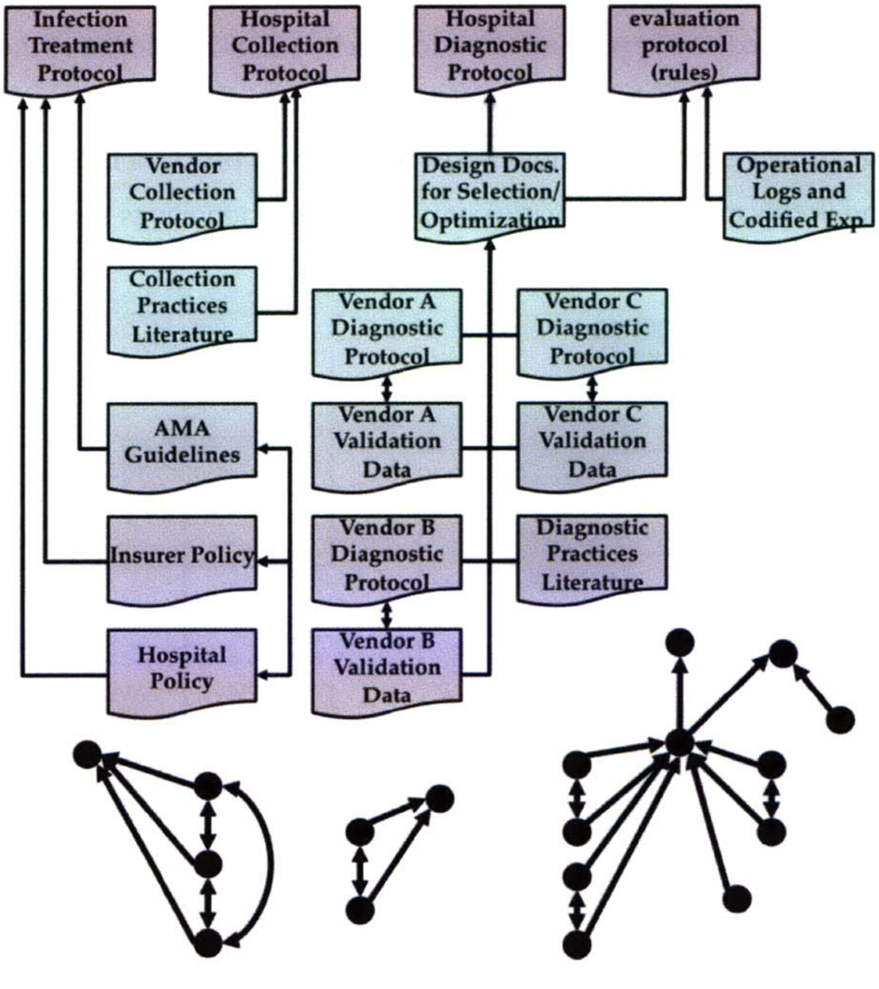

recommend antibiotics
request diagnostic
collect sample
perform diagnostic
evaluate results
recommend monitoring
Infection Treatment Protocol
Hospital Collection Protocol
Hospital Diagnostic Protocol
evaluation protocol (rules)
Vendor Collection Protocol
Design Docs. for Selection/ Optimization
Operational Logs and Codified Exp.
Collection Practices Literature
Vendor A Diagnostic Protocol
Vendor C Diagnostic Protocol
AMA Guidelines
Vendor A Validation Data
Vendor C Validation Data
Insurer Policy
Vendor B Diagnostic Protocol
Diagnostic Practices Literature
Hospital Policy
Vendor B Validation Data

Figure 47 - Interaction of Process and Information Domains

The elements contained in the information domain are connected with directional arrows indicating which elements are derived from which. In cases where elements are derived from each other (or are influenced) bidirectional arrows are shown. For example the infection treatment protocol is derived from AMA standard of care guidelines, insurer policy, and hospital policy but these three items also influence each other and are so represented. The diagram in Figure 48 extracts only the information domain elements and shows their network representation below. This representation is similar to the activity on node (AoN) representation but places information elements on nodes and the edges represent the genealogy (sources from which a particular element was derived). Each of the information domain elements can interact with organization domain individuals, and these interactions represent "familiarity" or "knowledge" of the particular information domain element. This is particularly important when analyzing the overall projection domain network as it impacts the "search" properties of the network. The "reachability" of information domain elements through either the way in which they are derived in the information domain or how well connected individuals are in the organizational domain is critical to the design, implementation, and troubleshooting of complex processes. This will be discussed later in this chapter.

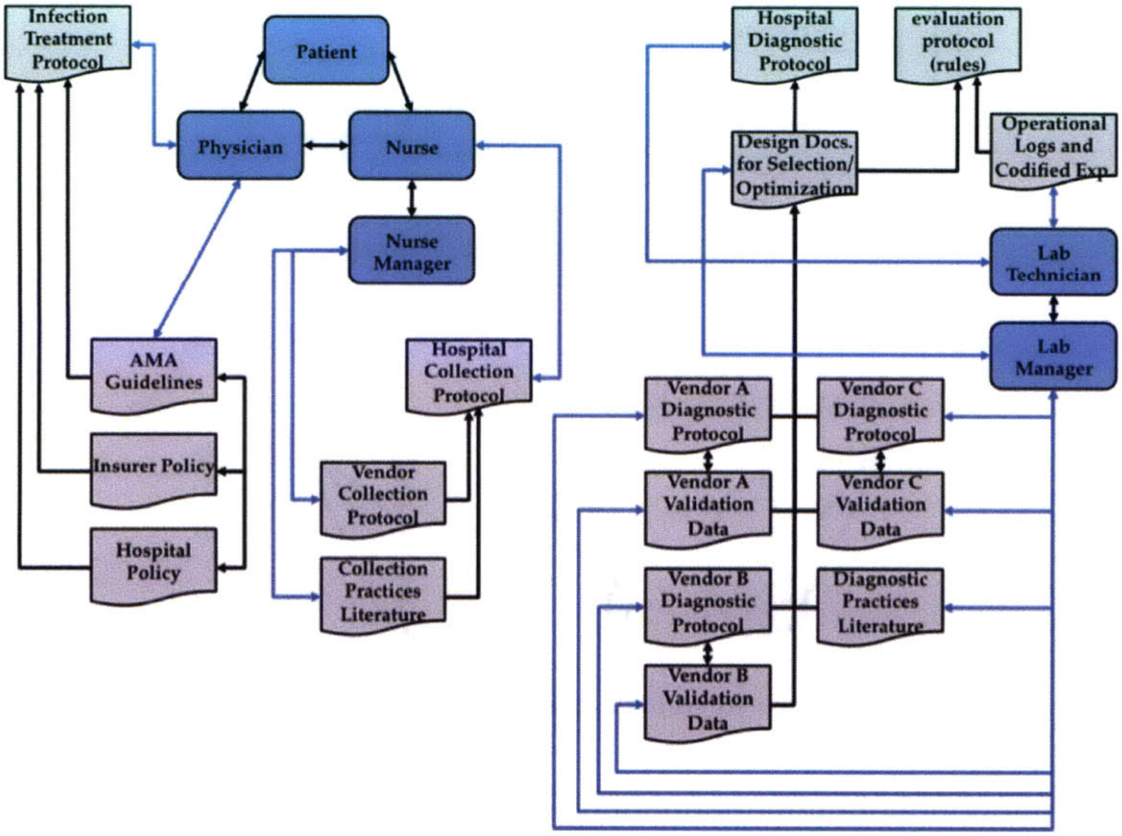


Infection Treatment Protocol
Hospital Collection Protocol
Hospital Diagnostic Protocol
evaluation protocol (rules)
Vendor Collection Protocol
Design Docs. for Selection/ Optimization
Operational Logs and Codified Exp
Collection Practices Literature
Vendor A Diagnostic Protocol
Vendor C Diagnostic Protocol
AMA Guidelines
Vendor A Validation Data
Vendor C Validation Data
Insurer Policy
Vendor B Diagnostic Protocol
Diagnostic Practices Literature
Hospital Policy
Vendor B Validation Data

Figure 48 - Information Domain as a Network

3.5 Projection Domain

Having defined the process, organizational, and information domains it is possible to define the projection domain. The connections between the individuals and process activities have already been shown, as have the connections between the information domain and the process. The only remaining pairing is the set of connections between the information domain and the organization domain, which represents the interaction between individuals and the explicit knowledge they are either aware of, responsible for, or subject to. The interaction between the organization domain and the information domain is conceptually shown in the diagram in Figure 49. Note that not all the elements in the information domain can be reached and certain information such as the knowledge of the protocols, and performance of vendors A, B, and C is held by the lab manager and not directly by the lab technician. However, the operational experience of the protocol is held by the lab technician and in this example explicitly codified in operational logs rather than as tacit knowledge.



Infection Treatment Protocol
Patient
Physician
Nurse
Nurse Manager
AMA Guidelines
Insurer Policy
Hospital Policy
Hospital Collection Protocol
Vendor Collection Protocol
Collection Practices Literature
Hospital Diagnostic Protocol
evaluation protocol (rules)
Design Docs. for Selection/ Optimization
Operational Logs and Codified Exp.
Lab Technician
Lab Manager
Vendor A Diagnostic Protocol
Vendor C Diagnostic Protocol
Vendor A Validation Data
Vendor C Validation Data
Vendor B Diagnostic Protocol
Diagnostic Practices Literature
Vendor B Validation Data

Figure 49 - Interaction of Organization and Information Domains

It is now possible to put all of the elements together in the projection domain using the graph representations of each of the domains and projecting them onto a single one. The projection domain is called a separate "domain" because it has different properties than the other domains in that every element and link is abstracted to be the "same" such that global network properties of the overall process can be calculated. It is of course possible to label and identify each type of node and link in the projection domain such that differences in links can be identified but a central goal of this thesis was to provide global properties of the process, which abstracting all elements onto the "projection domain" allows. Figure 50 is the combined projection domain for the example developed so far. This projection domain now contains all of the intra-domain connections but also the cross-domain connections for each of the pairings.

While many graph-based process analysis methods have focused strictly on the process (functional) domain it is clear from this example that although the process domain can have very little complexity, the supporting structures of information and organization contribute significant "hidden" complexity. And this additional complexity is not an artifact of the model developed here since all of the elements described in each of the domains exist in reality and are necessarily connected for the process to exist and operate. This model is an attempt to describe these connections such that they can be modeled and managed. As will be seen in Chapter 5, real-world processes have much greater complexity than is apparent from "process only" representations which may not include all of the other elements needed for a process including information and organization.

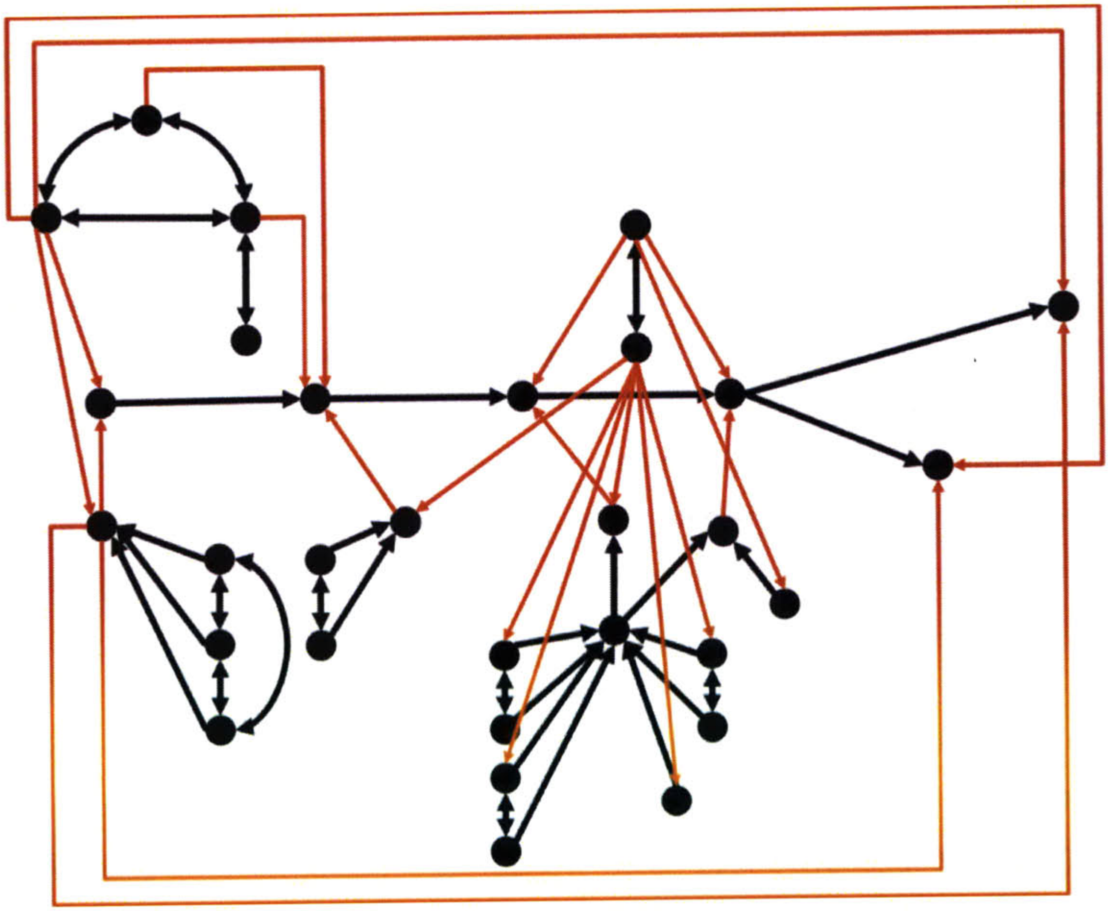


Figure 50 - Graph Representation of Projection Domain

3.6 Graph Properties

The first step to analyzing the projection domain matrix developed thus far is to translate it into a true graph representation. For convenience, the graph was input into NodeXL a prototype software package from Microsoft research that uses an Excel front end to input graph edges and vertices, and can then compute basic measures or output adjacency matrices for use on other platforms including Matlab, UCINET, and others. The same Powerpoint drawing developed thus far for the conceptual projection domain is represented in NodeXL in Figure 51 where red squares represent elements of the process domain, blue triangles, elements of the information domain, and green circles elements of the organization domain. This diagram is the same projection domain network developed thus far.

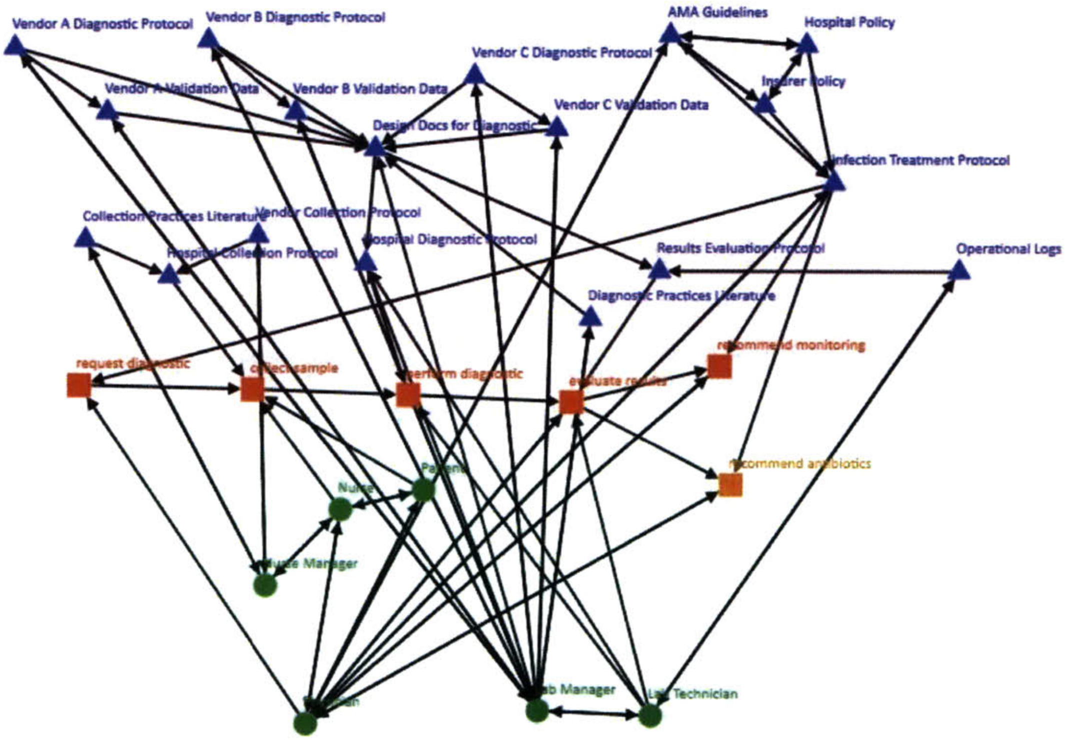

Vendor A Diagnostic Protocol
Vendor B Diagnostic Protocol
AMA Guidelines
Hospital Policy
Vendor C Diagnostic Protocol
Insurer Policy
Vendor A Validation Data
Vendor B Validation Data
Vendor C Validation Data
Design Docs for Diagnostic
Infection Treatment Protocol
Collection Practices Literature
Vendor Collection Protocol
Hospital Diagnostic Protocol
Hospital Collection Protocol
Results Evaluation Protocol
Operational Logs
Diagnostic Practices Literature
recommend monitoring
request diagnostic
collect sample
perform diagnostic
evaluate results
recommend antibiotics
Patient
Nurse
Nurse Manager
Lab Manager
Lab Technician

Figure 51 - Projection Domain with Individual Domains Highlighted

Using this formal graph representation, basic graph metrics can be computed, such as the number of elements, number of connections, average degree of the graph (number of connections per element), given in Equation 1. A higher average degree should correspond to a process where it is relatively easy to search for information and the stakeholders are well connected.

$$\bar{d}(G) = \frac{1}{|V|}\sum_{v \in V} d(v)$$

where G is the graph set, V is the set of all vertices, d(v) is the degree of vertex v

Equation 1 - Average degree for a graph

Using the properties of graphs, an equivalent formula for average degree using the set of Edges is given in Equation 2.

$$\bar{d}(G) = \frac{2 \cdot |E|}{|V|}$$

where E is the set of all edges in the graph and V is the set of all vertices in the graph

Equation 2 - Average degree as a function of edges and vertices

A higher average degree should also correspond to a more connected process network better able to absorb changes, but like any metric these should be balanced with others. It is possible that too high an average degree could lead to difficulties in managing the information flow and make the network also be resistant to change. The graph density: the number of edges over the maximum possible number of unique edges in a complete graph is given in Equation 3.

$$\Delta(G) = \frac{|E|}{|V| \cdot (|V| - 1)}$$

where G is the graph set, E is the set of Edges, and V is the set of vertices

Equation 3 - Graph Density

A very high density could mean the network has exceeded "bandwidth" limits for individuals as nearly every vertex is connected to every other. The shortest distance matrix d(u,v) between pairs of vertices can be found by search algorithms, and the overall average distance for a connected graph can then be computed from the distance matrix, given by Equation 4.

$$\bar{\delta}(G)=\frac{1}{|V|\cdot(|V|-1)}\sum_{\substack{u,v\in V\\u\neq v}}\delta(u,v)$$

where V is the set of all vertices, δ is the shortest path set between each vertex pair *(u,v)*

Equation 4 - Graph average distance

The average distance metric can provide insight as to the number of "steps" that an individual (or work instruction genealogy) would have to go through before reaching another element of the process. Shorter is likely better, but requires more vertices (nodes) as in the example of the process owner. Similarly, using the distance matrix, the diameter of a connected graph can be computed as the maximum distance between the sets of vertex pairs given by Equation 5.

$$diam(G)=\max\left\{\delta(u,v)\middle|u,v\in V\right\}$$

where G is the graph set, δ is the shortest path set between each vertex pair *(u,v)*

Equation 5 - Graph diameter

Other interesting measures can be calculated such as the "Cyclomatic complexity" M, which was originally developed to identify the maximum number of independent paths through software programs (McCabe 1976) as an estimate of their complexity, but can be applied here to the MDPM model as the maximum number of information exchange paths or the number of process paths, given by Equation 6.

$$M(G)=|E|-|V|+P$$

where E is the set of edges, V is the set of vertices, and P is the number of connected components in the graph

Equation 6 - Cyclomatic Number

Other measures can be defined based on the number of branch points in traditional Event Process Chains, although no explicit connectors are used in this model. However, branch points in the process domain could be used as a proxy for the "connectors" to indicate AND, OR, or XOR branches, such that the same definition applies as given in Equation 7.

$$\bar{d}_C(G) = \frac{1}{|C|}\sum_{c \in C} d(c)$$

where G is the graph set, C is the set of connectors, d(c) is the degree of connector c

Equation 7 - Average degree for EPC connectors

And a calculation of the maximum degree of the branch points (connectors) as given in Equation 8 can indicate the magnitude of the decision branches that need to be considered in the process.

$$\hat{d}_C(G) = \max\left\{d(c) \middle| c \in C\right\}$$

where G is the graph set, C is the set of connectors, d(c) is the degree of connector c

Equation 8 - Maximum degree of EPC connectors

Using these measures it is possible to calculate basic complexity measures of the example projection domain (and overall MDPM model) and compare these measures to changes that might be made to improve the communications performance of the process network or to reduce the complexity of the process. So for example, the addition of a process owner to communicate directly between the lab manager, the physician, and the nurse manager, as well as be aware of the key design and operational documents for the process can greatly improve the "search process" or "reachability" as given by a reduction in the average distance between nodes. In grid format, the original baseline process without a "process owner" is given first, and then the addition of a process owner in Figure 52 and Figure 53 below:

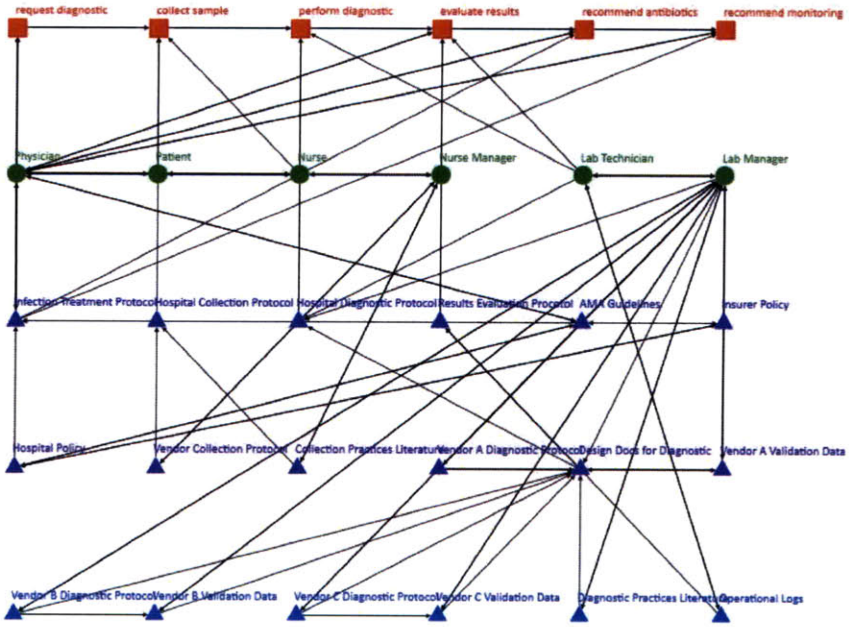

request diagnostic
collect sample
perform diagnostic
evaluate results
recommend antibiotics
recommend monitoring
Physician
Patient
Nurse
Nurse Manager
Lab Technician
Lab Manager
Infection Treatment Protocol
Hospital Collection Protocol
Hospital Diagnostic Protocol
Results Evaluation Procotol
AMA Guidelines
Insurer Policy
Hospital Policy
Vendor Collection Protocol
Collection Practices Literature
Vendor A Diagnostic Protocol
Design Docs for Diagnostic
Vendor A Validation Data
Vendor B Diagnostic Protocol
Vendor B Validation Data
Vendor C Diagnostic Protocol
Vendor C Validation Data
Diagnostic Practices Literature
Operational Logs

Figure 52 - Baseline Projection Domain

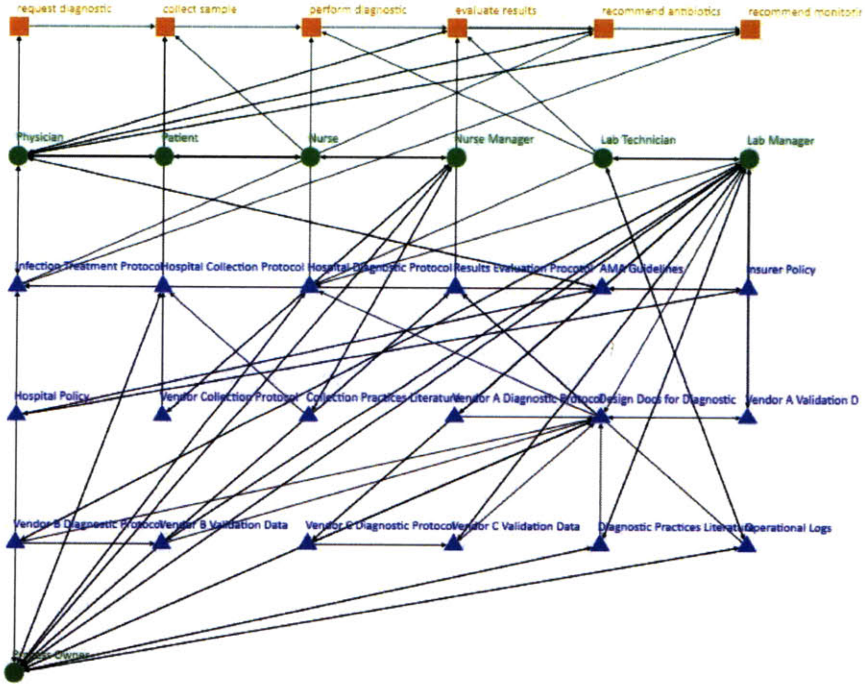

request diagnostic
collect sample
perform diagnostic
evaluate results
recommend antibiotics
recommend monitorir
Physician
Patient
Nurse
Nurse Manager
Lab Technician
Lab Manager
Infection Treatment Protocol
Hospital Collection Protocol
Hospital Diagnostic Protocol
Results Evaluation Procotol
AMA Guidelines
Insurer Policy
Hospital Policy
Vendor Collection Protocol
Collection Practices Literature
Vendor A Diagnostic Protocol
Design Docs for Diagnostic
Vendor A Validation D
Vendor B Diagnostic Protocol
Vendor B Validation Data
Vendor C Diagnostic Protocol
Vendor C Validation Data
Diagnostic Practices Literature
Operational Logs
Process Owner

Figure 53 - Projection Domain with Addition of a Process Owner

Other more sophisticated graph measures can be defined including some from information theory as discussed in chapter 2, such as the Shannon entropy of a graph as a function of the degree distribution. This is a measure of the randomness of a graph and is in units of "bits". The Shannon entropy of a graph is given in Equation 9.

$$I(G) = -\sum_{i=1}^{\max d} h_i \left(\log_2 (h_i)\right)$$

Where h_i is the fraction of nodes with degree i

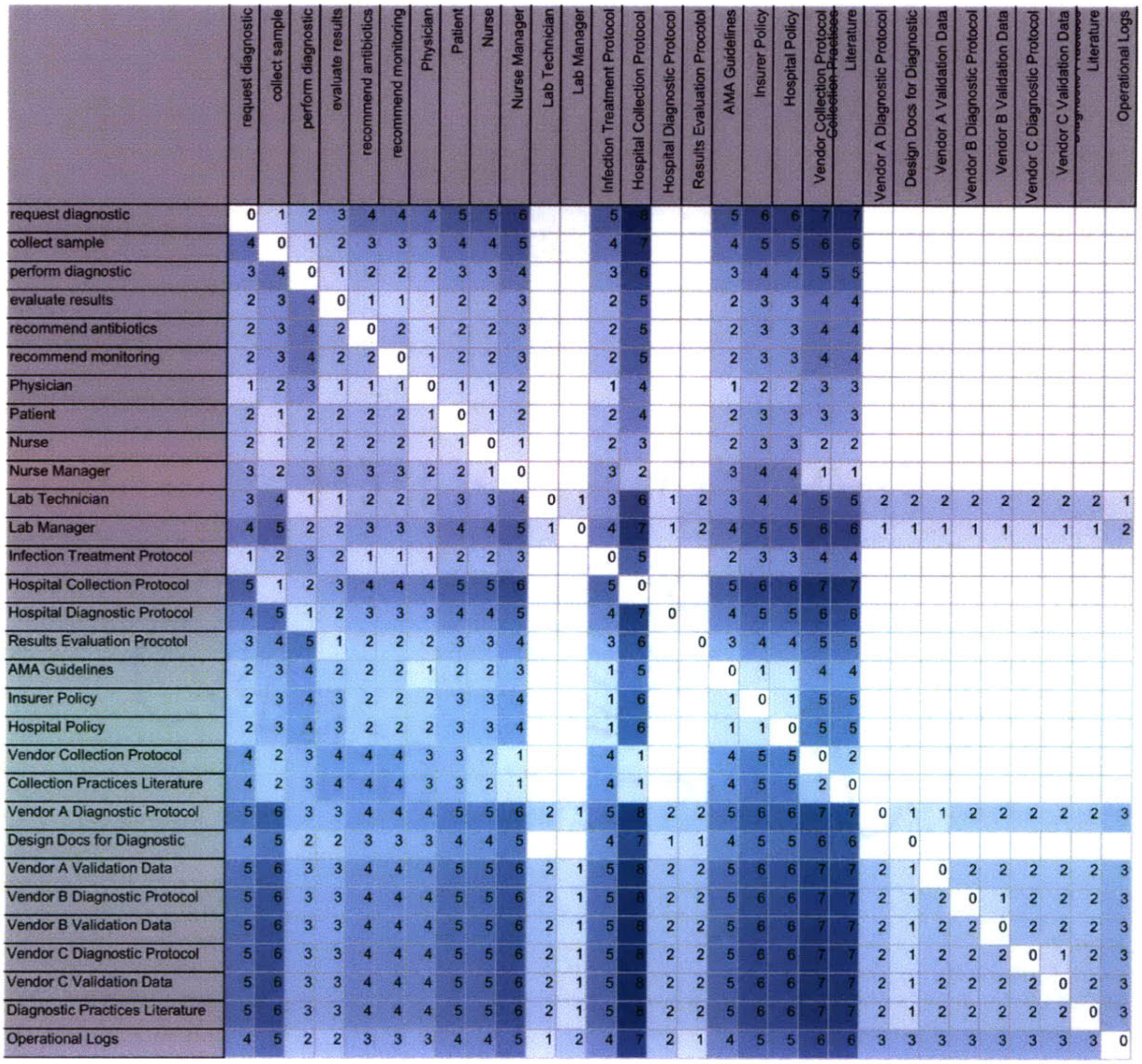
Equation 9 - Information (Structural) Entropy (from Shannon)

These measures are calculated for the baseline graph and for the one where a process owner has been added. The results are in Table 6:

	Baseline	**Baseline+Process Owner**
Average Degree	5.53	6.65
Cyclomatic Complexity	54	73
Size	30	31
Entropy	2.61	2.78
Max. Average Distance	**4.14**	**3.17**

Table 6 - Basic Metrics for Example Projection Domain

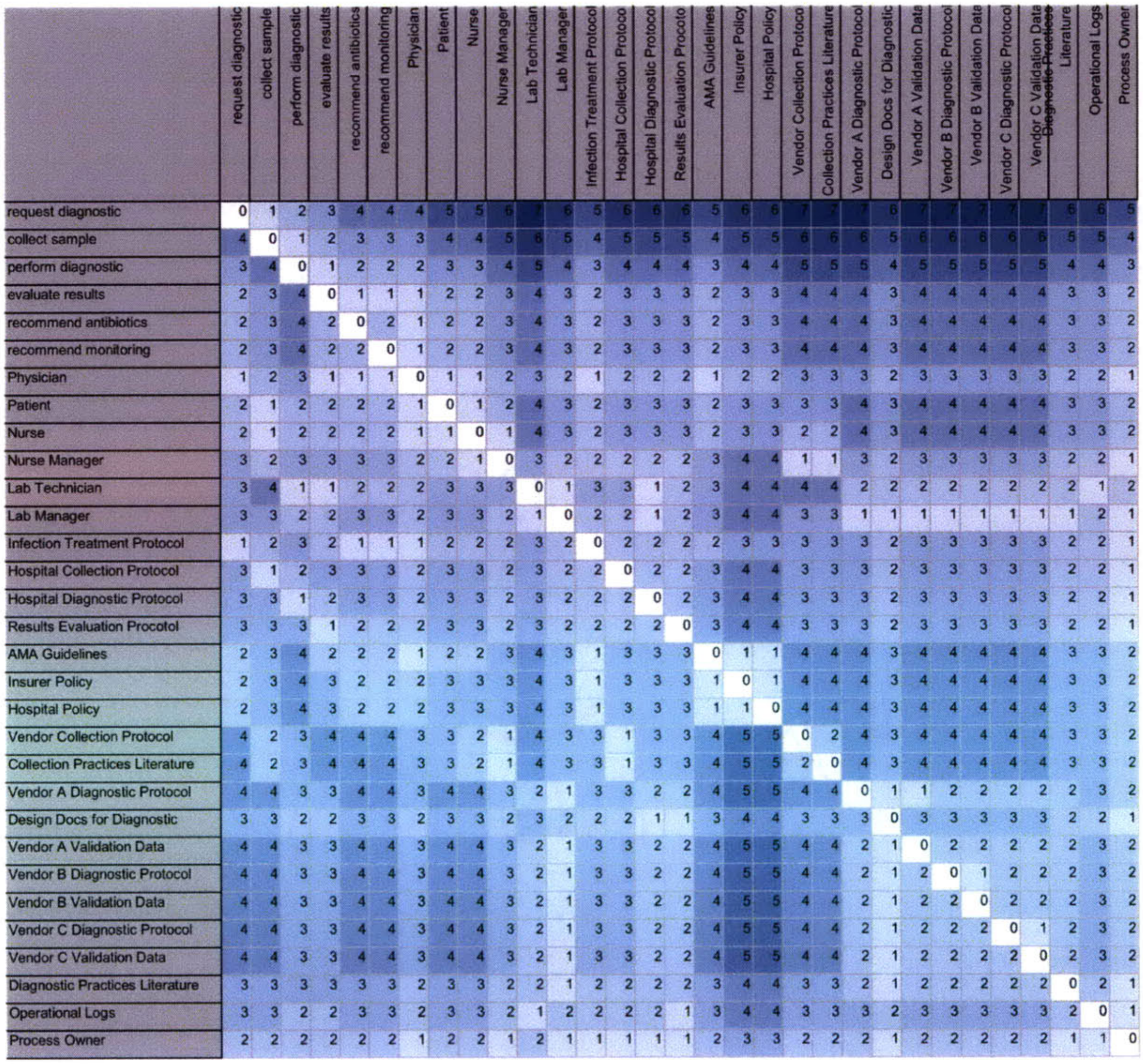
Adding a process owner to improve the connectivity of the graph reducing the maximum average distance between nodes to 3.17 from 4.14. There is some "cost" to this, as the cyclomatic complexity, the number of possible paths, increases significantly, but the benefit to reachability is significant. This can be clearly seen by calculating the geodesic (shortest path) distances across the network for every pair of nodes, as shown in the heat map in Figure 54:



	request diagnostic	collect sample	perform diagnostic	evaluate results	recommend antibiotics	recommend monitoring	Physician	Patient	Nurse	Nurse Manager	Lab Technician	Lab Manager	Infection Treatment Protocol	Hospital Collection Protocol	Hospital Diagnostic Protocol	Results Evaluation Procotol	AMA Guidelines	Insurer Policy	Hospital Policy	Vendor Collection Protocol	Collection Practices Literature	Vendor A Diagnostic Protocol	Design Docs for Diagnostic	Vendor A Validation Data	Vendor B Diagnostic Protocol	Vendor B Validation Data	Vendor C Diagnostic Protocol	Vendor C Validation Data	Diagnostic Practices Literature	Operational Logs
request diagnostic	0	1	2	3	4	4	4	5	5	6			5	8			5	6	6	7	7									
collect sample	4	0	1	2	3	3	3	4	4	5			4	7			4	5	5	6	6									
perform diagnostic	3	4	0	1	2	2	2	3	3	4			3	6			3	4	4	5	5									
evaluate results	2	3	4	0	1	1	1	2	2	3			2	5			2	3	3	4	4									
recommend antibiotics	2	3	4	2	0	2	1	2	2	3			2	5			2	3	3	4	4									
recommend monitoring	2	3	4	2	2	0	1	2	2	3			2	5			2	3	3	4	4									
Physician	1	2	3	1	1	1	0	1	1	2			1	4			1	2	2	3	3									
Patient	2	1	2	2	2	2	1	0	1	2			2	4			2	3	3	3	3									
Nurse	2	1	2	2	2	2	1	1	0	1			2	3			2	3	3	2	2									
Nurse Manager	3	2	3	3	3	3	2	2	1	0			3	2			3	4	4	1	1									
Lab Technician	3	4	1	1	2	2	2	3	3	4	0	1	3	6	1	2	3	4	4	5	5	2	2	2	2	2	2	2	2	1
Lab Manager	4	5	2	2	3	3	3	4	4	5	1	0	4	7	1	2	4	5	5	6	6	1	1	1	1	1	1	1	1	2
Infection Treatment Protocol	1	2	3	2	1	1	1	2	2	3			0	5			2	3	3	4	4									
Hospital Collection Protocol	5	1	2	3	4	4	4	5	5	6			5	0			5	6	6	7	7									
Hospital Diagnostic Protocol	4	5	1	2	3	3	3	4	4	5			4	7	0		4	5	5	6	6									
Results Evaluation Procotol	3	4	5	1	2	2	2	3	3	4			3	6		0	3	4	4	5	5									
AMA Guidelines	2	3	4	2	2	2	1	2	2	3			1	5			0	1	1	4	4									
Insurer Policy	2	3	4	3	2	2	2	3	3	4			1	6			1	0	1	5	5									
Hospital Policy	2	3	4	3	2	2	2	3	3	4			1	6			1	1	0	5	5									
Vendor Collection Protocol	4	2	3	4	4	4	3	3	2	1			4	1			4	5	5	0	2									
Collection Practices Literature	4	2	3	4	4	4	3	3	2	1			4	1			4	5	5	2	0									
Vendor A Diagnostic Protocol	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	0	1	1	2	2	2	2	2	3
Design Docs for Diagnostic	4	5	2	2	3	3	3	4	4	5			4	7	1	1	4	5	5	6	6		0							
Vendor A Validation Data	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	2	1	0	2	2	2	2	2	3
Vendor B Diagnostic Protocol	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	2	1	2	0	1	2	2	2	3
Vendor B Validation Data	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	2	1	2	2	0	2	2	2	3
Vendor C Diagnostic Protocol	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	2	1	2	2	2	0	1	2	3
Vendor C Validation Data	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	2	1	2	2	2	2	0	2	3
Diagnostic Practices Literature	5	6	3	3	4	4	4	5	5	6	2	1	5	8	2	2	5	6	6	7	7	2	1	2	2	2	2	2	0	3
Operational Logs	4	5	2	2	3	3	3	4	4	5	1	2	4	7	2	1	4	5	5	6	6	3	3	3	3	3	3	3	3	0

Figure 54 - Reachability Matrix for Baseline Example

Note that not all nodes are reachable (blank squares) in Figure 54, and the maximum distance is 8 between various elements of the projection domain network. For comparison, the geodesic distance heat map for the case where a process owner is added is shown in Figure 55 with clear benefits to reachability. All nodes are now connected and reachable in a maximum of 7 steps.



	request diagnostic	collect sample	perform diagnostic	evaluate results	recommend antibiotics	recommend monitoring	Physician	Patient	Nurse	Nurse Manager	Lab Technician	Lab Manager	Infection Treatment Protocol	Hospital Collection Protoco	Hospital Diagnostic Protocol	Results Evaluation Procoto	AMA Guidelines	Insurer Policy	Hospital Policy	Vendor Collection Protoco	Collection Practices Literature	Vendor A Diagnostic Protocol	Design Docs for Diagnostic	Vendor A Validation Data	Vendor B Diagnostic Protocol	Vendor B Validation Data	Vendor C Diagnostic Protocol	Vendor C Validation Data Diagnostic Practices	Literature	Operational Logs	Process Owner
request diagnostic	0	1	2	3	4	4	4	5	5	6	7	6	5	6	6	6	5	6	6	7	7	7	6	7	7	7	7	7	6	6	5
collect sample	4	0	1	2	3	3	3	4	4	5	6	5	4	5	5	5	4	5	5	6	6	6	5	6	6	6	6	6	5	5	4
perform diagnostic	3	4	0	1	2	2	2	3	3	4	5	4	3	4	4	4	3	4	4	5	5	5	4	5	5	5	5	5	4	4	3
evaluate results	2	3	4	0	1	1	1	2	2	3	4	3	2	3	3	3	2	3	3	4	4	4	3	4	4	4	4	4	3	3	2
recommend antibiotics	2	3	4	2	0	2	1	2	2	3	4	3	2	3	3	3	2	3	3	4	4	4	3	4	4	4	4	4	3	3	2
recommend monitoring	2	3	4	2	2	0	1	2	2	3	4	3	2	3	3	3	2	3	3	4	4	4	3	4	4	4	4	4	3	3	2
Physician	1	2	3	1	1	1	0	1	1	2	3	2	1	2	2	2	1	2	2	3	3	3	2	3	3	3	3	3	2	2	1
Patient	2	1	2	2	2	2	1	0	1	2	4	3	2	3	3	3	2	3	3	3	3	4	3	4	4	4	4	4	3	3	2
Nurse	2	1	2	2	2	2	1	1	0	1	4	3	2	3	3	3	2	3	3	2	2	4	3	4	4	4	4	4	3	3	2
Nurse Manager	3	2	3	3	3	3	2	2	1	0	3	2	2	2	2	2	3	4	4	1	1	3	2	3	3	3	3	3	2	2	1
Lab Technician	3	4	1	1	2	2	2	3	3	3	0	1	3	3	1	2	3	4	4	4	4	2	2	2	2	2	2	2	2	1	2
Lab Manager	3	3	2	2	3	3	2	3	3	2	1	0	2	2	1	2	3	4	4	3	3	1	1	1	1	1	1	1	1	2	1
Infection Treatment Protocol	1	2	3	2	1	1	1	2	2	2	3	2	0	2	2	2	2	3	3	3	3	3	2	3	3	3	3	3	2	2	1
Hospital Collection Protocol	3	1	2	3	3	3	2	3	3	2	3	2	2	0	2	2	3	4	4	3	3	3	2	3	3	3	3	3	2	2	1
Hospital Diagnostic Protocol	3	3	1	2	3	3	2	3	3	2	3	2	2	2	0	2	3	4	4	3	3	3	2	3	3	3	3	3	2	2	1
Results Evaluation Procotol	3	3	3	1	2	2	2	3	3	2	3	2	2	2	2	0	3	4	4	3	3	3	2	3	3	3	3	3	2	2	1
AMA Guidelines	2	3	4	2	2	2	1	2	2	3	4	3	1	3	3	3	0	1	1	4	4	4	3	4	4	4	4	4	3	3	2
Insurer Policy	2	3	4	3	2	2	2	3	3	3	4	3	1	3	3	3	1	0	1	4	4	4	3	4	4	4	4	4	3	3	2
Hospital Policy	2	3	4	3	2	2	2	3	3	3	4	3	1	3	3	3	1	1	0	4	4	4	3	4	4	4	4	4	3	3	2
Vendor Collection Protocol	4	2	3	4	4	4	3	3	2	1	4	3	3	1	3	3	4	5	5	0	2	4	3	4	4	4	4	4	3	3	2
Collection Practices Literature	4	2	3	4	4	4	3	3	2	1	4	3	3	1	3	3	4	5	5	2	0	4	3	4	4	4	4	4	3	3	2
Vendor A Diagnostic Protocol	4	4	3	3	4	4	3	4	4	3	2	1	3	3	2	2	4	5	5	4	4	0	1	1	2	2	2	2	2	3	2
Design Docs for Diagnostic	3	3	2	2	3	3	2	3	3	2	3	2	2	2	1	1	3	4	4	3	3	3	0	3	3	3	3	3	2	2	1
Vendor A Validation Data	4	4	3	3	4	4	3	4	4	3	2	1	3	3	2	2	4	5	5	4	4	2	1	0	2	2	2	2	2	3	2
Vendor B Diagnostic Protocol	4	4	3	3	4	4	3	4	4	3	2	1	3	3	2	2	4	5	5	4	4	2	1	2	0	1	2	2	2	3	2
Vendor B Validation Data	4	4	3	3	4	4	3	4	4	3	2	1	3	3	2	2	4	5	5	4	4	2	1	2	2	0	2	2	2	3	2
Vendor C Diagnostic Protocol	4	4	3	3	4	4	3	4	4	3	2	1	3	3	2	2	4	5	5	4	4	2	1	2	2	2	0	1	2	3	2
Vendor C Validation Data	4	4	3	3	4	4	3	4	4	3	2	1	3	3	2	2	4	5	5	4	4	2	1	2	2	2	2	0	2	3	2
Diagnostic Practices Literature	3	3	3	3	3	3	2	3	3	2	2	1	2	2	2	2	3	4	4	3	3	2	1	2	2	2	2	2	0	2	1
Operational Logs	3	3	2	2	3	3	2	3	3	2	1	2	2	2	2	1	3	4	4	3	3	3	2	3	3	3	3	3	2	0	1
Process Owner	2	2	2	2	2	2	1	2	2	1	2	1	1	1	1	1	2	3	3	2	2	2	1	2	2	2	2	2	1	1	0

Figure 55 - Reachability Matrix for Process Owner Example

What might this mean in a real process? If individuals in the organization domain are unaware of upstream activity information (whether it be changes, operational performance, or problems) it can make change, improvement, and troubleshooting difficult. Similarly if the information used to perform downstream steps has not been derived (or "aligned") with policies and previous experience (i.e. design reviews) the process will underperform, fail to satisfy stakeholder needs, and potentially fail. A side effect of low connectivity in the projection domain is that processes are more likely to resist change since it is difficult to evaluate the overall impact of changes at any given activity. This may cause processes to be “frozen” in sub-optimal configurations that have been demonstrated to work, but for which further optimization and change is difficult to manage due to a lack of connectivity (search) between the various organizational, process, and information domains.

3.7 Multi-Domain Process Matrix Metrics

A goal of this thesis was to provide a graph based model of complex processes such that many of the metrics used to understand graphs could be brought to bear in the analysis of the various process domains. Graph metrics used in this thesis can generally be classified in the following categories based on (Costa, Rodrigues et al. 2005):

- **Distance** metrics of graph “size” such as shortest (geodesic) paths between nodes
- **Structure** metrics of graph “organization” such as degree distribution or clustering
- **Entropy** metrics of graph “randomness” such as Shannon entropy of degree distribution

These metrics categories are further described in the following sections and are intended to provide a set of metrics that can be used in the comparison of multi-domain process matrices (MDPMs) associated with processes. The ability to measure properties of MDPM models is essential to their utility as an engineering tool in order to quantifiably analyze improvements to a process. With the set of metrics proposed in the following section it may be possible to design processes according to desired specifications and build Design of Experiments (DOE) models like those described by (Box and Liu 1999) for engineering systems that identify the key variables in the process subject to the response surface defined by the following metrics. The application of multi-domain matrices is still evolving however and greater experience and research will be needed to fully characterize the most relevant set of graph metrics to describe complex processes as in the MDPM model and similarly in the larger ESM model. The objective of this thesis was to provide a comprehensive set that could be used on the MDPM model and other similar models such as the ESM model. These metrics are shown to have utility by comparing improvements in the MRSA surveillance process test case in Chapter 5, but other test cases will be needed to converge on a useful set. Also while the focus of this thesis is on the study of complex processes, not all of these metrics are necessarily associated with complexity as some are simply properties of a process that can nevertheless be used analytically in designing and optimizing processes.

For each of the metrics proposed in this section a complete description is given including the metric name according to convention found in the relevant literature, a brief definition, the metric formula, its potential application in the analysis of MDPM domains, any potential limitations in the use and interpretation of the metric, a hypothesis for the interpretation of the metric, and related literature references. Again, the goal of the following sections are to provide a comprehensive description of available network metrics some of which will be used in Chapters 4 and 5, but the entire listing in this chapter can serve as a future reference for others wishing to use the MDPM methodology.

3.7.1 Structural Metrics

Structural metrics describe fundamental properties of the graph and its associated adjacency matrix including numbers of nodes, edges, graph density, and degree distribution. These metrics can provide process designers with the basic description of a process, such as the number of process steps (nodes in the process domain), how well individuals are connected to each other in the organizational domain (graph density), and the number of potential paths through a process (cyclomatic number). Each structure metric is described in the following tables.

Metric	S_V
Definition	Number of vertices (nodes) in the domain graph corresponding to the number of distinct elements

Formula	$S_V(G)=\|V\|$ where V is the set of vertices in graph G

Equation 10 – Number of Vertices Metric

Application	The greater the number of elements in the graph the higher the complexity is likely to be, but need to use with other measures. Can also be used to compare absolute sizes of element sets between processes (i.e. numbers of activities).
Limitations	Process domain graphs that are sequential can have large numbers of activities but low structural complexity. Similarly "small world" organizational networks may have large numbers of individuals but low communication complexity.
Hypothesis	High S_V corresponds to higher complexity
Related Work	(Costa, Rodrigues et al. 2005; Mendling 2009)

Metric	S_E
Definition	Number of edges in the domain graph
Formula	$S_E(G)=\|E\|$ where E is the set of vertices in graph G

Equation 11 - Number of Edges Metric

Application	The greater the number of connections in the graph the higher the complexity
Limitations	Process domain graphs that are sequential can have large numbers of activities but low structural complexity. Similarly "small world" organizational networks may have large numbers of individuals but low communication complexity.
Hypothesis	High S_V corresponds to higher complexity

Related Work	(Costa, Rodrigues et al. 2005; Mendling 2009)

Metric	Δ
Definition	Density of a graph. Number edges divided by the maximum possible number of edges given the number of nodes. Note that this is the directed graph density.
Formula	$\Delta(G)=\frac{\lvert E\rvert}{\lvert V\rvert\cdot(\lvert V\rvert-1)}$ Equation 3 - Graph Density Metric
Application	In a very dense graph in any of the domains, a large number of interconnections may be associated with a higher complexity. Can compare overall process MDPM models to evaluate "infrastructure".
Limitations	Ability to compare across graphs with varying numbers of vertices since density is proportional to $1/V^2$. Also, high density may not correspond to "complexity" since better communication may actually decrease complexity (increase reachability) to a point. Also may not be comparable across domains.
Hypothesis	High density corresponds to higher complexity and cost of maintaining links especially in organization and information domains.
Related Work	(Costa, Rodrigues et al. 2005; Mendling 2009)

Metric	M, Cyclomatic Complexity
Definition	Number of linearly independent paths through a graph

Formula $M(G)=|E|-|V|+P$

Where E is the set of edges and V the set of vertices in graph G, and P is the number of components (if graph is strongly connected)

Equation 12 - Cyclomatic Complexity Metric

Application A domain with a high cyclomatic complexity suggests there are a large number of linearly independent paths through the graph and therefore complexity in "knowing" which route to take. This measure is derived from software complexity analysis (McCabe 1976), and is directly relevant to the process domain although likely applicable to the others in some fashion as well. Note also the contribution of disconnected components, which can be substantial in organization and information domains for comparison between MDPM models.

Limitations High M graphs may not correspond to high complexity if alternative paths are not frequently used, or in organization domains where paths are not "pre determined"

Hypothesis High M corresponds to higher complexity

Related Work (McCabe 1976)

Metric $\bar{d}$

Definition Average degree of a vertex

Formula $\bar{d}(G)=\frac{2\cdot|E|}{|V|}$

Equation 1 - Average degree for a graph

Application A domain with a high average degree should be more complex as a greater number of connections must be maintained. A high average degree may also correspond to "bandwidth" problems as elements may not be able to communicate or process the information from many nodes they are connected to.

Limitations	High average degree graphs may not correspond to high complexity if reachability is improved. Also high average degree may be misleading if the graph is scale-free with a few very high degree nodes and many relatively low degree ones.
Hypothesis	High average degree corresponds to higher complexity and decreasing bandwidth of each element.
Related Work	(Costa, Rodrigues et al. 2005)

Metric	$P(k)$
Definition	Degree distribution of vertices (random binomial, poisson, power law, and others)
Formula	$P(k) \sim k^{-\gamma}$ where k is the degree and γ is the power law constant **Equation 13 - Power law degree distribution Metric**
Application	A domain with a power law distribution will have less complexity than a randomly distributed (binomial) one and be easier to navigate.
Limitations	Power law distributions may not correspond to lower complexity since some nodes may require hierarchical structures to deal with the many incoming and outgoing "connections". Many other distributions are possible, each with their own rationale.
Hypothesis	Information exchange is easier in power law "scale free" networks
Related Work	(Barabási, Dezsõ et al. 2003)

3.7.2 Distance Metrics

Distance metrics describe the spatial properties of the graph and its associated adjacency matrix including average geodesic distance between pairs, shortest (geodesic) distance between individual pairs, network efficiency, and network vulnerability. These metrics can provide process designers with knowledge about the relationship between nodes and the "information transmission" properties of the MDPM, which may be used in understanding how hard (or easy) it is to propagate change or troubleshoot problems in the process. Each structure metric is described in the following tables.

Metric	$\bar{l}$
Definition	Average geodesic distance between vertex pairs
Formula	$\bar{l} = \frac{1}{\lvert V \rvert \cdot (\lvert V \rvert - 1)} \sum_{i \neq j} d_{ij}$ where V is the set of graph vertices, and d is the shortest geodesic path between two vertices **Equation 14 - Average Geodesic Distance Metric**
Application	For a connected graph this measure provides an estimate of the average number of "hops" between vertex pairs and should correspond to the difficulty to transmit or receive information across the network. Distances can have different meanings in each of the domains, such as number of people along a communication path (organization) or number of process steps (process). This metric also works primarily for connected graphs as highly fragmented disconnected graphs will show a low average distance for each connected group but "infinite" distance between disconnected vertices. The organization and information domains are often highly fragmented so this metric may be difficult to interpret except for the overall projection domain.
Limitations	Efficient information transmission may make this measure less meaningful. Also, does not take into account vertex capacity (i.e. overload) or channel capacity.
Hypothesis	High value corresponds to higher difficulty of transmitting information, and potentially greater difficulty in controlling a complex process

Related Work (Latora and Marchiori 2001; Costa, Rodrigues et al. 2005)

Metric ε

Definition Average network "efficiency" of transmitting information

Formula

$$\varepsilon = \frac{1}{|V| \cdot \left(|V| - 1\right)} \sum_{i \neq j} \frac{1}{d_{ij}}$$

where V is the set of graph vertices, and d is the shortest geodesic path between two vertices

Equation 15 - Average Network Efficiency Metric

Application For a connected graph this measure provides a distance-weighted estimate of the efficiency to transmit or receive information across the network. The efficiency is assumed to be inversely proportional to the distance between pairs and provides a measure of the "degradation" of information if it passes through many levels (i.e. as in a hierarchy).

Limitations The falloff of efficiency may not be as severe as the inverse of the distance. Metric is likely only meaningful for connected graphs.

Hypothesis The farther away a node pair is on a graph, the more difficult it is to receive information and the lower the quality.

Related Work (Latora and Marchiori 2001; Costa, Rodrigues et al. 2005)

Metric μ

Definition Network vulnerability

Formula

$$\mu = \max_i \left(\frac{\varepsilon - \varepsilon_i}{\varepsilon} \right)$$

where ε is the overall network efficiency and ε_i is the efficiency minus the ith vertex

Equation 16 - Network Vulnerability Metric

Application For a connected graph this metric provides an estimate of the decrease in information transmission efficiency from the loss of critical nodes. Highly connected process networks will be presumably less vulnerable as there will be other "paths" that communication can go through although at some cost in maintaining all of these other connections.

Limitations Some of the domains in the MDPM model such as the actual activities in the process layer are "essential" and thus the vulnerability metric may not be as meaningful in the process domain.

Hypothesis A process network with a high vulnerability suggests much of the information travels through a few hubs, which should be supplemented with additional paths through other nodes.

Related Work (Latora and Marchiori 2005)

Metric	B
Definition	Distance weighted fragmentation

Formula

$$F_D = 1 - \frac{2\sum_{i>j} \frac{1}{d_{ij}}}{|V| \cdot (|V| - 1)}$$

where V is the set of graph vertices, and d is the shortest geodesic path between two vertices (1/∞ set to 0)

Equation 17 - Distance Weighted Fragmentation

Application	Measure of fragmentation, but weighted to account for greater "fragmentation" with longer path lengths. Accounts for the "limited horizon" of individuals who may not be able to "see" very far into the network, effectively fragmenting it. May be useful for the analysis of the organization domain.
Limitations	The falloff fragmentation effect due to distance may not be as severe as the inverse of the distance.
Hypothesis	The farther away a node pair is on a graph, the more difficult it is to receive information and the lower the quality effectively fragmenting that node from the rest of the network.
Related Work	(Borgatti 2006)

Metric	diam
Definition	Longest path between pairs of vertices in the graph
Formula	$diam(G) = \max\left\{\delta(u,v) \middle\| u,v \in V\right\}$ Equation 5 - Graph diameter metric
Application	In a process, where every step can impact the outcome, long information transmission paths produce higher the complexity and make it mode difficult to manage. Especially relevant to the process domain where the number of activities needed to be performed in "sequence" along a path suggests a greater probability of an individual link in the chain failing (and needing troubleshooting efforts).
Limitations	Process domain graphs that are sequential can have large diameters, but have low structural complexity. Can describe the complexity to information transmission in the organizational and information domains.

Hypothesis	High diam corresponds to higher complexity and the potential for alignment problems between stakeholders separated by many process steps (or other domain steps such as organization).
Related Work	(Mendling 2009)

3.7.3 Information Metrics

Information metrics in communications networks (and graphs) derive largely from (Shannon and Weaver 1948) where Shannon describes the information entropy as a measure of the amount of information needed to "know" the value of a random variable transmitted in a given communications channel. The Shannon entropy has been abstracted to describe the amount of information needed to fully specify or "know" the current state of a system and increases with the number of available states. An analogy might be the amount of information needed to fully specify the states of atoms in a solid near absolute zero will be less than the amount of information needed to specify the state of the same material at a much higher temperature in a gaseous state where many more states can exist. Applied to the MDPM model (and potentially to the ESM model as well) Shannon entropy ideas could be used to describe the amount of information needed to fully describe the "state" of a complex process including which individuals are talking to which, what process steps are being attended to, etc. The Shannon entropy ideas applied to network analysis have to a large extent focused on the "structural entropy" of a network, treating the degree distribution as the random variable and defining the amount of information that would be needed to "know" the exact degree distribution. These structural entropy ideas have been used to describe the robustness of networks and resilience as described in (Costa, Rodrigues et al. 2005).

However, the Shannon entropy concepts have been extended to describe the amount of information needed to "search" networks such as finding a particular node or finding a route through a network as described in (Sneppen, Trusina et al. 2004; Rosvall and Sneppen 2006). This "search information" can be thought of as the number of "yes/no" questions that an agent traversing the network between an origin and destination node would have to ask at each edge start on each node in the path to follow the shortest path. Thus at a node of degree 4, the agent would have just entered the node from one of the edges, but would have to "ask" for each of the other 3 edges "is this along the path for the shortest route?" and when one is yes, goes to that next node and so on until the destination is reached. This is done for every pair of nodes on the network and the sum is the "search information" for the network, measured in bits, so that the "yes/no" questions are a base 2 to the search information value for a network (which can be very large 2^{10} or more for some of the complex MDPM models in chapter 5). Most process networks will not have completely "blind" searches performed on them, and in many cases individuals in the organization domain will have an idea of who to talk to in order to obtain information. However, as processes become much more complex and people working on a process start to become only distantly connected (i.e. the technician in the lab will not usually be connected directly to the physician) measures of search information can provide a sense for how difficult it is to find information in an MDPM process network. This is essential for many of the process improvement, change, and troubleshooting activities in complex processes that may be more difficult (or not happen at all) if the "horizon" of information about other elements of the process is limited as described for other networks in (Rosvall and Sneppen 2006). Change propagation in complex process networks is equally affected by the "information horizon" as a small change in one part of the process, which does not appear to have an immediate or local impact can have a very significant impact downstream, yet if the network is not easily searched identifying the impacts a priori may be difficult and similarly troubleshooting the network from the affected downstream element may be just as difficult. Each of the entropy metrics is described in the following tables.

Metric	I
Definition	Graph entropy is a measure of the randomness of graph structure

Formula

$$I(G) = -\sum_{i=1}^{\max d} h_i \left(\log_2 \left(h_i \right) \right)$$

where h is the degree distribution as a funtion of i, over all degrees

Equation 18 - Graph Entropy (Structural)

Application — A domain with high graph entropy (measured in bits) suggests less structure tot eh organization of the domain, which should correspond to higher complexity as the network can be in many "states" leading to less effective search and communication

Limitations — Highly fragmented domains will have relatively low graph entropy (most of the nodes have similar degree) but the overall process will be harder to manage. Randomness in the degree distribution may not necessarily correspond to "higher randomness" of the process. High graph entropy may also mean the process is less vulnerable. This metric also assumes all nodes are in a single "state" with randomness coming from the degree distribution only which may not always be the case.

Hypothesis — High graph entropy corresponds to higher complexity and harder to manage processes as the structure is not well ordered.

Related Work — (Costa, Rodrigues et al. 2005; Latora and Marchiori 2005)

Metric — I_S

Definition — Graph entropy (with states) is a measure of the randomness of the process not just in structure but also in state. This includes the idea of "states" of the elements in the graph such as machines that may be idle, working, or down.

Formula

$$I_S(G) = -\sum_{j=1}^{\max s}\sum_{i=1}^{\max d} h_{ij}\left(\log_2\left(h_{ij}\right)\right)$$

where h is the degree and state distribution as a funtion of i and j, over all degrees and states

Equation 19 - Graph Entropy (Structural with States)

Application An MDPM model or domain with high graph and state (i.e. each element can be in multiple states) entropy (measured in bits) should correspond to greater complexity and difficulty to manage as the system has a much greater number of configurations to be searched leading to less effective search and communication. Alternatively this definition could also be extended to replace states with "transactions" and conditional probabilities for the interactions of each domain with each other (i.e. not all information needed for every transaction, therefore probability of certain nodes being useful under different conditions changes).

Limitations The states of the process will likely be Bayesian in nature, as each state will be influenced by the state of another, suggesting that the entropy may be less than the estimate from independent states would suggest.

Hypothesis High graph and state entropy corresponds to higher complexity and harder to manage processes as the structure and state space is not well ordered.

Related Work (Frizelle and Woodcock 1995; Basole and Rouse 2008)

Metric ξ

Definition Average search information, or amount of information needed to identify the shortest paths between two vertices.

Formula

$$\xi = \frac{1}{|V|^2} \sum_{ij} -\log_2 \sum_{\{p(i,b)\}} \frac{1}{k_i} \prod_{j \in p(i,b)} \frac{1}{k_j - 1}$$

where k is the vertex degree, p(i,b) is the set of all shortest paths from i to b, and V is the set of vertices in the graph

Equation 20 - Average Search Information Metric

Rationale

High information requirements to identify shortest paths in a network correspond to greater difficulty in managing change.

Limitations

The amount of information does not account for "Bayesian" behavior, such as once a path has been found, it will be followed and the next one will more closely match shortest path.

Hypothesis

High information requirements likely make the control of a complex process difficult. The ability to easily search and find information across the domains of the MDPM (and ESM) models is essential to the rapid troubleshooting and management of change in complex products and processes. A high search information value both overall and for each element may be useful in understanding how to design processes such that they can be more flexible and easier to improve.

Related Work

(Rosvall, Minnhagen et al. 2004; Rosvall, Gronlund et al. 2005)

4 Model Exploration

Using the metrics developed in chapter 3 it is possible to explore the model in further detail and identify sets of metrics relevant to each of the domains. The sets of metrics and MDPM model features developed in this chapter are then applied to the demonstration case of MRSA surveillance in Chapter 5.

4.1 Process Complexity

Typical processes have linear graph structures that would appear to have relatively low complexity. In addition, processes usually have well-defined start and end points with directional flow between each of the activities. The basic graph properties such as graph density, number of edges, average degree, and average distance are usually a linear function of the number of activities. Delving deeper into a process, however, can uncover significant complexity. For example, the linear process A through M shown in Figure 56 would appear to have very little complexity if each of the activities can only happen in one way, or "state", and the processing times are identical.

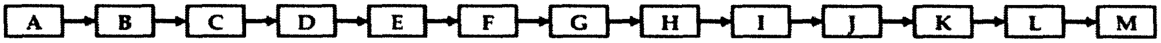


Figure 56 - Sequential Process

For simplicity the graphs used in the process model development examples so far have assumed single-states, but real processes can exist in multiple states and have variable processing times depending on the state and input. So for example the sample collection step for MRSA diagnosis will vary significantly depending on the location of the infection, the condition of the patient, and the type of patient attended to immediately prior. For simplicity, multiple states are not shown in this model, but could be added as additional matrices of the same dimension as the MDPM but each corresponding to a particular state. The increase in complexity for manufacturing processes from modeling states has been shown in (Zhu, Hu et al. 2008), who uses applies the ideas of state and graph entropy to describe the increase. Future extensions of the MDPM (and the ESM) model would likely require states and the modeling of multiple simultaneous processes, which introduces additional complexity. Multi-process multi-state complexity is discussed later in this chapter.

Certain conditions typical of real world processes can also introduce significant complexity to the process domain. For example, the idea of rework loops where an activity or series of activities must be repeated following a decision activity is represented in Figure 57. The corresponding activity on node graph will be topologically very different from the simple linear graph, adding additional edges, which increase graph entropy, density, and path length.

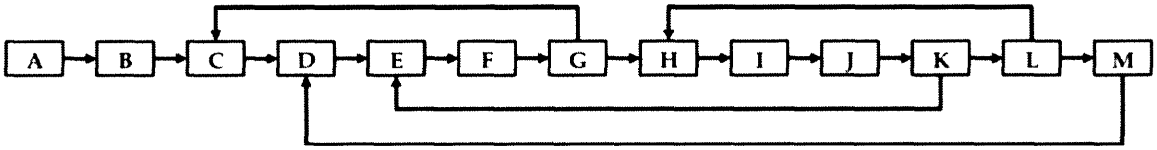


Figure 57 - Sequential Process with Rework

Similarly branch points can increase process domain complexity and eliminate "reachability" of nodes on different branches as shown conceptually in Figure 58, Figure 59, and Figure 60. Implied in these tree structures are decision rules associated with certain branch points. Although not all branch points need be decisions as it is possible that the product from one activity can be split between branches. Certain business process modeling languages explicitly model decision points as "connectors" such as EPCs. However for simplicity connectors are not modeled separately in this thesis.

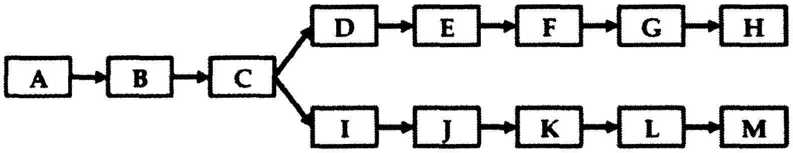


Figure 58 - Process with Single Branch

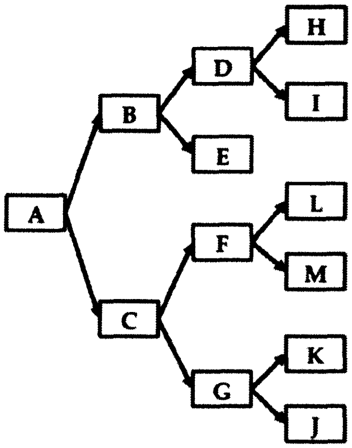


Figure 59 - Process with Multiple Branches

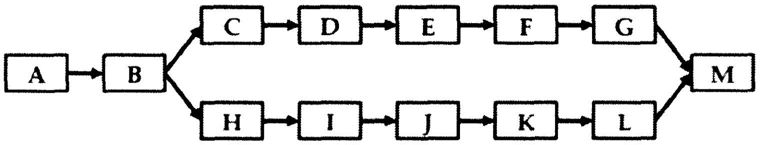


Figure 60 - Process with Branch and Rejoin

Process domain maps with greater numbers of tasks, paths, and loops are of course possible, but for the majority of diagnostic processes considered in this thesis, there are single start and end points and there is a single "objective" for each path in the process network. It may be necessary however to describe multiple "support" processes in the process domain, such as the gathering of production information, the preparation of reagents for use in an activity, or the training of personnel, and these processes can be disconnected in the process domain (but connected in other domains). In order to analyze the process domain, the set of metrics in Table 7 is proposed. The process domain also has distinct properties relative to other domains, as the density for any real world processes is unlikely to be very high (.05-0.2) as in practice this would correspond to enormous amounts of rework, branches, and re-processing not commonly found.

Metric	Feasible Values	Description
S_V	2-|V|	Number of activities: higher should equal greater complexity
S_E	1-|E|	Number of transfers: higher should equal greater complexity
Δ	0-1	Fraction of total possible transfers: higher = more complex
diam	1-|V|	Number of activities in longest process path (assuming no cycles)
$\bar{l}$	1-avg(l)	Average path length, longer if rework exists
B	0-1	Distance weighted fragmentation, long processes can seem disconnected from start to end
M	0-[$\|V\|^2$-2|V|+1]	Cyclomatic Complexity, more paths should = greater complexity
I	0-max(I)	Graph entropy (associated with structure), higher more complex

Table 7 - Proposed Process Domain Metrics

Typical sequential process graphs such as those depicted thus far will have densities less than 0.1 as can be seen from Figure 61 comparing 3 different process graph conditions.

