Local Networks: Theory and Applications

by

Markus M Möbius

B.A., Mathematics, Oxford University, 1994
M.Phil., Economics, Oxford University, 1996

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2000

Author ....
Department of Economics
May 15, 2000

Certified by ....
Glenn Ellison
Professor of Economics
Thesis Supervisor

Certified by ....
Abhijit Banerjee
Professor of Economics
Thesis Supervisor

Accepted by ....
Peter Temin
Elisha Gray II Professor of Economics
Chairman, Department Committee on Graduate Studies

Local Networks: Theory and Applications

by

Markus M Möbius

Submitted to the Department of Economics
on May 15, 2000, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Chapter 2 analyzes a simple evolutionary model of residential segregation based on decentralized racism which extends Schelling's (1972) well-known tipping model by allowing for local interaction among residents. The richer set-up explains not only the persistence of ghettos, but also provides a mechanism for the rapid transition from an all-white to an all-black equilibrium. The analysis of the model introduces a new technique to characterize the medium and long-run stochastic dynamics of processes with many agents.

Chapter 3 looks at the curious dynamics of telephone competition between AT&T and the 'Independents' at the turn of the century. Competition between these two non-interconnected networks was initially vigorous and reached its peak between 1905 and 1907 but then declined rapidly. My analysis is based on the observation that urban markets subdivide into social 'islands' along geographical and socio-economic dimensions: users are more likely to communicate with subscribers 'inside' their island than with those 'outside' it. In a simple evolutionary model I demonstrate how minority networks can thrive and preserve their market share at a low state of development but become eventually extinct as the industry matures.

Chapter 4 explores why the division of labor first increased enormously during industrialization but has decreased again since the 1970s through the increased use of job rotation, flat hierarchies and autonomous work teams. This striking pattern in the organization of work can be derived in a model where (a) technology and market size determine the degree to which products are standardized and (b) more customized products are subject to trends and fashions which make production tasks less predictable and a strict division of labor impractical. The model also explains changes in the demand for skilled labor over time, predicts the rise of multi-purpose vs. single-purpose machines in advanced industrial economies, and provides a mechanism for trade between similar countries to affect wage inequality.

Thesis Supervisor: Glenn Ellison
Title: Professor of Economics

Thesis Supervisor: Abhijit Banerjee
Title: Professor of Economics

Für meine Eltern

Acknowledgments

If somebody had told me during my last year of high school that I would first study mathematics at Oxford University and then go on to join a graduate economics program in the U.S. I would not have believed ıım. Looking back at my academic career so far I am reminded of a Brownian motion - a journey in which I made friends and met teachers who changed the course of my life in unpredictable ways. The sum of their impacts made me what I am, and I want to use this space to thank them for everything they have done.

I begin by thanking the unknown secretary at UCCA who sent me the wrong application form when I tried to apply to Cambridge University. I ended up reading mathematics at Oxford because it was too late to order a second form. Special thanks go to my German high-school teacher, Willie Hirmer, for writing a great letter of recommendation, and to the kind (and meanwhile probably discarded) admissions computer who randomly allocated me to Mansfield College. My tutors Janet Dyson, and Bob Coats showed me the strengths of the Oxford tutorial system. They constantly challenged me and were generous with their time and advice. Markwart, Ingrid, Anthea, and Johanna taught me how to live. The countless walks with Jessie Baird kept me sane, and the BBC breakfasts at Balliol with Alexander, Matthias, Amelie, Stephane, Frank, and Martin were the highlight of every Sunday morning. My friend George Hatjantonas introduced me to economics through numerous discussions. Although I made fun of his infatuation with that 'silly and imprecise' discipline he had the last laugh when I switched to economics after my B.A.

This transition turned out to be a bumpy ride but was made tolerable with the help of my friends Alexander Simkin, Mary Duffy, Pietro Stella, and Meir Yaish. John Muellbauer and Chris Harris were great teachers and advisors. I am truly grateful to Paul Klemperer who took a keen interest in my academic progress as my college advisor. He convinced me to take a second look at the M.I.T. graduate economics program when I was all set to accept a different offer. I am convinced that this was one of the best decisions of my life. Thanks, Paul and Meg, for your support and friendship.

I enjoyed wonderful guidance and support at M.I.T. In Glenn Ellison, Abhijit Banerjee, and Daron Acemoglu I met the best advisors one could hope for. Abhijit's comments helped me focus my papers much more sharply. Daron provided key insights for the third paper, and I relied on his patience and intuition when I 'tested' half-baked ideas on him. A special 'thank you' goes to Sendhil Mullainathan for our numerous discussions in his office and for his comments on my third paper. I am particularly indebted to Glenn for his outstanding support. Learning how to do research can be pretty tough, and requires confidence and good advice. Glenn gave me both. His thoughtful comments on countless drafts were instrumental in improving my papers, and influenced my research interests and thinking. I am also grateful to Susan Athey, Peter Diamond, and Peter Temin for discussions, to Lones Smith for his honest comments on an early draft of my first paper, and to Mike Piore for introducing me to labor economics in a non-standard way.

My friends and fellow students at M.I.T. and in Boston helped make the last few years enjoyable. In particular, I want to mention Max Amarante, Peter Benczúr, Kobi Braude, Alain Jousten, Petya Koeva, Botond Kőszegi, Jon Levin, Alejandro Neut, Dirk Niepelt, Deane Norton, Kathy Yuan, and Ksenia Yudaeva. Very special thanks go to my dear friends Botond and Kobi for stimulating discussions on economics in particular, and life in general.

This whole journey would not have been possible without the unwavering support of my family. I am deeply grateful to my parents who supported my decision to study abroad in every possible way. They were always there when I needed them most, with advice and encouragement. I was fortunate to grow up with two great brothers, Matthias and Michael, who put up with me all these years (it hasn't always been easy). I wish my grandmothers could attend my graduation. Oma Friedel passed away last year but I will never forget my countless childhood visits to Frankfurt, Gemüseeintopf, and subway adventures. Oma Vera would be proud to attend my graduation but is too sick. She gave me confidence by always believing in me, and I will always be proud of her.

And finally, Tanya, my super mushroom - who makes me feel alive. Pluto, the nineth planet of our solar system with his unusually large inclination, must have felt a similar sensation when some comet from outer space kicked him out of his orbit a couple of billion years ago. Thank you for being there and making my life beautiful. And for being such a wonderful shadow. And for making it all worthwhile.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In Chapter 2 I analyze a simple evolutionary model of residential segregation based on decentralized racism which extends Schelling's (1972) well-known tipping model by allowing for local interaction among residents. The richer set-up explains not only the persistence of certain ghettos, but also provides a mechanism for the rapid transition from an all-white to an all-black equilibrium.

On one-dimensional *streets* segregation arises once a group becomes sufficiently dominant in the housing market. However, the resulting ghettos are not persistent, and periodic shifts in the market can give rise to "avenue waves". On two-dimensional *inner-cities*, on the other hand, ghettos can be persistent due to the "encircling phenomenon" if the majority ethnic group is sufficiently less tolerant than the minority. I review the history of residential segregation in the US and argue that my model can explain the rapid rise of almost exclusively black ghettos at the beginning of the 20th century.

For the analysis of my model I introduce a new technique to characterize the medium and long-run stochastic dynamics. I show that *clustering* predicts the behavior of large-scale processes with many agents more accurately than standard stochastic stability analysis, because the latter concept overemphasizes the 'noisy' part of the stochastic dynamics.

Chapter 3 looks at the curious dynamics of telephone competition between AT&T and the 'Independents' at the turn of the century. Competition between these two non-interconnected networks was initially vigorous and reached its peak between 1905 and 1907 but then declined rapidly.

My analysis is based on the observation that urban markets subdivide into social 'islands' along geographical and socio-economic dimensions: users are more likely to communicate with subscribers 'inside' their island than with those 'outside' it. In a simple evolutionary model I demonstrate how minority networks can thrive and preserve their market share at a low state of development when the volume of telephone traffic between distinct islands is small. As telephone ownership becomes more common, traffic between the islands increases and the lack of interconnection forces business users to subscribe to both networks. Duplication imposes a direct *cost of incompatibility* on users and erodes public support for competition. In addition, duplication is asymmetric in the sense that business subscribers to the smaller network are more likely to get a second phone. This *duplication effect* makes minority islands vulnerable to invasion by the majority system such that standardization can arise endogenously at later stages of development. Without mandatory interconnection dual service competition was therefore bound to be a transitory phenomenon. I underpin the implications of my model with empirical evidence from a panel data set of US cities.

Chapter 4 explores why the division of labor first increased enormously during industrialization, but has again decreased since the 1970s as job roles have expanded both horizontally through greater use of job rotation and the merging of job classifications, and vertically by introducing flat hierarchies and autonomous work teams. This striking pattern in the organization of work can be derived in a model where (a) technology and market size determine the degree to which products are standardized and (b) more customized products are subject to trends and fashions which make production tasks less predictable and a strict division of labor impractical. In the craft economy customization is high as artisans use a general purpose, constant returns technology. In contrast, the machine economy exploits scale economies to produce large volumes of identical goods. At the onset of industrialization, the market supports only a small number of generic varieties which can be mass produced under a strict division of labor. Thanks to productivity growth, niche markets gradually expand, producers eventually move into customized production and the division of labor decreases again.

I discuss a number of important extensions. First, the model explains the changing demand for skilled labor over time if skills only provide comparative advantage under a weak division of labor. Second, I demonstrate that conventional calculations of the factor content of trade underestimate the impact of globalization because they do not take into account changes in product market competition induced by trade. Finally, the path of technological progress can be endogenously derived. This provides an explanation for the emergence of control technologies such as flexible multi-purpose machines and bar-coding.

Chapter 2

The Formation of Ghettos as a Local Interaction Phenomenon

2.1 Introduction

Residential segregation along ethnic and racial lines is a fact of life in the US and in many other countries. There is a substantial body of literature which documents the economic and social costs of segregation. Ethnic sorting retards inter-generational improvements for relatively disadvantaged groups[1], reduces empathetic connections between ethnic groups and therefore diminishes political support for redistribution (as in Cutler, Elmendorf, and Zeckhauser (1993)) and increases statistical discrimination because whites, for example, end up relying more on stereotypes of blacks instead of actual experience (Wilson 1987).

While we know a great deal about the outcomes of residential segregation on society the mechanism which gives rise to ghettos[2] in the first place is much less understood. This paper provides a simple theory of segregation based on decentralized racism which can explain the key empirical facts. First, ghettos historically developed fairly rapidly: in the US the core of the black ghettos formed between 1900 and 1920. Second, black ghettos in particular

[1] Borjas (1995) found that ethnicity has an external effect even after controlling for parental background and the socio-economic characteristics of a neighborhood. Neighborhood peers appear to affect the skills and norms of the young in particular, such as the probability of being involved in crime and the propensity of youths to be out of school or work (see, for example, Case and Katz (1991), and Glaeser, Sacerdote, and Scheinkman (1996)). Cutler and Glaeser (1997) compared the outcomes of blacks between cities and found that blacks in racially more segregated cities earn less income and are more likely to become single mothers or drop out of high school.

[2] The term "ghetto" is used nonpejoratively throughout the paper in order to denote a racially or ethnically segregated community.

tend to be very persistent over time once they cover a large contiguous geographical area, such as Harlem in New York City.[3] Ghettos are far less stable, on the other hand, if they extend only over a single street as examples from Chicago's avenues at the turn of the century show. My model abstracts away from other contributing factors to segregation, such as sorting by socio-economic differences and collective-action racism. I argue in the empirical part of my paper that these effects on their own cannot explain the key empirical facts about segregation in the US.

Existing models of segregation are typically variants of Schelling's (1972) influential tipping model which can generate multiple stable segregation equilibria.[4] But like most models with multiple equilibria, the basic tipping model suggests no mechanism for moving between an all-white equilibrium and a ghetto equilibrium. The theory can therefore explain the persistence but not the formation of ghettos. Perhaps surprisingly, these limitations can be overcome by allowing for a richer (and more natural) geometry of interaction between residents where agents care more about neighbors who are geographically close to their apartment than about residents further away.[5]

My model analyzes the location decision of two ethnic groups, which I refer to as 'blacks' and 'whites' for convenience. The model is described by four basic parameters: the tolerance levels of both groups α_b and α_w, their balance λ in the housing market and the geometry G of neighborhood interaction. The geometries I mainly consider are one-dimensional *streets* and two-dimensional *inner-cities* where residents only care about their direct neighbors. For completeness I also look at *bounded neighborhoods* where each resident is neighbor to every other resident in the area such that my model reduces to a variant of Schelling's tipping model.

Residents leave the residential area randomly at a fixed rate and are replaced by newcomers from the housing market where a share λ of apartment seekers are white. Most newcomers, however, exhibit mild ethnic preferences and only consider an apartment if

[3] Cutler, Glaeser, and Vigdor (1997) documented that the correlation across cities between segregation in 1890 and segregation in 1990 is as high as 50 percent. Residential segregation affected all ethnic minorities in the US to varying degrees. For African Americans, however, it is unique in its severity and persistence over several generations. Second- and third-generation non-black immigrants generally lived in much less segregated neighborhoods than their parents. For a good reference see chapter three in Taeuber and Taeuber (1965).

[4] For example, Galster (1990) and Cutler, Glaeser, and Vigdor (1997) use variants of the tipping model for their empirical studies.

[5] Incidentally Schelling himself sketched a two-dimensional model of segregation in his book *Micromotives and Macrobehavior* (1978).

they do not feel 'isolated', i.e. at most a share $\frac{1}{2} < \alpha < 1$ of their prospective neighbors is of a different ethnicity. A small share ϵ of newcomers are completely tolerant and do not care about the ethnicity of their neighbors. These non-discriminating residents provide the 'noise' which is necessary to move the process from one equilibrium to the next.

On streets the relative strength of ethnic preferences alone cannot give rise to segregation. Once fear of isolation interacts with the balance in the housing market, however, streets can rapidly turn into ghettos. A sudden rise in the share of blacks in the market can transform the residential area because blacks can 'invade' the street around small clusters of black residents which have formed on the street by chance. The 'contagious' dynamics of this transformation is reminiscent of the model studied by Ellison (1993), who demonstrated how local interaction can speed up convergence to the long-run equilibrium. My model highlights the importance of the increase in the share of African Americans in the housing market for the rise of black ghettos. This situation occurred at the beginning of the 20th century when African Americans started to migrate from the rural South to the booming industrial centers of the North.

Segregation on streets, however, is not persistent because the forces that give rise to ghettos are symmetric: as soon as whites dominate the housing market again, a ghetto can equally rapidly disintegrate. Temporary imbalances in the housing market can therefore give rise to periodic transformations of ethnic neighborhoods or "avenue waves". In the US there existed a natural source of variation in the housing market balance for non-black minorities at the turn of the century because large waves of immigrants from certain ethnic groups, such as Southern Europeans, Scandinavians or Russians, entered the country at different points in time. In section 2.5 I present evidence from Chicago which documents the occurrence of "avenue waves" along the main arteries as predicted by my model.

Ghettos can exhibit pronounced persistence, however, once the process evolves in a two-dimensional inner-city area. As long as blacks are sufficiently more tolerant than whites the process will again give rise to rapid segregation as soon as blacks dominate the housing market sufficiently. The resulting ghettos, however, will not disintegrate after subsequent changes in the balance of the housing market. The reason for this stasis is the "encircling phenomenon": in two dimensions, randomly forming clusters of white residents cannot expand as easily as on streets because white residents at the convex boundary of the cluster have on average more black than white neighbors. This prevents the kind of contagious

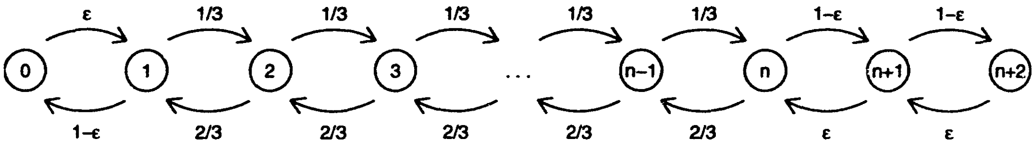


Figure 2-1: Illustration of the dynamics of a continuous-time random walk over the set of states $\{0, 1, .., n+2\}$

dynamics through which white clusters on streets can break up a black ghetto. Inner-city areas therefore behave like hybrids: they can be transformed into ghettos (like streets) which are subsequently stable (like the all-black equilibrium of Schelling's tipping model).

For the analysis of the medium and long-run behavior of my model I develop a new technique to characterize the stochastic evolution of large-scale dynamical systems, which I believe can be fruitfully applied to both existing and future models. The standard technique for understanding the long-run behavior of stochastic dynamics in evolutionary game theory has been stochastic stability analysis, which was introduced in the seminal woık by Young (1993) and Kandori, Mailath, and Rob (1993) and recently extended by Ellison (1999). The model of this paper, however, illustrates that stochastic stability can seriously mispredict the behavior of a process with many agents. Incidentally, these are exactly the environments where evolutionary reasoning seems most adequate.

The shortcomings of stochastic stability can be most easily explained with the help of a simple example. Figure 2-1 illustrates the dynamics of a random walk on the integers $\{0, 1, .., n+2\}$ in continuous time. Between states 1 and n the process evolves according to the 'undisturbed' dynamics, while between states 0 and 1 and between states n and $n+2$ the process is governed by 'noisy' dynamics. For any fixed size n only the state $n+2$ is stochastically stable in the sense that the process spends almost all its time at $n+2$ as the noise term ϵ becomes small. To see the intuition for this result, note that it takes two mutations to leave the basin of attraction of state $n+2$, but just one mutation to leave the basin of attraction of state 0.[6] This reasoning, however, does not take into account the nature of the undisturbed dynamics at all, even through it pushes the process away from the stochastically stable state.

[6] The basin of attraction is defined with respect to the undisturbed dynamics, i.e. $\epsilon = 0$.

Stochastic stability can thus lead us to mispredict the long-run behavior of the process. Assume that we want to find the process inside a small δ-neighborhood $[(1-\delta)\,n, n+2]$ of the stochastically stable state with probability $\gamma > 0$. It can be shown that the noise term ϵ then has to be smaller than $\left(\frac{3}{2}\right)^{-n}$. But as the size of the system increases stochastic stability will capture the dynamics adequately only for extremely small noise. Even more worrisome, the waiting time to reach the stable state $n+2$ becomes unrealistically large for such small ϵ.

Stochastic stability analysis therefore tends to work best for small-scale stochastic systems, while most evolutionary environments, such as residential neighborhoods, involve many interacting agents. For the preceding example one can in fact demonstrate that for a fixed noise term ϵ the process will spend almost all its time close to the state 0 as the size n of the system increases. The long-run evolution of the system is completely determined by the biased undisturbed dynamics of the process.

This insight immediately suggests an alternative technique to characterize stochastic dynamics. *Clustering* looks at the dynamics of a system as its size n increases and therefore takes into account both the disturbed and the undisturbed dynamics of a model. This makes clustering a more robust equilibrium concept than stochastic stability for systems with many agents.

The balance of the chapter is organized as follows. In the next section I lay out a general model for a residential segregation process and introduce the notion of clustering. The new technique is then applied to streets in section 2.3 and to inner-cities in section 2.4, in order to characterize the long-run equilibria of the model and to find bounds on the waiting time to reach those equilibria. In section 2.5 I discuss how my theory can help us to understand the formation of ghettos at the beginning of the 20th century in the US. I also present evidence of "avenue waves" from Chicago and describe in detail Harlem's transformation from a white upper-class neighborhood into a black ghetto between 1900 and 1930. The relationship between stochastic stability and clustering is explored in section 2.6 using the waiting time terminology introduced by Ellison (1999). In order to demonstrate the usefulness of clustering as a general technique I revisit a well-known application of stochastic stability by Ellison (1993) and illustrate how clustering can make the predictions of Ellison's paper robust to changes in the dynamics.

2.2 A Framework for Analyzing Segregation

This section introduces a simple evolutionary model of segregation and the notion of clustering which will be used to analyze the medium and long-run behavior of the resulting Markov process. In the case of bounded neighborhoods my model reduces to Schelling's (1972) tipping model, and I demonstrate why this setup describes the dynamics of segregation insufficiently. Although the tipping model allows for both an all-white and an all-black ghetto equilibrium, the process remains locked into basins of attraction around those equilibria. The medium-run dynamics is solely determined by the initial conditions, and there is no mechanism which gives rise to ghettos within a realistic time frame.

2.2.1 The Basic Setup

A residential area of size n consists of n residents $R = \{z_1, z_2, ..., z_n\}$. They form the vertices of a connected graph G defined through a symmetric neighborhood relation $G \subset R \times R$.[7] Each resident z has a natural neighborhood $N(z) = \{z' \,|(z, z') \in G\}$. I restrict attention to three possible residential geometries. *Bounded neighborhoods* $G_B(n)$ have a complete graph G such that individual neighborhoods coincide with the entire residential area. There are also two local geometries with easy intuitive representations: one-dimensional *streets* $G_S(n)$ and two-dimensional *inner-city areas* $G_C(n)$. On a street, residents are located on a circle with each agent having two neighbors on both sides. An inner-city area consists of a torus of size $\sqrt{n} \times \sqrt{n}$ such that each resident has four neighbors. Streets have the lowest possible connectivity of a regular connected graph, while bounded neighborhoods have the highest. Inner-cities take an intermediate position.[8] Both streets and inner-cities have intuitively related graphs of higher order if we allow for individual neighborhoods of radius $r > 1$ in the standard Euclidean norm. These geometries are denoted with $G_S^r(n)$ and $G_C^r(n)$ respectively. Figure 2-2 shows both a street and an inner-city with respective

[7] The graph is connected if any two residents are connected through a transitive chain of neighborhood relationships.

[8] The *connectivity* $C(G)$ of a finite graph G is defined as the lowest upper bound for the minimum length path connecting any two residents on the graph, i.e. $C(G) = \max_{z_i, z_j \in R} d(z_i, z_j)$ where $d(z_i, z_j)$ denotes the length of the minimum length path connecting z_i and z_j. The smaller $C(G)$ the better connected the graph. Bounded neighborhoods have $C(G_B) = 1$, a street $G_S(n)$ has connectivity $\left[\frac{n+1}{2}\right]$ and the inner city $G_C(n)$ has an intermediate connectivity of $2\left[\frac{\sqrt{n}}{2}\right]$.

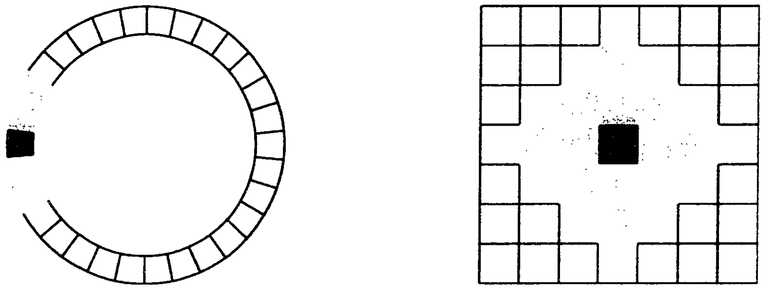


Figure 2-2: Street and inner-city geometries with individual neighborhoods of radius 2

neighborhoods of radius 2.[9] Note that for a large radius r the individual neighborhood of a resident comprises the entire residential area, and one again obtains the bounded neighborhood geometry.

The residential area is populated by two ethnic groups whom I refer to as 'blacks' and 'whites' throughout the paper. At each point in time the pattern of settlement is defined by a configuration $\eta : G_i \to \{0, 1\}$, where the values 0 and 1 denote a white and black resident respectively.[10] The residential area allows a total of 2^n configurations which form the configuration set Z. A cluster in some configuration η is defined as a connected set of residents of the same ethnic group.[11]

All residents in the area are assumed to have an area-specific socio-economic status. Time is continuous, and agents get 'lucky' according to an i.i.d. Poisson process at rate 1. 'Lucky' agents immediately become a member of the next highest socio-economic group and move out of the area because they can afford better housing. Vacant flats are occupied by newcomers from a large pool of prospective tenants. A share λ of those is white and a share $1 - \lambda$ is black.

A (small) share ϵ of prospective tenants are completely tolerant in the sense that they do not care about the ethnic composition of their individual neighborhoods. The remaining share $1 - \epsilon$ of prospective tenants have mild ethnic preferences because they are afraid of

[9] Looking at a circle and a torus respectively avoids the need to specify boundary conditions. All results hold for open linear streets and rectangular inner-cities, as well, once the decision rules for residents at the boundary are suitably adapted.

[10] I generally treat a lattice cell as a single resident. For densely populated cities such as New York City, however, it might be more adequate to interpret lattice cells as entire apartment blocks, as the owners of these buildings usually did not mix tenants of different racial groups.

[11] Connectedness is defined with respect to simple streets and inner-city geometries with radius of interaction $r = 1$.

isolation at their new apartments. All whites and all blacks have identical, group-specific tolerance levels α_w and α_b respectively. The tolerance level marks the maximum share of neighbors of a different ethnicity a prospective tenant is prepared to accept. I assume that $\frac{1}{2} \leq \alpha_i < 1$, i.e. agents are generally happy to live in integrated areas where both ethnic groups share the neighborhood equally.[12] I assume throughout the paper that the minority group ('blacks') is more tolerant than the majority group ('whites'), i.e. $\alpha_w \leq \alpha_b$.[13]

The housing market operates as follows. All prospective tenants have a basic willingness to pay WTP, which depends only on their socio-economic status and is equal for both whites and blacks. If a resident feels isolated, however, her willingness to pay decreases to $WTP - D$ for some $D > 0$. An apartment is then allocated amongst the highest bidders through randomization. One can derive a switching function which denotes the probability that the color of a tenant switches conditional on the previous tenant having moved out. I denote the share of black neighbors of a resident z in configuration η with $x(\eta, z)$ and the share of white neighbors with $y(\eta, z)$. The probability of a color switch g_w if the previous tenant was white then becomes:

$$g_w^\epsilon(x(\eta,z)) = \begin{cases} \frac{(1-\lambda)\epsilon}{(1-\lambda)\epsilon+\lambda} & \text{for } x < 1-\alpha_b \\ 1-\lambda & \text{for } 1-\alpha_b \leq x \leq \alpha_w \\ \frac{1-\lambda}{1-\lambda+\epsilon\lambda} & \text{for } x > \alpha_w \end{cases} \tag{2.1}$$

Analogously, the probability g_b for a switch from a black to a white tenant is:

$$g_b^\epsilon(y(\eta,z)) = \begin{cases} \frac{\lambda\epsilon}{\lambda\epsilon+1-\lambda} & \text{for } y < 1-\alpha_w \\ \lambda & \text{for } 1-\alpha_w \leq y \leq \alpha_b \\ \frac{\lambda}{\lambda+\epsilon(1-\lambda)} & \text{for } y > \alpha_b \end{cases} \tag{2.2}$$

Figure 2-3 illustrates the typical shape of the resulting switching functions in terms of the share of neighbors of the opposite color.

The evolution of the residential neighborhood can now be described by a continuous time Markov chain η_t on the space of configurations Z where η_0 is the initial configuration which is set by some historical accident.

[12] If $\alpha_i < \frac{1}{2}$ for both groups, segregation would be the socially efficient outcome.

[13] Empirical studies such as the General Society Survey reveal that whites discriminate more strongly than blacks (Cutler, Glaeser, and Vigdor 1997).

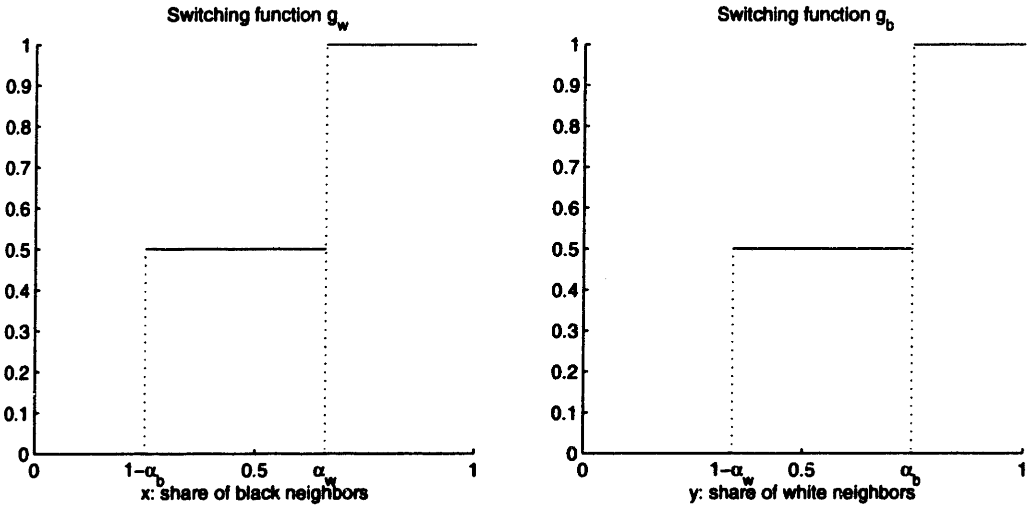


Figure 2-3: Switching functions for white/ black transition (g_w) and black/ white transition (g_b) for tolerance levels $\alpha_w = \frac{2}{3}$, $\alpha_b = \frac{3}{4}$, share of completely tolerant agents at $\epsilon = 0$ and the share of whites in society at $\lambda = 0.5$

Remarks on the Setup: 1. Which of the two local geometries approximates real-life residential areas best? It is natural to think of geographic entities, such as residential neighborhoods, in a two-dimensional setting. On the other hand, 'streets' might capture the neighborhood interaction on large avenues more appropriately.[14]

2. Prospective tenants in my model behave only in a boundedly rational manner and just take the contemporaneous ethnic balance of a neighborhood into account. Forward-looking rational agents with a positive discount factor should anticipate the probabilistic evolution of a residential area and possibly take into account additional available information such as the total ethnic balance of the area they move into. The computational requirements on prospective tenants become enormous, however, even for moderately large residential areas. I therefore adopt the myopia hypothesis which is commonly employed in evolutionary game theory (see, for example, Kandori, Mailath, and Rob (1993) and Young (1993)) and allows me to concentrate my analysis on the dynamics of segregation.

3. Fear of isolation gives rise to an S-shaped frequency distribution of tolerance levels within each ethnic group. The empirical evidence suggests that the tolerance distribution

[14] Residents on such major roads certainly had some preferences concerning the racial composition of side streets, but they presumably put greater weight on the residents living along the avenue: a majority of shops, public transport and institutions such as churches would be located along the avenues, making social interaction with residents there more likely.

is indeed highly non-linear and S-shaped (Galster 1990). The results of this paper carry through for more general tolerance distributions and richer models of the housing market as long as the tails of the reduced form switching functions are flat, i.e. whites (blacks) will mainly seek out white (black) neighborhoods.

4. Discrimination in this model operates only through destination selectivity of prospective tenants in the housing market. Schelling's (1972) original tipping model also allows agents to leave a neighborhood at an increased rate if they feel isolated.[15] This second channel can be easily incorporated in my model without changing the qualitative predictions. Destination selectivity, however, seems to be the more important channel, as moving costs are presumably higher than the search cost which is associated with excluding some apartments from further consideration. Furthermore, my main application of the model concerns the formation of ghettos at the turn of the century when city growth was rapid and the residential turnover rate was high.

5. The model can be regarded as a partial equilibrium building block for a richer general equilibrium model of a growing city consisting of many residential areas of different socioeconomic status. Historically, the frantic expansion of Northern cities in the US at the turn of the century was accompanied by a chain of succession and invasion. As wealthy middle class citizens in New York or Chicago gradually abandoned the city for the new suburbs, they were replaced by successful immigrants who had left behind their lower class origins. Their place, in return, was occupied by new immigrants and migrants from rural areas.

2.2.2 Characterizing the Stochastic Dynamics through Clustering

It is easy to see that the model has a unique ergodic distribution μ^{ϵ}_{∞} over the set of configurations Z which describes the long-run behavior of the system.[16] Therefore the long-run behavior of the system is independent of the initial conditions. But this observation is of little interest unless we find a way to classify the ergodic distribution. Will the process spend most of its time around segregation configurations or around mixed configurations? How does the equilibrium depend on parameters of the model, i.e. the geometry, the tolerance

[15] In his sketch of a local tipping model Schelling (1978) omits destination selectivity altogether.

[16] Appendix A.1 shows how to associate a Markov chain with the continuous time Markov process. The transition matrix P^{ϵ} of that chain is regular as $(P^{\epsilon})^n$ has no non-zero entries - each configuration of the geometry can be reached after n steps with positive probability (Kemeny and Snell 1960, Theorem 4.1.2). Therefore the process is ergodic (Kemeny and Snell 1960, Theorem 4.1.4).

levels of both groups and the balance in the housing market?

The standard technique for classifying the ergodic distribution μ_{∞}^{ϵ} is stochastic stability analysis, which was developed by Young (1993) and Kandori, Mailath, and Rob (1993). All results from this literature can be applied to my model if the small share ϵ of completely tolerant residents is interpreted as 'noise'. A configuration η is then called *stochastically stable* if $\lim_{\epsilon \to 0} \mu_{\infty}^{\epsilon} > 0$ for some fixed geometry $G_i(n)$ $(i = B, S, C)$.

I will demonstrate in section 2.6 that stochastic stability explains the long-run dynamics of the residential neighborhood process very poorly for large-scale residential neighborhoods. The intuition for this failure will be the same as for the example I gave in the introduction. Stochastic stability describes the process well only for extremely small ϵ. This requirement is troubling because the share of tolerant residents might be low but is certainly not negligible. Even more worrisome is the effect that such a small 'noise' parameter has on the waiting time before the stochastically stable configuration is reached for the first time. For large neighborhoods, convergence will be so unrealistically slow that the analysis will tell us nothing about the medium-run behavior of the process.

This insight quite naturally suggests an alternative to stochastic stability which fixes the noise term ϵ and instead considers very large residential areas, i.e. lets $n \to \infty$.[17] As the main parameter of interest is the ethnic balance in the residential area, I formally define the concept of *clustering* for the long-run share $\tilde{X}_n^{\epsilon}$ of black residents which is a scalar random variable.

Definition 1 *A sequence of random variables* $\left\{\tilde{X}_n\right\}$ *on the interval* $[0,1]$ *is said to cluster over the set* $I \subset [0,1]$ *if* $P\left(\tilde{X}_n \in I\right) \to 1$ *as* $n \to \infty$.

For example, we can interpret clustering of the residential neighborhood process on a street $G_S(n)$ around a black share close to 1 in the sense that large streets will become black ghettos in the long run.

Although clustering captures the long-run behavior of the process well, it does not tell us how fast a neighborhood I is reached over which the process clusters. Waiting times are a very useful measure for the speed of convergence to equilibrium, as was first emphasized by Ellison (1993) in the context of stochastic stability. With respect to clustering, the relevant measure is the maximum waiting time $W(n, I)$ in which the process reaches I for the first

[17] The type of geometry is assumed to be fixed when taking the limit.

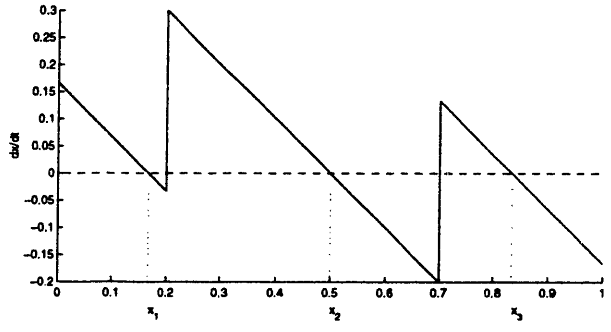


Figure 2-4: Graph of $\frac{dx}{dt}$ in the deterministic approximation to the residential neighborhood process on $G_B(n)$ for large n ($\alpha_b = 0.8$, $\alpha_w = 0.7$, $\lambda = 0.5$, $\epsilon = 0.2$)

time starting from any initial configuration:

$$W(n,I) = \max_{\zeta \in Z}\left[E\left(\min t \,|X(\eta_t) \in I \quad \text{and} \quad \eta_0 = \zeta\right)\right] \tag{2.3}$$

$X(\eta)$ here denotes the share of black residents in configuration η. Unless that waiting time remains bounded as n increases, the evolution of the process will be determined by the initial conditions rather than the long-run equilibrium.

2.2.3 Schelling's Tipping Model as a Benchmark

It is instructive to start the analysis of the residential neighborhood process for bounded neighborhoods $G_B(n)$ because the model becomes a variant of Schelling's (1972) well-known tipping model. The entire intuition for the behavior of the process on large bounded neighborhoods can be derived from the deterministic approximation of the change in the share of black residents $x(t)$:

$$\frac{dx}{dt} = (1-x)\, g_w^{\epsilon}(x) - x g_b^{\epsilon}(1-x) \tag{2.4}$$

Figure 2-4 shows the graph of $\frac{dx}{dt}$ and illustrates that the deterministic approximation of the process has, in general, multiple stable steady state equilibria: two segregation equilibria $x_1 = \frac{(1-\lambda)\epsilon}{(1-\lambda)\epsilon+\lambda}$ and $x_3 = \frac{1-\lambda}{1-\lambda+\lambda\epsilon}$, and possibly one integrated equilibrium $x_2 = 1-\lambda$.[18] The

[18] I assume that the share of completely tolerant residents ϵ is sufficiently small such that $x_1 < 1-\alpha_b$ and $x_3 > \alpha_w$. The integrated equilibrium might not exist if the housing market is sufficiently unbalanced, i.e. $\lambda > \alpha_b$ or $\lambda < 1-\alpha_w$.

corresponding basins of attraction are $B_1 = [0, 1 - \alpha_b)$,$B_2 = (1 - \alpha_b, \alpha_w)$, and $B_3 = (\alpha_w, 1]$, respectively.

The evolution of the deterministic approximation to the process is therefore entirely determined by the initial conditions, which is a highly unsatisfactory feature of bounded neighborhoods. The choice of the initial share of black residents x_0 is indeterminate without making arbitrary assumptions about the history of the process. While the model does well in explaining the persistence of ghettos, it suggests no mechanism for moving between the steady states.

The intuition which we gained from the deterministic approximation continues to hold for the stochastic model, as the next theorem shows. The stochastic drift will select one of the steady states in the long run, depending on the parameter values. For simplicity I restrict attention to the most interesting case where blacks dominate the housing market. In this case the residential area will turn into a black ghetto, and the process clusters around x_3 unless there are too many completely tolerant agents in the housing market such that the process clusters around the integrated steady state $x_2 = 1 - \lambda$. However, this characterization of the long-run behavior of the process is meaningless for its medium-run evolution. The process is tightly 'locked' in the basins of attraction around the three steady states, as the second part of the theorem shows. For all practical purposes the behavior of the process is indeed determined by the initial conditions. Therefore, the tipping model can not explain the formation of ghettos.

