Social Search in "Small-World" Experiments

Sharad Goel[1], Roby Muhamad[2], and Duncan Watts[1,2]
[1]Yahoo! Research, 111 West 40th Street, New York, NY 10018
[2]Department of Sociology, Columbia University, 420 West 118th Street, New York, NY 10027
goel@yahoo-inc.com, rm922@columbia.edu, djw@yahoo-inc.com

ABSTRACT

The "algorithmic small-world hypothesis" states that not only are pairs of individuals in a large social network connected by short paths, but that ordinary individuals can find these paths. Although theoretically plausible, empirical evidence for the hypothesis is limited, as most chains in "small-world" experiments fail to complete, thereby biasing estimates of "true" chain lengths. Using data from two recent small-world experiments, comprising a total of 162,328 message chains, and directed at one of 30 "targets" spread across 19 countries, we model heterogeneity in chain attrition rates as a function of individual attributes. We then introduce a rigorous way of estimating true chain lengths that is provably unbiased, and can account for empirically-observed variation in attrition rates. Our findings provide mixed support for the algorithmic hypothesis. On the one hand, it appears that roughly half of all chains can be completed in 6-7 steps—thus supporting the "six degrees of separation" assertion—but on the other hand, estimates of the mean are much longer, suggesting that for at least some of the population, the world is not "small" in the algorithmic sense. We conclude that search distances in social networks are fundamentally different from topological distances, for which the mean and median of the shortest path lengths between nodes tend to be similar.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Experimentation, Human Factors

Keywords

Social search, small-world experiment, attrition

1. INTRODUCTION

For forty years, the provocative "small-world" experiments of Stanley Milgram and colleagues [17, 25, 34] have been cited as evidence that everyone in the world is connected to everyone else via "six degrees of separation." This claim, however, can be interpreted in two very different ways. The first interpretation—which we call the "topological" version of the hypothesis—holds only that for a randomly chosen pair of individuals, there exists with high probability a short chain of intermediaries that connects them, where "short" is usually interpreted as proportional to the logarithm of the population size [38]. The second interpretation makes a much stronger claim—namely that ordinary individuals can effectively "navigate" these short chains themselves, with every individual having only local knowledge of the social network in question [1, 2, 13, 14, 21, 26, 33, 37]. For this reason, it has been labeled the "algorithmic" small-world problem [13, 14].

Distinguishing between the topological and algorithmic interpretations of the small-world hypothesis is important because each is relevant to different social processes [36]. The spread of a sexually-transmitted disease along networks of sexual relations, for example, does not require that participants have any awareness of the disease, or intention to spread it; thus for an individual to be at risk of acquiring an infection, he or she need only be connected in the topological sense to existing infectives. On the contrary, individuals attempting to "network"—in order to locate some resources like a new job [9] or a service provider [19]—must actively traverse chains of referrals; thus must be connected in the algorithmic sense. Depending on the application of interest, therefore, either the topological or algorithmic distance between individuals may be more relevant—or possibly both together.

Given the distinct implications of topological versus algorithm connectivity, it is also important to note that the two interpretations are supported by very different kinds of evidence. In recent years, hundreds of empirical studies of large-scale networks have been conducted across a number of domains, including not only social networks [8, 18, 20, 27], but also biological [12, 35, 38], technological [31, 38], organizational [16], and virtual networks [4]. Two near-universal findings of these studies are: (1) that a majority of individuals in a given population are connected in a single "giant component"; and (2) that the typical shortest path length connecting pairs of nodes within the giant component is on the order of the logarithm of the system size. In a recent study of a network of 180M instant messenger users (where a link is defined as two users nominating each other as IM "buddies"), for example, Leskovec and Horvitz [20] found that users were separated by a mean of 6.6 steps and a median of 7 steps. Taken together, therefore, these studies provide overwhelming evidence for the topological interpretation of the small-world hypothesis, confirming it

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

for networks spanning six orders of magnitude in size (from hundreds to hundreds of millions), and across many substantive domains.

Empirical evidence in favor of the algorithmic small-world hypothesis—that individuals can locate these paths—is, however, much less conclusive [15, 36]. A handful of classic sociological case studies—such as Lee's "The search for an abortionist" [19] and Granovetter's "Getting a job" [9]—along with numerous anecdotal examples from the growing literature of "networking" books [24, 28, 30]—suggest that in at least some circumstances individuals can indeed navigate their social networks to locate useful resources. It is unclear, however, whether these examples provide evidence for a general "searchability" property of social networks that allows even quite ordinary individuals to perform successful social searches under routine conditions, or instead represent a collection of special cases—either because the individuals in question, or their circumstances, are somehow atypical.

