COMPLEX DYNAMICS OF HUMAN ACTIVITY:
LANGUAGE, CITIES, COLLABORATION, AND BASEBALL

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Samuel Arbesman
August 2008

COMPLEX DYNAMICS OF HUMAN ACTIVITY:
LANGUAGE, CITIES, COLLABORATION, AND BASEBALL

Samuel Arbesman, Ph. D.

Cornell University 2008

A few areas of human activity are examined here, using a number of different types of mathematical and computational models. First, we examine networks of five languages of the world, with their connectivity derived from the sounds of the words in these languages. We explore the graph-theoretic properties of these networks, finding that these phonological language networks have common properties, and are in turn topologically distinct from other types of complex networks observed in the literature. In addition, we discuss what these common properties imply for how we process language and why natural language is structured the way it is. In addition, by examining the networks of English and Spanish, we explain a surprising difference in processing that was uncovered in some recent experiments, and discuss some more general implications of competition or facilitation between different modes of cognition.

We next explore a more macro-scale area of human activity: cities. Superlinear scaling in cities, which appears in sociological quantities such as economic productivity and creative output relative to urban population size, has been observed but not been given a satisfactory theoretical explanation. We provide a model for the superlinear relationship between population size and innovation found in cities, with a reasonable range for the exponent.

Next, we examine collaboration and innovation in the scientific world. We

attempt to understand how variations in ‘scientific distance’ among collaborators affect the degree to which that collaboration is a productive one. Using both mathematical models and empirical data, we explore the relationship between the scientific or social distance of collaborators and the fruitfulness of their output.

Last, we examine Joe DiMaggio’s 56-game hitting streak and look at its probability, using a number of simple models. And it turns out that, contrary to many people’s expectations, an extreme streak, while unlikely, is not unlikely to have occurred about once within the history of baseball. Surprisingly, however, such a record should have occurred far earlier in baseball history: back in the late 1800’s or early 1900’s. But not in 1941, when it actually happened.

BIOGRAPHICAL SKETCH

Samuel Arbesman was born on December 10, 1981, without a middle name. He was born in Cleveland, Ohio, little more than a decade after the Cayuhoga river caught fire due to being full of trash. Less than three years later, he and his family left the Midwest for another city in the American Rustbelt, that of Buffalo, New York. Sam considers Buffalo—referred to as the City of Lights, City of Good Neighbors, and, less commonly, Home of the Triscuit—his hometown.

Sam resided in the outskirts of Buffalo until college, whereupon he enrolled in Brandeis University in Waltham, a suburb of Boston. He majored in Computer Science and Biology, with a minor in Near Eastern and Judaic Studies. After graduation, he matriculated at Cornell University in Ithaca, where he completed Cornell's first PhD in Computational Biology in 2008. There, he slowly began his shift in interests into the burgeoning field of computational social science. Sam will continue this line of research as a postdoctoral fellow in the Department of Health Care Policy at Harvard Medical School.

He still doesn't have a middle name.

To my grandfather, Dr. Irwin Arbesman

ACKNOWLEDGMENTS

Acknowledgements are never complete. As soon as I think I've thanked everyone who has helped me over these four years (or longer), I realize there's someone else who needs to be mentioned. So, I hope I haven't left anyone out. But if I have, I extend my most sincere apologies, and I am more than happy to make it up by putting you in the second edition.

Thanks first go to Steve Strogatz. You're a wise mentor, a thoughtful advisor, and a friend. You helped me navigate the many changes in my research topics, and helped me find out where I fit within the world of academic research. You even helped get me my first paid writing gig. It's been great working with you, and I hope we continue to do so, even as I leave Cornell.

Thanks also go to the rest of my committee: Steve Ellner and Adam Siepel. You have been very understanding of my shifts in interests, and my switch to a 'higher-order' computational biology. I appreciate your guidance, advice and critical analysis of my research, which has helped to make my dissertation better.

I would also like to thank my collaborators who I worked with on the research that made it into here: Jon Kleinberg, Mike Vitevitch and Scott Page. You have enabled me to see the world differently, helped me to learn about new fields, and provided me with wonderful ideas. I knew that I would always enjoy collaborating, but you three made me certain of it.

Thanks also goes to the Computer Science department who accepted me, and then kindly let me switch fields.

I am also extremely indebted to the NSF IGERT Program in Nonlinear Systems, which I was a part of during my first two years of graduate school. IGERT taught me new things, but even more importantly, the program exposed me to a wide

variety of incredibly smart and interesting people, and provided a much-needed support structure for the students working on interdisciplinary research who would otherwise be strewn to the far corners of the Cornell campus. IGERT was a welcoming community, provided me with a great group of friends, and I am thankful that I had a chance to be a part of it.

This thesis was also made possible by the myriad elements of our contemporary popular culture, which are, frankly, too many to name.

I would also like to thank the National Science Foundation for their grant DMS 0412757 to Steven Strogatz, in addition to the NSF IGERT Fellowship in Nonlinear Systems. The NSF gave me the freedom to play with fun topics, try lots of things that didn't pan out, and pay me so I could eat. And, by extension, I would like to thank the American taxpayer. This dissertation was made possible by readers like you.

Also, I would like to thank Lauren Childs and Megan Owen, wonderful officemates (who are tolerant of numerous distractions) and great friends. You have endured the many research-related crises (or otherwise) that I have complained about. And everyone in the Center for Applied Mathematics: thanks. You welcomed me as one of your own—even if I refused to talk much about math—and provided me with a great environment to get a PhD. You made the whole graduate experience a fun and relaxing one. And of course, Dolores and Bridget: thanks for keeping CAM from falling apart at the seams, which we seemed to be bent on making happen. And Carly: you helped make the first PhD in Computational Biology a far less daunting experience than it could have been by smoothing out all the bureaucratic details that could have otherwise been insurmountable hurdles. Thanks.

Also, thanks to all of my friends, both in Ithaca and outside of Ithaca. You have all been great fun, provided so much support, and, frankly, just made these past four years (as well as the rest of my life) one long great hoot.

And of course, I would like to thank my family. My parents in particular provided a great deal of support and advice, about research, graduate student life, life in general, and why not to drop out of grad school. Mom and Dad – you’re wonderful and I love you both very much. You have far more than your fair shares of wisdom, and I hope some has rubbed off on me. And Josh, you’re the best brother a sibling could have. You too helped provide perspective and support throughout this entire graduate experience. I look forward to seeing you as a doctor as well (the real kind, that is).

And I would like to thank my grandfather, Irwin Arbesman, to whom I have dedicated my dissertation. Your ninety years of experience and wisdom have been dispensed well, and provided a great deal to think about while I have been a grad student. Not many students have nonagenarians who can productively brainstorm about their research with, let alone someone who is also such a great all-around guy. Thanks for making sure I always look at the world slightly differently, skeptically, and even a bit off-kilter (what I may safely term the Arbesman way); it has served me well. You’ve created quite a family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

While mathematics has been used for a long time to understand human behavior, only recently has there been the ability to examine large ensembles of interactions in a computational manner. Computational agent-based models, where agents interact with each other according to a set of rules, allow us to understand how micro-scale interactions create macro-level results. By allowing us to simulate emergent and nonlinear phenomena within the social sphere, we can gain insights that can guide our theoretical understanding. Within the past fifty years or so, scientists have begun to apply such computational approaches, often based on an understanding of complex systems, to many topics within the social sciences [38, 53, 55, 56, 69]. Here I attempt to provide some insight into a few areas of human activity, using a number of different types of models.

The first area I examine is how language is processed. While language is processed by individuals one at a time, it is more broadly a communal endeavor. Communication necessarily requires at least two people, and language evolution is a product of large-scale change within a society as a whole. In addition, language, while intriguing on its own, is as Steven Pinker notes in the subtitle of his recent book, 'a window into human nature' [52]. I explore language in Chapters 2 and 3. This work was carried out in collaboration with the cognitive scientist Michael Vitevitch and my adviser Steven Strogatz.

In Chapter 2, we examine networks of five languages of the world (English, Spanish, Mandarin, Basque, and Hawaiian), with their connectivity derived from the sounds of the words in these languages. We explore the graph-theoretic properties of these networks, finding that these phonological language networks have common properties, and are in turn topologically distinct from other types of complex networks observed in the literature, such as biological and social networks. In addition, the common properties among the phonological networks have a number of implications for how we process language and why natural language is structured the way it is.

Chapter 3 focuses on two languages—English and Spanish—and examines some recent paradoxical discoveries about how they are processed. It turns out that language, like all cognitive function, often requires many different types of processing simultaneously. When these processing types overlap, function is facilitated; otherwise, functions compete and slow down processing. In the case of spoken English, words that sound alike often have radically different meaning. In other words, English has little correlation between words that are phonologically similar and words that are semantically similar. On the other hand, the phonological and semantic networks for Spanish show a high degree of overlap. By examining the networks of English and Spanish, we explain a surprising difference in processing that was uncovered in some recent experiments, and discuss some more general implications of competition or facilitation between cognitive modules.

Chapter 4 turns to a more macro-scale area of human activity: productivity and innovation within cities. Cities provide an environment—fueled by the high density of human interaction—for creativity and innovation. Essentially, it is cities that provide the world with its ideas. In addition, in 2008, we reached an urban milestone for the planet. For the first time, over half the world's population is located within cities [49]. And a great deal of this growth is occurring within the developing world. Therefore, understanding the nature of human activity within cities is paramount.

There are many areas of research that can be examined within cities. Here we focus on the topic of innovation. Cities, in order to compete against one another, are moving from manufacturing centers to incubators of ideas [25]. The evidence indicates that urban idea generation is not capricious. Cities have recently been demonstrated to obey certain quantitative laws and are not as chaotic and messy as one might assume [11]. Superlinear scaling in cities, which appears in sociological quantities such as economic productivity and creative output relative to urban population size, has in fact been observed but not been given a satisfactory theoretical explanation. In Chapter 4, based on work done with Jon Kleinberg and Steven Strogatz, we provide a model for the superlinear relationship between population size and innovation found in cities, with a reasonable range for the exponent.

Of course, innovation is not a simple mathematical process either. We next provide a more detailed look at the relationship between innovation and collaboration. In Chapter 5, with Jon Kleinberg, Steven Strogatz, and Scott Page, we examine collaboration and innovation in the scientific world, using articles and patents as a guide.

While previous work has been done in the area of coauthorship, we attempt to understand how variations in ‘scientific distance’ among collaborators affect the degree to which that collaboration is a productive one [48]. One may assume that scientists who are farther apart in terms of their areas might have more to offer each other and be more productive. On the other hand, communication with colleagues in different disciplines can be far more difficult, because of differences in training, language, and scientific culture. Thus, obtaining a clearer notion of the empirical relationship is important.

We create a mathematical model of how innovation among collaborators might occur. Using this, as well as our empirical data, we explore the relationship between the scientific or social distance of collaborators and the fruitfulness of their output.

Chapter 6 expands on our opinion editorial in the New York Times, which dealt with a Monte Carlo study of Joe DiMaggio's 56-game hitting streak [8]. This investigation is intended as a case study of areas of human endeavor where streaks are important—such as performance in financial markets—and where skill and luck combine to influence performance. We examine Joe DiMaggio's streak—a record that has been said should never have happened—and look at its probability, using a number of simple models.

And it turns out that, contrary to many people's expectations, an extreme streak, while unlikely, is not unlikely to have occurred about once within the history of baseball. Surprisingly, however, such a record should have occurred far earlier in baseball history: back in the late 1800's or early 1900's. But not in 1941, when it actually happened.

Make no mistake: there is no comprehensive unified theory of humanity in all this. I have only examined a few areas of human activity—language, cities, collaboration, and baseball—using computational methods found within complex systems, mainly informed by network-based approaches. However, these small bits of insight will hopefully contribute to the steady chipping away at the unknown within the social sciences.

CHAPTER 2

THE STRUCTURE OF PHONOLOGICAL NETWORKS ACROSS MULTIPLE LANGUAGES[1]

Introduction

The results of numerous graph-theoretic analyses suggest that a number of principles may influence the emergent structures found in a wide variety of complex systems, including information, social, technological, and biological networks [4, 47, 63]. These unifying characteristics include small-world properties, distinct community structure, and scale-free distributions of the network connectivity.

Many aspects of language can be examined from a network perspective as well. Numerous studies have been conducted on semantic networks, where relationships in meaning have been made between words. These are often based on thesauri, word-associations in corpori or from academic databases [20, 44]. In addition, linguistic networks have been made from orthographic similarities of words (how words are spelled) [32]. Lastly, language can be viewed from the sounds of words (their phonological structure), where words that sound similar are neighbors. Whereas there is older work on small portions of phonological networks (nearest neighbors of words) [40], the first study of an entire language network only appeared this year, in 2008 [64].

In these phonological networks, words in a language are represented as vertices or nodes, and an edge is placed between them if the words sound similar to

[1] This work was completed in conjunction with Michael Vitevitch (University of Kansas) and Steven Strogatz (Cornell University).

each other (differing only by a single phoneme, or sound segment). For example, vertices representing the words *hand*, *send*, *sad*, *and*, and *stand* would all have edges connecting them to the vertex for the word *sand*; the meaning of the words is not used to place edges among vertices, such as in semantic networks [62]. Psycholinguistic studies suggest that several characteristics of the network influence cognitive processing, making this an especially intriguing network to examine [62, 64].

In examining English, Vitevitch found that its phonological network had a small giant component (the largest connected portion of the graph), with many other smaller components. This property is distinct from other complex networks observed in the literature. In addition, the degree distribution (the distribution of the number of edges per node) was not well modeled by a scale-free distribution, or a power law.

Here, we wanted to explore the generality of these results, by doing the first comparative study of multiple languages, using phonological networks. We examined some of the properties looked at by Vitevitch in English, as well as a number of others, and found that phonological networks all have certain properties distinct from other types of complex networks (such as biological and social networks).

Methods

The network structure of selected languages was examined to determine the generality of the network characteristics previously observed in English [64]. In addition to English, the following languages were examined: Spanish, Mandarin, Hawaiian, and Basque (see Table 2.1). Similar network characteristics across a variety of languages might hint toward principles that are common to all languages, whereas differences in network measures might provide a quantitative way to describe and categorize the languages of the world.

English is an Indo-European language from the Germanic branch, whereas Spanish comes from the Romance branch of the Indo-European family of languages. Mandarin, a Sino-Tibetan language, differs from English, Spanish, Hawaiian and Basque in that it also uses tones to convey word meanings (e.g., “fan” with a high level tone means sail, with a rising tone means trouble, with a dipping tone means turn, and with a falling tone means rice). Tone was not included in the phonological transcriptions, however. Hawaiian is an Austronesian language with a phoneme inventory (the number of consonants and vowels in the language) that is smaller than those found in English, Spanish, Mandarin, and Basque. Finally, Basque (or Euskara) is a linguistic isolate, meaning that it is not (or has not yet been identified as) a member of a given language family. Additional differences, such as those in morphology, exist among the languages that were selected for the present network analyses.

The phonological networks were constructed from a variety of sources. The English network contained the words from the Merriam-Webster Pocket Dictionary from 1964; this database has been used extensively in psycholinguistic studies [40]. The Hawaiian network was created in a similar manner using a Hawaiian Dictionary [31]. The words from the Spanish network consisted of the words in the LEXESP database [58], a large Spanish language corpus. The words in the Basque network were obtained in a manner similar to the words in the Spanish network [51]. The Mandarin network uses the words from Huang et al, 1997 [30].

Results

Table 2.1. Summary information of phonological networks in several languages. GC stands for Giant Component.

	English	***Spanish***	***Mandarin***	***Hawaiian***	***Basque***
Network Size (number of words)	19,323	122,066	30,086	2,578	99,321
Giant Component Size (percentage)	6,498 (0.34)	44,833 (0.37)	19,712 (0.66)	1,406 (0.55)	35,173 (0.35)
Assortative Mixing by Degree (r)	0.657	0.762	0.654	0.556	0.719
Average Shortest Path Length	2.7	4.3	6.5	3.2	4.4
Average Shortest Path Length (GC)	6.1	10.3	10.1	5.5	10.4
Average Shortest Path Length of random network (based on GC)	5.8	9.9	7.3	5.8	11.4
Clustering Coefficient	0.284	0.191	0.383	0.241	0.206
Clustering Coefficient of random network	8.35e-5	1.17e-5	8.55e-5	7.40e-4	1.21e-5
Transitivity	0.313	0.25	0.404	0.260	0.232
Ratio of Edges to Vertices	1.61	1.43	2.57	1.91	1.21
Ratio of Edges to Vertices (GC)	4.55	2.95	3.88	3.44	2.50

Size of the Giant Component and Network Robustness. The giant component sizes of the language networks were much smaller compared to other network structures discussed in the literature. Typically, the giant component contains approximately 80-90% of the vertices [48]. However, in the present networks, the proportion of vertices in the giant component was much smaller, with some networks having less than 50% of the vertices in the giant component. The proportion of vertices in the giant components for comparably sized random networks, containing 70-80% of the vertices, are also larger than the values for the language networks [12]. This difference in giant component size suggests that these phonological networks may be more robust to node removal due to more tightly connected components, and indicates the prevalence of smaller components in the networks.

To evaluate the robustness of the networks, vertices were removed in two ways: at random, and in decreasing order by degree (number of edges connected to a vertex). These results are shown in Figure 2.1. In scale-free networks, when vertices are randomly removed the mean shortest path length remains constant, whereas when vertices are removed in order of degree, the mean shortest path length increases dramatically [47]. In the language networks, however, both methods of node removal resulted in little to no change in the mean shortest path lengths. The shortest path lengths were calculated using a sampling technique where 1,000 nodes were chosen at random. Then, the distance to all other nodes (if part of the same component) were obtained and these paths lengths were then all averaged, to give an estimate of the shortest path length. This sped up the calculations considerably.

In addition, we examined the assortative mixing by degree of the language networks, which is a measure of the correlation of degree between neighboring nodes. As seen in Table 2.1, all of the language networks had large and positive correlations of the degrees of connected vertices, indicating that high degree vertices tended to be connected to each other. Newman [45] discussed how networks with assortative mixing by degree are more robust to vertex removal. In addition, the high assortative mixing is distinct from other types of networks: biological and technological networks often are disassortatively mixed, and social networks, which display assortative mixing, still have lower values of assortative mixing. Typical measures of assortativity for social networks are 0.1-0.3, and biological and technological networks are -0.1 to -0.2 [45]. On the other hand, phonological networks can be higher than 0.7.

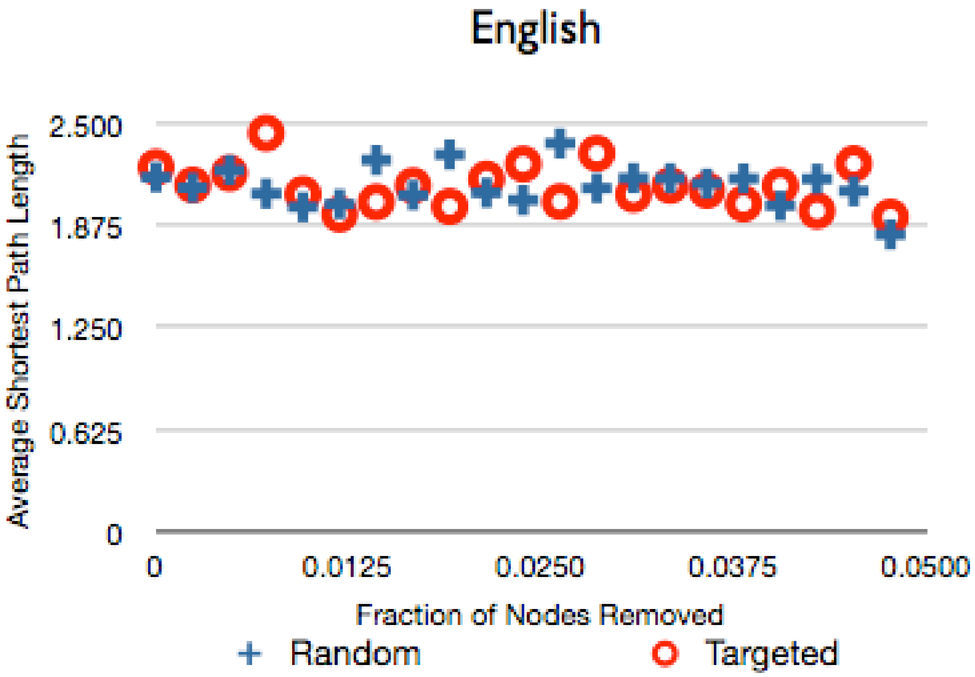

Figure 2.1. An example run of node removal in English, either random or in a targeted fashion (in order by degree). Up to 5% of the nodes were removed, and all languages showed similar patterns to the above results. In addition, when the simulations were done only for the giant component, a similar constant, though elevated, value of the average shortest path length was found.

Our findings suggest evidence of the robustness of the networks, and highlight the resilience of lexical processing in the face of injury to the language related areas of the brain (i.e., stroke). Specifically, there are neurological disorders such as various types of aphasia where the phonological network appears to be disrupted, so the general structure of language might then provide a certain minimization of these types of effects [43].

Alternatively, these features could be byproducts of the word formation process within language and not be indicative of language robustness. One argument in favor of these properties being artifacts is that there was likely little evolutionary pressure for their development, due to the rarity of strokes and similar injuries.