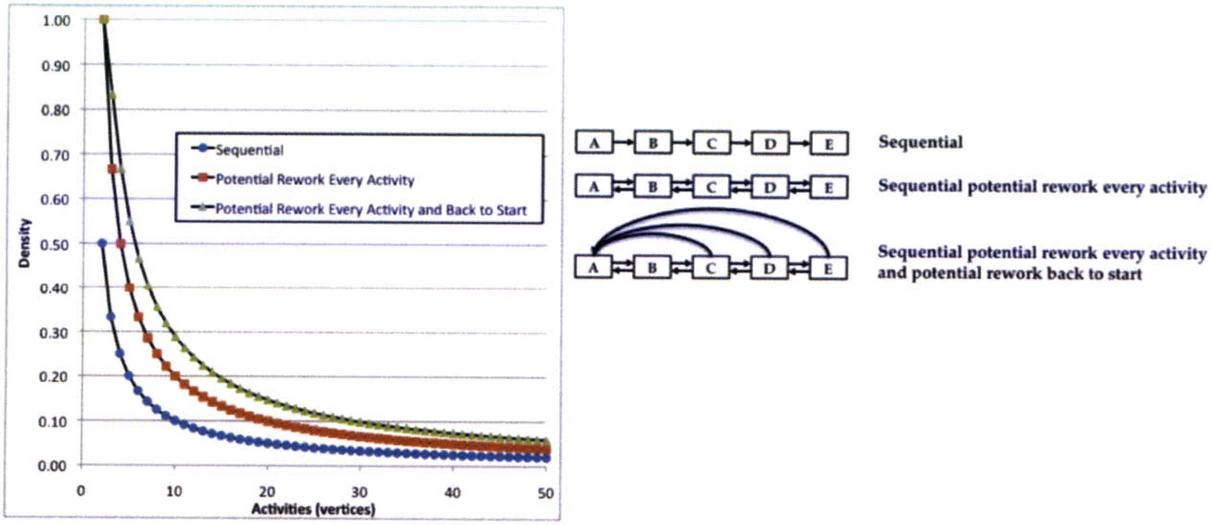


Figure 61 - Plot of Process Domain Density for Various Process Types

Process domain densities are in the majority of cases low, compared to networks where every node can connect to any other node, such as communications networks (the internet), or even social networks. Not every process step can logically be connected to another in the process domain, as for example a patient sample cannot go directly to the physician without actually going through the diagnostic process. Physical processes likely follow this pattern in the majority, although business processes where no "physical" steps are involved such as loan evaluation may have higher densities as they are more like communication networks (and in fact are processing information). The focus of this thesis however, is on physical healthcare diagnostic processes where activities must be connected in a certain order. However for all process domain networks, the number of states and paths can be significant and is an important measure of the true complexity of the process domain. States are not considered in this thesis, as they would require additional matrices to describe the state of each element as well as an additional physical object domain as in the ESM model to describe equipment and other infrastructure items in addition to people and information. However, modeling these additional states would allow for "behavioral" time-based simulation of the process as the state of elements (such as a machine that may be idle, under repair, or working) could be described. For healthcare diagnostic processes, the number of possible paths through the process domain is considered indicative of the "combinatorial" complexity faced by the organization and which must be captured in the work instructions. Thus, paths which are rarely used (i.e. a special request diagnostic using an older but still available technique) or potential paths (i.e. by mistake someone may skip an activity) must be dealt with by work instructions describing what to do in these rare paths, and by the organization being staffed to respond to them.

These metrics can serve as a guide to assess various properties of the process domain such that a designer could make tradeoffs between number of steps, alternative paths, and complexity as measured in various metrics. These process domain metrics must be used with caution as the process domain can contain disconnected activities (processes) that may only connect through other domains (i.e. individuals). The goal of the process model described in this thesis is to generate a complete view of the process and not just activities without the necessary work instructions and organization to carry them out. The information and organization domains are described next. These metrics are also only meaningful in the context of a matrix (graph) representation of a process such as the MDPM model or other MDM models.

4.2 Organizational Complexity

In contrast to process domain complexity, which is bound by the directionality and the largely sequential nature of many processes, the organizational domain can have a greater density and less directionality as many (but not all) of the links will be bi-directional. Additionally there may be many individuals that are not connected directly to an activity but which form part of the process either by managing people involved in activities, providing tacit knowledge, or who set policy that influences the work instructions. Traditional social network analysis metrics can be used to define the social relationships in this domain such as centrality, betweenness, path length, density, but also more sophisticated measures such as efficiency. An important concept of the organization domain is the full enumeration of all individuals required to manage a process, and is akin to stakeholder identification in current Enterprise Architecture methods as described in (Nightingale and Rhodes 2004). It is possible (and likely) in complex processes that the stakeholders are not directly connected to each other as part of the social network in the organization layer, but may influence policy on the information layer, which then influences other individuals, or they may be responsible for another activity on the process layer. For example, the technician preparing the reagents at the diagnostic vendor factory likely has no social network connection to the infectious disease physician treating a patient, but the two are linked through the diagnostic process and the actions of the technician can alter the behavior of the complex diagnostic process in unpredictable ways. What if the technician dispenses a slightly smaller amount into the kit, which in turn slightly decreases the sensitivity of the assay leading to a larger number of false negatives? This might lead the physician to lose trust in the diagnostic after a few close calls and start prescribing antibiotics without the diagnostic, expressing a lack of trust of the diagnostic through his or her social network, and eventually leading to widespread over prescription of antibiotics, which in turn can lead to increased resistance. This chain of events is one of many plausible unintended behaviors of complex healthcare processes, and the importance of mapping the interactions on the organizational domains is that these unintended scenarios can be better managed. To continue the example, what if a "process owner" as described in chapter 3, were introduced into the organization domain. This person might be able to connect the technician and the physician in far fewer steps, allowing for the possibility that the under-dispensing by the technician might be recognized as the root cause for the underperformance of the diagnostic and demonstrate this to the physician through a structured set of experiments. Good processes allow for organizations to rapidly identify and then "swarm" problems bringing multiple resources (and necessary stakeholders) to bear on a problem as described in (Spear 2008). But this is only possible if problems can be quickly identified, resources rapidly deployed to find the root cause of a problem, information about the root cause is added to the explicit body of knowledge, and process changes put in place to ensure it never happens again. Good processes should all share certain characteristics in the organization domain that facilitate this.

And what might these characteristics be? Well the average distance between nodes is likely small in high-performance Enterprises, but without too many links. This might be similar to the characteristics of "small world" networks where a few additional links can connect local neighborhoods to greatly reduce the average distance without the expense of too many connections linking vertices directly as in a complete graph. However, each of the neighborhoods must likely be very densely connected to deal with the enormous amount of information in managing a complex process, which in a computer network might require the use of "hubs" to reduce the number of connections any single node experiences. But real human "hubs" cannot have enormous numbers of connections (edges) as in the extremes of power law distributions. Real "process" networks have caps on how many simultaneous streams of information, actual work (in performing an activity), and communication (with others in the organization domain) can be sustained. This "cap" on the degree of vertices in a process network likely applies to all elements including the activities and the information elements. An activity with too many connections should be expanded into several sub-activities as too many individuals, activity inputs, and work instructions interacting simultaneously suggests a lack of control as no single individual or work instruction can be identified as "controlling" the activity. Likewise a work instruction in the information domain with too many individuals providing input, derived from too many other work instructions, and covering several activities likely means there are significant ambiguities and lack of "control". And how many is too many? This will depend on the "bandwidth" available but for individuals in process organizations the maximum interactions with other "nodes" would appear to be in the tens to perhaps a maximum of 100s, but very rarely in the 1,000s and above needed for a power law distribution.

How can this be resolved with human elements in the process? Hierarchies is one answer. These are commonly found in many organizations, and while much maligned serve a useful purpose in balancing the "small world" reachability in few steps with overload at a hub. For example, a single manager with 24 direct reports would likely be "overloaded" with just brief 30-minute meetings already taking 30% of a 40-hour workweek before doing any "actual" work. Dividing those same direct reports across 3 managers however reduces the number of interactions each one has to 9 (7 direct reports + 2 for each of the other managers), a much more manageable number. Note that extending the hierarchy in traditional "bureaucratic" ways does not reduce path length, but rather increases it. Rather, the idea here is that for process networks, "hierarchical" hubs are needed to attenuate hub overload as shown in Figure 62. These "hierarchical" hubs in this process model replace single vertex hubs that might be found in other "power law" networks. And empirically, these hierarchies are real features of processes, where coordinators, supervisors, managers, and directors, serve (when properly used) to reduce the individual degree of a hub member, but can effectively appear to be a large hub with only a small (1 hop in this case) penalty. In Figure 62 there are 24 edges associated with each hub, but the degree of the single vertex hub is 24, versus the hierarchical hubs whose average individual vertex degree is 9 (although with a 1 hop penalty for traversing the hub) and 8 with a true hierarchical hub (although the hop penalty increases to 2).

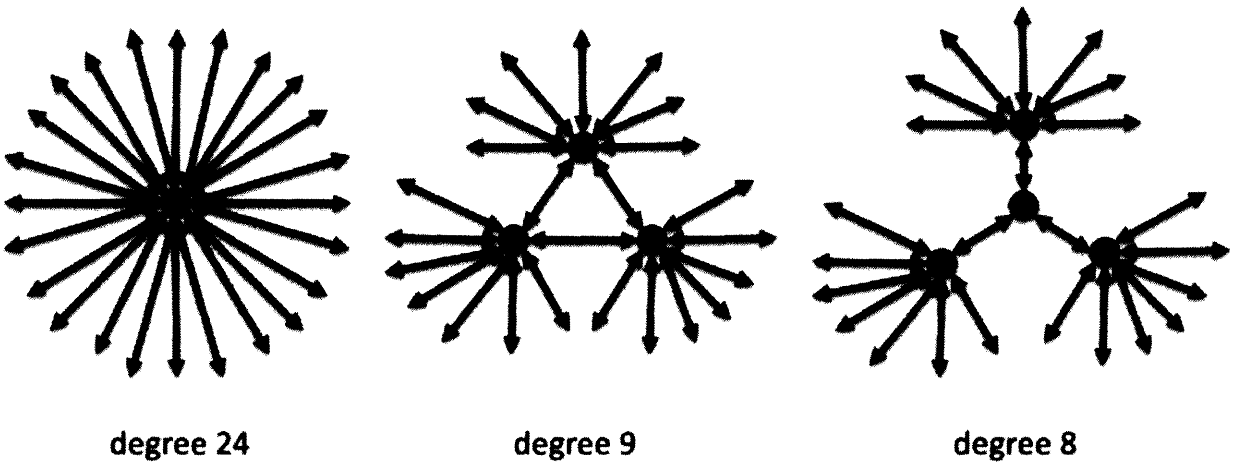


Figure 62 - Hub Structures

In order to analyze the organization domain, the set of metrics in Table 8 is proposed:

Metric	Feasible Values	Description
S_V	2-|V|	Number of individuals: higher should = more complexity
S_E	1-|E|	Number of connections: higher should = more complexity
Δ	0-1	Fraction of total possible connections: higher = more complex
$\bar{d}$	0.5-|V|-1	Average degree for a vertex, higher = more complex, less bandwidth per connection however
diam	0-|V|	Longest path between vertex pairs. Longer means communication more difficult
$\bar{l}$	0-$\bar{l}$	Average path length: lower = faster communication but at some expense to maintain connections
B	0-1	Distance weighted fragmentation, higher means more parts of the organization are "isolated"
F	0-1	Fraction of nodes that cannot reach each other, which means they can only reach each other indirectly though work on the process
Cn	1-|V|	Number of weakly connected components, how well connected is the network.

M	$0-[\lvert V\rvert^2-2\lvert V\rvert+1]$	Cyclomatic Complexity, number of potential communication paths
ε	0-1	Efficiency: higher = faster more direct communication but at some expense in terms of short distances.

Table 8 - Proposed Organization Domain Metrics

Because the elements of the organization domain are individuals, the properties of the domain are different than the process or information domains. For one, this organization domain has a much stricter "cap" on bandwidth, although as discussed previously, the other domains have similar caps. Ultimately the objective of this domain is to capture the full range of person-person communication needed to manage a complex process. And it is likely that the structure of this domain has a very strong influence on the structure of the process and information domains. For example, a hospital microbial diagnostic process staffed primarily by technicians who are familiar with older (culture-based) diagnostic technology, and who are not strongly connected to key stakeholders in the hospital organization, may not be able to absorb rapid change very easily. In contrast, a laboratory group with a highly connected lab manager and technicians that participate actively with physicians in comparing results will be more likely to absorb rapid change and to more easily resolve problems with a complex process by rapidly identifying problems and finding the necessary resources to solve them. The metrics described here can be used for this purpose, such as the average path length or efficiency indicating how well connected the organizational domain is.

Much of the additional complexity "hidden" in traditional process models is contained in this domain in the set of interactions necessary for the smooth operation of a process outside of the "routine" operations. While the direct interaction of a technician with an activity is a necessary part of the process, the occasional interaction of an expert in the activity (perhaps the activity designer) with the technician can improve the process by providing direct feedback from the technician or providing insight from the designer. These "occasional" interactions are not typically captured in process maps but do in fact occur and form the basis of much of the innovation, process improvement, and troubleshooting that occurs in real processes. Formal, but perhaps unrecognized, interactions must also be captured such as the organizational domain connections (or lack thereof) between insurance reimbursement analysts working for an insurer and the physicians and laboratory technicians on the front lines of medicine. The actual complexity of the organizational domain for healthcare processes is significantly greater than comparable processes in manufacturing due to the highly fragmented Enterprise of health care. A single hospital must produce a wide range of services from surgeries to checkups to diagnostics as opposed to a semiconductor factory that has been highly optimized to produce a single type of product.

There are a number of other measures from social network analysis that can be applied to better understand the MDPM model, such as centrality metrics like betweeness and closeness discussed in chapter 2. These will be used in Chapter 5 to identify MDPM nodes that are most "connected" to the rest of the process but not in a global network sense. The graph-theoretic definition of betweenness centrality is given in Equation 21 and closeness centrality is given in Equation 22.

$$C_B(k) = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$$

Where g_{ikj} is the set of all paths from i to j that pass through node k, and g_{ij} is the set of all paths between node i and j.

Equation 21 - Betweenness Centrality

$$C_C(k) = \frac{|V|-1}{\sum_{k \neq j} d_{kj}}$$

Where dkj is the set of geodesic shortest paths and V is the set of all nodes

Equation 22 - Closeness Centrality

Both of these centrality measures can be used to identify key elements in the MDPM model as will be shown in chapter 5.

4.3 Information Complexity

The main objective of the information domain is to represent the explicit information needed to manage the process. This can take the form of work instructions, design documents, operational logs, process data, metrics, policies, and other documents needed to support the process. The lack of many "elements" in the information domain for a complex process is likely a potential problem as it means the process is not sufficiently documented, monitored, or controllable. This also means that the process knowledge largely resides as tacit information in the organizational domain. Excessive numbers of elements however may unnecessarily increase the process complexity as each of these elements must be "processed" in some fashion either by an individual in the organizational domain or connected to another information domain object via an information mapping. It is worth noting that much like the connections in the organizational domain, the connections (edges) in the information domain represent links of various kinds between objects. The objective is to show the relationship between them, to in a sense understand the "genealogy". The edges represent various kinds of mappings from one information set to another, such as data to algorithms, readouts to information summaries, and policies to detailed day-to-day protocols. Much like all edges in the organizational domain represent communications, which can also be considered mappings of one data set (a person's knowledge) to another. This idea that all edges in each of the domains represent the mapping of information from one set to another is what enables the construction of the overall MDPM model which in the initial prototype version in this thesis considers the edges between all vertices to be information mappings including the process domain edges. Future refinements to the MDPM model could differentiate between the edges in order to show different types of relationships but this would require additional matrices each of the same dimension as the MDPM model they are derived from to show the "attributes" of each of the edges in the "master" adjacency matrix. This refinement could add significant richness to the model.

The information domain can provide significant insight into "hidden" information not commonly represented in process models. Much like the organizational domain, real world complex processes require far more information than the work instructions (protocols) associated with each activity. The hospital policies, FDA policies, the performance specifications for diagnostic kits, performance comparison documents of diagnostic kits, are all needed to manage the process. And understanding their interaction, and most importantly the information genealogy of work instructions directly associated with an activity is critical to managing the process. For example, imagine the case of having to troubleshoot a complex diagnostic process in healthcare. The first step might be to review the protocols to understand what variables are the most important. However the relationship between variables and experimental results for these are likely contained in the design documents the protocols were derived from. As part of the troubleshooting It might be important to understand the design envelope for a particular process step, which would then require access to the optimization documents associated with the protocol or perhaps vendor documentation indicating the performance limits of that particular process activity. In addition, it will be necessary to review any metrics or operator logs associated with that process activity, which requires understanding not just the logs or metrics but also the algorithms the metrics are derived from. Again all of this information is not visible in the very first level of a process representation but the policies and other documents can have a very significant impact on how the process is run, and most importantly is changed and improved. Much like the influence of a particular organizational structure on the structure of the process, the structure of the information domain can have a significant impact on the structure of the process.

The ideas of information entropy are most useful in the information domain, in particular measures associated with search. For example the average search information, or the amount of information needed to identify the shortest path between two vertices can be used to identify nodes that have the greatest "difficulty" in reaching targets through the particular network and will therefore not have good visibility over the impact of any changes or problems in their particular part of the process on the rest of the process. But why go through the trouble of enumerating all of the information associated with a process? The answer is that when the process is operating normally only the work instructions and some direct tacit knowledge from the individuals directly connected activities are needed. But as soon as changes, troubleshooting events, improvements, or new technologies are introduced there is a necessary search through the information associated with the process to understand the impact, manage the change, or find the root cause. This is why the enumeration and genealogy of the information is so critical: it provides the network to be searched. Table 9 contains some of the proposed complexity measures for the information domain, with a similar set of metrics to the organization domain but with different interpretations of these.

Metric	Values	Description
S_V	2-|V|	Number of information elements: higher should = more complexity
S_E	1-|E|	Number of connections: higher should = more complexity
Δ	0-1	Fraction of total possible connections: higher = more complex
$\bar{d}$	0.5-(|V|-1)	Average degree for a vertex, higher = more complex, information is highly interdependent
diam	0-|V|	Longest path between vertex pairs. Longer means search is more difficult
$\bar{l}$	0-$\bar{l}$	Average path length: lower = faster search but at some expense to maintain connections
B	0-1	Distance weighted fragmentation, higher means more elements of information are "isolated"
F	0-1	Fraction of nodes that cannot reach each other, which means they can only reach each other indirectly though work on the process
Cn	1-|V|	Number of weakly connected components, how well connected is the network.
M	0-[$\|V\|^2$-2|V|+1]	Cyclomatic Complexity, number of potential search paths
ε	0-1	Efficiency: higher = more complex but more efficient

Table 9 - Proposed Information Domain Metrics

4.4 Multi-Domain Complexity

The projection domain, which is also the entire graph described by the full MDPM matrix, captures all of the connections between elements of the different domains and thus can have significant complexity as measured by the metrics developed thus far. For the purposes of analysis the objects and links in the full MDPM matrix (the projection domain) are treated as equal although a relatively straightforward model extension could add rules and classes for the elements and relationships between them as discussed previously. But in order quantify the real complexity that exists in processes and measurement of the projection domain containing all of the elements in the information, organization, and process domains is a useful proxy. The projection domain will also, by definition, have a greater complexity than the domains it is derived from and metrics calculated on the full matrix describe the overall process properties including proxies for complexity. Also because many of the "hidden" connections between elements of the process are shown in the projection domain the complexity of this domain will be higher but more representative of the "true" process complexity.

The projection domain network is an abstraction of reality but measuring properties such as density, average distance, degree distribution, and reachability can provide significant insight into the performance of the network associated with the process. In addition, one can use the projection domain to identify "weak points" in the process where process activities are unmonitored, difficult to reach (in a network sense), individuals are overloaded, work instructions are not being properly updated to reflect changes elsewhere in the process, and other problems. These will be further discussed later in this chapter, in showing how this model can represent traditional process problems. The metrics used to analyze the projection domain are similar to those already described and shown in Table 10:

Metric	Values	Description
S_V	2-$\lvert V\rvert$	Number of elements: higher values should correspond to higher complexity and difficulty for change as every node must be "updated" or "pinged" in the case of troubleshooting
S_E	1-$\lvert E\rvert$	Number of connections: higher values should mean the network is more easily searched but come at some cost of complexity
Δ	0-1	Fraction of total possible connections: higher values should correspond to better communication and search but at some cost to maintain connections.
$\bar{d}$	0.5-($\lvert V\rvert$-1)	Average degree for a vertex, higher values may correspond to bandwidth limited processes that do not have more "bandwidth" to deploy on changes and troubleshooting
diam	0-$\lvert V\rvert$	Longest path between vertex pairs, a high value may correspond to long search and communications paths making change and troubleshooting difficult
$\bar{l}$	1-$\bar{l}$	Average path length: lower = faster search but at some expense to maintain connections
B	0-1	Distance weighted fragmentation, higher means more elements are "isolated"
F	0-1	Fraction of nodes that cannot reach each other, which means they can only reach each other indirectly though work on the process
Cn	1-$\lvert V\rvert$	Number of weakly connected components, how well connected is the network.
M	0-[$\lvert V\rvert^2$-2$\lvert V\rvert$+1]	Cyclomatic Complexity, number of potential communication paths
ε	0-2/V-1	Directed efficiency: larger = more complex but more efficient
μ	0-1	Network vulnerability, high if it is composed of a few hubs

ξ	0-ξ	Search information needed: higher values mean a network is harder to search and will make change and troubleshooting difficult
I	0-max(I)	Graph structure entropy: higher values may suggest the network is not well structured and may be vulnerable to failures of individual elements

Table 10 - Proposed Projection Domain Metrics

4.5 Multi-Process Complexity

One of the biggest contributors to process complexity is the need to run multiple processes simultaneously. This is especially common in the healthcare environment where nurses, physicians, laboratory technicians, and other staff are responsible for multiple process activities, their work instructions, and the necessary organizational interactions to perform them. The MDPM model can be extended to include other processes and ascertain the total "enterprise" complexity associated with running all of these simultaneous processes. Previously in this chapter the concept of "bandwidth" was discussed in the context of a single process, where a certain maximum number of connections for a given node (its degree) should not be exceeded in real processes. In a hospital enterprise for example with multiple processes the same idea applies, such as a nurse that is in the organization domain for two processes will have a higher "effective" degree as connections must be made to activities, work instructions, and individuals in both processes. Similarly information domain elements, such as hospital policies, and other "enterprise" documents will likely be shared, as shown conceptually by the overlaps in Figure 63, amongst multiple processes and these connections will have a higher degree for those nodes. Note that it is possible, and even likely that certain information domain elements can have a high degrees, but these are "one step removed" elements such as references or policies rather than direct work instructions (associated with an activity) which should closely match the activity number and degree.

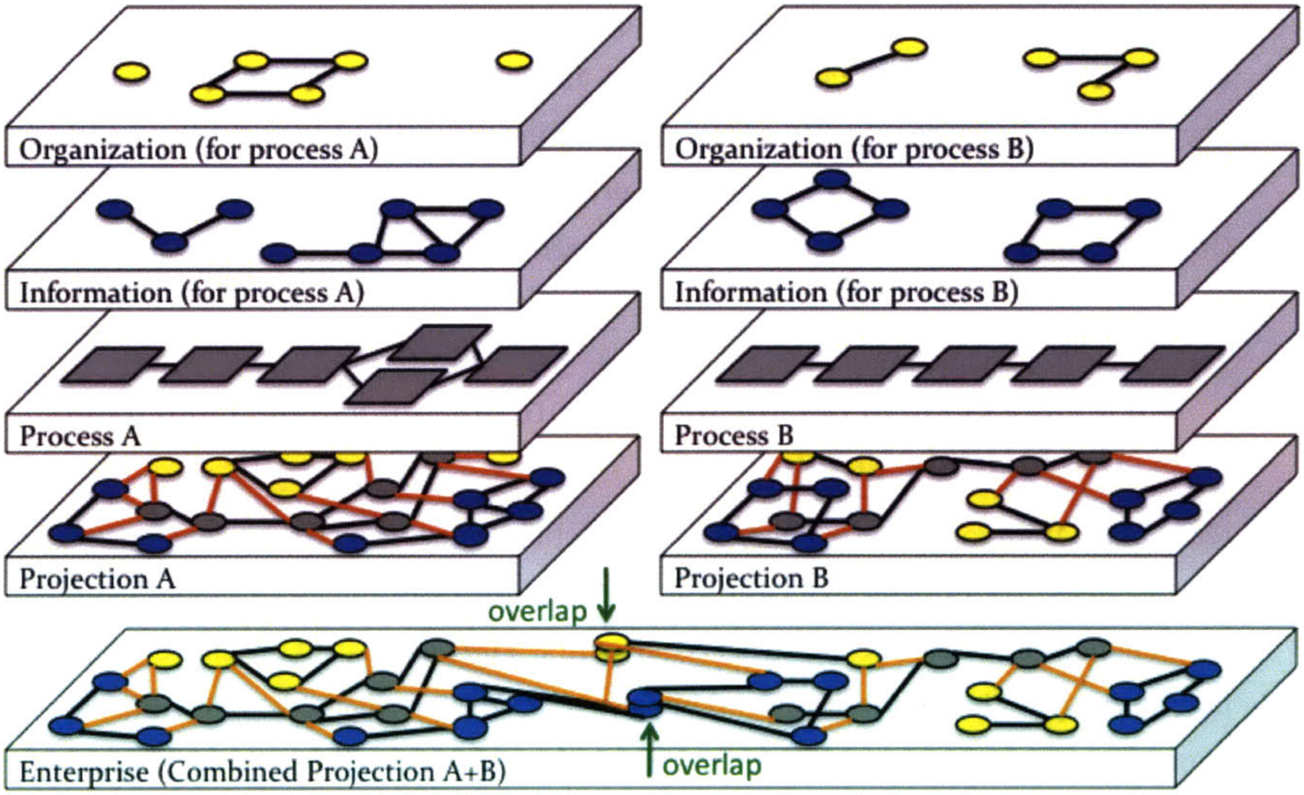

Organization (for process A)
Organization (for process B)
Information (for process A)
Information (for process B)
Process A
Process B
Projection A
Projection B
overlap
Enterprise (Combined Projection A+B)
overlap

Figure 63 - Conceptual Model of Multiple MDPM Model Interaction

While fully exploring the entire set of processes in the hospital is beyond the scope of this thesis, identifying which elements of a given process have significant demands from other processes is critical to understanding the performance of a single process. So for example one could imagine performing sensitivity analyses on key nodes of the projection domain to understand the process performance subject to varying "bandwidth" demands of these nodes. It is likely that for these key nodes a "fault tolerant" hierarchical and highly connected node structure would be better to replace a single node with high demand or potential for failure.

Additionally, this same idea of multi-process analysis could be used to better understand how to transfer processes to other locations. For example, the clinical microbiology organization of a rural hospital in Oregon will likely be very different than the organization of a research hospital in Boston, and yet it is important that the Oregon hospital receive up to date process and methods to combat rising antibiotic resistance. It is possible to use the MDPM model to map the process, information, and organization domains of the teaching hospital onto the Oregon rural hospital organization to identify which individuals need to talk to which and perhaps to re-arrange the process and information domains of the Oregon hospital to better match their (likely fixed) organizational structure.

4.6 Process Improvement Methodologies Viewed Through the MDPM Model

Many of the traditional process improvement methodologies can be represented using the MDPM model. For example, the basic lean idea that processes should be simple, but can and should continuously be improved requires first a process map of the entire value chain in terms of activities (process domain) but also of the many stakeholders (organization domain) and furthermore an understanding of what information flows between activities and individuals (information domain). Process improvement necessarily requires observation and identification of problems as a first step and organizations that do not have good means to rapidly identify problems and make sure they are resolved quickly have little hope of making improvements in complex processes. So, for example, a process where the technicians performing the activity have no communication to troubleshooting resources (perhaps a design team) and do not document their process observations would show up in the organization domain as disconnected individuals and the adjacency matrix would show that the technicians are not "reachable" from R&D or even perhaps senior management such that corrective actions may be taken. Contrast this to what an MDPM model would look like for a Toyota factory, where all of the "frontline" workers are reachable within a few nodes of senior management and "troubleshooting" resources such that any problems can be identified, escalated, and resolved quickly. In addition a Toyota factory MDPM model would show the existence of monitoring metrics documents and operational logs in the information domain and connected to the majority of the workforce such that performance and problem resolution are visible. Likely the Toyota factory would have a short average distance, higher density, but a much higher average efficiency such that problems throughout the process can be readily "reached" in few steps (in any of the 3 domains).

The model can also be compared to some of the "high-velocity organization" characteristics of market leading companies and organizations that deal with complex processes and products as described in a recent book by (Spear 2008), "Chasing the Rabbit" as four core capabilities. These are described next from the perspective of the MDPM model.

<u>Capability 1:</u> **Specifying Design to Capture Existing Knowledge and Building in Tests to Reveal Problems**

Applying best practices to processes and ensuring that their goals and the supporting organization are well aligned requires the identification of best practices, the organization, and the specific design documents for the process. The MDPM model can be used to identify the necessary communication links such that best practices are actually transmitted. For example, an organization may have a series of "experts" that have built processes before and ensuring that in the organization domain, these "experts" are reachable over short distances from the individuals at the front lines (those directly connected activities) will make possible the transmission of best practices. But as Spear points out it is not only about best practices but also about clearly specifying the work to be performed, the manner in which the work is to be performed, and what outcomes are expected. The information domain can be used to identify all of these "information" elements, or to identify their absence such that processes can be better specified from the start and all stakeholders can have visibility into these specifications to facilitate alignment. A corollary of this capability is that the process infrastructure should enable the rapid identification of problems as they occur which means a network with a high information transmission efficiency. Processes with long average distances in the projection domain will have difficulty in rapidly identifying problems as the number of steps between any two nodes (either information, organization, or process) will make it hard to transmit accurate (undistorted) information in a timely manner. Most importantly for the management of complex processes is that the individuals and work instructions directly on the front lines (directly associated with an activity) must be "supported" by the process infrastructure to provide short network distances to report problems and resolve them. This includes connections across the various activities as well as to the "design" and "troubleshooting" resources.

Capability 2 Swarming and Solving Problems to Build New Knowledge

Once a problem has been identified, the rapid deployment of resources to resolve the problem, identify its root cause, and document (as explicit knowledge) its resolution is essential to sustained process improvement. However, this is only possible if the additional resources are connected over relatively short network distances to first the identification of the problem and second the location of the problem in network space. For example, if no specific additional resources are shown in any of the model domains it is clear that resolution involving learning is not possible. But even if these resources are available they must be reachable in the projection domain network and any knowledge created must explicitly exist in the information domain as "resolution" logs or other information elements readily accessible to other elements of the network. Timely resolution of problems is also critical and complex processes as the "evidence" can have a very short lifespan making lengthy or delayed investigations problematic and inconclusive. This emphasis on rapid resolution requires high efficiency networks with relatively short path lengths, which must of course be balanced against individual bandwidths that cannot practically be exceeded but maybe balanced by hierarchical structures. For example, a problem with a particular diagnostic reagent might not normally involve a physician downstream of this problem and yet the assay quality can directly impact the physician. "Swarming" this problem might require the involvement of a physician, a technician, a lab manager, and perhaps a "process owner" to ensure the nature of the problem was identified and the resolution was both understood and satisfactory to the end-user, the physician. These types of connections can be explicitly documented and modeled in the projection domain.

Capability 3: Sharing New Knowledge Throughout the Organization

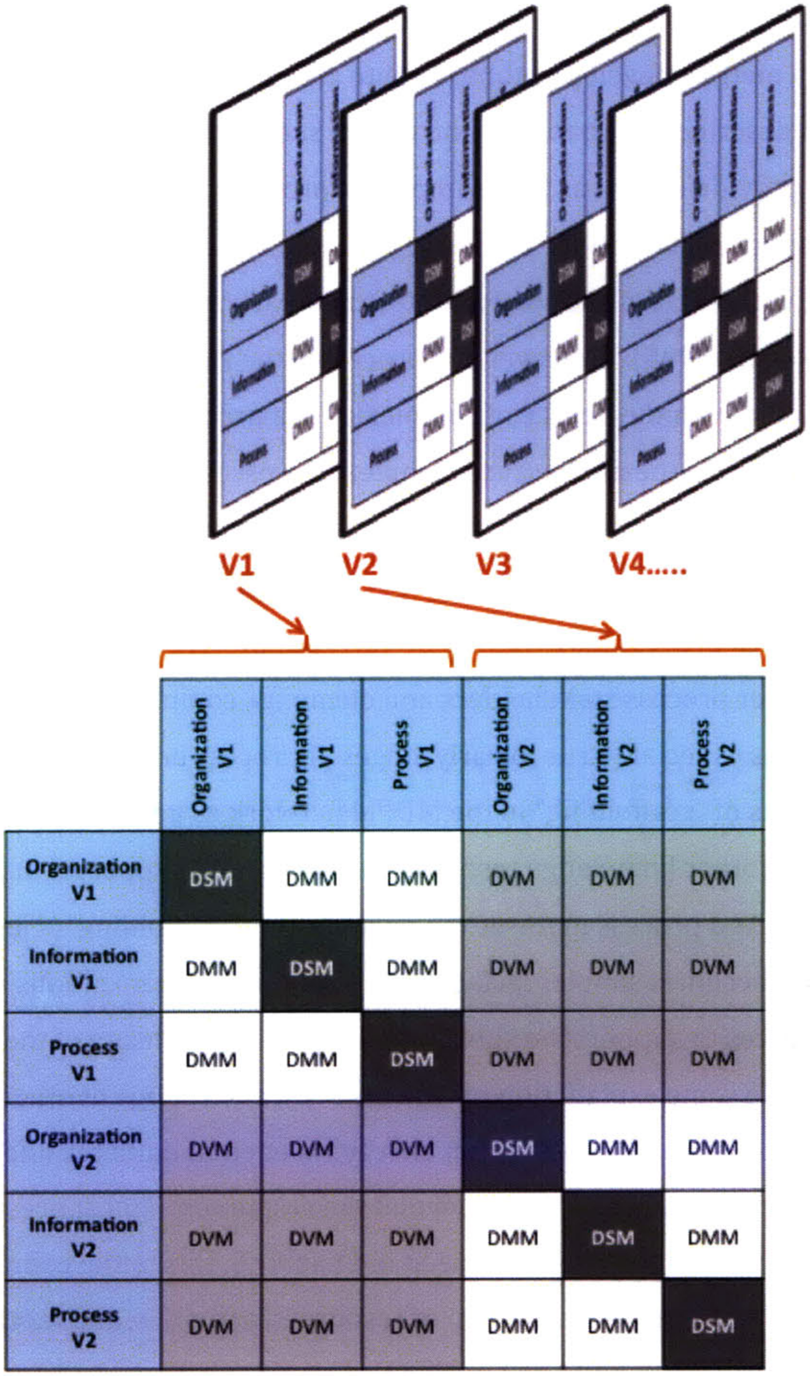
Once the root causes of problems and their resolution have been found it is essential that they be communicated and shared throughout the organization. While the MDPM model developed in this thesis focuses on a single process, the model can be used in an "enterprise" fashion to ensure local knowledge, both tacit and explicit, is shared across multiple processes in the enterprise as conceptually shown earlier in this chapter. However, this is important even across a single complex process where knowledge of the "local" resolution of a problem in an upstream activity may benefit the resolution of problems and downstream activities. In addition, the critically important work of transferring new and complex processes to other locations (i.e. between hospitals) requires that a formal "mapping" of the two processes be done to ensure knowledge will actually be transferred and shared. For example, a new diagnostic process developed in a teaching hospital with significant resources should be explicitly mapped onto a community hospital, which may not have the same resources, organizational structure, or distribution of skills and where some people may do double duty or some knowledge areas may not exist. The MDPM model can help identify these "holes", or areas that must be supplemented to facilitate the knowledge transfer.

Capability 4: Leading by Developing Capabilities 1, 2, 3

The greater the emphasis in network efficiency, reachability, and bandwidth utilization the better the capabilities of the organization will likely be. Continuously working to ensure that individuals on the "front line" of the process can rapidly flag problems and swarm them with resources from the "process infrastructure" will help ensure the process continues to improve. Further rationalizing the process to make sure individuals are correctly assigned (have sufficient bandwidth) and work instructions have the proper mechanisms for updating as changes happen is also critical and can be done with a network view of the process.

4.7 Process Evolution

The process examples and analysis developed thus far have mainly focused on static processes, but the time evolution of processes can also provide great insight into their structure and properties. The dynamic nature of processes has been implied in the previous section, which describes process improvement and change in existing processes to produce new processes. Mapping these changes can be readily accomplished by expanding the MDPM model matrix to include previous versions of processes and the "genealogy" of elements from one version to another. Figure 64 shows the conceptual evolution of a process from an initial version (V1) to a fourth (V4) and successive versions of the MDPM representation of each version of the process. Each of the process versions is captured in a "master" MDPM matrix, which includes all of the domain elements from each version of the process but also mappings from one version of the process to another, which are called "Domain Version Matrices" (DVM). These DVMs map the relationship of each element in a MDPM version to another. As processes evolve the dimension of the MDPM matrix will likely change as elements are added or subtracted in each of the domains and a master MDPM including all previous versions of a process would have dimensions equal to the sum of dimensions of each MDPM version contained within it. Also note that the DVM matrices are shown on both sides of the diagonal indicating version mapping relationships can exist in both "forward" and "reverse" directions. A “reverse” mapping might be used to show “reuse” of elements from previous versions or to show references to information elements in previous versions that current work instructions may impact if the previous version is still in use.



	Organization V1	Information V1	Process V1	Organization V2	Information V2	Process V2
Organization V1	DSM	DMM	DMM	DVM	DVM	DVM
Information V1	DMM	DSM	DMM	DVM	DVM	DVM
Process V1	DMM	DMM	DSM	DVM	DVM	DVM
Organization V2	DVM	DVM	DVM	DSM	DMM	DMM
Information V2	DVM	DVM	DVM	DMM	DSM	DMM
Process V2	DVM	DVM	DVM	DMM	DMM	DSM

Figure 64 - Evolution of Process Versions in Time and Representation as Domain Version Matrices

The evolution of processes can also be tracked with all of the metrics developed thus far to provide a rate of change for processes along key dimensions. This might be particularly helpful to understanding the dynamics of process lifecycles from original creation through development and into maturity as described in (Utterback 1996). For example, mapping some of the key metrics developed thus far onto process lifecycles like the Pilkington float glass evolution described in Chapter 2 and others such as the evolution of petrochemical refining might provide insight as to the impact of process complexity and structure on process innovation and adoption. Extensive literature exists as to the pattern of adoption for innovation, with a classification of adopters into early, middle, and late as notably described in (Rogers 1995). However, for processes the rates of adoption may also be strongly associated with the degree of complexity of a process as measured by the set of metric proposed in this thesis. Because complex processes require such significant resources spread over multiple groups, disciplines, locations, and resource levels, adoption may be less a function of the "desire" to adopt than a "feasibility" to adopt. Individual groups in a complex process may want to be early adopters, but without the coordination with all of the other process stakeholders and elements, complex process adoption may not be feasible. This is especially true in early stages of process development and adoption due to the need for lots of "scaffolding" in the MDPM network as additional design, optimization, and other knowledge resources are needed to support changes and troubleshooting. As a process matures the need for this "scaffolding" is lessened and the process network becomes simpler facilitating adoption for organizations without the additional resources to support the substantially higher complexity of the initial process. This might provide some explanation for the resistance to change of other complex processes in healthcare as the "activation energy" needed to build the additional scaffolding to manage changes is often beyond the resources of most organizations unless it is specifically provided for. As will be seen in Chapter 5, this initial complexity can be very high and the necessary combination of resources, activities, and motivation to build a new process can be hard to build and manage.

5 Case Study Model: DNA Sequencing Based MRSA Surveillance

In order to apply and test the MDPM model under real-life conditions a test case was sought that would have real-world impact. It would be difficult to evaluate the usefulness of a methodology such as the MDPM method without a real world case. As described in Chapter 2, MRSA infections are a serious threat to the health care system and yet molecular level surveillance has seen limited implementation despite the availability of many of the necessary technologies. While certain gene focused MRSA surveillance technologies such as PCR are starting to become more widespread their adoption has been relatively slow and more importantly do not provide a full picture of the genetic diversity in MRSA as they can only confirm the existence or absence of selected antibiotic resistance genes. Whole genome sequencing technologies can provide a full picture of the genetic diversity necessary for surveillance but many of these technologies are still in the early stages of their development requiring complex processes and infrastructure to support them. Yet, these technologies are also evolving very rapidly and while they may be complex, they are rapidly becoming cost feasible and even cost competitive with existing clinical microbiology surveillance processes. And with the rapid rise in MRSA infections and the enormous healthcare cost associated with them, the need for whole genome DNA sequencing surveillance processes is becoming critically necessary. The objective of this thesis was thus to design, develop, and implement a prototype whole genome DNA sequencing-based MRSA surveillance process for The Brigham and Women's Hospital, a large 746 bed research hospital in Boston.

However, this complex process spans numerous organizations, knowledge domains, activities, and information beyond the hospital most notably, the Broad Institute a leading genomics research center, but also various MIT departments, numerous technology vendors, and other organizations associated with MRSA, genomics, microbiology, and healthcare. Beyond being just a data gathering and modeling exercise, this process was actually built and implemented providing significant value to clinical microbiology research and our understanding of MRSA. However, in also being a real process, it has provided enormous insight into the MDPM model and the primary engineering systems research objective, which was to develop an analytic methodology for complex process design and analysis. This chapter details the design, development, design iteration, implementation troubleshooting, and initial pilot production of the MRSA DNA sequencing surveillance process from the perspective of the MDPM model and also develops the associated methodology to use the model.

5.1 Design Methodology

The MDPM model developed in Chapter 3 requires an associated methodology to be useful in analyzing complex processes. Borrowing from the CLIOS system analysis process described in (Dodder, Sussman et al. 2004) and Engineering Systems Analysis for Transformation (ESAT) described in (Nightingale 2009), a methodology is developed for process analysis using the MDPM model, called the MDPM methodology. This methodology follows some of the initial CLIOS methodology steps described in Chapter 2 to define the process, identify stakeholders, enumerate the process elements, and identify connections between elements. While CLIOS is mainly focused on systems, the key CLIOS step of "Identifying links among components and organizations" is also a key element of the MDPM process analysis methodology and the basis for the MDPM model. However, the MDPM methodology differs in that it includes steps to identify feasible processes and reduce the size of the design space to manage the complexity of the MDPM model to be analyzed. The analysis of the MDPM model itself also differs from the CLIOS methodology in using MDM and other network analysis tools. However, significant opportunity exists for integration of the MDPM methodology developed here specifically for processes with the broader systems analysis perspective of CLIOS. Similarly, the ESM model offers significant potential to capture many of the elements described by CLIOS as additional elements in the matrix although work would still be needed to capture some of the nuances of CLIOS such as different classes of links. These ideas will be further discussed in Chapter 7 indicating future research directions. The MDPM methodology is described in table Table 11 with notes describing the rationale of each step.

MDPM Methodology Step	Rationale
1.) Description of process, stakeholders, and goals	The first step in designing and analyzing a process is to describe its purpose, stakeholders, and general characteristics. Similar to CLIOS System Issue and Goal Identification and ESAT System Description and Stakeholder Identification.
2.) Identify macro process activities and define boundaries	Objective is to identify the macro level activities for the process described in step 1. Part of specifying the design space is also to define the system boundaries (or what will not be modeled). Similar to CLIOS Identification of Major Subsystems.
3.) Identify potential process alternatives and design space	The objective of this step is to identify all potential process design alternatives that could satisfy the goals but at "macro" level without detailing all of the elements in each. This step requires searching through available processes and technologies to identify existing processes that might be adapted, creating new ones from existing elements, or if none are available putting "placeholders" for missing elements corresponding to activities in step 2.
4.) Identification of design specifications and constraints	This step captures stakeholder preferences, physical constraints, and overall process specifications as set of variables in a vector, which can be used to numerically describe the properties of each macro process element.
5.) Selection of macro level feasible processes against design specification requirements	The network composed of the macro processes identified in Step 2 can then be analyzed using the property vectors, which define a "cost" to traverse the network. The minimum cost flow problem is solved for the network to find the most feasible process combination according to the overall system specifications and constraints.

MDPM Methodology Step	Rationale
6.) Enumeration of process elements onto detailed process map for selected process	Once a feasible process has been identified, this step expands each macro activity into its detailed elements and connections such that an MDPM model can be constructed. Similar to CLIOS step: Develop the CLIOS Diagram: Nesting, Layering, and Expanding
7.) MDPM creation from detailed process map	The MDPM model domains can then be populated from the detailed process map and the enumeration of elements corresponding to the rows and columns of the MDPM. Links between elements are also populated at this step as entries into the MDPM matrix.
8.) Analysis of MDPM model	The MDPM model can then be analyzed using the metrics sets developed in chapter 4 and graph based tools.

Table 11 - Description of the MDPM Methodology Steps

One of the primary goals of the case study demonstration was to build upon existing infrastructure such that the surveillance process could be readily integrated into routine clinical microbiology practice and could also be up and running in a much shorter time than a "from scratch" process. However, this required the identification of the existing infrastructure as well as the design alternatives (or design space) where no current infrastructure existed. This modular incremental approach has numerous advantages including the potential for greater ease of integration with existing processes, a greater focus on necessary innovation (rather than re-creating existing elements), potentially lower development costs (as only "new" elements are targeted), and potentially shorter cycle times. However, the reuse of existing elements also limits opportunities for global optimization, since the majority of elements in a process will be largely unchanged and can therefore lead to lower performance than a fully optimized system. In addition, reuse of existing elements does not necessarily mean that they will all function adequately when used in a new process and they must all be reviewed for suitability.

The system designer must therefore first identify the available process elements, and then review them for feasibility within an initial version of the process. Subsequent optimization is of course possible, but the primary goal of this design effort was to provide an initial working prototype of a MRSA surveillance system, which could begin to provide useful information and also serve as a starting point for additional optimization. A secondary but very significant benefit of this methodology is that it can filter significant complexity out of the initial process prototype. As will be seen later in the MDPM model development processes can have enormous inherent complexity without also introducing the additional complexity of process design alternatives. Thus, the methodology allows for the selection of a best feasible process according to a set of stakeholder preferences. The most feasible process is demonstrated and modeled with an MDPM, but the design space is much larger as there are multiple competing technologies for many of the intermediate MRSA surveillance process steps, each representing a highly complex process.

Selecting the "best feasible" processes from the available process elements greatly reduces the complexity that needs to be considered in the modeling of a process, but it is important to note that this complexity still exists in reality, as improvements in the alternatives may change the selection of which is "best" and designers must be aware of information or connected to individuals that can appraise these changes. Thus, in the prototype MDPM model of the surveillance system, the design alternatives discarded are still explicitly modeled, not as process steps, but as information elements that process owners must be aware of. The full complexity of design alternatives could be modeled in subsequent MDPM models if so desired, but in practice it is likely that system designers will prefer to first narrow the design space to a few (or a single) design alternative and then use the MDPM model to manage the complexity associated with even a single design alternative.

However, in reducing the design space this methodology also captures the stakeholder preferences such that they can be used in later evaluations for changes or new technologies.

The design methodology is summarized in Table 12 with individual steps divided into two major categories (feasibility + enumeration and network analysis). The "Process Feasibility and Element Enumeration" category contains all the methodology steps needed to identify the process goals, stakeholders, design space, to select a feasible process from the design space, and to enumerate all of the elements in the selected feasible process. The "Network Analysis" category contains the methodology steps needed to create the MDPM model from the process elements and analyze it.

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

Table 12 - MDPM Methodology Categories

Each of these steps is further described for the MRSA surveillance process. It is important to note that this process design methodology necessarily requires (and implies) significant stakeholder input in the form of requirements, description of goals, and understanding of the many interactions between layers. Throughout the following section, the table describing the steps in the methodology will be shown with the particular step highlighted.

5.1.1 Description of process, stakeholders, and goals

Category	Methodology Step
Process Feasibility and Element Enumeration	**1.) Description of process, stakeholders, and goals** 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

The fundamental goal of the MRSA surveillance process is to provide a cost feasible method to read the complete genome of MRSA microbes associated with hospital infections such that the genotypic characteristics can be tracked, new strains can be identified, and the continued adaptation of Staphylococcus aureus to antibiotics can be monitored. Implicit in this goal is that the process must be scalable to high throughput in order to match the large number of cases that may exist in a large hospital. The overall cycle time the process must also be relatively short (within weeks) in order to provide timely information to epidemiology efforts. The process must also provide highly reliable tracking and prevention of misidentification to ensure clinical histories (phenotypes) can be associated with the resulting whole genome genotypes. The resulting data must be consistent with existing MRSA diagnostic processes (cultures), be easily interpreted by medical staff (physicians), and have a high sensitivity and specificity. Where possible, the surveillance process should integrate with existing clinical microbiology workflows such that the process can be more easily adopted in multiple locations.

The MRSA surveillance process stakeholders were categorized into "internal" ones directly associated with the process and inside the MDPM model boundaries and "external" who can still impact the process but who are either indirectly associated with the process or outside the model boundaries. The internal vs. external stakeholder classification follows that of (Bartolomei 2007). The full list and description of external stakeholders is given in Table 14 and internal stakeholders in Table 14 along with a description of their role in the MRSA process. Note that the stakeholders are listed as organizations or groups in most cases rather than individuals as at this initial step of the methodology the goal is to identify the "macro" stakeholders with a detailed enumeration to the individual level once a feasible process has been identified. The surveillance process was developed at the Brigham and Women's Hospital, the Broad Institute, MIT, and with the support of a variety of technology vendors and research groups. The internal stakeholders described in Table 14 are represented by these organizations in this thesis, and the specific process preferences, constraints, and goals described in the remainder of the methodology are based on this specific subset of all the stakeholders. A national surveillance system would require a thorough analysis of the goals, constraints, and specifications of the entire set of stakeholders including the external ones, but given the prototype nature of the surveillance process described in this thesis, the specific stakeholders are taken as representative of the larger set of internal stakeholders.

Examples of internal stakeholders in the MRSA surveillance process are patients, hospital epidemiologists, physicians, clinical microbiologists, DNA sequencing staff, and key hospital administration such that data from the surveillance can be used to control potential outbreaks. Examples of external stakeholders are insurers (who would potentially benefit from this process through lower infection claim rates), drug manufacturers (who could use the data to design improved antibiotics), diagnostics manufacturers (who could design better diagnostics based on new strains), and government agencies such as CDC.

External Stakeholder	Description
Centers for Disease Control (CDC)	CDC has responsibility for national infectious disease surveillance systems including potential MRSA ones

External Stakeholder	Description
National Institutes of Health (NIH)	NIH oversees research efforts of various institutes investigating MRSA
American Medical Association (AMA)	Sets procedure codes for physicians and general guidelines including infectious disease
Infectious Diseases Society of America (IDSA)	Represents physicians, scientists, and other healthcare professionals specializing in infectious diseases. Sets general guidelines
American Society for Microbiology (ASM)	Provides training, guidelines, and information for microbiology labs including MRSA diagnostic techniques
Health and Human Services (HHS)	Oversees key government agencies such as FDA, CDC, NIH, and CMS (medicare) impacted by MRSA related
State Health Agencies	Responsibility for state infectious disease surveillance systems and control
Genomic Data Analysis Vendors	Provide software tools to annotate MRSA genomes following sequencing
DNA Extraction Vendors	Provide reagents and equipment to extract DNA from MRSA cells
LIMS Vendors	Provide software to control sample tracking throughout lab processes
Data Storage Vendors	Provide storage to handle the data produced by sequencing processes
Liquid Handling Automation Vendors	Provide robots to automate molecular biology and other process steps
Bacterial Diagnostic Vendors	Provide MRSA diagnostics based on PCR and other technologies
Clinical Microbiology Vendors	Provide reagents and equipment for standard clinical microbiology processes

External Stakeholder	Description
Insurers	Reimburse approved procedures including MRSA treatment but only select MRSA diagnostics
Hospital Administration	Set policies for hospital processes including MRSA treatment and surveillance but also responsible for overall budgeting
Hospital Pharmacy Staff	Responsible for ensuring appropriate antibiotic use and control
Hospital Services	Responsible for hospital infrastructure including sample transport, sterilization, and computer support
Antibiotic Researchers	Discover new antibiotics and virulence inhibitors to counteract MRSA
Antibiotic Vendors	Responsible for developing and commercializing new antibiotics to counteract MRSA

Table 13 - MRSA Surveillance Process External Stakeholders

Internal Stakeholder	Description
Patients	Ultimate beneficiaries of improved knowledge gained from MRSA surveillance
Primary Care Physicians	First to encounter many MRSA infections and prescribe antibiotics or order diagnostics
Nursing Staff	Primary patient care responders including obtaining samples for MRSA diagnostics
Specialty Physicians	Also encounter MRSA infections as part of other specialty care and prescribe antibiotics or order diagnostics

Internal Stakeholder	Description
Specialty Nursing Staff	Primary patient care responders also for specialty care including obtaining samples for MRSA diagnostics
Hospital Support Staff	Provide essential services such as transport of samples, replenishment of supplies, disinfecting of hospital equipment, etc.
Clinical Microbiology Lab Staff	Provide clinical microbiology sample processing for available MRSA diagnostics
Clinical Microbiology Researchers	Develop new methods to provide MRSA diagnostics and surveillance
DNA Sequencing Lab Staff	Provide DNA sequencing sample processing using available technology
DNA Sequencing Researchers	Develop new DNA sequencing processes and technology
Bio-Informatics Analysis Staff	Provide analysis of DNA sequence obtained for various MRSA samples using various technologies
Bio-Informatics Researchers	Develop new analysis methods for DNA sequence obtained using various technologies
DNA Sequencing Technology Sequencing Technology Vendors	Provide processes to sequence DNA including MRSA genomes
Alignment Algorithm Research Groups	Develop algorithms to align MRSA genomes to references in order to find mutations
Assembly Algorithm Research Groups	Develop algorithms to assemble MRSA genomes de novo without a reference
Genomic Data Analysis Research Groups	Develop algorithms to annotate MRSA genomes following sequencing
Bacterial Genomics Researchers	Discover mechanisms of resistance, virulence, and evolution in MRSA genomes

Internal Stakeholder	Description
Infectious Disease Researchers and Epidemiologists	Discover mechanisms of infection for MRSA microbes
Process Sponsors (Owners)	Manage overall process development and implementation for DNA sequence based MRSA surveillance

Table 14 - MRSA Surveillance Process Internal Stakeholders

5.1.2 Identify macro process activities and define boundaries

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals **2.) Identify macro process activities and define boundaries** 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

From the process description, stakeholders, and goals a macro level design space can be built identifying the "macro" activity categories the process must go through in order to deliver the genotype information from patient samples. The potential alternative processes can be mapped onto the general activity categories. The full enumeration of these design alternatives produces the "design space". This conceptual process map also serves a very important function as it defines the process boundaries that will be modeled to match external and internal stakeholder boundaries identified in the previous step. There are a number of activities in the MRSA surveillance process demonstrated in this thesis that are not explicitly modeled such as how the infections occurred, how the patients were admitted to the hospital, what their particular insurance and payment situation may be, and the particular procedures the patients might be undergoing. These activities also correspond to external stakeholders not considered in the MDPM model. The MDPM model could be used to model all of these other processes and the interactions between them as part of a larger "enterprise" model, but the goal of this thesis is to demonstrate the MDPM analysis on a single process and this initial boundary setting and selection is essential in reducing the modeling complexity. For the MRSA surveillance process the general activities (without macro design alternatives) are as shown in Figure 65, along with representative activities not modeled in the MDPM (and corresponding to external stakeholders).

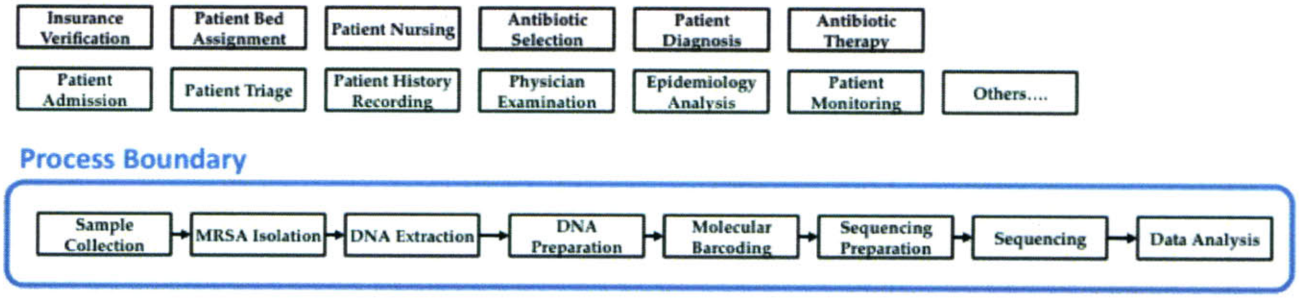


Figure 65 - Surveillance Process Boundary

The macro activities in the process are described in Table 15. With the process boundary defined and the general macro process steps identified, it is possible to construct the macro design space, which is the identification of design alternatives for each of the activities inside the process boundary identified in Figure 65.

Macro Process Step	Description
Sample Collection	Collect sample from patient, label, and transport to lab
MRSA Isolation	Process sample to obtain single MRSA isolate and grow culture
DNA Extraction	Lyse culture cells, extract DNA, and purify DNA
DNA Preparation	Prepare extracted DNA for sequencing according to technology

Macro Process Step	Description
Molecular Barcoding	Add molecular barcode to each DNA sample
Sequencing Preparation	Process barcoded sample into sequencing ready DNA
Sequencing	Read DNA sequence according to technology
Data Analysis	Analyze DNA sequence

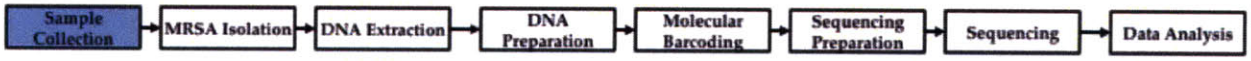
Table 15 - Steps in Macro Surveillance Process

5.1.3 Identify potential process alternatives and design space

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries **3.) Identify potential process alternatives and design space** 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

As described in the methodology description, the objective of this step is to identify all potential process design alternatives that could satisfy the goals but at a "macro" level without enumerating all of the detailed elements in each. This step requires searching through available processes and technologies to identify existing processes that might be adapted, creating new ones from existing elements, or if none are available, putting "placeholders" for missing elements corresponding to activities in step 2. The macro process steps and design alternatives for each are detailed in the following sections with a process diagram indicating which step is being described.

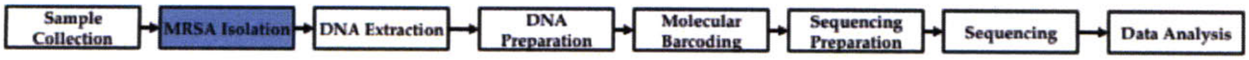
Sample collection



The first activity in the process to be modeled is sample collection from a patient infection. Again, the particular infection process or treatment process the patient might be undergoing is not modeled and the process starts with a physician request for surveillance. The surveillance process would likely happen in parallel to standard diagnostic requests, but physicians might request surveillance in especially difficult cases, as part of standard hospital procedures, or in response to outbreaks. From this request a nurse (or physician) would then collect a sample and send it to the clinical microbiology laboratory. Many of these procedures and processes are already in place for culture based diagnosis and likely require little change as they can already produce MRSA isolates in routine culture based diagnosis. However, a large variety of them exists corresponding to each potential infection location.

Available Processes: Current practice
Design Alternatives: None required

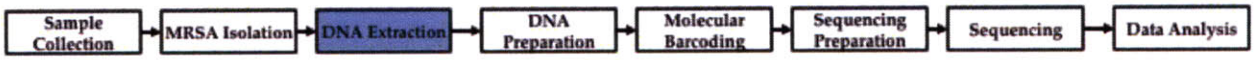
MRSA Isolation



From the sample the clinical microbiology laboratory would then prepare the sample, process it, and isolate a MRSA microbe. As with the clinical sample collection, the majority of clinical microbiology processes used for culture based diagnosis sample isolation remain nearly identical and current practice may be used. Some change may be required to satisfy the other "ilities" such as adaptability when sequencing other organisms, but this is not part of the "baseline" process. And in the case of needing to sequence other similar organisms, clinical microbiology procedures exist for their isolation, which can be used "off the shelf". Nevertheless, the variety of sample preparation procedures matches the variety in the upstream sample collection locations (infection sites).

Available Processes: Current practice
Design Alternatives: None required

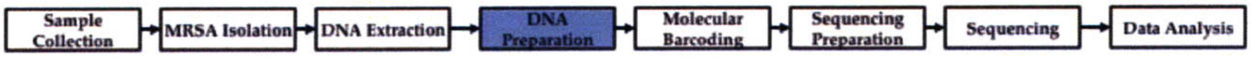
DNA extraction



The process begins to deviate from standard procedures at this point since for the majority of clinical microbiology procedures large amounts of DNA are not required. Processes exist, such as PCR diagnostics, that do require DNA extraction but this is usually in much smaller amounts and of lower quality. There exist a number of commercial kits for the extraction of DNA from tissues but these are not specifically designed for large quantities of high molecular weight DNA from MRSA microbes, which are notoriously tough with thick cell walls making them difficult to lyse. Development will likely be needed for this activity as there are some process elements to draw from but nothing that is directly "off the shelf" as in previous activities. In particular, methods to properly lyse the cells without damaging the DNA and then subsequently removing the DNA from the cellular debris will likely be required for a range of MRSA strains. Note that this activity is the first "placeholder" activity in that it is not part of an existing "off the shelf" process, but the "desired" activity is put into the process design space to determine if it is needed or the "standard" process will suffice.

Available Processes: Existing heat, enzymatic, and alkaline lysis methods
Design Alternatives: Optimization of current methods for MRSA application (**Placeholder**)

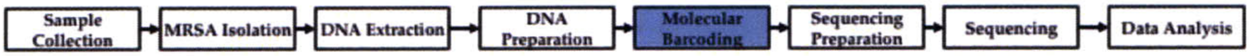
DNA preparation



This process activity category is a significant branch point in the process as the growing number of different sequencing technologies branch off from here, each requiring a specific DNA preparation process for use in their technology. The only "off the shelf" technology available is ABI Sanger sequencing originally developed for use in the human genome project, but likely prohibitively expensive for a whole genome surveillance application. Available emerging technologies include Roche 454, ABI SOLiD, Illumina, Helicos, and Polonator. A number of others are on the horizon further complicating the design challenge as these will each bring particular advantages, but only commercially available technologies are considered here. Each one of these requires a different DNA preparation process and the downstream process steps are thus "locked in" to that particular technology depending on the particular choice at this step. Selection of a single technology will be necessary for the initial prototype demonstration as the additional complexity introduced by running multiples of these highly complex sequencing processes is very significant and the relative performance of each is continuously changing. Starting with a single technology choice allows for optimization from that point rather than continuous change without a baseline.

Available Processes: ABI Sanger
Design Alternatives: Roche 454, Illumina, ABI SOLiD, Helicos, Polonator

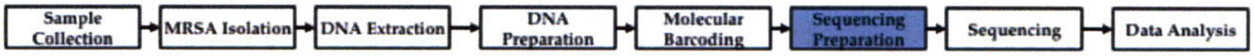
Modular Barcoding



One of the advantages of next-generation sequencing technologies is their enormous throughput, measured in the giga bases per run scale. These large batch sizes are also critical for the greater efficiency of these new sequencing technologies. Yet, MRSA microbial genomes are only 2.8 mega bases in size and therefore require some mechanism to multiplex many microbial genomes into a single sequencing run. This mismatch in batch size can be addressed by "barcoding" the microbial genomes at the molecular level such that they can be pooled together into a large-scale run but later the data can be disambiguated based on the barcodes. This is brand-new technology that must be developed and likely significant development will be needed. Note also that due to its lower throughput, no molecular barcoding method is required for the ABI Sanger process, but a method specific to each one is required for the other next generation processes. This is another example of a "placeholder" that must be placed in the design space to represent a not yet invented, available, or tested activity that is nevertheless necessary to build a process. The "target" specifications can then be put in place to determine if the overall process is feasible with this placeholder, and if it is, the feasibility of the individual placeholder can then be assessed.

Available Processes: ABI Sanger
Design Alternatives: Roche 454, Illumina, ABI SOLiD, Helicos, Polonator (**Placeholder**)

Sequencing preparation

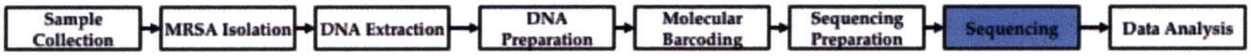


Once the barcodes have been added to each sample and these have been pooled, these "libraries" of MRSA microbial DNA can be put through the necessary molecular preparation such that they can be read on detection instruments. This is again specific to each technology and follows depending on the branch path chosen at the DNA preparation step. These processes are relatively straightforward in terms of their workflow structure but are continuously being optimized and improved as the competition between sequencing technologies intensifies. However, the current version of these processes is used for the process mapping exercise, and potential opportunities for complexity reduction are explored later in this chapter and in chapter 7.

Available Processes: ABI Sanger

Design Alternatives: Roche 454, Illumina, ABI SOLiD, Helicos, Polonator

Sequencing

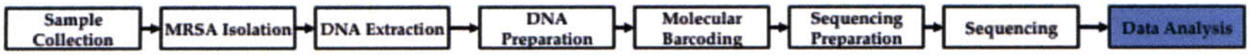


Once the DNA has been prepared for the particular sequencing technology, it must be "read" on the particular technology's detection instrument. While this is again relatively straightforward from a process mapping point of view, each of the steps is continuously being improved and optimized requiring significant "scaffolding" on the part of the organization. This step is also where the boundary from DNA molecules to biological data is crossed, as a number of algorithms process the raw signals coming from the detection instruments and process these into DNA sequences stored in computer files. And as before, the particular choice made at the DNA preparation step commits the process a particular branch up to and including the data type produced.

Available Processes: ABI Sanger

Design Alternatives: Roche 454, Illumina, ABI SOLiD, Helicos, Polonator

Data analysis



In this final step, the differing data types produced by each technology can be merged as they should all represent the fundamental biological information. However, complete equivalence is not practically feasible and each of the data types has different properties including error modes and quality. Algorithms must be designed to interpret these, identify false positives and false negatives, and properly report potentially important biological changes in the MRSA genomes. The data must of course also be associated with the original surveillance request such that the clinical observations can be correlated to the genotypic data. This is also an area where the state-of-the-art is rapidly changing and brand-new technology will be required, and the activity is represented as a "placeholder".

Available Processes: None
Design Alternatives: New Method Development (**Placeholder**)

Design Space

The macro level design space can be constructed from these elements showing the number of design alternatives. While this full design space (i.e. every detailed activity associated with the macro activities described) could be mapped in an MDPM model skipping this step, the goal of first producing a macro design space is to list potential design alternatives without exhaustively enumerating every activity. This is especially important due to the complexity of the processes and that some of the process alternatives may not be feasible when subject to stakeholder constraints. Thus, an initial "filtering" in the design process such that only the process most suited to the particular application is fully developed in the MDPM model can greatly reduce this design complexity and ensure resources are focused on the necessary optimizations and connecting technologies to make it work rather than working on multiple designs at once. As will be seen later even a single configuration (choice of process path) through the design space can represent enormous complexity. The synthesis view of the macro process design space is shown in Figure 66 in comparison the macro process, which can be thought of as the "generic" process one is seeking design alternatives for. The corresponding activities for each design alternative are also shown in Figure 66. Note also that the "placeholder" activities are marked in yellow showing activities that will have to be developed to "connect" the existing activities of the process. These activities will require specific attention in the initial phases of the process development to assess their technical feasibility. And this methodology allows for their early identification and verification that all other existing elements in the path are feasible.

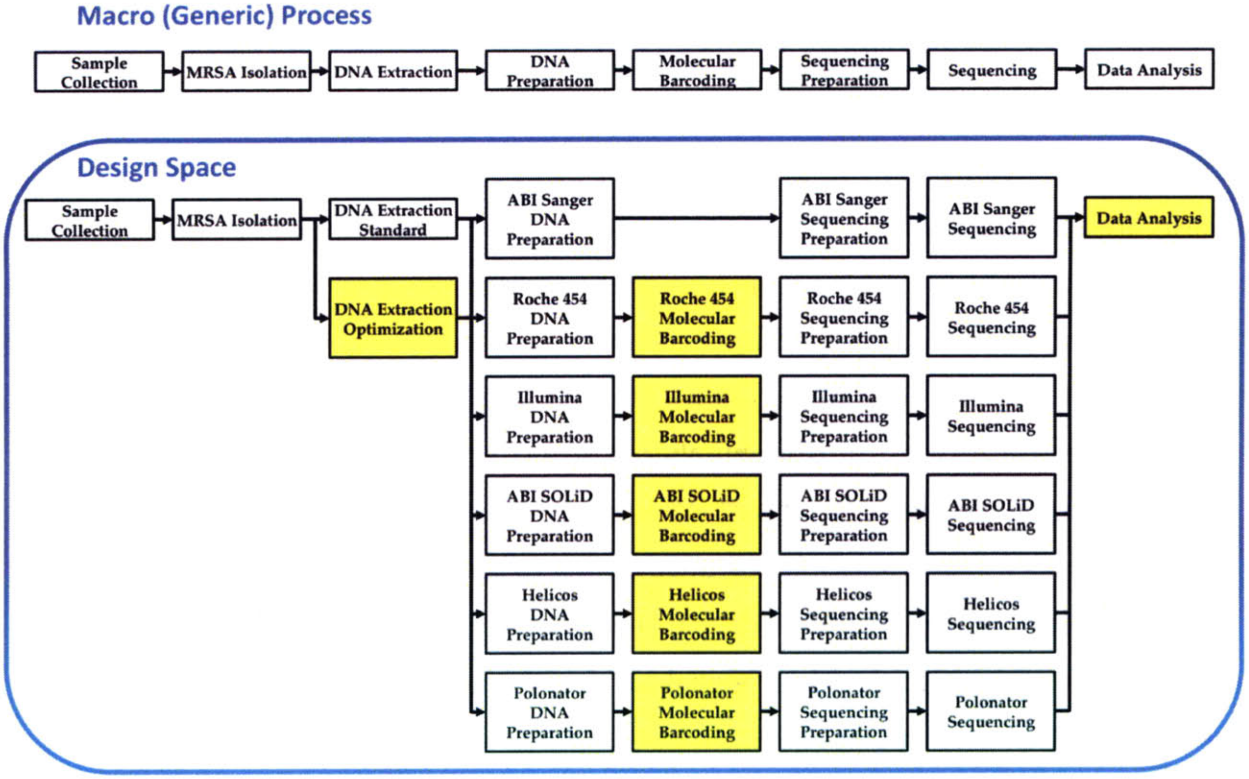

Macro (Generic) Process
Sample Collection
MRSA Isolation
DNA Extraction
DNA Preparation
Molecular Barcoding
Sequencing Preparation
Sequencing
Data Analysis
Design Space
Sample Collection
MRSA Isolation
DNA Extraction Standard
DNA Extraction Optimization
ABI Sanger DNA Preparation
ABI Sanger Sequencing Preparation
ABI Sanger Sequencing
Data Analysis
Roche 454 DNA Preparation
Roche 454 Molecular Barcoding
Roche 454 Sequencing Preparation
Roche 454 Sequencing
Illumina DNA Preparation
Illumina Molecular Barcoding
Illumina Sequencing Preparation
Illumina Sequencing
ABI SOLiD DNA Preparation
ABI SOLiD Molecular Barcoding
ABI SOLiD Sequencing Preparation
ABI SOLiD Sequencing
Helicos DNA Preparation
Helicos Molecular Barcoding
Helicos Sequencing Preparation
Helicos Sequencing
Polonator DNA Preparation
Polonator Molecular Barcoding
Polonator Sequencing Preparation
Polonator Sequencing

Figure 66 - Design Space of Surveillance Process Alternatives

5.1.4 Identification of design specifications and constraints

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space **4.) Identification of design specifications and constraints** 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

The MRSA surveillance process has a set of specifications and constraints that must be satisfied in order for it to be feasible. These are based on internal stakeholder interviews, initial design considerations, and comparison to existing surveillance systems. As described in section 5.1.1, the stakeholders for the prototype process developed in this thesis are a subset of all possible internal stakeholders but they are likely representative and the specification variables come from these. Among the specifications are throughput, genotypic information coverage, infrastructure requirements, cycle time, data quality, and most importantly cost per sample. Many of the process "ilities" are described here as the surveillance process must satisfy a number of factors such as scalability to higher throughputs, transferability to other locations (i.e. a San Francisco hospital), flexibility to accept samples from various sources, adaptability to produce genotype information for other microbes, and modularity such that change in one area does not require a global re-optimization of the process. These global specifications can be applied to each of the process categories described in the conceptual process map in order to filter (select) suitable existing processes or identify the need for new ones to satisfy the specifications and constraints. The set of specifications and constraints is described in the following section. The specifications and constraints are developed for a feasibility selection model which identifies the "cost" of each path in the design space network developed above and consist of a properties vector that calculates the sum of the product of the individual specifications (for that activity) times a weighting of the relative value of that activity. Thus in the feasibility selection model, the throughput constraints identify the maximum feasible throughput through each edge connecting two activities and the properties vector calculates the "cost" of traversing that edge according to the properties vector. Thus, the feasibility determination and selection of the path can be established as a minimum cost network flow model as will be described in the next section. Each of the specification variables and constraints on these variables is briefly described in Table 16 and fully detailed in the next sections. Note that these specifications apply to each *activity* in the macro process.