In recent years, a number of mathematical and simulation models [1, 2, 13, 14, 21, 26, 33, 37] have been proposed with the objective of making a theoretical case in favor of the generic searchability of social networks. Although suggestive, these results rely on some form of simulation—of the underlying social network, of the search procedure, or of both together—and thus they still do not constitute direct empirical evidence for the key claim of the algorithmic small-world hypothesis, that ordinary individuals in large social networks can routinely locate short paths. More seriously, perhaps, the models described above all tend to make certain homogeneity assumptions that treat all individuals as equivalent; hence, they have little to say about the impact on social search of well-known sources of heterogeneity and inequality in real social networks [5, 29]. Mathematical models, computer simulations, and isolated examples aside, therefore, the primary source of empirical evidence for the algorithmic small-world hypothesis continues to be that provided by small-world type experiments of the kind invented by Milgram.

How persuasive, then, are these findings? As Kleinberg [13] noted in his initial formulation of the algorithmic small-world problem, the experiment conducted by Travers and Milgram [25, 34] clearly demonstrated that at least some people are able to construct short chains, comprising an average of approximately five intermediaries, that connect distant senders from a chosen target individual. But as Travers and Milgram themselves emphasized, it was also the case that most of the chains that started—roughly 80% in their case—never reached the target. Subsequent experiments [6, 11, 17, 22, 23, 32] have demonstrated much the same pattern: On the one hand, chains that reach their targets tend to be short; but on the other hand, the rate of chain completion tends to be low. In Korte and Milgram's follow-up study of white senders in Los Angeles attempting to reach one of 270 black targets in New York [17], the average length of completed chains was 7, but the completion rate was only 13%. And in the most recently repeated small-world experiment, conducted by Dodds et. al. [6], the pattern was even more striking: Completed chains were only 4 steps long, but only 0.4% of about 24,000 chains that started (a total of 384) reached their targets.

References to "six degrees of separation" [10] typically emphasize the first of these results—that completed chains tend to be short—but here we wish to emphasize the second finding that the vast majority of chains never reach their ultimate target, and therefore that most of the desired data on chain length is effectively missing. Determining what chain attrition tells us about the searchability of social networks is therefore the main object of this paper, which is organized as follows. In the next section, we review the standard approach to handling chain attrition, and raise two possible objections that have not been resolved by previous studies. In Section 3, we describe data from two recent small-world experiments, and model attrition as a function of individual and relational attributes. In Section 4, we derive a general correction for missing data based on the classical statistical idea of importance sampling. In Section 5 we estimate chain length distributions based on the attrition model and the importance sampling estimator. We find that although the median chain length is consistent with earlier estimates, the mean is significantly longer than previously believed. Finally, in Section 6, we close with a discussion of the new evidence, concluding that the usual interpretation of the small-world hypothesis requires some caveats.

2. RELATED WORK

The conventional wisdom regarding chain attrition in small-world experiments, initially proposed by White [39], is that message chains terminate for reasons that are unrelated either to the topology of the underlying network, or to the search process. Rather, it is assumed that participants fail to pass on messages either because they are not sufficiently motivated to do so, or else because they fail to receive them in the first place. Message passing can therefore be viewed as a stochastic process that takes place on some network, where the only parameter governing the success of a search is the probability of termination at each step.[1]

A convenient feature of this stochastic attrition interpretation is that it can be invoked to derive estimates of the "true" length of chains—that is, the length distribution of chains that would have been observed had no attrition taken place. In particular, White [39] proposed the estimator

$$\hat{p}_l = \frac{\tilde{p}_l}{\prod_{j=0}^{l-2}(1-r_j)} \tag{1}$$

where $\hat{p}_l$ is an estimate of the "true" percentage of paths with length l, $\tilde{p}_l$ is the observed percentage of completed chains with length l, and r_j is the probability of attrition from step j to step $j+1$. In deriving this estimator, White assumed that individuals who knew the target directly would always pass on the message; thus "last-step attrition" would be zero. Although plausible, this assumption is not inevitable, nor was it based on empirical evidence. Dodds et al. [6] therefore removed this assumption, resulting in the estimator

$$\hat{p}_l = \frac{\tilde{p}_l}{\prod_{j=0}^{l-1}(1-r_j)} \tag{2}$$

which is identical to (1), except that the product in the denominator is from $j=0$ to $j=l-1$. Importantly, it follows from either estimator that the "true" average path length is longer than that for observed chains, for the simple reason that longer chains are more likely to terminate than shorter chains. For example, White estimated that in the absence

[1] In the original Travers and Milgram experiment, for example, roughly 75% of messages were passed on at each step, corresponding to a 25% "attrition rate."

of attrition, Milgram's experiment would have yielded chain lengths in the vicinity of eight steps, not six. The resulting estimates, however, are still "short" in the sense of being comparable to that of an equivalent random graph (i.e., of the same size and average density); thus the upshot of these methods, and the conclusion of most previous studies, is that as long as the stochastic failure assumption is reasonable, then it would appear that the observation of even a small fraction of short chains in small-world experiments is sufficient to support the algorithmic small-world hypothesis.

