An Algorithmic Approach to Social Networks

by

David Liben-Nowell

B.A., Computer Science and Philosophy, Cornell University, 1999
M.Phil., Computer Speech and Language Processing, University of Cambridge, 2000

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2005

Certified by .....
Erik D. Demaine
Assistant Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

An Algorithmic Approach to Social Networks

by

David Liben-Nowell

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

Social networks consist of a set of individuals and some form of social relationship that ties the individuals together. In this thesis, we use algorithmic techniques to study three aspects of social networks: (1) we analyze the "small-world" phenomenon by examining the geographic patterns of friendships in a large-scale social network, showing how this linkage pattern can itself explain the small-world results; (2) using existing patterns of friendship in a social network and a variety of graph-theoretic techniques, we show how to predict new relationships that will form in the network in the near future; and (3) we show how to infer social connections over which information flows in a network, by examining the times at which individuals in the network exhibit certain pieces of information, or interest in certain topics. Our approach is simultaneously theoretical and data-driven, and our results are based upon real experiments on real social-network data in addition to theoretical investigations of mathematical models of social networks.

Thesis Supervisor: Erik D. Demaine
Title: Assistant Professor

Acknowledgements

I get by with a little help from my friends.
— John Lennon (1940–1980)
Paul McCartney (b. 1942).

When I was applying to graduate school, I remember, I read maybe a dozen web pages with titles like "How to Succeed in Graduate School in Computer Science," or "Keeping Yourself Going towards a Ph.D." Each one said something like: you will feel lost; your resolve will waver; there will be massive hurdles that you will think are impossible to clear; you will doubt your abilities; you will think very hard about dropping out of your program. I also remember thinking that these web pages were total bunk. A half-decade later, I have, of course, accepted them as the gospel truth that they are. The reason that I am writing an acknowledgements page for a Ph.D. thesis at all is that my family and my friends and my fellow students and my mentors got me through these moments of doubt, even if they didn't know that was what they were doing at the time. To them I give my most sincere thanks. (The other thing that got me through these moments of doubt was reading acknowledgements pages in Ph.D. theses; for me, at least, they're always the best part. The other other thing, I suppose, is a strong stubborn streak, for which I think I owe dubious thanks to my family.)

Many thanks to my coauthors—especially Dan Gruhl, R. Guha, Jon Kleinberg, Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins, with whom the work in this thesis was done. (There's nothing that takes away the sting of a paper rejection quite like the hilarious email that Andrew is sure to send in response.) Thanks also to Seth Gilbert, Susan Hohenberger, April Rasala Lehman, and Matt Lepinski—the chance to work with other theory students was one of the biggest reasons that I ended up at MIT five years ago, and it has indeed been one of the highlights of my time here. (Also, Ryan O'Donnell should be on this list for the crosswords—not to mention for the hours we spent finding a way to scam the rules of every sport ever invented.) Special thanks to Susan—a coauthor, officemate, and friend; for somebody who disagrees with pretty much everything I've ever believed, you're okay by me. Thanks too to my committee members Piotr Indyk and David Karger, and my advisor Erik Demaine. I've learned a lot from many faculty members at MIT, but I've learned especially much from David and Erik. And, of course, Be Blackburn, who helped smooth over any emotional lows with corresponding sugar highs. I'd also like to thank the CS department at Cornell for getting me started on the road towards graduate study in computer science. There are a great number of people there who helped get me going, but I will single out only Jon Kleinberg. It has truly been a pleasure to count him as a mentor and collaborator, and he has always had time for my every question, big and small. I can't say enough to thank him.

There are far too many people to whom I owe personal thanks for me to list without incurring the dual risks of the boredom of readers and the omission of friends, so I will thank many in bulk—Paulina Street and all my unofficial roommates there (including those who also never lived there), my real roommates (especially the Brookline Street ones), my friends from the fields (especially Strange Blue, 7up, Vitamin I, Hoochie Blue, and Mr. T's House of Pity), everyone on what used to be the third floor (especially those with whom I wiled away so many afternoons throwing it down), and all of my old State High and Cornell friends who have stayed in touch. For those who don't fall into the above categories but to whom I owe my gratitude—you know who you are—think only that you are in a category of your own, and know that you are very much appreciated. There are a handful of people with whom I've shared every up and down, though, and I'd like to pay them very special thanks. KSW—mazel tov, slugger. RES—without you, I'd never have made it. AFL—for knowing me better than I know myself. You find the thalweg and I'll follow it with you. RCN, SPN—for always being there. RMD—for the words of wisdom when they were most needed. And, most of all, LSL—you've always been the first person I call, and it's not once been a mistake. My DSWM, no lie. Thank you for everything.

This thesis is dedicated to my grandfather, Jay Liben, chairman of the bellyaching department, who has influenced me far beyond my ability to express it.

— David Liben-Nowell, May 2005, Cambridge, MA.

Contents

List of Figures

Chapter 1

Introduction

INTRODUCTION, n. A social ceremony invented by the devil for the gratification of his servants and the plaguing of his enemies.
— *Ambrose Bierce (1812–c. 1914),*
The Devil's Dictionary.

With the rise of online networking communities like Friendster [62] and Orkut [151]—along with the popularization of the notions of "six degrees of separation" and the Kevin Bacon game [168]—the concept of a *social network* has begun to take hold in popular culture. A social network is simply a structure consisting of people or other entities embedded in a social context, with a relationship among those people that represents interaction, collaboration, or influence between entities. (One can consider a variety of types of this relationship: the mutual declaration of friendship, an email sent from one person to the other, the co-authorship of a scientific paper, and so forth.)

Significant work on social networks has been done by sociologists and social psychologists since the 1950's—see the historical survey in Wasserman and Faust's book [173], for example—though almost all of this research was carried out on small-scale networks, because of the inherent difficulty of getting reliable, large-scale data from subjects about their friendships. The now-trite claim that the internet has revolutionized the way we live is often a blatant exaggeration, but in the case of social-network research, it holds true: the internet has afforded a new opportunity to study social interactions on a much grander scale than was previously possible. We now have access to large-scale data on interactions and acquaintanceships that were unavailable on any appreciable scale as recently as a decade ago; the size of social networks that can be studied has exploded from the hundreds of people to the hundreds of thousands of people or even more.

In this thesis, we study several questions about social networks from an algorithmic perspective. Our approach is simultaneously theoretical and data-driven, and we make use of the results of real experiments on real social-network data in addition to theoretical investigations of models of such networks. We present three main contributions to the study of social networks:

(1) The "small-world" phenomenon—the observation that most people in the world are connected via short chains of intermediate friends—has been observed empirically, and a number of mathematical models have been proposed as potential explanations of it. However, none of the previously proposed models have been shown to explain small worlds both mathematically

and in real-world networks. In Chapter 2, we provide a new explanation of the small-world phenomenon, via a novel geography-based model of friendships that is both empirically and theoretically supported as an explanation of our small world.

(2) Social networks rapidly evolve over time, with new friendships forming through random encounters, introductions by mutual friends, and meetings through common interests. In Chapter 3, we explore mechanisms for predicting the formation of new relationships in a social network solely on the basis of the existing friendships in the network. Via the application of a variety of graph-theoretic techniques to the network of existing friendships, we are able to predict new interactions much more accurately than by random guessing, establishing that there is significant information implicit in the network that can be exploited to reveal important patterns of the network's future.

(3) Essentially all computational analysis of social networks takes the network itself as input, usually deriving it from personal interviews with people or automatically extracting it from a database of communication (like email or phone calls). In Chapter 4, we show how to infer social connections in a network, attempting to extract a social network via observation of the topics of conversation by the members of the network, and thereby via the inferred flow of information through the network.

(See Section 1.5 for a more detailed summary of our contributions.)

In the remainder of this chapter, we introduce some of the previous research on social networks and also some of the terminology and key concepts that will reappear throughout this thesis. For further background on social networks, a great many surveys are available. These sources include the general-interest books by Watts [175], Barabási [22], and Gladwell [68]; the more technical books of Wasserman and Faust [173], Scott [161], Buchanan [37], Dorogovtsev and Mendes [52] and Watts [174]; and survey articles by Newman [136, 142, 144], Albert and Barabasi [11], Strogatz [165] and Dorogovtsev and Mendes [51]. Finally, for a less mathematical perspective on the topic, the interested reader is directed to the play *Six Degrees of Separation* by John Guare [77], which was later turned into a movie starring Will Smith—who was in *Men in Black* with Tommy Lee Jones, who was in *JFK* with Kevin Bacon, giving Will Smith a Bacon number of at most two [168].

1.1 The Small-World Phenomenon

One of the first major achievements in the study of social networks occurred with the innovative experimental approach that was developed by Stanley Milgram in the 1960's [126, 170]. His goal, more or less, was to experimentally determine the length of the shortest chain of friends between a typical pair of people in the United States. (This question was previously considered by Pool and Kochen [44], whose results were published after Milgram's work appeared, but were written and distributed prior to his study.)

In the best-known small-world experiment, Milgram performed his study as follows. (Other studies by Milgram used different sources and targets.) First, a hundred subjects in Omaha, Nebraska were recruited, and each was sent a letter. Every letter-holder u was given instructions to forward the letter to some friend v of u—by Milgram's definition, a person v was considered to be a friend of u if u knew v on a first-name basis—with the eventual goal of reaching a friend of Milgram, a stockbroker living in Sharon, Massachusetts. Letter-holders were also given a small amount of

information about the target person, including his address, occupation, college and graduation year, and the like. Each subsequent recipient of the letter was given the same instructions and the same information about the target.

Of the chains that reached the target stockbroker—eighteen of the ninety-six that were actually started; four of the volunteer source subjects failed to initiate their chains—the average length of the path was approximately six, from which the popular notion of "six degrees of separation" was spawned. Note that over two-thirds of the letters did not actually reach the target, so the "six" here is misleadingly small: not only are shorter paths more likely to be completed if there is a fixed probability of attrition at every step of the chain, but a letter-holder may be even more likely to drop out of the experiment if he is far from the target. Judith Kleinfeld compellingly discusses some reasons for doubt in the widely accepted conclusions drawn from the Milgram study—e.g., the participation bias for volunteer subjects to be more social people and thus to be better connected in general, and the fact that in a previous small-world study performed by Milgram under 5% of chains actually reached the Harvard Divinity School student who was the target [106]. Later researchers were able to replicate the Milgram experiment on a smaller scale—across a university campus, or within a single ethnic community in a city—using similar methods [79, 98, 109, 119, 162], and a recent larger-scale small-world experiment conducted via email has shown similar results [49]. (See also the collection of surveys edited by Kochen [108].)

1.2 Explaining the Small World

Forty years after the original Milgram experiment, it is easy to think that Milgram's results are intuitive and obvious, but at the time that he performed his study the number "six" was a big surprise. When Milgram was reporting his results, the prevailing hunch in his audiences—including mathematically sophisticated audiences—was that the average length of chains from a random Omaha resident to a stockbroker near Boston would be quite large, possibly reaching into the triple digits. The small-world phenomenon does need an explanation.

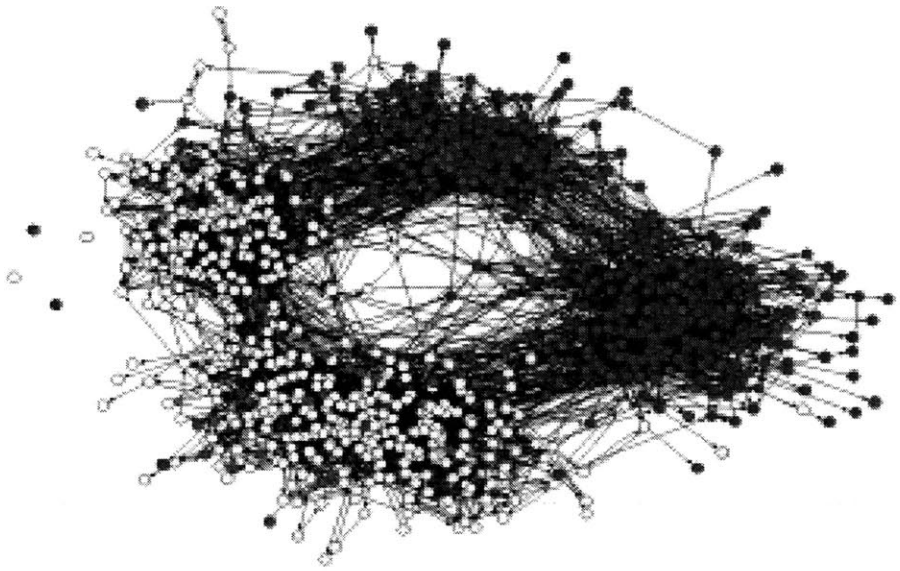
Before we discuss specific models that might explain the small-world phenomenon, we note that real social networks are *sparse* graphs. On average, a person has on the order of at most a thousand friends [108], and thus a realistic model cannot use a dense graph to explain the small world.

1.2.1 Erdős and Rényi: Random Graphs

Historically, the standard explanation given for Milgram's experimental results is simple: random graphs have small diameter. The standard random-graph model is the $G(n, p)$ random graph of Erdős and Rényi [56–58]: begin with a set of n nodes and, for each of the $\binom{n}{2}$ pairs of nodes in the graph, add an edge connecting them independently at random with some fixed probability p. Significant mathematical research has been carried out on $G(n, p)$ graphs (see, e.g., [16, 27]), showing that, among many other properties, random graphs have small diameter. (Paul Erdős, of course, plays another role in the history of social-network research: the mathematician's analogue of the Bacon game is the Erdős number, the graph distance from Erdős in the network of coauthorship, and calculating one's Erdős number—or, perhaps, selecting one's collaborators to minimize it—is a popular pursuit [40].)

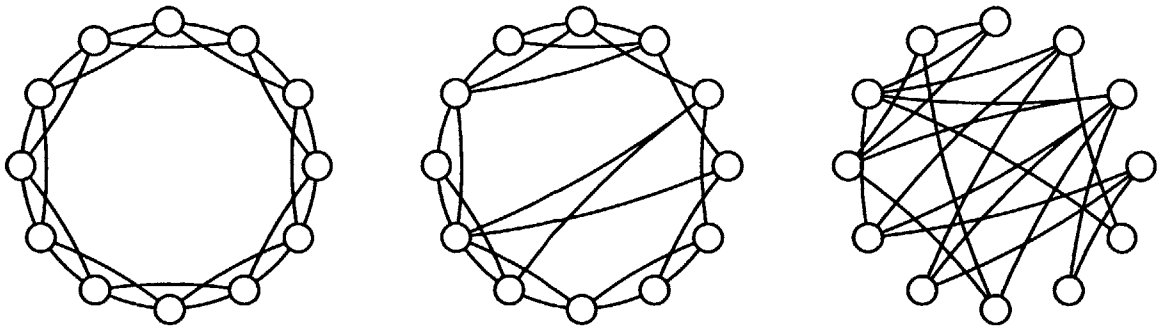
However, the relevance of random graphs to social networks is tenuous at best. See, for example, Figure 1-1, where we show two high-school social networks, of friendship and of dating [26, 131, 135].

(a) A high-school dating network. The darker (blue) nodes denote boys and the lighter (pink) nodes denote girls. Dating data are from Bearman, Moody, and Stovel [26]; the image was created by Newman [135].



(b) A network of high-school friendships, from the work of James Moody [131]. The lighter nodes denote white students (beige nodes), and darker nodes denote black students (green nodes) and students of other races (pink nodes). The vertical sparse cut is based on race; the horizontal sparse cut is based on age.

Figure 1-1: *Two high-school social networks—dating and friendship—derived from relationships self-reported by the students. Note that the visual placement of nodes depends only on edge data, without reference to any node characteristics.*



(a) The trivial rewiring probability $p = 0$, in which each node is connected to its nearest neighbors.

(b) An intermediate rewiring probability $p = 0.1$, where most edges still connect nearby neighbors, but a small fraction connect points to randomly chosen nodes.

(c) The trivial rewiring probability $p = 1$, in which each edge connects each node to a node chosen uniformly at random.

Figure 1-2: *The Watts/Strogatz social-network model of a randomly rewired ring lattice. A circle is formed of n people, each of whom is connected to the $\Theta(1)$ nearest neighbors in each direction. Then, with some probability p, each edge is "rewired" by replacing one of the edge's endpoints with a connection to a node chosen uniformly at random from the entire network.*

Even a cursory glance at these networks indicates that an Erdős/Rényi random graph is not a good model for these networks: the generation probability is very low for either a near-pseudotree (a connected graph formed by replacing the root node of a tree by a cycle) or a graph whose nodes can be partitioned into four densely connected sets that are sparsely connected to each other. More quantitatively, the *clustering coefficient* of a network (see Section 1.3.3) is a measure of the probability that two people who have a common friend will themselves be friends. The clustering coefficient of social networks is radically higher than in Erdős/Rényi random graphs when $p \ll 1$, which is necessary for the resulting graph to be sparse. This difference is partially explained by the tendency towards *homophily* in social networks: a person's friends are typically a lot "like" the person herself, in terms of race, geography, interests, and so forth. (Furthermore, the *degree distribution* of $G(n,p)$ random graphs is well approximated by a binomial distribution, which is a poor fit to real social networks. See Section 1.3.4.) Social networks are simply not well modeled by $G(n,p)$ random graphs.

1.2.2 Watts and Strogatz: The Rewired Ring Lattice

In the late 1990's, Duncan Watts and Steve Strogatz gave a new random graph model for social networks that simultaneously accounts for the two previously discussed properties found in real social networks, defining a model that yields both small diameter and high clustering coefficients [177]. Their work reinvigorated interest in the field of social networks and attracted researchers with a more mathematical inclination to the area.

Their network model is as follows. (See Figure 1-2.) Place n people on a k-dimensional grid—they focus on the case $k = 1$, but the model applies in higher dimensions as well—and connect each person u in the grid to all other people within a threshold Manhattan distance $\delta > 1$ of u. These are the "local neighbors" of person u in the network. Then some of the connections in the network

are randomly altered: with some probability p, for each edge $\langle u, v \rangle$ in the graph, randomly "rewire" the edge to become $\langle u, v' \rangle$, where v' is chosen uniformly at random from the set of all nodes in the graph. For $p = 0$, we have high clustering coefficients: for any two local neighbors u and v, most of the nodes that are near u are also near v. We also have a high diameter (polynomial in the population size), because the only edges in the graph are local links that allow hops that cover $O(1)$ distance in the grid. On the other hand, for $p \approx 1$, we have the low diameter of a random graph—the resulting graph is virtually identical to a $G(n, p)$ random graph, which, as discussed above, has low diameter—but we also have low clustering coefficients, again because the resulting graph is essentially a $G(n, p)$ random graph.

The main result of the study by Watts and Strogatz was the empirical discovery that the network has the best of both worlds for intermediate values of the rewiring probability p—that is, it exhibits both high clustering coefficients and low diameter. High clustering coefficients persist because most of the original local links in the network are unchanged (because the rewiring probability p is small), and, as above, for two nearby nodes u and v, the nodes close to u are mostly the same as the nodes close to v. Intuitively, the small diameter appears even for relatively small values of p because the random long-range links effectively form a random graph connecting small clusters of proximate nodes, and thus the cluster-level diameter is small. (Subsequent research on Watts/Strogatz-type graphs has typically considered adding new edges rather than rewiring preexisting edges, which is much more conducive to analysis and avoids the issue of disconnecting sets of nodes in the network (e.g., [102, 105, 136, 148, 149]).

Similar models have also been considered rigorously in the mathematics and computer-science literature; for example, the diameter of a graph consisting of an n-node cycle augmented by a random matching has been shown to be $\Theta(\log n)$ with probability approaching one as n tends towards infinity [28].

1.2.3 Kleinberg: A Navigable Small World

A major recent development in the modeling of social networks came about with the work of Kleinberg [102, 105], who made an important observation about the original Milgram small-world experiment that had seemingly been missed by previous researchers. Milgram's experiment shows *two* important things about social networks: not only do short chains of friends *exist* between arbitrary pairs of people in a social network, but the members of the network are able to *construct* short paths using only "local" information about their own friends. In other words, rather than viewing the Milgram experiment as a result about the magnitude of the diameter of a social network, one should view it as a result about the success of a particular type of routing algorithm when it is run on the graph.

Having made this observation, Kleinberg then proved that the Watts/Strogatz model cannot account for the success of Milgram's subjects in constructing short paths. Specifically, he showed that any *local-information algorithm*—formally, a routing algorithm that knows the layout of the grid, but not the specific rewirings that have occurred in the network—requires a polynomial number of steps to route messages in the network. Kleinberg then developed an extension of the Watts/Strogatz model, consisting of a network with a k-dimensional mesh of people, where each person knows his or her immediate geographic neighbors in every direction, and the probability of a long-distance link from u to v is proportional to $1/d(u, v)^\alpha$, for some constant $\alpha \geq 0$, where $d(u, v)$ denotes the Manhattan distance between u and v. Each node in the network chooses $\Theta(1)$ long-range links according to this distribution. (This model generalizes the edge-addition version

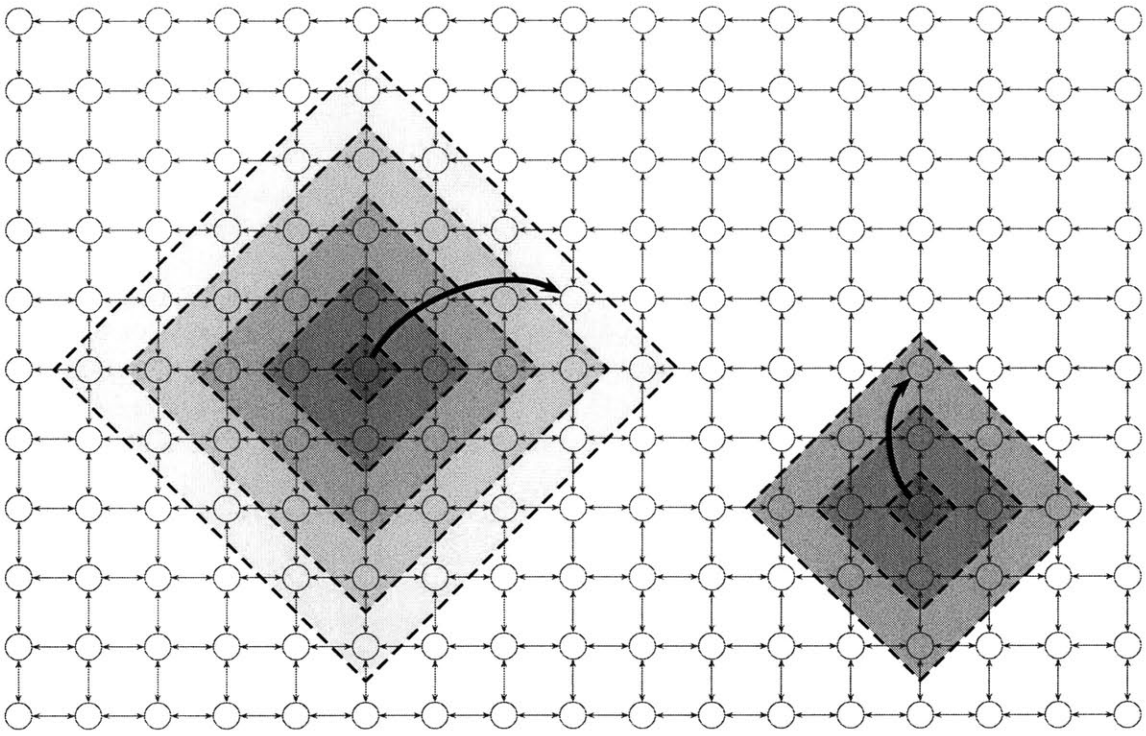


Figure 1-3: *The Kleinberg social-network model. Starting with a k-dimensional mesh (here $k = 2$), we connect each node to its nearest neighbor in each cardinal direction. Then, for each node u, we select a long-range link probabilistically, where the probability that u's long-range link is v is proportional to $d(u,v)^{-\alpha}$, for a parameter $\alpha \geq 0$, where $d(u,v)$ denotes the Manhattan distance between u and v. In the figure, the long-range links are shown for two of the nodes, along with dashed diamonds that show the distances to all nodes at least as close to the source as the randomly selected long-range link. Darker shading indicates a higher probability of a link.*

of the Watts/Strogatz model in which new edges are added with probability p, as setting $\alpha := 0$ yields a uniformly chosen long-range link.) See Figure 1-3.

The main result of Kleinberg's work is a characterization theorem that shows exactly when a local-information algorithm can construct short (that is, of length polylogarithmic in the population size) paths between nodes in the graph. The positive results (upper bounds on delivery time) derived by Kleinberg are based upon the *greedy algorithm*: for a person u currently holding a message eventually bound for a target individual t, person u selects as the next step in the chain the friend v of u who is closest to the target t in terms of Manhattan distance in the grid. Kleinberg's main result is the following theorem:

Theorem 1.2.1 (Kleinberg [102]) *In any n-person, k-dimensional mesh in which each person chooses $\Theta(1)$ long-range links according to the distribution $\Pr[u \to v] \propto d(u,v)^{-\alpha}$, and for any source person s and any target person t:*

1. *if $\alpha = k$, then the greedy algorithm will find a path of length $O(\log^2 n)$ with high probability.*
2. *if $\alpha \neq k$, then any local-information algorithm will find a path of length $\Omega(n^{\varepsilon})$, for some constant $\varepsilon > 0$, with high probability.*

Intuitively, if $\alpha > k$, then the long-range links are too close to the current message-holder and simply do not cover enough ground in terms of grid distance to reach the target in a small number of steps. Conversely, if $\alpha < k$, then the long-range links are too widely scattered to allow a local-information algorithm to "home in" on the target's location in the grid; although short paths are still likely to exist, the edges are not sufficiently clustered to guarantee that closing in the target in terms of grid distance means closing in on the target in terms of graph distance. We note that Kleinberg's positive results require only one long-range link to be chosen for every node in the network, but his negative results continue to hold for any constant number of long-range links.

Subsequent research has extended these results in a number of different directions. Kleinberg has shown analogous results in other underlying graph models in place of the grid, including hierarchies and "group structures" [103]. Barrière et al. [25] have proven that $\Theta(\log^2 n)$ steps are required for the greedy algorithm when $k = \alpha = 1$. Martel and Nguyen [124, 150] and Fraigniaud, Gavoille, and Paul [61] have shown that the expected diameter of a Kleinberg-style social network is $O(\log n)$ for any k and have given an algorithm to find paths of length $O(\log^{1+1/k} n)$ using some additional knowledge about the long-range links of nearby nodes in the graph. They have also shown that the greedy algorithm requires $\Omega(\log^2 n)$ steps in any Kleinberg-style network, for any k. Using *friend-of-a-friend routing*, in which each message-holder chooses the next recipient greedily on the basis of the grid locations of her friends *and* her friends' friends, achieves a $\Theta(\log n / \log\log n)$-step route with high probability when nodes have logarithmically many neighbors [123]. Finally, Slivkins has extended Kleinberg's results to metrics of low doubling dimension [163].

1.3 Properties of Social Networks

Now that we have defined the social-network models most relevant for the remainder of this thesis, we turn to a brief survey of some important properties of social networks that have been discussed in the literature. Some of these properties were mentioned in previous discussion—some extensively—but we include them again here for ease of reference.

Mathematically, a social network is just a graph $\langle V, E\rangle$, where the nodes represent people, and an edge $\langle u, v\rangle \in E$ denotes some type of social relationship between the people u and v. We use graph-theoretic terminology in the following, but readers whose background is in a different field can mentally translate "node" into *vertex*, *actor*, or *site*, and "edge" into *arc*, *tie*, *link*, or *bond*. We also note that "shortest path" is synonymous with *geodesic*.

1.3.1 Diameter

As discussed extensively in Sections 1.1 and 1.2, there are short chains of friends that connect a large fraction of pairs of people in a social network. It is not completely clear that social networks have a small diameter in the graph-theoretic sense of the longest shortest path; a recluse (Greta Garbo, J. D. Salinger) might even be disconnected from the rest of the world. Unfortunately, the literature on social networks tends to use the word "diameter" ambiguously in reference to at least four different quantities: (1) the *longest* shortest-path length, which is the true graph-theoretic diameter but which is infinite in disconnected networks, (2) the longest shortest-path length between connected nodes, which is always finite but which cannot distinguish the complete graph from a graph with a solitary edge, (3) the *average* shortest-path length, and (4) the average shortest-path length between connected nodes.

A wide variety of other intriguing networks have been shown to have small average shortest-path length, of which we highlight the world-wide web, for which the average (directed) shortest-path length was estimated to be under twenty by Albert et al. and Broder et al. [12, 36], and which is a partially social network that shares some (but not all) of the characteristics of the social networks discussed in this thesis.

1.3.2 Navigability

As discussed in Section 1.2.3, a further characteristic observed in real social networks is that they are *navigable* small worlds: not only do there exist short paths connecting most pairs of people, but using only local information and some knowledge of global structure—for example, each person u in the network might know the geographic layout of the United States and the geographic locations of only some target individual and each of u's own friends—the people in the network are able to construct short paths to the target.

Although no rigorous theoretical analysis of it has been given, we would be remiss if we proceeded without mentioning the small-world model defined by Watts, Dodds, and Newman [176] in which navigation is also possible. Their model is based upon multiple hierarchies (geography, occupation, hobbies, etc.) into which people fall, and a greedy algorithm that attempts to get closer to the target in any dimension at every step. Simulations have shown this algorithm and model to allow navigation, but no theoretical results have been established.

1.3.3 Clustering Coefficients

In Section 1.2.1, we discussed one of the limitations of the random-graph model of social networks, namely that the *clustering coefficient* of social networks is much higher than is predicted in a $G(n, p)$ random graph. Informally, the clustering coefficient measures the probability that two people who have a common friend will themselves be friends—or, in graph-theoretic terms, the fraction of triangles in the graph that are "closed."

There are a handful of different ways to measure the clustering coefficient in a graph formally; here we mention just one, for concreteness. The *clustering coefficient* for a node $u \in V$ of a graph $G = \langle V, E \rangle$ is $|E[\Gamma(u)]|/\binom{\Gamma(u)}{2}$—that is, the fraction of edges that exist within the neighborhood $\Gamma(u)$ of u, i.e., between two nodes adjacent to u. The clustering coefficient of the entire network is the average clustering coefficient taken over all nodes in the graph. (Other formalizations compute the probability that the third edge $\{v, w\}$ exists when we choose an ordered triple $\langle u, v, w \rangle$ such that $\{u, v\}, \{u, w\} \in E$ is chosen uniformly at random from the set of all such triples; the formalization described above gives relatively less weight to high-degree nodes than in this alternative approach.)

Typical social networks have clustering coefficients on the order of 10^{-1}, which is orders of magnitude greater than the clustering coefficient that a $G(n, p)$ random graph would exhibit when the link probability p is set in a reasonable range to approximate the sparsity of these networks [177]. A slightly more general way of viewing clustering coefficients is as a manifestation of the following phenomenon: for a pair $\langle u, v \rangle$ of people in a social network, the event that an edge between u and v exists is highly negatively correlated with the graph distance between u and v in the network with the possibly existent edge $\langle u, v \rangle$ deleted. (In a $G(n, p)$ graph, of course, the probability of the existence of the edge $\langle u, v \rangle$ is independent of the graph distance.) This correlation captures the notion that graph distance in the social network somehow captures social distance, and that

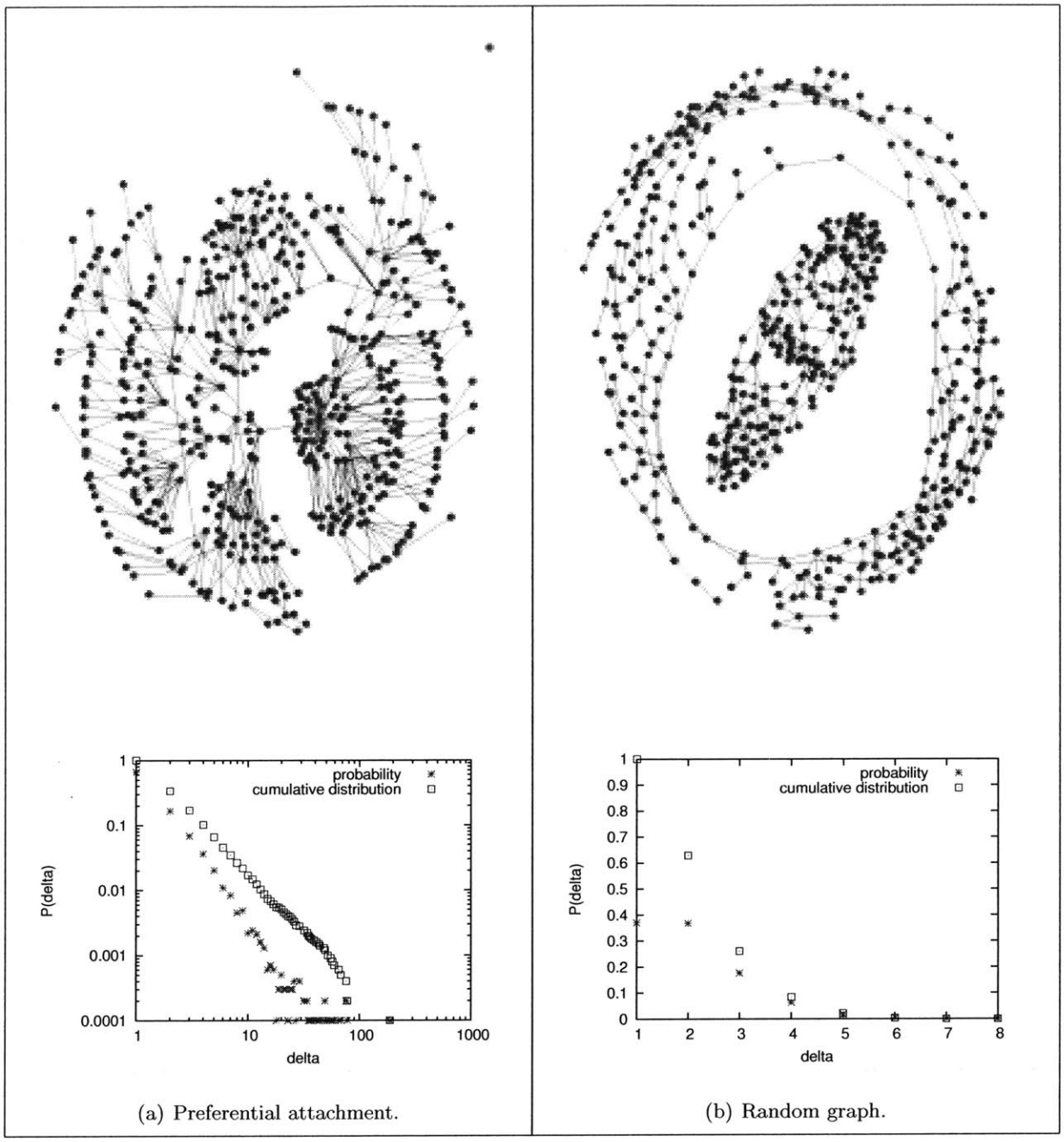
people separated by a smaller social distance are more likely to be friends. We discuss this idea extensively in Chapter 3.

1.3.4 Degree Distributions

The small-world models that we have discussed up until now have contained essentially homogeneous nodes; variations among nodes are relatively minor. In particular, the *degree distribution* of the network—that is, the proportion of nodes in the network who have a particular *degree* δ (i.e., the fraction of people with exactly δ friends), for every degree $\delta > 0$—has been essentially constant in the random graphs of Erdős and Rényi, Watts and Strogatz, and Kleinberg. Real social networks, however, show a wide heterogeneity in the popularity of their nodes, and their degree distributions are extremely poorly predicted by any of the models discussed above. In a typical social network, the proportion $P(\delta)$ of nodes with degree at least δ is reasonably well approximated by the *power-law distribution* $P(\delta) \propto \delta^{-\beta}$, for a constant $\beta > 0$, usually where $\beta \approx 2.1$–2.5. (Others have referred to networks exhibiting a power-law degree distribution as *scale-free* networks, or as exhibiting *Pareto*, *heavy-tailed*, or *Zipfian* degree distributions.) This phenomenon was noted for the degree distribution of the world-wide web [12, 112], and has also been observed in social networks [23, 137], along with some other interesting networks of various forms [17, 24]. In Figure 1-4, we show two randomly generated networks to aid in visualizing the difference between a power-law graph and an Erdős/Rényi-type random graph. (The power-law graph in the figure is generated according to the preferential-attachment model, described below.)

Mitzenmacher [128] has written a comprehensive survey of the literature on power-law distributions, especially noting that during the recent spate of research on social networks, the web, and other large-scale real-world networks, researchers have essentially rediscovered a vast literature regarding power-law distributions that has been studied under various guises many times in many different fields. He discusses the debate—which he notes has been refought in essentially every one of these disciplines—as to whether a power-law distribution or a *lognormal distribution* is a better model for the degree distribution in real-world graphs like social networks. (A random variable e^X follows a lognormal distribution if the random variable X follows a normal distribution. On a log/log plot of the probability density function, a lognormal distribution appears parabolic while a power-law distribution is linear.)

Perhaps the simplest model of social networks that matches the observation of power-law degree distribution simply generalizes the Erdős/Rényi random-graph model to produce the correct degree distribution. For each node u in the network, one can simply produce "stub" half-edges (with one endpoint fixed to be u) by choosing a degree $\mathsf{deg}(u)$ according to a power-law distribution, and then randomly connecting stub edges together to produce a graph (e.g., [9, 122, 129, 130, 147]). The important *preferential-attachment model* of Barabási and Albert gives another way of generating a power-law distribution [23]. This dynamic model works as follows. At every time step t, a new node u_t is introduced into the system and is endowed with exactly one edge. The neighbor of u_t is chosen probabilistically, with $\mathtt{Pr}[u_t \to v] \propto \mathsf{deg}(v)$. This model was shown via simulation to produce a power-law degree distribution [23], and has since been rigorously analyzed by Bollobás et al. [31], including a proof that preferential attachment yields a network with $\Theta(\log n)$ diameter (and $\Theta(\log n / \log\log n)$ if a constant $c > 1$ number of edges are added in the same manner for each new node) [30]. Power-law distributions are also generated by a copying model that has been advanced for the web, in which a new node can choose its neighbors by copying them from a node already in the network [104, 115].



(a) Preferential attachment.

(b) Random graph.

Figure 1-4: *Two example networks and their degree distributions. We generate random graphs with an average outdegree of one using two methods: the preferential-attachment model of Barabási and Albert and a fixed-outdegree variant of the $G(n,p)$ model of Erdős and Rényi. For each model, we show a randomly generated graph of $n = 500$ nodes, drawn by Ucinet [32], and the degree distribution for a randomly generated graph of $n = 10{,}000$ nodes. Notice that the degree distribution plot is log/log for preferential attachment and is linear for the random graph.*

1.3.5 Other Properties of Social Networks

Here we briefly highlight a few other properties that have been observed in real-world social networks, relating to the size of the network's connected components, correlations between the degrees of nodes that are neighbors ("assortative mixing"), and the structure of communities in the network. Newman and Park [146] have argued that high clustering coefficients and two of these properties—assortative mixing and community structure—are the crucial statistical properties that distinguish social networks from other types of networks.

Giant Component

One obvious corollary of the small-world effect ("there are short paths between most pairs of nodes") is that social networks must have a large connected component containing most of the nodes ("there are paths between most pairs of nodes"). This *giant component* typically encompasses the majority of the nodes in the network. The size of the largest component in a random graph has been well studied, in the mathematics and computer-science literature as well as in the physics literature. (In physics, the presence of a giant component is studied under the name of *percolation theory*.)

Assortative Mixing

While the degree distribution of a network gives some important information about its structure, the degree of a node is an inherently local quantity. One can also look at the relationship between the degrees of nodes depending on their location in the network. Do popular people tend to have popular or unpopular friends? The tendency of social networks—in contrast to biological and technological networks, which exhibit almost all of the social-network properties outlined above—is to exhibit *assortative mixing*, in which there is a positive correlation between the degree of a node u and the degree of a neighbor of u [110, 140, 143].