Specification Variable	Description
Throughput	Number of MRSA samples that can be processed per week
Cost	Cost per sample processed
Cycle Time	Time for activity to complete
Infrastructure	Infrastructure requirements for activity
Data Quality	Error rate contribution from activity
Data Completeness	Fraction of total genome that is covered by the activity
Scalability	Ability to rapidly increase throughput if needed
Transferability	Ease with which activity can be transferred
Flexibility	Ability of activity to respond to varying demand
Adaptability	Ability of activity to work with other microbes
Modularity	Degree to which activity is independent of other activities

Table 16 - Specification Variables for MDPM Model of MRSA Surveillance

Throughput

Constraint: >200 samples per week

Range: 0 to Hospital Patients per week

A fundamental requirement for the MRSA surveillance process is that large numbers of microbes must be sampled. Typically, a large academic hospital will have between 500-1,000 beds (source: AHA) but will perform several times that number in outpatient procedures (>4,000) and several times that number again in patient consultations every week (>12,000). Assuming a surveillance rate of 10% of all cases and on the order of 20,000 potential surveillance cases yields greater than 2,000 potential samples processed through the surveillance system every week. Individual hospitals may choose to lower the surveillance rate or may simply be smaller than a large academic hospital, but even so the minimum weekly throughput for a hospital MRSA surveillance system is on the order of 200 samples per week corresponding to a community hospital with ~200 beds and ~2,000 potential surveillance events.

Cost

Constraint: <$100 per sample

Range: 0 to 10 (normalized to highest cost activity at 10)

A key performance specification for the surveillance system is the cost per sample, which must compete with the cost of culture-based diagnostics that are on the order of $10-20 per sample. While DNA sequencing costs have decreased dramatically, these costs require relatively large "batch" sizes 2 to 3 orders of magnitude larger than the size of an individual microbial genome. This performance specification is especially important to the wider adoption of surveillance since even at a relatively modest throughput of 200 samples per week still represents a cost of over $1 million per year at $100 per sample.

Cycle Time

Constraint: <2 weeks

Range: 0 to 10 (normalized to highest cycle time activity at 10)

Although the focus of the surveillance process is not to replace more immediate diagnostic tests, the cycle time must still be relatively short for any epidemiological efforts to be timely. A total cycle time of less than two weeks from sample collection to analysis results is set as the minimum specification. Shorter cycle times are better and that the limit very short cycle times (within a day) could potentially replace current culture based and even PCR-based diagnostics.

Infrastructure

Constraint: None, specified in properties vector

Range: 0 to 10 (0=no infrastructure needed, 10=maximum)

The ease with which the surveillance process can be integrated and adopted into existing hospital infrastructures depends to a large extent on its requirements. Based on the initial process map, the fewer changes upstream in the "high transaction" environment of patient interaction the less infrastructure requirements. If physicians and nurses do not have to significantly change their workflows in order to provide surveillance samples, the easier it will be to integrate the process. Similarly if the majority of the existing clinical microbiology workflows can be "reused" the easier it will be to integrate the process. There is unavoidable additional infrastructure in the new technologies of DNA sequencing and analysis but if this can be contained to just those categories in the conceptual process map such that fewer links need to be made between activities in the process the easier it will be to adopt the process and preserve significant modularity and flexibility.

Data Quality

Constraint: None, specified in properties vector

Range: 0 to 10 (0=highest data quality, 10=minimum data quality)

Ensuring there are a minimum of false positives and false negatives in the genotype data is critical to establishing a robust and reliable surveillance process. In a 2.8Mb genome even a 1% error rate would lead to an unacceptably large 28,000 errors in the sequence data. More importantly, the error rate of the process must be significantly less than the mutation rate of the microbes such that true mutations can be identified which for E. Coli has been estimated at between 2 X 10^{-4} and 2 X 10^{-9} per replication per cell per position in the genome (Denamur and Matic 2006). In a single bacterial infection there may be millions of microbial cells at a 30min division rate in less than 12 hours, and for a genome size of 2.8Mb that means that every base in the genome of the infection colony will have had a "chance" for mutation in a 30 hour period which means that many of the bases investigated could be real mutations and the system error rate must be low enough to detect them. However, it appears that not all parts of the genome are equally susceptible to mutation, with some parts highly conserved and others subject to "hyper mutation" as described in (Blazquez and Eliopoulos 2003). The surveillance process is set initially to a specification of total "system" error rate less than .01% which is less than the estimated upper bound on mutation rate as estimated by (Denamur and Matic 2006). Note that this does not necessarily mean that each piece of data produced has that error rate, but that once the data has been filtered and processed the net data produced has that level of confidence that a particular base is correct. One way to do this is to "over cover" bases with relatively lower quality (i.e. ~1% error rate) bases but which have randomly distributed errors such that the probability of the same mistake being made multiple times is very low and the "true" signal from the mutation can be properly interpreted. This can of course affect the cost since more data might be needed if the sequencing technology has a higher intrinsic error rate. This "consensus" alignment is the basis for much of the analysis performed in chapter 6.

Data Completeness

Constraint: None, specified in properties vector

Range: 0 to 10 (0=complete data, 10=no data)

Unlike other diagnostic and surveillance processes to date the goal of this process is to provide data for the entire genome of a microbe rather than a small portion as in PCR, MLST, or restriction mapping. Visibility into the entire genome is important to identify novel mutations or gene transfer from other organisms, which methods that target known genes cannot find (such as PCR). The surveillance method should thus cover 100% of the genome and any plasmids contained within the microbe.

Scalability

Constraint: None, specified in properties vector

Range: 0 to 10 (0=completely scalable, 10=not scalable)

The process must be scalable such that if a hospital might need to increase its surveillance rate beyond 10% or need to provide services to associated facilities or other hospitals this can be readily accomplished. Ideally economies of scale might be achieved such that increasing scale actually lead to decreased costs, but most importantly the process could be scaled to sample 100% of the potential cases should the need arise. This might happen for example if a particularly dangerous new strain of MRSA emerged and the hospital might need to rapidly implement quarantine procedures or identify the source of the new strain.

Transferability

Constraint: None, specified in properties vector

Range: 0 to 10 (0=completely transferable, 10=not transferable)

The surveillance process must eventually be transferred to other locations and this will require that the majority of work instructions be in "explicit form" since tacit information transfer (i.e. mentoring) cannot be easily performed to more than one location. This also requires that the majority of process activities be relatively stable such that users without significant "scaffolding" infrastructure can operate the process without continuous changes. In order to provide a quantitative measure the process must have all activities documented (protocols exist) and must have operated in stable configuration for more than six months.

Flexibility

Constraint: None, specified in properties vector

Range: 0 to 10 (0=completely flexible, 10=not flexible)

The process must also have significant flexibility in order to accept samples from multiple sources (different departments of the hospital or different infection surveillance requests). In addition process must be flexible enough to operate under varying demand from each of the surveillance requests without losing efficiency. This can be quantified by saying that the process can accept greater than 50% of samples from any single source in a given week (i.e. 50% blood samples). Conversely the process should be able to accept less than 5% of its samples from a given source.

Adaptability

Constraint: None, specified in properties vector

Range: 0 to 10 (0=completely adaptable, 10=not adaptable)

The process should also be adaptable to produce information from other microbes as a common mechanism for new strains is horizontal gene transfer from another microbe. It is therefore important to have the ability to sequence other organisms that could potentially contribute resistance genes such as enterococci. The measure for this is the percent of the process that would need to be changed to accommodate other organisms with a target of less than 10% (as measured by the number of activities).

Modularity

Constraint: None, specified in properties vector

Range: 0 to 10 (0=completely modular, 10=not modular)

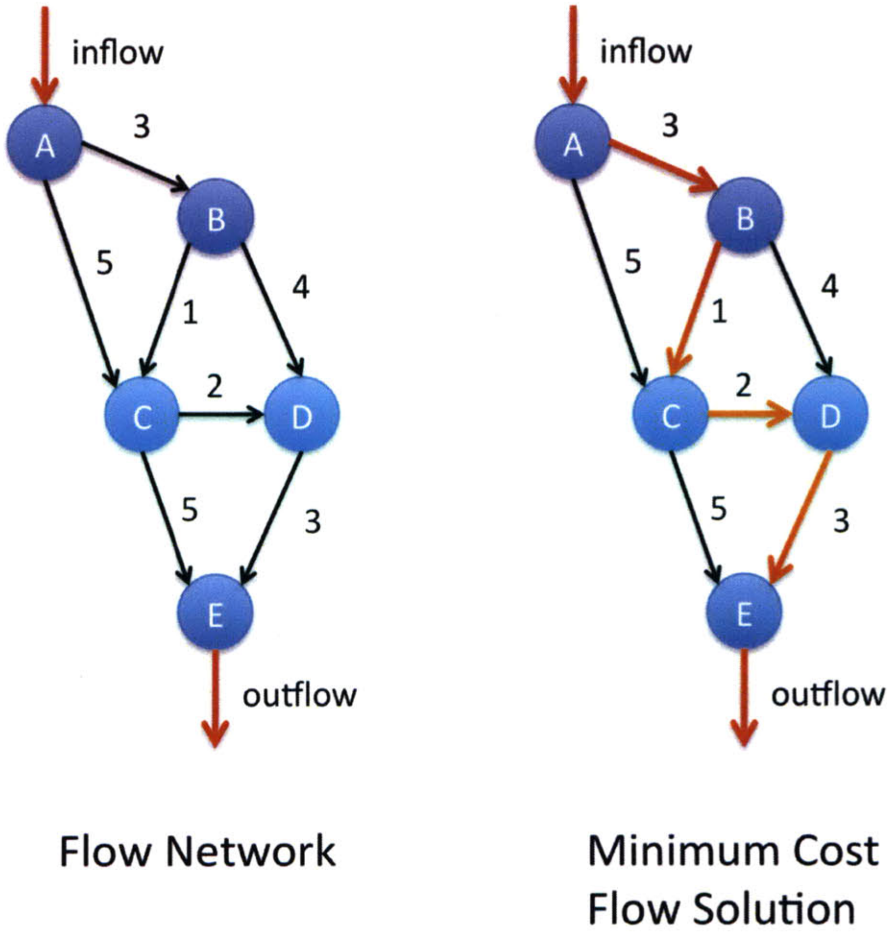
The degree to which changes in one section of the process impact another is a measure of the modularity for each section and overall the process. The measure used for the surveillance process is that less than 10% of the activities downstream of a module should need to change if there is sufficient modularity.

5.1.5 Selection of macro level feasible processes against design specification requirements

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints **5.) Selection of macro level feasible processes against design specification requirements** 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

With this defined set of specifications and identified constraints it is possible to construct a linear programming model to identify a feasible preferred process path based on "macro" information. The particular model to be used is commonly known as the minimum cost network flow (MCNF) model normally used to find optimal minimum cost routes through transportation networks as described in (Jensen and Bard 2003; Winston and Albright 2008). A generic MCNF model is conceptually illustrated in Figure 67 where flow enters node A, and has a cost associated with traversing each edge, before finally leaving at node E. The minimum cost flow is the path through the network whose sum of "costs" is lowest as shown in red in Figure 67. The MCNF algorithm is adapted here to identify the "preferred path" through the design space network developed in step 3 and shown in

Figure 66, where the "costs" for traversing each edge in the graph correspond to the particular set of specification variables associated with that activity. The model is not intended to provide an optimum path but rather identify which paths are feasible and also satisfy the greatest number of preferences according to the "specifications" vector composed of the individual specifications described in the previous section.



Flow Network

Minimum Cost Flow Solution

Figure 67 - Network Flow and Minimum Cost Solution

The standard formulation of the minimum cost network flow problem consists of a linear program for a directed, connected graph with m nodes and n arcs. At least one inflow node and one outflow node is specified as in Figure 67. Then the flow through a particular edge k from node i to node j is denoted as k(i,j) and thus the particular throughput through edge k is denoted by a decision variable x, as:

x_k = flow (throughput) through edge k(i,j) from node i to node j

The costs associated with flowing, or transporting, through edge k(i,j) is denoted by,

c_k = unit cost of flow through edge k(i,j)

The net flow balance variables are denoted by,

b_i = net supply (edge flow out – edge flow in) at node i
where $b_i>0$ if i is an inflow node, $b_i<0$ if bi is an outflow node, and $b_i=0$ for all other nodes

The maximum capacity of each edge is denoted by,

u_k = capacity of edge k(i,j)

And the set of edges entering node i, or leaving node i, are given by,

KL_i = set of edges leaving node i
KT_i = set of edges terminating at node i

With these definitions, the minimum cost network flow objective function can then be formulated in Equation 23 as the path (or paths) through the network that minimizes the overall cost to flow (or transport) inflows to the outflow nodes. Implicit in the formulation is the existence of an adjacency matrix defining the allowable paths through the network and of the same dimension as the "cost" matrix defining cost values c_k.

$$\min \sum_{k=1}^{n} c_k x_k$$

Equation 23 - Objective Function for Minimum Cost Flow Problem

subject to $\sum_{k \in KL_i} x_k - \sum_{k \in KT_i} x_k = b_i$ for all i = 1… m

$0 \leq x_k \leq u_k$ for all k=1…n

Equation 24 - Conservation of Flow for Minimum Cost Flow Problem

And for this application, it is assumed that all samples enter the process and leave with no accumulation or creation inside the network, thus the sum of all node balances must equal 0, as given in Equation 25.

$$\sum_{i=1}^{m} b_i = 0$$

Equation 25 - Flow Balance for Pure Network Flow Problem

This formulation of the MCNF can be adapted to find feasible processes within the design space graph constructed in step 3 made up of the macro activities, which have edge "costs" based on the properties vector defined by the specification variables identified in step 4 of the methodology and a certain number of samples flowing through them as indicated by the decision variables x, shown in Figure 68. Note that the MCNF formulation corresponds to an "Activity on Edge" AoE network as opposed to the "Activity on Node" AoN network used in the MDPM model in this thesis. Thus each edge contains the "cost" to performing that activity which can be interpreted as a "cost" to flow samples through that activity in the process. The costs are derived from the particular set of values in the specifications vector for that particular activity but adjusted by a weighting to indicate which of the process properties are most important to the stakeholders as shown in Figure 69. Thus rather than let cycle time have the same weight as modularity, the weighting allows for cycle time to have a higher importance according to the stakeholder preferences.

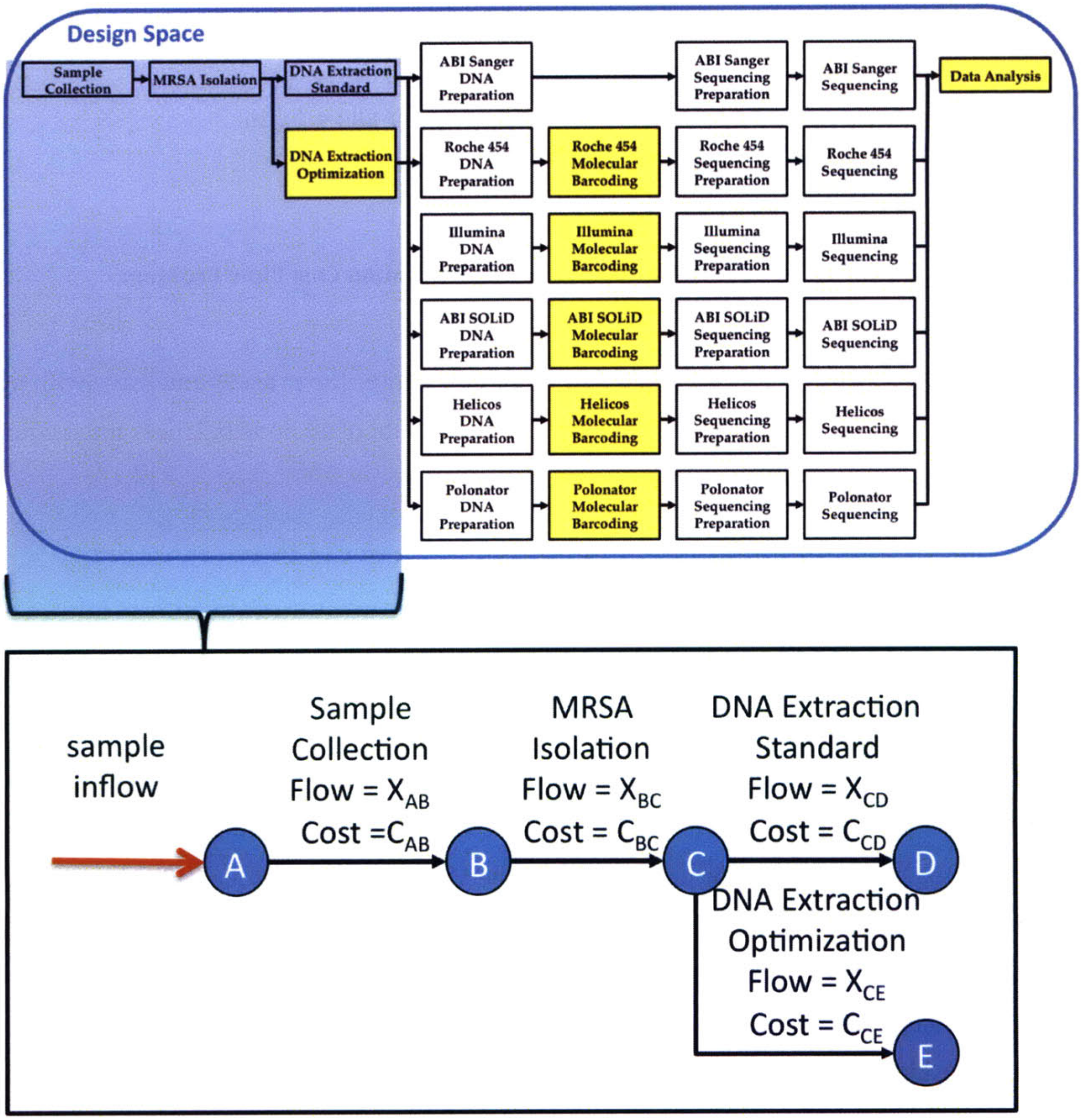

Design Space
Sample Collection
MRSA Isolation
DNA Extraction Standard
DNA Extraction Optimization
ABI Sanger DNA Preparation
ABI Sanger Sequencing Preparation
ABI Sanger Sequencing
Data Analysis
Roche 454 DNA Preparation
Roche 454 Molecular Barcoding
Roche 454 Sequencing Preparation
Roche 454 Sequencing
Illumina DNA Preparation
Illumina Molecular Barcoding
Illumina Sequencing Preparation
Illumina Sequencing
ABI SOLiD DNA Preparation
ABI SOLiD Molecular Barcoding
ABI SOLiD Sequencing Preparation
ABI SOLiD Sequencing
Helicos DNA Preparation
Helicos Molecular Barcoding
Helicos Sequencing Preparation
Helicos Sequencing
Polonator DNA Preparation
Polonator Molecular Barcoding
Polonator Sequencing Preparation
Polonator Sequencing
sample inflow
Sample Collection
Flow = X_{AB}
Cost =C_{AB}
MRSA Isolation
Flow = X_{BC}
Cost = C_{BC}
DNA Extraction Standard
Flow = X_{CD}
Cost = C_{CD}
DNA Extraction Optimization
Flow = X_{CE}
Cost = C_{CE}
A
B
C
D
E

Figure 68 - MDPM Methodology Design Space as a Minimum Cost Network Flow Problem

Only a portion of the flow network graph is shown in Figure 68 to highlight the correspondence between the "design space" process graph created previously and the flow graph to be used in finding feasible processes according to activity parameters and stakeholder preferences and constraints. It is important to emphasize once again that the use of MCNF model to filter out design alternatives is primarily required to reduce the complexity of the full MDPM graph to be developed in the following steps of the methodology. Each of the macro activities in the design space decomposes into a significantly more complex set of processes, individuals, and work instructions making the initial reduction of design alternatives necessary. However, the application of the MCNF design alternative is fully compatible with the MDPM model to be developed further as each of the detailed activities within the macro activities could be analyzed in the same way to perhaps to find a new improved path through a series of design changes to the baseline process. If new design alternatives (i.e. other processes) were to appear following an initial design the MCNF model could be re-applied at the macro level to find a feasible set. Note also that this thesis has emphasized the finding of "feasible" solutions rather than "optimal" ones since the level of detail at the macro level and the nature of new prototype processes means there will be large uncertainties as to the actual performance of these processes. The property values associated with each activity as shown in Figure 69 are estimates and can also have large uncertainties. Ultimately it is up to the process designers and owners to make a selection of which processes to pursue further into full MDPM modeling and real-world implementation, but the methodology developed here can help reduce the set of design options that need to be considered.

Specification Vector for MRSA Isolation Activity

Specification Variable	MRSA Isolation Value
Cost	10
Cycle Time	2
Infrastructure	1
Data Quality	2
Data Completeness	3
Scalability	5
Transferability	5
Flexibility	2
Adaptability	3
Modularity	2

X

Specification Weights Vector for Overall Process (stakeholder preferences)

Stakeholder Specification Weights	Value
Cost	0.4
Cycle Time	0.05
Infrastructure	0.1
Data Quality	0.05
Data Completeness	0.1
Scalability	0.05
Transferability	0.1
Flexibility	0.05
Adaptability	0.05
Modularity	0.05

= 5.7 = "Cost" C_{BC}

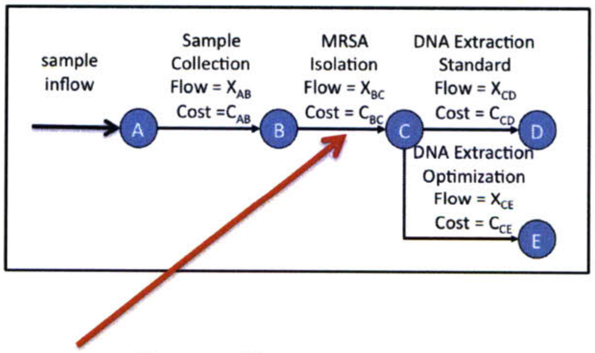


Figure 69 - Example of Edge Cost Calculation for Adapted MCNF Model

Another advantage of developing a graph based feasibility identification tool is that many processes are built from a series of existing "off the shelf" technologies joined together to form a process, but many alternatives may exist for each of the technologies. The MCNF enables the identification of feasible processes from each of these individual "modules" that may not have a priori been considered. The design space presented in Figure 66 contains relatively few modules except for the DNA sequencing technologies but it is possible to imagine that there could be many more alternatives for DNA extraction, data analysis, and perhaps sample collection in the future, making the number of possible permutations very large. The MCNF analysis allows for the "finding" of feasible paths through all of the possible combinations of modules (i.e. technology alternatives for each macro process activity) as illustrated by the conceptual "best path" in Figure 67.

The MCNF problem can then be solved using a linear programming algorithm, such as the one in Solver for Excel (Winston and Albright 2008). Some preparation and formatting is necessary first however, including the construction of vectors and matrices containing the structure of the graph itself as an adjacency matrix, each of the stakeholder weights, the activity costs, the capacity constraints, and the decision variables (i.e. the number of samples to be processed through each activity). The properties vector was purposely designed to be non-dimensional and for scales to correspond to "high" is "worse" such that higher costs can correspond to less desirable paths. Certain values such as actual monetary cost were normalized to the highest cost item such that they can also be onThe the same 0-10 scale. The weighting of each of the activity properties are given as fractions adding up to 1, such that the relative impact of each property can be specified according to the stakeholder preferences (i.e. if cost is more important). An adjacency matrix containing the costs can then be constructed, which when multiplied element by element with the decision variable matrix (flow quantities) and summed yields the total process cost, which is also the objective function as shown schematically in Figure 70 with the excel elements. The capacity constraints in Figure 70 define the maximum number of samples that could be processed through the system. In order to determine feasibility, all process paths are assumed to have the same maximum expected capacity, but the "cost" associated with an infrastructure to support that maximum capacity is reflected in the activity cost vector. This ensures that all processes are compared at the same maximum expected capacity (same scale). However, this could be modified to better reflect variation in capacity between processes.

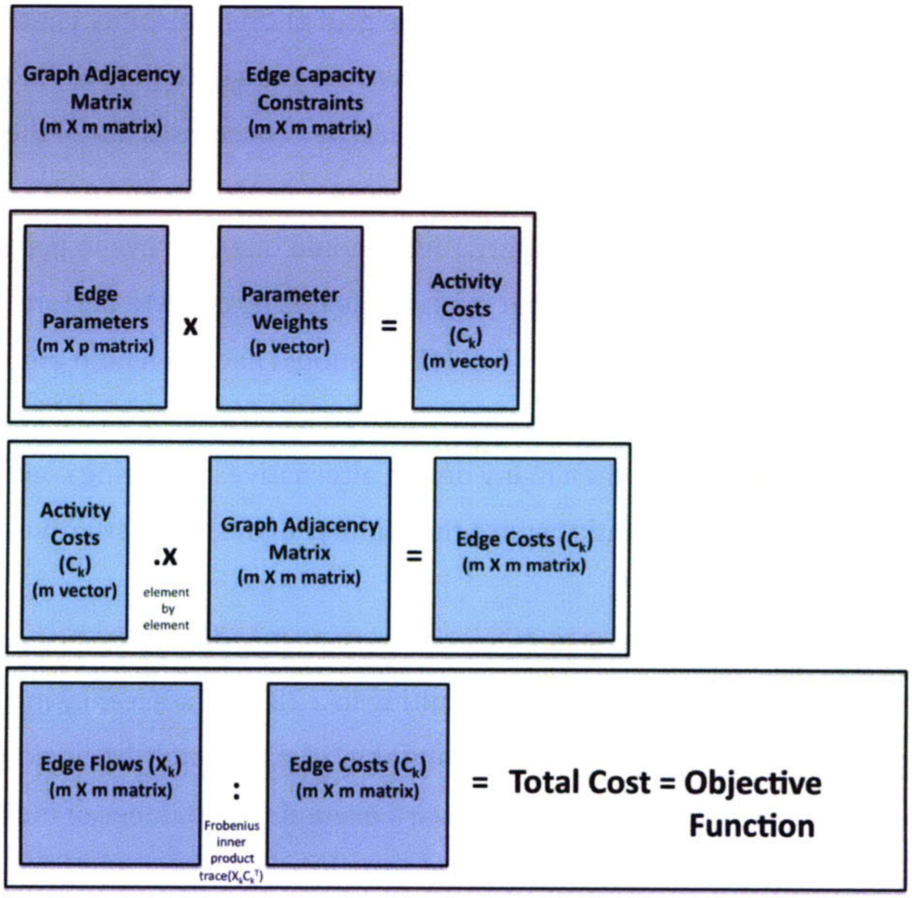

Graph Adjacency Matrix (m X m matrix)
Edge Capacity Constraints (m X m matrix)
Edge Parameters (m X p matrix)
X
Parameter Weights (p vector)
=
Activity Costs (C_k) (m vector)
Activity Costs (C_k) (m vector)
.X
element by element
Graph Adjacency Matrix (m X m matrix)
=
Edge Costs (C_k) (m X m matrix)
Edge Flows (X_k) (m X m matrix)
:
Frobenius inner product trace($X_kC_k^T$)
Edge Costs (C_k) (m X m matrix)
= Total Cost = Objective Function

Figure 70 - Minimum Cost Flow Algorithm Input Matrices and Setup in Excel and Risk Solver

The MCNF problem in Figure 70 can then be solved for a minimum objective function value subject to the capacity, non-negativity, and other constraints. The resulting solution corresponds to the process path in the design space network that best minimizes the set of "process costs" and is feasible. Note that for infeasible values are tested at the creation of the Edge Cost matrix, where if an activity has an infeasible value exceeding some limit in a particular parameter, the particular cost is set to an arbitrary high "infeasible" value exceeding the scale of 0-10 (10,000) such that "flow" could not traverse through that activity.

The resulting solution describes a process that is selected for further modeling in the MDPM methodology. The next step is detailed enumeration, filling in the "details" from the macro process activities. As previously stated, the goal of this feasibility check and initial selection is to reduce the number of design alternatives going into the MDPM model in order to reduce the model size. However, the "discarded" processes out of the selection are retained as elements in the information layer as process designers and participants will need to remain aware of these design alternatives should one improve in performance and become a better solution. Note again that in principle it is possible to perform the same selection process performed here once a full MDPM model had been built enumerating in detail every alternative. This might be worthwhile for "enterprise models" where multiple processes must be managed or when many design alternatives are being considered and an initial "filtering" cannot significantly reduce the number of design alternatives.

For the design space graph detailed in Figure 66 the specification variable values for each of the system activities were then filled in resulting in a 28 X 28 adjacency matrix such that the "m" activity dimension in Figure 70 is 28 and the "p" parameter dimension is 10. This also meant that for this relatively straightforward example the number of individual decision variables was 28 X 28 or 784, which is already a significant number. Simplification of the algorithm such that zero elements in the adjacency matrix are not considered would likely be required for larger models as this number of variables is already beyond the capability of the standard Excel solver and an advanced version[7] was needed.

An important note is that that the activity values and selected processes correspond to the suitability of the particular technology for a MRSA surveillance system only and do not reflect the suitability of a particular process path or technology for other purposes where a

[7] Risk Solver Platform, Frontline Systems

different set of values would need to be input. The weights in Table 17 were used to value the relative contribution of each global process property.

Dimension	Weight	Range	Constraint	Notes
Cost	0.4	0-10 (norm.)	< $100/unit	Most important consideration
Cycle Time	0.05	0-10 (norm.)	< 2 Weeks	Not critical for surveillance
Infrastructure	0.1	0-10		Important to adoption rate
Data Quality	0.05	0-10		Can be overcome with over sampling
Data Completeness	0.1	0-10		Important for whole genome
Scalability	0.05	0-10		Scale up to 100% sampling not likely
Transferability	0.1	0-10		Important to adoption rate
Flexibility	0.05	0-10		Not critical for surveillance
Adaptability	0.05	0-10		Not critical for MRSA focused system
Modularity	0.05	0-10		Not critical for surveillance

Table 17 - Preferences for MCNF Algorithm

The resulting MCNF model was solved and yielded the solution shown in Figure 71 with a feasible and preferred path going through the "optimized" DNA extraction and Illumina Sequencing. However, sensitivity analysis on the model showed that small variation in the weights shown in Table 17, could result in the Roche 454 technology being the feasible and preferred path, although both paths selected the optimized DNA extraction design alternative. The "second-best" Roche 454 alternative is shown in a different shade along with the lower process cost path in Figure 71.

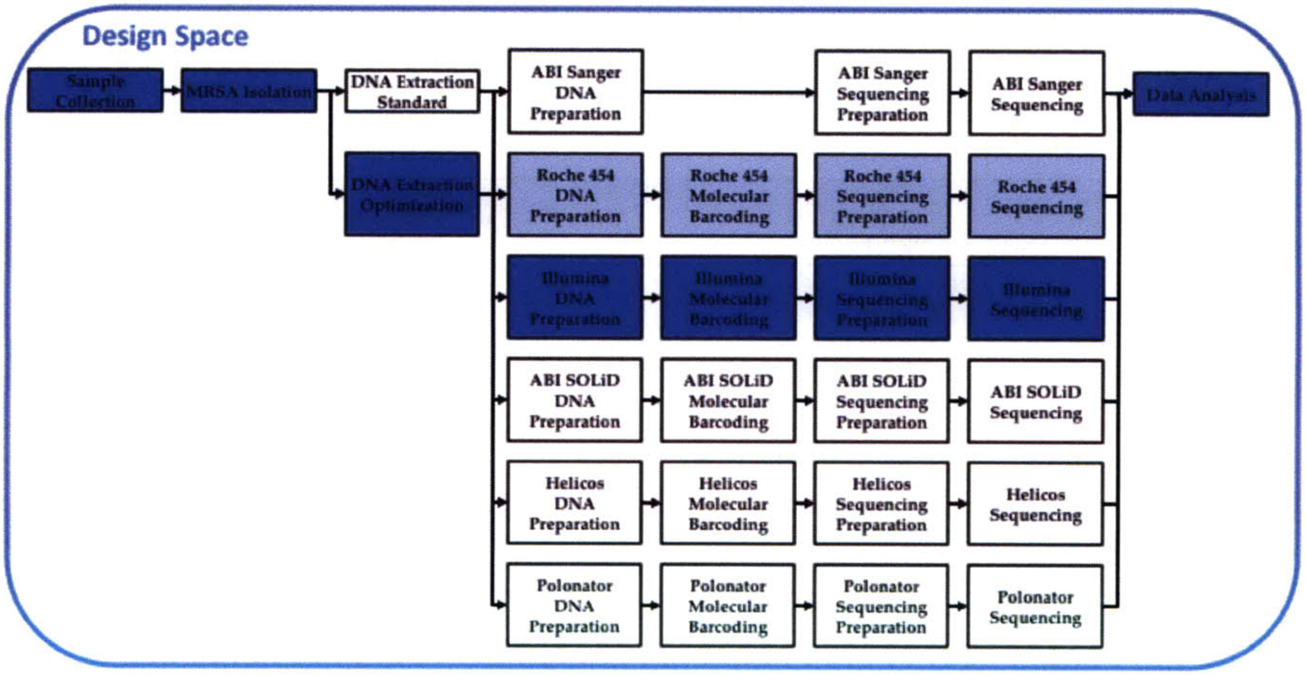

Design Space
Sample Collection
MRSA Isolation
DNA Extraction Standard
DNA Extraction Optimization
ABI Sanger DNA Preparation
ABI Sanger Sequencing Preparation
ABI Sanger Sequencing
Roche 454 DNA Preparation
Roche 454 Molecular Barcoding
Roche 454 Sequencing Preparation
Roche 454 Sequencing
Illumina DNA Preparation
Illumina Molecular Barcoding
Illumina Sequencing Preparation
Illumina Sequencing
ABI SOLiD DNA Preparation
ABI SOLiD Molecular Barcoding
ABI SOLiD Sequencing Preparation
ABI SOLiD Sequencing
Helicos DNA Preparation
Helicos Molecular Barcoding
Helicos Sequencing Preparation
Helicos Sequencing
Polonator DNA Preparation
Polonator Molecular Barcoding
Polonator Sequencing Preparation
Polonator Sequencing
Data Analysis

Figure 71 - Feasible and Next Feasible Surveillance Processes

The Roche 454 was a close second best despite a significant cost disadvantage per Gb of data produced. A sensitivity analysis revealed that a small change in the infrastructure variable weighting from 0.1 +/- 0.01 could change the outcome of the preferred path between Illumina and Roche 454. When considering all of the other "ilities" contained in the properties vector, the Roche 454 technology does have significant advantages in terms of infrastructure requirements, transferability, and cycle time which nearly make up the cost disadvantage deficit for this particular MRSA surveillance application when global system properties are considered. Given the relatively small size of MRSA genomes, the raw per Gb costs may not be as important as the other process "ilities" that could affect the speed and efficacy of the overall surveillance process. Marginal improvement of some of the Roche 454 system parameters could change the preferred path results and is a reason for the MDPM model to be developed in the next section to retain "second-best" paths as information elements the "process owners" must be aware of. More advanced technologies are on the horizon that could further improve on the design alternatives considered here, the impacts of which will be considered in Chapter 7. In addition, it is also possible to use the MCNF model in "reverse" to calculate what properties would be needed for process paths with certain desired characteristics and produce a "pareto front" of these alternatives to guide the design of new technologies and processes. This idea will be further explored in chapter 7. The feasibility analysis in this step of the MDPM methodology greatly reduces the number of alternative paths through the design space but also ensures "global" system constraints are met and a feasibility test has been performed. The design methodology then continues in the next section, which "fills in" the detail of the selected macro process.

5.1.6 Process Element Enumeration

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements **6.) Enumeration of process elements onto detailed process map for selected process**
Network Analysis	7.) MDPM creation from detailed process map 8.) Analysis of MDPM model

This section describes the identification of all necessary process elements to fully expand and flesh out the feasible preferred process path developed earlier. Each of the process elements described in this section represent the detailed decomposition of the macro process steps described previously. The particular process path selected at the macro level is shown in Figure 72, and each of the process activities will be further decomposed in the following sections. Also it is important to note that some description of the results from actually having run the process and knowledge from this is incorporated into the narratives for the enumeration of each of these macro activities. This is mainly done to help the reader, but the methodology would normally be run a priori without any knowledge of the downstream results, and likely would require iteration as new knowledge is found and is incorporated into the model.

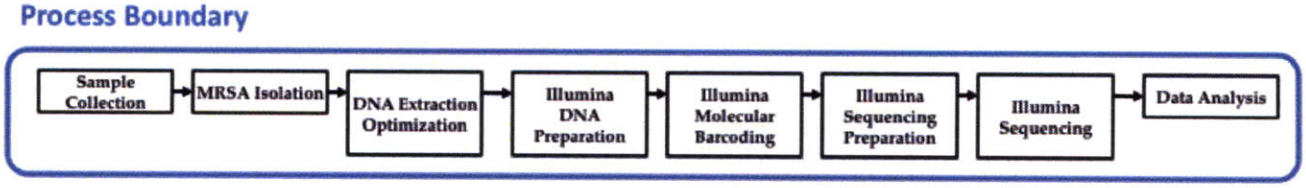
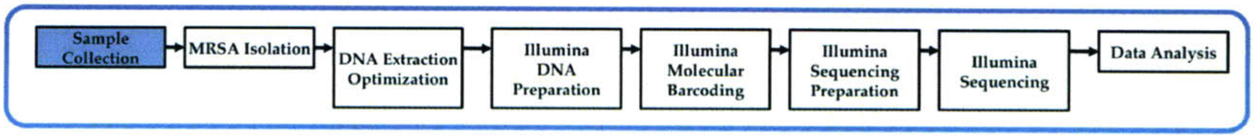


Figure 72 – Selected Feasible Process and Macro Process Elements Considered for Enumeration into MDPM Model

5.1.6.1 Sample Collection



The initial step in the process is the identification of a MRSA infection by a physician and their request for surveillance as shown in the decomposition of the macro step in Figure 73. Note that in every day hospital operation this first step is typically a request for diagnosis rather than surveillance, but in mapping this process the physician would request a "parallel" surveillance process to go along with any diagnostic request. A primary care physician or specialty physician will typically initiate the surveillance request starting the process. It is important to remember that MRSA can colonize skin and nares without infection but it is only when the microbe penetrates into the body that infections can occur. And these infections can occur throughout the body ranging from the bloodstream (septicemia) to internal organs to IV catheters. Therefore, nearly any "specialty" physician treating a patient for something else may also encounter a MRSA infection and require diagnosis and surveillance. This adds significant complexity to the process since in the organization layer the full range of specialty physicians, their nurses, and most importantly their patients must be modeled. For simplicity only a primary care physician and a "specialty" physician are modeled in the baseline process as representative of the "organization" of physicians and the node has an additional "organization" property which it represents an entire organization. This facilitates modeling but also visualization in the model, since it is not feasible to capture the detail of every organization the process interacts with, but it is important to model the interfaces between the process and these "organizations". In a network view the nodes with an "organization" property can be clearly highlighted to indicate interfaces with other networks and the complexity measures of the process can be adjusted to take into account the complexity that exists behind each of these nodes.

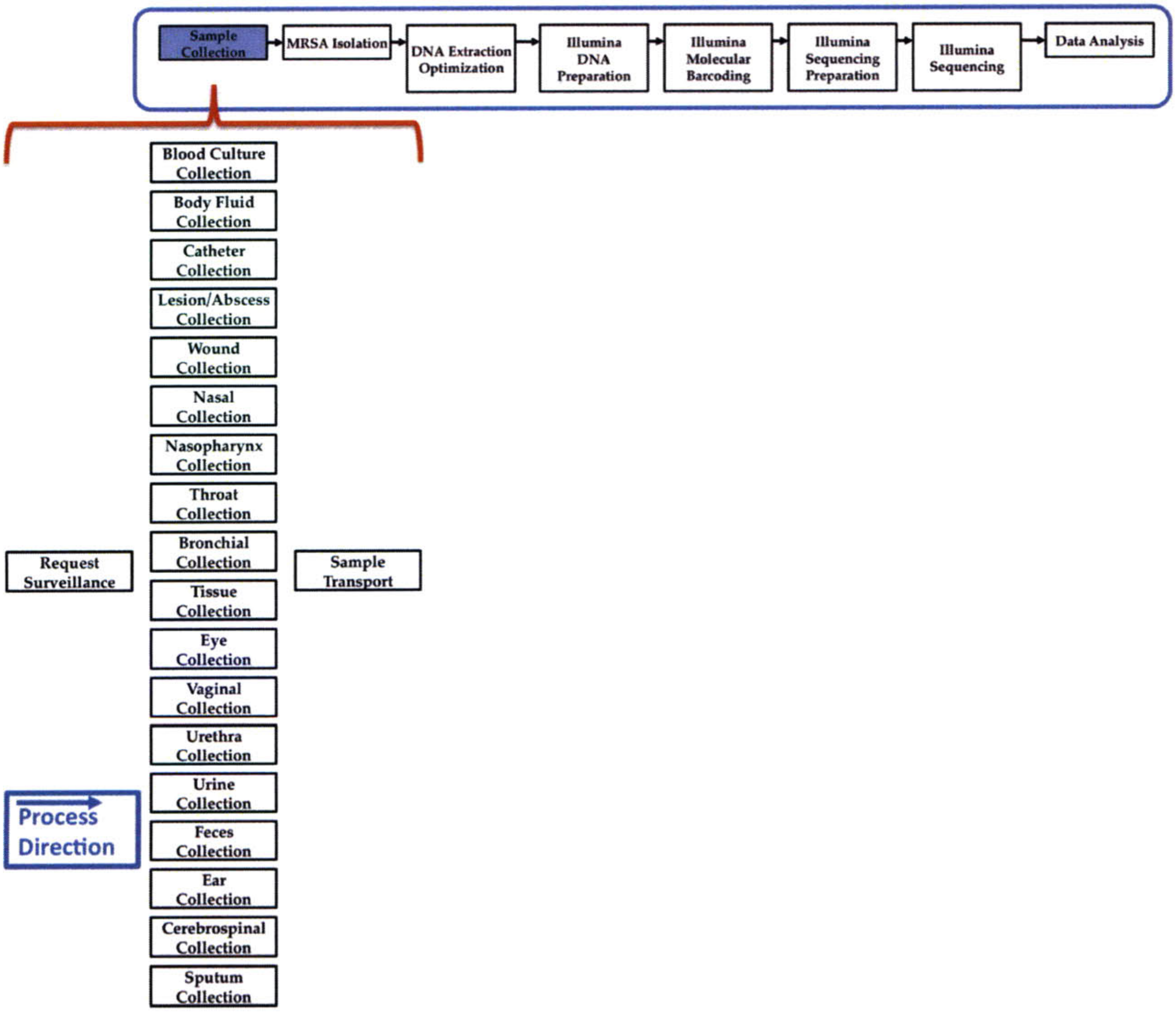


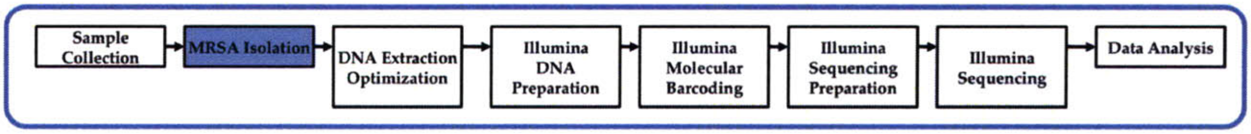
Figure 73 - Sample Collection Process Elements

Another important observation from the process map is that since MRSA infections can occur throughout the body, the collection of samples from each of these locations requires specialized work instructions, knowledge (tacit in the form of nurse experience), and are distinct activities significantly adding to the baseline complexity of the process. So for example collection of a MRSA sample from blood is very different from the collection from a deep tissue wound requiring different disinfection, collection, and preservation procedures. This diversity of sample collection activities is shown in Figure 73. The proper performance of these sample collection activities is critical because any contamination will lead to an improper surveillance result either identifying an incorrect organism, failing to detect MRSA, or incorrectly detecting MRSA when none is present.

The final step in the general sample collection category is the transport of the sample to the clinical microbiology laboratory. While this may seem like a trivial activity, the timing of it and handling during transport are also critical links in the process chain since delay in delivery can damage the sample or introduce the potential for contamination or sample misidentification. This is also a key observation for the importance of being able to "search" an entire process network since any step along the chain can cause a problem and unless the activities, individuals, and work instructions can be easily reached (or "searched") it is very difficult to manage or troubleshoot a complex process.

Sample collection is predominantly performed in the clinical domain largely associated with physicians, nursing staff, and other medical personnel interacting directly with the patient. Additional external stakeholders exist in this part of the process, represented by "organization" nodes such as the patient's insurance company, the hospital administration, the American Medical Association, and others. These external stakeholders are however not directly connected (in the organization domain) to many of the downstream process nodes and a key effort in building a sustainable surveillance process will be to map and maintain these connections.

5.1.6.2 MRSA Isolation



The next macro process activity is the isolation of the MRSA microbe from the collected samples which expands into a large number of detailed activities upon enumeration as shown in Figure 74. The first activity upon delivery is the triage of the sample depending upon what the suspected organism is, the collection method, and the intended assay as shown in Figure 74. While automated laboratory management systems can direct many of the samples to their appropriate preparation processes, a substantial amount of judgment (tacit knowledge) is still required typically on the part of a senior laboratory coordinator, manager, or technician. Once the sample has been appropriately routed, a specialty sample preparation is often required to match the specialty collection procedure used by the nursing staff. So for example a sample collected from bodily fluids may require centrifugation and other preparation before it can be used to inoculate a culture assay. The need for specialty preparations also adds significant complexity to the process but is again reflective of the "real" complexity since simply describing this activity as "sample prep" might simplify a process representation but would hide the real complexity behind it.

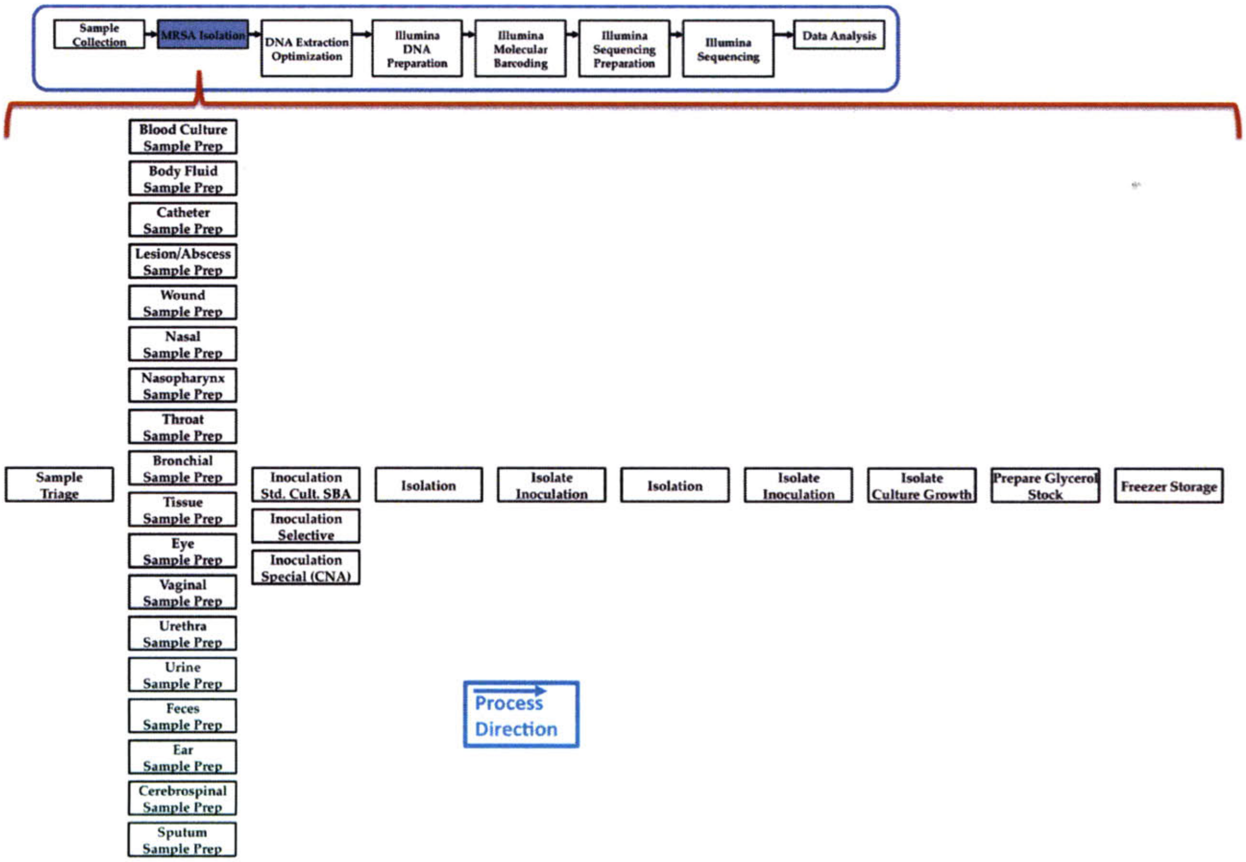


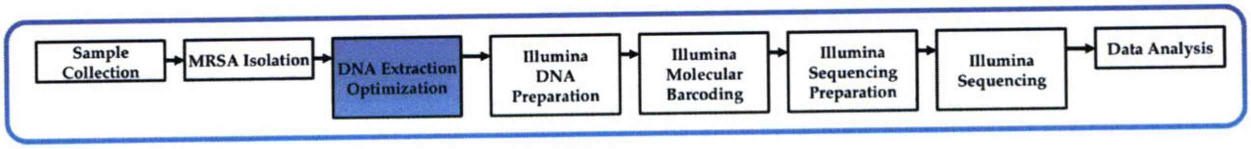
Figure 74 - MRSA Isolation Process Elements

Once the sample has been prepared and processed it can then be used to inoculate culture media such that the MRSA microbes can be identified and isolated as shown following the sample preparation steps in Figure 74. Significant operator discretion exists here as there are numerous choices for the type of media to be used, each with particular benefits and drawbacks. For example, selective media is designed to inhibit the growth of all other microbes except for MRSA, and while this might seem useful it could eliminate microbes critical for surveillance such as “near-MRSA” that are close to being resistant or have certain key resistance genes from other organisms. Alternatively, using nonselective media allows for the growth of all microbes and may hide the MRSA microbes in heavily “contaminated” samples such as those coming from the digestive system. Only three “choices” are modeled in the process enumeration for simplicity but in reality many more choices exist. The existence of these choices is modeled in the information domain, where lab managers will be aware of these other choices even though they might not exist as process domain choices since they may require specialized materials not routinely available. However, the existence of these choices in the information domain better represents the real complexity of these processes since it is possible that a lab manager may decide to modify a process domain activity based on knowledge from these "choices" (which may be thought of as design alternatives in traditional engineering). The inoculation step is also heavily technique dependent as the sample must be "streaked" across the media in a particular pattern, volume, and sterile technique. Poor technique will make it difficult to identify MRSA during the isolation activity.

Once the appropriate culture media has been inoculated it must be grown, typically overnight under specific conditions for temperature, agitation, and aeration. The resulting media can then be examined for the presence of MRSA under the microscope or sometimes visually if selective media was used. A single colony can then be isolated and used to inoculate a second media container in order to grow sufficient material, although now from a single MRSA isolate. Culture based diagnostics would stop here as the identification of MRSA following the first culture step would be sufficient to render a diagnostic test result. However, for any subsequent surveillance process the second growth step is necessary to produce an archival sample that with the addition of glycerol can be frozen. Aliquots from this frozen sample can then be taken and re-grown to produce additional material from the original isolate. This renewable resource can then be used for a variety of processes including DNA sequencing-based surveillance.

The organizational domain network associated with this part of the process was in its baseline (original configuration) not connected to many of the other organization domain networks. It so happened that at the Brigham and Women's Hospital, where this process was developed there existed a strong set of ties between the Microbiology Lab Director, who also happened to be a scientist in a research role at Harvard Medical School, and a research Physician who was interested in MRSA mutation discovery and research. These connections might not normally occur in other places where microbiology labs can be disconnected from medical research in the organization domain (i.e. only reporting sample results without additional relationships). The existence and development of these connections enabled this process to work but they were haphazard and often relied on the author's interest in creating the process. An important goal of the MDPM model is to enable future process designers to actively map the necessary relationships to manage complex processes, and the enumeration of all individuals and organizations interacting with the process is a key element of the MDPM model which will de described later in this chapter.

5.1.6.3 DNA Extraction



The next major macro process step is DNA extraction optimization, which was selected over the "standard" DNA extraction as it was expected that this activity would require significant optimization. The activity is called DNA extraction optimization recognizing that some work will be needed to make it perform better than the standard and is a "placeholder" activity as this macro process step did not exist prior to the development of this process. This is an example of how the process feasibility may be used to identify key process activities needed. The first detailed step in the decomposition of this process activity is to take small aliquots from the frozen MRSA samples, thaw them, and re-grow them to provide sufficient material for DNA extraction as shown in Figure 75. This re-growth is in a larger volume and liquid solution rather than solid media as is typical for diagnostic cultures. This part of the process starts to deviate from "standard" diagnostic processes, as large-scale high-quality DNA extraction from Staphylococcus aureus is not a routine procedure. While there are a number of protocols for Staphylococcus aureus DNA extraction many of them are for less demanding processes such as PCR which does not require large amounts of high molecular weight DNA (greater than 20 ug with average size greater than 20kb). In addition, the objective of this process was to develop a high throughput method that could use commercially available reagents were possible and would require relatively little specialized operator technique (tacit knowledge). However these requirements combined with significant variability in the MRSA microbes themselves led to serious process design problems that required significant iteration and knowledge sharing between the microbiology research team at Harvard, DNA extraction vendors, DNA extraction scientists, and the author. The design problems, their representation in the MDPM model, and their eventual resolution are described later in this chapter. However, it is important to point out here as well that the connections between the microbiology research group at Harvard, DNA extraction scientists, vendor organizations, and the associated explicit knowledge (such as vendor instructions and literature) does not normally exist and was developed as "scaffolding" to support the creation of this process.

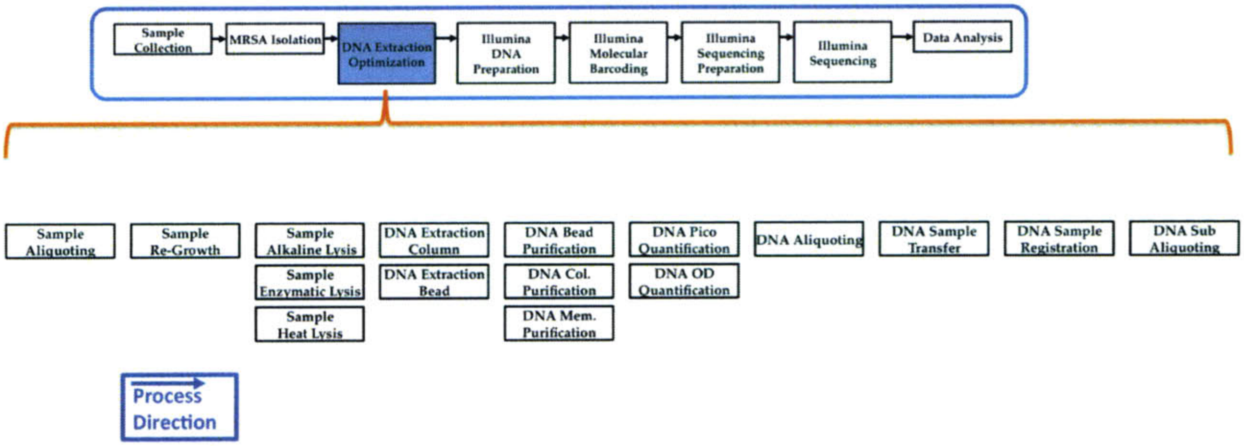


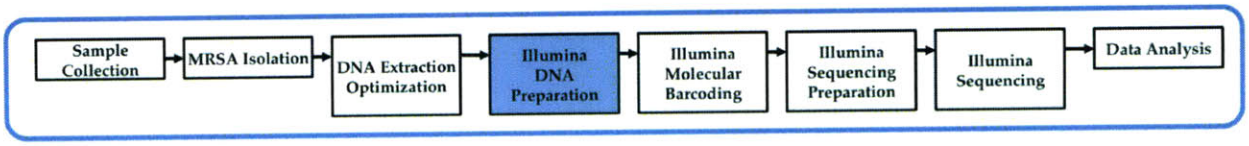
Figure 75 - DNA Extraction Process Elements

Downstream analysis of the process data found significant impacts from the quality of the DNA extraction emphasizing the need for a process network that can be easily reached and "searched". The variability in Staphylococcus aureus cell wall thickness made it necessary to optimize a protocol that would lyse the majority of these microbes (both very thick ones and relatively thin ones) without damaging the DNA of the relatively thin ones that would be exposed to much harsher conditions to ensure the very thick ones are always lysed. The damaged DNA would then show up as "missing" parts of the genome or would require significantly more material further complicating the extraction step. This is again an example of seemingly innocuous process steps having significant impacts downstream in complex processes and the need to integrate all activities to ensure local activities meet the global process constraints and requirements. Taking a page from Steve Spear's ideas in (Spear 2008) and translating them into the MDPM model, this problem was rapidly identified because there was a single "process owner", the author, who was only a few network hops away from any given part of the process and it was rapidly "swarmed" by bringing expert individuals and experimental attention to bear on the problem. This is again an example of the value of explicitly enumerating the process activities, the organizational elements, the information elements, and the connections (or lack thereof) between them in the MDPM model.

In the baseline process map three lysis alternatives are shown in Figure 75, and in fact all three are viable as tested in the development of this process and shown in the literature. There is discretion as to which method might be best in a particular setting (another hospital may prefer to optimize a different one) but the enzymatic lysis method was used in this process. In order to capture the real process complexity all three are shown. In addition to alternative DNA extraction methods, column and bead, are shown which also represent feasible process choices although the optimization necessary to develop the final work instructions (protocol) actually used is shown as "optimization documents" in the information domain. This is typical of how many of these protocols are developed where a series of design experiments and trial protocols will eventually lead to the final "production" protocol. However, there is enormous knowledge captured in these experiments and trial protocols that can guide troubleshooting or future design changes. And that is only if the knowledge of the people involved in the optimizations and trial protocols is formally captured as it otherwise resides as tacit knowledge with individuals in the organizational layer. In this case the information was explicitly captured and is represented by a "optimization documents" node with an additional knowledge base "property" signifying the node represents a deep and complex knowledge set not explicitly represented. In this sense the special knowledge base nodes are much like the "organizational nodes" that represent additional complexity not modeled but whose interface is critical to process operation. The special nodes can then be easily identified in the network representation such that the critical interfaces between the process and these knowledge bases can be identified.

The extracted DNA must be purified to eliminate any other cellular debris and extraction reagents. The DNA must then be quantified using one of a variety of methods to ensure the extraction has been successful and deliver the DNA at a known concentration in a stabilization buffer. Typically a smaller aliquot will be prepared for use in a sequencing process in order to preserve the larger amount of DNA for future use, and this aliquot is then handed off to start the next step of the process, library construction.

5.1.6.4 DNA Preparation



The extracted DNA sample can then be put into the start of the sequencing process. Because of the previous MCNF based feasibility selection, only a single DNA sequencing process is shown in this process map, but the other alternatives described in the design space shown in Figure 66 also exist. The major benefit of the design space reduction from previous methodology steps is that now only this single process path need be considered as part of the process. However, any of the existing and emerging sequencing technologies could rapidly displace the one shown across multiple performance dimensions. This multitude of design choices contributes to the effective complexity of the process since in this rapidly changing environment that the other "design alternatives" must be continuously considered by the organizational domain individuals responsible for the sequencing portions as well as the overall "process owners". These design alternatives are shown in the information domain as vendor documentation nodes and in the organizational domain as nodes associated with the vendor organizations. The purpose of these "alternative process" nodes is again to map the interfaces between the process and any other sources of information, organizational interaction, or in this case alternative processes.

But even for the single sequencing process selected the complexity of the process is substantial as measured by the number of individual steps necessary. Because of the high degree of competition and rapid evolution between the sequencing technologies many of these process steps are either semi-tested prototype protocols or for those fully tested subject to change as improvements or system changes occur. This requires an enormous amount of organizational "scaffolding" to test the changes, co-develop them with vendor organizations, and implement them. In addition, for particular applications a number of "integration" technologies must be developed in order to feed nonstandard samples such as the MRSA ones into processes not initially designed to handle them, or handle them at the high throughput necessary for a surveillance system.

The first step in DNA preparation for sequencing is to shear an aliquot of DNA from the previous step using either acoustic or hydrodynamic forces in order to break up the original 2.8 MB Staphylococcus aureus genome into small fragments that can be processed by the sequencing technology. The enumeration of detailed process activities for DNA preparation is shown in Figure 76. The particular choice of shearing technology and the exact settings used in the protocol are (as with most of the sequencing process steps) derived from significant optimization experiments and the knowledge base associated with these is often tacit knowledge contained in the operator's minds, but there also exist explicit optimization documents and protocols associated with them. In the MDPM model the links between the organizational domain and these optimizations are explicitly shown since in many cases the optimizations will have been performed by senior scientists and research staff not directly working on the process but may be not be "reachable" by the operator on the front lines through organizational connections. The operator however, may need to troubleshoot or identify the impact of changes on this particular process step and requires an efficient network to obtain the necessary information. Ensuring these connections are available for complex processes is key to the rapid identification of problems and also the "swarming" necessary to resolve them.

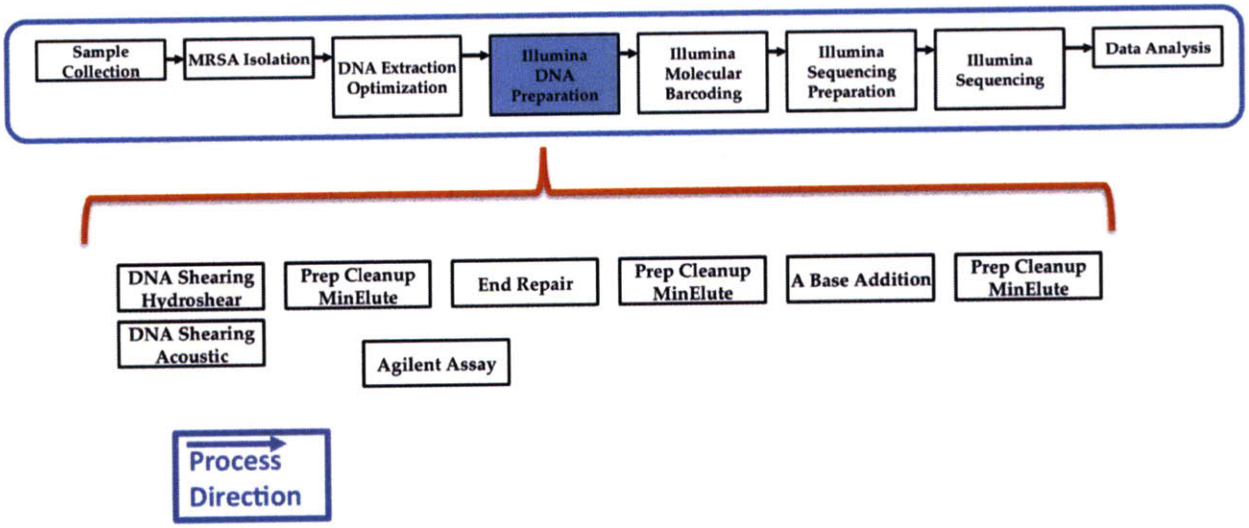


Figure 76 – DNA Preparation Process Elements

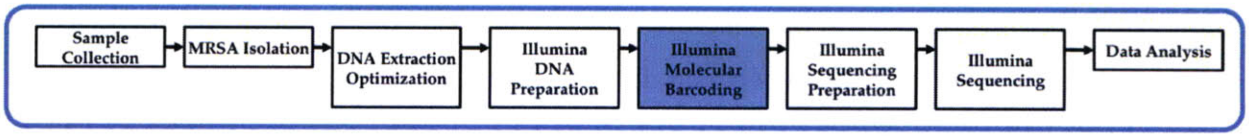
Following shearing are a series of steps to further prepare the genomic DNA into molecules that can be read by the selected sequencing technology as shown in Figure 76. It is important to note that not all of the detailed activities shown in Figure 76 are from the same vendor but are sourced from multiple vendors providing molecular biology “tools”. And although these tools are commonplace in molecular biology they still require significant operator skill or tacit knowledge and it is common to find in organizations dealing with molecular biology certain “golden hands” individuals who are especially skilled and adept at using these “tools”. Examples are the many “cleanups” necessary to remove various undesirable post reaction products at various steps. These cleanups can severely affect DNA yield, which can in turn affect the quality and amount of downstream data produced. It is thus especially important to map not only the explicit vendor instructions for these cleanups, the process steps at which they occur, but also the individuals associated with these steps and their tacit knowledge.

Significant efforts to automate many of the steps in this upstream sample prep have occurred in a number of genome sequencing Centers including the Broad Institute to match the rapidly rising sequencing throughput and not directly associated with this thesis. And from the MDPM model perspective these automation efforts transferred much of the tacit operator knowledge into explicit knowledge now codified in detailed robotic scripts and protocols in the information domain. In general, this thesis has built on elements of rapidly improving sequencing processes (intended for other uses such as cancer sequencing) and focused on developing the necessary "connections" between these elements to forge them into a MRSA focused surveillance process. These "connections" were primarily across boundaries such as the optimizations needed for DNA extraction described earlier, the molecular barcoding needed for multiplexing, and the mutation analysis algorithms. And this is likely representative of the development of new complex diagnostic processes in healthcare, which are built from parts of existing technologies and forged together by developing the necessary "connection" technologies. From this perspective it is especially important to have detailed process mapping methodologies such as the MDPM model, which can be used to identify the "holes" in complex processes such that robust and efficient "connection" technologies can be developed. This is a corollary to the "swarming" problems and much more like Steve Spear's "first capability" in focusing organizations at the start of a design process on the biggest and most important problems to ensure the highest probability of success. This would for example show process designers that additional "scaffolding" is needed around key "connections" that are not sourced from existing processes and will require significant resources and connected development iteration.

In many of the examples in this thesis of building this complex MRSA surveillance process the phrase "connected development" has been used. This is an important concept directly related to the fundamental goal of this thesis, which is to develop an analytical complex process design framework. Connected development is meant to describe the need for rapid local optimization but subject to global process constraints. So for example a team working on the optimization of cleanup steps in library construction must be able to "reach" downstream analysis individuals and data to ensure that any changes to not negatively impact the process despite perhaps being a "local optimum" for that process step. These connections are especially critical in lengthy complex processes such as the one described here since development and improvement can be slow or even impossible unless there is sufficient visibility from local teams into the overall process. Without this visibility it is likely that organizations will be very hesitant to make changes both on a global and local level since the impact of changes cannot easily be identified or related back to the groups making the changes. It is possible that the empirically observed "resistance to change" in complex processes is due to a lack of sufficient "scaffolding" or connections across the process to understand and manage changes leading to "complexity freezes" of processes in certain suboptimal but working conditions. Rather than risk potential failures, unintended behavior, or greater variability organizations will freeze processes as they may not have the necessary additional resources to build the "scaffolding" necessary to improve them.

The remainder of the DNA preparation activity includes the repair of DNA molecule ends following shearing such that adapters with known DNA sequences can be added to the genomic DNA molecules in order to facilitate their manipulation and enrichment. These steps are critical to one of the key "connecting" technologies in this process, the molecular barcoding of each of the original MRSA samples collected and isolated from patients which is describe next.

5.1.6.5 Molecular Barcoding



This activity was also a "placeholder" activity in the initial design space, as many of the procedures used in this step did not exist at the initial design and feasibility selection stages. The rapid decreases in DNA sequencing costs are largely associated with enormous increases in the amount of data produced for each instrument run. Where during the human genome project the average yield per run might be 62,000 high-quality bases the current (ca. 2009) average yield per run for the most efficient technologies is around 30 billion high-quality bases per run. This batch size is governed by the particular properties of the technologies but can generally be thought of as moving from one dimension in human genome project capillary technologies to two dimensions for current technologies.

