

Interest Networks: Understanding the Influence of Interesting People in an Organization

by

Julia Shuhong Ma

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 11, 2012, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

This thesis applies network theory to firms, their employees, and various aspects of the employees to understand diversity within an industry at both the firm-level and employee-level. We hypothesize that the interest diversity of a firm's employees can influence that firm's economic performance and growth. Using the LinkedIn API, we are able to collect ~600 employees (past and present) for 43 companies using a keyword search. This data is used to create a visualization of people's interests, an "Interest Space", which is a network graph of how interests are categorized and linked to each other. By analyzing the data from firms associated with the Media Lab, we begin to understand how individual interests affect success among a small network of companies. We research this through case studies of a few companies by analyzing their financial data and interests of their employees.


Thesis Advisor: Andrew B. Lippman

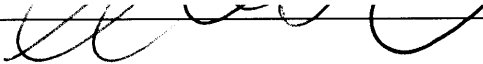
Title: Senior Research Scientist of Media Arts and Sciences, Program in Media Arts and Sciences


**Interest Networks: Understanding the Influence of
Interesting People in an Organization**

by

Julia Shuhong Ma

Thesis Advisor  _____
Andrew B. Lippman
Senior Research Scientist of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader  _____
Cesar A. Hidalgo
Assistant Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader  _____
Sepandar D. Kamvar
Associate Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Contents

Abstract	3
1 Introduction	13
1.1 Motivation	13
1.2 Contribution	14
1.3 Thesis Overview	16
2 Background	19
2.1 Previous Work from the Media Lab	19
2.2 Diversity and Organizations	22
2.3 Network Science Applied to Ecosystems	25
2.4 Social Networks for Research	27
2.5 Hypothesis	28
3 Data Collection	29
3.1 Methodology	29
3.2 Using the API	30
3.2.1 Choosing the Data Set	30
3.2.2 API Overview	31
3.2.3 Query Structure	33
3.2.4 Limitations of the API	34
3.3 Biases in the Data Set	35
3.4 A Spammer’s Profile	37
3.5 Extracting Relevant Data	37
4 Creating the Interest Space	39
4.1 Interest Popularity	39
4.2 Links Between Interests	43
4.3 Categorizing Interests	45
4.3.1 What is Divisi	46
4.3.2 Using Divisi to Categorize Interests	47
4.3.3 Manually Categorizing Interests	49
4.3.4 Categories	50

4.4	The Interest Space	51
5	Other Profile Data	55
5.1	Profile Completeness	55
5.2	Education	57
5.3	Age	58
5.4	Location	59
5.5	Skills	60
5.6	Languages	61
5.7	Positions	62
6	Case Studies	65
6.1	Research In Motion	65
6.1.1	The Company	65
6.1.2	Performance	66
6.1.3	Profiles and Interests	69
6.1.4	Discussion	72
6.2	Schneider Electric	73
6.2.1	The Company	73
6.2.2	Performance	74
6.2.3	Profiles and Interests	76
6.2.4	Discussion	78
6.3	Intel Corporation	79
6.3.1	The Company	79
6.3.2	Performance	80
6.3.3	Profiles and Interests	82
6.3.4	Diversity at Intel	84
6.3.5	Discussion	84
6.4	Summary	85
7	Conclusions and Future Work	87
7.1	LinkedIn as Research	87
7.2	The Interest Space	89
7.3	Company Interests	93
7.4	Final Thoughts	94

List of Figures

1-1	The Interest Space visualization	15
2-1	Social Energy visualization of the heat distribution	20
2-2	Social Energy graph showing the temperature fluctuation	21
2-3	The Glass Infrastructure	22
2-4	The Product Space	26
3-1	Screenshot of a LinkedIn profile	31
3-2	Screenshot of additional information on a LinkedIn profile	32
3-3	Screenshot of a spammer’s profile	38
4-1	Distribution of interests among profiles	40
4-2	Log binning graph of interest distribution	41
4-3	Interest popularity	42
4-4	Log binning graph of interest popularity	43
4-5	The Interest Space visualization	52
4-6	The interest categories are indicated by the color of the node.	52
4-7	Exploring the Interest Space	53
4-8	The Interest Space search function	54
4-9	The Interest Space neighboring interests	54
6-1	Research In Motion stock comparison chart	67
6-2	Research In Motion patent history comparison	68
6-3	Interest graph for Research In Motion	71
6-4	Schneider Electric stock comparison chart	74
6-5	Schneider Electric patent history comparison	75
6-6	Interest graph for Schneider Electric	77
6-7	Intel Corporation stock chart	80
6-8	Intel Corporation stock comparison chart	80
6-9	Intel Corporation patent history comparison	81
6-10	Interest graph for Intel	83
7-1	The Interest Space visualization	90
7-2	The main cluster of interests	91

7-3	The sports cluster of interests	91
7-4	Smaller clusters of interests	92

List of Tables

3.1	Base URLs for LinkedIn API	33
4.1	Top interests by popularity	41
4.2	Interests appearing once	42
4.3	Top linked pairs of interests	44
4.4	Similarity measures between interest pairs	48
4.5	Spread activation measures between interest pairs	48
4.6	Initial interest categories	49
4.7	Final interest categories	51
5.1	Number of profiles retrieved for the Media Lab sponsor companies . .	56
5.2	Number of profiles with information in a particular field	57
5.3	Most common profile locations	59
5.4	Most common profile locations for profiles that list interests	60
5.5	Top skills by popularity	61
5.6	Top languages by popularity	62
5.7	Companies of which we have the most profiles	63
6.1	Top interests of Research In Motion employees	69
6.2	Top pairs of interests for Research In Motion employees	70
6.3	Other companies for which Research In Motion employees have worked	71
6.4	The years of which Research In Motion employees started their positions	72
6.5	Top interests of Schneider Electric employees	76
6.6	Top pairs of interests for Schneider Electric employees	76
6.7	Other companies for which Schneider Electric employees have worked	77
6.8	The years which Schneider Electric employees started their positions .	78
6.9	Top interests of Intel employees	82
6.10	Top pairs of interests for Intel employees	82
6.11	Other companies for which Intel employees have worked	82
6.12	The years which Intel employees started their positions	83

Thank you Andy, my advisor, for your guidance, support, and inspirational thought quirks that made me think and laugh at the same time.

Thank you Cesar and Sep, my readers, for your advice and feedback. Your work is inspiringly beautiful and beautifully inspiring.

Thank you Mom and Dad, for your encouragement in all of my pursuits.

Thank you Elisha, for being the best and only sister, and always willing to proofread my writing.

Thank you Eyal, for your constant support and love. It's amazing to have someone who can give not only emotional support but also intellectual support.

Thank you Sherwin, my UROP, for keeping me excited about my project and making the lovely visualized Interest Space.

Thank you Viral Spaces (Matt Blackshaw, Boris Kizelshteyn, Kwan Hong Lee, Inna Lobel, Travis Rich, Dawei Shen, Eyal Toledano, Grace Woo, Polychronis Ypodimatopoulos), for being hands-down the best group in the Media Lab.

Thank you to all my friends inside and outside the Media Lab.

Now to become a Maestro of Science!!

Chapter 1

Introduction

We have become not a melting pot but a beautiful mosaic.

Different people, different beliefs, different yearnings,

different hopes, different dreams.

-Jimmy Carter

1.1 Motivation

Between 1920 and 1980, Bell Labs of AT&T was considered one of the most innovative organizations in the world [26]. From the transistor to the solar cell to the C programming language to cellular communications, the inventions that came out of Bell Labs changed the course of technology. These technologies are not just incremental research innovations but rather revolutionary technology that either disrupted an existing market economy or created a new one. We ask, what kind of environment enables a group of people to endlessly innovate? How do we measure and quantify the types of people that contribute to these kind of firms, and how do we justify increasing this quantity in other organizations?

Scott Page - a professor of complex systems, political science, and economics at the University of Michigan - introduces the groundbreaking idea that *diversity* breeds innovation. His book The Difference [44] reveals how creative thinking comes from groups of people that think different from each other rather than just groups of smart people. Using mathematical models and case studies, he describes how and when cognitively diverse groups will outperform intelligent groups of people. Although many organizations, schools and firms, have already implemented diversity policies, Page's research not only justifies these policies but also extends the definition of diversity to include the ways in which people think.

Diversity manifests itself in a myriad of ways, and the aspect we focus on is what people enjoy in their leisure time, their hobbies, their interests. We explore how these interests are related to each other and whether or not two interests are likely to occur in the same person. A wide variety of interests can be tied to a person's cognitive diversity, and we hypothesize that interest diversity can also impact an organization's performance.

1.2 Contribution

This thesis has three contributions. The first is a feasibility analysis of using an online social network to understand various aspects of an organization. Online social networks are not just extremely popular, but they receive a lot of traffic and use. We expect that using a social network as a data set may become more common, and we look at the advantages and limitations of them. We specifically use LinkedIn as our data set and delve deeply into whether or not their information can be useful towards research. While data sets from social networks can be large, various aspects such as privacy settings can alter the completeness of the data.

The second contribution is our creation of the Interest Space (figure 1-1). We gather

a list of possible interests a person puts in their LinkedIn profile and use the interests' popularity and co-occurrence frequency to create a network graph that shows how these interests are related. The interests are also categorized using Divisi, which is a semantic network that represents common-sense knowledge. This graph lets us visualize how strongly different types of interests co-occur.

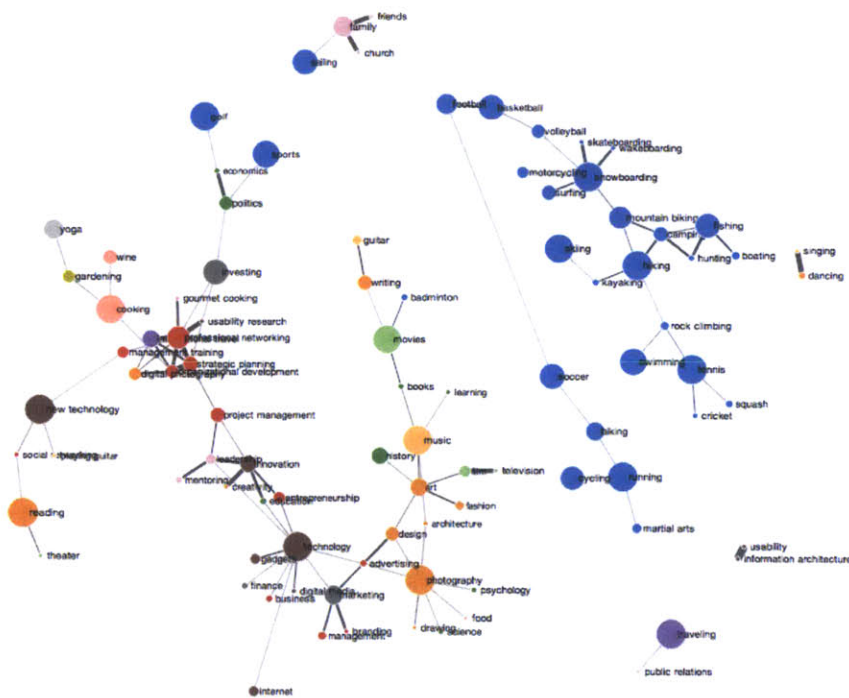


Figure 1-1: The Interest Space visualization

Lastly, we use the interest diversity of people working for particular firms to start to see when and how eclecticism affects an organization. We hypothesize that an organization can benefit from an increase in the interest diversity of their people, and we have evidence of this through case studies of a few companies.

1.3 Thesis Overview

The next chapter covers the background and previous work that led to this thesis. We discuss the related work in the Viral Spaces research group and how it guided our thinking. We highlight some relevant studies from the body of research of organizations and use their conclusions to justify our hypothesis and thought process. Since this thesis applies network theory to diversity in organizations, we cover previous work in both areas.

In chapter 3, we explain the data collection process. This thesis uses data gathered from LinkedIn's API, and we provide the method, an overview of the API usage, and limitations of the API that could affect our data. Then we describe the extraction and storage of relevant data and provide the online repository where all the data and scripts can be found.

Chapter 4 discusses the creation of the "Interest Space". After pulling out the interest nodes and links, we create a network map of the relatedness of the interests. The interests are not only linked together, they are also categorized using Divisi, a tool that allows users to find related concept and ideas in semantics. From this tool, we can measure relatedness/diversity of a set of interests, and use this measure to build the Interest Space, allowing us to visualize the realm of human activity where interests that are near each other are more related.

With such a large number of profiles, statistics and analysis on the data set are interesting to consider. Chapter 5 provides these statistics and data visualizations. From the geographic location distribution of profiles to the completeness of the profiles, we discuss various implications of the data and questions that may arise.

Although we have a large data set, the number of profiles for specific companies that list interests is still small. Since most companies in our data set do not have many

profiles, rather than doing a broad analysis of each company, we do case studies on a few companies that have a significant number of profiles. We analyze various aspects about the companies, from mission statements to financial data, in chapter 6.

Chapter 7 summarizes the thesis and provides conclusions and discussion on this work, and we conclude with ideas of future research that could be extended from this research.

Chapter 2

Background

A large body of research precedes this thesis. In this chapter, we summarize the previous work in this area and present the thought process that guided us to this work.

2.1 Previous Work from the Media Lab

This work is a part of the MIT Media Lab's Viral Spaces group [19], led by Dr. Andrew Lippman. Viral Spaces research focuses on creating technology for people to communicate both locally and at a distance. The group is interested in networks, not just in the social sense but also in a location-based sense, i.e. proximal networks. These projects explore ways for people to interact based on their intentions and devise systems that allow us to infer these intentions. These systems generate large amounts of data, and the challenge is to understand the data and present it to the users in ways that can affect their behavior. We describe a few of the Viral Spaces projects to illuminate the ideas that this thesis consists of.

Social Energy [18] is a browser-based project that visualizes real-time and historic thermostat data for the Media Lab building. By publicly displaying this information, we attempted to optimize people's behavior around the building to achieve energy efficiency. Multiple types of visualizations of the data (figure 2-1 and figure 2-2) were presented to the users to understand how people react to the same information in different formats. Although energy patterns were easily detected and users were able to see how their actions (e.g. opening an office window in summertime) affected a particular area, the greater impact was difficult to comprehend. Ultimately, this project tended more towards a diagnostic tool rather than an influence for social good. However, the themes of visualizing data in compelling ways and experimenting on the local environment pervade this thesis.

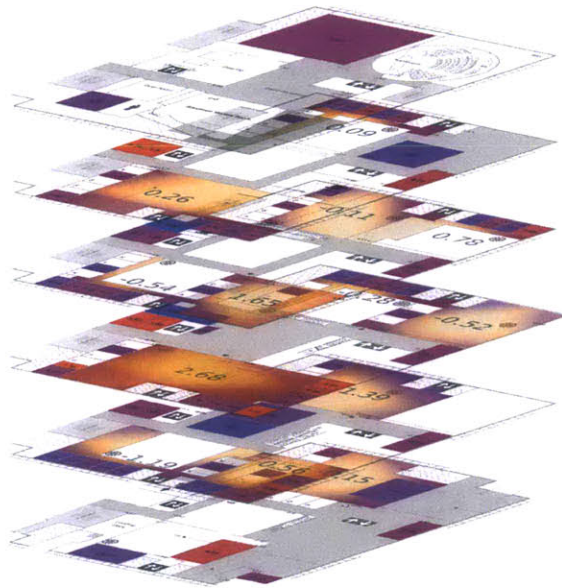


Figure 2-1: Social Energy visualization of the heat distribution between the different floors of the Media Lab building. The bluer areas indicate that the temperature is lower than the respective thermostat set point; the red areas indicate a higher temperature than the set point.

Social networks have a growing location-based aspect, and Viral Spaces explores these proximal networks. One such project, the Glass Infrastructure [27], is a social information window that consists of a networked set of large touch-screen monitors

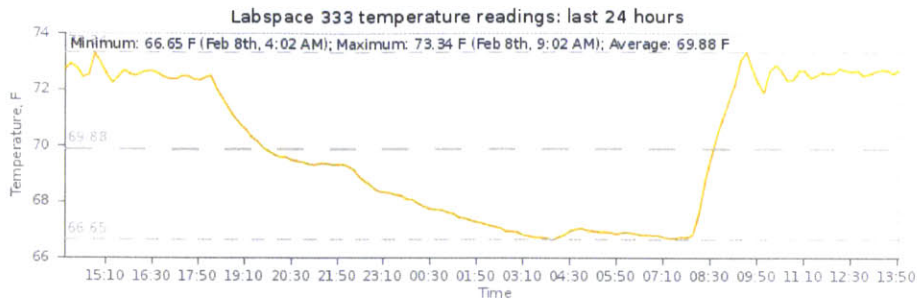


Figure 2-2: This Social Energy graph shows the temperature fluctuation over a 24 hr period.

with RFID tag readers that are placed in front of every lab area. Depending on a screen’s location, it displays a home page featuring the research groups of that lab space. People are able to browse through the research groups and learn about the people and the projects. The RFID readers allow the screen to detect the nearby users and can customize the display towards the user’s preferences. The system has a built-in recommendation system that uses Divisi (a tool we also use that facilitates reasoning by association over semantic networks) to categorize the projects by their descriptions. When a user indicates interest in particular projects by “charming” a project, these categories provide suggestions of similar projects that may be of interest to the user, and when two people simultaneously use the system, overlapping interests are highlighted on the display (see figure 2-3). The Glass Infrastructure connects people based on their location and their interests.