Furthermore, the time period within which such features could have evolved is short, on evolutionary time scales. On the other hand, it is possible that the relative infrequency of such debilitating linguistic disorders is due to the evolution of the robustness observed; these disorders would be much more common had this robustness not developed. Regardless, the extraordinary amount of robustness observed based on these common methods of node removal does seem intriguing and merits further examination.

Small-world properties. Although the languages differ in their history and linguistic characteristics, they all share a number of similarities in their network structure. An important commonality across the languages is that they all have the properties of a small-world network [69], that is, a high clustering coefficient and short vertex-to-vertex distance. The clustering coefficient can be calculated for each node (the average value of which is reported above in Table 2.1), and is the fraction of neighbors of a given node that are neighbors with each other. It is also known as network density. The vertex-to-vertex distance, also known as the shortest path length, is the shortest number of hops in a network to go from one node to another. Since these networks have many components, the shortest path length from one node to another is only calculated for nodes that are in the same component [47]. In addition, the mean shortest path length was calculated just within the giant component of each language.

As seen in Table 2.1, the values for the clustering coefficient are many orders of magnitude larger than what would be expected from a comparably sized random network—a network with the same number of nodes and edges—which can be calculated analytically [69]. The values of the clustering coefficient are also comparable to a similar measure referred to as transitivity, which is a more global measure of clustering [47].

On the other hand, the mean shortest path length of the language network's giant component, calculated using a random sample of 1,000 nodes, was similar to the mean shortest path length for comparably sized random networks, and significantly shorter than the overall number of nodes in the network, as seen in Table 2.1 [69]. The statistics of the giant component were used for comparable random networks, because the overall ratio of edges to nodes is far lower than within the giant component, due to the large number of islands in the networks.

It is thought that recognition and retrieval of words, while different, occur via a type of process known as spreading activation [14]. The idea is that as a word is being spoken or thought of, portions of one's linguistic network are activated (based on how much of the word has been already heard, for example). The activation of the rest of the network then proceeds outwards to the neighbors of these initially activated words, then to the neighbors of the neighbors, and so on, until the activation of the network is complete. In networks with small-world structures these sorts of processes tend to occur rapidly and robustly, due to the short path length, relative to the number of the nodes in the entire network. Given the small-world characteristics of the language network, it is not surprising that language processes are also rapid and robust.

However, it must be noted that, unlike in social networks, where it is clear what a distance of three friends is, for example, it is not entirely clear what the qualitative difference is between a distance of 5 and 6 within phonological networks. This is important when looking at the average shortest path lengths of the giant components of the different language networks. For instance, is it relevant that this value for Mandarin (10.1) is twice that of Hawaiian (5.5)? While it is likely that this number is most relevant relative to the size of the entire network (they are all orders of magnitude smaller than the size of the lexica examined), these differences might hint at more significant distinctions between the languages examined.

Furthermore, it might be that the small world property occurs fairly commonly in networks observed, and that it is less a relevant property of the language than simply an indicator that language is a fairly organic, unplanned construct. Therefore, we can actually attempt to determine if the path length within a network is an important property for language processing. A testable prediction is that words in the giant component will be processed at different speeds than those in the islands, due to the difference in how the spreading activation occurs.

Degree Distribution. The degree distributions of scale-free networks obey a power law function, $P(z) \sim z^{-\alpha}$. In contrast to many observed networks, we find that the language networks deviate from this behavior. Instead, they are reasonably fit to truncated power laws, similar to scientific co-authorship networks [48], as seen in Table 2.2 and Figure 2.2. A truncated power law, or a power law with an exponential cutoff, is defined as follows:

$$P(z) \sim z^{-\alpha} e^{-\frac{z}{z_c}}$$

Table 2.2 consists of the results to fit a truncated power law, an exponential and a stretched exponential to the degree distribution of each language. The truncated power law performs best overall, as exemplified by the high coefficients of determination (R^2). Figure 2.2 displays the degree distribution for each language, on a log-log scale, overlaid with the truncated power law best fit.

Table 2.2. Languages and coefficients of determination (R^2), for various functions.

Language	*Truncated Power Law*	*Stretched Exponential*	*Exponential*
English	0.998	0.962	0.624
Spanish	0.997	0.995	0.960
Mandarin	0.999	0.974	0.974
Hawaiian	0.997	0.970	0.884
Basque	0.999	0.999	0.963

Amaral et al. [6] found that if there is a constraint associated with the attachment of a new vertex (i.e., the vertex may only be able to accommodate a fixed number of edges), then a power law degree distribution, like that in the scale-free model proposed by Barabási and Albert [10], is not likely to be observed. In the language networks, a variety of constraints on word formation are present, such as the number of phonemes in the inventory of the language, the sequential arrangement of phonemes in words, the length of words, and the extent to which the language relies on morphemes (the smallest meaningful unit). All of these constraints limit the number of words that might be phonologically similar. Therefore, a truncated power law or similar distributions that decay faster than a traditional power law are reasonable as fits for the degree distributions in phonological networks.

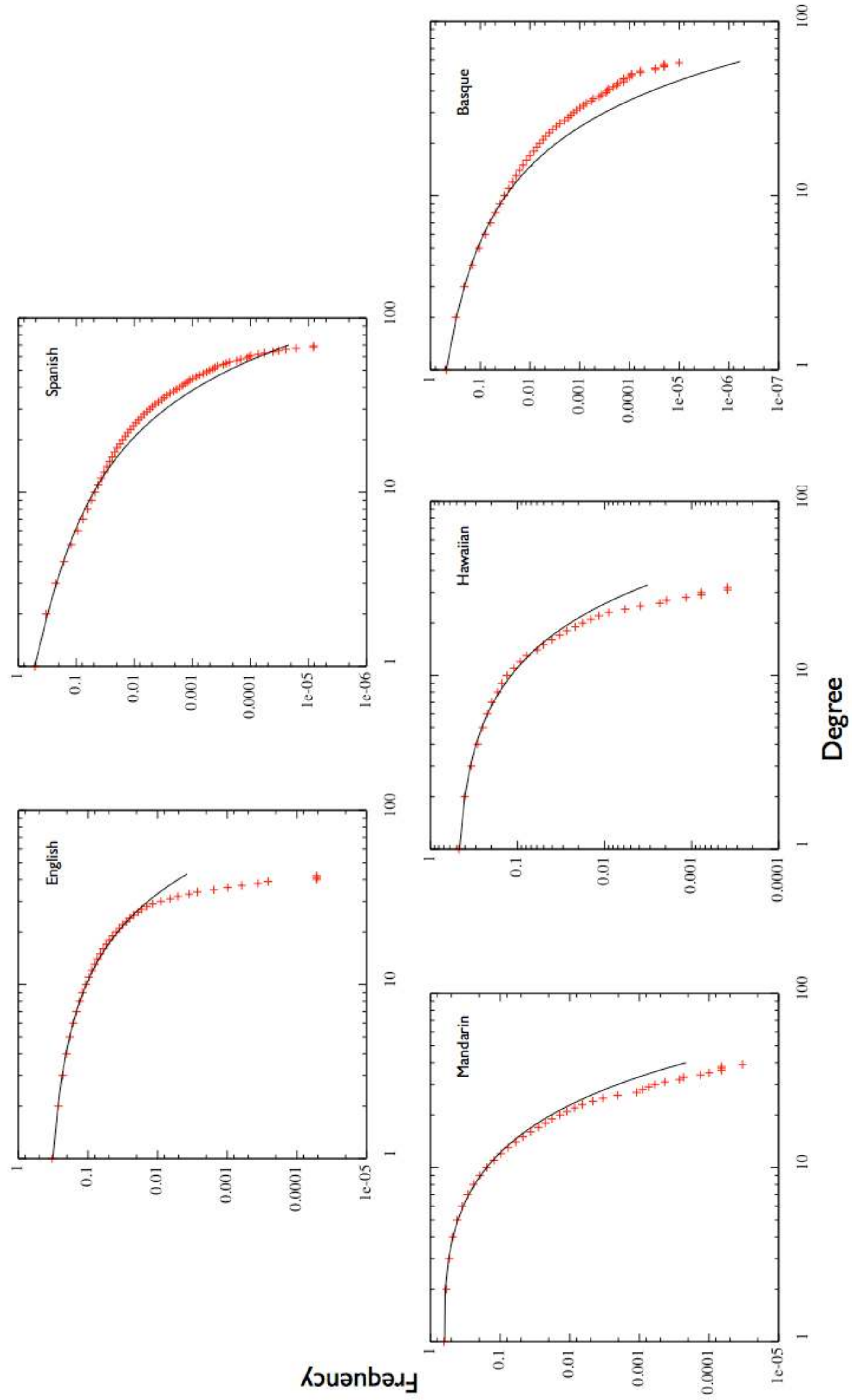

Figure 2.2. The degree distributions for all languages on log-log graphs. The black lines are the truncated power law best fits (since they overall performed the best). The final point for each distribution was not plotted, for legibility.

Characteristics of the other components. Recall that a smaller proportion of vertices in the language network compared to other types of networks were contained in the largest component. The remaining vertices (not counting the giant component) formed components that varied widely in size, and in fact obeyed a power law, as shown in Figure 2.3. The exponents are shown below, in Table 2.3.

Table 2.3. Exponents for powe -law fits of the distribution of the sizes of the islands for each language. The coefficients of determination (R^2) were all in excess of 0.999.

	English	*Spanish*	*Mandarin*	*Hawaiian*	*Basque*
exponent	3.75	2.21	4.94	4.17	2.48

Although a number of mechanisms might lead to a power law distribution [46], the power law distribution observed in the size of the components suggests models where a power law distribution is indicative of a phase transition. For example, within lattice percolation models, at the phase transition, ‘chunks’ of connected nodes within the system have a power law in their size distribution [59].

Kello and Beltz [32] argue that language is in a critical state, balanced between low memory effort and low disambiguation effort. Low memory effort is the extreme of a language having a single word encapsulating all meaning, while low disambiguation effort implies a language with separate words for every single concept. Natural language provides a balance, where there are many words in a lexicon, but where many words mean many different things. The power law distribution observed here of the sizes of connected portions within language, similar to what is seen in many physics-based models, might hint at such self-organizing criticality within language. Of course, such criticality is only one of many ways that a power law can occur [46], so further study is in order.

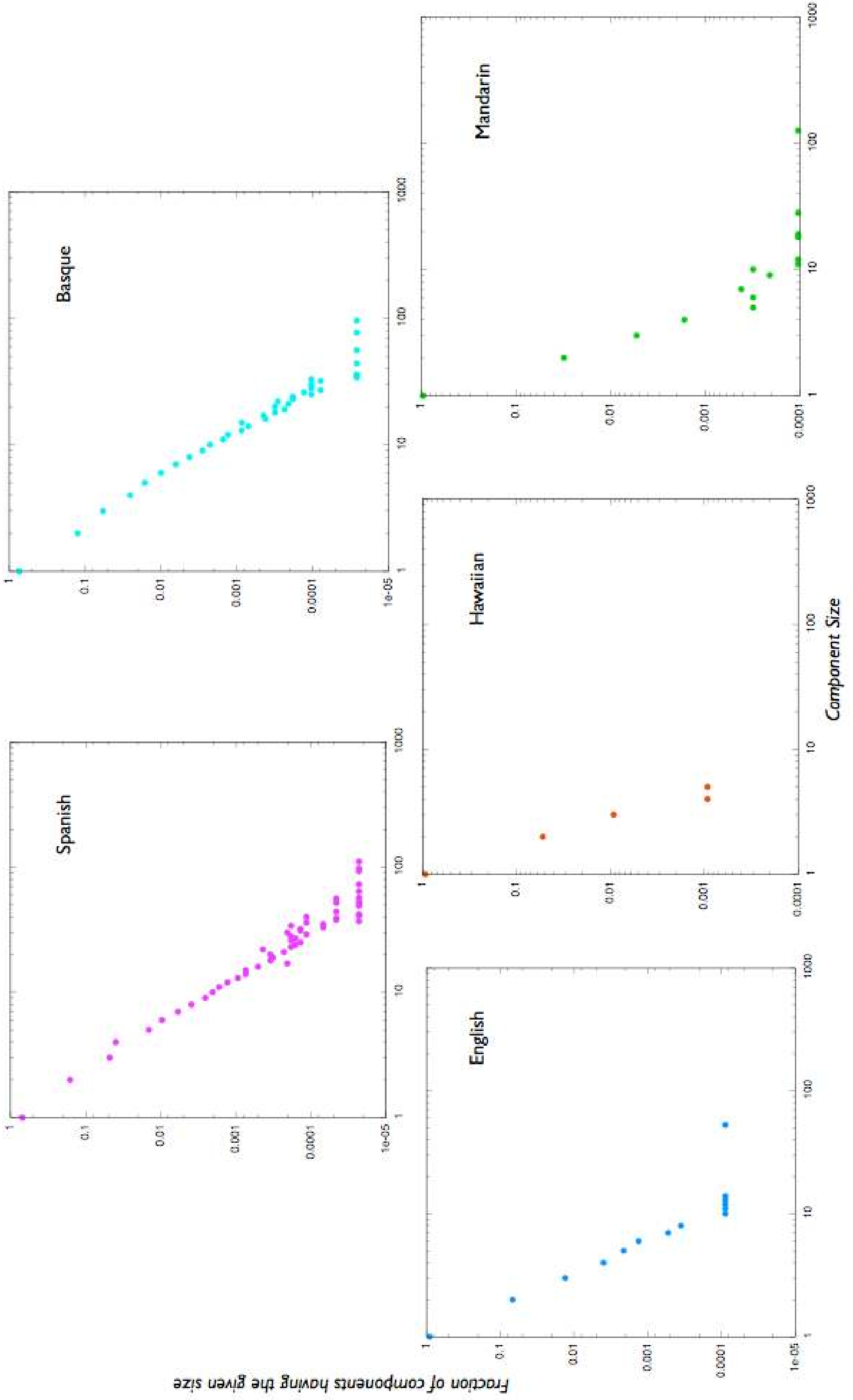

Figure 2.3. Size distribution of the smaller components (all components but the giant component).

Conclusion

The phonological networks of a variety of languages show a unique structure not found in other complex networks described in the literature. Despite coming from a diverse range of language families the networks all exhibited a common set of properties. Notably, the degree distribution is found to lie somewhere between a power law and an exponential distribution. The differences in degree distribution might provide a quantitative way to distinguish language families.

Furthermore, a small-world structure was observed, in conjunction with a number of other distinguishing characteristics. The giant components were far smaller than typically observed. The remaining vertices were distributed along a power law distribution of differently sized smaller components, or islands. The small sizes of the giant component together with the strong assortative mixing by degree and the robustness of the network to the removal of vertices is suggestive into the resilience of language processing in the brain, although further study is necessary.

Similarly, the power law distribution observed in the smaller component sizes suggests that language may be in a critical state akin to a system undergoing a phase transition. Researchers have posited that all natural languages must carefully balance opposing constraints, such as low memory effort and low disambiguation effort [32]. Observing a self-similar distribution in the size of the smaller component sizes is suggestive given the expectation of a critical state.

All of these observed characteristics hint at some deeper organization within language. Despite surface differences among languages, there are important commonalities that have implications for the processing of language in humans. The intriguing characteristics of these networks merit further investigation from network scientists as well as psycholinguistic researchers.

CHAPTER 3

WORD RECOGNITION AND THE STRUCTURE OF LANGUAGE NETWORKS[2]

Introduction

Information processing theories have advanced our understanding of the human mind, especially of the processes associated with language production and comprehension. In the processing of linguistic information, it has been generally assumed that a single model of language perception or production can account for processing in all languages. This fundamental assumption, however, has been challenged by several recent findings [16, 42].

Given a phonological network, using the same method of construction as in Chapter 2, we can look at the processing speed for word recognition relative to the degree, or number of neighbors, of a word. In English, it is well-known that a word with few similar sounding words, or neighbors, like pig (fig, wig, big, pin, pitch), is recognized more quickly and accurately than a word with many similar sounding neighbors, like cat (hat, fat, rat, mat, sat, cut, kit, cot, can, cap, calf; [40]). A word with low degree is considered to have a *sparse* neighborhood, while a word with a high degree is considered to have a *dense* neighborhood. Results from a study by Vitevitch et al. [66] found that words with a sparse neighborhood were responded to about 13 milliseconds more quickly than words with a dense neighborhood, on average (p-value < 0.05). In this study, multiple individuals were given multiple words to recognize (they had to identify a word as real or gibberish as quickly as possible). The statistical

[2] This work was completed in conjunction with Michael Vitevitch (University of Kansas) and Steven Strogatz (Cornell University).

test quantified the difference in processing speed for the two categories of words for each individual.

On the other hand, the opposite is true for Spanish: a word with a dense neighborhood is recognized *more* quickly and accurately than a word with a sparse neighborhood [65], with a difference in response time of about 21 milliseconds (p-value < 0.01).

What this means is that the processes involved in the recognition of words in English and Spanish operate in very different ways: in English, words with dense neighborhoods are processed more *slowly*, while in Spanish, words with dense neighborhoods are processed more *quickly*.

And viewed in terms of a network, words, or vertices, are processed differently depending on their degree. Due to this, it is natural to employ network analyses to gain further insight into the mechanism that underlies the processing difference observed in psycholinguistic studies of English and Spanish.

Although there are many similarities among the languages of the world (see the previous chapter), there are subtle differences that could account for the difference in processing. We constructed word networks based on phonological similarities of the lexicon, as mentioned above, and attempted to understand the reason behind the difference in processing.

Methods

The networks were characterized by smaller giant components than those typically observed in the literature. In fact, the giant components were small enough that a randomly chosen word was likely not to be found within the giant component. Given the prevalence of smaller components, or *islands*, examination of the constituents of the smaller components, in addition to the giant components, proved to be a constructive technique to expose subtle differences between the languages. We examined the giant components, as well as constituents of the 100 largest islands in the English and Spanish networks for insight into how subtle differences might explain the dramatic differences in processing.

Specifically, we looked at two properties of a word and its neighbors, to see if there was correspondence between word pairs (these properties are described below). Since one of the properties needed to be examined manually, we opted to use a sampling technique. We chose 100 words randomly from the giant component of each language, and then examined all of their neighbors, and then chose one word randomly from each of the 100 largest islands, and looked at each of their neighbors. By examining the similarity or difference in properties between a word and its neighbor, we hoped to better understand the reason for the difference in processing between English and Spanish.

Results

Position of Neighbor Overlap. The first property we examined was the position of neighbor overlap. Just as words that are connected by an edge differ by a single sound, the reverse can be examined: where are the connected words similar? Each word's phonological segments were divided into roughly three parts, and the similarity between a word and its neighbor was examined. In this way, each pair of connected words were categorized into one of three groups: beginning, end, and beginning and end (difference in the middle position) overlap. The table below gives examples of the three types of overlap:

Table 3.1. Overlap types, with the overlap shown in bold.

beginning	**sa**p & **sa**t
end	g**ave** & s**ave**
beginning and end	**t**i**p** & **t**a**p**

For English and Spanish, the overlap within the giant components were roughly similar, with no clear overlap category winning out, as shown in the top row of Figure 3.1.

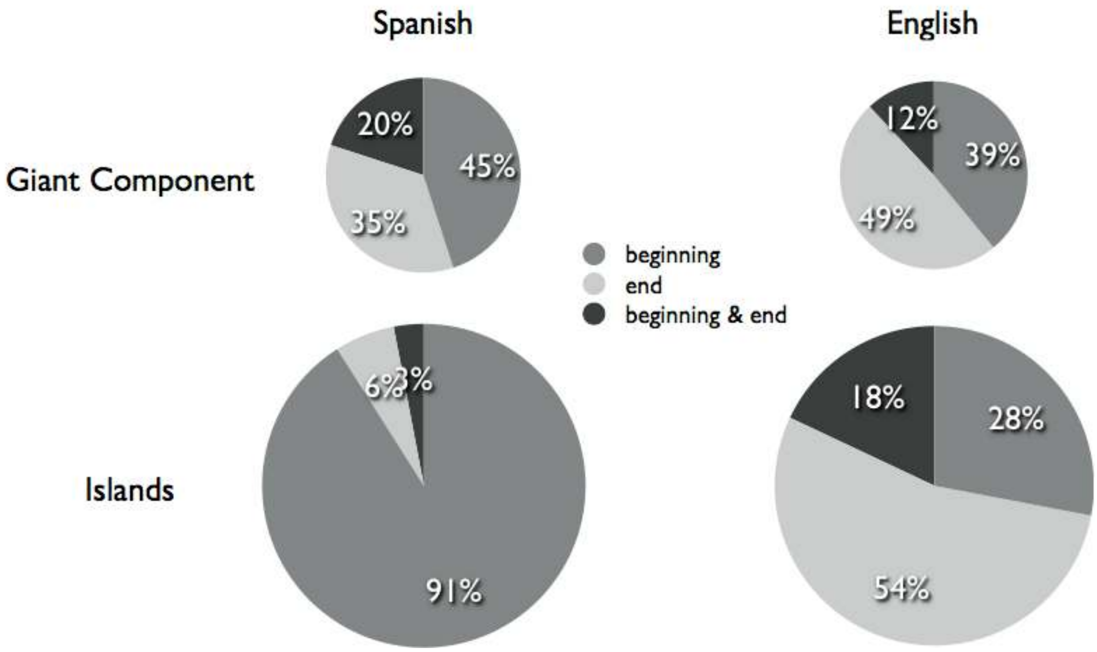

Figure 3.1. Distribution of overlaps in the neighbor word pairs examined.