These two-dimensional surface technologies gain greater efficiency the more densely DNA can be packed on them and the greater the surface area per run leading to larger and larger batch sizes. For comparison, the Staphylococcus aureus genome is only about 2.8 million bases compared to 30 billion bases (and growing) run sizes which means a single MRSA sample would be read over 10,000 times in a single run. In order to detect mutations with sufficient statistical significance each sample would only need to be read about 30 times. Most of these technologies are designed for high throughput human genome sequencing but this does not easily scale down to the relatively small microbial genomes. In order to take advantage of these efficiencies a means of multiplexing many microbial genomes on a single run is necessary, leading to the need for a "connecting" technology to accomplish this. Methods exist to add molecular barcodes to genomic DNA fragments in order to identify them, and this consists of ligating (joining) a known DNA fragment with a particular sequence (the barcode) to the genomic DNA fragments. When this combined DNA molecule is read the portion from the known fragment is the barcode and the actual DNA sequence is read immediately afterwards. Of course this molecular barcode must be added in such a way so as to not interfere with the other adapters and molecular manipulations needed to run genomic fragments through the sequencing process. This required a significant amount of "connected development" as not only did the molecular biology need to be worked out but also the software analysis tools that would correctly separate the barcodes in the downstream data such that the pooled samples could be correctly identified and their data binned. An initial version from the sequencing vendor was tested but this version proved to have problems in that adding the molecular barcodes also prevented certain portions of the genome from continuing through the process leading to holes in the sequence. These problems were only identified and a "swarm team" assigned to fix them because of the connected development model, which rapidly fed back results from the analysis (far downstream of the molecular biology) and this connected iteration continued until an alternative process was developed. The MRSA surveillance process used for the actual demonstration used this alternative method including the necessary downstream steps (both molecular biology and software).

Additional steps are needed for this molecular barcoding method to work such as removing the unused barcodes, size selecting the fragments within a certain range, enriching these fragments for the proper orientation of barcodes and adapters, additional cleanups, and a quantification of the amount and size of the resulting products. Once individual samples have received their molecular barcodes, these must be pooled together in equal molar concentrations such that the data produced by the sequencing run is equally distributed amongst all the samples barcoded as shown in Figure 77.

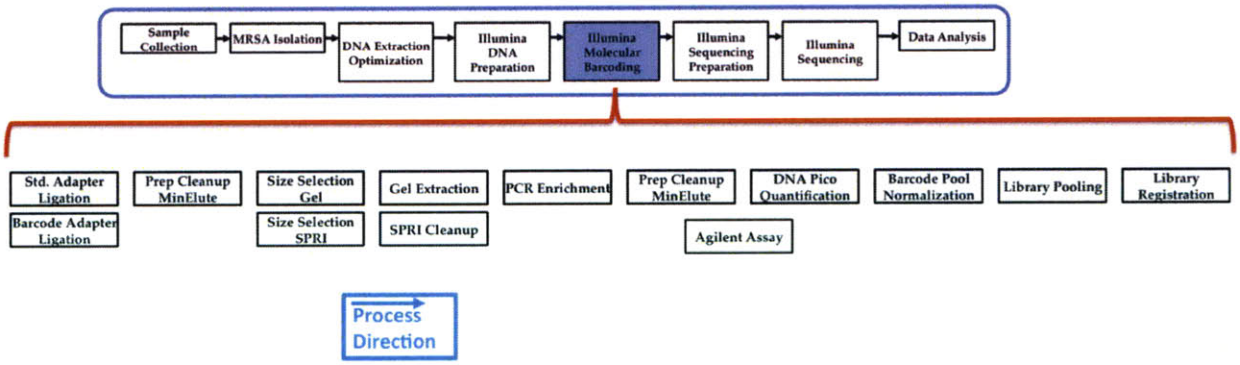
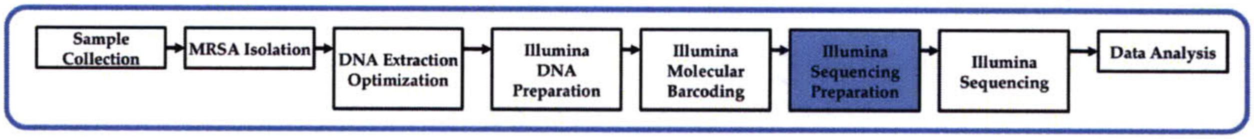


Figure 77 – Molecular Barcoding Process Element Enumeration

The molecular barcoding shown in Figure 77 is a key "connecting" technology as precise quantification and automation are needed to normalize each sample and pool in the correct amounts prior to the preparation of a "library" for sequencing. This required a significant amount of connected development as the readout for the pooling had to go through the entire process all the way to downstream analysis steps and back to the team working on the barcode pooling. Note that unlike abstract design processes, the particular order of physical process steps cannot easily be changed and some of the techniques used in DSM to re-order the sequence of process tasks may not apply directly as physical process activities cannot be easily moved together. However, once these optimizations for complete and automation developed barcode pools could be submitted to the next section of the process, preparation of the now molecularly barcoded samples for sequencing.

5.1.6.6 Sequencing Preparation



Once past molecular barcoding and pooling the process follows a relatively straightforward linear sequence of steps to add the barcoded sample fragments onto a glass slide and amplify them such that there are sufficient molecules to "see" on the sequencing instrument when fluorescent nucleotides are incorporated. However, as with the rest of the process most of the steps are still under substantial development as improvements are made or problems resolved. This requires a substantial "scaffolding" here as well since changes in the way these "DNA clusters" are made can have a very significant impact on the downstream data. The physical process steps are enumerated in Figure 78, and on a traditional flow-chart based process map would appear to be a relatively simple process. However, when additional domains are mapped onto these process activities in the MDPM model, such as the many individuals, work instructions, optimizations, vendor data, and other elements that are associated with these activities, the actual complexity of the process can be evaluated.

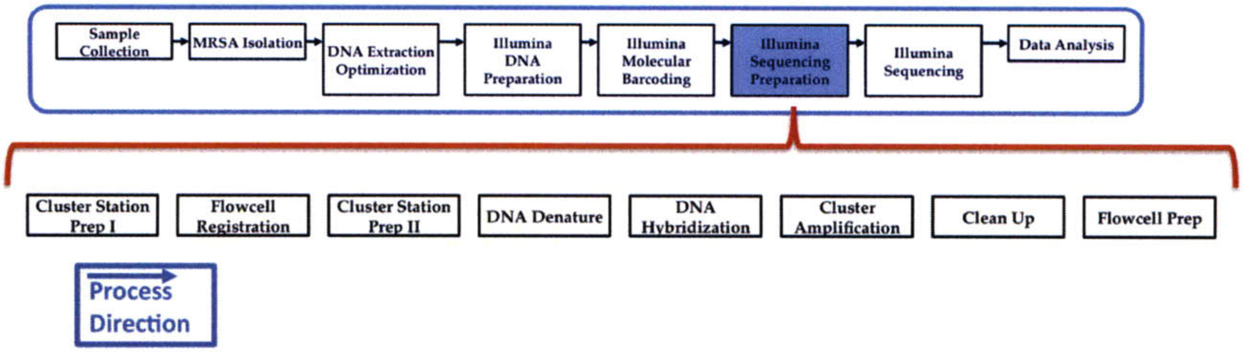
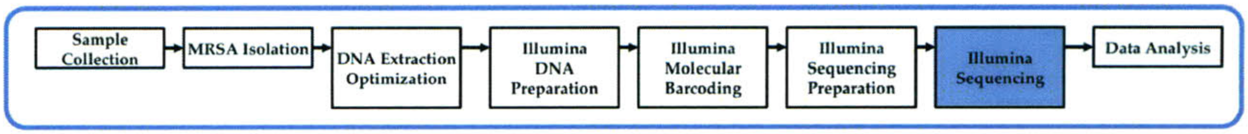


Figure 78 – Sequencing Preparation Process Elements

Throughout the process it is necessary to carefully track the genealogy of every sample to ensure the final downstream data can be precisely matched to the original sample without any mixups or errors. The surveillance process as designed will have a very significant throughput, in the thousands of samples per year, and additional infrastructure is needed to track these samples using standard (not molecular) barcodes on every sample container, intermediate tubes, glass slides, and any other physical sample transfer in the process starting with the initial sample collection. This laboratory information management system (LIMS) must keep track of the genealogy (sample transfer history) of every sample from patient to analysis data set. This adds significant complexity to the process but is not usually mapped in traditional process flow maps that focus only on the progress of the physical sample. The interaction of the organizational domain and the LIMS is essential in ensuring that the process can be controlled and using the MDPM model is a way to identify "holes" where the LIMS is not interacting with the process or the organization.

5.1.6.7 Sequencing



Once the DNA has been deposited onto glass slides and amplified, the Illumina instrument can be prepared to read the DNA molecules using a modified sequencing by synthesis (Sanger) chemistry. One disadvantage of the next generation high throughput sequencing technologies is that their final step, reading the DNA on the instrument, can be very long in the range of 6 to 14 days. This places tremendous emphasis on the careful setup and preparation of an instrument run since a failure during this time is often not recoverable and requires that any rework start many steps back in the process. The design and development of these instruments is continuously changing to address higher efficiency, process issues, and data quality, which requires significant process infrastructure, "scaffolding" to support. The sequencing portion of the process is also straightforward and linear in the process domain, but again this hides the significant associated complexity due to the ongoing change and optimization that will only be evident in the organizational and information domains. There are a significant number of sequencing process activities nevertheless as shown in Figure 79, and because of the length of a sequencing run, it is critical that all of these activities have high reliability and pass rates as rework is difficult or impossible to perform for many steps (i.e. returning to the previous step). This places a significant premium on ongoing process improvement to make the pass rate for each activity as high as possible, but this in turn requires considerable attention to the tacit knowledge operators may have as they perform these activities.

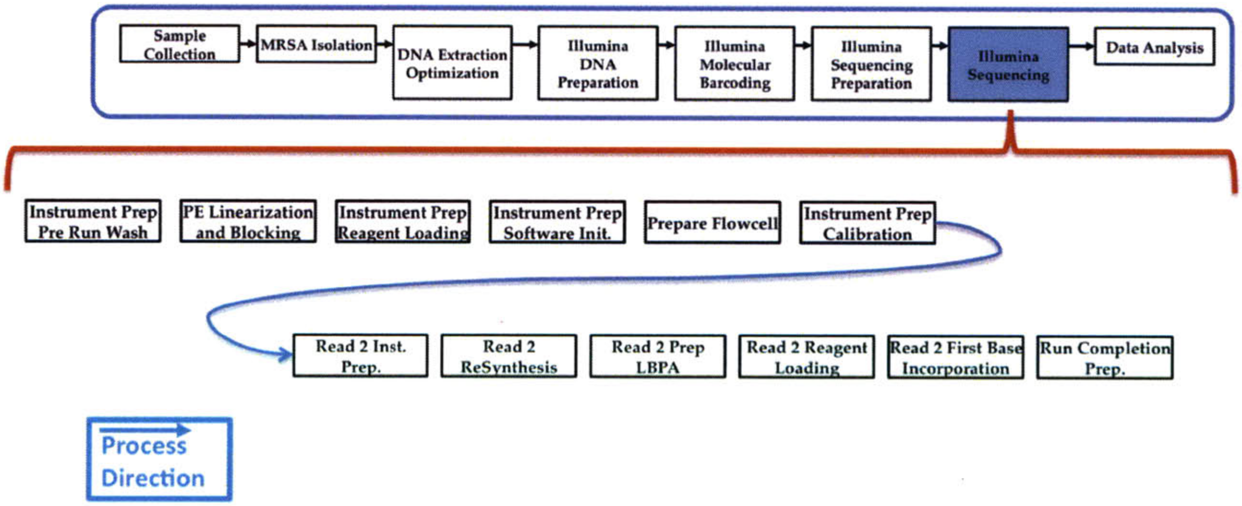
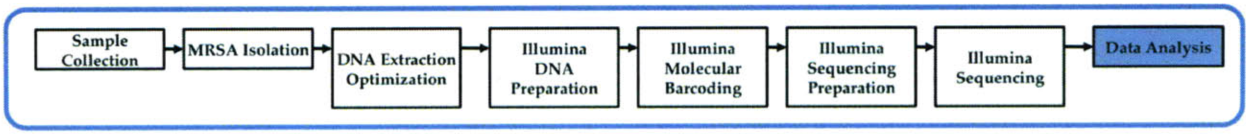

Sample Collection
MRSA Isolation
DNA Extraction Optimization
Illumina DNA Preparation
Illumina Molecular Barcoding
Illumina Sequencing Preparation
Illumina Sequencing
Data Analysis
Instrument Prep Pre Run Wash
PE Linearization and Blocking
Instrument Prep Reagent Loading
Instrument Prep Software Init.
Prepare Flowcell
Instrument Prep Calibration
Read 2 Inst. Prep.
Read 2 ReSynthesis
Read 2 Prep LBPA
Read 2 Reagent Loading
Read 2 First Base Incorporation
Run Completion Prep.
Process Direction

Figure 79 - Sequencing Process Elements

The instrument is also where the interface between physical DNA molecules and sequence data exists starting with the processing of images of the DNA molecules with fluorescent nucleotides for each base into the DNA sequence of each molecule. The instrument optics, image processing algorithms, and base calling (correlating a fluorescent signal to one of the four nucleotides) all directly affect the quality of the data and the ability of the MRSA surveillance process to correctly identify novel mutations and strains. Connected development is necessary to iterate through many different versions of instrument optics, image processing algorithms, and base calling algorithms. And once the raw base calls have been made data must be filtered to make sure it is valid. So for example, the molecular barcodes added many steps before must now be checked using error correcting algorithms such that any barcodes that do not match those originally put in (due to sequencing errors or bad quality in particular portions of the run) can be discarded. This requires that the original molecular barcodes be designed in such a way that error-correcting (Hamming) algorithms can take advantage of properly separated nucleotide sequences. For example one would not want barcodes to have nearly identical sequences as a single error would make them indistinguishable from one another but rather they must be designed to have a maximum information entropy such that every barcode is as different from another as possible (which makes them much more tolerant of errors while still being identifiable from one another). This shows the extent that a "connecting" technology must cover across the process to fully integrate the various steps. An alternative view of these "connecting" technologies is the technology invasiveness concept developed by (Smaling and de Weck 2007) but which here is applied to processes rather than product design.

5.1.6.8 Analysis



The final macro section in the process is the analysis of the raw data produced by the sequencing process. This activity is also a "placeholder" as many of the algorithms and software are under significant development. The analysis activity involves validating the data and filtering for low quality and errors. Much of the technology here is under active development and very much in the prototype stage as the very first versions were only developed recently for the human genome project and have had to be essentially redone for the new sequencing technologies as described in (Butler, MacCallum et al. 2008). In addition to the software provided by vendors, a number of other software packages exist, some commercial and some academic, with varying levels of support and emphasis for microbial analysis. This is the final and perhaps most critical "connecting" technology in the MRSA surveillance process since the algorithmic and computational problem is very significant but it is equally important to condense and synthesize the information such that end-users like physicians, epidemiologists, and clinical microbiologists can use this process to make effective observations and decisions on MRSA control. The high complexity of this first prototype process must eventually be reduced for broader implementation of surveillance. Yet, the process is still under active development and will continue to need significant "scaffolding" as changes and improvements continue to be made. Potential strategies to deal with this type of change in complex process are described in Chapter 7. The enumeration of the analysis process steps is shown in Figure 80, which describes the image analysis step that translates fluorescence from each nucleotide label into raw data to be interpreted as one of the four bases by the basecalling step followed by a filtering of this data to remove lower quality data. The filtered data can then be used to align individual "reads" corresponding to 50-100 base segments the instrument is able to read. There are millions of such reads, and a key limitation of next generation technologies is that the reads are very short compared to the size of the genomes being studied (i.e. 50 bases versus a 2.8Mb MRSA genome). This means that in order to fully assemble a genome from these short reads, over 30-100X the size of the genome being studied are needed in order to have enough overlap and statistical confidence in their correctness to put the pieces of the "puzzle" together. This is what is done at the alignment and assembly stages. The alignment step simply takes all of the reads and "places" them on top of a reference genome, such as an already sequenced MRSA, such that differences can be identified. Assembly is a much more demanding process where the reads are assembled *de novo* into a "new" genome without the assistance of a reference, and much of this algorithm technology using next generation data is under development. With the alignments or a fully assembled genome, biological, epidemiological, and surveillance activities can finally be performed as shown in Figure 80.

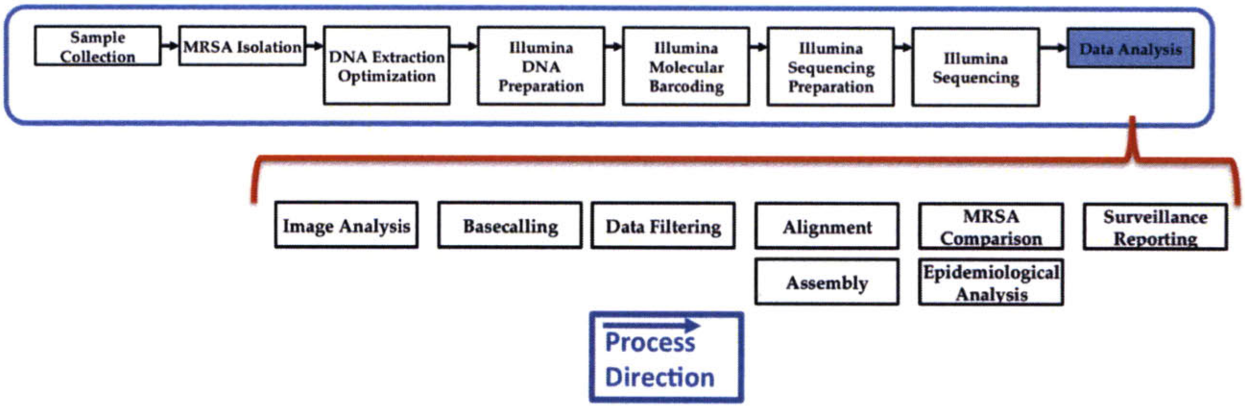


Figure 80 - Analysis Process Elements

Any potential discoveries in the data must be verified biologically by performing confirmatory experiments on preserved samples. These confirmatory experiments will be critical to the larger adoption of the process as it must be "trusted" and verified independently, but before this can happen it must also be transferred elsewhere such that the results can be reproduced and the medical community can verify the results. This later part of the process life cycle is beyond the scope of this thesis, but could conceptually also be modeled using the MDPM process model to facilitate the transfer and independent verification of this process.

5.2 Multi Domain Process Model (MDPM) Creation

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	**7.) MDPM creation from detailed process map** 8.) Analysis of MDPM model

With the detailed enumeration of process elements developed in the previous sections it is now possible to create the MDPM model for the MRSA surveillance process. To facilitate the understanding of the model application, the baseline process (without any improvements) is compared to the "V1" process (the one actually used in this thesis) in the following narrative. However, the general methodology would have a process designer use the MDPM methodology to first identify the "baseline" process elements and use this to iteratively find necessary additional elements or connections to make an improved process work. Thus, the "baseline" description below refers to the initial set of process elements identified in each of the domains that would need to be assembled into the version 1 (V1) initial process. The baseline description is useful in understanding the range of different elements that will need to be connected to form a working process and also provides a reference as to which elements are already connected. A process designer can thus focus on adding only new connections and elements rather than duplicating existing ones or adding unnecessary complexity.

5.2.1 Process Domain

The first step in building the process domain is to translate the detailed process map developed previously into an activity-on-node directed graph representing the process. Note that the previous enumeration was designed to produce the set of elements but these have to be connected to each other according to the process paths available through the process network. These are design choices or processing alternatives that exist and can be exercised. The design choices described here are specific to the particular macro process selected using the MCNF feasibility and element enumeration. However, reflecting the iterative nature of the MDPM methodology if new feasible detail level activities are found during initial testing, these are also reflected in the process domain. So for example, design alternatives for molecular barcoding, data analysis, and upstream DNA extraction are shown here as they represent alternative paths at the detail level. Building the process domain is relatively straightforward in that each activity corresponds to a node and it is most important to properly represent which nodes are connected to others particularly at branch points. Also, the baseline process layer network does not contain any rework loops as these were still largely unknown in the initial baseline stage. The most important goal of the initial iteration of the methodology is to capture all of the necessary activities and their relationships especially "connecting" activities spanning boundaries between processes joined together such as DNA extraction which connects MRSA isolation and downstream sequencing steps. The resulting process domain directed network graph representing the detailed enumeration of process activities carried out in section 5.1.6 is shown in Figure 81 along with the corresponding "macro" process activities.

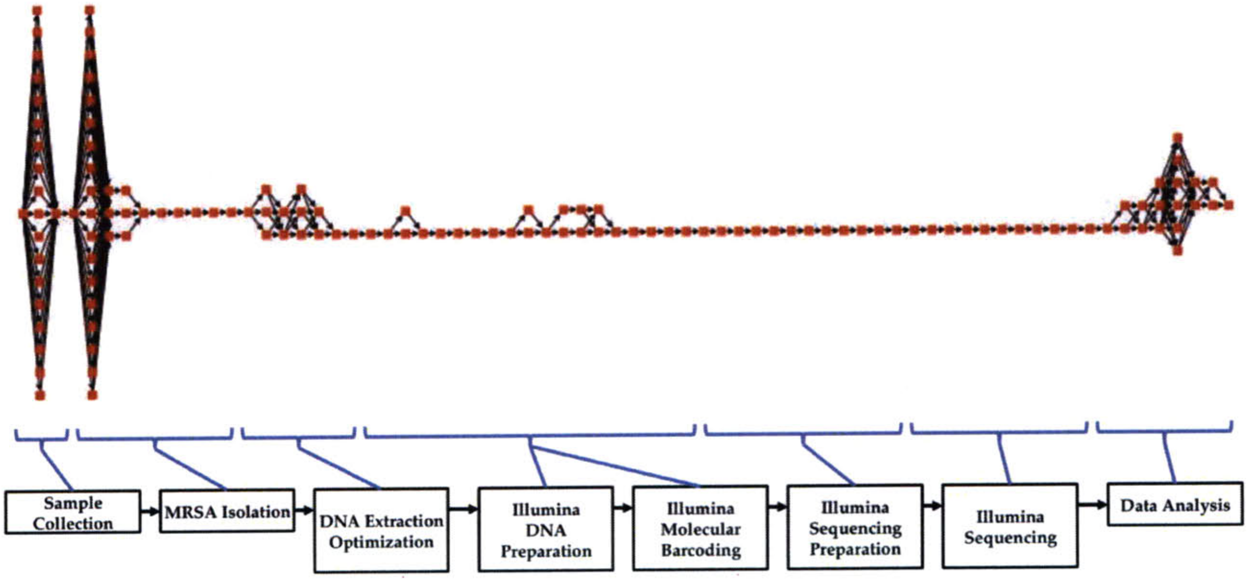


Figure 81 - Baseline MDPM Process Domain for Surveillance Process

As with many graph-based process representations, the baseline MRSA process domain has a number of sequential steps that would at first glance appear to have relatively little complexity. However, this is mainly because the design alternatives have been filtered out to a single feasible preferred path previously using the MCNF algorithm such that the construction of the MDPM model can be demonstrated and the complexity analysis focused on the multi-domain interactions for a single process. The MDPM model could be extended to include the other process paths, such as the Roche 454 design alternative, and this should likely be done in an enterprise MDPM model where both paths might be utilized. So for example, the speed advantages of the Roche 454 path might be combined with the greater per Gb efficiency of the Illumina process to produce a hybrid process that had both a fast cycle time for a subset of samples and a low overall cost. The process alternatives filtered out previously in this methodology are still explicitly described in this single process model in the information domain as they represent connections across the process boundary to other information, organization, and process networks. These alternatives (such as the Roche 454 process) are represented as information elements that process participants (and owners) must be aware of and if a true alternative becomes available (i.e.

is cost competitive or offers other advantages) then this process would be mapped as a branch point and parallel path along with its information and organizational domain elements.

Yet, even this single process contains enormous hidden complexity as will be shown in the construction of the MDPM model for this single process path. In addition, the baseline process domain graph shown in Figure 81 does not contain many of the real world features of processes, such as rework or queues that can greatly increase the complexity of even this single path. For comparison, the graph incorporating the potential rework loops is shown in Figure 82. Note that rework loops can only occur back to certain "buffer" points where the process may be restarted (i.e. MRSA can be re-grown from culture) but that most activities cannot go back to their immediate predecessor as most of these process activities are "irreversible" or do not have an available QC readout until much later in the process.

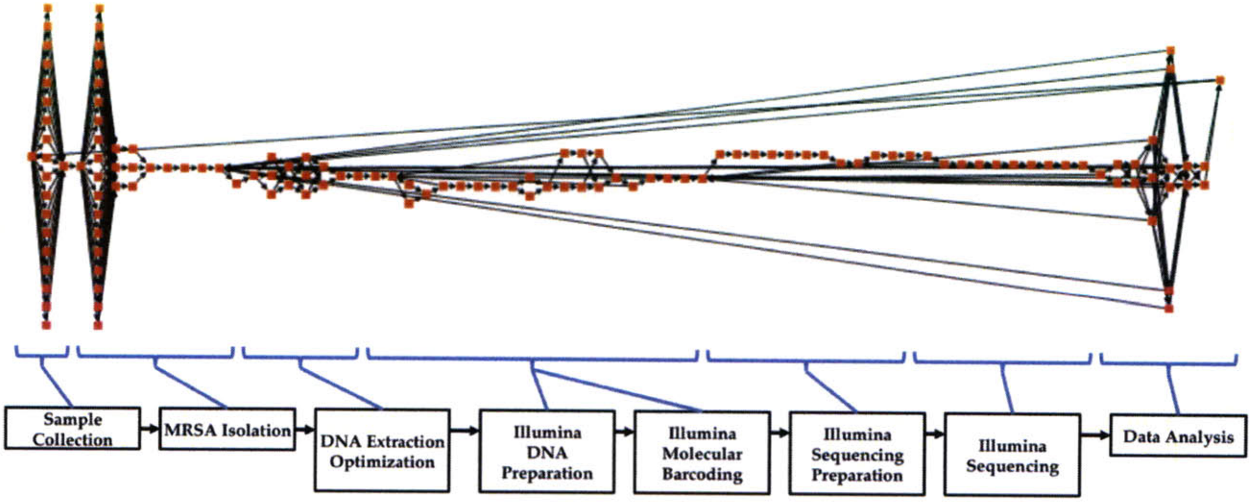


Figure 82 - MDPM Baseline Process Layer with Rework for Surveillance Process

Its is now possible to apply the graph metrics developed in section 4.1 to the process domain graphs developed in Figure 81 and Figure 82 which were built using Microsoft NodeXL and are stored as adjacency matrices. The associated metrics for the baseline and "rework" process domains are shown in Table 18 and the associated adjacency matrices are included in the appendix.

Metric	Baseline	W/Rework	Description
S_V	130	130	Number of distinct activities following enumeration
S_E	246	271	Number of connections between activities
Δ	0.015	0.016	Fraction of a complete graph
diam	69	71	Longest process path
$\bar{l}$	27.811	32.073	Average path length between two activities
B	0.946	0.915	Distance weighted fragmentation
M	117	142	Number of processes (defined paths) in domain

Table 18 - MDPM Process Domain Metrics for Demonstration Process

Of particular interest in the process domain is the diameter of the network, which represents the longest process path and is indicative of the process complexity. Note that the addition of rework loops significantly increases the average path length as a much larger set of potential processes is now available. The cyclomatic number is also particularly useful in describing the process complexity as it describes the number of linearly independent paths through the process layer, and adding rework loops greatly increases these (from 117 to 142). Because the process considered in this thesis has a relatively small number of decision points the cyclomatic number is relatively small, but processes with significantly more rework loops or decision branches would have a correspondingly higher cyclomatic number. As discussed previously the density of the process layer is also relatively low mainly due to the process properties, which require a certain precedence between nodes, directed relationships, and physical constraints for rework (i.e. not every node can be connected to any other).

5.2.2 Organizational Domain

Building on the process element enumeration, all of the organizational domain elements are also enumerated. Although in practice the organization and information elements are enumerated as part of the process element enumeration performed previously, only the full enumeration is shown in Figure 83 to explicitly separate each of the domains in the presentation of the methodology. The organizational map in Figure 83 shows the cumulative set of organization domain elements corresponding to the enumeration of process activities. As part of building the MDPM model, the individuals and organizations associated with each process step are defined with the same process boundaries. Thus, Figure 83 represents all of the individuals associated with the process layer but within the process boundaries. Not every process activity must be associated with an organizational or information domain node, but the lack of a connection between the process domain and information or organization domains means that a process activity is "unsupervised" and a problem in that activity or a change elsewhere in the process network affecting the activity could be a problem. In graph terms this means that process domain nodes not connected to the other two domains are not reachable and thus cannot be "seen" by other nodes on the graph. For example, this could have serious implication for troubleshooting as a problem associated with a node that was not "visible" would be difficult to identify or fix. Conversely, a change (or improvement) in a node that is no "visible" may have serious impacts elsewhere in the process, but that node itself may not be able to realize this impact. An example might be that given previously of a technician working in one of the reagent manufacturers is very far removed from the physician using the diagnostic which uses a particular reagent the technician makes. A small change in the reagent may have a serious consequence far downstream in the process. Note also that in Figure 83 there are a number of individuals

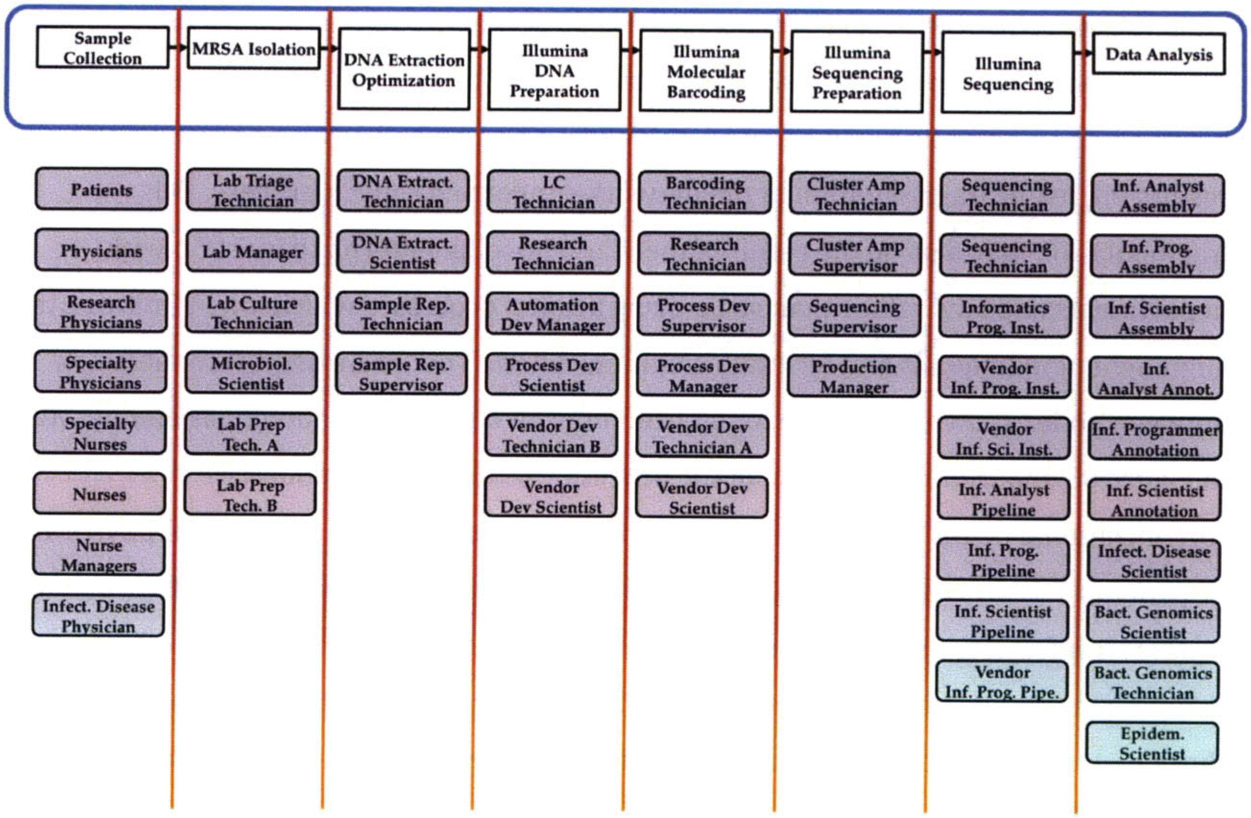


Figure 83 - Organization Domain Elements Enumerated from Macro Process

The organizational layer most closely represents traditional social network analysis as their association with the process drives relationships between individuals, but communication between them can also exist at random through friendships, peer relationships, and other social structures. Complex processes such as the whole genome MRSA surveillance process prototyped in this thesis can span numerous organizations and individuals that would not normally be connected. By chance some connections may exist, facilitating the integration of a process, but the goal of this thesis is to provide analytical tools to understand the degree of communication needed across all the domains, including the organizational one. Figure 84 shows the same set of enumerated individuals and organizations but adds the connections between them for the baseline process (before the MRSA surveillance process was built). It is readily apparent that the initial organizational domain network is highly fractured with a number of disconnected groups, and weakly connected individuals within these groups. The disconnectedness of the organizational domain network is partly a function of the number of disciplines that must be connected to build the surveillance process ranging from medical practice to microbiology to computer science. These "silos" represent significant barriers to the development of complex healthcare processes as the increasing specialization of each discipline or technology results in greater fragmentation of the network supporting a given process. This fragmentation may also be the result of the way a hospital enterprise is set up with departments and labs rather than structuring around processes which might improve the alignment of individuals and facilitate communication. The "pre-existing" connections between nodes shown here were found from interviews with the individuals enumerated in Figure 83 and from observation of the process. Note that some nodes are "organizational nodes" which represent multiple individuals, such as the set of physicians or patients and are marked as such. The goal is to provide a representation of the actual process in a feasible manner as detailed representation of all organizational nodes (i.e. mapping the connections of all the nursing staff) is only warranted if initial analysis shows those nodes to be particularly problematic (i.e. cannot easily be reached, contribute to long path lengths, etc.).

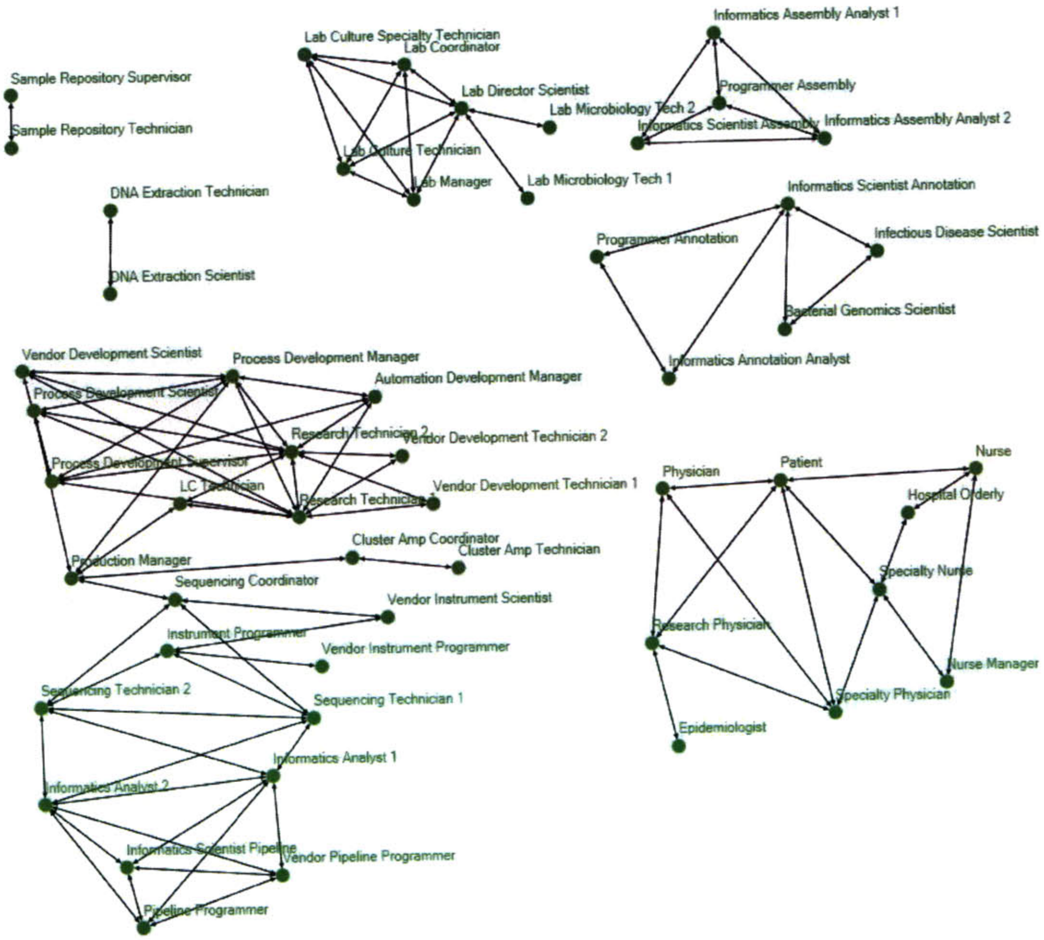

Sample Repository Supervisor
Sample Repository Technician
DNA Extraction Technician
DNA Extraction Scientist
Lab Culture Specialty Technician
Lab Coordinator
Lab Director Scientist
Lab Microbiology Tech 2
Lab Culture Technician
Lab Manager
Lab Microbiology Tech 1
Informatics Assembly Analyst 1
Programmer Assembly
Informatics Scientist Assembly
Informatics Assembly Analyst 2
Informatics Scientist Annotation
Programmer Annotation
Infectious Disease Scientist
Bacterial Genomics Scientist
Informatics Annotation Analyst
Vendor Development Scientist
Process Development Manager
Process Development Scientist
Automation Development Manager
Research Technician 2
Vendor Development Technician 2
Process Development Supervisor
LC Technician
Research Technician 1
Vendor Development Technician 1
Cluster Amp Coordinator
Cluster Amp Technician
Production Manager
Sequencing Coordinator
Vendor Instrument Scientist
Instrument Programmer
Vendor Instrument Programmer
Sequencing Technician 2
Sequencing Technician 1
Informatics Analyst 1
Informatics Analyst 2
Informatics Scientist Pipeline
Vendor Pipeline Programmer
Pipeline Programmer
Physician
Patient
Nurse
Hospital Orderly
Specialty Nurse
Research Physician
Nurse Manager
Specialty Physician
Epidemiologist

Figure 84 - Baseline MDPM Organization Domain for Demonstration Process

One of the most important steps in building the demonstration MRSA surveillance process was to connect all of the various groups and individuals such that the organizational domain could be a single connected network. Much like the example in Chapter 3, the author of this thesis served as "process owner" to facilitate communication across various "silos" and function as a hub for the design, initial tests, troubleshooting, and integrated analysis of the process. In addition other connections were necessary to facilitate communication across critical "connecting technologies" such as the communication between the DNA extraction scientist and the clinical microbiology laboratory. Additional connections were also needed for the molecular barcoding connecting technology as this technology impacted not just the molecular biology sample prep but also the necessary downstream data analysis software that must disambiguate (pull out) the individual barcodes in the data. All of the necessary additional connections are shown in red in Figure 85 below superimposed on the "baseline" fragmented organizational layer.

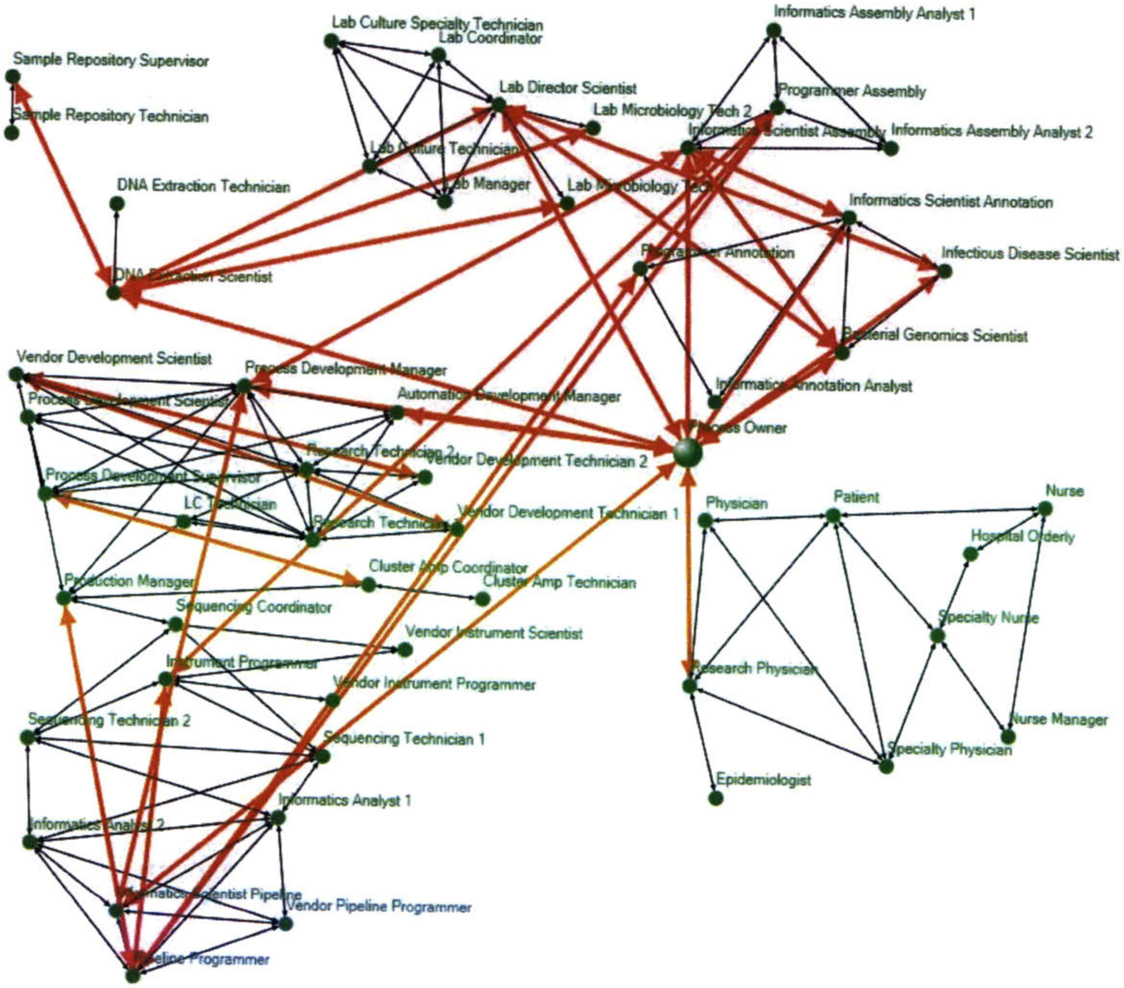

Lab Culture Specialty Technician
Lab Coordinator
Informatics Assembly Analyst 1
Sample Repository Supervisor
Sample Repository Technician
Lab Director Scientist
Programmer Assembly
Lab Microbiology Tech 2
Informatics Scientist Assembly
Informatics Assembly Analyst 2
Lab Culture Technician
Lab Manager
Lab Microbiology Tech
DNA Extraction Technician
Informatics Scientist Annotation
Programmer Annotation
Infectious Disease Scientist
DNA Extraction Scientist
Bacterial Genomics Scientist
Vendor Development Scientist
Process Development Manager
Informatics Annotation Analyst
Process Development Scientist
Automation Development Manager
Process Owner
Research Technician 2
Vendor Development Technician 2
Process Development Supervisor
LC Technician
Research Technician
Vendor Development Technician 1
Physician
Patient
Nurse
Hospital Orderly
Cluster Amp Coordinator
Cluster Amp Technician
Production Manager
Sequencing Coordinator
Specialty Nurse
Vendor Instrument Scientist
Instrument Programmer
Research Physician
Vendor Instrument Programmer
Sequencing Technician 2
Nurse Manager
Sequencing Technician 1
Specialty Physician
Epidemiologist
Informatics Analyst 1
Informatics Analyst 2
Informatics Scientist Pipeline
Vendor Pipeline Programmer
Pipeline Programmer

Figure 85 - V2 MDPM Organization Domain for Surveillance Process

The associated metrics for the organizational domain "baseline" and the first version prototype process "V1" are shown in Table 19. Most significantly the V1 organizational domain is a single connected component, which allows for the calculation of certain network metrics such as average path length and directed efficiency. The density of the organizational domain is also relatively low compared to the maximal density (1) of a complete graph where every node is connected to every other one. However, this reflects the real-world sparseness of organizations especially “multi-silo” processes such as the one assembled in this thesis.

Metric	Baseline	V1	Description
S_V	53	54	Number of organizational elements
S_E	184	242	Number of connections between elements
Δ	0.0668	0.0846	Fraction of total possible connections
$\bar{d}$	3.472	4.481	Average degree for a node
diam	7	8	Longest path between elements
$\bar{l}$	2.980	3.556	Average path length between two nodes (reachable)
B	0.873	0.642	Distance weighted fragmentation
F	0.745	0.000	Fraction of nodes that cannot reach each other
Cn	7	1	Number of weakly connected components
M	132	189	Number of communications paths in domain
ε	-	0.35	Directed efficiency

Table 19 - Organization Domain Metrics for Baseline and V1 Processes

Of particular use in the analysis of organizational domain is the efficiency metric, which gives a sense for how easy it is to communicate across the network in relatively few steps. This metric which is the sum of the inverses of the individual node shortest paths (geodesic distances) normalized by the maximum possible density better represents the “fall off” in efficiency with longer paths between individuals in the organizational layer. So for example a network with very long paths between nodes will have a very low efficiency compared to one with relatively shorter path lengths. A fully connected network would have an efficiency of 1.

Improved efficiency comes at some cost in terms of bandwidth as the average degree of each node must necessarily increase. In computer communications networks where bandwidth may not be limiting it is feasible to increase the number of connections for each node to improve efficiency, however in the "human" organizational domain this is likely not feasible as complex processes may have hundreds of individuals and increasing the number of connections for each node may not be possible. However, hierarchical structures such as those discussed in Chapter 4 offer some solution as many "connections" can be accepted with a relatively small increase in the number of additional nodes. Empirically this is a common solution in complex enterprises such as healthcare and is observed here as well with hierarchies of technicians, coordinators, supervisors, and managers. The metrics for the V1 process show that the efficiency of the organizational domain is not high (0.35) and the average path length would seem to be significantly long (3.556) making rapid communication and troubleshooting difficult.

5.2.3 Information Domain

The detailed enumeration of all information domain elements is given in the appendix as part of the adjacency matrix, but follows the enumeration of elements performed so far. The number of elements for the baseline model is 251 elements and 292 for the "V1" model. The information domain elements contain all of the protocols (work instructions) needed to carry out the activities described in the process domain. However, the information domain also contains the "genealogy" of these work instructions which are in their great majority derived from guidelines given by regulatory agencies or professional bodies such as the AMA, design documents which describe the choices made in selecting the particular parameters the protocols use, or most commonly are special cases of "general" instructions or protocols. In addition, the actual protocols routinely used in healthcare or research laboratories generally incorporate commercial kits or reagents with their own particular set of instructions (vendor instructions). Tracing all of these "genealogies" within the process model boundaries is key to the MDPM model analysis since the protocols describing an activity are often highly simplified work instructions containing only the necessary information to carry out the activity. However, in cases where the impact of a change elsewhere in the process might require modification to a particular protocol, it is necessary to "trace back" the protocol back to design documents or other reference documents to evaluate whether the change will require a change to the protocol. The design and reference documents are themselves not usually connected to the individuals performing the activity using the protocol derived from them, but may require specialized knowledge that only the activity designer may have.

An example might be the operation of a particular robotic liquid handling system in a diagnostic laboratory. While the technician will likely be very familiar with the detailed protocol to run the instrument, a change to the sample input material requirements might require reviewing the design and optimization documents (the design space) to see if the new input material changes are within the feasible design space. These design documents may only be familiar to design engineers and scientists and not "accessible" to the technician in a network sense. A connection would thus exist between the technician and the protocol but not necessarily between the technician and the design documents. The protocol would be shown in the information domain as having been derived from the design documents by a directed arrow. The design documents would then have separate connections to the scientists and engineers in the organizational domain. An important point emphasize is that there may or may not be connections between the design scientists and engineers and the technician, and if so change and troubleshooting will be much harder since the path lengths from the design documents and the practical (tacit) knowledge of the technician will be longer. This is an important feature of many process improvement systems such as the Toyota production system, which fosters direct communication between "frontline" workers and R&D knowledge and resources such that problems (troubleshooting) can be resolved rapidly and with a minimum of steps (in a network sense). The graphic in Figure 86 shows the baseline information domain with blue triangles representing information domain elements. As with organizational domain, the information domain also shows the high degree of fragmentation of information (knowledge) associated with the surveillance process. This again is a function of the many disciplines, technologies, and organizations spanned by the surveillance process and is indicative of the "silos" commonly encountered in assembling complex processes from highly specialized elements. Figure 86 should be viewed from the perspective of a "process owner" or an individual responsible for ensuring that all of the variables, parameters, and design spaces in each of these separate information networks can be joined together in a working fashion into a single process (and network).

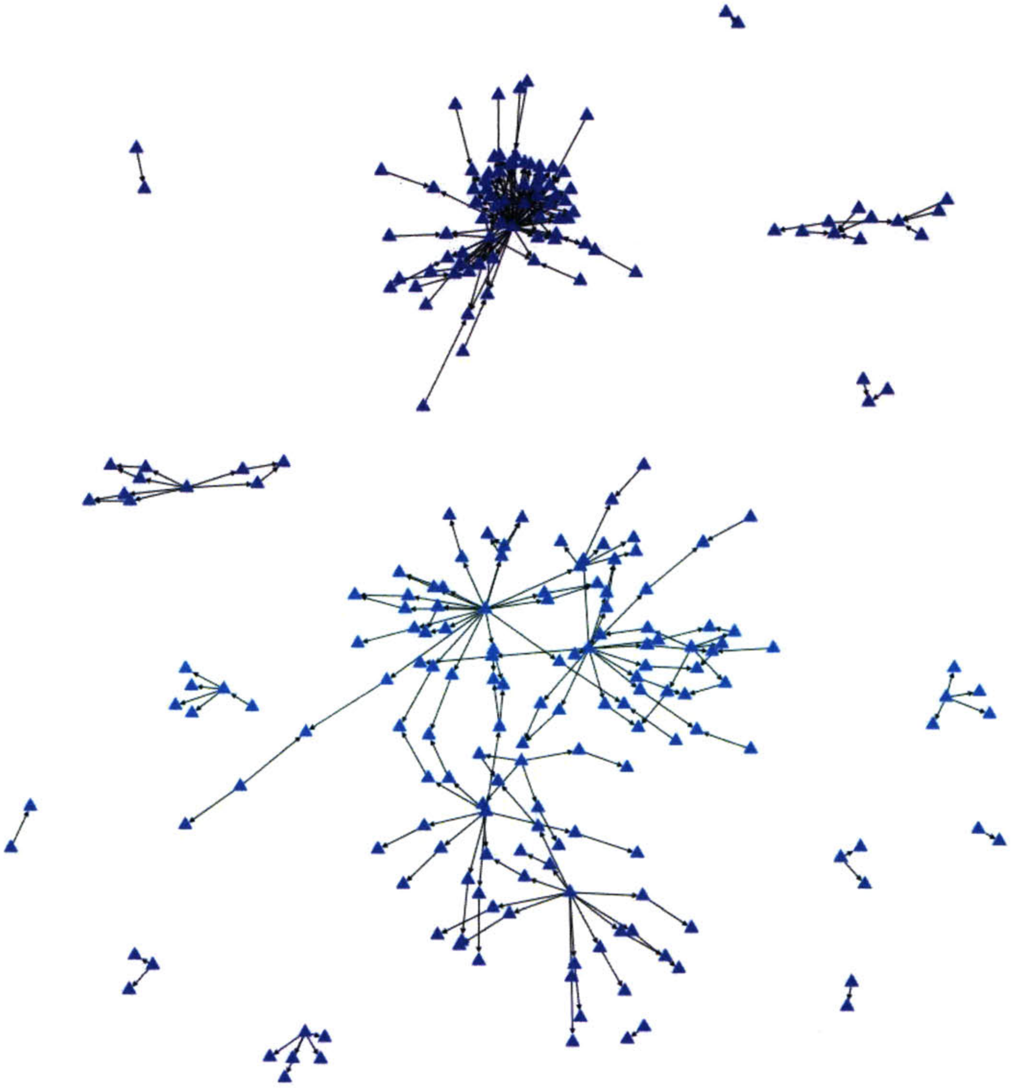


Figure 86 - MDPM Information Domain for Baseline Surveillance Process

In order to verify that the parameters and design spaces of each protocol are either adequate or need to be changed, the first thing a process designer might do is to perform a detailed review of every protocol according to the process specifications. Because of the many protocols that need to be reviewed and the multiple disciplines in which they fall, the process designer might choose to do "hierarchical" reviews in each discipline followed by an overall review integrating these into a single design document specifying what changes are needed in the existing protocols or which new protocols are needed. This is in fact what was done in this thesis to build a surveillance specific process. Each of the existing protocols was reviewed in their own area, such as sample prep protocols which were reviewed in a "molecular biology review" which verified that existing protocols would suffice or specified the new parameters that would have to be used in order to process MRSA genomes. Each of these individual reviews were then tied together into a master design document specifying the "V1" process which detailed whether existing protocols could be kept "as is", needed changes, or entirely new protocols were needed. Where protocols were kept "as is" a bidirectional connection between the review documents in the protocol was made, which in a network sense means that no additional "steps" are needed to reach these unrevised protocols. Where changes were needed, new protocols specific to the MRSA process were added as nodes and connected from the review documents. Similarly where entirely new protocols were needed, the review documents pointed to design and experimentation records (such as those performed for the molecular barcoding), which in turn produced the actual new protocols used. All of these items are shown in the graphic below, where the red lines indicate connections to the "reviews" and the maroon diamonds represent reviews and new protocols generated as part of the design process.

Note that showing the intermediate "reviews" suggests that the information domain network shown is the time evolution from the initial baseline network to the final network and thus might not represent the actual "final" process. However, it is important to recognize that the objective of the MDPM is to capture the real complexity of processes, especially when they fail and require troubleshooting or changed and require redesign. In both the troubleshooting and redesign cases the entire process network must be "searched" for either the problem in the troubleshooting case or affected nodes in the redesigned case. The search would extend to the original sources of the protocols in use and require that all of the design decisions, design spaces, and original "generic" work instructions be available on the network this can be accomplished in a manner similar to the "versioning" of MDPM models described in Figure 64. As processes mature it is likely that all of this extra "scaffolding" would no longer be needed, however in rapidly changing environments explicitly showing the scaffolding represents the true complexity and allows for search. It is also important to note that it is not only when processes are rapidly changing the scaffolding is needed, but likely also when a process must be implemented into many different locations, organizations, and infrastructure. No two hospital systems are alike and modifications will be needed to implement this process into each of those unique environments, which will in turn require further "search" of the network.

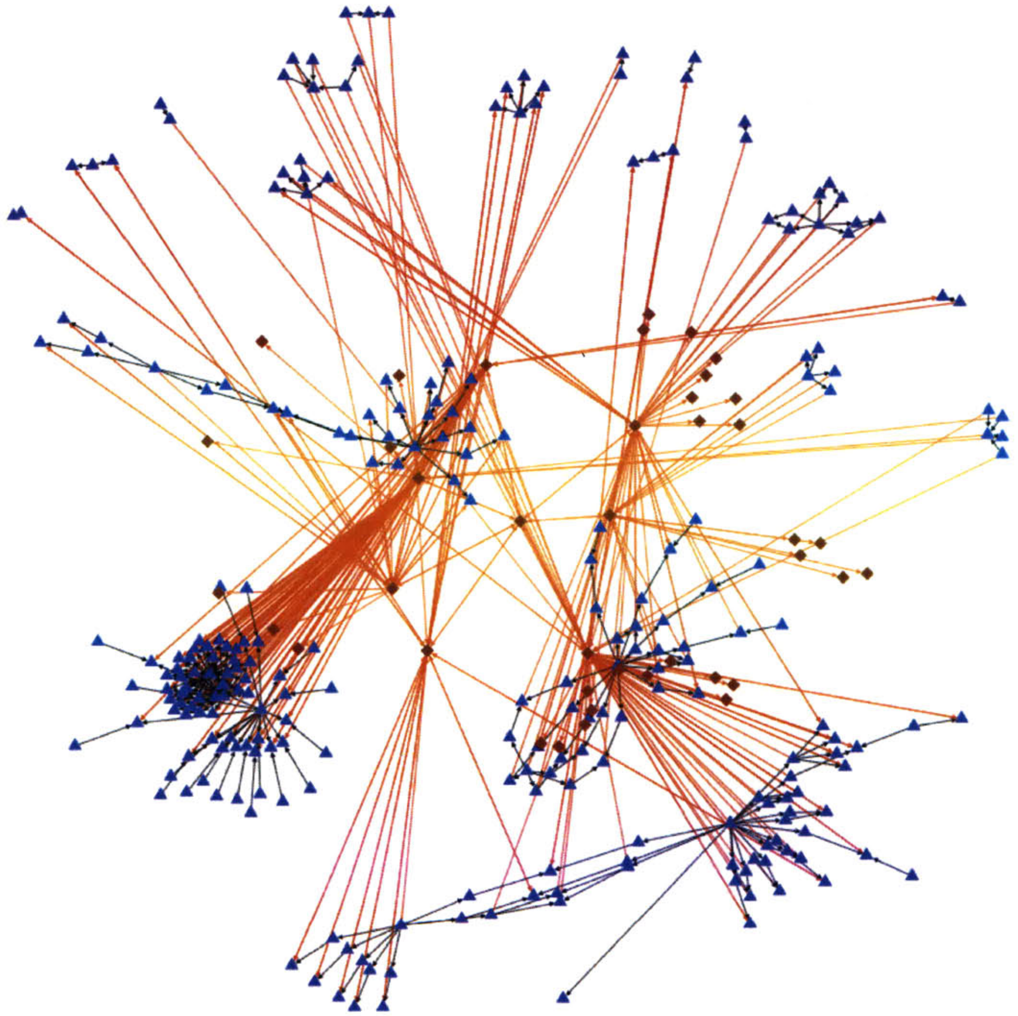


Figure 87 - MDPM Information Domain for V1 Process

Using this "V1" network is possible to calculate the information domain metrics in comparison to the baseline network as shown in Table 20. Of particular interest in the information domain is the degree of fragmentation before and after the process design, where the initial baseline fragmentation degree represents the fraction of the network in "silos" as does the number of components. Note that the directed efficiency of the information domain (0.12) is significantly lower than the efficiency of the organizational domain as short paths do not connect most elements. The low density of the information domain network also reflects this.

Metric	Baseline	V1	Description
S_V	251	292	Number of information elements
S_E	296	598	Number of connections between elements
Δ	0.0047	0.0070	Fraction of total possible connections
$\bar{d}$	1.681	2.048	Average degree for a node
diam	2	5	Longest path between information elements
$\bar{l}$	1.182	3.865	Average path length between two nodes (reachable)
B	0.995	0.879	Distance weighted fragmentation
F	0.836	0.000	Fraction of nodes that cannot reach each other
Cn	19	1	Number of weakly connected components
M	-	307	Number of paths in domain
ε	-	0.12	Directed efficiency

Table 20 - MDPM Information Layer Metrics for Demonstration Process

5.2.4 MDPM Analysis

Category	Methodology Step
Process Feasibility and Element Enumeration	1.) Description of process, stakeholders, and goals 2.) Identify macro process activities and define boundaries 3.) Identify potential process alternatives and design space 4.) Identification of design specifications and constraints 5.) Selection of macro level feasible processes against design specification requirements 6.) Enumeration of process elements onto detailed process map for selected process
Network Analysis	7.) MDPM creation from detailed process map **8.) Analysis of MDPM model**

Having populated the process, organization, and information domains it is possible to analyze the full MDPM matrix, which incorporates connections across domains as described in the MDPM model development in Chapter 3. The full MDPM matrix (also called the projection domain in this thesis) incorporates connections between elements (individuals) in the organizational domain and information domain such as the familiarity of a protocol by a technician or the familiarity of design documents by a research scientist. Connections between all domains are shown in the projection domain and a corresponding increase in the complexity of the network results, more closely approximating the "real" process complexity. The initial "baseline" network includes only baseline nodes and connections as developed in the previous sections. While the baseline organization and information domains by themselves showed many disconnected elements, in the projection domain they are connected through the process and across domains. This full baseline process network is shown in Figure 88 representing the much greater complexity of interaction across domains of process, organization, and information. This baseline network contains 437 elements (nodes) and 3,338 connections (edges) with a diameter of 17 and an average path length of 6.3. The efficiency of this network is rather low, 0.227, in comparison to the organization domain and the eventual "V1" network.

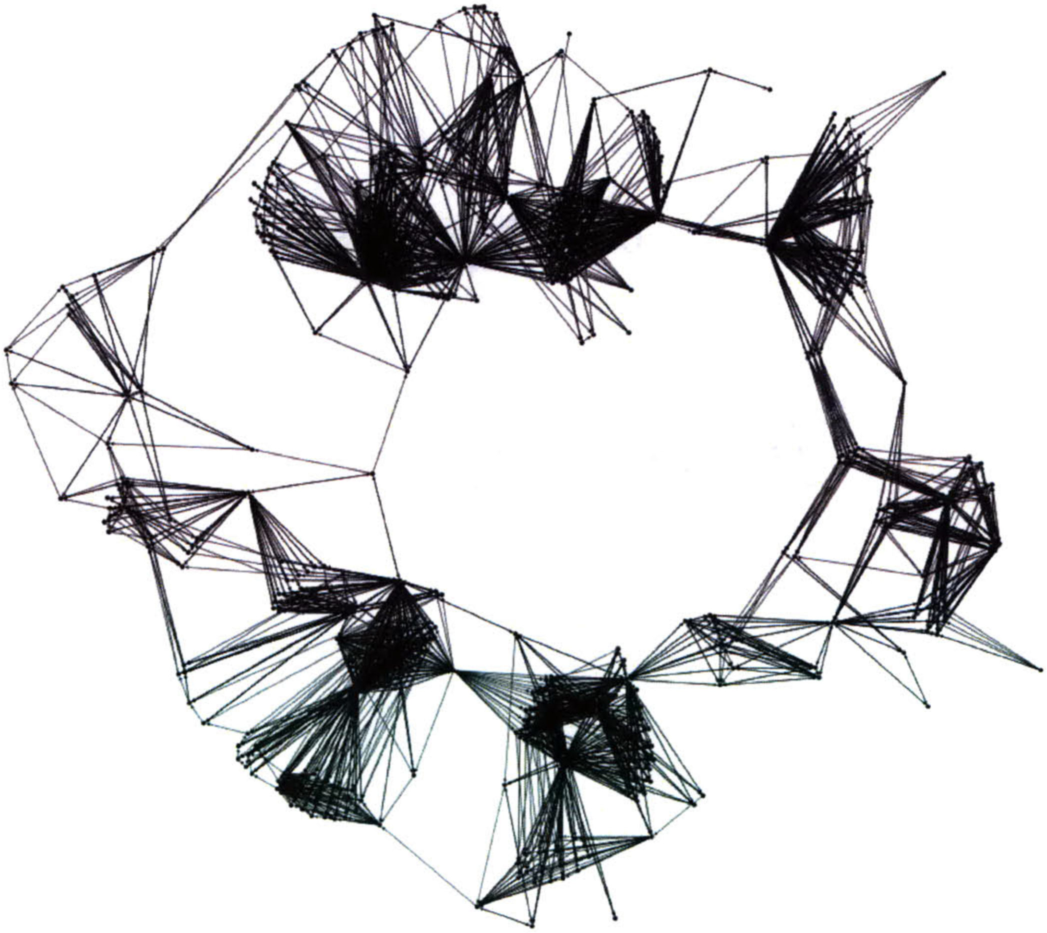


Figure 88 - Baseline MDPM Projection Domain for Surveillance Process

To highlight the additional nodes and edges needed to form the initial "V1" network, Figure 89 shows the full V1 network with the additional nodes and edges from the baseline network highlighted in red. At the center of Figure 89 is the process owner, but other nodes such as the process reviews are clearly visible some in hierarchical arrangements. The V1 network contains 480 elements and 4,289 connections with a diameter of 7 and an average path length of 3.6. The efficiency of the V1 network is much higher at 0.307, but additional design improvements could be made to further increase this value.

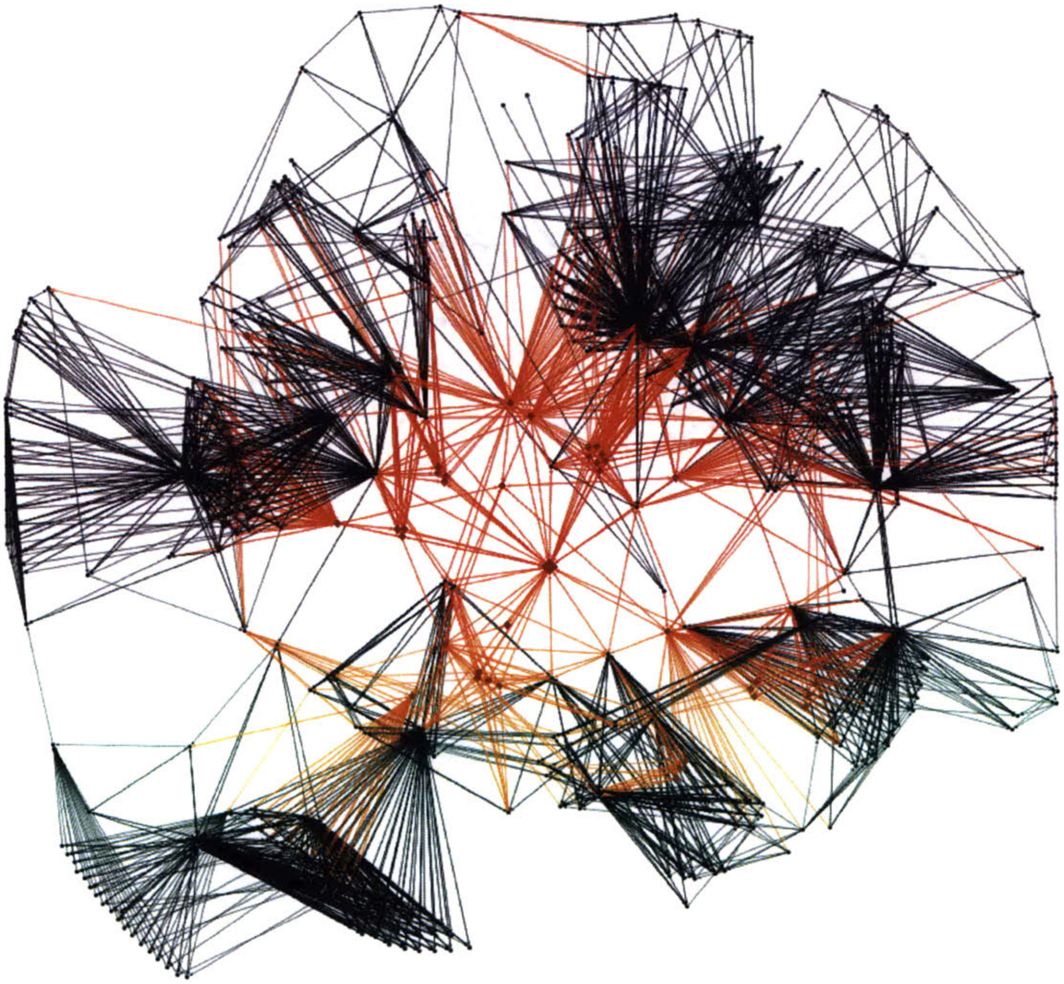


Figure 89 - MDPM Projection Domain Showing Baseline and V1 Processes

The individual domain elements can be highlighted on the full process network as shown in Figure 90 where red squares represent process elements, blue diamonds represent information elements, and green circles represent organization elements following the convention in Chapter 3. Certain clustering patterns emerge when the full V1 process network is shown highlighting each layer individually such as the concentration of information elements around certain individuals and the responsibility for multiple process steps by key individuals as well.

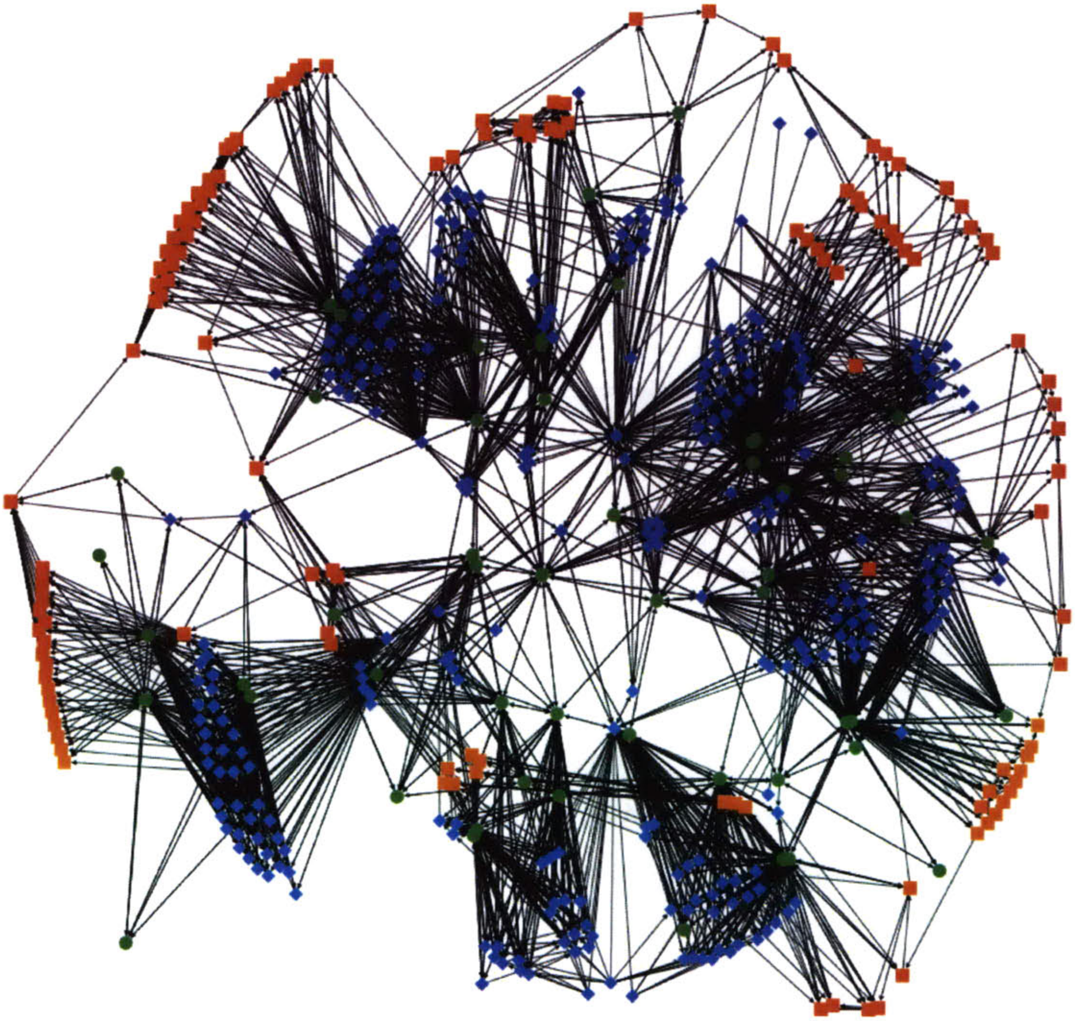


Figure 90 - MDPM with Domains Highlighted for Surveillance Process

Comparing the metrics between the baseline and the "V1" model, the additional network connections and notes needed to make the process work empirically (real world work to produce the data described in Chapter 6) do result in measurable improvements to the process network. Most notable is the increase in network efficiency and the reduction in average path length as shown in Table 21. The additional nodes and connections in the "V1" network were not added with the explicit goal of improving the network metrics, but rather simply were connections and elements made necessary in the design and development of the process to simply make it work. These connections were encountered as part of the initial design, as problems surfaced, or as part of an initial optimization. Because the objective of this thesis was to develop a network process analysis methodology by testing and validating against a real-world process, the MDPM model was developed simultaneously with the "V1" surveillance process. However, having developed the MDPM model it is possible to imagine using it to develop a process from "scratch". In this case the MDPM model would be used as part of the initial design phase to find a process that satisfied certain network metrics such as a minimum efficiency, average path length, and cyclomatic number.