Community Structure

The clustering coefficient is a measure of an extremely local notion of community, but larger-scale measures are also interesting and appropriate. (This is the traditional sense of "clustering" used in the computer-science community.) As one would expect, social networks tend to exhibit significant community-based organization, showing joint interests, membership in organizations, or fields of study [67, 113, 156].

1.4 Sources of Social Networks

The recent explosion of interest in social-network research, especially the newfound interest from computer scientists and statistical physicists, can be attributed in large measure to the proliferation of available large-scale data on social interactions. Here we briefly mention some of the types of networks derived from various social interactions that are now available for research.

1.4.1 Networks of Self-Reported Real-World Interactions

Traditional sociological research into social networks was based upon painstaking personal interviews with subjects, in which researchers would ask subjects to report the people with whom they

were friends. (One famous example of a study of this form is Zachary's Karate Club [184], where data on interactions among the thirty-four members of a soon-to-schism karate club were collected by interviews and observations over a two-year period.)

In addition to the inherent tedium of data collection, one faces an additional difficulty in data gathered from interviews and surveys: people have widely differing standards of what constitutes a "friend." (Milgram largely handled this problem by explicitly defining a friend as a person with whom one is on a mutual first-name basis, but even this standard leaves open some matters of interpretation.) For example, in recent research on networks of sexual contacts [118], one observes a power-law-like distribution in the number of partners that a subject has had within the past year, but the number of partners reported by men is consistently larger than the number reported by women. Although this difference could conceivably reflect reality, it seems far more likely that this difference is the result of different subjective interpretations of "partner," perhaps because of societal pressure for men to report higher numbers and women to report lower numbers.

1.4.2 Communication Graphs

One automatic way to extract social networks is to define friendship in terms of communication and extract edges from records of that communication.

Call graphs [1, 9] are one natural basis for deriving a communication-based network. Here nodes are telephone numbers, and a (directed) edge is recorded from u to v, with a timestamp t, if a call is placed from u to v at time t. There are some problems of node conflation here, as more than one person may be associated with the same phone number, and call data are not always readily available outside of telecommunications companies.

One can also use *email networks* [3, 53, 145] to record social interactions, analogously to call graphs. Here there are issues of privacy, so most studies have taken place on relatively limited email datasets that are not publicly available. An interesting recent development on this front was the late-2003 release by the Federal Energy Regulatory Commission of around 1.5 million emails from the top executives at Enron, containing about 150 Enron users and about 2000 total people [55, 107].

1.4.3 Collaboration Networks

Another source of data on interactions is *collaboration networks*: an (undirected) edge exists between u and v if they have collaborated on a mutual project. One can consider many different types of projects, but the most-studied collaboration networks are based on the *Hollywood-actor collaboration graph* [10, 17, 23, 147, 177], where actors who jointly appear in a movie are connected by an edge, and the *academic collaboration graph* [21, 138, 139], where researchers who coauthor an academic paper are connected by an edge. (*Citation networks*, in which the nodes are papers connected by directed edges from each paper to the papers that it cites, are not social networks, in that their edges point only backwards in time, and do not really represent a social relationship between their endpoints.)

A collaboration network as we have described it here is a derived structure that loses some of the information present in the original data. For a lossless representation of a collaboration network, a set of collaborations can be represented by a bipartite graph of people and projects, where a person u is connected to a project p if and only if u worked on p. The collaboration graph as described above looks only at the people in the bipartite graph, connecting u to v exactly if there is a length-two path connecting u and v in the bipartite graph.

1.4.4 The Web and Blogs

The web itself has some strongly social characteristics—the "my friends" section of many a personal home page being an explicit example—but overall the link structure is only partially social. Particular types of web pages, though, have much more purely social link structure. A *blog* (an abbreviation of "web log") is an online diary, often updated daily (or even hourly), typically containing reports on the user's personal life, reactions to world events, and commentary on other blogs. Links on blog pages are often to other blogs read by the author, or to the blogger's friends, and thus the link structure within blogging communities is an essentially social relationship [113, 114]. Thus social networks can be derived from blog-to-blog links on a blogging community website.

1.5 Contributions of this Thesis

In this thesis, we consider three questions arising regarding social networks: inferring a social network, analyzing its existing friendships, and predicting its future friendships. More specifically, this thesis considers the following: (1) connecting the theoretical modeling of small-world phenomena to observed data about real-world social networks by developing a new small-world model that matches the observations both empirically and analytically, (2) using the set of currently existing friendships in the network as the basis for making predictions about future friendships that will form in the near future, and (3) extracting an implicit social network by observing the inferred flow of information among people.

1.5.1 The Geographic Basis of Small Worlds

Stanley Milgram's experiments were the origins of a voluminous body of literature on the small-world problem, both experimental and theoretical. However, there is a disconnect between the theoretical models discussed previously and the real-world experiments—do people actually form friendships according to the distribution required by Theorem 1.2.1? Does Kleinberg's theorem explain the observed small-world experimental results? In Chapter 2, we explore the relationship between, on one hand, existing theoretical models of the small world and, on the other, linking behavior in a 1,300,000-person social network extracted from the explicit listings of friendships in the LiveJournal blogging community, `www.livejournal.com`.

In the LiveJournal system, each blogger explicitly lists her geographic location and each other blogger whom she considers to be a friend. About 500,000 LiveJournal bloggers list a hometown locatable within the continental United States. Using these data, we simulate the Milgram experiment, confirming that the LiveJournal network does indeed form a small world, navigable by the natural geographically greedy algorithm. Thus, because a local-information algorithm produces short paths, and because people live on a two-dimensional grid (formed by lines of longitude and latitude on the earth's surface), we expected to find that $\alpha = 2$, as per Theorem 1.2.1. Surprisingly, though, this relationship does *not* hold; the fit is more like $\alpha \approx 1$, which Kleinberg's results imply would not produce a navigable small world. This apparent discrepancy is resolved because people in the LiveJournal network do not live on a *uniform* grid—population density varies significantly from location to location across the United States. Thus, for a random person u in the network, the number of people who live within a radius-δ circle centered at u grows roughly linearly with δ, rather than quadratically, as would occur in a uniform grid. We introduce the notion of *rank-based friendship*, in which the probability that u has a friend v is inversely proportional to the number

of people who live closer to u than v does. We show analytically that rank-based friendship alone is sufficient to account for the small-world phenomenon observed in the LiveJournal network, and we also show empirically that rank-based friendship holds in the LiveJournal network.

1.5.2 The Link-Prediction Problem

Social networks are highly dynamic objects; they grow and change quickly over time through the addition of new edges, signifying the appearance of new interactions in the underlying social structure. Understanding the mechanisms by which they evolve is a fundamental question that is still not well understood, and it forms the motivation for the work described in Chapter 3.

We define and study a basic computational problem underlying social-network evolution, the *link-prediction problem*: given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'. In effect, the link prediction problem asks: what does a social network tell us about its own future? To what extent can the evolution of a social network be modeled using purely graph-theoretic features of the network itself? New edges in a social network, of course, form for a variety of reasons, many of which appear to be fully independent of the network structure. However, one also senses that a large number of new friendships are hinted at by the topology of the network: two people who are "close" in the network will have friends in common, and will travel in similar circles; this closeness suggests that they themselves are more likely to meet and become friends in the near future. Our goal is to make this intuitive notion precise, and to understand which measures of "proximity" in a network lead to the most accurate link predictions. We find that a number of proximity measures—led by certain fairly subtle measures of proximity—lead to predictions that outperform chance by factors of 40 to 50 on several social networks derived from collaborations among physicists in several different subfields, indicating that a social network does contain significant latent information about its own future.

1.5.3 Inferring a Social Network

Virtually all of the social-network literature discussed above takes as a given the fully specified network of interest. While this assumption is realistic for communication or collaboration networks, the connections in such networks are often viewed merely as a proxy for the "true" friendship relation about which we care most. In Chapter 4, we discuss the *inference* of social-network edges on the basis of overt behavior of people in the network.

A frequent real-life social-network-induced phenomenon is the transfer of a *meme*—a unit of cultural knowledge or information, like the purchase of a new electronic gizmo, interest in a particular political cause, or a joke forwarded through email—from one friend to another. We are interested in the possibility of inferring social connections between people by the inference of the transfer of memes from one person to another. Bloggers typically publish frequently, and they often write about the content of the blogs of their friends in addition to news about the world and their personal lives. The temporal resolution of blog postings and the culture of blogging about topics about which one's friends are blogging make blogspace fertile ground for this type of inference. We will report on experiments demonstrating that a social-network edge between two people u and v can be inferred with a reasonable degree of reliability using only information about temporally proximal postings about the same topics in u and v's blogs.

Chapter 2

The Geographic Basis of Small Worlds

Without geography, you're nowhere.
— Jimmy Buffett (b. 1946).

We live in a "small world," where two arbitrary people are likely connected by a short chain of intermediate friends. With knowledge of their own friends and scant information about a target individual, people can construct such a chain, by successively forwarding a message towards the target. In Chapter 1, we detailed both experimental studies verifying this property in real social networks and theoretical models that have been advanced to explain it. However, there is a disconnect between the experimental and the theoretical: existing theoretical models have not been shown to accurately reflect actual linking behavior in real-world social networks. In this chapter, we show that existing models do not capture the geographic distribution of friendships in a real, large-scale, online social network, making previous theoretical results inapplicable in explaining its small-world behavior. We then introduce a new, richer model of *rank-based friendship* relating geography and social-network friendship, in which the probability of befriending a particular candidate is inversely proportional to the number of geographically closer candidates. In the online social network, we show that approximately one third of the friendships are independent of geography—i.e., uniformly distributed with respect to geographic distance—and that the remainder are closely modeled by rank-based friendship. Further, we prove that any social network constructed via rank-based friendship is a navigable small world: for any population densities in the network, the path constructed by the geographically greedy algorithm from any source to a randomly chosen target has expected length polylogarithmic in the size of the population.

2.1 Introduction

Anecdotal evidence that we live in a "small world," where arbitrary pairs of people are connected through extremely short chains of intermediary friends, is ubiquitous. Sociological experiments, beginning with the seminal work of Milgram and his coworkers [98, 109, 126, 170], have shown that a *source* person can transmit a message to a *target* via only a small number of intermediate friends,

The work described in this chapter was performed jointly with Jasmine Novak, Ravi Kumar, and Andrew Tomkins at IBM Almaden, and with Prabhakar Raghavan at Verity, Inc., and appears in a manuscript entitled "Geographic Routing in Social Networks."

using only scant information about the target's geography and occupation; in other words, these networks are *navigable small worlds*. On average, the successful messages passed from source to target through six intermediaries; from this experiment came the popular notion of "six degrees of separation." (See Section 1.1.)

As part of the recent surge of interest in networks, there has been active research exploring strategies for navigating synthetic and small-scale social networks (e.g., [3, 5, 6, 99, 102, 103, 105, 176]), including routing based upon common membership in groups, popularity, and geographic proximity, the property upon which we focus. In both the Milgram experiment and a more recent email-based replication of Dodds et al. [49], one sees the message geographically "zeroing in" on the target step by step as it is passed on. Furthermore, subjects report that geography and occupation are by far the two most important dimensions along which they choose the next step [98], and geography tends to predominate in early steps of the chain [49]. This observation leads to an intriguing question: what is the connection between friendship and geography, and to what extent can this connection explain the navigability of large-scale real-world social networks? Of course, adding dimensions other than geography to routing strategies—especially once the chain has arrived at a point geographically close to the target—can make routing more efficient, sometimes considerably [3, 4, 170, 176]. However, geography appears to be the single most valuable dimension for routing, and we are thus interested in understanding how powerful geography alone may be.

In this chapter, we investigate the relationship between friendship and geography, combining a theoretical model of path discovery in social networks with empirical measurements of a large, real, online social network to validate and inform the theoretical results. First, a simulation-based study on a 500,000-person online social network reveals that routing via geographic information alone allows people to discover short paths to a target city. Second, we empirically investigate the relationship between geography and friendship in this network. The proportion of links in a network that are between two entities separated by a particular geographic distance has been studied in a number of different contexts: the infrastructure of the Internet [63, 71, 181], small-scale email networks within a company [3, 4], transportation networks [63], and wireless-radio networks [127]; here we study this relationship in a large-scale social network. We discover that approximately 70% of friendships in our social network are derived from geographical processes, but that existing models that predict the probability of friendship solely on the basis of geographic distance are too weak to explain these friendships, rendering previous theoretical results inapplicable. Finally, we propose a new density-aware model of friendship formation called *rank-based friendship*, relating the probability that a person befriends a particular candidate to the inverse of the number of closer candidates. We prove that the presence of rank-based friendship for *any* population density implies that the resulting network will contain discoverable short paths to target individuals. Rank-based friendship is then shown by measurement to be present in the large social network. Thus, a large online social network is observed to exhibit short paths under a simple geographical routing model, and rank-based friendship is identified as an important social network property, present in the network studied, whose presence implies short paths under geographic routing.

2.2 The LiveJournal Social Network

We begin our investigation into the navigability of large-scale social networks by first extracting a social network from the information provided by the millions of users of a blogging web site and then performing empirical investigations of that network.

The social network that we consider is composed of bloggers in the LiveJournal online community, found online at `www.livejournal.com`. As of mid-January 2005, over five million users had registered LiveJournal accounts, and almost one million of these users had updated their accounts within the past week [120]. Our data are drawn from the results of a crawl of the LiveJournal site that was performed during February 2004. At the time that this crawl was performed, there were 1,312,454 registered users in LiveJournal. (Incidentally, the explosion of popularity of the LiveJournal community is an example of the diffusion of an innovation that has spread largely through social-network links; many people join a community like LiveJournal because their friends have also joined it. See Chapter 4.)

The LiveJournal system is predominantly a blogging site, but it furthermore allows its users to specify additional personal information as part of their accounts. Each LiveJournal blogger explicitly provides a *profile*, including his or her geographic location, date of birth, a list of topical interests (entered in a freeform text input box with suggestions for standard formatting), and a list containing each other blogger whom he or she considers to be a friend. We define the *LiveJournal social network* on the basis of these explicitly listed friendships: the nodes are the approximately 1.3 million bloggers, and $\langle u, v \rangle$ is an edge of the network if blogger u has explicitly listed blogger v in her list of friends. Note here that friendship is a directed concept—person v may not list u as a friend even if $\langle u, v \rangle$ is an edge—but 80% of links are reciprocal. (In all of our experiments, we respect the directionality of the edges.) In this network, there are 18,154,999 directed edges, an average of just under fourteen links per person. The clustering coefficient in the network is 0.2. (Recall from Section 1.3.3 that the clustering coefficient is the probability that two different friends of a person u are themselves friends.) Further investigation into the structural and statistical properties of this network have been carried out by Kumar, Novak, Raghavan, and Tomkins [114], and we refer the interested reader to their work for more high-level properties of the network.

Our interest in the LiveJournal network lies in the fact that it contains both explicit listings of friendships—the edges of our social network—and explicit listing of additional data that describe certain properties of the people in the network. Of the 1.3 million bloggers in the system in February 2004, there are 495,836 in the continental United States who listed a hometown and state that we find in the USGS Geographic Names Information System [171]. Thus we are able to map almost 500,000 LiveJournal users to a geographic location specified in terms of longitude and latitude. There are many more users who enter partial geographic information ("Pennsylvania, United States") or who enter their cities in ways that are unrecognizable in the database ("LR, Arkansas, United States" for Little Rock), but we are limited to those users who enter precise and unabbreviated geographic locations for the experiments described below. Note that the resolution of our geographic data is limited to the level of towns and cities—more precise locations, perhaps at the level of street addresses, would be valuable data, but are not available on the LiveJournal site. Thus, our discussion of routing is from the perspective of reaching the home town or city of the destination individual. That is, we study the problem of "global" routing, in which the goal is to direct a message to the target's city; once the proper locality has been reached, a "local" routing problem must then be solved to move the message from somewhere in the target city down to the target individual him or herself. There is evidence in the work of Milgram [126] and Dodds et al. [49] that, in real-world message-passing, routing of messages tends to first zero in on the correct city geographically, subsequently moving toward the target individual using a wide set of potential non-geographic factors, like hobbies or profession.

In the remainder of this chapter, we focus our attention on these roughly 500,000 people, i.e.,

those who have provided a fully specified, locatable hometown in the United States. The number of distinct geographic locations that are home to at least one LiveJournal user is 11,387. Thus the population of an average city (i.e., the expected population of a city chosen uniformly at random from the 11,000 locations) is just under 44 people. City populations roughly follow a power-law distribution, and range from one (in 2778 distinct locations) to just under 8000 (in Seattle, Washington). The next-most populous cities are, in descending order, Los Angeles, Chicago, New York, Houston, and San Diego, each with over 5000 $\approx$ 1% of LiveJournal residents. Because of skewed distribution of city populations, an average person in the network lives in a city containing significantly more than 44 LiveJournal users: the population of the city of an average LiveJournal user (i.e., the expected population of the city in which a uniformly chosen LiveJournal user lives) is 1306. There are 3,959,440 friendship links in the directed LiveJournal network, an average of about eight friends per user. (Notice that even treating each city as a "supernode" in the graph, there are still only $8 \cdot 44 = 352$ edges leaving an average city. Thus an average city c is connected to at most 352 of 11,386 other cities, even disregarding the fact that a large fraction of the friends of the inhabitants of city c live in city c themselves.)

In Figure 2-1, we show both indegree and outdegree distributions for the LiveJournal social network, both for all 1,300,000 people in the entire system and for the 500,000 people we geographically locate within the United States. The power-law degree distribution that has been predicted by a number of social-network growth models, including the preferential-attachment model of Barabási and Albert [23] and the copying models of Kleinberg et al. [104, 115], would yield a straight line in a log/log plot. For the LiveJournal network, the indegree plot in Figure 2-1 is more linear than the corresponding outdegree plot, but both plots appear far more parabolic than linear; this shape is the signature of a lognormal distribution, and not a power-law distribution. (See Section 1.3.4 and also the survey of Mitzenmacher [128] for a discussion of the long history of the power-law-versus-lognormal debate in a number of different fields.) The plots in Figure 2-1 seem to support the claim that social networks have lognormal degree distributions, though this evidence is relatively weak support for that position.

This network also exhibits many important structural properties observed in other social networks (see, e.g., the descriptions by Newman [142] or Wasserman and Faust [173]), of which we will highlight one additional feature. The LiveJournal network has a giant component—that is, a strongly connected component of the graph comprising most of the nodes in the network (see Section 1.3.5)—that contains 384,507 people, or 77.6% of the network.

Finally, in the sequel, we will need to compute the geographic distance between two nodes in the network. Over the range of distances about which we will be concerned, treating the earth as flat can cause non-negligible errors in distance computations. Throughout, our distance calculations will be done under the assumption that the earth is a perfect sphere of radius 6367 kilometers, ignoring the planet's oblateness. For two people u and v in the network, we let $d(u, v)$ denote the geographic distance between them, as the crow flies.

2.3 Geographic Routing and Small-World Phenomena

In this section, we describe the results of running a simulated version of the Milgram small-world experiment on the LiveJournal social network, using only geographic information to choose the next message-holder in a chain. Our simulation should not be viewed as a replication of real-world experiments studying human behavior such as those of Milgram [126] or Dodds et al. [49], but

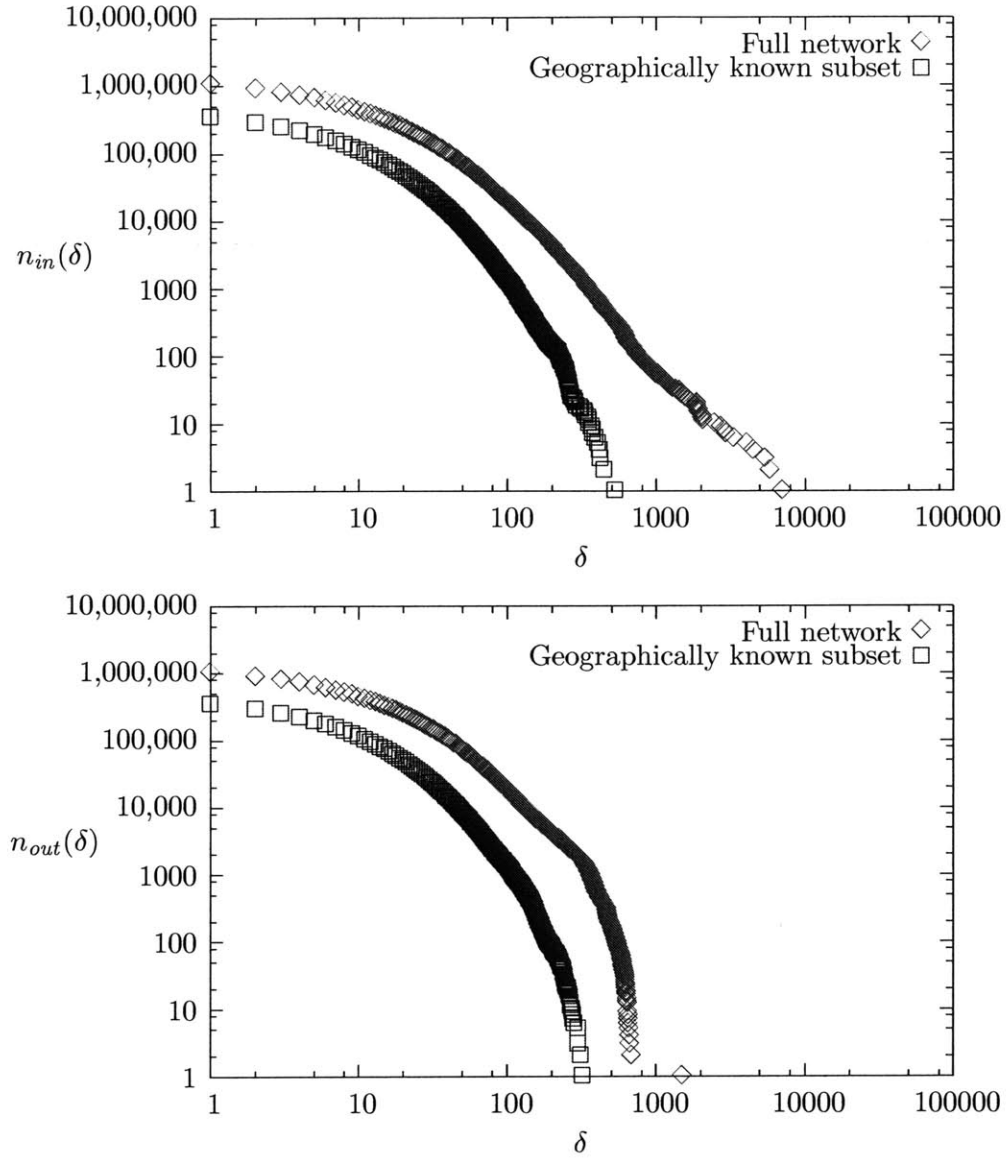


Figure 2-1: *Indegree and outdegree distributions in the LiveJournal social network. For a given degree δ, the plots give the numbers $n_{in}(\delta)$ (and $n_{out}(\delta)$, respectively) of people in the LiveJournal social network whose indegree (and outdegree, respectively) is at least δ. Plots are shown for both the full 1,300,000-node network and for the 500,000-node network of users who list locatable hometowns in the United States.*

rather as an investigation into what would be *possible* for people participating in a message-passing experiment in such a network. This simulation may be viewed as a thought experiment upon the network, with two goals.

First, we wish to tease apart the complexity of routing messages within a network from the voluntary participation of people in the network. The original Milgram experiment [126, 170] had only a hundred source subjects, and only eighteen chains reached the target stockbroker. The largest subsequent small-world experiment—performed over email by Dodds, Muhamad, and Watts [49]—considered just under 25,000 chains, of which under 400 completed. This low completion rate is consistent with the exponential decay of success probability that would result from each link in the chain abandoning the process with some moderate probability. However, the probability of a letter-holder abandoning a chain is plausibly positively correlated with the hop-distance to the target—a friend of the target will surely just complete the chain—and thus there is cause for concern that the experiment is biased towards completing much shorter paths than the true average length. (Kleinfeld discusses a number of reasons for doubting the conclusions of the Milgram experiment [106].) By simulating the message-passing experiment through the simulation of the transmission of a letter along social-network edges defined by human subjects (instead of relying on the human subjects to transmit the letter themselves), we can test purely graph-theoretic properties of the social network rather than testing the motivation and interest of the subjects in the study. The simulation-based experiment that we carry out here does not suffer from the same participation bias as in previous experiments—any social-network experiment in which people volunteer to be subjects may be biased towards having participants who think of themselves as "social"—but we do still face the issue that the LiveJournal world consists of a more uniform set of people than the world as a whole (in terms of age, socioeconomics, geography, and exhibitionism, for example), and thus that the LiveJournal world may be "smaller" than the real world.

Second, in our simulation we consider purely geographical information in choosing the next step in the chain—specifically using the *geographically greedy algorithm* [102, 105], which we denote by GeoGreedy. Suppose that a person u in the network is currently holding the message. If person u wishes to eventually reach a target individual t, then she considers her set of friends and chooses as the next step in the chain the person $v \in \{\text{friends of } u\}$ who is geographically closest to t. Stanley Milgram himself observed a remarkable geographic "zeroing in" on the target as chains progressed in his original experiment, and geography was observed to be an important dimension for forwarding in the recent replication of Dodds, Muhamad, and Watts. Simulation allows us to explore geography's freestanding power in these small-world results. Of course, the cost of a simulation-based approach is that real small-world experiments demonstrate something fascinating about what people actually do when faced with the challenge of routing to a target in social network, and their algorithm is obviously something different from—and more complicated than—the geographically greedy algorithm.

An additional motivation for our simulation is that, as described above, the previous largest experiment comprised 24,163 initial letter sources, of which 384 (only 1.6% of the chains) were completed. Our simulation study is at least an order of magnitude larger, both in the number of people and the number of completed chains.

The results of the simulated experiment are shown in Figure 2-2. When sources and targets are chosen randomly from the network, we find that the chain successfully reaches the city of the target individual in about 13% of the trials, with a mean chain length of slightly more than four, averaged over the completed chains. Our results are similar to the success rates from Milgram's

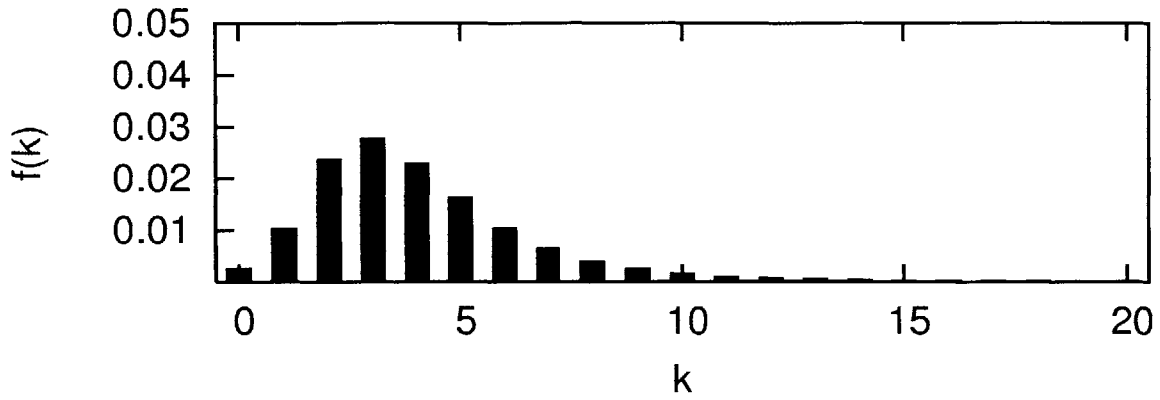


Figure 2-2: *Results of the geographically greedy algorithm on the LiveJournal social network. A source individual s and a target individual t are chosen randomly from the network; at each step the message is forwarded from the current message holder u to the friend v of u geographically closest to t. If $d(v,t) > d(u,t)$, then the chain is considered to have failed. We show the fraction $f(k)$ of chosen pairs in which the chain reaches the city where t lives in exactly k steps. Here 12.78% of chains completed, with median length 4, mean length $\mu = 4.12$, and standard deviation $\sigma = 2.54$ for completed chains. The results shown are averaged over 500,000 independent samples of source/target pairs.*

original experiment, where 18% of chains were completed (with an average length of just under six), and to the results of Dodds et al., where the average length of completed chains was 4.01 (and 1.6% of the chains were completed). This simulation therefore suggests that no information beyond geography is necessary for the discovery of many short chains in social networks.

Note that a chain terminates if the current message-holder u does not have any friends closer to the target than u herself. Recall that we choose a source and a target independently and uniformly at random from the network and then use the geographically greedy algorithm to try to reach the city of the target. Because we do not have any intra-city geographic information, we do not attempt to reach the target him or herself; without geography as a guide we cannot use this algorithm to route to the correct part of town. There is some debate in the sociology community about whether the geographic mechanisms that people use to route a letter to the correct city can also operate at the smaller scale within cities, or if other notions of proximity are required within cities. (We might hope to use some sort of induced metric on the interests that bloggers list in their profiles to guide the search for the target once the message has arrived in the target city.) For a target individual t chosen uniformly at random from the network, the average population of t's city is 1306, and always under 8000; we postulate that the same process that narrows the search from 500,000 users across the United States to 1300 in a city can further narrow the search to the target individual in a small number of additional steps, possibly by using more precise geographic locations or by other means [3, 79, 119, 162].

There is another reasonable model under which we can carry out this simulation, by treating "dead-end" nodes in a chain differently. One reason that a chain may fail to reach the target is that friendship lists of people in the LiveJournal network are incomplete, both in that people with friendship lists in their profiles may omit friends and also in that people may fail to include a friendship list in their profiles at all. Here is a breakdown of the nodes in the network based upon

whether their indegree δ_{in} and outdegree δ_{out} are nonzero:

	$\delta_{in} > 0$	$\delta_{in} = 0$	total
$\delta_{out} > 0$	398,011	19,828	417,839
$\delta_{out} = 0$	31,997	46,000	77,997
total	430,008	65,828	495,836

Thus, for a randomly chosen source s and target t, the probability that either $\delta_{out}(s) = 0$ or $\delta_{in}(t) = 0$ is $1 - 430{,}008 \cdot 417{,}839/495{,}836^2 \approx .269$. Thus over a quarter of all randomly chosen source/target pairs in this network are doomed from the start. Furthermore, about 6.5% of users (31,997) list no friends but are listed as the friend of at least one user; if these users are uniformly distributed, then a would-be chain of length six will hit one about a third of the time: the probability that one of the five would-be intermediary nodes is a dead end is $1 - (0.935)^5 \approx .285$. Together, these dead ends would cause around 55% of would-be chains of length six to terminate without reaching their target.

If a person has no close friends geographically nearer to the target, and has sufficient incentive to complete the chain, then in a real message-passing experiment she might pass the message along to a more distant acquaintance—her next-door neighbor, say, or some other "weak" friend to whom she is not sufficiently emotionally close to list in her LiveJournal profile—hoping that this weak friend could then complete the chain. To model this phenomenon, we also run the simulation with the following modification: if the current message-holder u has no friends closer to the target than u herself, then she randomly chooses another person v in her city and forwards the message to v. (If we had finer-grained data on bloggers' geographic locations, instead we would have u forward the message to her closest geographic neighbor; lacking this data, we use a random person in the city as a proxy for her neighbor.) The chain now fails only if there are no other people in the current message-holder's city who have not already held the message.

The results of this simulation are shown in Figure 2-3, and we show the proportion of successful paths compared to the number of "jump" links used in the chain in Figure 2-4. In this setting, we find that over 80% of chains reach the target, with an average path length of around seventeen and a median path length of twelve. This large increase in the fraction of completed paths supports claims of the importance of "weak links" in tying together a social network [49, 73] and suggests that failed chains are largely attributable to lack of incentive, rather than to an absence of navigability [106].

2.4 Geographic Friendships

In Section 2.3, we established that the geographically greedy algorithm executed on the LiveJournal social network under a restrictive model of friendship allows 13% of paths to reach their destination city in an average of about four steps, which is comparable to the end-to-end completion rates of earlier experiments on real human subjects [49, 126]. Because such a restrictive routing scheme enjoys a high success rate, a question naturally arises: is there some special structure relating friendship and geography that might explain this finding?

In Figure 2-5(a), we examine the relationship between friendship and geographic distance in the LiveJournal network. For each distance δ, we display the proportion $P(\delta)$ of pairs $\langle u, v \rangle$ separated by distance $\delta = d(u, v)$ who are friends. As δ increases, we observe that $P(\delta)$ decreases, indicating that geographic proximity increases the probability that two people are friends. (We note that

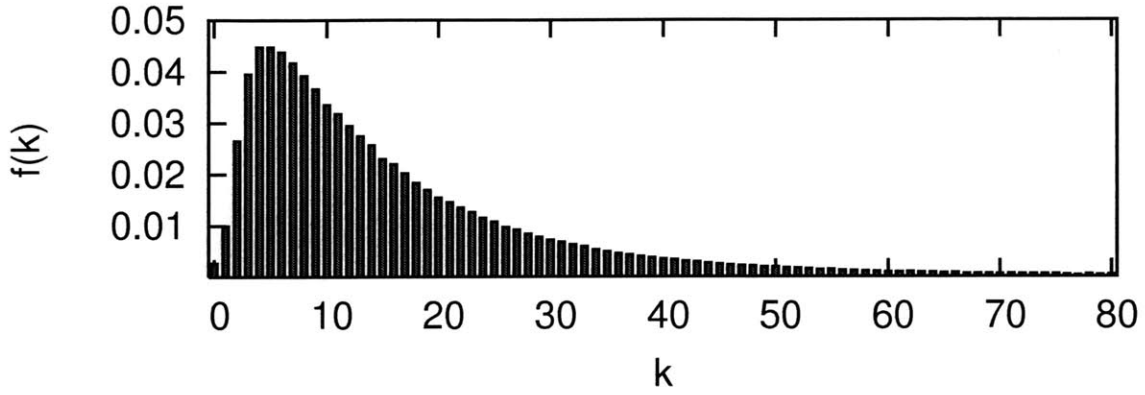


Figure 2-3: *Results of the geographically greedy routing algorithm on the LiveJournal social network with "weak-link" forwarding. A source s and a target t are chosen randomly from the network; at each step the message is forwarded from the current message-holder u to the friend v of u geographically closest to t. Here, if $d(v,t) > d(u,t)$ then u picks a random person in the same city as u as the next recipient of the message. The chain fails only if there is no such person available. As before, we show the fraction $f(k)$ of chosen pairs in which the chain reaches t's city in exactly k steps. We find that 80.16% chains completed, with a median of 12, a mean $\mu = 16.74$, and standard deviation $\sigma = 17.84$ for completed chains. The results shown are averaged over 500,000 independent samples of source/target pairs.*

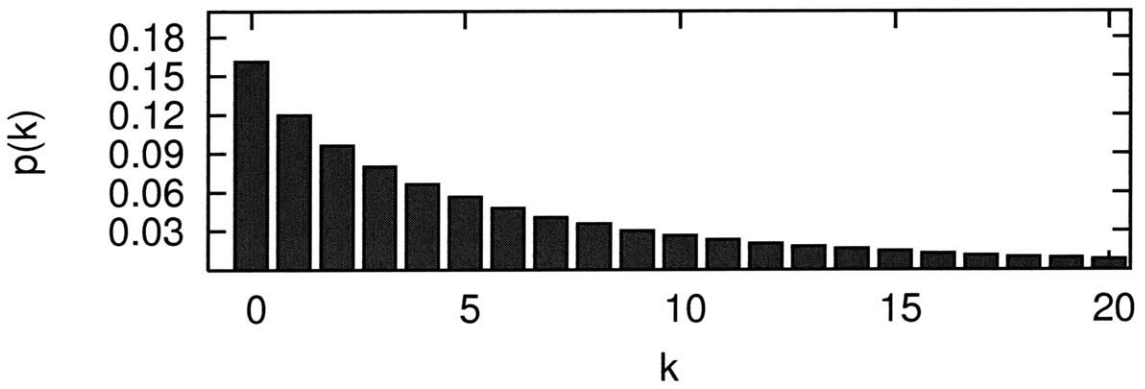
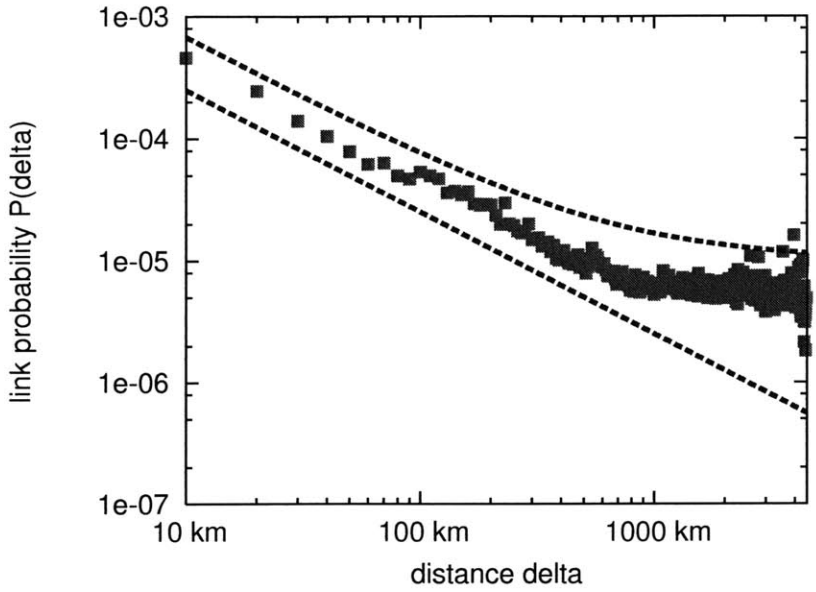
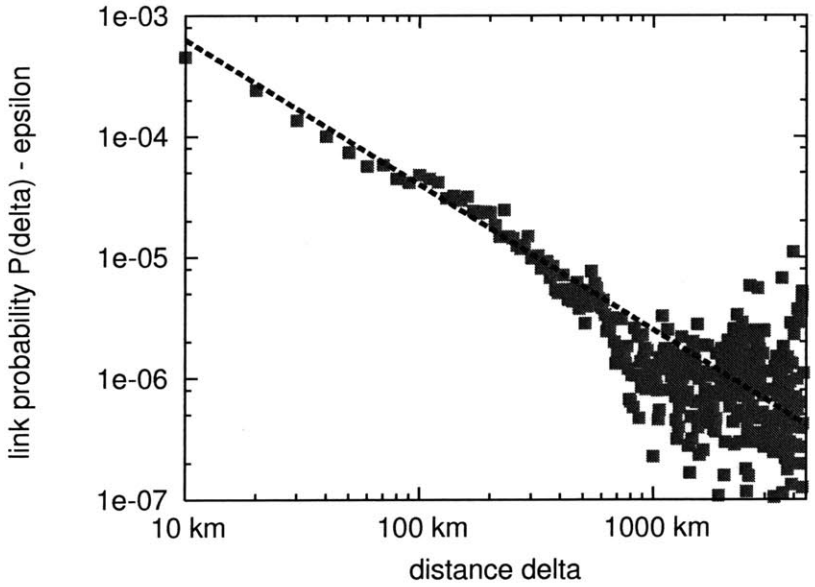


Figure 2-4: *Number of "weak links" used in the geographically greedy routing algorithm on the LiveJournal social network with "weak-link" forwarding reported in Figure 2-3. We consider each of the 80.16% of the chains that successfully connect their randomly chosen source and target. For each $k > 0$, we show the proportion $p(k)$ of these successful chains that use exactly k random jumps within the city of the current message-holder.*



(a) The fraction of friends among all pairs of people separated by a geographic distance of exactly δ in the LiveJournal network. The dotted lines correspond to $P(\delta) \propto 1/\delta$ and $P(\delta) \propto \varepsilon + 1/\delta$.



(b) The fraction of pairs of people separated by distance exactly δ who are friends, less $\varepsilon = 5 \times 10^{-6}$. The dotted line corresponds to $P(\delta) \propto 1/\delta^{1.2}$.