An alternative interpretation of results from small-world experiments, however, takes direct issue with the stochastic attrition assumption, contending that low completion rates should be seen not merely as missing data, but as evidence that most pairs of individuals are not connected by short, navigable paths. For example, Kleinfeld [15] has argued that

> The research on the small-world problem suggests not a counter-intuitive triumph of social research, but an all-too familiar pattern: We live in a world where social capital, the ability to make personal connections, is not wide-spread and more apt to be a possession of people of higher social status. Many small worlds do exist, such as scientists with worldwide connections or university administrators at a single campus. Rather than living in a "small, small world," we may live in a world that looks a lot like a bowl of lumpy oatmeal, with many small worlds loosely connected and perhaps some small worlds not connected at all.

Objections of this kind raise two distinct ways in which low completion rates may be systematically related to the searchability of the network itself. First, it may be the case simply that some people are good at conducting social searches, while others are not. Exceptionally determined or resourceful individuals, that is, may be able to construct short paths to distant targets; but most people lacking such "social capital" [5, 29] are effectively isolated from one another. A second objection, compounding the first, is that chains terminate because most source-target pairs live in effectively separate populations that can only be reached via long or otherwise undiscoverable paths—that is, "many small worlds, loosely connected." Thus, pairs of individuals who are already "close", in the sense of sharing social, geographical, and demographic attributes, can find each other, but "distant" pairs cannot. Because, moreover, many more pairs are distant than close, the bulk of message chains should not be expected to complete—as indeed is observed in small-world experiments.

Although distinct, both objections call into question the veracity of the estimators (1) and (2). Specifically, the first objection implies that chain attrition is not merely a function of unrelated extrinsic factors, as assumed by the stochastic attrition hypothesis, but rather reflects variability among individuals. Accordingly, one would expect chain attrition to correlate strongly with individual-level attributes such as education, income, and so on, that are typically related to social capital [5, 29]. Meanwhile, the second objection—that many source-target pairs live in "distant" groups—implies that long chains are not being counted in the estimation procedure, and that the resulting estimators are therefore biased. Because the estimators (1) and (2) fail to account for heterogeneity in attrition, and also neglect to provide any guarantee that they are unbiased, one might suspect that the "true" chain length estimates are higher—possibly much higher—than conventional wisdom allows.

In what follows, we examine data from two small-world experiments more systematically than in previous studies, and introduce an estimator that is (a) provably unbiased, and (b) allows for heterogeneous attrition rates. We find support for each of the interpretations discussed above: Specifically, the median "true" chain length (i.e. in a hypothetical experiment with zero attrition) can be estimated robustly to be about 6-7; however, the mean of this distribution is likely to be much larger. In other words, at least half of all chains that start would be expected to reach their targets within about six steps; but at least some chains should be expected to be very long, consistent with the claim that some pairs of individuals are extremely unlikely to be able to locate each other.

3. MODELING ATTRITION

Our data come from two recent experiments—one of which has been reported on previously [6]—that were designed to replicate Milgram's small-world method, except using email instead of physical packets, a change that permitted access to a much larger and more diverse population than was available to Milgram, and at lower cost [3]. The first of these experiments was conducted between December 2001 and August 2003; and the second version followed immediately thereafter, and ran until December 2007. In the first experiment, 98,865 people from 168 countries initiated 106,295 chains directed at 18 targets in 13 countries. In the second experiment, 85,621 people from 163 countries participated in 56,033 chains, directed at 21 targets in 13 countries. In both cases, participants were mostly from the United States and Western Europe, were predominantly white and Christian, and were largely young, college-educated, middle-class professionals.[2]

The results exhibit the same combination of short path lengths and low completion rates that typify small-world

[2]Five targets were recruited directly by members of the research teams, and the remaining targets were selected from more than 4,000 volunteers (recruited via the website) with the aim of achieving as diverse a pool as possible. Initial senders were recruited from among people who had heard about the experiment from the media or by word of mouth, where, in order to obtain as many participants as possible, there was no attempt to control the characteristics of senders. Participants registered at the website and were asked to reach target persons, with the restriction that they could only advance the chain through people whom they already knew. The following message was then emailed to the chosen recipients:

> We are testing the idea that everyone in the world can be reached through a short chain of social acquaintances. Your objective is to use your social connections to move a message "closer" to a particular "target" person.