However, when the overlap within the islands is examined, a difference becomes clear. The bottom row of Figure 3.1 reveals that over 90% of the overlap within Spanish is a beginning overlap, while in English, the islands there is still a reasonable portion of end and beginning and overlap.

What does this mean? Having islands comprised of elements with predictable phonological overlap could facilitate the retrieval of similar words. Therefore, the structure of Spanish might allow for more rapid retrieval of word forms that follow this predictable pattern as compared to English.

Semantic relationships affect processing. Next, we looked at the neighbor word pairs in both languages to see if a word and its neighbor are semantically related (this was the part that required manual examination). The results are below.

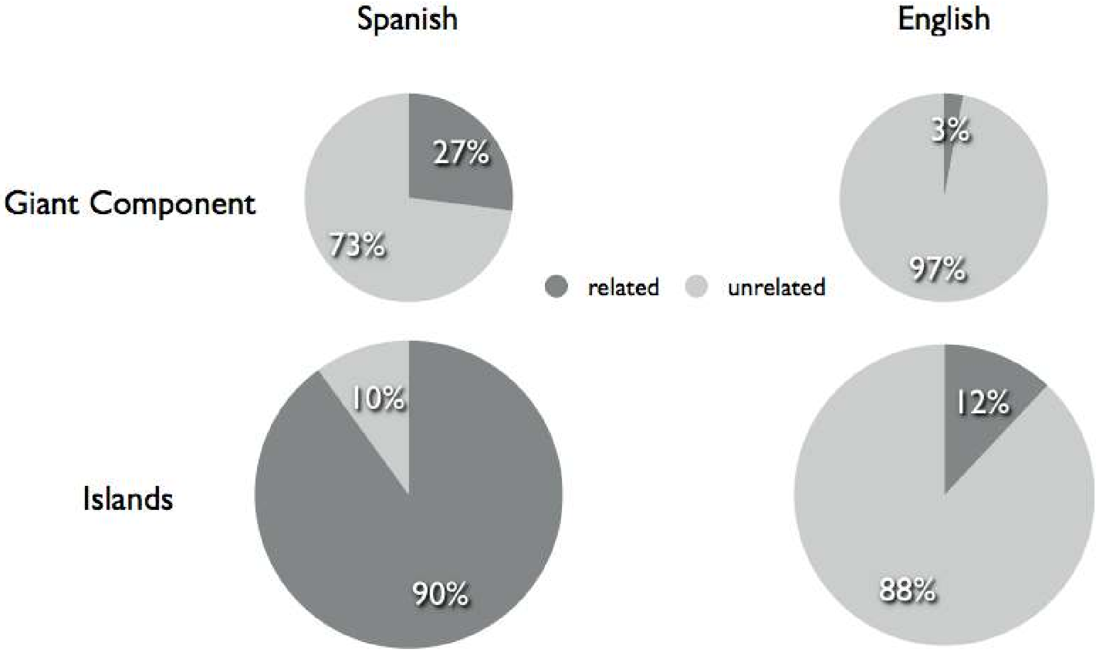

Figure 3.2. Semantic relationships between neighbor word pairs.

As can be seen, in both the islands and giant component, Spanish has a great deal more semantic similarity between neighbors than that found in English. Why might this be? Linguists generally categorize languages into one of two types, in terms of how the meanings of words are modified. Derivational morphology changes the meanings of words by adding a morpheme (or unit of meaning), such as changing the adjective *good* to the noun *goodness*. In contrast, in inflectional morphology a morpheme is added to tag the word with additional meaning, such as person, number or tense. For example, going from *dog* (singular) to *dogs* (plural) is an inflectional change.

Having word pairs that are not only similar in phonology, but also similar in meaning (i.e., more inflectional) might facilitate the retrieval of the correct word form. This is what is found in Spanish, a language that is highly inflectional. It is so inflectional, that all of the islands in Spanish contained inflectional forms of one or two verbs. For example, the largest Spanish island, containing 111 words, is comprised entirely of conjugations of preguntar (to ask) and presentar (to present; e.g., presentar, presenta, presente, presenten).

On the other hand, if a word is surrounded by words that are phonologically similar but unrelated in meaning (such as what is found in more often in English), recognition of the spoken word might have to overcome the competition among its phonological neighbors. Figure 3.3 illustrates this difference between typical islands in English and Spanish.

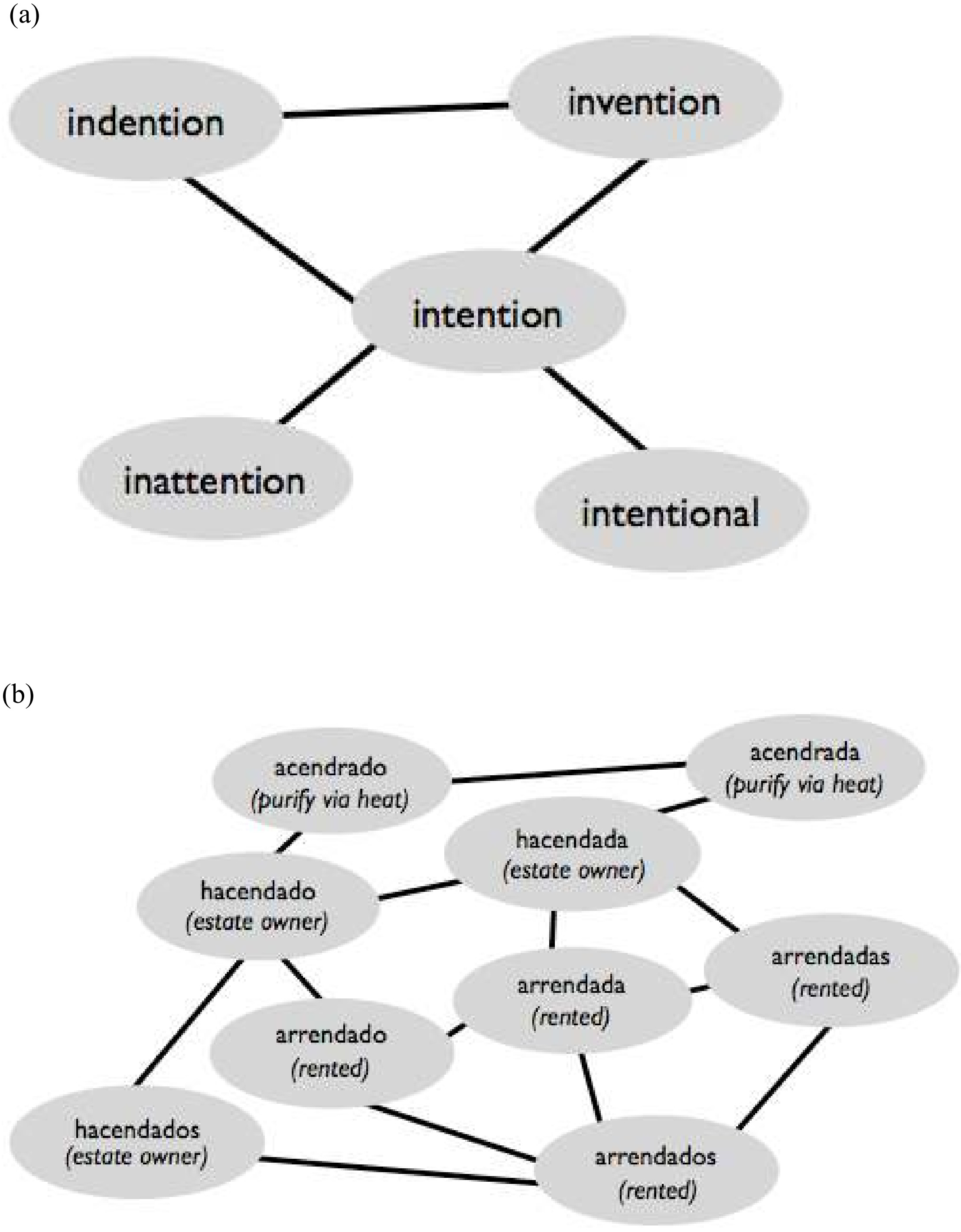

Figure 3.3. (a) A small island from English. Only the words 'intention' and 'intentional' are related semantically (and, in fact, are related derivationally). (b) An island from Spanish (with the root meanings below each word). While there are nine words, it can be seen that there are only three root meanings.

Discussion

In this way, we can explain the difference in processing between the languages. Languages similar to Spanish, due to their confluence of phonological and semantic networks, can be expected to facilitate recognition of words with dense neighborhoods. On the other hand, languages which are uncorrelated in these cognitive networks should be expected to perform similarly to English and have slowed processing of words with dense neighborhoods. This is a prediction that can be readily tested for other languages.

However, it must be noted that there is an important limitation within the present study. The English dataset, while a classic one, suffers from a fairly important shortcoming that is particularly relevant here: it lacks nearly all inflections. Due to this, the difference between Spanish and English related to the semantic similarity of neighbors is artificially increased.

Luckily, there is a dataset for English, known as CELEX, which does include the inflected forms of words [9]. It is quite likely that the results found here will be qualitatively similar if redone with this different dataset (due to the highly inflectional nature of Spanish, in comparison to English). Nonetheless, a careful repetition of the analysis with the CELEX data is in order, and will be completed prior to submission of this chapter for publication. Furthermore, the distinction in the psycholinguistic experiments between words with dense and sparse neighborhoods can be reexamined using a different dataset as a check to be certain the results are not changed greatly using different lexica.

Assuming that the results obtained here are valid, this concept of overlap and facilitation among distinct modes of cognitive processing has larger implications. If the environmental input can be processed multiple ways, depending on the similarity in cognitive modules, processing can be facilitated or impeded.

For example, there is the well-known Stroop task. In this task, a number of color words are presented to the subject, with the letters of each word colored (see Figure 3.4). The goal of the task is to say as quickly as possible, what color the letters are in each word.

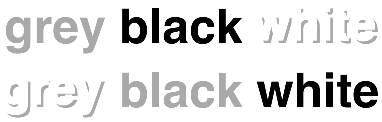

Figure 3.4. The Stroop task for greyscale readers. The top line facilitates the recitation of the text colors, while the bottom line hinders recitation of the text colors.

For example, when the word ‘black’ is written in black letters, it’s easy to say ‘black’ quickly. But when it’s written with grey letters, the conflict slows down your mental processing and it takes a bit longer to say ‘grey’.

Similarly, the orthography of languages can be examined. Orthography of languages, or the way the language is written, can fall somewhere on the spectrum between deep and shallow. For example, Italian has a shallow orthography since the way you write a word is invariably the way you pronounce it. On the other hand, English has a remarkably deep orthography. This is illustrated by the joke brainteaser ‘GHOTI’ [57]. How do you pronounce this word? Why ‘fish’, of course! The ‘gh’ from ‘enough’, the ‘o’ from ‘women’ and the ‘ti’ from ‘nation’ together yield the piscine pronunciation. Clearly this is a silly example. But it turns out that languages with shallow orthographies can be read more quickly [61], providing an additional example of how overlap of distinct modes of cognition can facilitate processing.

The present analyses highlight the advantage of looking at the constituents of each language network and of how difference in overlaps between network structures may lead to observable differences in language processing. While only English and Spanish were considered here, an examination of the phonological and semantic networks other languages of the world, and their similarity or dissimilarity, can help us better understand how facilitation and competition functions in word recognition. This hints more generally at how much processing variation might exist among the languages of the world, due to differences in processing modules.

CHAPTER 4

AN EXPLANATION OF SUPERLINEAR SCALING FOR INNOVATION IN CITIES[3]

Introduction

It has been known for nearly a hundred years that living things obey scaling relationships. Max Kleiber first recognized that the metabolism of different organisms scale according to their masses, raised to a three-quarters power [33]. More recently, Geoffrey West and his colleagues have provided a theoretical explanation for this scaling law, and for many other allometric laws found in biology [70, 71]. Their theory is based upon the fractal-shaped branching structures within all organisms (such as circulatory systems) that provide energy to every part of organisms. They argue that the larger the organism, the more efficient the system that can be constructed to provide energy, thereby yielding this sublinear exponent.

More recently, West and his team examined a variety of properties of cities. They found that cities, which have long been compared to living things, obey scaling relationships as well [11]. Similar to living things, cities have economics of scale, yielding sublinear scaling for such quantities as the number of gas stations within a city as a function of its population. In other words, you need fewer gas stations per person, in a bigger city. Examples of such scaling laws are shown in the upper portion of Table 4.1.

[3] This work was completed in conjunction with Jon Kleinberg (Cornell University) and Steven Strogatz (Cornell University).

On the other hand, cities also exhibit superlinear scaling, which appears in relation to sociological quantities. As shown in the lower part of Table 4.1, properties of cities related to economic productivity and creative output have exponents that are all found to cluster between 1 and 1.5, with the mean around 1.2. Thus, the productivity *per person* increases as a city gets larger. However, this superlinear scaling has not been given a satisfactory mathematical explanation.

Table 4.1. A variety of urban quantities and their exponents. For instance, if y denotes the number of gas stations in a city of population N, the data in [11] show $y = cN^{\alpha}$, with $\alpha \approx 0.77$. Taken from Bettencourt et al [11].

Urban Indicators (y)	*Exponent (α)*
Gasoline stations	0.77
Gasoline sales	0.79
Length of electrical cables	0.87
Road surface	0.83
New patents	1.27
Inventors	1.25
Private R&D employment	1.34
'Supercreative' employment	1.15
R&D establishments	1.19
R&D employment	1.26
Total wages	1.12
Total bank deposits	1.08
GDP	1.15

Here we suggest a theoretical explanation for the superlinear relationship between population size and innovation found in cities, with a reasonable range for the exponent. Due to the sociological nature of the variables being measured, it is natural to use a network model of a city, since it is reasonable to assume that network effects must underlie the superlinear effect.

Model and Results

We first assume that all social interactions and relationships are arranged in a hierarchical tree structure [37, 68]. Picture a binary tree, or in general, a tree where each branch splits into *b* new branches. For example, in a city, each person is in a household, and there are many households on a block, and many blocks in a neighborhood, and so forth. Or the grouping could be based on your family tree, or corporations, or many other ways to group individuals. While in reality each individual belongs to many independent hierarchies, here we simplify it as a single hierarchy, with branching number *b* (greater than or equal to 2). We view the total system as a city, meaning that a city of population *N* represents a single tree that contains *N* leaves. On top of the tree structure, which serves to determine the social distance among nodes, a graph is placed showing who actually knows who. Thus, social connections are created, with each individual having the same number of outgoing connections. These connections form the basis for a city's productivity.

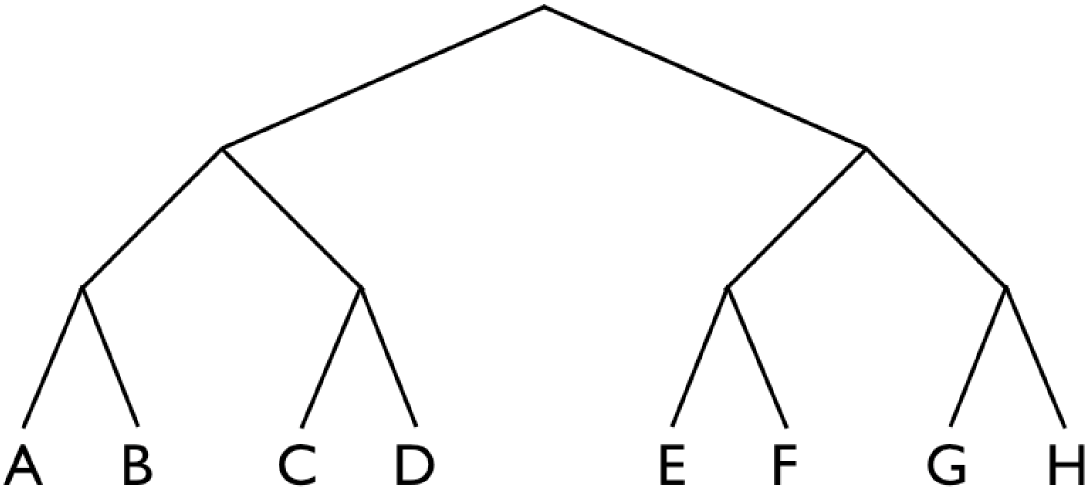

Figure 4.1. Network representation of the social structure of a city's inhabitants.

The total creative productivity of the city can then be determined by calculating the total productivity of a single person. To do this, we simply multiply the productivity of a single person by the number of individuals in the entire city to yield the productivity of the city. To calculate the total productivity of a single person, three separate effects must be considered: (1) the probability of connecting to someone at distance *d*; (2) the number of available people at distance *d*; (3) the creative output that is obtained by linking to a single person at distance *d*. Multiplying these together gives the productivity due to one person linking to all of his collaborators at distance *d*, as seen below:

$$\left[\frac{\#\,\text{contacts at distance } d}{\#\ \text{ of people at distance } d}\right]\cdot\left[\#\ \text{ of people at distance } d\right]\cdot\left[\frac{\text{output at distance } d}{\text{contact at distance } d}\right]$$

By summing this term over all distances, the total creative contribution of a single individual is obtained. The functional form of each term in the above recipe for calculating the productivity of a single individual is discussed below.

Taking the first term, the social connections between collaborators are constructed such that the likelihood of forming a connection at a certain social distance drops off exponentially fast with distance. That is, the probability of a connection being made between nodes of a social distance *d* (where *d* is the height of the first common internal node) is assumed proportional to $b^{-\alpha d}$, where α is a tunable parameter greater than or equal to zero.

It's clear that the connection probability should decay with social distance, but why exponentially? We have assumed that the social network tree is self-similar at all levels (values of *d*). Since the tree is self-similar, it makes sense to have the function also be self-similar (scale-free) with respect to the value of *d*, and doing this yields an exponential function (this assumption is relaxed in the next section).

Since at each increase in d there are exponentially more potential contacts to interact with, we multiply the above function by a second term, b^d, which means that as we increase d, while the likelihood of making a connection decays, there are exponentially more contacts to make. To keep things simple, we suppose connections are only made between residents of the city (connections outside a city are viewed as far less fruitful for productivity, and are ignored).

Lastly, the usefulness of a social connection within a city is assumed to vary with its social distance. For example, let's assume that there is a benefit as social distance increases. This can be explained as being due to the fact that individuals that are socially distant are exposed to different ideas and experiences, and that collaboration between two more socially distant individuals is more productive than interaction between ones that are closer. However, the value of a social connection is left open, and simply assumed to be proportional to $b^{\beta d}$, where β is a tunable parameter that can hold any value (even negative values, allowing the value of a connection to decrease with distance). An exponential function is reasonable here as well, if we assume that a connection's innovation potential is proportional to the number of individuals that lie between you and the individual on the other end of the connection in social space. This is explained as being due to the potential for all of these in-between individuals to provide fodder for innovation (such as by interacting with you in some manner). This assumption is also relaxed in the next section.

Therefore, as the social distance between two individuals increases, the number of individuals in between you and your contact grows exponentially, implying that the productivity of that connection will also grow exponentially. Note that if β is zero, then all connections are equally beneficial.

As the number of connections becomes large, the value of the total productivity $P(N)$ of the social connections within a city is given by

$$P(N) = N\sum_{d=1}^{\log_b N} b^{-\alpha d} b^{d} b^{\beta d}$$

In summary, the first term, $b^{-\alpha d}$, is the probability of connecting at distance *d*. The second term, b^{d}, is the number of nodes at distance *d*. And the final term, $b^{\beta d}$, is productivity per connection. So, when these are multiplied together and then summed for each distance, they yield the average productivity per node for the entire network. When multiplied by *N* we get the productivity for the entire network.

This can be summed exactly since it's just a finite geometric series, and we get the following solution:

$$p(N) = N\left(\frac{b^{\beta+1}}{b^{\beta+1} - b^{\alpha}} N^{\beta-\alpha+1} - 1\right).$$

For large values of *N*, we find $P(N)$ is proportional to $N^{\beta-\alpha+2}$, if $\alpha < 1+\beta$. Whereas, if $\alpha > 1+\beta$, the function $P(N)$ becomes linear, because the geometric series converges to a constant as *N* becomes large:

$$N\sum_{d=1}^{\log_b N}\left(b^{\beta-\alpha+1}\right)^{d} \xrightarrow[N>>1]{} N\cdot\frac{b^{\beta-\alpha+1}}{1-b^{\beta-\alpha+1}}$$

This is found to be in good agreement with numerical evaluation of the above summation, as can be seen in Figure 4.2.

The growth of the productivity function with distance is not essential. The key is, rather, that as a city increases in size, it is more likely to contain socially distant contacts. Thus, even if β is slightly negative (meaning that more distant connections are *less* productive), as long as $\beta - \alpha + 1$ is greater than zero, the densification of the social network due to the increasing size of the city means that the productivity will grow superlinearly. For the special case when β is zero (all connections are equally beneficial), the exponent $\beta - \alpha + 2$ must lie strictly between 1 and 2, which is where all the measured exponents for urban innovation lie. This is because we are assuming that $0 < \alpha < 1+\beta$ to get superlinear behavior, which means that α is between 0 and 1.