Theorem 1 *Consider a residential neighborhood process on $G_B(n)$ with initial share of black residents x_0 in one the three basins of attraction B_i $(i = 1, 2, 3)$. Blacks dominate the housing market, i.e. $\lambda < \frac{1}{2}$.*

1. *(Long-run behavior) The process clusters around any neighborhood of the ghetto steady state x_3 if $\frac{\epsilon^{1-\alpha_w}}{1-\lambda+\lambda\epsilon} < 1$, i.e. the share of completely tolerant apartment seekers is sufficiently small (and x_2 exists). Otherwise, the process clusters around the integrated steady state $x_2 = 1 - \lambda$.*

2. *(Medium-run behavior) The process reaches a δ-neighborhood of the steady state x_i before it can leave the basin of attraction with probability approaching* 1 *as $n \to \infty$. The conditional waiting time for this event is bounded above by some finite W_δ. Moreover, the waiting time to reach a neighborhood of the steady state chosen in the long-run*

Table 2.1: Comparison of waiting times $W(n, I)$ for reaching the neighborhood $I = [0.97, 1]$ of the ghetto steady state $x_3 = 0.99$ ($\alpha_b = \frac{3}{4}$, $\alpha_w = \frac{2}{3}$, $\epsilon = 0.05$, $\lambda = 0.2$)

Size of area	n=100	n=200	n=300	n=400
Waiting time $W(n, I)$	21	137	627	4908

Estimated standard errors for the waiting times are 10% or less.

is of the order $A(x_0)^n$ where $A(x_0) > 1$ if the process starts outside the basin of attraction of that steady state.[19]

Proof: see appendix A.3

Example: A little numerical example illustrates the irrelevance of the long-run equilibrium for the medium-run behavior of the process. I consider the case where the black and white tolerance levels are $\alpha_b = \frac{3}{4}$ and $\alpha_w = \frac{2}{3}$ respectively, and 5 percent of all agents are completely tolerant ($\epsilon = 0.05$). I assume that the neighborhood has initially been in an all-white steady state ($\lambda = 1$) when an influx of blacks into the housing market occurs, causing the share of whites in the market to fall to 20 percent. The all white steady state and the all black steady states are $x_1 = 0.17$ and $x_3 = 0.99$ respectively; there is no integrated steady state. Theorem 1 tells us that under these circumstances the process clusters around any neighborhood I of the ghetto steady state, say $I = [0.97, 1]$. How long will it take until the bounded neighborhood has turned into a ghetto? Table 2.1 shows the results from a simulation for neighborhoods of various sizes.[20] The data nicely confirm the theory, as the waiting times increase rapidly with the size n of the bounded neighborhood. In the medium run, the behavior of the process on even moderately large residential areas is therefore entirely determined by the fact that the process started from an all-white configuration. As a response to the dominance of blacks in the housing market their share in the area will increase rapidly to about 17 percent, i.e. the steady state value x_1. The process will then oscillate around this *meta equilibrium*, but is unlikely to escape its basin of attraction within any realistic time frame.

[19] *I assume that $1 - \alpha_w < \lambda < \alpha_b$ such that a bounded neighborhood can stay integrated around x_2 in the medium run. If $\lambda < 1 - \alpha_w$ the process will reach its long-run equilibrium in finite time for $x_0 > 1 - \alpha_b$. If $\lambda > \alpha_b$ the process will reach its long-run equilibrium in finite time only for $x_0 > \alpha_w$.*

[20] Note that the maximum waiting time $W(n, I)$ coincides in this case with the waiting time of reaching I starting from the all-white configuration.

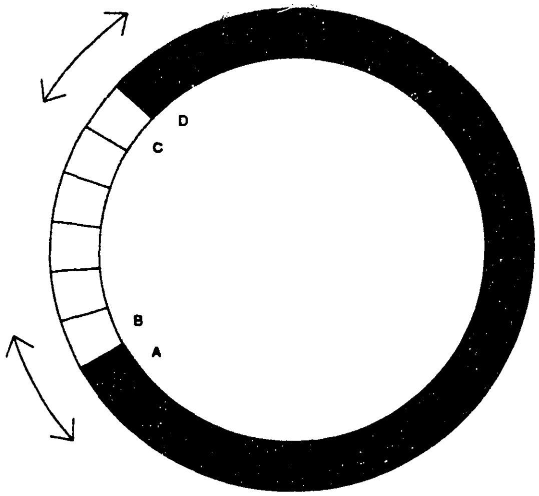


Figure 2-5: Black ghetto on a street $G_S^{r=1}(n)$ with a single white cluster: if residents A to D move, members of both racial groups are equally interested in vacant apartments.

2.3 Rapid Segregation and "Avenue Waves" on Streets

The previous section demonstrated that the original tipping model exhibits stasis around its steady states. It is noteworthy that the persistence of the all-white and all-black segregation steady states is preference based: the apartment seekers of the minority group feel isolated and avoid the residential area. In particular, an increased presence of blacks in the housing market does not trigger the transformation of the area into a ghetto in the medium run.

On streets, on the other hand, the residential neighborhood process behaves in a radically different manner because it gives a role to the balance in the housing market. Black (white) segregation on streets is upheld in the long run because the share of blacks (whites) in the housing market exceeds a critical level. Moreover, this mechanism lets the process reach its long-run equilibrium rapidly. In contrast to the standard tipping model, streets do not behave differently in the medium run and in the long run. Streets, therefore, provide a mechanism for moving between all-white and all-black equilibria through changes in the composition of the housing market.

I begin my analysis with a heuristic argument in order to illustrate why streets become ghettos in the long run if blacks sufficiently dominate the housing market. The intuition is cleanest for the case where the tolerance levels are close to $\frac{1}{2}$. For simplicity I only look at simple streets $G_S^{r=1}(n)$, although the argument is readily generalized to higher-order streets. A ghetto on such a street will occasionally face invasion by small white clusters of completely tolerant apartment seekers, as illustrated in figure 2-5. Under the assumption that the share ϵ of tolerant agents is small, vacant apartments *inside* of black and white clusters are almost always taken only by black and white residents respectively due to the assumption on the tolerance levels. However, if apartments at the boundary of the cluster become vacant (such as A or D) apartment seekers from both ethnic groups will be interested in them. The boundaries of the black cluster therefore move according to a random walk with absorption (the process ends if one of the clusters vanishes). The drift of this random walk is solely determined by the composition of the housing market, i.e. λ. As blacks dominate the housing market the white cluster is likely to shrink rather than grow.

We can now ask the question, what would happen to this cluster if the circle was infinite and it could not interact with other random white clusters? Standard theory tells us that the cluster would die out with probability 1, and its expected maximum length would be finite and determined by the negative drift only.[21] From this observation we can conclude that white clusters form and die independently from one another to a first approximation, as long as the share ϵ of completely tolerant agents is small and the size n of the street is large. Therefore, the equilibrium share of white residents in the ghetto can be derived by calculating the average length of white clusters which originate from some fixed apartment z on the street.[22] The length of such a cluster forms a random walk with transitions rates as indicated in figure 2-6. The probability x that the originating apartment is black then becomes:[23]

$$x = \left(1 + \epsilon \frac{\lambda\left(2\lambda + 1\right)}{1 - \lambda} + \epsilon \frac{2\lambda^2}{\left(1 - 2\lambda\right)\left(1 - \lambda\right)}\right)^{-1} \tag{2.5}$$

From formula 2.5 one can immediately deduce that if the share of blacks in the housing market is greater than 50 percent and the share of completely tolerant agents is small the street will become a ghetto.

The next theorem makes this heuristic argument precise and generalizes it for the case where the tolerance levels are not necessarily close to $\frac{1}{2}$. If the share of whites in the market falls below some critical level $\hat{\lambda}$, the residential neighborhood process on a street will turn

[21] These are standard results from random walk theory (Stirzaker 1994, section 5.6).

[22] I invoke the law of large numbers here.

[23] The expression for x is a first order approximation in ϵ.

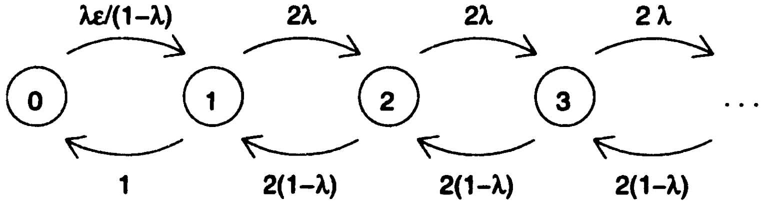


Figure 2-6: Transition rates for the length of a white cluster originating at a single apartment z on the street $G_S^{r=1}(n)$

into a ghetto in the long run and cluster around a black share of $x = 1$. On the other hand, if the share of whites exceeds some critical level $\tilde{\lambda}$ the process clusters around the all-white equilibrium.

Theorem 2 *Given is a street $G_S^r(n)$ with group tolerance levels α_w and α_b. Then there exist critical values $0 < \hat{\lambda}(r, \alpha_w, \alpha_b) \leq \tilde{\lambda}(r, \alpha_w, \alpha_b) < 1$ such that the following holds.*

1. *If blacks sufficiently dominate the housing market ($\lambda < \hat{\lambda}$) the street becomes a black ghetto in the long run, i.e. the process clusters on the interval $[x_b^*(\epsilon), 1]$ with $\lim_{\epsilon \to 0} x_b^*(\epsilon) = 1$.*

2. *If whites sufficiently dominate the housing market ($\lambda > \tilde{\lambda}$) the street becomes a white ghetto in the long run, i.e. the process clusters on the interval $[0, x_w^*(\epsilon)]$ with $\lim_{\epsilon \to 0} x_w^*(\epsilon) = 0$.*

Proof: see section 2.3.2

The description of the long-run equilibrium of the process is, of course, only relevant if it is reached reasonably quickly. Fortunately, this is the case on streets, as the next lemma shows. The intuition can be again derived for a simple street $G_S^{r=1}(n)$ where blacks dominate the housing market. I have demonstrated how white clusters tend to shrink rather than grow in such an environment. The argument can be flipped around in the sense that black clusters have to grow rather than shrink. Assume that the street is initially all-white. It is useful to divide the street up into k segments of some fixed size N. On each segment black clusters form after a waiting time of about $\frac{B}{N\epsilon}$. Such clusters can subsequently expand with positive drift. The problem is complicated by the fact that black clusters can be broken up by randomly forming white clusters. Ignoring this issue for the moment the black cluster would take over the neighborhood after some waiting time of the order AN (see, for example, lemma 6 in appendix A.2). The total waiting time until a segment becomes a ghetto is therefore of the order $\frac{B}{N\epsilon} + AN$. If the segments are large enough, each of them stays mostly black subsequently due to theorem 2. As $n \to \infty$ (i.e. $k \to \infty$) one can invoke a form of the central limit theorem in order to show that the waiting time $W(n, I)$ to reach some neighborhood I of the black ghetto equilibrium is of the order $O(1)$.

Lemma 1 *Consider a street $G^T_S(n)$ and assume that the share of whites in the housing market falls below the critical level $\hat{\lambda}$. If ϵ is sufficiently small the waiting time until the share of blacks reaches exceeds $1 - \delta$ satisfies $W(n, [1 - \delta, 1]) = O(1)$. An analogous result holds if the share of whites exceeds the critical level $\tilde{\lambda}$.*

Proof: see appendix A.6

It should be pointed out that the residential neighborhood process responds fast to changes in the composition of the housing market in *both* directions. While a street can quickly turn into a black ghetto, this development can reverse just as rapidly as soon as whites dominate the housing market again. In some sense, streets do too well in explaining the formation of black ghettos: if the composition of the housing market is highly volatile we should observe "avenue waves" instead of highly persistent ghettos. In section 2.5 I provide evidence for such waves in the case of Chicago. To summarize, while Schelling's tipping model lacks a mechanism for ghetto formation but does well in terms of ghetto persistence the reverse is true for streets. One, therefore, would like a 'hybrid' geometry between streets and bounded neighborhoods which can explain both phenomena. I argue in the next section that inner-cities provide such an environment.

The remainder of this section is devoted to some Monte Carlo simulations, which illustrate the fast response of streets to changes in the housing market, and to the proof of theorem 2. I discuss the steps of the proof in some detail because it introduces the *coupling* technique, a highly useful device for understanding Markov processes on a lattice.

2.3.1 Simulation Results on Critical Behavior and Speed of Adjustment

Theorem 2 and lemma 1 characterize the long-run and medium-run behavior of the residential neighborhood process on streets qualitatively. They do not permit us to numerically calculate the critical imbalance in the housing market which gives rise to ghettos, or the waiting time until convergence.[24] Simulations are therefore essential to assess the relevance of the theory. First, are the critical imbalances realistic, i.e. sufficiently bounded away from $\{0,1\}$? Second, what does fast convergence in the medium run mean? Time is measured in my model in terms of tenant generations. Waiting times $W(n,I)$ of the order of 100, for example, translate into centuries when measured in real time.

The first round of simulations is aimed at finding the critical imbalance in the housing market such that a street becomes a ghetto in the long run. Figure 2-7 shows the long-run share of black residents depending on the share λ of whites in the housing market for streets with radius of interaction $r = 2$ and $r = 3$. The critical imbalance of the housing market is encouragingly close to 50 percent when both groups are equally tolerant such that $\hat{\lambda} = 1 - \tilde{\lambda}$. For tolerance levels close to $\frac{1}{2}$ the street turns into a black ghetto if blacks control more than 50 percent of the market as expected. But even if both groups can tolerate having up to 75 percent (80 percent) of their neighbors be of a different ethnicity, the share of blacks only has to exceed 70 percent (80 percent) for a black ghetto to arise. If blacks are strictly more tolerant than whites the critical values shift accordingly. A street with radius 2, for example, and tolerance levels of $\alpha_w = \frac{1}{2}$ and $\alpha_b = \frac{3}{4}$ will turn into a black ghetto as soon as the share of blacks exceeds 30 percent. Nevertheless, even though blacks are far more tolerant than whites, a black ghetto will dissolve again as soon as at least 80 percent of all apartment seekers are white.

Next, I look at the medium-run evolution of the residential neighborhood process in a setup where 5 percent of agents are completely tolerant, blacks constitute 80 percent of the housing market and the radius of interaction is again $r = 2$ or $r = 3$.[25] For the simulations in figure 2-8 I assume that both groups have equal tolerance levels which can only slow down the transformation of the area into a black ghetto, as compared to the case where whites

[24] An exception is the case where the tolerance levels of both ethnic groups are close to $\frac{1}{2}$. The heuristic argument of the previous section established that $\hat{\lambda} = \tilde{\lambda} = \frac{1}{2}$.

[25] The share $\lambda = 0.2$ has been chosen such that $\lambda < \hat{\lambda}$ in all cases.

Table 2.2: Comparison of waiting times $W(n,[0.9,1])$ until more than 90 percent of residents are black starting from an all white neighborhood on a street $G_S^{r=2}(n)$ with $\lambda = 0.2$, $\alpha_b = \frac{3}{4}$, $\alpha_w = \frac{2}{3}$.

Expected Wait $W(n,[0.9,1])$

	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.05$
n=100	9.32	7.16	5.84	4.86
n=1000	10.30	7.64	5.98	4.98
n=10000	10.02	7.84	6.00	5.00
n=20000	10.00	7.90	6.00	5.00

Estimated standard errors for the waiting times are 5% or less.

are strictly less tolerant than blacks.[26] The speed in which the all-white area is transformed into a black ghetto is impressive. It takes less than 5 tenant generations until more than 75 percent of the street has become black if blacks can tolerate having whites be two thirds or more of their neighbors. If blacks are less tolerant they will still populate about 50 percent of the area after 5 tenant generations.

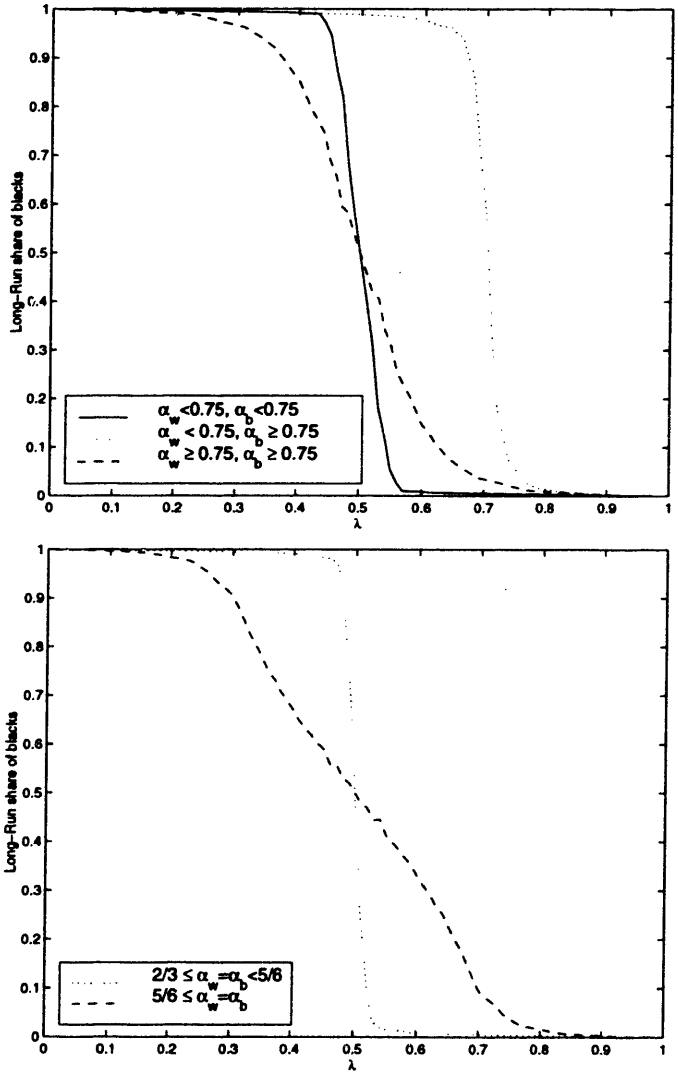
How does the speed of convergence depend on the share ϵ of tolerant agents and the size n of the street? Table 2.2 lists the expected waiting times $W(n,[0.9,1])$ until 90 percent of all residents on the street are black for the geometry $G_S^{r=2}(n)$ and ϵ varying between 0.5 percent and 5 percent. The results indicate that the waiting time until convergence does not increase with the size of the street and depends only weakly on the share of tolerant agents.[27] Because of lemma 1, one would expect the waiting time not to depend on the size of street for large n. The simulations demonstrate, however, that the lemma holds even on fairly small streets.

Example (continued from section 2.2.3): By repeating the numerical exercise for streets one can directly compare the different behavior of the residential neighborhood process on bounded neighborhoods and on streets. I again assume that the black and white tolerance levels are $\alpha_b = \frac{3}{4}$ and $\alpha_w = \frac{2}{3}$ respectively, and that 5 percent of all agents are completely tolerant. The street is initially all-white when an influx of blacks into the market occurs ($\lambda = 0.2$). Table 2.3 reveals that convergence is now rapid even though the street is much

[26] The less tolerant whites are, the easier it is for black clusters to expand on the street.

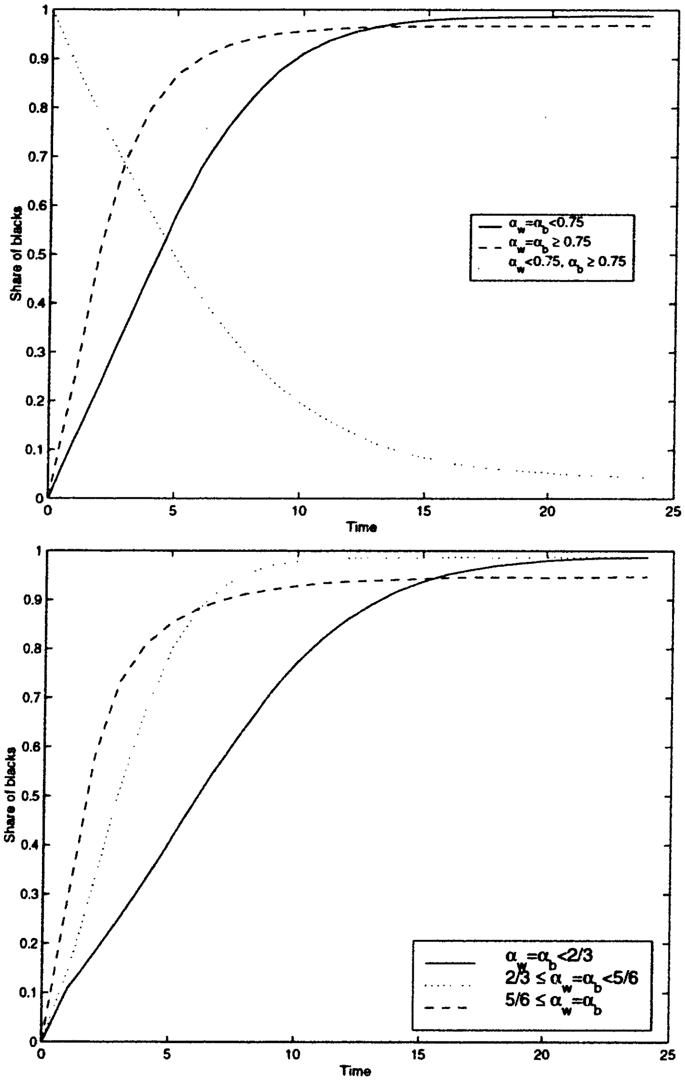
[27] Note that if $\epsilon = 0.005$ a vacant apartment with only white neighbors will switch color with a small probability of 2 percent. Still, it will only take 10 generations until 90 percent of all residents are black.

Figure 2-7: Dependence of long-run share of black residents on the share λ of whites in the housing market for streets with radius of interaction $r = 2$ (top) and $r = 3$ (bottom).



The size of the street has been set at $n = 8100$ and the share of tolerant agents at $\epsilon = 0.01$. The 'long-run share of blacks' was defined as the share of blacks at time $t = 10,000$. In all cases the street started off from a random configuration with 50 percent of residents being black.

Figure 2-8: Evolution of the share of black residents on a street $G_S^{r=2}(8100)$ (top) and $G_S^{r=3}(8100)$ (bottom). The process starts from an all-white neighborhood ($\lambda = 0.2$, $\epsilon = 0.05$) whenever blacks and whites are equally tolerant, and from an all black neighborhood ($\lambda = 0.8$, $\epsilon = 0.05$) if blacks are strictly more tolerant than whites.



Estimated standard errors are 5 percent or less.

Table 2.3: Comparison of waiting times $W(n,[0.97,1])$ until 97 percent of all agents are black on various streets $G_S^r(5000)$ ($\alpha_b = \frac{3}{4}$, $\alpha_w = \frac{2}{3}$, $\epsilon = 0.05$, $\lambda = 0.2$)

Radius of interaction	r=2	r=4	r=6	r=8
Waiting time $W(n,I)$	6.94	7.16	7.68	7.56

Estimated standard errors for the waiting times are 5 percent or less.

larger than the bounded neighborhoods I considered in the previous section. The example also illustrates why the resulting black ghetto lacks persistence on streets. If the balance in the housing market is reversed (i.e. 80 percent of all apartment seekers are white) the street $G_S^{r=2}(5000)$ will become 50 percent white within 5 tenant generations and 75 percent white within 10 generations, as the graph in figure 2-8 (top) illustrates.

2.3.2 Proof of Theorem 2

The proof utilizes some novel techniques from the theory of interacting particle system.[28] The argument proceeds in three steps. First, I propose a simplified Markov process σ_t on the street. Second, I show that the original process η_t and the simplified process σ_t can be *coupled* such that the hypothesis of theorem 2 only has to be proved for the simpler process σ_t. This last step is accomplished through lemma 3. Without loss of generality I restrict attention to the case when streets become black ghettos i.e. $\lambda < \frac{1}{2}$.[29]

The original neighborhood process is difficult to analyze because black clusters continuously form and break up when tolerant white tenants move into a vacant apartment. The simpler process σ_t limits and 'tags' all potential black clusters. Intuitively, the new process makes it both harder for new black clusters to form and easier for existing clusters to break up. The process is therefore biased against blacks in a monotonic fashion: if ghettos develop in the simplified process they should certainly develop in the original process, too.

Formally, the process σ_t is defined as follows. The street is divided up into k segments of fixed length N such that $n = kN$.[30] Residents move out at rate 1 and the process starts from an initial configuration σ_0 where all residents are white. The evolution of the process follows the same switching rules as before with the following qualifications:

[28] Liggett (1985) provides a thorough introduction to this branch of probability theory.

[29] In this section it is no longer assumed that blacks are more tolerant than whites.

[30] I abstract away from integer constraints. I will take $n \to \infty$ and keep N fixed such that the contribution of a single segment of length less than N will vanish by the law of large numbers.

1. I assume that a resident regards any neighbors outside his segment as white. This implies that the dynamics of the process within each segment develops independently from other segments.

2. If a black cluster within the same segment already exists only the (at most two) adjacent white neighbors of the cluster can switch. This guarantees that no seeds for new disjointed clusters can be generated.

3. If a cell switches from black to white such that it divides a cluster up into two separate clusters, the smaller one dies.[31] Together with the previous rule, this assumption ensures that at any point in time at most one black cluster exists within each segment.

These rules completely define the evolution of the process starting from σ_0.

In general, coupling is simply a construction of two stochastic processes on a common probability space - in this case I construct a coupled process (σ_t, η_t) such that the two marginals of the process are the original process η_t and its simpler counterpart σ_t. In order to be of any interest the two processes cannot move independently but must be related in some nontrivial way. I define a simple partial order on the set Z of configurations of the street $G_S(n)$ which allows me to compare the two processes at any point in time:

$$\sigma \leq \eta \quad \text{if and only if} \quad \sigma(z) \leq \eta(z) \quad \text{for all residents } z \tag{2.6}$$

I assume that both processes start to evolve from the same initial all-white configuration $\eta_0 = \sigma_0$. The next lemma shows that there exists a coupled process (σ_t, η_t) such that the original process 'dominates' the new process monotonically, i.e. the inequality $\sigma_t \leq \eta_t$ holds with probability 1 for all $t \geq 0$.

Lemma 2 *There is a coupling (σ_t, η_t) such that both marginal processes start to evolve from the same all-white configuration $\eta_0 = \sigma_0$ and $\sigma_t \leq \eta_t$ holds with probability 1 at any point in time.*

Proof: see appendix A.4

One can immediately conclude that $E(f(\sigma_t)) \leq E(f(\eta_t))$ for each increasing function f

[31] In the case of a tie I assume that the cluster clockwise to the right of the switching cell dies.

on the space of configurations.[32] The share of blacks $X(\eta_t)$ of the residential neighborhood process at any time t, then first-order stochastically dominates the share of blacks $X(\sigma_t)$ of the simplified process.[33] Therefore, the claim in theorem 2 only has to be established for the simplified process σ_t.

I exploit the observation that each segment of the street develops independently in the simplified process. Inside the initially all-white segment, a black cluster of length $[2r(1-\alpha_b)]^+$ will eventually form, which is the minimum length for the cluster to be stable under the undisturbed dynamics ($\epsilon = 0$).[34] Black house-seekers now show interest in the apartments surrounding this minimally stable cluster and the ends will start to move like a random walk with drift under the undisturbed dynamics. The drift is solely determined by the balance of the housing market.

I denote the expected long-run share of blacks in a segment of length N with $E_b(\epsilon)$. As $n \to \infty$ the number of segments becomes arbitrarily large, and by the law of large numbers and lemma 2 we can conclude for the long-run share $\tilde{X}_n$ of blacks in the original process that $P\left(\tilde{X}_n \in (2E_b(\epsilon)-1, 1]\right) \to 1$. In order to finish the proof of theorem 2, the next lemma shows that the expected share of blacks in a segment can get arbitrarily close to 1 for sufficiently small ϵ and large N.

Lemma 3 *There is an upper bound $\hat{\lambda}(r, \alpha_w, \alpha_b) > 0$ such that for each $\lambda < \hat{\lambda}$ and each $\delta > 0$ there is an $\bar{\epsilon}$ such that for all $\epsilon < \bar{\epsilon}$ there is some N such that the expected share of blacks $E_b(\epsilon)$ in the segment of length N fulfills $E_b(\epsilon) > 1 - \delta$.*

Proof: see appendix A.5

The intuition for lemma 3 can be most easily outlined by using the language of stochastic stability analysis (Kandori, Mailath, and Rob 1993, Young 1993). If the share of blacks in the housing market is sufficiently large, the single black cluster living in a large segment of length N always exhibits a positive drift, i.e. is more likely to grow rather than to shrink. The process has essentially two 'limit sets' under the undisturbed dynamics: if the share of

[32] Note that $E(f(\sigma_t)) = \int f(\sigma_t)\, d(\sigma_t, \eta_t)$ and $E(f(\eta_t)) = \int f(\eta_t)\, d(\sigma_t, \eta_t)$ because the coupled process has marginals η_t and σ_t. By construction $\sigma_t \leq \eta_t$ with probability 1 and therefore $f(\sigma_t) \leq f(\eta_t)$ with probability 1. This implies that $\int f(\sigma_t)\, d(\sigma_t, \eta_t) \leq \int f(\eta_t)\, d(\sigma_t, \eta_t)$.

[33] For a proof, define the following increasing function f_{x_0} indexed by each possible share of blacks: $f_{x_0}(\eta)$ is 0 if the share $X(\eta)$ of blacks in the configuration η is below x_0 and equals $X(\eta)$ otherwise.

[34] With $[x]^+$ I denote the smallest integer which is greater than or equal to x.

blacks is $x = 0$ the process can leave the basin of attraction only after a minimally stable cluster of size b has formed. On the other hand, if the share of blacks lies in a neighborhood I of $x = 1$ the process will escape that neighborhood only after a huge waiting time because of the positive drift pushing the process towards $x = 1$. Therefore both the shares $x = 0$ and $x \in I$ form 'limit sets' of the undisturbed dynamics. For any intermediate shares the dynamics will push the process rapidly into I, e.g. there are no further limit sets. It takes b 'mutations' until the basin of attraction of the limit set $x = 0$ can be left. The 'limit set' I can only be exited through a sequence of tolerant whites who move to vacant apartments inside the single black cluster. Each such 'mutation' cuts the length of the black cluster by at most half. Even after $b + 1$ consecutive mutations its length will still be approximately $2^{-(b+1)}N$ and the undisturbed dynamics will push the process back into the 'limit set' I. It therefore takes fewer mutations to reach I than to leave I and we expect the process to spend most of its time inside I.

2.4 Rapid Segregation and Ghetto Persistence in Inner-City Areas

Ghettos can develop rapidly on streets in response to shifts in the housing market, as the previous section demonstrated. This mechanism overcomes the stasis in Schelling's (1972) original tipping model where the transition from an all-white steady state to a black ghetto does not occur within a realistic time frame even when the bounded neighborhood has only moderate size. However, the effect works in both directions, and periodic changes in the composition of the housing market give rise to "avenue waves". In order to reconcile the observed persistence of black ghettos in US cities with the dynamics of the model on streets one has to assume that African Americans have dominated the low-income housing market for the last 100 years. The data does not support this assertion because the migration of blacks to the cities has leveled off while other ethnic minorities (notably Mexican Americans) have grown at a far greater rate in recent decades. In the light of this evidence how can we explain the continued persistence of black ghettos even though blacks face far more competition in the housing market?

In an 'ideal' model the mechanism that gives rise to ghettos should be uni-directional, i.e. ghettos form rapidly but break up slowly. Inner-city areas can provide exactly such an environment if blacks are sufficiently more tolerant than whites. The following assumption on the tolerance levels of blacks and whites ensures that inner-cities preserve the ghetto formation mechanism of streets while making segregation persistent as on bounded neighborhoods.

Assumption 1 *Blacks can tolerate whites constituting 75 percent or more of their neighbors ($\alpha_b \geq \frac{3}{4}$), while whites can only tolerate blacks making up slightly more than 50 percent of their neighbors ($\alpha_w < \frac{1}{2} + \frac{r}{m}$).*[35]

The assumption will hold for the rest of this section.

Just as on streets, randomly forming black clusters can expand, quickly take over the inner-city area and therefore give rise to ghettos as long as blacks dominate the housing market sufficiently. If whites subsequently make up the majority in the housing market, however, random white clusters are hindered in their expansion. The two-dimensional geometry adds a 'geometric' drift that lets small white clusters shrink. This effect can be strong enough to completely counteract the pressure from the housing market which induces white clusters to expand on streets.

2.4.1 Rapid Formation of Black Ghettos in Inner-Cities

It can be easily checked that a black square cluster of size $(r+1) \times (r+1)$ can expand under the undisturbed dynamics, i.e. it can take over the inner-city area with positive probability even if there are no tolerant black house-seekers ($\epsilon = 0$).[36] This observation essentially guarantees that the inner-city will turn into a black ghetto, both in the medium and in the long run, if blacks dominate the housing market.

The formal proof of this claim exploits the results of the previous section by breaking the inner-city up into 'stripes' of width $r+1$, shown in figure 2-9. Although a stable $(r+1) \times (r+1)$ cluster can expand in all four directions in an inner-city, the ghetto formation mechanism will work even if it could only expand in East/ West direction. Each stripe will then behave very much like a street: stable clusters of length $r+1$ form and take over the stripe within a waiting time of order $O(1)$ if blacks dominate the housing market. Moreover,

[35] *The size of an individual neighborhood on the inner-city area $G_C^r(n)$ is m, i.e. $|N(z)| = m$. If the radius of interaction is $r = 1$ ($r = 2$) we have $m = 4$ ($m = 12$) and whites can tolerate at most two (seven) black neighbors.*

[36] Recall, that on streets the minimally stable black cluster had length $[2r(1-\alpha_b)]^+$.

the process will cluster around a black share of $x = 1$ on each segment and, hence, on the entire inner-city area. The proof of the next theorem goes through this reasoning in greater detail.

Theorem 3 *Given is an inner-city area $G_C^r(n)$ with group tolerance levels α_w and α_b satisfying assumption 1. Then there exists some critical value $0 < \hat{\lambda}(r, \alpha_w, \alpha_b) < 1$ such that the following holds when blacks dominate the housing market sufficiently ($\lambda < \hat{\lambda}$).*

1. *The inner-city becomes a black ghetto in the long run, i.e. the process clusters on the interval $[x_b^*(\epsilon), 1]$ with $\lim_{\epsilon \to 0} x_b^*(\epsilon) = 1$.*

2. *If the share ϵ of tolerant agents is sufficiently small the waiting time until the share of blacks exceeds $1 - \delta$ satisfies $W(n, [1 - \delta, 1]) = O(1)$.*

Proof: The proof is easiest outlined for the case $r = 1$. The inner-city area is 'sliced' up into 'stripes' of length N and width $r + 1 = 2$ (see figure 2-9). As in section 2.3.2 I construct a simplified process σ_t which evolves independently on each stripe. The switching rules are translated in a straightforward manner: they only differ in their emphasis on 'stripes' instead of 'segments'.

1. A resident regards any neighbors outside her stripe as white. Therefore, the dynamics of the process within each stripe develops independently from other stripes.

2. If a black cluster within a stripe exists, only adjacent white neighbors can switch to black. A cluster has to 'fill up' vertically before it can expand horizontally. This rule implies that the single black cluster always fills up the full width of the stripe. The length of the single black cluster on a stripe can therefore be treated in the same way as the size of the single black cluster on a street segment.

3. If a cell switches from black to white inside the single black cluster the cluster is cut in two, and the shorter half is eliminated. This ensures that at any point in time there exists at most one black cluster within each segment.

Due to assumption 1, a black cluster can only be invaded by whites under the undisturbed dynamics at its two boundaries. Any apartment inside a black cluster with a distance of at least r from either boundary has a black neighborhood share of $\frac{1}{2} + \frac{r}{m}$

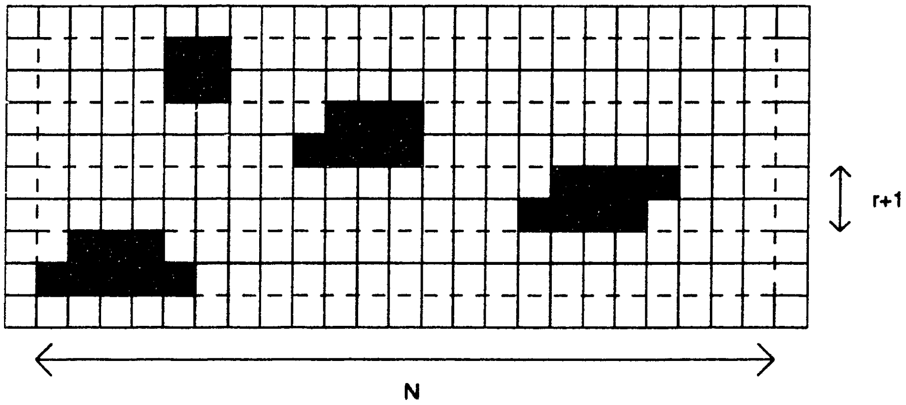


Figure 2-9: Four $N \times (r+1)$ stripes on an inner-city; the coupled process evolves independently within each stripe

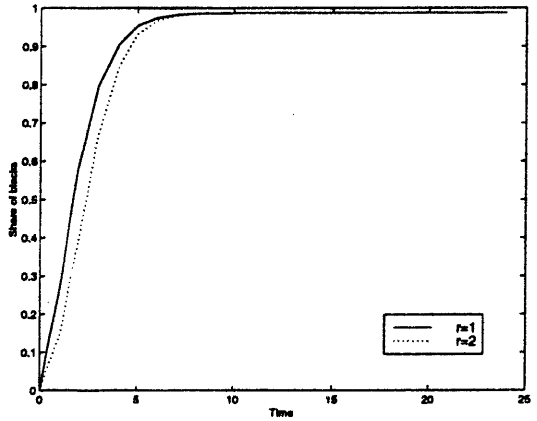
which exceeds the white tolerance level α_w. Apartments at the boundary of the cluster, on the other hand, have a black neighborhood share of at least 25 percent, which makes them acceptable to all black house-seekers. Each cluster behaves like a cluster on a street segment: it can expand under the undisturbed dynamics with a drift depending on the composition of the housing market, and it can only be broken in half by rare ϵ-jumps. The proofs of theorem 2 and lemma 1 can be easily adapted to establish the existence of $\hat{\lambda} > 0$ and fast convergence. QED

Monte-Carlo simulations confirm that black ghettos form as rapidly in inner-cities as they do on streets. Figure 2-10 illustrates the medium-run evolution of the residential neighborhood process in inner-cities with radius of interaction $r = 1$ and $r = 2$ where, again, 5 percent of agents are tolerant and blacks constitute 80 percent of the housing market. In both cases I have set the white and black tolerance levels at $\alpha_w = \frac{7}{12}$ and $\alpha_b = \frac{3}{4}$ such that they just satisfy assumption 1. Within five tenant generations 90 percent of all residents are black, which is almost the same waiting time I obtained for streets (see figure 2-8).

2.4.2 "Encircling" and Persistence of Ghettos in Inner-Cities

On streets the mechanism that gives rise to ghettos is fully reversible. In response to an influx of white apartment-seekers, white clusters can form and rapidly break up the black ghetto. The fact that blacks have a far higher tolerance level than whites only matters

Figure 2-10: Evolution of the share of black residents in the inner-city areas $G_C^{r=1}$ (8100) and $G_C^{r=2}$ (8100). The process starts from an all white neighborhood with tolerance levels $\alpha_b = \frac{3}{4}$ and $\alpha_w = \frac{7}{12}$ ($\lambda = 0.2$, $\epsilon = 0.05$).



Estimated standard errors are 5 percent or less.

insofar as whites have to dominate the housing market relatively more in order to break up the all-black equilibrium ($\tilde{\lambda} > 1 - \hat{\lambda}$). However, the residential neighborhood process behaves in a qualitatively different way in inner-cities. 'Small' white clusters can nolonger expand under the undisturbed dynamics.