Upon receiving this message, recipients had to verify whether they knew the senders and then send the message on to another person whom they knew; and this process continued until the target was reached. Thus, they created a series of message chains from initiators to targets. The experiment also recorded demographics data and relational attributes (type, origin, strength of relationship, and reasons for choosing recipients) between senders and recipients.

experiments. The completion rates for these experiments, however, were much lower than those recorded by Milgram and his colleagues: whereas Travers and Milgram experienced roughly 20% completion, in Experiment 1, a total of 491 chains (0.5%) successfully reached their targets[3]; and in Experiment 2, the completion rate was 0.1% (61 chains). These ultra-low completion rates are directly attributable to the peculiar design of the small-world method, in which chain completion rates diminish exponentially with chain length.[4]

In order to address the first critique, above, that the estimators (1) and (2) do not account for individual-level heterogeneity, we first need to model attrition in terms of individual-level attributes, like socioeconomic status, gender, education, and age, that are typically associated with differences in social capital [5, 29]. Unfortunately, measuring the impact of individual attributes on chain attrition is complicated by another feature of small-world experiments: The majority of people who did not continue chains never came to the experiment website; thus we do not have data on their individual attributes. As a substitute, therefore, we instead estimate the "next-step continuance" probability: Given an individual A who forwards the message to B (and assuming B is not the ultimate target), we estimate the probability that B continues the chain. In this sense, we are effectively modeling the "search ability" of our participants (i.e., their ability to choose someone who will pass along the message). As such, our model also includes relational attributes (between A and B) and is adjusted for the target. In total, we analyzed 88,875 sender-recipient pairs[5], of which 32% of pairs comprised recipients who forwarded messages (continued links), and 68% comprised recipients who did not (terminated links).

We estimate the next-step continuance probability by logistic multilevel regression (also known as "hierarchical linear modeling"), a standard statistical tool for modeling data with group structure [7]. In our case, for example, "high school," "college" and "graduate school" would be separate groups in the category of education. Multilevel regression can be thought of as a compromise between two extreme approaches to pooling data from different groups: no pooling and complete pooling. No pooling corresponds to treating different groups within the same category as unrelated; thus one would fit distinct parameters for each group (e.g. high school versus college) without imposing any relationship amongst them. By contrast, complete pooling effectively corresponds to ignoring the presence of groups within a particular category; for example, treating all individuals as the same regardless of educational status. In the middle ground of multilevel models, one allows for the possibility that groups within a category are related, without specifying a hard constraint on the strength of their relationship. Specifically, our model is of the form

$$\mathbb{P}(y_i = 1) = \text{logit}^{-1}(\gamma + \beta_{\text{nonwhite}} X_{\text{nonwhite},i} + \beta_{female} X_{female,i} + \alpha_{j_1[i]} + \alpha_{j_9[i]})$$

where the outcome variable y_i indicates the next-step continuance, γ is the intercept, the two β terms are fixed effects for female and nonwhite participants respectively, and the $\alpha_{j_k[i]}$ correspond to the nine group effects. For each category k (e.g., education), $j_k[i]$ is the group (e.g., high school, college, graduate school, etc.) of the i^{th} response, and we model the group parameters within each category as coming from a normal distribution: $\alpha_{j_k[i]} \sim N(0, \sigma_k^2)$. The subscript $j_k[i]$, in other words, can be thought of as a mapping from a particular response i to a particular group $j_k[i]$ within category k.

Our model therefore includes a total of 66 parameters: an overall intercept term; one parameter each for gender (male/female) and race (white/non-white); 54 distinct attribute parameters, which in turn are grouped into nine categories[6] (age, education, work field, work position, income, strength of relationship with message recipient, reason for choosing recipient, origin of relationship with recipient, and target); and one variance parameter for each category. The estimated attrition rate for a particular individual (e.g., female, 30-39 years old, with a college degree, etc.) is obtained by adding together the relevant terms in the model, and taking logit^{-1} of the sum. The variance parameters σ_k^2 indicate how closely related the groups are within each category k, and themselves are inferred from the data; large variance corresponds to weak association of groups (i.e. no pooling), and small variance corresponds to strong association (i.e. complete pooling).