We intended to utilize the Glass Infrastructure for our data collection as users can indicate their interests through their charms, but the data was sparse and unorganized. The peak usage of the Glass Infrastructure occurred during a Media Lab sponsor event attended by hundreds of people. However, over a period of three days, only ~100 simultaneous usage events occurred, so we looked for other data sources for people’s interests.

This thesis analyzes the data from the professional social network LinkedIn, specifi-

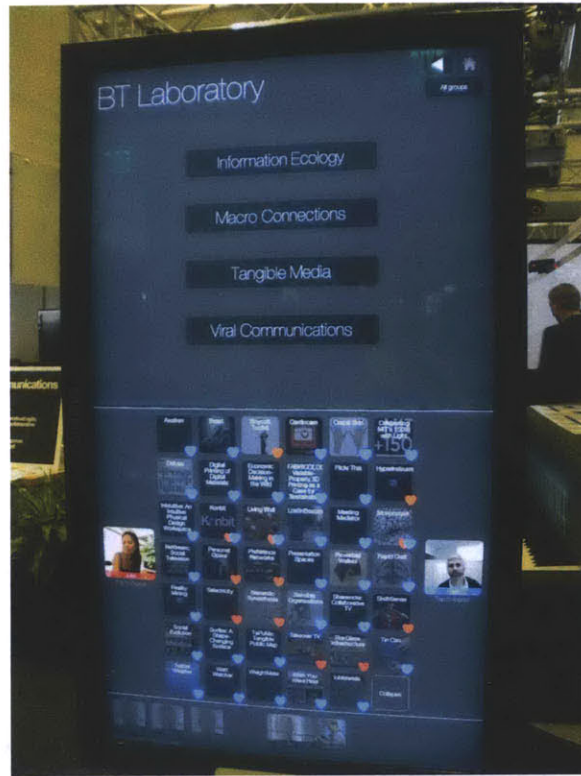


Figure 2-3: The Glass Infrastructure shows the common interests between people in the vicinity.

cally interests of people at particular companies. We not only visualize the connections between people's interests but also attempt to answer how people's interests and the diversity of these interests affect the organizations they are a part of.

2.2 Diversity and Organizations

The body of work in the field of organizational research and innovation is extensive. Here we give an overview of previous studies in this area and describe how this research has influenced our work.

There are two main classes of problems that people try to solve: well-defined problems that have a single guaranteed solution and ill-defined problems that have multiple non-

guaranteed solutions [35]. Prior to the 21st century, the most frequent way of solving problems was collecting a group of people with high IQs to work on the problem [36]. However, research has shown that this method is best used for well-defined problems [47]. Bigger problems are typically less well-defined, and rather than relying on just ability, it will require ingenuity and innovation to tackle these.

Joseph McGrath's model of group development [38] introduced the idea that different groups can reach the same outcome even if they follow different paths in problem solving. His theory breaks down the inputs of a group into classes: properties of group members, properties of the group structure, properties of the task, and properties of the environment. Our focus on team diversity falls into the category of studying the properties of group members.

Organizational demography is defined as the relationship between the diversity of a team's demography and the outcome of their problem solving. Diversity in an organization can mean many different things, from demographic variables like age and race to educational background to cognitive diversity, but the diversity policies of most institutions and organizations are looking to either admit students or hire people of diverse demographics. They assume (and there is evidence to support this) that having a group of individuals of different race, gender, culture, age, etc. prevents discrimination and promotes a healthy work environment of tolerance, inclusion, and community. However, diversity does not always add value to an organization. Horwitz et al did an overview of the research done on diversity in teams, and the results are mixed and inconclusive [31]. For example, it was shown that demographic diversity does not influence the effectiveness of entrepreneurial teams.

There are many studies that research how the make-up and diversity of an organization affect its performance. Teams, or an organization the size of a company, that are diverse are theoretically better than teams that are uniform, and many organizations have employment policies in place to purposely increase hiring diversity [24].

In his pivotal book, “The Difference” [44], Scott Page demonstrates that cognitive diversity can indeed contribute to a team’s success on certain problems. He says that cognitive diversity comes in the forms of different paradigms (schema for modeling the world), different values (which outcomes should be maximized), and different processes (approaches to putting it all together). Problems that are ill-defined, like developing good policies, are more difficult to solve and are better suited to benefit from a cognitive diverse team, and the more diverse, the better. He gives examples of situations where a group of normal people with a diverse set of view points will outperform a group of like-minded experts. Most diversity policies focus on demographic diversity, which Page states is not necessarily correlated with cognitive diversity. However, diversity is, by definition, differences, and these differences can also lead to conflict within the team [34]. Different values can lead to less group commitment. The primary alternative theory is that homogeneous teams perform better because of their shared characteristics (whether it be demographics or way of thinking), which then contributes to team cohesion and thus success [33].

Chowdhury [23] seems to reinforce Page’s thesis that demographic diversity does not correlate with team performance unless tied with cognitive diversity. They looked at various demographic diversity aspects, including age, gender, background, and cognition and showed that these factors do not contribute to the performance of entrepreneurial teams. However, team process variables, specifically group commitment and cognition, positively contributed to the team’s effectiveness. With all these differing results, it appears that the landscape of what contributes to a company or team’s success is complex and still not completely understood [31].

This thesis builds upon this body of work. We postulate that people’s interests reflect their cognitive diversity as interests and hobbies are solutions to the problem of extraneous time. We extract a variety of interests from people’s LinkedIn profiles and create a measure of interest diversity. This is then analyzed against the performance

of the companies these people work for in case studies.

2.3 Network Science Applied to Ecosystems

Network science studies the relationships and connectedness of networks. These networks can be created by data from systems that are obvious networks, like social networks, or extracted from systems that can be represented in network form. We review some studies that use network science to analyze complex systems and consider how to apply the analysis in this thesis.

In global economics, the diversity and ubiquity of a country's products has a direct influence on its economic success [25, 29, 30]. Hidalgo et al state that the more products a country exports, the more knowledge it is able to hold, which then improves its economic future. In addition, the ubiquity of its products also determines its role in the global market. In fact, the diversity and ubiquity of a country's product space is a strong predictor of its economic success. They create a network model of the export products to visualize the "Product Space" where the node size relates to the market size of the product, the node colors are determined by the category of export, and the links between the nodes indicate their proximity value, which is a measure of the ability of a country to produce that export given they can produce the nearby products. They observe that when countries export a new product, they tend to move to nearby nodes of similar categories.

When two countries have similar levels of development and export production, their economic future can be modeled through the Product Space. If one country's exports are more focused on the production of core products (such as machinery and electronics), they are able to diffuse and move through the network faster than a country without the ability to produce these products. The Product Space is able to explain

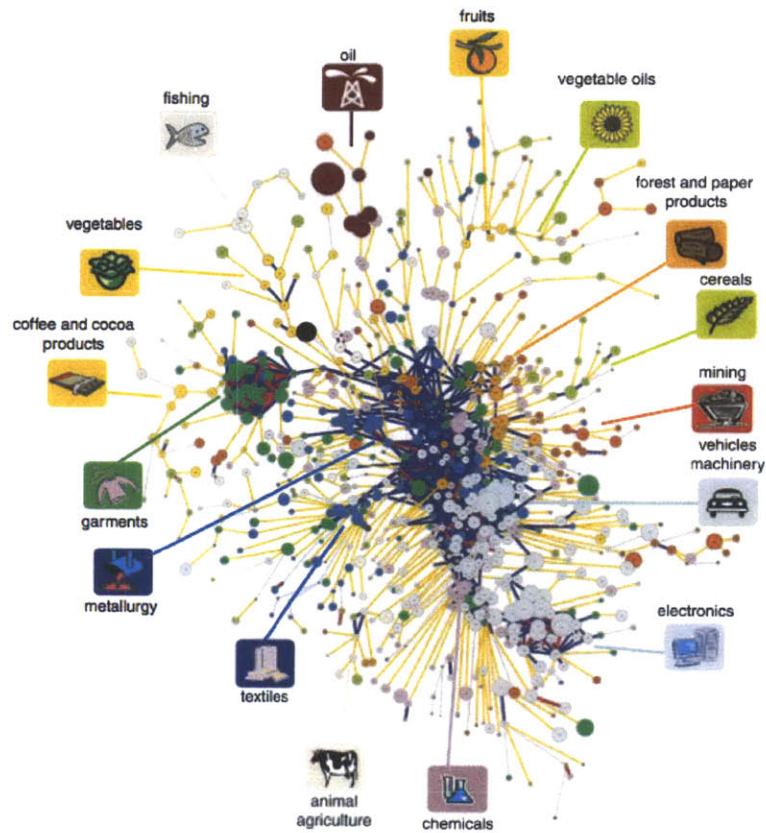


Figure 2-4: The Product Space shows how the world’s exports are connected [25].

the current economic discrepancies between countries that started with similar export conditions.

Ter Wal and Boschma [53] found that firms have previously been modeled with clusters, but these models were insufficient in capturing the internal capabilities of the firms. They created a network model that describes the knowledge-based interaction among firms and show that it can also account for proximal clustering effects. Other models look at a firm’s *dynamic capabilities* [51] to better understand how a firm can maintain a competitive advantage in the current market, describe a firm’s role as a knowledge and skill aggregator in the innovation process [42, 43], or show that the value of social capital depends on the number of people doing similar work [21]. A person’s skill set need not change as they move from job to job, or even from industry

to industry. For example, an engineer specializing in aerodynamics for airplanes can easily adapt his knowledge to work on aerodynamics for cars. This network that connects industries by their overlapping skills is called an *industry space*, and it has been shown to predict a firm's diversification abilities [41].

Can we model people and their interests in a similar way? We make the analogy of companies and organizations to countries and the interests of their employees to the exports. While this is not a perfect analogy, it gets us thinking about how individuals and their personal pursuits can influence the organizations they are a part of. Perhaps we can tease out whether or not there are benefits to a company for hiring people with diverse interests.

Thus we apply this model to the interests of people gathered from LinkedIn. This approach of using network theory to understand the economic success of individual firms or sectors of industry with respect to an individual's diversity, defined by their personal interests, is new.

2.4 Social Networks for Research

Social networks, a recent phenomenon in the past decade, are a ripe new way of accessing large amounts of data for research. Facebook [5], with 800 million users, has been used in many studies on network effects and social graphs. Facebook and LinkedIn are sometimes studied together to compare and contrast aspects of the different networks [45, 50] However, very few studies have relied on LinkedIn as their only source of data.

LinkedIn, with 150 million users [9] and a developing API, provided the basis and means to pull large amounts of data for research. Particularly for those interested in

people and the organizations they work for, LinkedIn can be a valuable resource to verify organizational theory.

2.5 Hypothesis

This thesis applies network theory to firms, their employees, and various aspects of the employees to understand diversity within an industry at both the firm-level and employee-level. We hypothesize that the interest diversity of a firm's employees can influence that firm's economic performance and growth. Using the LinkedIn API, we are able to collect ~600 employees (past and present) for 43 companies using a keyword search. This data is used to create a visualization of people's interests, an "Interest Space", which is a network graph of how interests are categorized and linked to each other. By analyzing the data from firms associated with the Media Lab, we begin to understand how individual interests affect success among a small network of companies. We research this through case studies of a few companies by analyzing their financial data and interests of their employees.

Chapter 3

Data Collection

3.1 Methodology

Because we focused on firms and the people that make up the firms, we needed to select a list of firms for the study.

The MIT Media Lab is inherently a diverse place. The faculty and students are selected based on their research interests, and in particular, they are selected for diverse research interests. Thus, the Lab is an extremely eclectic group of people ranging from designers to engineers to musicians to architects, and the projects that are dreamed and created reflect this diversity. The sponsor companies of the Lab are also diverse, ranging from financial companies like Fidelity Investments to toy manufacturers like Hasbro, so we begin with this set of companies. In order to obtain financial data for the companies, we then limited the sponsor companies to those that are public.

This list of companies provided us a way to search LinkedIn for their employees. We gathered these profiles to analyze as our data set.

3.2 Using the API

3.2.1 Choosing the Data Set

We chose to pull profile data from LinkedIn for a variety of reasons. First, LinkedIn has a RESTful (representational state transfer)¹ API (application programming interface), which allows us to query for structured data (more details under the API Overview). Rather than using a web crawler to gather profile data, the API is designed so that we can request specific information from specific profiles, including company profiles and people profiles. By searching with a keyword or a profile ID, we can get any information field, provided that the user's privacy settings are open. The API also allows us to search for a set of profiles that match a query, for example: searching for profiles of people that have worked at company A. The API is well documented online, with an active forum moderated by LinkedIn developer advocates that answer posts for help and issues within a day.

Second, LinkedIn has a large user base. As of February 2012, LinkedIn crossed 150 million users [9]. This is crucial for our purposes as we need to be able to get a large percentage of people profiles from specific companies, and with LinkedIn's use so widespread, we can safely pull a large enough data set. Having a large data set, however, does not eliminate the bias of the types of people that join social networks, but we may ameliorate it. This is addressed under the Biases in the Data Set section.

Next, the profiles on LinkedIn are segregated into fields that we are concerned with (figures 3-1 and 3-2). We are interested in people's employment history, education history, age, location, interests, and skills. These are all fields that LinkedIn provides with their API. On the other hand, a social network like Google+ does not have these

¹An API is considered RESTful if it follows the REST-style architecture. In general, this means that the interface between a client and the server is uniform and that the communication is stateless. Many web APIs are RESTful as they provide a scalable way to access the data.

specific fields, so it would be extremely hard to gather the relevant information.

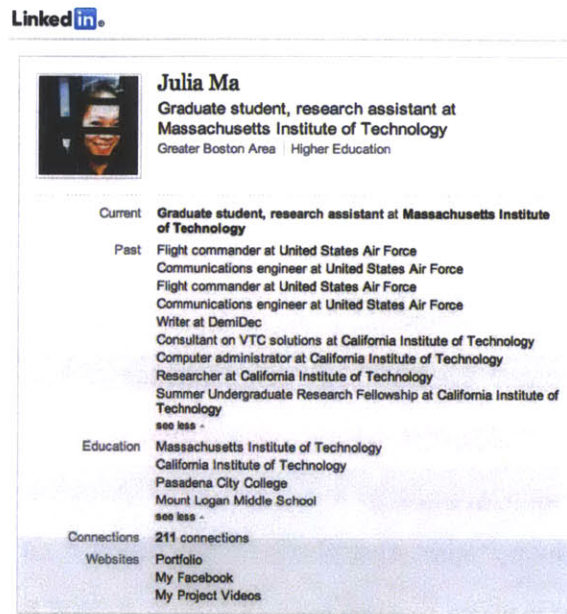


Figure 3-1: Screenshot of a LinkedIn profile

Lastly, LinkedIn is a professional network. Facebook is a much larger, more active social network than LinkedIn, but LinkedIn brands itself as the social network of professionals. People use LinkedIn to market themselves in a professional way, and the profiles take the place of resumes. LinkedIn allows companies and recruiters to use their data to find potential employees, and thus people use LinkedIn as a place of business. Its use is more formal than Facebook as there are no games or apps, poking or liking, or pictures. As a result, profiles are generally focused on facts that people feel comfortable sharing with a potential employer. This is a positive feature, but we also address the inherent biases that this results in below.