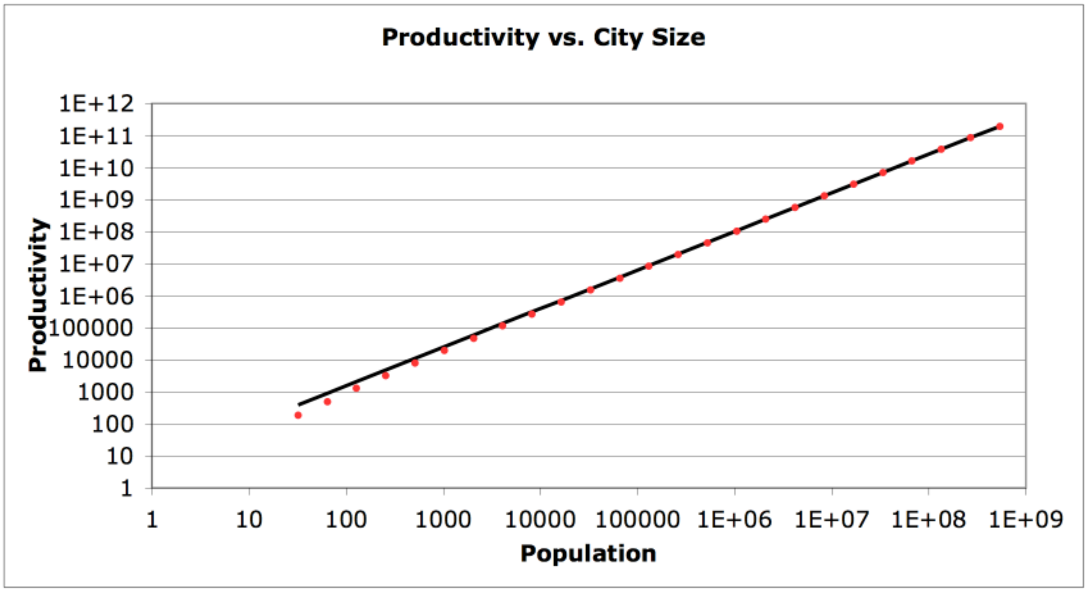

Figure 4.2. Simulation and fit for *P(N)*. The points show the value of *P(N)*, calculated for $\beta = 0.3$ and $\alpha = 1.1$. The least-squares approximation of the exponent is 1.205, and the expected value is $\beta - \alpha + 2 = 1.2$, for large values of N.

Of course, other parameter relationships are also capable of yielding the expected range of superlinear exponents and it is likely that β is greater than zero. If β is greater than zero, while there is an exponentially *decreasing* probability of connecting to someone at a "social distance" *d* away from you, if you do connect with such a person, you get a creative benefit that is exponentially *increasing* in *d*. In general, this model is a reasonable explanation for the values observed within cities related to productivity and innovation, and can be fit properly to explain the superlinear exponents observed within cities.

Expansion of the Analysis

The assumptions of exponentials for the three functions that make up the sum discussed above are stringent ones. What happens if we relax these assumptions?

Using a numerical simulation, each of the components of the sum can be modified, and we can graph the resulting scaling relationship and see if it remains superlinear. And in fact, the model is robust under a variety of situations. For example, instead of using an exponential for the number of nodes at distance *d*, if we use a 'linearithmic' function ($d \cdot \log(d)$), the resulting function asymptotically approaches a superlinear function, as seen in Figure 4.3. By doing this, we are implicitly changing the structure of the social distance tree, such that the number of nodes no longer grows exponentially with distance. A similar superlinear result can be obtained by replacing the creative benefit function with a linearithmic function and leaving the other two functions exponential.

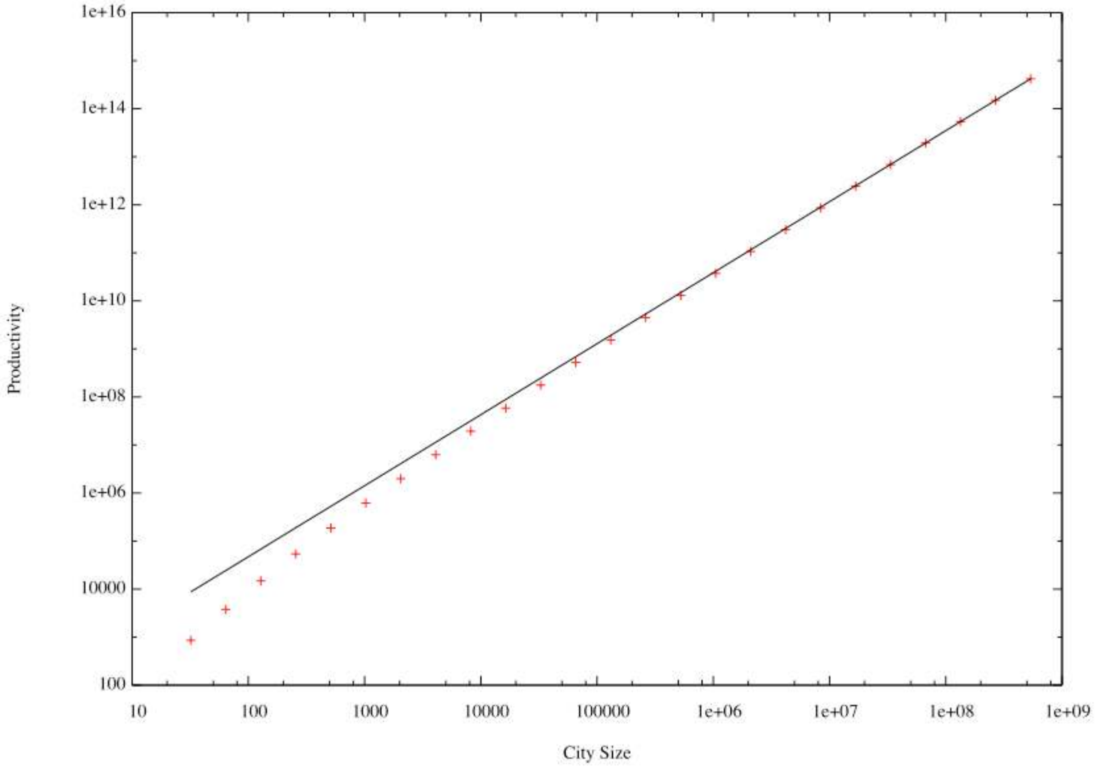

Figure 4.3. A 'linearithmic' function. Using the function $b^{-\alpha d} \cdot (d \cdot \log(d)) \cdot b^{\beta d}$ as the term within the sum for *P(N)* (where $\alpha = 0.5$ and $\beta = 0.9$), the resulting function mimics a power law, with an exponent of 1.48.

In fact, even if all three functions are linear, the sum still grows superlinearly with *N*, as seen in Figure 4.4. Since, if the function is proportional to d^3, using the Euler-MacLaurin summation, we find that $P(N) \approx N \cdot (\log_b N)^4$, which grows a bit faster than linearly.

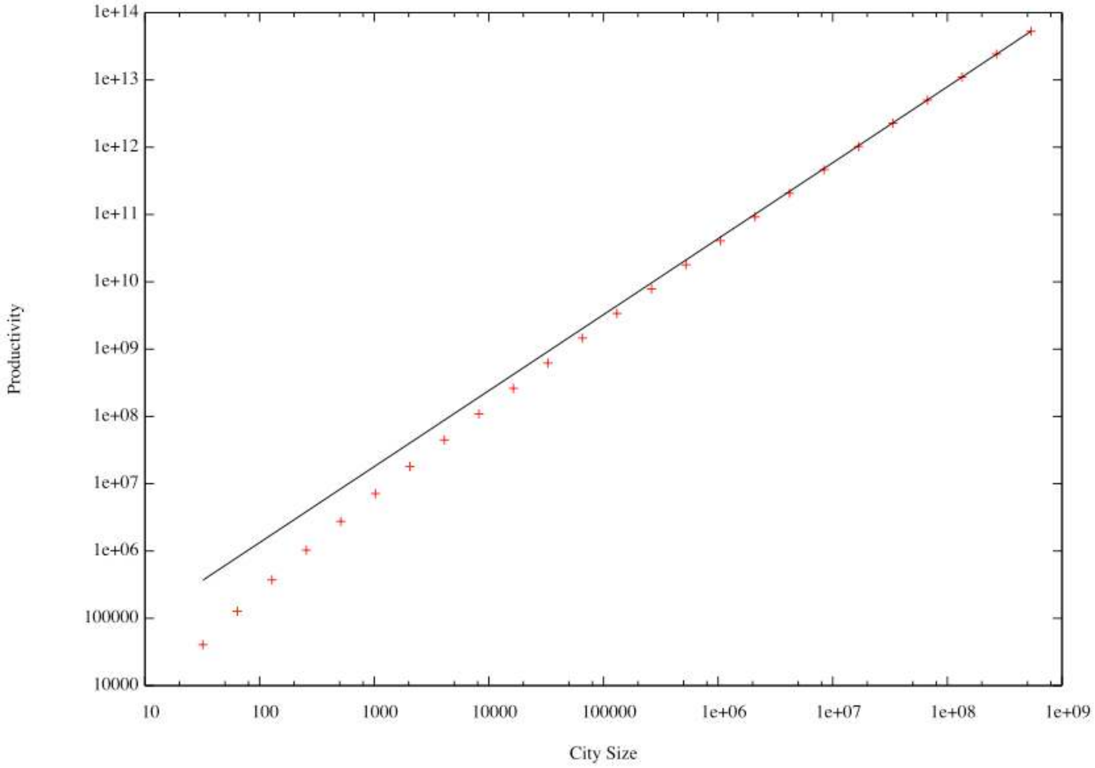

Figure 4.4. Linear Functions. Using the function $(50 - \alpha d) \cdot (1 + \beta d) \cdot (2d)$ as the term within the sum for *P(N)* (where $\alpha = 0.4$ and $\beta = 0.1$), the resulting function still mimics a power law, with an exponent of 1.13.

However, if the average productivity per node grows with *N* but only at the rate log(*N)*, then the rate of growth is only slightly superlinear, mimicking a power law exponent of 1.05, as seen in Figure 4.5. Slightly faster than logarithmic growth for the summation appears to be required for superlinear growth of *P(N).*

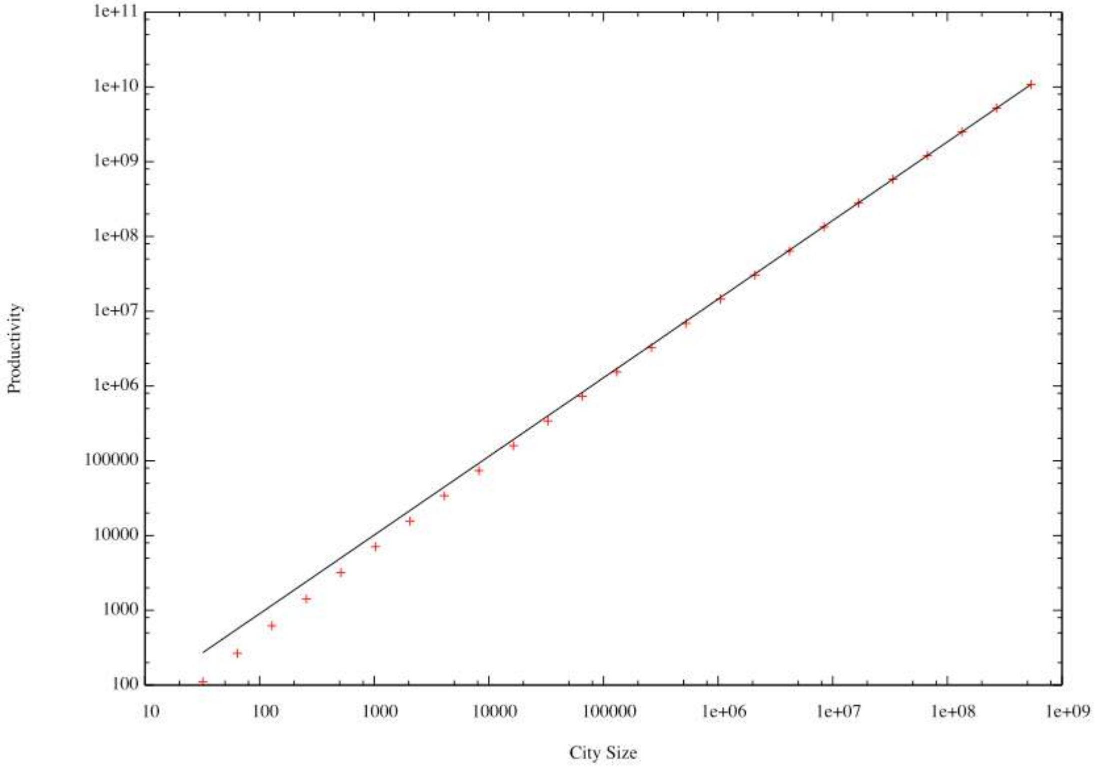

Figure 4.5. The function $P(N) = N \cdot \log_b N$ mimics a power law, with an exponent of 1.05.

What can be seen is that using fairly loose assumptions, superlinear growth can be obtained. Notably, these functions need not be power laws. They can simply be superlinear functions (such as a $P(N) \approx N \cdot (\log_b N)^x$), that mimic power laws. Therefore, this could also be true of the city productivity data observed: the productivity functions observed are superlinear, but are not necessarily power laws, which further measurement might help resolve.

Discussion

Ultimately, what are most important are the distant ties present in large cities. These long-distance ties (due to α), which are prevalent in a higher proportion when there is a larger population, provide the potential for productive social interactions.

Granovetter’s classic paper “The Strength of Weak Ties” examines this explicitly [27]. As part of his study, Granovetter examines the structure of Boston’s West End and its inability to organize against a neighborhood urban renewal project, which included the large-scale destruction of buildings to make room for new residential high-rises [15]. While the West End contained many strong ties between individuals due to most individuals being lifelong residents, these strong ties often resulted in cliques, where everyone was connected within a single group. Crucially, however, Granovetter argues that there were few, if any, ties between these strong-knit communal cliques. Since personal ties are necessary for information spread and organizational ability (‘people rarely *act* on mass-media information unless it is also transmitted through personal ties’ [27], p. 1374), the inhabitants of the West End would have had a great deal of difficulty in organizing their opposition to the municipal project. In contrast, if there had been interaction throughout the social hierarchy, such as between communities within the neighborhood, they might have been able to organize successfully.

Along these lines, Charlestown, a similar Boston neighborhood, was able to successfully organize against urban renewal. Granovetter argues that there was a rich interconnection between different communities, allowing for wider coordination within the community. Similarly, productivity in general is governed by the presence of these longer ties, only available in large cities, or any community that allows for large-scale interaction.

Of course, any good model must be testable in order for it to rise above the level of a pleasant story. That is why in Chapter 5, we indirectly attempted to determine what the value of β is (it seems to be around zero). In addition, given a network of social interaction for a city, its hierarchical social structure can be determined [13], to see if it conforms to the type of growth with distance that is discussed above. This has not yet been done, but it should not be difficult, given the relevant dataset.

In addition though, there are other possible explanations for this superlinear scaling in cities. For example, it could be that larger cities have a larger proportion of more highly educated individuals, which is enough to yield increased productivity per capita. Or it could be that larger cities simply have a greater transient population, which provides more fodder for different ways of thinking about the world, yielding a higher rate of productivity per individual. By distinguishing between our model based on social interaction and other competing models, we can get a better sense of how good our model is. But how can this be done?

While cities do exhibit superlinear scaling for a variety of quantities, many cities do not lie exactly on the predicted curve for a given property, based on a curve of best fit. For example, some cities will produce more patents than expected, while others will produce far fewer than expected, given their population. By looking at the pattern of social interaction in the underperforming cities as compared to the overperforming cities, we can determine how reasonable our model is. And by examining how well other models can predict this type of variation, as opposed to ours, we can determine what is the likeliest explanation for superlinear scaling within cities.

Nonetheless, as argued above, the presence of socially distant ties within a single city can be a powerful force. By using simple assumptions about social interactions, we gain a useful tool in understanding the mathematical behavior of innovation and productivity in cities.

CHAPTER 5

A SIMPLE MODEL OF COLLABORATION AND INNOVATION[4]

Introduction

According to conventional wisdom, intellectual and cognitive diversity is good. Bringing together people with different backgrounds and different ways of thinking about the world helps find problems and provide solutions [50]. The claim that cognitive diversity promotes increased productivity and problem solution has been documented in a variety of areas: corporate management, organizational structures, business idea generation and so forth [29, 36, 72]. Of course, it must be understood that increased diversity need not always lead to higher productivity. What appears to be at work here is that collaboration at a higher distance yields a higher variance; diverse people are just as likely to create something wonderful as they are to reach impasses due to their differences, accomplishing nothing [50]. But on the whole, diversity appears to be beneficial for solving problems.

When it comes to understanding research within academia, not as much is known. We recognize that collaborative academic research is beneficial in general: larger groups produce research that is more highly cited [73]. This is true even when accounting for self-citation. So, groups can do science well. But while we know that increased intellectual diversity helps in business, is this true within science? The answer should be obvious: of course it does! Why else would be interdisciplinarity be hyped so much?

[4] This work was completed in conjunction with Scott E. Page (University of Michigan), Jon Kleinberg (Cornell University), and Steven Strogatz (Cornell University).

Interdisciplinarity has become a very trendy word, and areas of research that bridge departmental gaps are increasingly important [3, 19]. The National Science Foundation even has grants specifically designed for interdisciplinarity [18]. Clearly there is the sense that interdisciplinary research is important and beneficial. And as I am a byproduct of interdisciplinary research, I too recognize the truth to this, and its value.

However, attempting to demonstrate that increased scientific distance between collaborators increases the scientific impact of the collaboration is difficult. One may assume that scientists who are farther apart in terms of their areas might have more to offer each other and be more productive. However, this is not necessarily the case, and a clearer notion of the empirical relationship is important.

To understand these issues, we examined two datasets—a coauthorship graph with the number of citations for scientific publications, and a similar dataset for United States patents—and attempted to tease out the relationship between distance and scientific impact. Next, we examined a series of computational models of collaboration and innovation, in order to better understand under what sorts of circumstances any meaningful relationship can actually be unearthed from the data.

We find that it is very difficult to measure the beneficial effects (if any) of interdisciplinarity, at least in these data sets and models. This is due to a number of factors, based on how research is often conducted.

Methods

Two datasets were used in this analysis: arXiv data from Paul Ginsparg [24], and United States patent data from the National Bureau of Economic Research [28]. The arXiv data consisted of preprints in various branches of the physical sciences, and the patterns of authorship and citations among them. Preprints from the following areas were used: astrophysics; general relativity and quantum cosmology; high energy physics—experiment; high energy physics—lattice; high energy physics—phenomenology; high energy physics—theory; nuclear experiment; nuclear theory. The papers were derived from the early 1990's to late 2000's, with the specific years varying from area to area (we also used a subset of the data from 1995-2000, and found similar results). The patent data came from the years 1975 to 1985, in order to control for changes in citation styles (later patents cite more often).

The following procedure was carried out for both datasets (with suitable cleaning). Papers and patents were handled separately, but are referred to as *items* more generally. The names of authors of all the items were extracted and a coauthorship graph was constructed, where nodes are authors and the edges connect individuals who have coauthored an item. Coauthors have at most a single edge connecting each other, even if they have worked on many papers together.

Names were more easily derived from the patent data (it was more structured), so the coauthorship network is necessarily noisy. This problem is exacerbated by the fact that multiple individuals often have the same name. Newman encountered a similar problem, but found that multiple mechanisms for network construction based on names do not have a marked effect on the network properties [48]. We were unable to remove self-citation from the arXiv data due to the form in which we received it, so no attempt was made to remove self-citation from the patent data, for consistency.

For a given item, the two relevant quantities being measured are scientific distance and the impact of each item. The impact of an item was determined through the number of citations of an item from within the dataset. The distance between coauthors was calculated using the coauthorship network as follows: the edge between two coauthors is removed, and then the length of the shortest path between them is determined, as seen in the figure below. This distance is referred to as the gap distance (Watts discussed a similar metric, calling it the 'range' of the edge that was removed [67]).