Figure 2-5: *The relationship between the probability that u lists v as a friend and the geographic distance $d(u, v)$ between u and v. For distance δ, we plot the proportion $P(\delta)$ of friendships among all pairs u, v in the LiveJournal social network with $d(u, v) = \delta$. Distances are rounded down to multiples of ten kilometers. The number of pairs $\langle u, v \rangle$ with $d(u, v) = d$ is estimated by computing the distance between 10,000 randomly chosen pairs of people in the network.*

this relationship holds true even in the virtual LiveJournal community; at first blush, geographic location might have very little to do with the identity of a person's online friends, but this plot verifies that geography remains crucial in online friendship. While it has been suggested that the impact of distance is marginalized by communications technology [38], there is a large body of research showing that proximity remains a critical factor in effective collaboration, and that the negative impacts of distance on productivity are only partially mitigated by technology [97].) However, for distances larger than about 800 to 1000 kilometers, we find that the δ-versus-$P(\delta)$ curve approximately flattens to a constant probability of friendship between people, regardless of the geographic distance between them.

The shape of the curve in Figure 2-5(a) can be explained by postulating a *background probability* ε of friendship that is independent of geography, so that the probability that two people who are separated by a geographic distance δ are friends is modeled as $P(\delta) = \varepsilon + f(\delta)$, for some constant $\varepsilon > 0$ and some function $f(\delta)$ that varies as the distance δ changes. That is, we may model friendship creation by the union of two distinct processes, one that includes all geography-dependent friendship-creation mechanisms (like meeting while at work in the same office building or while playing on opposing intramural ultimate-frisbee teams), and the other that includes non-geographic friendship-creation processes (like meeting online through a shared interest). A model for "geographic" friendships should incorporate the observation that as the distance δ increases, the probability $f(\delta)$ of geographic friendship should decrease. Notice that such a model will still account for some friendships between people who are geographically far apart, so we cannot simply equate "geographic" friendships with "nearby" friendships. Figure 2-5(a) shows that $P(\delta)$ "flattens out" for large distances: that is, the background friendship probability ε dominates $f(\delta)$ for large separating distances δ. In our data, the baseline probability of friendship is estimated as $\varepsilon \approx 5.0 \times 10^{-6}$ from Figure 2-5(a). Thus, an average person in the LiveJournal network has eight friends, of whom $500{,}000 \cdot \varepsilon \approx 2.5$ are formed by non-geographic processes, and the remaining 5.5 friendships are formed by geographic processes.

Because the non-geographic friendship-formation processes are by definition independent of geography, we can remove them from our plot to reveal only the geographic friendships. In Figure 2-5(b), we show the plot of geographic distance δ versus the probability $f(\delta) = P(\delta) - \varepsilon$ of a geographic friendship between people separated by distance δ. The plot of geographic-friendship probability as a function of geographic distance shows that $f(\delta)$ decreases smoothly as δ increases. Our computed value of ε implies that just over two-thirds of the friendships in the network are generated by geographic processes. Of course, the on-average 2.5 non-geographic friends may represent the realization of deep and complex mechanisms in their own right and may themselves explain small-world phenomena and other important properties of social networks. Here, though, we show that they are not required to give such an account: in succeeding sections, we will use only the on-average 5.5 geographic links per LiveJournal user to give a sufficient explanation of the navigable small-world phenomenon.

2.5 Geographic Models for Small Worlds

The experiments described in Section 2.3 establish that geography alone is sufficient to account for much of the small-world phenomenon, modulo the fact that our data limit us to routing to the city of the target, not to the target him or herself. Furthermore, the geographically greedy algorithm suffices as a local-information algorithm allowing people to navigate the social network efficiently.

We have shown that approximately two-thirds of the friendships in the LiveJournal network are geographic, and thus that geographic processes are a dominant aspect of friendship, even in an online social network. In this section, we examine geographically motivated friendship models to attempt to account for these empirical results.

The natural starting points for this investigation are the recently developed models of Watts and Strogatz [177] and Kleinberg [102, 105], discussed in Sections 1.2.2 and 1.2.3, and the model of Watts, Dodds, and Newman [176]. Although the Watts/Strogatz model produces social networks with short connecting chains, the goal of this model was only to explain the existence of short paths between pairs of people, and it does not attempt to give an explanation of why social networks are navigable. The model of Watts et al. [176] accounts for navigability by assigning individuals to locations in multiple hierarchical dimensions; two individuals are socially similar if they are nearby in any dimension. The authors give examples of geography and occupation as dimensions, so their model may be viewed as an enclosing framework for our geography-specific results. However, the generality of their framework does not specifically treat geographic aspects and leaves two open areas that we address. First, while interests or occupations might be naturally hierarchical, geography is far more naturally expressed in two-dimensional Euclidean space—embedding geographic proximity into a tree hierarchy is not possible without significant distortion [15]. Second, while they provide a detailed simulation-based evaluation of their model, there is no theorem in their work, nor any direct empirical comparison to a real social network.

Because we here are interested in explaining the navigability of the LiveJournal social network via the geographically greedy algorithm, we will concentrate on the work of Kleinberg. We briefly review here his social-network model, which is described in detail in Section 1.2.3. The model is based on a k-dimensional mesh of people, where each person knows his immediate geographic neighbors in every cardinal direction, and the probability of a long-distance link from u to v is proportional to $1/d(u,v)^\alpha$, for a constant $\alpha \geq 0$. Kleinberg's characterization theorem shows that short paths (that is, paths of length polylogarithmic in the size of the network) can be discovered in these networks by GeoGreedy if $\alpha = k$; more surprisingly, he proved that this is the only value of α for which these networks are navigable.

In our initial experimentation with this data set, we set out to verify the distance-versus-link-probability relationship from Kleinberg's results that we expect to find in the LiveJournal social network, given the demonstrated navigability shown in Section 2.3. Specifically, because the earth's surface is two dimensional, we anticipated that we would find that the friendship probability between two people in the network would fall off as the inverse square of the geographic distance between them, which would suffice to explain the fact that GeoGreedy was able to find short paths. Figure 2-5(a) displays the results of this experiment, showing the best fit for the probability of friendship as a function of distance over all the data in the network. More specifically, the experiment was performed as follows:

- 10,000 pairs of people in the LiveJournal social network were chosen uniformly at random to estimate the number of pairs of people in the network who are separated by a geographic distance of exactly δ kilometers.

- For each friendship link between people u and v in the network, we compute the distance $\delta = d(u,v)$ between them, and increment the count of friends at distance δ.

- Throughout the experiment, we round distances down to multiples of ten kilometers.

Examining Figure 2-5, one sees that the hypothesis that $\alpha \approx 2$ is highly unsupported. The troubling aspect of this experiment is that the friendship probability is best modeled as $1/d^{\alpha}$ for $\alpha \approx 1$, an exponent that cannot result in a navigable social network based on a two-dimensional mesh, according to Kleinberg's theorem. Yet the LiveJournal network is clearly navigable, as shown in our simulation of the geographically greedy routing algorithm.

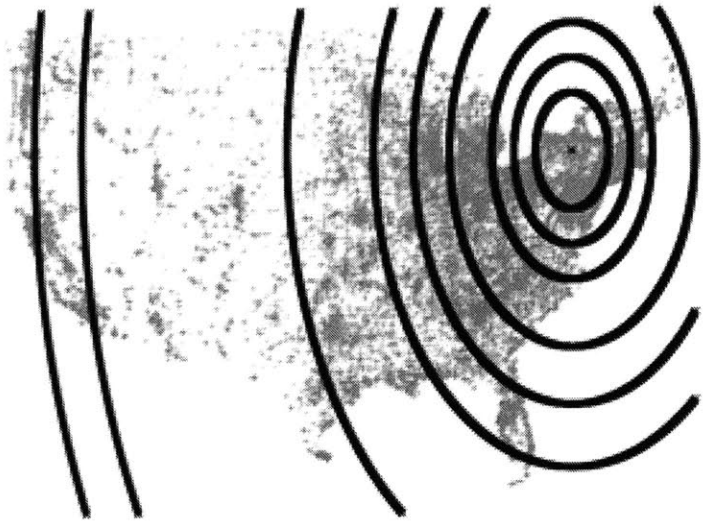
This seeming contradiction is explained by a large variance in population density across the LiveJournal social network, which is thus poorly approximated by the uniform two-dimensional mesh of Kleinberg's model. Figure 2-6 explores issues related to population patterns in more detail. Figure 2-6(a) shows concentric circles representing bands of equal population centered on Ithaca, New York. Under uniform population density, the width of each band should shrink as the distance from Ithaca increases. In the LiveJournal dataset, however, the distance between annuli actually gets larger instead of smaller. For other evidence of nonuniformity, note that purely distance-based predictions imply that that the probability of a friendship at a given distance should be constant for different people in the network. Figure 2-6(b) explores this concern, showing a distinction between East Coast (the states from Maine to Virginia) residents and West Coast (the states from Washington to California) residents in terms of probability of friendship as a function of distance. The East Coast plot is more like the national average, showing that the link probability for people separated by a geographical distance of δ is proportional to $1/\delta^{\alpha}$ for $\alpha \approx 1$, but the West Coast shows a link probability proportional to $1/\delta^{\alpha}$ for $\alpha \approx 0.5$. Thus an accurate geographic model of friendship for the LiveJournal network must be based upon something more than the simple geographic distance between two people; no uniform description of friendship as a function of distance will accurately predict the behavior in all regions of the network.

To summarize, we have shown that any model of friendship that is based solely on the distance between people is insufficient to explain the geographic nature of friendships in the LiveJournal network. In particular, current models [105, 177] do not take into account the organization of people into cities of arbitrary location and population density and thus cannot explain the success of the message-passing experiment. We therefore seek a network model that reconciles the linkage patterns in real-world networks with the success of the geographically greedy routing algorithm on these networks. Such a model must be based on something more than distance alone.

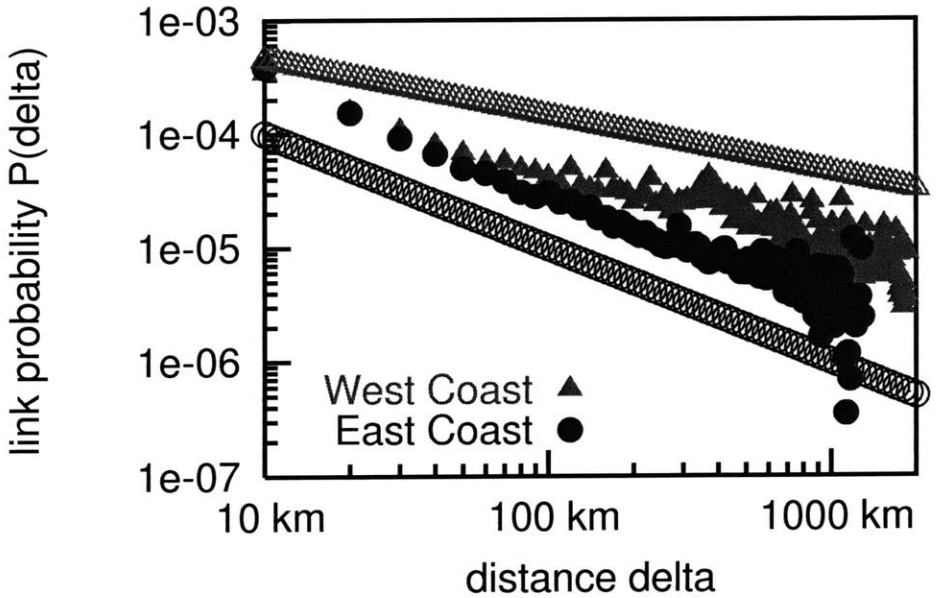
2.6 Population Networks and Rank-Based Friendship

We explore the idea that a simple function of probability of friendship that combines distance and density may apply uniformly over the network. Consider, for example, a person A and another person B who lives 500 meters away from A. In rural Minnesota, A and B are probably next-door neighbors, and very likely to know each other; in Manhattan, there may be tens of thousands of people who live closer to A than B does, and the two are unlikely to have ever even met. This discrepancy suggests why geographic distance alone is insufficient as a basis for a geographical model of social-network connections. Instead, our model uses as the key geographic notion the idea of *rank*: when examining a friend B of A, the relevant quantity is the number of people who live closer to A than B does. (There are some related notions in the literature on geometric data structures [46, 86].)

We model the probability that A and B are geographic friends as inversely proportional to B's rank with respect to A. Under this *rank-based friendship* model, the probability of a link from A to B depends only on the number of people within distance $d(A, B)$ of A and not directly on the



(a) A dot is shown for every distinct longitude/latitude pair home to at least one blogger in the LiveJournal network. We show concentric circles centered on Ithaca, New York so that the population of each successive circle increases by 50,000 people. Notice that the gap between the 350,000- and 400,000-person circles encompasses almost the entire Western United States.



(b) The same relationship as in Figure 2-5, restricted to people living on the West Coast (California, Oregon, and Washington) and the East Coast (from Virginia to Maine). The dotted lines correspond to $P(d) \propto 1/d$ and $P(d) \propto 1/d^{0.5}$.

Figure 2-6: *Evidence of the nonuniformity of population density in the LiveJournal network.*

geographic distance itself; thus the high degree of nonuniformity in population density that we observe in the LiveJournal data fits naturally into this framework. While either distance- or rank-based models may be appropriate in some contexts, we will show in this section and in Section 2.7 that (1) analytically, rank-based friendship implies that the geographically greedy routing scheme will find short paths in any social network; and (2) empirically, the LiveJournal social network exhibits rank-based friendship. First, in this section, we introduce a formal model for geographic social networks and formally define rank-based friendship. We then prove the main theorem of this chapter, showing that rank-based friendship is a sufficient explanation of the navigability of any grid-based geographic social network that adheres to it.

2.6.1 A Model of Population Networks

There are two key features that we wish to incorporate into our social-network model: *geography* and *population density*. We will first describe a very general abstract model for social networks that supports this pair of features; in later subsections we examine a concrete grid-based instantiation of this model.

Definition 2.6.1 (Population network) *A* population network *is a quintuple* $\langle L, d, P, \mathsf{loc}, E\rangle$ *where*

- L *is a finite set of* locations *$(\ell, s, t, x, y, z, \ldots)$;*
- $d : L \times L \longrightarrow \mathbb{R}^+$ *is an arbitrary* distance function *on the locations;*
- P *is a finite ordered set of* people *$(u, v, w, \ldots)$;*
- $\mathsf{loc} : P \longrightarrow L$ *is the* location function, *which maps people to the location in which they live; and*
- $E \subseteq P \times P$ *is the set of* friendships *between people in the network.*

The ordering on people is required only to break ties when comparing distances between two people. Let $\mathscr{P}(L)$ denote the power set of L.

Let $\mathsf{pop} : L \longrightarrow \mathbb{Z}^+$ denote the *population* of each point in L, where $\mathsf{pop}(\ell) := |\{u \in P : \mathsf{loc}(u) = \ell\}|$. We overload notation and let $\mathsf{pop} : \mathscr{P}(L) \longrightarrow \mathbb{Z}^+$ denote the population of a subset of the locations, so that $\mathsf{pop}(L') := \sum_{\ell \in L'} \mathsf{pop}(\ell)$. We will write $n := \mathsf{pop}(L) = |P|$ to denote the total population and $m := |L|$ to denote the total number of locations in the network.

Let $\mathsf{density} : L \longrightarrow [0, 1]$ be a probability distribution denoting the *population density* of each location $\ell \in L$, so that $\mathsf{density}(\ell) := \mathsf{pop}(\ell)/n$. We similarly extend $\mathsf{density} : \mathscr{P}(L) \longrightarrow [0, 1]$ so that $\mathsf{density}(L') = \sum_{\ell \in L'} \mathsf{density}(\ell)$. Thus $\mathsf{density}(L) = 1$.

We extend the distance function to accept both locations and people in its arguments, so that we have the function $d : (P \cup L) \times (P \cup L) \longrightarrow \mathbb{R}^+$ where $d(u, \cdot) := d(\mathsf{loc}(u), \cdot)$ and $d(\cdot, v) := d(\cdot, \mathsf{loc}(v))$ for all people $u, v \in P$.

When comparing the distances between people, we will use the ordering on people to break ties. For people $u, v, v' \in P$, we will write $d(u, v) < d(u, v')$ as shorthand for $\langle d(u, v), v\rangle \prec_{lex} \langle d(u, v'), v'\rangle$, where $\prec_{lex}$ denotes the standard lexicographic ordering on pairs. This tie-breaking role is the only purpose of the ordering on people.

2.6.2 Rank-Based Friendship

Following the navigable-small-world model of Kleinberg [102, 105], each person in the network will be endowed with exactly one *long-range link*. (Note that under Kleinberg's model, people can have more than one long-range link, but when the number of long-range links per person is $\Theta(1)$, the network behaves qualitatively as if there were exactly one long-range link per node. Thus we restrict our attention to the one-link case in the following.) We diverge from the model of Kleinberg in the definition of our long-range links. Instead of distance, the fundamental quantity upon which we base our model of long-range links is *rank*:

Definition 2.6.2 (Rank) *For two people $u, v \in P$, the* rank of v with respect to u *is defined as*

$$\mathsf{rank}_u(v) := |\{w \in P : d(u,w) < d(u,v)\}|.$$

Note that because we break ties in distance consistently according to the ordering on the people in the set P, we have the following: for any $i \in \{1, \ldots, n\}$ and any person $u \in P$, there is exactly one person v such that $\mathsf{rank}_u(v) = i$. We now define a model for generating a *rank-based* social network using this notion:

Definition 2.6.3 (Rank-based friendship) *For each person u in the network, we generate one long-range link from u, where*

$$\Pr[u \text{ links to } v] \propto \frac{1}{\mathsf{rank}_u(v)}.$$

Intuitively, one justification for rank-based friendship is the following: in order to be befriended by a person u, person v will have to compete with all of the more "convenient" candidate friends for u, i.e., all people w who live closer to u than v does. Note that, for any person u, we have $\sum_v 1/\mathsf{rank}_u(v) = \sum_{i=1}^{n} 1/i = H_n$, the nth harmonic number. Therefore, by normalizing, we have

$$\Pr[u \text{ links to } v] = \frac{1}{H_n \cdot \mathsf{rank}_u(v)}. \tag{2.1}$$

(We would have quantitatively identical behavior if we allowed the number of long-range links from a person u to vary, by, for every node v, independently at random adding a long-range link from u to v with probability $1/H_n \cdot \mathsf{rank}_u(v)$, as in (2.1). This modification introduces variation in nodes' outdegree, but has no effect on our technical results.) Again, under rank-based friendship, the only geographic dependence of the probability of a link from u to v is on the number of people within distance $d(u,v)$ of u, and there is no direct dependence on the geographic distance itself. (Link probabilities also depend on the entire population size, due to the normalization factor.)

One important feature of this model is that it is *independent* of the dimensionality of the space in which people live. For example, in the k-dimensional grid with uniform population density and the L_1 distance on locations, we have that $|\{w : d(u,w) \leq \delta\}| \propto \delta^k$, so the probability that person u links to person v is proportional to $d(u,v)^{-k}$. That is, the rank of a person v with respect to a person u satisfies $\mathsf{rank}_u(v) \approx d(u,v)^k$. Thus, although our model has been defined without explicitly embedding the locations in a metric space, our rank-based formulation gives essentially the same long-distance link probabilities as Kleinberg's model for a uniform-population k-dimensional mesh.

2.6.3 Rank-Based Friendship on Meshes

In the following, our interest will lie in population networks that are formed from meshes with arbitrary population densities. Let $L := \{1, \ldots, q\}^k$ denote the points on the k-dimensional mesh, with length q on each side. (The restriction that the mesh must have the same length on each side is for simplicity of notation only; our results immediately generalize to the setting of a mesh of size $\langle q_1, \ldots, q_k \rangle$.) We write $x = \langle x_1, \ldots, x_k \rangle$ for a location $x \in L$. We will consider Manhattan distance (L_1 distance) on the mesh, so that $d(\langle x_1, \ldots, x_k \rangle, \langle y_1, \ldots, y_k \rangle) := \sum_{i=1}^{k} |x_i - y_i|$. The only restriction that we impose on the people in the network is that $\mathsf{pop}(\ell) > 0$ for every $\ell \in L$—that is, there are no ghost towns with zero population. This assumption will allow us to avoid the issue of disconnected sets in what follows. (The assumption of no ghost towns guarantees a property required for proof—for any current message-holder u and any target t, there must be a friend of u that guarantees progress towards t. Other less restrictive conditions may suffice as well, but the no-ghost-town condition the easiest conceptually, and is the easiest to formulate.)

Thus a *mesh population network* is fully specified by the dimensionality k, the side length q, the population P (with an ordering to break ties in interpersonal distances), the friendship set E, and the location function $\mathsf{loc} : P \longrightarrow \{1, \ldots, q\}^k$, where for every location $\ell \in \{1, \ldots, q\}^k$, there exists at least one person $u_\ell \in P$ such that $\mathsf{loc}(u_\ell) = \ell$.

Following Kleinberg's model of navigable small worlds, we include *local links* in E for each person in the network. For now, we assume that each person u at location $\ell^{(u)} = \mathsf{loc}(u)$ in the network has a local link to some person at the mesh point in each cardinal direction from $\ell^{(u)}$, i.e., to some person at each of the $2k$ points $\langle \ell_1^{(u)}, \ldots, \ell_{i-1}^{(u)}, \ell_i^{(u)} \pm 1, \ell_{i+1}^{(u)}, \ldots, \ell_k^{(u)} \rangle$, for any coordinate $i \in \{1, \ldots, k\}$. Thus, for any two people u and v, there exists a path of length at most qk between them, and, more specifically, the geographically greedy algorithm will find a path of length no longer than qk.

In a *rank-based mesh population network*, we add one long-range link to E per person in P, where that link is chosen probabilistically by rank, according to (2.1).

2.6.4 The Two-Dimensional Grid

For simplicity, we first consider the two-dimensional grid, where we have $L := \{1, \ldots, q\} \times \{1, \ldots, q\}$, and thus $m = |L| = q^2$. We may think of the two-dimensional grid as representing the intersection of integral lines of longitude and latitude, for example.

In this section, we will show that the geographically greedy algorithm on the two-dimensional grid produces paths that are on average very short—more precisely, that the expected length of the path found by the geographically greedy algorithm is bounded by $O(\log^3 n)$ when the target is chosen randomly from the population P. Formally, the geographically greedy algorithm GeoGreedy proceeds as follows: given a target t and a current message-holder u, person u examines her set of friends, and forwards the message to the friend v of u who is geographically closest to the target t. First, we need some definitions:

Definition 2.6.4 (L_1-Ball) *For any location $x \in L$ and for any radius $r \geq 0$, let*

$$B_r(x) = \{y \in L : d(x, y) \leq r\} = \{y \in L : |x_1 - y_1| + |x_2 - y_2| \leq r\}$$

denote the L_1-ball of radius r centered at location x.

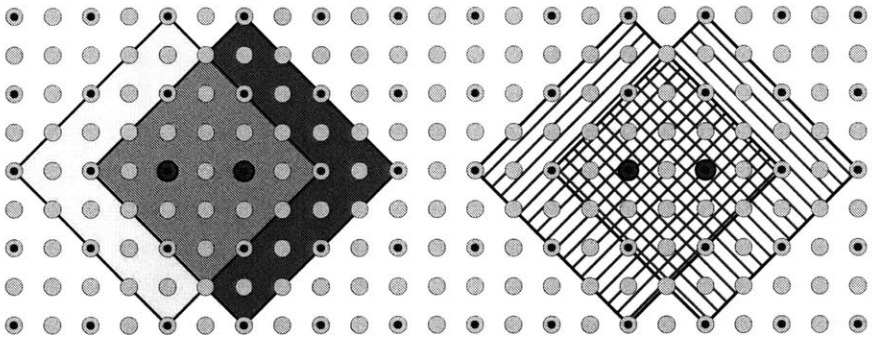


Figure 2-7: *A sketch of the radius-r balls covering the grid. Here we show the L_1 balls of radius 4 that are in the covering set formally defined in Definition 2.6.5, i.e., the radius-4 balls centered at the points in $\mathscr{C}_4 = \{z : 2z_i \bmod 4 = 0\} = \{z : z_i \text{ is even}\}$. The points of $\mathscr{C}_4$ are the gridpoints with a dark (blue) dot in them. Four such points are darkened, and the balls centered at those points are shown. On the left, the two balls are shown in yellow and red, with the overlap region shown in orange. For those viewing this figure in grayscale, the two balls on the right are shown with differing hatching patterns, and the overlap is crosshatched.*

Before we proceed to the technical details of the proof, we begin with a brief high-level (and oversimplified) outline of the proof and some of the intuition of the analysis. First, we claim that the expected number of hops taken by GeoGreedy before we reach a point halfway from the source to the target is $O(\log^2 n)$; after $O(\log n)$ such halvings, we will reach the target itself.

To establish the claim, we show that the probability p of reaching a "good" point—that is, one within distance $d(s,t)/2$ of the target—in any step of GeoGreedy is at least $p \geq \frac{\mathsf{density}(B_{d(s,t)/2}(t))}{\mathsf{density}(B_{4d(s,t)}(t)) \cdot H_n}$, by upper bounding the rank of the $\mathsf{pop}(B_{d(s,t)/2}(t))$ "good" nodes by the population of the larger ball. Then the expected number of nodes that GeoGreedy encounters before reaching a good node is $H_n \cdot \frac{\mathsf{density}(B_{4d(s,t)}(t))}{\mathsf{density}(B_{d(s,t)/2}(t))}$. We then approximate these balls by rounding the ball centers to regular gridpoints and rounding ball radii to powers of two. Taking expectations over the choice of t cancels the denominator of the above expression, and we then show that the resulting expression is $O(\log^2 n)$ because there are only logarithmically many different ball radii.

The remainder of this section is devoted to making the above intuition precise. We consider an exponentially growing set $\mathscr{R} := \{2^i : i \in \{1, 2, 4, \ldots, 128\lceil \log q \rceil\}\}$ of ball radii, and we place a series of increasingly fine-grained collections of balls that cover the grid:

Definition 2.6.5 (Covering Radius-r Ball Centers) *For any radius $r \in \mathscr{R}$, let the set*

$$\mathscr{C}_r := \{z \in L : 2z_1 \bmod r = 2z_2 \bmod r = 0\}$$

be the set of locations z such that z_i/r is half-integral for $i \in \{1, 2\}$.

For each radius $r \in \mathscr{R}$, we will consider the set of radius-r balls centered at each of the locations in $\mathscr{C}_r$. (See Figure 2-7.) We begin with a few simple facts about these L_1-balls:

Fact 2.6.6 (Only a small number of balls in $\mathscr{C}_r$ overlap) *For each radius $r \in \mathscr{R}$:*

1. *For each location $x \in L$, we have that $|\{z \in \mathscr{C}_r : d(z, x) \leq r\}| \leq 25$.*

2. *For each location* $z \in \mathscr{C}_r$, *we have that* $|\{z' \in \mathscr{C}_{r/2} : B_{r/2}(z') \cap B_r(z) \neq \emptyset\}| \leq 169$.

Proof. For the first claim, note that if $|z_1 - x_1| > r$ or if $|z_2 - x_2| > r$, then $d(z, x) > r$, and z is not an element of the set of relevance. Thus every $z \in \mathscr{C}_r$ such that $d(z, x) \leq r$ must fall into the range $\langle x_1 \pm r, x_2 \pm r \rangle$. There are at most five half-integral values of z/r that can fall into the range $[b, b + 2r]$ for any b, so there are at most twenty-five total points $z \in \mathscr{C}_r$ that satisfy $d(x, z) \leq r$.

For the second claim, notice that any ball of radius $r/2$ that has a nonempty intersection with $B_r(z)$ must have its center at a point z' such that $d(z, z') \leq 3r/2$. Thus the only $z' \in \mathscr{C}_{r/2}$ that could be in the set of relevance must have $z'_i \in [z_i - 3r/2, z_i + 3r/2]$ for $i \in \{1, 2\}$ and have $2z'_i/(r/2)$ be half-integral. As in the first claim, the number of half-integral values of $2z'/r$ that can fall into the range $[b, b + 3r]$ is at most thirteen for any b. Thus there can be at most 169 total points $z' \in \mathscr{C}_{r/2}$ so that $B_{r/2}(z') \cap B_r(z) \neq \emptyset$. □

Fact 2.6.7 (Relation between balls centered in L and in $\mathscr{C}_r$) *For each location* $x \in L$ *and for each radius* $r \in \mathscr{R}$*:*

1. *There exists a location* $z \in \mathscr{C}_r$ *such that* $B_{r/2}(x) \subseteq B_r(z)$.
2. *There exists a location* $z' \in \mathscr{C}_{r/2}$ *such that* $B_{r/2}(z') \subseteq B_r(x)$ *and* $x \in B_{r/2}(z')$.

Proof. For the first claim, let $z \in \mathscr{C}_r$ be the closest point to x in $\mathscr{C}_r$. Note that $x_1 \in [z_1 - r/4, z_1 + r/4]$; otherwise x would be strictly closer to either $\langle z_1 - r/2, z_2 \rangle \in \mathscr{C}_r$ or $\langle z_1 + r/2, z_2 \rangle \in \mathscr{C}_r$. Similarly we have $x_2 \in [z_2 - r/4, z_2 + r/4]$. Therefore we have $d(x, z) = \sum_{i \in \{1,2\}} |x_i - z_i| \leq r/2$. Let $y \in B_{r/2}(x)$ be arbitrary. Then by the triangle inequality we have $d(z, y) \leq d(z, x) + d(x, y) \leq r/2 + r/2 = r$. Thus we have $y \in B_r(z)$, which proves the claim.

For the second claim, let $z' \in \mathscr{C}_{r/2}$ be the closest point to x in $\mathscr{C}_{r/2}$. By the same argument as above, we have $d(x, z') \leq r/4$. Immediately we have $x \in B_{r/2}(z')$. Let $y \in B_{r/2}(z')$ be arbitrary. Then $d(x, y) \leq d(x, z') + d(z', y) \leq r/4 + r/2 < r$, and $y \in B_r(x)$, which proves the claim. □

Let x and y be two arbitrary locations in L. In what follows, we will use the size of the smallest ball in $\bigcup_{r \in \mathscr{R}} \{B_r(z) : z \in \mathscr{C}_r\}$ that includes both x and y as a ceiling-like proxy for $d(x, y)$, and as the measure of progress towards the target. We will also need a large ball from $\{B_r(z) : z \in \mathscr{C}_r\}$ that includes both x and y and also includes a large ball centered at y.

Definition 2.6.8 (Minimum enclosing-ball radius) *For two arbitrary locations* $x, y \in L$, *let* $\mathsf{mebr}(x, y)$ *("minimum enclosing-ball radius") denote the minimum* $r \in \mathscr{R}$ *such that, for some* $z \in \mathscr{C}_r$, *we have* $x, y \in B_r(z)$.

Fact 2.6.9 (Relating distance and minimum enclosing-ball radius) *For any* $x, y \in L$, *let* $r := \mathsf{mebr}(x, y)$. *Then we have* $2r \geq d(x, y) \geq r/4$.

Proof. Let $z \in \mathscr{C}_r$ be such that $x, y \in B_r(z)$, and note that, by definition, there is no $z' \in \mathscr{C}_{r/2}$ such that $x, y \in B_r(z')$. The first direction is easy: by the triangle inequality, we have that $d(x, y) \leq d(x, z) + d(z, y) \leq r + r = 2r$. For the other direction, suppose for a contradiction that $d(x, y) \leq r/4$. Let $z^* \in \mathscr{C}_{r/2}$ be such that $B_{r/4}(x) \subseteq B_{r/2}(z^*)$, as guaranteed by Fact 2.6.7.1. But then we have $x, y \in B_{r/4}(x)$ because $d(x, y) \leq r/4$, which implies that $x, y \in B_{r/2}(z^*)$, which in turn contradicts the minimality of r. □

Thus, in the path from any source $s \in L$ to any target $t \in L$ found by GeoGreedy, the path will always remain inside the ball $B_{d(s,t)}(t) \subseteq B_{2 \cdot \mathsf{mebr}(s,t)}(t)$.

Definition 2.6.10 (Sixteenfold enclosing ball) *Let $x, y \in L$ be an arbitrary pair of locations, and let $r = \mathsf{mebr}(x, y)$. Let $\mathsf{sebc}(y, r)$ ("sixteenfold-enclosing-ball center") denote the location $z^*_{y,r} \in \mathscr{C}_{16r}$ such that $B_{8r}(y) \subseteq B_{16r}(z^*_{y,r})$ whose existence is guaranteed by Fact 2.6.7.1.*

Lemma 2.6.11 (Relationship between ball population and rank) *Let $s, t \in L$ be an arbitrary source/target pair of locations. Let $r = \mathsf{mebr}(s, t)$, and let $z^* = \mathsf{sebc}(t, r)$. Let $x, y \in L$ be arbitrary locations such that $x \in B_{2r}(t)$ and $y \in B_{r/8}(t)$, and let $u, v \in P$ be arbitrary people such that $\mathsf{loc}(u) = x$ and $\mathsf{loc}(v) = y$. Then $\mathsf{rank}_u(v) \leq \mathsf{pop}(B_{16r}(z^*))$.*

Proof. First, we note

$$\begin{array}{rcll} d(x,y) & \leq & d(x,t) + d(t,y) & \text{triangle inequality} \\ & \leq & 2r + r/8 & \text{assumptions that } x \in B_{2r}(t) \text{ and } y \in B_{r/8}(t) \\ & = & 17r/8. & \end{array} \tag{2.2}$$

We now claim the following:

$$\text{for any location } \ell \in L, \text{ if } d(x,\ell) \leq d(x,y), \text{ then } d(z^*,\ell) \leq 16r. \tag{2.3}$$

To prove (2.3), let ℓ be an arbitrary location so that $d(x,\ell) \leq d(x,y)$. Then we have

$$\begin{array}{rcll} d(t,\ell) & \leq & d(t,y) + d(y,x) + d(x,\ell) & \text{triangle inequality} \\ & \leq & d(t,y) + d(y,x) + d(x,y) & \text{assumption that } d(x,\ell) \leq d(x,y) \\ & \leq & r/8 + d(y,x) + d(x,y) & \text{assumption that } y \in B_{r/8}(t) \\ & \leq & r/8 + 17r/8 + 17r/8 & (2.2) \\ & = & 35r/8. & \end{array}$$

Then we have that $\ell \in B_{35r/8}(t) \subseteq B_{8r}(t) \subseteq B_{16r}(z^*)$ by the definition of $z^* = \mathsf{sebc}(t, r)$, which proves (2.3). Now, by definition of rank, we have that

$$\begin{array}{rcl} \mathsf{rank}_u(v) & \leq & |\{w \in P : d(u,w) \leq d(u,v)\}| \\ & = & \sum_{\ell \in L : d(x,\ell) \leq d(x,y)} \mathsf{pop}(\ell) \\ & = & \mathsf{pop}(\{\ell \in L : d(x,\ell) \leq d(x,y)\}) \\ & \leq & \mathsf{pop}(\{\ell \in L : d(\ell,z^*) \leq 16r\}) \\ & = & \mathsf{pop}(B_{16r}(z^*)) \end{array}$$

where the second inequality follows from (2.3). □

We are now ready to prove the main technical result of this section, namely that the geographically greedy algorithm will halve the distance from the source to the target in an expected number of steps that is polylogarithmic in the population size, for a randomly chosen target person.

Lemma 2.6.12 (GeoGreedy halves distance in expected polylogarithmic steps) *Let $s \in L$ be an arbitrary source location, and let $t \in L$ be a target location randomly chosen according to the distribution $\mathsf{density}(\cdot)$. Then the expected number of steps before the geographically greedy algorithm started from location s reaches a point in $B_{d(s,t)/2}(t)$ is $O(\log n \log m) = O(\log^2 n)$, where the expectation is taken over the random choice of t.*

Proof. Let $r_t := \mathsf{mebr}(s,t)$, and let $z_t := \mathsf{sebc}(t, r_t)$ so that

$$z_t \in \mathscr{C}_{16r_t} \text{ and } B_{8r_t}(t) \subseteq B_{16r_t}(z_t). \tag{2.4}$$

Let z'_t be the location whose existence is guaranteed by Fact 2.6.7.2 such that

$$z'_t \in \mathscr{C}_{r_t/16} \text{ and } B_{r_t/16}(z'_t) \subseteq B_{r_t/8}(t) \text{ and } t \in B_{r_t/16}(z'_t). \tag{2.5}$$

Putting together (2.4) and (2.5), we have the following two facts:

$$B_{r_t/16}(z'_t) \subseteq B_{r_t/8}(t) \subseteq B_{8r_t}(t) \subseteq B_{16r_t}(z_t) \tag{2.6}$$

$$t \in B_{r_t/16}(z'_t). \tag{2.7}$$

By Fact 2.6.9, we know that $d(s,t)/2 \geq r_t/8$. Thus it will suffice to show that the expected number of steps before GeoGreedy started from location s lands in $B_{r_t/8}(t) \subseteq B_{d(s,t)/2}(t)$ is $O(\log n \log m)$.

Suppose that we start GeoGreedy at the source s, and the current point on the path found by the algorithm is some person $u \in P$ at location $x_u = \mathsf{loc}(u)$. By definition, every step of GeoGreedy decreases the distance from the current location to the target t, so we have that

$$d(x_u, t) \leq d(s,t) \leq 2r_t. \tag{2.8}$$

We refer to a person u as *good* if there exists a long-range link from that person to any person living in the ball $B_{r_t/8}(t)$. Let $\alpha_{u,t}$ denote the probability that a person $u \in P$ living at location $x_u = \mathsf{loc}(u) \in L$ is good. Then

$$\alpha_{u,t} = \sum_{v:\mathsf{loc}(v) \in B_{r_t/8}(t)} \frac{1}{\mathsf{rank}_u(v) \cdot H_n} \geq \sum_{v:\mathsf{loc}(v) \in B_{r_t/8}(t)} \frac{1}{\mathsf{pop}(B_{16r_t}(z_t)) \cdot H_n} = \frac{\mathsf{pop}(B_{r_t/8}(t))}{\mathsf{pop}(B_{16r_t}(z_t)) \cdot H_n}$$

by the definition of good, by Lemma 2.6.11 (which applies by (2.8)), and by the definition of $\mathsf{pop}(\cdot)$. Noting that the lower bound on $\alpha_{u,t}$ is independent of u, we write

$$\alpha_t := \frac{\mathsf{pop}(B_{r_t/8}(t))}{\mathsf{pop}(B_{16r_t}(z_t)) \cdot H_n} \leq \alpha_{u,t}.$$

Thus the probability that u is good is at least α_t for every person u along the GeoGreedy path. Furthermore, each step of the algorithm brings us to a new node never before seen by the algorithm because the distance to t is strictly decreasing until we reach node t. Thus the probability of finding a good long-range link is independent at each step of the algorithm until it terminates. Therefore, the expected number of steps before we reach a good person (or t itself) is at most $1/\alpha_t$.

We now examine the expected value of $1/\alpha_t$ for a target location $t \in L$ chosen randomly according to the distribution $\mathsf{density}(\cdot)$:

$$\begin{aligned}
\mathbf{E}_{t \in L \sim \mathsf{density}(\cdot)}[1/\alpha_t] &= \sum_t \mathsf{density}(t) \cdot \frac{1}{\alpha_t} \\
&= \sum_t \mathsf{density}(t) \cdot \frac{\mathsf{pop}(B_{16r_t}(z_t)) \cdot H_n}{\mathsf{pop}(B_{r_t/8}(t))} \\
&= H_n \cdot \sum_t \mathsf{density}(t) \cdot \frac{\mathsf{density}(B_{16r_t}(z_t))}{\mathsf{density}(B_{r_t/8}(t))} \\
&\leq H_n \cdot \sum_t \mathsf{density}(t) \cdot \frac{\mathsf{density}(B_{16r_t}(z_t))}{\mathsf{density}(B_{r_t/16}(z'_t))}.
\end{aligned}$$

The equalities follow from the definition of expectation, the definition of α_t, and from the fact that $\mathsf{density}(\cdot) = \mathsf{pop}(\cdot)/n$. The inequality follows from the definition of z'_t in (2.5), using the fact that $B_{r_t/16}(z'_t) \subseteq B_{r_t/8}(t)$ and the monotonicity of $\mathsf{density}(\cdot)$.