Metric	Baseline	V1	Description
S_V	437	480	Number of elements
S_E	3,338	4,289	Number of connections
Δ	0.0175	0.0187	Fraction of total possible connections (cross-domain)
$\bar{d}$	8.011	8.944	Average degree for a vertex
dmax	92	100	Maximum Vertex Degree
diam	17	7	Longest path (cross-domain)
$\bar{l}$	6.343	3.636	Average path length between two nodes (reachable)
B	0.773	0.693	Distance weighted fragmentation
F	0.000	0.000	Fraction of nodes that cannot reach each other
Cn	1	1	Number of weakly connected components
M	2,902	3,810	Number of processes (defined paths) in domain
ε	0.227	0.307	Directed efficiency
	0.0030	0.0023	Network vulnerability
ξ	14.629	11.280	Average search information needed
I	3.9447	4.4065	Graph structure entropy

Table 21 - Projection Domain Metrics for MDPM Demonstration

The heavy reliance of the V1 model on the "process owner", also the thesis author, as a hub to effectively reduce the average path length is not necessarily a sustainable model for the development of other processes in healthcare. In this case, the requirements for a process owner familiar with molecular biology, microbiology, computer science, process design, genomics, and social network analysis allowed for the relatively rapid implementation and integration of the disparate elements needed to make the surveillance process work. However, this is not necessarily a common resource and presents significant barriers to process improvement and change in healthcare unless specific enterprise structures are designed to overcome the fragmentation of organizations, knowledge, and resources. A clear view of this is Figure 91, which plots the betweenness centrality of each of the nodes in the V1 process network calculated using UCINET (Borgatti, Everett et al. 2002) a social network analysis program. The nodes with the highest betweenness centrality are highlighted in red indicating they are on the shortest paths of the majority of the node pairs in the network. Note that there are nodes in both the organization and information layers with high betweenness centrality, such as the molecular biology process review (information layer element) and the process owner (organization layer element, and also the element highest value of betweenness centrality). A corollary of having these high betweenness centrality nodes is that they are also likely the best "transfer" nodes such that the surveillance process could be transferred to another organization and location as they can best connect to the rest of the network to "get" any necessary information. The MDPM model could in a transfer scenario be used to map the target organization for the transfer, identify the key nodes in both (organization and information) such that transfer of the process had the highest likelihood of success as measured by network metrics such as average path length and efficiency over the combined "source" and "target" networks. Because there are both organization and information nodes with the highest betweenness centrality values, the transfer of a process to another organization would require access to both the individuals perhaps in some formal training process, but also require the "target" organization to systematically review (and understand) the design documents such as the molecular biology process review and the sequencing sample prep process reviews such that any changes specific to the target organization's structure could be evaluated by them.

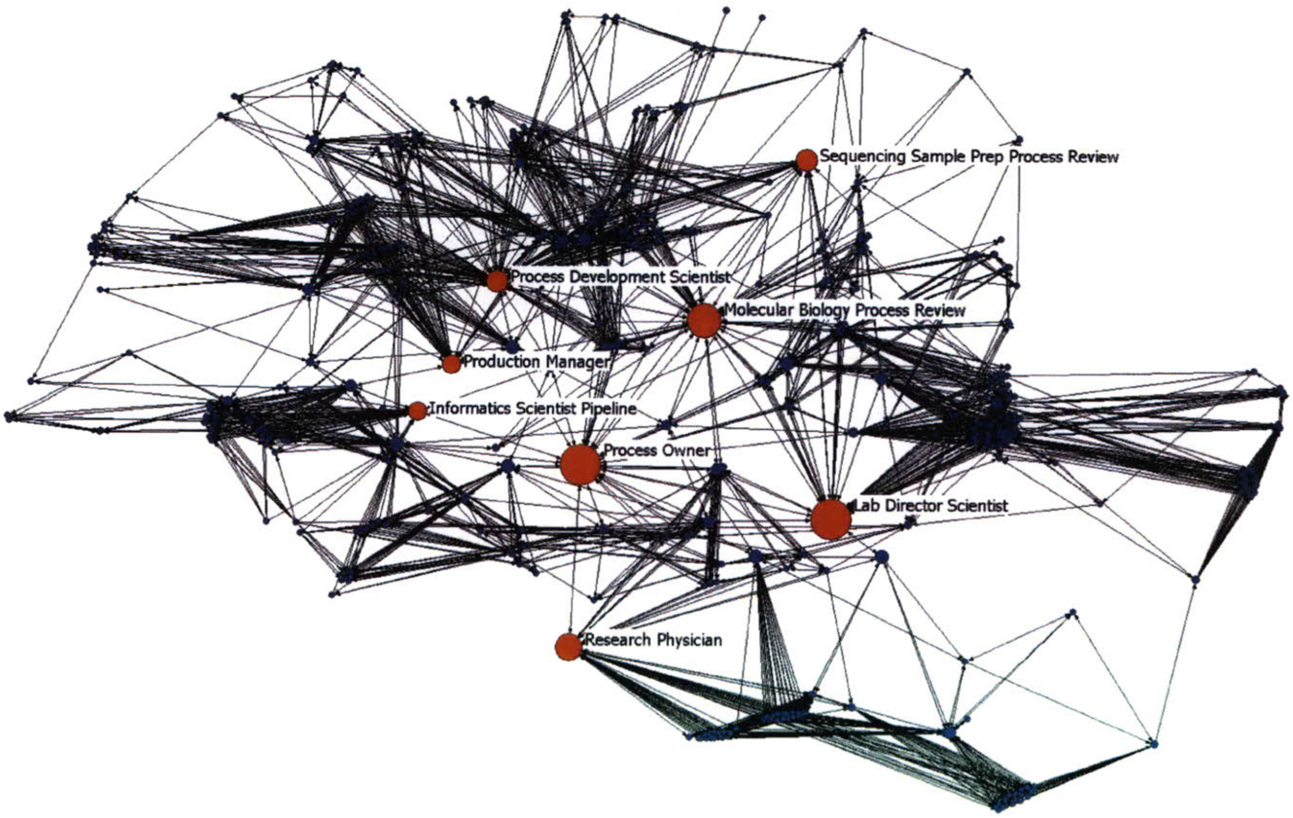

Sequencing Sample Prep Process Review
Process Development Scientist
Molecular Biology Process Review
Production Manager
Informatics Scientist Pipeline
Process Owner
Lab Director Scientist
Research Physician

Figure 91 - Betweenness Centrality for the MDPM V1 Surveillance Process Model

An alternative view of the network is degree centrality, where network elements (nodes) with the highest degree are highlighted, the premise being that in social structures high degree nodes are well-connected and have high "brokerage" abilities. The figure below shows the V1 network with node size as a function of degree centrality, and with the highest degree centrality nodes highlighted in red. In the context of the projection domain of the MDPM model high degree nodes may represent "network" nodes as described previously in Chapter 3 such as a "nurse staff" representing the nursing organization at a hospital rather than individually modeling each nurse. However, when individual nodes that are not "network" nodes have high degree this may represent a bandwidth overloading of that node where the hierarchical structure may better deal with an excessive number of connections. For example, in the V1 model the "process development scientist" node has the highest degree centrality partly because of the many dimensions of the sequencing process that are changing simultaneously and the need to integrate them. In some sense this process development scientist serves a similar role to the process owner but within the sequencing portion of the surveillance process. It is worth noting again that the surveillance process is built on existing elements (and processes), which are in continuous flux, the sequencing process being one of them. The management of this change requires a significant amount of scaffolding, which would likely be removed as processes were stabilized and a "V2" process was built that further simplified the complexity found in the V1 process. A V2 process might therefore remove the process development scientist in exchange for multiple nodes (additional staff) that could better deal with the many process connections needed. These multiple nodes might be arranged in a hierarchical fashion with the existing process scientist in the middle such that the individual "bandwidth" of each node is not exceeded.

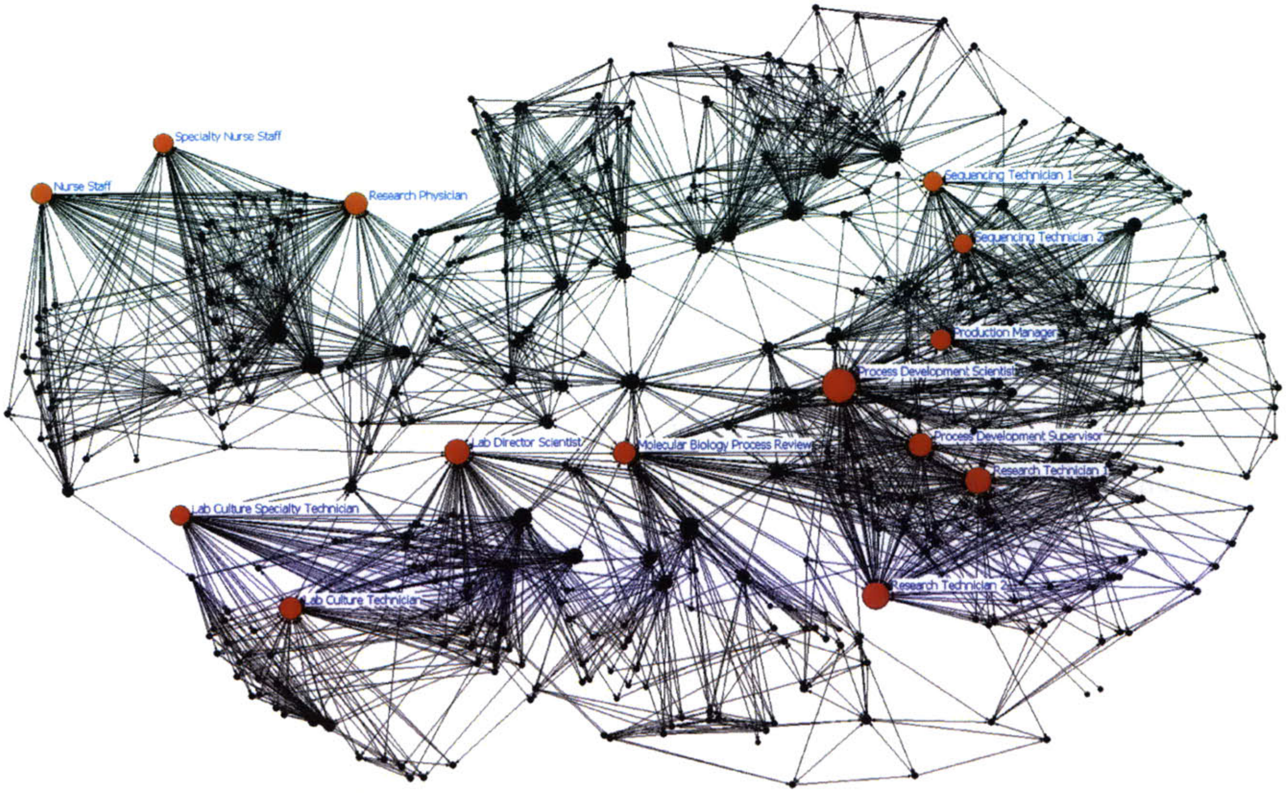


Figure 92 - Degree Centrality for the V1 MDPM Surveillance Process Model

Once a process has been described using the MDPM methodology, it is possible to apply a wealth of network analysis algorithms to better understand the patterns and potential weaknesses (and strengths) of the process. A particularly useful metric is the "search information" required to traverse a network, as proposed by (Rosvall, Gronlund et al.

2005). The basic idea is that in order to find another node within a network there are a series of "yes/no" questions that must be asked at each node in order to pick the shortest path route. This can be thought of as the product of the probabilities at each node along the shortest path of picking the right shortest path edge. The negative base 2 log of the product of these probabilities along the shortest path corresponds to the amount of information in bits needed to find the shortest path. This search information can be calculated from every node to every other node in the MDPM model such that the amount of information needed for every node pair shortest path can be found. Nodes with a very high search information value should correspond to nodes that are particularly difficult to manage in troubleshooting or process change cases. As a reminder of how the search information is calculated, Equation 20 - Average Search Information Metric is presented again.

$$\xi = \frac{1}{|V|^2} \sum_{ij} - \log_2 \sum_{\{p(i,b)\}} \frac{1}{k_i} \prod_{j \in p(i,b)} \frac{1}{k_j - 1}$$

where k is the vertex degree, p(i,b) is the set of all shortest paths from i to b, and V is the set of vertices in the graph

Equation 20 - Average Search Information Metric

A high search information value is interpreted to mean that the impact of a change at that node on the rest of the network is difficult to assess and similarly the effect of a problem at that node on the rest of the network will be difficult to identify. The search information for every node in the baseline and V1 networks was calculated (methods shown in the appendix) and the resulting values were plotted on the network nodes and edges as a color scale with the lowest values in black and the highest values in red as shown in Figure 93. The baseline network shows many nodes in each of the layers with high search information values especially around some of the more fragmented (siloed) parts of the process such as the upstream sample collection by a variety of nursing staff and parts of the DNA extraction process. Overall the baseline network has a significantly higher average search information requirement of 6,392 "questions" compared to 5,414 for the V1 network. These values mean that on average each shortest path would require this many "questions" (i.e. asking at each node: is this edge the correct one on the shortest path to node Y?). While in practice, the troubleshooter of a process would likely have a sense for which nodes are connected to which, making the search an informed one rather than a blind one, the blind search information values are valuable to identify which nodes should be better connected such that they can be readily searched for impact or problems.

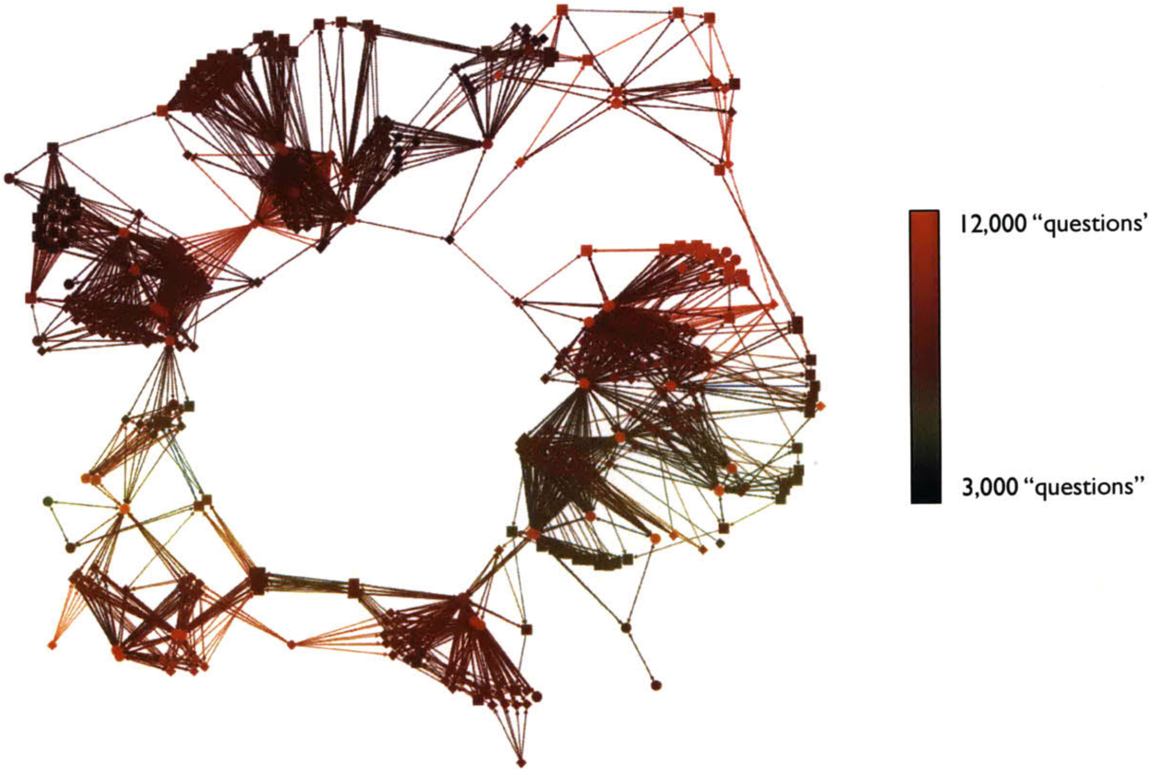


Figure 93 - Baseline MDPM Search Information

In comparison the V1 network has a much lower average search information value as can be seen in the network plot in Figure 94 on the same scale as the baseline network plot. Of note are certain process elements not immediately considered in the initial design, such as the triage of samples coming into the clinical microbiology laboratory from collection throughout the hospital. This activity has a significantly higher search information value and upon reflection could correspond to a significant failure mode for the surveillance process since improper triage could lead to missing or incorrectly processed samples, the results of which would not be apparent until much later in the process. Similarly any simple triage workflow changes or protocol changes could have a very significant downstream impact given the number possible paths downstream of that node. Note that the term downstream only refers to the shortest path from that node to another rather than an actual process direction. So for example the process design scientist also has a high search information value suggesting that certain parts of the process, such as the nursing and physician associated process steps are not readily "visible" from that node and yet changes in sample collection practices or other "difficult to see" nodes could readily impact process activities associated with the process scientist. Again this is a design tool that could be used to design a V2 process that further reduced the search information necessary such that troubleshooting and process change could be greatly facilitated.

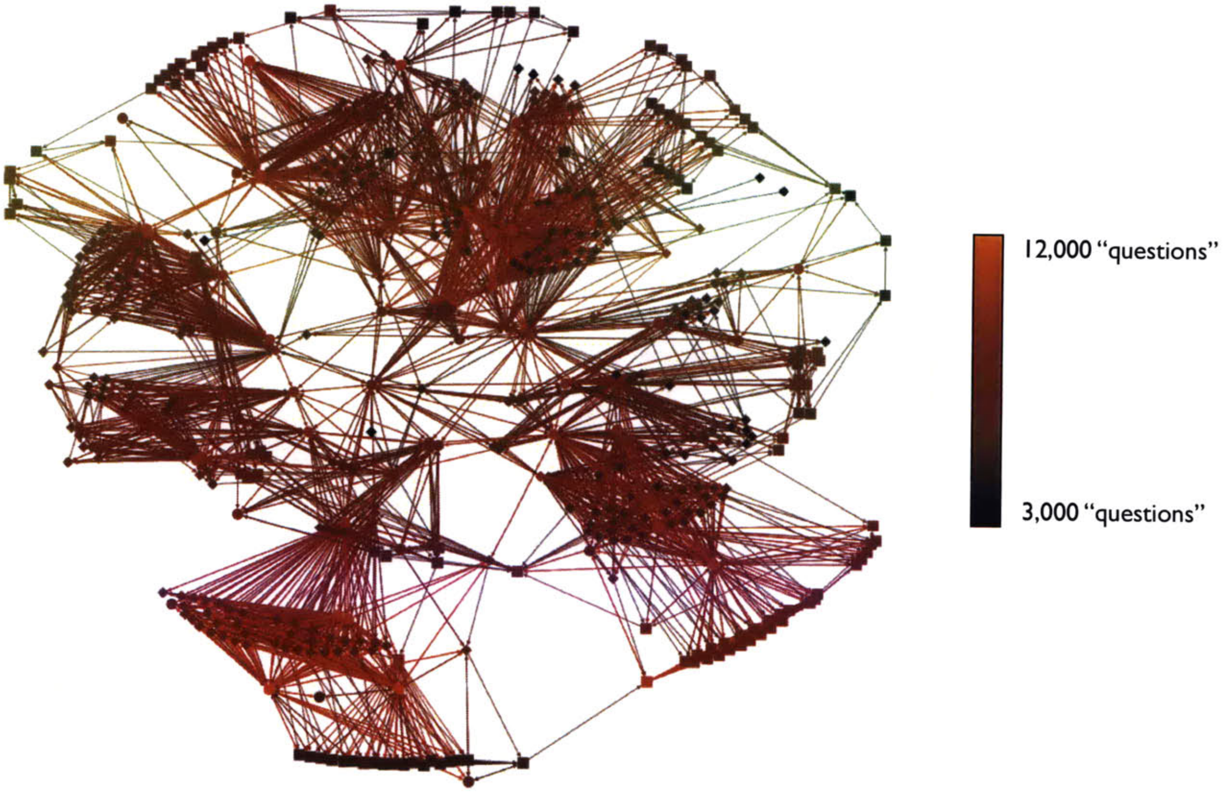


Figure 94 - V2 MDPM Search Information

The properties that make networks robust are also likely to contribute to lower the search information needed, although at the cost of additional "scaffolding" in terms of additional nodes and edges. For example having multiple alternative shortest paths (i.e. same shortest path length but different routing) can decrease the search information needed as there are now many more chances to find a shortest path and the sum of the products of these probabilities along each path will greatly lower the average search information for the network.

5.3 Comparison to Culture Based Diagnostics

For comparison, the process and organizational domains for culture based diagnostics are shown in Figure 95. While this process has the same up front complexity in sample collection and preparation, the downstream processes are much simpler and require significantly less scaffolding than the DNA sequencing based process. There are a number of reasons for this lower complexity, principal among them is that culture-based diagnostics span fewer fields than DNA-based sequencing being almost entirely contained within microbiology. During their development these culture-based diagnostics did of course require the involvement of other disciplines, especially in the semi automated high throughput versions found in clinical microbiology laboratories. However, most of these processes are mature and there is limited need for the direct involvement of these other disciplines (such as automation engineering) in the day-to-day operation of a clinical microbiology lab. These processes are still connected to automation resources but through much longer network distances such as the connections the service technician assigned to a lab might have through his or her supervisors and across to the R&D group within the particular vendors organization. Through continued improvement and targeted complexity reduction however it is possible that the DNA sequencing based process could reach the much lower complexity of this process as is described in Chapter 7.

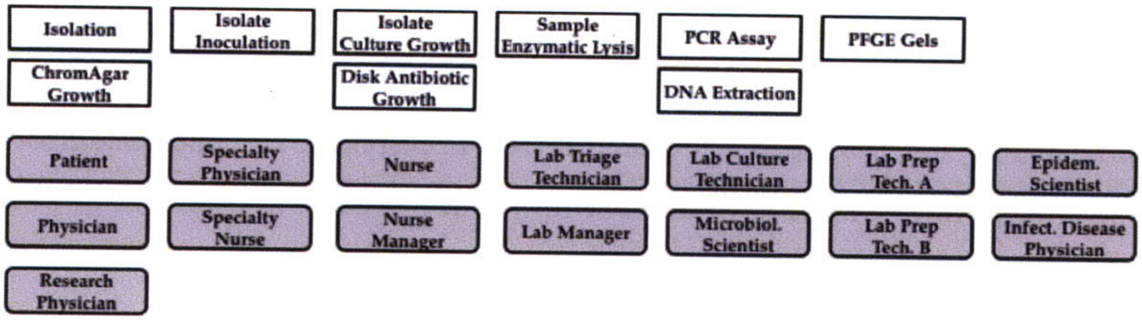


Figure 95 - Culture Based Diagnostics Process and Organizational Elements

5.4 Connecting Technologies

One of the most critical elements in building the V1 process was the identification and development of the necessary connecting technologies such as DNA extraction, modular bar coding, and data analysis to integrate the various existing process elements. In the feasibility selection phase of the methodology these are also called "placeholder" activities, as they describe process elements not yet available that are nevertheless necessary to connect the entire process together. These can also be though of as technologies that will be inserted into process to modify them to another purpose, and has similarities to the work by (Smaling and de Weck 2007) relating the impact of technology insertion and product design using DSM analysis. The key difference is that diagnostic process steps cannot easily be rearranged and in some cases cannot physically be changed in terms of their logical precedence. Similarly the organizations in which healthcare processes occur have certain structures related to the performance of other processes (i.e. other diagnostics or patient care procedures) that dictate certain organizational structures, which are unlikely to be changed for the benefit of a single process. An expansion of the MDPM model to the enterprise-level might allow for the global optimization necessary but in many cases the critical connecting technology must be inserted within a "fixed" process and organizational structure. It is especially important therefore to map the impact of a connecting technology (or technology insertion) into the overall network. Figure 96 maps each of the key surveillance process connecting technologies into the process network with the corresponding edges colored in red showing the extent to which these technologies affect the rest of the network. The connecting technologies considered are DNA extraction shown as green circles, modular barcoding shown as yellow squares, genomic data analysis shown as blue diamonds, and process design shown as aqua triangles. Note that for this figure the standard process, information, and organization symbols and colors are different.

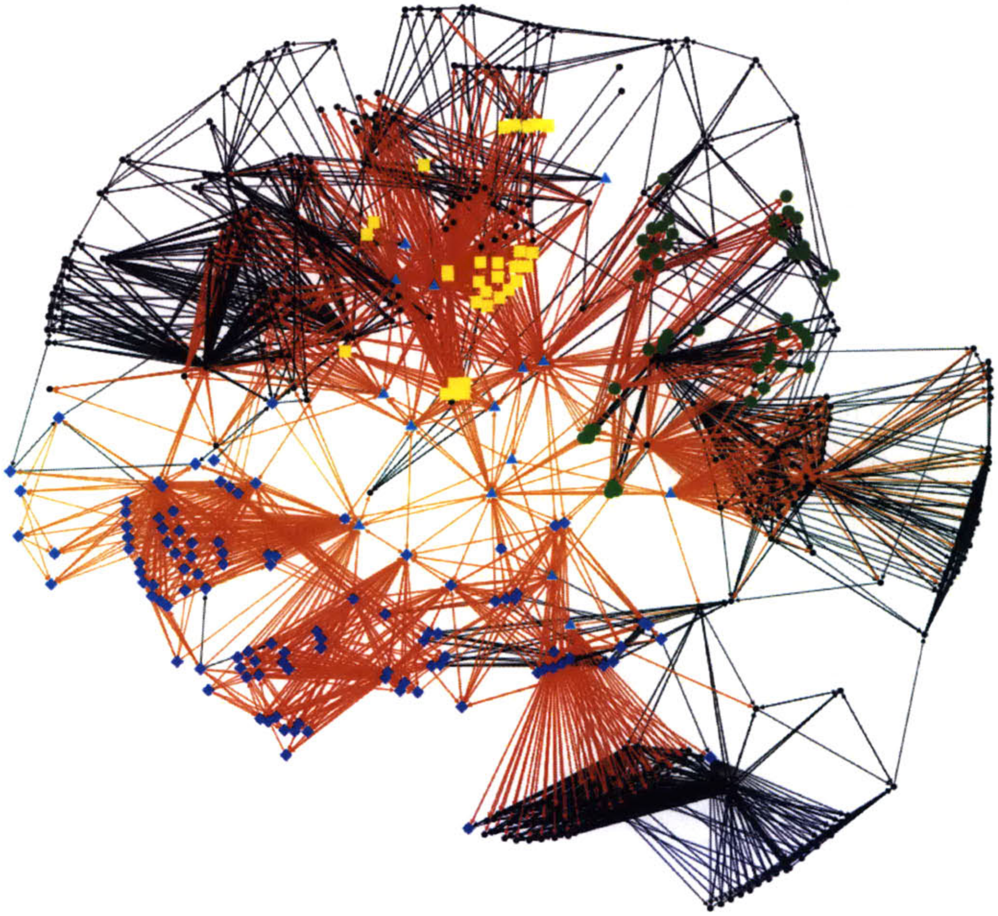


Figure 96 - MDPM Connecting Technologies for V1 Surveillance Process

DNA Extraction

Molecular Barcoding

Genomic Data Analysis

Process Design

From Figure 96 it is apparent that the "connecting technologies" have a significant impact on the rest of the process network and with each other, showing how widespread the impact of a new technology or change can be on a complex process. Descriptions of the technical details of each of these connecting technologies along with their impact on the MDPM network are given in the following sections.

5.4.1 DNA Extraction

One of the hardest problems in building the V1 process turned out to be what initially was thought to be the easiest: DNA extraction from the MRSA microbes. Many clinical microbiology labs routinely prepare MRSA samples for PCR-based assays but these are significantly less demanding of the quality and quantity of DNA produced with pico grams of fragmented DNA being sufficient to amplify the relatively small PCR targets (in the hundreds of bases) these assays target. In contrast next generation DNA sequencing requires 10 to 20 micro grams of DNA in large (greater than 20,000 base) single molecules. The impact "footprint" of the DNA extraction connecting technology is shown in Figure 97, where the DNA extraction related nodes in the organizational, information, and process domains are highlighted as solid green circles and immediate connections in red.

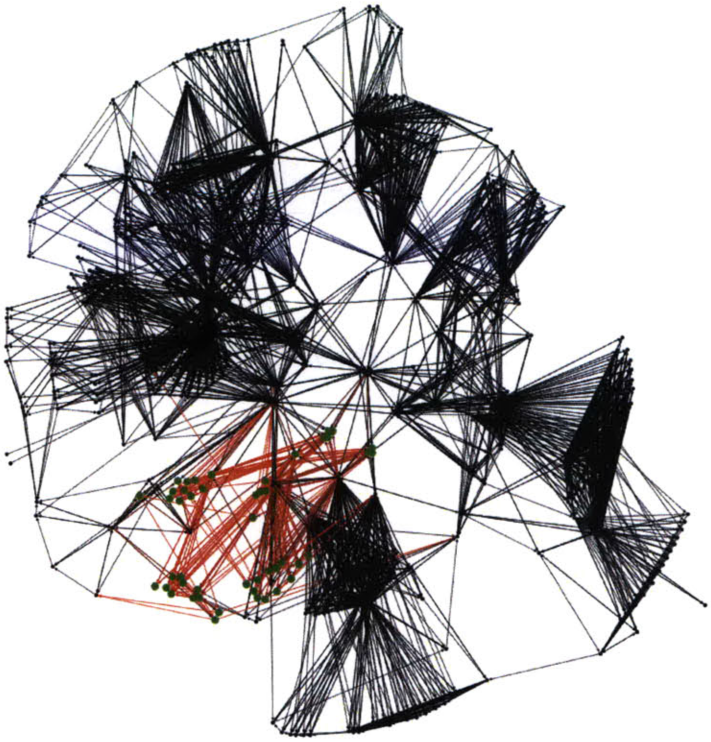
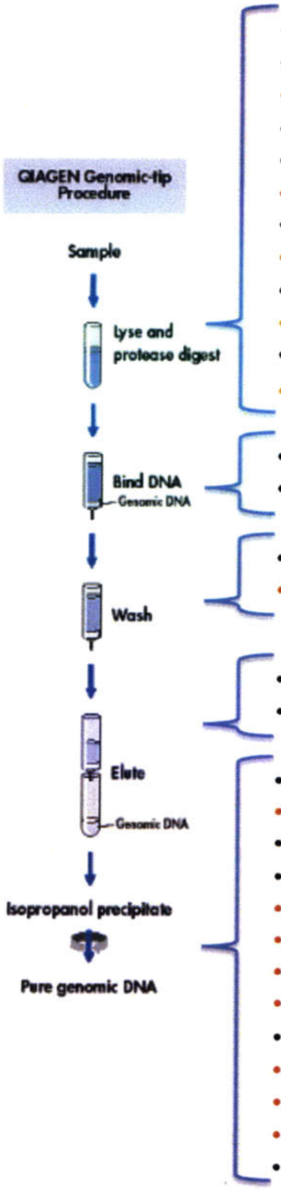


Figure 97 - Direct Impact of DNA Extraction on V1 Surveillance Process

A key observation from Figure 97 is that there are relatively few "long range" connections to this technology in the process network and it would appear to be fairly "modular" as most of the green nodes connect to each other with a few edges extending into the process network. However, this "modularity" could also be interpreted as a silo, where the DNA extraction portion of the process is not easily "reached" other than through a few connections. An observation from the network analysis but also an empirical one from developing the process is that these local "neighborhoods" must be formally tied together rather than relying on informal ad hoc connections to convey critical information. This means that instead of relying only on social network structures such as (Watts 2003) "small worlds" or (Granovetter 1973) "strength of weak ties", formal mechanisms must be put in place to also review elements of the information domain, and to actively "test" the process using industrial design of experiments methods such as those in (Box and Liu 1999). This might mean that instead of the relatively isolated DNA extraction portion of the process, that early identification of this key technology as an issue might allow for its "packaging" as a much better integrated module with formal design reviews to connect the information layer elements, a specific set of intra-process stakeholders, and a key set of "connected" experiments to determine the impact of changes in this part of the process on the rest of the process network. These ideas will be explored later in this chapter in describing potential improvements to produce a "V2" process.

The very first iteration of the process used a "standard" molecular biology protocol based on the leading vendor, Qiagen, which sells column based DNA extraction kits in routine use throughout the world for high molecular weight DNA applications. The particular product used was the Genomic tip 100/G and associated buffers from Qiagen. The 100/G product comes with a detailed guidebook detailing how it should be used to prepare DNA samples from a variety of sources, and how these samples can then be processed through the 100/G tips. Despite the kit being sold as an easy to use product, there are a large number of steps in the Qiagen DNA extraction process that require extreme user skill and tacit knowledge. For example, one of critical steps in obtaining DNA is to carefully wash with ethanol a hard to see and easily disturbed pellet on the side of a glass tube after centrifugation, which requires great skill on the part of the technician (as the author was able to verify in performing a number of these extractions). The detailed process steps in the protocol are shown in Figure 98 adapted from the vendor manual (Qiagen 2001). Steps requiring tacit knowledge (and technician skill) highlighted in red.

Qiagen Procedure for Bacterial DNA Extraction
100/G Midi Prep (25-90ug DNA in 20-160Kb Yield)



- Grow Sample: MRSA Culture in LB Media (2.2×10^{10} cells) **[tacit] [variable]**
- Prepare Necessary Reagents
- Pellet MRSA cells by centrifugation at 3000-5000g for 5-10 min. **[tacit]**
- Discard supernatant
- Resuspend pellets in 3.5ml of buffer B1 (w / Rnase A)
- Vortex to resuspend thoroughly **[tacit]**
- Add 80ul of lysozyme stock solution and 100ul of Proteinase K solution
- Incubate at 37C for at least 30 min. depending on how well resuspended **[tacit]**
- Add 1.2ml of buffer B2
- Mix or vortex thoroughly for few seconds **[tacit]**
- Incubate at 50C for 30 min.
- Verify lysate is clear, otherwise extend incubation and/or centrifuge **[tacit]**

- Equilibrate 100/G tip with 4ml buffer QBT and allow to empty by gravity flow
- Vortex sample for 10s @ max. speed and apply it to tip **[tacit]**

- Wash tip with 2 x 7.5ml of buffer QC by gravity flow
- Perform additional wash with 1 x 7.5ml of buffer QC if needed **[tacit]**

- Elute genomic DNA with 1 x 5ml of buffer QF pre-warmed to 50C
- Collect in 10ml glass centrifuge tube

- Precipitate DNA by adding 3.5ml room temperature isopropanol
- Mix by inverting the tube 10-20 times **[tacit]**
- Mark outside of tube
- Centrifuge >5000g for at least 15m at 4C
- Locate pellet, which may be difficult to see **[tacit]**
- Carefully wash the loosely attached pellet with 2ml of cold 70% ethanol **[tacit]**
- Carefully wash again if needed with 2ml of cold 70% ethanol **[tacit]**
- Vortex briefly **[tacit]**
- Centrifuge at >5000g for 10 min. at 4C
- Carefully remove the supernatant without disturbing the pellet **[tacit]**
- Air dry for 5-10 min. without over-drying **[tacit]**
- Resuspend the DNA in 1ml TE carefully by rinsing walls to recover all DNA **[tacit]**
- Dissolve DNA on a shaker at 55C for 1-2h or overnight **[tacit]**

Figure 98 - Qiagen 100/G Genomic Tip Protocol Showing Tacit Knowledge Steps in Red (Adapted from Qiagen)

Figure 98 highlights the need to look across domains in complex processes, since despite there being a documented protocol, or work instruction, in the information domain, the organizational domain connections are critical (containing the tacit knowledge). This become even more important in the transfer and scaling of complex processes, since it would not be sufficient to simply hand over the protocol shown in Figure 98 without some degree of training or interaction with the technicians that are familiar with the process and know the "tricks" to make it work as part of their tacit knowledge. The steps in black in Figure 98 represent explicit instructions with little or no judgment or skill required. A key variable in the use of the QIAGEN system is to produce a cell culture without too many cells (not more than 2.2 x 10^{10} for the 100/G) so that the genomic tip is not overloaded. However, this requires careful calibration of the growth media to match the growth rate of the particular cells and also depends on the type of cell being lysed, as the thick cell wall of MRSA means there is more cellular debris, which can lead to clogging of the tip. Typically the QIAGEN process can be optimized for a particular organism, where are all the cells will be similar leading to similar yields. However, for this surveillance process the input MRSA cells are expected to be different with varying growth rates and potentially cell wall thicknesses. The many downstream tacit steps compound the initial sample variability, as they require adjustment based on technician observations, skill, and judgment. All this leads to a highly variable process. A key paragraph in the vendor manual is the following:

> *".... protocols are optimized for use with fixed cell densities corresponding to the capacity of the QIAGEN Genomic-tip used. Overloading tips with DNA from an excessive number of cells (too much culture volume) will lead to reduced performance of the system. .... Please note that culture volumes may differ for other species of bacteria."*

The Qiagen method was initially used "off the shelf" to extract DNA from an initial panel of MRSA microbes totaling 168 attempts on a panel of 150 distinct samples obtained with anonymous labels from the Channing Laboratory at Harvard Medical School. The expected DNA yield was in the range of 25-90ug of DNA per the manufacturer. Figure 99 shows the actual DNA extraction yield obtained in micro grams of DNA from each of these attempts with the region in red denoting the acceptable yield range for the sequencing process. While some samples performed within the expected range, most produced very little yield and others produced none at all. The combination of input sample variability and a large number of also variable (tacit) steps in the protocol produce a highly variable and difficult to control process as shown in Figure 99 with only a few of the samples having sufficient DNA to continue the process.

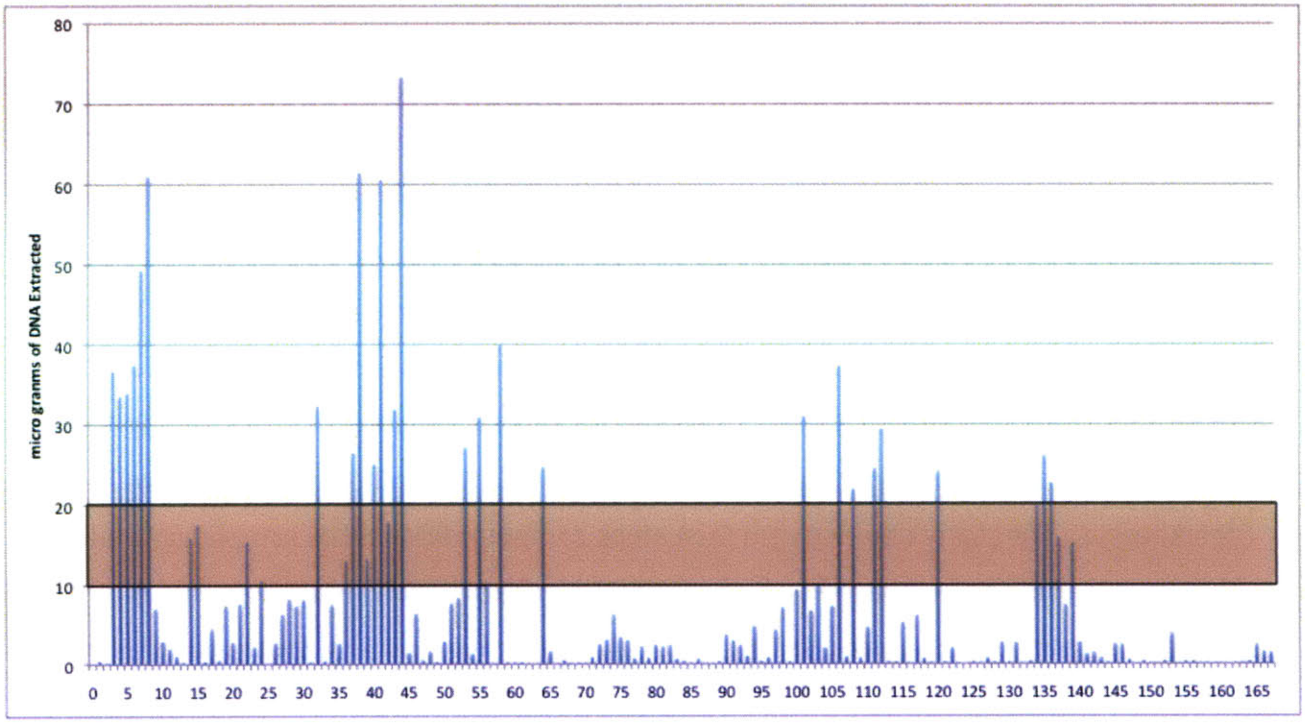


Figure 99 - MRSA DNA Extraction Yield from Un-optimized Qiagen 100/G Process

However, one interesting observation resulting from this initial effort was that certain samples appear to produce reproducible DNA yields suggesting certain properties of the microbes themselves might be the cause for the observed variation rather than variability in the technique or reagents. A possible explanation may be that each of these strains have either different cell wall thicknesses or different growth rates that fall within the feasible parameter range for the process. Different cell wall thicknesses would lead to some cells lysing under a given set of conditions and others with thicker cell walls not lysing at all which would explain the zero yield results. The other hypothesis is that differential growth rates will lead to some samples growing far more cells than others and potentially clogging the DNA extraction filters also explaining the zero yield attempts. In order to identify the root cause of the issue a design of experiments series was run testing each of the key variables in this part of the process. A series of DOEs were eventually required to zero in on conditions that would work for a variety of different samples, and a representative DOE is shown in Figure 100 developed by a Marcia Lara, a scientist in the Broad Institute specializing in DNA extraction along (also the DNA extraction scientist "node" in the organizational domain) with the clinical microbiology laboratory staff at the Brigham and Women's Hospital, and the thesis author.

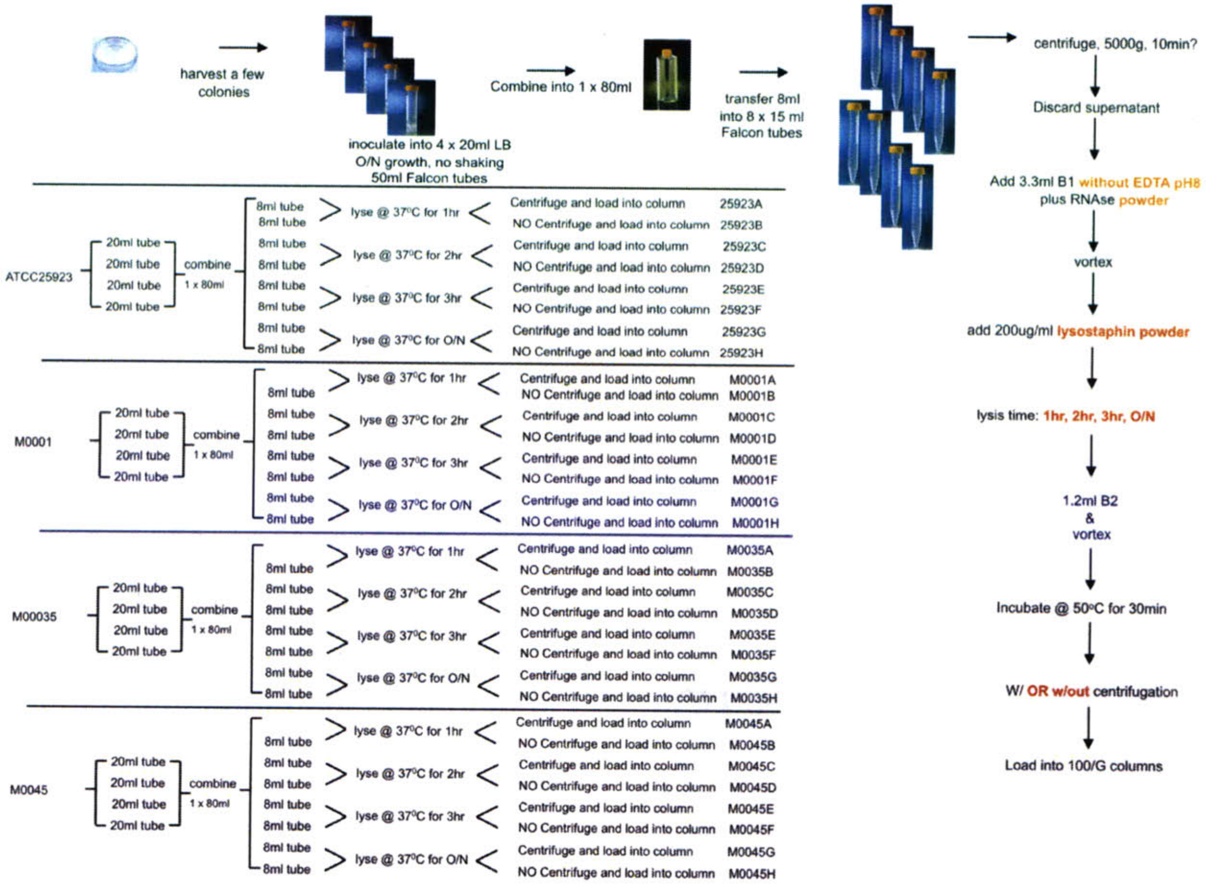
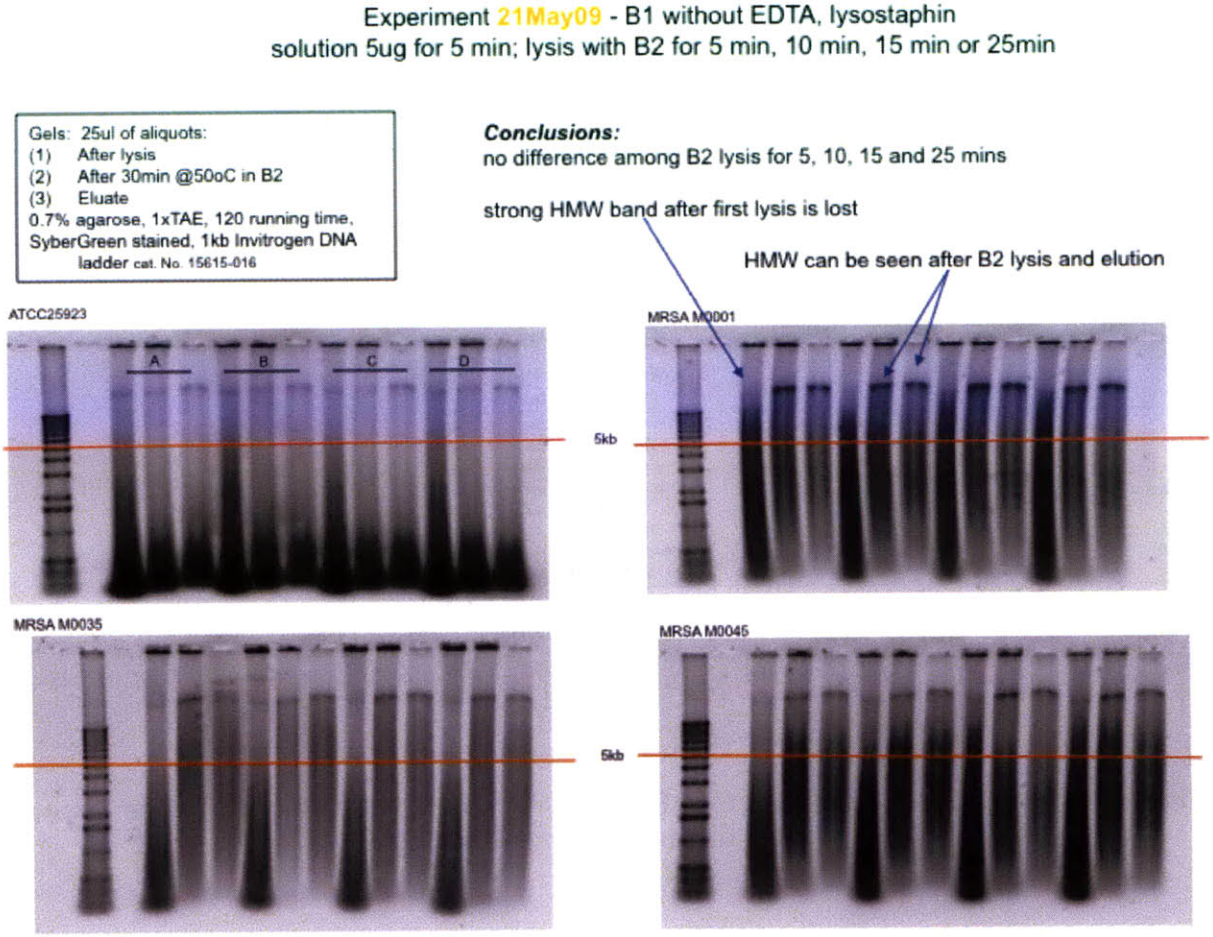

harvest a few colonies
inoculate into 4 x 20ml LB
O/N growth, no shaking
50ml Falcon tubes
Combine into 1 x 80ml
transfer 8ml into 8 x 15 ml Falcon tubes
centrifuge, 5000g, 10min?
Discard supernatant
Add 3.3ml B1 without EDTA pH8 plus RNAse powder
vortex
add 200ug/ml lysostaphin powder
lysis time: 1hr, 2hr, 3hr, O/N
1.2ml B2 & vortex
Incubate @ 50ºC for 30min
W/ OR w/out centrifugation
Load into 100/G columns
ATCC25923
20ml tube
20ml tube
20ml tube
20ml tube
combine 1 x 80ml
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
lyse @ 37ºC for 1hr
lyse @ 37ºC for 2hr
lyse @ 37ºC for 3hr
lyse @ 37ºC for O/N
Centrifuge and load into column 25923A
NO Centrifuge and load into column 25923B
Centrifuge and load into column 25923C
NO Centrifuge and load into column 25923D
Centrifuge and load into column 25923E
NO Centrifuge and load into column 25923F
Centrifuge and load into column 25923G
NO Centrifuge and load into column 25923H
M0001
20ml tube
20ml tube
20ml tube
20ml tube
combine 1 x 80ml
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
lyse @ 37ºC for 1hr
lyse @ 37ºC for 2hr
lyse @ 37ºC for 3hr
lyse @ 37ºC for O/N
Centrifuge and load into column M0001A
NO Centrifuge and load into column M0001B
Centrifuge and load into column M0001C
NO Centrifuge and load into column M0001D
Centrifuge and load into column M0001E
NO Centrifuge and load into column M0001F
Centrifuge and load into column M0001G
NO Centrifuge and load into column M0001H
M00035
20ml tube
20ml tube
20ml tube
20ml tube
combine 1 x 80ml
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
lyse @ 37ºC for 1hr
lyse @ 37ºC for 2hr
lyse @ 37ºC for 3hr
lyse @ 37ºC for O/N
Centrifuge and load into column M0035A
NO Centrifuge and load into column M0035B
Centrifuge and load into column M0035C
NO Centrifuge and load into column M0035D
Centrifuge and load into column M0035E
NO Centrifuge and load into column M0035F
Centrifuge and load into column M0035G
NO Centrifuge and load into column M0035H
M0045
20ml tube
20ml tube
20ml tube
20ml tube
combine 1 x 80ml
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
8ml tube
lyse @ 37ºC for 1hr
lyse @ 37ºC for 2hr
lyse @ 37ºC for 3hr
lyse @ 37ºC for O/N
Centrifuge and load into column M0045A
NO Centrifuge and load into column M0045B
Centrifuge and load into column M0045C
NO Centrifuge and load into column M0045D
Centrifuge and load into column M0045E
NO Centrifuge and load into column M0045F
Centrifuge and load into column M0045G
NO Centrifuge and load into column M0045H

Figure 100 - MRSA DNA Extraction DOE Example Varying Lysostaphin Incubation Time and Post incubation Centrifugation

The particular DOE shown in Figure 100 varied the incubation time following the addition of lysostaphin to digest the thick MRSA cell wall, as well as centrifugation or not post incubation.   The 4 incubation conditions, and +/- centrifugation were attempted on each of four MRSA samples, yet this experiment did not find a condition that worked for all four samples. A large number of similar experiments were performed varying nearly all of the parameters in Figure 98, and some conditions were found to deliver sufficient DNA yield, but not sufficient quality as is described next.

A second key requirement in the DNA extraction process was the need for large size fragments (>5Kb) as these may be needed for the downstream computational assembly. Further optimization was required to deliver samples in the required size range with a representative example of the results shown in Figure 101 showing the sizes DNA delivered for a particular set of process parameters compared to the 5Kb desired.



Experiment 21May09 - B1 without EDTA, lysostaphin solution 5ug for 5 min; lysis with B2 for 5 min, 10 min, 15 min or 25min

Gels: 25ul of aliquots:
(1) After lysis
(2) After 30min @50oC in B2
(3) Eluate
0.7% agarose, 1xTAE, 120 running time, SyberGreen stained, 1kb Invitrogen DNA ladder cat. No. 15615-016

Conclusions:
no difference among B2 lysis for 5, 10, 15 and 25 mins

strong HMW band after first lysis is lost

HMW can be seen after B2 lysis and elution

Figure 101 - MRSA DNA Extraction DOE Results

The images in Figure 101 show the gel bands associated with each of 4 different MRSA samples under several conditions of a DOE. The red 5Kb line denotes molecules of that size and dark bands extending above the red line have higher molecular weight and are preferable. One result of the optimization of lysis was that too strong of a lysis condition would degrade the DNA presumably because thin-walled MRSA would lyse easily under the more stringent conditions required to lyse the thicker cell walled organisms leading to DNA damage for the relatively thin-walled MRSAs. Significant optimization of this process eventually yielded a set of conditions that worked for 48 MRSAs in the original panel, although significant work would be required to develop a "V2" process that could yield sufficient DNA for any MRSA that could be input into the process. This is again, why DNA extraction is a critical "connecting" technology, where a failure in delivering sufficient DNA of adequate quality has significant impact throughout the process network all the way to the eventual analysis of the genomic data.

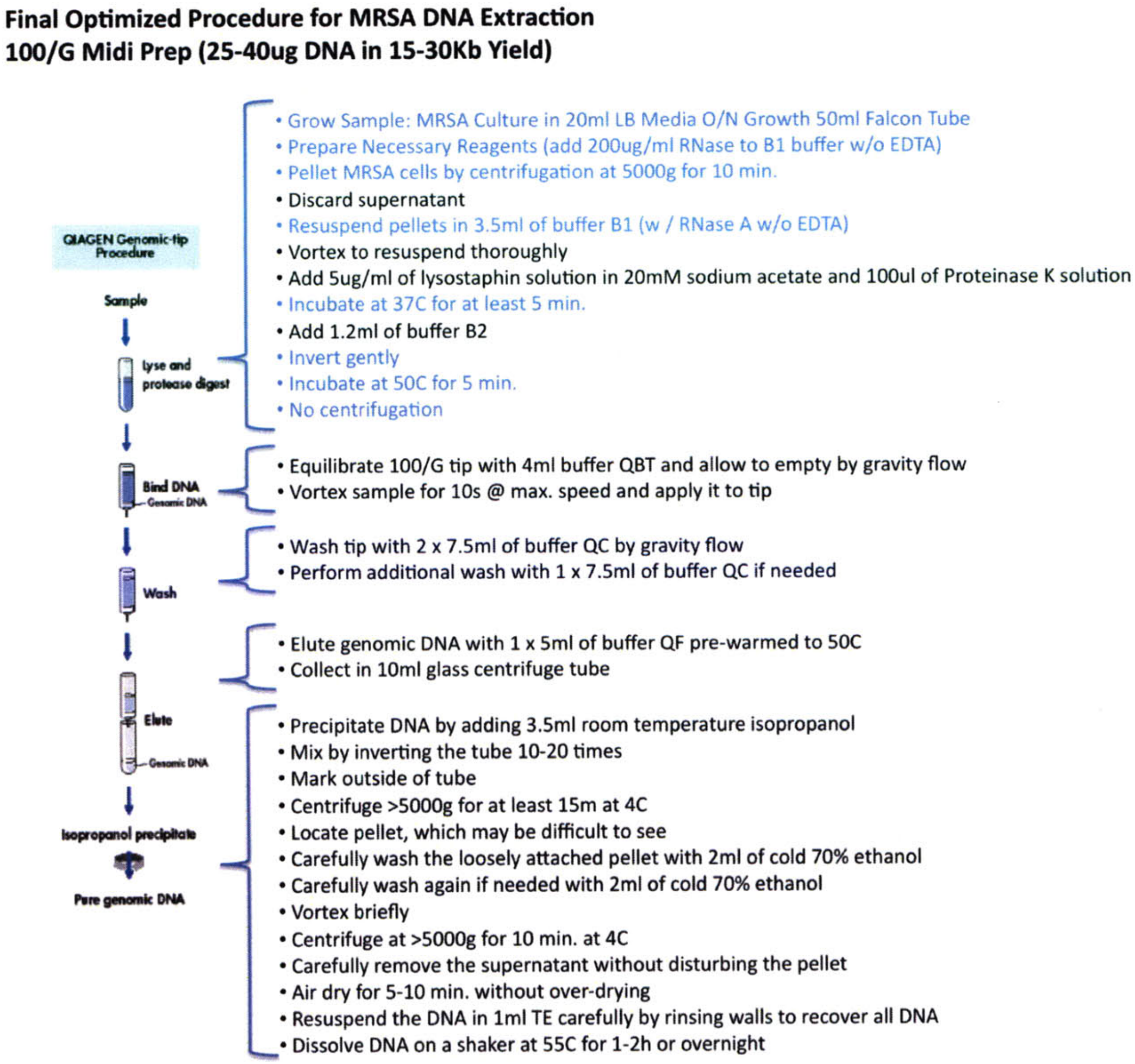
The DOEs were conducted to test variables in the lysis portion of the process in Figure 98 and those tested are listed in Table 22 to find a set of conditions that would work for MRSA samples.

DOE Variable	Rationale
Culture Volume (8, 10, 20 ml)	Too many cells causing clogging
% of culture growth used (80%, 100%)	Too many cells causing clogging, debris
Transfer of inoculate to new tube (y/n)	Losing some cells due to transfer
Lysis buffer B1 with EDTA (y/n)	EDTA may inhibit Lysostaphin
Lysostaphin conc. ug/ml (200, 100, 50, 10, 5, 1)	Too much or too little Lysostaphin
Lysostaphin with 5 min. boiling (y/n)	Boiling may be needed to open cells
Vortexing or inverting gently after B2 addition	Vortexing may damage DNA
First B1 lysis time min. (O/N, 60, 45, 30, 25, 5, 1)	Lysis time may not be sufficient, or too long causing DNA damage
Second B2 lysis time (45m, 30m, 25m, 15m,	Lysis time may not be sufficient, or too

DOE Variable	Rationale
10m, 5m)	long causing DNA damage
Centrifugation after second lysis (y/n)	Centrifugation may damage DNA

Table 22 - DOE Variables Tested for MRSA Lysis Optimization

The series of variables listed in Table 22 were used in a series of DOEs that eventually produced the optimized process described in Figure 102 with steps that changed highlighted in blue.



Final Optimized Procedure for MRSA DNA Extraction
100/G Midi Prep (25-40ug DNA in 15-30Kb Yield)

- Grow Sample: MRSA Culture in 20ml LB Media O/N Growth 50ml Falcon Tube
- Prepare Necessary Reagents (add 200ug/ml RNase to B1 buffer w/o EDTA)
- Pellet MRSA cells by centrifugation at 5000g for 10 min.
- Discard supernatant
- Resuspend pellets in 3.5ml of buffer B1 (w / RNase A w/o EDTA)
- Vortex to resuspend thoroughly
- Add 5ug/ml of lysostaphin solution in 20mM sodium acetate and 100ul of Proteinase K solution
- Incubate at 37C for at least 5 min.
- Add 1.2ml of buffer B2
- Invert gently
- Incubate at 50C for 5 min.
- No centrifugation

- Equilibrate 100/G tip with 4ml buffer QBT and allow to empty by gravity flow
- Vortex sample for 10s @ max. speed and apply it to tip

- Wash tip with 2 x 7.5ml of buffer QC by gravity flow
- Perform additional wash with 1 x 7.5ml of buffer QC if needed

- Elute genomic DNA with 1 x 5ml of buffer QF pre-warmed to 50C
- Collect in 10ml glass centrifuge tube

- Precipitate DNA by adding 3.5ml room temperature isopropanol
- Mix by inverting the tube 10-20 times
- Mark outside of tube
- Centrifuge >5000g for at least 15m at 4C
- Locate pellet, which may be difficult to see
- Carefully wash the loosely attached pellet with 2ml of cold 70% ethanol
- Carefully wash again if needed with 2ml of cold 70% ethanol
- Vortex briefly
- Centrifuge at >5000g for 10 min. at 4C
- Carefully remove the supernatant without disturbing the pellet
- Air dry for 5-10 min. without over-drying
- Resuspend the DNA in 1ml TE carefully by rinsing walls to recover all DNA
- Dissolve DNA on a shaker at 55C for 1-2h or overnight

Figure 102 - Optimized MRSA DNA Extraction Protocol (blue steps indicate a change)

While the technical work associated with the DOEs used to find feasible DNA extraction conditions was a critical part of the process development, so was fostering the connection of the Broad Institute DNA extraction scientist to the clinical microbiology laboratory at BWH. Numerous discussions, e-mails, and data exchanges were required to bridge these two "silos" of expertise, and it was only through the joint effort enabled by the "network connection" that this was possible. This connection is shown in the V1 organization layer, and through it, the clinical microbiology laboratory obtained the necessary DNA extraction expertise (tacit) and protocols (explicit) in order to develop a MRSA specific protocol. This is now a key resource for the community as it enables other researchers to perform extractions for DNA sequencing on a variety of MRSA strains and provides a framework for further optimization. The author of this thesis, functioning as the process owner and process designer provided the necessary DOE expertise to ensure conditions could be rapidly screened and promising conditions could be further refined without resorting to one factor at a time experiments that would have taken much longer and been more expensive than a DOE approach and may not have revealed interactions between factors such as the impact of longer lysis times on DNA degradation (molecular weight). The V1 network allowed for this scaffolding to join 3 fields (silos): process design (DOEs), microbiology, and molecular biology into a working process. Explicitly mapping these interactions is one of the key advantages of the MDPM model.

5.4.2 Molecular Barcoding

The second key connecting technology was the development of molecular barcoding such that multiple samples can be run on a single sequencing run. Because of the high data yield per run of an instrument (>20Gb) and the relatively small genome size of a MRSA microbe (~2.8Mb) it is necessary to barcode multiple samples such that they can be run simultaneously. Note that it is not the entire genome being barcoded, but rather the DNA fragments produced from the DNA preparation steps which are much smaller fragments (200-400 bases) produced from the genomic DNA in the original MRSA sample culture. These small fragments are the DNA molecules actually read by the instrument and because the technology is only able to read between 50-100 bases accurately, it is necessary to have many overlapping fragments. However, if each of the DNA preparations can be barcoded, then the fragments derived from these could be pooled together and placed on a single instrument run. A conceptual schematic of two fragments each from two different samples, but with unique barcodes corresponding to each sample is shown in Figure 103.

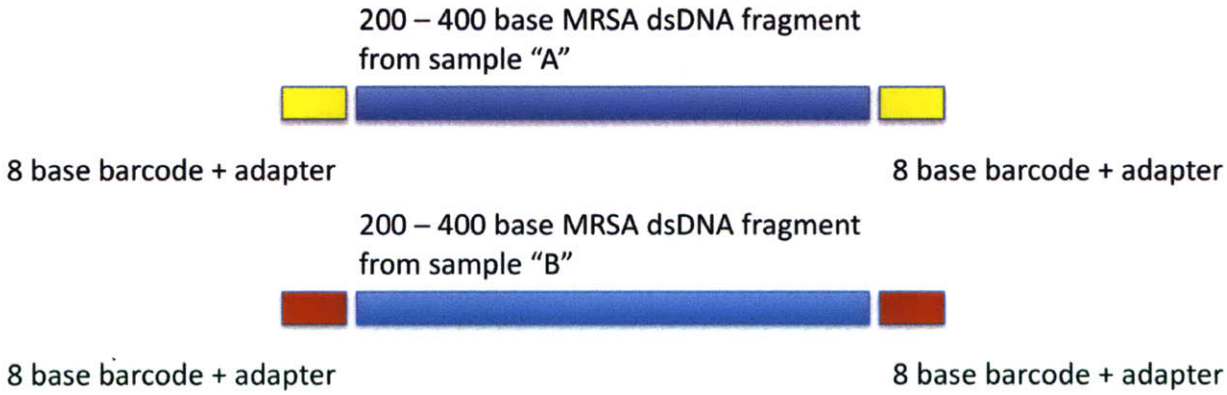

200 – 400 base MRSA dsDNA fragment from sample "A"
8 base barcode + adapter
8 base barcode + adapter
200 – 400 base MRSA dsDNA fragment from sample "B"
8 base barcode + adapter
8 base barcode + adapter

Figure 103 - DNA Fragment Barcoding Conceptual Schematic for Two Distinct Samples

Although the conceptual schematic in Figure 103 appears simple and straightforward, molecular barcoding is fraught with potential problems and requires extensive support throughout the process. For example, the downstream data analysis will have to be set up such that it can identify individual barcodes accurately and can bin the data appropriately according to the original sample ID. The barcodes themselves will have to be designed in such a way that they do not overlap and are tolerant of sequencing errors (i.e. if one base is mistakenly read, the remaining bases are still sufficient to identify the barcode) accomplished with error correcting codes. The direct impact of molecular barcoding (i.e. all nodes in all domains directly associated with barcoding) is shown in Figure 104 with bar coding nodes labeled as yellow squares and 1-hop connections in red and the impact is significant in multiple domains.

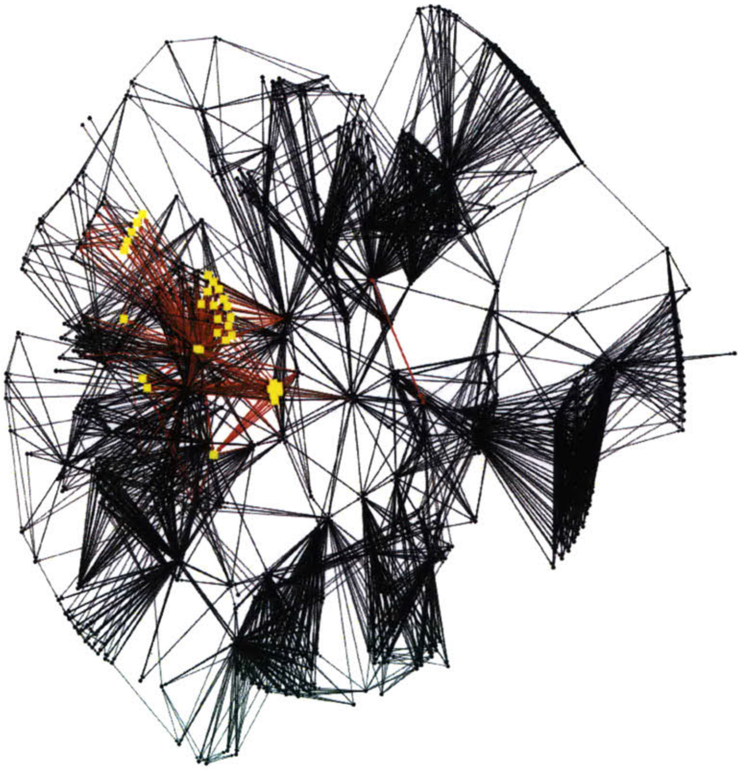
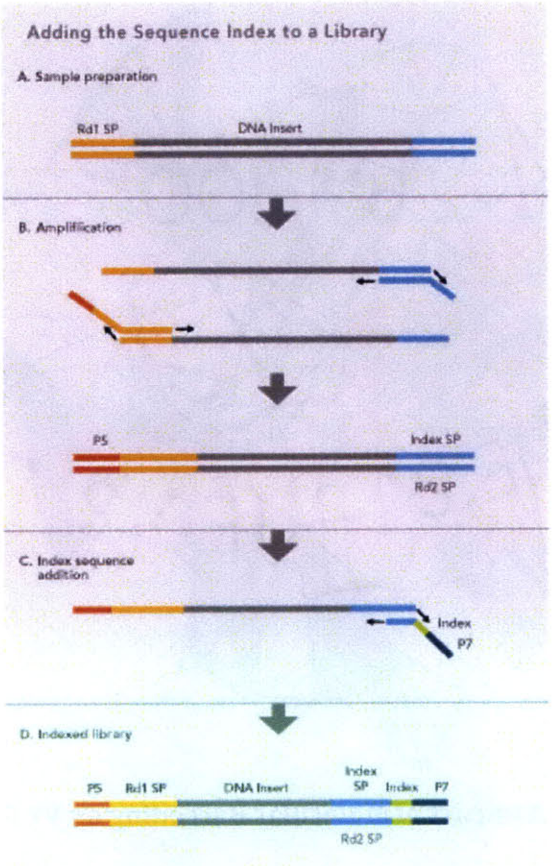


Figure 104 - Direct Impact of Molecular Barcoding on V1 Surveillance Process

A fundamental requirement of the barcoding step is that it does not change or bias the DNA sequence data in introducing the barcode. Also barcoding must be done in such a way that each of the samples introduced into the pool are represented in equal molar concentrations so that one sample doesn't dominate the data produced in a sequencing run. Thus if there are 100 barcoded samples in a run, each sample should represent 1% of the data produced. Because of its obvious benefits to pool multiple samples molecular barcoding is an active area of development for many applications beyond MRSA surveillance and a commercially available reagent kit is available from Illumina[8]. This kit is labeled "method A" in the MDPM process domain and the barcode labeling mechanism is shown in Figure 105.