I consider a white cluster inside a black inner-city ghetto to be 'small' if it is "encircled" by black residents, i.e. it does not span the residential area.[37] Such a cluster can be covered by a rectangle which is convex in the two-dimensional geometry. For this reason, each apartment vacated by a black resident along the boundary of this rectangle has more black than white neighbors. More precisely, a vacant apartment outside the rectangle has a black neighborhood share of at least $\frac{1}{2} + \frac{r}{m}$ which exceeds the tolerance level of whites. Close to the corners of the rectangle the black share even approaches 75 percent. Therefore, the white cluster can never expand beyond the rectangle unless tolerant house-seekers move in along the boundaries. Black house-seekers, on the other hand, can easily invade the white cluster under the undisturbed dynamics ($\epsilon = 0$). Obviously, the cluster has to die out in this environment and can never take over the inner-city regardless of the balance in the housing market as the following lemma shows.

[37] Formally, I call a cluster of residents "encircled" in an inner-city with radius of interaction r if the cluster can be covered by a rectangle with width and length not exceeding $\sqrt{n} - 1 - r$. This ensures that the dynamics of the process along the boundary is not influenced by the finiteness of the inner-city.

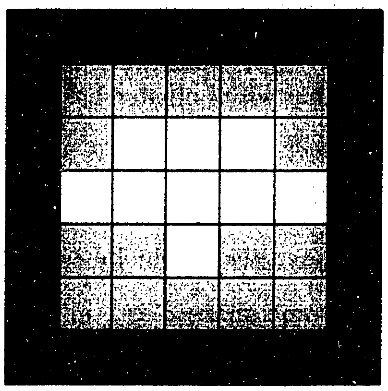


Figure 2-11: Isolated white cluster in an inner-city area $G_C^{r=1}(49)$. The cluster can be covered by a 5×5 rectangle (lightly shaded).

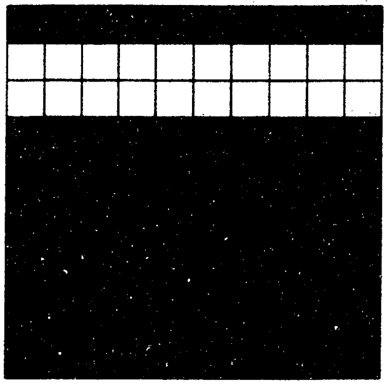


Figure 2-12: 'Large' non-encircled white cluster in an inner-city with radius of interaction $r = 1$

Lemma 4 *Under assumption 1 and without the presence of tolerant agents* ($\epsilon = 0$) *an "encircled" white cluster in an inner-city area $G_C^r(n)$ will die out almost surely for any balance in the housing market* $0 < \lambda < 1$.

Proof: see appendix A.7

A white cluster can therefore only survive under the undisturbed dynamics and lie outside the basin of attraction[38] of the black ghetto configuration if it is 'large' and nolonger encircled. The stable 'large' cluster shown in figure 2-12 serves as an example.

[38] The basin of attraction $D(\Omega)$ of some subset of configurations $\Omega \subset Z$ is the set of configurations from which the undisturbed process reaches an element of Ω with probability 1.

The minimum number of completely tolerant house-seekers who have to move into the black ghetto in order to form a non-encircled white cluster grows with $\sqrt{n}$. On streets, on the other hand, the minimally stable white cluster has only size $[2r(1-\alpha_w)]^+$, which does not depend on the size of the street. Intuitively, it should therefore take longer to leave a black inner-city ghetto than leave a black street ghetto if the share ϵ of tolerant agents is small. The next theorem confirms this insight by comparing the waiting times until the share of white residents exceeds some share $\delta < 1$ in an inner-city and on a street *of the same size n.*

Theorem 4 *Given are an inner-city area $G_C^r(n)$ and a street $G_S^r(n)$ of equal size n. Assumption 1 on the black and white tolerance levels holds, and the share ϵ of tolerant agents is small. Then the waiting time until the share of whites exceeds δ satisfies $W_C(n,[0,1-\delta]) \geq \epsilon^{-\left[\frac{\sqrt{n}}{r+1}\right]}$ in the inner-city area and $W_S(n,[0,1-\delta]) \sim \epsilon^{-[2r(1-\alpha_w)]^+}$ on the street.*

Proof: see appendix A.8

The ratio $\frac{W_C(n,[0,1-\delta])}{W_S(n,[0,1-\delta])}$ of the waiting times to leave an inner-city and street ghetto respectively, can become arbitrarily large if there are few tolerant house-seekers and the size of the residential area is large. The "encircling phenomenon" therefore lends persistence to black inner-city ghettos, and makes the ghetto formation process uni-directional. Inner-cities are in some sense a 'hybrid' geometry because they combine features of streets and bounded neighborhoods.[39]

Theorem 4 does not allow us to assess if the increase in the persistence of ghettos in inner-cities compared to streets is quantitatively significant. In particular, the share ϵ of tolerant agents should not be unrealistically small for the effect to apply. For this purpose I have relied on simulations in order to calculate the waiting times $W(0,[0,0.5])$ until at least 50 percent of all residents are white. Table 2.4 compares the waiting times for the street/ inner-city pairs $\left(G_S^{r=2}(n), G_C^{r=1}(n)\right)$ and $\left(G_S^{r=6}(n), G_C^{r=2}(n)\right)$, respectively. These pairs have been chosen so that individual neighborhoods have equal size in both geometries,

[39] A comparison of the waiting times to leave a black bounded neighborhood ghetto and an inner-city ghetto of equal size n also reveal that inner-cities lie somehow 'between' streets and bounded neighborhoods. It takes $[(n-1)(1-\alpha_w)]^+$ 'mutations' to leave the basin of attraction of a black ghetto in a bounded neighborhood. As in theorem 4 one can then show that $\frac{W_B(n,[0,1-\delta])}{W_C(n,[0,1-\delta])} \to \infty$ as $\epsilon \to 0$ because n grows faster than $\sqrt{n}$. Inner-cities are therefore more persistent than streets but less persistent than bounded neighborhoods.

Table 2.4: Comparison of waiting times $W(n, [0, 0.5])$ until more than 50 percent of residents are white starting from a black ghetto on a street and an inner-city area with individual neighborhoods of equal size m ($\lambda = 0.8$, $\alpha_b = \frac{3}{4}$, $\alpha_w = \frac{7}{12}$, $n = 8100$).

Expected Wait ($m = 4$)

	$\epsilon = 0.005$	0.01	0.015	0.02	0.03	0.04	0.05
$G_S^{r=2}(n)$	57.4	26.9	17.3	12.8	8.5	6.8	5.4
$G_C^{r=1}(n)$	594.5	63.7	27.3	16.9	9.8	7.1	5.8

Expected Wait ($m = 12$)

	$\epsilon = 0.005$	0.01	0.015	0.02	0.03	0.04	0.05
$G_S^{r=6}(n)$	1618	267.4	94.0	48.5	19.9	11.4	8.1
$G_C^{r=2}(n)$	$> 3 \times 10^4$	2912.3	169.0	66.4	21.1	11.8	8.0

Estimated standard errors for the waiting times are 5% or less.

i.e. $m = 4$ and $m = 12$. In both cases I have again set the white and black tolerance levels at $\alpha_w = \frac{7}{12}$ and $\alpha_b = \frac{3}{4}$. I have assumed that 80 percent of all house seekers are white, and I have varied the share of tolerant agents between one half of a percent and 5 percent. This implies that a vacant apartment with only black neighbors will switch color with a probability between 2 percent and 17 percent which I consider reasonably large. The waiting times for the inner-cities are consistently larger than the corresponding waiting times on streets. For $\epsilon = 0.02$ the inner-cities are more than 30 percent more persistent while for $\epsilon = 0.01$ the difference is at least three-fold.

2.5 Historical Evidence

This section evaluates the empirical relevance of my model for explaining segregation in the US. The main prediction of the theory is the rapid formation of ghettos on streets and in inner-cities as a response to shifts in the housing market. As a natural first-pass test I compare the degree of residential segregation over the last 200 years with changes in the relative demand for housing between African Americans and non-blacks. This exercise provides broad support for the model. In order to test the more subtle implications of the theory I take a closer look at the rise of the black ghetto in Harlem. In particular, I find that the dynamics of Harlem's transformation conform with the contagious growth process predicted by the theory, which has random black clusters form and then expand along the boundaries. Although I concentrate on the residential separation between African Americans and non-blacks for most of this section I provide evidence for "avenue waves" in Chicago which involved both African Americans and European immigrants.

2.5.1 The Rise of the Black Ghetto at the Turn of the Century

After the Civil War race relations improved, and northern cities such as Cleveland, Philadelphia and Chicago established integrated schools, hospitals and colleges.[40] Throughout the 19th century segregation between African Americans and non-blacks was relatively low, and the average African American lived in a ward that was only 20 percent black (Cutler, Glaeser, and Vigdor 1997, table 2). As far as spatial concentrations of blacks existed, they had not stabilized yet: in New York, for example, the principal clusters of black concentration moved repeatedly over the century.

Between 1860 and 1890 the share of blacks in fact decreased in many of the booming northern cities, and blacks made up only 2.5 percent of the population in the North and Midwest in 1890. Blacks started to leave the South and migrate to the booming North in significant numbers only after 1890.[41] Even then, the black growth rate just about matched the rate of increase of the general population, such that the share of blacks in the housing market was likely to be low. This combination of a low degree of segregation, shifting clusters of black concentration and an insignificant presence in the housing market before the turn of the century is consistent with my model when the share of blacks is non-critical, i.e. smaller than the share $1-\hat{\lambda}$ which has to be exceeded to give rise to ghettos.

This picture changed radically with the outbreak of the first World War in 1914. Immigration to the US fell off sharply and never recovered after the war due to the immigration restriction enacted in the 1920s. The manufacturing industry continued to demand cheap labor, and companies began to dispatch labor agents to the South in order to convince more African Americans to move north. These factors resulted in a massive population movement between 1916 and 1919 known as the *Great Migration*[42] that continued well into

[40] Cleveland, for example, integrated schools in 1871 (Kusmer 1976).

[41] It is puzzling that few blacks left the South between 1865 and 1890 even though economic conditions were poor. Kusmer (1976) argues that the first black generation born in freedom had a different perspective from their parents and sought to exploit improved economic opportunities more actively.

[42] In these years alone the number of African Americans doubled in Cleveland, tripled in Chicago and increased more than sixfold in Detroit (Kusmer 1976).

the second half of the 20th century. The annual growth rate of the black urban population in the North was 3.1 percent between 1910 and 1940 and 4.4 percent between 1940 and 1970 (Cutler, Glaeser, and Vigdor 1997, table 2).

Competition for housing intensified because the black community grew at a much faster rate than the general population in northern cities. In Chicago, for example, 18.5 percent of the net inflow of newcomers between 1920 and 1930 were African American, and in Cleveland the corresponding share was an impressive 36.2 percent.[43] As blacks were poorer than the average newcomer, they presumably dominated the market for apartments at the lower end of the socio-economic scale far more than these numbers indicate. It is likely that the share of blacks did not even have to exceed 50 percent to transform a residential area into a black ghetto[44] because evidence from surveys[45] shows consistently that blacks have a higher tolerance level than whites. Therefore, it is perfectly plausible that the share of blacks in northern cities was high enough during this period to trigger the rapid formation of ghettos in inner-cities and on streets as predicted by the theory.

In the wake of the Great Migration, northern US cities, in particular, became indeed much more segregated. In 1940 the average African American lived in a residential area that was 37.6 percent black and by 1970 that share had increased to almost 70 percent. Cutler, Glaeser, and Vigdor (1997) found in a sample of 313 US cities that only 5 cities had ghettos in 1910 but more than a third had one by 1970.[46] Most of these almost exclusively black neighborhoods formed around the principal black cluster of concentration that happened to exist before the Great Migration.

In my model decentralized racism is the sole transmission channel that translates the conditions in the housing market during and after the first World War into changes in the level of segregation. Although other factors undoubtedly contributed to this process, I argue that decentralized racism was the most significant channel during the early formative years. First, sorting by socio-economic differences explains less than half of the observed variation in segregation indices between neighborhoods, even in the 1950s (Taeuber and

[43] See table 1 in Spear (1967) for Chicago and table 1 in Kusmer (1976) for Cleveland.

[44] In section 2.3.1 I found that a street $G_S^{r=2}$ with black and white tolerance levels $\alpha_b = \frac{3}{4}$ and $\alpha_w = \frac{1}{2}$ becomes a ghetto as soon as the share of blacks exceeds 30 percent.

[45] Cutler, Glaeser, and Vigdor (1997) cite evidence from the General Society Survey.

[46] The authors characterize a city as having a ghetto if the index of dissimilarity is greater than 0.6 and the index of isolation exceeds 0.3 (see the paper for details on calculating these standard measures).

Taeuber 1965). Second, collective-action racism[47] could not prevent the invasion of white neighborhoods at the onset of the Great Migration, as the example of Harlem shows. African Americans only made up a small share of the population, and the relatively peaceful race relations were only gradually overshadowed by resentment due to continued black migration from the South. Even neighborhoods that remained largely white during this period experienced scattered instances of black families moving in and out according to an unpredictable pattern.[48] Such noisy "reshuffling" of blacks in 'white ghettos' does not fit a theory of collective racism but is consistent with my model in the case of up-scale neighborhoods, for example, which few blacks could afford and where whites would face little competition in the housing market.

There is stronger evidence that collective-action racism played a more significant role by the middle of the century in sustaining segregation. By then, ghettos were already well-established, and a formal and informal institutional framework consolidating segregation had developed. Cutler, Glaeser, and Vigdor (1997) find that in 1940, blacks paid relatively more for equivalent housing than whites in more segregated cities, as compared to less segregated cities. While this observation is consistent with some degree of collective-action racism, it disappeared from the data by 1990. Nowadays whites pay more for equivalent housing than blacks in more segregated areas, suggesting that decentralized racism is again the driving force behind continuing segregation.

Segregation has slightly declined since the 1970s because formerly all-white neighborhoods have become more racially mixed. But almost exclusively black ghettos persist and show very little sign of change. The stability of ghettos is surprising because blacks no longer dominate the housing market as they did in the aftermath of the Great Migration. After 1970 the black community in the northern cities only increased at an annual rate of 0.9 percent. While the share of blacks has stagnated other ethnic groups have shown vigorous growth. In particular, the share of Hispanics in US cities doubled between 1970 and 1990 to 10.3 percent.[49] Inner-cities can provide an explanation for the longevity of black ghettos

[47] An important type of collective-action racism were racial zoning or restrictive covenants that excluded blacks from particular residential areas (Massey and Denton 1993).

[48] Smith (1959) compared the census data for New Haven between 1940 and 1950 and found that 76 blocks with a black share of less than 10 percent became all-white, while black families moved into 72 formerly all-white blocks. One third of these new blocks were contiguous to those which they were replacing and the rest were scattered throughout the city and lacked any spatial pattern.

[49] Mexican Americans are also highly segregated and live in neighborhoods with a Mexican share of 50.3 percent (Borjas 1995, table 4).

despite the fact that African Americans face far more competition in the housing market. The theory also suggests that today's inner-city ghettos are unlikely to disappear in the near and even medium term unless they are forcibly broken up by some policy intervention, such as urban redevelopment.

2.5.2 Harlem's Transformation into a Black Ghetto

Harlem was an affluent suburb of New York City in the 19th century and became the largest black ghetto in the US by 1920. The various stages in the spatial growth of the ghetto are well documented, which allows me to directly test the dynamic predictions of my theory on streets and in inner-cities, i.e. the growth of randomly forming black cluster around their boundary.[50]

New York City provides an almost ideal environment for the application of my model. The residential turnover rate was high because former peripheral neighborhoods, such as Harlem, were continually redeveloped as the metropolis expanded northwards on Manhattan island. Harlem, for example, was a rural village and became incorporated only in 1873. By 1886, the three lines of the elevated railroad came as far north as 129th Street, and a massive building boom in the 1880s made Harlem a preferred residential area for New York City's white upper- and upper-middle-classes. Lower Harlem experienced an influx of Eastern European Jews in the 1890s and of Italians before 1890 (see the map of Harlem in figure 2-13).

Like other northern cities, New York was not particularly segregated in the 19th century. African Americans did not dominate any single neighborhood, and the principal clusters of black concentration moved repeatedly up the West Side over the course of the century.[51] In 1890, for example, six wards in Manhattan had a substantial black population of between 2,000 and 4,000.

There were few scattered black families in Harlem before the turn of the century. Most of them were servants and lived at the periphery of white Harlem. Blocks occupied by African Americans in 1902 are marked in figure 2-13. African Americans entered Harlem in greater

[50] This section is based on Osofsky's (1963) comprehensive history of Harlem (especially chapters 5-8).

[51] In the early 19th century many blacks lived in the Five Points district on the site of the present City Hall. By 1860 the district was overwhelmingly Irish and the largest cluster of African Americans could be found in Greenwich Village. Between 1880 and 1890 their numbers declined as the district became predominantly Italian. San Juan Hill and the "Tenderloin" (between 20th Street and 53rd Street) emerged as the most populous black residential areas.

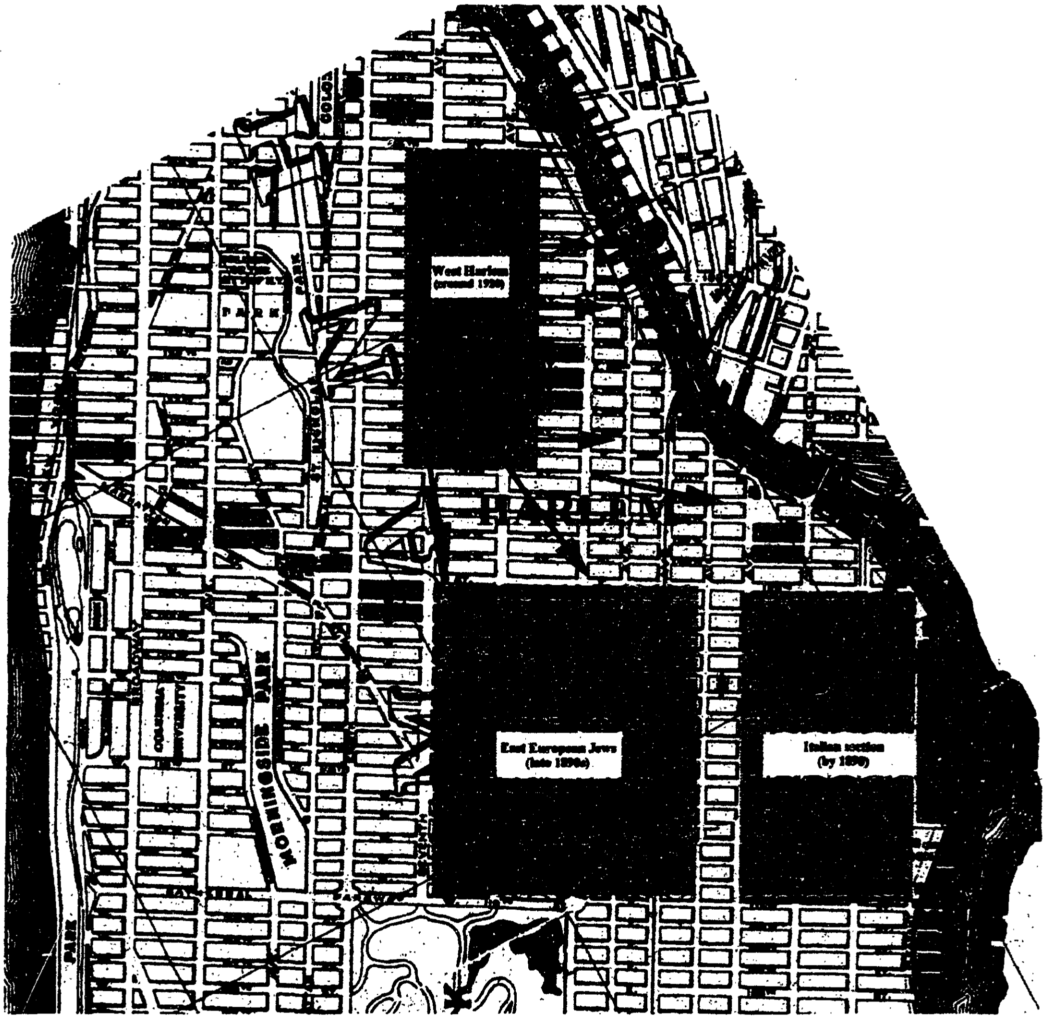


Figure 2-13: Harlem was an essentially all-white middle and upper-class neighborhood in the 19th century. Lower Harlem saw an influx of Eastern European Jews in the 1890s and had an Italian section in the south-east. Only a few scattered black families lived in blocks at the periphery of central Harlem which are marked black in the figure. West Harlem became a black ghetto until 1920 (darkly shaded). The ghetto expanded in the 1920s to the east and south as indicated by the arrows. Blacks lived as far south as 110th Street by the end of the decade.

numbers during the years 1900-1914, when the black population of Manhattan doubled. Development in West Harlem north of 130th Street had been slow in the 1890s because the area lacked public transportation. The construction of the Lenox subway line up to 145th Street in the years 1898 to 1904 set off a building boom and massive speculation in property along Lenox and Seventh Avenue where entire new apartment blocks were built. By 1904-5 the bubble burst and realtors woke up to the fact that there was insufficient demand for these high-quality apartments. Landlords began to compete intensely for tenants and some of them started to open their apartment houses to blacks. Demand for decent housing was strong amongst African Americans during this period. More and more blacks migrated to New York City, and established blacks were displaced from their old living quarters in Lower Manhattan as the business district expanded.[52]

This combination of factors colluded to establish the initial black cluster of residents in Harlem which then expanded dramatically in the aftermath of the Great Migration. Osofsky (1963, p. 17) emphasizes that Harlem's black colony would most likely have been a passing phenomenon just like previous clusters of black concentration in Lower Manhattan without the enormous influx of black migrants after 1900:

> The most important factor underlying the establishment of Harlem as a Negro community was the substantial increase of Negro population in New York City in the years 1890-1914. That Harlem became the specific center of Negro settlement was the result of circumstance; that *some* section of the city was destined to become a Negro ghetto was the inevitable consequence of the Negro's migration from the South.

African Americans took over West Harlem in a striking geographical pattern which resembles contagious growth. There was a clear 'color line' that separated the southward advancing black settlement from established white residents. This type of dynamics is not only predicted by my theory but, more generally, suggests that decentralized local interaction between residents should be at the heart of any realistic model of segregation. Landlords attempted to stop the black invasion through collective-action racism and successively signed restrictive agreements on West 140th, 137th, 135th, 131st, 129th Streets

[52] Many black apartment blocks disappeared when Pennsylvania Station was built in the Tenderloin at the beginning of the century. In 1914 African Americans occupied about 1,100 different houses within a 23 block area of Harlem and 80% of the whole black population lived in Harlem by then.

etc., which obliged them not to rent to blacks. Each of these local arrangements ultimately failed because sooner or later some landlord would 'panic-sell' and the coalition collapsed. Restrictive agreements also invited 'block-blusters' to test the strength of support for unified action. These speculators bought single apartment houses on an all-white street and invited black tenants. Adjoining white owners then had to re-purchase the apartment house at inflated prices in order to evict the black tenants again.

By 1920, an almost exclusively black ghetto had formed north of 130th Street and West of 5th Avenue as shown in figure 2-13. The ghetto expanded further to the east and south and completely crowded out the white residents in central Harlem. By the end of the 1920s African Americans lived as far south as 110th Street and the ghetto consolidated in the subsequent decades as black migration continued.

2.5.3 "Avenue waves" in Chicago

I have emphasized that ghettos are not persistent on streets because the ghetto formation process is reversible. The model therefore predicts "avenue waves" in response to periodic shifts in the housing market. In the 19th century there existed a natural source of variation because different groups of immigrants entered the country at different points in time. Immigration in the first half of the century was characterized by waves from Northern Europe while in the latter half of the century immigrants from Southern and Eastern Europe dominated. If one assumes that most immigrants arrived poor in the US and climbed the social hierarchy at similar rates one would expect that different ethnic waves of immigrants joined the housing market for residential areas of a particular quality at different points in time. This, in turn, induced shifts in the balance λ of the housing market.[53]

Burgess (1928) recorded the resulting avenue waves in Chicago as shown in figure 2-14. He notes that "the great arterial business streets of the city have been and remain the highways of invasion". Of particular interest are the A and B waves where 'new' immigrant groups (Hungarians, Italians and Poles) crowd out 'older' immigrants (Germans and Scandinavians).

[53] The "avenue waves" effect should be strongest for streets at the lower socio-economic end of the market. The diffusion rate at which residents move into better neighborhoods differs by ethnicity and between individuals such that ethnic waves should become increasingly intermingled.

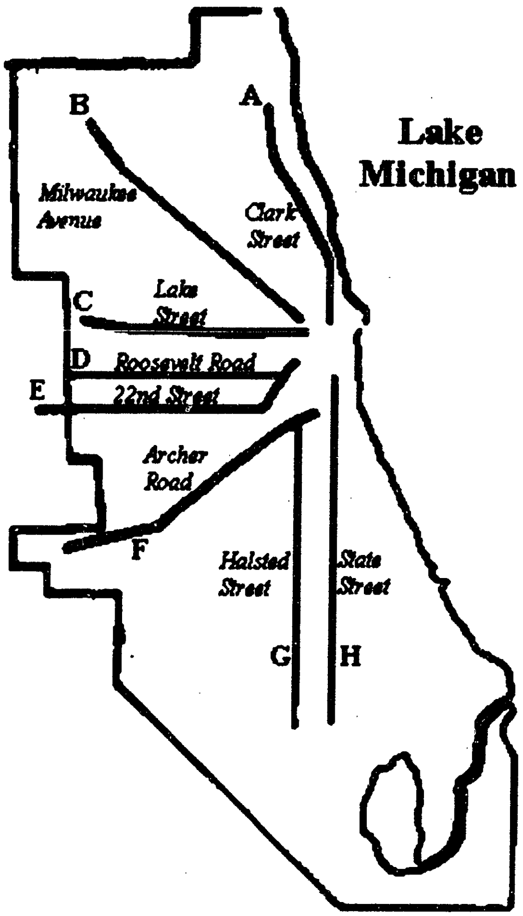


Figure 2-14: Chicago's avenue waves (Burgess 1928) - Germans/Scandinavians followed by Hungarians/ Italians (A), Germans/ Scandinavians followed by Poles (B), invasion of African Americans (C), invasion of Russian Jews (D), invasion of Czechs (E), Polish invasion (F), invasion of Irish (G), Invasion of African Americans (H)

2.6 The Relationship between Stochastic Stability and Clustering

My analysis characterized the evolution of the residential neighborhood process through clustering rather than the standard stochastic stability techniques developed by Kandori, Mailath, and Rob (1993) and Young (1993). This section explores the relationship between both techniques and concludes that stochastic stability can seriously mispredict the long-run behavior of a stochastic system because it ignores too much information about its undisturbed dynamics. The problem is most severe for the large-scale systems which we typically encounter in evolutionary environments. In the context of my model I show that stochastic stability fails to predict the rise of ghettos on streets because it does not take into account the balance in the housing market.

But it is my hope that clustering will enhance our understanding of both future and existing evolutionary models. I present two examples which are intended to illustrate the robustness of clustering versus stochastic stability. First, I consider a simple model of imitation where agents live on a general (not necessarily regular) class of graphs and hold one of two possible 'opinions'. Whereas stochastic stability predicts that agents synchronize their opinions in the long run, clustering analysis reveals that, to the contrary, a system of many agents will typically be in disagreement. Second, I revisit a well-known large population coordination game which has been studied by Ellison (1993). Although both clustering and stochastic stability predict the long-run behavior of the large-scale system correctly, in this case the standard technique is vulnerable to seemingly innocuous changes in the dynamics.

I will rely on a characterization of stochastic stability recently introduced by Ellison (1999). His waiting-time approach, unlike the "tree-surgery" arguments used in the earlier papers, not only makes the reasoning behind stochastic stability arguments very transparent but also allows me to clearly identify the key weakness of this concept. Ellison considers a general 'model of evolution' (Z, P, P^{ϵ}) with a state space Z and a Markov process defined over Z in discrete time[54] with 'disturbed' transition matrix P^{ϵ} and 'undisturbed' transition matrix P. The matrix is assumed to be ergodic for each $\epsilon > 0$ and, P^{ϵ} is continuous in ϵ

[54] The residential neighborhood process is defined in continuous time. Appendix A.1 illustrates how such a process can be transformed into a corresponding discrete time process such that all of Ellison's results carry over.

such that $P^0 = P$. Ellison then defines a cost function $c : Z \times Z \to R^+ \cup \infty$ such that for all pairs of states $\eta, \eta' \in Z$, $\lim_{\epsilon \to 0} P^{\epsilon}_{\eta\eta'} / \epsilon^{c(\eta,\eta')}$ exists and is strictly positive if $c(\eta, \eta') < \infty$ (with $P^{\epsilon}_{\eta\eta'} = 0$ for sufficiently small ϵ if $c(\eta, \eta') = \infty$). Intuitively, the cost of transition can be thought of as the number of independent mutations necessary for it to occur.

Ellison introduces two new concepts, the radius and the coradius, which he uses to bound the waiting times required to leave and enter the basin of attraction of a union of limit sets[55] $\Omega \subset Z$. The radius $R(\Omega)$ describes the minimum cost of leaving the basin of attraction $D(\Omega)$ and is a measure of the persistence of the process when it rests at Ω. Formally, Ellison defines a path out of $D(\Omega)$ as a sequence of distinct states $(\eta_1, \eta_2, .., \eta_T)$ with $\eta_1 \in \Omega$, $\eta_t \in D(\Omega)$ for $1 < t < T$ and $\eta_T \notin D(\Omega)$. The set of all these paths is denoted $S(\Omega, Z - D(\Omega))$. The radius can then be defined as

$$R(\Omega) = \min_{(\eta_1,..,\eta_T) \in S(\Omega, Z - D(\Omega))} \sum_{t=1}^{T-1} c(\eta_t, \eta_{t+1}). \tag{2.7}$$

The coradius $CR(\Omega)$, on the other hand, captures the length of time necessary to reach the basin of attraction of Ω starting from any other state by counting the number of intermediate mutations:[56]

$$CR(\Omega) = \max_{\eta_1 \notin \Omega} \min_{(\eta_1,...\eta_T) \in S(\eta_1, \Omega)} \sum_{t=1}^{T-1} c(\eta_t, \eta_{t+1}) \tag{2.8}$$

A combination of a large radius $R(\Omega)$ and a small coradius $CR(\Omega)$ ensures that the process reaches the basin of attraction $D(\Omega)$ quickly, but is very reluctant to leave it. Building on this intuition Ellison can prove the following theorem:

Theorem 5 *The union of limit sets Ω is stochastically stable if $R(\Omega) > CR(\Omega)$. The waiting time to leave the basin of attraction $D(\Omega)$ is $W(\Omega, Z - D(\Omega), \epsilon) \sim \epsilon^{-R(\Omega)}$, and the process reaches a long-run equilibrium after a waiting time of $W(\eta, \Omega, \epsilon) = O\left(\epsilon^{-CR(\Omega)}\right)$ for any $\eta \notin \Omega$.*

Proof: see Ellison (1999)

[55] The limit sets or recurrent classes of a stochastic system are the sets of states which can persist in the long run absent noise or mutations ($\epsilon = 0$).

[56] Ellison also defines the modified coradius $CR^*(\Omega)$ which bounds the waiting time until convergence more precisely. For the purpose of comparing clustering and stochastic stability, however, is suffices to use the simple coradius.

The waiting time approach has the advantage of giving a bound on the rate of convergence to the long-run equilibrium. In models of local interaction the coradius typically remains small even if the system is large (as in Ellison (1993)). This observation is then interpreted as evidence that local interaction 'speeds up' convergence to the long-run equilibrium.

Unfortunately, the coradius can give a very misleading picture of how fast the process actually reaches the basin of attraction $D(\Omega)$. The random walk example discussed in the introduction illustrates the problem nicely (see figure 2-1). It is easily shown that only the state $n+2$ is stochastically stable because it takes at most one mutation to reach its basin of attraction, but two mutations to leave it (i.e. $R(\{n+2\}) = 2$ and $CR(\{n+2\}) = 1$). Theorem 5 also indicates that the process reaches its long-run equilibrium quickly because the waiting time is of the order $O\left(\frac{1}{\epsilon}\right)$ independent of the 'size' n.

The example is simple enough to carry out a more careful analysis. Although the process will spend almost all its time at $n+2$ as $\epsilon \to 0$, it is instructive to calculate how small ϵ has to be depending on the size n of the system in order to find the process in a small δ-neighborhood $[(1-\delta)n, n+2]$ of the long-run equilibrium with probability $\gamma > 0$. The following condition on ϵ and n has to be satisfied:[57]

$$\frac{2\frac{\epsilon}{1-\epsilon}\left(\left(\frac{1}{2}\right)^{(1-\delta)n}-\left(\frac{1}{2}\right)^{n}\right)+\left(\frac{1}{2}\right)^{n-1}+\left(\frac{1}{2}\right)^{n-1}\frac{1-\epsilon}{\epsilon}}{1+2\frac{\epsilon}{1-\epsilon}\left(1-\left(\frac{1}{2}\right)^{n}\right)+\left(\frac{1}{2}\right)^{n-1}+\left(\frac{1}{2}\right)^{n-1}\frac{1-\epsilon}{\epsilon}}\geq\gamma$$

For large n this condition becomes approximately

$$\left(\frac{1}{2}\right)^{n-1}\frac{1-\epsilon}{\epsilon}\geq\frac{\gamma}{1-\gamma}, \tag{2.9}$$

which requires that $\epsilon < \bar{\epsilon}_n = \left(\frac{3}{2}\right)^{-n}$.[58]

Therefore, stochastic stability describes the long-run behavior of the random walk well on large systems only if the noise term ϵ is extremely small. Using the technique of appendix A.2 it is straightforward to show that the waiting time until convergence is at least of the order 3^n even though it takes just a single mutation to reach the δ-neighborhood. Convergence to the stochastically stable equilibrium is therefore anything but fast.[59]

[57] The probability of finding the random walk in the δ-neighborhood can be calculated as in appendix A.3.

[58] Note, that otherwise $\left(\frac{1}{2}\right)^{n-1}\frac{1-\epsilon}{\epsilon} < \left(\frac{1}{2}\right)^{n-1}\left(\frac{3}{2}\right)^{n} \to 0$ as $n \to \infty$.

[59] The estimate for the waiting time is the product of the waiting time until a single mutation occurs (which

What has gone wrong? Stochastic stability analysis essentially ignores the nature of the undisturbed dynamics. In my simple example it suffices that there is *some*, however small, positive probability of reaching state n starting from state 1. This transition has zero cost attached to it and hence does not enter the calculation of the coradius. But the larger the size n of the system, the more the undisturbed dynamics pushes the process away from state n, and the error term ϵ has to decrease at an exponential rate in order to sustain the predictions of stochastic stability. The coradius formula therefore fails for two reasons in predicting the waiting time until convergence. First of all, the single mutation to reach $D(\{n+2\})$ requires a waiting time that increases exponentially with the size of the system. Second, overcoming the negative drift of the undisturbed dynamics between states 1 and n will require a waiting time that also increases exponentially in the size.

For the purpose of characterizing the long-run behavior of a dynamic system, stochastic stability analysis takes the wrong limit by fixing the size n of the system and letting $\epsilon \to 0$. While we typically think of the noise term ϵ as small, we also want it to be sufficiently bounded away from 0 such that the stochastic system does not get 'stuck' in intermediate limit sets in the medium run. At the same time we usually want our results to hold primarily for environments with many agents, due to the bounded rationality assumption buried in almost all evolutionary models. Agents behave myopically or use rules of thumb because their computational abilities are assumed to be limited. This simplification in the decision-making process is particularly compelling for models with local interaction, such as my residential neighborhood process, because the number of possible states increases exponentially in the size of the system.

Clustering describes the long-run and medium-run behavior of a stochastic process more adequately by taking the 'correct' limit $n \to \infty$. The perturbation ϵ is kept fixed such that clustering takes both the disturbed and the undisturbed dynamics of the process into account. In my simple example it can be easily checked that the process clusters around any δ-neighborhood of state 0. Moreover, the process reaches the neighborhood quickly as the waiting time satisfies $W(n, [0, \delta n]) \sim n$. Clustering therefore completely reverses the predictions of the standard analysis.

is at least $\bar{\epsilon}_n^{-1}$) and the waiting time to reach the δ-neighborhood (which is of the order 2^n as the ratio of the probability for a downward-jump and the probability of an upward jump is 2 under the undisturbed dynamics).

2.6.1 Example I: Formation of Black Ghettos on Streets

I next demonstrate that stochastic stability describes the evolution of the residential neighborhood process very poorly on streets, which vindicates my choice of clustering over the standard technique for the analysis of the model.

The only limit sets of the process on streets are the all-white and all-black configuration. It is straightforward to determine the radius and coradius of the black ghetto configuration η_b. The process will leave the basin of attraction of the ghetto only once a minimally stable white cluster of length $w = [2r(1-\alpha_w)]^+$ has formed, i.e. at least w completely tolerant white house-seekers have settled on the street. Therefore, one can deduce that $R(\{\eta_b\}) = w$. Similarly, an all-white neighborhood can turn into a black ghetto with positive probability only in the presence of a minimally stable black cluster of length $b = [2r(1-\alpha_b)]^+$ which tells us that $CR(\{\eta_b\}) = b$.

Because blacks are assumed to be more tolerant than whites, we know already that $b \leq w$. If blacks are sufficiently more tolerant and/or the radius of interaction r is sufficiently large the inequality becomes strict. In this case the black ghetto is stochastically stable according to theorem 5 and will be reached quickly in the medium run because the waiting time is $O\left(\epsilon^{-b}\right)$ on a street of fixed size n.[60]

But these conclusions contradict the findings of section 2.3 where I showed that the process can cluster both around the all-black *and* the all-white configuration depending on the ethnic composition of the housing market. Intuitively, stochastic stability fails for exactly the same reasons as in the simple random walk example I considered previously. The radius/ coradius reasoning focuses on a 'sideshow' of the dynamics by looking at the emergence of minimally stable clusters of minority residents. Unless the share ϵ of tolerant agents is unrealistically small, such clusters will always arise quickly. For large and even moderately large streets, however, all the 'action' comes from the undisturbed dynamics which governs the evolution of the street *after* minimally stable clusters have formed. The ethnic composition of the housing market then determines whether white or black clusters can expand with positive drift. Clustering captures this effect, while stochastic stability does not.

[60] The bound on the waiting time does not depend on the size of the system which is interpreted as 'fast' convergence.

2.6.2 Example II: Imitation and Coordination

The following simple model of imitation between communicating agents on a graph provides a further illustration for the weak predictive power of stochastic stability when applied to large-scale systems. It is also of interest in its own right because the result holds for a wide class of non-regular graphs.

I consider n agents who live on some connected graph of order q.[61] I refer to this class of graphs as 'proper' graphs. This can be a street, an inner-city area or some more complicated structure such as a simple street with an even number of agents where each agent has an additional randomly drawn third neighbor. Agents choose to hold exactly one of two possible 'opinions' which I denote 0 and 1. Time is continuous, and all agents revise their opinion each time their Poisson 'alarm clock' goes off at rate 1. With (small) probability ϵ they listen to an exogenous signal telling them to change their opinion. Otherwise, they sample one of their q neighbors and imitate her action.