We now reindex the summation to be over radii and ball centers rather than over targets t. Recall that $z_t \in \mathscr{C}_{16r_t}$ and $z'_t \in \mathscr{C}_{r_t/16}$, and that $B_{r_t/16}(z'_t) \subseteq B_{16r_t}(z_t)$ by (2.6), and therefore that $z'_t \in B_{16r_t}(z_t)$. Thus, we have that

$$\begin{aligned}
\mathrm{E}_{t\in L\sim\mathsf{density}(\cdot)}[1/\alpha_t] &\leq H_n \cdot \sum_t \mathsf{density}(t) \cdot \frac{\mathsf{density}(B_{16r_t}(z_t))}{\mathsf{density}(B_{r_t/16}(z'_t))} \\
&\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \sum_{z'\in\mathscr{C}_{r/16}:z'\in B_{16r}(z)} \frac{\mathsf{density}(B_{16r}(z))}{\mathsf{density}(B_{r/16}(z'))} \sum_{t:z'_t=z'} \mathsf{density}(t) \\
&\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \sum_{z'\in\mathscr{C}_{r/16}:z'\in B_{16r}(z)} \frac{\mathsf{density}(B_{16r}(z))}{\mathsf{density}(B_{r/16}(z'))} \sum_{t\in B_{r/16}(z')} \mathsf{density}(t)
\end{aligned}$$

where the last inequality follows from (2.7). But then

$$\begin{aligned}
\mathrm{E}_{t\in L\sim\mathsf{density}(\cdot)}[1/\alpha_t] &\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \sum_{z'\in\mathscr{C}_{r/16}:z'\in B_{16r}(z)} \frac{\mathsf{density}(B_{16r}(z))}{\mathsf{density}(B_{r/16}(z'))} \sum_{t\in B_{r/16}(z')} \mathsf{density}(t) \\
&= H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \sum_{z'\in\mathscr{C}_{r/16}:z'\in B_{16r}(z)} \frac{\mathsf{density}(B_{16r}(z))}{\mathsf{density}(B_{r/16}(z'))} \cdot \mathsf{density}(B_{r/16}(z')) \\
&= H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \sum_{z'\in\mathscr{C}_{r/16}:z'\in B_{16r}(z)} \mathsf{density}(B_{16r}(z)) \\
&= H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \mathsf{density}(B_{16r}(z)) \cdot |\{z' \in \mathscr{C}_{r/16} : z' \in B_{16r}(z)\}|.
\end{aligned}$$

Now we are almost done: by applying Fact 2.6.6.2 a constant number of times, we have that

$$|\{z' \in \mathscr{C}_{r/16} : z' \in B_{16r}(z)\}| = O(1). \tag{2.9}$$

Furthermore, we have $\sum_{z\in\mathscr{C}_{16r}} \mathsf{density}(B_{16r}(z)) \leq 25$: by Fact 2.6.6.1, there are at most twenty-five balls in $\mathscr{C}_r$ that include any particular location, so we are simply summing a probability distribution with some "double counting," but counting each point at most twenty-five times. Thus we have

$$\begin{aligned}
\mathrm{E}_{t\in L\sim\mathsf{density}(\cdot)}[1/\alpha_t] &\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \mathsf{density}(B_{16r}(z)) \cdot |\{z' \in \mathscr{C}_{r/16} : z' \in B_{16r}(z)\}| \\
&\leq H_n \cdot O(1) \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{C}_{16r}} \mathsf{density}(B_{16r}(z)) \\
&\leq H_n \cdot O(1) \cdot \sum_{r\in\mathscr{R}} 25 \\
&= H_n \cdot O(1) \cdot 25 \cdot |\mathscr{R}| = H_n \cdot O(\log q) = O(\log n \log m).
\end{aligned}$$

because $|\mathscr{R}| = \Theta(\log q) = \Theta(\log m)$. □

In the case of uniform population density, the value of $\alpha_{u,t} = \Omega(1/\log n)$ is independent of s and t, and the greedy algorithm finds an s-t path of length $O(\log^2 n)$ with high probability [102, 105].

Theorem 2.6.13 (GeoGreedy finds short paths in all 2-D meshes) *For any two-dimensional mesh population network with n people and m locations, the expected length of the search path found by* GeoGreedy *from an arbitrarily chosen source location s to the location t of a uniformly chosen target individual is $O(\log n \log^2 m) = O(\log^3 n)$.*

Proof. Immediate by inductive application of Lemma 2.6.11: the expected number of hops required before moving to a node s' with $d(s',t) \leq d(s,t)/2$ or t itself is $O(\log n \log m)$; by repeating this process $O(\log(\max_{s,t} d(s,t))) = O(\log qk) = O(\log q^k) = O(\log m)$ times, we must arrive at the target node t itself. □

2.6.5 The k-Dimensional Grid

In this section, we extend the results of the previous section to higher-dimensional mesh population networks. Consider a k-dimensional grid, with side length q, so that $L = \{1,\ldots,q\}^k$ and hence that $m = |L| = q^k$. Let $\kappa := 2^{\lceil \log k \rceil}$ denote the smallest exact power of two greater than k.

The proof in this section follows the same basic outline as the two-dimensional result in Section 2.6.4, but the results are affected by the dimensionality of the grid. Here, we use L_∞ balls (i.e., L_1 squares; we use the terms interchangeably) instead of L_1 balls in our covering set.

Definition 2.6.14 (L_∞-ball of radius r) *For any location $x \in L$ and for any radius $r \geq 0$, let*

$$B_r^\infty(x) = \{y \in L : \forall i \in \{1,\ldots,k\}, |x_i - y_i| \leq r\}$$

denote the L_∞-ball (L_1-square) of radius r centered at location x.

We cover the mesh with a series of increasingly fine-grained collections of squares:

Definition 2.6.15 (Centers of covering radius-r squares) *For any radius $r \in \mathscr{R}$, let the set*

$$\mathscr{S}_r := \{z \in L : \forall i \in \{1,\ldots,k\}, 2z_i \bmod r = 0\}$$

be the set of locations z such that z_i/r is half-integral for every coordinate $i \in \{1,\ldots,k\}$.

Analogues to the facts from Section 2.6.4 continue to hold here, with very similar proofs; hence in this section we give terser arguments for the analogous facts.

Fact 2.6.16 (Only a small number of squares in $\mathscr{S}_r$ overlap) *For each radius $r \in \mathscr{R}$:*

1. *For each location $x \in L$, we have that $|\{z \in \mathscr{S}_r : x \in B_r^\infty(z)\}| = 2^{O(k)}$.*
2. *For each $z \in \mathscr{S}_r$ and $\alpha \in \mathbb{Z}^+$, we have that $|\{z' \in \mathscr{S}_{r/2^\alpha} : B_{r/2^\alpha}^\infty(z') \cap B_r^\infty(z) \neq \emptyset\}| = 2^{O(\alpha k)}$.*

Proof. (Analogous to Fact 2.6.6.1 and 2.6.6.2.) There are $O(1)$ half-integral values of z_i/r in the range $[x_i - r, x_i + r]$, and thus at most $O(1)^k = 2^{O(k)}$ points in $\{z \in \mathscr{S}_r : x \in B_r^\infty(z)\}$. Similarly, for the second claim, note that any $z_i' \notin [x_i - 2r, x_i + 2r]$ cannot possibly have $B_{r/2^\alpha}^\infty(z') \cap B_r^\infty(z) \neq \emptyset$. In a range of width $4r$, the number of half-integral values of $z_i/(r/2^\alpha) = 2^\alpha \cdot z_i/r$ is $O(2^\alpha)$. Thus there are at most $O(2^\alpha)^k = 2^{O(\alpha k)}$ points in $\{z' \in \mathscr{S}_{r/2^\alpha} : B_{r/2^\alpha}^\infty(z') \cap B_r^\infty(z) \neq \emptyset\}$. □

Fact 2.6.17 (Relation of balls centered in L and squares centered in $\mathscr{S}_r$) *For each $x \in L$ and for each radius $r \in \mathscr{R}$:*

1. *There exists a location $z \in \mathscr{S}_r$ such that $B_{r/2}(x) \subseteq B^\infty_r(z)$.*
2. *There exists a location $z' \in \mathscr{S}_{r/2\kappa}$ so that $B^\infty_{r/2\kappa}(z') \subseteq B_r(x)$ and $x \in B^\infty_{r/2\kappa}(z')$.*

Proof. (Analogous to Facts 2.6.7.1 and 2.6.7.2.) For the first claim, let $z \in \mathscr{S}_r$ be the closest point to x in $\mathscr{S}_r$. We have $x_i \in [z_i - r/4, z_i + r/4]$; otherwise x would be strictly closer to the point $z \pm e_i \cdot (r/2) \in \mathscr{S}_r$, where e_i is the unit vector with value one in the ith coordinate and zero elsewhere. For arbitrary $y \in B_{r/2}(x)$, we have $|x_i - y_i| \leq r/2$, so $|z_i - y_i| \leq |z_i - x_i| + |x_i - y_i| = 3r/4$, so we have $y \in B^\infty_r(z)$.

For the second claim, let $z' \in \mathscr{S}_{r/2\kappa}$ be the closest point to x in $\mathscr{S}_{r/2\kappa}$. By the same argument as above, we have $|x_i - z'_i| \leq r/8\kappa$. Immediately we have $x \in B^\infty_{r/2\kappa}(z')$. Let $y \in B^\infty_{r/2\kappa}(z')$ be arbitrary. Then $d(x,y) \leq d(x,z') + d(z',y) = \sum_{i \in \{1,\ldots,k\}} |x_i - z'_i| + |z'_i - y_i| \leq k \cdot (r/8\kappa + r/2\kappa) < k \cdot (r/\kappa) \leq r$, and we have $y \in B_r(x)$. □

Definition 2.6.18 (Enclosing squares) *For any pair of locations $x, y \in L$:*

1. *Let $\mathsf{mesr}(x,y)$ ("minimum enclosing-square radius") denote the minimum $r \in \mathscr{R}$ such that, for some $z \in \mathscr{S}_r$, we have $x, y \in B^\infty_r(z)$.*
2. *Let $r = \mathsf{mesr}(x,y)$. Let $\mathsf{sesc}(y,r)$ ("16κ-fold-enclosing-square center") denote the location $z^*_{y,r} \in \mathscr{S}_{16\kappa r}$ such that $B_{8\kappa r}(y) \subseteq B^\infty_{16\kappa r}(z^*_{y,r})$ whose existence is guaranteed by Fact 2.6.17.1.*

Fact 2.6.19 (Relating distance and minimum enclosing-square radius) *For any $x, y \in L$, let $r := \mathsf{mesr}(x,y)$. Then we have $2kr \geq d(x,y) \geq r/4$.*

Proof. (Analogous to Fact 2.6.9.) Let $z \in \mathscr{S}_r$ be such that $x, y \in B^\infty_r(z)$. Then $d(x,y) \leq d(x,z) + d(z,y) = \sum_{i \in \{1,\ldots,k\}} |x_i - z_i| + |y_i - z_i| \leq k \cdot (2r)$. For the other direction, suppose for a contradiction that $d(x,y) \leq r/4$. Let $z^* \in \mathscr{S}_{r/2}$ be such that $B_{r/4}(x) \subseteq B^\infty_{r/2}(z^*)$, as guaranteed by Fact 2.6.17.1. Thus $x, y \in B_{r/4}(x) \subseteq B^\infty_{r/2}(z^*)$, contradicting the minimality of r. □

Thus, in the path from any source $s \in L$ to any target $t \in L$ found by GeoGreedy, the path will always remain inside the ball $B_{d(s,t)}(t) \subseteq B_{2k \cdot \mathsf{mesr}(s,t)}(t)$.

Lemma 2.6.20 (Relationship between square population and rank) *Let $s, t \in L$ be an arbitrary source/target pair of locations. Let $r = \mathsf{mesr}(s,t)$, and let $z^* = \mathsf{sesc}(t,r)$. Let $x, y \in L$ be arbitrary locations such that $x \in B_{2kr}(t)$ and $y \in B_{r/8}(t)$, and let $u, v \in P$ be arbitrary people such that $\mathsf{loc}(u) = x$ and $\mathsf{loc}(v) = y$. Then $\mathsf{rank}_u(v) \leq \mathsf{pop}(B^\infty_{16\kappa r}(z^*))$.*

Proof. (Analogous to Lemma 2.6.11.) Exactly as in that proof, we have $d(x,y) \leq d(x,t) + d(t,y) \leq 2kr + r/8 \leq 17kr/8$ and, for an arbitrary location ℓ so that $d(x,\ell) \leq d(x,y)$, we have $d(t,\ell) \leq d(t,y) + d(y,x) + d(x,\ell) \leq d(t,y) + 2d(x,y) \leq r/8 + 34kr/8 \leq 35kr/8$. Then, as before, we have

$$\text{for any location } \ell \in L, \text{ if } d(x,\ell) \leq d(x,y), \text{ then } \ell \in B^\infty_{16\kappa r}(z^*) \tag{2.10}$$

because if $d(x,\ell) \leq d(x,y)$, then $\ell \in B_{35kr/8}(t) \subseteq B_{8\kappa r}(t) \subseteq B^\infty_{16\kappa r}(z^*)$ by the definition of $z^* = \mathsf{sesc}(t,r)$. By definition of rank, we have that $\mathsf{rank}_u(v) \leq \mathsf{pop}(\{\ell \in L : d(x,\ell) \leq d(x,y)\}) \leq \mathsf{pop}(B^\infty_{16\kappa r}(z^*))$, where the second inequality follows from (2.10). □

Lemma 2.6.21 (GeoGreedy halves distance in expected polylogarithmic steps) *Let $s \in L$ be an arbitrary source location, and let $t \in L$ be a target location randomly chosen according to the distribution* $\mathsf{density}(\cdot)$. *Then the expected number of steps before the geographically greedy algorithm started from location s reaches either t or a point in $B_{d(s,t)/2}(t)$ is $O(2^{O(k \log k)} \log n \log m) = O(2^{O(k \log k)} \log^2 n)$, where the expectation is taken over the random choice of t.*

Proof. Let $r_t := \mathsf{mesr}(s,t)$. Fact 2.6.19 implies that it suffices to show that the expected number of steps before GeoGreedy started from location s lands in $B_{r_t/8}(t) \subseteq B_{d(s,t)/2}(t)$ is $O(2^{O(k \log k)} \log n \log m)$. Let $z_t := \mathsf{sesc}(t, r_t)$, and let z'_t be the location whose existence is guaranteed by Fact 2.6.17.2 such that the following facts hold:

$$z_t \in \mathscr{S}_{16\kappa r_t} \text{ and } B_{8\kappa r_t}(t) \subseteq B^{\infty}_{16\kappa r_t}(z_t) \tag{2.11}$$

$$z'_t \in \mathscr{S}_{r_t/16\kappa} \text{ and } B^{\infty}_{r_t/16\kappa}(z'_t) \subseteq B_{r_t/8}(t) \text{ and } t \in B^{\infty}_{r_t/16\kappa}(z'_t). \tag{2.12}$$

Assembling (2.11) and (2.12), we have

$$B^{\infty}_{r_t/16\kappa}(z'_t) \subseteq B_{r_t/8}(t) \subseteq B_{8\kappa r_t}(t) \subseteq B^{\infty}_{16\kappa r_t}(z_t) \qquad \text{and} \qquad t \in B^{\infty}_{r_t/16\kappa}(z'_t). \tag{2.13}$$

Start GeoGreedy at the source s. For any person $u \in P$ encountered by the algorithm, we have by the definition of GeoGreedy and by Fact 2.6.19 that

$$d(\mathsf{loc}(u), t) \leq d(s,t) \leq 2kr_t. \tag{2.14}$$

We refer to a person u as *good* if there exists a long-range link from that person to any person living in the ball $B_{r_t/8}(t)$. Let $\alpha_{u,t}$ denote the probability that a person $u \in P$ living at location $x_u = \mathsf{loc}(u) \in L$ is good. Then

$$\begin{aligned}
\alpha_{u,t} &= \sum_{v:\mathsf{loc}(v) \in B_{r_t/8}(t)} \frac{1}{\mathsf{rank}_u(v) \cdot H_n} \\
&\geq \sum_{v:\mathsf{loc}(v) \in B_{r_t/8}(t)} \frac{1}{\mathsf{pop}(B^{\infty}_{16\kappa r_t}(z_t)) \cdot H_n} \\
&= \frac{\mathsf{pop}(B_{r_t/8}(t))}{\mathsf{pop}(B^{\infty}_{16\kappa r_t}(z_t)) \cdot H_n} \quad =: \quad \alpha_t
\end{aligned}$$

by the definition of good, by Lemma 2.6.20 (which applies by (2.14)), and by the definition of $\mathsf{pop}(\cdot)$. As before, each step of the algorithm brings us to a new node never before seen by the algorithm until we reach node t. Thus whether a step is good is independent at each step, and the expected number of steps before we reach a good person (or t itself) is at most $1/\alpha_t$.

We now examine the expected value of $1/\alpha_t$ for a target location $t \in L$ chosen as the location of a person uniformly drawn from the population:

$$\begin{aligned}
\mathrm{E}_{t \in L \sim \mathsf{density}(\cdot)}[1/\alpha_t] = \sum_t \mathsf{density}(t) \cdot \frac{1}{\alpha_t} &= \sum_t \mathsf{density}(t) \cdot \frac{\mathsf{pop}(B^{\infty}_{16\kappa r_t}(z_t)) \cdot H_n}{\mathsf{pop}(B_{r_t/8}(t))} \\
&\leq H_n \cdot \sum_t \mathsf{density}(t) \cdot \frac{\mathsf{density}(B^{\infty}_{16\kappa r_t}(z_t))}{\mathsf{density}(B^{\infty}_{r_t/16\kappa}(z'_t))}
\end{aligned}$$

by (2.13), using the fact that $B^{\infty}_{r_t/16\kappa}(z'_t) \subseteq B_{r_t/8}(t)$ and the monotonicity of $\mathsf{density}(\cdot)$. We now reindex the summation to be over radii and square centers rather than over targets t. Recall that $z_t \in \mathscr{S}_{16\kappa r_t}$ and $z'_t \in \mathscr{S}_{r_t/16\kappa}$, and that $B^{\infty}_{r_t/16\kappa}(z'_t) \subseteq B^{\infty}_{16\kappa r_t}(z_t)$ by (2.13), and therefore that $z'_t \in B^{\infty}_{16\kappa r_t}(z_t)$. Thus, we have that

$$\begin{aligned}
\mathrm{E}_{t\in L\sim\mathsf{density}(\cdot)}[1/\alpha_t] &\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{S}_{16\kappa r}} \sum_{z'\in\mathscr{S}_{r/16\kappa}:z'\in B^{\infty}_{16\kappa r}(z)} \frac{\mathsf{density}(B^{\infty}_{16\kappa r}(z))}{\mathsf{density}(B^{\infty}_{r/16\kappa}(z'))} \sum_{t:z'_t=z'} \mathsf{density}(t) \\
&\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{S}_{16r}} \sum_{z'\in\mathscr{S}_{r/16\kappa}:z'\in B^{\infty}_{16\kappa r}(z)} \frac{\mathsf{density}(B^{\infty}_{16\kappa r}(z))}{\mathsf{density}(B^{\infty}_{r/16\kappa}(z'))} \sum_{t\in B^{\infty}_{r/16\kappa}(z')} \mathsf{density}(t)
\end{aligned}$$

where the second inequality follows from (2.13). But then

$$\begin{aligned}
\mathrm{E}_{t\in L\sim\mathsf{density}(\cdot)}[1/\alpha_t] &\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{S}_{16\kappa r}} \sum_{z'\in\mathscr{S}_{r/16\kappa}:z'\in B^{\infty}_{16\kappa r}(z)} \frac{\mathsf{density}(B^{\infty}_{16\kappa r}(z))}{\mathsf{density}(B^{\infty}_{r/16\kappa}(z'))} \cdot \mathsf{density}(B^{\infty}_{r/16\kappa}(z')) \\
&= H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{S}_{16\kappa r}} \mathsf{density}(B^{\infty}_{16\kappa r}(z)) \cdot |\{z' \in \mathscr{S}_{r/16\kappa} : z' \in B^{\infty}_{16\kappa r}(z)\}|.
\end{aligned}$$

Now we are almost done: by Fact 2.6.16.2, we have that

$$|\{z' \in \mathscr{S}_{r/16\kappa} : z' \in B^{\infty}_{16\kappa r}(z)\}| = 2^{O(k\log(256\kappa^2))} = 2^{O(k\log k)}. \tag{2.15}$$

Furthermore, we have $\sum_{z\in\mathscr{S}_{16kr}} \mathsf{density}(B^{\infty}_{16kr}(z)) = 2^{O(k)}$; Fact 2.6.16.1 implies that we only "double count" each point $2^{O(k)}$ times. Thus we have

$$\begin{aligned}
\mathrm{E}_{t\in L\sim\mathsf{density}(\cdot)}[1/\alpha_t] &\leq H_n \cdot \sum_{r\in\mathscr{R}} \sum_{z\in\mathscr{S}_{16\kappa r}} \mathsf{density}(B^{\infty}_{16\kappa r}(z)) \cdot |\{z' \in \mathscr{S}_{r/16\kappa} : z' \in B^{\infty}_{16\kappa r}(z)\}| \\
&= H_n \cdot 2^{O(k\log k)} \cdot 2^{O(k)} \cdot |\mathscr{R}| \\
&= H_n \cdot 2^{O(k\log k)} \cdot O(\log q) \;=\; O(\log n \cdot \log m \cdot 2^{O(k\log k)}),
\end{aligned}$$

which proves the lemma. □

Theorem 2.6.22 (GeoGreedy finds short paths in all k-D meshes) *For every k-dimensional mesh population network with n people and m locations, the expected length of the search path found by* GeoGreedy *from an arbitrarily chosen source location s to the location t of a uniformly chosen target individual is $O(2^{O(k\log k)}\log n\log^2 m) = O(2^{O(k\log k)}\log^3 n)$.*

Proof. Immediate by inductive application of Lemma 2.6.21. □

2.6.6 Recursive Population Networks

The model of population networks that we have defined in previous sections accounts for the geographic locations of people down to the "city" level of precision. That is, many people are potentially colocated in the model defined above. (In our experimental examination of the LiveJournal network, the locations are cities and towns in the United States.) In this section, we describe a recursive model for geographic locations that allows further specification of geographic

locations, and that allows the routing of messages all the way to an *individual*, rather than just to that individual's city or town.

Our motivation for this recursive model is the following. Suppose that we are attempting to route a message from a source person, say A, to a target person, say Be, located in the sixth floor of the Gates Tower of the Stata Center on the MIT campus in Cambridge, Massachusetts. A plausible route for the message from A to Be is that it would first be routed to someone living in the Boston area, then to someone in Cambridge, then someone associated with MIT, then someone in Stata, and then someone on G6. From there, routing to Be herself is easy; essentially everyone on the floor will know Be. (Of course, as we discussed previously, geography is not the only way in which real subjects will attempt to reach a designated target; here, though, we are considering the implications of purely geographic routing.) This example suggests a recursive model for geographic location; after the message reaches the correct "city" (Cambridge), it is then routed to the correct "neighborhood" (MIT), and so forth, until the target herself has been reached.

In this section, we will primarily be focused on recursive mesh-based networks. The best intuition for this type of network is routing to a target in Manhattan: first we route along the longitude/latitude grid to reach the city itself, then we route along the street grid of Manhattan to reach the correct block, then we route along the (one-dimensional) grid of floors within an apartment building, and at last we route on the grid of apartments within each floor until we reach the target herself.

The RPN Model

We will begin with a population network as described in Section 2.6.1. However, in the earlier model, each location $\ell \in L$ of the network represented a collection of colocated individuals. In the extended model, each $\ell \in L$ represents *either* a single individual or a population subnetwork. Formally, we define a recursive population network (RPN) as follows:

Definition 2.6.23 (Recursive Population Network) *A* recursive population network (RPN) *is a sextuple* $\langle L, d, P, \mathsf{loc}, E, M\rangle$ *where*

- $\langle L, d, P, \mathsf{loc}, E\rangle$ *is a population network as per Definition 2.6.1; and*
- $M : L \longrightarrow P \cup \{\text{RPNs}\}$, *as follows. Let* $P_\ell := \{u \in P : \mathsf{loc}(u) = \ell\}$.
  - *If* $|P_\ell| = 1$, *then* $\{M(\ell)\} = P_\ell$.
  - *If* $|P_\ell| \geq 2$, *then* $M(\ell) = \langle L_\ell, d_\ell, P_\ell, \mathsf{loc}_\ell, E_\ell, M_\ell\rangle$ *is an RPN with* $|L_\ell| \geq 2$.

The structure of an RPN, then, is as follows: we have a population network consisting of people living at various locations, with a distance function describing the geographic separation between locations. For each location ℓ in which strictly more than one person lives, we have, recursively, a population network for the people living in location ℓ. We think of an RPN as a tree of population networks, where each network $\langle L, d, P, \mathsf{loc}, E\rangle$ has "child networks" for each location $\ell \in L$ with $|P_\ell| \geq 2$. The leaves of the tree are the locations with a single resident.

Again, in the following we limit our attention to rank-based recursive population networks on meshes, so that every population network in our RPN satisfies the conditions described in previous sections. Specifically, we require that every location $\ell \in L$ satisfies $\mathsf{pop}(L) = |\{u : \mathsf{loc}(u) = \ell\}| \geq 1$; no locations are empty. Distances between locations in any population network in the tree are given

by Manhattan distance. We allow the meshes to differ in dimension and in size (i.e., k and q from the previous sections can vary from grid to grid), though we require that each grid be nontrivial, in the sense that there must be at least two distinct locations (and thus at least two people) in it.

We will require some additional notation. Let $\rho = \langle L, d, P, \mathsf{loc}, E, M\rangle$ be an RPN. Define $\mathcal{M}(\rho)$ to be the set of population networks contained in ρ—that is,

$$\begin{aligned}\mathcal{M}(p) &= \emptyset \\ \mathcal{M}(\langle L, d, P, \mathsf{loc}, E, M\rangle) &= \{\langle L, d, P, \mathsf{loc}, E\rangle\} \cup \bigcup_{\ell \in L} \mathcal{M}(M(\ell)).\end{aligned}$$

Define the tree of the meshes $\mathcal{M}(\rho)$ as above, so that the meshes formed on the locations in a population network N are the children of N in the tree. Let $\mathsf{parent}(N)$ denote the parent mesh in this tree for every mesh not at the top level. Define $\mathsf{depth} : \mathcal{M} \longrightarrow \mathbb{Z}^+$ to denote mesh depth in the tree, so the top-level mesh has depth one, and every other mesh N has depth given by $\mathsf{depth}(N) = 1 + \mathsf{depth}(\mathsf{parent}(N))$.

For an individual $u \in P$, let $N_u = u$ denote the leaf node of the tree where u is the lone person living at a location. For a mesh $N \in \mathcal{M}(\rho)$, we write $u \in N$ to denote that N_u is in the subtree rooted at N—or, in other words, that $u \in P_N$ where $N = \langle L_N, d_N, P_N, \mathsf{loc}_N, E_N\rangle$. Write $\mathsf{depth}(u) := \mathsf{depth}(N_u)$, and for any index $i \in \{1, \ldots, \mathsf{depth}(u)\}$, we write $\mathsf{mesh}_i(u)$ to denote the unique mesh at depth i in the tree such that $u \in \mathsf{mesh}_i(u)$. Finally, for two individuals $u, v \in P$, let $\mathsf{LCA}(u, v)$ denote the least common ancestor of u and v in the tree—that is, the deepest (i.e., smallest population) mesh N in the tree such that $u, v \in N$.

As in the nonrecursive model, we let $\mathsf{pop} : L \longrightarrow \mathbb{Z}^+$ denote the population $|\{u \in P : \mathsf{loc}(u) = \ell\}|$ of each point on L, and let $\mathsf{pop} : \mathscr{P}(L) \longrightarrow \mathbb{Z}^+$ denote $\mathsf{pop}(L') := \sum_{\ell \in L'} \mathsf{pop}(\ell)$. Let $\mathsf{density}(\ell) := \mathsf{pop}(\ell) / \left(\sum_{\ell' \in L} \mathsf{pop}(\ell')\right)$ and $\mathsf{density}(L') = \sum_{\ell \in L'} \mathsf{density}(\ell)$ denote population density, so that $\mathsf{density}(L) = 1$. We write $n := \mathsf{pop}(L) = |P|$ to denote the total population. We will further overload notation, and let $\mathsf{pop}(N)$ denote the population of a mesh in the tree, so that $\mathsf{pop}(N) = |P_N|$, where $N = \langle L_N, d_N, P_N, \mathsf{loc}_N, E_N\rangle$.

Distance in RPNs. As in previous sections, we need to be able to calculate which of two people is closer to a specified source. In computing distances, we consider only the coarsest-resolution grid in which the locations are distinct—that is, distances between different cities are computed based upon the distance between the cities, but distances within the same city are computed on the basis of the distance between neighborhoods. For example, we treat the distance from Winslow, Arizona to the Upper West Side as identical to the distance from Winslow to Midtown, but two people living in Greenwich Village are closer to each other than either is to a resident of Harlem. More formally, we extend the distance function from locations to people as follows.

For people $u, v \in P$, let $N := \mathsf{LCA}(u, v)$, where $N = \langle L_N, d_N, P_N, \mathsf{loc}_N, E_N\rangle$. (Note that $u, v \in P_N$ and that $\mathsf{loc}_N(u) \neq \mathsf{loc}_N(v)$.) We define $d(u, v) := \langle -\mathsf{depth}(N), d_N(\mathsf{loc}_N(u), \mathsf{loc}_N(v))\rangle$, and we use standard lexicographic ordering on pairs to compare distances. That is, for people $u, v, w \in P$, we have that $d(u, v) \leq d(u, w)$ if and only if one of the following two conditions holds:

(i) $\mathsf{depth}(\mathsf{LCA}(u, v)) > \mathsf{depth}(\mathsf{LCA}(u, w))$—that is, there is a deeper mesh in the tree containing both u and v than the deepest mesh containing both u and w; or

(ii) $N = \mathsf{LCA}(u, v)) = \mathsf{LCA}(u, w)$ and $d_N(\mathsf{loc}(u), \mathsf{loc}(v)) \leq d_N(\mathsf{loc}(u), \mathsf{loc}(w))$—that is, there is no mesh containing $\{u, v\}$ or $\{u, w\}$ that is deeper than the deepest mesh containing all three

people $\{u, v, w\}$, and furthermore the distance between the locations of u and v in N is smaller than the distance between the locations of u and w in N.

That is, the distance between two people is given by the height of their least common ancestor, with ties being broken by the distance separating their locations at the highest level of the tree in which their locations are distinct.

As before, we will need to break ties among pairs of people separated by the same distance. Again, we do so by postulating a total ordering on the people in the network, but here we require a the stronger assumption that this total ordering is derived by choosing a permutation of $\{1, \ldots, n\}$ uniformly at random from the space of all such permutations.

Neighbors in RPNs. As with a simple population network, we assume that each individual has local neighbors in each cardinal direction in the grid. The notion of a local neighborhood, however, is complicated by the fact that we allow the dimensionality of the grids to vary within an RPN.

The only property that we require of our local neighbors is the following: for an arbitrary source person u_s and an arbitrary distinct target person u_t, there must be a neighbor v of u_s such that $d(u_s, u_t) > d(v, u_t)$. One way to achieve this property is as follows: for an RPN ρ, let k^* be the maximum dimensionality of every grid that appears in ρ. Now consider a particular person $u \in P$. To compute the local neighbor of u to the east, say, we proceed as follows: continue to walk up the tree starting from N_u until we reach a mesh N such that there is a location directly to the east of $\mathsf{loc}_N(u)$ in the grid N. (That is, we walk up the tree until u is not on the eastern border of the current grid.) Select an arbitrary resident of that location as the eastern neighbor of u. If no eastern neighbor can be found, then u does not have such a neighbor (as in the case of a single mesh). We compute a local neighbor for each node in each of the $2k^*$ cardinal directions. (A node u may have a local neighbor v that is u's local neighbor in more than one cardinal direction.)

Rank-Based RPNs

As before, we introduce one long-range link per node in an RPN, chosen according to rank-based friendship. Thus every person in an RPN has a constant number of local neighbors if the dimensionality of every mesh contained in the RPN is constant, and one long-distance neighbor chosen according to the rank-based-friendship formula (2.1).

Routing on Rank-Based RPNs

In this section, we will prove that GeoGreedy finds short paths in expectation in any rank-based grid RPN. We first observe that the results of the previous section imply that the paths found by GeoGreedy are short in any "shallow" RPN—i.e., an RPN corresponding to a tree of small depth—and then prove that these results carry over to "deep" trees as well.

Fact 2.6.24 *For any individual $u \in P$ and any depth $i \leq \mathsf{depth}(u)$, all elements of $\mathsf{mesh}_i(u)$ are closer to u than any element outside $\mathsf{mesh}_i(u)$ is to u.*

Proof. Immediate by the definition of distance. □

An immediate consequence of this fact is that the path found by GeoGreedy aiming for a target t will never leave $\mathsf{mesh}_i(t)$ once it enters this subtree.

We can now observe that the expected time required to reach a target t drawn uniformly from the population P is upper bounded by $O(\log^3 n \cdot \mathsf{depth}(t))$, by Theorem 2.6.22: in expectation we reach the target location in any particular mesh in $O(\log^3 n)$ steps; we must find the correct location $\mathsf{depth}(t)$ times before we have arrived at the target person herself. Thus if no location contains more than a $(1-\varepsilon)$-fraction of the population—that is, that $\mathsf{density}(\ell) < 1-\varepsilon$ for every location ℓ in every mesh in the RPN—then $\mathsf{depth}(t) = O(\log n)$, and thus the expected length of the GeoGreedy search path is $O(\log^4 n)$ when we take the dimensionality of the meshes contained in the RPN to be constant. In the following, we will remove this restriction on the depth of the RPN and derive the same result.

Let $\langle L, \hat{d}, P, \mathsf{loc}, E, M\rangle$ be a grid-based RPN, and let d be the L_1-based distance function derived as described above. Let u_s be an arbitrarily chosen source person, and let u_t be a target person chosen uniformly at random from all people in P. Let $N_{\mathsf{LCA}} := \mathsf{LCA}(u_s, u_t)$, let $\delta_{\mathsf{LCA}} := \mathsf{depth}(N_{\mathsf{LCA}})$ be its depth, and let $P_{\mathsf{LCA}} := \mathsf{pop}(N_{\mathsf{LCA}})$ be its population. We are interested in the number of steps taken by the geographically greedy algorithm in routing from u_s to u_t.

Intuitively, our proof proceeds as follows. In the mesh N_{LCA}, we begin at some location ℓ_s and we wish to reach some location ℓ_t. We claim that within a polylogarithmic number of steps we will reduce by a factor of two the number of people closer to the target than the current message-holder is. There are two cases. If $\mathsf{pop}(\ell_t) \leq |P_{\mathsf{LCA}}|/2$ (i.e., the subpopulation containing the target is not too big), then simply reaching ℓ_t as per Theorem 2.6.22 constitutes considerable progress towards the target. On the other hand, if $\mathsf{pop}(\ell_t) > |P_{\mathsf{LCA}}|/2$, then any node encountered on the GeoGreedy path has a good probability ($\Omega(1/H_n)$) of linking to one of the $|P_{\mathsf{LCA}}|/2$ people closest to u_t. Thus in $O(\log n)$ steps, with high probability we reach one of $|P_{\mathsf{LCA}}|/2$ people closest to u_t, which is also considerable progress towards the target. In either case, we have reduced by a factor of two the number of people closer to the target than the current message-holder; a logarithmic number of repetitions of this process will find the target individual herself.

To formalize this argument, consider running GeoGreedy starting from person u_s until the completion of the following two-phase operation:

Phase 1 ("Halfway there"): Run GeoGreedy starting from u_s until we reach a person v such that either (i) $v \in \mathsf{mesh}_{\mathsf{depth}(u_t)-1}(u_t)$—i.e., the mesh that directly contains the target u_t—or (ii) $\mathsf{rank}_{u_t}(v) \leq \mathsf{pop}(\mathsf{LCA}(u_s, u_t))/2$.

Phase 2 ("One level deeper"): Run GeoGreedy starting from v until we reach a person w such that either $w = u_t$ or $\mathsf{depth}(\mathsf{LCA}(w, u_t)) > \mathsf{depth}(\mathsf{LCA}(v, u_t))$.

We wish to show two things about this operation: first, after a logarithmic number of iterations of the two-phase operation, GeoGreedy will reach the target u_t; and, second, each iteration requires only a polylogarithmic number of steps.

Lemma 2.6.25 *Suppose we run one iteration of the two-phase process, starting from source person u_s and aiming for target person u_t, where the message holders at the end of Phase 1 and Phase 2 are, respectively, people v and w. Then either $w = u_t$ or $\mathsf{pop}(\mathsf{LCA}(w, u_t)) \leq \mathsf{pop}(\mathsf{LCA}(u_s, u_t))/2$.*

Proof. Note that if Phase 1 terminates when we reach a person v in case (i), such that $v \in \mathsf{mesh}_{\mathsf{depth}(u_t)-1}(u_t)$, then by definition Phase 2 terminates when we reach the target $w = u_t$. Otherwise, suppose that Phase 1 terminates when we reach a person v such that

$$\mathsf{rank}_{u_t}(v) \leq \mathsf{pop}(\mathsf{LCA}(u_s, u_t))/2 \tag{2.16}$$

and Phase 2 terminates when we reach a person w such that $\mathsf{depth}(\mathsf{LCA}(w, u_t)) > \mathsf{depth}(\mathsf{LCA}(v, u_t))$. Recall again Fact 2.6.24: in a GeoGreedy path, the distance from the current person on the path to the target is always decreasing, and thus every person in $\mathsf{LCA}(w, u_t)$ is closer to u_t than v was, and thus than u_s was as well. Note then that

$$\mathsf{rank}_{u_t}(v) > \mathsf{pop}(\mathsf{LCA}(w, u_t)). \tag{2.17}$$

Thus, by assembling (2.16) and (2.17), we have $\mathsf{pop}(\mathsf{LCA}(u_s, u_t))/2 \geq \mathsf{rank}_{u_t}(v) > \mathsf{pop}(\mathsf{LCA}(w, u_t)$, which proves the lemma. □

Lemma 2.6.26 *Consider an RPN ρ in which every mesh in $\mathcal{M}(\rho)$ has dimensionality $\Theta(1)$. Suppose we run one iteration of the two-phase process, starting from an arbitrarily chosen source person u_s and aiming for a target person u_t chosen uniformly at random from the population. Then the expected number of steps to complete the two-phase operation is $O(\log^3 n)$, where the expectation is taken over the random choice of target u_t.*

Proof. Recall that every person encountered along the GeoGreedy path will be inside $N_{\mathsf{LCA}} := \mathsf{LCA}(u_s, u_t)$. Let $P^* := \{v \in N_{\mathsf{LCA}} : \mathsf{rank}_{u_t}(v) \leq |\mathsf{pop}(N_{\mathsf{LCA}})|/2\}$ denote the set of "good" people, so that if the path reaches any person $v \in P^*$, then Phase 1 will terminate. Notice that because the $|\mathsf{pop}(N_{\mathsf{LCA}})|$ people in N_{LCA} are all closer to $u_t \in N_{\mathsf{LCA}}$ than any person outside of N_{LCA}, we have that $|P^*| = |\mathsf{pop}(N_{\mathsf{LCA}})|/2$. Similarly, for any person $u' \in N_{\mathsf{LCA}}$ appearing along the GeoGreedy path, we have that

$$\mathsf{rank}_{u'}(v^*) \quad \leq \quad |\mathsf{pop}(N_{\mathsf{LCA}})| \quad \forall v^* \in P^*. \tag{2.18}$$

Thus we know that

$$\mathtt{Pr}\left[u' \rightarrow v^*\right] \geq 1/|\mathsf{pop}(N_{\mathsf{LCA}})| \cdot H_n \tag{2.19}$$

by (2.1). There are $|\mathsf{pop}(N_{\mathsf{LCA}})|/2$ distinct elements $v^* \in P^*$, and thus

$$\mathtt{Pr}\left[\exists v^* \in P^* : u' \rightarrow v^*\right] \geq |\mathsf{pop}(N_{\mathsf{LCA}})|/2|\mathsf{pop}(N_{\mathsf{LCA}})| \cdot H_n = 1/2H_n \tag{2.20}$$

because the events of (2.19) are negatively correlated (as only one link is picked for each u'). Therefore the expected number of trials (i.e., elements u' found on the GeoGreedy path) before we find a person u' who links to P^* is at most $2H_n = O(\log n)$. Furthermore, each trial is independent, because every step moves to a person strictly closer to u_t. By the Chernoff bound, then, after $O(\log n)$ trials, with probability at least $1 - O(1/n)$, we reach P^*. Because no GeoGreedy path can exceed n in length, this high-probability result implies that the expected length of the GeoGreedy path in Phase 1 is $O(\log n)$.