Table 1 shows the standard deviations σ_k for the nine categories[7] (first column). Although these σ_k capture the typical effect of the category on attrition, they do so on the logit scale, which is difficult to interpret. To aid interpretation, therefore, these values are translated in the second column to a probability scale that is relative to the baseline attrition rate. For example, differences between targets account for a 2% absolute change in attrition rates, while differences in the education of senders accounts for 3%. Since the baseline attrition rate is 30%, a 2% absolute change corresponds to a 7% relative change. Furthermore, as discussed below, these differences are amplified by the correlation of attributes to-

[3]Although we have used the same raw data as Dodds et al. [6], changes in our coding and some subsequent cleaning of the data have resulted in somewhat different figures for number of participants and completed chains. There are three main reasons for these discrepancies. First, we now include a number of chains that were originally excluded by Dodds et al [6] because they contained unidentified individuals. Second, Dodds et al. [6] required actual emails to be sent between two people in order to establish a connection. After closer examination, however, we found that there were people-especially those directly adjacent to targets-who received more than one message but who had forwarded only one. Here we count a connection whenever we know that it exists from previous email exchanges; thus we have a higher total number for both incomplete and completed chains. Finally, we have added three demographic variables (ethnicity, work industry, and work position) that were inaccessible previously due to many participants answering the "other" category. We solved the problem by checking manually all uncategorized answers and putting them into relevant categories.

[4]For example, if 100,000 chains are initiated with a 25% constant attrition, in six removes there are 1780 chains left, whereas with a 67% attrition rate, only 129 survive.

[5]By contrast, we note that there were only 491 completed chains; thus by studying links instead of completed chains, we dramatically increase the number of relevant data points.

[6]We found religion, country, type of relationship, and current chain length were not statistically significant predictors.

[7]Although is is difficult to compute overall standard errors for the variance parameters, Table 2 states standard errors for each group-level coefficient within every category.

Table 1: Category standard deviation parameters from a multilevel logistic regression model of next-step continuance probabilities. Attrition is stated as typical deviation from the baseline of 30% for white males.

Category	σ_k	Attrition
Target	0.09	± 0.02
Age	0.12	± 0.02
Relationship Origin	0.04	± 0.01
Income	0.07	± 0.01
Work Position	0.03	± 0.01
Work Field	0.08	± 0.02
Reason for Choosing Recipient	0.05	± 0.01
Relationship Strength	0.11	± 0.02
Education Level	0.14	± 0.03

gether with the compounding effects of chain-based message propagation.

Table 2 develops this analysis, where now we examine the effects of individual and relational attributes on the next-step continuance probability, relative to the baseline attrition of 30% (for typical white males). Each row in Table 2 corresponds to a group (e.g., "college" and "18-29") within the nine attribute categories (e.g., "education" and "age"), as well as the overall intercept, and the two fixed effects for females and non-whites. For any given group (i.e., row in the table), the first table column is the estimated coefficient for that group[8], along with its associated standard error; and the second column gives the corresponding effect on the next-step continuance rate.

Consistent with our interpretation of Table 1, Table 2 reveals a small but significant range of attrition rates, with individuals possessing the attributes of high social capital associated with higher continuance probabilities: possessing a graduate education, for example, increased pass-along by 4% above the baseline, whereas having only a high-school education diminished it by 3%; and whereas participants earning over $100,000 were 2% more likely than average to pass along a chain, those earning less than $25,000 were 1% less likely. As one might suspect, individual-level heterogeneity in attrition rates tends to be correlated across attributes (e.g., high education is associated with high income); thus the overall distribution of estimated attrition rates for participants is considerably greater than is indicated by any single group effect, with attrition rates varying from 60% to 80%, as shown in Figure 1.

Overall, therefore, our analysis shows that high status individuals are more likely to pass along messages to friends who again pass them along, and that these differences, once compounded over multiple attributes, can be large. One might also note, however, that the distribution in Figure 1 is sharply peaked around the mean; thus although individual differences can be large, they are typically small. Regardless, the homogeneity assumption used in estimators (1) and (2) is clearly invalid. To understand the relation between attrition and chain completion, therefore, we require an estimator that can account for heterogeneous attrition rates. We also wish to address the second criticism raised in Sec-

[8]Since gender and race are fixed effects in our model, we use white males as a baseline group.

Table 2: Coefficient estimates from a multilevel logistic regression model of next-step continuance probabilities. The probabilities are stated as deviation from the baseline of 30% for white males.