Stochastic stability suggests that society should hold unanimous opinions most of the time. It takes one mutation to leave the unanimous configurations η_0 and η_1 where all agents hold either opinion 0 or opinion 1. On the other hand, a path leading to an unanimous configuration has cost 0. Therefore, the radius $R(\{\eta_0, \eta_1\})$ exceeds the coradius $CR(\{\eta_0, \eta_1\})$ and theorem 5 applies.

However, Monte Carlo simulations do not confirm this prediction, but suggest, to the contrary, that typical societies are perfectly 'confused'. For the numerical analysis I call a society 'unanimous' if the long-run share $\tilde{X}_n$ of agents with opinion 1 lies either in the interval $[0, 0.1]$ or the interval $[0.9, 1]$. Society is called 'confused' if the share $\tilde{X}_n$ lies in the interval $[0.45, 0.55]$, i.e. society is almost evenly divided into two camps. Table 2.5 compares the respective probability of society being unanimous or confused for streets $G_S^{r=2}(n)$ of varying size n and exogenous signals ϵ of different strength. Society is well described as unanimous only if both the size of the graph *and* the disturbance term ϵ are small. As the size of society increases agents hold very rarely the same opinions and society becomes more and more confused. 'Confusion' will be more pronounced if agents are more likely to listen to the exogenous signal, but even if this event occurs very rarely ($\epsilon = 0.01$) society will be confused at least 60 percent of the time for $n \geq 1000$.

[61] A graph has order q if q edges meet at each node.

Table 2.5: Comparing the probability p_U that society is 'unanimous' with the probability p_C that society is 'confused' when agents imitate the opinions of their neighbors. Agents live on a street $G_S^{r=2}(n)$ of varying size n and listen to exogenous signals ϵ of different strength.

$\epsilon = 0.1$

	$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 1000$
p_U	0.326	0.002	0.000	0.000	0.000
p_C	0.104	0.275	0.423	0.552	0.899

$\epsilon = 0.01$

	$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 1000$
p_U	0.836	0.168	0.030	0.001	0.000
p_C	0.022	0.131	0.216	0.300	0.602

Estimated standard errors are 0.001 or less for all estimates. Society is called 'unanimous' if $\tilde{X}_n$ lies in the interval $[0, 0.1]$ or $[0.9, 1]$, and is called 'confused' if $\tilde{X}_n$ lies in the interval $[0.45, 0.55]$.

The failure of stochastic stability analysis can be again traced back to its inability to take into account the intermediate dynamics, and its overemphasis on the noisy dynamics close to the two unanimous configurations. This can be most clearly seen by considering the special case of a complete graph of size n where each agent has $n - 1$ neighbors. For large n the process is well described by the deterministic approximation of the change in the share of agents $x(t)$ holding opinion 1:

$$\begin{aligned}
\frac{dx}{dt} &= \text{share agents with opinion } 0 \times \text{Prob. of switching from 0 to 1} \\
&- \text{share agents with opinion } 1 \times \text{Prob. of switching from 1 to 0} \\
&= (1 - x)\left[\epsilon + (1 - \epsilon)x\right] - x\left[\epsilon + (1 - \epsilon)(1 - x)\right] \\
&= \epsilon(1 - 2x) \qquad (2.10)
\end{aligned}$$

This differential equation has a unique stable steady state at $x^* = \frac{1}{2}$, which suggests that the imitation process clusters around x^* in the case of complete graphs. Note that the imitation effect cancels out because it is linear: the probability of changing one's mind is proportional to the number of neighbors with different opinion. The exogenous signal then pushes the process towards x^*.

The next theorem shows that this observation extends to *any* increasing sequence of proper graphs and that the process converges fast.[62]

Theorem 6 *Consider the imitation process on some sequence $G(n)$ of proper graphs. The share of agents holding opinion* 1 *clusters around any neighborhood of $x^* = \frac{1}{2}$. The process reaches some δ-neighborhood of x^* after a waiting time of $W(n, [x^* - \delta, x^* + \delta]) = O(1)$.*

Proof: see appendix A.9

If we believe that typical societies are large we should indeed observe them to be 'confused' most of the time. The numerical results in table 2.5 reassure us that societies do not have to be unduly large for theorem 6 to hold.

2.6.3 Example III: Revisiting Ellison's (1993) Model of Local Interaction and Coordination

Finally, I demonstrate how clustering can enhance our understanding of existing evolutionary models whose dynamics was characterized through stochastic stability analysis. Ellison (1993) examined how agents in large populations learn to play a 2 × 2 coordination game, shown in figure 2-15. Strategy A is assumed to be risk-dominant. Agents live on a street with radius of interaction r.[63] Time is discrete, and in each period t agents play with probability $1 - 2\epsilon$ the best response to the average play of their neighbors at period $t - 1$. With probability 2ϵ they choose one of the two strategies A and B at random with 50-50 probability. This system has the two limit sets η_A and η_B with all agents playing strategy A and B, respectively.

Because of risk dominance, the best response of a player will be action A if at least some fraction $q^* < \frac{1}{2}$ of her neighbors play A. This implies that a cluster of at most $r + 1$ agents playing strategy A can expand contagiously under the undisturbed dynamics and take over the entire street. The limit set η_A has therefore a large basin of attraction, while the limit set η_B has a small one. The radius/ coradius reasoning then quickly establishes that all agents play A in the long run (Ellison 1999).

[62] Note, that theorem 2 and theorem 3 only hold for an increasing sequence of streets and inner-cities, respectively.

[63] Kandori, Mailath, and Rob (1993) look at the case of uniform interaction.

	A	B
A	a,a	c,d
B	d,c	b,b

Figure 2-15: Stage game with strategy A as the risk-dominant strategy: $(a-d) > (b-c)$

Stochastic stability predicts the same long-run behavior as clustering, as I will show shortly. Nevertheless, this prediction is not very robust because tiny changes in the dynamics of the model can make η_B the stochastically stable equilibrium. Consider, for example, the following modification: in each period an agent faces a different stage game with a very small probability $\sigma > 0$ where she receives an additional payoff g from playing B against a neighbor who also plays B. I assume that g is large enough, such that this agent would play B if at least one of her neighbors played B in the previous period. Intuitively, we would not expect this change to have a major effect on the behavior of the model. After all, a cluster of $r+1$ agents playing strategy A is still very likely to expand, even though the growth of the cluster is no longer contagious. But it still grows with a strong positive drift and the modified system very much resembles the residential neighborhood process on a street with the cluster of agents playing A corresponding to the minimally stable cluster of black residents and blacks dominating the housing market. The next theorem confirms this intuition.[64]

Theorem 7 *Consider Ellison's modified population coordination game on a street $G_S^r(n)$. There exists some critical value $0 < \tilde{\sigma} < 1$ such that the following holds for $\sigma < \tilde{\sigma}$.*

1. *Most agents play strategy A in the long run, i.e. the share of agents who play strategy A clusters on the interval $[x_A^*(\epsilon), 1]$ with $\lim_{\epsilon \to 0} x_A^*(\epsilon) = 1$.*
2. *For ϵ sufficiently small the waiting time until the share of agents playing strategy A exceeds $1-\delta$ satisfies $W(n, [1-\delta, 1]) = O(1)$.*

[64] The second part of the theorem contains theorem 3 in Ellison (1993) as a special case.

Proof: The proof is exactly analogous to the proofs of theorem 2 and lemma 1 in section 2.3.2 and is therefore omitted. Action A (B) corresponds to a resident being black (white) and the parameter σ plays the role of the balance in the housing market.

Note, that the theorem also holds for Ellison's original setup which corresponds to the special case $\sigma = 0$. Hence the small modification of the dynamics has no discontinuous effect on the evolution of the model. However, it has a dramatic effect on the basins of attraction of the two limit sets η_A and η_B. In particular, the size of the basin $D(\eta_A)$ is much smaller as the process can escape from configuration η_A as soon as two neighbors play strategy B. Therefore, η_B has a coradius $CR(\eta_B) = 2$ while the radius is still $R(\eta_B) = [q^*2r] + 1$. If the radius of interaction is sufficiently large, the configuration η_B, rather than η_A, is stochastically stable according to theorem 5.

The sensitivity to innocuous changes in the dynamics of a model is a worrisome feature of stochastic stability. The problem arises, because the standard technique ignores the nature of the dynamics *after* the process has left the basins of attraction of the limit sets η_A and η_B. This part of the dynamics is not significantly influenced by small changes in σ, which explains why the predictions of clustering remain unaffected.

2.7 Conclusion

This chapter outlines a new theory to understand the rise and the persistence of ghettos in US cities. I build a simple evolutionary model which is completely described by the geometry of the residential area, the tolerance level of both ethnic groups and the balance in the housing market. I analyze in what way these parameters interact to lead to rapid segregation and found that the balance in the housing market is the determining factor. Furthermore, I prove that black ghettos can be very persistent in large inner-city areas.

My model can be viewed as a version of Schelling's (1972) tipping model with a richer non-uniform geometry of interaction. Exploring the implications of a local interaction set-up has been the domain of game theorists rather than applied researchers.[65] This is regrettable for two reasons. First, most social networks are local in the sense that the vast majority of

[65] Important exceptions are the work by Glaeser, Sacerdote, and Scheinkman (1996) on crime and social interaction, and chapter 3 of this thesis on competition in the telephone industry around the turn of the century.

agents interact with and care about only a small subset of the population. Local networks are therefore a far more natural modeling environment than the uniform geometry. Second, many surprising and empirically significant effects arise from the local interaction setup. One of the first insights of this kind was Ellison's (1993) observation that local interaction can hugely speed up the convergence to the long-run equilibrium. In the context of my model this effect manifests itself in the different mechanisms that uphold a black ghetto on a street compared to a bounded neighborhood. In the uniform geometry a black ghetto is persistent because white residents feel isolated and refuse to enter the residential area. On streets, on the other hand, the ghetto is upheld because blacks dominate the housing market. "Avenue waves" such as observed in Chicago are only possible within the local interaction setting.

I hope that the new techniques developed in this paper will facilitate the analysis of models with local interaction. The standard radius/ coradius reasoning is, in some sense, too successful in simplifying the analysis of a dynamic model. It ignores a great deal of information about the intermediate dynamics and overemphasizes the dynamics around the limit sets. For large-scale systems this imbalance can lead to poor predictions of the medium- and long-run behavior of a process. In these cases clustering can prove to be a safer and more robust tool to understand the evolution of the system.

Chapter 3

Death through Success: The Rise and Fall of Independent Telephony at the Turn of the Century

3.1 Introduction

The dominant position of AT&T in the American telecommunication industry until its break-up in the early 1980's masks the fact, that the company had once faced a formidable challenge to its monopoly at the turn of the century. After the loss of its patent protection in 1893 thousands of small independent telephone companies formed and within a decade almost half of all telephones in the US were operated by the Independents. They served regional rural and urban markets and competed fiercely with AT&T for subscribers. The Independents interconnected amongst each other but not with their Bell rivals which gave rise to widespread dual service competition, particularly in the cities: urban subscribers who wanted to communicate with all users in a city had to buy service from both AT&T and its independent competitor. Between 1905 and 1907 the independent movement peaked after years of double-digit growth and entered an equally rapid decline in subsequent years. By the mid-1920s dual service competition had been eliminated and AT&T was again controlling the entire US telephone network either directly or indirectly through dependent regional sublicencees.

This paper explores the reasons for the initial rise and subsequent decline of the independent movement. My argument rests crucially on the observation, that urban markets subdivide into social 'islands' along geographical and socio-economic dimensions. Communication between agents in the urban social network is characterized by local interaction because users are more likely to communicate with subscribers 'inside' their island than with those 'outside' it. In a simple evolutionary model I then demonstrate how the initial low state of development allows new firms to enter islands of users which are not yet being served by AT&T. Competition is welcomed initially by residents and businesses as it leads to a marked decrease in rates and because the amount of telephone traffic between distinct islands is small. These conditions allow minority networks to thrive and preserve their market share at the early stages of technology diffusion.

As telephone ownership becomes more common, however, the lack of interconnection imposes an increasing cost on business subscribers in particular who eventually have to subscribe to both competing networks in order to talk to outside customers. This rising *cost of incompatibility* erodes the public support for 'wasteful' competition and leads to calls for consolidation in the industry and compulsory interconnection of the rival networks. But I can identify a second and more subtle threat to the long-term survival of the minority system. Duplication is asymmetric in the sense that business subscribers to the smaller network are more likely to get a second phone. This effect opens minority islands to the majority network because new subscribers in those islands are now more likely to subscribe to the larger system. Once a minority cluster has been 'invaded' the majority system will expand and ultimately take over the entire island. When the state of development is sufficiently advanced standardization therefore arises endogenously in my model due to this *duplication effect*.

The existing literature develops two polar interpretations of the rise and decline of Independents based on either global network effects or predatory strategies pursued by AT&T. Neither explanations can explain on its own the curious dynamics of telephone competition.

Early historians of the telephone industry such as Anderson (1907), Stehman (1925) and more recently Brooks (1975) tended to adopt the view of AT&T, that telephony is a natural monopoly. According to this interpretation the telephone business is all about connecting any user with everybody else. Hence two unconnected systems are a major nuisance and will eventually lead to consolidation. While this story can explain the decline of dual service competition it cannot account for its initial success. For global network effects to apply the state of development is irrelevant. Some simple predictions of the natural monopoly view are also at odds with the data. The number of calls per residential telephone for example did not seem to increase dramatically in cities as the telephone diffused in contrast to what we would expect from a model with global network effects.

Recent research by Gabel (1969), Lipartito (1989) and Weiman and Levin (1994) has taken a less benign view of AT&T. These authors claim that AT&T regained monopoly power through a conscientious campaign, a combination of preemptive investment into long-distance telephony, rate wars and pricing below average cost, strategic acquisitions, sub-licensing and regulatory capture.

While AT&T presumably had predatory intentions, proponents of this view tend to apply it too indiscriminately. Most of the evidence comes from the territory of the Southern Bell Telephone Company, where Independents were weak and telephone development was low. There the Bell company managed to divide the independent movement through an early policy of sub-licensing independent exchanges in non-competed territories. I believe that the story in this paper can complement our understanding of the decline of competition in the rest of the country, especially the highly developed North Central states such as Indiana, Illinois and Ohio, where the predatory argument is less convincing. There a substantial number of independent regional systems survived the sub-licensing era until 1913 and every third US city was still contested by rival, incompatible networks. Nevertheless, telephone subscribers and regulators pressed for further consolidation in the industry and dual service competition disappeared largely until the end of World War I.

The contribution of this paper is to revive the role of mass market forces in the selection of a telephone standard. Furthermore, my story can explain the entire dynamics of network competition and its long-run tendency towards standardization.

The chapter is organized as follows. In section 3.2 I compare the evolution of the AT&T and the independent networks at the turn of the century. I discuss the emergence and subsequent decline of dual service competition in the cities. Section 3.3 draws on various sources to illustrate the local structure of the urban social network. Residential subscribers made most of their calls to a limited number of 'friends' and local stores while businesses in return served particular social islands within a city. Taking these communication patterns into account I develop a model based on local interaction in section 3.4. The diffusion of telephone technology will increase the cost of incompatibility and lead to an increase in the duplication rate of businesses. Moreover, the duplication effect can directly precipitate the decline of the minority system. I also show how interconnection between the carriers could have preserved dual service competition by making duplication unnecessary. In section 3.5 I under-pin the implications of my model with empirical evidence.

3.2 The Growth of the AT&T and Independent Telephone Networks

The American Bell company had a monopoly in the US telephony market until 1893/94 when its principal patents expired.[1] Entry occurred almost immediately with 18 independent commercial systems being established in 1893 and 80 in 1894. By 1900 independent companies were founded at a rate of 500 a month and by the end of 1902 a total number of 3,113 Independents crowded the industry.

Competition led to a dramatic decrease in telephone rates of often more than 50% even in towns without competition.[2] The lower cost of telephone service accelerated the adoption of the technology dramatically. Between 1893 and 1902 the subscriber base increased tenfold. Table 3.1 illustrates the explosivc growth of both AT&T and independent telephone networks after 1894. The Independents, however, managed to gain market share at the expense of the AT&T subsidiaries and by 1907 they controlled half of all the telephones in the US.

The rival systems followed very different strategies for expanding their networks. Before 1894 AT&T concentrated on developing the business districts in the cities and larger towns. It had also begun to build a nascent long-distance network connecting those key cities in the 1880's.[3] Local exchanges were built expensively and of high quality in order to incorporate them later easily into the long-distance network. As a consequence rural areas and small towns were often left completely undeveloped. With the onset of competition AT&T started

[1] American Bell was a holding company and held controlling stakes in regional Bell sublicencees such as the Southern Bell Telephone Company (SBT). American Bell conducted its long-distance business through a separate subsidiary which was founded in 1885 as AT&T. In 1899 all assets of American Bell were transferred to AT&T.

[2] Weiman and Levin (1994) examine the case of Southern Bell and confirm that the company lowered their rates for exchanges where it anticipated entry.

[3] Langdale (1978) documents the growth of AT&T's long-distance network.

Table 3.1: Total growth of telephone system (dual service denotes percentage of cities with population of 5,000 and more which have competing systems).

	All systems	Bell system		Cities with
		Share	annual growth	dual service
1894	270,381	100%	1885-94: 6.26%	2%
1897	NA	NA	NA	23%
1902	2,371,044	55.6%	1895-1902: 22.06%	55%
1904	NA	NA	NA	60%
1907	6,118,578	51.2%	1903-1907: 18.11%	57%
1912	8,729,592	58.3%	1908-1912: 9.81%	37%

Sources: Telephone Census (1912), table 2; Mueller (1997), table 6-2; Warren Report (1938), table 33

to provide service in most urban areas and invested aggressively in its long-distance network. The expansion of its network under the slogan "One System, One Policy" proceeded even in regions where only a minority of users subscribed to Bell and toll lines had to be operated at a loss.

While AT&T pursued a 'top-down' approach to network building, the Independents did the reverse. The first wave of entry occurred in the rural areas and small towns with no telephone service. In a second wave around 1900 the bigger towns and cities were invaded and AT&T was challenged on its established territory. The increase of cities with dual service from 2% in 1894 to 60% a decade later is documented in table 3.1. The populous North Central states with their large rural populations became the hotbed of competition. In 1902 of the 3,113 commercial independent exchanges 1,694 were located in Illinois, Indiana, Ohio, Kansas, Michigan, Minnesota, Missouri, Iowa and Wisconsin.[4] The concentration in telephone ownership in the North Central states increased from slightly more than 4% of the population in 1902 to more than 10% in 1907 and 13% in 1912. Gabel (1969) noted, that growth during the competitive era was both intensive and extensive, leading to higher telephone saturation in cities and bringing telephone service to rural and suburban areas.

Due to the lack of a coordinating strong center the Independents could not match AT&T's investment in long-distance lines. Independents had to establish exchanges in a critical number of town and cities within a region before they could start to invest heavily in toll lines. Most users, however, had no demand for ultra-long distance calls. Between

[4] Telephone Census (1902), table 10

1902 and 1912 at least 97% of all calls were between points in the same city.[5] About 90% of long-distance calls originated and ended within a 50 miles radius.[6] Interstate calls were still too expensive for most subscribers and the telegraph provided a cheaper alternative for sending ultra-long distance messages.[7]

Independent exchanges could therefore survive by forming regional systems with a an interconnecting network of toll lines that could satisfy most of their subscribers' demand for long-haul connections. This process started by 1906 in those parts of the country where Independents were most strongly represented such as the North Central states. Mergers gave rise to independent regional cperators which owned exchanges in typically ten to thirty key cities and had exclusive connecting contracts with adjoint Independents. In Indiana, Ohio and Michigan the independent movement even managed to establish dedicated long-distance operators.[8]

Both AT&T and the Independents refused to interconnect their networks at the beginning of the competitive era. AT&T did not want to accommodate their competitors in contested cities and allow the Independents to free-ride on its heavy investments in long-distance lines. The company also feared that interconnection would allow dual service competition to continue indefinitely which was incompatible with its principle of "One System, One Policy". The Independents on the other hand believed that interconnection would slow down the growth of their own networks and prevent them from building a true nation-wide alternative to AT&T's system. Exclusive control of rural exchanges and many small towns gave them a powerful negotiating tool in order to secure franchises in the bigger cities. Moreover, interconnection would have halted the development of their own nascent long-distance network. The New Long Distance Company of Indiana for example prohibited its member exchanges from connecting to different long-distance carriers for 99 years (MacMeal 1934, p. 174).

[5] Telephone Census (1912), table 19

[6] In Chicago 90% of all messages in and out of Chicago were transmitted within a radius of 100 miles (Committee on Gas, Oil and Electric Light 1907, p. 120). In Indiana 89% of all messages originating or terminating at a population center stayed within a radius of 35 miles (*The American Telephone Journal* July 5, 1902: p. 15).

[7] In 1902 a three minute station to station call from New York to Philadelphia cost $0.55 and to Chicago $5.45 (about 5% of the *yearly* flat business rate in Chicago) according to the Bureau of the Census (1975, p. 784). A 10 word telegraph message on the other hand cost just $0.25 and $0.50 respectively in 1908 (p. 790). The invention of the electronic repeater made long-distance telephony more affordable after 1915.

[8] The Independents managed to handle 28% of all US long-distance traffic in 1907 and presumably a much higher share in the North Central states.

AT&T modified its strategy after 1900 as the independent movement continued to prosper despite the lack of interconnection. In order to prevent the formation of regional independent systems AT&T offered interconnection to independent companies in *non-competed* territories which controlled 55% to 60% of all independent telephones. Those so called sublicencees enjoyed a local monopoly but were not allowed to interconnect with any independent exchange. Sub-licensing therefore fragmented the independent movement and made it more difficult for the Independents to build viable regional systems with a satisfactory toll network. The Southern Bell Telephone company pioneered sub-license arrangements between 1900 and 1903, which effectively halted the independent movement in the South at an early state of development (Weiman and Levin 1994). In 1907 AT&T extended the sub-license policy to the rest of the country and by 1913 87% of the non-competing independent telephones were interconnecting with AT&T. Sub-licensing isolated many independent operators in dual service markets and contributed to rapid consolidation of competing exchanges between 1907 and 1913. In those cases the independent company was either acquired by its Bell competitor or, alternatively, bought the Bell exchange and became a sub-licensee if it had a commanding lead.

Unlike in the South, however, AT&T did not manage to stamp out dual service in the rest of the country, especially the North Central states where the independent movement was strongest. In 1913, 37% of cities with populations over 5,000 still had dual exchanges down from 59% at the beginning of 1907, and 34% of all independent telephones were still unconnected to the Bell system (75% of them in competed territories) (Mueller 1997, p. 110-12). Many regional systems rooted in larger cities had managed to survive thanks to their extensive toll networks.

AT&T's heavy-handed attempt to unify the telephone system was meanwhile coming under increasing attack. In the face of anti-trust legislation the company was forced to back away from its pursuit of universal service and announced the "Kingsbury Commitment" in 1913. Under this arrangement AT&T opened up its long-distance lines to independent exchanges for calls outside a 50 miles radius. Furthermore, the company agreed to stop acquiring competing independent exchanges in the more than 1,200 cities and towns with dual service. This moratorium on further acquisitions was the most significant part of agreement because it left intact the remaining regional independent operators in direct competition with Bell. The primary aim of the arrangement was therefore to preserve dual service competition on the exchange level.

To summarize by 1913 about 13% of the total number of communities and more than a third of all cities with more than 5,000 people had dual service competition, and the regulatory tools were now in place to prevent further monopolization of the telephone industry. The Kingsbury commitment in deed seemed to stabilize the number of non-connected independent telephones between 1913 and 1917. But these statistics mask the fact that the industry continued to converge towards standardization with dual service territories being replaced by regulated regional monopolies.

Consolidation proceeded after the Kingsbury Commitment by swapping territories instead of acquisition. City Councils, federal and state authorities had to agree to those transactions in order to waive the Kingsbury commitment, and in many cases voters expressed their desire to unify the service through a referendum. Dual service was therefore gradually eliminated with the implicit or explicit permission of telephone users and gave rise to regional independent monopolies. During the late period of telephone competition the pressure to abolish user fragmentation through consolidation was coming from the demand side rather than predatory policies of AT&T. The history of the late competitive era suggests, that the replacement of dual service markets by regional independent and Bell monopolies was an inevitable process given the lack of interconnection in dual service territories. AT&T's abuse of market power in the 1900's therefore secured its dominant position in the post-competitive era but did not in itself bring to an end competition in local telephony.

In the following I will formally explore the relationship between the state of telephone development and the feasibility of having competing and non-connecting telephone networks in a single city. In my model the local nature of telephone communication makes competition an acceptable option to most subscribers when the concentration of telephone ownership is low. When the telephone market becomes saturated, however, user fragmentation will become more costly to the average business user. Furthermore, an increase in telephone duplication by businesses can push the minority system towards extinction.

3.3 The Local Nature of Telephone Communication

Before we can model competition between rival systems we first have to understand the structure of telephone communication. If we think of the various subscribers as nodes in a *social network* we want to understand in other words the geometry of this network.[9] I will distinguish two basic types of telephone subscribers - businesses and residential customers. In the following I will describe in a stylized fashion the way in which both groups communicated with each other. This will allow me to build a generic social network in the next section that incorporates both types of users.

From the sparse historical statistics on residential communication patterns we can conclude that the bulk of telephone conversations were exchanged with various businesses (about 25%) and a small number of 'friends' (about 30% to 50%). In 1909 a Bell manager listened in on a sample of conversations at a residential Seattle exchange. He found, that 20% of calls were orders to stores and other businesses, 20% were made from homes to family-owned businesses, 15% were social invitations and 30% were "pure idle gossip" (Fischer 1992, p. 79).[10] AT&T research has shown that nowadays about half of all calls from a given residence go to just five numbers while about 75% of all messages are social calls to friends and family.[11] Given that today most social calls are sent to a very limited circle of friends we should expect the same to hold for the roughly 50% of social calls recorded in the Seattle data.

The importance of a small number of social neighbors can help to explain the curious fact, that during the competitive era the number of calls per telephone remained virtually unchanged. This phenomenon can be confirmed both for state statistics as well as city data.[12] If subscribers would really care about the total number of people they can commu-

[9] Various studies such as Kern (1983) and Meyerowitz (1985) claim that the telephone itself changed the social network by weakening local ties in favor of extra-local ones and by eradicating the concepts of space and distance. In the most comprehensive study on this topic, however, Fischer (1992) finds that the telephone did not open any new social contacts. On the contrary telephony was an anti-modern technology in the sense that it reinforced existing social relationships.

[10] Telephone companies initially fought their residential customers over social calls, which they saw as an unnecessary burden on their networks. Only after World War I did AT&T discover sociability as a selling point for their service.

[11] When an accident in 1975 knocked out thousands of telephones in New York City 70% of all users declared in a subsequent survey, that they had missed most of all social calls. In a 1985 poll Californians estimated that about 75% of all their calls are for social purposes (Fischer 1992, p. 226).

[12] Between 1902 and 1912 the number of calls per telephone in the US and the North Central states was at about seven calls per day (Bureau of the Census 1912). In 1907 the number of calls per measured rate telephone in the city of Chicago was also seven (Committee on Gas, Oil and Electric Light 1907).

nicate with, we would expect a steady increase in the number of calls per telephone. In the telephone census of 1902 (p. 30) we are offered the following explanation:

> Experience shows that this does not happen, because of what is termed the "acquaintance factor". In every community each individual is acquainted with and transacts business among a certain limited group; and while such circles of acquaintances overlap and the business increases more rapidly than is indicated by a simple arithmetical ratio, it does not increase quite as fast as the square of the population.

Social bonds are not randomly distributed within the population of a given city. Subscribers are differentiated along several dimensions such as social status, place of residence, ethnicity and the establishment they work at. MIT students are more likely to be acquainted with fellow MIT students rather than Boston University students. Faculty members are more likely to call fellow faculty rather than undergraduates. Subscribers therefore fall into one or several 'social islands' such as the island of MIT undergraduates for example. The 'friends' of a subscriber in the social network are much more likely to be found within the same social island rather than outside it. Given the importance of social calls we would expect that residential subscribers within the same island tend to coordinate their network choice.[13] Brooks (1975, p. 110), for example, reports:

> In Minneapolis, for example - according to the recollection of a survivor of the competitive era there - the Bell exchange, being the longer - established, was the exchange of the socially elite, while the competing Tri-State Telephone Company was for just about everybody else.

In the next step I determine the position of business subscribers in the social network of a city. I already remarked that an estimated 25% of all residential messages went to stores and other business establishments.[14] During the early years of telephony business and industry were in fact the primary users of the new technology. The rapid diffusion of

[13] Even at the beginning of the century the telephone was already used by a substantial share of people of each social class. Fischer (1992) finds that in the three Californian cities of Palo Alto, Antioch and San Rafael about 50% of managers and professionals had telephones in 1910, but also 20% of all commercial white collar households and 10% of blue collar workers.

[14] Fischer (1992, p. 361) suggests an estimate for the share of functional calls from residential telephones of between 20% and 60%.

Table 3.2: Telephone distribution and duplication rates in Louisville, 1910

Size of Business	Both Phones	Home Only	Bell Only	Duplication Rate (%)	Subscriber Rate (%)
Large-scale					
Telegraph	4	0	0	100	100
Gas and Electric	4	0	0	100	100
Fast Freight	11	1	0	92	100
Railroads	21	2	2	87	100
Banks	25	2	2	86	100
Hotels	21	6	0	78	100
Sub-total	86	11	4		
Medium-scale					
Druggists	83	69	3	53	100
Coal dealers	46	42	9	47	100
Insurance	65	46	36	44	NA
Dentists	35	44	3	42	63
Liquor dealers	43	56	18	37	NA
Plumbers	25	45	1	35	74
Attorneys	85	109	90	30	78
Butchers	19	47	7	26	NA
Dry goods	15	36	6	26	21
Groceries	182	466	62	25	NA
Sub-total	598	970	235		
Small-scale					
Billiard/ Bowling	2	10	0	16	NA
Carpenters	11	55	9	14	50
Barber shops	1	6	1	12	NA
Bakers	9	61	9	11	39
Saloons	64	487	19	11	87
Tailors	8	60	9	10	NA
Churches	3	12	14	10	NA
Residences	900	5,449	3, 971	9	20
Sub-total	998	6,140	4,032		

Source: Mueller (1997, Table 7-1). 'Home' refers to the independent Home Telephone Company while 'Bell' denotes the Cumberland Telephone Company, a Bell licensee.

telephony during the competitive era, however, shifted the balance soon towards residential users. By 1907 the share of residential telephones[15] had already risen to about 50% in the cities and by 1920 the proportion was close to 70%.[16]

How did businesses use the telephone? We can gain valuable insights from a breakdown of the telephone subscribers in Louisville in 1910 which is given in table 3.2. There the Cumberland Telephone Company (Bell) was competing against the independent Home Telephone Company.

Business users demand connections to their residential customers and to their supplier. These upstream businesses are usually large-scale establishments which had a duplication rate of close to 100%. They were generally the first ones to order a second telephone, but made up just 1.5% of all telephone subscribers in the city. Due to the high degree of duplication amongst suppliers medium-scale and small-scale businesses presumably made their network choice dependent on the choices of their residential customers rather than their upstream suppliers.

Several features of the Louisville survey support this hypothesis. The simple presence of single-phone grocers for example connecting to either the independent or the Bell exchange indicates, that some grocers mainly served subscribers to one of the exchanges. While most single-phone businesses favored the independent company by a ratio of at least 5:1, the Bell company was almost even amongst insurance agents and attorneys. The longer established Bell exchange was likely to serve more upper-class citizens who were amongst the first telephone subscribers in a city before the onset of competition. These observations lead to the conclusion, that business users mainly served customers within a particular social island.[17] Their choice of telephone network therefore depended to a greater extent on the choices of residents within those islands rather than on the choices of subscribers elsewhere. The high degree of duplication amongst medium-scale businesses, however, suggests that some business users either served customers of several social clusters with different competing

[15] The share of business telephones overstates the share of business subscribers. Large businesses often had private branch exchanges and extension telephones. In Chicago for example the share of business telephones in 1909 was 52%. If I correct for private branch telephones the share drops to 38% (*Electrical World* 1909: p. 1072,1504).

[16] In Columbus the Citizens Company reported a 55% share of business telephones in 1900 and only 39% in 1907 (Johnston 1908, p. 8). In Indianapolis (Ind) 39% of subscribers were business users in 1904 but only 31% ten years later (*The American Telephone Journal* 1904: p. 309).

[17] Fischer (1992, p. 180) reports further evidence in his study of three Californian towns. Florists catering to the middle class had phones whereas corner stores serving working class neighborhoods did not have one due to the lower subscriber rate amongst blue-collar workers.

networks or that they cared sufficiently about customers outside their main customer base and ordered a second phone.

I have presented evidence for the local character of the social network that underlies the demand for telephone conversations. Independent entrants had to exploit this structure in order to gain an initial foothold in a town. In the light of the stylized facts developed above we would expect that network building should have first targeted a group of close neighbors within a social island and the associated businesses serving that island. The following account from a publication of the New England Telephone and Telegraph Company (1908, p. 11) describes the process of soliciting the initial group of subscribers for a new exchange exactly according to that mechanism.[18]

> The origin of the Independent movement is somewhat typical ... The subscribers to the Bell system are not solicited to subscribe to the new system, at least not at first. A man who has no telephone is asked if he wouldn't like one ... [He] desires to know how many subscribers he can communicate with ... he need not bind himself except conditionally upon a certain number of subscribers being secured. Are there any of his friends with whom he or his family would especially desire to communicate? If names are given, these people are seen and assured that Mr. A, the first one visited, is going to put in a telephone. It is not difficult in this way to secure a number of signatures to conditional contracts ... The next step is with this list of names to go to some local dealer ... The canny canvasser has ascertained from those whom he has previously visited ... where they do their trading, who is their butcher, baker or grocery man. The dealer sees the possibility of increased orders ... Thus the signatures are secured.

Entry into the telephone market was considered to be easy at the onset of competition when telephone concentration was low and many social islands were yet undeveloped. Theodore Gary, president of the Interstate Independent Association, for example writes in 1907:

> There is a place for two telephone companies in about every community of 10,000 population and up, occupied by a Bell company, because it generally fails to

[18] A surprisingly small number of subscribers was necessary to establish the first exchange. The manager of the Eureka, California, exchange for example reported that "my hardest work was getting my first twelve subscribers", which was regarded as the minimum number for an exchange (Fischer 1992, p. 64). It is noteworthy that telephone companies used solicitors mainly during the start-up phase. As soon as demand outstripped supply solicitors were laid off.

> develop a territory. Competition is desirable unless there is a development by one company of one telephone to five or six persons in the community. With full development there is very little room or demand for competition.[19]

3.4 A Model for Telephone Competition

I will now introduce a formal model that takes the communication patterns of business and residential users of an urban area into account. I will start with a basic setup that allows me to discuss the impact of network growth on the cost of incompatibility. I then add some 'noise' to the dynamics which will give rise to the duplication effect. Finally I demonstrate that interconnection between competing systems could have preserved dual service competition indefinitely by avoiding duplication.

3.4.1 The Basic Setup

I consider a city that consists of an infinite number of social islands.[20] Each island is made up of n residents and m types of businesses such as grocery, butcher, physician etc. Every resident has a number of 'friends' within her social island with whom she communicates. Specifically, I assume that residents are located along a circle such that each of them is friends with her two direct neighbors.[21] Every resident is a 'core customer' for each type of business serving that island. Figure 3-1 shows the resulting network linking residents and businesses of a social island.

Time is discrete and each resident exchanges a total number of N messages per time period through the social network. Communication between the various agents is assumed to be symmetric, i.e. any two sets of agents send and receive an equal amount of messages between each other. I assume for simplicity that the demand for communication is inelastic with respect to the cost of communication: technological innovations such as the telephone will therefore reduce the cost of communication but will not increase demand for it. A share α of those N messages are exchanged with friends and a share β_j with type j businesses such that $\beta = \sum_{j=1}^{m} \beta_j = 1 - \alpha$. The businesses are indexed in descending order of importance,

[19] Theodore Gary in *Telephony* December, 1907: p. 338, quoted by Bornholz and Evans (1983, p. 19)

[20] Assuming an infinity of islands is analytically convenient as it makes the model non-stochastic at the aggregate level.

[21] The particular graph mapped out by these bonds can be more general and even random. My results would still go through qualitatively.

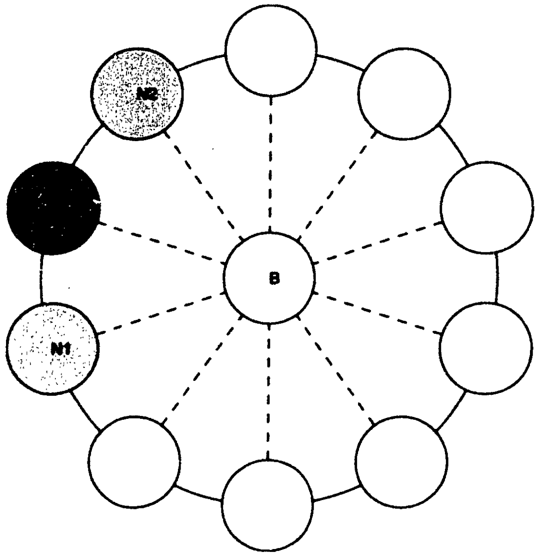


Figure 3-1: Social island of size 10 - resident R is 'friends' with residents N1 and N2. Business users B serve the entire island.

i.e.

$$\beta_1 > \beta_2 > ... > \beta_m.$$

We might expect druggists for example to receive more messages than attorneys such that $\beta_{Druggist} > \beta_{Attorney}$. A resident who contacts a type j business will deal with probability $1 - \gamma$ with a local business and with probability γ with any of the outside businesses. Each type j business therefore exchanges $\beta_j nN$ messages with customers and with probability $1 - \gamma$ such a message is exchanged with a core customer.

I assume $0 < \gamma < \frac{1}{2}$, e.g. residents do most of their shopping at their local stores but occasionally frequent outside businesses. Blue-collar workers for example might visit upper-class stores if these are geographically conveniently located. A residents from one suburb might have a favorite physician who lives in a different part of the city. Clearly γ will depend on physical characteristics of the city. For large cities with a higher degree of suburbanization we would expect a lower γ than for a small town where different socio-economic classes live closely together.

Communication is costly for residents and businesses. A resident has to spend a total amount of time $T < 1$ for sending all her messages: she has to write a letter to a friend or walk to a store for making an order. Her total time endowment is 1 and she has an income of y in each time period, which she can spend on consumption C or telephone service. Two competing companies A and B offer access to their networks at a rate K per time period.[22] I assume that the subscription rates are fixed for all time periods which allows me to focus on the demand-side when I analyze the growth of both systems: in particular this precludes strategic interaction between the two telephone carriers. While the assumption might seem extreme, rate schedules were in fact changed infrequently and usually matched by the competitor. Telephone companies generally operated under a franchise from the city council in order to use public roads for their line construction. Such a franchise was only granted if the entrant offered competitive rates: the subscription rate K was usually calculated as the cost of laying and maintaining a telephone wire between a subscriber's place of residence and the central office of the exchange plus some reasonable rate of return on the investment.[23] Increases in the rates required the permission of the local authorities which was a time-consuming process.