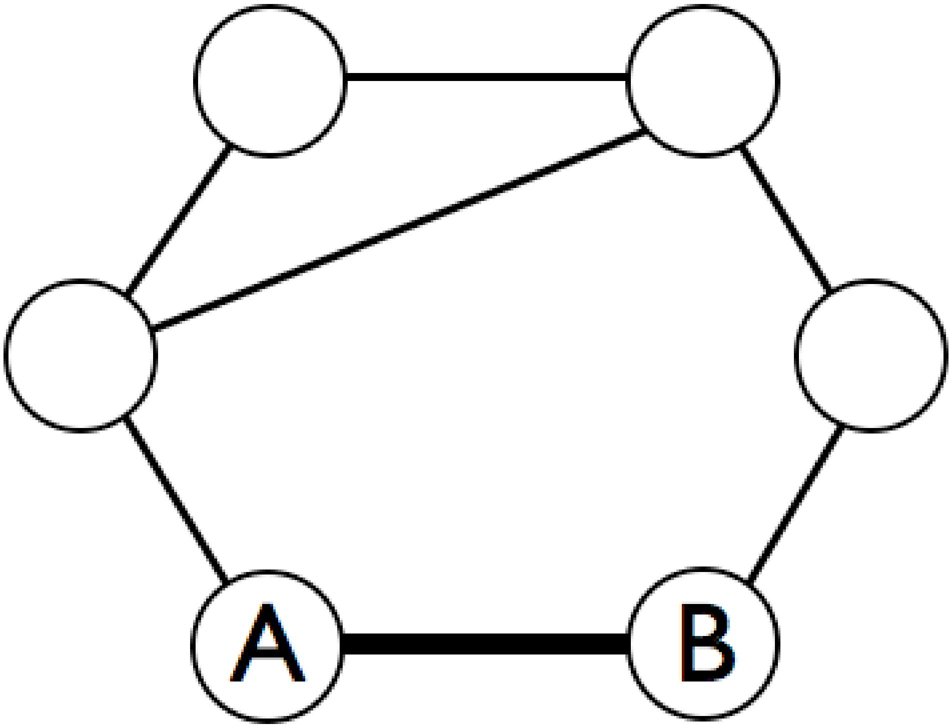

Figure 5.1. The gap distance between nodes A and B is 4. This is calculated through the removal of the thick edge, and then calculating the shortest path between points A and B.

Only items with two coauthors were examined, for ease of analysis. For all of the items in each dataset, the gap distance of the coauthors was calculated, along with the number of citations. For each distance the mean, median and standard error of the mean of the number of citations for all of the papers with that gap distance was calculated.

Empirical Results

The means, medians and standard errors of the means for both the arXiv and U.S. patents are visible in Figures 5.2 and 5.3.

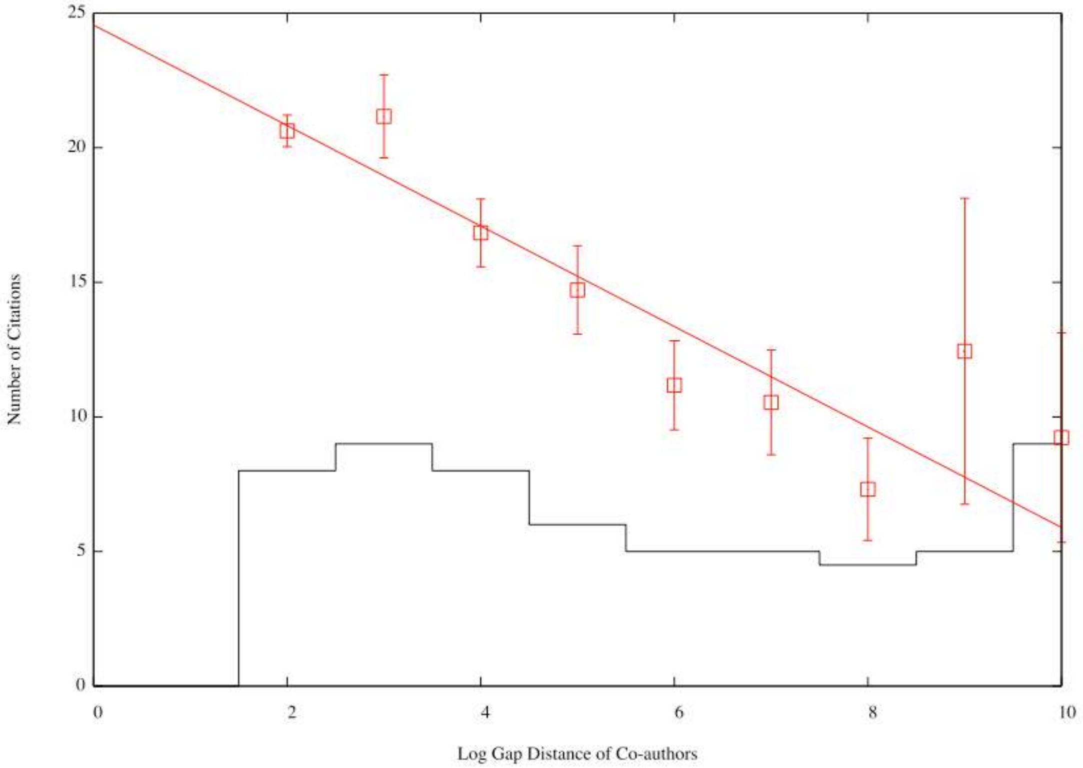

Figure 5.2. arXiv Data. The black curve is the *median* number of citations for each gap distance, while the points show the *mean* number of citations. The bars are the standard errors for each mean. The line of best fit for the data (straight line) has a slope of -1.866 and a p-value << 0.001. Similar results were obtained when only the years 1995-2000 are used. A fairly similar graph is also obtained when the y-axis is logarithmic.

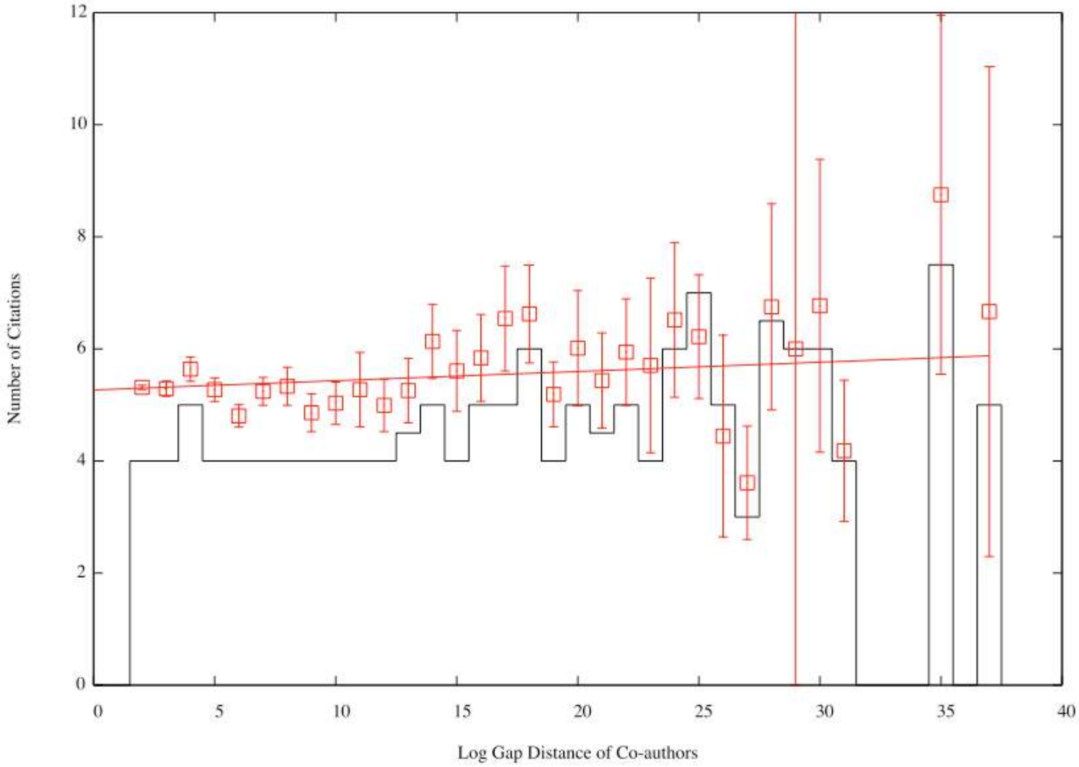

Figure 5.3. Patent Data. The black curve is the *median* number of citations for each gap distance, while the points show the *mean* number of citations. The bars are the standard errors for each mean. The line of best fit for the data (straight line) has a slope of 0.0036 and a p-value of 0.0005. A fairly similar graph is obtained when the y-axis is logarithmic.

As can be seen in Figure 5.3, the patent data contains no clear correlation between gap distance between collaborators and the impact of their patents. On the other hand, the arXiv data of Figure 5.2 has an unexpected pattern: a steady decrease in impact as distance increases, with the possibility of the impact beginning to increase again after some threshold is crossed in distance. Assuming these effects are real, how could they be explained? Well, one story that can be told is that when people are more similar, they can help each other more. But suddenly, at a certain scientific distance, interdisciplinarity kicks in and the impact of the product of the collaboration between two individuals far apart begins to increase.

However, how can we know if these correlations and differences in behavior between the datasets are genuine, or simply spurious? To explore this, we created a series of simple models that represent the process of collaboration between scientists with different scientific distances.

Models of Collaboration

A number of models were constructed with the same basic features (a flow chart depicting a rough overview of the process used in the model is shown in Figure 5.4.). There are *M* individuals, who are problem solvers. The problem solvers are all embedded within a network structure, which specifies the scientific relationship and similarity between individuals. Each individual *i* has z_i problems that require a collaborator in order to be solved.

At each timestep in the model, each individual *i* chooses a collaborator from among the other individuals, who has a probability *c* of being able to help solve the problem. There are two possible methods of choosing a collaborator: a collaborator is chosen at random from the *M* problem solvers (*Random*); or a collaborator is chosen via a network-based search algorithm (*Network*). This algorithm is essentially breadth-first search, where individuals in the network are evaluated as possible collaborators in order of increasing distance, one individual at each timestep. Essentially, the individuals are ordered by distance from *i*, and are evaluated in order, one per timestep. Each problem has a given impact or value, if solved, which is assumed proportional to the number of citations the item later receives, after it's published or patented. These values, v_i , are assumed to be drawn at random from a probability distribution *F*.

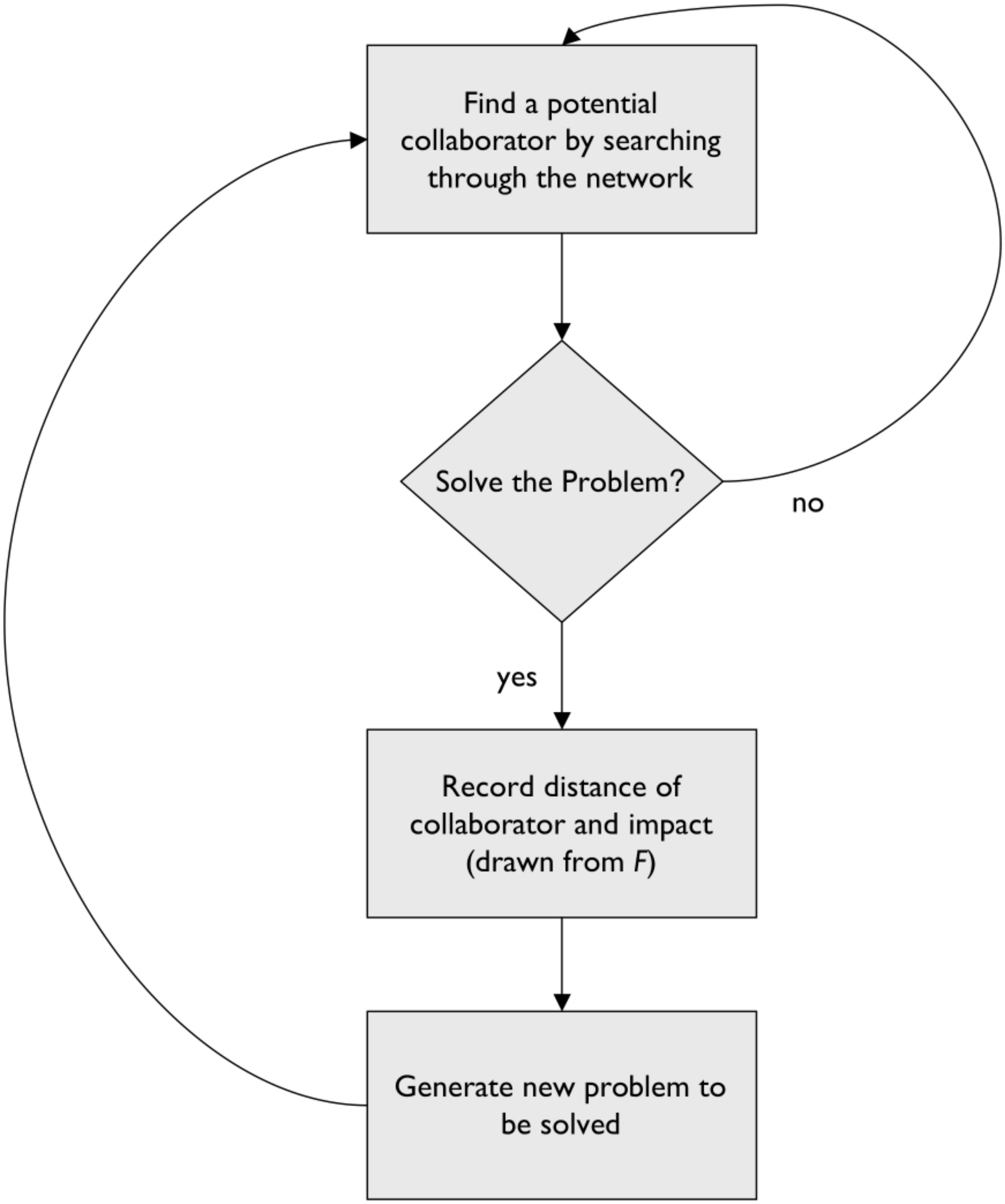
Find a potential collaborator by searching through the network
Solve the Problem?
no
yes
Record distance of collaborator and impact (drawn from *F*)
Generate new problem to be solved

Figure 5.4. A flow chart depicting what occurs at each timestep for each individual *i*, for each of their problems $z_{i.}$until the process is terminated (via a tunable parameter for the maximum number of potential collaborators that can be reached).

In addition, the probability c, relative to distance, of the newly formed team being able to solve the problem is modeled in one of three ways. Most simply, we could consider what happens if c is independent of the potential coauthor and is a constant value (*Uniform*). A second possibility would be that science favors specialization (*Specialization*). In this case, the probability that a collaborator can help solve the research problem decreases as the distance increases between the two potential coauthors. In that case, as a simple first approximation, the probability of finding a solution is taken to decrease linearly:

$$p(d) = \frac{20 - \alpha d}{100}$$

Distance d is measured as the length of the shortest path between the two problem solvers, and α is a parameter that can be allowed to vary, to change the strength of the function.

The third possibility would be that science favors diversity (*Diversity*). In this case, the probability that a collaborator can help solve the research problem increases as the distance increases between the two potential coauthors. Now the probability of finding a solution is assumed to increase linearly with distance:

$$p(d) = \frac{10 + \beta d}{100}$$

There are then six possible models, where one of the two 'collaborator choice' methods is used, and one of the three 'probability of success' models is used.

The structure of the social network for potential collaborators used in this model can take a variety of shapes and types. We used a similar network structure to what was seen in Chapter 4, where individuals are arranged within a hierarchical tree structure. Distance is the height of the nearest common internal node. It was felt that this would be a reasonable model as it has been used in previous research [68], and sufficiently simple for modeling.

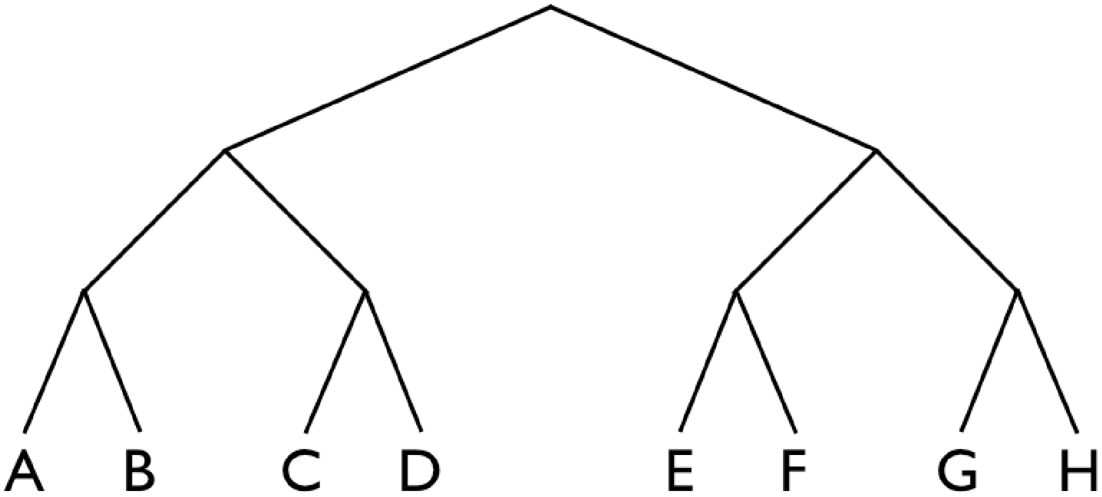

Figure 5.5. The social structure used within the coauthorship models. The distance is the distance to the nearest common internal node. For example, A and B are at a distance of 1, while A and E are at a distance of 3.

In order to determine what an appropriate distribution for the number of citations *F* is, we need to estimate its shape from the data. A number of functions were examined as possible fits, as shown in Table 5.1. Both datasets are well approximated by a log-normal distribution. In addition, both datasets have right tails approximated by power law distributions: the arXiv data follows a power law distribution with a slope of about -1.5, and the patent data follows a power law with a slope of about -4.3, as seen in Figures 5.6 and 5.7.

Table 5.1. Coefficients of determination (R^2) for a number of functions to the citation distribution data.

	Power -Law	Log-Normal	Power Law for Tail (9 citations and above)	Log-Normal for Tail (9 citations and above)	Log-Normal for cumulative distribution
arXiv Data	0.945	0.998	0.988	0.998	0.998
Patent Data	0.656	0.985	0.975	0.976	0.996

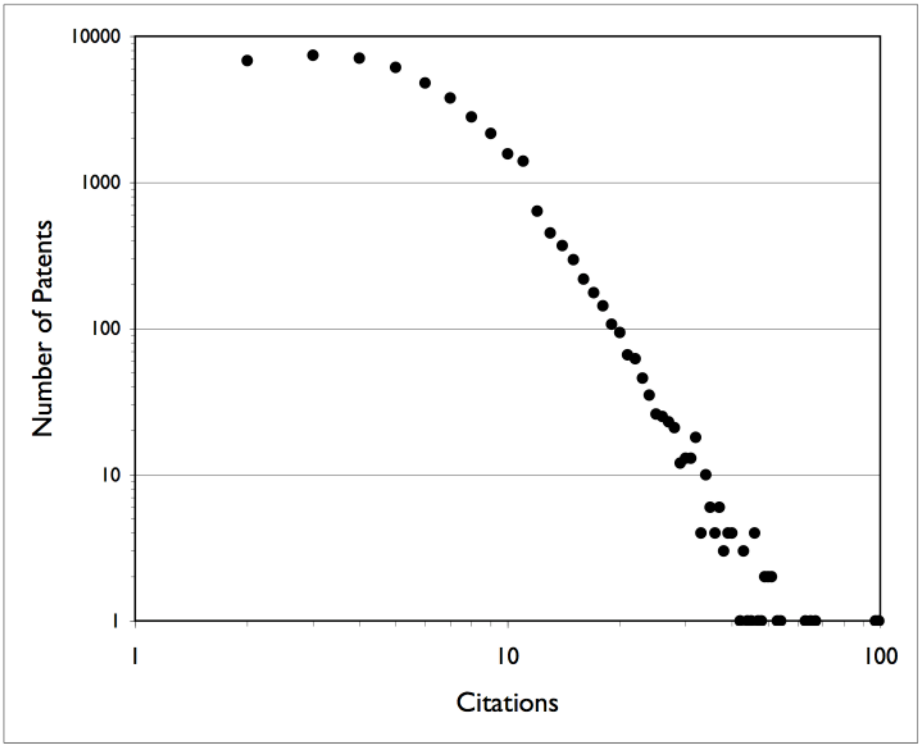

Figure 5.6. The distribution of the number of citations for the patent data. The right tail follows a power law distribution with a slope of -4.31 and an R^2 = 0.975.

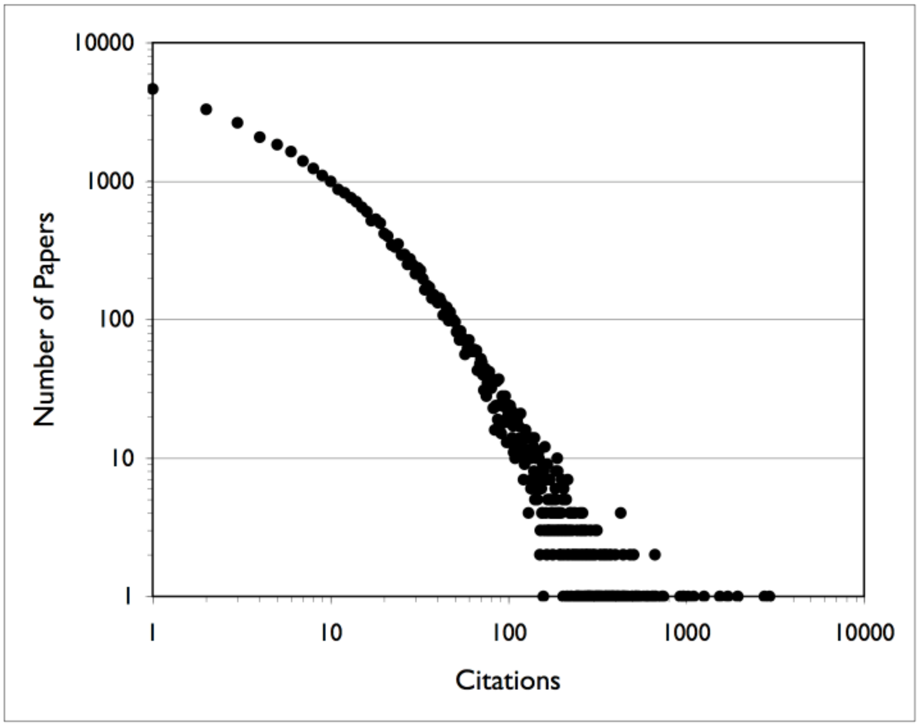

Figure 5.7. The distribution of the number of citations for the arXiv data. The right tail follows a power law distribution with a slope of -1.51 and an R^2 = 0.988.