3.2.2 API Overview

The LinkedIn API [10] is a RESTful API, which allows clients to initiate requests for information to a server. The API requires the client application to authenticate to

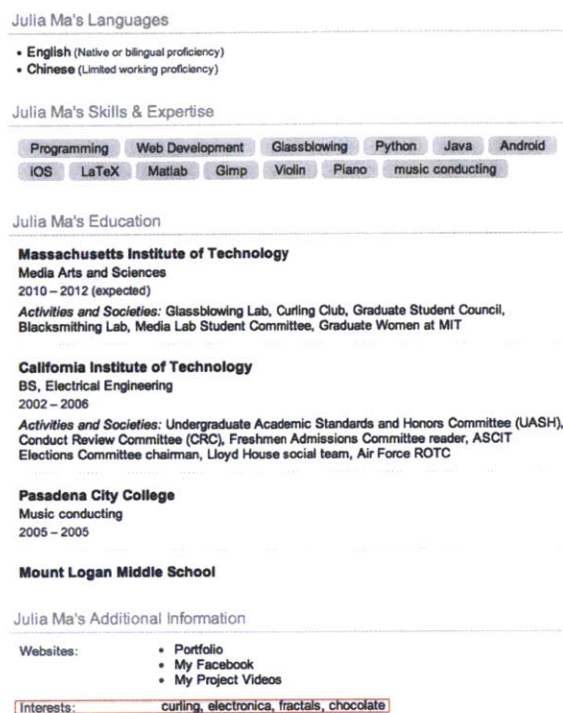


Figure 3-2: Screenshot of additional information on a LinkedIn profile with the “interests” field highlighted

a specific user account for API access as most applications are centered around the user’s own profile. Our application is not user-centered, however, as we are gathering profiles from searches, but we are still required to authenticate the application. Our code is based off of the Python-LinkedIn module [14] with modifications to fit our purposes.

LinkedIn uses Open Authentication (OAuth) [11] as its authentication protocol. OAuth provides users the ability to allow a 3rd-party application to access their account without revealing security information, such as passwords. We use an OAuth library for Python called “oauth2”, which provides a simple Python interface for OAuth [15]. Each time an API instance is initialized, the user is required to authenticate the application before any other API calls can be made. This means that if a change needs to be made in our application, the entire authentication process needs to be redone.

Get people profile fields	http://api.linkedin.com/v1/people/
Get company profile fields	http://api.linkedin.com/v1/companies/
Search people profiles	http://api.linkedin.com/v1/people-search
Search company profiles	http://api.linkedin.com/v1/company-search/

Table 3.1: Base URLs for LinkedIn API

3.2.3 Query Structure

An API query requires the base URL, fields, and parameters.

The base URL changes depending on the type of query. Table 3.1 shows the base URLs for our various queries.

The fields of a query is a formatted string that indicates the profile fields we want returned. For example, if we wish to retrieve a person’s profile ID, date of birth, skills, and interests, we use the “people” base URL with the field string “:(id,date-of-birth,skills,interests)”. To retrieve specific fields from a company’s profile, we use the “companies” base URL, and the field string could be “:(id,name,industry,company-type,locations:(is-headquarters,address:(country-code,state,city,postal-code)))”. All possible people and company profile fields are documented in the LinkedIn API.

Parameters are only used for search queries. They indicate the constraints of the search. For example, if we wish to search for people who list “Google” as their current employer, we use the “people-search” base URL with the parameter string “?company-name=google¤t-company=true”. If we are looking for the company profile for “Google”, we use the “company-search” base URL with the parameter string “?keywords=google”. All parameters for the search queries can be found in the LinkedIn API documentation.

Since the API access requires a user’s authentication, the queries are normally limited to that user’s accessible network, which defined by LinkedIn as three connections from current user. This means that all out-of-network profiles are normally not accessible

to that user. To get around this, the LinkedIn API requires applications to pass an additional authorization token before allowing access to out-of-network profiles. However, even if we receive out-of-network profile access for a particular profile, we are still subject to that profile's privacy settings. The majority of the people profiles we studied were out-of-network profiles.

3.2.4 Limitations of the API

The LinkedIn API, while it suits our purposes, has a few downsides that other researchers should consider before using it.

We have described the OAuthentication feature for the API in the previous section. This requirement forces an application developer to register a key with the API before making queries and centers the application onto the current user. While this does not affect the searches we did, it means that we are unable to simply crawl the site for data.

The LinkedIn API limits the number of possible search results we can access. Starting as a free Basic account user, the maximum number of search results we could access was 100. As this was not sufficient for our needs, we upgraded the account to an Executive account to access up to 700 search results. This cost \$99.95 a month. While we believe this increased number of search results was sufficient for our needs, it is possible that the results were somehow skewed depending on how LinkedIn's internal system does the search. Perhaps the search goes by degree-distance or activeness of profile. This we are unable to know. Something to note, though, is that the increase in search results not only affected our own account, but also the user accounts that used our application. This is important regarding throttle limits, which is discussed below.

LinkedIn protects itself from spam applications by implementing throttle limits for

the users and applications. The throttle limit for most queries, including searches and profile requests, is 1000 queries per day. We encountered this limit when we tried to access our data, which is about 20,000 profiles. To get around this, we crowd-sourced the profile access to multiple user accounts over a period of a few days. For example, with 1000 queries per user per day, it would take four days to access all the profiles if we had five user accounts. If there was an error in the profile access, the process had to be repeated. This was a very time-consuming part of the research. Query results made from different user accounts can also differ slightly in the order the profiles appear, which can cause confusion and difficulty in the process.

Following this method of data retrieval, by using 1000 user accounts to make 1000 daily queries, it would take five months to retrieve the entire 150 million profiles on LinkedIn. In this case, it would make sense to request a “firehose” or unlimited access to LinkedIn’s API.

Even with access to out-of-network profiles, we were still subject to particular profiles’ privacy settings. If that person has completely restricted access to their profile, then we were unable to collect data from that profile. While the number of restricted profiles is relatively low, it may have still influenced the data.

3.3 Biases in the Data Set

Our data set comprises of almost 20,000 people profiles. However, even with so many profiles, there are biases that we address here.

The main bias is that since our data set comes from LinkedIn, we obviously select for users that are on LinkedIn versus those who are not. This is a self-selection as people opt in to social networks. From using just one social network, we do not have a way of getting information about people who are not on it. A more comprehensive data

set would have to come from the companies themselves surveying their employees, but the feasibility of obtaining that data is not possible or necessary on our time scale and for this research. We acknowledge that our data is biased towards more technologically-oriented people.

Another aspect to consider is the people who set their privacy settings such that we cannot access their profiles through the API. We do not know what kind of bias this could introduce. Perhaps these people are more technologically savvy and know the how to limit profile access, or perhaps these people are less technologically savvy and trust the web less and thus restrict access to their profiles.

Given that a person has a LinkedIn profile and that we have access to it via the API, there are still many profiles that are not completely filled out. We find the profiles via a profile search using a company as a parameter, so if a person does not list their work history or has an incorrect work history (spelling error or otherwise), we are unable to include them in our data set. Again we hypothesize that people who have incomplete profiles may be less familiar with social network technology, and thus our data set does not consider them.

While our data set is comprised of at most 700 profiles per company, as that is the API search limit, each company's search results differed. The number of profiles we have per company ranges from 33 at EMC Corporation to 686 at Denso Corporation, with most companies having either ~ 100 profiles or ~ 600 profiles. This may reflect a bias of the types of people who work at those companies, the types of people who have a LinkedIn profile, and the type of access we have through the API.

3.4 A Spammer’s Profile

As we created the Interest Space (described in Chapter 4), we noticed an anomaly in the interest data. Among the most popular interests were “job”, “networker”, “toplinked”, “invites”, “lion”, “lion500”, and “linkedin”.² This on its own was not interesting as it is possible, though odd, that many people would list “linkedin” as an interest. But we also looked at the top paired links, and the top pairs were every combination of these interests, which raised a flag.

When we looked closer at these interests, we noticed that one profile in particular listed these interests multiple times. The other fields of this profile were also concerning as they advertised tax assistance and anecdotal stories of people using a tax service. We concluded that this profile was a spammer’s profile and removed it from the data set. This eliminated all the top links in question. We also confirmed that the rest of the profiles were clean and valid.

This made us realize that there is sometimes a conflict between a user’s goals and what the network is trying to achieve. Even if this profile was a realistic profile, the user is attempting to promote themselves or a service that is unrelated to the social network. While this particular network tries to connect people for business purposes, this user exploits the connectivity for his own use.

3.5 Extracting Relevant Data

For each query, the LinkedIn API returns an XML-formatted profile object. We breakdown the object into the fields we are concerned about and create a dictionary-

²LION (and LION500) is LinkedIn Online Networking, which is a group of people that actively encourage connections with other members, even if there is no business or other established relationship. TopLinked is a service built on top of LinkedIn that allows users to create lists of people and promote themselves for a fee.

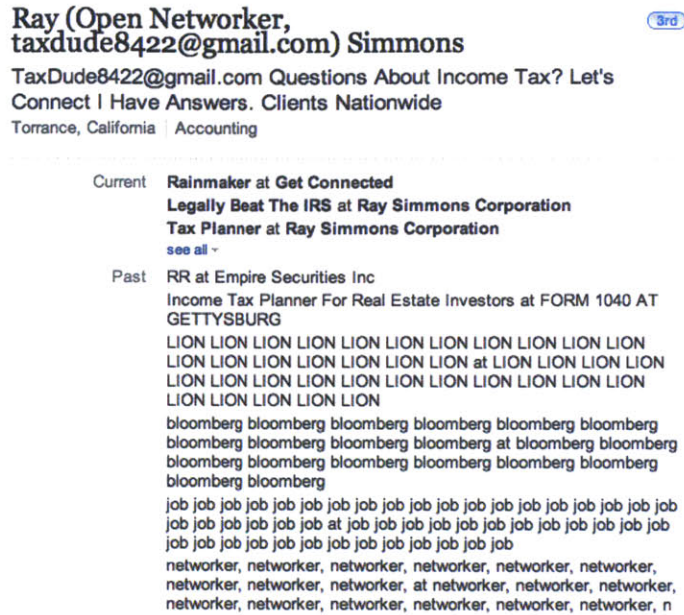


Figure 3-3: Screenshot of a spammer’s profile

formatted profile object. We use a dictionary object because the fields can then be keyword-indexed and we can export the profiles into a JSON file for easy storage.

The data analysis is done using Python scripts on the profiles.

The full data set and code can be found at: <https://github.com/joulesm>

Chapter 4

Creating the Interest Space

Interests are shared by many people and people have many interests. We create an Interest Space, which is the network of relatedness of these interests. This network is populated using the data gathered from LinkedIn profiles.

4.1 Interest Popularity

We collected all the interests and number of occurrences of the interest from the profiles. In LinkedIn, the “interest” field is a short-form text area, so we parsed out each interest from a long string which is comma delimited. Not every profile lists interests, either due to the incompleteness of the profile or privacy settings. Our data set is comprised of the 2122 profiles that list interests. A total of 5359 interests were collected from these profiles.

The distribution of interests was uneven. The number of interests listed in a profile ranged from one to 75. The average number of interests per profile, discounting the profiles that did not list any interests, was five. Figure 4-1 shows the distribution of

the number of profiles that had certain number of interests. For example, 313 profiles only listed one interests, while there was only one profile that had 75 interests. The number of profiles with one interest and the number of profiles with three interests are similar, but there is a sharp drop in the number of profiles with just two interests. This suggests that users who took the time to fill out more than one interest had a tendency to continue listing more. We also plot this on a log-log scale (figure 4-2) and we see after a certain number of interests, the curve follows a power law.

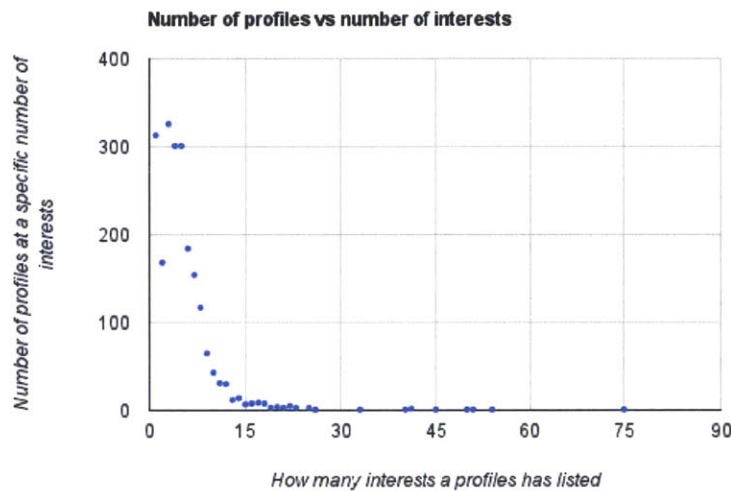


Figure 4-1: The distribution of interests among the profiles was uneven. Most people listed less than 10 interests.

As we built the list of interests, we kept track of all unique interests and counted how many times each interest appears. Not only did this help us determine the popularity of a particular interest, but also the unpopularity of an interest. The unpopularity of an interest is interesting in two ways. One is that interests that would seem relatively commonplace like “alternative music” is actually not listed more than once in this set of profiles. This may reflect the professional environment that LinkedIn creates. The other result of an unpopular interest is eliminating the need for extensive natural language processing algorithms on the list of interests. A spelling mistake of “runnig” can be easily removed from the end list of interests we work with. We also eliminate

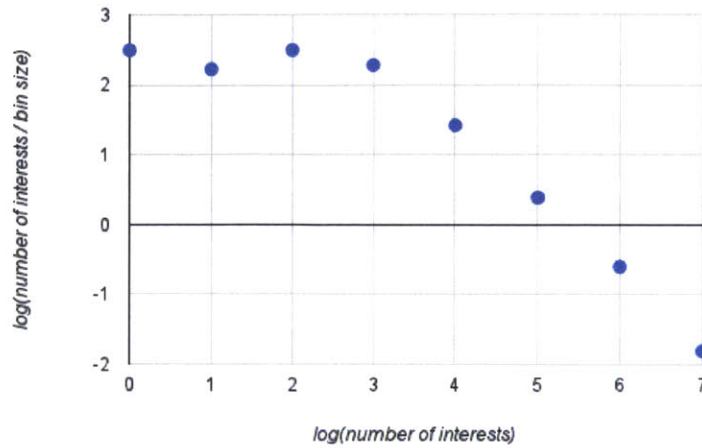


Figure 4-2: We use logarithmic binning to better understand the interest distribution. We divide the x-axis of figure 4-1 into bins of exponentially increasing sizes and count the number of profiles that fall into each bin. In this graph, the x-axis is the \log_2 of the number of interests and the y-axis is \log_{10} of the number of interests scaled by the bin size.

the disadvantage of using a free-form text field as some profiles have full sentences and descriptions in the interests, example “social media and how it is changing the world”.

Interest	Popularity	Interest	Popularity
running	115	skiing	121
new technology	122	golf	129
music	135	photography	140
reading	161	traveling	309

Table 4.1: Top interests by popularity

The most popular interests are listed in table 4.1. These are the interests that people in industry have converged on. Most interests appeared only once. In fact, of the 5359 interests, 4587 of them appeared only once. A sample of these interests are listed in table 4.2.

barefoot running	idiosyncrasies	single malts
social recognition	death	sound design
energy technology	agile development	bible
semiconductors	interactive marketing	motivating people

Table 4.2: Interests appearing once

The number of interests that are extremely popular is small and the number of interests that are extremely rare is big. We plotted the popularity of the interests against the number of interests at that popularity in figure 4-3. For example, there are 4587 interests that occur only once and there are 43 interests that appear five times. The graph uses a log-scale on the vertical axis. We also plot the log-log graph of the popularity in figure 4-4 and see that the curve follows a power law.