If a resident sends a share z of her messages over the phone she only spends $(1-z)\,T$ of her time for the remaining messages. I assume that her per period utility function $U(C,t)$ in consumption and time has the usual properties and that the following condition holds:

$$U(y-2K,1) < U(y,1-T) < U(y-K,1) \tag{3.1}$$

The left inequality tells us that a resident would never duplicate as she could always do better by having no telephone at all.[24] The right inequality ensures that a resident will certainly subscribe to a network if everyone else has a telephone. We can therefore define a threshold m^r as

$$U(y,1-T) = U(y-K,1-T+m^rT).$$

[22] For simplicity I assume that the marginal cost of sending a message over the phone is 0. Most cities (especially the smaller ones) had in fact unlimited service, but it is straightforward to incorporate a message cost c in my model on top of the base rate K.

[23] The city council of Chicago for example launched its own investigation on telephone rates and service and forced the Bell monopolist to accept a lower schedule in 1907 (Committee on Gas, Oil and Electric Light 1907).

[24] The average cost of Bell residential service in 1905 was about 5% of a basic manufacturing worker's earnings. For this reason most residential users could not afford two telephones in their homes (Fischer 1992).

A resident will then enjoy a higher per period utility from subscribing to a network l ($l = A, B$) rather than not subscribing at all if $z_l > m^r$. In each time period an agent can take an action $a_r \in \{0, A, B\}$.

A business user on the other hand has to pay a cost v_b for every message to a customer which is not sent by phone. We can think of v_b as the cost of sending a telegram or a messenger. A business chooses an action $a_b \in \{0, A, B, AB\}$ in each time period, i.e. it either uses no phone, subscribes to network A (B) or chooses dual service. Assume that network l gives access to a share z_l of customers and that a business of type j sends $M_j = \frac{\beta_j nN}{2}$ messages per time period. The per period costs of the various strategies of a business can then be calculated as follows:

$$\begin{aligned}
\pi_0 &= M_j v_b \\
\pi_A &= M_j (1 - z_A) v_b + K \\
\pi_B &= M_j (1 - z_B) v_b + K \\
\pi_{AB} &= M_j (1 - z_A - z_B) v_b + 2K
\end{aligned}$$

It will be again convenient to introduce a threshold level $m_j^b = \frac{K}{M_j v_b}$. A business will then realize a per period gain by subscribing to company l ($l = A, B$) as soon as it can reach at least a share m_j^b of customers via its network.

In each time period residents and businesses choose the action which maximizes their utility and minimizes their communication costs for that period.[25] This *best-response* dynamics is widely used in the evolutionary game theory literature[26] and a reasonable behavioral assumption in my model. It is unlikely that telephone users pursued strategic goals when choosing telephone service because the costs of switching a carrier were very low.[27] They presumably compared the benefits of the different systems and then decided to subscribe to the network which suited their communication needs best.

At time $t = 0$ the two competing telephone companies enter the city by soliciting a number of initial subscribers. With probability $\mu < \frac{1}{2}$ a social island is occupied by

[25] I assume that agents stick to last period's strategy in the case of a tie. This convention will not influence the results.

[26] Examples include Kandori, Mailath, and Rob (1993) and Ellison (1993).

[27] Subscribers usually had to sign contracts on an annual basis. There were no one-time connection fees except of mileage fees in a few cities for residents who lived very far away from the central office (Committee on Gas, Oil and Electric Light 1907, p. 172-).

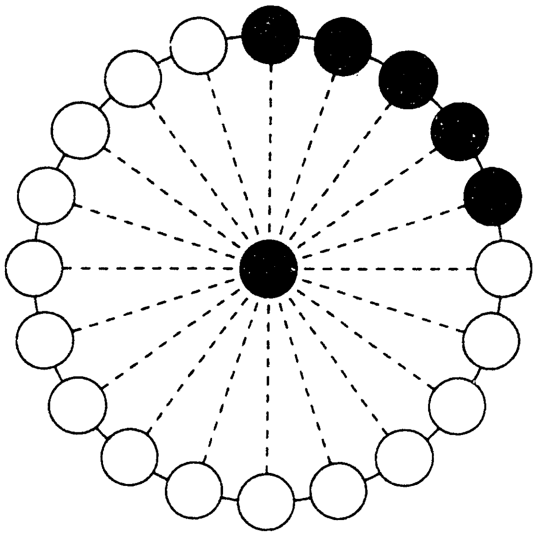


Figure 3-2: Social island of size $n = 20$ and an initial cluster of size $k = 5$.

company A (the minority company) and with probability $1-\mu$ by company B (the majority company). An entering company signs up a connected cluster of $k < n$ residents and all the businesses of the respective social island (see figure 3-2). The initial concentration of telephone ownership is therefore $x(0) = \frac{k}{n}$. After entry solicitors are laid off and the diffusion of telephone ownership follows the best-response dynamics.

Throughout I impose the following parameter restrictions:

$$\frac{\alpha}{2} + (1-\gamma)\beta \;>\; m^r > \beta \tag{3.2}$$

$$(1-\gamma)x(0) \;>\; m_j^b \tag{3.3}$$

$$\frac{\gamma}{2} \;>\; m_j^b \qquad \text{for } j \in \{1,..,m\} \tag{3.4}$$

Condition 3.2 assures that a resident will subscribe if the local business and at least one friend subscribe to the same company and will not subscribe if no friend uses the company. Therefore the initial cluster of subscribers will expand along its boundaries. This condition is quite intuitive - otherwise the initial cluster would either shrink instead of expand and the telephone technology would die out or all non-subscribers could buy telephone service at time $t = 1$ and the social space could fill instantaneously. By imposing condition 3.3 I make the entry story consistent and the entering company will be able to sign up the local businesses after having secured the conditional contracts from the initial cluster of subscribers.

The dynamics of this simple model is straightforward. Within each social island the initial group of subscribers will grow along the boundaries and the concentration of telephone ownership at time t will be $x(t) = \frac{k+2t}{n}$. Business subscribers to the minority company ('minority businesses') will eventually get dual service once the telephone concentration crosses the threshold

$$y_A^j = \frac{m_j^b}{(1-\mu)\gamma} \tag{3.5}$$

for a type j business. Because of condition 3.4 we know that $y_A^j < 1$ for all types of businesses such that minority businesses will surely duplicate once the state of development is sufficiently advanced. At any point in time denote the type of the marginal duplicating minority business with

$$j_A(t,\mu) = \max_{j\in\{1,..,m\}} y_A^j < x(t) \tag{3.6}$$

and set $j_A = 0$ if no business duplicates. Analogously, business subscribers to the majority company ('majority businesses') choose dual service once the telephone concentration surpasses

$$y_A^j = \frac{m_j^b}{\mu\gamma} \tag{3.7}$$

and one can define the type $j_B(t,\mu)$ of the marginal duplicating majority business accordingly. The duplication rate $w_A(t,\mu)$ of minority businesses therefore becomes $w_A(t,\mu) = \frac{j_A(t,\mu)}{m}$ and in the same way I define the duplication rate $w_B(t,\mu)$ of majority businesses. Clearly, the minority duplication rate is decreasing in the market share of the minority system while the reverse holds for the the majority duplication rate. Furthermore, the minority businesses will duplicate at least as much as majority businesses for all time periods t and any market share μ of the minority network:

$$w_A(t,\mu) \geq w_B(t,\mu) \tag{3.8}$$

Duplication, however, has no effect on the network choice of new subscribers: they will always choose the established system within their social island as it allows them to communicate both with the local business community *and* their friends. In the worst case for the minority system ($\mu = 0$) such a resident can send a share $\frac{\alpha}{2} + (1-\gamma)\beta$ of her messages over the minority system but only a share β over the majority network.

This simple model is already useful to analyze the *cost of incompatibility* for business users as they cannot reach those customers outside their social island who subscribe to the rival network. For each business type j I calculate the loss $C_j^b(t,\mu)$ to the *average* business subscriber under the assumption that interconnection would not result in a change in the subscription rate K:

$$\begin{aligned} C_j^b(t,\mu) &= (1-\mu)\min(K, \mu\gamma x(t) M_j v_b) \\ &+ \mu \min(K, (1-\mu)\gamma x(t) M_j v_b) \end{aligned} \tag{3.9}$$

The cost of incompatibility depends therefore on the state of development and the market share of the minority company. The next result shows that this cost is increasing in both variables.

Lemma 5 *The average loss to businesses $C_j^b(t,\mu)$ increases over time as the technology diffuses and with the market share μ of the minority company. If type j businesses have started to duplicate in a city the loss to the average type j business is at least $2\mu K$.*

Proof: see appendix B.1

In the following I apply lemma 5 in a back of the envelope calculation to find a lower bound on the cost of incompatibility for businesses in Louisville. From table 3.2 we know that the market share of the minority company (here the Bell system) amongst residents was about 40%. Assuming that each social island has a similar number of associated type j businesses I conclude that most duplicate users first subscribed to the minority company. Therefore the duplication rate for medium-scale minority businesses varied between 80% and 100%. From lemma 5 I therefore deduce that the average cost of competition to medium-scale businesses in Louisville was at least $2 \times 0.40 \times 0.8K = 0.64K$.

Unless business users in Louisville expected a 64% rate increase from a monopolistic provider they are likely to have rejected dual service competition in favor of consolidation in the industry. Such large increases were hard to push through, however, because telephone companies operated on a franchise from the city council. In Louisville both networks in fact consolidated in 1910, when the Bell subsidiary bought the Home Telephone Company. After the Kingsbury Commitment in 1913 rate hikes were even harder to achieve as consolidations could only proceed if all relevant user groups consented.

Despite the high cost of incompatibility in mature markets businesses are likely to have welcomed competition at low states of development when $C^b(t, \mu)$ was negligible. Therefore my cost of incompatibility argument can provide an explanation why business users would gradually turn against competition. We would expect that the biggest users with the highest duplication rate (i.e. a low index) would be the first ones to oppose competition. This seems to have been in deed the case in most cities. Furthermore, the cost of competition should be highest in cities where the rival systems have a comparable size even though the duplication rate might be low overall.

Residential users on the other hand, especially subscribers to the majority system, will be able to conduct an increasing share of their communication over the telephone thanks to increasing duplication. If they support competition at time $t = 0$ they should certainly support it at later time periods. While initially both residents and businesses share the (small) cost of competition, the burden shifts more and more to duplicating business subscribers over time.

3.4.2 Chance Adopters and the Duplication Effect

Telephone technology diffuses in my model because residents at the boundary of the expanding initial cluster subscribe. While I want to preserve this 'contagious growth' dynamics as the main force behind network expansion, there are many reasons why some residents might get a telephone without waiting for their friends to subscribe first. Advertising might convince a resident to become an early adopter even if the share of messages which she can send over the network is below her threshold level m^r. Alternatively, a small cluster of residents might coordinate and simultaneously subscribe - in my model two friends could form a seed for a new expanding cluster.

I will therefore add some 'noise' to the model: in each period the threshold m^r of a non-subscribing resident will decrease to 0 for at least two time periods with a small probability ϵ.[28] She will therefore certainly subscribe to the network that offers her most connections.

[28] Lowering the threshold level for two time periods ensures that the resident does not give up telephone

Those *chance adopters* provide the seeds for further development because residents can now join at the boundaries of more and more clusters as time proceeds.

What difference do chance adopters make in the model? Let us first take a look at chance adopters in minority islands. Without any business duplication a chance adopter will always subscribe to the minority system because

$$(1-\gamma)\beta + \mu\gamma\beta > (1-\mu)\gamma\beta \tag{3.10}$$

with $\gamma < \frac{1}{2}$. If duplication has occurred on the other hand the chance adopter can conduct a fraction of local business calls through the majority network. Condition 3.10 for choosing the minority system then becomes

$$\begin{aligned}(1-\gamma)\beta + \mu\gamma\beta + (1-\mu)\gamma \sum_{i=1}^{j_B(t,\mu)} \beta_i \quad &> \quad (1-\gamma)\sum_{i=1}^{j_A(t,\mu)} \beta_i \\ &+ \quad (1-\mu)\gamma\beta + \mu\gamma \sum_{i=1}^{j_A(t,\mu)} \beta_i, \end{aligned} \tag{3.11}$$

where $j_A(t,\mu)$ and $j_B(t,\mu)$ denote the marginal business types which take dual service in minority and majority islands respectively. The inequality can be simplified and we obtain:

$$\frac{1-(1-\mu)\gamma}{(1-\mu)\gamma} > \frac{\sum_{j_B(t,\mu)+1}^{m} \beta_i}{\sum_{j_A(t,\mu)+1}^{m} \beta_i} \tag{3.12}$$

Two polar cases illustrate this condition nicely. First, assume that both companies have almost equal market share ($\mu = \frac{1}{2}$) and second, let the majority company completely dominate the market ($\mu = 0$). In the former case $j_A = j_B$ and condition 3.12 is fulfilled at any point in time - chance adopters in minority islands will therefore always choose the minority network. In the latter case, however, no majority business will see a need to duplicate at any point in time and $j_B(t,0) = 0$ but all minority businesses will eventually duplicate. This implies, that condition 3.12 will surely be violated beyond a certain state of development and chance adopters in minority islands will then subscribe to the majority network. I will call this phenomenon the *duplication effect*. Such a subscriber will provide the seed of a new majority cluster inside the minority island. Residents at its boundary without a telephone

service in the subsequent period. After two periods the new seed has grown sufficiently to be self-sustaining.

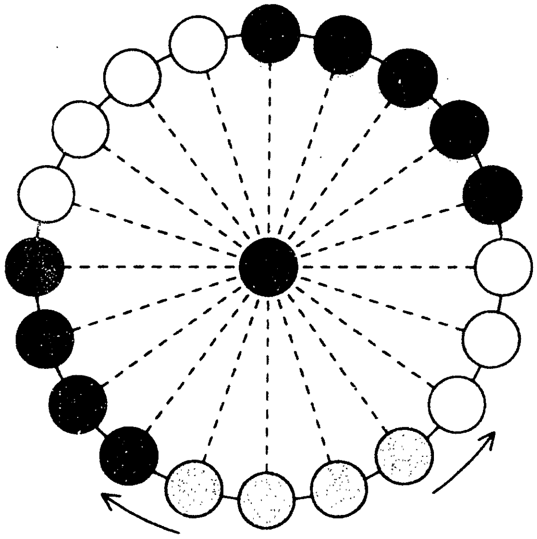


Figure 3-3: Social island of minority company has been invaded by the majority company (lightly shaded): majority cluster expands on the expense of the minority company

will now also choose the majority system: due to condition 3.11 they can conduct at least a share $(1-\gamma)\beta$ of their business calls through the majority network. Therefore they can send a share $\frac{\alpha}{2}+(1-\gamma)\beta$ of their messages over the telephone which justifies subscription due to condition 3.2.

Note, that chance adopters in majority islands will always choose the majority system regardless of the state of development. Minority businesses are at least as likely to duplicate as majority businesses such that a chance adopter can always reach more businesses through the majority network.

To summarize, asymmetric duplication by majority and minority businesses weakens the hold of the minority company on its islands. Without duplication new subscribers have a powerful incentive to stick to the established minority network as they cannot conduct their local business calls over the majority network. With increasing duplication the relative size of both networks in the city becomes more important and can induce new subscriber in minority islands to defect to the majority network. The larger network will therefore be able to 'invade' the market niches of its rival.

Such an invading cluster will act as a 'Trojan horse' and eventually take over an entire social island as long as condition 3.12 continues not to hold. Minority subscribers at the boundary of a majority cluster will defect to the majority network: they are indifferent between both systems for their social calls but will prefer the greater share of businesses they can reach through the larger system.

The dynamics of telephone diffusion can now be informally described as follows. At a low state of development both networks will hold on to their respective islands. Occasional chance adopters will imitate the network choice of other residents in the island. Gradually the increase in telephone ownership will force business subscribers to duplicate which makes minority islands vulnerable to invasion. Eventually condition 3.12 can be violated and chance adopters in minority islands will defect to the majority system. In this case the minority cluster will eventually be taken over by the majority system.

The next theorem formally summarizes the results for the disturbed model. For this purpose I introduce some further notation. The concentration of telephone ownership within each island now evolves stochastically, but the concentration of telephone ownership $x(t)$ within the city grows deterministically due to the law of large numbers. The market share of the minority system $\mu_n(t)$ is no longer fixed because minority islands can get invaded beyond a certain state of development. In the long-run, however, the dynamics settles down, every agent has a telephone and social space consists of 'pure' minority and majority islands. I denote the long-run market share of the minority system with μ_n^∞. Note, that at time $t = 0$ we have $x(0) = \frac{k}{n}$ and $\mu_n(0) = \mu$ just as in the undisturbed model.

Theorem 8 *The concentration of telephone ownership in the city at time t satisfies*

$$x(t) = 1 - \max\left[\frac{n-k-2t}{n}(1-\epsilon)^{t^2}, 0\right]. \tag{3.13}$$

There is a critical market share $\mu^* > 0$ for the minority system and a critical state of development $y^*(\mu) < 1$ for $\mu < \mu^*$ such that for any state of development $x(t) > y^*(\mu)$ all chance adopters will subscribe to the majority network. The market share $\mu_n(t)$ of the minority system is decreasing in time and $\lim_{n\to\infty}\mu_n^\infty = 0$ for $\mu < \mu^*$.

Proof: The probability that a single household has no telephone at time t can be calculated as

$$\frac{n-k-2t}{n}(1-\epsilon)^t \prod_{i=1}^{t}(1-\epsilon)^{2(t-i)}$$

The first two terms of this expression denote the probabilities that the household is not part of the expanding original cluster and that the household has not been a chance adopter up to time t. The remaining terms describe the probabilities that the two neighbors at a distance i of the household have not been chance adopters up to time $t - i$. This expression can then be simplified to obtain formula 3.13.

The second part of the theorem is illustrated by figure 3-3. QED

The duplication effect complements the cost of incompatibility argument of the previous section. While the latter cost will be most severe for the average business user when both networks are of similar size, the duplication effect will be strongest if the market share of the minority network is small.

I now continue the back of the envelope calculation for Louisville from the previous section in order to demonstrate that the duplication effect was likely to apply in US cities during the peak of the independent movement. I inferred, that the market share of the minority network in this city (here the Bell system) was fairly high at about 40%, and the duplication rate amongst minority businesses was between 80% and 100%, while it was close to 0 for majority stores. Using these parameters I can now calculate that for $\gamma < 0.28$ condition 3.12 is violated and all chance adopters would choose the home company.[29] In other words the duplication effect was like.y to be relevant unless the social islands were very isolated from each other.

In order to get a feeling for the impact of the duplication effect I simulated the model for a city with social islands of size n and $\epsilon = 0.01$. The minority network has initial market share of $\mu = 20\%$ and the initial telephone concentration is $x(0) = 5\%$. I also assume that all minority businesses start to duplicate when the telephone concentration reaches the threshold $y_A = 30\%$ ($y_A = 50\%$). This implies, that no majority business will ever want to

[29] Condition 3.12 will only hold if

$$\begin{aligned} \gamma &< \left[(1-\mu)\left(\frac{\sum_{j_B+1}^{j_A}\beta_i + \sum_{j_A+1}^{m}\beta_i}{\sum_{j_A+1}^{m}\beta_i} + 1\right)\right]^{-1} \\ &\leq \left[(1-\mu)\left(\frac{1-w_B}{1-w_A} + 1\right)\right]^{-1}. \end{aligned} \tag{3.14}$$

For the last step we use the fact that the parameters β_i are decreasing in i such that

$$\frac{\sum_{j_B+1}^{j_A}\beta_i}{\sum_{j_A+1}^{m}\beta_i} \geq \frac{(j_A - j_B)\beta_{j_A}}{(m - j_A)\beta_{j_A}}.$$

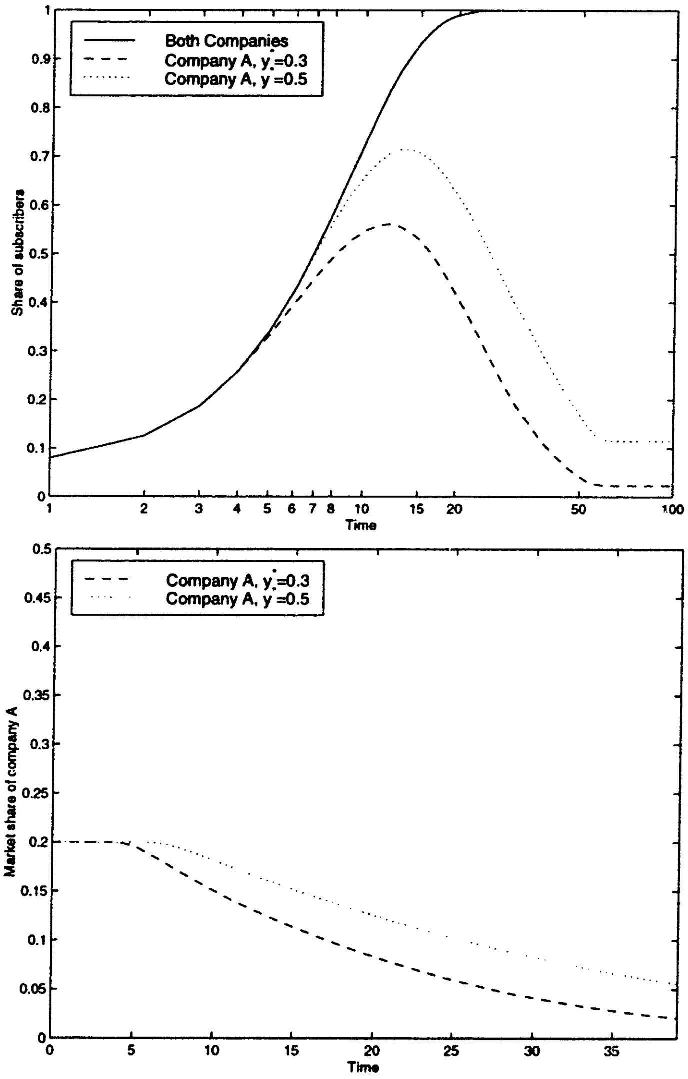


Figure 3-4: Average share of subscribers to company A and B in a minority social island for $y_A = 30\%$ and $y_A = 50\%$ (top with log time axis) and evolution of market share of minority system (bottom): $\epsilon = 0.01$, $n = 100$, $x(0) = 0.05$, $\mu = 0.2$

get a second phone at any state of development because $y_B = \frac{1-\mu}{\mu} y_A > 1$. Therefore the critical state of development beyond which all chance adopters choose the majority network is $y^* (\mu) = 30\%$ ($y^* (\mu) = 50\%$). Figure 3-4 illustrates that the market share of the minority company is initially stable. As soon as minority businesses start to duplicate, however, the minority social islands get invaded. Although both systems still grow in absolute terms the market share of the minority system is decreasing as new subscribers are more likely to join the majority company. Eventually the minority company will suffer a net outflow of subscribers as their users will increasingly defect to the majority company. Although most islands will get invaded a few minority islands will be spared such that the long-run market share μ_n^∞ of the minority system is bounded away from 0.

It is particularly interesting to observe that a marked slowdown in the growth rate of telephone diffusion approximately coincides with the absolute decline of the minority company. This phenomenon is very intuitive: once the social islands are filling up the minority company cannot offset any longer the defection of its customers to the majority company by wooing first-time subscribers. While the majority company initially expands through the growth of the entire market, it subsequently grows by increasing its market share at the expense of its rival.

Remark: In evolutionary game theory the introduction of noise into the model has become a common tool for selecting between multiple equilibria. Kandori, Mailath, and Rob (1993) and Young (1993) introduced the concept of a stochastically stable equilibrium by analyzing the behavior of the system in the limit when $\epsilon \to 0$ while keeping the 'size' n of the system constant. For my model stochastic stability is of no interest, however, as $\mu_n^\infty \to \mu$ as $\epsilon \to 0$.

3.4.3 Saving Dual Service Competition through Interconnection

My model has grim predictions for the long-term survival of dual service competition. If the competing operators are of similar size the cost of incompatibility becomes unacceptably high and business subscribers will prefer a monopoly. Otherwise the duplication effect can drive the minority system towards extinction. In this section I show that general interconnection could have preserved dual service competition in the long run by making duplication unnecessary.

Assume that all subscribers can place a call on the rival network by paying a small interconnection fee c per message. Minority businesses would then never duplicate as long as the cost of a second telephone exceeds the total expense on interconnection charges, e.g.[30]

$$K > \gamma (1 - \mu) M_j c. \tag{3.15}$$

Because the fixed rate K includes the cost of maintaining an extra wire while interconnection just requires a link-up of the exchanges of both companies this condition presumably held.

Without duplication each chance adopter minimizes the interconnection charges by simply adopting the same system as all other residents in the island. Interconnection would therefore have both reduced the cost of incompatibility and eliminated the duplication effect. The market share of the minority system would not decrease over time and its long-term survival would be ensured.

My model therefore implies that the lack of interconnection was the single most important factor for explaining the decline of dual service competition. Although regulators repeatedly tried to introduce interconnection between competing carriers AT&T appealed successfully against those attempts at the courts (see section 3.5).

3.5 Empirical Evidence for the Decline of Competition due to Market Saturation

In this section I test how well my model can explain the actual dynamics of network competition in US cities at the turn of the century. The main variable of interest is the duplication rate of business subscribers. First of all it is the only observable variable which allows us to calculate at least a lower bound for the cost of incompatibility. Second, under asymmetric duplication chance adopters can defect to the majority system which will directly erode the market share of the smaller network. In order to successfully explain the rise and decline of the independent movement the observed duplication rate should satisfy the following properties:

1. Business duplication increases over time such that the cost of incompatibility is initially low but rising.

[30] Note, that the total interconnection charge will be largest for minority businesses when the market share μ is close to 0.

2. The duplication rate eventually becomes large enough such that the resulting high cost of incompatibility makes subscribers prefer a regulated monopoly to dual service competition.

3. Duplication is sufficiently asymmetric in cities where the minority system has a small market share such that the duplication effect applies.

Business duplication rates in fact fulfilled these three properties as I will show in the following. I then present evidence that pressure to consolidate dual service territories came in deed from the demand side as my model would predict.

3.5.1 The Evolution of Business Duplication

Table 3.3 documents the evolution of network competition between the Bell incumbent and an independent rival system for eight Northern cities over at least two time periods. The table summarizes the shares of exclusive subscribers and dual service users. All cities with the exception of Louisville experienced rapid annual growth of the telephone networks which significantly increased the concentration of subscribers in the urban population.

The total rate of duplication increased for half of the cities over time and stayed roughly constant for the rest. In the context of the model, however, only the business duplication

[31] The annual rates are given for business/residential unlimited single-party service respectively. Although the Bell rates in Louisville, Kansas City, Minneapolis and St. Paul seem a lot higher than the independent rates, subscribers could choose two-party and four-party service at rates comparable to independent single-line rates. The Independents on the other hand did not offer multi-party lines in those cities.

[32] The projected duplication rate (in %) has been calculated by dividing the number of duplicates by the reference value B_C. The reference value for the number of 'businesses' in a city has been defined as 50% of all subscribers in July 1905.

[33] Telephone concentration refers to telephones per inhabitant. The population data for 1900 and 1910 is taken from the 1910 census. Intermediate values were interpolated.

[34] *Telephone Magazine* January 1903: p. 6

[35] All data for the year 1905 is taken from the Supplemental Telephone Report (Merchants' Association of New York 1905, p. 3-7).

[36] Anderson (1907, p. 28)

[37] Johnston (1908, p. 7f, 24)

[38] The Indianapolis 1903/1904 data is taken from *The American Telephone Journal* May 14, 1904: p. 309.

[39] For data on the number of exclusive and duplicate subscribers to the New company see Anderson (1907, p. 14). Anderson also mentions that the Bell subscriber base is *about* as big as the independent one. In 1907, however, Bell had only 9,800 subscribers and the New company 11,500 (Committee on Gas, Oil and Electric Light 1907, p. 192). I therefore scaled down the number of Bell subscribers by 20%.

[40] The 1907 data is taken from the Chicago report(Committee on Gas, Oil and Electric Light 1907, p. 172-).

[41] Mueller (1997, p. 82)

[42] Instead of the 1905 rate schedule for Minneapolis and St. Paul and the 1903 schedule for Indianapolis I used 1902 rates quoted in *The American Telephone Journal* April 12, 1902: p. 229.

[43] Anderson (1907, p26)

[44] Rate schedule for 1906 quoted from *Soundwaves* November 1906: p. 433.

Table 3.3: Panel data documenting the evolution of duplication in eight US cities

City		Total	Market share (in %)			Proj.	Conc.	Rates (in US$) [31]		
Bell	Year/	sub-	Bell	Dupli-	Indep.	dupl.	(in %)	Bell	Indep.	Dual
Independent	Month	scribers	only	cations	only	[32]	[33]			
Cleveland	1903/1 [34]	21,000	52.4	*14.3*	33.3	20.3	4.8	48/-	48/36	96/-
Cleveland Tel.	1905/7 [35]	29,560	43.4	*11.9*	44.7	23.8	6.2			
Cuyahoga Tel.	1906/9 [36]	42,000	45.2	*21.4*	33.3	60.8	8.4	84/54	72/48	156/102
Columbus, O	1905/7	12,904	42.8	*10.2*	47.0	20.4	8.3	72/42	40/24	112/66
Central Union	1908/7 [37]	21,340	46.1	*10.1*	43.8	33.4	12.3	54/27	40/24	94/51
Citizens Tel. Comp.										
Indianapolis	1903/10 [38]	11,980	42.6	*11.8*	45.6	21.8	6.2	40/24	40/24	80/48
Central Union	1904/4	12,716	39.5	*10.4*	50.1	20.4	6.5			
New Tel. Comp.	1905/7	12,965	29.5	*12.0*	58.4	24.0	6.3			
	1906/3 [39]	17,000	41.2	*11.8*	47.1	30.9	8.1	54/24	40/24	94/48
Kansas City	1905/7	18,078	46.1	*15.9*	38.0	31.8	8.6			
Missouri Tel. Comp.	1907/1 [40]	32,920	49.0	*13.7*	37.3	49.9	14.8	96/36	54/36	150/72
Home Tel. Comp										
Louisville, Ky	1905/7	14,343	43.8	*17.6*	38.6	35.2	6.7	96/-	48/24	144/-
Cumberland T&T	1910 [41]	16,263	37.9	*18.0*	44.1	40.8	7.3	96/36	48/24	144/60
Home Tel. Comp										
Minneapolis	1905/7 [42]	20,147	50.6	*15.1*	34.4	30.2	7.8	84/60	48/30	132/90
Northwestern Tel.	1907/1	24,998	43.2	*16.9*	40.0	41.9	9.2	90/54	48/30	138/84
Tristate Tel. Comp										
St. Paul, Minn	1905/7	11,124	67.7	*13.3*	18.9	26.6	5.8	84/60	48/30	132/90
Northwestern Tel.	1907/1	14,834	50.5	*19.2*	30.3	51.2	7.4	90/54	48/30	138/84
Tristate Tel. Comp										
Toledo, O	1905/7	9,301	25.0	*10.2*	64.8	20.4	6.1	54/27	48/26	102/53
Central Union	1906/9 [43]	13,300	24.8	*25.6*	49.6	73.2	8.5		52/32 44	
Home Comp.										

rate is of interest. I will in fact concentrate on duplication by medium-scale businesses such as grocers, druggists and physicians because they most closely resemble the 'businesses' in my model. They were amongst the first users to subscribe when a telephone company entered their social island and also made up most of the duplicates. I construct an estimate $w(t, \mu)$ for the rate of duplication amongst those businesses which I call *projected duplication* and list in table 3.3 for each city. Appendix B.2 provides the details for calculating that estimate.

The rapid diffusion of telephone technology led to an enormous increase in the duplication rate of businesses. For four out of eight cities the share of dual business users was above 50% by 1907.[45] Therefore duplication in deed increased over time as required by the model and reached a large enough share to present a burden to business subscribers. Although I do not have information on the duplication rate of minority businesses and therefore cannot calculate a lower bound as in lemma 5, I can use the total business duplication rate w for that purpose:

$$C^b(t, \mu) \geq w(t, \mu) K \tag{3.16}$$

Hence the cost of competition to the average business subscriber increased in four cities to more than 50% of the base rate K. Businesses are likely to have opposed competition in those cities unless they feared enormous rate increases due to consolidation. Such rate hikes were unlikely especially after the Kingsbury Commitment when AT&T could only consolidate in a dual service city by seeking the consent of relevant user groups such as the Chamber of Commerce.

While table 3.3 illustrates that duplication was a growing and significant concern to business subscribers the data is unsuitable for looking at the duplication effect because the competing networks in six out of eight cities were of comparable size. In this case the duplication rate of minority and majority businesses will be similar and chance adopters in a minority cluster would have no incentive to defect to the majority system (i.e. condition 3.12 holds). I instead use data on a cross-section of 14 cities in 1905 where the Bell exchange had a commanding lead (see table 3.4). The maximal market share of the minority system is

[45] These numbers seem plausible when we compare them to the detailed telephone census of Louisville in table 3.2. Although this city experienced only very low telephone growth the duplication rate amongst medium-scale businesses was already between 30% and 50%.

Table 3.4: Sample of cities with a dominant system in 1905

City	Market share (in %)			Share of duplicating independent subscribers
	Bell only	Duplications	Indep. only	
Atlanta, Ga	72	18	10	64
Buffalo, NY	66	18	16	53
Columbus, Ga	71	21	8	72
Elgin, Ill	74	7	19	27
Fall River, MA	68	16	16	50
Harrisburg, Pa	70	13	17	43
Mobile, Ala	78	9	13	41
N. Bedford, MA	72	13	15	46
Norfolk, Va	64	18	18	50
Philadelphia, Pa	78	16	6	73
Pittsburgh, Pa	69	14	17	45
Portsmouth, Va	80	11	9	55
St Paul, Minn	68	13	19	41
Syracuse, NY	70	10	20	33

Source: Merchants' Association of New York (1905, p. 5-6): Supplemental Telephone Report. The table includes all the cities from the original source where the minority system had a market share of less than 35%. In all 14 cases the Bell exchange had the commanding lead.

less than 35% in all cases.[46] For each type j business the threshold telephone concentrations y_A^j and y_B^j at which minority and majority businesses duplicate satisfy:

$$\frac{y_A^j}{y_B^j} = \frac{\mu}{1-\mu} \qquad \text{and hence} \qquad y_B^j > 2y_A^j$$

Minority businesses therefore should duplicate much earlier than majority businesses. The duplication rates w_A and w_B for minority and majority businesses respectively depend, however, on the distribution and the support of business communication shares β_j. The more similar the communication demand of the various business types, the larger the asymmetry between the duplication rate of minority and majority businesses. If the β_j have support $\left[\underline{\beta}, 2\underline{\beta}\right]$ for example all minority businesses in my cross-section of cities would duplicate before any majority business does so.

Table 3.4 illustrates that in half of all cities subscribers to the independent network duplicated with a probability of more than 50%. Under the assumption that the communi-

[46] The maximal market share of the minority system is the sum of the duplication rate and the share of exclusive independent subscribers.

cation shares β_j are sufficiently bunched together and that on average half of all subscribers were businesses in 1905 these probabilities imply that most minority businesses duplicated while few of the majority businesses did so. Business duplication was therefore sufficiently asymmetric for the duplication effect to apply.

The duplication effect will impede further growth of the minority system and reduce its market share. But recall that in the numerical simulations the minority system was only shrinking in absolute terms once the social space was filling up and the growth rate in telephone concentration was declining. Evidence for such a slowdown due to gradual market saturation after 1907 can be found in table 3.1. The rate of telephone diffusion halved from 1903-7 to 1907-12. According to this argument telephone companies with a small market share were likely to see their subscriber list shrink during the late competitive era after 1913.

3.5.2 Growing Public Opposition to Dual Service Competition

If both competing systems are of similar size we would expect that the rise in the cost of incompatibility should be reflected in growing public opposition to dual service competition. Public opinion in deed gradually shifted by 1907. Town councils which had welcomed the Independents in the early competitive era became more careful in granting new franchises. Herbert Casson (1910, p. 191), an early historian of the telephone industry, explained the change in public opinion:

> Most people fancied that a telephone system was practically the same as a gas or electric light system, which can often be duplicated with the result of cheaper rates and better service. They did not for years discover that two telephones in one city means either half service or double cost.

In January 1907 motions were introduced in the state legislatures of Nebraska, Missouri and Kansas for compulsory local interconnection between companies doing business in the same city or village. Twenty-two states finally passed interconnection bills between 1910 and 1913 which empowered utility commissions to order interconnection if telephone users of some locality demanded it. Enforcement was cumbersome, however, and AT&T's opposition to interconnection in dual service cities meant that these laws were almost never applied.

Municipal governments in Kansas City, Cleveland and Indianapolis began their own investigations of the telephone situation and recommended local interconnection. In Cleveland the city council described duplication a "nuisance" in 1908 (Mueller 1997, p. 121). This is unsurprising given that at least 60% of medium-scale businesses subscribed to both networks by 1906 already (see table 3.3). But AT&T rejected interconnection without consolidation and support of an outright merger was often tempered by the fear of large rate increases. The Kingsbury Commitment made consolidation a more attractive options for dual cities after 1913 as it gave them enough leverage to negotiate future telephone rates before consolidation.

A case study of consolidation in Southern California by Mueller illustrates particularly nicely the political dynamics which worked against dual service in a region where both networks divided the market roughly equally (Mueller 1997, p. 140-43). By 1916 Bell operated 11 exchanges in Los Angeles with 67,000 stations while the independent company had 60,300 subscribers on 14 exchanges. Both networks had extensive toll connections in the region. In 1910 political agitation against duplicate service began and in 1915 a municipal referendum was held. A 5:1 majority of voters expressed their support for compulsory interconnection which would have left competition between both companies intact. This option turned out to be too expensive as the rival systems were technically incompatible.[47] The city council then decided to insist on consolidation under rate regulation and the companies merged in 1917. A survey conducted by an economics student at the University of Southern California in 1916 found that business users rather than residents were the driving force behind calls for consolidation as we would expect from the theoretical model: 100% of business subscribers in his sample were troubled by being unable to reach people on another network, but 34% of housewives were never troubled at all.

3.5.3 User Convergence to the Dominant System

In cities with one dominant network the minority system could lose subscribers due to the duplication effect during the late competitive era. A case study by Mueller (1997, p. 137-140) on telephone competition in western New York shows how subscriber convergence to the larger system forced consolidation and ended dual service competition.

[47] This does not imply that a general policy of interconnection for the telephone industry would have been expensive as the rival systems would have been technically compatible from the start.

The Federal Telephone Co. owned 35 independent exchanges including system in Buffalo, Rochester and Jamestown. In 1916, the manager of the Buffalo-based company, Burt G. Hubbell observed a tendency among subscribers in dual towns to gravitate towards the larger of the two systems. This problem was particularly acute in Buffalo, where Bell outnumbered the independent by a ratio of 3:1 and where the subscriber list of the Independent was shrinking. Consequently, the region was being subdivided into Bell and independent towns which prevented communication between those communities. Furthermore, this development threatened the viability of the independent system which saw its regional market increasingly fragmented. Hubbell noted:

> The natural tendency of the public to patronize the company with the largest number of subscribers ... has led to a segregation into telephone districts in each of which one of the two competitors has usually acquired a great predominance of subscribers.

Hubbell had failed to stop the migration of subscribers to the Bell system even after heavy advertising campaigns. The Bell system agreed to consolidate but required the approval of a majority of telephone users in order to apply for a waiver under the Kingsbury Commitment. After the Buffalo Chamber of Commerce had agreed on a new rate structure which left rates in the middle and the bottom of the communication hierarchy unchanged, the competitors swapped territories and thereby eliminated dual service competition. Bell acquired the Buffalo exchange while the Independent gained a local monopoly over Rochester and Jamestown.