Attributes	Coef. (S.E.)	Probability
Age		
Under 17	0.038 (0.11)	0.01
18-29	0.14 (0.06)	0.03
30-39	0.090 (0.06)	0.02
40-49	-0.068 (0.06)	-0.01
50-59	-0.071 (0.06)	-0.02
Above 60	-0.13 (0.07)	-0.03
Education Level		
Graduate school	0.18 (0.08)	0.04
College/University	0.014 (0.08)	0.0
High school	-0.14 (0.08)	-0.03
Elementary school	-0.048 (0.11)	-0.01
Work Field		
Media/Advertising/Arts	0.098 (0.05)	0.02
Education/Science	0.059 (0.04)	0.01
IT/Telecommunication	-0.018 (0.05)	0.0
Government	-0.056 (0.05)	-0.01
Other	-0.084 (0.04)	-0.02
Work Position		
Specialist/Technical	0.028 (0.03)	0.01
Student	0.016 (0.03)	0.0
Other	0.00049 (0.02)	0.0
Unemployed/Retired	-0.0045 (0.03)	0.0
Executive/Manager	-0.040 (0.02)	-0.01
Income		
More than $100,000	0.076 (0.04)	0.02
$50,000 - $100,000	0.052 (0.04)	0.01
$25,000 - $49,999	-0.0078 (0.04)	0.0
$2,000 - $24,999	-0.056 (0.04)	-0.01
Less than $2000	-0.064 (0.05)	-0.01
Relationship Strength		
Extremely close	0.13 (0.06)	0.03
Very close	-0.013 (0.05)	0.0
Fairly close	0.05 (0.05)	0.01
Casually	-0.0093 (0.05)	0.0
Not close	-0.16 (0.07)	-0.03
Reason for Choosing Recipient		
Profession	0.033 (0.04)	0.01
Education	0.031 (0.04)	0.01
Work brings contact	0.020 (0.04)	0.0
Geography	-0.010 (0.03)	0.0
Other	-0.074 (0.03)	-0.01
Relationship Origin		
Work	0.043 (0.03)	0.01
School	0.025 (0.03)	0.0
Internet	0.014 (0.03)	0.0
Mutual friend	-0.013 (0.03)	0.0
Relative	-0.028 (0.03)	-0.01
Other	-0.041 (0.03)	-0.01
Fixed Effects		
Intercept	-0.85 (0.12)	NA
Female	-0.063 (0.025)	-0.01
Nonwhite	-0.13 (0.041)	-0.03

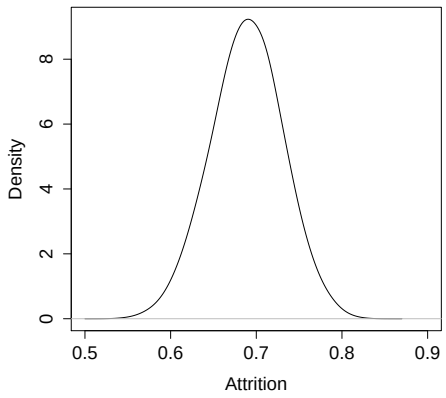

Figure 1: The estimated distribution of attrition over individuals. Average attrition 0.7.

tion 2, which is that the estimators (1) and (2) have not been proven to be unbiased, and therefore may not account correctly for the presence of long, but unobserved chains.

4. CORRECTING FOR MISSING DATA

We address both these problems by introducing a general technique to correct for missing data based on the classical statistical idea of importance sampling. Although at a high level importance sampling is similar to estimators (1) and (2), its more rigorous formulation permits us to guarantee that estimates are unbiased, and to specify errors on the estimates. It also allows us to incorporate heterogeneous attrition rates in a straightforward manner

Intuitively, our estimation procedure works in the following manner. For each non-missing data point (i.e., completed chain), we estimate the likelihood of that data point having been observed. We then re-weight the non-missing values to account for the fact that some of the data points were more likely to have been observed (i.e., the shorter chains) than others (i.e., the longer chains). More formally, we consider the space of all possible paths between all pairs of individuals. In this hypothetical space, there may be many paths that connect any two given individuals, and so we reasonably assume that some paths are more likely to be traversed than others. In an ideal small world experiment without attrition, we would repeatedly observe the lengths of these random paths between random pairs of individuals. In the more general setting, we suppose there is a discrete outcome space Ω (e.g., the space of paths) that is equipped with a probability P (e.g., P describes the likelihood of selecting any particular path amongst all paths between all pairs of individuals). An experimental trial corresponds to observing an outcome (e.g., a path) $\omega \in \Omega$ drawn with probability P.