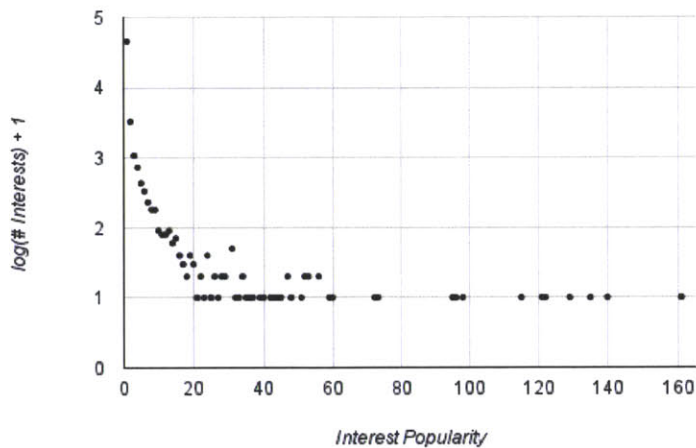


Figure 4-3: Popularity of interests vs number of interests at that popularity

The graph shows that the overall popularity of specific interests is extremely low. Most interests are listed in less than 10 profiles. Not counting the interests that only occur once, the median popularity is 3. That means half of all the interests show up

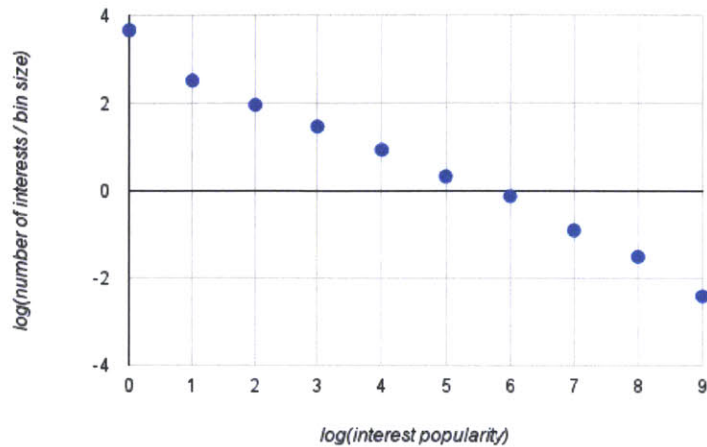


Figure 4-4: We use logarithmic binning to better understand the interest popularity of figure 4-3. The x-axis is the \log_2 of the interest popularity and the y-axis is \log_{10} of the number of interests at that popularity scaled by the bin size.

in only two or three profiles. Since LinkedIn is a professional social network, perhaps people who use it are looking for ways that distinguish them from other possible hires. By listing interests that are less common, headhunters and recruiters looking at the profile may find the person unique and consider them for a position over someone who has generic interests.

4.2 Links Between Interests

Once we have the number of occurrences of each interest, we use only the interests that appear more than twice (we call these “frequent interests”) to then pull co-occurring interests. We choose this popularity threshold because interests that appear only once have at most one link and this link cannot be significant. Also given the median popularity of the interests is three and we are interested in popular links as well, the threshold of three suffices. As infrequent interests make up the majority of the nodes,

Interest Pair	Popularity	Interest Pair	Popularity
reading, travel	29	photography, travel	26
technology, travel	22	music, travel	22
reading, travel	19	golf, travel	18
movies, reading	18	music, reading	18

Table 4.3: Top linked pairs of interests

we eliminate them from the network so that we can better observe the shape of the graph.

We look at each profile with interests again, and if there are pairs of interests per profile that are both frequent interests, this pair is saved as co-occurring interests. We also keep a tally of how many times specific pairs of co-occurring interests appear. This tally will be used in our calculations of how significant particular links are.

The top linked pairs are listed in table 4.3.

These links only measure the number of times the interests occur in the same profile, but this does not necessarily indicate the significance of the links. For instance, it is possible that “travel” and “reading” may be inherently related because people who travel a lot also read while on the plane. This would mean that even though these interests have a high co-occurrence, this link is less significant than it seems.

We use a Pearson binary correlation ϕ (also known as Pearson’s r) on the interests to calculate the significance of the links [46]. Pearson’s r is defined as the covariance of the two variables divided by the product of their standard deviations. This coefficient will be a value between -1 and 1. We calculate Pearson’s r using formula 4.1. N_a and N_b are the number of times interest a and interest b appear respectively. $N_{(a \cap b)}$ is the number of times interest a and interest b co-occur, and N_p is the total number of people with interests.

$$\phi = \frac{N_{(a \cap b)} \cdot N_p - N_a \cdot N_b}{\sqrt{(N_p - N_a) \cdot (N_p - N_b) \cdot N_a \cdot N_b}} \quad (4.1)$$

From the Pearson correlation, we can calculate a t -statistic (formula 4.2), which is a measure of how extreme a statistic is. The higher the t -statistic, the more significant the link is.

$$t = \frac{\phi \cdot \sqrt{N_p - 2}}{\sqrt{1 - \phi^2}} \quad (4.2)$$

A ϕ value of 0.664 has a t -statistic of 18.58. We convert this to a p -value, a statistical measure of significance or a probability of obtaining this extreme of a t -statistic. Using a standard t -distribution table, we see that $t=18.58$ corresponds to a p -value of less than .0025. Extremely significant!

We can pick a p -value threshold to limit the number of significant links. A typical p -value threshold in statistics is .05, which corresponds to a ϕ value of 0.29.

4.3 Categorizing Interests

Divisi is a library for Python that adds common sense analogy reasoning to applications. We use Divisi to create initial categories for the interests and then manually categorize the remaining interests. These categories cluster similar interests together and help us better understand the Interest Space.

4.3.1 What is Divisi

The Common Sense Computing (CSC) Initiative [2] is a project that collects millions of common sense statements to develop tools to give intelligence to computers. Thousands of people contribute statements like “Ragtime is a style of music”, which are used to create a semantic network that ties concepts together through analogies. This kind of information, when given to a computer application, allows computers to better understand human text, language, and intent.

Divisi [4] is one of these tools developed by the CSC. It is a Python library that connects ideas through the semantic network. It not only provides a numerical measure of similarity of two concepts, it can predict features and create categories with confidence measures.

Divisi calculates the similarity between two concepts as a dot product of the features of these concepts. Similar features in the semantic network are vectors that point in similar directions. These vectors are normalized to the unit vector, and the similarity becomes a value between -1 and 1. For example, the similarity of “dog” and “cat” is .837, which means these two concepts have similar features. “Music” and “poetry” have a similarity of 0.532 while “Italian” and “robot” have a similarity of -0.105.

Another type of similarity that Divisi can calculate is called “spread activation”. Spread activation is more about the relatedness of two concepts than how similar they are. For example, concepts like “sad” and “cry” are not highly similar as they have different features and properties, such as parts of speech and usage. However, these concepts are very related, and Divisi’s spread activation can provide this measure. This calculation starts with our input concepts and takes into account any neighboring concepts that contribute to the relatedness of our inputs. For example, the similarity of “Italian” and “Spanish” is very low, a 0.0009, but the spread activation is .994. That is because concepts similar to “Spanish” are “talk”, “music”, and

“conversation”, but taking into account neighboring concepts, Divisi figures out that “Spanish” and “Italian” are both adjectives and languages.

Divisi provides a way to create categories of concepts from which it infers the shared properties of these concepts. By giving Divisi a set of concepts that we deem related (a category) and another concept that we want to categorize, it can give us a confidence rating of how related the lone concept is to the category. For example, we can create a category with “car”, “bus”, “train”, and “bicycle”. The concept “skateboard” into this category gives a confidence rating of .943 while “cat” has a confidence rating of .082.

4.3.2 Using Divisi to Categorize Interests

We use Divisi to categorize our list of interests from the LinkedIn profiles.

We start by first sorting out the interests that are already found in Divisi. Some items like “human computer interaction” are too specific and technical to be in Divisi, while others like “skydiving” will need to be corrected to “skydive”. We will correct for both cases further on.

For the interests found in Divisi, we then calculate the similarity and spread activation measures for each pair of interests. The top 10 pairs and their confidence numbers for each method are in table 4.4 and table 4.5.

Interest 1	Interest 2	Similarity	Interest 1	Interest 2	Similarity
theater	movies	0.444	jazz	music	0.489
travel	traveling	0.490	education	entertainment	0.490
art	dance	0.530	art	music	0.595
music	dance	0.708	faith	church	0.824
theater	theatre	1.005	religion	church	2.156

Table 4.4: Similarity measures between interest pairs

Interest 1	Interest 2	SA measure	Interest 1	Interest 2	SA measure
nanotechnology	karate	0.972	online	japanese	0.972
nanotechnology	japanese	0.975	faith	religion	0.977
friends	online	0.982	friends	japanese	0.983
nanotechnology	online	0.984	friends	nanotechnology	0.990
italian	spanish	0.994	ice hockey	badminton	0.998

Table 4.5: Spread activation measures between interest pairs

From the table, we can see that using spread activation seems to highly correlate concepts like “friends” and “nanotechnology” together, which is not a good result. Thus we start by using the similarity correlations to create categories.

We group together all concepts that have a confidence measure of greater than .2. This value was chosen by trial and error. The initial groupings created in this manner are in table 4.6.

These initial groupings are quite good and it is easy to see the relatedness of the various interests. We use Divisi to create categories from these groupings. Here is the initial “sports” category: sport, basketball, soccer, entertainment, football, baseball, ski. The initial “religion” category includes: faith, religion, church. The relatedness

sport, basketball, soccer, entertainment, football, baseball, ski
guitar, acoustic guitar, music, classical music, poetry, piano, jazz, art, dance, design
education, research, literature, news, science
theater, cinema, television, movies, theatre
business, job
faith, religion, church
food, coffee, wine, apple

Table 4.6: Initial interest categories

of “tennis” to the “sports” category is 0.608 while the relatedness of “tennis” to the “religion” category is -0.005. We do this for all remaining uncategorized interests, pick the most related category, and if the relatedness is above a certain threshold, that interest is inserted into the category. To be clear, we have labeled the categories by hand, like “sports”, but “sports” can also be an interest in this category.

We repeat this process until the remaining uncategorized interests are not related enough to automatically fall into the categories.

4.3.3 Manually Categorizing Interests

After using Divisi to create initial categories, the remaining interests are either not related enough (according to Divisi) to a category, too technical for Divisi to understand the meaning, or not in Divisi’s knowledge base at all. These interests are manually categorized by human judgment.

We split a couple categories into more specific ones. For example, the category of “sports” include “basketball”, “baseball”, “tennis”, “ski”, “kayak”, and “yoga”. Looking at these interests, we realize that “basketball”, “baseball”, and “tennis” are not just sports but also games, while “ski”, “kayak”, and “yoga” are physical activities that do not involved score-keeping. We split the “sports” category into “game” and “sport”.

Another example of splitting categories is the “art” category. It includes “poetry”, “photography”, “drawing”, “guitar”, “jazz”, and “opera”. We notice that the latter three interests are music-centered, so we split a “music” category out of the “art” category.

We categorize the technology-related interests together, as those were non-understandable by Divisi. “Web”, “internet”, and “electronics” form the beginnings of this category.

As a human that can comprehend the subtleties of language, we run into the issue of whether we should categorize the interests based on our own knowledge or based on what we think Divisi would have been able to understand. For example, “blackberry” is one of the interests that Divisi was unable to categorize. We understand “blackberry” to most likely refer to the line of smartphones, but Divisi would most likely categorize it in “food”. Since a user would be more likely to list “blackberry” as an interest to mean the phone and not the fruit, we decide to include “blackberry” into the “technology” category because we have the advantage of knowing the intentions of a profile.

4.3.4 Categories

The final categories used in the visualization are listed in table 4.7. When we create the Interest Space, each category is colored differently in the visualization, allowing us to see whether or not similar interests are strongly linked.

sports	sports, golf, cricket, squash, tennis, swimming, rock climbing, hiking, kayaking, camping, skiing, hunting, fishing, boating, mountain biking, snowboarding, surfing, skateboarding, motorcycling, wakeboarding, volleyball, basketball, football, soccer, biking, running, martial arts, cycling, badminton, sailing
art	art, writing, fashion, design, photography, digital photography, architecture, drawing, reading, dancing, creativity
music	music, guitar, playing guitar, singing
learning	learning, education, politics, economics, science, psychology, history, books
movies	movies, film, television
business	business, advertising, branding, management, entrepreneurship, strategic planning, organizational development, project management, professional networking, social networking, management training
food	food, cooking, gourmet cooking, gardening, wine
travel	traveling, international travel
technology	technology, new technology, internet, gadgets, digital media, innovation, usability, usability research, artificial intelligence, computer vision, machine learning, information architecture
people	family, friends, church, leadership, mentoring, public relations

Table 4.7: Final interest categories

4.4 The Interest Space

To better understand how various interests relate to each other, we created a network visualization of the interests.

With each interest as a node and the Pearson correlation of the paired interests as links, we create a network graph. The popularity of an interest translates into the size of the node, and the strength between interests becomes the strength of the link. The color of the nodes is determined by the category of the interest (figure 4-6). In this way, we can see how the various interests are related to each other while also seeing how often they co-occur (figure 4-5).

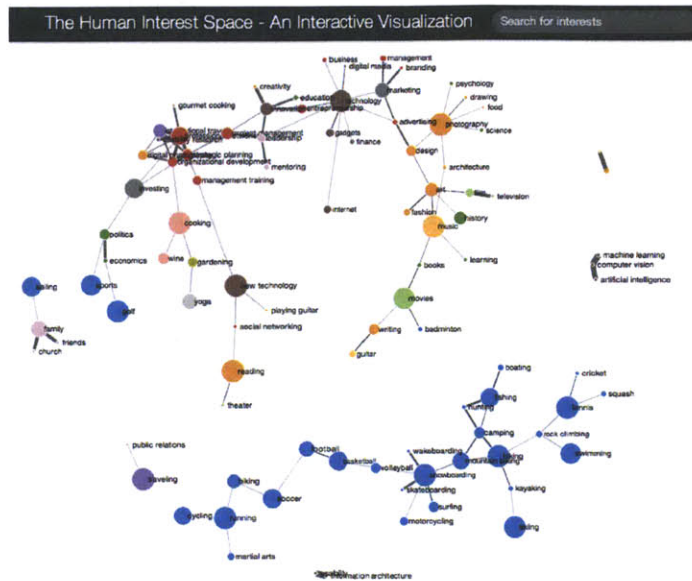


Figure 4-5: The Interest Space visualization

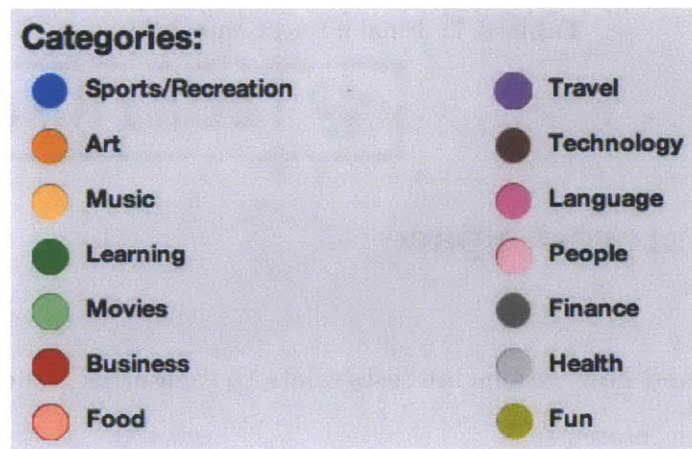


Figure 4-6: The interest categories are indicated by the color of the node.

The visualization is a network of these nodes and links is created using Javascript libraries [8, 1, 3], and users are able to experience and interact with the visualization through a browser (figure 4-7).

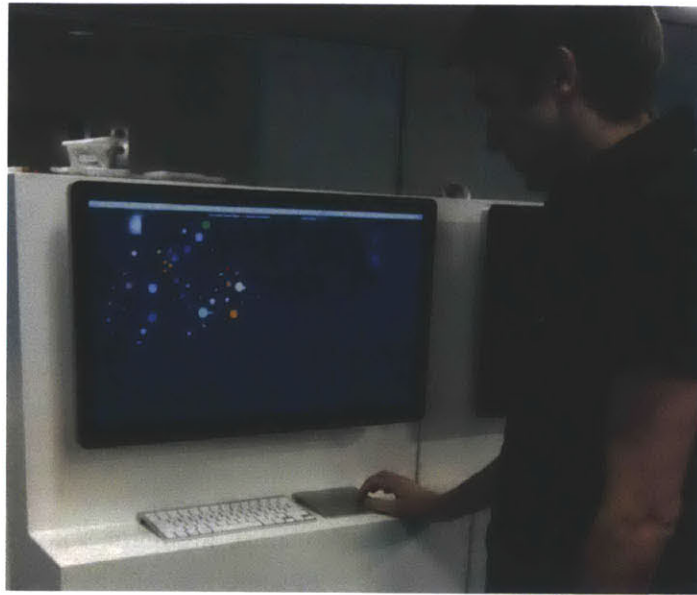


Figure 4-7: The Interest Space allows the user to explore how interests are connected.

The visualization simulates the physics between objects linked by springs, thus the visualization clusters into a circle. We can move the nodes around to better understand the connected graph, and we immediately see that there are two large linked clusters of interests.

We provide a few functions to make the visualization easier to navigate. First, the visualization has a search function, giving the user the ability to find nodes that he is particularly interested in. If the interest is found, the node will highlight itself and its closest neighbors (figure 4-8). This lets the user immediately see the closest interests to his own.