For Phase 2 of the two-phase operation, we start the GeoGreedy walk in a population network $\langle L, d, P, \mathsf{loc}, E\rangle$ at a person $v \in P$, and we must continue the path until we arrive at a person w such that $\mathsf{loc}(w) = \mathsf{loc}(u_t)$. Notice that u_t is chosen uniformly at random among all elements of P, by definition. Therefore, by Theorem 2.6.22, we have that the expected length of the GeoGreedy path to reach the target location $\mathsf{loc}(u_t)$ is $O(\log^3 n)$ because the dimensionality of the mesh is constant.

Therefore the expected length of the GeoGreedy path in one iteration of the two-phase operation is $O(\log n + \log^3 n) = O(\log^3 n)$. □

We can now state the main theorem:

Theorem 2.6.27 *Let $\rho = \langle L, d, P, \mathsf{loc}, E, M \rangle$ by an arbitrary grid-based RPN with $n = |P|$ people such that every population network $N \in \mathcal{M}(\rho)$ has dimensionality $k_N = \Theta(1)$. For an arbitrary source person $u_s \in P$ and a target person $u_t \in P$ chosen uniformly at random from P, we have that the expected length of the* GeoGreedy *path from u_s to u_t is $O(\log^4 n)$.*

Proof. The number of steps required for one iteration of the two-phase operation is $O(\log^3 n)$, by Lemma 2.6.26. Clearly, after $\log n$ iterations of the two-phase operation, we must reach the target u_t, because we have $|\mathsf{pop}(\mathsf{LCA}(w, u_t))| = 1$ for the current person w on the GeoGreedy path, by Lemma 2.6.25. □

2.7 Geographic Linking in the LiveJournal Social Network

We return to the LiveJournal social network to show that rank-based friendship holds in practice. The relationship between $\mathsf{rank}_u(v)$ and the probability that u is a friend of v shows an approximately inverse linear fit for ranks up to approximately 100,000, as shown in Figure 2-8(a). Because the LiveJournal data contain geographic information limited to the level of towns and cities, our data do not have sufficient resolution to distinguish between all pairs of ranks. (Specifically, the average person in the LiveJournal social network lives in a city with a population of 1306.) Thus in Figure 2-8(b) we show the same data, but with the displayed probabilities averaged over a range of 1306 ranks. (Due to the logarithmic scale of the rank axis, the sliding window may appear to apply broad smoothing; however, recall that each source contains ranks for each of approximately 500,000 points, so smoothing by a window of size 1300 causes a point to be influenced by only the one third of one percent of the closest points on the curve.) This experiment validates that the LiveJournal social network does exhibit rank-based friendship, and thus yields a sufficient explanation for the navigable-small-world properties observed experimentally. In Figures 2-8(c) and 2-8(d), we show the same data with the same ε-correction as in Figure 2-5(b), in which we subtract the $\varepsilon \approx 5.0 \times 10^{-6}$ probability of non-geographic friendship that is independent of rank, so the resulting plot shows the relationship between rank and geographic-friendship probability.

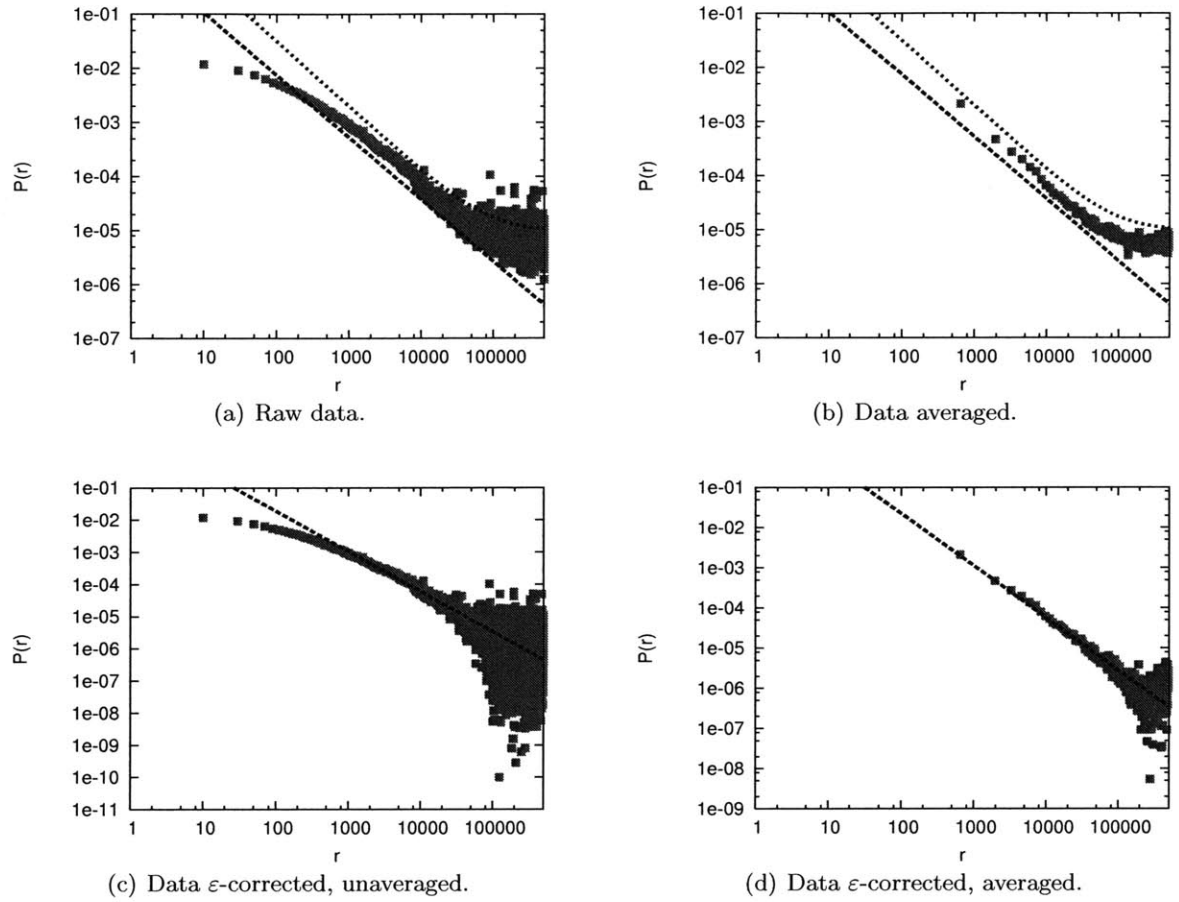
In Figure 2-9, we plot the same data as in Figure 2-8(b), but restricted to the East and West coasts. Notice that the slopes of the lines for the two coasts have nearly the same slope, and are much closer together than the distance/friendship-probability slopes shown in Figure 2-6(a).

The natural mechanisms of friendship formation, whatever they may be, result in rank-based friendship: people in aggregate have formed relationships with almost exactly the optimal connection between friendship and rank that is required to produce a navigable small world. In a lamentably imperfect world, it is remarkable that people form friendships so close to the perfect distribution for navigating their social structures.

2.8 Future Directions

In this chapter, we have given a strong characterization of the nature of geographic friendship in social networks, from both empirical and theoretical perspectives. There are a number of interesting directions for future work, again both theoretical and practical.

In the simulated experiments in this chapter, we aimed only to reach the town or city of the target individual. From the real-world experiments of Milgram [126] and Dodds et al. [49] on message-passing, there is significant evidence that an effective routing strategy typically begins by



(a) Raw data.

(b) Data averaged.

(c) Data ε-corrected, unaveraged.

(d) Data ε-corrected, averaged.

Figure 2-8: *The relationship between the probability of a link between u and v and v's rank with respect to u. We plot the probability $P(r)$ of a link from a randomly chosen source u to the rth-closest node to u in the LiveJournal network, averaged over 10,000 independent source samples. If there is a link from u to one of the nodes $S_\delta^u = \{v : d(u,v) = \delta\}$, where the people in S_δ^u are all tied at ranks $r+1, \ldots, r+|S_\delta^u|$, then we count a $(1/|S_\delta^u|)$-fraction of a link for each of these ranks. The dotted lines in panels (a) and (b) show $P(r) \propto 1/r^{1.15}$ and $P(r) \propto \varepsilon + 1/r^{1.2}$. In panels (a) and (c), we show data for every twentieth rank. Because we do not have fine-grained geographic data, on average the ranks r through $r+1305$ all represent people in the same city; thus we have little data to distinguish among these ranks. In panels (b) and (d), we show the same data averaged into buckets of size 1306: for each displayed rank r, we show the average probability of a friendship over ranks $\{r-652, \ldots, r+653\}$. In panels (c) and (d), we replot the same data as in (a) and (b), correcting for the background friendship probability $\varepsilon = 5.0 \times 10^{-6}$: we plot the rank r versus $P(r) - \varepsilon$. In panel (c), the dashed line corresponds to $P(r) - \varepsilon \propto r^{-1.25}$; in panel (d), the dashed line corresponds to $P(r) - \varepsilon \propto r^{-1.3}$.*

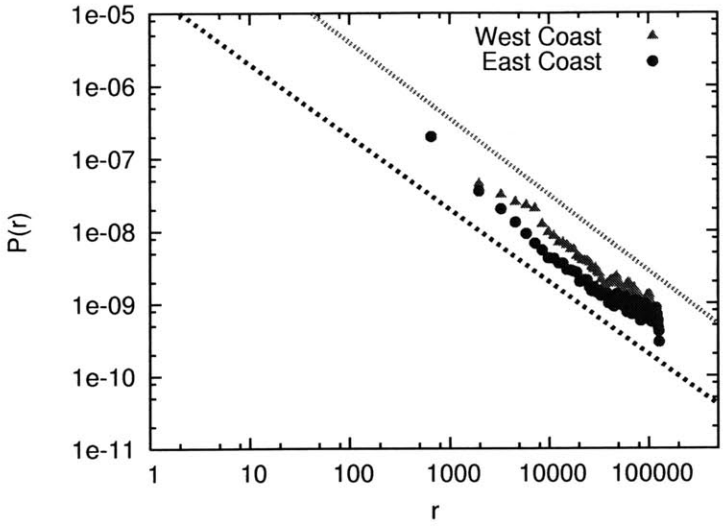


Figure 2-9: *As in Figure 2-8, we plot the relationship between the probability $P(r)$ that person v is a friend of person u and the rank $r = \mathsf{rank}_u(v)$ of v with respect to u. Here we restrict the data to people living on the West Coast and the East Coast, respectively. Ranks are averaged over a range of ranks of size 1306. The dotted lines correspond to $P(r) \propto r^{-1.00}$ and $P(r) \propto r^{-1.05}$.*

making long geography-based hops as the message leaves the source, and typically ends by making hops based on attributes other than geography. Thus there is a transition from geography-based to non-geography-based routing at some point in the process. The recursive version of our theorem shows that short paths are constructible using the geographically greedy algorithm through any level of resolution in which rank-based friendship holds. However, within a sufficiently small geographic distance, more effective routing strategies are likely to be based upon non-geographic properties, such as occupation, organization membership, or hobbies. In the LiveJournal data, we have listings of the topical interests of the bloggers, and we might hope to first induce a metric on topics based upon topic cooccurrence and then use this metric as the basis of a "local" interest-greedy routing algorithm (in place of GeoGreedy). Using interests in this manner has been previously considered in purely synthetic data by Watts, Dodds, and Newman [176]; understanding this style of approach in the LiveJournal data, and establishing provable guarantees on the length of the path resulting from using an interest-greedy algorithm, could be highly revealing.

The rank-based friendship model that we have introduced here has two desirable properties: (1) it holds, at least approximately, in a real social network, and (2) it gives a theoretical explanation of the small-world phenomenon. It also serves to partially explain the high clustering coefficients seen in real social networks: if v and w both have low ranks with respect to u, then they will have relatively low ranks with respect to each other as well. Understanding *why* rank-based friendship should hold is another interesting open direction. One possible explanation is that the interests of the LiveJournal users (e.g., Lynah Rink, the Boston Red Sox, Trent Lott) have natural "geographic centers"; the geographic distribution of friendships may result from the right kind of distribution of interests of people in the network. Another possibility is that longer-range geographic friendships arise because of the distribution of the distances that people typically move when they relocate. (It is curious that friendship probability continues to decrease as far away as 800 to 1000 kilometers;

this phenomenon may be explained by the tendency of people to stay on the same coast when they move.) It is also interesting to understand the limitations of rank-based friendship. For example, are there mechanisms beyond rank-based friendship that are required to account for increased clustering coefficients? What is the effect of measuring proximity by as-the-crow-flies distance instead of by, for example, driving distance?

From a theoretical perspective, our theorems in this section show that $\mathbf{E}_{u_t}[|\mathsf{GeoGreedy}(u_s, u_t)|]$ is polylogarithmic in the size of the population for any source person u_s. On the other hand, the results of Kleinberg for uniform populations show that, with high probability, the length of the path $\mathsf{GeoGreedy}(u_s, u_t)$ has polylogarithmic length for *any* u_s and u_t when friendship probabilities are chosen according to the correct distance-based distribution. It appears that there may be population distributions for which the "for all" condition cannot be achieved in our context, probably by creating a recluse who is very unlikely to be reached by long-range links. Finding and formalizing such an example (or proving that GeoGreedy always finds a short path for any target) is an interesting open direction.

The assumption that there are no empty locations in our network—for every location $\ell \in L$, there exists at least one person $x \in P$ such that $\mathsf{loc}(x) = \ell$—is required in our results to guarantee that GeoGreedy never gets "stuck" at a person u without a local neighbor closer to the target than u herself is. Investigating the limitations of GeoGreedy in a model with zero-population locations (like lakes and deserts in the real world) is an intriguing direction, and would eliminate the most unrealistic limitation in our model. The problem of geographic routing via local-information algorithms in general, and geographic routing around obstacles in particular, has been previously considered in the wireless-networking community (see, e.g., the work of Brad Karp and his coauthors [92, 93, 100]). It is an interesting question as to whether these results, where there is typically a threshold on the geographic distance that a message can be passed because of physical constraints in the communication technology, can be adapted to the social-network setting. There has also been potentially applicable work in computational geometry on establishing (worst-case) bounds on geometric routing algorithms that attempt to transmit a packet from a source to a target in a planar graph embedded into the plane (e.g., [33, 34]).

Finally, our theoretical results in this chapter can be viewed as an extension of Kleinberg's theorem to a dimension-independent model that actually holds in reality. There have been some recent theoretical results extending and refining Kleinberg's result, and we might hope to be able to make analogous improvements to our result. "Friend of a friend" routing, in which the knowledge of the current message-holder u is augmented to include the set of friends S_v of each friend v of u—and u then chooses as the next step in the chain his friend v such that $\min_{w \in S_v} d(w, t)$ is minimized—has been shown to reduce the length of discovered paths to $\Theta(\log n / \log \log n)$ (with logarithmic degree) from $\Theta(\log^2 n)$ [123]. Nguyen and Martens have also exhibited local-information routing algorithms that produce paths that are closer in length to the diameter of the network [124, 150]. Slivkins [163] has extended Kleinberg's result to networks with low doubling dimension. We hope that these theoretical results may be applicable in the rank-based models considered in this chapter.

Chapter 3

The Link-Prediction Problem

Prediction is very difficult, especially of the future.
— Niels Bohr (1885–1962).

Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future? We formalize this question as the *link-prediction problem*, and develop approaches to link prediction based on measures for analyzing the "proximity" of nodes in a network. Experiments on large coauthorship networks suggest that information about future interactions can be extracted from network topology alone, and that fairly subtle measures for detecting node proximity can outperform more direct measures.

3.1 Introduction

As part of the recent surge of research on large, complex networks and their properties, a considerable amount of attention has been devoted to the computational analysis of social networks. The availability of large, detailed datasets encoding such networks has stimulated extensive study of their basic properties, and the identification of recurring structural features. (See Chapter 1.) Social networks are highly dynamic objects; they grow and change quickly over time through the addition of new edges, signifying the appearance of new interactions in the underlying social structure. An understanding of the mechanisms by which they evolve is still not well developed, and it forms the motivation for the work described in this chapter. We define and study a basic computational problem underlying social-network evolution, the *link-prediction problem*: given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'.

In effect, the link-prediction problem asks the following fundamental question: to what extent can the evolution of a social network be modeled using features *intrinsic to the network itself?* Consider a coauthorship network among scientists, for example. There are many reasons, many exogenous to the network, why two scientists who have never written a paper together will do so in the next few years: for example, they may happen to become geographically close when one of

The work described in this chapter was done jointly with Jon Kleinberg. An abbreviated version of this joint work appears as "The Link Prediction Problem for Social Networks," *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM'03)*, November 2003, pp. 556–559.

them changes institutions, or they may happen to meet when they both attend a conference. Such collaborations can be hard to predict. But one also senses that a large number of new collaborations are hinted at by the topology of the network: two scientists who are "close" in the network will have colleagues in common and will travel in similar circles; this social proximity suggests that they themselves are more likely to collaborate in the near future. Our goal is to make this intuitive notion precise and to understand which measures of "proximity" in a network lead to the most accurate link predictions. We find that a number of proximity measures lead to predictions that outperform chance by factors of forty to fifty, indicating that the network topology does indeed contain latent information from which to infer future interactions. Moreover, certain fairly subtle measures—involving infinite sums over paths in the network—often outperform more direct measures, such as shortest-path distances and numbers of shared neighbors.

We believe that a primary contribution of the work described in this chapter is in the area of network evolution models. While there has been a proliferation of such models in recent years—see, for example, the work of Jin et al. [90], Barabási et al. [21], and Davidsen et al. [43] for recent work on collaboration networks, or the survey of Newman [144]—they have generally been evaluated only by asking whether they reproduce certain global structural features observed in real networks. As a result, it has been difficult to evaluate and compare different approaches on a principled footing. Link prediction, on the other hand, offers a very natural basis for such evaluations: *a network model is useful to the extent that it can support meaningful inferences from observed network data.* One sees a related approach in recent work of Newman [137], who considers the correlation between, on one hand, the output of certain network growth models (like preferential attachment) and, on the other hand, data on the appearance of edges of coauthorship networks.

In addition to its role as a basic question in social network evolution, the link-prediction problem could be relevant to a number of interesting current applications of social networks. Increasingly, for example, researchers in artificial intelligence and data mining have argued that a large organization, such as a company, can benefit from the interactions within the informal social network among its members; these interactions serve to supplement the official hierarchy imposed by the organization itself [95, 157]. Effective methods for link prediction could be used to analyze such a social network and suggest promising opportunities for interaction or collaboration that have not yet been identified within the organization. In a different vein, research in security has recently begun to emphasize the role of social-network analysis, largely motivated by the problem of monitoring terrorist networks; link prediction in this context allows one to conjecture that particular individuals are working together even though their interaction has not been directly observed [111].

The link-prediction problem is also related to the problem of inferring missing links from an observed network: in a number of domains, one constructs a network of interactions based on observable data and then tries to infer additional links that, while not directly visible, are likely to exist [69, 155, 166]. This line of work differs from our problem formulation in that it works with a static snapshot of a network, rather than considering network evolution; it also tends to take into account specific attributes of the nodes in the network, rather than evaluating the power of prediction methods based purely on the graph structure.

We now turn to a description of our experimental setup, in Section 3.2. Our primary focus is on understanding the relative effectiveness of network-proximity measures adapted from techniques in graph theory, computer science, and the social sciences, and we review a large number of such techniques in Section 3.3. Finally, we discuss the results of our experiments in Section 3.4.

3.2 Data and Experimental Setup

Suppose we have a social network $G = \langle V, E \rangle$ in which each edge $e = \langle u, v \rangle \in E$ represents an interaction between u and v that took place at a particular time $t(e)$. Such time-stamped interactions are the basis of many interesting social networks, like those where an edge $\langle u, v \rangle$ denotes an underlying social relationship like "u went on a date with v at time t" or "u emailed v at time t." (Note that in the email-network example, unlike in the social networks we consider here, friendship is a directed notion, though we can choose to discard the directions of edges in such a network.) If there are multiple interactions between u and v, we will record these interactions as a set of distinct parallel edges, each of which joins u and v with a potentially different timestamp. For two times t and $t' > t$, let $G[t, t']$ denote the subgraph of G consisting of all edges with a timestamp between t and t'.

Here, then, is a concrete formulation of the link-prediction problem. Let t_0, t_0', t_1, and t_1' be four times, where $t_0 < t_0' \leq t_1 < t_1'$, and let $G = \langle V, E \rangle$ be a social network. We give a link-prediction algorithm access to the network $G[t_0, t_0']$; it must then output a list of edges not present in $G[t_0, t_0']$ that are predicted to appear in the network $G[t_1, t_1']$. We refer to $[t_0, t_0']$ as the *training interval* and $[t_1, t_1']$ as the *test interval.*

Of course, social networks grow through the addition of nodes as well as edges, and it is not sensible to seek predictions for edges whose endpoints are not present in the training interval. Thus, in evaluating link-prediction methods, we will generally use two parameters $\kappa_{training}$ and κ_{test} and define the set Core to be all nodes incident to at least $\kappa_{training}$ edges in $G[t_0, t_0']$ and at least κ_{test} edges in $G[t_1, t_1']$. We will then evaluate how accurately the new edges between elements of Core can be predicted. (We only attempt to predict the formation of edges between nodes that persist throughout the entire duration of the training and test intervals, although predicting when a node will stop acquiring friends—e.g., retire from scientific research—is an interesting question that we do not explore here.)

3.2.1 Data Sets for Link-Prediction Experiments

We now describe our experimental setup more specifically. We work with coauthorship networks G obtained from the author lists of papers in the physics e-Print arXiv, `www.arxiv.org` [64, 65]. The arXiv is an electronic server for papers, primarily in physics. Authors submit their own papers to the server, without peer review. When an author uploads a paper, he or she designates a subfield into which the paper falls. In our experiments, we consider five different sections of the arXiv: astrophysics, condensed matter (incidentally the arXiv section used by physicists when they write papers about social networks), general relativity and quantum cosmology, high-energy physics—phenomenology, and high-energy physics—theory. See Figure 3-1 for statistics on the sizes of each of the five networks that we derive from the arXiv.

Some heuristics were used to deal with occasional syntactic anomalies, and authors were identified by first initial and last name, a process that introduces a small amount of noise due to multiple authors with the same identifier [138]. The errors introduced by this process appear to be minor, and performance should only improve with a more accurate listing of the authors.

Now consider any one of these five graphs. We define the training interval to be the three years [1994, 1996], and the test interval to be [1997, 1999]. Let A denote the set of all authors who have written a paper during the entire period of relevance, from 1994 to 1999. We denote the subgraph $G[1994, 1996]$ on the training interval by $G_{collab} := \langle A, E_{old} \rangle$, and use E_{new} to denote the set of

	training period			Core		
	authors	papers	edges	authors	$\|E_{old}\|$	$\|E_{new}\|$
astro-ph	5343	5816	41852	1561	6178	5751
cond-mat	5469	6700	19881	1253	1899	1150
gr-qc	2122	3287	5724	486	519	400
hep-ph	5414	10254	17806	1790	6654	3294
hep-th	5241	9498	15842	1438	2311	1576

Figure 3-1: *The five sections of the arXiv from which coauthorship networks were constructed:* `astro-ph` *(astrophysics),* `cond-mat` *(condensed matter),* `gr-qc` *(general relativity and quantum cosmology),* `hep-ph` *(high energy physics—phenomenology), and* `hep-th` *(high energy physics—theory). The training period consists of the three-year period from 1994 to 1996, and the test period consists of the three-year period from 1997 to 1999. The set* Core *denotes the subset of the authors who have written at least* $\kappa_{training} = 3$ *papers during the training period and* $\kappa_{test} = 3$ *papers during the test period. The sets* E_{old} *and* E_{new} *denote edges between a pair of* Core *authors that first appear during the training and test periods, respectively. The elements of* E_{new} *are the new collaborations that we seek to predict.*

edges $\langle u, v \rangle$ such that $u, v \in A$, and u, v coauthor a paper during the test interval but not the training interval. The author pairs in E_{new} are the new interactions that we are seeking to predict. In our experiments on the arXiv, we can identify which authors are active throughout the entire period on the basis of the number of papers published and not on the number of coauthors. Thus we define the set Core to consist of all authors who have written at least $\kappa_{training} := 3$ papers during the training period and at least $\kappa_{test} := 3$ papers during the test period.

3.2.2 Evaluating a Link Predictor

With the above definitions in hand, we can now fully describe the link-prediction problem and the way in which we measure performance.

Each link predictor p that we consider takes as input the graph G_{collab} and outputs a ranked list L_p of pairs in $(A \times A) - E_{old}$; the list L_p contains its predicted new collaborations, in decreasing order of confidence.

For our evaluation, we focus on the set Core. Define $E^*_{new} := E_{new} \cap (\text{Core} \times \text{Core})$ to be the set of all new core/core interactions, and define $n := |E^*_{new}|$ to be the number of these interactions—that is, the total number of possible correct answers. Our performance measure for predictor p is then determined as follows: from the ranked list L_p, we take the first n pairs in Core $\times$ Core, and determine the size of the intersection of this set of pairs with the set E^*_{new}.

3.3 Methods for Link Prediction

Now that we have formally defined the link-prediction problem, in this section we survey an array of methods for link prediction. For concreteness, we describe all of the predictors in the following terms. All the methods assign a connection weight $\text{score}(x, y)$ to pairs of nodes $\langle x, y \rangle$ based on the input graph G_{collab}, and then produce a ranked list in decreasing order of $\text{score}(x, y)$. Predictions

are made according to the ordering in this list; thus, our predictors can be viewed as computing a measure of proximity or "similarity" between nodes x and y, relative to the network topology. Although we use these proximity measures for link prediction, they of course can be used for a variety of other applications like collaborative filtering [116] and identifying a "connection subgraph" that best explains the relationship between two nodes in the social network [60].

The link-prediction methods outlined here are in general adapted from techniques used in graph theory and social-network analysis; in many cases, these techniques were not designed to measure node-to-node similarity and hence need to be modified for this purpose. Figure 3-2 summarizes the basic measures we explore in this chapter. We discuss these basic predictors in more detail in Sections 3.3.2 and 3.3.3, and in Section 3.3.4 we also discuss some meta-level approaches to link prediction that extend these basic predictors.

We note here that some of the measures discussed in this section are designed only for connected graphs, and $\mathsf{score}(x, y)$ may be undefined for nodes x and y that are disconnected in G_{collab}. All of the graphs that we consider in our experiments—and indeed virtually all social networks—have a giant component, a single connected component containing the majority of the nodes; it is therefore natural to restrict the predictions for these measures to this component.

3.3.1 The Graph-Distance Predictor

Perhaps the most basic approach to measuring proximity of nodes in a social network is by measuring the *graph distance* between them. That is, we rank pairs $\langle x, y \rangle$ by the length of the shortest path connecting them in G_{collab}. Such a measure follows the notion that collaboration networks are small worlds, in which individuals are related through short chains. (See Chapter 2.) In keeping with the notion that we rank pairs in *decreasing* order of $\mathsf{score}(x, y)$, we define $\mathsf{score}(x, y)$ here to be the *negation* of the shortest-path length.

Pairs with shortest-path distance equal to one are joined by an edge in G_{collab}, and hence they belong to the training edge set E_{old}. For all of our graphs G_{collab}, there are well more than n pairs with a shortest-path distance of two, so our shortest-path predictor simply selects a random subset of these distance-two pairs.

3.3.2 Predictors Based on Node Neighborhoods

For a node x, let $\Gamma(x)$ denote the set of neighbors of x in G_{collab}. A number of approaches to link prediction are based on the idea that two nodes x and y are more likely to form a link in the future if their sets of neighbors $\Gamma(x)$ and $\Gamma(y)$ have large overlap. This style of approach follows the natural intuition that such nodes x and y represent authors with many colleagues in common, who hence are more likely to come into contact themselves.

The notion of friendship formation through common acquaintances has been used to justify the high clustering coefficients of social networks [174], and Jin et al. [90] and Davidsen et al. [43] have defined abstract models for network growth using this principle, in which an edge $\langle x, y \rangle$ is more likely to form if edges $\langle x, z \rangle$ and $\langle z, y \rangle$ are already present for some z.

Common Neighbors

The most direct implementation of this idea for the link-prediction problem is the *common-neighbors predictor*, under which we define $\mathsf{score}(x, y) := |\Gamma(x) \cap \Gamma(y)|$, the number of neighbors that x and y have in common.

Basic graph-distance method:	
graph distance	(negated) length of shortest path between x and y
Methods based upon node neighborhoods:	
common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Adamic/Adar	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log\lvert\Gamma(z)\rvert}$
preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Methods based upon the ensemble of all paths:	
Katz_β	$\sum_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\mathsf{paths}_{x,y}^{\langle\ell\rangle}\rvert$ where $\mathsf{paths}_{x,y}^{\langle\ell\rangle}$:= {paths of length exactly ℓ from x to y} with multiplicities in the weighted case, so that, e.g.: *weighted:* $\mathsf{paths}_{x,y}^{\langle 1\rangle}$:= number of connections between x, y. *unweighted:* $\mathsf{paths}_{x,y}^{\langle 1\rangle} := 1$ iff x and y are connected.
hitting time $\sim$ stationary-normed	$-H_{x,y}$ $-H_{x,y} \cdot \pi_y$
commute time $\sim$ stationary-normed	$-(H_{x,y} + H_{y,x})$ $-(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$ where $H_{x,y}$:= expected time for random walk from x to reach y π_y := stationary distribution weight of y (proportion of time the random walk is at node y)
rooted PageRank_α	stationary distribution weight of y under the following random walk: with probability α, jump to x. with probability $1-\alpha$, go to a random neighbor of the current node.
SimRank_γ	the fixed point of the following recursive definition of $\mathsf{score}(x, y)$: $\mathsf{score}(x,y) = \begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} \mathsf{score}(a,b)}{\lvert\Gamma(x)\rvert\cdot\lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$

Figure 3-2: *Summary of basic predictors for the link-prediction problem. For each listed predictor, the value of* $\mathsf{score}(x, y)$ *is shown; each predictor predicts pairs* $\langle x, y\rangle$ *in descending order of* $\mathsf{score}(x, y)$*. The set* $\Gamma(x)$ *denotes the neighbors of the node* x *in* G_{collab}*. See Sections 3.3.2 and 3.3.3.*

The common-neighbors predictor captures the notion that two strangers who have a common friend may be introduced by that friend. This introduction has the effect of "closing a triangle" in the graph and feels like a common mechanism in real life. Newman [137] has computed this quantity in the context of collaboration networks, verifying a positive correlation between the number of common neighbors of x and y at time t, and the probability that x and y will collaborate at some time after t.

Jaccard's Coefficient

The Jaccard coefficient—a similarity metric that is commonly used in information retrieval [159]—measures the probability that both x and y have a feature f, for a randomly selected feature f that *either* x or y has. If we take "features" here to be neighbors in G_{collab}, then this measure captures the intuitively appealing notion that the *proportion* of the coauthors of x who have also worked with y (and vice versa) is a good measure of the similarity of x and y. Formally, the *Jaccard predictor* uses the following measure:

$$\mathsf{score}(x, y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}.$$

Frequency-Weighted Common Neighbors (Adamic/Adar)

Adamic and Adar [2] consider a measure similar to Jaccard's coefficient, in the context of deciding when two personal home pages are strongly "related." To do this calculation, they compute features of the pages, and define the similarity between two pages to be the following:

$$\mathsf{similarity}(x, y) := \sum_{z \,:\, \text{feature shared by } x, y} \frac{1}{\log(\mathsf{frequency}(z))}.$$

This measure refines the simple counting of common features by weighting rarer features more heavily. This approach suggests the measure $\mathsf{score}(x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$, which we will simply call the *Adamic/Adar predictor.*

The Adamic/Adar predictor formalizes the intuitive notion that rare features are more telling; documents that share the phrase "for example" are probably less similar than documents that share the phrase "clustering coefficient." If "triangle closing" is a frequent mechanism by which new edges form in a social network, then for x and y to be introduced by a common friend z, person z will have to choose to introduce the pair $\langle x, y \rangle$ from the $\binom{|\Gamma(z)|}{2}$ pairs of his friends; thus an unpopular person may be more likely to introduce a particular pair of his friends to each other.

Preferential Attachment

Preferential attachment has received considerable attention as a model of the growth of networks [128]. The basic premise is that the probability that a new edge involves node x is proportional to $|\Gamma(x)|$, the current number of neighbors of x. (See Section 1.3.4.)

Newman [137] and Barabási et al. [21, 89] have further proposed, on the basis of empirical evidence, that the probability of coauthorship of x and y is positively correlated with the product of the number of collaborators of x and y. In our context, this notion corresponds to the *preferential-attachment* measure $\mathsf{score}(x, y) := |\Gamma(x)| \cdot |\Gamma(y)|$.

3.3.3 Methods Based on the Ensemble of All Paths

A number of methods refine the notion of shortest-path distance by implicitly considering the ensemble of *all* paths between two nodes. Here, we describe the adaptation of several such methods to the link-prediction problem.

Most of the methods that we discuss in this section can be calculated by computing the inverse of an $|A|$-by-$|A|$ matrix, where A denotes the set of authors in G_{collab}. In our experiments, we compute these inverses using the built-in Matlab `inv()` function; this calculation was one of our biggest computational bottlenecks.

Exponentially Damped Path Counts (Katz)

Katz [94]—originally motivated by the problem of measuring social standing based upon an explicit set of "endorsements" of one person by another—defines a measure that directly sums over this collection of paths, exponentially damped by length to count short paths more heavily. This idea leads to the *Katz predictor*:

$$\mathsf{score}(x, y) := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\mathsf{paths}_{x,y}^{\langle \ell \rangle}|,$$

where $\mathsf{paths}_{x,y}^{\langle \ell \rangle}$ denotes the set of all length-ℓ paths from x to y. (A very small β yields predictions much like common neighbors, because paths of length three or more contribute very little to the summation.) It is straightforward to verify that the matrix of scores is given by $(I - \beta M)^{-1} - I$, where M is the adjacency matrix of the graph:

$$\begin{aligned}
\mathsf{score} &= \sum_{\ell=1}^{\infty} \beta^{\ell} M^{\ell} = \beta M \left[\sum_{\ell=0}^{\infty} \beta^{\ell} M^{\ell} \right] = \beta M [I + \mathsf{score}] \\
\mathsf{score} + I &= \beta M [I + \mathsf{score}] + I \\
(\mathsf{score} + I)(I - \beta M) &= I \\
\mathsf{score} &= (I - \beta M)^{-1} - I.
\end{aligned}$$

This formulation allows us to compute the Katz measure efficiently. Note that the matrix $I - \beta M$ that we invert is sparse because the original adjacency matrix is sparse in a social network.

We consider two variants of this Katz measure: (1) *unweighted*, in which $M_{x,y} = \mathsf{paths}_{x,y}^{\langle 1 \rangle} = 1$ if x and y have collaborated and 0 otherwise, and (2) *weighted*, in which $M_{x,y} = \mathsf{paths}_{x,y}^{\langle 1 \rangle}$ is the number of times that x and y have collaborated. We also let the parameter β take on a range of different values.

Hitting Time and Commute Time

A *random walk* on G_{collab} starts at a node x and iteratively moves to a neighbor of x chosen uniformly at random [134]. The *hitting time* $H_{x,y}$ from x to y is the expected number of steps required for a random walk starting at x to reach y. Because the hitting time is not in general symmetric, it is also natural to consider the *commute time* $C_{x,y} := H_{x,y} + H_{y,x}$. Both of these measures serve as natural proximity measures, and hence (negated) can be used as $\mathsf{score}(x, y)$.

One difficulty with hitting time as a measure of proximity is that $H_{x,y}$ is quite small whenever y is a node with a large stationary probability π_y, regardless of the identity of x. (That is, for a

node y at which the random walk spends a considerable amount of time in the limit, the random walk will soon arrive at y, almost no matter where it starts. Thus the predictions made based upon $H_{x,y}$ tend to include only a few distinct nodes y.) To counterbalance this phenomenon, we also consider *normalized* versions of the hitting and commute times, by defining $\mathsf{score}(x,y) := -H_{x,y} \cdot \pi_y$ or $\mathsf{score}(x,y) := -(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$.

We compute the hitting-time matrix using the algorithm of Tetali [167], which allows the computation of the matrix H via a single matrix inversion, with the stationary distribution and the matrix of transition probabilities as input. The transition probabilities are straightforwardly computable from the adjacency matrix M of the graph, and the stationary-distribution weight of a node u in an undirected graph $G = \langle V, E \rangle$ is given by $\Gamma(u)/2|E|$. We note that the matrix that we invert in the Tetali algorithm again preserves the sparsity of the adjacency matrix M, which improves the efficiency of this computation.

PageRank and Rooted PageRank

Another difficulty with using the measures based on hitting time and commute time is their sensitive dependence to parts of the graph far away from x and y, even when x and y are connected by very short paths. A way of counteracting this difficulty is to allow the random walk from x to y to periodically "reset," returning to x with a fixed probability α at each step; in this way, distant parts of the graph will almost never be explored.

Random resets form the basis of the *PageRank* measure for web pages [35, 152], and we can adapt it for link prediction as follows. Define $\mathsf{score}(x,y)$ under the *rooted-PageRank predictor* with parameter α to be the stationary probability of y in a random walk that returns to x with probability α at each step, moving to a random neighbor of the current node with probability $1 - \alpha$. Similar approaches have been considered for *personalized PageRank*, in which one wishes to rank web pages based both on overall "importance" (the core of PageRank) and relevance to a particular topic or individual, by biasing the random resets towards topically relevant or bookmarked pages [82, 83, 88, 91].

Computing predictions for the rooted-PageRank measure requires us to compute the stationary distribution of a different random walk for each "root" node x, because the random resets are to a different root in the random walk for each $\mathsf{score}(x,\cdot)$. Each of these computations requires an eigenvector computation on the transition matrix for the random walk, and incurs a significant computational load. We might be interested in using a rooted-PageRank style of predictor using the hitting time $H_{x,y}$ in the x-rooted random walk instead of the stationary-distribution weight of node y in this walk; however, computationally the load for computing the inverses of $|A|$ different $|A|$-by-$|A|$ matrices—where, again, A denotes the set of authors—is too excessive to be feasible. (The literature on personalized PageRank is largely concerned with making the stationary-distribution computation efficient through approximation; because our social networks are significantly smaller than the web, we can afford the stationary-distribution computations, but the hitting-time calculations are too computationally demanding.)

Similarity of Nodes Based on Neighbors' Similarity: SimRank

SimRank [87] was developed as way to compute node similarity in contexts in which there are "structural-context" indicators of similarity. SimRank is a fixed point of the following recursive definition: two nodes are similar to the extent that they are joined to similar neighbors. Numerically,

this notion is specified by defining

$$\begin{aligned} \mathsf{similarity}(x,x) &:= 1 \\ \mathsf{similarity}(x,y) &:= \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \mathsf{similarity}(a,b)}{|\Gamma(x)| \cdot |\Gamma(y)|} \end{aligned}$$

for some parameter $\gamma \in [0,1]$. We then define $\mathsf{score}(x,y) := \mathsf{similarity}(x,y)$. Computation of SimRank scores is done iteratively, by unwinding the above recursive definition [87].

SimRank can also be interpreted in terms of a random walk on the collaboration graph: it is the expected value of γ^ℓ, where ℓ is a random variable giving the time at which two synchronized random walks that are started from x and y first arrive simultaneously at the same node.

3.3.4 Higher-Level Approaches

We now discuss three "meta-approaches" that can be used in conjunction with any of the methods discussed above. The goal of these higher-level approaches is to reduce the amount of noise inherent in the data; we hope that noise-reduction techniques can improve the performance of the basic predictors above.