In the case of missing data, however, we do not always get to see the uncorrupted outcome $\omega \in \Omega$ (e.g., the path). To model this missing data situation, we suppose that after an outcome ω is drawn, it is observed with probability $Q(\omega)$ and reported missing with probability $1 - Q(\omega)$. In our small-world setting, "missing" corresponds to an incomplete chain, and due to attrition, $Q(\omega)$ is generally smaller for longer paths—longer "true" paths are more likely to be "observed" as incomplete chains[9]. Hence, the outcome $\omega \in \Omega$ is ultimately "observed to complete" with probability $P(\omega)Q(\omega)$. In particular, summing over all possible outcomes, a non-missing value is observed on a given trial with probability $\sum_\omega P(\omega)Q(\omega)$, and a missing values is observed with probability $1 - \sum_\omega P(\omega)Q(\omega)$. Let $X_1, \ldots, X_n$ denote n such independent trials (in this case, n is the total number of chains that are initiated), where X_i is either an uncorrupted observation (e.g., a completed chain) or a missing value (e.g., an incomplete chain). That is, $X_i \in \Omega \cup \{\texttt{NA}\}$ where NA denotes a missing value.

The "true" expected chain length without attrition μ can be expressed as a weighted average over the space of all paths (higher weights correspond to more likely paths):

$$\mu = \sum_\omega f(\omega)P(\omega)$$

where $f(\omega)$ denotes the length of the path ω[10]. Without missing values (i.e., $Q(\omega) = 1$), the usual unbiased estimator for μ is the sample average $\frac{1}{n}\sum_{i=1}^n f(X_i)$. With missing values, (i.e., $Q(\omega) < 1$), averaging over all the non-missing values in the sample biases our estimate toward outcomes that we are more likely to observe[11]. We adjust for this biased observation of outcomes by re-weighting samples by their inverse probability of observation, an idea based on the classical statistical technique of importance sampling. This re-weighting results in an unbiased estimator, as given in Theorem 1.

THEOREM 1. *In the general setting described above, an unbiased estimate of the mean $\mu = \sum_\omega f(\omega)P(\omega)$ is given by*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^m \frac{f(X_{k_i})}{Q(X_{k_i})}$$

where $X_{k_1}, \ldots, X_{k_m}$ are the m observed, non-missing values, and $Q(\omega)$ is the probability that ω is observed uncorrupted after it has been sampled.

PROOF. First extend f to a function $\bar{f}$ defined on $\Omega \cup \{\texttt{NA}\}$ (where NA denotes a missing value), by setting $\bar{f}(\omega) = f(\omega)$ for $\omega \in \Omega$ and $\bar{f}(\texttt{NA}) = 0$. Then we can rewrite the estimator as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \frac{\bar{f}(X_i)}{Q(X_i)}$$

where the sum is taken over all samples (including the missing values). Since the samples X_i are identically distributed,

[9]We emphasize that our use of the term "observed" corresponds to its usage in probability theory—that is, as a random trial in an experiment—and does not imply that the chain is observed to complete. Indeed, the point is that most chains that are "observed" in the statistical sense do not complete. Thus one should think of all chains as being observed, where some (a minority) are "observed to complete," while others (a majority) are "observed not to complete."

[10]More generally, we might be interested in the expectation of an arbitrary function f.

[11]We emphasize again that this problem is particularly egregious in the case of estimating chain lengths, since the probability of observing a chain tends to decrease exponentially with its length; thus we are much more likely to see short chains, and hence underestimate mean chain length.

$\mathbb{E}\hat{\mu} = \mathbb{E}[\bar{f}(X_i)/Q(X_i)]$. Finally, since ω is observed non-missing with probability $P(\omega)Q(\omega)$, and $\bar{f}(\texttt{NA}) = 0$, we have

$$\mathbb{E}\left[\frac{\bar{f}(X_i)}{Q(X_i)}\right] = \sum_{\omega\in\Omega}\frac{f(\omega)}{Q(\omega)}P(\omega)Q(\omega) = \sum_{\omega\in\Omega} f(\omega)P(\omega) = \mu.$$

Hence, $\hat{\mu}$ is unbiased. □

Theorem 1 shows that $\hat{\mu}$ generates an unbiased estimate of $\mu = \sum_\omega f(\omega)P(\omega)$ for any function f. When $f(\omega)$ is the length of a chain ω, then $\hat{\mu}$ estimates the mean chain length.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{m}\frac{L(X_{k_i})}{Q(X_{k_i})} \tag{3}$$

where $X_{k_1}, \ldots, X_{k_m}$ are the m observed completed chains, $L(\omega)$ is the length of chain ω, and n is the total number of complete and incomplete chains in the sample. Theorem 1 can also be used to estimate the entire chain length distribution. Set $f_i(\omega) = 1$ if ω is a chain of length i, and $f_i(\omega) = 0$ otherwise. Then the expectation of f_i is given by

$$p_i = \sum_{\omega\in\Omega} f_i(\omega)P(\omega) = \sum_{\omega \text{ is of length } i} P(\omega).$$