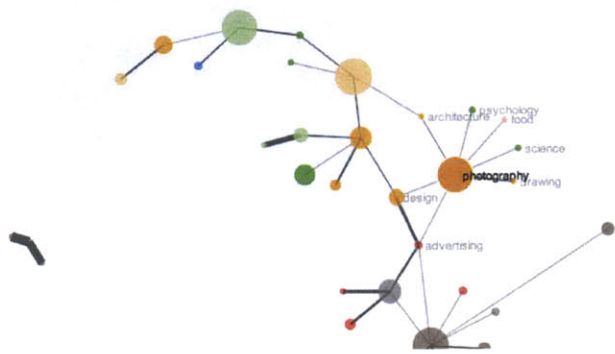


Figure 4-8: The visualization allows the user to highlight a particular interest.

The visualization also has a suggestion engine. The user highlights a few nodes to submit, and the visualization will use those nodes to generate interest suggestions (figure 4-9).

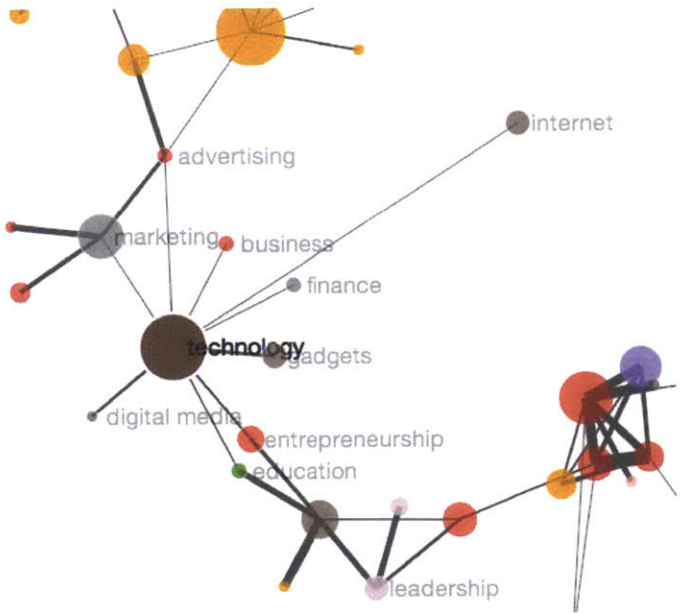


Figure 4-9: When a user hovers over an interest, all connected interests are highlighted as well.

Chapter 5

Other Profile Data

5.1 Profile Completeness

From the list of companies that sponsor the Media Lab, we select the ones that are public and on LinkedIn. This gives us 43 companies to begin with. We retrieve people profiles for these companies and result in 19448 profiles. Table 5.1 shows the distribution of number of profiles for each company:

Company	# Profiles	Company	# Profiles
Aegis Media	677	AOL	667
Bank of America	663	Best Buy	675
BT	644	Cisco Systems	654
Comcast	678	Denso	686
DirecTV	696	EMC	633
ESPN	104	Fujitsu	673
Glaxosmithkline	103	Google	99
Hasbro	586	Humana	108
Intel Corporation	654	Interdigital Communications	106
Intuit	106	KT	102
LG Electronics	676	Marvell	118
Motorola Mobility	642	News Corporation	100
Nokia	654	Northrop Grumman	106
Panasonic	682	Qualcomm	105
Research in Motion	663	RR Donnelley	107
Samsung	577	Sanofi Aventis	686
Sberbank	107	Schneider Electric	624
Shell	660	Sony	637
Steelcase	663	Telecom Italia	106
Time Warner Inc.	109	Toshiba	679
Verizon	651	VF Corporation	676
Volkswagen	106	TOTAL	19448

Table 5.1: Number of profiles retrieved for the Media Lab sponsor companies

There seems to be two modes of the number of profiles per company, hovering around 100 and 600. The LinkedIn API limits the number of search results to 700 profiles,

but of the 43 companies, we were able to retrieve only ~ 100 profiles for 15 of them. There are no companies where we had 400 profiles or anything else in between 100 and 600, so we postulate that there is an API limitation that we were not aware of.

We did not have complete profile information for every profile. This was either because not every profile had all the fields filled in or there were privacy restrictions set by the user. Table 5.2 shows the number of profiles that had specific fields completed.

Profile field	Number of profiles	Profile field	Number of profiles
interests	2122	skills	1797
age	16	education	1
languages	326	positions	17187
location	19448	total	19448

Table 5.2: Number of profiles with information in a particular field

5.2 Education

We wanted to not only compare profile interests with the organizations that person works for, but also see if other factors can be a diversity measure. Cognitive diversity includes having different perspectives, ways of representing the situation, or different heuristics, approaches to solving a problem. Since these perspectives and heuristics can come from a person’s background, we are interested in studying their educational background.

Unfortunately, because the vast majority of the profiles we retrieved were not within three degrees of separation of our network, the API limited our access to pulling education information. The only profile in our set that had education information was a person that was directly connected to our network. This profile was retrieved because

they had worked at Samsung, which is one of the sponsor companies. Unfortunately, this profile did not list any interests, so it was not included in most of our analysis.

5.3 Age

While thinking about how people can develop different perspectives, we understood that the amount of life experience a person has will influence their outlook. Thus we included “age” in our list of profile attributes.

LinkedIn does not specifically have an “age” field, but some profiles include information in the “date-of-birth” field. Unfortunately, this field is also restricted access through the API. We attempted to overcome this restriction by extrapolating the age from a person’s last listed education. We assume a person was 22 at their graduation date and 18 when they entered schooling. Thus if someone has no “end-date” to their education, we assume they are still in college and that their age to be 18 at the given “start-date”, and calculate their age from this year. If someone does have an “end-date”, then we assume their age at graduation to be 22 and calculate their current age.

This method of extrapolating age data is rough. We may not know the type of degree their most recent education is. Their most recent school could be for a masters, PhD, or other advanced degree. Many people also don’t start college at 18 and graduate at 22. Unfortunately, since we were only able to get one profile that had an education field, this was not an assumption we could test with our data. In the end, we had just 15 profiles with a “date-of-birth” field. These plus the profile that had “educations” gave us only 16 profiles with “age”.

Location	Number of profiles
San Francisco Bay Area	1059
Greater New York City Area	1048
Greater Los Angeles Area	618
Greater Boston Area	607
Greater Minneapolis-St. Paul Area	514
Washington D.C. Metro Area	458
Greater Philadelphia Area	438
London, United Kingdom	412
Finland	374
Kitchener, Canada Area	345

Table 5.3: Most common profile locations

5.4 Location

Alfred Weber’s theory of industrial location suggests that industries tend towards that minimize the cost of transportation of raw materials and final products. Silicon Valley is the iconic success story of a location-based innovation hub [22]. Regional stories of economic success are often linked to the existing networked environment of that region. As our data has the location information of people, we explore whether there is a regional component to the profiles that list interests.

Every profile has a location listed as this is a required field and the API has no restrictions on the location field. Table 5.3 shows the top 10 locations of our data.

We see that the people in these profiles predominantly live in the United States. This is probably representative of the fact that most of the sponsor companies are headquartered in the USA as well. However, the maximum number of profiles we have for any particular company is less than 700, so there are clear clusters of people from various companies in San Francisco and New York City. Together with Los Angeles and Boston, profiles from either California or New England make up 17% of our data.

For just the 2122 profiles that have interests listed, table 5.4 shows the top locations. Interestingly, the top two locations have not changed positions, and eight of the top

10 from table 5.3 are in table 5.4. Looking at just the location parameter, the subset of profiles that have interests listed is relatively representative of the larger data set.

Location	Number of profiles
San Francisco Bay Area	183
Greater New York City Area	130
Greater Los Angeles Area	74
Kitchener, Canada Area	72
Greater Boston Area	71
Greater Philadelphia Area	47
Washington D.C. Metro Area	43
London, United Kingdom	43
Greater Grand Rapids, Michigan Area	36
Toronto, Canada Area	36

Table 5.4: Most common profile locations for profiles that list interests

5.5 Skills

We considered building a corresponding “Skills Space” in the same vein as the Interest Space, but the skills are job-related. We are less interested in what people do *at* work than what people do while *not at* work. Thus while we give an overview of the skills in our data set, we do not deeply explore the skills network.

Skills may seem like a similar field to interests, but LinkedIn treats them differently. The skills field is not a short-form text field, but rather it is collection of “skill” objects that include a skill ID and optional fields of proficiency level and number of years of experience.

A total of 4489 skills were collected from the 1797 profiles that had the field. The top skills are listed in table 5.5.

Skill	Popularity	Skill	Popularity
product management	206	strategic planning	200
business strategy	193	product development	188
program management	188	marketing strategy	185
cross-functional team leadership	175	telecommunications	164

Table 5.5: Top skills by popularity

As we see from the most popular skills, skills are very work-oriented. Although “cooking” can be considered a skill, a person would most likely not list it in their skills field unless that was related to their work. Since our profile retrieval was based on profile searches of companies that sponsor the Media Lab, the likelihood of professional chefs in our data set is small.

Also, “fun” is a skill listed by 2 people.

5.6 Languages

Languages could be a subset of skills, but LinkedIn breaks this field out separately. Like skills, the language field is a collection of structured “language” objects that includes a language ID and optional proficiency level. 64 unique languages were collected from the 326 profiles that had the language field and the most spoken ones are listed in table 5.6.

There are a few issues with the language field. One is that it does not properly account for dialects of languages. For instance, “Chinese” is normally assume to

Language	Popularity	Language	Popularity
English	261	Spanish	105
French	103	German	77
Mandarin Chinese	47	Italian	33
Japanese	32	Portuguese	27

Table 5.6: Top languages by popularity

be Mandarin Chinese. However, profiles that list Chinese have various methods to indicate the dialect, such as “Mandarin”, “Mandarin Chinese”, “Chinese - Mandarin”, all of which we had to sort manually. Another issue is that the language field does not account for a profile in a foreign language. For example, a profile written in German would list “English” as “Englisch”. Again, since the number of languages was pretty small, we sorted these manually.

Interestingly, someone knows how to communicate through “wingdings”.

5.7 Positions

A person’s work history is listed in the “positions” profile field. Of the 19448 profiles in our data set, 17187 of them have work history. This is because although we get a profile through a search result, that profile’s privacy settings may restrict the fields we are able to gather. For out-of-network profiles, the LinkedIn API also restricts the number of past positions we can access. Thus the work history of some profiles may not be complete. In fact, for an out-of-network profile, unless the person is currently working at the sponsor company, we do not know any information about their time there. It is extremely limiting to not have a complete work history of people’s profiles. In relation to cognitive diversity, it would be interesting to see whether many people cross industry sectors to do a similar job. In this sense, a person doing marketing at an internet company could also do marketing at a clothing company. Perhaps a

Company	# of profiles	Company	# of profiles
Research In Motion	650	Bank of America	611
Sanofi Aventis	588	Best Buy	569
Schneider Electric	561	Cisco Systems	492
DirecTV	480	EMC	472
Steelcase	424	LG Electronics	387

Table 5.7: Companies of which we have the most profiles

company that has more people who cross sectors manages to be more flexible and adaptable to changing economies.

The LinkedIn profile field is a collection of “position” objects that contain information about the person’s work title, position summary, the company name, start and end dates of the position, and whether the position is current. If the position is current, the end date of the position is blank.

The vast majority of the profiles had less than 10 positions, with 80% of the profiles only having one position. The average number of positions per profile is 1.3.

Since a person who currently works at Samsung may have worked for other companies, we have position information about non-sponsor companies. Table 5.7 lists the companies of which we have the most profiles. As expected, the top companies are sponsor companies. We have a total of 6420 companies gathered from the profiles, but due to the non-uniformity of company naming, there are many duplicates. For instance, “Toshiba Corporation”, “Toshiba American Medical Systems”, and “Toshiba Business Solutions” are clearly all part of Toshiba, but enough people list them separately that we did not combine these into one company. Companies that have branches in various countries were also troublesome.

The first company that isn’t a Media Lab sponsor is Microsoft, with 16 profiles. Next is the MIT Media Lab which has 14 profiles.

As we see in the next chapter, we will study the profiles of Research In Motion,

Schneider Electric, and Intel Corporation in depth. Interestingly, while Intel had the third most profiles that listed interests, proportionally, we did not have very many total Intel profiles. We will explore the work history of each company in the case studies and the interests of these companies' employees.

Chapter 6

Case Studies

Although we retrieved almost 20,000 profiles through LinkedIn's API, the actual number of profiles that have interests listed is much smaller. We break down the profiles even more by considering clusters of profiles that share an employer. The amount of data, specifically profiles with interests, we have for a particular company ends up being around 100, two orders of magnitude less than our original data set. Thus we select three companies that we have the most data for and explore these profiles in depth. Each company is different, and we study their history, stock performance, number of patents, and other aspects that may reveal connections to our research.

6.1 Research In Motion

6.1.1 The Company

Research In Motion (RIM) is a telecommunications company, headquartered in Waterloo, Canada, that manufactures software and hardware devices for the wireless

market. They were founded in 1984 and went public in 1998. Their signature product is the BlackBerry, a smartphone that is best known for its encryption capabilities, instant messaging to other BlackBerry users, and pioneering wireless email [28].

In 2011, RIM's smartphones make up 3% of the market share of the world's mobile devices [6], making them 6th place in mobile device manufacturing. In just the smartphone market, RIM's devices had 19% of the market share in 2010 which declined to 12% in 2011.

Here is some basic information about Research In Motion:

Mission statement: Research In Motion is a leading designer, manufacturer and marketer of innovative wireless solutions for the worldwide mobile communications market. Through the development of integrated hardware, software and services that support multiple wireless network standards, RIM provides platforms and solutions for seamless access to time-sensitive information including email, phone, text messaging (SMS and MMS), Internet and intranet-based applications. RIM technology also enables a broad array of third party developers and manufacturers to enhance their products and services with wireless connectivity to data.

Company size: 16,000+ employees

Locations: Headquartered in Canada; other locations in USA, Europe, Middle East, Africa, Asia

6.1.2 Performance

As stock price has been used as a tool for managerial compensation for a company's CEO [39], we can use a company's stock performance as a rough performance measure [20]. Although stock performance may not be the best measurement because other factors can affect the price, a consistently growing stock value indicates strong market



Figure 6-1: Stock chart of Research In Motion (blue), Apple (red), HTC (yellow)

performance. Better performance measures such as Return On Investment, Return On Equity and Economic Value Analysis [32] are difficult to obtain over a consistent period of time, whereas stock market performance can be retrieved from financial websites through either a search or an API.

By looking at RIM's stock chart (figure 6-1, blue line), we see that they suffered a huge loss at the end of 2008 and nearly four years later, they have not recovered. The main cause of the steep drop in 2008 was the global economic collapse, and RIM's competitors, like Apple, also suffered. But while Apple's stock has since skyrocketed, RIM stays low. However, other RIM competitors, like HTC, seem to be doing similarly to RIM.

Although the economy initiated RIM's slump, other factors contributed to the decline. In the end of 2008, RIM announced a new BlackBerry line, the Storm, that would directly compete with the Apple iPhone and Google Android devices. Prior to the BlackBerry Storm, RIM's devices all included a hardware keyboard. But as the popularity of the iPhone and Androids increased, RIM needed a touchscreen-only handset to compete. The Storm was extremely late to the game as Apple had already released two versions of the iPhone before the Storm was announced, and the reports from both the reviewers and consumers on the device were negative. Unfortunately,

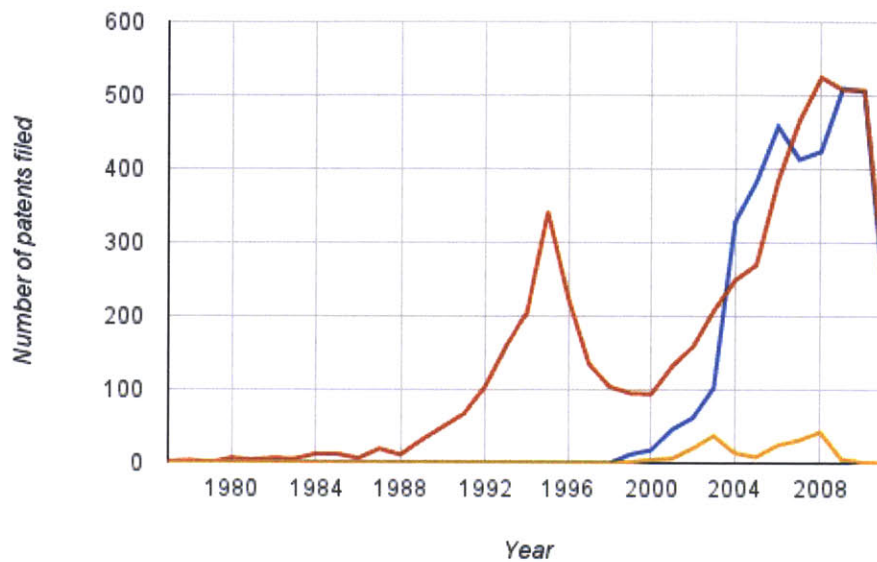


Figure 6-2: Number of patents filed each year for Research In Motion (blue), Apple (red), HTC (yellow)

the research and development costs of the Storm were significant. This cast a large amount of market doubt on RIM, and the company still struggles to regain its foothold in the communications industry [28].