[8] http://www.illumina.com/technology/multiplexing_sequencing_assay.ilmn



(A) During sample preparation, adapters are ligated to the DNA fragments. One adapter contains the sequencing primer site for application read 1 (Rd1 SP). (B) Prepared samples are amplified via PCR using two universal primers. One primer contains an attachment site (P5) for the flow cell, while the other contains the sequencing primer sites for the index read (Index SP) and for application read 2 (Rd2 SP). (C) A third primer in the PCR adds the Index as well as a second flow cell attachment site (P7) to the PCR product shown in step 2. (D) The indexed library is ready for sequencing using the Genome Analyzer system.

Figure 105 - Illumina Multiplex Sequencing Barcode Labeling Mechanism (from Illumina)

The vendor kit relies on a 3 primer PCR, where adapters are added to DNA fragments, followed by PCR with tailed primers, one of which in turn contains a priming site for a third primer containing a barcode. This kit enables the barcode labeling of 12 individual samples, and because each fun in an Illumina instrument is divided into 8 lanes, corresponds to 96 individual samples per instrument run, which still corresponds to over 110X coverage of each MRSA genome for a 30Gb run. This coverage is likely excessive for mutation finding, but may not be for de novo assembly due to the large amount of "overlap" needed by the reads to uniformly cover the genome and assemble the overall puzzle. However, as yields continue to increase for next generation sequencing technologies, greater numbers of barcodes will be required than those initially provided.

The "method A" vendor kit was tested as part of the initial process verification, and one concern appeared to be the number of PCR cycles needed to affix the 3rd primer barcode onto the sample molecules, as high numbers of PCR cycles can introduce significant bias by preferentially amplifying certain DNA molecules over others leading to a reduced diversity of "unique" molecules. These vendor barcoding kits were intended for a number of applications, but they were initially tested on the MRSA surveillance samples (and V1 process) developed here. In the first test pools of 12, 24, and 48 samples were sequenced together (additional barcode primers were made to augment the vendor provided number). The barcoded data was then separated into individual bins corresponding to each sample and a typical result is shown in Figure 106 for a 50Kb region of the genome of a single sample. Each grey bar represents the reading (sampling) of an individual DNA fragment coming from the original sample measuring 75 bases each, the data was plotted using the IGV viewer[9]. The alignment algorithm then places these fragments onto their correct location on the MRSA genome by comparing to a reference (this is why they "pile up" as in this example the genome was covered over 100X, which means that on average 100 individual "reads" should land at each corresponding position of the genome. The colored lines in Figure 106 represent differences from the reference at each base location, called single nucleotide polymorphisms (SNPs). The assembly process by comparison is much harder as that algorithm attempts to put all of these puzzle pieces together without the assistance of a reference (i.e. as if this were a de novo organism). In Figure 106 spikes can clearly be seen where molecules "pile up" as part of the sequencing sample preparation process, but the "spikiness" is made worse by the PCR-based method A, which relies on a high number of PCR cycles to barcode label causing certain molecules that amplify with slightly greater efficiency to be greatly over represented. Of greater concern is the opposite

[9] Integrative Genomics Viewer (IGV) from the Broad Institute: http://www.broadinstitute.org/igv

case, where molecules that are not as efficient will not be represented leading to areas with much lower coverage in the genome as shown by the dips in the coverage plot.

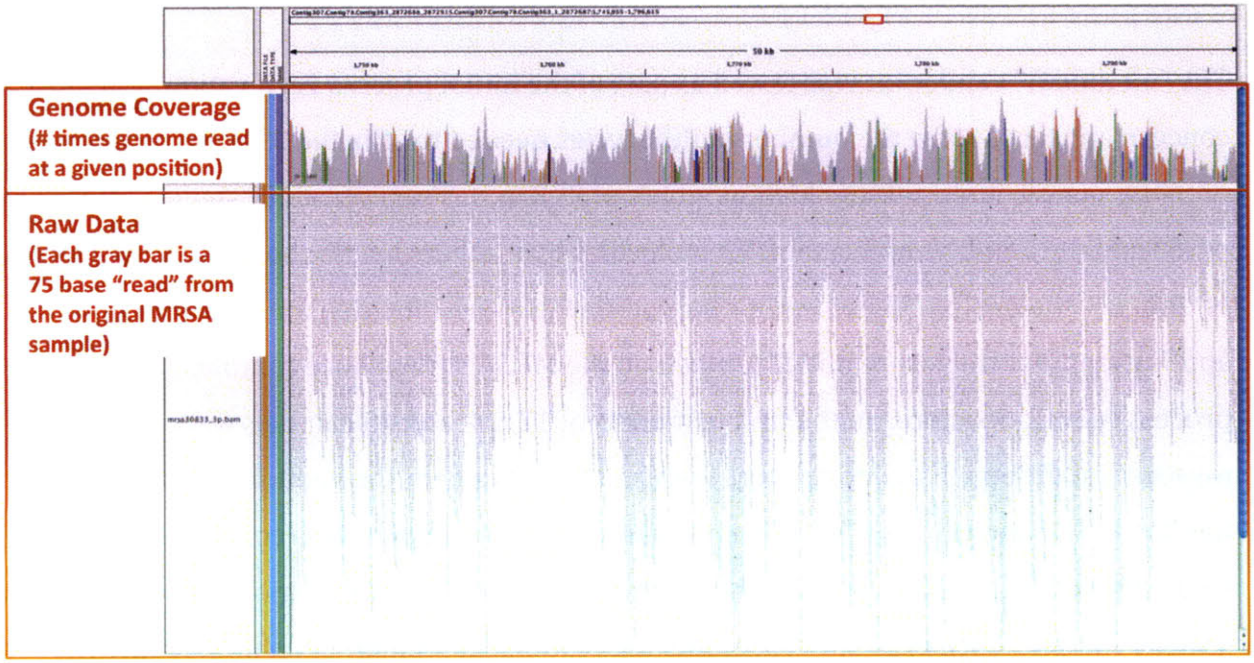

50 kb
1,750 kb
1,760 kb
1,770 kb
1,780 kb
1,790 kb
Genome Coverage
(# times genome read at a given position)
Raw Data
(Each gray bar is a 75 base "read" from the original MRSA sample)

Figure 106 - Sequence Coverage from Vendor Barcoding Method

The dips in coverage did not occur throughout the genome but happened frequently enough to be noticeable by eye as in Figure 106. Further zooming in to one of these regions and comparing it to a sample sequenced without barcoding shows that in difficult to sequence regions the PCR based barcoding can make coverage dips much worse, as can be seen in Figure 107 comparing a 40Kb region of a MRSA sample sequenced without PCR-based barcoding at the top and a PCR-based barcoded one at the bottom plotted using the Argo[10] viewer. The barcoded sample has no molecules represented in certain regions, and while the non-barcoded sample also shows a dip in this region, it is still able to provide coverage in that region of the genome. These holes in coverage are obviously problematic as no data can be obtained from them, and critical gene changes perhaps leading to resistance or virulence would not be seen by the downstream analysis.

[10] Argo Genome Browser from the Broad Institute: http://www.broadinstitute.org/annotation/argo/

PCR-Based barcode method comparison to non-barcoded sequence:
Top: S. aureus reference lane (single lane high coverage >200X)
Bottom: S. aureus barcoded sample (~120X from 3 separate development lanes)
Difficult region but barcoding appears to make it worse

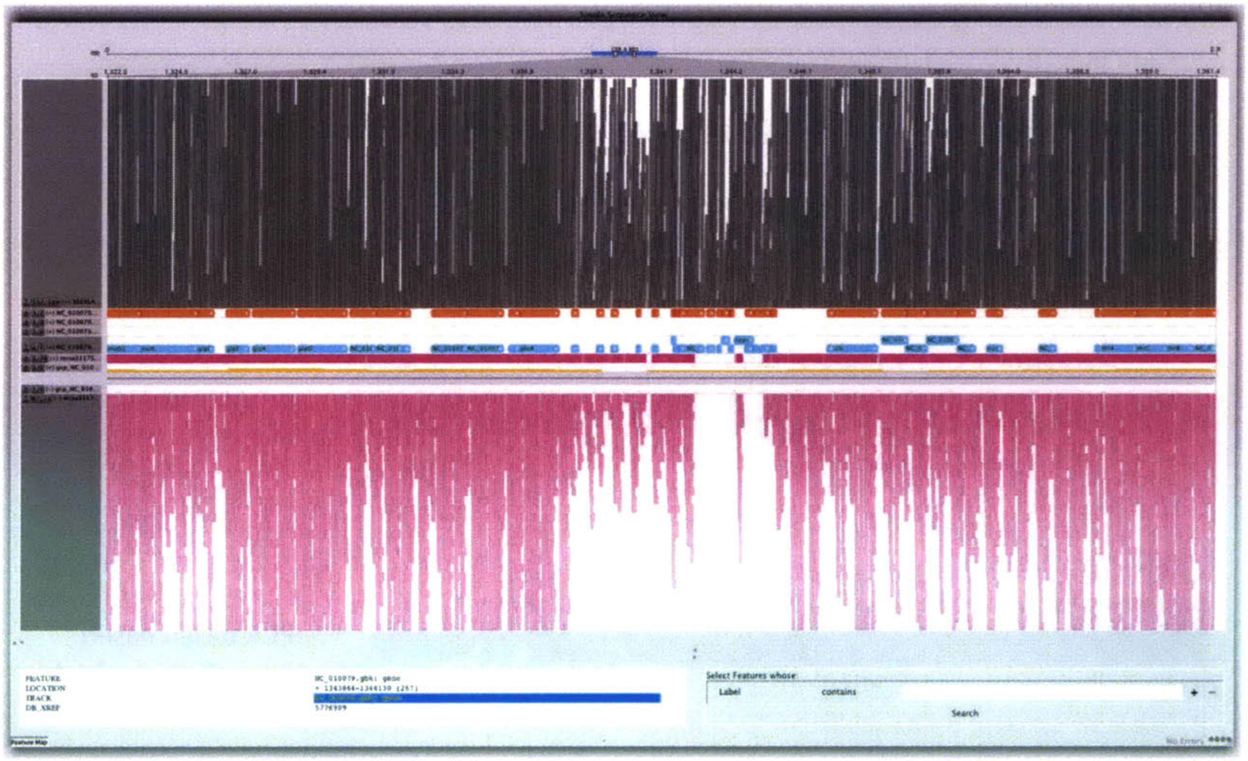
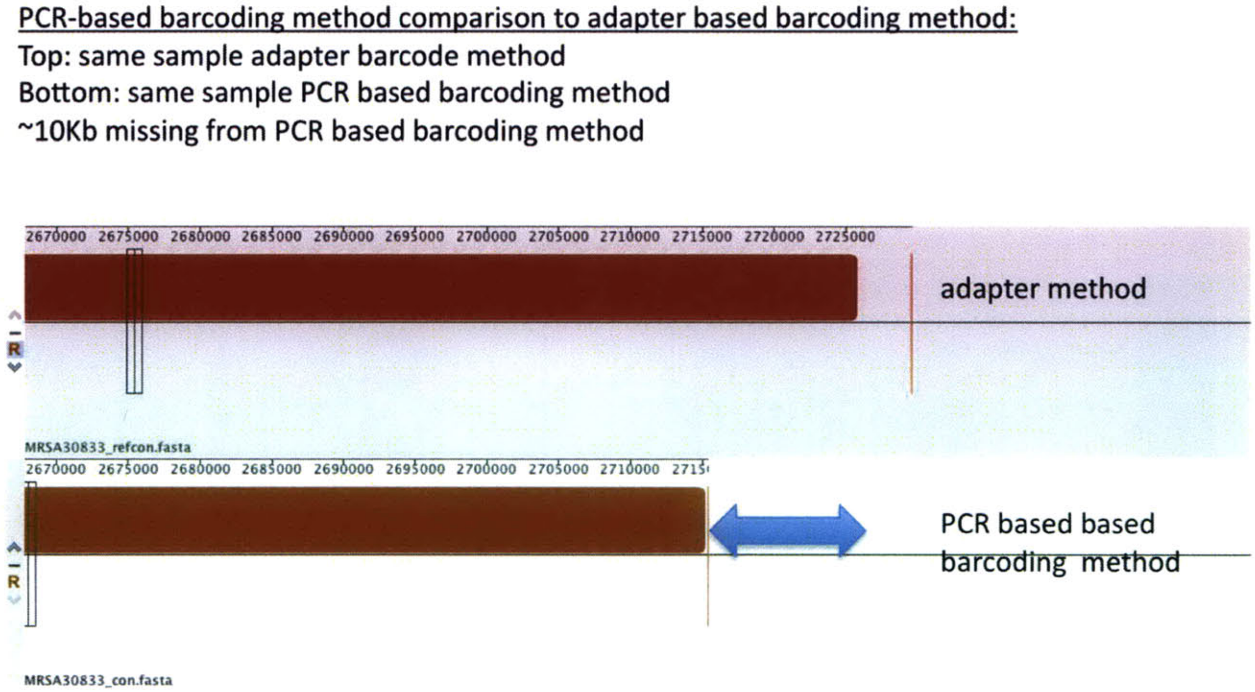


Figure 107 – 40Kb Sequence Coverage from PCR Based Barcoding Method (Bottom) Compared to Non-Barcoded Sequence (Top)

Because of the issues found in the initial barcode "method A", which nevertheless has certain workflow advantages, a barcode "method B" was developed which does not rely on PCR for the barcode labeling but rather includes it as part of the standard adapter ligation necessary as part of the sample prep. The "troubleshooting" and process change needed to identify this problem, devise a solution, and test flowed through the various domains and connections described in the V1 MDPM model, taking advantage of a number of cross-organization and cross-domain connections to speed the development cycle as this method is critical not just for the surveillance system but also for many other applications including cancer sequencing. The results of the new process using adapters such that the barcode is directly incorporated can be seen in a comparison of the amount of the genome captured as data using both methods for the same sample in Figure 108. The adapter method captures

10kb more than the original PCR based method. This difference in sequence length is likely due to some marginally covered regions in normal sequencing dropping out with the PCR-based method, but the same effect of reduced coverage in other sections can also affect the ability to identify SNPs.



PCR-based barcoding method comparison to adapter based barcoding method:
Top: same sample adapter barcode method
Bottom: same sample PCR based barcoding method
~10Kb missing from PCR based barcoding method

Figure 108 - Sequence Coverage from Adapter and PCR Based Methods for the Same Sample, Showing PCR-Based Method Fails to Capture over 10Kb of MRSA Genome

5.4.3 MRSA Genome Analysis

A third key connecting technology is the analysis of the genomic data produced. This is an area of enormous investment and development by everyone associated with DNA sequencing as the rapid increase in data production has forced a related development of new algorithms to handle the massive amounts of data but also the new properties of the data as described in (Shendure and Ji 2008). New sequencing technologies produce much smaller data fragments, or "reads" ranging from 25 bases up to 400 bases. The original capillary based sequencing instruments (the Sanger technology) produced far fewer reads but each was often in excess of 700 bases which greatly facilitated the putting together of the "puzzle pieces" of a genome which in the MRSA case is greater than 2.8 million bases. This means that with the shorter reads there are far more puzzle pieces but also because these pieces are shorter they are harder to place uniquely. The V1 surveillance process developed here takes advantage of very recently developed algorithms (ca. 2008-2009) such as the Maq aligner written by and described in (Li, Ruan et al. 2008) and the recently developed velvet assembler described in (Zerbino and Birney 2008). As described in the previous section an aligner places reads on an existing reference but cannot build "new sequence" where the reference ends and thus is only useful to look at variation in comparison to known genes and reference genomes.

The results of a Maq alignment were shown in Figure 106, Figure 107, and Figure 108 to compare the PCR-based and adapter based barcoding methods. A detailed view of the raw data coming from the process is illustrative as shown in Figure 109 which shows 75base reads as gray bars corresponding to the top and bottom strand of the DNA sample. The colored boxes represent differences from the reference genome and it is clear there is a real SNP corresponding to an G to A change in the top strand shown by the green boxes in the top strand (and the corresponding C to T blue boxes in the bottom strand). Note however that the data is "noisy" with many errors towards the end of reads (random colored boxes). The alignment software such as Maq must discriminate between the real SNPs such as the G to A change in Figure 109 and the noise in found elsewhere. One way to address this is through coverage or "reading" the genome multiple times such that statistical models can filter out noise based on multiple observations. This is why consistent and high coverage is so important and why the lower coverage due to the PCR-based barcoding method is problematic, as in low coverage regions a noise error may look like a read mutation without additional data to confirm.

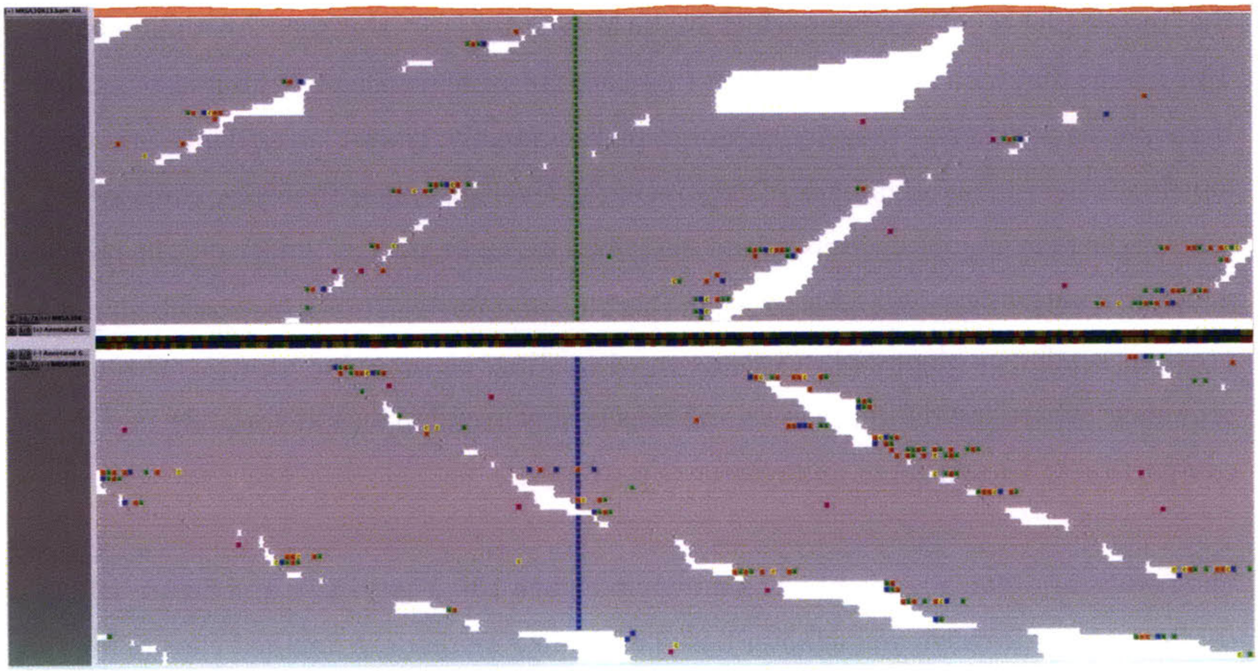


Figure 109 - Raw Sequencing Data From a Region of the MRSA Genome

An assembler is significantly more complex than an aligner as it must assemble the puzzle pieces without any reference and only based on how they overlap and fit together. The data shown in Figure 109 were aligned using a reference that places each of the reads uniquely on a position in the reference. The conceptual differences between an aligner and an assembler are shown in Figure 110, which compares the resulting genome from each. Note that if there are no reads covering a reference position but this position lies within the reference (an already sequenced MRSA genome) the missing bases can be "inferred". However, if there are reads falling outside the reference, the aligner will not be able to place them as there is no reference to map onto. In comparison, the assembler (velvet and others) can stitch these reads together without the need for a reference although a much more complex algorithm and additional data are needed.

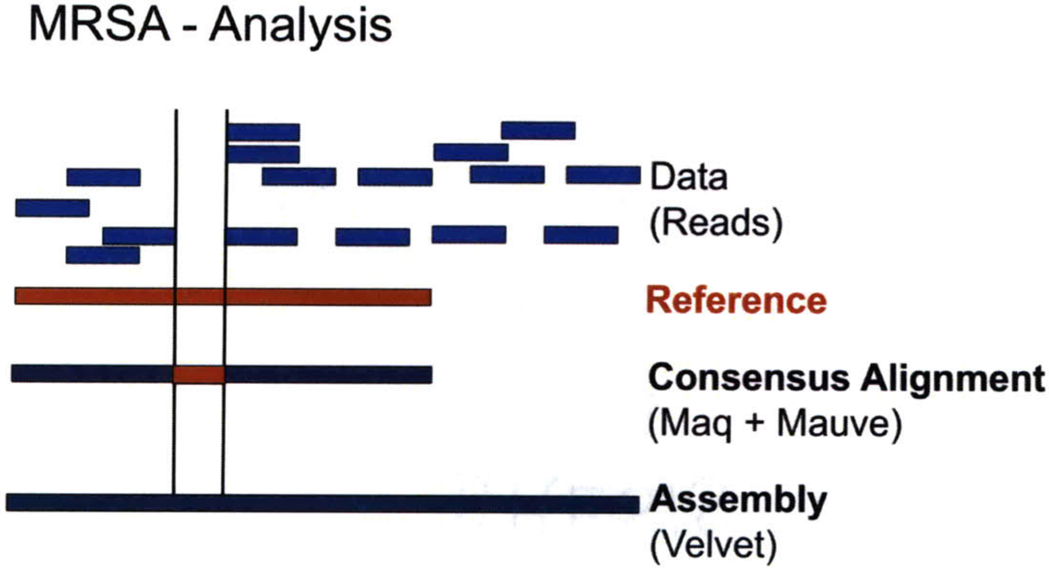


Figure 110 - Conceptual Comparison of an Aligner and an Assembler in Producing Full MRSA Genome Sequences

The Maq aligner used in this thesis were developed at the Sanger center in the UK, but is still very much prototype code requiring significant development and in an MDPM network sense an enormous amount of scaffolding. The velvet assembler was written at the European Bioinformatics Institute (EMBL-EBI), which is part of the European Molecular Biology Laboratory. Velvet is even more of a prototype with limited testing and support. Other newer algorithms are rapidly emerging and these are labeled as methods A, B, and C for each of the alignment process and assembly processes as placeholders for these new developments. The impact of the genome analysis connecting technology is significant covering all domains and spanning multiple process domain steps as can be seen in Figure 111.

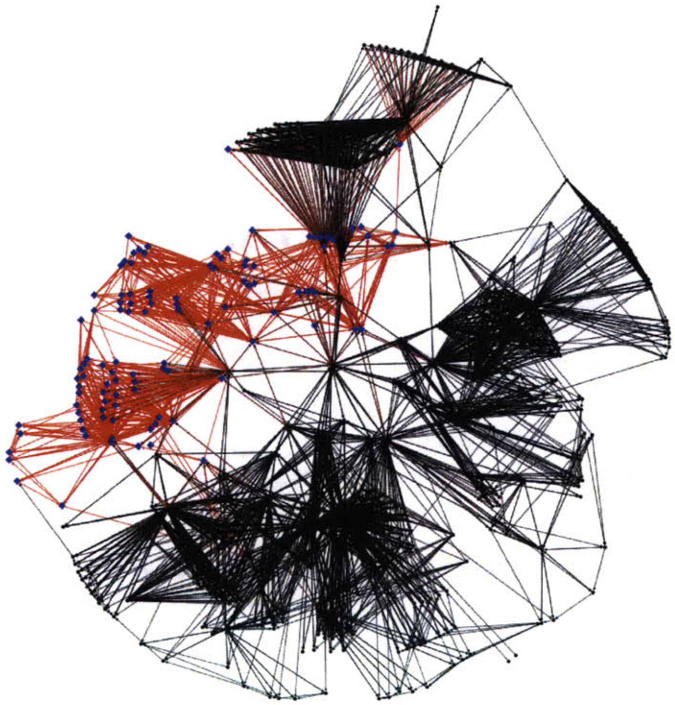


Figure 111 - Direct Impact of Genome Analysis Connecting Technology on V1 Surveillance Process

Much more work will be needed to improve the alignment and assembly algorithms such that they can reliably work at high throughput and on a wide range of clinical samples. Currently the analysis of each genome must be manually fine tuned with a surprising amount of tacit knowledge despite it being in a "software" and algorithms technology area. Some of the suggested improvements are described in the V2 surveillance process later in this chapter.

5.5 Process Analysis

A number of observations to improve the V1 process can be made based on the MDPM analysis thus far including specific mechanisms to deal with troubleshooting, technology change, transferability, and other process "ilities" such as modularity and scalability. These observations are described using the MDPM matrix and graph concepts to devise structural changes to the V1 process to improve its "big P" process performance across a number of system properties. Among the most important general observations of the V1 process are the following:

Connected development

The insertion of new technology and process change are essential capabilities of modern healthcare processes despite their complexity. It is important to recognize that any change or new technology any particular portion of the process can have unforeseen consequences elsewhere in the process outside the "visibility" of the individuals directly associated with the change or technology insertion. It is critical to provide this general "system" visibility to every part of the process to facilitate rapid improvement and minimal impact on the overall process. Examples of connected development include the molecular barcoding and DNA extraction technologies in this chapter, which were facilitated by design reviews including all other “nodes” in the MDPM that were impacted by the particular technology. Analytically this might be done by identifying the set of nodes in a MDPM associated with a particular change, such as DNA extraction in Figure 97, and treating them as a "cluster" to see whether the cluster is "well-connected" or not to the rest of the network.

Cross discipline, cross-group training

Beyond simply facilitating communication between individuals in the organizational domain, there is a need for actual training and education across process disciplines, organizations, and process activities. So for example it is very different for a sample tracking software designer to sit in a meeting with laboratory personnel discussing the features the software should have than for the software designer to actually spend some time in the laboratory performing some of the work. Similarly educating the molecular biologists on the analysis algorithms used can bring enormous rewards as the molecular biology might then be adjusted to significantly improve the algorithm performance and vice versa. In addition to cross discipline training, the particular ways in which groups perform their work (both explicit in terms of policies and tacit in terms of culture) are critical to the rapid communication and problem solving in a MDPM network. Despite the increasing trend towards knowledge specialization and "silos" efforts on the part of complex process owners to educate individuals about the other "silos" they interact with is a necessary activity.

Cross discipline, connected development, and organizational responsibility

In addition to training it is important to assign specific individuals to manage the interfaces between disciplines, technology insertions, process change, and organizations. All of the "connected technology" projects in this thesis to integrate the overall process were managed by the thesis author, however this may not be feasible in many complex processes. Yet, specific "ownership" is exceedingly important and could be accomplished by assigning "project managers" to manage each of the connected development projects and perhaps reporting to each other in a hierarchical structure such that efficient communication could still be maintained. The boundaries between disciplines likely require the same "ownership" such that education of individuals on either side of the boundary can be maintained and directed. And similarly ownership of the boundaries across organizations in the MDPM can also be exceedingly valuable. In some sense rather than rely on ad hoc individuals that happen to have a high betweenness or network centrality, these individuals could be "engineered" into the MDPM network to lie at the boundaries of disciplines, organizations, and process change projects.

Cross discipline, connected development, and organizational information review

A key process management task is to perform design reviews to ensure that process changes and cross discipline interfaces properly reflect the available information from potentially impacted nodes. For example, the DNA extraction can impact a number of other protocols and these should be carefully reviewed to ensure any changes in DNA extraction are properly reflected. The difference here is proactive review of protocols (and other information including policies) when implementing changes rather than assuming the changes "modular" or has little impact. Forcing reviews is an element of the cross discipline education, but also serves to ensure explicit information is connected in the information domain and is consistent. If this is not done complex processes may rely on ad hoc communications channels in the organizational domain to obtain information and propagate, change which can be very difficult in large complex processes. This "review" process is what was done in going from the "baseline" to the V1 process and the specific review documents, such as for molecular biology, become elements in the information domain connecting baseline protocols and new versions that needed to change for the V1 process.

Process maps and troubleshooting communication

A related process infrastructure item is the need for detailed process maps, perhaps even publishing the MDPM model within an organization such that individuals could rapidly navigate the process network to find answers or identify problems. This could greatly reduce the "search information" needed during troubleshooting or process change scenarios since the paths would be mapped. In combination with specific "owners" of cross discipline boundaries and other key individuals specifically identified to serve as "hubs" to facilitate communication and the rapid deployment of resources to “swarm” problems the MDPM model could be a very useful tool to ensure processes can be easily navigated.

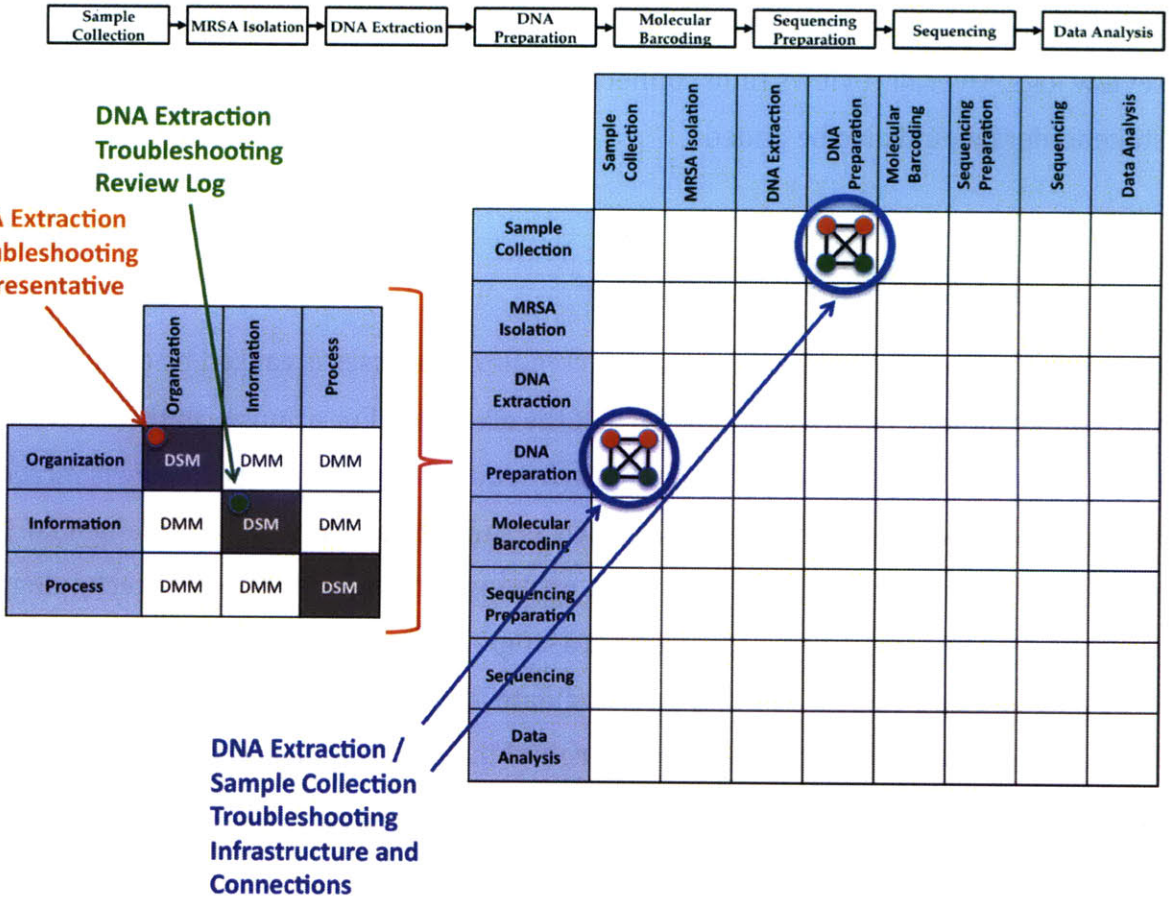
Systematic Overall Process Testing

A final observation from the development of the V1 process is that many complex processes rely on "passive" testing to identify problems or find process improvements. This means that because of the complexity of a process troubleshooting only happens when there are problems rather than purposely "testing" the process under controlled circumstances to identify failure modes, performance limits, and potential improvements. This is also true of technology insertions where a variety of conditions should likely be tested against the response of the system in keeping with the "connected development" ideas and to verify the overall system response to "local" changes. This would mean using many of the ideas from design of experiments including (Box and Liu 1999) to iteratively measure the response of the overall system to variables in particular activities and make improvements. However, this requires much of the process infrastructure described previously such as an overall process owner, a hierarchy of individuals at key interfaces, efficient communication, and the identification of key process variables.

5.5.1 Troubleshooting

One of the most important "nonstandard" tasks in complex processes is troubleshooting, which because of the large number of elements increases significantly in probability and time taken from process owners. The V1 process is an example of a complex process and in routine operation would exhibit many failure modes not yet seen in the initial demonstration in this thesis. However, it is possible to use the MDPM model to anticipate the need for significant troubleshooting and to build in infrastructure to facilitate communication and rapid resource deployment to resolve problems. In addition, the learning from these troubleshooting events should be captured as an explicit information element in the information domain. The V1 MDPM graph has a relatively long average path length (3.636) as given in Table 21 and also a correspondingly low network efficiency (0.307). The goal is to increase the network efficiency without resorting to excessive numbers of costly to maintain links. So for example it is important that a technician in the DNA extraction portion of the process be able to have a relatively short path length to genome analysis in order to troubleshoot any potential missing parts of the genome due to problems in DNA extraction. However, this does not need to be direct social connections, but rather some systematic means of providing additional paths and hubs to facilitate communication but also allow for systematic review.

One other important concept in troubleshooting is that the infrastructure to enable it should be in place before a troubleshooting event rather than a reaction to a problem. The suggested mechanism for doing this combines operational logs from chemical engineering best practices, lean process improvement, and the graph ideas presented in the MDPM model. The basic idea is to designate specific individuals that can serve as "hubs" that can reach nodes in the MDPM graph within 1-2 hops. These individuals are then connected to each other by weekly process review meetings where an overall process log built from "troubleshooting problems" from each of their neighborhoods are discussed and resolved. The process log provides for an explicit knowledge document describing the problems in each area, but because it is visible to the entire process network within 1 to 2 hops (going through the individuals acting as hubs) it serves to connect the entire process. In addition, the explicit assignment of individuals to serve as "hubs" or "troubleshooting representatives" also greatly facilitates the search of the network for information. These "troubleshooting representatives" might be the same ones identified to serve as cross discipline or cross organization integration owners or could be distinct individuals. Standard network clustering algorithms can readily be used to identify the "local" neighborhoods that the troubleshooting representatives should serve, or alternately a process designer could use the MDPM model to reorganize these clusters to further facilitate communication. This overall idea is shown in Figure 112, which shows the MDPM model, but with elements sorted such that the original macro process activities are grouped together. It is then possible to add a troubleshooting representative and a review log to each domain within the activity.


Sample Collection
MRSA Isolation
DNA Extraction
DNA Preparation
Molecular Barcoding
Sequencing Preparation
Sequencing
Data Analysis
DNA Extraction Troubleshooting Review Log
DNA Extraction Troubleshooting Representative
DNA Extraction / Sample Collection Troubleshooting Infrastructure and Connections

	Organization	Information	Process
Organization	DSM	DMM	DMM
Information	DMM	DSM	DMM
Process	DMM	DMM	DSM

Figure 112 - Use of MDPM Model to Build Troubleshooting Infrastructure and Connections

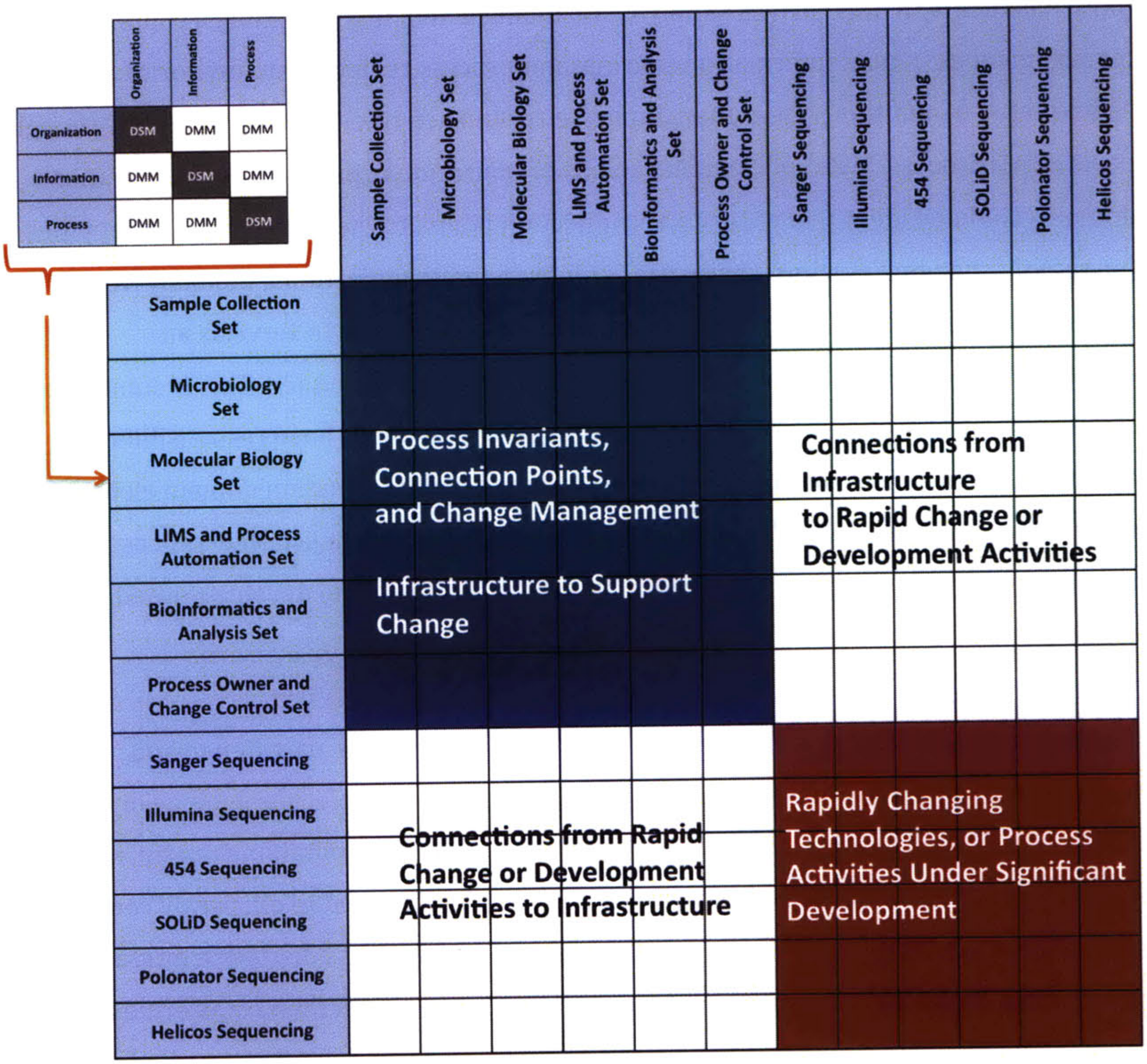
Each of the process representatives in each activity can then be formally connected to others, such as the sample collection example shown in Figure 112 such that they could rapidly identify problems between each of their activities and serve as "hubs" to connect other elements (in all 3 domains) to each other. The troubleshooting logs serve a similar purpose in formally capturing the problems, and it is important that they also capture the resolutions such that the organization can learn. The review logs also serve as an important communication tool allowing for asynchronous communication not through the organizational domain, such that anyone in the network can see issues as they come up. In addition, the review logs serve as a critical source of information for process improvement and change as they should describe things to improve but also insights into the performance of the process. Much of this was done in developing the V1 process but a V2 process would significantly expand on the designation of key individuals in each "cluster" (which will likely correspond to a process activity but may not) and also add explicit review logs. Overall reviews (and connections) with the process owner and key stakeholders would also be added.

5.5.2 Design for Change

One of the biggest difficulties in managing complex processes is caused by the differential rates of technology improvement in each of the activities. For example in the surveillance system designed in this thesis the rate of DNA sequencing technology change is incredibly rapid, yet some of the process activities such as sample collection have remained fairly stable for several decades. Strategies must be devised to take advantage of technology improvement in complex processes where a single activity change can have unforeseen consequences and costs to continuously evaluate the impact of a new technology on the overall process and perhaps re-optimize around the change can be cost prohibitive and infeasible. The amounts of "scaffolding" as described previously would have to significantly increase in order to accommodate the rapid evaluation and insertion of new technologies. This is still an issue for the surveillance process designed in this thesis since there are significant improvements on the horizon for existing sequencing technologies but also entirely new sequencing technologies.

One solution to this problem might be to find process "invariants" that do not change either because the technology is fairly mature or because they have purposely been designed that way. So for example the existing V1 surveillance process could be "partitioned" such that invariants were identified and grouped together such that they could be managed separately from the rapidly changing parts of the process. This might mean taking the sample collection, initial MRSA isolation, culture growth, sample storage, and other infrastructure activities such as supply-chain management and designing a common management structure, information sharing, and process review structure for these "invariants". In addition, rather than duplicate certain activities within the rapidly changing parts of the process, "sets" of elements across organizational, process, and information domains could be identified to serve as connection points to the corresponding activities in the rapidly changing portion. These sets (groups in the organizational domain) well connected to each other and to the process invariants by design to serve as an infrastructure on which rapidly changing process elements are added. This is similar to the concept of "troubleshooting representatives" where individuals have responsibility to communicate problems, but in this case organization and information domain elements are grouped together to provide "hubs" to which new processes connect. New processes are then brought in but required to connect to these hubs rather than develop new ad hoc connections.

Figure 113 illustrates this idea where the MDPM matrix could be partitioned and restructured to include "sets" around key disciplines such as molecular biology or activities such as sample collection. Each of these sets is a "mini" MDPM containing organizational, information, and process domain elements. The sets are designed to efficiently connect to each other using many of the ideas developed so far such as cross discipline interaction and process change and troubleshooting reviews. New processes can then be inserted as modules that need to connect to the process infrastructure. In the MDPM methodology the feasibility selection step using the MCNF algorithm was designed to filter out a number of alternatives but it is possible an organization would want to simultaneously evaluate processes or may need to operate multiple processes as the data from each may have different properties. But most importantly in the case where an initial process has been selected and is operating but a new more efficient process must be evaluated the organization must be able to insert the new process and if the technology is changing rapidly this will happen very frequently. The area in red in Figure 113 could thus contain all of the feasible design alternatives, such as the 454, SOLiD, Illumina, and other upcoming technologies.



	Organization	Information	Process
Organization	DSM	DMM	DMM
Information	DMM	DSM	DMM
Process	DMM	DMM	DSM

Sample Collection Set
Microbiology Set
Molecular Biology Set
LIMS and Process Automation Set
BioInformatics and Analysis Set
Process Owner and Change Control Set
Sanger Sequencing
Illumina Sequencing
454 Sequencing
SOLiD Sequencing
Polonator Sequencing
Helicos Sequencing
Process Invariants, Connection Points, and Change Management
Infrastructure to Support Change
Connections from Infrastructure to Rapid Change or Development Activities
Connections from Rapid Change or Development Activities to Infrastructure
Rapidly Changing Technologies, or Process Activities Under Significant Development

Figure 113 - MDPM Analysis to Identify Process Change Infrastructure and Invariants to Support Rapid Change in a Subset of the Process

This idea could also be used to develop a framework for "process options" where additional investment might be made in infrastructure or "scaffolding" to support anticipated technology change and ensure that the overall process was able to rapidly adapt and improve. Net present value analysis could straightforwardly be performed in terms of the time between technology insertion and the corresponding gain of additional "revenue" from faster and more efficient adoption the change. The "change infrastructure" may even be less than what would have existed if the overall process had not been rationalized around invariants and rapid change. But if the cost were higher this would be subtracted from the additional revenue gained to find whether the NPV is positive in which case the "option" is worth it. "Process options" could also be used to value the additional infrastructure associated with troubleshooting as described in the previous section where the cost of the infrastructure would be weighed against improvements in performance and reduced failures or downtime.

5.5.3 Other Process "ilities"

Other process properties can also be examined using the MDPM model such as scalability, transferability, flexibility, adaptability, and modularity. Many of these parties rely on efficient communication within the process both in the organizational and information domains. In addition these properties also rely on the identification of "clusters" within the MDPM matrix that are modular in nature. For example, one of the key process properties for the surveillance system is the ease with which it can be transferred to other locations. This transferability property is related to which nodes within the MDPM model are best connected (have the highest betweenness centrality) such that these nodes can serve as the interfaces to corresponding nodes in the target MDPM (the location where the process is to be transferred). The nodes with the highest betweenness centrality for the V1 process were identified previously in this chapter, but these were not purposely designed for ease of transfer. With a MDPM model of a process it would be possible to identify key nodes to serve as transfer connection points to the MDPM model of the target organization. These "transfer nodes" in the organizational domain are very similar to, and likely the same as, the "troubleshooting representatives" or the individuals responsible for change management in process "clusters". However, transfer will also require connections in the information domain such as ensuring that the target organization has incorporated not just the final activity protocols but also some of the design knowledge contained in process design documents such as optimizations, process reviews, troubleshooting logs, and other explicit knowledge.

Similarly, the modularity of a process could be improved by identifying the connections associated with a "cluster" to the rest of the process and rewiring these such that they went through a "hub" which ensured that the cluster remains efficiently connected but is accessed through a specified set of "channels". In some sense the ideas in the previous section of finding process invariants could be directly applied to redesigning a process to be significantly more modular with benefits to technology insertion and process improvement. Modularity is also related to the adaptability of the process to changing inputs and deliverables. For example, the surveillance system designed in this thesis could be used to monitor microbes other than MRSA and it is likely that the main changes that would need to be made are in the sample collection, microbe isolation, and DNA extraction "modules" such that the rest of the process would remain the same. The MDPM model could thus be used to engineer in the necessary modularity and adaptability by designing the process to contain "connected modules" that preserve efficient network communication within the module and across the process to other modules.

5.6 Potential Improvements

There are a significant number of improvements that should be made to the “V1” surveillance process based on the technical insight gained in running it, but also based on insight gained from modeling the process using the MDPM model. Six projects were identified that would lead to a “V2” improved process that addressed key issues found with the V1 process. In order to specify the technical and MDPM based integration “project sheets” detailing various elements of the project were made and are shown in appendix 1, based on the template in Table 23.

Project #	**Project Title**
Description	Background on the project and key objectives
Internal Stakeholders	Who are the key internal stakeholders that must be directly involved with the project? (within the MDPM network). Including any "change owners" or "troubleshooting representatives"
External Stakeholders	Who are the key external stakeholders that must be directly involved with the project?
Deliverables	What should the project accomplish?
Project Owner	Who within the MDPM network will be responsible for the project?
Process Domain Integration	What process domain elements will be potentially affected by the project? Where in the process domain does the project connect? What additional process domain activities are needed?
Organizational Domain Integration	In addition to key stakeholders what organizational domain elements must be connected to ensure the project is "visible" in a few hops to the rest of the MDPM network. What additional organizational domain elements and connections are needed?
Information Domain Integration	What information elements will be needed to ensure the project can be visible to the MDPM network and what reviews (i.e. protocols) must be performed of existing knowledge domain elements to ensure project is integrated?
Process Testing Needed	What tests must be performed on the overall process to assess the impact of variation from elements in the project and similarly what variables within the project elements are affected by variation from other process elements. (DOEs of expected variation in project process domain elements against response variables in the overall process.)

Table 23 - Process Improvement Project Template with MDPM Integration

The template in Table 23 was filled out for each of the six projects identified for a V2 process, and these are detailed in appendix 1. A key goal of specifying these projects was both to address technical improvements in a V2 surveillance process but also to show how the MDPM model might be used to help speed the development cycle for process improvement by "designing" projects from the start to integrate with the existing MDPM network such that their impact can be well managed and all parts of a complex process be aware of the potential impacts of the project (or technology insertion). Brief summaries of each of the projects are provided here, with details in appendix 1.

Project 1: MRSA Sample Diversity

One of the main goals of the surveillance system is to provide a true representation of the MRSA diversity that exists in hospital infections. The sampling used as an initial demonstration in this thesis was confined to a single outbreak in a single hospital, however a much broader sampling will be needed to support epidemiological analysis. The goal of this project is thus to build a much more diverse collection of MRSA samples representing multiple hospitals, regions, and case histories. This project would ideally reuse existing infrastructure for strain typing performed at local hospital, state health department, and national CDC levels. This collection would provide for a proper estimate of the "pan-genome" of MRSA or the sum total of all mutations seen in MRSA throughout the country. This would ideally be an ongoing process as MRSA will continue to evolve and a constantly updated sampling should be performed.

Project 2: DNA Extraction Robustness

As described previously in this thesis DNA extraction is a highly variable process for MRSA samples, and while it has been optimized for the particular set of samples used in this thesis, it is likely that the variation seen in a much larger set would require re-optimization of this process activity. The goal of this project would be to find a new optimized protocol that can cover the widest possible range of MRSA samples. It is likely that additional measurements may have to be taken such as carefully controlling cell growth, and potentially automating a number of the steps to remove any operator variability. In addition new technologies and processes should be explored as alternatives to the current protocol. This project would then perform a series of design of experiments runs to find a robust set of parameters that could be used as a general protocol. Because of its impact on the rest of the process the same "connected development" using change owners and other mechanisms will be needed to make sure the impact of the project is well managed on the rest of the surveillance process.

Project 3: Advanced Molecular Biology

One major limitation of the current "V1" surveillance process is the size of DNA fragments used in preparing the DNA for sequencing, which are in the 300-500 base pair range. However, in order to link genomic regions with high similarity and assemble the genomes out of the short reads (100bp long) it is necessary to construct much longer specialty DNA fragments that can span greater distances than 300-500 base pairs, ideally into the 4,000-10,000 and even longer but which can still be read using the available sequencing technology. Methods for doing this are so-called jumping libraries, but there may be others. This has significant technical difficulty especially given the high throughput nature of the surveillance process and will require tight integration with algorithms and other process elements.

Project 4: Configure Process for Rapid DNA Sequencing Technology Change

There is expected to be significant technology change in the DNA sequencing portion of the surveillance process and the V2 process should be configured to deal with this change. This project would identify the process invariants such that these could be managed as a group and efficient connections made to the new sequencing processes as they emerge rather than ad hoc connections for each process. In addition "connection points" would be identified within the MDPM matrix such that new technologies could be rapidly absorbed without having to "make" new connections. This might mean the formation of "clusters" around key disciplines such as molecular biology that could oversee related steps in each of the new processes (rather than have each create a new one). Similarly each new process would have explicit representatives for each area that were within 1-2 hop network reach of the new process elements to facilitate integration.

Project 5: Build Data Analysis Pipeline

Most of the data analysis performed in the V1 process was performed on prototype software with significant manual intervention and little optimization. To make the surveillance process truly useful a significant amount of work will have to go into building robust and automated data analysis pipelines that have been extensively validated and optimized for MRSA. These pipelines must be able to align data from these samples against the entire set of MRSA genomes to date as well as perform a verification of the data through assembly in order to find novel genes or horizontal gene transfer from other organisms that simple alignment would not find. The overall accuracy and specificity of the algorithms must be evaluated as a function of the data input to specify the degree of coverage (i.e. how much oversampling of the genome is needed, 30X, etc.) to provide sufficient accuracy. Most importantly the resulting data and analyses must be easy to use for epidemiological studies such that outbreaks or new strains can be readily identified from the data. Another requirement is that the data analysis must be entirely backwards compatible with existing surveillance methods such as PFGE, MLST, and 16s such that data can be integrated with existing databases and the transition to the new process can be greatly facilitated. Attention to this final user interface is critical to help speed adoption of the process. In addition, sample tracking must be provided such that original clinical histories can be associated with the final surveillance data. All of this represents a very significant amount of work and impacts many elements of the MDPM.

Project 6: Verify Biology

An essential part of the surveillance process must be to "close the loop" with the biology, which is to compare actual biological results (phenotype) with the DNA data (genotype) being produced. For example, measuring the Minimum Inhibitory Concentration (MIC) of antibiotics for MRSA strains found to have mutations, is important to develop an understanding of which mutations are important and which are not. Other biological tests are also needed such as growth rates, cell wall thickness, and perhaps virulence measures. Correlation of the phenotype with the genotype will help realize enormous value from the surveillance process as new diagnostics can then be developed based on these correlations.

This will require going back to samples stored in freezers as a product of the MRSA isolation portion of the process.

6 Case Study Results and Process Observations

This chapter examines the results from the initial set of MRSA samples run through the V1 process and derives policy recommendations based on the MDPM analysis performed in Chapter 5 and the biological results. Much of the epidemiology of MRSA is believed to be clonal in nature meaning that a few distinct MRSA strains dominate the populations found in healthcare settings. The strain designations in use today come from a large-scale typing effort by the CDC using pulsed field gel electrophoresis (PFGE) to characterize 957 oxacillin antibiotic resistant or susceptible *S. aureus* isolates chosen at random from the CDC strain collection as described in (McDougal, Steward et al. 2003). These isolates originally came from hospital outbreaks, food-borne disease, and community acquired infections and at the limited resolution of PFGE grouped into 8 major clusters of isolates as shown in Figure 114 where the restriction digest pattern of each major strain type is shown along with the CDC assigned name for that pulse field type (PFT).

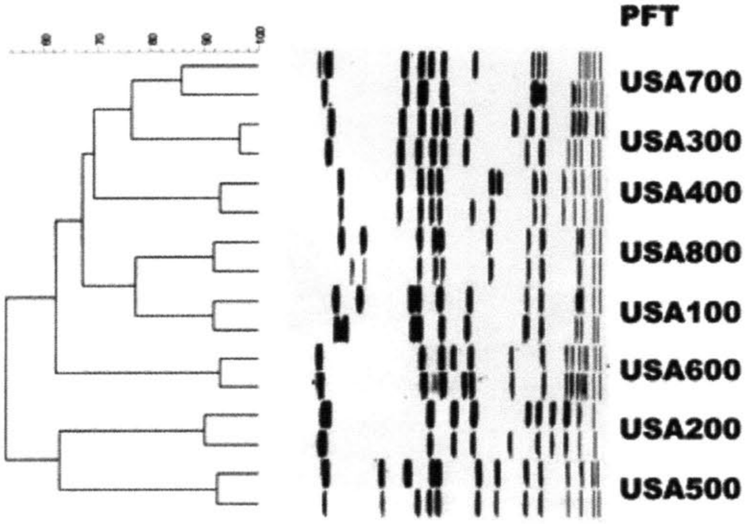


Figure 114 - Major MRSA Pulse Field Type (PFT) CDC Names (from McDougal et. al. 2003)

Each row in Figure 114 corresponds to a cluster of genomes that have a similar PFGE pattern. The dark bands are unique fragments of that genome cluster separated by size, where the fragments have been cut by a restriction enzyme *Sma*I that only cuts in the middle of the specific DNA sequence "CCCGGG". In comparison to the single nucleotide resolution of the surveillance process developed in this thesis, the smallest fragment that can be resolved in Figure 114 is >40,000 bases (the MRSA genome is ~2.8Mb in size). An enormous number of significant genetic changes can occur inside a 40,000 base fragment with even single nucleotide changes leading to very different phenotypic characteristics. Of the types classified by (McDougal, Steward et al. 2003), USA100 was the most prevalent at the time (2003), but the USA300 type associated with Community Acquired MRSA (CA-MRSA) is displacing others as the prevalent type as reported in (Gould 2006).

It is worth emphasizing that within a PFGE type "family" there can exist many specific strains with varying characteristics of virulence and antibiotic resistance further emphasizing the need for single nucleotide resolution. For example a recent study by (Kennedy, Otto et al. 2008) looked at genome variation using a probe based genome comparison method on 10 MRSA sample isolates finding that 8 of the 10 had very few single nucleotide polymorphisms (SNPs) between them suggesting they were closely related (likely having a common ancestor) and most importantly had a similar phenotype in terms of their virulence. However, 2 of the isolates had significantly reduced mortality in a mouse sepsis model despite also having relatively few changes (SNPs) from the others. As (Kennedy, Otto et al. 2008) describes, this suggests that even a few genetic changes can lead to significant phenotype changes despite all of the strains being classified as USA300 by PFGE. If small genetic changes can lead to significant genotype changes this suggests that the current classification system may be inadequate to describe the real diversity that exists and many of the surveillance policies which have to date assumed that relatively few "types" accounted for much of the MRSA population will have to be revisited. However, this cannot be known until a significant amount of data is gathered using a system such as the improved "V2" surveillance process described in chapter 5, or perhaps another technology. The first element of building a large-scale surveillance system will be to gather a significant amount of data in a "pilot" process that can answer what the true variability of the MRSA genome is at the nucleotide level. This chapter describes some of the initial results obtained using the "proof of principle" V1 process, describes what changes would need to be made for a "V2" process in further detail (including connections to the projects described in appendix 1), and describes a roadmap for further improvements in MRSA surveillance taking into account the expected rapid technology change.

6.1 MRSA Genome

The Methicillin Resistant Staphylococcus aureus (MRSA) genome consists of a single circular chromosome ranging from 2.8 to 2.9Mb in length, depending on the strain. The typical genome contains approximately 2,600 genes and as with most bacteria the majority of the genome represents coding sequence (~82%) as shown in Figure 115 for a genome of the Community Associated (CA-MRSA) USA300 sub type. MRSA strains typically also contain 1-3 circular plasmids of approximately 25Kb in length. Only a small set of genes distinguishes MRSA from normal Staphylococcus aureus, including the necessary antibiotic and virulence genes and many of these genes that are largely grouped in mobile genetic elements such as chromosomal cassettes, plasmids, and pathogenicity islands. These mobile elements can be exchanged between strains through horizontal gene transfer, and represent a key mechanism in the spread of resistance and virulence genes in *S. aureus* populations. Community associated MRSA (USA300) for example likely acquired many of its initial virulence adaptations through horizontal gene transfer although the population now appears to be clonal in nature as described in (Diep, Gill et al. 2006). A key difference in CA-MRSA is the existence of Panton-Valentine leukocidin, which is a cytotoxin that attacks leukocytes, the human white blood cells responsible for eliminating microbial infections. The CA-MRSA USA300 is a relatively new microbe not seen before the year 2000 (Carleton, Diep et al. 2004), yet has rapidly spread into the community (outside healthcare settings) as an apparent combination of hospital MRSA and "normal" S. aureus strains in the community. The rapid emergence of the USA300 strain suggests surveillance methods would need rapid identification (within weeks to months) of new strains such that screening and isolation procedures might be put in place. Note also that many of the techniques now in place such as PFGE can have difficulty differentiating even between clones with the same PFGE profile (such as USA300) as can be seen by the long black lines around the periphery of the genome in Figure 115 showing the cut sites for the PFGE fragments and the very long fragments (containing many genes and potential differences that result. A recent example of this limitation is described in (Larsen, Goering et al. 2009) who describe two phenotypically different clones with an identical USA300 PFGE profile and recommend using sequence based identification methods such as MLST although this method only looks for variation in a small fraction of the genome as shown in Figure 115.

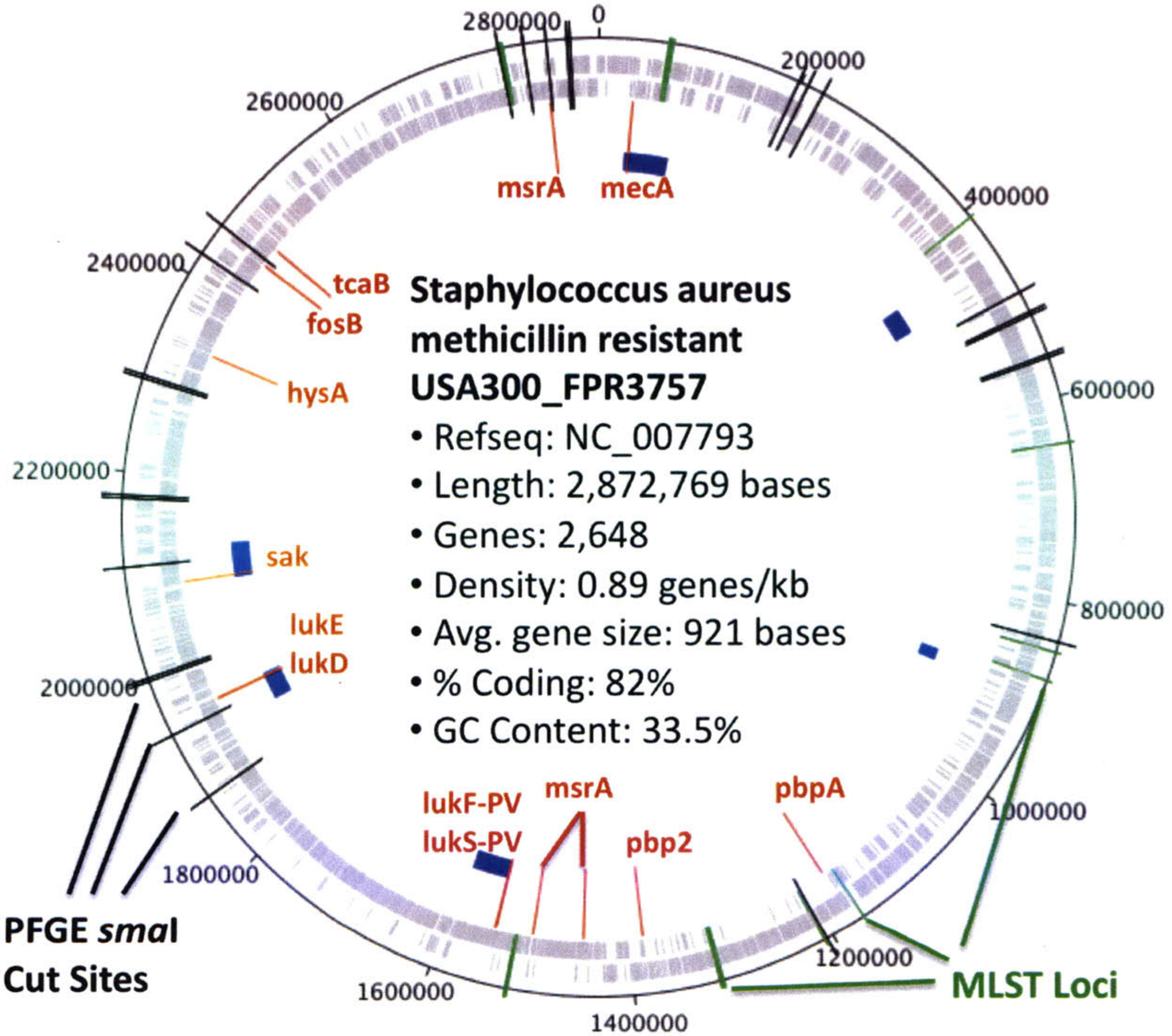


Figure 115 - MRSA Genome Showing General Characteristics, Selected Resistance and Virulence Genes in Red, Pathogenicity Islands in Blue, smaI Cut Sites in Black (corresponding to PFGE fragments), and MLST Sequence Typing Loci in Green. Note the limited amount of information provided by MLST and PFGE.

Fortunately the CA-MRSA USA300 type appears to be susceptible to antibiotics outside the beta-lactam group (i.e. methicillin, oxacillin) such as clindamycin, which inhibits protein synthesis. However, nosocomial MRSA continues to be resistant to nearly all common antibiotic classes except for Vancomycin, the current standard of treatment. The intense focus on resistance has accelerated the development and approval of new antibiotics, but many are still in early phases of development and FDA approval, and in the meantime strains of MRSA have already been found with resistance to Vancomycin. These Vancomycin Resistant Staphylococcus aureus (VRSA) strains appear to be derived from MRSA (ST5-MRSA-II) as reported in (Enright 2003) and gained their Vancomycin resistance genes through horizontal gene transfer from Vancomycin Resistant *Enterococci* (VRE) an entirely different microbial species that emerged around 1985 and is increasingly common in healthcare settings leading to higher likelihoods of gene transfer between VRE and MRSA as has already been seen in patients with simultaneous infections as described in (Chang, Sievert et al. 2003). The first VRSA cases were reported around 1996-1997 (Kuroda, Ohta et al. 2001; Hiramatsu 2004) and while the spread of VRSA appears to be fairly limited, it may be due to its lack of the necessary virulence adaptations found in CA-MRSA types such as USA300. The necessary set of genetic elements are "tested" and available in each of VRSA, MRSA, and CA-MRSA to make a true super microbe resistant to all common antibiotics including Vancomycin but also highly virulent and easily spread. Approximately 75% of *S. aureus* genomes appear to be highly conserved part of the "core" genome shared by all strains and containing essential functions such as metabolism as described in (Lindsay and Holden 2004). However, fully 25% of the genome appears contain accessory genes including mobile genetic elements such as pathogenicity islands, chromosomal cassettes, plasmids, and transposons that can transfer horizontally between strains. Most of the virulence and resistance genes are found on these mobile elements facilitating their transmission between microbes and facilitating a higher adaptability of the population as a whole rather than a single individual. Selected virulence and resistance genes are detailed in Table 24 and a subset of CA-MRSA USA300 specific ones also shown in Figure 115 for illustration. Numerous additional genes beyond those listed in Table 24 are involved in resistance and virulence mechanisms including the regulation of key proteins.

Gene	Function	Description
mecA	penicillin binding protein 2 prime	Low β lactam affinity peptidoglycan biosynthesis (cell wall construction, low affinity neutralizes methicillin)
pbpA	penicillin binding protein 1	peptidoglycan biosynthesis (cell wall construction, antibiotic target)
pbp2	penicillin binding protein 2	peptidoglycan biosynthesis (cell wall construction, antibiotic target)
msrA	methionine sulfoxide reductase A	cell wall stress stimulon (cell wall repair under antibiotic stress)
lukF-PV	Panton-Valentine leukocidin	leukotoxin (white blood cell attack)
lukS-PV	Panton-Valentine leukocidin	leukotoxin (white blood cell attack)
lukE	leukotoxin lukE	leukotoxin (white blood cell attack)
lukD	leukotoxin lukD	leukotoxin (white blood cell attack)
blaZ	beta-lactamase precursor	beta lactamase production element (degrades antibiotic)
sak	staphylokinase precursor	plasmin production element (disrupts host fibrin clots that localize infection)
ermA	rRNA methylase	23S rRNA modification (lowers macrolide antibiotic affinity for 23S rRNA allowing continued protein production. Neutralizes erythromycin and clindamycin)
tetM	tetracycline resistance protein	ribosome protection from translation inhibition (neutralizes tetracycline)
hysA	hyaluronate lyase precursor	hyaluronan cleavage (tissue invasion by degradation of host connective tissue)
fosB	fosfomycin resistance protein FosB	Cysteine addition to Fosfomycin (neutralizes antibiotic)
tcaB	TcaB protein	element of peptidoglycan terminal D-alanyl-D-alaynine lipid II linked precursor modification

Gene	Function	Description
		(prevents teicoplanin antibiotic binding to cell wall building blocks)
vanA	vancomycin/teicoplanin A-type resistance protein VanA	element of peptidoglycan terminal D-alanyl-D-alaynine lipid II linked precursor modification (prevents vancomycin antibiotic binding to cell wall building blocks)

Table 24 - Selected Virulence and Resistance Genes in S. aureus

The ability of microbes to rapidly exchange genetic information is central to their evolutionary success and the existence of numerous genetic elements encoding for many different phenotype adaptations insures that there will always be a large "pool" to draw from as illustrated in Table 24. Under appropriate conditions including antibiotic selection, simultaneous infection of different pathogens, and increased infection rates the probability of strains incorporating all of the resistance and virulence elements listed in Table 24 is possible as has already been seen by VRSA strains, although fortunately these have not yet included many of the virulence adaptations. A key goal of the surveillance system would be to first identify the complete set of all genetic elements in all known strains of antibiotic resistance Staphylococcus aureus including MRSA and VRSA. This catalog would provide enormous insight into which genetic elements constitute the "core" genome and which are the “accessory” genome not essential for the basic functioning of Staphylococcus aureus but containing much of the specialized virulence and resistance machinery. Building a database of all these genetic elements would also greatly facilitate the identification of any truly novel mutations found in new strains as the existing elements from the database could be "subtracted" from the new sequence to identify horizontal gene transfer or other novel changes. The Staphylococcus aureus genome is also not infinitely variable as certain essential genes and regulatory structures must exist for the microbe to function and mutations are changes in these regions are mostly lethal. Building a database across many strains of S. aureus would provide insight into what the "rules" for genome modification are perhaps allowing for the development of alternate control strategies not associated with antibiotics. The database would certainly be a very valuable resource to antibiotics development as new targets may be discovered or insight into modifications of existing antibiotics may be obtained. This database can be thought of as the sum total of observed changes in all S. aureus genomes, sometimes called the "pan-genome" as described in (Medini, Donati et al. 2005).

6.2 MRSA Genome Comparison

To emphasize this point, four S. aureus genome sequences were compared in Figure 116 with the mauve color showing the shared "core" genome, areas in white correspond to regions not aligned with the other genomes and unique that particular strain, and other colors representing regions shared by genomes sharing that same color. As can be seen in Figure 116 two strains TCH1516 and FPR3757 both classified as USA300 share most of their genome but also have significant unique regions not shared with the other. Note that the MRSA genome has extensive differences with the USA300 genomes, as would be expected since the MRSA genome lacks much of the virulence machinery in USA 300 strains such as Panton-Valentine leukocidins and in turn the USA300 genomes lack the additional antibiotic resistance genes found in "standard" nosocomial MRSA. These genomes were downloaded from the National Center for Biotechnology Information (NCBI) and aligned using Mauve[11] multiple alignment software. Also included in the alignment is a Methicillin Sensitive Staphylococcus aureus (MSSA) at the top, and the shared regions with MRSA can be seen light green although interestingly the MSSA genome also shares significant regions with the USA300 strains. The MSSA genome also contains a re-arrangement where a portion of its genome has "moved" to a different location and orientation (i.e. direction on the DNA strand) in the MRSA genome as shown in Figure 116. It is worth noting that each of these four genomes were sequenced using older generations of technology, the "Sanger Sequencing" described as a design space alternative in chapter 5. This technology is extremely expensive compared to the design goals for the proposed surveillance system,

[11] Mauve multiple genome alignment software from http://asap.ahabs.wisc.edu/software/mauve/

but it is capable of delivering fully detailed genomes without any gaps (i.e. missing sequence) partly due to its longer read length, but also due to the extensive manual curation performed on these genomes, which would not be feasible in the high throughput process being proposed. The cost of producing each of the genomes to the level of detail and annotation (i.e. verifying the identity and corresponding sequence of every gene) is several orders of magnitude above what would be needed for a surveillance system. The surveillance system proposed would need to deliver similar results but at a fraction of the cost.