3.6 Conclusion

In this chapter I argue that dual service competition was bound to be a transitory phenomenon in the absence of interconnection between the rival networks. When the rival networks were of comparable size the cost of incompatibility eventually resulted in political pressure to unify the service. With one dominant network on the other hand subscribers could gravitate towards the larger system due to the duplication effect. In both cases the dynamics of competition gave rise to regional monopolies.

This insight allows us to reinterpret the predatory policies which AT&T pursued before the Kingsbury Commitment. Rather than destroying dual service competition AT&T

secured itself the dominant position in the post-competitive era by taking control of most regional monopolies. In particular AT&T weakened the independent movement to an extent that it never managed to build a rival long-distance network. When technological progress made long-distance telephony affordable to most subscribers after World War I, AT&T could reap the full benefits of defending its monopoly in this market.

Furthermore, my model causally links the eventual demise of competition with the lack of interconnection between competing networks. It is likely that AT&T's refusal to interconnect rather than predatory pricing was most damaging to competition in the telephone industry.

Chapter 4

The Evolution of Work

4.1 Introduction

During the first two thirds of the 20th century the organization of manufacturing work relied on a sharp division of labor where workers performed a narrow set of tasks according to detailed job descriptions. The principles of job design in the mass production economy were outlined by Frederick Taylor (1911, p. 21) in his theory of "scientific management":

> Under our system a workers is told just what he is to do and how he is to do it. Any 'improvement' he makes upon the orders given to him is fatal to his success.

In the last 30 years departures from this *Taylorist* organization of manufacturing work have become increasingly common. Job roles are expanding both horizontally through job rotation and the merging of narrow job descriptions into broad job classifications, and vertically by introducing flat hierarchies and autonomous work teams.[1]

In many ways these innovative forms of work organization in the *New Economy* resemble those of a much earlier era, namely the pre-industrial artisan economy where skilled craftsmen worked on a product from start to finish. The technologies used by the carriage maker of the 19th century and the team worker in a Japanese transplant car factory might differ

[1] Osterman (1994, 1998) found in a representative sample of US establishments that 23.8 percent of companies had job rotation in place in 1992 with at least half of all production workers involved while 39.8 percent of companies organized their workforce in teams. By 1997 the use of job rotation had more than doubled to 56.4 percent. Pil and McDuffie (1996) analyzed a matched sample of assembly plants as part of the International Motor Vehicle Project and reported that 15.7 percent of employees were involved in teams in 1989 but 46.3 percent in 1996.

enormously. However, in terms of their work experience they have far more in common with each other than with an assembly line worker in Ford's Model T plant in the 1920s.

How can we explain this pattern in the organization of work over time? A formal framework helps to clarify the question. The organization of work Ω and the skill mix S are inputs in the firm's production function $F(\Omega, S, E)$ while the set of parameters $E = \{\tau, P\}$ describe the exogenous technological and product market environment in which the firm operates.[2] I am interested in the mechanism which translates changes in firms' environment into new forms of work organization.

Most of the modern labor literature has little to say about the organization of work and instead looks at the complementarity of technological progress and skill requirements over time. Can an analogous hypothesis of "organization-biased" technological change explain the decreasing division of labor in the New Economy? Evidence from case studies on work reorganization suggests that the answer is no.[3] The introduction of innovative forms of work organization does generally not require a new type of production technology, but rather emphasizes the need to use existing technology in a new way. Instead, pressure to reorganize seems to come mainly from the demand side. Osterman (1998) found that the best predictors for the adoption of innovative work systems are the intensity of product market competition and a company's decision to compete on the basis of quality and product variety rather than price. The importance of the product market for the organization of work was first noticed by Piore and Sabel (1984) who argued that stable product markets were a prerequisite for the mass production economy.

Building on this early work I develop a formal model which allows me to explain the evolution of work from artisan production, over mass production to the New Economy. I start from the premise that technology determines the degree of variety or *customization* in product markets. Greater product variety implies a less predictable product demand

[2] A growing body of empirical evidence suggests that the organization of work is a choice variables of firms. Industry studies show that the *same* technology can be combined with widely different forms of work organization. Wilkinson (1983) and Giordano (1992) analyzed the adoption of computer numerically controlled metal cutting machines in the engineering industry. They can be either programmed by engineers in a central planning department or by machine operators themselves. In the apparel industry sewing machine technology has remained essentially unchanged for the last 30 years as a growing number of companies have introduced team assembly since the late 1980s (see Abernathy et al. (1999)).

[3] Formally, technical change is organization-biased if $\frac{\partial^2 F}{\partial\Omega\partial\tau} > 0$. Bresnahan, Brynjolfsson, and Hitt (1999) find evidence for complementarity between information technology and work reorganization. I do not regard such 'IT-enabled organizational change' as a form of organization-biased technological change because the set of computer users in a company is typically distinct from the set of workers who are involved in work reorganization.

mix because producers become subject to unanticipated trends and fashions. Uncertainty about the composition of demand in return makes production tasks less predictable and favors a flexible organization of work with a weak division of labor. In contrast, if products are standardized production tasks are perfectly predictable and the division of labor is low. Because of its significance for the rest of the paper I restate this link between the degree of customization and work organization as a separate principle:

Taylor's Principle: *The division of labor is determined by the extent of standardization in the product market.*

In my model it is technology which ultimately determines the organization of work with the product market P acting as the transmission mechanism. In other words, firms in my model face a production function of the form $F(\Omega, S, P(\tau))$ and there is no direct effect of technology on the organization of work.

This set-up allows me to explain changes in the organizational of work by making minimal assumptions about the characteristics of the underlying production technology. The artisan economy used general purpose tools and a constant returns to scale technology. A customer could describe the specifications of a good exactly to the artisan who produced diverse output. The degree of customization was therefore high and independent of the extent of the market. The lack of standardization limited the division of labor in artisan production and skilled craftsmen performed most of the intermediate production steps themselves.

The machine economy uses dedicated special purpose equipment to produce identical items at low marginal costs. Production exhibits increasing returns to scale and the degree of customization depends positively on the size of the market. At the onset of industrialization machine production could only support a small number of product varieties. Therefore, US manufacturers began to actively pursue the standardization of product markets towards the end of the 19th century.[4] These efforts made the output mix predictable during the mass production era, and allowed companies to assign workers to narrowly defined tasks.

[4] Landes describes how US metal working companies were the first to adopt uniform shapes and sizes, and imposed them by fiat on manufacturing clients and consumers from the 1880s on (see Landes (1969), p 315). When Henry Ford started to mass produce his Model T he famously declared that customers could have their car in any color they wanted as long as it was black.

The mass production economy started to reach its limits in the 1960s when niche markets for more customized varieties of a basic product had become large enough to attract new entrants. Product proliferation[5] in the mature machine economy offers consumers a similar degree of customization as the early artisan economy but also gives rise to uncertainty about the mix of varieties.[6] Producers are implementing innovative production system, frequently referred to as Just in Time or Lean Production system, in order to deal with the greater uncertainty about the composition of product demand.[7] Job classifications in these systems are typically broader than in mass production facilities. The Toyota production system, for example, groups machines in cells on the shop floor instead of separating them by function. Workers are no longer assigned to a particular machine but to a cell.

A simple extension of my model can provide a non-technological explanation for shifts in the relative demand for skilled labor over time. I assume that skilled workers are more flexible than unskilled workers in the sense that they have a higher average productivity when performing more than one task. High-skilled workers then enjoy a comparative advantage over low-skilled workers in the artisan and the New Economy but demand for flexible labor decreases during the mass production era when production tasks are very predictable. Technological progress and the relative demand for skilled labor are therefore negatively correlated during industrialization but increasingly positively correlated during the rise of the New Economy. My model generates the historic pattern of capital/skill complementarity which has been reported for the US economy by Goldin and Katz (1998). But unlike the literature on skill-biased technological change, I do not have to make any special assumptions about the direction of technological change during different time periods because my model does not assume any direct effect of technological change on skills (e.g. $\frac{\partial^2 F}{\partial S \partial \tau} = 0$).

[5] Abernathy et al. (1999) document product proliferation in the apparel sector. For example, men's shirts were a commodity product up to the 1960s when more than 70 percent of all shirts were white and had a standard cut. That proportion had decreased to 20 percent in 1986. Similarly, in car manufacturing the number of different platforms used as structural under-bodies for product families such as the Oldsmobile increased from 24 in 1955 to 69 in 1973 and 91 in 1986 (Womack (1989), table 7).

[6] The increase in demand uncertainty and the subsequent need to clear unwanted inventories has led to a significant change in pricing practice for consumer goods starting in the late 1960s as more products were sold at mark-down. The dollar value of total mark-downs (on all merchandise sold in department stores) as a percentage of sales increased from 5.2 percent in 1955, to 6.1 percent in 1965, 8.9 percent in 1975, and 16.1 percent in 1984 (Pashigian and Bowen 1991).

[7] Kelly (1982) surveys case studies of work reorganization in the 1960s and 1970s in mass production plants. Companies typically cited *line balancing* problems (uneven workloads under a stochastic demand mix) as the main motivation for abandoning traditional assembly line production.

An analysis of the impact of globalization on labor markets provides another application of my theory. Trade between similar countries accelerates the rise of the New Economy as it increases the size of the market and therefore promotes product proliferation. I demonstrate that conventional calculations based on the factor content of trade underestimate the effects of trade on wage inequality because they do not take into account changes in product market competition induced by trade.

Finally, the model can be used to endogenize the path of technological progress. During the last 30 years new control technologies became available which gave rise to re-toolable multi-purpose machines on the production side, and information technologies such as bar codes and point of sale information processing on the distribution side. The demand for control arises naturally in my model as the machine economy matures and the demand mix becomes less predictable. The theory also implies that multi-purpose machines and information technology have different feedback effects on the organization of work and on skill requirements.

Taylor's principle is the central assumption of my model and closely resembles the famous insight by Adam Smith (1776) that the division of labor is determined by the extent of the market. At the onset of industrialization both principles coincide as improved means of transportation create mass markets for standardized goods. However, the traditional theory cannot explain the observed decrease of the division of labor in the New Economy. Modification of the basic Smithian model can at best explain a slowdown in the division of labor, for example, by introducing coordination costs (as in Becker and Murphy (1992)). My paper is also related to recent work by Thesmar and Thoenig (1999) who interpret product market instability as a high rate of creative destruction in a model of Schumpeterian growth. Globalization and an increase in the supply of skilled labor after 1960 can increase that rate. Skilled workers leave production for research which increases the skill premium.

The balance of the chapter is organized as follows. Section 4.2 introduces the basic model and derives the pattern of work organization over time. Section 4.3 demonstrates how the model can generate capital/skill substitutability during industrialization and accelerating capital/skill complementarity as the machine economy matures. Section 4.4 discusses the impact of globalization. Section 4.5 endogenizes the path of technological progress. Section 4.6 concludes.

4.2 The Basic Model

My formal framework builds on the now standard Dixit-Stiglitz (1977) model of monopolistic competition, but allows for a more elaborate demand and production system. Consumers do not simply purchase products but can choose between different varieties, or degrees of customization of each product. On the production side there are both monopolistic machine producers, and perfectly competitive artisans. This extension allows me to characterize the evolution of work during industrialization as well as during the rise of the New Economy.

4.2.1 Product Varieties

There is a continuum of consumers $C = [0,1]$ who buy products on the unit interval $P = [0,1]$. Each product has m customizable *features* $\Xi = \{\xi_1, \xi_2, .., \xi_m\}$. At any point t in continuous time each consumer $c \in C$ has a preference profile $(\xi_1(c,t), \xi_2(c,t), .., \xi_m(c,t))$ over all features.

Consumers' preferences for each feature ξ_i are distributed according to an i.i.d. process with mass function $g_{i,t}$ and support $\{A_{i,t}, B_{i,t}, C_{i,t}\}$. Each of these three values corresponds to a 'trend' or 'fashion'. While producers know the set of possible trends in advance they cannot perfectly predict which trends will materialize. Formally, the mass function $g_{i,t}$ can take the form $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$, $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$ or $\left(0, \frac{1}{2}, \frac{1}{2}\right)$ with equal probability, e.g. only two of the three possible trends turn out to be successful and each prospective trend fails with probability $\frac{1}{3}$.

Producers can either customize a feature ξ_i with the three possible trends $A_{i,t}$, $B_{i,t}$ and $C_{i,t}$ or they can leave it uncustomized (indicated by the value U). A product with a profile of customized features is called a *variety* and is said to have degree of customization d if exactly d features are customized. For the sake of simplicity I assume that features can only be customized sequentially, i.e. feature ξ_i can only be customized after ξ_{i-1} has been customized.[8] The single variety without any customization is called the *generic* variety and there are 3^m *fully customized* varieties. In total, the set V of varieties for each product category $s \in P$ has size $\frac{1}{2}\left(3^{m+1} - 1\right)$.

[8] For example, in the case $m = 3$ the three varieties with degree of customization $d = 1$ are $(A_{1,t}, U, U)$, $(B_{1,t}, U, U)$ and $(C_{1,t}, U, U)$.

4.2.2 Consumer Demand

A consumer prefers varieties which have a greater number of customized features matching her preference profile. However, she will attach no value at all to a variety with unwanted customized features. A variety with degree of customization d is said to 'flop' if any of the d targeted trends is unsuccessful. Hence the generic variety is not subject to trends while a partially customized variety flops with probability

$$1 - \left(\frac{2}{3}\right)^{d}$$

which increases in the degree of customization d.

This specification embodies the idea that product proliferation increases uncertainty about the mix of varieties demanded by consumers.[9] Products with few customized features may be uninspiring but demand for them is fairly predictable. Taste shifts will hardly matter because they occur mainly within the targeted consumer groups. Varieties become more vulnerable to trends as they are designed to target smaller niche markets. Taste shifts will occur between rather than within targeted consumer groups which gives rise to endogenous demand uncertainty.

At any point in time a consumer can buy a quantity $x_d(s,t)$ of some variety with d customized features matching her preference profile.[10] Consumers have a CES utility function of the following form:

$$\begin{aligned} U &= \int_0^{\infty} x(t) \exp(-\delta t)\, dt, \text{where} \\ x(t) &= \left[\int_0^1 \left(\sum_{d=0}^{m} \mu^d x_d(s,t)\right)^{\rho} ds\right]^{\frac{1}{\rho}} \qquad (4.1) \end{aligned}$$

Good are substitutes ($0 < \rho < 1$) and consumers prefer varieties with a greater degree of customization ($\mu > 1$).

I will show that in equilibrium all consumers buy product varieties in industry s with the same degree of customization $d(s,t)$ at price $p(s,t)$. The aggregate price level $p(t)$ and

[9] The total demand for all varieties in a product class $s \in [0,1]$ will be stable in the model.

[10] In equilibrium not all varieties might be available to consumers.

the total demand for all varieties of product $s \in P$ can then be derived as follows:

$$p(t) = \left[\int_0^1 \left(\frac{p(s,t)}{\mu^d(s,t)}\right)^{\frac{\rho}{\rho-1}}\right]^{\frac{\rho-1}{\rho}} \tag{4.2}$$

$$x(s,t) = x(t)\left[\mu^{d(s,t)}\right]^{\frac{\rho}{1-\rho}}\left(\frac{p(s,t)}{p(t)}\right)^{-\frac{1}{1-\rho}} \tag{4.3}$$

Remark: A consumer's preference for more customized products is independent of her income because the demand function is homothetic. A bigger market alone is therefore insufficient for greater market segmentation.[11] In my model, the level of customization depends on the interaction between production technology and market size.

4.2.3 Artisan Production

Each consumer/worker supplies one unit of labor inelastically. There are three stages of production, namely the *entry stage* at time $t.0$, the *implementation stage* at time $t.1$ and the *production stage* at time $t.2$.[12] Table 4.1 illustrates the sequence of actions taken at each stage. At the entry stage workers have to decide whether they want to become self-employed artisans or industrial workers in a competitive labor market. Plants purchase machines needed for manufacturing during the implementation stage and artisans/ production workers are assigned to preliminary production tasks. The tastes of consumers, however, are only revealed at the production stage when artisans and plants finally decide which varieties should be produced. I assume that the manufacturing process differs for each product variety in both artisan and industrial production: the set of possible production tasks in industry $s \in P$ is therefore indexed by the set of varieties V.

Artisans use a constant returns to scale production technology based on general purpose tools which can be costlessly acquired during the implementation stage. Artisan technology is completely flexible, i.e. it can be used to produce any variety. One unit of output requires c_A units of artisan labor. Artisans decide in the production phase which variety to produce and sell their products in a competitive market.

[11] There is a literature on the 'hierarchy of needs' which essentially assumes that poor people want basic products while rich consumers want more customized varieties. However, the anthropological evidence does not support this hypothesis as Piore and Sabel (1984) point out.

[12] The sequential timing of these three stages would be better captured by a discrete time version of my model in which producers would plan at time $t-1$ and produce at time t. For the sake of simplicity I have 'merged' all phases into period t. The basic intuition of the model is unaffected by this assumption.

Table 4.1: Entry, implementation and production stage in artisan and machine production

	Artisan Production	**Machine Production**
Entry Stage		- incumbent/ entrant play entry game
Implementation Stage	- production workers prepare for task	- hire workers - build machines - production workers prepare for task
Production Stage	- produce variety	- switch workers - produce variety

In order to discuss the organization of work I introduce the following *organizational index.*

Definition 2 *The organizational index $\Upsilon(s,t)$ measures the probability that in the production stage a production worker/ artisan performs the task she has been assigned to in the implementation phase.*

By measuring the attachment of a worker (artisan) to a task the organizational index captures the extent of the division of labor in industry s. The index reflects Taylor's principle that the division of labor is determined by the extent of standardization. If the degree of customization is low there is little demand uncertainty. This makes production tasks predictable and the organizational index assumes a value close to 1. In contrast, under full customization a worker has to anticipate a large variety of potential tasks. The low degree of division of labor is reflected in an organizational index close to 0.

Clearly, artisans will always produce fully customized products because artisan technology can produce any variety at the same cost. The artisan economy therefore exhibits a low degree of division of labor and the organizational index takes the value $\Upsilon(s,t) = \left(\frac{2}{3}\right)^m$.

4.2.4 Machine Production

Machine technology relies on dedicated equipment to produce large quantities of identical goods at low marginal costs. However, efficiency comes at the cost of inflexibility. I make the (extreme) assumption that each product variety requires an extra machine which has to be installed during the implementation stage. A labor input of $k(t)$ is needed to develop and install this machine.[13] Each unit of output requires an additional labor input of $c_M(t)$ during the production stage.[14] I assume that due to general technological progress both the fixed cost $k(t)$ and the marginal cost $c_M(t)$ decrease at the same rate θ. This specification implies that the ratio of the average cost of producing x_1 and x_2 units does not change over time while productivity improves. For simplicity, I also assume that a machine producer either manufactures all varieties of a certain degree of customization or none.[15]

As in the standard Dixit-Stiglitz (1977) model we would expect a firm to sell its output at a mark-up of $\frac{1}{\rho}$ over marginal production costs. However, the firm might face competition from the producer of less customized varieties. The following condition ensures that such lower quality varieties will never succeed because consumers value customized features sufficiently:[16]

$$\mu > \frac{1}{\rho} \tag{4.4}$$

In each industry there is an incumbent and free entry of firms. Shares in the incumbent are equally owned by consumers. Incumbent and entrants play the following game during the entry period. The incumbent has a first-mover advantage and commits to producing all varieties of degree of customization d_I *unless* an entrant decides to enter the market with more customized varieties. The entrant observes the design decision of the incumbent and commits with probability y to produce varieties with degree of customization $d_I + 1$ in which case the incumbent drops his production plans.[17]

In the implementation stage the winner of the entry game hires workers from a com-

[13] I define a 'machine' fairly broadly. I assume that it consists of all sunk investments which a firm has to make before it can begin the large-scale production of a new variety. In the automobile industry, for example, a company has to commission expensive design studies and prototypes before it can install any physical equipment. This final step does not necessarily involve the construction of a green-field plant because the body of a new car model can be produced on existing pressing machines after re-tooling.

[14] Machine technology exhibits increasing returns to scale because the average cost of a unit of output decreases with the scale of production.

[15] This assumption is not essential. It simplifies the set-up because the total demand for each product is deterministic even though the product mix is uncertain. Companies which produce all varieties of a certain degree of customization can then reassign production workers internally rather than 'trade' them in a secondary labor market.

[16] If the competitor prices its variety at marginal cost the up-market producer can charge consumers a price up to $c_M \mu$ before they defect to the less customized variety. But profits are maximized at a price of $\frac{c_M}{\rho}$ which is *below* that limit.

[17] Note, that it would not make sense for the entrant to produce the same varieties as the incumbent. Bertrand competition would erode all potential profits. It will also become clear that the entrant would not want to enter with a more customized variety.

petitive labor market and builds machines. Consumer tastes are realized at the production stage and producers can react to the news by assigning production workers to different tasks. The total demand in each sector is deterministic and a company can therefore reassign production workers internally when it responds to the realization of trends. However, the investment into a machine is sunk even if the corresponding variety is never produced.

One further condition ensures that incumbents would not always want to produce fully customized varieties. A necessary condition is that producers face a decrease in the expected demand for each variety. Since the total demand for a product increases by a factor $\mu^{\frac{\rho}{1-\rho}}$ by adding one more customized feature (see expression 4.3) and the number of varieties increases by a factor of 3 the condition can be calculated as:

$$\frac{\mu^{\frac{\rho}{1-\rho}}}{3} < 1 \tag{4.5}$$

Condition 4.5 expresses the tradeoff between customization and increasing returns. In each industry incumbents have to commit to a degree of customization such that entry generates either no profits or entrants are just indifferent between entering and staying out.

The organizational index can be calculated as

$$\Upsilon(s,t) = \left(\frac{2}{3}\right)^{d(s,t)}.$$

Although the producer who survived the entry game invests in machines to produce all $3^{d(s,t)}$ potential varieties only the $2^{d(s,t)}$ successful ones will be produced. An industry is said to engage in *mass production* if it produces only the generic variety, i.e. $d(s,t) = 0$. In this case the division of labor is high since there is no demand uncertainty. If niche markets are large enough an industry can offer fully customized products and the organization of work resembles that under artisan production.

4.2.5 Characterizing the Dynamics

In the remainder of this section I derive the evolution of the economy over time. I assume that time starts at $t = 0$ and I calibrate the model such that all workers are artisans initially. I then show that the economy goes through two basic transitions. During industrialization machines gradually replace artisan technology. However, product markets are still small and the nascent machine economy can only support mass-produced generic varieties. The weak division of labor under artisan production therefore gives way to a strict Taylorist work organization. At later stages of development product markets fragment as companies target increasingly narrow niches. A New Economy emerges which eventually offers the same degree of customization as the artisan economy. Although vastly more productive the work organization in the New Economy is the same as under artisan production. The era of mass production appears as an intermediate stage in economic development when increasing returns constrain the depth of customization in the machine economy.

I assume that machine producers of generic varieties have sufficiently low marginal costs to compete from the start with any artisan:

$$\mu^m \frac{c_m(0)}{\rho} < c_A \tag{4.6}$$

The marginal mass producer faces the following demand for her generic product variety:

$$x_{C0}(t) = \frac{E\rho}{c_M(t)\,\omega} \left(\mu^m \frac{c_M(t)}{c_A \rho}\right)^{-\frac{\rho}{1-\rho}}, \tag{4.7}$$

where $E = 1$ is the total income of consumers once we take the wage as the numeraire. Mass production will be unprofitable as long as the level of demand for the generic variety does not justify the expense of building a dedicated machine:[18]

$$x_{C0}(t) < \frac{k(t)}{c_M(t)} \frac{\rho}{1-\rho} = A \tag{4.8}$$

If machines are sufficiently expensive (i.e. the fixed cost of a machine is sufficiently large relative to the marginal cost) this condition will be fulfilled at time $t = 0$ and mass producers will stay out of the market.

However, over time the output of the marginal mass producer increases as machines continue to improve due to technological progress. Eventually, entry occurs as soon as $x_{C0}\left(t_1^{C0}\right) = A$ which marks the onset of *industrialization*. From then on artisans in more and more industries will become displaced as mass production spreads through the economy.

Theorem 9 *The share y of industrialized sectors increases until the entire economy has*

[18] Note, that $\frac{k(t)}{c_M(t)}$ is constant over time because both the fixed and the marginal cost decrease at the same rate.

industrialized at time t_4^{CO}. The division of labor becomes stricter as the organizational index increases from $\Upsilon\left(s, t_1^{C0}\right) = \left(\frac{2}{3}\right)^m$ to $\Upsilon\left(s, t_4^{C0}\right) = 1$.

Proof: see appendix C.1 ($a = 1$)

In the mass production economy the marginal producer of a variety with one customized feature faces demand $x_{C1}(t)$ which can be calculated as:

$$x_{C1}(t) = \frac{\mu^{\frac{\rho}{1-\rho}}}{3}\frac{E\rho}{c_M(t)} \tag{4.9}$$

Initially, customized production is unprofitable because the volume of demand does not cover the fixed cost of investing into a machine:

$$x_{C1}(t) < A \tag{4.10}$$

As technology improves the niche markets for customized varieties eventually become large enough to attract entrants at time t_1^{C1} when $X_{C1}\left(t_1^{C1}\right) = A$. From then on customized production spreads and mass markets dissolve until mass production is completely replaced. Now a new cycle starts and the economy moves to the next stage of customization in an analogous fashion.

Theorem 10 *At time $t_1^{C(d+1)}$ producers of variety d start to face entry from competitors who offer varieties with degree of customization $d+1$. The probability y of entry is increasing over time until the entire economy has moved to producing varieties with degree of customization $d+1$ at time $t_4^{C(d+1)}$. During each cycle tne division of labor becomes less strict as the organizational index decreases from $\Upsilon\left(s, t_1^{C0}\right) = \left(\frac{2}{3}\right)^d$ to $\Upsilon\left(s, t_4^{C0}\right) = \left(\frac{2}{3}\right)^{d+1}$.*

Proof: see appendix C.2 ($a = 1$)

4.3 The Emergence of Capital Skill Complementarity

In section 4.2 I demonstrated how a model based on Taylor's principle can explain the organization or work. Another important aspect in the evolution of work is the changing relative demand for skilled labor. A simple extension of my model can map the results on the organization of work into predictions about the relative demand for skilled labor. I first outline the main idea and discuss some of the related literature before I introduce the formal model.

4.3.1 Skills, Flexibility and the Organization of Work

The main facts to be explained are a decrease in the demand for skilled labor during industrialization in the 19th century followed by a gradual increase in relative demand during the first two-thirds of the 20th century and an acceleration of this trend since the 1970s.[19] My model replicates this pattern if I add the assumption that skilled workers are more *flexible* than unskilled workers. I call a worker flexible if her productivity does not depend on the production task to be performed while inflexible workers achieve high productivity only for the subset of production tasks with which they are familiar. There are strong theoretical reasons to believe that skill and flexibility are correlated: skilled workers have either acquired customary knowledge of a number of tasks through experience, or they have an abstract understanding of the entire production process and can therefore deduce the work content of unfamiliar tasks autonomously.

Flexible workers enjoy no comparative advantage when the division of labor is high because inflexible workers can prepare for the production task which they are likely to perform. This "cost of labor" argument was first made by Charles Babbage (1835, p. 175-176) who realized that the increasing division of labor under industrialization eroded the position of the skilled worker:

> ...the master manufacturer by dividing the work to be executed into different processes, each requiring different degrees of skill or force, can purchase exactly that precise quantity of both which is necessary for each process; whereas if the whole work were executed by one workman, that person must possess sufficient skill to perform the most difficult, and the sufficient strength to execute the most laborious, of the operations into which the art is divided.

Flexibility has again become an important quality in the rise of the New Economy. Caroli and Van Reenen (1999) analyze a sample of British and French firms and find that organizational change decreases the demand for unskilled labor. Direct evidence about the

[19] See Goldin and Katz (1995, 1998), and Goldin and Margo (1992) for pre-1960s evidence, and Bound and Johnson (1992), Berman, Bound and Grilliches (1994), and Katz and Murphy (1992) for evidence on acceleration.

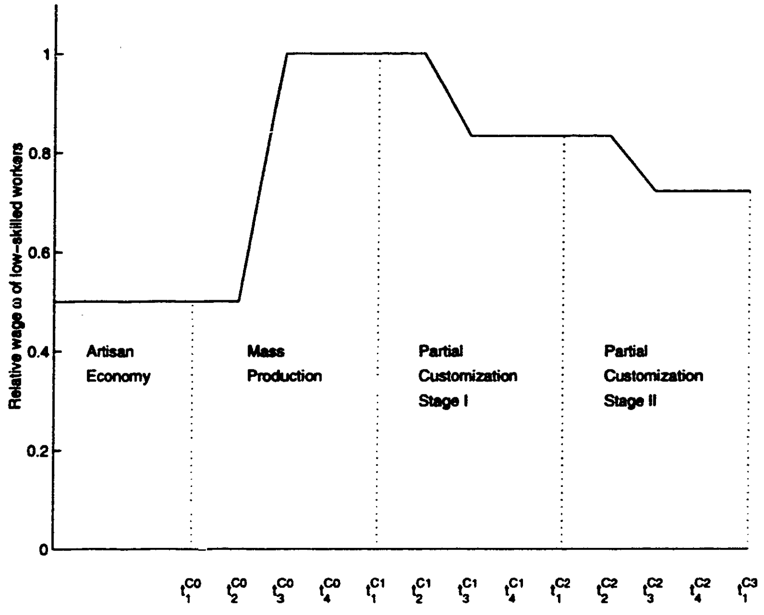


Figure 4-1: Evolution of income inequality in the extended model

positive effect of innovative forms of work organization on skill requirements in the US has been collected by Capelli and Rogovsky (1994). Case studies by Murnane, Levy and Autor (1999) and Zell (1997) also demonstrate that companies which undergo organizational change provide better training and apply a more discriminating selection process.

An outside observer who would try to interpret changes in the demand for skilled labor over time might conclude that capital and skills were substitutes during industrialization but that they increasingly complement each other as the machine economy matures. I want to emphasize that this interpretation would be wrong in the context of my model: there is no direct complementarity between skills and technology. Instead technological change affects skill requirements only indirectly with the product market acting as transmission mechanism. In particular, I do not have to invoke skill-biased technological change to explain the recent increase in the demand for skills as most of the labor literature does.[20]

[20] Bartel and Lichtenberg (1987), and Galor and Tsiddon (1997) argue that capital-skill complementarity arises because skilled workers are better in implementing new technologies. Acemoglu (1998) suggests that technology complements skills not by nature but by design and demonstrates how an increase in the supply of skilled workers can induce directed technological change.

4.3.2 Extending the Basic Model

I assume that there are two types of workers: a share α of the workforce is flexible and provides a full unit of labor at any task. The remaining share $1-\alpha$ of workers are inflexible. They can *prepare* for exactly one production task in the implementation phase (see table 4.1). If they perform this particular task in the production phase they are as productive as skilled workers. However, at any other task they only provide a labor input of $a < 1$.

I assume that the productivity advantage of flexible workers is not too large:

$$a > \frac{1}{\mu} \tag{4.11}$$

Condition 4.11 ensures that both flexible and inflexible artisans will always produce fully customized varieties. Income inequality in the economy is completely characterized by the relative wage ω of inflexible artisans/ workers. In the artisan economy the relative wage can be calculated as:[21]

$$\omega = w(m) = \left(\frac{2}{3}\right)^m + \left(1 - \left(\frac{2}{3}\right)^m\right) a \tag{4.12}$$

Condition 4.4 in the basic model ensured that machine producers of more customized varieties do not face competition from less customizing producers. This condition has to be modified because competitors can use cheap inflexible workers in production:[22]

$$\mu a > \frac{1}{\rho} \tag{4.13}$$

Furthermore, I now assume that of the $k(t)$ units of labor required for installing a machine, $\alpha k(t)$ workers have to be skilled and $(1-\alpha)k(t)$ workers unskilled. Assuming a Leontief production function for machines makes the model particularly easy to solve.

The organization of work follows the same dynamics in the extended model as in the basic model. However, the division of labor will now determine the relative wage of unskilled workers which in return influences the tradeoff between more and less customized

[21] A slight complication arises from the fact that inflexible artisans face income uncertainty because they can only prepare for a task successfully with probability $\left(\frac{2}{3}\right)^m$. I assume that workers have access to perfect income insurance in order to avoid this complication.

[22] The relative wage of low-skilled workers is $\omega = a$ in the worst case. Note, that condition 4.13 implies condition 4.11.

production.

As before, the marginal mass producer will not enter as long as the market for generic varieties is small (note, that $\omega = w(m)$ and that the marginal mass producer only uses cheap unskilled labor):

$$x_{C0}(t) < \frac{k(t)}{c_M(t)} \frac{\rho}{1-\rho} \frac{\alpha + (1-\alpha)\omega}{\omega} = A\frac{\alpha + (1-\alpha)\omega}{\omega} \tag{4.14}$$

After entry at time t_1^{C0} the following theorem describes the process of industrialization in the extended model.

Theorem 11 *Inflexible workers gradually switch into mass production until all of them are employed in industry at time t_2^{C0}. From then on the relative wage of inflexible workers will start to increase until unskilled workers earn the same wage as artisans ($\omega = 1$). The remaining artisans will become production workers subsequently such that all workers are employed in mass production at time t_4^{C0}.*

Proof: see appendix C.1

Figure 4-1 describes the path of income inequality during industrialization. Initially, flexible workers will continue as artisans because they do not have a productivity advantage over unskilled workers in standardized mass production. As machine technology becomes increasingly productive the relative demand for artisans decreases once all low ability workers moved into manufacturing. This process erodes the relative wage of artisans and eventually equalizes it. From then on skilled artisans are willing to move into manufacturing until mass production has spread across all sectors.

Mass producers will initially face no entry because the niche markets for more customized varieties are too small (note, that $\omega = 1$ and that the marginal producer therefore only uses flexible labor):

$$x_{C1}(t) < A[\alpha + (1-\alpha)\omega] = A \tag{4.15}$$

Eventually, niche markets become large enough to attract entrants at time t_1^{C1}. The next theorem describes the subsequent emergence of the New Economy as a series of cycles in which producers customize more and more features of their products and upskill their labor force in the process.

Theorem 12 *The relative wage of unskilled workers at time t_1^{Cd} is $w(d-1)$. Flexible workers will gradually switch into producing goods with depth of customization d until all of them are employed in the more sophisticated industries at time t_2^{Cd}. From then on the relative wage of unskilled workers will start to decrease until it has reached the level $w(d)$ and firms in the more sophisticated industry are indifferent between employing flexible or unskilled production workers at time t_3^{Cd}. Unskilled workers will start to move into the more sophisticated industries until the entire economy only produces goods with level of customization d at time t_4^{Cd}.*

Proof: see appendix C.2

Figure 4-1 illustrates the rise in inequality as the New Economy emerges. The wage of inflexible workers has to fall once all flexible production workers have moved into the more sophisticated sectors and demand for flexible labor outstrips supply. This process continues until the relative wage reflects the comparative advantage of flexible workers under less predictable manufacturing conditions.

4.4 The Impact of Globalization

This section explores the impact of trade on the organization of work and the demand for skilled labor. Unlike in standard Heckscher-Ohlin models, trade between *similar* countries (i.e. intra-OECD trade) can increase the returns to skills in my model. The prevailing view that globalization did not affect the distribution of income in the US might therefore be premature because it only focuses on the factor content of trade but ignores changes in product market competition induced by trade.

In my model two identical countries trade with each other to take advantage of increased market size and scale economies in machine production.[23] This motivation to trade has been explored by the New Trade literature in order to explain phenomena such as intra-industry trade (see Krugman (1981), Dixit and Norman (1980) and Ethier (1982)). My model adds to this list the possibility that trade between equals promotes a more flexible organization

[23] It is not necessary that consumers in both countries have the same tastes, i.e. follow the same trends. As long as there is some overlap between preferences trade will increase the average degree of customization. This condition is likely to hold as a simple example demonstrates: Volkswagen's New Beetle is manufactured in Mexico for both the North American and the European markets.

of work and an increase in the demand for skilled labor. The opening of the US economy to world trade, in particular trade with OECD countries, could therefore have accelerated the rise of the New Economy.

This insight is significant, because most attempts to quantify the impact of globalization on wage inequality in the US have focused on the factor contents of trade. In this context only trade with less developed countries which have a relatively large pool of unskilled workers matters. However, the volume of such trade is too small to have a sizable effect on the US wage distribution as Katz and Murphy (1992) and Berman, Bound and Grilliches (1994) showed. Although total trade as a fraction of GDP more than doubled in the 1970s most of this expansion affected trade with high-wage countries. The share of US manufacturing imports from low-wage countries in manufacturing value-added only increased from 5.7 percent in 1960 to 5.1 percent in 1978 and 10.9 percent in 1990.[24] In contrast, imports from high-wage countries increased from 0.8 percent in 1960 to 13.2 percent in 1978 and 19.8 percent in 1990.

Krugman (1995) concluded that we should think of the OECD as one large closed economy and dismiss trade with LDCs as a significant force behind the widening income distribution in the US. However, it would be wrong to put forward pervasive skill-biased technological change as the only logical explanation. My model predicts the transition from mass production towards New Economy to occur in all mature economies even in the absence of biased technological progress.[25] Intra-OECD trade might well have accelerated this transition and even triggered it in some industries. Direct empirical support for this view comes from Osterman (1994) who found that firms are more likely to introduce ınnovative forms of work organization if they compete on international markets.

[24] Low-wage countries are those with a monthly wage less than or equal to 50 percent of the US monthly wage. See table 3 in Sachs and Shatz (1994).

[25] When comparing the rate of up-skilling amongst advanced nations Berman, Bound and Machin (1997) found similar cross-industry patterns. While they interpreted these results as evidence for pervasive skill-biased technological change the data is also consistent with my model if the degree of uncertainty about the demand mix in each industry is correlated amongst countries.

4.5 Technological Progress and the Emergence of Control

There is widespread agreement that the path of technological progress has changed systematically in the last 30 years by providing a greater degree of *control.*[26] Up to the 1950s machines were so specialized that the cost of retooling was enormous. Starting in the 1960s control technologies gradually improved. This gave rise to *multi-purpose machines* (i.e. numerically controlled and computer numerically controlled machines) which are directly used in manufacturing, and *information technology* which is mainly used to coordinate the distribution of goods (i.e. bar codes, point of sale information). Starting with Milgrom and Roberts (1990) comparative statics comparisons based on the level of control have become a standard exercise in a young literature which relies on supermodular production functions to explain the clustering of business practices such as outsourcing, lean production and integrated and process development.

In this section I demonstrate how improvements in control arise naturally in my model. It is intuitively obvious that the demand for flexible technology should increase as the economy matures and the demand mix becomes less predictable. However, the precise mechanism differs for multi-purpose machines on the one hand, and information technology on the other hand. The former complement the rise of the New Economy but are not essential since the production system can usually be made more productive by using existing technology differently (such as grouping machines in cells rather than by function). In contrast, information technology not only complements the rise of the New Economy but also enables it because it gives companies the ability to administer demand uncertainty effectively. By distinguishing between these different types of new technologies my model can make sense of the empirical findings of Doms, Dunne and Haltiwanger (1997) who found that the use of information technologies are correlated with workers' skills both in the cross-section and the time series while the use of multi-purpose machines is correlated with skill requirements only in the cross-section. The model also explains the correlation reported by Bresnahan, Brynjolfsson and Hitt (1999) between the adoption of information technology and the use of innovative forms of work organization.