Low-Rank Approximation

Because the adjacency matrix M can be used to represent the graph G_{collab}, all of our link-prediction methods have an equivalent formulation in terms of this matrix M. In some cases, this formulation was noted explicitly above (for example in the case of the Katz similarity score), but in many cases the matrix formulation is quite natural. For example, the common-neighbors method consists simply of mapping each node x to its row $r(x)$ in M, and then defining $\mathsf{score}(x,y)$ to be the inner product of the rows $r(x)$ and $r(y)$.

A common general technique when analyzing the structure of a large matrix M is to choose a relatively small number k and compute the rank-k matrix M_k that best approximates M with respect to any of a number of standard matrix norms. This computation can be done efficiently using the singular-value decomposition, and it forms the core of methods like *latent semantic analysis* in information retrieval [45]. Intuitively, working with M_k rather than M can be viewed as a type of "noise-reduction" technique that generates most of the structure in the matrix but with a greatly simplified representation.

In our experiments, we investigate three applications of low-rank approximation: (i) ranking by the Katz measure, in which we use M_k rather than M in the underlying formula; (ii) ranking by common neighbors, in which we score by inner products of rows in M_k rather than M; and—most simply of all—(iii) defining $\mathsf{score}(x,y)$ to be the $\langle x,y \rangle$ entry in the matrix M_k.

Unseen Bigrams

An analogy can be seen between link prediction and the problem in language modeling of estimating frequencies for *unseen bigrams*—pairs of words that cooccur in a test corpus, but not in the corresponding training corpus (see, e.g., the work of Essen and Steinbiss [59]). In this problem, the data consist of a collection of pairs of words that appear in adjacent positions in training text, and one aims to extend frequency estimates for these observed pairs to frequency estimates for unobserved pairs. (For example, if one observes that "a" and "the" have extremely similar distributions in the

training set—say, both appear very frequently before words like "friend" and "predictor"—then one can infer from a high frequency of "a matrix" that "the matrix" should have high frequency, even if the bigram "the matrix" never appears in the training set.)

Following ideas proposed in that literature ([117], for example), we can augment our estimates of $\mathsf{score}(x, y)$ using values of $\mathsf{score}(z, y)$ for nodes z that are "similar" to x. Specifically, we adapt this approach to the link-prediction problem as follows. Suppose we have values $\mathsf{score}(x, y)$ computed under one of the measures above. Let $S_x^{\langle\delta\rangle}$ denote the δ nodes most related to x under $\mathsf{score}(x, \cdot)$, for a parameter $\delta \in \mathbb{Z}^+$. We then define enhanced scores in terms of the nodes in this set:

$$\begin{aligned} \mathsf{score}^*_{unweighted}(x, y) &:= \left|\Gamma(y) \cap S_x^{\langle\delta\rangle}\right| \\ \mathsf{score}^*_{weighted}(x, y) &:= \sum_{z \in \Gamma(y) \cap S_x^{\langle\delta\rangle}} \mathsf{score}(x, z). \end{aligned}$$

Clustering

One might seek to improve on the quality of a predictor by deleting the more "tenuous" edges in G_{collab} through a clustering procedure, and then running the predictor on the resulting "cleaned-up" subgraph. Because our predictors themselves are measures of node proximity, we base our clustering on the output of the predictor itself.

Consider a measure computing values for $\mathsf{score}(x, y)$. We compute $\mathsf{score}(u, v)$ for all edges in E_{old}, and delete the $(1 - \rho)$ fraction of these edges for which the score is lowest, for some parameter $\rho \in [0, 1]$. We now recompute $\mathsf{score}(x, y)$ for all pairs $\langle x, y\rangle$ on this subgraph; in this way we determine node proximities using only edges for which the proximity measure itself has the most confidence.

3.4 Results and Discussion

In this section, we discuss the performance of our predictors in predicting new links in the physics collaboration networks. It rapidly becomes apparent that the formation of any particular new edge is unlikely, and many collaborations form (or fail to form) for reasons outside the scope of the network; thus the raw performance of our predictors is relatively low. To more meaningfully represent predictor quality, we use as our baseline a *random predictor* that simply randomly selects pairs of authors who did not collaborate in the training interval as its predictions. A random prediction is correct with probability between 0.15% (`cond-mat`) and 0.48% (`astro-ph`).

Figures 3-3 and 3-4 show each predictor's performance on each arXiv section, in terms of the factor improvement over random. Figures 3-5, 3-6, and 3-7 show the average relative performance of several different predictors versus three baseline predictors—the random predictor, the graph-distance predictor, and the common-neighbors predictor. There is no single clear winner among the techniques, but we see that a number of methods significantly outperform the random predictor, suggesting that there is indeed useful information contained in the network topology alone. The Katz measures and its variants based on clustering and low-rank approximation perform consistently well; on three of the five arXiv sections, a variant of Katz achieves the best performance. Some of the very simple measures also perform surprisingly well, including common neighbors and, especially, the Adamic/Adar measure.

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability that a random prediction is correct		0.475%	0.147%	0.341%	0.207%	0.153%
graph distance (all distance-two pairs)		*9.6*	*25.3*	*21.4*	*12.2*	*29.2*
common neighbors		***18.0***	***41.1***	***27.2***	***27.0***	***47.2***
preferential attachment		4.7	6.1	7.6	*15.2*	7.5
Adamic/Adar		*16.8*	***54.8***	***30.1***	***33.3***	***50.5***
Jaccard		*16.4*	***42.3***	19.9	***27.7***	*41.7*
SimRank	$\gamma = 0.8$	*14.6*	*39.3*	*22.8*	*26.1*	*41.7*
hitting time		6.5	23.8	*25.0*	3.8	13.4
hitting time, stationary-distribution normed		5.3	23.8	11.0	11.3	21.3
commute time		5.2	15.5	***33.1***	*17.1*	23.4
commute time, stationary-distribution normed		5.3	16.1	11.0	11.3	16.3
rooted PageRank	$\alpha = 0.01$	*10.8*	*28.0*	***33.1***	*18.7*	*29.2*
	$\alpha = 0.05$	*13.8*	*39.9*	***35.3***	*24.6*	*41.3*
	$\alpha = 0.15$	*16.6*	***41.1***	***27.2***	***27.6***	*42.6*
	$\alpha = 0.30$	*17.1*	***42.3***	*25.0*	***29.9***	*46.8*
	$\alpha = 0.50$	*16.8*	***41.1***	*24.3*	***30.7***	*46.8*
Katz (weighted)	$\beta = 0.05$	3.0	21.4	19.9	2.4	12.9
	$\beta = 0.005$	*13.4*	***54.8***	***30.1***	*24.0*	***52.2***
	$\beta = 0.0005$	*14.5*	***54.2***	***30.1***	***32.6***	***51.8***
Katz (unweighted)	$\beta = 0.05$	*10.9*	***41.7***	***37.5***	*18.7*	***48.0***
	$\beta = 0.005$	*16.8*	***41.7***	***37.5***	*24.2*	***49.7***
	$\beta = 0.0005$	*16.8*	***41.7***	***37.5***	*24.9*	***49.7***

Figure 3-3: *Performance of the basic predictors on the link-prediction task defined in Section 3.2. See Sections 3.3.1, 3.3.2, and 3.3.3 for definitions of these predictors. For each predictor and each arXiv section, the displayed number specifies the factor improvement over random prediction. Two predictors in particular are used as baselines for comparison: graph distance and common neighbors. Italicized entries have performance at least as good as the graph-distance predictor; bold entries are at least as good as the common-neighbors predictor. See also Figure 3-4.*

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability a random prediction is correct		0.475%	0.147%	0.341%	0.207%	0.153%
graph distance (all distance-two pairs)		*9.6*	*25.3*	*21.4*	*12.2*	*29.2*
common neighbors		***18.0***	***41.1***	***27.2***	***27.0***	***47.2***
Low-rank approximation: Inner product	rank = 1024	*15.2*	***54.2***	***29.4***	***34.9***	***50.1***
	rank = 256	*14.6*	***47.1***	***29.4***	***32.4***	***47.2***
	rank = 64	*13.0*	***44.7***	***27.2***	***30.8***	***47.6***
	rank = 16	*10.1*	21.4	***31.6***	***27.9***	*35.5*
	rank = 4	8.8	15.5	***42.6***	*19.6*	23.0
	rank = 1	6.9	6.0	***44.9***	*17.7*	14.6
Low-rank approximation: Matrix entry	rank = 1024	8.2	16.7	6.6	*18.6*	21.7
	rank = 256	*15.4*	*36.3*	8.1	*26.2*	*37.6*
	rank = 64	*13.8*	***46.5***	16.9	***28.1***	*40.9*
	rank = 16	9.1	21.4	*26.5*	*23.1*	*34.2*
	rank = 4	8.8	15.5	***39.7***	*20.0*	22.5
	rank = 1	6.9	6.0	***44.9***	*17.7*	14.6
Low-rank approximation: Katz ($\beta = 0.005$)	rank = 1024	*11.4*	*27.4*	***30.1***	***27.1***	*32.1*
	rank = 256	*15.4*	***42.3***	11.0	***34.3***	*38.8*
	rank = 64	*13.1*	***45.3***	19.1	***32.3***	*41.3*
	rank = 16	9.2	21.4	***27.2***	*24.9*	*35.1*
	rank = 4	7.0	15.5	***41.2***	*19.7*	23.0
	rank = 1	0.4	6.0	***44.9***	*17.7*	14.6
unseen bigrams (weighted)	common neighbors, $\delta = 8$	*13.5*	*36.9*	***30.1***	*15.6*	***47.2***
	common neighbors, $\delta = 16$	*13.4*	*39.9*	***39.0***	*18.6*	***48.8***
	Katz ($\beta = 0.005$), $\delta = 8$	*16.9*	*38.1*	*25.0*	*24.2*	***51.3***
	Katz ($\beta = 0.005$), $\delta = 16$	*16.5*	*39.9*	***35.3***	*24.8*	***50.9***
unseen bigrams (unweighted)	common neighbors, $\delta = 8$	*14.2*	*40.5*	***27.9***	*22.3*	*39.7*
	common neighbors, $\delta = 16$	*15.3*	*39.3*	***42.6***	*22.1*	*42.6*
	Katz ($\beta = 0.005$), $\delta = 8$	*13.1*	*36.9*	***32.4***	*21.7*	*38.0*
	Katz ($\beta = 0.005$), $\delta = 16$	*10.3*	*29.8*	***41.9***	*12.2*	*38.0*
clustering: Katz ($\beta_1 = 0.001, \beta_2 = 0.1$)	$\rho = 0.10$	7.4	*37.5*	***47.1***	***33.0***	*38.0*
	$\rho = 0.15$	*12.0*	***46.5***	***47.1***	*21.1*	*44.2*
	$\rho = 0.20$	4.6	*34.5*	19.9	*21.2*	*35.9*
	$\rho = 0.25$	3.3	*27.4*	20.6	*19.5*	17.5

Figure 3-4: *Performance of various meta-approaches discussed in Section 3.3.4 on the link prediction task defined in Section 3.2. As before, for each predictor and each arXiv section, the given number specifies the factor improvement over random prediction. Again, italicized entries have performance at least as good as the graph-distance predictor; bold entries are at least as good as the common-neighbors predictor. See also Figure 3-3.*

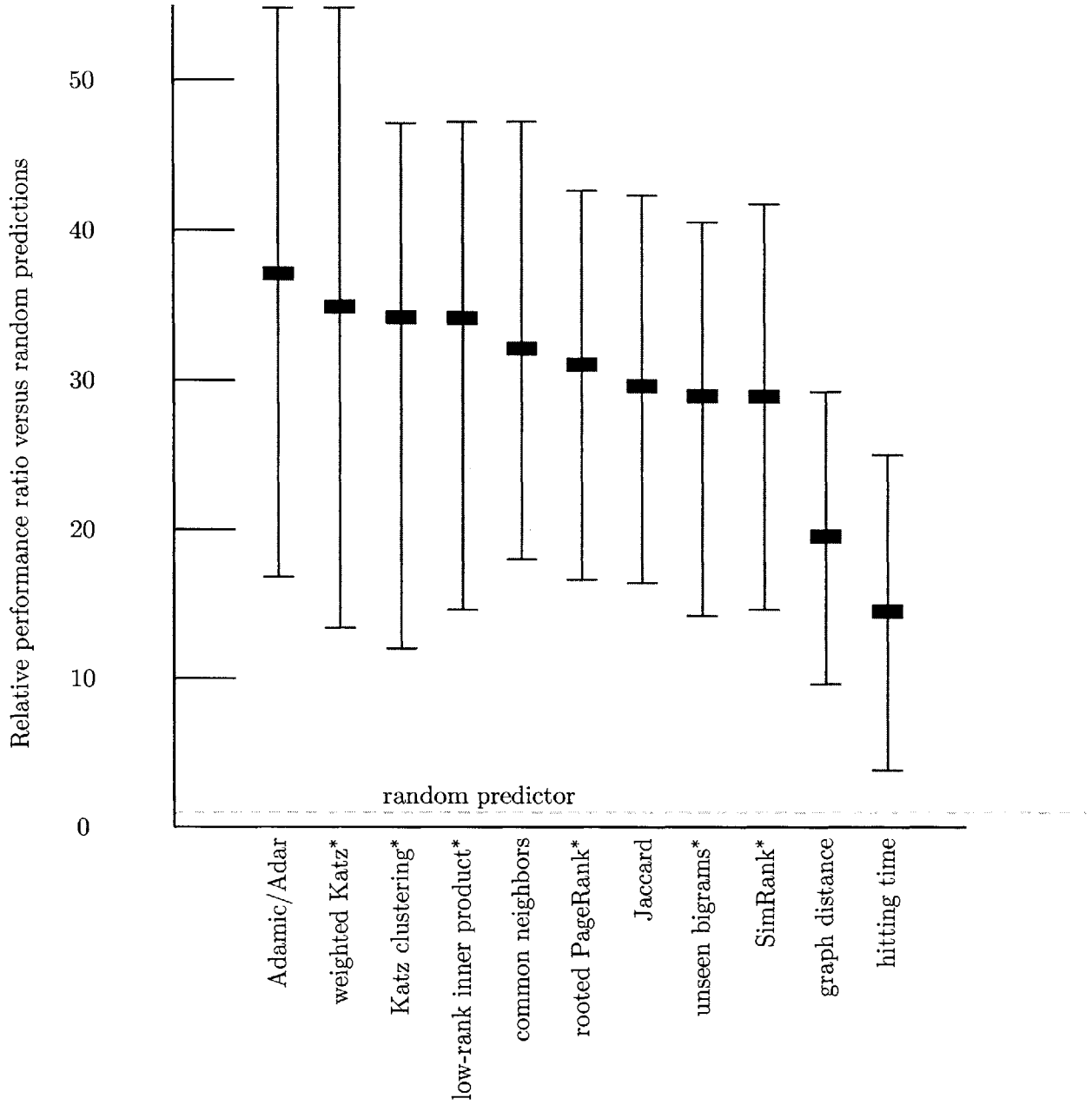


Figure 3-5: *Relative average performance of various predictors versus random predictions. The value shown is the average ratio over the five datasets of the given predictor's performance versus the random predictor's performance. The error bars indicate the minimum and maximum of this ratio over the five datasets. The parameters for the starred predictors are: (1) for weighted Katz,* $\beta = 0.005$*; (2) for Katz clustering,* $\beta_1 = 0.001, \rho = 0.15, \beta_2 = 0.1$*; (3) for low-rank inner product,* rank $= 256$*; (4) for rooted PageRank,* $\alpha = 0.15$*; (5) for unseen bigrams, unweighted common neighbors with* $\delta = 8$*; and (6) for SimRank,* $\gamma = 0.8$*.*

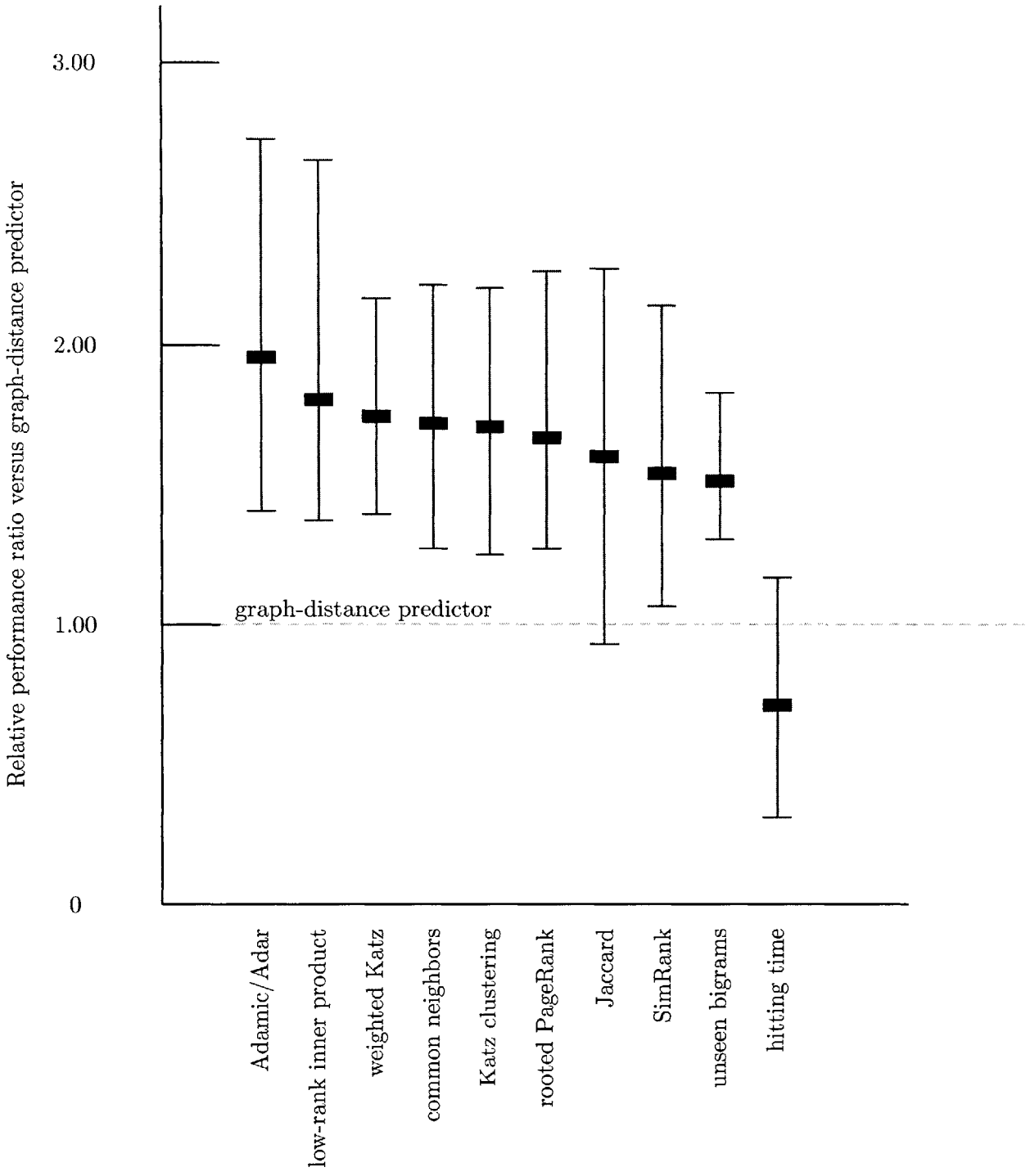


Figure 3-6: *Relative average performance of various predictors versus the graph-distance predictor. The plotted value shows the average taken over the five datasets of the ratio of the performance of the given predictor versus the graph-distance predictor; the error bars indicate the range of this ratio over the five datasets. All parameter settings are as in Figure 3-5.*

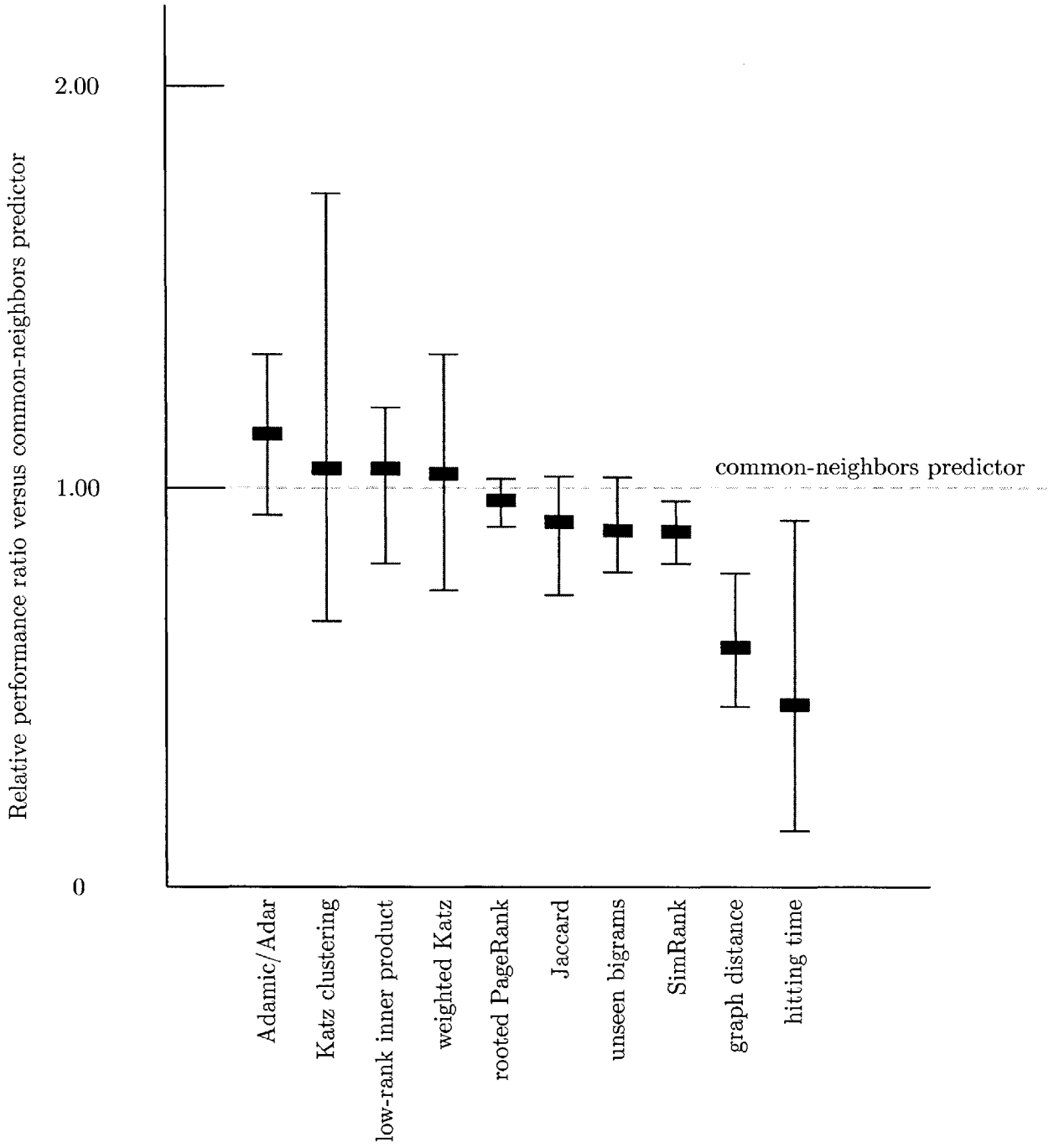


Figure 3-7: *Relative average performance of various predictors versus the common-neighbors predictor, as in Figure 3-6. Error bars display the range of the performance ratio of the given predictor versus common neighbors over the five datasets; the displayed value gives the average ratio. Parameter settings are as in Figure 3-5.*

3.4.1 The Small-World Problem

It is reassuring that even the basic graph-distance predictor handily outperforms random predictions, but this measure has severe limitations. Extensive research has been devoted to understanding the small-world problem in collaboration networks—i.e., accounting for the existence of short paths connecting virtually every pair of scientists [138]. (See Section 1.1.) This small-world property is normally viewed as a vital fact about the scientific community (new ideas spread quickly, and every discipline interacts with and gains from other fields), but in the context of our prediction task we come to a different conclusion: the small-world problem is really a problem.

The shortest path between two scientists in wholly unrelated disciplines is often very short, but also very tenuous. (These tenuous "weak ties" are exactly the connections between social groups that Mark Granovetter found to be crucial in organizing a community or in getting a job [73, 74].) For example, we might hope that small Erdős numbers would distinguish researchers in mathematics and computer science, but to take one particular (but not atypical) example, the developmental psychologist Jean Piaget has as small an Erdős Number—three [40]—as most mathematicians and computer scientists. Overall, the basic graph-distance predictor is not competitive with most of the other approaches studied; our most successful link predictors can be viewed as using measures of proximity that are robust to the few edges that result from rare collaborations between fields.

3.4.2 New Links without Common Neighbors: The Distance-Three Task

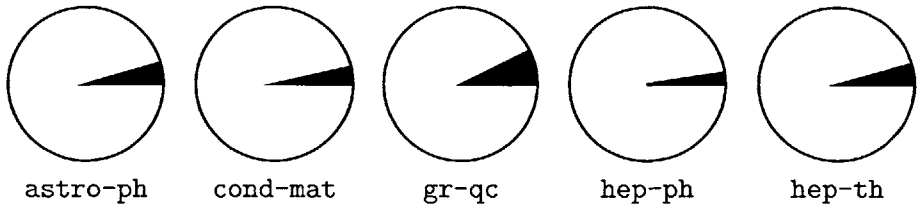
The small-world problem suggests that there will be many pairs of authors separated by a graph distance of two who will not collaborate, because proximity in graph distance does not imply proximity in the space of research interests. However, we also observe the dual problem: many pairs of authors who collaborate during the test period are separated by a graph distance larger than two during the training period. Between 71% (`hep-ph`) and 83% (`cond-mat`) of new edges form between pairs at distance three or greater. See Figure 3-8 for full statistics on the relationship between pairs of authors separated by a graph distance of two and the probability that a new collaboration will form between them.

Because most new collaborations occur between authors who are not separated by a graph distance of two in G_{collab}, we are also interested in how well our predictors perform when we disregard all new collaborations between distance-two pairs. By definition, nodes separated by a graph distance of more than two have no neighbors in common in G_{collab}, and hence this task essentially rules out the use of methods based on common neighbors. The performance of the other measures is shown in Figures 3-9 and 3-10.

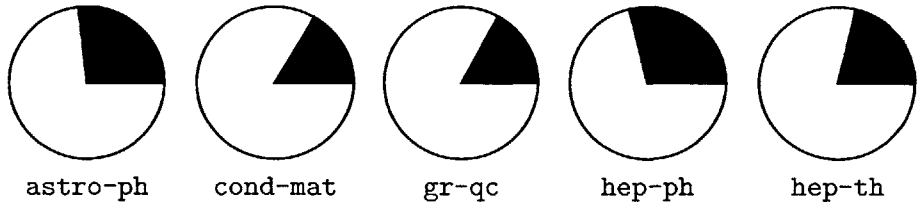
The graph-distance predictor (i.e., predicting all distance-three pairs) performs between three and eight times better than random, and is consistently beaten by virtually all of the predictors: SimRank, rooted PageRank, Katz, and the low-rank-approximation and unseen-bigram techniques. The unweighted-Katz and unseen-bigram predictors have the best performance (as high as about thirty times random, on `gr-qc`), followed closely by weighted Katz, SimRank, and rooted PageRank.

3.4.3 Similarities among Predictors

Not surprisingly, there is significant overlap in the predictions made by the various methods. In Figure 3-11, we show the number of common predictions made by ten of the most successful measures on the `cond-mat` graph. We see that Katz, low-rank inner product, and Adamic/Adar are quite similar in their predictions, as are (to a somewhat lesser extent) rooted PageRank, SimRank,



(a) Proportion of pairs $\langle u, v \rangle$ separated by graph distance two in G_{collab} that collaborate during the test period—i.e., the probability that a randomly chosen distance-two pair collaborates.



(b) Proportion of pairs who newly collaborate in the test period that are separated by a graph distance of two in G_{collab}—i.e., the probability that a randomly chosen newly collaborating pair had a common neighbor in the training period.

	astro-ph	cond-mat	gr-qc	hep-ph	hep-th
# pairs at distance two	33862	5145	935	37687	7545
# new collaborations at distance two	1533	190	68	945	335
# new collaborations	5751	1150	400	3294	1576

(c) The raw data.

Figure 3-8: *The relationship between distance-two pairs in G_{collab} and pairs who collaborate for the first time during the test period.*

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
graph distance (all distance-three pairs)		3.1	5.5	8.4	3.8	8.4
preferential attachment		3.2	2.6	8.6	4.7	1.4
SimRank	$\gamma = 0.8$	6.0	14.4	10.6	7.7	22.0
hitting time		4.4	10.2	13.7	4.5	4.7
hitting time—normed by stationary distribution		2.0	2.5	0.0	2.6	6.7
commute time		4.0	5.9	21.1	6.0	6.7
commute time—normed by stationary distribution		2.6	0.8	1.1	4.8	4.7
rooted PageRank	$\alpha = 0.01$	4.6	12.7	21.1	6.5	12.7
	$\alpha = 0.05$	5.4	13.6	21.1	8.8	16.6
	$\alpha = 0.15$	5.4	11.9	18.0	11.1	20.0
	$\alpha = 0.30$	5.9	13.6	8.5	11.9	20.0
	$\alpha = 0.50$	6.4	15.2	7.4	13.1	20.0
Katz (weighted)	$\beta = 0.05$	1.5	5.9	11.6	2.3	2.7
	$\beta = 0.005$	5.8	14.4	28.5	4.3	12.7
	$\beta = 0.0005$	6.3	13.6	27.5	4.3	12.7
Katz (unweighted)	$\beta = 0.05$	2.4	12.7	30.6	9.1	12.7
	$\beta = 0.005$	9.2	11.9	30.6	5.1	18.0
	$\beta = 0.0005$	9.3	11.9	30.6	5.1	18.0

Figure 3-9: *The performance of the basic predictors on the distance-three task. We show the performance of predictors when correct answers are restricted to the edges in E_{new} for which the endpoints were at distance three or more in G_{collab}. Methods based on common neighbors are not appropriate for this task. See Section 3.4.2, and see also Figure 3-10 for more results.*

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
graph distance (all distance-three pairs)		3.1	5.5	8.4	3.8	8.4
Low-rank approximation: Inner product	rank = 1024	2.3	2.5	9.5	4.0	6.0
	rank = 256	4.8	5.9	5.3	10.2	10.7
	rank = 64	3.9	12.7	5.3	7.1	11.3
	rank = 16	5.4	6.8	6.3	6.8	15.3
	rank = 4	5.4	6.8	32.8	2.0	4.7
	rank = 1	6.1	2.5	32.8	4.3	8.0
Low-rank approximation: Matrix entry	rank = 1024	4.1	6.8	6.3	6.2	13.3
	rank = 256	3.8	8.5	3.2	8.5	20.0
	rank = 64	3.0	11.9	2.1	4.0	10.0
	rank = 16	4.6	8.5	4.2	6.0	16.6
	rank = 4	5.2	6.8	27.5	2.0	4.7
	rank = 1	6.1	2.5	32.8	4.3	8.0
Low-rank approximation: Katz ($\beta = 0.005$)	rank = 1024	4.3	6.8	28.5	6.2	13.3
	rank = 256	3.6	8.5	3.2	8.5	20.6
	rank = 64	2.9	11.9	2.1	4.3	10.7
	rank = 16	5.1	8.5	5.3	6.0	16.0
	rank = 4	5.5	6.8	28.5	2.0	4.7
	rank = 1	0.3	2.5	32.8	4.3	8.0
unseen bigrams (weighted)	common neighbors, $\delta = 8$	5.8	6.8	14.8	4.3	24.0
	common neighbors, $\delta = 16$	7.9	9.3	28.5	5.1	19.3
	Katz ($\beta = 0.005$), $\delta = 8$	5.2	10.2	22.2	2.8	18.0
	Katz ($\beta = 0.005$), $\delta = 16$	6.6	10.2	29.6	3.7	15.3
unseen bigrams (unweighted)	common neighbors, $\delta = 8$	5.6	5.1	13.7	4.5	21.3
	common neighbors, $\delta = 16$	6.4	8.5	25.4	4.8	22.0
	Katz ($\beta = 0.005$), $\delta = 8$	4.2	7.6	22.2	2.0	17.3
	Katz ($\beta = 0.005$), $\delta = 16$	4.3	4.2	28.5	3.1	16.6
clustering: Katz ($\beta_1 = 0.001, \beta_2 = 0.1$)	$\rho = 0.10$	3.5	4.2	31.7	7.1	8.7
	$\rho = 0.15$	4.8	4.2	32.8	7.7	6.7
	$\rho = 0.20$	2.5	5.9	7.4	4.5	8.0
	$\rho = 0.25$	2.1	11.9	6.3	6.8	5.3

Figure 3-10: *The performance of the meta-predictors on the distance-three task. We show the performance of predictors when correct answers are restricted to the edges in E_{new} for which the endpoints were at distance three or more in G_{collab}. Methods based on common neighbors are not appropriate for this task. See Section 3.4.2, and see also Figure 3-9 for more results.*

	Adamic/Adar	Katz clustering	common neighbors	hitting time	Jaccard's coefficient	weighted Katz	low-rank inner product	rooted PageRank	SimRank	unseen bigrams
Adamic/Adar	1150	638	520	193	442	1011	905	528	372	486
Katz clustering		1150	411	182	285	630	623	347	245	389
common neighbors			1150	135	506	494	467	305	332	489
hitting time				1150	87	191	192	247	130	156
Jaccard's coefficient					1150	414	382	504	845	458
weighted Katz						1150	1013	488	344	474
low-rank inner product							1150	453	320	448
rooted PageRank								1150	678	461
SimRank									1150	423
unseen bigrams										1150

Figure 3-11: *The number of common predictions made by various predictors on the* `cond-mat` *dataset, out of 1150 predictions. Parameter settings are as in Figure 3-5.*

and Jaccard. Hitting time is remarkably unlike any of the other nine in its predictions, despite its reasonable performance.

The number of common *correct* predictions made by these methods shows qualitatively similar behavior; see Figure 3-12. It would interesting to understand the generality of these overlap phenomena, especially because certain of the large overlaps do not seem to follow obviously from the definitions of the measures.

Another interesting direction for future work is using different predictors in combination. When we consider two distinct predictors, and examine the common predictions made by both, the proportion of correct predictions is typically higher than the proportion of correct predictions made by either one alone. However, this difference is typically relatively small, and finding a way to leverage multiple predictors is an interesting open question.

3.4.4 Similarities among Datasets

It is harder to quantify the relationships among the datasets, but this relationship is a very interesting issue as well. One perspective is provided by the methods based on low-rank approximation: on four of the datasets, their performance tends to be best at an intermediate rank, while on `gr-qc` they perform best at rank 1. See Figure 3-13, for example, for a plot of the change in the performance of the low-rank-approximation matrix-entry predictor as the rank of the approximation varies. This observation suggests a sense in which the collaborations in `gr-qc` have a much "simpler" structure than in the other four datasets.

One also observes the apparent importance of node degree in the `hep-ph` collaborations: the

	Adamic/Adar	Katz clustering	common neighbors	hitting time	Jaccard's coefficient	weighted Katz	low-rank inner product	rooted PageRank	SimRank	unseen bigrams
Adamic/Adar	92	65	53	22	43	87	72	44	36	49
Katz clustering		78	41	20	29	66	60	31	22	37
common neighbors			69	13	43	52	43	27	26	40
hitting time				40	8	22	19	17	9	15
Jaccard's coefficient					71	41	32	39	51	43
weighted Katz						92	75	44	32	51
low-rank inner product							79	39	26	46
rooted PageRank								69	48	39
SimRank									66	34
unseen bigrams										68

Figure 3-12: *The number of* correct *common predictions made by various predictors on the* `cond-mat` *dataset, out of 1150 predictions. The diagonal entries indicate the number of correct predictions for each predictor. Parameter settings are as in Figure 3-5.*

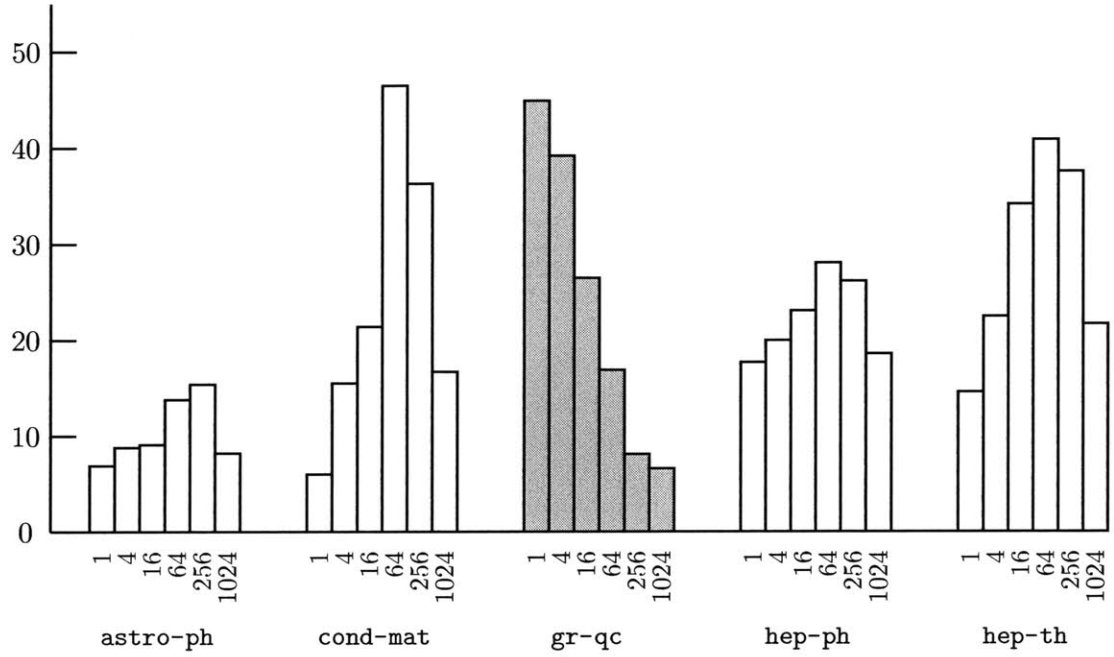


Figure 3-13: *Relative performance of the low-rank-approximation matrix-entry predictor for various ranks on each arXiv section. For each arXiv section, the performance of this predictor with ranks* 1, 4, 16, 64, 256, *and* 1024 *is shown, in increasing order from left to right. In all datasets except* `gr-qc`, *predictor performance peaks at an intermediate rank, whereas for* `gr-qc`, *performance strictly decreases as the rank increases.*

preferential-attachment predictor—which considers only the number (and not the identity) of a scientist's coauthors—does uncharacteristically well on this dataset, outperforming the basic graph-distance predictor.

Finally, it would be interesting to make precise a sense in which `astro-ph` is a "difficult" dataset, given the low performance of all methods relative to random, and the fact that none beats simple ranking by common neighbors. We will explore this issue further below when we consider collaboration data drawn from other fields.

3.4.5 The Breadth of the Data

We also have considered three other datasets: (1) the proceedings of the theoretical-computer-science conferences Symposium on the Theory of Computing (STOC) and Foundations of Computer Science (FOCS), (2) the papers found in the Citeseer (`www.citeseer.com`) online database, which finds papers by crawling the web for any files in postscript form, and (3) all five of the arXiv sections merged into one. These datasets grant some insight as to the inherent difficulty of the task, and how well our predictors are really doing on the arXiv sections.

Consider the performance of the common-neighbor predictor on these datasets:

	STOC/FOCS	arXiv sections	all arXiv's	Citeseer
common neighbors, factor over random	6.1	18.0—41.1	71.2	147.0

Performance versus random swells dramatically as the topical focus of our data set widens. That is, when we consider a more diverse collection of scientists, it is fundamentally easier to group scientists into fields of study (and outperform random predictions, which will usually make guesses between fields). When we consider a sufficiently narrow set of researchers—e.g., STOC/FOCS—almost any author can collaborate with almost any other author, and there seems to a strong random component to new collaborations. (In extensive experiments on the STOC/FOCS data, we could not beat random guessing by a factor of more than about seven.)