Research In Motion's situation is exactly the scenario our research tries to improve. By understanding how the composition of an organization influences their success, we can help companies hire the right people to help them thrive and adapt to economic downturns and the constantly changing market.

One could make the case that a possible measure of innovation is the number of patents filed each year. If an organization is truly doing new things, inventing products, and driving the industry, they would patent all of their ideas and creations to keep their edge in the market. On the other hand, if a company is doing poorly, perhaps they will file more patents to keep their intellectual property safe. We will look

at the number of patents filed per year by each company to see if they are indicators to the company’s performance.

We use Google Patent search [7] with a filter on the filing year, shown in figure 6-2. The oldest patents we were able to find for RIM were from 1999. Between 2003 and 2006, the number of patents filed increased threefold. These numbers indicate that RIM was still in the forefront of their industry. Even in 2009, when they reduced their company size by 10%, they peaked in the number of filed patents at 510. The number of patents filed is comparable to Apple’s numbers starting in 2003. We believe that the drop in numbers in 2011 is due to incomplete information in the search database given how recent 2011 was. This drop is seen in the graphs for all three companies.

6.1.3 Profiles and Interests

Of the profiles that list interests, 136 hold current or past positions at Research In Motion. These profiles list a total of 459 interests, with the most popular ones in table 6.1. Further down the list of popular interests are interests that are relevant to RIM the company. Interests such as “wireless”, “gadgets”, and “blackberry” occur in multiple profiles. This should be encouraging to RIM as some of their employees are passionate about the products they work on.

Interest	# of profiles	Interest	# of profiles
traveling	13	hockey	12
reading	11	new technology	10
technology	10	snowboarding	9
running	9	music	9
photography	8	mountain biking	8
skiing	8	hiking	8

Table 6.1: Top interests of Research In Motion employees

We also find the pairs of interests that co-occur within a profile. Since the data size is much smaller, the number of links is also smaller, 326 unique pairs total. We can see that the links are much weaker as only two pairs appeared three times and 28 pairs that appear twice. Table 6.2 lists a few of the more interesting pairs that appear twice.

Interest pair	# of profiles	Interest pair	# of profiles
technology, traveling	3	traveling, yoga	3
blogging, investing	2	hockey, new technology	2
scuba diving, skiing	2	fitness, traveling	2
photography, snowboarding	2	cooking, yoga	2
cricket, table tennis	2	music, photography	2

Table 6.2: Top pairs of interests for Research In Motion employees

Figure 6-3 shows the interest network for the Research In Motion employees. We see three distinct clusters of interests. The main cluster is primarily sports, along with music and photography. Another cluster is just “traveling”, “yoga”, and “cooking”. The last cluster is mainly technology-related interests. While the clusters are smaller than the clusters in the large interest space (discussed in Chapter 7), this seems to suggest a “geeks” vs “jocks” types of people in the workplace.

Of these 136 profiles, 120 of them have only one work position listed. All but one profile are current employees of Research In Motion. The other companies that these profiles list are shown in 6.3.

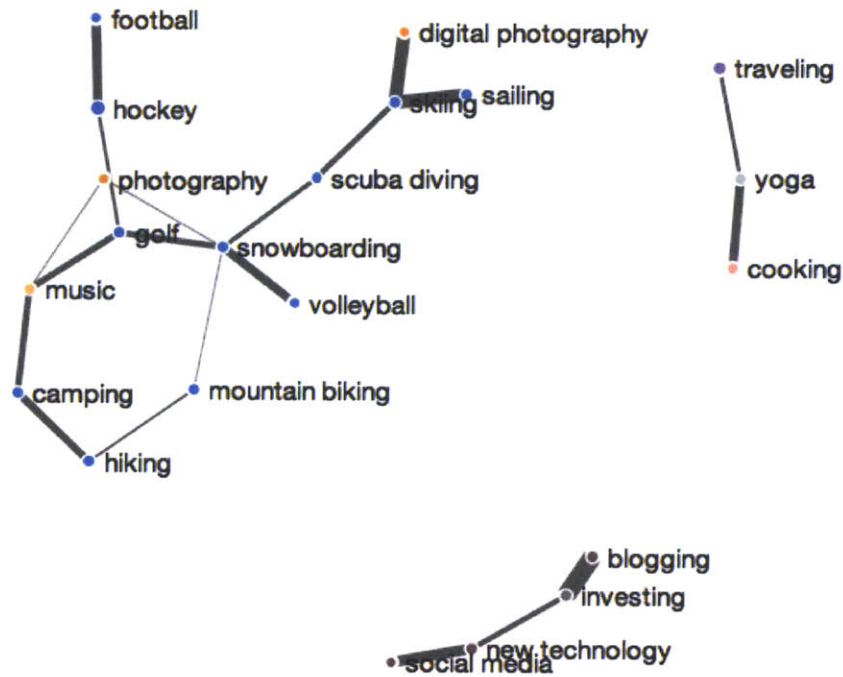


Figure 6-3: Interest graph for Research In Motion

Hanov Solutions Inc.	PolicyPak.com	Cartmell Properties
Elive Consulting	SVP Waterloo Region	OpenDBM Database Marketing
PG Enterprises	Scoopal Inc.	Junior Achievement of Waterloo Region
Iotum	QNX Software Systems	Composure Business Services
Seenopsis	University of Waterloo	Camz Software Enterprises

Table 6.3: Other companies for which Research In Motion employees have worked

The one profile that is not a current Research In Motion employee worked for RIM two years ago.

We extract the length of time an employee has worked at RIM by looking at the year they started working there (table 6.4).

Year started	# of profiles	Year started	# of profiles
2005	1	2006	0
2007	1	2008	7
2009	17	2010	28
2011	72	2012	10

Table 6.4: The years of which Research In Motion employees started their positions

6.1.4 Discussion

A user of Quora¹ posted the question “What would you do if you were the CEO of Research In Motion?” [17] which, at the time of this writing, has 12 detailed answers. Many of the answers suggest that since RIM is extremely behind in the smartphone market, they should suspend their work in the individual consumer business and refocus their efforts on the needs of enterprises, i.e. large-scale environments of companies.

The underlying theme is that RIM had a great product but only provided incremental improvements over the years. When a competitor appeared, their reaction was slow and poorly executed. This kind of challenge to their business should be met with more than just iterating on their past. They stopped innovating at the peak of their success.

The small amount of interest links in this data set limits our ability to conclude whether or not Research In Motion has benefited from an eclectic set of employees. Their products used to lead the market and owning a BlackBerry was almost a status symbol among business people. However, their inability to create new technology

¹Quora [16] is a question and answer website. The content is created and edited by the community of users.

that effectively competes with the iPhone and Android has cost their company. Thus if our hypothesis proves correct, they will greatly benefit from hiring people who can think differently to develop radically different products.

6.2 Schneider Electric

6.2.1 The Company

Schneider Electric (SE), headquartered in France, is an energy management company that creates technology focused on the safety, reliability, and efficiency of buildings. It is an old company, founded in 1836, that has reinvented itself over time to adapt to the world's changing economies. It is truly a global company as it has locations on six continents.

Schneider Electric began in the steel industry, at the height of the industrial revolution, working on train tracks, ships, and heavy machinery. After the world wars, the company was reorganized and focused toward construction and steel products for the civilian sector. Near the end of the 20th century, SE started expanding into the electrical industry, and after some acquisitions of other companies, Schneider became the second largest electricity distributor of Europe.

SE now develops products that improve the electrical and energy use of large buildings. From power monitoring systems to room occupancy sensors to surge protection devices, Schneider Electric leads the industry with its wide range of products.

Here is some basic information about Schneider Electric:

Mission statement: At Schneider Electric, we provide turnkey solutions to reduce the energy and operational inefficiencies of your building systems. We fulfill the critical

role of first uncovering the real causes of facility issues. Clients' needs come first and accountability on all levels is simply a way of doing business. Enduring performance provides more than just peace of mind. By focusing intently on diagnosis and education, Schneider Electric ensures that the projects developed minimize owners' risk while delivering relevant business value. We deliver solutions. We Deliver Enduring Performance.

Company size: 140,000+

Locations: Headquartered in France; other locations in North and South America, Europe, Africa, Asia, and Australia; operates in over 100 countries

6.2.2 Performance



Figure 6-4: Stock chart of Schneider Electric (blue), Siemens AG (red), and General Electric (yellow)

Schneider Electric began publicly trading in 1999. In figure 6-4 (blue line), we see that SE's stock performance has had ups and downs, but in general, it trends upwards. It seems to be doing better than General Electric and about the same as Siemens AG, both being market competitors of SE. The slumps in 2001 and 2008 are most likely due to the global recessions. As SE has focused on home and business energy efficiency, the relevancy and need of their products will be stable through the years and possibly increase as the energy crisis grows more urgent.

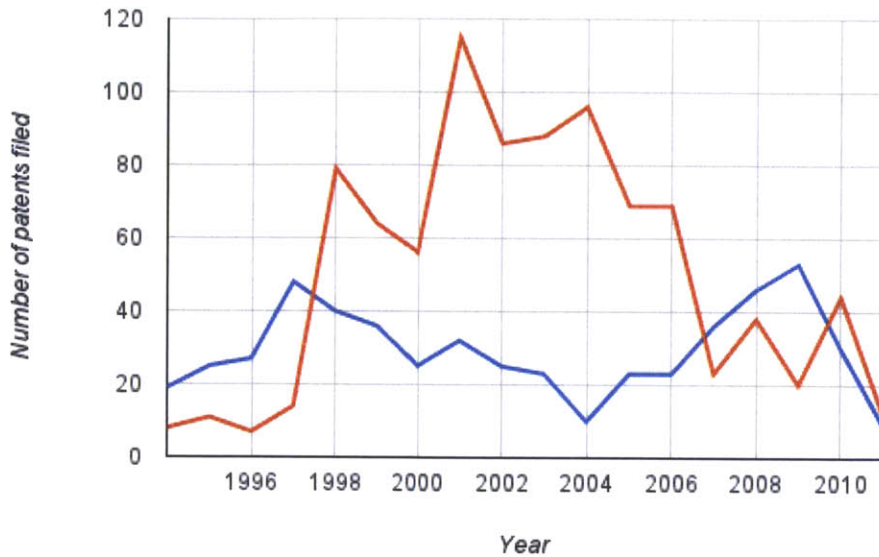


Figure 6-5: Number of patents filed each year for Schneider Electric (blue) and Siemens AG (red). We did not include General Electric as it is an American company and has filed an order of magnitude more patents than Schneider Electric.

Because SE’s products span a wide variety of sectors, they have not one but many competitors. This actually works in their favor as their competitors do not directly compete with them in every market, just a few. This is greatly different from Research In Motion’s situation, where they focused on one really successful product, which caused their decline when that product was challenged. The old adage “don’t put all your eggs in one basket” describes these two situations perfectly.

Schneider Electric’s patent history (figure 6-5) is less revealing than it was for Research In Motion. For one, since SE is a global company with a large European base, they likely file less patents in the United States. The peak is 53 patents in 2009, which is still an extremely low number for a technology company in the energy and manufacturing sector. The patent numbers for Siemens AG, a German company in the same industry, is also shown. General Electric, an American company, is not

Interest	# of profiles	Interest	# of profiles
traveling	7	golf	6
reading	5	professional networking	4
photography	4	music	4
fishing	4	swimming	3
international travel	3	basketball	3
mountain biking	3	strategic planning	3

Table 6.5: Top interests of Schneider Electric employees

Interest pair	# of profiles	Interest pair	# of profiles
golf, traveling	3	golf, mountain biking	3
fishing, golf	2	international travel, professional networking	2
fishing, traveling	2	professional networking, strategic planning	2
golf,reading	2		

Table 6.6: Top pairs of interests for Schneider Electric employees

shown as their numbers are an order of magnitude larger.

6.2.3 Profiles and Interests

Of the profiles that list interests, 62 hold current or past positions at Schneider Electric.

We collected 271 unique interests from these profiles and 59 unique pairs. The most common interests are listed in table 6.5, and the most co-occurring pairs are in table 6.6. As the number of profiles in this set is about half of the number of profiles we had for Research In Motion, the number of interests and interest pairs are also greatly reduced.

Figure 6-6 shows the interest network for the Schneider Electric employees. We see two distinct clusters of interests. The main cluster is primarily sports, along with “music”. The other cluster is work-related interests, along with “photography”. This also supports the “geeks” and “jocks” theory.

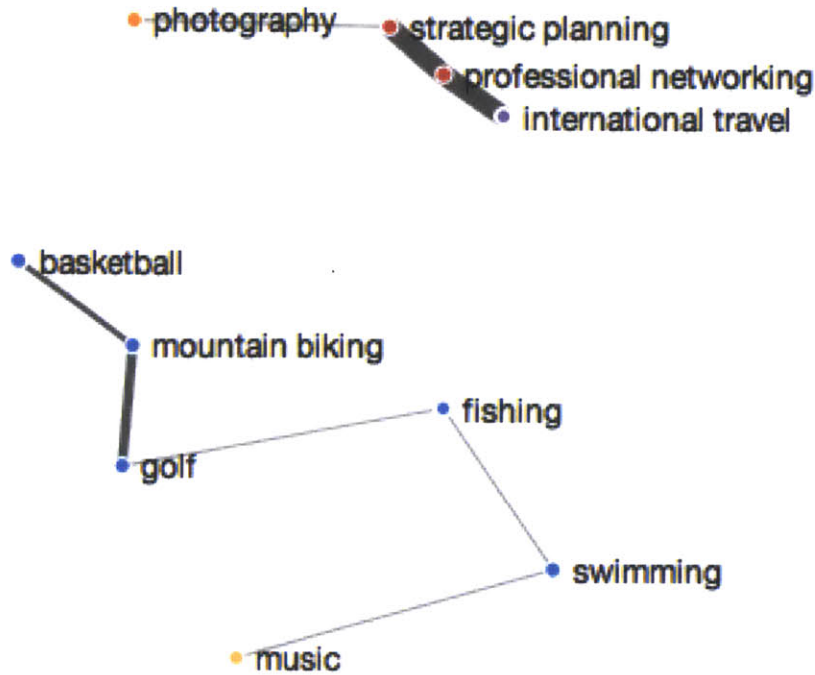


Figure 6-6: Interest graph for Schneider Electric

Sustainable Business Hub	Sustainable Healthcare	Grenoble Ecole de Management
--------------------------	------------------------	------------------------------

Table 6.7: Other companies for which Schneider Electric employees have worked

Five profiles listed more than one work position, but most of these held different positions at Schneider Electric. Table 6.7 lists the three other companies these SE employees have worked at. Only one profile is not a current SE employee. This is a clear limitation of the LinkedIn API as it has restricted our data set to mostly current employees.

Excluding the person that is not a current Schneider Electric employee, table 6.8 shows the years that the current employees started their positions. We postulate that the peak years, 2010 and 2011, are a result of the increase in LinkedIn usage of

Year started	# of profiles	Year started	# of profiles
1997	1	2005	3
2007	3	2008	1
2009	7	2010	22
2011	23	2012	1

Table 6.8: The years which Schneider Electric employees started their positions

younger employees. Since SE is an established company of almost two centuries, it's unlikely that our limited data reflects the actual hiring pattern.

6.2.4 Discussion

The size of our data set for Schneider Electric is difficult to work with as there aren't many profiles. However, their company profile of interests is slightly different from the set of interests of Research In Motion. While many of the top interests are the same ("traveling", "reading", "photography"), SE employees do not list "technology" and "new technology" like RIM employees do. Alternatively, interests such as "professional networking" and "strategic planning" do not appear in RIM's set of interests. While RIM is certainly a technology company, SE's interests seem to imply that these SE profiles are less technologists and more business-oriented people.

Unlike the RIM employees, very few Schneider Electric employees listed interests that matched their company's interests. Looking at some of the interests of the whole data set, "energy efficiency", "alternative energy", "green buildings", and "efficient homes" were listed, but only a few of these were SE employees². It would be interesting to see if this is true of SE employees as a whole, not just the ones we have profiles for.

²Other companies whose employees had these interests were Shell, Catenon Worldwide Executive Search, Amerigon, and Eleven LLC.