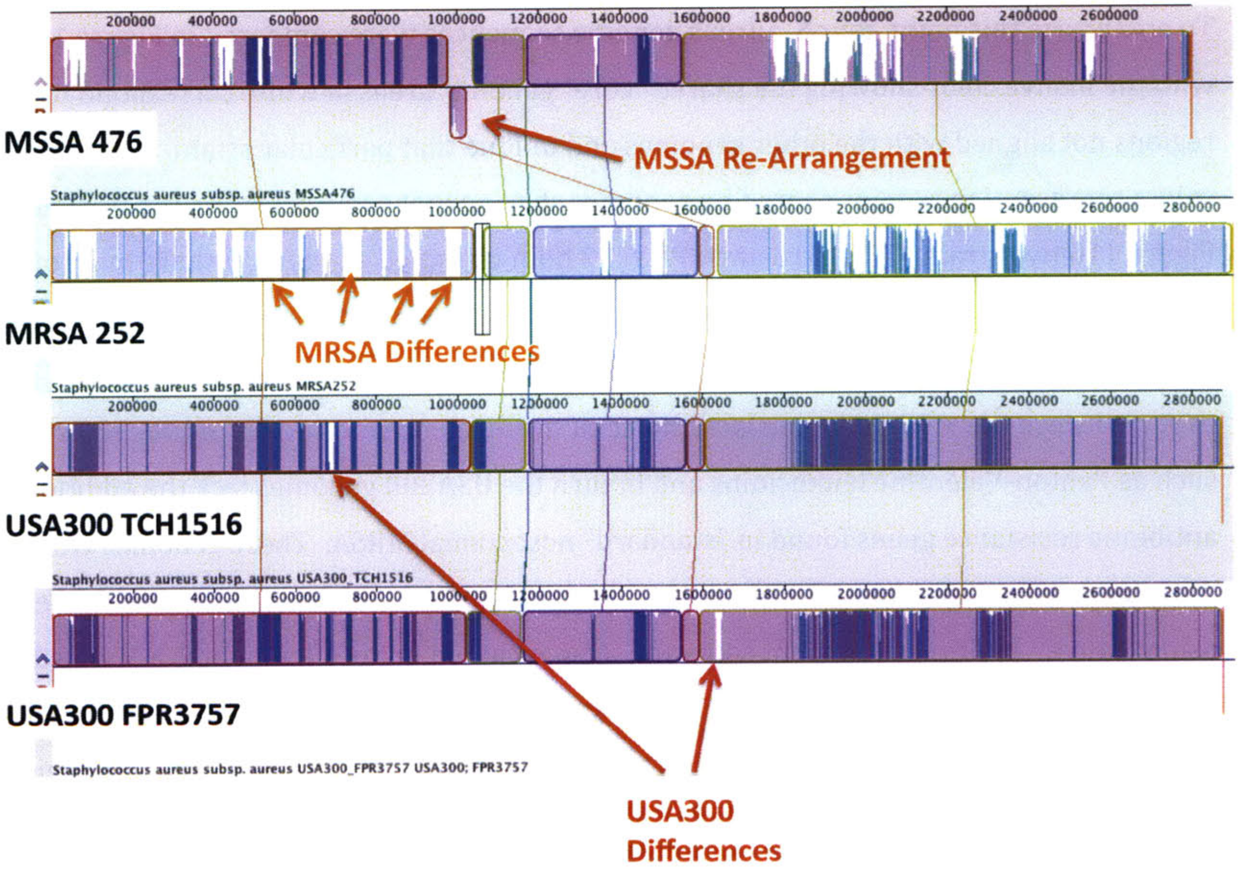


Figure 116 - Comparative Multiple Alignment of Fully Annotated MSSA, MRSA, and USA300 Genomes from NCBI. Note areas in white are unique to that genome, mauve areas are common to all genomes, and individual colors are common to each subset of genomes. Note also the difference in total length of genomes with MSSA being the shortest and MRSA the longest. Two USA300 genomes classified as identical by PFGE show significant differences.

6.3 Case Study Data Set

The case study sample set used in this thesis was obtained from a MRSA collection at the Channing Laboratory of Harvard Medical School consisting of anonymized MRSA patient cases from the Brigham and Women's Hospital in Boston of which 150 samples were selected at random from a year long time period. Samples were collected as pairs from both the nares and patient bacteremia to evaluate any genetic differences between colonizing and infecting MRSA strains. MRSA populations may be heterogeneous and only a few microbes may lead to infection with particular adaptations, or it may be that under certain selection conditions MRSA may evolve from a benign colonizing version to a more virulent infection version. Although MRSA populations are clonal in nature, a cross section of hospital cases is expected to show some variation, as some will be nosocomial clonal strains and other strains contracted outside the hospital or during a previous intervention. Sequencing the entire set would yield insight into the population dynamics especially in terms of correlating the observed phenotypes with genotypes. Differing clinical outcomes from the same MRSA genotype would suggest host or treatment differences are responsible for the outcomes. And if the opposite is observed, differing genotypes should correlate to varying clinical outcomes. It is important to note that all of these samples exhibited the exact same phenotype and all are classified as MRSA by standard clinical microbiology procedures where they showed growth in culture dishes containing methicillin antibiotic and by would thus all be classified as the "same" by normal procedures.

6.4 Case Study Process Attempts

All 150 samples were run through the V1 process described in chapter 5, although significant attrition resulted from the DNA extraction process steps as described in chapter 5. An initial set of 44 samples out of the 150 was chosen to test molecular barcoding methods and each of these samples were sequenced with 75base paired end reads to a coverage in excess of 100X the average MRSA genome size. 12 experiments were performed using the PCR-based barcoding method where each of the 44 samples were pooled and normalized using 2 different methods of DNA quantification and normalization (qPCR and picogreen) as well as mixing different pools in different sets of 12, 22, and 44. Both DNA measurement methods showed significant variability, and additional development will be required to develop a method for pool normalization. The data presented in the analysis sections that follow use the amalgamated data from each of these experiments to deliver >100X coverage of each sample. Following evaluation of the data and the development of an alternative barcoding method an additional set of experiments was performed using a subset of 12 (out of the 44 used for PCR-based barcoding tests) to verify the operation of the adapter-based barcoding process as described in chapter 5.

The analysis of these 12 samples is presented next as these were the best performing product of the V1 process following optimization of DNA extraction and barcoding. The V2 process would be fully capable of sequencing all 150 samples using optimized procedures based on the process learning from initial testing of the V1 process. The 12 samples sequenced in the final iteration of the V1 process have in excess of 100X average sequencing coverage of each sample genome in 75 base paired end reads. Note that only a single type of small insert library was made for the sequencing attempts, which requires only ~1-5ug of DNA, although significantly more DNA was specified out of the DNA extraction process (>20ug). The reasons for this are that additional library types will likely be required to fill in gaps in the assembly of the genomes, which would require significantly more DNA. In addition, the need for rework can be significant which can require going back to the sample DNA for successive attempts. This happened multiple times in the testing of the barcode methods and shows the need for an integrated process network where upstream and downstream steps can be coordinated.

The resulting data from the V1 process for each of the resulting 12 samples is summarized in Table 25 showing the number of "passing filter reads" or passing signal processing QC DNA molecules read from each sample, the number of these that were aligned to a reference genome, the resulting raw coverage (unaligned), the aligned coverage (against a reference genome), and the average error rate in each read. It is important to note that the strain type of these samples was not known in advance (i.e. whether they were MRSA, CA-MRSA USA300, or others), and in order to run the alignment process a single reference had to be chosen, which for convenience was an annotated USA300 type (genbank CP000730.1 corresponding to USA300_TCH1516 concatenated with plasmid pSR1 genbank AF167161.1) also shown in Figure 116. One issue with choosing a single reference is that only reads that are close matches to regions in the reference genome will be aligned and others will be excluded. This is why assembly methods are likely the long-term solution, or alternatively multiple alignments may be possible although this can be computationally intensive as well. As more genomes are added to the databases, multiple alignments become possible, although a key issue is that only a few of these genomes have been carefully annotated (i.e. the structure and identity of every gene has been verified) and using these greatly facilitates the finding of features on the genome.

In the case of the V1 process, most of the genomes aligned only matched the reference by about ~81% as shown in Table 25 although even despite these limitations, the variability between strains can be seen as one of the genomes fully matched 85% of the USA300 reference suggesting it is closer to the reference and may have some of the virulence adaptations in USA300. As part of a V2 process additional alignments should likely be performed should assembly not be available, such that a "best match" can be found within the known strains. As shown in Figure 116 the core genome of MRSA strains can be relatively small with significant variation across major strain types but also between supposedly identical ones such as USA300. And as mentioned previously as the database of genomes grows (i.e. the pan-genome) the ease with which improved alignments can be found also increases. Another thing to note in Table 25 is the variability in coverage despite careful attempts at equimolar pooling. The DNA quantification QC measurements are significantly variable and would have to be refined in a V2 process. This is a clear example of the need for connected development as the solution to the variability would require connections between the downstream analysis such as the one presented here and the laboratory optimization and design of the barcoding process. The lack of robustness of the V1 process is also apparent in the pooling error for one of the attempts where twice the volume was added as a technician mistake resulting in double coverage and opportunities for "troubleshooting representatives" to solve issues.

Sample	Passing Filter (PF) Reads	PF Reads Aligned	% PF Reads Aligned	PF HQ Bases	PF HQ Aligned Bases	Raw Coverage (X of Genome)	Aligned Coverage (X of Genome)	Avg. PF HQ Error Rate (per read)
1	4,624,214	3,745,828	81%	335,993,336	272,170,199	120	97	3.3%
2	5,135,974	4,130,735	80%	372,910,467	299,922,530	133	107	3.3%
3	5,809,068	4,739,816	82%	423,522,838	345,566,677	151	123	3.1%
4	5,108,604	4,133,021	81%	371,422,767	300,492,678	133	107	3.0%
5	6,887,632	5,576,740	81%	502,000,525	406,457,024	179	145	3.0%
6	10,574,792	8,669,612	82%	771,074,261	632,155,665	275	226	2.9%
7	4,843,010	3,938,483	81%	352,736,257	286,855,850	126	102	3.1%
8	7,411,530	5,937,507	80%	542,065,159	434,257,930	194	155	3.1%
9	3,549,962	2,860,380	81%	254,652,958	205,186,486	91	73	3.3%
10	2,638,954	2,093,008	79%	189,184,539	150,046,099	68	54	2.4%
11	4,794,760	4,071,487	85%	348,584,752	296,001,945	124	106	2.8%
12	2,610,898	2,000,266	77%	190,391,314	145,862,945	68	52	2.7%

Table 25 - Summary Data Results of Case Study Samples on V1 Process (alignment to USA300_TCH1516)

6.5 Variation Found

All of the samples contained the mecA gene as would be expected based on their phenotype. The mecA gene showed no modifications in any of the samples appearing in its established conformation. As has been discussed previously, the raw data produced by the process is noisy, containing errors towards the end of reads as the signal intensity decreases relative to the background and it is harder to discriminate the signal from a correct base. This leads to errors that can be overcome by statistical analysis using multiple observations to confirm their validity. As an example a region around the mecA gene in one of the samples is shown in Figure 117 with the underlying data and coverage for the region. The noise from sequencing errors can be seen, randomly distributed as colors through the reads, where grey represents the same sequence as the reference. To emphasize the statistical approach, a true SNP is also shown in Figure 117 in blue, where all raw data reads contain this difference from the reference making it unlikely that it is a random error and can be visually seen as a straight line connecting all reads. The initial alignment of the reads was performed using Maq as shown in Figure 117, but these had to be "collapsed" into a single "consensus" sequence where individual SNPs are statistically evaluated against random error using the package SAMTOOLS[12] developed by (Li, Handsaker et al. 2009). This final consensus sequence is then used as the sequence for each of the strains. Note that this approach has limitations in that only the sequence aligned to the reference will appear in the consensus sequence and is the reason both more references will be needed but also improved assembly algorithms that can piece together genomes without the need for alignment. Both of these issues are addressed by projects for the V2 process in Appendix 1.

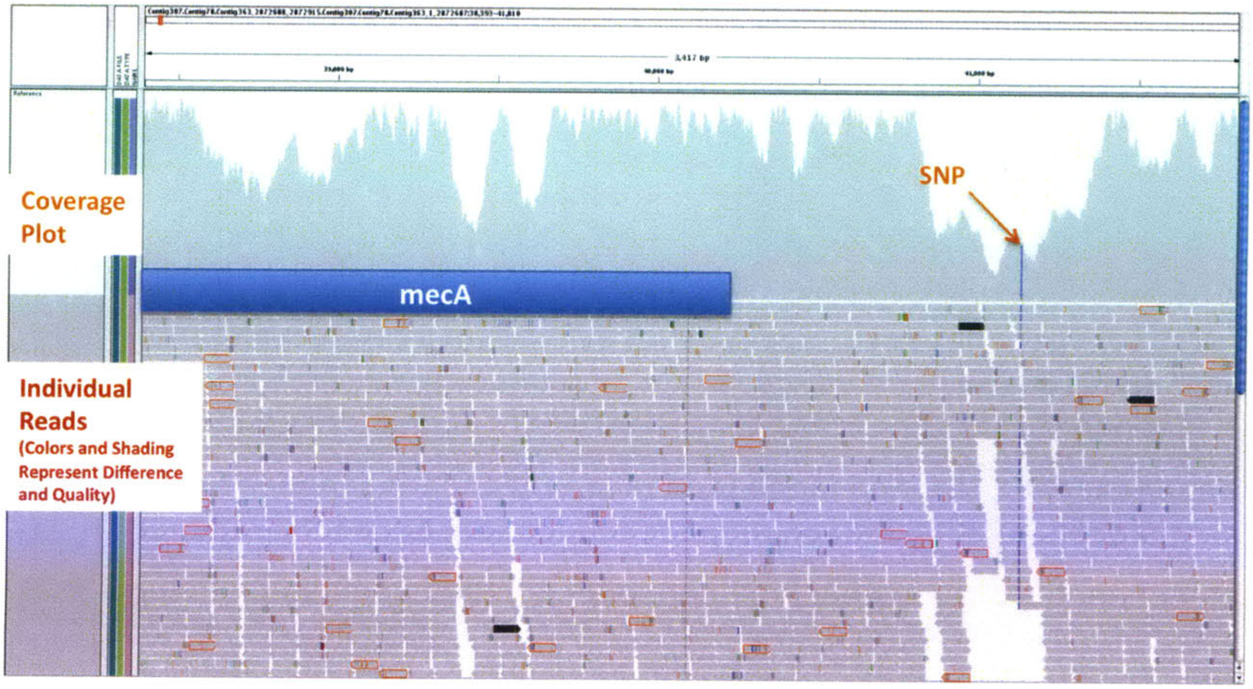


Figure 117 - Read Data Around mecA Gene for a Sample in the Case Study Showing Coverage and a SNP. Colors indicate differences from reference sequence, color shading indicates confidence of difference, and red boxes indicate pairing mismatches.

The mecA gene occurs as part of the Staphylococcus Chromosomal Cassette (SCCmec) mobile genetic element, and is not considered part of the core genome. However, the

[12] SAMTOOLS from http://samtools.sourceforge.net/

SCCmec does exist in certain patterns defined as SCCmec types (I-V), which include the mecA gene but also others such as the ccr complex of recombinase encoding genes that provide mobility to the SCC cassette. In Figure 118 the 12 case study sample sequences and the reference USA300 TCH1516 genome are aligned using Mauve showing the sequence in common as mauve, the unique sequence to a particular genome as white, and other colors showing sequence shared by a subset of genomes. The mecA gene is highlighted and its corresponding location in all genomes highlighted by a box. The ccrB gene is also highlighted and it is interesting to note that it is not found in all the samples suggesting sample 31195 is a different strain with a different SCCmec region and type as shown in Figure 118. Given the diversity of samples collected some variation between strains expected although the change in SCCmec cassette is substantial and additional comparison against other strains would be needed to determine the SCCmec type (or a novel type if no existing match can be found).

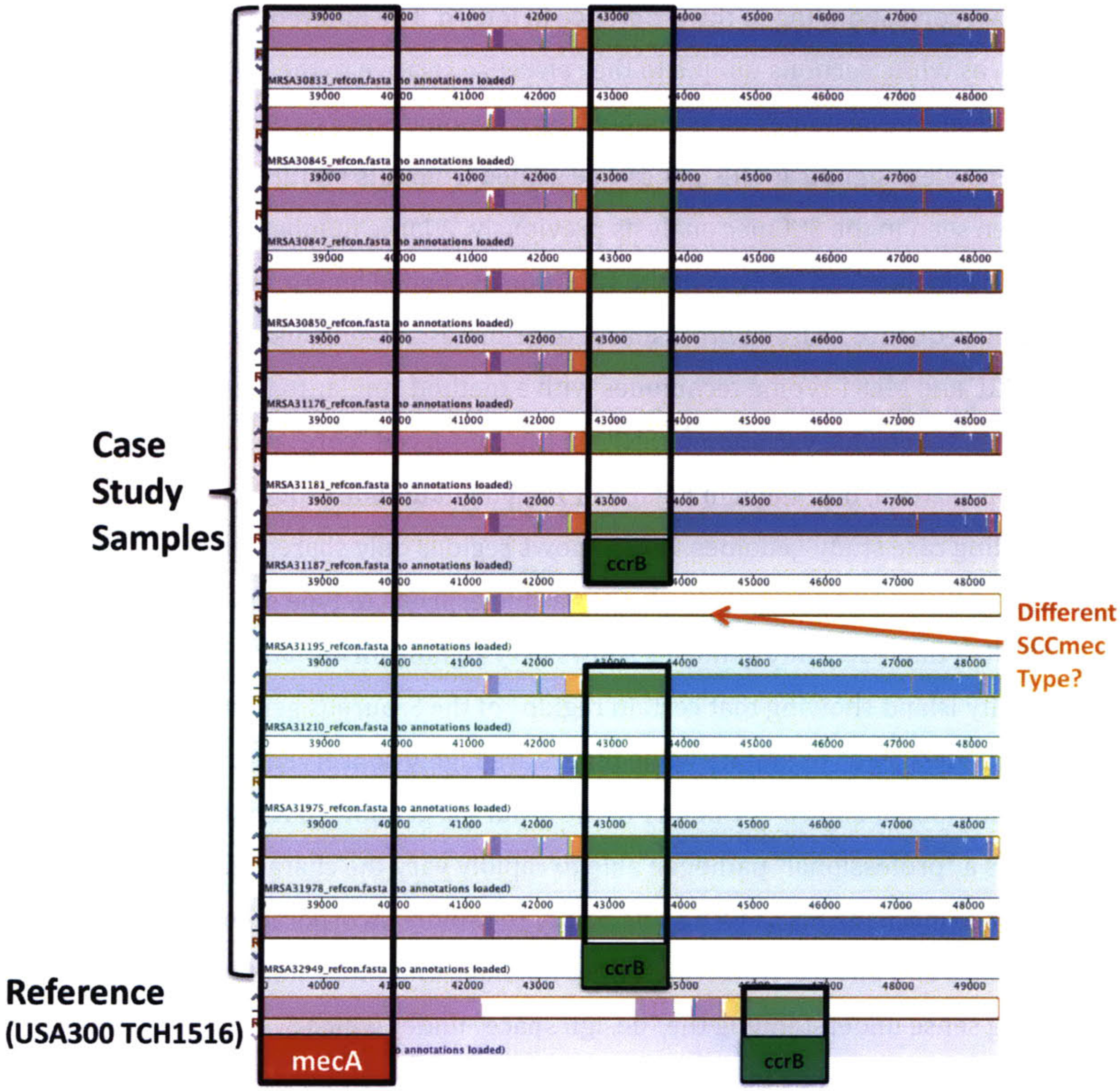


Figure 118 - Comparison of 12 Case Study Strains and Reference in the mecA region of the SCCmec element. The mecA gene is highlighted in red and the ccrB gene in green. Common regions to all genomes are in mauve, and other colors indicate shared regions to a subset of genomes.

A multiple alignment was performed across the entire set of case study genomes and the results are shown in Figure 119 with the standard Mauve color convention showing the majority of case study genome sequences are shared with each other and with the USA300 reference. However, it appears that many of the initial case study samples are substantially different than the USA300 reference, most likely hospital MRSA. This can be seen from the difference in aligned length between the reference in Figure 119 at the bottom and the case study samples. Reads from the case study samples that do not align to the reference will not be shown corresponding to the difference in length. The differences in the reference can be seen as white sections unique to the reference relative to the case study samples. Additional reference genomes or assembly based genome analysis would address this issue. Nevertheless significant insight can be gained from this multiple alignment as has already been seen in the SCCmec analysis previously. A large number of differences between the genomes can be seen in the variation in aligned length, areas unique to each genome, and sequence shared with a subset of genomes. It would likely be possible to replace PFGE and MLST typing techniques with a method similar to the analysis performed in this section to look for variation or SNPs in the conserved "core" region of the genome. For example, boxes a, b, c, and d in Figure 119 highlight unique sequence in each of the corresponding case study genomes. Box d shows regions only shared by each of the three genomes highlighted which could be used as a "fingerprint" to type strains. Finally, box f spanning all of the case study genomes shows very significant variation around a pathogenicity island showing that certain regions of the S. aureus genome are hyper variable and "designed" to mutate while others are highly conserved as described in (Lindsay and Holden 2004). The ability of the S. aureus genome to rapidly adapt is essential to its role as a "professional" pathogen able to rapidly vary the characteristics of virulence and resistance mechanisms as well as its ability to rapidly incorporate novel genes from other microbes and phages. However, it does not have unlimited ability to mutate and in an engineering sense understanding the "design space" under which S. aureus can change its genome and adapt would be invaluable to its control. An essential element of understanding the design space is to catalogue many more genomes as well as to build models for the gene interactions.

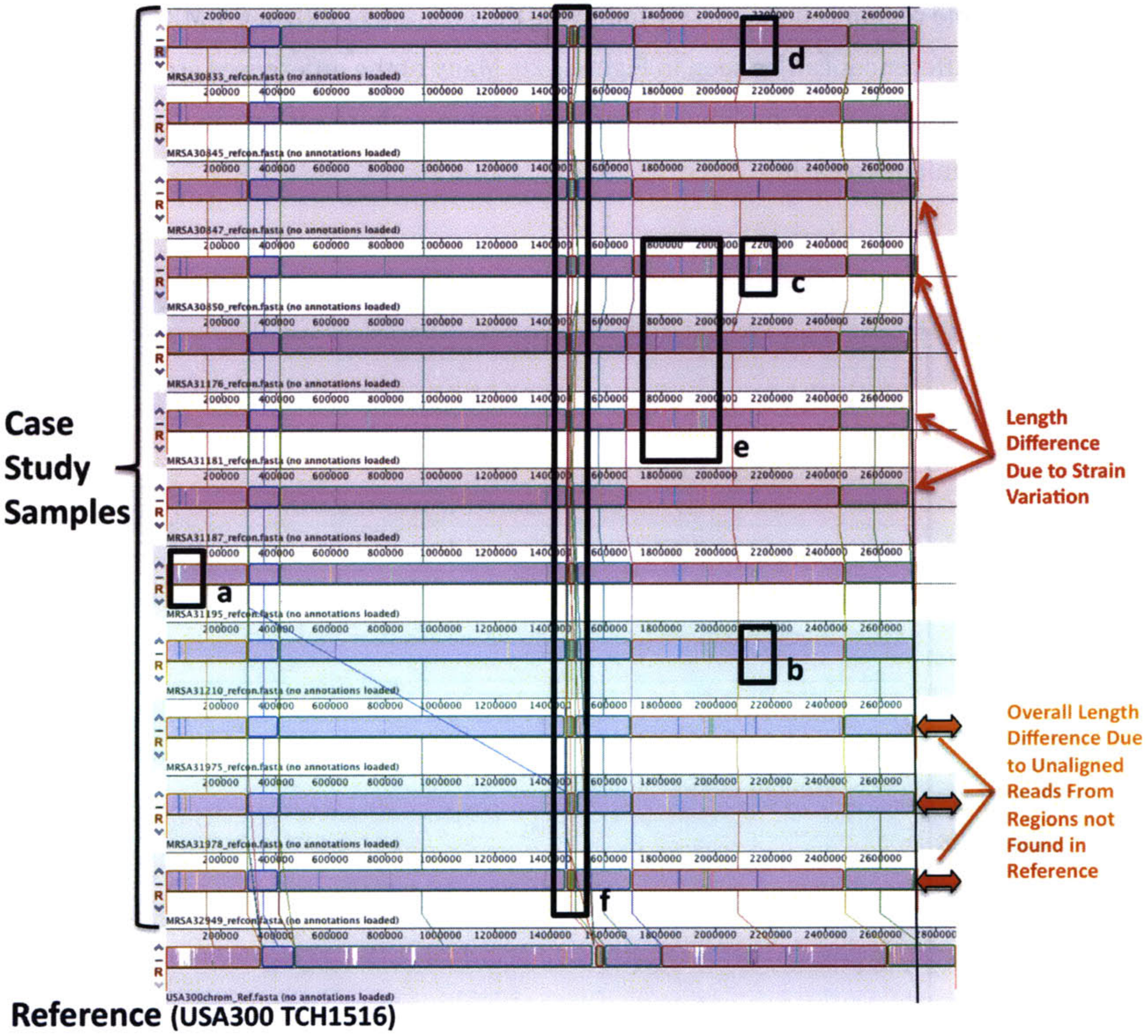

Case Study Samples
200000
400000
600000
800000
1000000
1200000
1400000
1600000
1800000
2000000
2200000
2400000
2600000
2800000
MRSA30833_refcon.fasta (no annotations loaded)
MRSA30845_refcon.fasta (no annotations loaded)
MRSA30847_refcon.fasta (no annotations loaded)
MRSA30850_refcon.fasta (no annotations loaded)
MRSA31176_refcon.fasta (no annotations loaded)
MRSA31181_refcon.fasta (no annotations loaded)
MRSA31187_refcon.fasta (no annotations loaded)
MRSA31195_refcon.fasta (no annotations loaded)
MRSA31210_refcon.fasta (no annotations loaded)
MRSA31975_refcon.fasta (no annotations loaded)
MRSA31978_refcon.fasta (no annotations loaded)
MRSA32949_refcon.fasta (no annotations loaded)
USA300chrom_Ref.fasta (no annotations loaded)
a
b
c
d
e
f
Length Difference Due to Strain Variation
Overall Length Difference Due to Unaligned Reads From Regions not Found in Reference
Reference (USA300 TCH1516)

Figure 119 - Comparison of 12 Case Study Samples and Reference Genome. Boxes a, b, c, d highlight unique sequence to that particular genome, box e highlights shared sequence to a subset of genomes, and box f highlights a hyper-variable region

The hyper variable region in box f in Figure 119 is magnified in Figure 120 showing only an 80,000 base region of the multiple alignment. The reference genome is a USA300 strain that contains additional virulence genes such as Panton-Valentine leukocidin (PVL) encoded by the lukF-PV and lukS-PV genes, which are also shown highlighted in red. Note that the reference genome region where the PVL genes exist is shown as white "unique" sequence only found in the reference, which suggests that the case study samples are not community associated MRSA (CA-MRSA). However, despite not having PVL genes, the hyper variable region is immediately following the location of PVL in CA-MRSA and significant differences can be seen in Figure 120. Many of the differences are shared as can be seen from the coloring implying at least two genomes share each region, although some regions are unique to each.

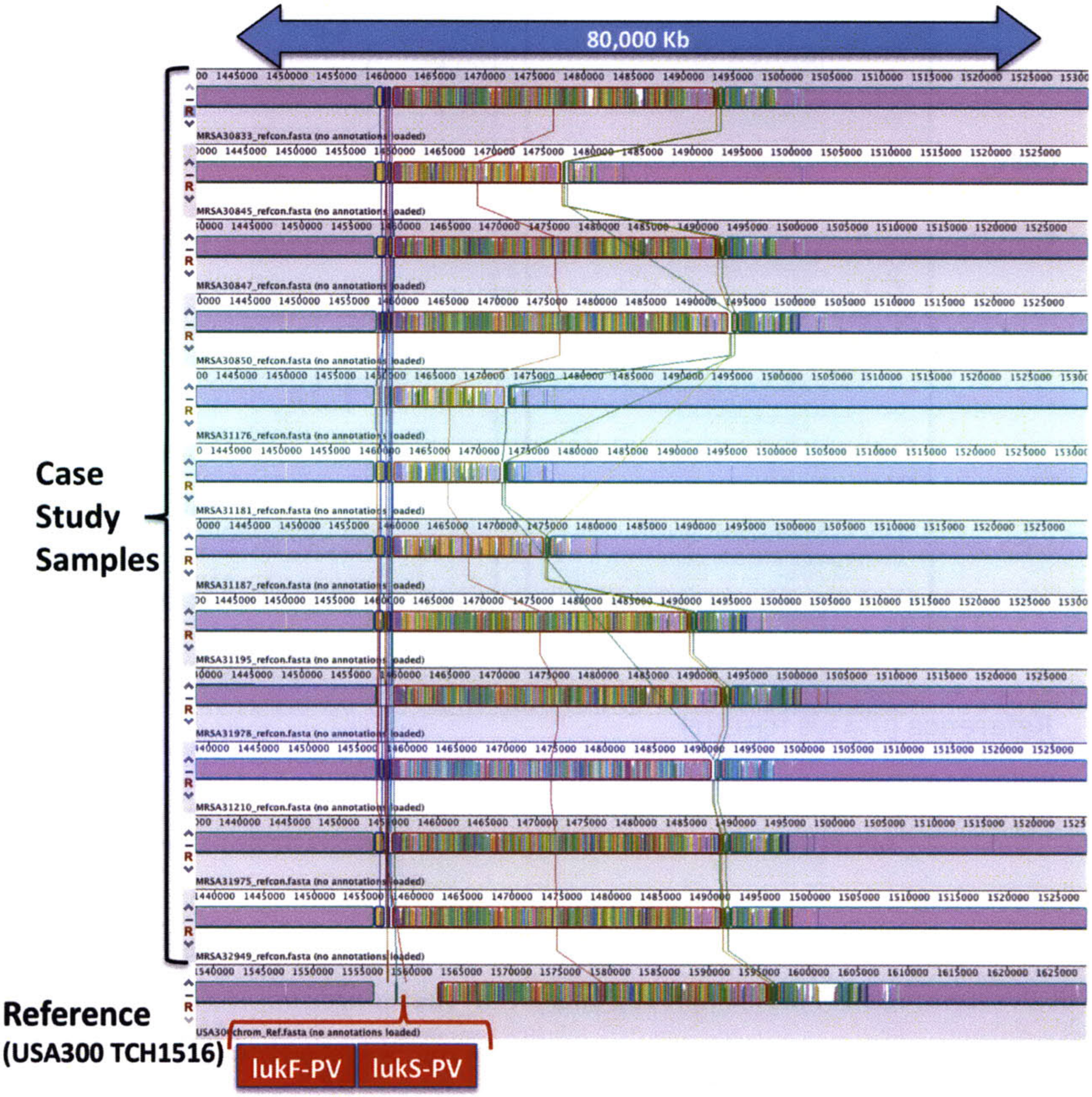

80,000 Kb
Case Study Samples
MRSA30833_refcon.fasta (no annotations loaded)
MRSA30845_refcon.fasta (no annotations loaded)
MRSA30847_refcon.fasta (no annotations loaded)
MRSA30850_refcon.fasta (no annotations loaded)
MRSA31176_refcon.fasta (no annotations loaded)
MRSA31181_refcon.fasta (no annotations loaded)
MRSA31187_refcon.fasta (no annotations loaded)
MRSA31195_refcon.fasta (no annotations loaded)
MRSA31978_refcon.fasta (no annotations loaded)
MRSA31210_refcon.fasta (no annotations loaded)
MRSA31975_refcon.fasta (no annotations loaded)
MRSA32949_refcon.fasta (no annotations loaded)
Reference
(USA300 TCH1516)
USA300chrom_Ref.fasta (no annotations loaded)
lukF-PV
lukS-PV

Figure 120 - Comparison of Case Study Genomes and Reference Around Hyper Variable Virulence Region (box f). The Panton-Valentine leukocidin genes in the reference USA300 genome are highlighted although these are not found in the samples.

Note that many sequence fragments are shared between the genomes, and they are significantly re-arranged in a different order leading to the multi-color bands observed and indicating that significant gene diversity exists but from a "building block" re-arrangement rather than novel unique sequence, although some unique sequence is also seen in Figure 120. The particular region shared between the reference and the case study samples in Figure 120 was searched for in the NCBI databases to determine if its function had been elucidated or annotated, but most of the proteins encoded by the region are classified as "hypothetical proteins" meaning the sequence is theoretically able to encode proteins but their function or existence has not yet been confirmed. The ability of S. aureus to contain hyper variable regions associated with virulence likely gives it a significant competitive advantage as it seems to have been designed for rapid "trial and error" experimentation where novel proteins (such as cytotoxins and others) can be produced and if successful that particular clone will have competitive advantage. These regions have been described as "gene nurseries" where novel changes can be tried. The successful clone will not be satisfied with any temporary success however, and has the machinery for its offspring to continue to experiment. Much like Toyota and other organizations that have adopted lean techniques, S. aureus owes its evolutionary success to continuous process improvement. And there is much to be learned from the way S. aureus manages change in that it has grouped "invariants" or the core genome into well controlled and efficient gene sets, but has also designated certain regions for rapid change and variation. It is interesting to compare the change management ideas described in Figure 113 and shown again here with the way S. aureus has also confined change to certain hyper variable regions as shown in Figure 120. In the case of S. aureus the main objective is rapid adaptation to changing conditions and because of its ability to quickly reproduce, many "experiments" are possible in a relatively short amount of time. In large scale complex engineering systems it is not possible to have many copies of the system to experiment, but it may be possible to confine change to certain smaller manageable sections such that many more iterations of these components or modules can be achieved in critical areas of system performance. S. aureus appears to have optimized its core genome such that it is robust, shares genes with other organisms so that it will not be susceptible to antibiotic attack (i.e. elements are not unique to S. aureus), and is highly efficient such that the virulence and resistance machinery can be allowed to be very flexible and adaptable as a system design objective. Engineering systems and processes might be designed in similar ways taking a page from millions of years of evolution.

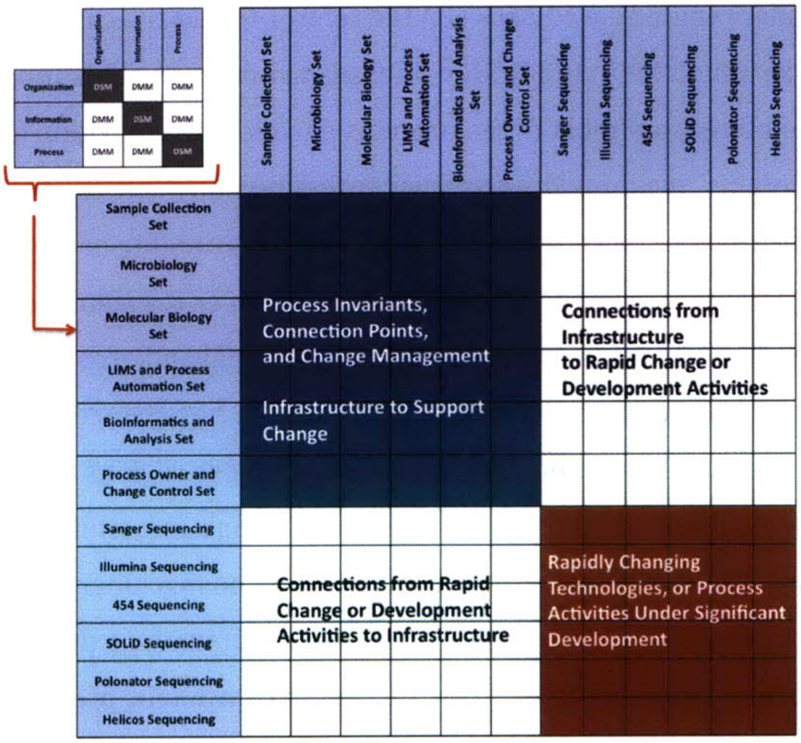

Organization
Information
Process
DSM
DMM
Sample Collection Set
Microbiology Set
Molecular Biology Set
LIMS and Process Automation Set
Bioinformatics and Analysis Set
Process Owner and Change Control Set
Sanger Sequencing
Illumina Sequencing
454 Sequencing
SOLiD Sequencing
Polonator Sequencing
Helicos Sequencing
Process Invariants, Connection Points, and Change Management
Infrastructure to Support Change
Connections from Infrastructure to Rapid Change or Development Activities
Connections from Rapid Change or Development Activities to Infrastructure
Rapidly Changing Technologies, or Process Activities Under Significant Development

Figure 113 - MDPM Analysis to Identify Process Change Infrastructure and Invariants to Support Rapid Change in a Subset of the Process

6.6 Implications for V2 Surveillance Process

The significant variation found in the case study samples and that reported elsewhere (Monecke, Berger-Bachi et al. 2007; Mwangi, Wu et al. 2007; Kennedy, Otto et al. 2008) support the urgent need for a whole genome high throughput surveillance system. However, it is clear the V1 version designed as a proof of principle requires significant improvement into a V2 version before it could be used as a "production process". This follows the iterative development of engineering processes where an initial conceptual design or model is formulated perhaps based on initial observations, this initial design is tested to produce actual experimental data and results on its performance, and this information is then applied to develop an improved version and so on. This follows ideas for statistical process improvement described in (Box 1999). In this case, the initial conceptual V1 process design (models) followed the case study testing (data) as conceptually shown in Figure 121.

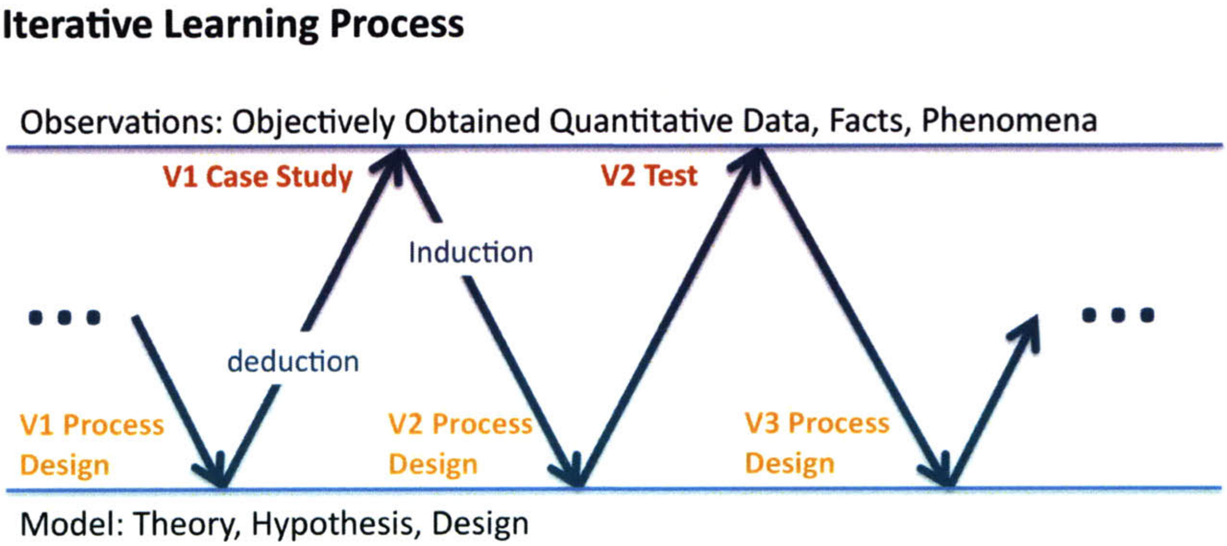

Iterative Learning Process
Observations: Objectively Obtained Quantitative Data, Facts, Phenomena
V1 Case Study
V2 Test
Induction
deduction
V1 Process Design
V2 Process Design
V3 Process Design
Model: Theory, Hypothesis, Design

Figure 121 –Iterative Learning Process for V1 Surveillance Process and Case Study (adapted from Box 1999, Magee)

The lessons learned from the case study can be applied to refine the potential improvement projects described in chapter 5 to produce a V2 process. The cycle would of course continue, incorporating additional improvements in future versions. There are clear areas for improvement based on the case study results and these are detailed in the next section describing the improvement projects needed, which are also detailed using the MDPM model to ensure these projects will be well integrated into the process. However, the surveillance process will also need to integrate with the healthcare system and the external stakeholders such as antibiotic researchers, CDC, and others who will need to interact with the process and data produced. An extension of the ESM model to include the MDPM model can be used to explicitly map the interactions with these external stakeholders such that the process is well integrated with their needs and conversely receives input from these external sources.

6.6.1 Projects

The following set of projects necessary for the development of a V2 process were described in terms of the MDPM model and process observations in chapter 5, but they are described in more detail here based on the biological observations in the case study. The full detail of each project is given in appendix 1.

Project 1: MRSA Sample Diversity

Having a larger database of sequenced MRSA genomes allows for improved alignment of sequences as well as better identification of genetic differences between strains. As was seen in the case study examples, variation between MRSAs and the likelihood that hospitals will see multiple strains as patients can travel between hospitals or may be infected with CA-MRSA, makes it essential that a much larger set of MRSA genomes be sequenced above the less than 20 available in public databases. A larger set would also allow for the tracking of geographical diversity, as well as an improved understanding of the degree of variation that can exist across the entire MRSA genome, which will be important for both diagnostics and antibiotics manufacturers. This project would necessarily have to occur before any routine surveillance to test the majority of samples and conditions within a V2 process to ensure they can be accurately identified and the process works for all sample types. For example, the DNA extraction conditions would need to be worked out for all samples as well as analysis tools to cope with the classification of many closely related organisms. A potential source is CDC strain collection or other comprehensive set of MRSA samples which contained the expected diversity in order to test the process. In addition it will be valuable to collect samples from international sources in order to track the variation outside the US, which through air travel could easily spread. Additionally sample collections such as the one used in this thesis that compare MRSA samples that colonize individuals versus those that actually produce the infection will be valuable. Finally, time series sample collections for MRSAs exposed to multiple antibiotics would be extremely valuable to identify the evolutionary mechanisms of MRSA.

Project 2: DNA Extraction Robustness

Much more work will be needed to optimize the DNA extraction conditions for a wide variety of samples, but it may be possible to use genetic information to correlate phenotype changes (i.e. a thicker cell wall) with genotypic changes (modified genes or regulation). Ensuring sufficient DNA of high quality is obtained is critical for the additional V2 processes needed to support assembly based surveillance processes that do not rely only on alignment to known sequenced genomes, particularly in terms of finding novel mutations. New technologies are available to do this such as the DNA extraction of MRSA in a microfluidic chip as described in (Mahalanabis, Al-Muayad et al. 2009) a type of approach which could completely transform the way current DNA extraction is performed and ensuring robustness against a variety of samples.

Project 3: Advanced Molecular Biology

One of the main difficulties in assembling de novo genomes is the need to span long distances (>2,000 bases) with molecules of DNA that are read at each end. The data in the V1 process consisted of a 75-200 base molecule size as shown in Figure 122, which is not sufficient to span certain regions of the genome that have a high similarity. This prevents the logical ordering of DNA sequence such that only small fragments (<10,000 bases) can be constructed from the reads. Developing the connecting technology to build high throughput "libraries" of sequencing technology compatible DNA molecules that were originally separated on the genome by >2,000 bases is a key technical challenge. Achieving larger separation sizes is technically difficult due to problems in manipulating larger DNA molecules and the need to convert these larger DNA molecules into smaller molecules (i.e. <300 bases) that delete the intervening DNA and only retain the ends of the original molecule (which were separated by a much longer distance i.e. >2,000). Complex molecular biology development will be needed. This is a critical "connecting technology" for the V2 process because it will involve molecular biology, algorithm development, and DNA extraction. The importance of this will be seen in project 5 describing the algorithms needed.

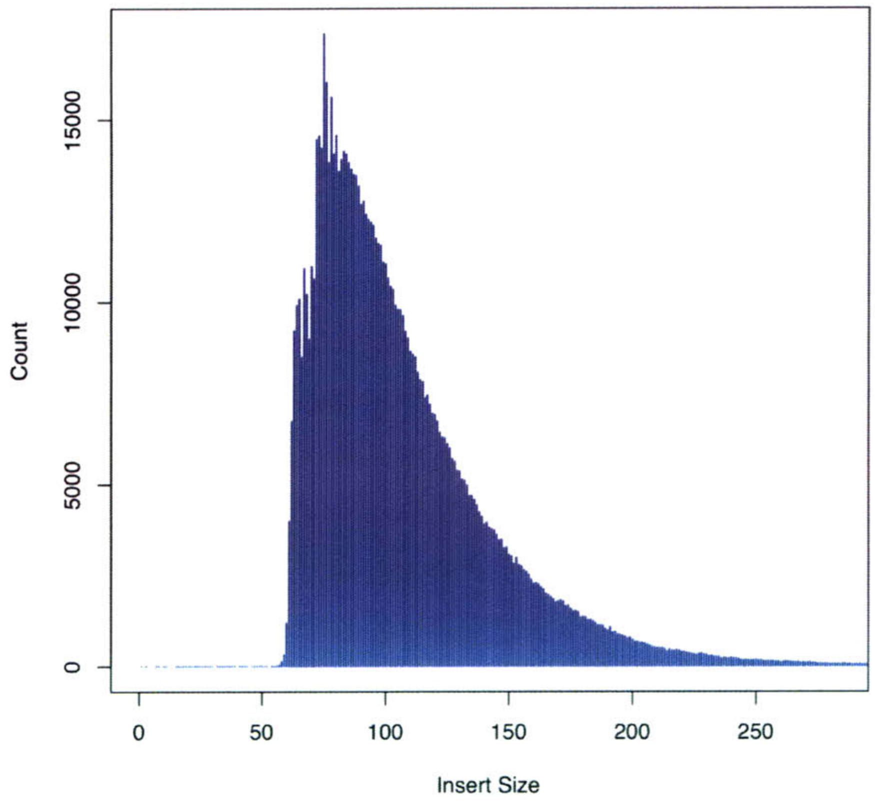


Figure 122 - DNA Read Separation in V1 Process (distance between 75 base reads on a single molecule)

Project 4: Configure Process for Rapid DNA Sequencing Technology Change

While a single sequencing technology was selected for the initial case study using the MCNF selection process, it is likely that a V2 process could use multiple technologies each with their own particular benefit. This places greater emphasis on the ability of the process to incorporate change much like S. aureus confines change to certain "hyper variable" regions. For example, a hierarchical process might be developed where samples are pooled together initially under the expectation that they are similar using the Illumina process and only if significant variation is seen in this set would more expensive sequencing processes be deployed such as 454 or an Illumina hybrid method that captured longer sequences. It will be important that technology change in sequencing occur as connected development to projects 3 and 5 especially.

Project 5: Build Data Analysis Pipeline

Perhaps the most important element of the V2 surveillance process will be the data analysis software and infrastructure since much of the V1 process relied on prototype tools or modify tools not primarily intended for microbial analysis. The many algorithms needed to translate raw data into alignments with high confidence SNPs or complete genome assemblies still require significant manual intervention and tuning in the V1 process. More importantly the analysis of the data in the V1 process is still "raw" since even if all of the SNPs are correct or the genome assembly is complete there is still no information about the function of the encoded genes or their correlation to phenotypes. Thus, a fully integrated V2 data analysis pipeline would not only process all the raw sequence data in automated ways with no need for manual tuning but would also deliver interpretation of the results such as classification of SCCmec types or correlation and genealogy with existing MRSA strains. Backwards compatibility would also be needed such that users could verify results using traditional techniques such as MLST and PFGE.

To highlight the additional development needed in the data analysis pipeline but also the associated connected development the 12 case study samples were run through one of the current state-of-the-art assembly programs called Velvet[13] developed by (Zerbino and Birney 2008). The results are shown at the top of Figure 123, where the resulting genome from velvet is compared to a reference MRSA genome. The highly fragmented assembly is evident from the many individual fragments that align in a disordered fashion. This is due to the lack of longer insert libraries such as those that would result from project 3, which could link all of these fragments into their logical order. The fragments from the raw velvet assembly were then re-ordered using the reference to help identify their correct location on the genome and the result is shown in the bottom of Figure 123, where it can also be seen that the overall length of the genomes are similar. The assembly method recovers the majority of the unaligned reads, which were discarded in the V1 process as shown in Figure 119. The re-ordering of fragments based on a reference was done here to illustrate the value in assembly based methods but would not be recommended in practice as it would

[13] Velvet Assembler: http://www.ebi.ac.uk/~zerbino/velvet/

encounter similar limitation to alignment based feature finding in that only things known a priori are found, but not potentially novel re-arrangements or structures. Improved similar methods combined with the advanced molecular biology in project 3 would enable the routine production of high-quality assembled genomes such that novel mutations could be found rather than differences to a reference. This is especially important for the identification of horizontal gene transfer between organisms such as the exceedingly dangerous VRSA strains that combine MRSA with the Enterococcus VRE vancomycin resistance.

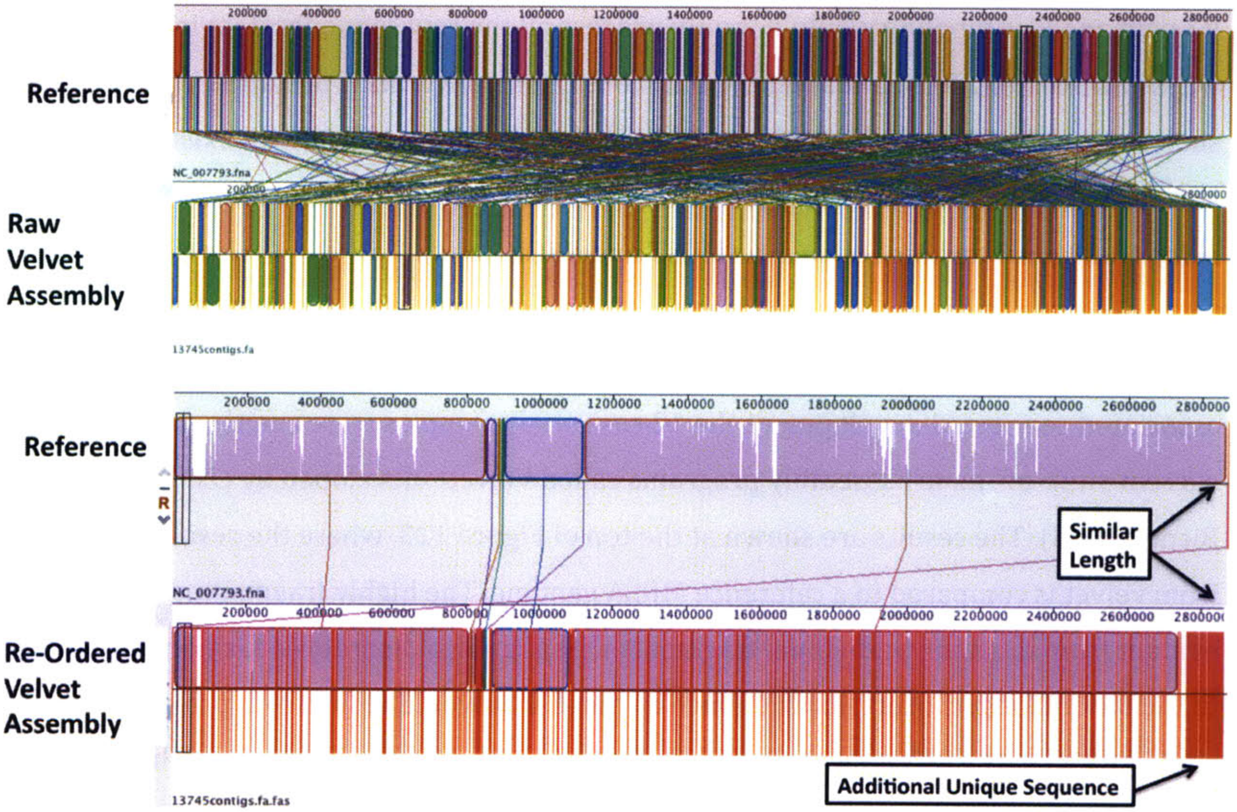


Figure 123 - Velvet Assembly Alignment Showing Fragmentation Due to Lack of Linking Information (Top) and Re-Ordering Based on Reference Order (Bottom) Red Bars Correspond to Boundaries of Original Contigs (fragments) in the Raw Assembly.

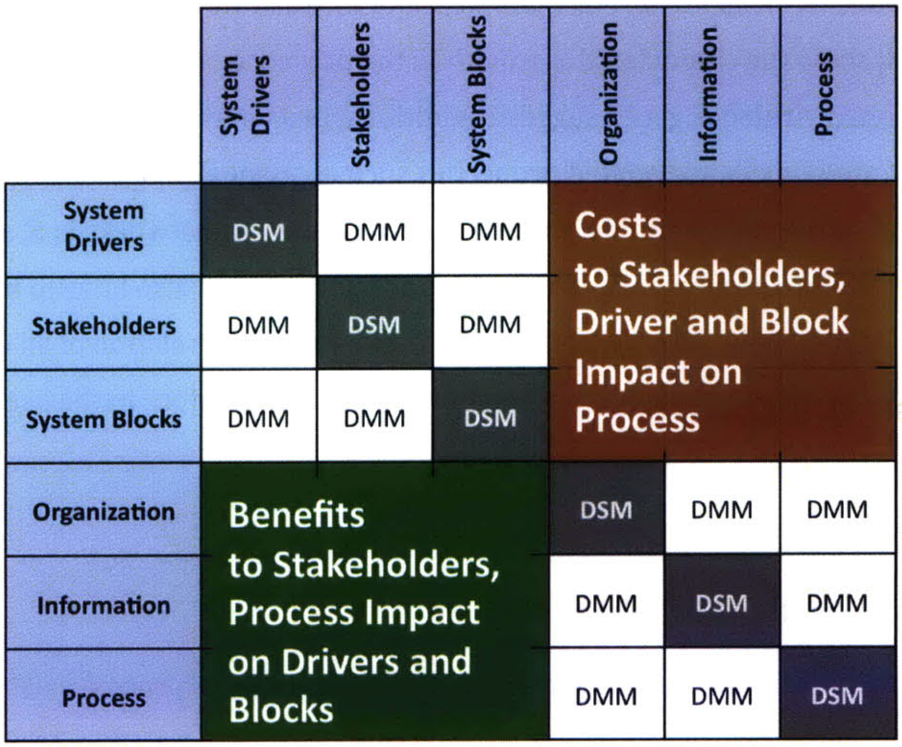
But even with high-quality complete sequences, such as those in Figure 116, this data still says very little about any potential correlations with phenotype or individual gene function and also do not describe the relationships between sequenced organisms in terms of their genealogy. A significant additional layer of software would be needed to transform this data into useful epidemiological and clinical information that could be acted upon by hospitals and physicians. This would require the involvement of the external stakeholders associated with this project as detailed in appendix 1 such that this additional software layer could be well integrated internally through the MDPM but also externally to the stakeholders.

Project 6: Verify Biology

Based on the variation seen, additional phenotype tests will likely be required to verify biology encoded in the genome. For example, all of the case study strains contain the mecA gene and will show the characteristic growth in the presence of antibiotic, however the wide variation in virulence genes suggests additional tests may be required to determine the extent of other virulence related properties such as tissue damage, colonization potential, and others. Deeper understanding of the regulatory networks in S. aureus will also help determine the effect of genome changes but these will also have to be verified in the lab such as metabolic activity, growth rates, and others. Correlation of all of these phenotype properties to genotype information will be a key feature in the usefulness of a V2 process.

6.6.2 Healthcare System

Beyond the technical challenges to developing a V2 surveillance process exist challenges in integrating the surveillance process with the larger health care system in terms of its ownership, funding, locations, and perceived benefit. Because the currently feasible surveillance process (even in an improved V2 configuration) spans many organizations, disciplines, and process elements it is unlikely to spontaneously arise and would likely require significant effort to initially develop it and continue to operate it. Securing the support of external stakeholders through a detailed MDPM analysis would be an essential additional "project" to make sure the necessary resources were available and any blocking issues addressed. This might be done as a "cost-benefit" analysis where the links between external stakeholders in the process were mapped to ensure a net positive benefit to stakeholders and sufficient resources to build and operate the process. This is conceptually shown in Figure 124 where the Engineering Systems Matrix (ESM) has been modified to include the MDPM but also the system drivers, stakeholders, and a new category called "system blocks" associated with key external elements that may be blocking issues such as regulation, public perception, and others. The area highlighted in green represents the benefits of the process to stakeholders but also any impact from the process to system drivers and system blocks. The area in red represents the cost to stakeholders of the process, but also the impact of system drivers and system blocks on the process.



	System Drivers	Stakeholders	System Blocks	Organization	Information	Process
System Drivers	DSM	DMM	DMM	Costs to Stakeholders, Driver and Block Impact on Process		
Stakeholders	DMM	DSM	DMM			
System Blocks	DMM	DMM	DSM			
Organization	Benefits to Stakeholders, Process Impact on Drivers and Blocks			DSM	DMM	DMM
Information				DMM	DSM	DMM
Process				DMM	DMM	DSM

Figure 124 - ESM and MDPM Based Conceptual Stakeholder Cost Benefit Matrix

The framework conceptually illustrated in Figure 124 could be analytical in nature where elements of the matrix have values corresponding to the actual benefits or costs created as a non-dimensional dot product of normalized values and a vector of weights similar to the costs developed for the MCNF model. This would allow for the analytical evaluation (but also a framework for discussion) of the costs and benefits of the surveillance process. Thus, difference between benefits and costs should identify a net benefit (or not). Ensuring that there are sufficient resources to develop and operate the process is part of the V2 process development as the scope of the V2 process is significantly larger and likely cannot be developed using just minimal resources as for the proof of principle V1 process in this thesis. The ESM and MDPM based stakeholder cost benefit analysis is proposed as an important analytical tool to secure the necessary resources and ensure a net benefit to stakeholders. However, in addition to identifying the costs and benefits basic familiarity and understanding of the surveillance process data will be needed for research physicians and epidemiologists to use (and trust) the data. Their involvement and buy-in of the projects detailed in appendix 1 will be necessary in addition to their contribution in terms of analysis and use of the surveillance system.

7 Future Work

Improvement in complex processes requires continuous iteration to identify necessary changes, devise changes, and test them. The knowledge gained from the previous iteration can then be used to design an improved process, which then serves as the next step in the improvement cycle. The "V1" process developed in Chapter 5 and tested in Chapter 6, allowed for the identification of necessary changes (such as the need for assembly based methods), which were identified as the series of projects defined in appendix 1. Implementation of these projects would lead to a "V2" version with much greater capability. However, the iteration should not stop there, as MRSA surveillance will require far more than a single prototype process addressing the needs of a hospital. The surveillance processes should continue to expand in successive versions to address more hospitals, communities, states, and eventually a nationwide system. Deployment of surveillance processes to the much larger number and type of locations will require the redesign of processes to be simpler, cheaper, and highly integrated with each other. Additionally, emphasis on process and publication and cost reduction would likely lead directly to the development of DNA sequencing based processes that could replace current culture based MRSA diagnostics. This would allow for "real-time" surveillance as cases appear, and while these systems may not be initially as comprehensive as "full" surveillance that could serve to collect enormous quantities information not available today. Perhaps most importantly, ongoing improvement and optimization could produce robust and very low cost surveillance systems that could be deployed in Third World locations providing true "frontline" surveillance in key areas where new pathogens may emerge. The ultimate goal is then to integrate all these elements into a global surveillance system that would allow for real-time tracking of MRSA evolution, outbreaks, and control measures. A detailed roadmap for these successive process versions is presented in this chapter.

However, many of the tools developed in this thesis could apply to other areas beyond healthcare where the analysis of processes is essential. Complex manufacturing processes such as those in aerospace, pharmaceuticals, semiconductors, and other industries could readily be modeled using the MDPM methodology in order to assess their complexity and manage it. Application of these ideas is not limited to "physical" processes but could also be used to understand "business processes" such as those in the financial services industry. More generally existing research on enterprise architecture could be integrated with the process specific models developed in this thesis to better understand both the enterprise and its processes (i.e. many of the elements outside the boundary of the current analysis). Additional research will likely be needed to apply the MDPM models developed in this thesis and other related ones such as the ESM matrix to many additional cases in order to identify general characteristics and design rules for complex systems (and processes). A key general improvement is the development of software to facilitate the capture and analysis of complex process elements in each of the domains as well as the need for "versioning" to map the change of the network through time. The second key improvement would be the explicit quantification of the additional "scaffolding" described in this thesis to support either rapid process change or significant improvement. The scaffolding could be described as "process options" which process designers and owners could evaluate in terms of their return on overall process properties such as flexibility, adaptability, and NPV for the process overall. These ideas for future work are also presented in this chapter.

7.1 Surveillance Roadmap

The proof of principle surveillance process developed in this thesis should be just the starting point for a series of process iterations leading to first a comprehensive national MRSA surveillance system and eventually to a global surveillance system. The high interconnectedness of modern society means that new pathogens can emerge across the country or across the planet and be transported very quickly to nearly any other point on the planet through the air transportation system. However, despite the need for substantially increased surveillance, the initial V1 case study suggests the process still requires additional functionality despite its high complexity. Significant effort will be needed to first add the additional functionality into a V2 process but looking beyond that it would likely be possible to follow on with a V3 process that focused on simplifying the process and removing much of the "scaffolding" needed for the V1 and V2 processes. And yet removing the scaffolding within itself require additional resources before the final "streamlined" V3 process was fully deployed and tested. This suggests that much greater process complexity will be required between the current proof of principle V1 to an operational V3 process and this potentially large increase in the number of elements in each domain may limit the speed and usability of developing these additional iterations unless sufficient resources are secured. This might be thought of as a "complexity barrier" where significant additional resources are needed to transform a current process into a future one that may be simpler and more efficient but requires a temporary increase in development and testing resources to "cross" the barrier. Table 26 describes additional process versions to address the incorporation of radical new technologies (the proposed V3 process assumes only existing commercially available technologies) and the need for 3rd world surveillance. However, a conceptual representation of the "complexity barrier" for these process iterations is shown in Figure 125 where the cyclomatic number (edges-vertices+1) is plotted for the two data points in this thesis (the baseline and V1 processes) and notionally for successive iterations of design, implementation, and testing.

	Description	**Process Knowledge**	**MRSA Knowledge**
V1	**Proof of Principle**	Basic Performance Identify Improvements MDPM Model	Variation in Genome Diversity Observations Analyses Needed
V2	**Pilot Process**	Address V1 Issues Additional Features High Throughput Operation	MRSA Pan-Genome Sampling Rate Needed Phenotype Correlations
V3	**Production Ready (Existing Technologies)**	Scaffolding Removal (Org/Info) Robust Process Transferable Process Multiple Location Network	Regional Variation Multi-Infection Studies Meta-Genomics Related Organisms
V4	**Production Ready (New Technologies)**	Process Domain Optimization National Surveillance National Network	Host Interaction Population Dynamics Outbreak Tracking
V5	**Diagnostic Integration**	Culture Diagnostic Replacement Surveillance Data Integration Integration with Network	Diagnostic Validation Prevalence (screening) Correlation to Cases

	Description	Process Knowledge	MRSA Knowledge
V6	**3rd World Surveillance**	Low Cost, Robust Process Surveillance Data Integration Integration with Network	3rd World Variation Selective Pressure Horizontal Gene Transfer
V7	**Global Surveillance**	Integrated Processes Surveillance Data Integration Integration into Single Network	Global Changes MRSA Evolution Global Outbreak Tracking

Table 26 - Surveillance Process Roadmap of Future Versions

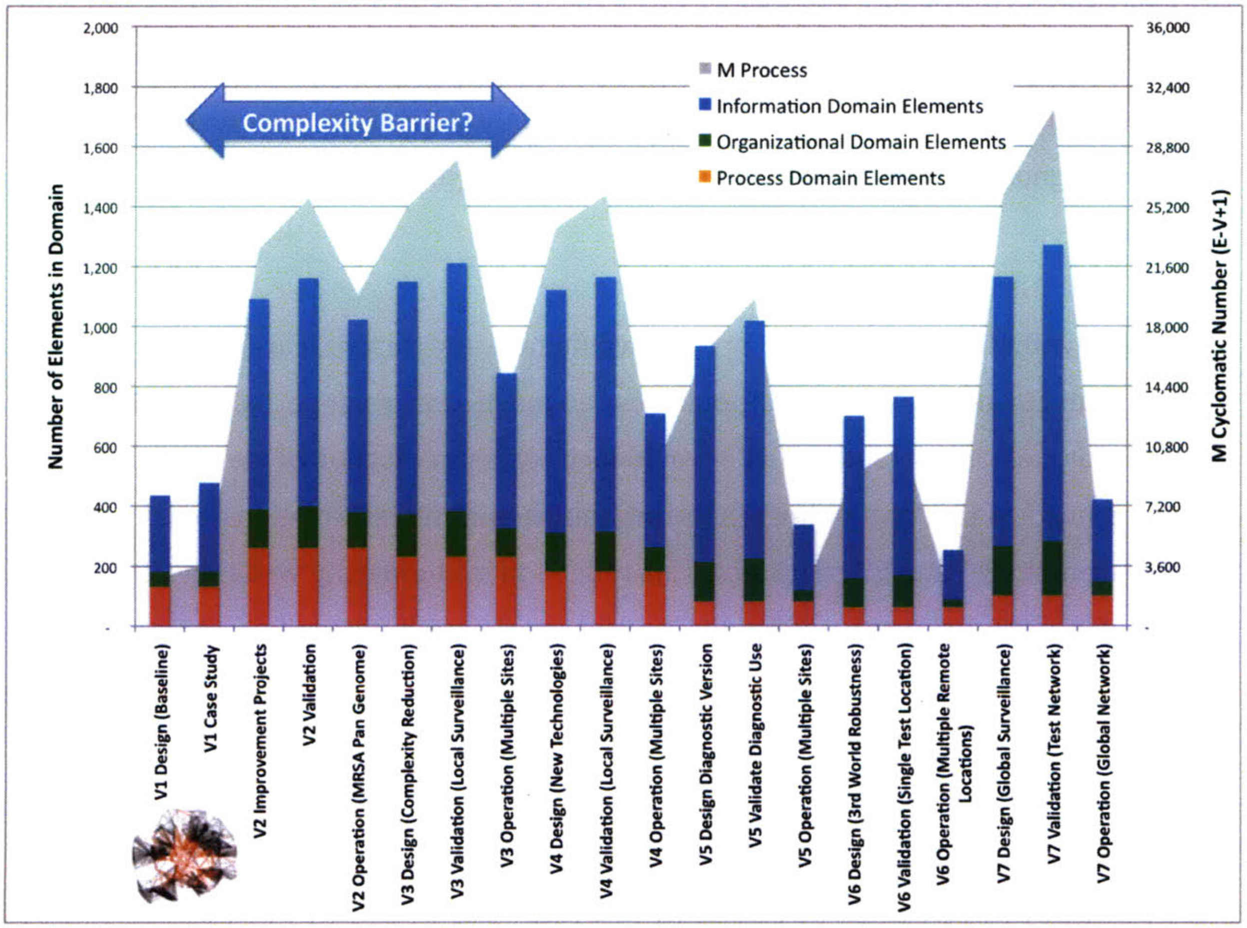


Figure 125 - Conceptual Representation of Cyclomatic Complexity Increases for Successive Surveillance Process Versions. Process Networks at Same Density (0.02).

The V1 process has already been extensively described in this thesis and the V2 process would incorporate the projects described in chapters 5 and 6 to address issues found in the testing of the V1 process and incorporate desired improvements. The V2 process would allow for a vast increase in our knowledge of MRSA as it would have the scale and ability to sequence entire collections of MRSA samples such as the CDC's. Obtaining the full sequence of these diverse collections would allow for the first full MRSA "pan-genome" describing a sampling of the variability seen to date across the United States. The pan-genome survey would also provide an estimate of the sampling rate needed in a national surveillance system based on the genetic divergence observed in the collections. Most importantly the V2 process would enable the correlation of phenotype characteristics (such as virulence measurements) with the genotype found such that only genetic information might be used in future diagnostics. The V2 would only rely on existing commercially available technology.

Continuing with the roadmap, a V3 process would take any lessons learned from the V2 operation and incorporate any additional desired functionality not perfected in the V2 process. But most importantly, the V3 would focus on simplification of the process by removing much of the "scaffolding" in the organizational and information domains "left behind" from the V1 and V2 processes. This reduction in complexity would be critical for the transfer of this process to additional locations such as other hospitals or research centers. The V2 process is not contemplated as a "multi-location" process due to the significant complexity is likely to have and the need to have a correspondingly large and complex network at a potential additional location. The V3 process would thus focus on providing "transfer links" similar to the change representatives were troubleshooting representatives described in Chapter 5 who would engage the transfer stations and ensure most of the process was a few network hops away. Additionally, as this V3 process would be focused on replication such that it could be transferred, it should also be made robust to minimize the troubleshooting needed at destination organizations and would thus require very significant "scaffolding" during its development to test a number of failure modes and eliminate them. The V3 process would allow for the observation of regional variation in

MRSA as well as the tracking of patients with multiple infections over time. Most importantly the V3 would have sufficient additional capacity to sequence related organisms that may be critical to the development of MRSA through horizontal gene transfer such as Vancomycin resistant Enterococcus. Additionally, the V3 process could be used to perform "meta genomics" surveillance where samples might be pooled together and sequenced to greatly increase the number of samples that could be reserved. The pools would be compared to the reference "pan genome" subtracting known genes and only flagging novel sequence corresponding to new genes or gene is being transferred from other organisms. The V3 process is also expected to only use commercially available current technology.