It is an interesting challenge to formalize the sense in which the STOC/FOCS collaborations are truly intractable to predict—i.e., to what extent information about new collaborations is simply not present in the old collaboration data.

3.5 Future Directions

Although the predictors we have discussed perform reasonably well, even the best predictor on its best dataset (the Katz-clustering predictor on the `gr-qc` dataset) is correct on only 16.1% of its predictions. It seems that there may be considerable room for improvement in performance on this task, and finding ways to take better advantage of the information in the training data is an interesting open question. Another issue is to improve the efficiency of the proximity-based methods on very large networks; fast algorithms for approximating the distribution of node-to-node distances may be one approach [153]. (The Citeseer dataset discussed above is significantly larger than an arXiv section, and it is large enough to make computational costs prohibitive.)

The graph G_{collab} is a lossy representation of the data; we can also consider a bipartite collaboration graph B_{collab}, with a vertex for every author and paper, and an edge connecting each paper to each of its authors. (We can of course reconstruct G_{collab} from B_{collab}.) The bipartite graph contains more information than G_{collab}, so we may hope that predictors can use it to improve

performance. The size of B_{collab} is much larger than G_{collab}, making experiments expensive, but we have tried using the SimRank and Katz predictors—which apply directly to B_{collab} as well—on smaller datasets (`gr-qc`, or shorter training periods). Their performance does not seem to improve, but perhaps other predictors can fruitfully exploit the additional information contained in B_{collab}.

Similarly, our experiments treat all training period collaborations equally. There is additional temporal information in the data, and perhaps one can improve performance by treating more recent collaborations as more important than older ones. This strategy is particularly valuable with longer training periods. One could also tune the parameters of the Katz predictor, e.g., by dividing the training set into temporal segments, training β on the beginning, and then using the end of the training set to make final predictions. (This approach, of course, explicitly disregards the observed evolution of the collaboration process over the decades-long time scale [75].)

There has also been relevant work in the machine-learning community on *estimating distribution support* [160]: given samples from an unknown probability distribution P, we must find a "simple" set S so that $\Pr_{x \sim P}[x \notin S] < \varepsilon$. We can view training-period collaborations as samples drawn from a probability distribution on pairs of scientists; our goal is to approximate the set of pairs that have a positive probability of collaborating. There has also been some work in machine learning on classification when the training set consists only of a relatively small set of positively labeled examples and a large set of unlabeled examples, with no labeled negative examples [183]. It is an open question whether these techniques can be fruitfully applied to the link-prediction problem.

One might also try to use additional information, such as the titles of papers (or their full text), to identify the specific research area of each scientist and use areas to predict collaborations. In the field of bibliometrics, for example, Melin and Person [125] and Ding, Foo, and Chowdhury [48] have observed institutional and geographic correlations in collaboration; it would be interesting to attempt to use affiliations as a component of a predictor. The role of geographic influence in friendship formation is discussed in Chapter 2, and it is an interesting question as to how to apply rank-based friendship to link prediction. On the other hand, one might hope that geographic information, e.g., is simply latently present in the graph G_{collab}—exactly because of the geographic role in the formation of old edges in the training set—and thus that any geography-based predictors can work directly with the graph itself; it is likely, however, that the use of explicit geographic information can improve predictor performance.

Finally, as discussed in Sections 3.4.4 and 3.4.5, one gets a strong sense that a large fraction of new edges are essentially unpredictable given the training data, both from qualitative examination of the data and from the quantitative comparison across datasets. (Seemingly, almost any STOC/FOCS core author can form a link with almost any other such researcher; exactly which edges form appears to be mostly random.) Finding a way to formalize this randomness, and to quantify the best performance that can be expected of any predictor on a particular dataset, is a fascinating direction for future study.

Chapter 4

Inferring a Social Network

Those who try to lead the people can only do so by following the mob.
— Oscar Wilde (1854–1900).

Throughout the previous chapters of this thesis, we have taken the underlying social network as a given: from an input graph $G = \langle V, E \rangle$ consisting of a set V of people and the friendships E between them, what interesting properties can we discover about the relationship between E and particular attributes of the nodes, or the way that G changes over time in relation to its current edges? The networks upon which our studies have been based—explicit "friend" links among LiveJournal users and academic coauthorship among physicists—have strong social components, but are not perfect representations of social interaction. Explicit friend listings rely on the varying definition of "friend" from person to person; coauthorship is a proxy for a real social relationship—something more like "collaboration" than like "coauthorship"—that can have both false positives (researchers who have coauthored a paper without any true social interaction, perhaps without ever having even met each other) and false negatives (researchers who have collaborated but who have never published a coauthored paper).

In this chapter, we discuss algorithms that attempt to *infer* a social network, with "a influences b" as the underlying social relationship that we attempt to infer. We then use these induced edges to identify the people who are bellwethers of the network, those who originate ideas that spread widely in the network. Our inferences are based upon observations of the "infection" of various people in the network with various topics—that is, the public discussion of a topic by a person who may have become interested in the topic via the influence of a friend. By observing that person u frequently talks about a subject soon before person v does, we have evidence for the existence of a social-network edge between u and v. Although this approach to social-network inference can lead to erroneous conclusions, we will independently validate our inferred edges against other evidence of influence, discovering that topic-based inference can be accurate and helpful when there is no explicitly provided social network.

The work described in this chapter was done jointly with Dan Gruhl, R. Guha, and Andrew Tomkins. This work appears as "Information Diffusion through Blogspace," *SIGKDD Explorations*, 6(2):43–52, December 2004. A previous version of the paper appears in *Proceedings of the Thirteenth Annual World Wide Web Conference (WWW'04)*, May 2004, pp. 491–501.

4.1 Introduction

Our eventual goal in this chapter is the inference of the edges of a social network, via the observation of the behavior of the people in the network. To infer these edges, we concentrate on the propagation of a *meme*—the term coined by Richard Dawkins for a minimal "unit of culture" transmitted from person to person—among bloggers.

Before the dramatic growth of the web in general and blogging specifically, technological barriers in publication were the major barrier to the propagation of information; someone wishing to spread a new idea through society would typically have to find some mass distributor (major media, book-publishing houses, etc.) to broadcast the idea. Blogs provide a new mechanism for the spread of memes at the grassroots level, allowing bloggers to potentially reach and influence a great many readers. Of course, as Robert Browning wrote in "Andrea del Santo," a man's reach should exceed his grasp; what we explore in our meme-tracking studies is the extent to which each person's grasp is exceeded by this expanded reach. To achieve our inference goals, we have two distinct tasks to accomplish with our data. First, we must identify the memes that appear to be propagating through our blogs and, second, we must track the spread of these memes to identify pairs $\langle u, v \rangle$ of nodes such that many memes are propagated from u to v. (Such a pair $\langle u, v \rangle$ will be an inferred edge in our network.)

We discuss our approach to topic identification in Section 4.4—for our purposes, it turned out that the best topic-tracking techniques were naïve approaches based on simple regular expressions. With the set of topics in hand, in Section 4.5 we discuss the individuals in our network. Then in Section 4.6 we develop a model for information diffusion based on the theory of the spread of infectious diseases; the parameters of the model capture how a new topic spreads from blog to blog. We give an algorithm to learn the parameters of the model based on real data and then apply the algorithm to real and synthetic blog data. As a result, we are able to identify particular individuals who are highly effective at contributing to the spread of "infectious" topics. First, though, we discuss related work in Section 4.2 and describe the blog dataset in detail in Section 4.3.

4.2 Related Work

There is a rich literature surrounding questions of propagation through networks that is relevant to our work, from a variety of fields ranging from thermodynamics to epidemiology to marketing. We provide here a broad survey of the area, with pointers to more detailed survey works where possible, and give some details of recent work on disease propagation that is closest in spirit to the models that we present. Before we turn to the more relevant study of probabilistic models, in which propagation is not guaranteed to occur to a node's neighbors, we note that the spread of information has been studied extensively from an algorithmic perspective in the context of *gossiping* and *broadcasting* [85] in a variety of networks.

4.2.1 Information Propagation and Epidemics

Much of the previous research investigating the flow of information through networks has been based upon the analogy between the spread of disease and the spread of information in networks. This analogy brings centuries of study of epidemiology to bear on questions of information diffusion. (See, for example, the book of Bailey [19] for some of the extensive work in this field.)

Classical disease-propagation models in epidemiology are based upon the cycle of disease in a host: a person is first *susceptible (S)* to the disease. If he or she is then exposed to the disease by an infectious contact, the person becomes *infected (I)* (and *infectious*) with some probability. The disease then runs its course in that host, who is subsequently *recovered (R)* (or *removed*, depending on the virulence of the disease). A recovered individual is immune to the disease for some period of time, but the immunity may eventually wear off. Thus a *SIR model* applies for diseases in which recovered hosts are never again susceptible to the disease—as with a disease conferring lifetime immunity, like chicken pox, or a highly virulent disease from which the host does not recover—while a *SIRS model* applies for the situation in which a recovered host eventually becomes susceptible again, as with influenza. In blogspace, one might interpret the SIRS model as follows: a blogger who has not yet written about a topic is exposed to the topic by reading the blog of a friend. She decides to write about the topic, becoming infected. The topic may then spread to readers of her blog. Later, she may revisit the topic from a different perspective and write about it again.

Girvan et al. [66] study a SIR model *with mutation*, in which a node u is immune to any strain of the disease that is sufficiently close to a strain with which u was previously infected. They observe that for certain parameters it is possible to generate periodic outbreaks, in which the disease oscillates between periods of epidemic outbreak and periods of calm while it mutates into a new form. In blogspace, one could imagine the mutation of the topic of Arnold *qua* movie star into the topic of Arnold *qua* governor. (We observe this kind of ebb and flow in the popularity of various "spiky chatter"-type memes. See Section 4.4.)

Early studies of propagation took place on "fully mixed" or "homogeneous" networks in which a node's contacts are chosen randomly from the entire network. Recent work, however, focuses on more realistic models based on social networks. In the Watts/Strogatz [177] model of small-world networks, Moore and Newman [132] are able to calculate the minimum transmission probability ε for which a disease will spread from one seed node to infect a constant fraction of the entire network (known as the *epidemic threshold*).

As with many social networks, the social network defined by blog-to-blog links displays a power-law degree distribution [113], and we now review some previous research on epidemic spreading on power-law networks. (See Section 1.3.4.) Pastor-Satorras and Vespignani [154] analyze an SIS model of computer-virus propagation in power-law networks, showing that—in stark contrast to random or regular networks—the epidemic threshold is *zero*, so an epidemic will always occur if an infection begins. These results can be interpreted in terms of the robustness of the network to random edge failure, as follows. Suppose that each edge in the network is deleted independently with probability $1 - \varepsilon$; we consider the network "robust" if most of the nodes are still connected. It is easy to see that nodes that remain in the same component as some initiator v_0 after the edge-deletion process are exactly the same nodes that v_0 infects according to the disease-transmission model above. This question has been considered from the perspective of *error tolerance* of networks like the Internet: what happens to the network if a random $(1 - \varepsilon)$-fraction of the links in the Internet fail? Many researchers have observed empirically and analytically that power-law networks exhibit extremely high error tolerance [13, 29, 39, 41, 42].

In blogspace, however, many topics propagate without becoming epidemics, so a model with an epidemic threshold of zero is inappropriate. One refinement is to consider a more accurate model of power-law networks. Eguíluz and Klemm [54] have demonstrated a non-zero epidemic threshold under the SIS model in power-law networks produced by a certain generative model that takes into account the high clustering coefficient found in real social networks. Another refinement is to

modify the transmission model. Wu et al. [180] consider the flow of information through real and synthetic email networks under a model in which the probability of infection decays as the distance from the initiator v_0 increases. They observe that meme outbreaks under their model are typically limited in scope—unlike in the corresponding model without decay, where the epidemic threshold is zero—exactly as one observes in real data. Newman et al. [145] have also empirically examined the simulated spread of email viruses by examining the network defined by the email address books of a user community. Finally, Newman [141] is able to calculate properties of disease outbreaks, including the distribution of outbreak sizes and the epidemic threshold, for an SIR model of disease propagation.

4.2.2 The Diffusion of Innovation

The spread of a piece of information through a social network can also be viewed as the propagation of an *innovation* through the network. (For example, the URL of a website that provides a new, valuable service is such a piece of information.) In the field of sociology, there has been extensive study of the *diffusion of innovation* in social networks, examining the role of *word of mouth* in spreading innovations. At a particular point in time, some nodes in the network have adopted the innovation, and others have not. Two fundamental models for the process by which nodes adopt new ideas have been considered in the literature:

- *Threshold models [72].* Each node u in the network chooses a *threshold* $t_u \in [0, 1]$, typically drawn randomly from some probability distribution. Every neighbor v of u has a nonnegative *connection weight* $w_{u,v}$ so that $\sum_{v \in \Gamma(u)} w_{u,v} \leq 1$, and u adopts the innovation if and only if $t_u \leq \sum_{v \in \mathsf{Adopters} \cap \Gamma(u)} w_{u,v}$, where $\Gamma(u)$ denotes the graph-theoretic neighborhood of node u.
- *Cascade models [70].* Whenever a social contact $v \in \Gamma(u)$ of a node u adopts the innovation, then u adopts with some probability $p_{v,u}$. (In other words, every time a person close to u adopts, there is a chance that the person u will decide to "follow" v and adopt as well.)

In the *independent cascade model* of Goldenberg, Eitan, and Muller [70], we are given a set of N nodes, some of which have already adopted. In the initial state, some nonempty set of nodes is "activated." At each successive step, some (possibly empty) set of nodes becomes activated. The episode is considered to be over when no new activations occur. The nodes are connected by a directed graph with each edge $\langle u, v \rangle$ labeled with a probability $p_{u,v}$. When node u is activated in step t, each node v that has an arc $\langle u, v \rangle$ is activated with probability $p_{u,v}$. This influence is independent of the history of all other node activations. (If v is not activated in that time step, then u will never activate v.) The *general cascade model* of Kempe, Kleinberg, and Tardos [96] generalizes the independent cascade model—and also simultaneously generalizes the threshold models described above—by discharging the independence assumption.

Kempe et al. [96] are interested in a related problem with a marketing motivation: assuming that innovations propagate according to such a model, and given a number k, find the set S_k^* of k "seed" nodes that maximizes the expected number of adopters of the innovation if the members of S_k^* adopt initially. (One can then give free samples of a product to S_k^*, for example.)

4.2.3 Game-Theoretic Approaches

The propagation of information through a social network has also been studied from a game-theoretic perspective, in which one postulates an increase in utility for players who adopt the

new innovation (or learn the new information) if enough of their friends have also adopted. (For example, each player chooses whether to switch from video tape to DVDs; a person with friends who have made the same choice can benefit by borrowing movies.) In blogspace, sharing discussion of a new and interesting topic with others in one's immediate social circle may bring pleasure or, potentially, increased social status.

Morris [133] and Young [182] consider a setting like the following coordination game: in every time step, each node in a social network chooses a *type* $\{0, 1\}$. Here we interpret players of type 1 to have adopted the meme. Each player i receives a positive payoff for each of its neighbors that has the same type as i, in addition to an intrinsic benefit that i derives from its type. (Each player may have a distinct utility for adopting, depending on his inherent interest in the topic.) Suppose that all but a small number of players initially have type 0. Morris and Young explore the question of whether type 1's can "take over" the graph if every node i chooses to switch to type 1 with probability increasing as the number of i's neighbors that are of type 1 increases.

There has also been work in the economics community on models of the growth of social networks when an agent u can selfishly decide to form a link with another agent v, who may have information that u desires to learn. There is a *cost* borne by u to establishing such a link, and a *profit* for the information that u learns through this link. This research explores properties of the social network that forms under this scenario [20, 80].

4.3 Corpus Details

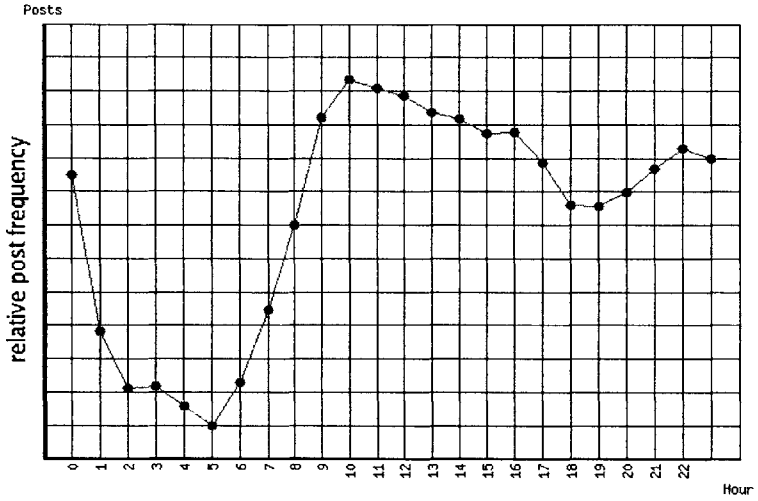
The data that we require for social-network inference must have information both about the author's identity and about the time of writing. We make use of *RSS feeds*, a technology by which an increasingly large number of publishers—both personal publishers and major media sources—distribute their posts [81, 101]. (RSS is officially an abbreviation of *RDF (Resource Description Framework) Site Summary*, but can also stand for *rich site summary* or *really simple syndication.*) RSS was originally developed to support user personalization in Netscape's Netcenter portal, by allowing the user to see the list of headlines published by the *New York Times* or Slashdot, e.g., directly on his or her portal homepage. RSS has now been adopted by the blogging community as a simple mechanism for syndication. The consistent formatting of temporal information present in RSS makes it ideal for our purposes.

Our data were collected by crawling 11,804 RSS blog feeds daily, for the one-month period from mid-September 2003 through mid-October 2003. In total, we collected 401,021 postings from blogs. In addition, we also tracked traditional media via hourly crawls of fourteen RSS channels from `rss.news.yahoo.com`. These sources helped us to identify when topics were being driven by major media or real-world events, as opposed to arising within blogspace itself. Blog entries were stored in WebFountain [76, 178] and processed to standardize the format of the date and text.

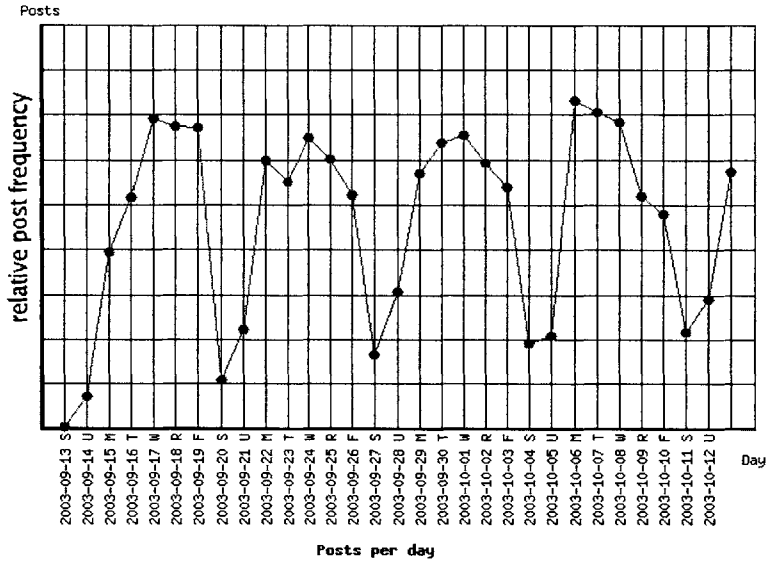
See Figure 4-1 for the profile of blog postings within a day and from day to day, normalized by the poster's time zone. Observe the strong effect of the standard work week on post volume, both within and between days.

4.4 Identification of Topics

In this section, we discuss the identification and tracking of *topics* in our data for the purpose of inducing edges in the social network. We are interested in the short-term ebb and flow of topic



(a) Number of postings by time of day.



(b) Number of postings by day of the week.

Figure 4-1: *Frequency of blog posting by time of post. For the month-long data-collection period, we show the number of blog postings both by the time of the day and by the day of the week. Times and days are normalized to the local time of the poster.*

interest (on the scale of days or weeks)—and not the longer-term shifts in emphases that occur on the scale of years—as these short-term variations are the kinds of things more likely to be caused endogenously, by the influence of one blogger on another.

4.4.1 Topic Identification and Tracking

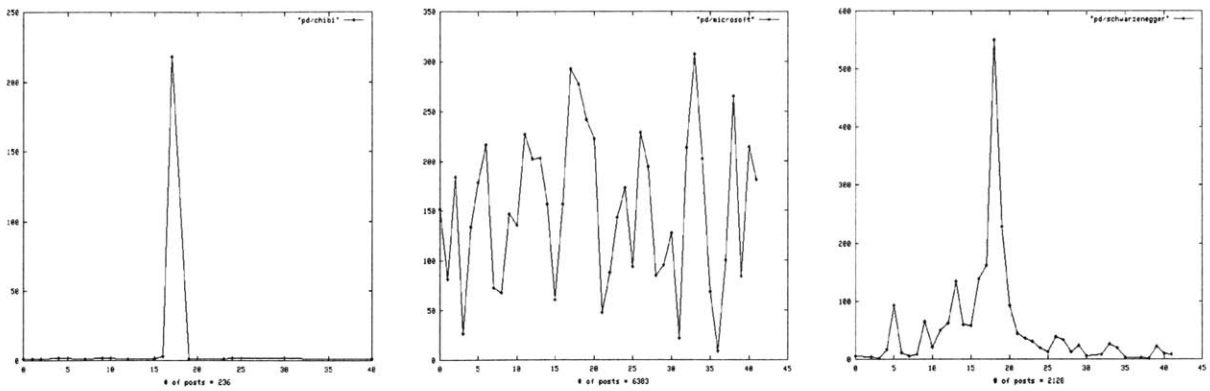
The field of *topic detection and tracking* has studied the problem of topic identification in depth for a number of years; for some recent background, see the papers from NIST's evaluation workshop [169] or, for example, the book edited by James Allan [14]. Our requirements are somewhat different from those in the topic-tracking literature; instead of high precision or recall for a particular topic, we need representatives of the various types of topics in blogspace. We have thus evaluated a handful of simple pattern-based techniques, choosing the ones that were most effective given our goals, and then manually validated different subsets of this broader set for use in particular experiments.

Our initial attempts at topic tracking all failed to yield sufficient data for tracking the spread of topics; we mention them here because their failure revealed some unexpected gaps in our intuition regarding blogspace. Our biggest initial hope for topics was links to unusual web pages. (Around the time of this study, a movie of a Rube-Goldberg-type assembly of a Honda was posted on the web, and links to it were propagated across the social network; we thought that this type of propagation was a fairly common phenomenon.) Of the over 100,000 distinct URLs to which links appeared in our dataset, however, only 700 appeared more than nine times, so we did not have nearly enough link data for significant study. We next considered recurring sequences of words using sequential pattern mining [8], but again found under 500 such recurrent sequences, many of which represented automatically generated server text, or common phrases such as "I don't think I will" and "I don't understand why." During our data-collection period, there was one non-trivial block-reposting meme that appeared frequently:

> *aoccdrnig to rscheearch at an elingsh uinervtisy it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer is at the rghit pclae.*

Although this phrase is exactly the right sort of meme for topic tracking, its uniqueness doomed this approach to finding topics. We then turned to references to entities defined in the TAP ontology [78]. This provided around 50,000 instances of references to 3700 distinct entities, but, again, fewer than 700 of these entities occurred more than 10 times.

The majority of the topics that we used in our experiments were based on much simpler approaches. We derived around 20,000 individual terms as topics, using a modified version of the *tfcidf* (*term frequency, cumulative inverse document frequency*) as a filter: we considered any term that, for some day t, appeared at least ten times on day t and exceeded by a factor of three the average number of occurrences of that term on previous days. (We used these topics mostly to analyze topic structure in experiments not reported in this thesis, as they were mostly unrelated to social-network inference.) Finally, we used a naïve pattern for proper nouns—any repeated sequence of uppercase words surrounded by lowercase text—which yielded 11,000 distinct topics, over half of which occurred at least ten times. These proper-name topics were used as the input to our edge-inference algorithm. (The URL and phrasal topics were too sparse to use for these purposes, and the individual-term topics appeared to be less likely to be spread via social-network edges.)



(a) Chibi (b) Microsoft (c) Schwarzenegger

Figure 4-2: *Three types of topic patterns: the topic "Chibi" is a single spike; "Microsoft" is spiky chatter; and "Schwarzenegger" is mostly chatter, with one spike. For each, the number of posts on the topic is shown for each day of our data-collection window.*

4.4.2 Characterization of Topic Structure

Although our interest in blogspace topics is primarily focused on tracking topics for the purposes of edge inference in the social network, we will digress briefly and discuss some of the structure of the topics that we find in our dataset.

In Figure 4-2, we show plots of the number of daily posts on three different topics, *Microsoft*, *Schwarzenegger*, and *Chibi* (a term meaning "small" or "young" adopted from the Japanese, used most frequently in the context of anime, where there was even a character named Chibi Chibi [172, 179]; or, apparently, a type of sewing needle). Based upon on our observations of these and other topics, we can loosely characterize topic traffic into *chatter*, sustained discussion usually internally driven within blogspace, and *spikes*, sharp rises in posting frequency often associated with external events. (Chibi is a pure spike, Microsoft has spiky chatter, and Schwarzenegger is mostly chatter with one large spike.)

On rare occasions, chatter can achieve *resonance*—someone posts something to which many people respond, and the story spikes without a real-world cause. The formation of order (a spike) out of chaos (chatter) has been observed in a variety of situations [164], though observation of our data reveals that this happens very rarely in blogspace. (The "aoccdrnig to rscheearch ..." meme is an example; the topic came out of nowhere, spiked, and died in about two weeks, with most postings over a four-day period.)

4.5 Characterization of Individuals

Having identified the topics on the basis of which we will infer social-network edges, we turn to the *individuals* who make up the network. In this section, we discuss a classification of individuals into a small number of categories indicating their typical role in the spread of a new topic. In Section 4.6 we formulate a model the propagation of topics from person to person through blogspace, and we present and validate an algorithm for inducing model parameters.

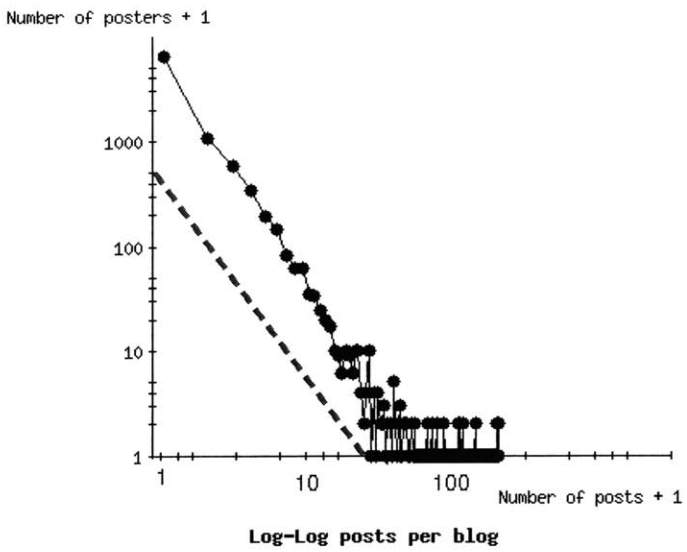


Figure 4-3: *Distribution of the number of posts per blogger. We show the number of bloggers who post exactly k blog entries during our data-collection window, for each k. The dashed line is proportional to $1/k^2$.*

4.5.1 Blogging Volume

Before we turn to the interaction between individual bloggers and topics, we first plot the distribution of the number of postings made by different users during our data-collection window in Figure 4-3. One would expect to see power-law-like behavior here [128], and indeed this distribution is well-approximated by a power law. The power law that we observe in the plot of blogger output in our dataset has an exponent around two—in the figure, we plot a reference line with the number of bloggers who produce k postings proportional to k^{-2}, and it appears to be a very good fit. *Lotka's Law* [121], one of the first published observations of a power-law distribution, asserts that the distribution of the number of papers published by a scientific researcher is power-law distributed, with an exponent of approximately two; it is interesting to note that almost the same distribution holds in the productivity of bloggers.

4.5.2 Blogger Roles

We now wish to understand more about the various roles that a particular blogger may play in the spread of a meme through blogspace. To do so, we adopt a set of simple predicates on topics that will allow us to associate particular posts with different parts of the "life cycle" of the topic. Once we have this categorization of posts, we can examine whether certain individuals are associated with a particular stage of the life cycle.

Identifying Topic Regions

For a fixed topic, we refer to the daily post rate for day i as the number of blog entries posted on day i that are on that topic (i.e., match the pattern for the topic, as described in Section 4.4). The mean daily post rate is the daily post rate for the topic averaged over all days i. We also refer to

	RampUp	RampDown	MidHigh	Spike
Total posts this region	1733	3300	12453	55624
Number of associated users	20	55	157	310

Figure 4-4: *The association of users to various topic regions. For each type of region, we show the total number of posts that are associated with that region, as defined in Section 4.5.2, and the number of users associated with each region. A user u is associated with a region if his or her number of region posts is at least four and exceeds by three standard deviations the expected number of region posts that a random user matching u's daily output would have produced.*

certain segments of the post mass on the topic—e.g., the first 20% of the post mass consists of all topical posts from the beginning of the data-collection window until the first day by which at least a fifth of the total posts on the topic had appeared. We consider the following predicates:

RampUp: The first 20% of the posts on a topic are categorized as a *RampUp region* if (1) for every day that occurs during the first 20% of the post mass on the topic, the number of posts on the topic are below the topic's overall mean daily post rate μ, and (2) the average posting rate per day during the period encompassing the first 20% of the post mass is smaller than $\mu - \sigma/2$, where σ is the standard deviation of the daily posting rate on the topic throughout the data-collection window.

RampDown: The last 20% of the posts on a topic are categorized as a *RampDown region* analogously: all days during this period must have posting rates below the overall topic mean μ, and the average within the region must be at least half of a standard deviation below μ.

MidHigh: Similarly, the middle 25% of the posts on a topic are categorized as a *MidHigh region* if (1) all days during this period have more posts than the topic mean μ, and (2) the average posting rate during this period exceeds μ by at least half of a standard deviation.

Spike: A *Spike region* occurs if there is a day in which the number of postings on a topic exceeds the topic mean μ by at least two standard deviations, and encompasses the range from the spike to the nearest inflection point below μ in each direction.

Of the topics that we consider, 3.7% exhibit a RampUp region, 5.1% exhibit a RampDown region, 9.4% exhibit a MidHigh region, and 18.2% exhibit a Spike. Notice that definition of the predicates depends fundamentally on the overall topic mean and standard deviation across the entire data-collection window, and thus whether a particular topic satisfies a particular predicate could change if the collection window were altered.

Associating Individuals with Topic Regions

With the topic regions now defined, we can attempt to associate particular individuals in the network with these regions. Perhaps the most interesting association is with the RampUp region: a person u in the network who is disproportionately likely to post during a RampUp phase may well play a causal role in the topic's spread. A person of this description would be a prime candidate as a seed for viral marketing [50, 96, 158].

Because our regions are defined in terms of a month-long data-collection window, we must take some care in associating individuals with topic regions. For example, because RampUp regions are associated with the first 20% of the postings on a topic during the month-long period, a person who happened to blog more actively during the first few days of the month could automatically appear to be a RampUp poster. To avoid this difficulty, in performing this association we recalculate the fraction of RampUp posts on a particular day. For a day i during the collection period, let $p_{up}(i)$ denote the fraction of posts on day i that are in the RampUp region for some topic. For user u, let the pair $\langle \mathsf{total}_i(u), \mathsf{rampup}_i(u) \rangle$ denote, respectively, the total number of posts made by user u on day i, and the total number of RampUp posts made by user u on day i. We compare the total number $\sum_i \mathsf{rampup}_i(u)$ of RampUp posts made by user u versus the expected number $\sum_i p_{up}(i) \cdot \mathsf{total}_i(u)$ of RampUp posts made by a "random" user who matches u's daily output but has no bias for or against RampUp posts. We can also compute the standard deviation of the "random" user's RampUp output, as this quantity is just the sum of independent Bernoulli random variables, one for each post, with success probability varying from day to day. In Figure 4-4, we show users associated with each region. Our criterion for association is that the number of region posts by the user u must exceed the expected number of posts for u's "random" user by more than three standard deviations. We associate a relatively small percentage of users with each region, but find a relatively substantial correlation; these associations may help to identify the "trend setters" (RampUp), the "trend propagators" (Spike and MidHigh), and the "sheep" (RampDown).

4.6 Inferring the Social Network

Having identified topics that are discussed in our blog dataset, we now wish to infer social-network connections by watching the way that particular topics spread. To do so, we describe a model of the spread of a topic from blog to blog and then give an algorithm based on the model to induce the network. Because we do not have access to direct information about what caused a particular person to begin blogging about a particular topic at a particular time, we must infer the cause from the surface form of the information, the sequence in which thousands of topics spread across blogspace. Our algorithm processes these sequences and extracts the most likely communication channels to explain the propagation, based on the underlying model.

Our model is analogous to models of disease propagation studied in the epidemiology literature, in which an individual can become "infected" by a topic and can then pass that topic along to others with whom she has close contact. In our arena, close contact is a directed concept—person u may read the blog of person v, but v may not even know of u's existence, let alone read u's blog. The major difference here is that, unlike in the typical approach in disease-propagation research, we are not trying to predict the next person who will come down with the flu. Rather, we are trying to infer who gave the flu to whom, on the basis of the times at which people in the network started sneezing. (This type of approach is also sometimes followed in epidemiology—as in the search for the origin of an outbreak, like the one for Typhoid Mary—but it seems to be less common that attempting to predict future infections.)

4.6.1 Model of Individual Propagation

We derive our formal model from the independent cascade model of Goldenberg, Libai, and Muller [70], which has been generalized by the general cascade model of Kempe, Kleinberg, and

Tardos [96]. Intuitively, we wish to capture the following in our model. Whenever a friend v of person u has written about a topic on his blog, when person u subsequently reads v's blog she may become interested in the topic and write about it herself.

We are given a fixed set of n nodes, corresponding to the bloggers in the network. An *episode* begins when one or more authors initially write about a topic, and the episode ends when no new posts on the topic appear for a period of time exceeding the *timeout threshold*, a parameter of the model. In the initial state of an episode, some nonempty set of bloggers has written about the topic. In each successive timestep, some (possibly empty) set of bloggers writes about the topic. We present the model in the conceptually simpler SIR framework, in which authors do not write multiple postings on a topic; in Section 4.7, we consider extending the model to the more appropriate SIRS framework, which better models authors' repeated writings on the same topic.

Under the independent cascade model, the authors are connected by a directed graph, where each edge $\langle v, w \rangle$ is labeled with a *copy probability* $\kappa_{v,w}$. When author v writes an article at time t, each node w that has an edge from v to w writes an article about the topic at time $t+1$ with probability $\kappa_{v,w}$. This influence is independent of the history of whether any other neighbors of w have written on the topic. (The general cascade model can be seen as generalizing this model by eliminating the assumption of independence of history.) We add to our model the notion of a delay between v's writing on a topic and w's reading of v's post, and we further model the fact that a user may visit certain blogs more frequently than others. Specifically, we add the *reading probability* $r_{u,v}$ as an additional edge parameter, denoting the probability that u reads v's blog on a particular day, independent of her reading history. (Thus we approximate the waiting time before person u reads person v's blog by an geometrically distributed random variable.) We introduce a second parameter on each edge because there are significant and variable delays between topical infections in the network, and we want a propagation model that can match this observation.

Formally, the parameters of our model are (1) the timeout threshold, and (2) for every pair $\langle u, v \rangle$ of nodes in the graph, the reading probability $r_{u,v}$, which is assumed to be zero if the edge $\langle u, v \rangle$ does not exist, and the copy probability $\kappa_{u,v}$. A propagation episode for a particular topic occurs as follows in our model. If the topic "appears" at vertex u on day t—i.e., u has written about the topic on day t—then we compute the probability that the topic will propagate from u to a neighboring vertex v as follows:

- choose a delay d from a geometric distribution with parameter $r_{u,v}$. (Node v reads u's blog on any particular day with probability $r_{u,v}$; thus the delay until v reads the u's post on the topic is geometrically distributed with parameter $r_{u,v}$.)

- with probability $\kappa_{u,v}$, node v then chooses to write about the topic on day $t+d$. (If v reads the topic and chooses not to copy it, then v will never copy that topic from u; there is only one opportunity for a particular topic to propagate along a particular edge.)

Alternatively, one may imagine that once u is infected, node v will become infected with probability $\kappa_{u,v} r_{u,v}$ on any given day, but once the $r_{u,v}$ coin comes up heads, no further trials are made. See Section 4.7 for some extensions to the model.

This fully describes the propagation model, given the parameters of the transmission graph. We now turn to our question of primary interest in this chapter: given a set of episodes—that is, for many topics, the nodes at which the topic appeared and the times at which it did so—we would like to infer the set of edges that exist in the graph, along with the values of κ and r.

4.6.2 Induction of the Transmission Graph

In the following, we make a *closed-world assumption* that all occurrences of a topic except the first are the result of communication via edges in the network. This assumption is obviously an oversimplification, but the probabilistic nature of our inferences lessens the impact of "real-world" infections that we mistakenly categorize as blog-to-blog infections. (In Section 4.7, we discuss weakening this assumption by introducing an "outside-world" node into the model.)

For the purposes of the algorithm, a *topic* is a URL, phrase, name, or any other representation of a meme that can be tracked from page to page. We gather all blog entries that contain a particular topic into a list $[(u_1, t_1), (u_2, t_2), \ldots, (u_k, t_k)]$, where u_i is a blog and t_i is the first time at which blog u_i contained a reference to the topic, and the list is sorted by increasing t_i. We refer to this list as the *topic trace* for the topic.

We wish to use topic traces for a large number of topics to infer the existing edges in the network. Because social networks are in general very sparse graphs—the average degree of a node is typically quite small, though of course the degree distribution is usually heavily skewed—we need to identify the roughly linear number of edges that exist among the candidate set of $\Theta(n^2)$ edges. However, the number of topics is very small compared to the size of the candidate set, and we suffer from the limited amount of data. In our algorithm, we will make critical use of the following observation: suppose that blog u appears in a topic trace and blog v *does not* appear later in the same sequence. This situation gives us negative evidence about the existence of an edge between u and v: if v were a regular reader of u's blog with a reasonable copy probability, then sometimes memes discussed by u should appear in v's blog. Thus, we gain information from both the presence and absence of entries in the topic trace.

We present an EM-like algorithm [47] to induce the parameters of the transmission graph. We first compute a "soft assignment" of each new topic infection to the edges that may have caused it, and then we update the edge parameters to increase the likelihood of the assigned infections. Suppose that we have an initial guess at the value of the reading probability $r_{u,v}$ and the copy probability $\kappa_{u,v}$ for each pair $\langle u, v \rangle$ of nodes, and we wish to improve our estimates of these values. We adopt the following two-stage process:

Soft-Assignment Step: Using the current version of the transmission graph, compute the probability that topic j traversed the $\langle u, v \rangle$ edge, for each topic j and each pair $\langle u, v \rangle$ such that v was infected with topic j after u was.

Fix a particular topic j, and consider the topic trace for topic j. For each node v in the topic trace, consider any node u that precedes v in the sequence. We compute the probability $p_{u,v}$ that topic j would have been copied from u to v, given the delay between u and v in the sequence, and then normalize by the sum of these probabilities to compute the posterior probability that node u was the source of infection for node v. That is, letting $\delta_{u,v,j}$ denote the delay in days between u and v in the topic trace for j, we set

$$p_{u,v} := \frac{r_{u,v}(1 - r_{u,v})^{\delta_{u,v,j}} \kappa_{u,v}}{\sum_{w:\delta_{w,v,j} \geq 0} r_{w,v}(1 - r_{w,v})^{\delta_{w,v,j}} \kappa_{w,v}}.$$

To improve efficiency in practice, we require propagation to occur within thirty days and consider only the twenty nodes w that are closest to v in the topic trace.