6.3 Intel Corporation

6.3.1 The Company

Intel Corporation was founded in 1968 in Mountain View, California. Since its beginning, their abilities to make semiconductors has kept them in the forefront of technological innovation. Their main products are still microprocessors and memory storage chips, and, sticking true to the law named after their founder³, they release faster and faster chips every year.

Intel's chips are interesting in that while they are essential to a computer, they are only a piece of the necessary hardware. Thus consumers did not necessarily know the value of Intel's products. It is similar to making the engines of all the cars; a consumer would not know the manufacturers of all the component pieces. In 1991, Intel started the "Intel Inside" marketing campaign, where all the computers that used Intel processors advertised an Intel sticker. This made Intel a household name in computer technology.

Here is some basic information about Intel:

Mission statement: Delight our customers, employees, and shareholders by relentlessly delivering the platform and technology advancements that become essential to the way we work and live.

Company size: 100,000+

Locations: Headquartered in California, USA; other locations in North and South America, Asia, Europe, the Middle East, and Africa; operates in 45+ countries

³Gordon Moore, the founder of Intel, has a rule of thumb that the number of transistors placed on a single chip will double roughly every two years. This is known as Moore's Law and has held true since the 1965 paper that detailed this prediction. [40]

6.3.2 Performance

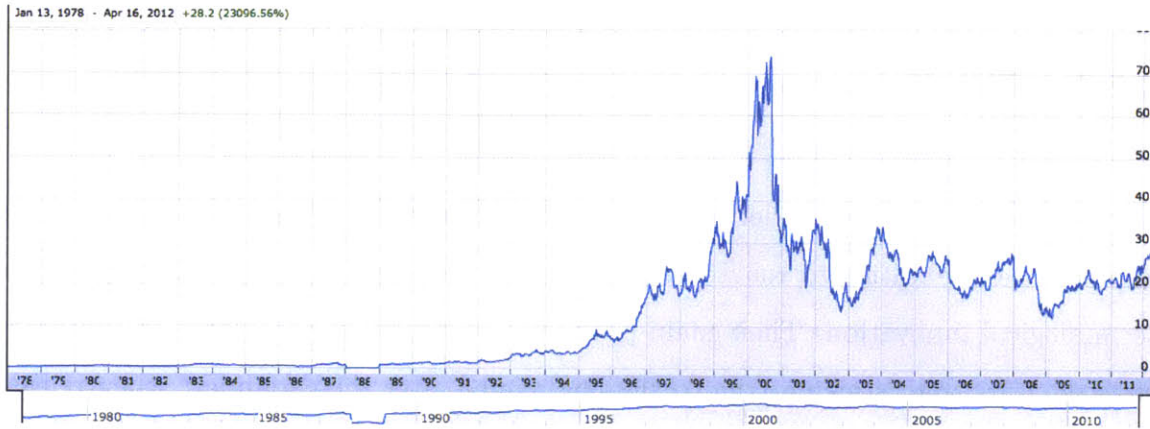


Figure 6-7: Intel Corporation stock chart

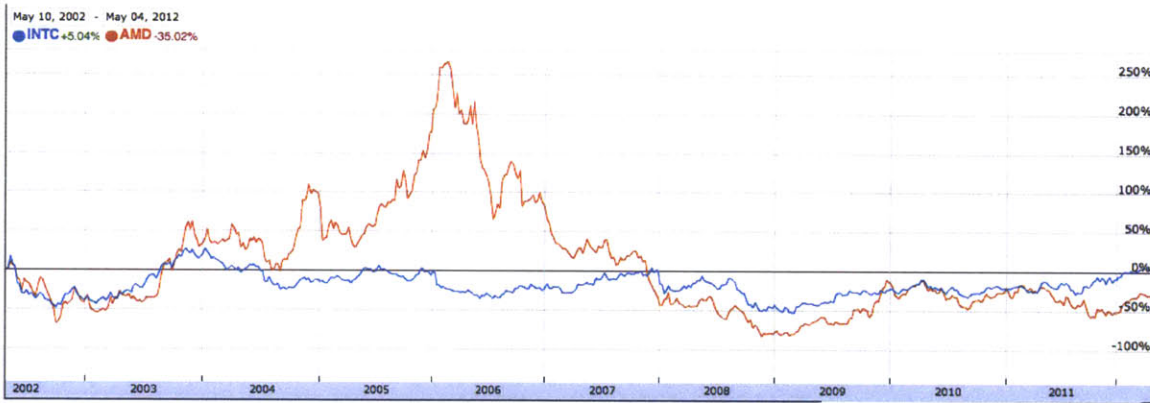


Figure 6-8: Stock chart of Intel Corporation (blue) and AMD (red)

Intel went public in the late 1970s, but its stock value held solidly low for almost 20 years (figure 6-7). When the personal computer market rose in the late 1980s and boomed in the 1990s, many of these computers had Intel-manufactured hardware inside. The “Intel Inside” campaign had direct impact on Intel’s success, and Intel’s stock exploded exponentially until the internet bubble burst in 2000.

Overall, their stock performance is relatively stable with a slow if constant total growth. Their main competitor, AMC, did better in the mid-2000s, but both are about the same now (figure 6-8). The recession in 2008 barely affected them, and

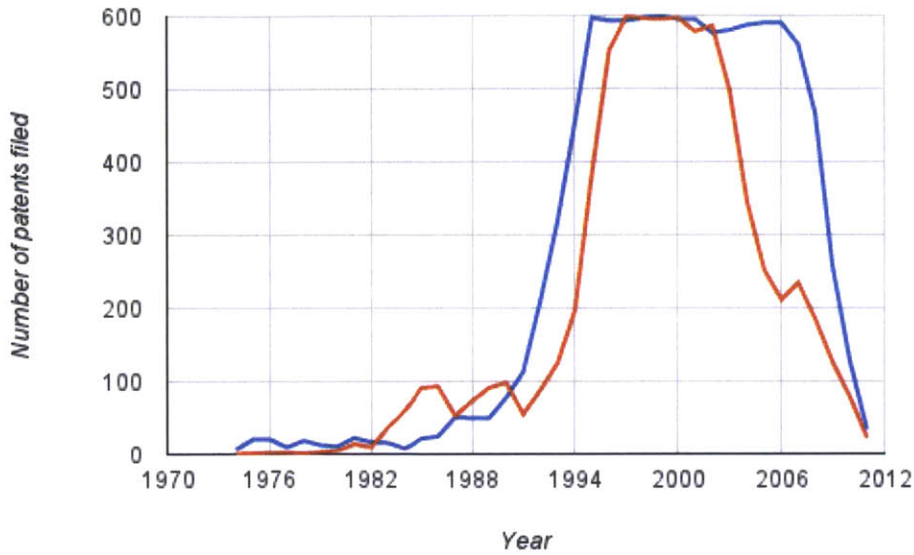


Figure 6-9: Number of patents filed each year for Intel Corporation (blue) and AMD (red)

by being the component supplier of the basics of technology, they can weather many economic downturns.

The data showing patent numbers over the years (figure 6-9) for Intel and AMD is also illuminating. We see a rapid growth in number of patents filed beginning in the early 1990s, just before the stock performance growth begins. The graph seems to abruptly peak at 600, and since the numbers for this graph come from Google search results, we suspect that this is a limiting aspect of the search engine as we do not expect that a company would cap their own patent submissions. The drop in numbers after 2009 is also suspicious, and rather than speculate on whether this is indicative of Intel's business, we assume this reflects the amount of time it takes for patents to be indexed by the search engine.

Interest	# of profiles	Interest	# of profiles
traveling	10	reading	8
photography	5	snowboarding	4
new technology	4	skiing	4

Table 6.9: Top interests of Intel employees

Interest pair	# of profiles	Interest pair	# of profiles
reading, traveling	4	reading, tennis	2
basketball, snowboarding	2	management training, tennis	2
photography, traveling	2	professional networking, tennis	2
soccer, traveling	2		

Table 6.10: Top pairs of interests for Intel employees

6.3.3 Profiles and Interests

Of the profiles that list interests, 56 hold current or past positions at Intel Corporation.

We collected 229 unique interests from these profiles and 55 unique interest pairs, the most occurring of each are in tables 6.9 and 6.10, respectively. Since the number of profiles we have for Intel is small, the popularity of interests and frequency of links are also small.

Figure 6-10 shows the interest network for the Intel employees. Since the number of interests is so small, only one cluster appeared. This consists primarily of sports interests along with “family”, “reading”, “traveling”, and “management training”. The outlier interest here is “management training”, as this is the only interest that is work-related.

14 profiles had work histories that included more than the current position, but only

Your Business Maven	Nerium International	United Way California Capital Region
L. Michele Designs	Startup Weekend	Whiskipedia

Table 6.11: Other companies for which Intel employees have worked

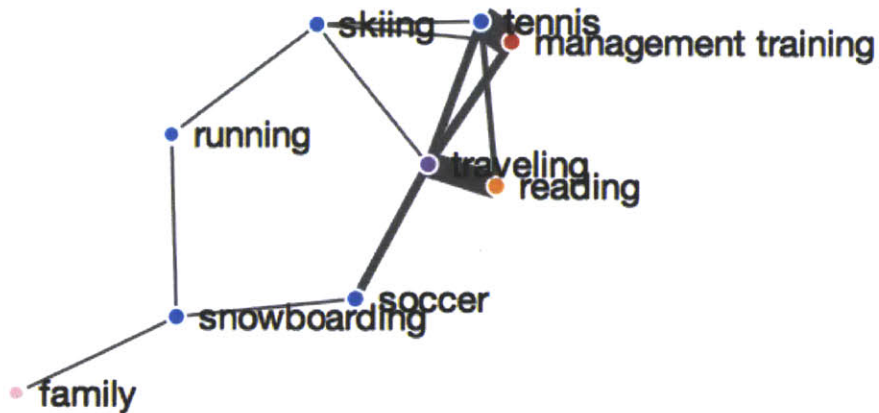


Figure 6-10: Interest graph for Intel

Year started	# of profiles	Year started	# of profiles
2005	2	2006	1
2007	1	2008	2
2009	4	2010	13
2011	26	2012	7

Table 6.12: The years which Intel employees started their positions

one profile was not a current Intel employee. Most of the previous work descriptions were previous positions also held at Intel. Table 6.11 lists the other companies that these Intel employees worked at.

Similar to Research In Motion and Schneider Electric, many employees seem to have joined Intel in 2010 and 2011. Given that we are unable to retrieve complete work histories, it is difficult to conclude anything from this pattern. We speculate that perhaps the type of people who use LinkedIn are younger employees who are just out of the job market, or perhaps the popularity growth of LinkedIn in these past couple years biases our data towards people who started new jobs recently.

6.3.4 Diversity at Intel

Diversity, in the demographic sense, is incredibly important to Intel. They support internal employee diversity groups based on race, nationality, religion, and sexual identity. For three years, they received a 100% rating on the Corporate Equality Index from the Human Rights Campaign.

Intel clearly understands the importance of diversity in innovation as it is highlighted on their website. They realize that since they are a global company, they need to be as diverse as their customers in order to stay competitive. “This worldwide perspective helps us anticipate and provide for the growing needs of a changing marketplace” - Intel.com

6.3.5 Discussion

Unlike Research In Motion, Intel employees did not have any interests listed that match the company’s products. The closest interest is “new technology”, which is a vague and broad term. In the complete set of interests, “microprocessors” and “semiconductors” were in a few profiles, but none of these were Intel employees.

Our data is far from complete enough to make a conclusion about employee interests and their impact on a company’s success. However, we raise more questions that this data, if complete, would be able to answer. For instance, is it advantageous for the employees to have interests that match the company’s product? We hypothesize that this is a measurable benefit if the employees are personally interested about their line of work. This would lead to greater work satisfaction and passion for that company, rather than employees that just consider it a way to make a living. However, we do not know if these in-line-with-company interests would contribute to an overall cognitive diverse group. Perhaps the ideal makeup is a mix of both people who have

in-company interests and those who don't. This may allow for the greatest amount of cognitive diversity that fuels innovation yet still provides the expertise and focus on the company's products.

6.4 Summary

These three companies were chosen solely for the number of profiles with interests we had. Although the products of these companies are all related to technology, they span different sectors of technology: mobile devices, energy, and semiconductors. We looked at the stock performance, number of patents filed over the years, and the interest graph for each company.

The most interesting result is the “jocks” vs “geeks” separation in the individual company interest graphs. In both Research in Motion and Schneider Electric, there were distinct clusters between sports-related interests and job-related interests. We will discuss the clusters for the large interest graph in the next section, and it would be fascinating to see if this holds true for other companies.

Chapter 7

Conclusions and Future Work

This thesis has presented the implementation process of using an online social network for research, the creation of the “Interest Space”, and case studies of three companies in the context of their employees’ interests. In this chapter, we discuss the implications of this research and possible future directions of the work.

7.1 LinkedIn as Research

LinkedIn, with its 150 million members, is a wealth of data for organizational research. Since we targeted specific companies for our profile search, we retrieved almost 20,000 profiles for our data set through the LinkedIn API. “Interests” is not a field that most people fill out, and in our data set, only 2122 (about 11%) of profiles listed interests. This was sufficient for our purposes in creating an “Interest Space” as these profiles had 5359 unique interests.

Using the LinkedIn API to access profile information is not be the best way to get data. Although our resulting data set was large in size, the restrictive nature of the

API limited our abilities to build a complete data set. Many profiles have privacy settings to protect their information, but the API is even more restrictive than the privacy settings. A user may be able to see another profile’s full information while the API only allows us to retrieve a subset of that information. This is most likely to prevent web crawlers from abusing the API. However, this prevents us from using the available data to the full potential for research. We would advocate for open data APIs, at least for research institutions.

Being able to access the full extent of LinkedIn’s data would be greatly beneficial to this research, but we also would like to know what the threshold of completeness is statistically similar to the complete data. In general, the greater the sample size, the better the results mirror the full population. Given all the biases we recognize in section 3.3, the effects of these biases can be reduced with a sufficiently large sample size.

We would like to compare the network of LinkedIn to other social networks, like Facebook or Orkut. Facebook has a much larger user base, but the Facebook environment is different than LinkedIn’s. Facebook has emphasized the “social” aspect of a social network. All activity is exposed to a user’s connections, and Facebook has created a platform for social gaming. Our guess is that a higher percentage of Facebook profiles list interests than LinkedIn, while a smaller percentage have a complete work history, but this will need to be verified.

Orkut [12] is a smaller social network that is relatively unknown in the United States. It has 66 million users worldwide, with the majority of them in Brazil [13]. Since the majority of our profiles were located in the United States, comparing the LinkedIn profiles with the Orkut profiles would be interesting as we would be able to see how different cultures affects the data.

Online social networks are a phenomenon of the last decade, and APIs have only recently become available and popular. With the rise of big data research [37], online

networks have a wealth of data for researchers to better understand human behavior and organization. While APIs are a great way to access some of the information, the best way would be to petition the social network company for open access to the data for research.

7.2 The Interest Space

From the LinkedIn profiles that had interests, we created an “Interest Space” network that visualizes how interests are shared among people (figure 7-1). Interests that appear together a significant amount of times have strong links in the network. The interests are also colored based on a categorization we did using natural language processing.

The first thing we notice about the Interest Space is that the interests are clearly divided into two groups. The first group (figure 7-2) includes 58 interests, with “technology” having the most links. The majority of these interests appear to revolve around job activities, like “marketing”, “professional networking”, and “innovation”. Even “photography” can be considered a job-related interest as it can relate to photographers, artists, and designers.

The second largest cluster (figure 7-3) has 26 interests, and the thing to note is that every interest was categorized as a “sport”. In fact, the only sports interests that are not in this cluster are “badminton”, “golf” and “sports”. This leads us to speculate that there is a clear distinction in the LinkedIn profiles between so-called “jocks” and “geeks”. Since the two interest clusters are separate, we can imagine that the “jocks” are the types of people who consider their jobs as a means to a living and love sports in their free time, and the “geeks” are people who are passionate about their careers, where they happen to be interested in activities that relate to their job and thus love what they do.