The cumulative distribution of the number of citations has been found elsewhere to conform to a log-normal distribution, and for smaller datasets to appear similar to a power law distribution, so these results here are consistent [54]. Note that both distributions are heavy-tailed. For simplicity, a power law was used within the simulations. A power law for impact that was independent of distance between collaborators was used for the results below, but we also examined a function where the slope of the power law varies with distance between collaborators (the exponent of the curve becomes more shallow as distance increases). The implications of these different functions will be discussed.

Model Results

Table 5.2. Collaboration model results. One standard deviation above and below the mean of the slopes of the linear regressions for 1000 iterations of each of the six models.

	Uniform	*Diversity*	*Specialization*
Random	(-266.5,274.5)	(-240.8,243.2)	(-37.6,38.6)
Network	(-307.1,317.5)	(-263.2,266.0)	(-48.6,49.4)

Since the impact of collaboration in the model can vary linearly with distance, as a very simple initial expectation we attempted to see if these linear functions were reflected in the data. To do this, we attempted to fit the results of each simulation—the collaboration distances versus impact—to a linear best fit. The results for simulations for each of the six models are above, in Table 5.2. As can be seen above, these are not linear functions, since the linear regressions are terrible (the R^2 values are generally 0.001 or smaller).

For each model, 128 individuals are in each network, with 20 problems being solved by each individual. At each timestep then, all 2,560 problems are attempted to be solved through collaboration. If successful, the distance between the collaborators and the impact of the solution is recorded, as drawn from distribution F. Each simulation ran for 100 timesteps, with a problem solving cutoff of 100 attempts per problem. For each simulation, from 20,000 to more than 100,000 points (i.e., solved problems), depending on the model, were used for the linear regressions. The slope of the line of best fit for these points was recorded, and to get a handle on the range of the slope, each simulation was repeated 1,000 times.

An example run under the same conditions, with line of best fit, is shown in Figure 5.8. As is clear here as well, the results being generated are not linear. We discuss the reason for this shape in the next section.

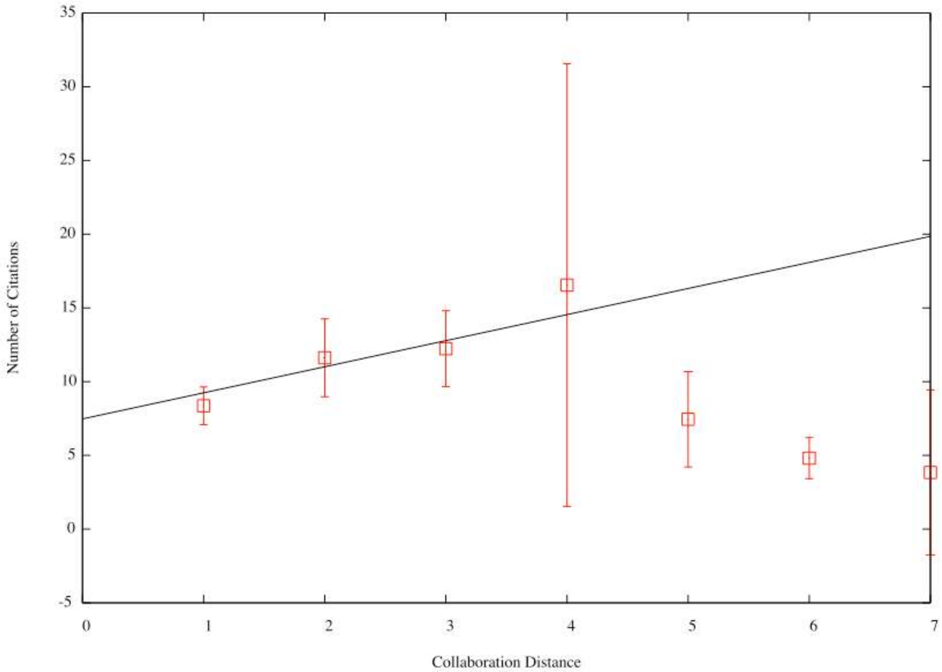

Figure 5.8. Sample run of the collaboration model, with the parameters given above. The points are the mean impact at each distance, and the line is a linear fit. The bars are the standard errors for each mean. The network search algorithm and a function that favors diversity were both used. In this run, the line of best has a positive slope, even though the mean value of citations is beginning to decrease with distance, after an initial increase.

Discussion and Mathematical Model

As can be seen in Table 5.2, there is no clear pattern to be discerned from the simulated results. The best fit lines can either be negative or positive, not giving a good sense of the relationship between distance and impact. This is qualitatively similar to the results derived from the empirical data. And in fact, sometimes it can even appear that the impact of collaboration decreases with distance even if that is not the case.

In this model we have assumed that impact is independent of distance, as specified by the function *F*. Therefore, it is not unexpected that there should be little correlation with distance. This is bolstered by the combination of two other factors: the heavy-tailed distribution of impact (corresponding to the number of eventual citations), and the likelihood of interacting with someone who is nearby for collaboration. Since the likelihood of interacting with someone farther away decreases with distance, individuals collaborate more frequently with closer individuals. Therefore, there are many more collaborations of short distance than those of long distance.

Since the impact distribution follows a distribution with heavy tails, the tail of the distribution (that is, the large values corresponding to the eventual number of citations) is only sampled when there is sufficient sampling from the distribution. Since there are more close collaborations, the mean number of citations at close distances will therefore be larger.

Since this model does not seem to yield much more interesting results than what was put into it, we modified the function *F*, which specifies the impact of the solved problems. As alluded to earlier in the chapter, we also tested a model where the slope of the power law became more shallow as the distance between collaborators of solved problems increased. This means that the more diverse the team is, the more likely the impact is to be higher (the mean of the function increases, since the likelihood of sampling the tail increases). The hope was that there might be some conditions under which the impact function increases with distance, but combined with the decreased number of problems solved by teams that are socially distant, it would appear that the impact decreases with distance (or is uncorrelated), like what was seen in the data.

However, after attempting some simulations with this modified model, we received results where the correlation between distance and impact was always positive, regardless of the six models used. Once again, we received little more than what we inputted into the model. At this stage, the results are therefore very exploratory and tentative, with additional work needed to provide a model that is less circular and more satisfying in its results.

Nonetheless, it turns out that the model with our first impact function (where the impact function is independent of distance) can in fact be solved, and that the mean number of citations will decrease with distance, in the limit of large samples. Let’s begin with the following definition of terms:

n_d = number of individuals at distance *d*

p_d = probability of a team whose members are of distance *d* solve a problem

q_d = probability of a problem being solved at distance *d*

a_d = probability of arriving at distance *d*

We want to find a formula for q_d, and for a network-based approach we have the following equations, with explanations:

$$q_d \mid a_d = 1 - (1 - p_d)^{n_d}$$

The probability of a problem being solved at distance *d*, conditioned on getting to that distance, is 1 minus the probability of no individual at distance *d* solving it.

$$a_d = [1-(q_{d-1}|a_{d-1})] \cdot a_{d-1}$$

The probability of arriving at distance *d* is the probability you arrived at distance *d*–1 multiplied by the probability that the problem wasn't solved at distance *d*–1.

$$q_d = a_d \cdot (q_d | a_d)$$

The probability of solving a problem at distance *d* is equal to the probability that you solve a problem given that you get to that distance, multiplied by the probability you arrive at distance *d*.

Putting all of this together, we get the following recursion:

$$q_d = \left((1-p_d)^{n_d}\right) \cdot \left(1-(1-p_{d-1})^{n_{d-1}}\right) \cdot a_{d-1} \text{ and } a_1 = 1$$

This simplifies to the following equation:

$$q_d = \left(1-(1-p_d)^{n_d}\right) \cdot \prod_{i=1}^{d-1} (1-p_i)^{n_i}$$

The idea is that q_d is proportional to the number of solved problems at distance *d*. In addition, the larger the sample size being drawn from a heavy-tailed power law distribution, the larger the mean, if the mean is infinite. Even if the mean is finite (such as for many heavy-tailed distributions), as long as the variance is large, the sample mean takes a long time to converge to the mean. Therefore, the values of q_d should give us a sense of how citation impact varies with the social distance between collaborators.

This function can be plotted, and the shape of p_d can be modified to be either increasing, decreasing, or constant. It turns out that for all of these (at least linear changes), given the hierarchical social structure of the network, we get the following general shape (the black line is a linear fit), as seen in Figure 5.9:

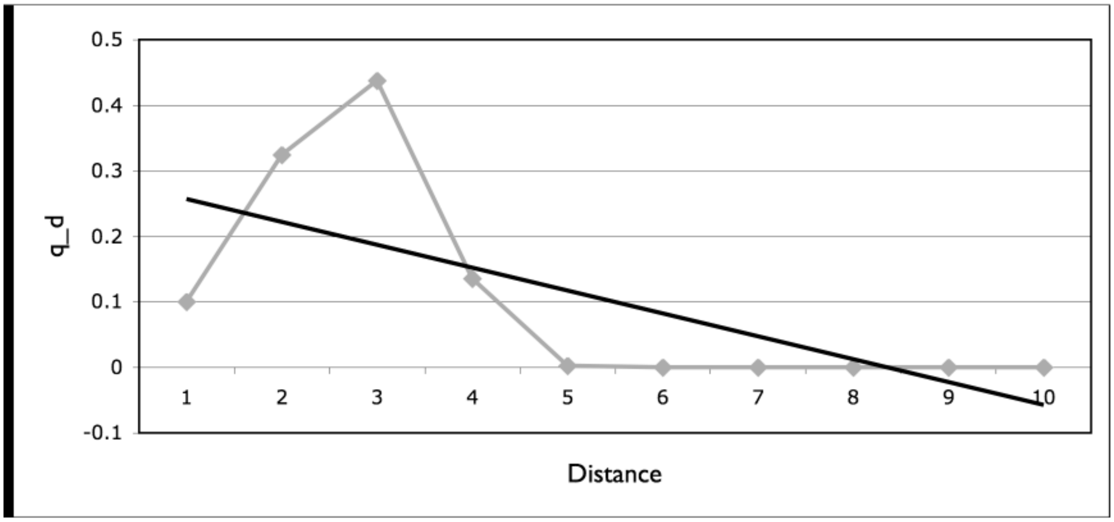

Figure 5.9. q_d versus distance (grey curve). Here the underlying probability p_d of a collaborator at distance *d* solving a problem, is assumed to increase with distance, consistent with a function that favors diversity. Even so, the general trend of the calculated function q_d is downward. The line of best fit to q_d shown in black, confirms this.

For comparison, Figure 5.10, plots data from a run of the model (suitably scaled) of the numbers of solved problems at each distance *d*, for when there is no scholarly benefit to distance. The fit is quite good:

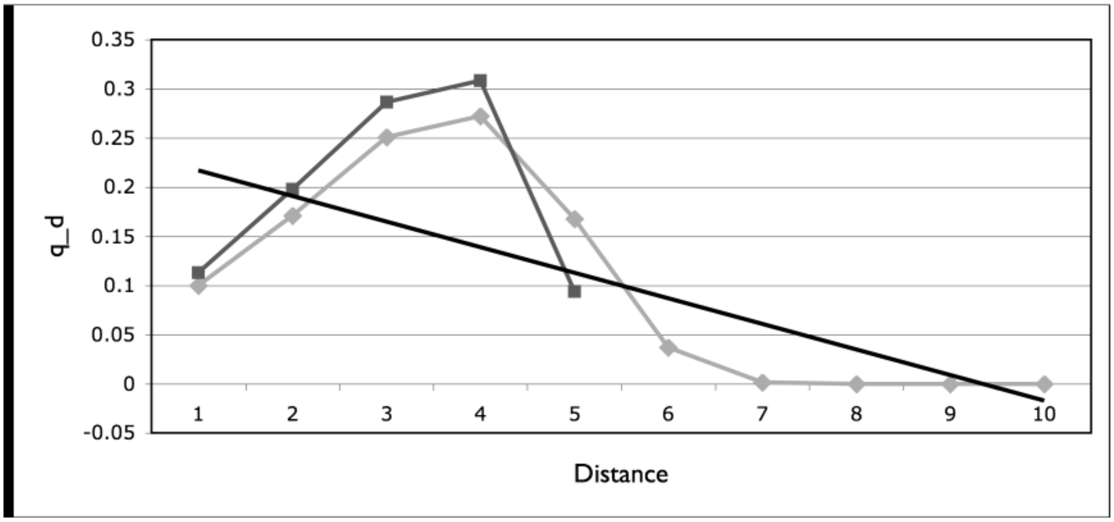

Figure 5.10. q_d versus distance. Here p_d is assumed constant with distance. Grey curve, model prediction for q_d; black curve, simulation results; black line, best linear fit.

These results confirm what we suspected. More problems are solved earlier in the search (at smaller *d*). In terms of proving that q_d eventually decreases with *d*, we want to show that, for large values of *d*, $q_d < q_{d-1}$, which, when simplified, is equivalent to showing that

$$\left(1-\left(1-p_d\right)^{n_d}\right)\cdot\left(1-p_{d-1}\right)^{n_{d-1}} < \left(1-\left(1-p_{d-1}\right)^{n_{d-1}}\right).$$

If *d* becomes large, and we are using our branching tree structure, then n_d becomes large. Let $s_d = (1-p_d)^{n_d}$. Then the desired inequality reduces to

$$(1-s_d)\cdot s_{d-1} < 1-s_{d-1}.$$

But since s_d → 0 as *d* and n_d → ∞, this inequality holds for *d* sufficiently large.

Conclusions

What do our results suggest for understanding collaboration and innovation? As discussed already, our model is quite preliminary and requires a great deal of additional work. It is very difficult to determine what the relationship between distance and impact is from our model, but this is primarily due to the impact function being independent of distance. And thus far we have been unable to create a model that yields a consistently declining or uncorrelated relationship between impact and distance, even if the function *F* has an increasing relationship between impact and distance. Further work on our model is required to explain the patterns found in our datasets.

In addition, these models, whatever form they will eventually take, are not the final words on interdisciplinarity. Interdisciplinary work often provides insights that are valuable to more than one area, and could not have been reached by working within a single discipline (as we have been looking at with the arXiv data, for example). Furthermore, whole new areas of science would not exist without the presence of people working between disciplines (such as computational biology). Simply counting the number of citations is only an approximation of understanding impact, and more complex models are necessary. Interdisciplinarity appears to be very important within science, but the results here imply that it might be much harder to quantify that we had thought previously.

CHAPTER 6

A MONTE CARLO APPROACH TO JOE DIMAGGIO AND STREAKS IN BASEBALL[5]

The incredible record of Joe DiMaggio in the summer of 1941 is unparalleled. No one has come close—before or since—to equaling his streak of hitting safely in 56 games in a row. People have gone even further and stated that it is the only record in baseball (or perhaps even in all of sports) that never should have happened, statistically speaking: while other records can be explained by expected outliers over the long and varied history of professional baseball (nearly 150 years), DiMaggio's record stands alone [26].

In addition, streaks are of more general interest than the parochial domain of baseball. For example, Bill Miller's Legg Mason Value Trust mutual fund succeeded in beating the S&P 500 for fifteen years in a row [35]. Streaks clearly have a certain correspondence in relation to skill, and teasing out the difference between luck and skill is important. Therefore, a better understanding of streaks is a useful exercise.

We wondered whether Joe DiMaggio's record was more likely than people might think. To study this, we constructed a series of simple Monte Carlo simulations, using a comprehensive baseball statistics database from 1871 to 2004 [34]. The first and simplest model is as follows:

Each player, for each season that he played, was characterized by a few numbers: the number of games played, number of plate appearances, and number of hits in the season. Number of plate appearances is the sum of at-bats, times walked,

[5] This work was completed in conjunction with Steven Strogatz (Cornell University), some of which has already appeared in the *New York Times* [8].

being hit by a pitch, sacrifice hits, and any other way in which a player appears at the plate and either has a chance for a hit, or is denied that chance (such as being walked or being hit by a pitch). Our data contained nearly all of the categories for plate appearances, so the values used are as similar as possible to the actual number of plate appearances in an entire season. Then, we calculated each player's probability of getting at least one hit in any game in which he played for that year as follows:

M = number of games

n = number of plate appearances per game = number of plate appearances/M

p = number of hits/number of plate appearances

The probability of not getting a hit for a single plate appearance is 1-p. Assuming that successive plate appearances within a single game are independent, the probability of not getting a hit for an entire game is then $(1-p)^n$. So, the probability of getting at least one hit in a single game is $1-(1-p)^n = g$. Similar calculations were done in a paper by Michael Freiman [22].

Given g for all players and for all seasons (this was only calculated for players who had at least 300 or 450 plate appearances in the season, as seen below), we can simulate a player's streaks for a season. We randomly generate G games, and count how many successful games in a row there were. This is done for all players and for all seasons, effectively simulating an entire baseball history. We tabulate the long streaks in each history, including the longest streak, when it occurred, and who had it. Of course, the actual probabilities for streaks of various lengths can be calculated exactly for this simple model, but we later modify the model in various ways, so a computational approach is most reasonable [39]. To gain reliable statistics, we performed these simulations 10,000 times.

For this model, a streak of 56 games or longer occurs about 49% of the time. Figure 6.1 shows the histogram of the longest streaks, in each of the 10,000 simulated histories of baseball. In this and all other instances in the chapter, we take an occurrence of more than 5% in our simulations—the conventional number in hypothesis testing—to be indicative that this could happen by chance reasonably often.

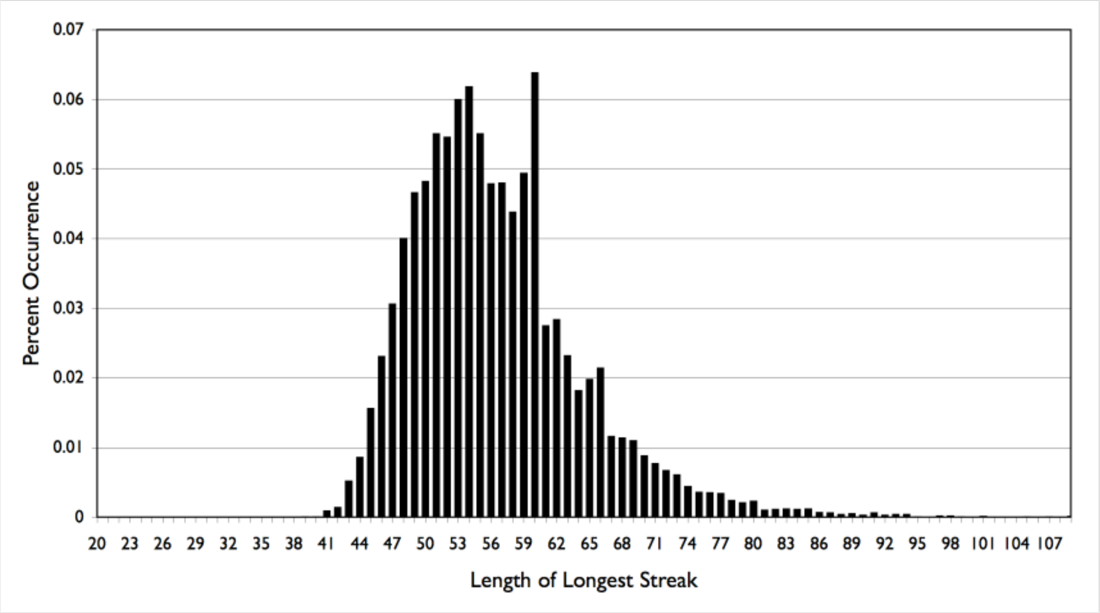

Figure 6.1. Distribution of the longest streaks in 10,000 simulated runs of baseball history. The curve is roughly bell-shaped with a longer tail to the right, favoring longer streaks. The most frequent outcome was a record streak of 60 games, but this is due to a short, high impact season of Ross Barnes in 1873. Adding up all the heights of the columns for streaks 56 games or longer yields about 49% of all the data, meaning that streaks as long as DiMaggio's are expected about 49% of the time.

The most frequent outcome was a record streak of 60 games. This is due to a single player in a single season: Ross Barnes in 1873. Ross Barnes, who had an incredibly high probability of getting at least one hit in any game (0.946) only played 60 games that season. These two facts combined to yield a large number of simulated histories where Barnes hit in every game in the entire season, giving us the spike at 60.