[26] Bell (1972) suggested a useful classification of technological progress. He argued that up the 1950s technological innovation strived to improve labor productivity through advances in the *transformation* of workpieces (i.e. mechanical looms, pressing machines) and their *transfer* between work stations (assembly lines, pumps). Engineers began to address the control dimension only in the 1960s.

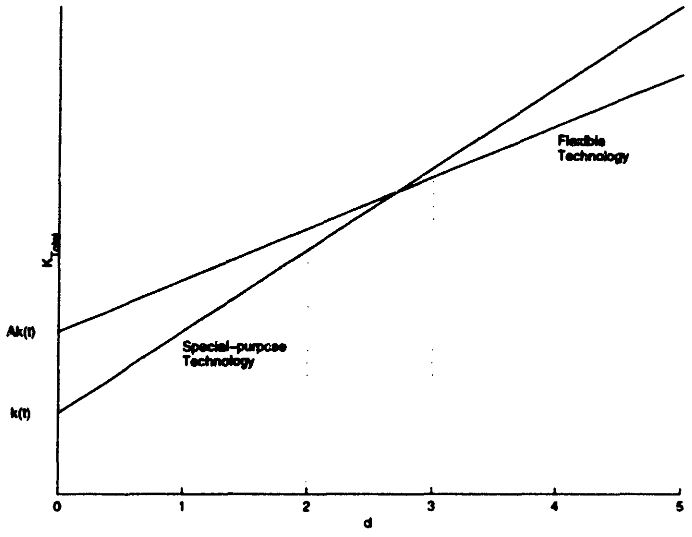


Figure 4-2: Comparison of the total fixed cost of flexible and special-purpose machines for different degrees of customization (on log-scale with $A = 2$): flexible equipment is more cost effective at a degree of customization $d \geq 3$

4.5.1 Multi-Purpose Machines

In my model dedicated *special-purpose machines* become an increasingly risky investment as the economy matures. If an industry offers varieties with a degree of customization d such a machine will be obsolete with probability $\left(\frac{2}{3}\right)^d$ after consumer trends have realized. Capacity utilization (the ratio of expected to maximum volume of production) will therefore decrease rapidly as industries offer more customized varieties.

A natural extension of the model gives companies the option to install *multi-purpose machines.* I continue to assume that producers can build special-purpose machines which require a total labor input of $k(t)$ workers and which have to be scrapped if the respective variety flops. Alternatively they can install flexible machines which can be re-tooled exactly once at the production stage and which can produce any variety with degree of customization d.[27] Although flexible machines will always be fully utilized their versatility comes at a price. I assume that multi-purpose tools are more expensive than standard machines and require a total labor input of $Ak(t)$ with $A > 1$.

Under mass production there will be no demand for flexible technology because the

[27] This condition can be relaxed. Assuming a single opportunity to re-tool assures that even flexible machines cannot produce two distinct varieties at the same time.

specifications of a product are entirely predictable. Multi-purpose machines only become valuable once the economy starts to offer varieties with a greater degree of customization d and uncertainty about the product demand mix increases. It is easy to show that multi-purpose equipment become more cost-effective than standard machines once the capacity utilization rate of special-purpose machines has dropped sufficiently:

$$\left(\frac{2}{3}\right)^d < \frac{1}{A} \tag{4.16}$$

There exists a critical depth of customization d^* such that firms will choose flexible technology over special-purpose machines for all $d \geq d^*$ (see figure 4-2). Theorem 12 continues to characterize the evolution of work even in the extended model.[28]

What can we learn from this richer set-up? First of all, production technology with a greater degree of control emerges endogenously in my model as an increasingly unpredictable demand mix erodes the cost advantage of special purpose machines. Second, the option to install multi-purpose machines will induce machine producers to offer more customized varieties sooner. Greater control therefore accelerates the rise of the New Economy.

Third, the model can shed light on the puzzling observation by Doms, Dunne and Troske (1997) that the use of advanced manufacturing techniques (in particular, multi-purpose machines) explains some of the cross-sectional variations in the demand for skilled labor but little of the time-series variation. This can be seen by introducing some heterogeneity into the model. I assume that not all varieties are equally predictable because one of the two successful trends for each of the m features is known to producers at the implementation stage. This reduces the degree of demand uncertainty for all varieties which incorporate one or more known trends.[29] Skilled, flexible workers and multi-purpose machines are then utilized in the production of 'risky' varieties with few known trends while 'safe' varieties are produced by unskilled workers on special-purpose equipment. However, the demand for more flexible workers will increase in both risky and relatively safe industries because of

[28] Condition 4.5 has to be strengthened in order to make sure that firms do not suddenly start to produce fully customized products when they switch to flexible technology:

$$\frac{\mu^{\frac{\rho}{1-\rho}}}{2} < 1 \tag{5*}$$

[29] I continue to assume that machine producers either manufacture all varieties of a certain degree of customization or none.

continuing market fragmentation. Controlling for multi-purpose machines in a time series regression will then only pick up the difference in the rate of up-skilling between adopters and non-adopters which is not clearly signed.

4.5.2 Information Technology

In the artisan economy the production and distribution operations are usually integrated. Customers can walk into an artisan shop and describe the exact specifications of a variety. The craft economy therefore never produces 'flops'. In contrast, economies of scale lead to the concentration of production in the machine economy and goods reach customers only after they have traversed an elaborate distribution system. Goods are no longer made to order and producers bear the risk of accumulating inventories of 'flopped' varieties. As long as industry produces standardized varieties this risk is small because the demand mix is predictable. The main logistical challenge of the mass production system is to create and efficiently supply mass markets for machine produced goods rather than to track consumer tastes. Mass retailers such as department stores and mail-order houses placed orders well in advance and shipments were large and of low frequency.[30]

This system started to run into problems in the late 1960s as a result of ever greater product proliferation. The demand mix became less predictable and retailers found it more difficult to match their inventories to consumers' tastes. They held an increasing number of 'flops' in their inventories which had to be marked down for sale. The dollar value of mark-downs (of all merchandise sold in department stores) almost tripled from 6.1 percent in 1965 to 16.1 percent in 1984 (see Pashigian and Brown (1991)).

The problems of mass retailing in the maturing machine economy can be easily analyzed in my model. Producers cannot adjust the mix of varieties because demand information is only revealed to them *after* they have manufactured all 3^d varieties. Hence, a variety has to be marked down with probability $1-\left(\frac{2}{3}\right)^d$. Product markets continue to fragment over time but at a slower rate than in the standard model.[31] However, the organization of work

[30] In the apparel market, for example, these transactions typically occurred eight to ten months before the beginning of each season (see Abernathy et. al. (1999)).

[31] Customizing one more feature will increase the effective unit labor input for each *successful* variety by 50 percent because producers take the risk of mark downs into account. Formally, it can be calculated as $\left(\frac{3}{2}\right)^d c_M(t)$. Therefore, the expected demand for each more customized variety decreases by a factor $\frac{\left(\frac{2}{3}\mu\right)^{\frac{\rho}{1-\rho}}}{3}$ which exceeds the contraction of demand in the standard model.

remains the same as under mass production because firms cannot switch workers between production lines for lack of information. Flexible workers do not enjoy a comparative advantage over unskilled workers and earn the same wage.

In order to respond to fashions and market trends in time, the distribution system has to collect, process and relay information about the demand mix back to suppliers. The development of bar codes, scanners and electronic data interchange (EDI) are a rational response of the distribution system to the increased uncertainty in the product market. In the late 1970s a new breed of lean retailers began to take advantage of these information technologies in an attempt to improve inventory management. Wal-Mart, for example, no longer *pushes* inventories to consumers through promotions and other discounts. Instead, the company lets customers *pull* their orders: Wal-Mart collects point of sale information from its various stores in real time which is used to rapidly replenish 'hits' and discontinue 'flops' without holding a large stock of inventory.

The adoption of information technology has a number of testable implications in my model. First of all, firms now find it profitable to implement a more flexible organization of work which allows them to adapt their output mix rapidly. Organizational change in return increases the demand for flexible workers and the skill premium. Second, there is an increase in the degree of customization because producers no longer manufacture unprofitable 'flops'. Empirical support for the implied complementarity between firms' adoption of information technologies, greater customization and innovative forms of work organization includes Bresnahan, Brynjolfsson and Hitt (1999).

It is instructive to compare the impact of information technologies with the previously discussed adoption of flexible equipment. Whereas multi-purpose machines merely complement the rise of the New Economy, information technology acts as the catalyst which enables it. This is consistent with the findings of Berman, Bound and Grilliches (1994), and Autor, Katz and Krueger (1998) that investments in information technology on the industry level explain some of the time series variation in the demand for skilled workers even though the adoption of multi-purpose machines does not.

4.6 Conclusion

My model builds on the wave of empirical research in the 1990s which explored the relationship between technological progress and the transformation of the workplace. This literature successfully demonstrated that new technologies, workplace reorganization and skill requirements have been complements since the 1970s.

However, the typical paper in this literature follows a methodology which makes it problematic to infer organization-biased and skill-biased technological change from this evidence. It assumes a reduced form production function of the form $F(\Omega, S, \tau)$, throws in controls for the various dimensions of technological progress and estimates the strength of the complementarities. Because no attempt is made to understand the precise transmission mechanism that explains the positive cross-partials in the reduced form production function, it is unsurprising that the estimated coefficients have been unstable over time. In particular, the cross-partials were negative during the industrial revolution which gave rise to the Taylorist organization of work and replaced skilled artisans with unskilled machine operators.

In contrast, I explicitly model the transmission of technological progress through the product market environment in which firms operate. This set-up allows me to explain the historic U-shaped evolution of work organization from artisan to New Economy. Moreover, I can derive the impact of distinct technological innovations on the demand for skilled workers within a unified framework. The model promotes the view that the era of mass production was a transitory phenomenon, a period in which the scale economies embodied in machine production limited the degree of product customization.

Appendix A

A.1 Transforming the Residential Process into a Discrete Time Markov Chain

It is often easier to work with the discrete time counterpart of a continuous-time Markov process on the state set Z. In particular, results from the stochastic stability literature can be applied directly.

The discrete time Markov process is constructed as follows. The discrete 'clock' is scaled so that time increases in increments of $\frac{1}{n}$. In each period exactly one of the n residents in the residential is randomly selected, moves out and is replaced by a newcomer from the housing market. The switching functions are assumed to be the same as for the continuous-time process. I introduce the convention that the configuration η_z is obtained from η by inverting the ethnicity of the resident at size z and leaving the other residents unchanged. The transition matrix P^ϵ then becomes:

$$P^\epsilon(\eta,\mu) = \begin{cases} \frac{1}{n} g_w^\epsilon(x(\eta,z)) & \text{if } \mu = \eta_z \text{ , } \eta(z) = 0 \\ \frac{1}{n} g_b^\epsilon(y(\eta,z)) & \text{if } \mu = \eta_z \text{ , } \eta(z) = 1 \\ 1 - \frac{1}{n}\sum_{i=1}^{n}(1-\eta(z_i))\, g_w^\epsilon(x(\eta,z_i)) - & \\ -\frac{1}{n}\sum_{i=1}^{n}\eta(z_i)\, g_b^\epsilon(y(\eta,z_i)) & \text{if } \mu = \eta \\ 0 & \text{otherwise} \end{cases}$$

Note, that each site will be chosen once per time unit just as in the continuous-time process. For this reason the long-run ergodic distribution and all waiting times derived for the discrete time model are the same as for the original continuous-time process.

The 'undisturbed' transition matrix P is obtained from P^ϵ by setting $\epsilon = 0$. The triple

(Z, P, P^{ϵ}) describes a model of evolution with noise as specified by Ellison (1999) and his results for characterizing waiting times apply.

A.2 Results on Random Walks with Drift

For the following theorem I assume that time is discrete and that time increases in increments of $\frac{1}{n}$ as in appendix A.1.

Lemma 6 *Consider a random walk on the integers between* 0 *and* $n > 0$. *The process moves up with probability* α *and down with probability* β *where* $\alpha + \beta \leq 1$ *and* $\alpha > \beta$. *Starting from* $0 < k < n$ *the process will reach* 0 *before it reaches* n *with probability*

$$p_k = \frac{\left(\frac{\beta}{\alpha}\right)^k - \left(\frac{\beta}{\alpha}\right)^n}{1 - \left(\frac{\beta}{\alpha}\right)^n}.$$

The waiting time of reaching n *- conditional on* n *being reached before* 0 *- is bounded above by* $\frac{1}{\alpha-\beta} + o\left(\frac{1}{n}\right)$.

Proof: The conditional probability p_k has to fulfill the following standard difference equation for $0 < k < n$:

$$\begin{aligned} p_k &= \alpha p_{k+1} + \beta p_{k-1} + (1 - \alpha - \beta) p_k \\ p_n &= 0 \\ p_0 &= 1 \end{aligned} \tag{A.1}$$

For the second part of the lemma denote the conditional waiting time (measured in discrete time periods) starting from $0 < k \leq n$ with w_k. The following equations have tc be fulfilled (for the last one remember that the process is conditioned *not* to jump to 0):

$$\begin{aligned} w_k &= \alpha (w_{k+1} + 1) + \beta (w_{k-1} + 1) + (1 - \alpha - \beta)(w_k + 1) \\ w_n &= 0 \\ w_1 &= \frac{\alpha}{1-\beta}(w_2 + 1) + \frac{1-\alpha-\beta}{1-\beta}(w_1 + 1) \end{aligned} \tag{A.2}$$

This system can be solved such that $w_1 \approx \frac{n}{\alpha-\beta} + const.$ For the result to follow note, that $w_k \leq w_1$ and that each discrete time period has duration $\frac{1}{n}$. QED

A.3 Proof of Theorem 1

Because of the global geometry the space of configurations can be collapsed onto the reduced state space $Z' = \left\{0, \frac{1}{n}, \frac{2}{n}, .., 1\right\}$ representing the possible share of blacks in the bounded neighborhood. The discrete counterpart of the residential neighborhood process is now described by the following Markov matrix P^ϵ (see appendix A.1 for construction):

$$P^\epsilon\left(x, x'\right) = \begin{cases} (1-x)\, g_w^\epsilon(x) & \text{if } x' = x + \frac{1}{n} \\ x g_b^\epsilon(1-x) & \text{if } x' = x - \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

The process has an ergodic distribution μ_n on a bounded neighborhood of size n. For notational convenience I denote the probability of jumping from $\frac{m}{n}$ to $\frac{m-1}{n}$ with a_m and the probability of jumping from $\frac{m}{n}$ to $\frac{m+1}{n}$ with b_m. The ergodic distribution then has to fulfill the stationarity condition:

$$\begin{aligned} (a_m + b_m)\, \mu_n\left(\frac{m}{n}\right) &= a_{m+1}\mu_n\left(\frac{m+1}{n}\right) \\ &+ b_{m-1}\mu_n\left(\frac{m-1}{n}\right) \quad \text{for } 0 < m < n \\ a_1\mu_n\left(\frac{1}{n}\right) &= b_0\mu_n(0) \\ b_{n-1}\mu_n\left(\frac{n-1}{n}\right) &= a_n\mu_n(1) \end{aligned} \tag{A.3}$$

One can then show by induction:

$$\mu_n\left(\frac{m}{n}\right) = \frac{a_{m+1}}{b_m}\mu_n\left(\frac{m+1}{n}\right) \tag{A.4}$$

Recall, that there are at most three black shares where $a_m = b_m$:

$$\begin{aligned} x_1 &= \tfrac{(1-\lambda)\epsilon}{(1-\lambda)\epsilon+\lambda} \\ x_2 &= 1-\lambda \qquad \text{if } 1-\alpha_w < \lambda < \alpha_b \\ x_3 &= \tfrac{1-\lambda}{1-\lambda+\lambda\epsilon} \end{aligned}$$

It can be easily checked that the random walk exhibits a drift towards x_1 on $B_1 = [0, 1 - \alpha_b)$, towards x_2 on $B_2 = (1 - \alpha_b, \alpha_w)$ and towards x_3 over $B_3 = (\alpha_w, 1]$.

This observation is sufficient to show that the process will be found with probability approaching 1 inside any neighborhood of $\{x_1, x_2, x_3\}$. Consider any δ-neighborhood of x_1 for example (i.e. $I = (x_1 - \delta, x_1 + \delta)$). For $x < x_1 - \frac{\delta}{2}$ one can deduce that

$$\frac{a_{m+1}}{b_m} \leq C_{\epsilon,\delta} < 1,$$

while for $x_1 + \frac{\delta}{2} < x < 1 - \alpha_b$ the following holds:

$$\frac{b_m}{a_{m+1}} \leq C'_{\epsilon,\delta} < 1.$$

Let $\tilde{C}_{\epsilon,\delta} = \max\left(C_{\epsilon,\delta}, C'_{\epsilon,\delta}\right)$. This implies that $\mu_n(x) \leq \left[\tilde{C}_{\epsilon,\delta}\right]^{\frac{\delta n}{2}}$ for any $x \in B_1 - I$. Therefore the probability of finding the process inside $B_1 - I$ is at most

$$(1 - \alpha_b - 2\delta)\, n \left[\tilde{C}_{\epsilon,\delta}\right]^{\frac{\delta n}{2}},$$

which tends to 0 as $n \to \infty$. The same exercise can be repeated for x_2 and x_3 which establishes the claim.

Next, I show that the process has a vanishing probability weight around x_1. Using formula A.4 repeatedly one can derive:[1]

$$\begin{aligned} \mu_n(x_1) &= \mu_n(x_3) \prod_{m=x_1 n}^{(1-\alpha_b)n} \frac{\lambda^{\frac{m}{n}}}{(1-\lambda)\left(1-\frac{m}{n}\right)\epsilon} \times \\ &\times \prod_{m=(1-\alpha_b)n}^{\alpha_w n} \frac{\lambda^{\frac{m}{n}}}{(1-\lambda)\left(1-\frac{m}{n}\right)} \prod_{m=\alpha_w n}^{x_3 n} \frac{\lambda^{\frac{m}{n}}\epsilon}{(1-\lambda)\left(1-\frac{m}{n}\right)} \end{aligned} \tag{A.5}$$

We know that $1 - x_1 < x_3$ as $\lambda < \frac{1}{2}$ which allows us to simplify the expression:

$$\mu_n(x_1) = \mu_n(x_3) \left(\frac{\lambda}{1-\lambda}\right)^{(1-2x_1)n} \epsilon^{(\alpha_b - \alpha_w)n} \prod_{m=(1-x_1)n}^{x_3 n} \frac{\epsilon \lambda^{\frac{m}{n}}}{(1-\lambda)\left(1-\frac{m}{n}\right)} \tag{A.6}$$

Note, that $\epsilon \lambda x_3 \leq (1-\lambda)(1-x_3)$. Therefore every term in the product on the left is less

[1] All expressions hold up to an integer constraint. As n becomes large the finite number of misplaced terms in the product have a vanishing influence and are therefore omitted.

than 1 and one obtains the inequality:

$$\mu_n(x_1) \leq \mu_n(x_3)\left[\left(\frac{\lambda}{1-\lambda}\right)^{(1-2x_1)n} \epsilon^{\alpha_b-\alpha_w}\right]^n \tag{A.7}$$

As $\alpha_b > \alpha_w$ or $\lambda < \frac{1}{2}$ one finds again that

$$\mu_n(x_1) \leq [F_\epsilon]^n$$

for $F_\epsilon < 1$. More generally, one can repeat the exercise for any x' in some small δ-neighborhoods of x_1 because the inequality $1 - x_1 < x_3$ is strict. One then obtains

$$\mu_n(x') \leq [F_{\epsilon,\delta}]^n$$

for $F_{\epsilon,\delta} < 1$. As before this implies that the probability weight inside the δ-neighborhood vanishes as $n \to \infty$.

It remains to be determined whether the process clusters around x_2 or around x_3. Using formula A.4 again one obtains:

$$\mu_n(x_2) = \mu_n(x_3)\left(\frac{\lambda}{1-\lambda}\right)^{(x_3-x_2)n} \prod_{m=x_2 n}^{x_3 n} \frac{\frac{m}{n}}{1-\frac{m}{n}} \epsilon^{(x_3-\alpha_w)n} \tag{A.8}$$

We now use the fact that[2]

$$\left[\prod_{m=x_2 n}^{x_3 n} \frac{\frac{m}{n}}{1-\frac{m}{n}}\right]^{\frac{1}{n}} = \frac{x_3^{x_3}(1-x_3)^{1-x_3}}{x_2^{x_2}(1-x_2)^{1-x_2}} + o\left(\frac{1}{n}\right).$$

This implies, that

$$\mu_n(x_2) = \mu_n(x_3)\left[\tilde{F}_\epsilon\right]^n \tag{A.9}$$

[2] Note, that

$$\left[\prod_{m=x_2 n}^{x_3 n} \frac{\frac{m}{n}}{1-\frac{m}{n}}\right]^{\frac{1}{n}} = \exp\left(\int_{x_2}^{x_3} \ln\frac{t}{1-t}dt\right) + o\left(\frac{1}{n}\right).$$

where

$$\begin{aligned}\tilde{F}_\epsilon &= \left(\frac{\lambda}{1-\lambda}\right)^{x_3-x_2} \frac{x_3^{x_3}\left(1-x_3\right)^{1-x_3}}{x_2^{x_2}\left(1-x_2\right)^{1-x_2}} \quad \epsilon^{x_3-\alpha_w} + o\left(\frac{1}{n}\right) \\ &= \frac{\epsilon^{1-\alpha_w}}{1-\lambda+\lambda\epsilon} + o\left(\frac{1}{n}\right).\end{aligned}$$

As before one can show that the process vanishes on some δ-neighborhood of x_2 (x_3) if $\tilde{F}_\epsilon < 1$ ($\tilde{F}_\epsilon > 1$). The process clusters then over x_3 (x_2).

Finally, it is easy to show that the medium-run behavior of the process is determined by the initial conditions alone. Consider for example the case where the initial share of black residents satisfies $x_1 + \delta < x_0 < 1 - \alpha_b$. Using lemma 6 one can deduce that the probability of reaching $1 - \alpha_b$ before reaching $x_1 + \delta$ goes to zero exponentially as $n \to \infty$ and the conditional waiting time for reaching the δ-neighborhood is bounded above by some finite W_δ. QED

A.4 Proof of Lemma 2

In both marginal processes a switch at some cell z is associated with an 'event'. For the original process η_t that event is simply the switch of color at that particular cell. For the simplified process σ_t on the other hand only a white to black switch always involves just one cell. A black to white switch might cause a number of cells to the right or left to switch too, and in this case I call the joint switching of all those cells the event corresponding to the switch at z.

Think of two identical streets $G_S(n)$ such that σ_t moves on street A and η_t on street B. At any point in time and for any cell z the cells on both streets switch as follows. If $\sigma_t(z) \neq \eta_t(z)$ the cells will switch independently and trigger off the associated event at the rates specified by the switching rule. If $\sigma_t(z) = \eta_t(z)$ both cells will flip together and trigger off the associated events with as large a rate as possible consistent with the requirement that each process flips at the correct rate. In other words if z switches on street A with rate c_1 and on street B with rate c_2 (WLOG assume $c_1 < c_2$) then they flip together with rate c_1 and on street B the cell will additionally flip at an independent rate of $c_2 - c_1$. This procedure defines a coupled process (σ_t, η_t) with the correct marginal processes σ_t and η_t.[3]

[3] This coupling is a more elaborate variant of the basic Vasershtein coupling for spin systems (Liggett

It remains to be shown that $\sigma_t \leq \eta_t$ with probability 1 at any point in time t. At random times t_i $(i = 0, 1, 2, ..)$ cells on street A or street B switch $(t_0 = 0)$. There are two types of switches - joint flips which trigger off the associated events simultaneously on both streets and independent switches which involve either only street A or street B. I show by induction on i that $\sigma_{t_i} \leq \eta_{t_i}$ for all $i \geq 0$ which implies of course that $\sigma_t \leq \eta_t$ for all $t \geq 0$.

$i = 0$ The claim is true by assumption because $\sigma_0 = \eta_0$.

$i \rightarrow i+1$ Assume the claim holds for i such that $\sigma_{t_i} \leq \eta_{t_i}$. First, assume the switch at time t_{i+1} at cell z is an independent switch on either street A or street B. If $\sigma_{t_i}(z) \neq \eta_{t_i}(z)$ we must have $\sigma_{t_i}(z) = 0$ and $\eta_{t_i}(z) = 1$. The associated event then involves in any case just the cell z and therefore $\sigma_{t_{i+1}} \leq \eta_{t_{i+1}}$. If $\sigma_{t_i}(z) = \eta_{t_i}(z) = 1$ the independent switch must have occurred on street A because the flip rate from black to white is increasing in the share of white neighbors of cell z and $y(\sigma_{t_i}, z) \geq y(\eta_{t_i}, z)$. The associated event flips one or more cells on street A from black to white such that $\sigma_{t_{i+1}} \leq \eta_{t_{i+1}}$. If $\sigma_{t_i}(z) = \eta_{t_i}(z) = 0$ the independent switch must have occurred on street B for analogous reasons. Cell z is now occupied by a black resident such that again $\sigma_{t_{i+1}} \leq \eta_{t_{i+1}}$.

Finally assume that the switch at time t_{i+1} at cell z is a simultaneous switch on both streets. If $\sigma_{t_i}(z) = \eta_{t_i}(z) = 1$ then on street A possibly some additionally cells switch from black to white so that certainly $\sigma_{t_{i+1}} \leq \eta_{t_{i+1}}$. If $\sigma_{t_i}(z) = \eta_{t_i}(z) = 0$ only the cell z on both streets changes from white to black so that again $\sigma_{t_{i+1}} \leq \eta_{t_{i+1}}$. This proves the inductive hypothesis. QED

A.5 Proof of Lemma 3

I will prove the lemma in detail for simple streets $(r = 1)$ where the minimally stable cluster has size 1. At the end I sketch how the proof generalizes to streets with radius of interaction greater than 1.

I use coupling once more in order to further simplify the process σ_t which is now restricted to a single segment. I start by replacing the space of configurations Z over which σ_t evolves. Because each segment contains at most one black cluster by construction, a configuration is completely described by the share of blacks in the set $Z' = \left\{0, \frac{1}{N}, .., 1\right\}$ and the position of the first clockwise black resident in the set $\tilde{Z} = \{1, 2, .., N\}$. We can then

1985, chapter 3).

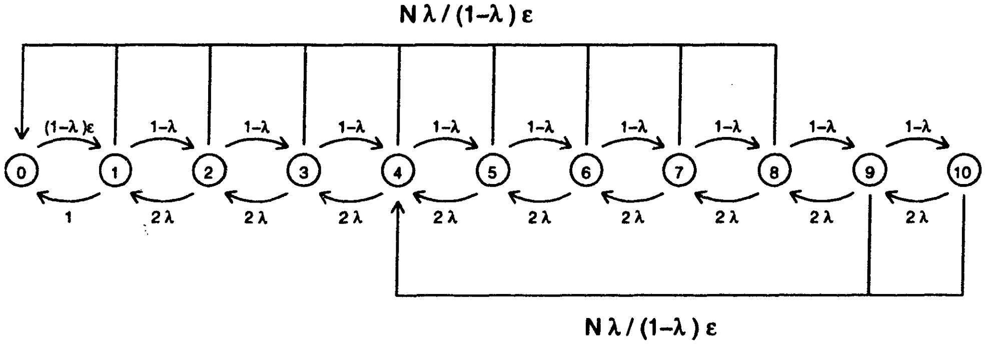


Figure A-1: Dynamics of the random walk ξ_t^δ for $N = 10$ and $\delta = 0.1$

say that σ_t evolves on $Z' \times \tilde{Z}$.

For a fixed $\delta > 0$ I construct a continuous-time random walk ξ_t^δ on Z' with the following transition rate $c_\xi(x_1, x_2)$ between the states of Z':

$$c_\xi(x_1, x_2) = \begin{cases} (1-\lambda)\epsilon & \text{if } x_1 = 0 \text{ and } x_2 = \frac{1}{N} \\ 1-\lambda & \text{if } x_1 > 0 \text{ and } x_2 = x_1 + \frac{1}{N} \\ 2\lambda & \text{if } x_1 > \frac{1}{N} \text{ and } x_2 = x_1 - \frac{1}{N} \\ 1 & \text{if } x_1 = \frac{1}{N} \text{ and } x_2 = 0 \\ N\frac{\lambda}{1-\lambda}\epsilon & \text{if } x_1 \geq 1-\delta \text{ and } x_2 = \frac{1-\delta}{2} \\ N\frac{\lambda}{1-\lambda}\epsilon & \text{if } x_1 < 1-\delta \text{ and } x_2 = 0 \\ 0 & \text{otherwise} \end{cases}$$

Figure A-1 illustrates the dynamics of this random walk. It resembles a simple random walk with a drift determined by λ with the added capability of making 'large' jumps. With probability $N\frac{\lambda}{1-\lambda}\epsilon$ the process jumps to an intermediate state $\frac{1-\delta}{2}$ if the share of black residents is larger than $1-\delta$ and to 0 otherwise.

The random walk ξ_t^δ and the process σ_t are now coupled in the following way. Both evolve from the initial states 0 and $(0,1)$ respectively, i.e. from an 'all-white' configuration. Any transition which increases (decreases) the share of blacks ξ_t^δ and $X(\sigma_t)$ is called an 'upward' jump ('downward' jump). I further distinguish between (small) downward jumps due to the undisturbed dynamics ('normal jumps') and those (potentially large ones) which are caused by the disturbance ('ϵ-jumps'). For the coupled process (ξ_t^δ, σ_t) the transitions are linked by the following rules:

1. Both processes jump upwards independently if $\xi_t^\delta \neq X(\sigma_t)$ and otherwise jump simultaneously with as large a possible rate consistent with the requirement that both jump at the correct rate.

2. For normal downward jumps adopt the same convention as for upward jumps. For downward ϵ-jumps let both processes jump simultaneously with as large a possible rate consistent with the requirement that both jump at the correct rate.

The definition of the random walk ξ_t^δ ensures that the share of blacks is always less likely to increase and more likely to contract compared to the process σ_t. Using the same technique as in appendix A.4 one can then show that coupling preserves the inequality $\xi_t^\delta \leq X(\sigma_t)$ at any point in time with probability 1.

Therefore, it is sufficient to show that there exists $\bar{\epsilon}$ such that for all $\epsilon < \bar{\epsilon}$ the expected long-run share of blacks $E_b(\epsilon)$ of the random walk ξ_t^δ fulfills $E_b(\epsilon) > 1 - 2\delta$ for some N. The result can only be true if the random walk has a positive drift i.e. $\lambda < \frac{1}{3} \leq \hat{\lambda}$. Denote the expected waiting time to jump out of the interval $(1-\delta, 1]$ starting from $x = 1$ with W_{out}. In this case the process can be found either at $x = 1 - \delta$ or at $x = \frac{1-\delta}{2}$. Hence the expected waiting time to reach again the state $x = 1$ is at most the expected waiting time to get from $\frac{1-\delta}{2}$ to $x = 1$ which I denote with W_{in}. The expected long-run share of blacks can then be bounded below as follows:[4]

$$E_b(\epsilon) \geq (1-\delta)\frac{W_{out}}{W_{in} + W_{out}} \tag{A.10}$$

We now just have to find a lower bound for W_{out} and an upper bound for W_{in}.

I set $\epsilon = N^{-4}$. For large N the waiting time to jump out of the interval $(1-\delta, 1]$ through single downward jumps grows exponentially with N due to the positive drift of the random walk while the waiting time to leave the interval through an ϵ-jump is just $\frac{1-\lambda}{N\lambda\epsilon}$ which is only of the order $O(N^3)$. Therefore only the latter event matters and W_{out} can be bounded below as follows:

$$W_{out} \geq \frac{1}{2}\frac{1-\lambda}{\lambda}N^3 \tag{A.11}$$

Using the same techniques as in appendix A.2 the waiting time $\hat{W}$ for reaching $x = 1$

[4] The process spends at least a share $\frac{W_{out}}{W_{in}+W_{out}}$ of the time in the interval $(1-\delta, 1]$.

starting from $\frac{1-\delta}{2}$ *conditional on no ϵ-jumps occurring* is bounded above by $\frac{N}{1-3\lambda} = AN$.[5] The probability that an ϵ-jump occurs is at most

$$AN \times N\frac{\lambda}{1-\lambda}\epsilon,$$

which is of the order $O\left(N^{-2}\right)$. In the worst case the process ends up at $x = 0$ after such an ϵ-jump. The waiting time to reach $x = 1$ conditional on no further ϵ-jump occurring can the be calculated as[6]

$$\frac{1}{(1-\lambda)(1-3\lambda)\epsilon} + \frac{N}{1-3\lambda} + O\left(N^{-1}\right).$$

Because ϵ-jumps can only occur for $x > 0$ the probability for such an event is as before of the order $O\left(N^{-2}\right)$. Therefore the unconditional waiting time $\tilde{W}$ to reach $x = 1$ starting from $x = 0$ is bounded above by

$$\frac{1}{1 - O\left(N^{-2}\right)} \times \quad \text{conditional waiting time.}$$

The various estimates allow me to bound the waiting time W_{in} from above as follows:

$$\begin{aligned} W_{in} &\leq \hat{W} + AN^2\frac{\lambda}{1-\lambda}\epsilon\tilde{W} \\ &\leq AN + AN^2\frac{\lambda}{1-\lambda}\epsilon \times 2\left[\frac{A}{(1-\lambda)\epsilon} + AN\right] \\ &\leq AN + 2A^2N^2\frac{\lambda}{(1-\lambda)^2} + 2A^2\frac{\lambda}{1-\lambda}\frac{1}{N} \end{aligned} \qquad (A.12)$$

Plugging the bounds for W_{out} and W_{in} into expression A.10 delivers:

$$E_b(\epsilon) \geq (1-\delta)\left(1 - O\left(N^{-1}\right)\right) \qquad (A.13)$$

for $\epsilon = N^{-4}$. But this proves the lemma: simply choose $\overline{N}$ large enough such that $E_b(\epsilon) > 1 - 2\delta$ for all $N > \overline{N}$ and take $\overline{\epsilon} = \overline{N}^{-4}$.

Finally I briefly discuss how to generalize the proof to radii of interaction $r > 1$. There

[5] More precisely, it is $\frac{N(1-\delta)}{2(1-3\lambda)} + O\left(N^{-1}\right)$.

[6] The first term captures the waiting time to jump out of $x = 0$ corrected for the fact that the process visits $x = 0$ on average $\frac{1}{1-3\lambda}$ times before reaching $x = 1$. The second term corresponds to the simple waiting time for reaching $x = 1$ if the process would start at $x = \frac{1}{N}$.

are two complications. First, the single black cluster on the segment grows by increments of 1 but can shrink by up to r in a single transition under the undisturbed dynamics of the process σ_t. The share of whites λ in the housing market therefore has to be low enough such that the associated random walk ξ_t^δ has a positive drift. Second, the size of the minimally stable cluster is now generally some $b > 1$ and the dynamics requires b 'mutations' in order to jump out of the basin of attraction of $x = 0$. The approximation A.12 for W_{in} will not work any longer as ϵ does not cancel. This problem can be overcome by choosing more intermediate states for the random walk ξ_t^δ to which the process can move after a downward ϵ-jump. Simply define:

$$\begin{aligned}
x_1^* &= \frac{1-\delta}{2} \\
x_2^* &= \frac{x_1^* - \delta}{2} \\
&\dots \\
x_b^* &= \frac{x_{b-1}^* - \delta}{2}
\end{aligned}$$

I then postulate that after an ϵ-event the random walk jumps to x_1^* if $\xi_t^\delta > 1 - \delta$ and to x_i^* if $x_{i-2}^* - \delta \geq \xi_t^\delta > x_{i-1}^* - \delta$ for $1 < i \leq b + 1$ (I set $x_0^* = 1$ and $x_{b+1}^* = 0$). The process will therefore reach $x = 0$ from x_1^* before it reaches $x = 1$ only with a probability of order $O\left(\epsilon^b\right)$ such that my approximation for W_{in} holds again. QED

A.6 Proof of Lemma 1

For the proof I exploit the results from the proof of theorem 2 by taking $\epsilon = N^{-4}$ again. Through judicious coupling I constructed a random walk ξ_t^δ such that $\xi_t^\delta \leq X(\eta_t)$ with probability 1. Therefore it is sufficient to prove the hypothesis of the lemma for k independent random walks ξ_t^δ and let $k \to \infty$.

The expected waiting time W until the random walk ξ_t^δ reaches $x = 1$ starting from a share of blacks $x = 0$ can be bounded using the results from the previous section:

$$W \leq B\epsilon^{-b} + AN$$

The random walk can be subsequently found outside the $\frac{\delta}{2}$-neighborhood of $x = 1$ with probability $O\left(N^{-1}\right)$ (see appendix A.5). This implies that after a time $\frac{2W}{\delta}$ the process has reached $x = 1$ once with probability $1 - \frac{\delta}{2}$ and can therefore be found inside the $\frac{\delta}{2}$ neighborhood with probability $1 - \frac{\delta}{2} - O\left(N^{-1}\right)$.[7]

For large k the normal approximation of the binomial distribution can be used to show that the share of blacks on the street is at least $1 - \frac{\delta}{2} - \frac{2}{\sqrt{k}} - O\left(\frac{1}{N}\right)$ with probability 0.95 after time $\frac{2W}{\delta}$.[8] So we simply choose $k > \frac{16}{\delta^2}$ and the share of blacks has to be at least $1 - \delta$ after time $\frac{2W}{\delta}$ with probability 0.95. Therefore the expected time to reach the δ neighborhood of $x = 1$ is at most $\frac{1}{1-0.95}\frac{2W}{\delta}$. QED

A.7 Proof of Lemma 4

Consider an encircled cluster C of white residents such as shown in figure A-2 which forms the initial configuration of the residential neighborhood process, i.e. $\eta_0 = \chi_C$ where $\chi_C(z) = 0$ iff $z \in C$. By definition this cluster can be covered by a suitable rectangle S and $\eta_0 = \chi_C \geq \chi_S$.

First, I prove that $\eta_t \geq \chi_S$ for all $t > 0$ such that the white cluster can never 'break out' of S. This is equivalent to showing that after a sequence of individual residents switching at random times $t_0 = 0, t_1, t_2, .., t_i, ..$ the inequality $\eta_{t_i} \geq \chi_S$ holds. This claim is proved by induction. For $i = 0$ it is true by assumption. Assume it holds for $i - 1$. Then the switch from $\eta_{t_{i-1}}$ to η_{t_i} involves either a resident $z \in S$ or $z \notin S$. In the former case we are fine. The latter case is not possible, as for all $z \notin S$ and the configuration χ_S the share of black neighbors $x(\chi_S, z)$ of z is at least $\frac{1}{2} + \frac{r}{m}$. By the inductive hypothesis $\eta_{t_{i-1}} \geq \chi_S$ and hence $x\left(\eta_{t_{i-1}}, z\right) \geq x(\chi_S, z)$. Due to assumption 1 and the absence of tolerant agents no white house-seekers will ever be interested in moving to z.