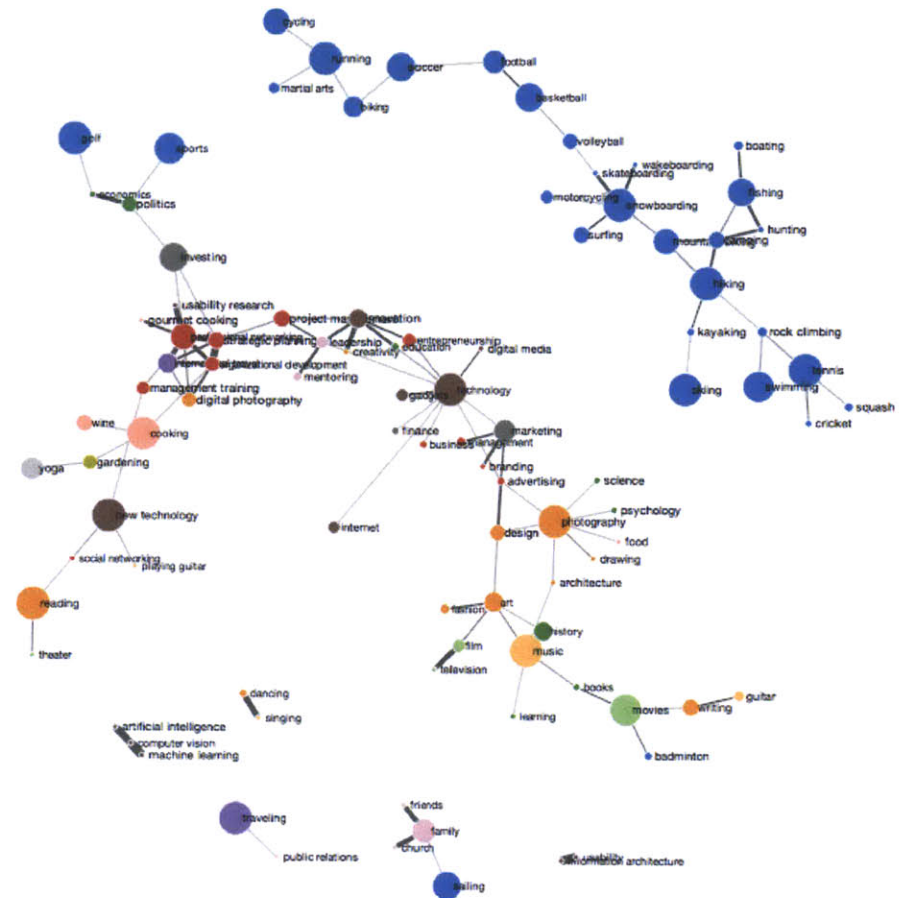


Figure 7-1: The Interest Space visualization

For future analysis, we would like to see if the “jocks” and “geeks” theory has other indicators in the profiles. For example, there may a significant difference between the types of jobs they hold such as management or engineering positions. Or perhaps there is a difference in the average length of time at each job, which can be an indicator for job satisfaction. If these distinctions hold true, then we will be able to profile people based on just their interests, which can be a powerful way to observe human resource related trends within a company or an industry.

An interesting thing to note is that two of the sports interests are in the large cluster. One is that “golf” is connected to the network through “economics”. This seems to reinforce the stereotype that golf is primarily a businessman sport. The generic

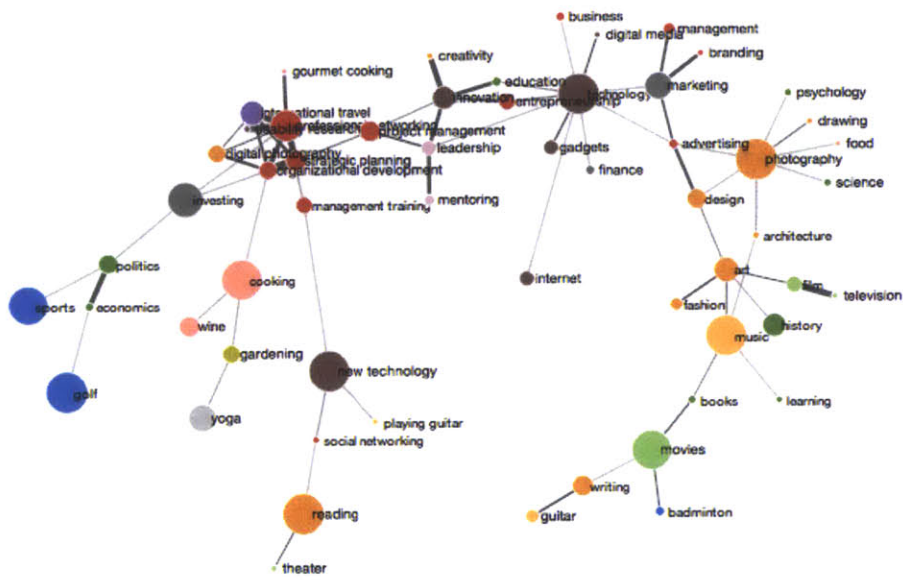


Figure 7-2: The main cluster of interests includes many job-related interests.

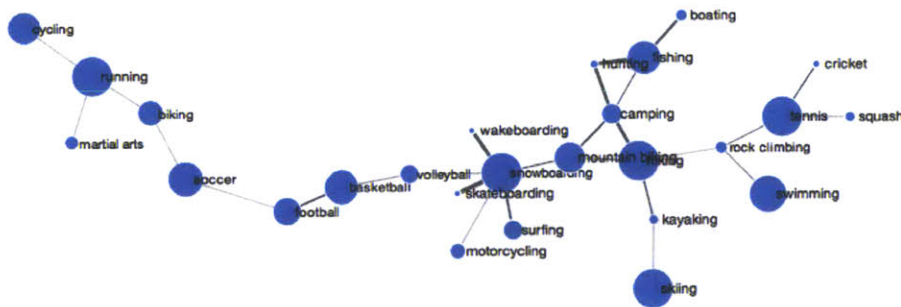


Figure 7-3: The sports cluster of interests

interest of “sports” is connected to the rest of the network through “politics”, both of which are general conversation topics, and all of these interests are connected to the main part of the network through “investing”. This leads us to believe that this isolated branch of interests came from those working in the financial sector.

Other than the two large interest clusters, the rest of the significant links are found in 13 interests that are not connected to each other (figure 7-4). These are mostly pairs. Two of the clusters are specifically related to technology: “usability” and “information architecture”; “artificial intelligence”, “computer vision”, and “machine learning”.

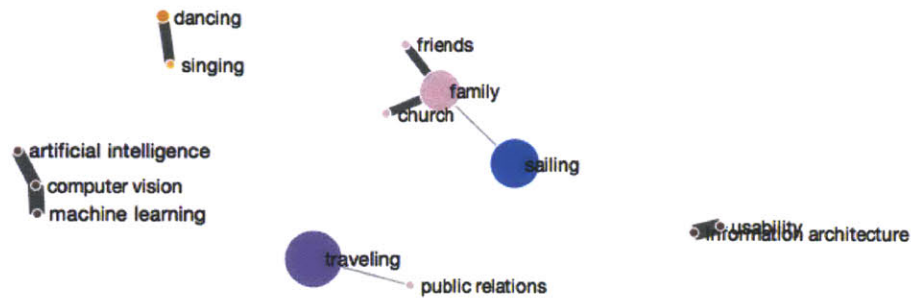


Figure 7-4: Smaller clusters of interests

“Singing” and “dancing” form a pair, which seem to be performance-related interests, although we would have expected these to connect to “theater” or something similar.

An additional feature of the network graph that we would like to implement is creating a sliding threshold of the significant links. This would allow us to better see how strong the links are, and perhaps at a slightly lower threshold, the entire network would be connected.

We can also look at how the interests are “unrelated” or inversely related. Which interests most commonly do *not* occur together and why do these pairs happen? Is there something about one interest that would make someone less likely to participate in the other one?

The interest categories are helpful for us to visualize how the nodes are related. Follow-on work would be to create a measurement of how specific an interest is. Some interests, like “sports”, are very generic while others, like “digital photography”, are quite specific. This distinction of specificity can be useful in determining the level of a person’s interest in something. For example, a person with an interest of “sports” may be more likely to be a sports spectator while someone interested in “tennis” may actually play the sport as a hobby.

This distinction in specificity is even more important when the generic and specific version of an interest are in different places on the graph. For example, “photog-

raphy” and “digital photography” are connected to different nodes. “Photography” is clustered with other art-related interests like “design” and “architecture”, while “digital photography” is linked to “organizational development”, “professional networking”, and “international travel”. We speculate that those who list “photography” are ones who actually use photography in their work while those that indicate interest in “digital photography” are hobbyists in photography and primarily do non-art related work.

Although the Interest Graph was created with a small subset of the LinkedIn profiles, it can already give us immediate insights into who these people are. We have assumed some characteristics about the people who list these interests, and the next step would be to verify these assumptions with more robust data.

7.3 Company Interests

The big picture of this work is exploring enterprise agility, to understand how companies can better adapt their organizations to solve big problems, including innovating during economic downturns. The three case studies we did had mixed results. Research In Motion’s employees tended to have interests that were in sync with the company’s products, like “new technology” and “gadgets”, but their business has declined in the last few years. The employees of Schneider Electric, however, did not have interests that matched their company’s interests, although this does not seem to have affected their stock performance. Intel’s patent filing history seemed to foreshadow their growth, and again, their employees’ interests were uninteresting.

These results are inclusive as three case studies is not enough to indicate a trend. We were limited by the amount of data retrieved and these companies were chosen as we had the most number of profiles for them. Immediate future work would be to expand the case studies to include more companies from a larger variety of industries.

This would allow us to investigate what level of granularity is possible and useful to quantify interest diversity and understand its impact, and thus better generalize any trends and implications for a wider audience.

7.4 Final Thoughts

This thesis is a step in better understanding the role of diversity, specifically people's interests, in the success of organizations. We hope the eventual impact of this research is that companies can place a quantitative value on the activities of their employees when they are *not* at work. For example, we can imagine that the employees of Wal-Mart and Google are different types of people not necessarily in the jobs that they do, as both companies can have sales people or computer engineers, but in what they choose to do in their free time. In analyzing LinkedIn profile data, we are trying to tease out this difference and quantify it in terms of company productivity and innovation.

Perhaps companies can even encourage their employees to not only seek education and training for job-related tasks, but also a wide range of activities. The skills an employee gets from their hobbies (patient practicing for a musical instrument, understanding of heat distribution from glassblowing, or gestural significance from orchestral conducting, etc.) do not always directly relate to their job, but these skills may be indirectly related or unconsciously used during brain-storming sessions or task executions. We expect that these benefits are measurable and will impact future organizational research.

Bibliography

- [1] Bootstrap - simple and flexible html, css, and javascript for popular user interface components and interactions. <http://twitter.github.com/bootstrap>.
- [2] Common sense computing initiative. <http://csc.media.mit.edu/>.
- [3] D3.js - a small free javascript library for manipulating documents based on data. <http://mbostock.github.com/d3>.
- [4] Divisi: a sparse toolkit for python. <http://csc.media.mit.edu/docs/divisi2/index.html>.
- [5] Facebook. <http://www.facebook.com>.
- [6] Gartner says sales of mobile devices in second quarter of 2011 grew 16.5 percent year-on-year; smartphone sales grew 74 percent. <http://www.gartner.com/it/page.jsp?id=1764714>.
- [7] Google advanced patent search. http://www.google.com/advanced_patent_search.
- [8] jQuery - a fast and concise javascript library that simplifies html document traversing, event handling, animating, and ajax interactions for rapid web development. <http://jquery.com>.

- [9] LinkedIn announces 150 millionth user. <http://blog.ukfast.co.uk/2012/02/12/linkedin-announces-150-millionth-user>.
- [10] LinkedIn developers API. <https://developer.linkedin.com>.
- [11] Open authentication. <http://oauth.net>.
- [12] Orkut - social networking and discussion site operated by google. <http://www.orkut.com>.
- [13] Orkut search traffic. <http://www.alexa.com/siteinfo/orkut.com>.
- [14] Python LinkedIn. <http://code.google.com/p/python-linkedin>.
- [15] Python OAuth2. <https://github.com/synedra/python-oauth2>.
- [16] Quora. <http://www.quora.com>.
- [17] Quora: What would you do if you were the ceo of research in motion? <http://www.quora.com/What-would-you-do-if-you-were-the-CEO-of-Research-in-Motion>.
- [18] Social Energy. <http://vimeo.com/joulesm/social-energy>.
- [19] Viral Spaces research group. <http://viral.media.mit.edu>.
- [20] Jeffrey M. Bacidore, John A. Boquist, Todd T. Milbourn, and Anjan V. Thakor. The search for the best financial performance measure. *Financial Analysts Journal*, 53(3):11–20, May 1997.
- [21] Ronald S. Burt. The contingent value of social capital. *Administrative Science Quarterly*, 42:339–365, 1997.
- [22] Manuel Castells and Peter Hall. *Technopoles of the World: The making of 21st century industrial complexes*. Routledge, London, 1994.

- [23] Sanjib Chowdhury. Demographic diversity for building an effective entrepreneurial team: Is it important? *Journal of Business Venturing*, 20(6):727–746, Nov 2005.
- [24] Dennis J. Devine, Laura D. Clayton, Jennifer L. Philips, Benjamin B. Dunford, and Sarah B. Melner. Teams in organizations: Prevalence, characteristics, and effectiveness. *Small Group Research*, 30(6):678–711, Dec 1999.
- [25] C. A. Hidalgo et al. The product space conditions the development of nations. *Science*, 317:482–487, 2007.
- [26] Jon Gertner. *The Idea Factory*. Penguin Press, United States of America, 2012.
- [27] Catherine Havasi, Richard Borovoy, Boris Kizelshteyn, Polychronis Ypodimatopoulos, Jon Ferguson, Henry Holtzman, Andrew Lippman, Dan Schultz, Matthew Blackshaw, Greg Elliott, and Chaki Ng. The Glass Infrastructure: Using common sense to create a dynamic, place-based social information system. *Association for the Advancement of Artificial Intelligence*, 2011.
- [28] Jesse Hicks. Research, no motion: How the blackberry ceos lost an empire. <http://www.theverge.com/2012/2/21/2789676/rim-blackberry-mike-lazaridis-jim-balsillie-lost-empire>.
- [29] C. A. Hidalgo and R. Hausmann. Network view of economic development. *Developing Alternatives*, 12(1):5–10, 2008.
- [30] C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009.
- [31] S. Horwitz and I. Horwitz. The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, 33(6):9871015, 2007.

- [32] John Hagel III, John Seely Brown, and Lang Davison. The best way to measure company performance. *Harvard Business Review*, Mar 2010.
- [33] Susan E. Jackson, Karen E. May, and Kristina Whitney. Understanding the dynamics of diversity in decision-making teams. *Team Effectiveness and Decision Making in Organizations*, pages 204–261, 1995.
- [34] Karen A. Jehn, Gregory B. Northcraft, and Margaret A. Neale. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly*, 44(4):741–763, Dec 1999.
- [35] David H. Jonassen. Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45(1):65–94, 1997.
- [36] Kalle Lyytinen and Daniel Robey. Learning failure in information systems development. *Information Systems Journal*, 9(2):85–101, April 1999.
- [37] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, May 2011.
- [38] Joseph E. McGrath. *Groups: Interaction and Performance*. Prentice Hall, Englewood Cliffs, NJ, 1984.
- [39] Todd T. Milbourn. The executive compensation puzzle: Theory and evidence. *IFA Working Paper*, 1996.
- [40] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), April 1965.
- [41] Frank M. H. Neffke and Martin S. Henning. Skill-relatedness and firm diversification. *Papers on Economics and Evolution*, 2009.

- [42] I. Nonaka, R. Toyama, and A. Nagata. A firm as a knowledge-creating entity: A new perspective on the theory of the firm. *Industrial and Corporate Change*, 9:1–19, 2000.
- [43] B. Nooteboom. *Learning and Innovations in Organizations and Economics*. Oxford University Press, Oxford, UK, 2000.
- [44] Scott E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, Princeton, NJ, 2007.
- [45] Zizi Papacharissi. The virtual geographies of social networks: A comparative analysis of facebook, linkedin, and asmallworld. *New Media Society*, 11(2):199–220, Feb 2009.
- [46] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, February 1988.
- [47] Gregory Schraw, Michael E. Dunkle, and Lisa D. Bendixen. Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology*, 9(6):523–538, December 1995.
- [48] A. J. Scott. Industrial organization and the logic of intra-metropolitan location: I. theoretical considerations. *Economic Geography*, 59(3):233–250, July 1983.
- [49] A. J. Scott. Industrial organization and location: Division of labor, the firm, and spatial process. *Economic Geography*, 62(3):215–231, July 1986.
- [50] Meredith M. Skeels and Jonathan Grudin. When social networks cross boundaries: A case study of workplace use of facebook and linkedin. *Proceedings of the ACM 2009 international conference on Supporting group work*, May 2009.
- [51] David J. Teece, Gary Pisano, and Amy Shuen. Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7):509–533, Aug 1997.

- [52] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111(2):447–504, Sept 2005.
- [53] Anne L.J. Ter Wal and Ron A. Boschma. Co-evolution of firms, industries and networks in space. *Regional Studies*, 2008.