Parameter-Update Step: For each pair u and v, recompute $r_{u,v}$ and $\kappa_{u,v}$ based on the posterior probabilities computed above, so that $1/r_{u,v}$ is the expected delay in topics copied from u to v and $\kappa_{u,v}$ is the ratio between the expected number of u's topics copied and read by v.

We now formally describe the updates for a fixed u and v. Let S_{copied} denote the set of topics j such that topic j appeared first at node u and subsequently at node v, and let $S_{uncopied}$ denote the set of topics j such that u was infected with topic j but v was never infected with the topic. For each topic $j \in S_{copied}$, we require as input the pair $\langle p_j, \delta_j \rangle$, where p_j is the posterior probability, computed above, that u infected v with topic j, and δ_j is the delay in days between the appearance of the topic in u and in v. For every topic $j \in S_{uncopied}$, we require as input the value δ_j, where δ_j days elapsed between the appearance of topic j at node u and the end of our data-collection window. We can then recompute updated versions of $r_{u,v}$ and $\kappa_{u,v}$ as

$$r_{u,v} := \frac{\sum_{j \in S_{copied}} p_j}{\sum_{j \in S_{copied}} p_j \delta_j} \qquad \kappa_{u,v} := \frac{\sum_{j \in S_{copied}} p_j}{\sum_{j \in S_{copied} \cup S_{uncopied}} \Pr[r \leq \delta_j]}$$

where $\Pr[a \leq b] = (1-a)(1-(1-a)^b)$ is the probability that a geometric distribution with parameter a has value $\leq b$. Given the p_j's, the updated $1/r_{u,v}$ is the expected delay in topics copied from u to v, and the updated $\kappa_{u,v}$ is the ratio of the expected number of topics at u that were copied by v to the expected number of such topics read by v. (In calculating the κ values, we work under the simplifying assumption that each infection has a single cause—that is, that if u is infected by a topic j, then there is exactly one node v that is responsible for u's infection.)

We can iterate this two-step process: in the first step, we use our model of the graph to guess how data traveled; in the second, we use our guess about how data traveled to improve our model of the graph. For our data sets, the values of r and κ converge within two to five iterations, depending on the data, to a vector of values within 1% of the limiting value under the L_2 norm.

4.6.3 Validation of the Algorithm

In this section, we validate the algorithm described in the previous section both on synthetic data derived from simulated propagation on an artificial graph and on real blog data based on the topics described in Section 4.4.

Validation for Synthetic Data

For the synthetic validation, we create synthetic propagation networks, using the synthesized graphs to generate a set of topic traces generated according to the model described in Section 4.6.1, and then run our inference algorithm. We then compare the inferred edges to the actual edges in the synthetic graph.

The synthetic graphs here are based on a slight variation on Erdős/Rényi random graphs, which are in general not very good models of social networks, as discussed in Section 1.2.1. However, they are simple graphs, easy to generate, and are a first cut at validating our algorithm. We fix a number n of vertices and a degree d; each vertex selects exactly d neighbors uniformly with replacement from the vertex set, and parallel edges and self-loops are excised. The edge parameters are set to $r := 2/3$ and $\kappa := 1/10$ for each of the edges in our graph. For each graph, we simulate

	Reading Probability		Copy Probability	
τ	$\mu_{\tilde{r}}$	$\sigma_{\tilde{r}}$	$\mu_{\tilde{\kappa}}$	$\sigma_{\tilde{\kappa}}$
20	0.718	0.175	0.141	0.455
40	0.703	0.157	0.107	0.039
60	0.694	0.134	0.103	0.034

Figure 4-5: *Edge parameters inferred on the low-traffic synthetic benchmark. For a 1000-node graph with each node connected to three randomly chosen neighbors, with reading probability $r = 2/3$ and copy probability $\kappa = 0.1$ on every edge, we seed τ targets at each node, allow them to propagate, and then infer edges and edge parameters $\tilde{r}$ and $\tilde{\kappa}$ via our algorithm. The mean and standard deviations for $\tilde{r}$ and $\tilde{\kappa}$ are shown.*

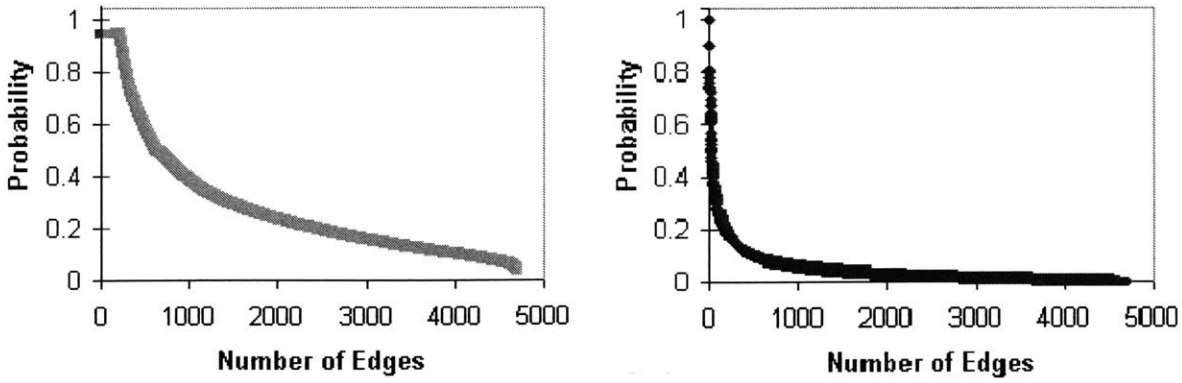
the propagation mechanism according to our model: for each node in the graph, we seed τ distinct topics and let each topic propagate throughout the graph. We ran experiments with two different settings for the parameters: (1) the *low-traffic synthetic benchmark*, where $n = 1000$, $d = 3$, and $\tau \in \{20, 40, 60\}$, and (2) the *high-traffic synthetic benchmark*, where $n = 500$, $d = 9$, $\tau = 20$. The high-traffic case is harder for our algorithm, because topics tend to spread much further through the graph. (For a seed node, the expected number of directly infected neighbors is $d\kappa$, which is 0.3 in the low-traffic case and 0.9 in the high-traffic case.)

Low-Traffic Synthetic Benchmark. In the low-traffic benchmark, we ran multiple simulations with the number τ of seed topics per node ranging from 20 to 60. (Note that the probability of a topic being picked up by any particular neighbor of the seed node is only 0.1, so the probability that the topic dies without infecting any of the three neighbors is about 25%; thus the number of non-trivial topics per node ranges from about five to about fifteen.)

We ran our inference algorithm on the resulting topic traces, producing an edge set $\tilde{E}$ with edge parameters $\tilde{r}$ and $\tilde{\kappa}$. We define $\tilde{E}$ as the set of pairs $\langle u, v \rangle$ so that there was an inferred probability of at least 0.1 of propagation from u to v, for at least three distinct topics. In this case, two iterations of the algorithm suffice for convergence.

We then compare our results to the original propagation network. For edges, we count any edge in the symmetric difference of E and $\tilde{E}$ as incorrect. For the $\tau = 60$ case, the algorithm correctly inferred 2663 of the roughly 3000 edges, or about 90% of the edges, with an additional four nonexistent edges erroneously inferred. (The number of edges is not exactly 3000 edges because of the self-loop and parallel-edge deletion.) The values of $\tilde{r}$ and $\tilde{\kappa}$ are quite close to the correct values r and κ, especially as τ increases. See Figure 4-5.

High-Traffic Synthetic Benchmark. In the high-traffic benchmark, where we have degree nine instead of three, topics propagate much further through the graph; the number of nodes infected with a topic ranges from one to just over 200. With $\tau = 20$ topics seeded at each node, our algorithm correctly identifies almost all edges in E (to within 1%), and erroneously identifies a further ~9% spurious edges. (The more pervasive nature of topic propagation in this graph causes many more infections to appear possible because topic traces are in general longer, and thus spurious edges are more likely to come out of our algorithm.) The estimated $\tilde{r}$ values have mean 0.73 with standard deviation 0.12, and the $\tilde{\kappa}$ values have mean 0.08 and standard deviation 0.03. As with the low-



(a) Distribution of $\tilde{r}_{u,v}$ values.

(b) Distribution of $\tilde{\kappa}_{u,v}$ values.

Figure 4-6: *The distribution of edge parameters as inferred from topics spreading through blogspace. We run the inference algorithm on the RSS data, using the roughly 7000 proper-name topics for which under 10% of the RSS occurrences are in mass media. The inferred values of mean propagation delay ($\tilde{r}$) and copy probability ($\tilde{\kappa}$) are shown.*

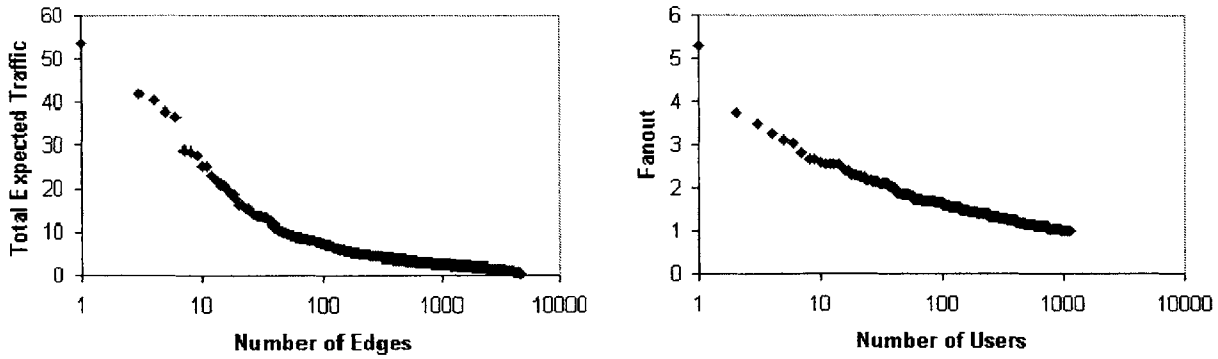
traffic benchmark, then, the inferred values of $\tilde{E}$, $\tilde{r}$, and $\tilde{\kappa}$ are very close to the original values in the synthetic network.

Validation for Real Data

Now that we have validated the algorithm on synthetically generated data, we turn to the main goal of the chapter, inferring social-network edges among the bloggers in our dataset.

Of the topics identified in Section 4.4, we focus on the proper-name topics. As discussed in Section 4.6.2, we want to impose the assumption that the spread of the topic through blogspace is wholly the result of blog-to-blog infection. To make this simplifying assumption closer to accurate, we limit our set of topics to those proper-name topics for which at least 90% of the occurrences in our RSS dataset are in blogs, rather than in mainstream-media content. This leaves us with a set of about 7000 topics. With these topics in hand, we can run our inference algorithm to induce edges between bloggers. Figure 4-6 shows the distributions of $\tilde{r}$ and $\tilde{\kappa}$ inferred by the algorithm. Most edges have an expected propagation delay ($1/\tilde{r}$) of under five days; the mean $\tilde{r}$ is 0.28 and the standard deviation is 0.22. Copy probabilities are quite low, with mean 0.04 and standard deviation 0.07, indicating that even bloggers who commonly read from another source are selective in the topics they choose to write about. To validate the edges inferred by our algorithm, we performed two experiments:

- We examined the top 100 blogs reported by `http://blogstreet.com`, a clearinghouse website that gives information about blogs and their relationships. Of these 100 blogs, our RSS-generated dataset contained 70. Using our inferred edge parameters, we can rank individual nodes in the network based upon the expected amount of traffic that flows through each node (i.e., the expected number of topics that are copied from a particular node). Of the 70 nodes in our dataset, 49 were in the top 10% of blogs in our analysis, 40 were in the top 5%, and 24 were in the top 1.2%. (Note that this validation is relatively weak evidence for the success



(a) Expected number of topics copied across an edge.

(b) Expected number of infections from a seed node (for a single topic), including the seed.

Figure 4-7: *Expected number of infections from an edge or a person. For the 7000 topics from which we induce our inferred transmission graph, we compute the expected number of topics that are copied across a single edge; for each person in the graph, we compute the expected number of people infected by a topic originating at a particular node u.*

of our algorithm, in that we might simply be identifying the most popular blogs in terms of total traffic; the precise edges that we identify might still be inaccurate.)

- In the inferred graph, we can compute for each edge the *expected traffic* across it—that is, the expected number of topics that were transmitted across that edge. We found the 200 edges that had the highest expected flow and hand-examined a random sample of this set. In 90% of these edges, we found a real link of some sort between the two blogs. The links we found include direct hypertext links to the other blog in a blog entry, the presence of the other blog in the *blogroll* (a collection of "favorite blog links" that appears on most blogs), or mention of the name or userid of the other blogger. (There are well under a hundred links in a typical blogroll—i.e., a typical blogroll links to under 1% of the 11,804 blogs in our dataset—so this proportion of links suggests that we are finding evidence of real influence between blogs.) Notice that our algorithm did not have access to any of these data and inferred these edges using the proper-name topics alone.

The nature of the inferences we seek makes it very difficult to directly observe the topical spread that we postulate, but these indirect observations suggest that our algorithm is extracting real social-network edges from the spread of topics through blogspace.

4.6.4 Discussion of the Inferred Transmission Graph

In this section, we will mention a few of the interesting properties of the inferred transmission graph. At convergence, our algorithm reports about 4000 edges that have traffic across them. Distributional statistics on the edge parameters are shown in Figure 4-6; here we discuss some of the unusual edges and nodes in the graph.

For each edge $\langle u, v \rangle$ in the inferred graph, we can compute the expected traffic across that edge (that is, the number of topics that v writes about because he read about them on u's blog). In Figure 4-7(a), we show the distribution of the expected traffic of edges in the network. Of the 7000

topics present in the data, a popular edge might account for between ten and fifty infections in expectation; the median edge has one or two topics that propagate across it.

The expected-traffic analysis identifies the edges that are most crucial in spreading topics throughout the network. We can perform a similar analysis on the nodes of the graph to identify people who in expectation trigger large outbreaks of a topic. Define the *fanout* of person u as the expected number of infections directly caused by u, assuming that a topic is initiated by u. (That is, we compute the expected number of neighbors of u who will copy a topic from u, and add one for u herself.) The distribution of fanouts is shown in Figure 4-7(b). Typical fanouts are less than two, causing a topic to die out without spreading to encompass the entire graph because less than one new person is infected by each infected person, but there are some users with much higher fanouts. The maximum-fanout person at the far left of Figure 4-7(b), with fanout 5.3, is a standout; she is a classic "connector" in the sense of *The Tipping Point* [68] with a huge collection of friends, a broad set of interests, and an intelligent and up-to-date blog.

Finally, we note that the interaction between people and topics is an interesting direction worthy of further exploration. In the low-traffic synthetic graph, 357 of the simulated topic propagations eventually reach four or more nodes in the graph. Although topics were started at all nodes in the graph, and on average hit only 5 users, there is a single user who is present in 42 of the 357 episodes, and there are eighteen users present in at least 20 of the 357 episodes. These eighteen users appear to play a crucial role in topic propagation.

4.7 Future Directions

The inference algorithm that we have presented in this chapter appears to perform quite well in inferring links representing influence between bloggers. The validations that we presented suggest that we have been able to uncover real interactions via the tracking of memes. These results are promising first steps in unearthing a social network that is only implicitly observable through the behavior of its members. Our real-world validations were based upon comparing our inferred links with observable links, and we found a high correlation between the two. For most interesting applications, though, these validating data are unavailable—this unavailability being the reason for attempting to infer links in the first place—so it is difficult to find ways to evaluate exactly how accurate our inferences are.

Before we discuss possible extensions to the techniques and questions specifically addressed in this chapter, we note that the type of inferences that we have explored here are closely related to Bayesian networks, which are (acyclic) networks representing causation that been well-studied in the machine-learning community. (See the survey of Heckerman [84].) The potential for cycles in our setting makes the relationship to Bayesian networks less direct, but there may still be valuable technical connections to be drawn to that literature.

One direction for future research on information propagation through blogspace (and thus on social-network inference) is centered on improvements to the model of propagation presented in Section 4.6.1. Some important extensions may include the following:

- Most blogspace topics do not travel exclusively through blogspace; rather, they are real-world events that are partially covered in traditional media. Some of the apparent blog-to-blog infections on a topic may in fact be media-to-blog infections. (Our model can be extended by introducing a "real-world" node to represent traditional media.)

- In our model of meme transmission, the probability of an infection of node v by node u depends only on the parameters $r_{u,v}$ and $\kappa_{u,v}$ and is independent of the topic itself. Because certain topics are inherently more interesting than others, the model can be made more realistic with copy probabilities differing from topic to topic. We introduce the *stickiness* of each topic, which controls the probability that the topic will "stick" with v. (Stickiness of a topic is analogous to *virulence* of a disease.) The probability of infection when v reads u's blog now becomes $\kappa_{u,v} \cdot \textit{stickiness}$ instead of just $\kappa_{u,v}$. Inducing stickiness values via the EM-like algorithm is possible, but the likelihood equations appear quite complicated, and the estimation would be computationally expensive.

Another interesting direction for these inferences is their application to marketing [50, 96, 158]: by inferring the network of influence, we have gained information about the most important members of the network from the perspective of influence. Kempe et al. [96] are interested in finding the best set of consumers to "seed" with a free sample of a product given a social network; a potentially valuable marketing approach is using our inference algorithm as a preprocessing step to infer the network upon which their algorithm would be run.

Since (or simultaneously with) the work that is described here, there have also been at least two similar studies—one performed by researchers at HP Labs [7], and one that involves the tracking of a meme *injected* into blogspace by Arbesman [18]; we hope that the observation of the evolving topics of interest to the members of a network will continue to be a valuable mechanism for social-network inference and understanding.

Chapter 5

Conclusions and Future Directions

You can't build a reputation on what you are going to do.
— *Henry Ford (1863–1947).*

In this thesis, we have considered three algorithmic questions that arise in social networks. How can empirical small-world/navigability properties be explained from a theoretical perspective? How can the future of a network of friendships be predicted solely from the current state of the network? How can the links of a social network be inferred solely from observations about the changing topical interests of its members?

In the previous chapters of this thesis, we have highlighted a number of specific open directions that follow naturally from the problems contained therein. The most closely related future directions appear there. Here, we mention as an interesting, broad future direction for study a single prevailing theme that arises implicitly in each of the main technical chapters of this thesis. In these chapters, we introduced models of a social network's "evolution" in one way or another—the creation of friendships on the basis of geographic mechanisms, the creation of friendships on the basis of existing friendships, and the creation of topical interests on the basis of friendships. In each case, in a sense, we have considered a set of observations (friendships in LiveJournal, newly created friendships in the arXiv, postings on a topic in blogs) and attempted to explain them using our model. Many of these observations, however, are wholly unrelated to the phenomenon that we are modeling; some friendships form because of random processes independent of geography and existing friendships, and some people become interested in topics because of the topics and not because of their friends. One of the challenges faced in this thesis is finding a way to quantify the limited role of the modeled phenomenon—as in, for example, our discovery that roughly a third of LiveJournal friendships are independent of geography. Similarly characterizing the proportion of new friendships that are independent of old friendships would be of great interest; doing so could also help to resolve the question of characterizing the sense in which the `astro-ph` dataset is "hard" compared to the others in terms of link prediction, which is one of the most interesting open questions from Chapter 3.

The present is an exciting time in the algorithmic study of social networks: as more and more data on social interactions become available, from electronic databases and online communities, the role of algorithmic thinking becomes more and more important in social-networks research. Even

"efficient" algorithms for the LiveJournal social network (Chapter 2) and for the link-prediction problem (Chapter 3) required hours or even days to run on high-end desktop machines. We have taken a data-driven approach in this thesis, by carefully studying large social networks, finding anomalous or intriguing aspects of them, and then using algorithmic techniques to attempt to understand them. The opportunity for future study in this style remains large, as new and larger data sets become available. Within the past two years, the Enron email dataset has been released; during February and March 2005 alone, over 800,000 new users have joined the LiveJournal community. These new and expanding datasets give us the chance to explore new questions about social networks. Computer science—and theoretical computer science in particular—has much to contribute to answering them.

Bibliography

[1] J. Abello, P. M. Pardalos, and M. G. C. Resende. On maximum clique problems in very large graphs. In J. Abello and J. Vitter, editors, *External Memory Algorithms*, volume 50 of *DIMACS Series on Discrete Mathematics and Theoretical Computer Science*, pages 119–130. American Mathematical Society, 1999.

[2] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, July 2003.

[3] Lada A. Adamic and Eytan Adar. How to search a social network. Submitted to *Social Networks*, 2004. Available as `cond-mat/0310120`.

[4] Lada A. Adamic and Bernardo A. Huberman. Information dynamics in a networked world. In Eli Ben-Naim, Hans Frauenfelder, and Zoltan Toroczkai, editors, *Complex Networks*, number 650 in Lecture Notes in Physics, pages 371–398. Springer, 2004.

[5] Lada A. Adamic, Rajan M. Lukose, and Bernardo A. Huberman. Local search in unstructured networks. In S. Bornholdt and H. G. Schuster, editors, *Handbook of Graphs and Networks: From the Genome to the Internet*, chapter 13. Wiley-VCH, 2002.

[6] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Physical Review E*, 64(046135), 2001.

[7] Eytan Adar, Li Zhang, Lada A. Adamic, and Rajan M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference (WWW'04)*, May 2004.

[8] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering (ICDE'95)*, pages 3–14, April 1995.

[9] Bill Aiello, Fan R. K. Chung, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10:53–66, 2001. Appears under the title "A random graph model for massive graphs" in *Proceedings of the 32nd Annual Symposium on the Theory of Computation (STOC'00).*

[10] Réka Albert and Albert-László Barabási. Topology of evolving networks: Local events and universality. *Physical Review Letters*, 85(24):5234–5237, 11 December 2000. Available as `cond-mat/0005085`.

[11] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–97, 2002. Available as `cond-mat/0106096`.

[12] Réka Albert, Hawoong Jeong, and Albert-László Barabási. The diameter of the world wide web. *Nature*, 401:130–131, 1999. Available as `cond-mat/9907038`.

[13] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, July 2000.

[14] James Allan, editor. *Topic Detection and Tracking: Event-based Information Organization.* Kluwer International Series on Information Retrieval. Kluwer Academic Press, 2002.

[15] Noga Alon, Richard M. Karp, David Peleg, and Douglas West. A graph-theoretic game and its application to the k-server problem. *SIAM Journal on Computing*, 24(1):78–100, 1995.

[16] Noga Alon and Joel H. Spencer. *The probabilistic method.* Wiley-Interscience, New York, second edition, 2000.

[17] Luís A. Nunes Amaral, Antonio Scala, Marc Barthélémy, and H. Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences USA*, 97:11149–11152, 2000. Available as `cond-mat/0001458`.

[18] Samuel Arbesman. The memespread project: An initial analysis of the contagious nature of information in online networks. Manuscript, April 2004.

[19] Norman T. J. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications.* Griffin, London, 2nd edition, 1975.

[20] Venkatesh Bala and Sanjeev Goyal. A strategic analysis of network reliability. *Review of Economic Design*, 5:205–228, 2000.

[21] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaboration. *Physica A*, 311(3–4):590–614, 2002.

[22] Albert-László Barabási. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life.* Perseus Publishing, Cambridge, 2002.

[23] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 15 October 1999.

[24] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, pages 50–59, May 2003.

[25] Lali Barrière, Pierre Fraigniaud, Evangelos Kranakis, and Danny Krizanc. Efficient routing in networks with long range contacts. In *Proceedings of the 15th International Symposium on Distributed Computing (DISC'01)*, October 2001.

[26] Peter S. Bearman, James Moody, and Katherine Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 110:44–91, 2004.

[27] Béla Bollobás. *Random Graphs.* Cambridge University Press, second edition, 2001.

[28] Béla Bollobás and Fan Chung. The diameter of a cycle plus a random matching. *SIAM Journal on Discrete Mathematics*, 1:328–333, 1988.

[29] Béla Bollobás and Oliver Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35, 2003.

[30] Béla Bollobás and Oliver Riordan. The diameter of scale-free random graph. *Combinatoricas*, 24(1):5–34, 2004.

[31] Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. The degree sequence of a scale-free random graph process. *Random Structures and Algorithms*, 18(3):279–290, May 2001.

[32] S. P. Borgatti, M. G. Everett, and L. C. Freeman. Ucinet for windows: Software for social network analysis, 2002.

[33] Prosenjit Bose, Andrej Brodnik, Svante Carlsson, Erik D. Demaine, Rudolf Fleischer, Alejandro López-Ortiz, Pat Morin, and J. Ian Munro. Online routing in convex subdivisions. *International Journal of Foundations of Computer Science*, 12(4):283–295, August 2002. Special issue of selected papers from the 11th Annual International Symposium on Algorithms and Computation, 2000 (ISAAC'00).

[34] Prosenjit Bose and Pat Morin. Online routing in triangulations. *SIAM Journal on Computing*, 33(4):937–951, 2004.

[35] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[36] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Proceedings on the 9th International World Wide Web Conference (WWW9)*, June 2000.

[37] Mark Buchanan. *Nexus: Small Worlds and the Groundbreaking Science of Networks.* W. W. Norton & Co., New York, 2002.

[38] Frances Cairncross. *The Death of Distance: How the Communications Revolution Will Change Our Lives.* Harvard Business School Press, Boston, 1997.

[39] Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85:5468–5471, 2000. Available as `cond-mat/0007300`.

[40] Rodrigo De Castro and Jerrold W. Grossman. Famous trails to Paul Erdős. *Mathematical Intelligencer*, 21(3):51–63, 1999.

[41] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, 85(21):4626–4628, November 2000. Available as `cond-mat/0007048`.

[42] Paolo Crucitti, Vito Latora, Massimo Marchiori, and Andrea Rapisarda. Efficiency of scale-free networks: Error and attack tolerance. *Physica A*, 320:622–642, 2003.

[43] Jörn Davidsen, Holger Ebel, and Stefan Bornholdt. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, 88(128701), 2002.

[44] Ithiel de Sola Pool and Manfred Kochen. Contact and influence. *Social Networks*, 1:1–48, 1978.

[45] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[46] Erik D. Demaine, John Iacono, and Stefan Langerman. Proximate point searching. *Computational Geometry: Theory and Applications*, 28(1):29–40, May 2004. Special issue of selected papers from the 14th Canadian Conference on Computational Geometry, 2002.

[47] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.

[48] Ying Ding, Schubert Foo, and Gobinda Chowdhury. A bibliometric analysis of collaboration in the field of information retrieval. *International Information and Library Review*, 30:367–376, 1999.

[49] Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 8 August 2003.

[50] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 57–66, 2001.

[51] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advanced Physics*, 51(4):1079–1187, 2002.

[52] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of networks: Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[53] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(035103), 2002.

[54] Víctor M. Eguíluz and Konstantin Klemm. Epidemic threshold in structured scale-free networks. *Physical Review Letters*, 89(108701), 2002. Available as `cond-mat/0205439`.

[55] Enron email dataset. Available at `http://www.cs.cmu.edu/~enron/`.

[56] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[57] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institue of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[58] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.

[59] Ute Essen and Volker Steinbiss. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'92)*, volume 1, pages 161–164, 1992.

[60] Christos Faloutsos, Kevin S. McCurley, and Andrew Tomkins. Fast discovery of connection subgraphs. In *Proceedings of the 10th International Conference on Knowledge discovery and Data Mining (KDD'04)*, pages 118–127, 2004.

[61] Pierre Fraigniaud, Cyril Gavoille, and Christophe Paul. Eclecticism shrinks even small worlds. In *Proceedings of the 23rd Annual Symposium on Principles of Distributed Computing (PODC'04)*, pages 169–178, July 2004.

[62] Friendster. `http://www.friendster.com`.

[63] Michael T. Gastner and M. E. J. Newman. The spatial structure of networks, 2004. Available as `cond-mat/0407680/`.

[64] Paul Ginsparg. First steps towards electronic research communication. *Computers in Physics*, 8(4):390–396, 1994.

[65] Paul Ginsparg. Winners and losers in the global research village (invited address). In *Conference on Electronic Publishing in Science, UNESCO*, Paris, 19–23 February 1996. Available at `http://arxiv.org/blurb/pg96unesco.html`.

[66] Michelle Girvan, Duncan S. Callaway, M. E. J. Newman, and Steven H. Strogatz. A simple model of epidemics with pathogen mutation. *Physical Review E*, 65(031915), 2002. Available as `nlin.CD/0105044`.

[67] Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA*, 99:7821–7826, 2002.

[68] Malcolm Gladwell. *The Tipping Point: How little things can make a big difference.* Little Brown & Co., 2000.

[69] Debra S. Goldberg and Frederick P. Roth. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences USA*, 100(8):4372–4376, April 2003.

[70] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[71] Sean P. Gorman and Rajendra Kulkarni. Spatial small worlds: New geographic patterns for an information economy. *Environment & Planning B*, 31:273–296, 2004.

[72] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1987.

[73] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[74] Mark S. Granovetter. The strength of weak ties: A network theory revisited. In P. V. Marsden and N. Lin, editors, *Social Structure and Network Analysis*, pages 105–130. Sage, Beverly Hills, 1982.

[75] Jerrold W. Grossman. The evolution of the mathematical research collaboration graph. In *Proceedings of the Southeast Conference on Combinatorics, Graph Theory, and Computing*, March 2002.

[76] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a WebFountain: An architecture for very large-scale text analytics. *IBM Systems Journal*, 43(1):64–77, 2004.

[77] John Guare. *Six Degrees of Separation: A Play.* Vintage, New York, 1990.

[78] R. Guha and Rob McCool. TAP: a semantic web platform. *Computer Networks*, 42(5):557–577, 2003.

[79] J. M. Guiot. A modification of Milgram's small world method. *European Journal of Social Psychology*, 6:503–507, 1976.

[80] Hans Haller and Sudipta Sarangi. Nash networks with heterogeneous agents. Working Paper Series E-2001-1, Virginia Tech, 2003.

[81] Ben Hammersley. *Content Syndication With RSS.* O'Reilly & Associates, 2003.

[82] Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing PageRank. Technical report, Stanford University, June 2003.

[83] Taher H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, July/August 2003.

[84] David Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, March 1995.

[85] Sandra M. Hedetniemi, Stephen T. Hedetniemi, and Arthur L. Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18:319–349, 1988.

[86] John Iacono and Stefan Langerman. Proximate planar point location. In *Proceedings of the 19th Symposium on Computational Geometry (SoCG'03)*, pages 220–226, June 2003.

[87] Glen Jeh and Jennifer Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD'02)*, July 2002.

[88] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings on the 12th International World Wide Web Conference (WWW12)*, pages 271–279, May 2003.

[89] H. Jeong, Z. Neda, and A.-L. Barabási. Measuring preferential attachment for evolving networks. *Europhysics Letter*, 61:567–572, 2003.

[90] Emily M. Jin, Michelle Girvan, and M. E. J. Newman. The structure of growing social networks. *Physical Review E*, 64(046132), 2001.

[91] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Exploiting the block structure of the web for computing PageRank. Technical report, Stanford University, March 2003.

[92] Brad Karp. *Geographic Routing for Wireless Networks.* PhD thesis, Harvard University, Cambridge, MA, October 2000.

[93] Brad Karp and H. T. Kung. GPSR: greedy perimeter stateless routing for wireless networks. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking (Mobicom'00)*, pages 243–254, August 2000.

[94] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, March 1953.

[95] H. Kautz, B. Selman, and M. Shah. ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM*, 30(3), March 1997.

[96] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (KDD'03)*, 2003.

[97] Sara Kiesler and Jonathon N. Cummings. What do we know about proximity in work groups? A legacy of research on physical distance. In Pamela Hinds and Sara Kiesler, editors, *Distributed Work*, pages 57–82. MIT Press, Cambridge, 2002.

[98] P. Killworth and H. Bernard. Reverse small world experiment. *Social Networks*, 1:159–192, 1978.

[99] B. J. Kim, C. N. Yoon, S. K. Han, and H. Jeong. Path finding strategies in scale-free networks. *Physical Review E*, 65(027103), 2002.

[100] Young-Jin Kim, Ramesh Govindan, Brad Karp, and Scott Shenker. Geographic routing made practical. In *Proceedings of the 2nd Annual Symposium on Networked Systems Design and Implementation (NSDI'05)*, May 2005.

[101] Andrew King. The evolution of RSS. `http://www.webreference.com/authoring/languages/xml/rss/1/`.

[102] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd Annual Symposium on the Theory of Computation (STOC'00)*, May 2000.

[103] Jon Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS'01)*, 2001.

[104] Jon Kleinberg, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The web as a graph: Measurements, models, and methods. In *Proceedings of the 9th International Conference on Combinatorics and Computing (COCOON'99)*, pages 1–18, 1999.

[105] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 24 August 2000.

[106] Judith Kleinfeld. Could it be a big world after all? The "six degrees of separation" myth. *Society*, 39(61), April 2002.

[107] Bryan Klimt and Yiming Yang. Introducing the Enron corpus. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, 2004.

[108] Manfred Kochen, editor. *The Small World: A Volume of Recent Research Advances Commemorating Ithiel de Sola Pool, Stanley Milgram, Theodore Newcomb.* Ablex, Norwood (NJ), 1989.

[109] C. Korte and S. Milgram. Acquaintance links between white and negro populations: Application of the small world method. *Journal of Personality and Social Psychology*, 15:101–118, 1970.

[110] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(066123):1–14, 2001.

[111] Valdis Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, Winter 2002.

[112] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493, 1999.

[113] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *Proceedings on the 12th International World Wide Web Conference (WWW12)*, pages 568–576, 2003.

[114] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, December 2004.

[115] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS'00)*, 2000.

[116] Chuck Lam. SNACK: incorporating social network information in automated collaborative filtering. In *Proceedings of the 5th Annual Conference on Electronic Commerce (EC'04)*, pages 254–255, May 2004.

[117] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 25–32, June 1999.

[118] Fredrik Liljeros, Christofer R. Edling, Luís A. Nunes Amaral, H. Eugene Stanley, and Yvonne Åberg. The web of human sexual contacts. *Nature*, 411:907–908, 21 June 2001. Available as `cond-mat/0106507`.

[119] Nan Lin, Paul W. Dayton, and Peter Greenwald. The urban communication network and social stratification: a "small world experiment". In B. D. Ruben, editor, *Communication Yearbook*, volume 1, pages 107–119. Transaction Books, New Brunswick, 1978.

[120] LiveJournal Statistics web page. Available at `http://www.livejournal.com/stats.bml`. Downloaded 20 January 2005.

[121] Alfred J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323, 1926.

[122] Tomasz Luczak. Component behavior near the critical point of the random graph process. *Random Structures and Algorithms*, 1:287–310, 1990.

[123] Gurmeet Singh Manku, Moni Naor, and Udi Wieder. Know thy neighbor's neighbor: the power of lookahead in randomized P2P networks. In *Proceedings of the 36th Annual Symposium on the Theory of Computing (STOC'04)*, pages 54–63, June 2004.

[124] Chip Martel and Van Nguyen. Analyzing Kleinberg's (and other) small-world models. In *Proceedings of the 23rd Annual Symposium on Principles of Distributed Computing (PODC'04)*, pages 179–188, July 2004.

[125] G. Melin and O. Persson. Studying research collaboration using co-authorships. *Scientometrics*, 36(3):363–377, 1996.

[126] Stanley Milgram. The small world problem. *Psychology Today*, 1:61–67, May 1967.

[127] Leonard E. Miller. Distribution of link distances in a wireless network. *Journal of Research of the National Institute of Standards and Technology*, 106(2):401–412, March/April 2001.

[128] Michael Mitzenmacher. A brief history of lognormal and power law distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[129] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–180, 1995.

[130] M. Molloy and B. Reed. The size of the largest component of a random graph on a fixed degree sequence. *Combinatorics, Probability and Computing*, 7:295–306, 1998.

[131] James Moody. Race, school integration, and friendship segregation in America. *American Journal of Sociology*, 107:679–716, 2001.

[132] Cristopher Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61:5678–5682, 2000. Available as `cond-mat/9911492`.

[133] Stephen Morris. Contagion. *Review of Economic Studies*, 67:57–78, 2000.

[134] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[135] M. E. J. Newman. Gallery of network images. `http://www-personal.umich.edu/~mejn/networks/`.

[136] M. E. J. Newman. Models of the small world. *Journal of Statistical Physics*, 101:819 841, 2000.

[137] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(025102), 2001. Available as `cond-mat/0104209`.

[138] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA*, 98:404–409, 2001.

[139] M. E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. Appears (in two parts) in *Physical Review E*, 64(016131 & 016132), 2001. Available as `cond-mat/0011144`.

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

[141] M. E. J. Newman. The spread of epidemic disease on networks. *Physical Review E*, 66(016128), 2002. Available as `cond-mat/0205009`.

[142] M. E. J. Newman. The structure and function of networks. *Computer Physics Communications*, 147:40–45, 2002.

[143] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(026126), 2003.

[144] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[145] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(035101), 2002.

[146] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68(036122), 2003. Available as `cond-mat/0305612`.

[147] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(026118), 2001. Available as `cond-mat/0007235`.

[148] M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263:341–346, 1999.

[149] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60:7332–7342, 1999.

[150] Van Nguyen and Chip Martel. Analyzing and characterizing small-world graphs. In *Proceedings of the Symposium on Discrete Algorithms (SODA'05)*, January 2005.

[151] Orkut. `http://www.orkut.com`.

[152] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Working Paper, November 1999.

[153] Christopher Palmer, Phillip Gibbons, and Christos Faloutsos. ANF: A fast and scalable tool for data mining in massive graphs. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD'02)*, July 2002.

[154] Romauldo Pasto-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, April 2001.

[155] A. Popescul and L. Ungar. Statistical relational learning for link prediction. In *Proceedings of the Workshop on Learning Statistical Models from Relational Data at the International Joint Conference on Artificial Intelligence (IJCAI'03)*, 2003.

[156] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences USA*, 101:2658–2663, 2004.

[157] P. Raghavan. Social networks: From the web to the enterprise. *IEEE Internet Computing*, 6(1):91–94, January/February 2002.

[158] Matt Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 61–70, 2002.

[159] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[160] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, 1999.

[161] John Scott. *Social Network Analysis: A Handbook.* Sage, London, 2000.

[162] R. L. Shotland. *University communication networks: The small world method.* Wiley, New York, 1976.

[163] Aleksandrs Slivkins. Distance estimation and object location via rings of neighbors. In *Proceedings of the 24th Annual Symposium on Principles of Distributed Computing (PODC'05)*, July 2005. To appear.

[164] Steven Strogatz. *Sync: The emerging science of spontaneous order.* Hyperion, 2003.

[165] Steven H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 8 March 2001.

[166] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *Advances in Neural Information Processing Systems (NIPS'03)*, December 2003.

[167] Prasad Tetali. Design of on-line algorithms using hitting times. *SIAM Journal on Computing*, 28:1232–1246, 1999.

[168] Brett Tjaden, Patrick Reynolds, et al. The Oracle of Bacon. `http://www.cs.virginia.edu/oracle/`.

[169] Topic Detection and Tracking (TDT-2004). `http://www.nist.gov/TDT`.

[170] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[171] United States Geological Survey. 2000 census. http://geonames.usgs.gov, 2000.

[172] Urban Dictionary. http://www.urbandictionary.com.

[173] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

[174] Duncan J. Watts. *Small Worlds.* Princeton University Press, 1999.

[175] Duncan J. Watts. *Six Degrees: The Science of the Connected Age.* W. W. Norton & Company, New York, 2003.

[176] Duncan J. Watts, Peter Sheridan Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 17 May 2002.

[177] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[178] WebFountain. http://www.almaden.ibm.com/WebFountain/.

[179] Wikipedia. http://en.wikipedia.org.

[180] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups. Manuscript, 2003. Available as cond-mat/0305305.

[181] Soon-Hyung Yook, Hawoong Jeong, and Albert-Lásló Barabási. Modeling the Internet's large-scale topology. *Proceedings of the National Academy of Sciences USA*, 99(21):13382–13386, 15 October 2002.

[182] H. Peyton Young. The diffusion of innovation in social networks. Sante Fe Institute Working Paper 02-04-018, 2002.

[183] Hwanjo Yu, ChengXiang Zhai, and Jiawei Han. Text classification from positive and unlabeled documents. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03)*, pages 232–239, November 2003.

[184] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.