Second, I prove that all configurations $\eta \geq \chi_S$ except the black ghetto configuration η_b are transient states of the Markov process. This establishes the result because the black ghetto is absorbing. It is sufficient to show that there exists a positive transition probability from any configuration $\eta \geq \chi_S$ to η_b. This will be guaranteed if for all configurations $\eta \geq \chi_S$ of cluster mass m there exists a positive transition probability to a configuration η' of cluster

[7] A random variable X with expectation $E(X)$ has to fulfill $P\left(X \leq \frac{E(X)}{\delta}\right) \geq 1 - \delta$.

[8] For a random variable X the following relation holds: $P(|X - E(X)| \leq 2\sigma) \geq 0.95$, where σ denotes the standard deviation.

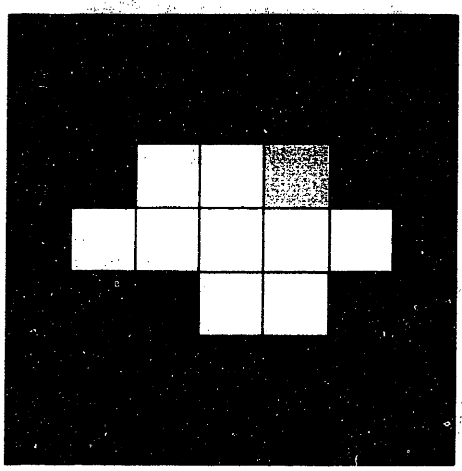


Figure A-2: White "encircled" cluster in an inner-city with radius $r = 1$: white resident at upper right-hand corner (shaded) has two white and two black neighbors

mass $m - 1$ (i.e. a single white resident is replaced by a black resident) - in this case there exists with positive probability a sequence of m switches from η to η_b.

Consider a configuration $\eta \geq \chi_S$ of cluster mass $m > 0$. Choose any natural Cartesian coordinate system on the geometry G_C^r. Then find the upper right-hand corner z^* of the cluster (see figure A-2). The share of white neighbors is at most $\frac{1}{2}$. Therefore, black house seekers will not feel isolated and show interest in the flat and the ethnicity of the resident at flat z^* will switch with positive probability of at least $1 - \lambda$. QED

A.8 Proof of Theorem 4

It is sufficient to count the number of 'mutations' (i.e. completely tolerant house-seekers) that are necessary to leave the basin of attraction of the black ghetto configuration η_b. Ellison (1999) defines that number to be the radius $R(\eta_b)$ and the result follows directly from his theorem 2 (see appendix A.1 for transforming the residential neighborhood process into an equivalent discrete time process).

On streets the result is trivial because a cluster of size $[2r(1 - \alpha_w)]^+$ is minimally stable. In inner-cities it suffices to show that $\left[\frac{\sqrt{n}}{r+1}\right] - 1$ mutations do not allow the process to leave the basin of attraction of the black ghetto. The inner-city can be divided up into $k = \left[\frac{\sqrt{n}}{r+1}\right]$ full horizontal stripes H_i $(i = 1..k)$ and equally many vertical stripes V_j $(j = 1..k)$ of width $r + 1$. Assume that $\left[\frac{\sqrt{n}}{r+1}\right] - 1$ tolerant agents moved into the inner-city up to time

t. Then there must be at least one horizontal stripe H_{i^*} where no mutations has occurred. Due to assumption 1 all residents on that stripe have to be black (no white cluster could have invaded that stripe from outside). For the same reason there must be at least one vertical stripe V_{j^*} where all residents are still black at time t. But this implies that all existing white residents at time t can be covered by a rectangle of length and width at most $\sqrt{n} - r - 1$. Therefore, the set of whites is "encircled" and the configuration η_t is in the basin of attraction of the black ghetto by lemma 4. Hence at least $\left[\frac{\sqrt{n}}{r+1}\right]$ mutations are necessary to jump out of $D(\eta_b)$. QED

A.9 Proof of Theorem 6

I will analyze the discrete time counterpart of the imitation process with transition matrix P^ϵ which can be derived as in appendix A.1. I index all the $\binom{n}{m}$ configurations $\eta \in Z$ where exactly m agents have opinion 1 by

$$\left\{\eta_{m,1}, \eta_{m,2} ... \eta_{m,\binom{n}{m}}\right\}.$$

I also introduce a reduced state space $Z' = \left\{0, \frac{1}{n}, \frac{2}{n} .., 1\right\}$ representing the possible shares of agents having opinion 1. The ergodic distribution μ_n^I of the imitation process on the graph $G(n)$ induces a corresponding distribution μ_n on the reduced state space defined by

$$\mu_n\left(\frac{m}{n}\right) = \sum_{i=1}^{\binom{n}{m}} \mu_n^I(\eta_{m,i}).$$

Next, I derive a transition matrix P_n describing a Markov chain on Z' which generates the ergodic distribution μ_n. For any configuration $\eta \in Z$ I index all $v(\eta)$ configurations which generate η when a single agent changes her opinion from 0 to 1 by

$$\left\{\eta^{-,1}, \eta^{-,2} ... \eta^{-,v(\eta)}\right\}.$$

Similarly, I index all $w(\eta)$ configurations which generate η when a single agent changes her opinion from 1 to 0 by

$$\left\{\eta^{+,1}, \eta^{+,2} ... \eta^{+,v(\eta)}\right\}.$$

For the ergodic distribution μ_n^I the probability 'outflow' from some configuration $\eta_{m,i} \in Z$ has to equal the probability 'inflow':[9]

$$\sum_{\mu \neq \eta_{m,i}} P^\epsilon\left(\eta_{m,i}, \mu\right) \mu_n^I\left(\eta_{m,i}\right) = \sum_{j=1}^{w(\eta_{m,i})} P^\epsilon\left(\eta_{m,i}^{+,j}, \eta_{m,i}\right) \mu_n^I\left(\eta_{m,i}^{+,j}\right) + \sum_{j=1}^{v(\eta_{m,i})} P^\epsilon\left(\eta_{m,i}^{-,j}, \eta_{m,i}\right) \mu_n^I\left(\eta_{m,i}^{-,j}\right)$$

After summing over the $\eta_{m,i}$ configurations one obtains:

$$\begin{aligned} &\left(\frac{\sum_i \mu_n^I\left(\eta_{m,i}\right) \sum_j P^\epsilon\left(\eta_{m,i}, \eta_{m,i}^{-,j}\right)}{\mu_n\left(\frac{m}{n}\right)} + \frac{\sum_i \mu_n^I\left(\eta_{m,i}\right) \sum_j P^\epsilon\left(\eta_{m,i}, \eta_{m,i}^{+,j}\right)}{\mu_n\left(\frac{m}{n}\right)}\right) \mu_n\left(\frac{m}{n}\right) \\ &\qquad = \frac{\sum_i \mu_n^I\left(\eta_{m+1,i}\right) \sum_j P^\epsilon\left(\eta_{m+1,i}, \eta_{m+1,i}^{-,j}\right)}{\mu_n\left(\frac{m+1}{n}\right)} \mu_n\left(\frac{m+1}{n}\right) \\ &\qquad + \frac{\sum_i \mu_n^I\left(\eta_{m-1,i}\right) \sum_j P^\epsilon\left(\eta_{m-1,i}, \eta_{m-1,i}^{+,j}\right)}{\mu_n\left(\frac{m-1}{n}\right)} \mu_n\left(\frac{m-1}{n}\right) \end{aligned} \tag{A.14}$$

I denote the probability of jumping from $\frac{m}{n}$ to $\frac{m-1}{n}$ with a_m and the probability of jumping from $\frac{m}{n}$ to $\frac{m+1}{n}$ with b_m such that equation A.14 becomes:

$$\left(a_m + b_m\right) \mu_n\left(\frac{m}{n}\right) = a_{m+1} \mu_n\left(\frac{m+1}{n}\right) + b_{m-1} \mu_n\left(\frac{m-1}{n}\right)$$

The following transition matrix P_n gives rise to a Markov chain with ergodic distribution μ_n:

$$P_n\left(x, x'\right) = \begin{cases} a_m & \text{if } x = \frac{m}{n} \text{ and } x' = \frac{m-1}{n} \\ b_m & \text{if } x = \frac{m}{n} \text{ and } x' = \frac{m+1}{n} \\ 0 & \text{otherwise} \end{cases}$$

The Markov chain described by P_n has the same structure as the process discussed in appendix A.3. I can therefore use formula A.4:

$$\mu_n\left(\frac{m}{n}\right) = \frac{a_{m+1}}{b_m} \mu_n\left(\frac{m+1}{n}\right)$$

[9] Note, that the imitation process can reach $\eta_{m,i}$ only due to a single switch by one agent.

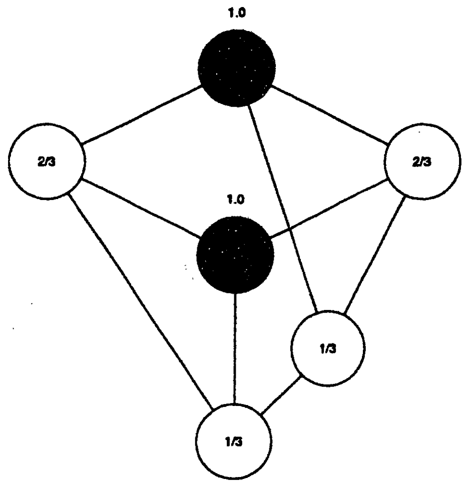


Figure A-3: Proper graph of order $q = 3$ and a configuration with exposure $e(\eta) = 2$

It is advantageous to transform the parameters in the following way. I define $\mu_n^*\left(\frac{m}{n}\right) = (a_m + b_m)\mu_n\left(\frac{m}{n}\right)$, $a_m^* = \frac{a_m}{a_m+b_m}$ and $b_m^* = \frac{b_m}{a_m+b_m}$. Formula A.4 is preserved by these transformations:

$$\mu_n^*\left(\frac{m}{n}\right) = \frac{a_{m+1}^*}{b_m^*}\mu_n^*\left(\frac{m+1}{n}\right) \tag{A.15}$$

The ratio $\frac{a_{m+1}^*}{b_m^*}$ can be bounded by exploiting the linearity of imitation. For any configuration η I define weights w_z for each agent z on the graph such that whenever (z, z') is an edge of the graph and agents z and z' use different actions then $\frac{1}{q}$ is added to the weights of both agents. These weights are exactly the switching probabilities of each agent under the undisturbed imitation process. From the construction and the properness of the graph it is immediately clear that the sum of weights of agents with opinion 0 equals the sum of weights of agents with opinion 1. This observation formalizes exactly the intuition that the imitation effect 'cancels out'. I call the sum of weights the *exposure* $e(\eta)$ of the configuration η. Clearly $0 \leq \frac{e(\eta)}{n} \leq \min(x, 1-x)$. Figure A-3 demonstrates the algorithm on a small graph of order 3.

I can now conveniently express the probability that the process jumps from some given configuration η with m agents having opinion 1 to some configuration η' where $m+1$ agents have opinion 1:

$$(1-\epsilon)\frac{e(\eta)}{n}+\epsilon(1-x)$$

Similarly, the probability that the process jumps from η to η'' where $m-1$ agents have opinion 1 can be calculated as:

$$(1-\epsilon)\frac{e(\eta)}{n}+\epsilon x$$

I can then deduce

$$\begin{aligned} a_m &= (1-\epsilon)h(x)+\epsilon x \\ b_m &= (1-\epsilon)h(x)+\epsilon(1-x), \end{aligned} \tag{A.16}$$

where $h(x)=\frac{\sum_{i=1}\mu_n^I(\eta_{m,i})e(\eta_{m,i})}{n\mu_n(m)}$. Note, that

$$0 \leq h(x) \leq x.$$

Hence, for $x<\frac{1}{2}$ the following inequalities must hold:

$$\begin{aligned} a_m^* &= \frac{(1-\epsilon)h(x)+\epsilon x}{2(1-\epsilon)h(x)+\epsilon} \leq \frac{1-\epsilon}{2}+\epsilon x \\ b_m^* &= \frac{(1-\epsilon)h(x)+\epsilon(1-x)}{2(1-\epsilon)h(x)+\epsilon} \geq 1-x \\ \frac{a_{m+1}^*}{b_m^*} &\leq \frac{\frac{1-\epsilon}{2}+\epsilon x}{1-x}+O\left(\frac{1}{n}\right) \end{aligned} \tag{A.17}$$

For $x<\frac{1}{2}-\frac{\delta}{2}$ the following uniform bound holds:

$$\frac{a_{m+1}^*}{b_m^*} \leq \frac{\frac{1-\epsilon}{2}+\epsilon\frac{1-\delta}{2}}{\frac{1+\delta}{2}}+O\left(\frac{1}{n}\right) \leq C_{\epsilon,\delta}<1$$

Due to the symmetry of the ergodic distribution and repeated use of formula A.15 one can conclude that $\mu_n^*(x) \leq [C_{\epsilon,\delta}]^{\frac{\delta n}{2}}\mu_n^*\left(\frac{1}{2}-\frac{\delta}{2}\right)$ for $\left|x-\frac{1}{2}\right|>\delta$. Recall, that $a_m+b_m>\epsilon$. Therefore the process will be found within the δ-neighborhood of $x^*=\frac{1}{2}$ with a probability of at least

$$1-\frac{1-2\delta}{\epsilon}n\left[C_{\epsilon,\delta}\right]^{\frac{\delta n}{2}},$$

which tends to 1 as $n \rightarrow \infty$. Therefore, the process clusters around x^*.

The second part of the theorem is now easy. The waiting time until the δ-neighborhood is reached is at most as large as the corresponding waiting time of a process which satisfies

$$\frac{a^*_{m+1}}{b^*_m}=C_{\epsilon,\delta}$$

for $x < \frac{1}{2} - \frac{\delta}{2}$ and starts from $x = 0$. Because this random walk has a positive drift the waiting time is $O(n)$ (see appendix A.2). Note, that the discrete 'clock' increases at increments of $\frac{1}{n}$ and the result follows immediately. QED

Appendix B

B.1 Proof of Lemma 5

$C_j^b(t,\mu)$ is clearly increasing in x and therefore in t. For the monotonicity in μ we are fine, if we can show that $C_j^b(t,\mu) \leq C_j^b(t,\mu+\epsilon)$ for a small ϵ and for all t. For the threshold duplication concentrations the following inequality holds:

$$y_A^j(\mu) < y_A^j(\mu') < y_B^j(\mu') < y_B^j(\mu),$$

where $\mu' = \mu + \epsilon$. For $x(t) < y_A^j(\mu)$ the claim is obvious as the function $\mu(1-\mu)$ is increasing in μ for $\mu < \frac{1}{2}$. For $y_A^j(\mu) < x(t) < y_A^j(\mu')$ we get:

$$\begin{aligned} C_j^b(t,\mu) &= (1-\mu)\mu\gamma x(t) M_j v_b + \mu(1-\mu)\gamma y_A^j(\mu) M_j v_b \\ &< (1-\mu')\mu'\gamma x(t) M_j v_b + \mu'(1-\mu')\gamma x(t) M_j v_b \\ &= C_j^b(t,\mu') \end{aligned}$$

The claim is again obvious for $y_A^j(\mu') < x(t) < y_B^j(\mu')$. So we are only left with $y_B^j(\mu') < x(t) < y_B^j(\mu)$. For this case consider:

$$\begin{aligned} C_j^b(t,\mu) &= (1-\mu)\mu\gamma x(t) M_j v_b + \mu K \\ &< (1-\mu)K + \mu K \\ &= C_j^b(t,\mu') \end{aligned}$$

Finally assume that duplication has occurred in the city. If both minority and majority businesses duplicate the lower bound is trivial. Otherwise we obtain

$$\begin{aligned} C_j^b(t,\mu) &= (1-\mu)\mu\gamma x(t) M_j v_b + \mu K \\ &\geq \mu K + \mu K \\ &= 2\mu K \end{aligned}$$

This finishes the proof. QED

B.2 Construction of Business Duplication Rate in Table 3.3

I *define* the number of 'businesses' B_C in a given city C to be 50% of all subscribers in that city in July, 1905. Given that the share of business subscribers varied between 40% and 50% for US cities around 1905 I have chosen a generous estimate. I then calculate the projected duplication rate amongst businesses by expressing the number of duplicates in all previous and later years as a share of B_C. This methodology allows me to compare duplication rates across time and cities by introducing a common reference point.

Some comments are necessary in order to justify this construction.

1. The vast majority of the duplicates were in deed business telephones. On the basis of its survey of 39 cities in 1905, the Merchants' Association of New York (1905, p. 3) found that almost 85% of dual subscribers were businesses. Most of the residential duplicates were in fact physicians and other professionals who would be classified as businesses in my model rather than as residents.[1]

2. Most subscribing businesses after 1905 were either residents or small-scale businesses with low communication demand (i.e. small β_j). Therefore B_C should include all large-scale and medium-scale businesses which subscribed early and received the bulk of all business communication from residents.

3. A 'business' in my model serves a particular social island - medium and small scale stores fit that description but not large-scale businesses. I will neglect the latter group in the data. Large-scale businesses had a duplication rate of close to 100% but made up only a

[1] In Kansas City almost all residential duplicates were physicians' phones (Committee on Gas, Oil and Electric Light 1907, p. 194).

small share of duplicates: in Louisville, for example, only 7.5% of dual users were large-scale businesses in 1910 (Mueller 1997, p. 82). As many of those establishments were up-stream suppliers of medium and small-scale businesses they did not exchange many messages with residents. Due to the extremely high rate of duplication they would not have influenced the network choice of residents in any case.

4. In the context of the model an increase of business duplication is causally linked to the state of development in the city (i.e. the telephone concentration $x(t)$). For this interpretation to be correct we have to exclude other factors which can explain this phenomenon. In particular, a decrease of telephone rates over time would make duplication cheaper and we might expect the business duplication rate to increase even in the absence of network growth. I have therefore included information on the rate schedules of the various companies for unlimited business and residential service in table 3.3. With the exception of Columbus the cost of dual service in fact increased over time.[2] Therefore my business duplication rate is downward biased by not controlling for the subscription rate which works in favor of the model.

[2] Both Bell and Independent operators reported that the telephone industry exhibited decreasing returns to scale. Larger exchanges had to lay their cables underground as the poles could no longer support the increasing weight of the wires. The number of switchboards increased more than proportionally to the number of subscribers as a company had to provide for possible connections between *all* subscribers.

Appendix C

C.1 Proof of Theorems 9 and 11

For the proof of theorem 9 simply set $a = 1$, e.g. all workers are of high-ability. I assume that a share y of the economy utilizes mass production. The price level $p(t)$ in the economy and the demand $x_A(t)$ for artisan goods and $x_{C0}(t)$ for industrial goods can be derived from equations 4.2 and 4.3:

$$\begin{aligned} p(t) &= \frac{c_M(t)\,\omega}{\rho}\left[y+(1-y)\left(\mu^m\frac{c_M(t)\,\omega}{c_A\rho}\right)^{\frac{\rho}{1-\rho}}\right]^{\frac{\rho-1}{\rho}} && \text{(C.1)} \\ x_{C0}(t) &= \frac{E\rho}{c_M(t)\,\omega}\frac{1}{y+(1-y)\left(\mu^m\frac{c_M(t)\omega}{c_A\rho}\right)^{\frac{\rho}{1-\rho}}} && \text{(C.2)} \\ x_A(t) &= x_{C0}(t)\,(\mu^m)^{\frac{\rho}{1-\rho}}\left(\frac{c_M(t)\,\omega}{c_A\rho}\right)^{\frac{1}{1-\rho}} && \text{(C.3)} \end{aligned}$$

During industrialization machine producers are indifferent between entering mass production or staying out. Therefore, they have to make zero profits and condition 4.14 holds:

$$x_{C0}(t) = A\frac{E}{\omega}$$

At the onset of industrialization the demand for high-skilled workers in the artisan industry exceeds supply and the wage differential is $w(m)$, the same as in the artisan economy. The zero profit condition will then determine the share of industrializing sectors:

$$A = \frac{\rho}{c_M(t)}\frac{1}{y+(1-y)\left(\mu^m\frac{c_M(t)\omega}{c_A\rho}\right)^{\frac{\rho}{1-\rho}}} \tag{C.4}$$

Due to condition 4.6 the left hand side of this expression is decreasing in y and $c_M(t)$. Hence technological progress promotes industrialization.[1]

At some time t_2^{C0} all [illegible]-skilled workers switched to industrial production while the demand for artisans continues to decrease. However, artisans will not enter industry yet because they would have to accept the wages of low-skilled workers.[2] Instead, the wage levels of both groups will gradually equalize. During this process the zero profit condition C.4 continues to hold. Furthermore, the ratio of high-skilled artisans and low-skilled industrial production workers equals the relative share of both groups:

$$\frac{(1-y)\, c_A x_A(t)}{y\, c_M(t)\, x_{C0}(t)} = \frac{\alpha}{1-\alpha}$$

This condition can be simplified:

$$c_M(t)^{\rho}\omega = \left(\frac{\alpha}{1-\alpha}\right)^{1-\rho} \frac{1}{\rho} \left(\frac{c_A}{\mu^m}\right)^{\rho} \left(\frac{y}{1-y}\right)^{1-\rho} = D\left(\frac{y}{1-y}\right)^{1-\rho} \tag{C.5}$$

If we define an auxiliary variable $z = c_M\omega$ we can rewrite the two conditions C.4 and C.5 in reduced form as

$$A = f(c_M, y, z) \tag{4a}$$

$$D = g(c_M, y, z), \tag{5a}$$

with $\frac{\partial f}{\partial c_M} < 0$, $\frac{\partial f}{\partial y} < 0$, $\frac{\partial f}{\partial z} < 0$, $\frac{\partial g}{\partial c_M} < 0$, $\frac{\partial g}{\partial y} < 0$, $\frac{\partial g}{\partial z} > 0$.

We can then deduce that industrialization proceeds during wage equalization as

$$\frac{dy}{dc_M} = \frac{\frac{\partial g}{\partial c_M}\frac{\partial f}{\partial z} - \frac{\partial g}{\partial z}\frac{\partial f}{\partial c_M}}{\frac{\partial g}{\partial z}\frac{\partial f}{\partial y} - \frac{\partial g}{\partial y}\frac{\partial f}{\partial z}} < 0 \tag{C.6}$$

The relative wages of low-skilled workers will in deed increase as one can immediately see from equation C.5.

At time t_3^{C0} the wages of workers will have equalized. High-skilled workers are now

[1] Note, that the entry decisions of mass producers are strategic substitutes. Every new entrant lowers the demand faced by other mass producers because goods are substitutes. This guarantees uniqueness of the equilibrium.

[2] High-skilled workers do not enjoy a comparative advantage in producing generic goods because the division of labor is high.

indifferent between staying on as artisans or becoming industrial production workers. They will gradually switch into mass production until the entire economy has industrialized at time t_4^{C0}.

It is important to note that throughout the process of industrialization no machine producer would wish to customize a variety. Due to condition 4.5 her revenue from the production of the variety would be less than the revenue of a mass producer but her cost of producing a dedicated machine would be the same. As mass producers just break even customized varieties would be unprofitable. QED

C.2 Proof of Theorems 10 and 12

For the proof of theorem 10 simply set $a = 1$, e.g. all workers are of high-ability. Without loss of generality I concentrate on the demise of mass production. In a share y of sectors incumbent mass producers face entry by firms which offer more customized varieties. The price level $p(t)$ in the economy and the expected demand $x_{C0}(t)$ for generic goods and $x_{C1}(t)$ for varieties with degree of customization $d = 1$ can be derived from equations 4.2 and 4.3:

$$
\begin{aligned}
p(t) &= \frac{c_M(t)\,\omega}{\rho}\left[1 - y + y\,(\mu\omega)^{\frac{\rho}{1-\rho}}\right]^{\frac{\rho-1}{\rho}} && \text{(C.7)}\\
x_{C0}(t) &= \frac{E\rho}{c_M(t)\,\omega}\frac{1}{1 - y + y\,(\mu\omega)^{\frac{\rho}{1-\rho}}} && \text{(C.8)}\\
x_{C1}(t) &= x_{C0}(t)\,\frac{\mu^{\frac{\rho}{1-\rho}}}{3}\omega^{\frac{1}{1-\rho}} && \text{(C.9)}
\end{aligned}
$$

During transition entrants have to make zero profits and condition 4.15 holds:

$$x_{C1}(t) = A\left[\alpha + (1-\alpha)\,\omega\right]$$

The income of consumers consists of labor income and profits made by incumbent mass producers, e.g. $E = \alpha + (1-\alpha)\,\omega + \Pi$. Profits can be derived as follows:

$$
\begin{aligned}
\Pi &= (1-y)\left[x_{C0}(t)\,c_M(t)\,\omega\frac{1-\rho}{\rho} - k(t)\left[\alpha + (1-\alpha)\,\omega\right]\right] && \text{(C.10)}\\
&= (1-y)\left[\alpha + (1-\alpha)\,\omega\right]k(t)\left[\frac{1}{\frac{(\mu\omega)^{\frac{\rho}{1-\rho}}}{3}} - 1\right] && \text{(C.11)}
\end{aligned}
$$

We can then rewrite total consumer income as:

$$E = [\alpha + (1-\alpha)\omega]\left[1 + (1-y)\,k(t)\left[\frac{1}{\frac{(\mu\omega)^{\frac{\rho}{1-\rho}}}{3}} - 1\right]\right] \tag{C.12}$$

After time t_1^{C1} the supply of high-skilled workers exceeds demand in the mass production economy and wages are equal for both types of workers. The zero profit condition determines the share of sectors y with customized production and can be expressed as

$$A = \frac{\rho}{3}\frac{\frac{z}{c_M(t)} + (1-y)\frac{k(t)}{c_M(t)}(3-z)}{1-y+yz} \tag{C.13}$$

with the help of the auxiliary variable $z = (\mu\omega)^{\frac{\rho}{1-\rho}}$. This condition can be written in reduced form as

$$A = f(c_M, y, z) \tag{13a}$$

with $\frac{\partial f}{\partial c_M} < 0$, $\frac{\partial f}{\partial y} < 0$, $\frac{\partial f}{\partial z} > 0$.[3] Because z is fixed ($\omega = 1$) technological progress implies an increase in the share y of customized sectors in the economy.

At some time t_2^{C1} all high-skilled production workers are employed in the customized sectors and the labor market tightens as a result. The relative wage of high-skilled workers then has to increase. During this process the zero profit condition C.13 continues to hold. Furthermore, the ratio of high-skilled production workers in the customized sectors and low-skilled mass production workers equals the relative share of both groups:

$$\frac{y c_M(t)\, x_{C1}(t)}{(1-y)\, c_M(t)\, x_{C0}(t)} = \frac{\alpha}{1-\alpha} \tag{C.14}$$

This condition can be written in reduced form as

$$F = g(y, z) \tag{14a}$$

with $\frac{\partial g}{\partial y} > 0$ and $\frac{\partial g}{\partial z} > 0$.

[3] The right hand side of expression 13a decreases in y because of condition 4.11. The derivative of the expression with respect to z has the same sign as $1 - (1-y)\,k(t) - 3yk(t)$ which is the mass of production workers in the economy and therefore positive.

Combining this condition with condition 14a we can deduce that the share of customized sectors will continue to increase during the process of wage widening.

At time t_3^{C1} relative wages have reached the level $w(1)$ and reflect the productivity difference between low-skilled and high-skilled workers in the customized industries. Low-skilled workers will gradually leave mass production until the entire economy produces varieties with degree of customization $d = 1$ at time t_4^{C1}. QED

Bibliography

ABERNATHY, F. H., J. T. DUNLOP, J. H. HAMMOND, AND D. WEIL (1999): *A Stitch in Time.* Oxford University Press, New York.

ACEMOGLU, D. (1998): "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality," *Quarterly Journal of Economics*, 113(4), 1055–1089.

ANDERSON, G. (1907): "Telephone Competition in the Middle West and Its Lessons for New England," prepared for New England Telephone and Telegraph Company.

AUTOR, D. H., L. F. KATZ, AND A. B. KRUEGER (1998): "Computing Inequality: Have Computers Changed the Labor Market," *Quarterly Journal of Economics*, 113, 1169–1213.

BABBAGE, C. (1835): *On the Economy of Machinery and Manufactures.* C. Knight, London.

BARTEL, A., AND F. LICHTENBERG (1987): "The Comparative Advantage of Educated Workers in Implementing New Technologies," *Review of Economics and Statistics*, 69, 1–11.

BECKER, G. S., AND K. M. MURPHY (1992): "The Division of Labor, Coordination Costs, and Knowledge," *Quarterly Journal of Economics*, 107, 1137–60.

BELL, R. (1972): *Changing Technology and Manpower Requirements in the Engineering Industry.* Sussex University Press, Brighton.

BERMAN, E., J. BOUND, AND Z. GRILLICHES (1994): "Changes in the Demand for Skilled Labor within US Manufacturing: Evidence from the Annual Survey of Manufacturers," *Quarterly Journal of Economics*, 109, 367–397.

BERMAN, E., J. BOUND, AND S. MACHIN (1997): "Implications of Skill-Biased Technological Change: International Evidence," Working Paper 6166, National Bureau of Economic Research.

BORJAS, G. J. (1995): "Ethnicity, Neighborhoods, and Human-Capital Externalities," *American Economic Review*, 85, 365–389.

BORNHOLZ, R., AND D. S. EVANS (1983): "The Early History of Competition in the Telephone Industry," in *Breaking Up Bell: Essays on Industrial Organization and Regulation*, ed. by D. S. Evans. North-Holland, New York.

BOUND, J., AND G. JOHNSON (1992): "Changes in the Structure of Wages in the 1980s: An Evaluation of Alternative Explanations," *American Economic Review*, 82, 371–392.

BRESNAHAN, T. F., E. BRYNJOLFSSON, AND L. M. HITT (1999): "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," Working Paper 7136, National Bureau of Economic Research.

BROOKS, J. (1975): *Telephone.* Harper and Row, New York.

BUREAU OF THE CENSUS (1902): *Telephones and Telegraphs.* Government Printing Office, Washington.

——— (1912): *Telephones and Telegraphs.* Government Printing Office, Washington.

——— (1975): *Historical Statistics of the U.S., Colonial Times to 1970*, II. Government Printing Office, Washington.

BURGESS, E. W. (1928): "Residential Segregation in American Cities," *Annals of the American Academy of Political and Social Science*, 140, 105–115.

CAPELLI, P., AND N. ROGOVSKY (1994): "New Work Systems and Skill Requirements," *International Labor Review*, 133, 204–220.

CAROLI, E., AND J. V. REENEN (1999): "Skills and Organizational Change: Evidence from British and French establishments in the 1980s and 1990s," mimeo.

CASE, A. C., AND L. F. KATZ (1991): "The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths," Working Paper 3705, National Bureau of Economic Research.

CASSON, H. N. (1910): *The History of the Telephone.* A.C. McClurg & Co., Chicago.

COMMITTEE ON GAS, OIL AND ELECTRIC LIGHT (1907): "Telephone Service and Rates," Report to the City Coucil of Chicago.

CUTLER, D., D. ELMENDORF, AND R. ZECKHAUSER (1993): "Demographic Characteristics and the Public Bundle," in *On the Role of Budgetary Policy during Demographic Changes*, ed. by B. Wolfe. International Institute of Public Finance, Paris, France.

CUTLER, D. M., AND E. L. GLAESER (1997): "Are Ghettos Good or Bad?," *Quarterly Journal of Economics*, 112, 827–872.

CUTLER, D. M., E. L. GLAESER, AND J. L. VIGDOR (1997): "The Rise and Decline of the American Ghetto," Working Paper 5881, National Bureau of Economic Research.

DIXIT, A., AND V. NORMAN (1980): *Theory of International Trade.* Cambridge University Press, Cambridge.

DIXIT, A. K., AND J. E. STIGLITZ (1977): "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, 67, 297–308.

DOMS, M., T. DUNNE, AND K. R. TROSKE (1997): "Workers, Wages and Technology," *Quarterly Journal of Economics*, 112, 253–290.

ELLISON, G. (1993): "Learning, Local Interaction, and Coordination," *Econometrica*, 61, 1047–1071.

——— (1999): "Basins of Attraction, Long Run Equilibria, and the Speed of Step-by-Step Evolution," *Review of Economic Studies*, forthcoming.

ETHIER, W. J. (1982): "National and International Returns to Scale in the Modern Theory of International Trade," *American Economic Review*, 72, 389–405.

FISCHER, C. S. (1992): *America Calling: A Social History of the Telephone to 1940.* University of California Press, Berkeley.

GABEL, R. (1969): "The Early Competitive Era in Telephone Communication, 1893-1920," *Law and Contemporary Problems*, pp. 340–69.

GALOR, O., AND D. TSIDDON (1997): "Technological Progress, Mobility, and Economic Growth," *American Economic Review*, 87(3), 363–382.

GALSTER, G. C. (1990): "White Flight from Racially Integrated Neighborhoods in the 1970s: the Cleveland Experience," *Urban Studies*, 27, 385–399.

GIORDANO, L. (1992): *Beyond Taylorism: Computerization and The New Industrial Relations.* St. Martin's Press, New York.

GLAESER, E. L., B. SACERDOTE, AND J. A. SCHEINKMAN (1996): "Crime and Social Interactions," *Quarterly Journal of Economics*, 111, 507–48.

GOLDIN, C., AND L. F. KATZ (1995): "The Decline of Non-Competing Groups: Changes in the Premium to Education, 1890 to 1940," Working Paper 5202, National Bureau of Economic Research.

——— (1998): "The Origins of Technology-Skill Complementarity," *Quarterly Journal of Economics*, 113, 693–732.

GOLDIN, C., AND R. A. MARGO (1992): "The Great Compression: The Wage Structure in the United States at Mid-Century," *Quarterly Journal of Economics*, 107, 1–34.

JOHNSTON, G. R. (1908): *Some Comments on the 1907 Annual Report of the American Telephone and Telegraph Company.* International Independent Telephone Association, Chicago.

KANDORI, M., G. MAILATH, AND R. ROB (1993): "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, 61, 29–56.

KATZ, L. F., AND K. M. MURPHY (1992): "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 107, 35–78.

KELLY, J. E. (1982): *Scientific Management, Job Redesign and Work Performance.* Academic Press.

KEMENY, J. C., AND J. L. SNELL (1960): *Finite Markov Chains.* D. Van Nostrand Company, Princeton.

KERN, S. (1983): *The Culture of Time and Space, 1880-1918.* Harvard University Press, Cambridge.

KRUGMAN, P. R. (1981): "Intraindustry Specialization and the Gains from Trade," *Journal of Political Economy*, 89, 959–973.

——— (1995): "Technology, Trade and Factor Prices," Working Paper 5355, National Bureau of Economic Research.

KUSMER, K. L. (1976): *A Ghetto Takes Shape.* University of Illinois Press, Chicago.

LANDES, D. S. (1969): *The Unbound Prometheus.* Cambridge University Press, Cambridge.

LANGDALE, J. V. (1978): "The Growth of Long-Distance Telephony in the Bell System: 1875-1907," *Journal of Historical Geography*, 4, 145–59.

LIGGETT, T. M. (1985): *Interacting Particle Systems.* Springer Verlag, New York.

LIPARTITO, K. (1989): "System Building at the Margin: The Problem of Public Choice in the Telephone Industry," *Journal of Economic History*, 49, 323–36.

MACMEAL, H. B. (1934): *The Story of Independent Telephony.* Independent Pioneer Telephone Association, Chicago.

MASSEY, D., AND N. DENTON (1993): *American Apartheid: Segregation and The Making of the Underclass.* Harvard University Press, Cambridge.

MERCHANTS' ASSOCIATION OF NEW YORK (1905): *Supplemental Telephone Report: Further Inquiry into Effect of Competition.* Merchants' Association, New York.

MEYEROWITZ, J. (1985): *No Sense of Place: The Impact of Electronic Media on Social Behavior.* Oxford University Press, New York.

MILGROM, P., AND J. ROBERTS (1990): "The Economics of Modern Manufacturing: Technology, Strategy and Organization," *American Economic Review*, 80(3), 511–528.

MUELLER, M. L. (1997): *Universal Service: Competition, Interconnection, and Monopoly in the Making of the American Telephone System.* MIT Press.

MURNANE, R. J., F. LEVY, AND D. AUTOR (1999): "Technological Change, Computers and Skill Demands: Evidence from the Back Office Operations of a Large Bank," mimeo.

NEW ENGLAND TELEPHONE AND TELEGRAPH COMPANY (1908): "Competition in Telephony," .

OSOFSKY, G. (1963): *Harlem: The Making of a Ghetto.* Harper & Row, New York.

OSTERMAN, P. (1994): "How Common Is Workplace Transformation and Who Adopts It?," *Industrial and Labor Relations Review*, 47(2), 173–188.

——— (1998): mimeo.

PASHIGIAN, B. P., AND B. BOWEN (1991): "Why are Products Sold on Sale?: Explanations of Pricing Regularities," *Quarterly Journal of Economics*, 106, 1015–1038.

PIL, F., AND J.-P. MACDUFFIE (1996): "The Adoption of High Involvement Work Practices," *Industrial Relations*, 35(3), 423–455.

PIORE, M. J., AND C. F. SABEL (1984): *The Second Industrial Divide.* Basic Books.

SACHS, J. D., AND H. J. SHATZ (1994): "Trade and Jobs in U.S. Manufacturing," *Brookings Papers on Economic Activity*, 0(1), 1–84.

SCHELLING, T. C. (1972): "A Process of Residential Segregation: Neighborhood Tipping," in *Racial Discrimination in Economic Life*, ed. by A. Pascal. Lexington Books, Lexington, MA.

——— (1978): *Micromotives and Macrobehavior.* Norton and Company, New York.

SMITH, A. (1776): *An Inquiry into the Nature and Causes of the Wealth of Nations.* Whitestone, Dublin.

SMITH, JR., B. (1959): "The Reshuffling Phenomenon: A Pattern of Residence of Unsegregated Negroes," *American Sociological Review*, 24, 77–79.

SPEAR, A. H. (1967): *Black Chicago.* University of Chicago Press, Chicago.

STEHMAN, J. W. (1925): *The Financial History of the American Telephone and Telegraph Company.* Houghton Mifflin, Boston.

STIRZAKER, D. (1994): *Elementary Probability.* Cambridge University Press, Cambridge.

TAEUBER, K. E., AND A. F. TAEUBER (1965): *Negroes in Cities.* Aldine Publishing Company, Chicago.

TAYLOR, F. W. (1911): *The Principles of Scientific Management.* Harper & Brothers, New York.

THESMAR, D., AND M. THOENIG (1999): "Creative Destruction and Firm Organization Choice: A New Look into the Growth Inequality Relationship," mimeo.

U.S. FEDERAL COMMUNICATIONS COMMISSION (1938): *Proposed Report: Telephone Investigation.* Government Printing Office, Washington.

WEIMAN, D. F., AND R. C. LEVIN (1994): "Preying for Monopoly? The Case of Southern Bell Telephone Company: 1894-1912," *Journal of Political Economy,* 102(1), 103–126.

WILKINSON, B. (1983): *The Shopfloor Politics of New Technology.* Heinemann Educational, London.

WILSON, W. J. (1987): *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* University of Chicago Press, Chicago, IL.

WOMACK, J. P. (1989): *The US Automobile Industry in an Era of International Competition: Performance and Prospects.* MIT Press, Cambridge, Mass., the working papers of the mit commission on industrial productivity edn.

YOUNG, H. P. (1993): "The Evolution of Conventions," *Econometrica,* 61, 57–84.

ZELL, D. (1997): *Changing by Design.* Cornell University Press, Ithaca.