The above model, which I shall term Model A (one with constant probabilities), overestimates certain long streaks, as exemplified by Ross Barnes acting as an outlier. So, Model A was then modified to improve its robustness and its similarity to actual baseball history. We added game-to-game variation to *g,* due to variation in number of plate appearances in games and opposing pitching ability. The true amount of variation due to pitching ability or variable number of plate appearances is more complicated to estimate, especially due to a lack of more detailed game-by-game data for the entire dataset, as discussed later. So, for a rough approach, we ran the simulations with a 10% and 20% uniform variation in *g* for each game. The 10% variation model is Model B, and the 20% variation model is Model C.

Table 6.1. Model names and descriptions.

Model	*Description*
Model A	Constant *g* and a minimum of 300 plate appearances in a season for inclusion
Model B	*g* varied by up to 10% and a minimum of 300 plate appearances in a season for inclusion
Model C	*g* varied by up to 20% and a minimum of 300 plate appearances in a season for inclusion
Model B (450 PA cutoff)	*g* varied by up to 10% and a minimum of 450 plate appearances in a season for inclusion
Model B (post-1905)	*g* varied by up to 10% and a minimum of 300 plate appearances in a season for inclusion, and only data from 1905 onwards was used

To test the validity of all of these models, we compared their predictions against the observed number of shorter streaks in baseball history. The distribution of streaks of length 30 or greater follows what is approximately an exponential distribution [1], as shown in Figures 6.2 and 6.3.

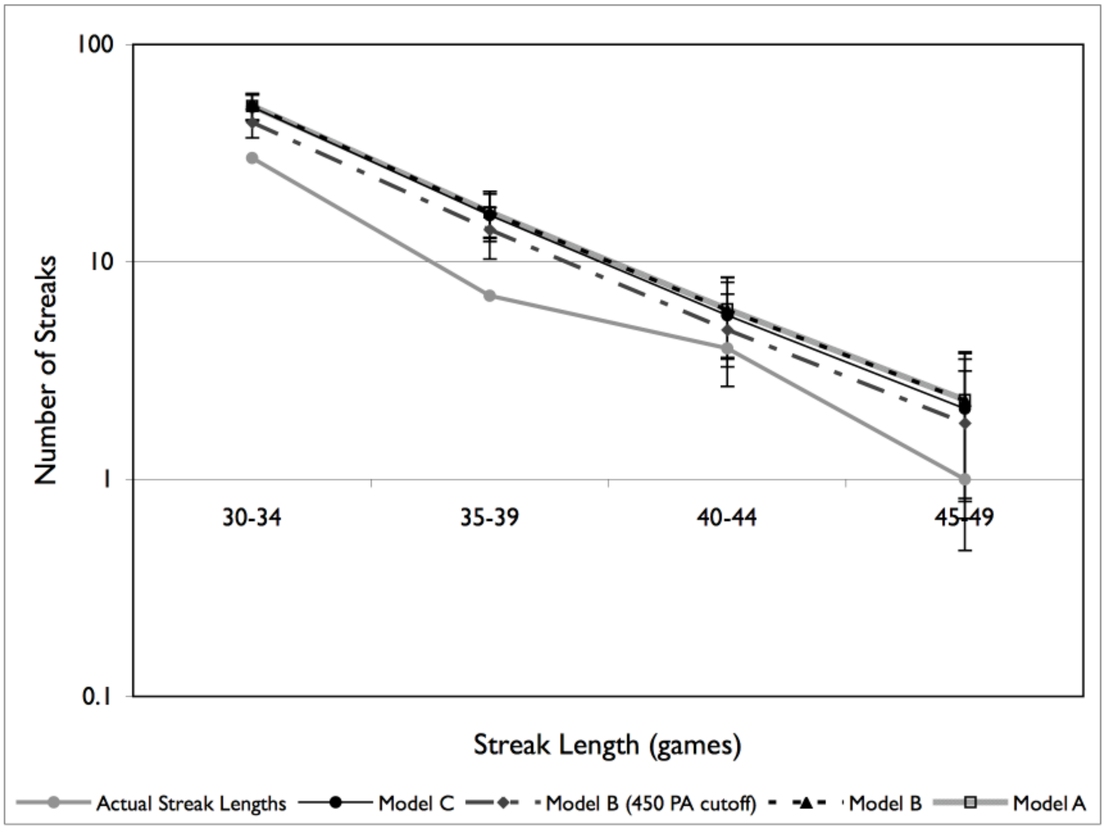

Figure 6.2. The distribution of streaks (aside from DiMaggio's) for all of baseball to the present [1]. The following models are included: Model B, using all available data and a minimum number of 300 plate appearances in a season; Model B, using all available data and a minimum number of 450 plate appearances in a season; and Model C, using all available data and a minimum number of 300 plate appearances in a season. Models A, B, and C all yield extremely similar results and overlap quite a bit in the figure.

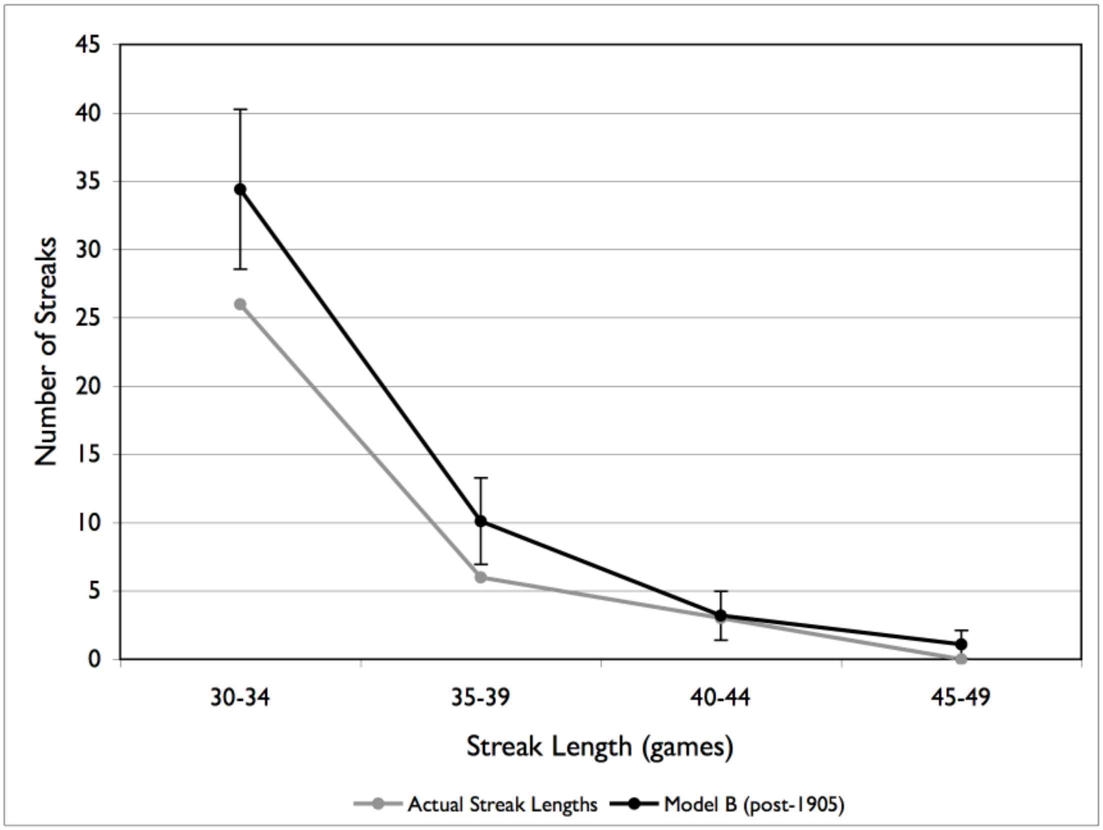

Figure 6.3. The distribution of streaks (aside from DiMaggio's) from 1905 to the present [1]. Model B, using data only from 1905 until 2005, is included for comparison.

It is found that the simulated histories also have exponential distributions for the length of the streaks. The slope varies as variation is added, the minimum number of plate appearances to be included in the simulation is increased, or the number of years being examined is limited. However, it can be seen that the results are similar to what is found in 'real' baseball, and that the functional form is preserved by these simple models.

Specifically, all the Models are within a single standard deviation for the number of actual streaks of lengths 40-44 and 45-49. For the shorter streaks, the models overestimate, by about 50%, yielding same order-of-magnitude results. These might be improved by refining the model, as discussed later.

Furthermore, adding variation turns out to change very little of the results, and actually increases the correspondence between the streak length data and the simulations, without changing the probability of outlier streaks, such as DiMaggio's, appreciably.

The true amount of variation due to pitching ability or variable number of plate appearances is not known, however it could be estimated. One method to examine both of these is to use box score data, which is available in an electronic format for the past fifty years or so [2]. By examining the variation in at-bats, we could get a better estimate of plate appearance variation. In addition, by examining the variation in hitting ability against different pitchers, we could estimate the range for hitting ability for an average player. In this case though, the simulations were run with a 10% and 20% uniform variation in *g* for each game, as an approximation. As seen in Table 6.2, the probability of a DiMaggio streak was still non-trivial.

In addition, since, as can be seen from Figure 6.1 and 6.4, there are certain spikes due to outliers in the early days of the game, a simulation using data only for 1905 and later was conducted. This is considered to be the more modern era of baseball. Using this more limited data sample, the spike of 60 game streaks due to Ross Barnes in the early 1870's is eliminated and the curve becomes smoother (Figure 6.5). Streaks become less likely, but DiMaggio-like streaks (56 games or longer) still occur nearly 20% of the time. Please see Table 6.2 and Figures 6.4-6.7 for all results.

Table 6.2. Probabilities of DiMaggio-like Streaks in a variety of models. All demonstrate a non-trivial likelihood of such an extreme streak.

	Model A	*Model B*	*Model B (post-1905)*	*Model C*	*Model B (450 PA min.)*
Probability of DiMaggio-Like Streak	0.49	0.49	0.18	0.39	0.38

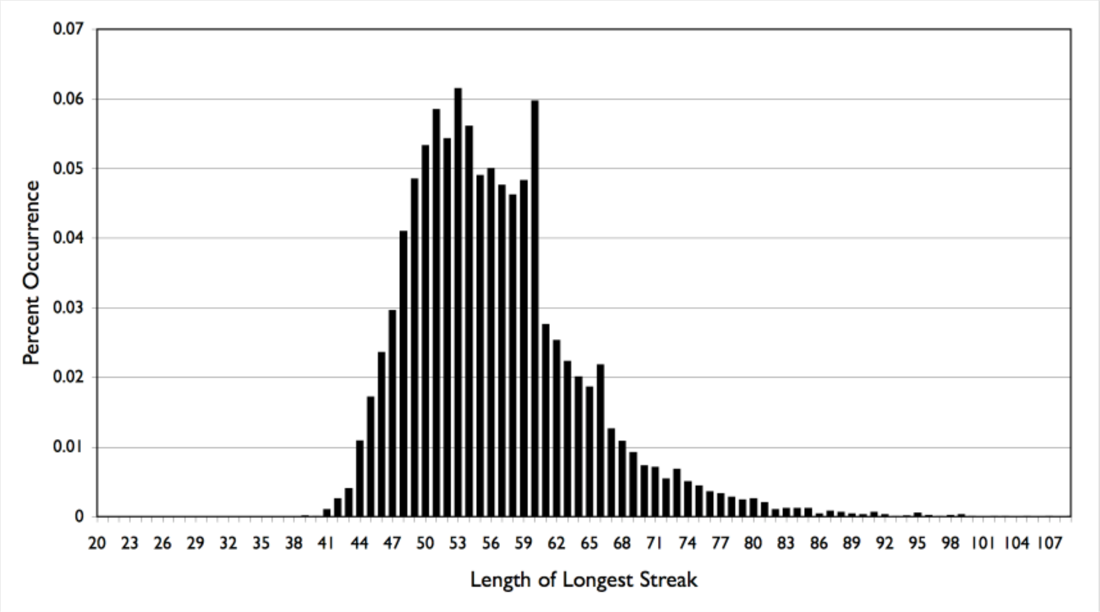

Figure 6.4. Streak length histogram for Model B.

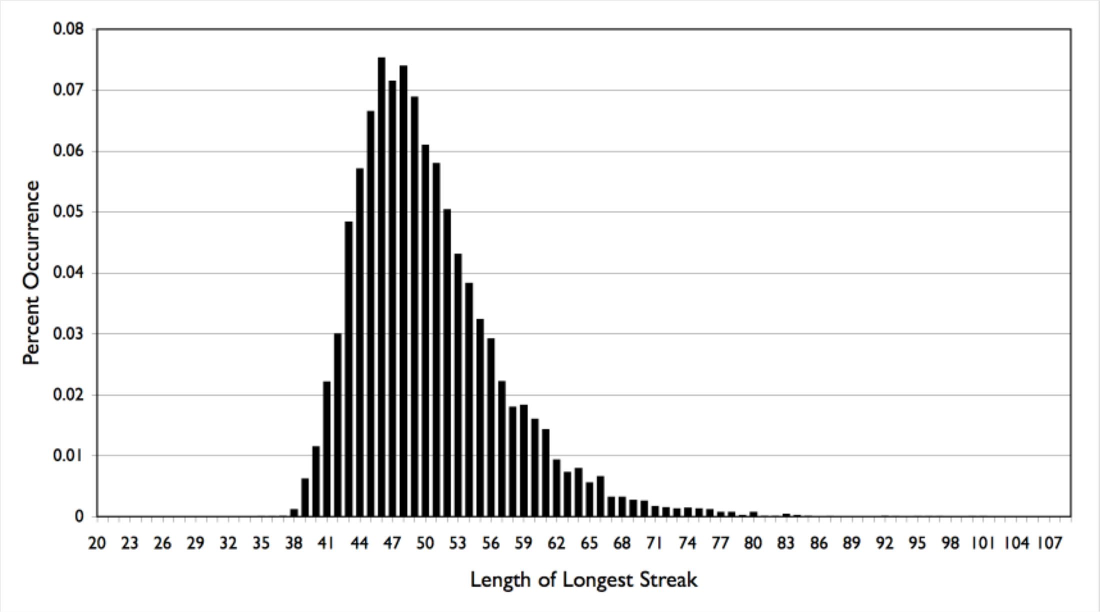

Figure 6.5. Streak length histogram for Model B (1905 or later).

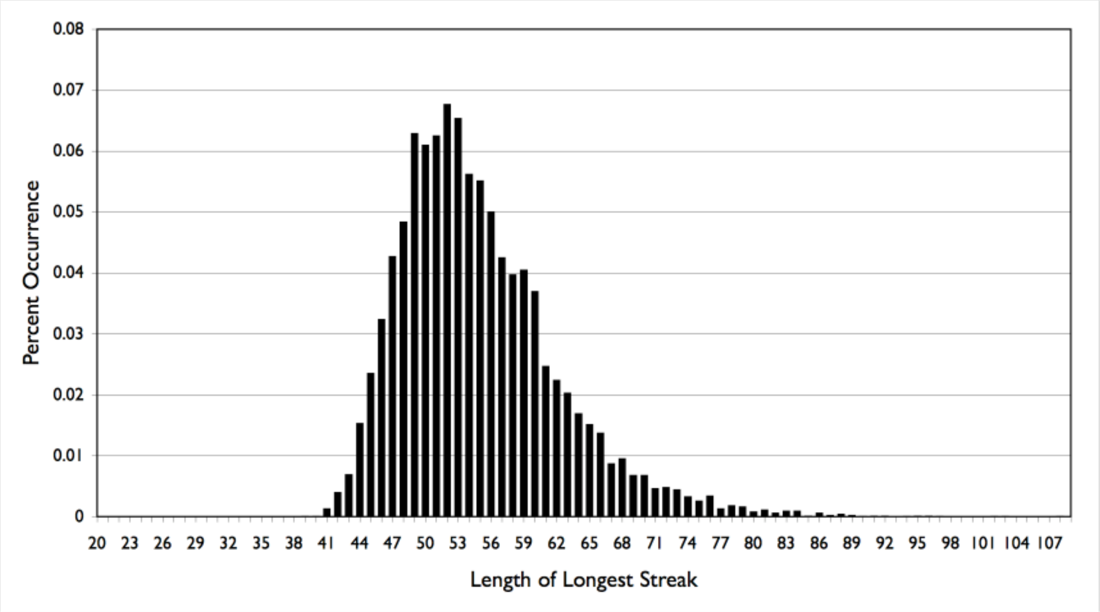

Figure 6.6. Streak length histogram for Model C.

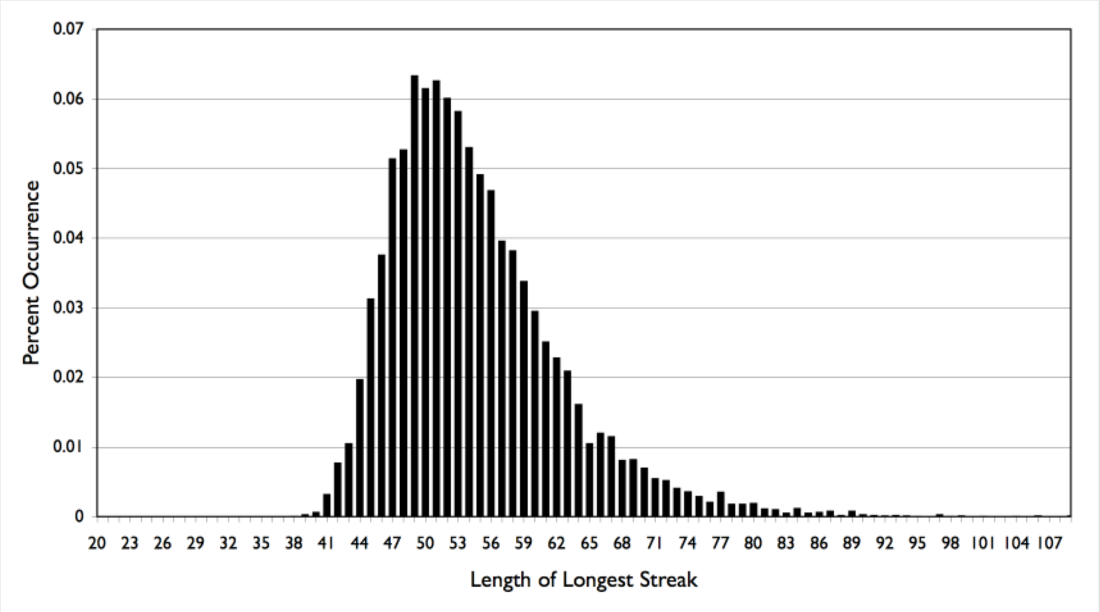

Figure 6.7. Streak length histogram for Model B (450 plate appearance minimum).

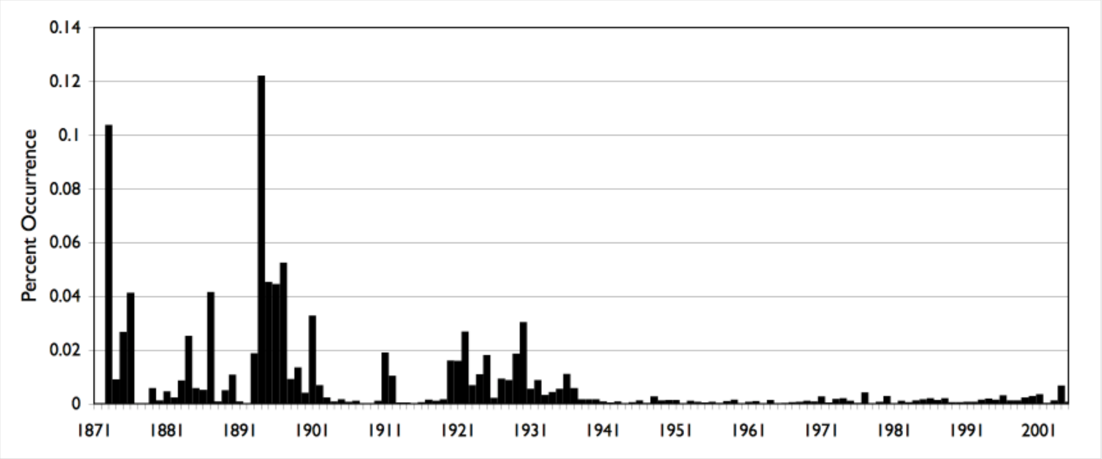

Figure 6.8. Timeline for streak likelihood (Model A). This graph shows when the longest streaks occurred in our simulations. The record was likeliest to have been set in 1894, when Hugh Duffy batted .440 and had a 91 percent chance of hitting successfully in each game. Other likely periods for the record to be set were in the early teens (Ty Cobb's era) and throughout the 20's and 30's. Joe DiMaggio's miracle year of 1941 was a relative dry spell, making his achievement all the more stunning. The other models have relatively similar yearly histograms.