However, radical new technologies are being developed in early development stages that could radically change the process domain of the surveillance process described in this thesis. These new technologies include microfluidics based devices that could replace many of the steps performed in the V1, V2, and V3 processes with highly integrated "chips" with minimal operator involvement. For example, (Mahalanabis, Al-Muayad et al. 2009) describe a microfluidics based device to lyse MRSA cells mechanically without the need for the enzymatic processing that proved to be a key development problem and connecting technology for DNA extraction described in Chapter 5. This device would replace many of the steps and require perhaps the single technician with a corresponding decrease in the number of organizational and information domain elements. Additionally, new DNA sequencing technologies are becoming available that greatly increase the speed, completeness, and quantity of data potentially making same-day results possible. These technologies also have the potential to greatly reduce the number of process domain elements with a corresponding reduction in information and organizational domain elements. The development of processes using these technologies however would require significant initial scaffolding as the technical complexity in building these new devices is very significant although the resulting processes are much simpler. The incorporation of all these new technologies with the specific aim of further complexity reduction and process optimization is called the "V4" process as shown in Table 26. The V4 process would allow for much more extensive deployment due to its reduced complexity (and reduced infrastructure needs) leading to the potential for national surveillance. However, going along with national surveillance would be the need for a national system to collect and process all of the relevant information and coordinates any changes or improvements in each of the distributed processes. Thus, while the "local" process complexity may be greatly reduced the overall "system" complexity could grow significantly and would have to be carefully managed to avoid unnecessary complexity. The V4 might also allow for the sequencing of hosts to identify any host-MRSA interactions not previously observed (i.e. some individuals may be better able to resist MRSA or the reverse). The more excessive deployment would also allow for outbreak tracking and the study of population dynamics (i.e. how the overall MRSA population is changing).

The same technologies used for the V4 process might also be used for the development of highly efficient diagnostics to replace current culture-based methods. These diagnostics might be "simplified" lower performance processes than the full-scale surveillance processes but nevertheless sufficient for diagnostic use, particularly when combined with ongoing surveillance that could inform what the diagnostics should look for (i.e. new virulence genes). While, PCR is currently available for this function, these tests only look for the presence of a few select resistance genes to identify MRSA. A DNA sequencing-based diagnostic would look in several of the pathogenic regions of the genome (if not the whole genome itself) to identify the particular strain precisely and better tailor the therapy the patient might receive. These diagnostics should be integrated with the overall surveillance network providing even more information. The development of these diagnostics however would require very significant "scaffolding" as the validation would need to be extensive and the diagnostics would have to be proven as highly reliable.

One of the most important and often overlooked areas of surveillance is in poorer 3rd world countries that suffer from an enormous burden of infectious disease. The high prevalence of disease in these countries also makes them a breeding ground for new strains of pathogens many of which could spread very rapidly. In addition to efforts to improve health care in these countries, surveillance could be a valuable tool to improve the ability of these countries to deal with epidemics but also serve as frontline surveillance tools for the identification of new pathogens. However, the development of robust and low-cost surveillance processes for these countries would require significant effort to greatly reduce the complexity of the processes such that they could operate in locations with very limited infrastructure but which would nevertheless need to be connected to the overall surveillance network. This process is called the "V6" process as shown in Table 26.

Finally, the ultimate goal must be to integrate all of these elements such as diagnostics, V4 new technology surveillance processes, and 3rd world surveillance into a single global surveillance system. This would allow for the efficient use of resources to identify emerging MRSA strains, limit their spread, and perhaps develop ways to eliminate them. The global tracking of pathogen changes would be a powerful tool when integrated with not just MRSA but other microbial pathogens such that coordinated efforts could be made to manage global antibiotic distribution, management, and development. Global strategies to cycle antibiotics might be effective in controlling the spread of resistance involving the use of sophisticated supply-chain models to coordinate antibiotic distribution with surveillance data.

The processes described thus far (V2-V7) are intended to be elements of large-scale surveillance systems, which can have significant complexity just from the need to manage and communicate with many surveillance locations. For example, a national surveillance system would need to cover all major hospitals, key clinics in remote or high-risk areas (low income or high drug use), and importantly cover animal production (cattle, swine, etc.) which are exposed to antibiotics and could harbor resistant microbes. These locations could be more than 10,000 alone without counting the possibility of including diagnostic elements. A global surveillance system might increase this number by an order of magnitude or more to cover 3rd world locations plus equivalent hospitals and clinics in developed countries. The complexity of such a system could be very high causing difficulties to manage, change, and quickly troubleshoot elements of the system in similar ways to those described in this thesis. Two aspects of this global surveillance system deserve further discussion in the context of this thesis. First, what should the control infrastructure look like to manage such a system? And second, what is the impact of the complexity of each individual process element on the overall complexity of the system?

This is shown conceptually in Figure 126, where a hierarchical structure for control of the system is used and the V1-V7 processes are each at multiple locations.

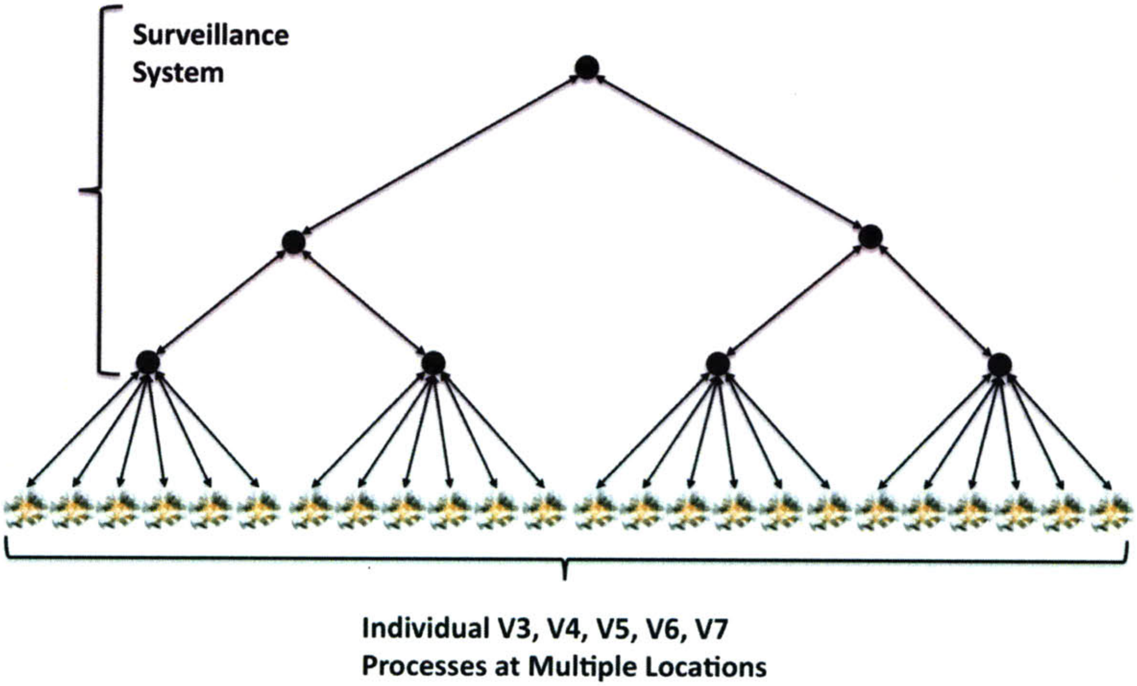

Surveillance System
Individual V3, V4, V5, V6, V7 Processes at Multiple Locations

Figure 126 - Conceptual Diagram Showing a Hierarchical Structure for a Regional Surveillance System

A hierarchical control system can connect a large number of elements just a few nodes depending on the branching ratio and number of levels as given by Equation 26.

$$N = \frac{b^L - 1}{b - 1}$$

where N is the number of Nodes, b is the branching ratio, and L is the number of levels

Equation 26 - Number of Nodes for a Hierarchical Structure

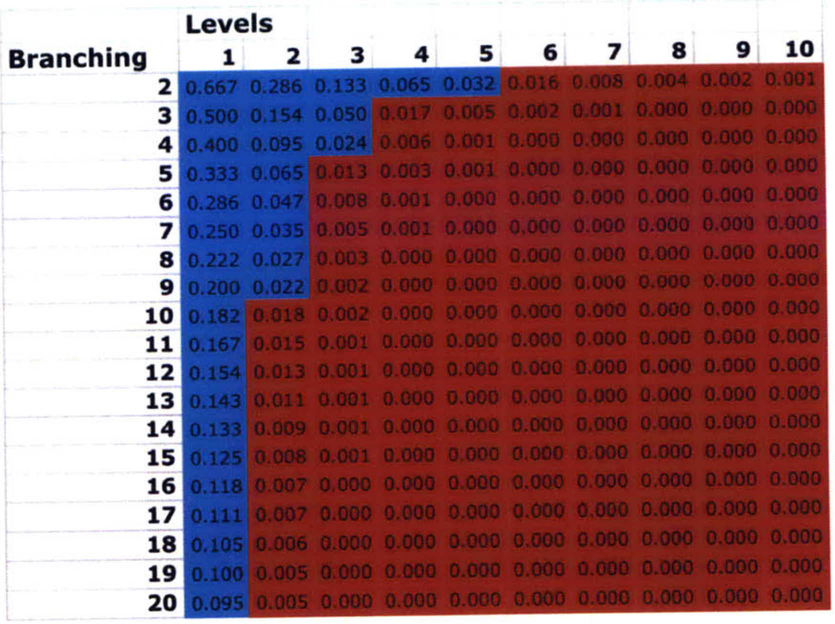
However, despite their advantages for connecting large numbers of nodes, hierarchical structures are extremely sparse, rapidly falling below the network densities observed in the MDPM model developed in this thesis. The density for a hierarchical structure can be derived from Equation 26 where the number of edges will be twice one less than the number of nodes, and compared to the maximum density an equivalent number of nodes could achieve in a complete graph. The density for a hierarchical structure was derived and is given in Equation 27.

$$\underset{hierarchical}{\Delta} = \frac{2(b-1)}{b^L - 1}$$

where b is the branching ratio, and L is the number of levels. Bi-directional edges.

Equation 27 - Density of a Hierarchical Structure Compared to A Complete Graph with the Same Number of Nodes

Compared to the observed densities in the MDPM of the V1 process, hierarchical structures are exceedingly sparse, and for comparison the density of a hierarchical structure compared to the ~0.02 observed density of the V1 model is shown in Figure 127. Areas in blue are the same or equal to or greater than a 0.02 density, while areas in red are below this value and drop exponentially with increasing numbers of levels and branching ratios.



	Levels									
Branching	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	0.667	0.286	0.133	0.065	0.032	0.016	0.008	0.004	0.002	0.001
3	0.500	0.154	0.050	0.017	0.005	0.002	0.001	0.000	0.000	0.000
4	0.400	0.095	0.024	0.006	0.001	0.000	0.000	0.000	0.000	0.000
5	0.333	0.065	0.013	0.003	0.001	0.000	0.000	0.000	0.000	0.000
6	0.286	0.047	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000
7	0.250	0.035	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000
8	0.222	0.027	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.200	0.022	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.182	0.018	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	0.167	0.015	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.154	0.013	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
13	0.143	0.011	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0.133	0.009	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	0.125	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	0.118	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	0.111	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18	0.105	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.100	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	0.095	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 127 - Density of a Hierarchical Structure with Increasing Hierarchical Levels and Branching Ratios. Areas in red are below (0.02) density.

The degree to which the overall surveillance system might be connected is critically important to its design. As has been shown in this thesis each process can have significant complexity and a design question is how best to replicate these processes in multiple locations such that the overall system complexity is reduced. For example, it may be that each location has just a few MDPM elements (<5%) that are unique to that location due to the particular organizational structure, available infrastructure, or stakeholder needs. The rest of the elements could be identical in a "copy exactly" fashion. Yet even a 5% difference in each location multiplied by 10,000 or more locations means there will be many elements unique to each location making change and troubleshooting very difficult and requiring significantly more overhead on the part of the surveillance system network. Shown in Figure 126 is a hierarchical network, which can link all of the processes with relatively few levels, however this network is exceedingly fragile and can quickly become congested since all "messages" must travel up through the hierarchy potentially losing information in the process and then back down to another process element. This would be problematic for the management of complex processes and is likely that additional connections would be needed to facilitate learning (i.e. incorporating improvements discovered at various locations) across the network. Analyses similar to those in (Dodds, Watts et al. 2003), where the communications performance of various control structures are evaluated should be performed on the surveillance system. As found in (Dodds, Watts et al. 2003), multiscale networks which combine the advantages of hierarchies with the shorter average paths of "small world" networks where additional links are inserted will likely be the optimal solution. This is shown conceptually in Figure 128.

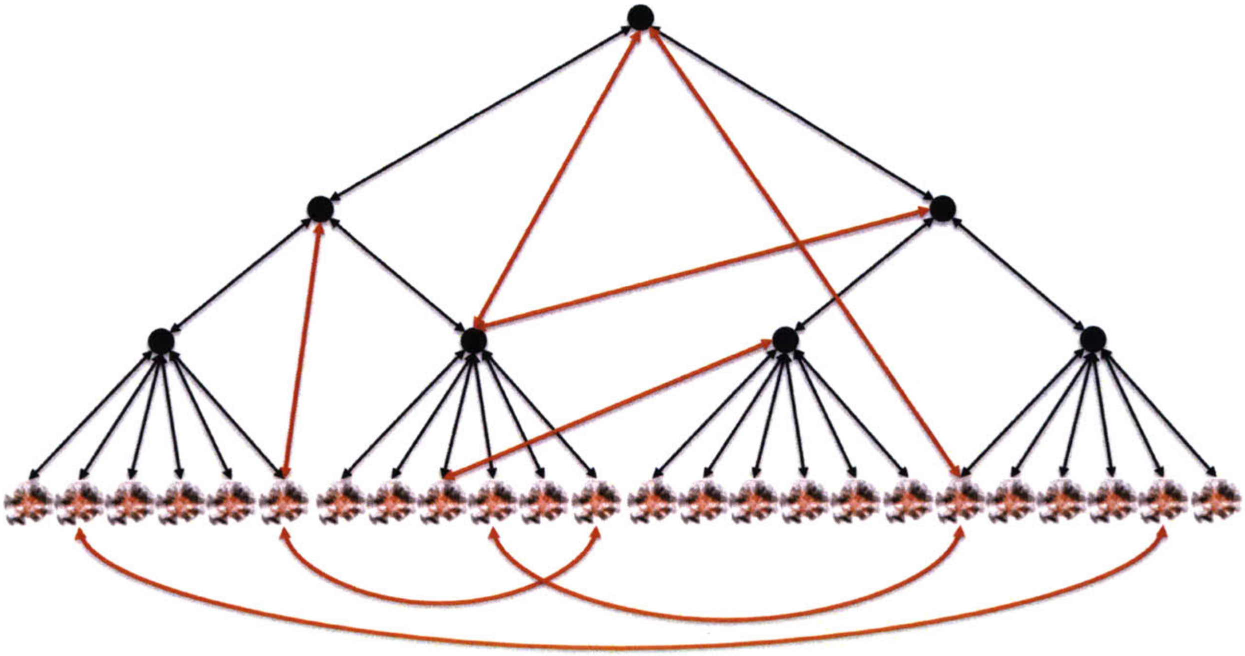


Figure 128 - Conceptual Representation of a Multiscale Control Structure for the Surveillance System Incorporating Hierarchical Structure Supplemented by Additional Links in Red

7.2 Additional Surveillance Dimensions

In addition to the microbial genome, which has been the emphasis of this thesis, there are a number of other dimensions that should be considered in developing a "V2" and successive versions of surveillance systems. Our knowledge of the microbial world we inhabit is very limited as described in chapter 1, however new projects such as the Human Microbiome Project to sequence all bacterial species associated with humans (living in our bodies) will vastly increase our knowledge. In addition, sequencing of more human genomes will provide insight into the role host immunity plays in infection. The true "microbial state" of an individual will likely encompass not just the individual organism (i.e. MRSA) but also the state of other organisms with which it might share genetic information, the genetic composition of the host, environmental factors, and several other dimensions. An ideal system would provide information on each of these additional dimensions beyond those considered in this thesis such that an accurate representation of all factors associated with infection can be formed. An ideal surveillance system would include the additional dimensions listed below:

Microbe: System pathogen target (i.e. MRSA)

- **Phenotype:** the observable physical characteristics of an organism including visual, metabolic, and clinical properties resulting from its genotype and environment
- **Genotype:** the particular genetic composition of an individual organism
- **Pangenotype:** the genetic composition of closely related strains, species, and other microbial organisms with which an organism can readily exchange genetic information

Host: Pathogen target (i.e. human)

- **Phenotype:** the observable physical characteristics of the host including visual, metabolic, and clinical properties resulting from its genotype and environment.
- **Genotype:** the genetic and epigenetic composition of the host
- **Microbiotype:** the particular microorganism composition of the host (i.e. such as from Human Microbiome Project Data)

Environment: Microbe and host

- **Epidemiological State:** the conditions including patient history, environment, interaction, and current medical condition that affect infection
- **Healthcare State:** the set of policies related to treatment of bacterial infection including type of antibiotic commonly prescribed (selective pressure)

This represents an ideal and complete representation (and data set) to satisfy the needs of additional stakeholders, including individual physicians treating patients, but also enabling system-wide control approaches. The feasibility of various surveillance system designs should also be evaluated in terms of how completely they fulfill each of these dimensions. Note that these would be "systems" rather than processes as the data for these dimensions would come from multiple sources. A systems approach for a microbial surveillance system is essential. Further description of each dimension follows:

Microbial Pangenotype

One of the key mechanisms of antibiotic resistance in microbes is horizontal gene transfer from one strain or even a different species whereby novel genes are acquired. This happens routinely, and in fact the idea of a single "species" of microbe may not be correct as microbes really exist as communities with a distributed genetic instruction set ensuring the survival of the community if not individual species. The rapid development of antibiotic resistance demonstrates this process very clearly and in order to develop a truly effective surveillance system it may be necessary to also perform surveillance of other bacteria closely associated or living in the same environment as the target pathogen (i.e. MRSA). The recent start of the Human Microbiome Project, which aims to identify and sequence all microbial species, will help identify many of these intermediaries although some are likely to exist outside human hosts. This community of microbes is the pangenotype, defined as: the set of microbes in contact with and able to exchange genetic information with the target organism. Of particular interest currently are situations where MRSA may come in contact with Vamcomycin resistant enterococci (VRE) with the potential to exchange genetic information and become a staphylococcus strain resistant to nearly all antibiotics (VRSA) as described previously. The only feasible technology to perform surveillance of the pangenome are next generation sequencing methods such as 454 and Illumina, although significant work is needed to develop the upstream microbial acquisition, isolation, and preparation prior to sequencing. The only currently available approach is based on meta-genomic protocols where entire samples (i.e. blood) are sequenced as a pool without regard for isolation of individual microbes to identify any novel genes. This approach could help identify genes that may be in the microbial community with the possibility for exchange, but does not allow for the isolation and phenotypic evaluation of the microbe. The Human Microbiome Project is likely to develop much of the necessary technology however as the needs are similar. The key pangenotype identification processes required (some of which do not exist, except as prototypes) are:

- Meta-genomic sequencing
- 16s sequencing
- Single cell isolation
- Single cell sequencing

Host Phenotype

A surveillance system will require correlation with the state of the patient, including all clinical information to determine the effects of different populations, disease states, and co-factors (smoking, etc.) on the original infection and outcome of treatment. The host phenotype description consists of clinical observation and a complete medical record. The key host phenotype identification process required is:

- Clinical observation and recording

Host Genotype

The genetic profile of a patient is likely to have a significant correlation to the incidence of microbial infection, and a complete surveillance system would also provide this information. It is possible certain metabolic or immune system changes may make hosts better suited to fight off infection and this understanding would help develop both novel antibiotics but also a vaccine. The technology to do this also relies on next generation sequencing, but would also require new and not yet commonly available processes to read both the DNA sequence but also the epigenetic state of cells. These epigenetic modifications switch on or off the genes in cells and are responsible for the cell differentiation. It is very likely that in addition to changes in the basic DNA sequence these modifications may provide advantages to certain hosts that could be exploited for vaccine or antibiotic development. A surveillance system will require correlation with the genotype of the host including both genomic and epigenomic information. The key host genotype identification processes required are:

- Whole genome sequencing
- Whole epigenome sequencing
- Expression analysis

Host Microbiotype

Human beings co-exist with an enormous number of microbes, nearly 10^{15} cells containing an order of magnitude more genes than are in the human genome by various estimates. The Human Microbiome Project intends to catalog all of these species and eventually catalog them. This project will likely change the definition of bacterial species, since the entire population probably exists as a single pangenome with various strains and species sharing significant similarity but diverging in small but critical ways. The particular composition of an individual's microbiome is likely to have a very significant impact on their susceptibility to disease and in particular microbial infection. Some species may "out-compete" pathogens preventing them from infecting individuals. In others, a different microbiome may assist infection if certain species are not present. A complete surveillance system would also provide a catalog of the host microbiome such that the presence or absence of certain species could be correlated with infection. The host microbiome is also the prime reservoir of commensal microbes with which pathogenic microbes can exchange information and is therefore an element of the "pangenotype". As with the pangenotype the only feasible methods to produce a host catalog are sequencing based approaches using either the conserved 16s region to serve as a "fingerprint" of the species, or full genome sequencing. The key microbiotype identification processes required (some of which do not exist, except as prototypes) are:

- Meta-genomic sequencing
- 16s sequencing

- Single cell isolation
- Single cell sequencing

Epidemiological Environment

A surveillance system must also provide information on key epidemiological variables occurring simultaneously either in time, location, population, or exposure. These variables are collected in the host phenotype data set but include others that may be relevant such as cohort (i.e. older patients), location (i.e. ICU patients), or exposure (i.e. a doctor not washing hands regularly). The extensive data that could be collected in the other dimensions is made more valuable by carefully collecting and correlating other variables to the emergence of new or different pathogens (or pathogenic mutations). Key epidemiological processes are:

- Epidemiological data collection
- Epidemiological data aggregation

Healthcare Environment

A surveillance system must provide the particular information of treatment and outcomes for patient treatment. Knowledge of which antibiotics were prescribed in a particular case but also as policy (affecting the pangenome) is necessary to draw correlations between healthcare policies and observed effects in the surveillance data. Additionally, knowledge of diagnostic tests performed help validate their accuracy. These variables are collected in the healthcare environment data set but could include others such as hospital hand-washing policies, clinical staff education, and others to help correlate the observed surveillance data with healthcare policies. Key processes in this dimension are:

- Healthcare data collection
- Healthcare data aggregation

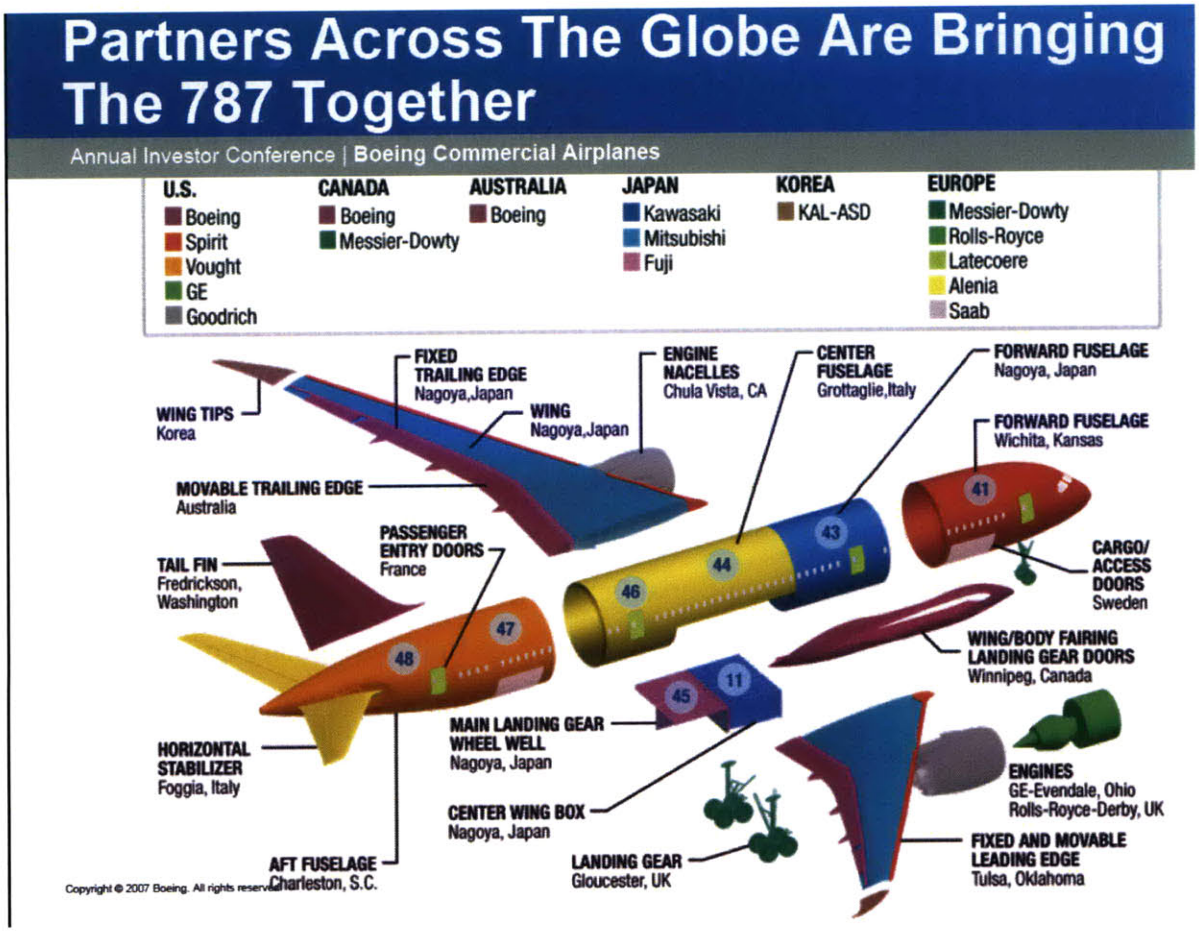
7.3 Complex Processes

While the main focus of this thesis has been on microbial antibiotic resistance surveillance as an example of a complex process there are many others in healthcare and many more in other areas of engineering. For example, much of our modern infrastructure relies on complex but largely unseen processes such as water treatment, power generation, and petrochemical production. The technology and objectives of these processes may be different but they all share the same types of elements as those described in the MDPM model developed in this thesis. All these processes are composed of an intricate series of activities that must be performed in order to deliver a product or service. Individuals and organizations interact with each of these activities as operators, designers, and managers providing a control structure for the process but also significant tacit information. These processes also have explicit information in the form of protocols, work instructions, and other documents. The interaction of all these elements across the three domains identified in this thesis, organizational, information, and process could be analyzed using the same MDPM methodology and model described in this thesis. Many of these processes will be embedded in systems, which will require expansion of the scope of the MDPM to also include external stakeholders and certain other elements not considered in the MDPM. Connection of the MDPM with the additional dimensions considered in the Engineering Systems Matrix (ESM) would facilitate the analysis of larger systems containing processes.

To extend the application of the MDPM model to some of these other processes additional work would be needed to facilitate the entry and analysis of process data as well as significantly improved visualization techniques. Most important are the abilities to represent evolution of an MDPM model as changes occur and new versions of individual elements are created. In the example developed in this thesis, successive process versions were distinctly labeled but this was done manually, and software to manage change as individual work instructions change or processes are reorganized will be needed. Additionally in terms of troubleshooting there will also likely be a need to add a physical "equipment and infrastructure" domain to describe any machines or devices needed to perform a particular activity as these may also contribute to failure modes and must be "searched" across the network in much the same way as protocols or operators.

Further description of the connections between elements could also improve the MDPM model as not all connections between individuals will be the same and could be highlighted to represent this. Similarly, connections between information elements may be of different types and have very different bandwidths, which could further improve the model. This also partly addresses a need for a very robust methodology to capture the richness of all these different connections including both computer science data mining methods for the information domain and also formal interview methods from sociology. This may also require the creation of a few additional element types such as nodes representing organizations rather than individuals.

A concrete example of an alternative application of the MDPM model could be the design, development, and manufacturing of Boeing's new 787 aircraft whose value chain was explicitly designed to span multiple groups, disciplines, companies, and even continents. A schematic of where each of the 787 components is made and which organization is responsible for them is provided in Figure 129. While potentially more efficient in terms of leveraging the resources of multiple suppliers to help with capital and design resources, the complexity of the design and manufacturing process was greatly increased leading to problems with supply chain coordination, change propagation, and design difficulties well documented in the press. The program is now more than two years late and while many of the difficulties may be attributable to the product complexity of this highly advanced and technologically innovative jet, it is possible that these were compounded by a lack of visibility into the complexity of the process designing and building it. Application of a MDPM model to the 787 design and manufacture may have shown the need for much shorter average path lengths between nodes and certain "hidden" nodes with very high search information values that could be difficult to troubleshoot or conversely certain nodes where changes will propagate through the entire system making them critically important to manage. The breadth of organizations and disciplines also might have benefited from the "troubleshooting representatives" or "change owners" described in chapters 5 and 6. These individuals could have served as long links with their counterparts in other organizations to greatly reduce the average network distance but also to proactively search for information their local network "neighborhood" might need. In the information domain, a network-based change propagation tool such as the MDPM might have helped to identify which changes had the most impact and which parts of the aircraft could be "modularized" with less concern about changes elsewhere impacting those systems. The careful monitoring of change requests to determine the true impact of each change could also be performed using an MDPM model to identify "hot spots" of change or alternatively to make sure areas not considered have been reviewed. This is analogous to the protocol reviews performed in the surveillance system to make sure other protocols were not impacted by a change in any particular one.



Partners Across The Globe Are Bringing The 787 Together

Annual Investor Conference | Boeing Commercial Airplanes

U.S.	CANADA	AUSTRALIA	JAPAN	KOREA	EUROPE
■ Boeing	■ Boeing	■ Boeing	■ Kawasaki	■ KAL-ASD	■ Messier-Dowty
■ Spirit	■ Messier-Dowty		■ Mitsubishi		■ Rolls-Royce
■ Vought			■ Fuji		■ Latecoere
■ GE					■ Alenia
■ Goodrich					■ Saab

Figure 129 - Boeing 787 Suppliers as an Example of a Complex Process (from Boeing)

7.4 Process Options

Throughout this thesis there has been description of the additional elements needed to create, design, and implement complex processes as "scaffolding" which can be greatly reduced during the operation of a fully robust and tested process. However, this scaffolding has an additional cost, which might be quantified using real options methods. The scaffolding of a process can be thought of as an option "in" the process as it enables the process owner to change or troubleshoot the process at some future date likely increasing its value. The process owner has the option to exercise the change or not, but not the obligation to do so. The MDPM model could be used to explicitly quantifying the additional resources needed for the scaffolding, which might be compared to the potential value in terms of flexibility and process availability. Integration of the MDPM model with process based cost modeling where each activity in the process domain could have an associated cost along with the "fixed cost" of the organizational and information domains could also yield a powerful analysis methodology. And the Minimum Cost Network Flow (MCNF) methodology developed in Chapter 5 could readily be used for cost analysis but also for the valuation of real options where a comparison between the "cost" of producing a certain amount of products can be compared under multiple scenarios such as one with "scaffolding" and one without. The total cost across a given path can then be used as the production function in a valuation model to understand the relative benefit of different design approaches. The combination of the MDPM model in the MCNF based valuation tool could yield an especially valuable area of future research in developing "process options" that could quantify the value of additional flexibility in process areas with rapidly changing technologies. Complex processes such as the one cited in this thesis are composed of underlying process elements whose technologies are changing at different rates requiring certain parts of the process to absorb much faster "clock speeds" (Fine 1998) than others and requiring additional flexibility (and scaffolding) in those areas as shown in Figure 113. The additional resources or scaffolding needed to ensure change in those areas can be well-managed and integrated to the rest of the process is a process option since the future of the technology is not known, but a process with additional resources to accommodate this change will be more adaptable and have a correspondingly higher value (which must be balanced against the cost of the additional resources). This is the concept behind process options and it is hoped additional research will be performed in this area.

Appendices

Appendix 1: Potential Improvement Projects to Surveillance Process

Project #1	MRSA Sample Diversity
Description	One of the main goals of the surveillance system is to provide a true representation of the MRSA diversity that exists in hospital infections. The sampling used as an initial demonstration in this thesis was confined to a single outbreak in a single hospital, however a much broader sampling will be needed to support epidemiological analysis. The goal of this project is thus to build a much more diverse collection of MRSA samples representing multiple hospitals, regions, and case histories. This project would ideally reuse existing infrastructure for strain typing performed at local hospital, state health department, and national CDC levels. This collection would provide for a proper estimate of the "pan-genome" of MRSA or the sum total of all mutations seen in MRSA throughout the country. This would ideally be an ongoing process as MRSA will continue to evolve and a constantly updated sampling should be performed.
Internal Stakeholders	Process Owner, Process Change Owners, Epidemiologists, Bacterial Genomics Scientists, Physicians, Clinical Microbiologists, Genome Annotation Scientists
External Stakeholders	CDC, NIH, IDSA, ASM, State Health Agencies, Bacterial Diagnostic Vendors, Antibiotic Researchers, Antibiotic Vendors
Deliverables	Diverse collection of MRSA samples with associated case histories and continuously updated
Project Owner	As this project would lie largely outside the process boundaries, the owner should be someone connected to the various external stakeholders perhaps at CDC.

Project #1	MRSA Sample Diversity
Process Domain Integration	This project would lie outside the process boundaries as currently defined, but would require integration with sample tracking systems, analysis, and potentially DNA extraction.
Organizational Domain Integration	Some process changes may be required to re-optimize the DNA extraction process to handle the wider diversity of samples, so coordination with the "change owner" or "troubleshooting representative" in that area would be needed.
Information Domain Integration	A process review and integration document would need to exist in order to ensure all parts of the process were aware of the potential changes, could react to them, or confirm no changes needed in any of the explicit work instructions (protocols) etc.
Process Testing Needed	Because of the significant information exchange and tracking that might be needed for this project, testing all of the systems with "dummy" samples would likely be needed. This would verify that sample histories, IDs, and physical samples could be processed without error throughout the process incorporating this project. In addition, characterization of the various samples to ensure they can be run through the existing culture growth, DNA extraction, and sequencing protocols would be necessary followed by a design review.

Project #2	DNA Extraction Robustness
Description	As described previously in this thesis DNA extraction is a highly variable process for MRSA samples, and while it has been optimized for the particular set of samples used in this thesis, it is likely that the variation seen in a much larger set would require re-optimization of this process activity. The goal of this project would be to find a new optimized protocol that can cover the widest possible range of MRSA samples. It is likely that additional measurements may have to be taken such as carefully controlling cell growth, and potentially automating a number of the steps to remove any operator variability. In addition new technologies and processes should be explored as alternatives to the current protocol. This project would then perform a series of design of experiments runs to find a robust set of parameters that could be used as a general protocol. Because of its impact on the rest of the process the same "connected development" using change owners and other mechanisms will be needed to make sure the impact of the project is well managed on the rest of the surveillance process.
Internal Stakeholders	Process Owner, Process Change Owners, DNA Extraction Scientist, Clinical Microbiology Staff, Clinical Microbiology Researchers
External Stakeholders	DNA Extraction Vendors, Liquid Handling Automation Vendors, ASM, Bacterial Diagnostic Vendors
Deliverables	Robust Protocol that Can Extract >20ug of DNA in >30Kb sizes for a variety of MRSA samples
Project Owner	DNA Extraction Scientist
Process Domain Integration	This project would replace the current DNA extraction connecting technology with an improved process, however, it must be made physically compatible with upstream and downstream activities (i.e. the DNA must be useable to the downstream process and free of any

Project #2	DNA Extraction Robustness
	contaminants from processing).
Organizational Domain Integration	Process changes may be required to re-optimize upstream and downstream of the DNA extraction process, so coordination with the "change owner" or "troubleshooting representative" in that area would be needed.
Information Domain Integration	A process review and integration document would need to exist in order to ensure all parts of the process were aware of the potential changes, could react to them, or confirm no changes needed in any of the explicit work instructions (protocols) etc.
Process Testing Needed	Beyond the DOEs needed to optimize the DNA extraction process itself, an additional set of DOEs will be needed to test the response of the rest of the process to the expected variation (and any extreme conditions) produced by this part of the process.

Project #3	**Advanced Molecular Biology**
Description	One major limitation of the current "V1" surveillance process is the size of DNA fragments used in preparing the DNA for sequencing, which are in the 300-500 base pair range. However, in order to link genomic regions with high similarity and assemble the genomes out of the short reads (100bp long) it is necessary to construct much longer specialty DNA fragments that can span greater distances than 300-500 base pairs, ideally into the 4,000-10,000 and even longer but which can still be read using the available sequencing technology. Methods for doing this are so-called jumping libraries, but there may be others. This has significant technical difficulty especially given the high throughput nature of the surveillance process and will require tight integration with algorithms and other process elements.
Internal Stakeholders	Process Owner, Process Change Owners, Alignment Algorithm Research Groups, Assembly Algorithm Research Groups, DNA Sequencing Technology Vendors, DNA Sequencing Lab Staff
External Stakeholders	Other DNA Sequencing Technology Vendors, Genomics Researchers, Molecular Biology Vendors
Deliverables	Parallel process that can deliver ready to sequence fragments with ends of each molecule spanning >4,000 bases
Project Owner	DNA Sequencing Researcher (Molecular Biology)
Process Domain Integration	This project would add a parallel process in addition to the current DNA preparation and would then add an additional data type through analysis. The activity must be mad compatible with existing upstream processes (DNA extraction) and downstream analysis.
Organizational Domain Integration	Process changes may be required to re-optimize upstream and downstream of the DNA extraction process, so coordination with the "change owner" or "troubleshooting representative" in that area

Project #3	**Advanced Molecular Biology**
	would be needed.
Information Domain Integration	A process review and integration document would need to exist in order to ensure all parts of the process were aware of the potential changes, could react to them, or confirm no changes needed in any of the explicit work instructions (protocols) etc.
Process Testing Needed	A significant number of DOEs will be needed to ensure the process is robust, but further tests will be required to determine the impact of variability and certain failure modes on downstream analysis and similarly the impact of upstream variability and failure modes in DNA extraction on this part of the process.

Project #4	Configure Process for Rapid DNA Sequencing Technology Change
Description	There is expected to be significant technology change in the DNA sequencing portion of the surveillance process and the V2 process should be configured to deal with this change. This project would identify the process invariants such that these could be managed as a group and efficient connections made to the new sequencing processes as they emerge rather than ad hoc connections for each process. In addition "connection points" would be identified within the MDPM matrix such that new technologies could be rapidly absorbed without having to "make" new connections. This might mean the formation of "clusters" around key disciplines such as molecular biology that could oversee related steps in each of the new processes (rather than have each create a new one). Similarly each new process would have explicit representatives for each area that were within 1-2 hop network reach of the new process elements to facilitate integration.
Internal Stakeholders	Process Owner, Process Change Owners, Alignment Algorithm Research Groups, Assembly Algorithm Research Groups, DNA Sequencing Technology Vendors, DNA Sequencing Lab Staff
External Stakeholders	Other DNA Sequencing Technology Vendors, Genomics Researchers, Molecular Biology Vendors
Deliverables	Parallel process that can deliver ready to sequence fragments with ends of each molecule spanning >4,000 bases
Project Owner	DNA Sequencing Researcher (Molecular Biology)
Process Domain Integration	This project would add a parallel process in addition to the current DNA preparation and would then add an additional data type through analysis. The activity must be mad compatible with existing upstream processes (DNA extraction) and downstream analysis.

Project #4	Configure Process for Rapid DNA Sequencing Technology Change
Organizational Domain Integration	Process changes may be required to re-optimize upstream and downstream of the DNA extraction process, so coordination with the "change owner" or "troubleshooting representative" in that area would be needed.
Information Domain Integration	A process review and integration document would need to exist in order to ensure all parts of the process were aware of the potential changes, could react to them, or confirm no changes needed in any of the explicit work instructions (protocols) etc.
Process Testing Needed	A significant number of DOEs will be needed to ensure the process is robust, but further tests will be required to determine the impact of variability and certain failure modes on downstream analysis and similarly the impact of upstream variability and failure modes in DNA extraction on this part of the process.

Project #5	**Build Data Analysis Pipeline**
Description	Most of the data analysis performed in the V1 process was performed on prototype software with significant manual intervention and little optimization. To make the surveillance process truly useful a significant amount of work will have to go into building robust and automated data analysis pipelines that have been extensively validated and optimized for MRSA. These pipelines must be able to align data from these samples against the entire set of MRSA genomes to date as well as perform a verification of the data through assembly in order to find novel genes or horizontal gene transfer from other organisms that simple alignment would not find. The overall accuracy and specificity of the algorithms must be evaluated as a function of the data input to specify the degree of coverage (i.e. how much oversampling of the genome is needed, 30X, etc.) to provide sufficient accuracy. Most importantly the resulting data and analyses must be easy to use for epidemiological studies such that outbreaks or new strains can be readily identified from the data. Another requirement is that the data analysis must be entirely backwards compatible with existing surveillance methods such as PFGE, MLST, and 16s such that data can be integrated with existing databases and the transition to the new process can be greatly facilitated. Attention to this final user interface is critical to help speed adoption of the process. In addition, sample tracking must be provided such that original clinical histories can be associated with the final surveillance data. All of this represents a very significant amount of work and impacts many elements of the MDPM.
Internal Stakeholders	Process Owner, Process Change Owners, Alignment Algorithm Research Groups, Assembly Algorithm Research Groups, DNA Sequencing Technology Vendors, Bio-Informatics Analysis Staff, Bio-

Project #5	**Build Data Analysis Pipeline**
	Informatics Researchers, Genomic Data Analysis Research Groups, Bacterial Genomics Researchers, Infectious Disease Researchers, Epidemiologists, Physicians, Clinical Microbiology Researchers
External Stakeholders	CDC, IDSA, NIH, ASM, State Health Agencies, Genomic Data Analysis Vendors, LIMS vendors, Bacterial Diagnostics Vendors, Hospital Administration
Deliverables	Integrated data analysis pipeline that can identify MRSA strains, find novel mutations, assemble full MRSA genomes to find novel genes or transfers, correlate the data with case histories, provide backward compatible surveillance data (i.e. PFGE fingerprints), and facilitate epidemiological tracking.
Project Owner	Epidemiologist (with some training in genome analysis)
Process Domain Integration	This project integrates all of the analysis activities and provides for additional processing steps to deliver epidemiologically useful data. Significant change to current process activities will be required as well as adding new process activities. The algorithms changes may impact upstream process activities as well.
Organizational Domain Integration	Coordination with the "change owners" or "troubleshooting representative" in each area will be essential as the algorithm changes will require close coordination with upstream process activities and the individuals responsible for them particularly in the molecular biology areas.
Information Domain Integration	A process review and integration document will need to exist in order to ensure all parts of the process were aware of the potential changes, could react to them, or confirm no changes needed in any of the explicit work instructions (protocols) etc. In addition, visibility and understanding of the necessary algorithm changes and types of data needed must be communicated to the rest of the process perhaps through online documentation, etc.

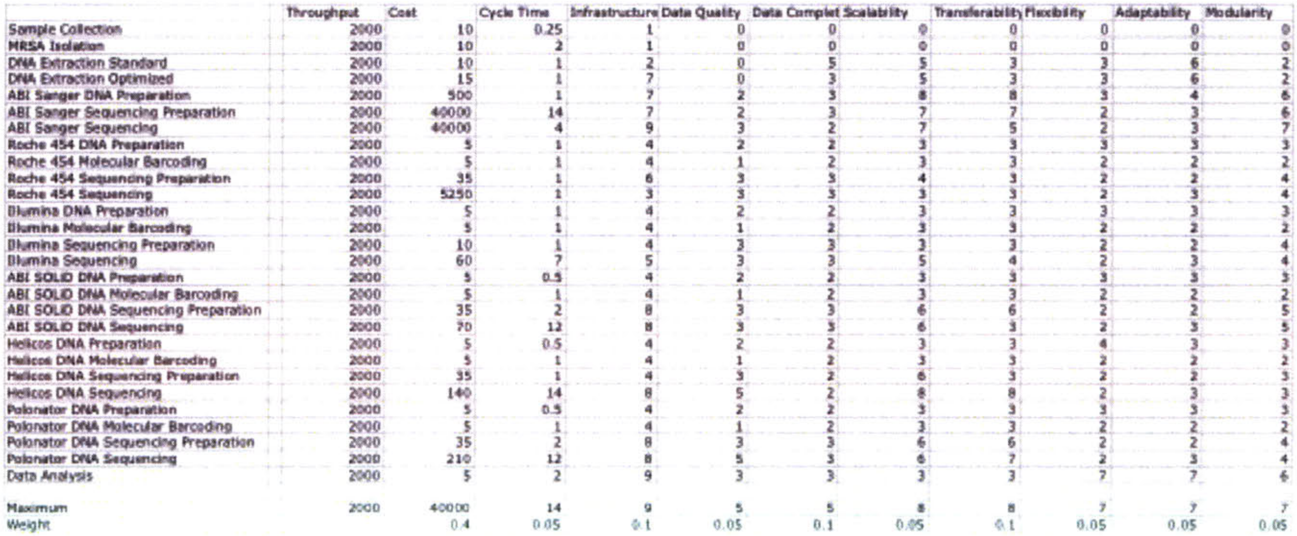
Project #5	Build Data Analysis Pipeline
Process Testing Needed	Extensive testing with reference (i.e. known MRSAs) will be needed to ensure the process is robust and delivers the correct information. Tests to verify the performance of the analysis portion of the process under potential upstream failure modes will be important to ensure the algorithms and software are robust, but also that the overall process can deal with both normal process variation and "out of specification" conditions.

Project #6	Verify Biology
Description	An essential part of the surveillance process must be to "close the loop" with the biology, which is to compare actual biological results (phenotype) with the DNA data (genotype) being produced. For example, measuring the Minimum Inhibitory Concentration (MIC) of antibiotics for MRSA strains found to have mutations, is important to develop an understanding of which mutations are important and which are not. Other biological tests are also needed such as growth rates, cell wall thickness, and perhaps virulence measures. Correlation of the phenotype with the genotype will help realize enormous value from the surveillance process as new diagnostics can then be developed based on these correlations. This will require going back to samples stored in freezers as a product of the MRSA isolation portion of the process.
Internal Stakeholders	Process Owner, Process Change Owners, Bio-Informatics Analysis Staff, Genomic Data Analysis Research Groups, Bacterial Genomics Researchers, Infectious Disease Researchers, Epidemiologists, Physicians, Clinical Microbiology Researchers
External Stakeholders	CDC, IDSA, NIH, ASM, State Health Agencies, Genomic Data Analysis Vendors, Bacterial Diagnostics Vendors, Hospital Administration, Antibiotics Researchers, Antibiotics Vendors
Deliverables	Database showing the correlation of genotype data from the surveillance system to phenotype data obtained from strains.
Project Owner	Bacterial Genomics Researcher
Process Domain Integration	This project adds additional process steps at the MRSA isolation step, where an isolate would be grown and subjected to a variety of phenotype assays such that these results could be later correlated. These additional process activities would occur in parallel. And are not

Project #6	Verify Biology
	directly connected to downstream processing other than for the correlation of the data and integration into a database.
Organizational Domain Integration	Coordination with other parts of the process is not as essential for this activity, but some information may be useful, such as cell wall properties, and the owner of the project should connect in a reporting mode to "change owners".
Information Domain Integration	A process review document should exist for the project, which is made visible to the rest of the process, but most importantly the data format and sample tracking should integrate with the downstream genotype analysis to build a database containing both sets of data.
Process Testing Needed	Limited process testing is needed in that this will be a "parallel" process other than to verify sample tracking accuracy (i.e. making sure the same sample is tested for phenotype and genotype. Validation of the accuracy of the phenotype assays is likely necessary to standardize them across locations, and also to understand the tolerance of the data (error bounds).

Appendix 2 MATLAB Routines

```
%Process network variable calculator
%This program calculates the key graph metrics for process networks RNicol
addpath('/MATLAB/matlab_bgl','/MATLAB/Thesis');
process_network=xlsread('/MATLAB/Thesis/baseline_wPO.xls');
%geodesic_distance=graphallshortestpaths(sparse(process_network),'Directed',true)
geodesic_distance=all_shortest_paths(sparse(process_network));
edges=num_edges(process_network)
vertices=num_vertices(process_network)
avg_dist=(1/((vertices)*(vertices-1)))*sum(sum(geodesic_distance))
glob_efficiency=(1/((vertices)*(vertices-
1)))*(sum(sum(ones(vertices,vertices)./(geodesic_distance+eye(vertices))))-vertices)
undirected_density=(2*edges)/(vertices*(vertices-1))
%gcoor=circle_graph_layout(process_network)
%gplot(process_network,gcoor)
for k=1:1:vertices;
  minus_vertex_geodesic_dist=geodesic_distance;
  minus_vertex_geodesic_dist(k,:)= [];
  minus_vertex_geodesic_dist(:,k)= [];
  vertex_eff=(1/((vertices-1)*(vertices-2)))*(sum(sum(ones(vertices-1,vertices-
1)./(minus_vertex_geodesic_dist+eye(vertices-1))))-(vertices-1));
  vulnerability(1,k)=(glob_efficiency-vertex_eff)/glob_efficiency;
end;
Network_Vulnerability=max(vulnerability)
degree(1,:)=sum(process_network,1); % in degree, sum of columns
degree(2,:)=sum(process_network,2); % out degree, sum of rows
degree(3,:)=degree(1,:)+degree(2,:);
degree_bins=[1:1:max(degree(3,:))];
degree_dist=histc(degree(3,:),degree_bins);
%Entropy
degree_nonzeros=nonzeros(degree_dist);
deg_dist_entropy=0;
for k=1:1:size(degree_nonzeros);
  deg_dist_entropy=-((degree_nonzeros(k)/vertices)*log2(degree_nonzeros(k)/vertices))+deg_dist_entropy;
end;
```



	Throughput	Cost	Cycle Time	Infrastructure	Data Quality	Data Complet	Scalability	Transferability	Flexibility	Adaptability	Modularity
Sample Collection	2000	10	0.25	1	0	0	0	0	0	0	0
MRSA Isolation	2000	10	2	1	0	0	0	0	0	0	0
DNA Extraction Standard	2000	10	1	2	0	5	5	3	3	6	2
DNA Extraction Optimized	2000	15	1	7	0	3	5	3	3	6	2
ABI Sanger DNA Preparation	2000	500	1	7	2	3	8	8	3	4	6
ABI Sanger Sequencing Preparation	2000	40000	14	7	2	3	7	7	2	3	6
ABI Sanger Sequencing	2000	40000	4	9	3	2	7	5	2	3	7
Roche 454 DNA Preparation	2000	5	1	4	2	2	3	3	3	3	3
Roche 454 Molecular Barcoding	2000	5	1	4	1	2	3	3	2	2	2
Roche 454 Sequencing Preparation	2000	35	1	6	3	3	4	3	2	2	4
Roche 454 Sequencing	2000	5250	1	3	3	3	3	3	2	3	4
Illumina DNA Preparation	2000	5	1	4	2	2	3	3	3	3	3
Illumina Molecular Barcoding	2000	5	1	4	1	2	3	3	2	2	2
Illumina Sequencing Preparation	2000	10	1	4	3	3	3	3	2	2	4
Illumina Sequencing	2000	60	7	5	3	3	5	4	2	3	4
ABI SOLiD DNA Preparation	2000	5	0.5	4	2	2	3	3	3	3	3
ABI SOLiD DNA Molecular Barcoding	2000	5	1	4	1	2	3	3	2	2	2
ABI SOLiD DNA Sequencing Preparation	2000	35	2	8	3	3	6	6	2	2	5
ABI SOLiD DNA Sequencing	2000	70	12	8	3	3	6	3	2	3	5
Helicos DNA Preparation	2000	5	0.5	4	2	2	3	3	4	3	3
Helicos DNA Molecular Barcoding	2000	5	1	4	1	2	3	3	2	2	2
Helicos DNA Sequencing Preparation	2000	35	1	4	3	2	6	3	2	2	3
Helicos DNA Sequencing	2000	140	14	8	5	2	8	8	2	3	3
Polonator DNA Preparation	2000	5	0.5	4	2	2	3	3	3	3	3
Polonator DNA Molecular Barcoding	2000	5	1	4	1	2	3	3	2	2	2
Polonator DNA Sequencing Preparation	2000	35	2	8	3	3	6	6	2	2	4
Polonator DNA Sequencing	2000	210	12	8	5	3	6	7	2	3	4
Data Analysis	2000	5	2	9	3	3	3	3	7	7	6
Maximum	2000	40000	14	9	5	5	8	8	7	7	7
Weight		0.4	0.05	0.1	0.05	0.1	0.05	0.1	0.05	0.05	0.05

Bibliography

Achtman, M. and M. Wagner (2008). "Microbial diversity and the genetic nature of microbial species." Nature Reviews Microbiology **6**(6): 431-440.

Anderson, R. M. (1991). "Discussion: the Kermack-McKendrick epidemic threshold theorem." Bulletin of mathematical biology **53**(1): 1-32.

Barabási, A., Z. Dezsõ, et al. (2003). Scale-free and hierarchical structures in complex networks. AIP Conference Proceedings.

Barber, M. (1961). "Methicillin-resistant staphylococci." Journal of Clinical Pathology **14**(4): 385.

Bartolomei, J. E. (2007). "Qualitative knowledge construction for engineering systems: extending the design structure matrix methodology in scope and procedure."

Bartolomei, J. E., D. E. Hastings, et al. (2006). Screening for Real Options" In" an Engineering System: A Step Towards Flexible System Development--PART I: The Use of Design Matrices to Create an End-to-End Representation of a Complex Socio-Technical System. INCOSE Conference on System Engineering Research, Los Angeles, CA.

Basole, R. C. and W. B. Rouse (2008). "Complexity of service value networks: conceptualization and empirical investigation." IBM Systems Journal **47**(1): 53.

Bass, F. M. (1969). "A New Product Growth Model for Consumer Durables." Management Science **15**: 215-227.

Blazquez, J. and G. M. Eliopoulos (2003). "Hypermutation as a factor contributing to the acquisition of antimicrobial resistance." Clinical Infectious Diseases **37**(9): 1201-1209.

Borgatti, S. P. (2005). "Centrality and network flow." Social Networks **27**(1): 55-71.

Borgatti, S. P. (2006). "Identifying sets of key players in a social network." Computational & Mathematical Organization Theory **12**(1): 21-34.

Borgatti, S. P., M. G. Everett, et al. (2002). "Ucinet for Windows: Software for social network analysis." Harvard: Analytic Technologies.

Box, G. (1999). "Statistics as a Catalyst to Learning by Scientific Method. Part II- A Discussion." JOURNAL OF QUALITY TECHNOLOGY **31**(1): 16-29.

Box, G. and P. Liu (1999). Statistics as a Catalyst to Learning by Scientific Method. Part I An Example. JOURNAL OF QUALITY TECHNOLOGY. **31:** 1-15.

Bradley, J. A. and A. A. Yassine "ON THE USE OF NETWORK ANALYSIS IN PRODUCT DEVELOPMENT TEAMS." Urbana **51**: 61801.

Bradley, J. A. and A. A. Yassine (2008). "A Multi-Domain Analysis Framework for Product Development." Urbana **51**: 61801.

Browning, T. and R. Ramasesh (2007). A survey of activity network-based process models for managing product development projects. Production and Operations Management. **16:** 217-240.

Butler, J., I. MacCallum, et al. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Research.

Cardoso, J., J. Mendling, et al. (2006). A discourse on complexity of process models. LECTURE NOTES IN COMPUTER SCIENCE. **4103:** 117.

Carleton, H. A., B. A. Diep, et al. (2004). "Community-adapted methicillin-resistant Staphylococcus aureus (MRSA): population dynamics of an expanding community reservoir of MRSA." The Journal of infectious diseases **190**(10): 1730-1738.

Cegelski, L., G. Marshall, et al. (2008). The biology and future prospects of antivirulence therapies. Nat Rev Micro.

Chambers, H. (2005). Community-Associated MRSA-Resistance and Virulence Converge. New England Journal of Medicine.

Chang, S., D. M. Sievert, et al. (2003). "Infection with vancomycin-resistant Staphylococcus aureus containing the vanA resistance gene." The New England Journal of Medicine **348**(14): 1342.

Christoffersen, R. (2006). Antibiotics—an investment worth making? Nat Biotechnol. **24:** 1512-1514.

Clatworthy, A. E., E. Pierson, et al. (2007). Targeting virulence: a new paradigm for antimicrobial therapy. Nature Chemical Biology. **3:** 541-548.

Costa, L. d. F., F. A. Rodrigues, et al. (2005). Characterization of complex networks: A survey of measurements. arXiv. **cond-mat.dis-nn**.

Curtis, T. P., W. T. Sloan, et al. (2002). "Estimating prokaryotic diversity and its limits." Proceedings of the National Academy of Sciences of the United States of America **99**(16): 10494.

Davenport, T. H. (1993). Process innovation: reengineering work through information technology, Harvard Business School Pr.

Denamur, E. and I. Matic (2006). "Evolution of mutation rates in bacteria." Molecular Microbiology **60**(4): 820.

Diep, B., S. Gill, et al. (2006). Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. The Lancet.

Diep, B. and M. Otto (2008). The role of virulence determinants in community-associated MRSA pathogenesis. Trends in Microbiology.

Dodder, R. S., J. B. McConnell, et al. (2006). "The CLIOS Process." Unpublished Manuscript.

Dodder, R. S., J. M. Sussman, et al. (2004). The Concept of the CLIOS Process: Integrating the Study of Physical and Policy Systems Using Mexico City as an Example. Massachusetts Institute of Technology Engineering Systems Symposium, Cambridge, MA.

Dodds, P., D. Watts, et al. (2003). Information exchange and the robustness of organizational networks. Proceedings of the National Academy of Sciences.

Dong, Q. (2002). Predicting and managing system interactions at early phase of the product development process. Mechanical Engineering, Massachusetts Institute of Technology. **PhD**.

Dori, D. (2002). Object-process methodology: a holistics systems paradigm, Springer.

Enright, M. (2003). The evolution of a resistant pathogen – the case of MRSA. Current Opinion in Pharmacology. **3:** 474-479.

Eppinger, S. D., D. E. Whitney, et al. (1990). "Organizing the tasks in complex design projects." Lecture Notes in Computer Science **492/1991**: 229-252.

Feldmann, C. G. (1998). The practical guide to business process reengineering using IDEF0, Dorset House Publishing Co., Inc. New York, NY, USA.

Filler, A. G. (2009). "The History, Development and Impact of Computed Imaging in Neurological Diagnosis and Neurosurgery: CT, MRI, and DTI." Nature Precedings.

Fine, C. H. (1998). Clockspeed: Winning industry control in the age of temporary advantage, Basic Books.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496.

Freeman, L. C. (1979). "Centrality in social networks conceptual clarification." Social Networks **1**(3): 215-239.

Frizelle, G. and E. Woodcock (1995). "Measuring complexity as an aid to developing operational strategy." International Journal of Operations and Production Management **15**: 26-26.

Giaglis, G. M. (2001). "A taxonomy of business process modeling and information systems modeling techniques." International Journal of Flexible Manufacturing Systems **13**(2): 209-228.

Giffin, M., O. de Weck, et al. (2009). "Change propagation analysis in complex technical systems." Journal of Mechanical Design **131**: 081001.

Gilbreth, F. B. and L. M. Gilbreth (1921). Process Charts: First Steps in Finding the One Best Way to Do Work. ASME Annual Conference.

Goldstine, H. H. and J. Von Neumann (1947). Planning and coding of problems for an electronic computing instrument, Institute for Advanced Study.

Gonzales, R., D. Malone, et al. (2001). Excessive Antibiotic Use for Acute Respiratory Infections in the United States. CLIN INFECT DIS.

Gould, I. (2006). Community-acquired MRSA: can we control it? The Lancet. **368:** 824-826.

Graham, B. B. (2004). Detail process charting: speaking the language of process, Wiley.

Granovetter, M. S. (1973). "The strength of weak ties." American journal of sociology **78**(6): 1360.

Gruhn, V. and R. Laue (2006). Complexity metrics for business process models. 9th international conference on business information.

Hammer, M. and J. Champy (2003). Reengineering the corporation: A manifesto for business revolution, Collins Business.

Hammer, M. and S. Stanton (1999). "How process enterprises really work." Harvard Business Review **77**: 108-120.

Harbarth, S. (2006). Control of endemic methicillin-resistant Staphylococcus aureus-recent advances and future challenges. Clinical Microbiology & Infection. **12:** 1154.

Heller, M., D. Jäger, et al. (2004). A management system for dynamic and interorganizational design processes in chemical engineering. Computers and Chemical Engineering. **29:** 93-111.

Higuchi, R., G. Dollinger, et al. (1992). "Simultaneous amplification and detection of specific DNA sequences." Bio/Technology **10**(4): 413-417.

Hiramatsu, K. (2004). Elucidation of the Mechanism of Antibiotic Resistance Acquisition of Methicillin-Resistant Staphylococcus Aureus (MRSA) and Determination of Its Whole Genome Nucleotide Sequence JMAJ-Japan Medical Association Journal.

Howell, A., K. Hanson, et al. (2002). "Engineering to business: Optimizing asset utilization through process engineering." Chemical engineering progress **98**(9): 54-63.

Hutchison, C. A. (2007). DNA sequencing: bench to bedside and beyond. Nucleic Acids Research. **35:** 6227-6237.

J. Carrington, P., J. Scott, et al. (2005). Models and methods in social network analysis, Cambridge University Press.

Jensen, P. A. and J. F. Bard (2003). Operations research: models and methods, John Wiley & Sons Inc.

Kennedy, A., M. Otto, et al. (2008). Epidemic community-associated methicillin-resistant Staphylococcus aureus: Recent clonal expansion and diversification. Proceedings of the National Academy of Sciences.

Kennedy, M. (2004). A brief history of disease, science and medicine: from the ice age to the genome project, Writers' Collective.

Kermack, W. O. and A. G. McKendrick (1991). "Contributions to the mathematical theory of epidemics. Further studies of the problem of endemicity." Bulletin of mathematical biology **53**(1): 89-118.

Klevens, R., M. Morrison, et al. (2007). Invasive Methicillin-Resistant Staphylococcus aureus Infections in the United States. JAMA.

Klevens, R. M., M. A. Morrison, et al. (2007). "Invasive methicillin-resistant Staphylococcus aureus infections in the United States." JAMA: The Journal of the American Medical Association **298**(15): 1763.

Kristiansen, J. E., O. Hendricks, et al. (2007). Reversal of resistance in microorganisms by help of non-antibiotics. Journal of Antimicrobial Chemotherapy. **59:** 1271-1279.

Kuroda, M., T. Ohta, et al. (2001). "Whole genome sequencing of meticillin-resistant Staphylococcus aureus." The Lancet.

Larsen, A. R., R. Goering, et al. (2009). "Two distinct clones of MRSA with the same USA300 PFGE profile-A potential pitfall for identification of USA300 CA-MRSA." Journal of Clinical Microbiology.

Latora, V. and M. Marchiori (2001). Efficient Behavior of Small-World Networks. arXiv. **cond-mat**.

Latora, V. and M. Marchiori (2005). "Vulnerability and protection of infrastructure networks." Physical Review E **71**(1): 15103.

Lewis, T. G. (2009). Network Science: Theory and Applications, Wiley.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome research **18**(11): 1851.

Lindsay, J. and M. Holden (2004). Staphylococcus aureus: superbug, super genome? Trends in Microbiology. **12:** 378-385.

Liu, W. and K. Fang (2006). Using IDEF0/Petri net for ontology-based task knowledge analysis: The case of emergency response for debris-flow. HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES. **39:** 76.

Mahalanabis, M., H. Al-Muayad, et al. (2009). "Cell lysis and DNA extraction of gram-positive and gram-negative bacteria from whole blood in a disposable microfluidic chip." Lab on a Chip **9**(19): 2811-2817.

Maurer, M. and U. Lindemann (2008). The Application of the Multiple-Domain Matrix. IEEE International Conference on Systems, Man and Cybernetics 2008, Singapore.

Maurer, M. S. (2007). Structural Awareness in Complex Product Design, Verlag Dr. Hut.

McCabe, T. (1976). "A complexity measure." IEEE Transactions on software Engineering: 308-320.

McDougal, L. K., C. D. Steward, et al. (2003). "Pulsed-field gel electrophoresis typing of oxacillin-resistant Staphylococcus aureus isolates from the United States: establishing a national database." Journal of Clinical Microbiology **41**(11): 5113.

Medini, D., C. Donati, et al. (2005). The microbial pan-genome. Current Opinion in Genetics & Development. **15:** 589-594.

Mellmann, A., A. W. Friedrich, et al. (2006). "Automated DNA sequence-based early warning system for the detection of methicillin-resistant Staphylococcus aureus outbreaks." PLoS Medicine **3**(3): 348.

Mendling, J. (2006). Testing density as a complexity metric for EPCs. German EPC workshop on density of process models.

Mendling, J. (2009). Metrics for Process Models, Springer.

Milgram, S. (1967). "The small world problem." Psychology today **2**(1): 60-67.

Monecke, S., B. Berger-Bachi, et al. (2007). Comparative genomics and DNA array-based genotyping of pandemic Staphylococcus aureus strains .... Clinical Microbiology & Infection.

Mullis, K. B. and F. A. Faloona (1987). "Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction." Methods in enzymology **155**: 335.

Mwangi, M., S. Wu, et al. (2007). "Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing." Proceedings of the National Academy of Sciences **104**(22): 9451-9456.

Nagl, M., B. Westfechtel, et al. (2003). Tool support for the management of design processes in chemical engineering. Computers and Chemical Engineering. **27:** 175-197.

Nathan, C. (2004). "Antibiotics at the crossroads." Nature **431**(7011): 899-902.

Nightingale, D. (2009). Principles of Enterprise Systems. Second International Symposium on Engineering Systems.

Nightingale, D. J. and D. H. Rhodes (2004). Enterprise systems architecting: Emerging art and science within engineering systems. MIT Engineering Systems Symposium 2004.

Norton, J. A. and F. M. Bass (1987). "A diffusion theory model of adoption and substitution for successive generations of high-technology products." Management Science **33**(9): 1069-1086.

Noskin, G. A., R. J. Rubin, et al. (2007). "National trends in Staphylococcus aureus infection rates: impact on economic burden and mortality over a 6-year period (1998-2003)." Clinical Infectious Diseases **45**(9): 1132-1140.

Osorio, C. A., D. Dori, et al. (2009). "COIM: AN OBJECT-PROCESS BASED METHOD FOR ANALYZING ARCHITECTURES OF COMPLEX, INTERCONNECTED, LARGE-SCALE SOCIO-TECHNICAL SYSTEMS."

Owen-Smith, J. and W. W. Powell (2004). "Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community." Organization Science **15**(1): 5-21.

Porter, R. (2001). The Cambridge illustrated history of medicine, Cambridge Univ Pr.

Qiagen (2001). "QIAGEN genomic DNA handbook." Wiena: Qiagen.

Queck, S., M. Jameson-Lee, et al. (2008). "Quorum-Sensing System: Insight into the Evolution of Virulence Regulation in Staphylococcus aureus." Molecular Cell.

Rogers, E. M. (1995). Diffusion of innovations, Free press.

Rosvall, M., A. Gronlund, et al. (2005). Searchability of Networks. arXiv. **cond-mat.dis-nn**.

Rosvall, M., P. Minnhagen, et al. (2004). Navigating Networks with Limited Information. arXiv. **cond-mat.dis-nn**.

Rosvall, M. and K. Sneppen (2006). Networks and our limited information horizon. Arxiv preprint cond-mat/0604036.

Sargent, R. (2005). Process systems engineering: A retrospective view with questions for the future. Computers and Chemical Engineering. **29:** 1237-1241.

Shannon, C. E. and W. Weaver (1948). "A mathematical theory of communication." Bell Syst. Tech. J **27**: 379-423.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Smaling, R. and O. de Weck (2007). Assessing risks and opportunities of technology infusion in system design Note: An early version of this manuscript was presented as a conference paper: RM Smaling and OL de Weck, Fuzzy Pareto frontiers in multidisciplinary system architecture analysis, AIAA-2004-4553, 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Albany, NY, August 30-September 1, 2004. Syst. Engin. **10**.

Smaling, R. M. (2005). "System architecture analysis and selection under uncertainty."

Sneppen, K., A. Trusina, et al. (2004). Hide and seek on complex networks. arXiv. **cond-mat.dis-nn**.

Soulsby, L. (2007). "Antimicrobials and animal health: a fascinating nexus." Journal of Antimicrobial Chemotherapy **60**(5): 1184.

Spear, S. J. (2005). "Fixing Health Care from the Inside, Today." Harvard Business Review **83**(9): 78-91.

Spear, S. J. (2008). Chasing the Rabbit, McGraw-Hill.

Staccini, P., M. Joubert, et al. (2001). Modelling health care processes for eliciting user requirements: a way to link a quality paradigm and clinical information system design. International Journal of Medical Informatics. **64:** 129-142.

Statistics, N. C. f. H. (2009). Health, United States, 2008.

Sterman, J. (2002). All models are wrong: reflections on becoming a systems scientist. System Dynamics Review. **18:** 501-531.

Steward, D. V. (1965). "Partitioning and tearing systems of equations." Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis **2**(2): 345-365.

Sussman, J. M., S. P. Sgouridis, et al. (2005). "New Approach to Transportation Planning for the 21st Century: Regional Strategic Transportation Planning as a Complex Large-Scale Integrated Open System." Transportation Research Record: Journal of the Transportation Research Board **1931**(-1): 89-98.

Tokars, J. I., C. Richards, et al. (2004). "The changing face of surveillance for health care-associated infections." Clinical Infectious Diseases **39**(9): 1347-1352.

Utterback, J. M. (1996). Mastering the dynamics of innovation, Harvard Business School Pr.

Van der Aalst, W. M. P. (1998). "The application of Petri nets to workflow management." Journal of Circuits Systems and Computers **8**: 21-66.

Walsh, C. (2003). Antibiotics: actions, origins, resistance, ASM Press.

Walsh, C. (2003). Where will new antibiotics come from? Nature Reviews Microbiology. **1:** 65-70.

Warfield, J. N. (1973). "Binary Matrices in System Modeling." IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS **3**(5).

Wasserman, S. and K. Faust (1994). Social network analysis: methods and applications, Cambridge University Press.

Watts, D. J. (1999). "Networks, Dynamics, and the Small-World Phenomenon." American journal of sociology **105**(2): 493-527.

Watts, D. J. (2003). Small worlds: the dynamics of networks between order and randomness, Princeton Univ Pr.

Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of small-world networks." Nature **393**(6684): 440-442.

Winston, W. L. and S. C. Albright (2008). Practical management science, South-Western Pub.

Womack, J. P., D. T. Jones, et al. (1990). The machine that changed the world, Rawson Associates New York.

Zerbino, D. R. and E. Birney (2008). "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821.

Zhu, X., S. Hu, et al. (2008). Modeling of manufacturing complexity in mixed-model assembly lines. Journal of Manufacturing Science and Engineering. **130:** 051013.