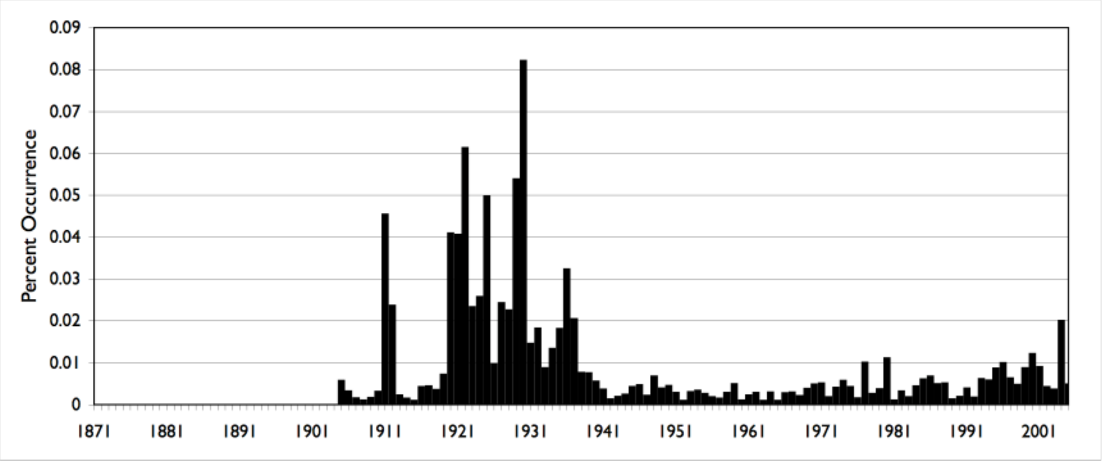

Figure 6.9. Timeline for streak likelihood (Model B, post-1905).

The surprising thing found in the simulation is when DiMaggio’s streak occurred, as seen in Figures 6.8 and 6.9. It should have occurred far earlier in the history of baseball, back in the late 1800’s or early 1900’s. But not in 1941, or similar years. Years with a similar probability of 1941 or worse of holding the record account for only about 5% of the records (Model A).

Table 6.3. The ten players who are likeliest to hold the record for longest streak, for each model. The fraction of the simulations they each hold the streak is included, as *P(Streak)*.

Model A

Player	*P(Streak)*
Ross Barnes	0.1035
Willie Keeler	0.0646
Hugh Duffy	0.048
Jesse Burkett	0.0465
Sam Thompson	0.0368
Ed Delahanty	0.0335
Nap Lajoie	0.0333
George Sisler	0.0291
Ty Cobb	0.0286
George Wright	0.0284

Model B (post-1905)

Player	*P(Streak)*
George Sisler	0.0723
Ty Cobb	0.0685
Rogers Hornsby	0.0395
Al Simmons	0.0386
Bill Terry	0.0288
Paul Waner	0.0217
Ichiro Suzuki	0.0217
Chuck Klein	0.0217
Joe Jackson	0.021
Harry Heilmann	0.0197

Model B

Player	*P(Streak)*
Ross Barnes	0.0905
Willie Keeler	0.0651
Hugh Duffy	0.0488
Jesse Burkett	0.0448
Sam Thompson	0.0411
Ed Delahanty	0.0363
George Sisler	0.0311
Nap Lajoie	0.0305
Ty Cobb	0.0295
George Wright	0.0288

Model B (450 PA Cutoff)

Player	*P(Streak)*
Willie Keeler	0.0806
Jesse Burkett	0.0567
Hugh Duffy	0.0566
Ed Delahanty	0.0494
Nap Lajoie	0.0389
Tip O'Neill	0.0388
George Sisler	0.0382
Ty Cobb	0.0357
Sam Thompson	0.0338
Al Simmons	0.0219

Model C

Player	*P(Streak)*
Willie Keeler	0.0512
Ross Barnes	0.0458
Jesse Burkett	0.0431
Sam Thompson	0.037
Hugh Duffy	0.0343
George Sisler	0.0339
Ed Delahanty	0.0338
Nap Lajoie	0.0308
Ty Cobb	0.0296
Al Simmons	0.0207

On the other hand, while DiMaggio is not the likeliest person to hold the record (he is over forty-seventh most likely in Model A), there is not a single likely player. As seen in Table 6.3, the top three players—Ross Barnes, Willie Keeler, and Hugh Duffy—only account for about 21.6% of the record-holding streaks in our simulations, even in Model A, which overestimates the likelihood (especially for Ross Barnes). And players of DiMaggio's caliber or less account for nearly 25% of the streaks.

So, while no single player is especially likely to hold the record, it is likely that an extreme streak would have occurred. This subtle probabilistic concept has been discussed by Diaconis and Mosteller within the context of calculating the probability of a double-lottery winner [17]. While the probability of a certain individual winning the lottery twice is extremely small, the fact that someone somewhere will do this (since there are many people who buy lottery tickets on a regular basis) is virtually assured. Diaconis and Mosteller like to think about this in terms of their law of truly large numbers: "With a large enough sample, any outrageous thing is likely to happen."

Richard Feynman once made a similar point. He walked into a lecture hall and noted that he just saw the most amazing thing: a car with the license plate ANZ 912; what are the odds of that? [21]. In our model, while there are some players more likely to hold the streak record, there is a 25% chance that the 'unlikely' players (DiMaggio's likelihood or below) are the ones with the record.

Lastly, we checked the distribution of the difference in length between the longest streak in a simulation, and the second-longest streak. Stephen Jay Gould, in his article, 'The Streak of Streaks', notes that 'DiMaggio's fifty-six–game hitting streak is ridiculously and almost unreachably far from all challengers (Wee Willie Keeler and Peter Rose, both with forty-four, come second)' [26]. Is it true that DiMaggio's streak is much farther away from the other streaks than we might expect, or is this a red herring?

Using Models A and B (post-1905) as representative of the models, we checked to see if Gould’s thinking was correct. And it turns out that in Model A, nearly 20% of the simulations had a difference between the longest and second-longest streak of 12 games or more. A histogram of the differences is shown in Figure 6.10. For Model B (post-1905), the number was 13% of the simulations for a difference of 12 games or more. So, DiMaggio as an apparent outlier is not actually true, and this model is able to replicate the observed large differences in streak length.

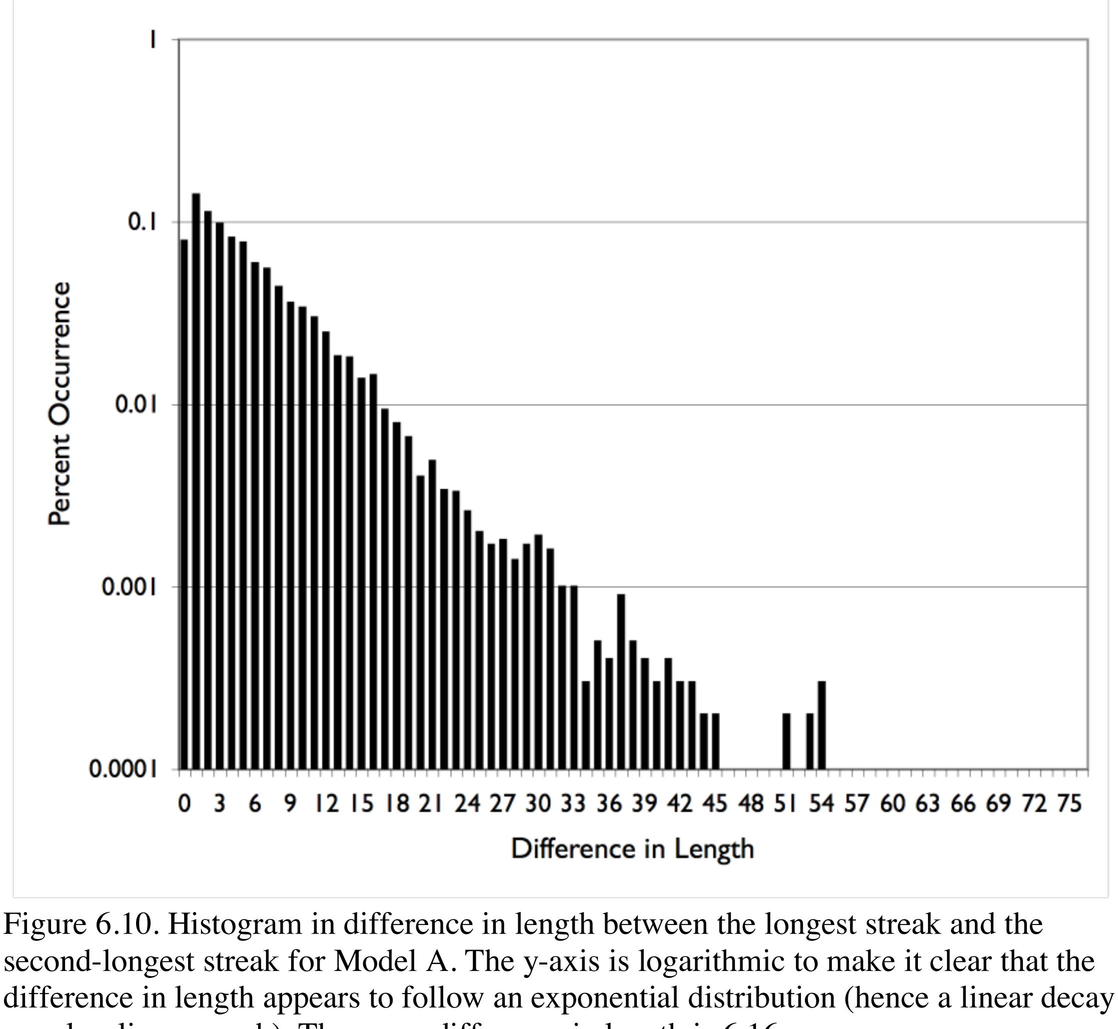

Figure 6.10. Histogram in difference in length between the longest streak and the second-longest streak for Model A. The y-axis is logarithmic to make it clear that the difference in length appears to follow an exponential distribution (hence a linear decay on a log-linear graph). The mean difference in length is 6.16.

While an effective model and one that provides surprising results, this is of course still an extremely crude model for baseball. Adding the factor of opposing team abilities would presumably change the results, though it is unclear how. There is a model that has recently been developed which looks at streaks of wins between opposing teams in baseball, and this would be a reasonable model to adapt for pairs of pitchers and batters [60].

Furthermore, examination of the psychological effects of a streak has to be taken into account. For example, Roger Maris's hair fell out in large chunks while chasing Babe Ruth's home run record [7]. Stress can certainly affect you.

This is related to the issue of the independence assumption, that is, whether or not a hit in a game is independent of the games before it. Trent McCotter has attempted to examine this, by taking the time series of games and asking whether or not there was a hit in the game [41]. This yields a series of successes and failures. To test the independence assumption, he shuffled the order of the successes to see if the randomized streaks are longer or shorter than expected. He found that the randomized streaks are somewhat shorter than expected, arguing against independence in the real data and hinting at a psychological component or some other factor. Intriguingly, this is different than what has been found in relation to the Hot Hand phenomenon in basketball [23], or even an analysis of successful at-bats [5], which in both cases yielded streaks indistinguishable from random chance.

However, if this model is a moderately realistic one, and it seems to be, then it provides a statistically informed check to our collective baseball intuition. Joltin' Joe's record, while certainly incredible, is in fact not that unlikely within the long history of baseball. But that he did it is certainly still achievement indeed.

REFERENCES

[1] Baseball Almanac. http://www.baseball-almanac.com/feats/feats-streak.shtml. Accessed April 28, 2008.

[2] Baseball Reference. http://www.baseball-reference.com/. 2000-2008.

[3] New science building will foster interdisciplinary efforts. http://www.expressnews.ualberta.ca/article.cfm?id=7433. Accessed May 18, 2008.

[4] R. Albert and A. L. Barabási, *Statistical mechanics of complex networks*, Reviews of Modern Physics, 74 (2002), pp. 47-97.

[5] S. C. Albright, *A Statistical Analysis of Hitting Streaks in Baseball*, Journal of the American Statistical Association, 88 (1993), pp. 1175-1183.

[6] L. A. N. Amaral, A. Scala, M. Barthélémy and H. E. Stanley, *Classes of small-world networks*, Proceedings of the National Academy of Sciences of the United States of America, 97 (2000), pp. 11149-11152.

[7] R. Angell, *Green*, The New Yorker. April 7, 2008. http://www.newyorker.com/talk/comment/2008/04/07/080407taco_talk_angell.

[8] S. Arbesman and S. Strogatz, *A Journey To Baseball's Alternate Universe*, New York Times. March 30, 2008.

[9] R. H. Baayen, R. Piepenbrock and L. Gulikers, CELEX2. 1996.

[10] A. L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509-512.

[11] L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert and G. B. West, *Growth, innovation, scaling, and the pace of life in cities*, Proceedings of the National Academy of Sciences of the United States of America, 104 (2007), pp. 7301-7306.

[12] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman and S. H. Strogatz, *Are randomly grown graphs really random?*, Physical Review E, 64 (2001).

[13] A. Clauset, C. Moore and M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453 (2008), pp. 98-101.

[14] A. M. Collins and E. F. Loftus, *A spreading activation theory of semantic memory*, Psychological Review, 82 (1975), pp. 407-428.

[15] M. Collins, *Born Again*, Boston Globe. August 7, 2005.

[16] A. Culter, J. Mehler, D. Norris and J. Segui, *The syllable's differing role in segmentation of French and English*, Journal of Memory and Language, 25 (1986), pp. 385-400.

[17] P. Diaconis and F. Mosteller, *Methods for Studying Coincidences*, Journal of the American Statistical Association, 84 (1989), pp. 853-861.

[18] L. E. Douglas, Interdisciplinary Grants in the Mathematical Sciences. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5299&org=DMS. Accessed May 15, 2008.

[19] S. R. Eddy, *"Antedisciplinary" Science*, PLoS Computational Biology, 1 (2005), pp. e6.

[20] R. Ferrer-i-Cancho and V. S. Ricard, *The small world of human language*, Proceedings of the Royal Society B: Biological Sciences, 268 (2001), pp. 2261-2265.

[21] R. P. Feynman, *The Meaning of It All: Thoughts of a Citizen-Scientist*, 1999. pp. 81.

[22] M. Freiman, *56-Game Hitting Streaks Revisited*, The Baseball Research Journal, 31 (2003), pp. 11-15.

[23] T. Gilovich, R. Vallone and A. Tversky, *The hot hand in basketball: On the misperception of random sequences*, Cognitive Psychology, 17 (1985), pp. 295-314.

[24] P. Ginsparg, arXiv.org. 2007. Accessed 2007.

[25] E. L. Glaeser, *Can Buffalo Ever Come Back?*, City Journal (2007). http://www.city-journal.org/html/17_4_buffalo_ny.html.

[26] S. J. Gould, *The Streak of Streaks*, The New York Review of Books. August 18, 1988. http://www.nybooks.com/articles/4337.

[27] M. Granovetter, *The Strength of Weak Ties*, American Journal of Sociology, 78 (1973), pp. 1360-1380.

[28] B. H. Hall, A. B. Jaffe and M. Trajtenberg, *The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools*, NBER Working Paper 8498 (2001).

[29] L. R. Hoffman, *Homogeneity of Member Personality and Its Effects on Group Problem-Solving*, Journal of Abnormal and Social Psychology, 58 (1959), pp. 27-32.

[30] S. Huang, X. Bian, G. Wu and C. McLemore, LDC Mandarin Lexicon.

[31] H. P. Judd, The Hawaiian Language and Hawaiian-English Dictionary: a complete grammar.

[32] C. T. Kello and B. C. Beltz, *Scale-free networks in phonological and orthographic wordform lexicons*, in I. Chitoran, C. Coupé, E. Marsico and F. Pellegrino, eds., *Approaches to Phonological Complexity*, Mouton de Gruyter, in press.

[33] M. Kleiber, *Body size and metabolism*, Hilgardia, 6 (1932), pp. 315-353.

[34] S. Lahman, The Lahman Baseball Database. http://www.baseball1.com. December 28, 2005.

[35] P. R. LaMonica, *Bill Miller's New Streak*, CNNMoney.com. December 4, 2007.

[36] K. Laursen, V. Mahnke and P. Vejrup-Hansen, *Do Differences Make a Difference? The Impact of Human Capital Diversity, Experience and Compensation on Firm Performance in Engineering Consulting*. Unpublished Work.

[37] J. Leskovec, J. Kleinberg and C. Faloutsos, *Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations*, Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (2005).

[38] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang and M. A. Nowak, *Quantifying the evolutionary dynamics of language*, Nature, 449 (2007), pp. 713-716.

[39] A. Looney, *How improbable is DiMaggio's 56 game hitting streak?* Berkeley, 2008. Personal Communication.

[40] P. A. Luce and D. B. Pisoni, *Recognizing spoken words: The neighborhood activation model*, Ear and Hearing, 19 (1998), pp. 1-36.

[41] T. McCotter, *Hitting Streaks Don't Obey Your Rules*, The Baseball Research Journal (in press).

[42] J. Mehler, J. Y. Dommergues, U. Frauenfelder and J. Segui, *The syllable's role in speech segmentation*, Journal of Verbal Learning and Verbal Behavior, 20 (1981), pp. 298-305.

[43] H. Monoi, Y. Fukusako, M. Itoh and S. Sasanuma, *Speech sound errors in patients with conduction and Broca's aphasia*, Brain and Language, 20 (1983), pp. 175-194.

[44] A. E. Motter, A. P. S. de Moura, Y.-C. Lai and P. Dasgupta, *Topology of the conceptual network of language*, Physical Review E, 65 (2002), pp. 065102.

[45] M. E. J. Newman, *Assortative mixing in networks*, Physical Review Letters, 89 (2002), pp. 208701.

[46] M. E. J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics, 46 (2005), pp. 323-351.

[47] M. E. J. Newman, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167-256.

[48] M. E. J. Newman, *The structure of scientific collaboration networks*, Proceedings of the National Academy of Sciences of the United States of America, 98 (2001), pp. 404-409.

[49] T. A. Obaid, *State of World Population 2007: Unleashing the Potential of Urban Growth*, United Nations Population Fund (2007).

[50] S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.*, Princeton University Press, 2007.

[51] M. Perea, M. Urkia, C. J. Davis, A. Agirre, E. Laseka and M. Carreiras, *E-Hitz: A word frequency list and a program for deriving psycholingustic statistics in an agglutinative language (Basque)*, Behavior Research Methods, 38 (2006), pp. 610-615.

[52] S. Pinker, *The Stuff of Thought: Language as a Window into Human Nature*, Viking, 2007.

[53] A. Rapoport, *Mathematical models of social interaction*, in R. D. Lace, R. R. Bush and E. Galanter, eds., *Handbook of Mathematical Psychology*, John Wiley and Sons, 1963, pp. 493-579.

[54] S. Redner, *Citation Statistics from 110 Years of Physical Review*, Physics Today, 58 (2005), pp. 49-54.

[55] T. Schelling, *Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices with Externalities*, The Journal of Conflict Resolution, 17 (1973), pp. 381-428.

[56] T. C. Schelling, *Micromotives and Macrobehavior*, W. W. Norton, 2006.

[57] J. Scobbie, What is "ghoti"? http://alt-usage-english.org/excerpts/fxwhat04.html. Accessed April 29, 2008.

[58] N. Sebastián-Gallés, M. A. Martí-Antonín, M. F. Carreiras-Valin ã and F. Cuetos-Vega, *Lexesp. Léxico informatizado del espanõl.*, Edicions de la Universitat de Barcelona (2000).

[59] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters, and Complexity*, Clarendon Press, Oxford, 2006. pp. 293-294.

[60] C. Sire and S. Redner, *Understanding Baseball Team Standings and Streaks*. http://aps.arxiv.org/abs/0804.1110. Unpublished Work.

[61] L. H. Spencer and J. R. Hanley, *Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales*, British Journal of Psychology, 94 (2003), pp. 1-28.

[62] M. Steyvers and J. Tenenbaum, *The Large Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth*, Cognitive Science, 29 (2005), pp. 41-78.

[63] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268-276.

[64] M. S. Vitevitch, *What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval?*, Journal of Speech, Language, and Hearing Research, 51 (2008), pp. 408-422.

[65] M. S. Vitevitch and E. Rodríguez, *Neighborhood density effects in spoken word recognition in Spanish*, Journal of Multilingual Communication Disorders, 3 (2005), pp. 64-73.

[66] M. S. Vitevitch, M. K. Stamer and J. A. Sereno, *Word length and lexical competition: Longer is the same as shorter*. Unpublished Work.

[67] D. J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton, 1999.

[68] D. J. Watts, P. S. Dodds and M. E. J. Newman, *Identity and search in social networks*, Science, 296 (2002), pp. 1302-1305.

[69] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440-442.

[70] G. B. West, J. H. Brown and B. J. Enquist, *A General Model for the Origin of Allometric Scaling Laws in Biology*, Science, 276 (1997), pp. 122-126.

[71] J. Whitfield, *In the Beat of a Heart: Life, Energy, and the Unity of Nature*, Joseph Henry Press, 2006.

[72] K. Y. Williams and I. Charles A. O'Reilly, *Demography and Diversity in Organizations: A Review of 40 Years of Research*, in B. M. Staw and L. L. Cummings, eds., *Research in Organizational Behavior*, 1998, pp. 77-140.

[73] S. Wuchty, B. F. Jones and B. Uzzi, *The Increasing Dominance of Teams in Production of Knowledge*, Science, 316 (2007), pp. 1036-1039.