That is, p_i is the probability that a randomly chosen chain has "true" length i. Applying Theorem 1 to this indicator function f_i yields an unbiased estimate of p_i:

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{m}\frac{f_i(X_{k_j})}{Q(X_{k_j})} = \frac{1}{n}\left[\sum_{X_k \text{ has length } i}\frac{1}{Q(X_k)}\right]. \tag{4}$$

Computing $\hat{p}_i$ for $i = 1, 2, \ldots$ gives an approximation of the true chain length distribution, and in particular, allows us to estimate the true median chain length in an idealized world without attrition.

Theorem 1 shows that $\hat{\mu}$ is unbiased; Corollary 1, below, derives the variance of $\hat{\mu}$, and in particular, shows that the variance of $\hat{\mu}$ increases as the probability of observation $Q(\omega)$ decreases. That is, when data are more likely to go missing, our estimates are understandably more variable.

COROLLARY 1. *The variance of $\hat{\mu}$ satisfies*

$$Var(\hat{\mu}) = \frac{1}{n}\left[\left(\sum_{\omega\in\Omega} f^2(\omega)\frac{P(\omega)}{Q^2(\omega)}\right) - \mu^2\right].$$

PROOF. Using the notation of Theorem 1, by independence,

$$\text{var}(\hat{\mu}) = \frac{1}{n}\text{var}\left(\frac{\bar{f}(X_1)}{Q(X_1)}\right).$$

Since $\hat{\mu}$ is unbiased,

$$\begin{aligned}\text{var}\left(\frac{\bar{f}(X_1)}{Q(X_1)}\right) &= \mathbb{E}\left(\frac{\bar{f}^2(X_1)}{Q^2(X_1)}\right) - \mu^2 \\ &= \left(\sum_{\omega\in\Omega} f^2(\omega)\frac{P(\omega)}{Q^2(\omega)}\right) - \mu^2\end{aligned}$$

and the result is shown. □

In Corollary 1, the variance of $\hat{\mu}$ is expressed in terms of the "true" probability of selection $P(\omega)$, which is not usually available in applications. Hence, the variance cannot generally be computed directly by this formula. Furthermore, in practice variability in estimates of the mean is further increased by noise in estimates of attrition (i.e., noise in estimates of $Q(\omega)$). We account for these two sources of variance by a modified bootstrap sampling procedure described in Section 5.

5. ESTIMATING CHAIN LENGTHS

Applying the attrition model of Section 3 together with the missing data correction of Section 4, we now estimate the "true" chain length distribution based on the small-world data. We start by examining the Travers and Milgram [34] data, and the Dodds et al. [6] data, under the assumption of homogeneous attrition; and then proceed with the main analysis, an investigation of chain length with heterogeneous attrition.

Homogeneous Attrition. For estimation based on the homogeneous attrition model, we assume that individuals terminate chains (i.e., they do not forward messages) with probability r, independent of their personal attributes; hence, individuals independently forward messages with probability $1 - r$. Since a chain ω of length $L(\omega)$ reaches its target only if each individual in the chain forwards the message, in the homogeneous attrition model the chain is observed as complete with probability $Q(\omega) = (1 - r)^{L(\omega)}$.

We generate generate confidence intervals for our estimates by bootstrap sampling. From the original set of n complete and incomplete chains, we first resample n chains with replacement, generating a bootstrap sample S_1, where we note that in the sample some of the original chains appear more than once while others do not appear at all. Repeating this resampling procedure $k = 10,000$ times produces k bootstrap samples $S_1, \ldots, S_k$ where each sample S_i is itself a random resampling of the original data. These randomly generated sets of chains simulate what would have been observed had we been able to repeat the entire experiment k times; thus we can obtain k estimates of the "true" mean chain length[12]. The confidence interval for the mean chain length is then defined by the range of the middle 95% of these k bootstrap estimates.

In the Travers and Milgram data, we empirically find an attrition rate of $r = .25$, and so have $Q(\omega) = (.75)^{L(\omega)}$. Applying the estimator (3) with this choice of $Q(\omega)$ yields an estimated "true" mean of 11.8 (95% CI: 8.5-15), compared with the observed mean 6.2. The estimator (4) correspondingly allows us to approximate the entire chain length distribution, from which a median of 7 (95% CI: 6-7) is computed, in approximate agreement[13] with the previously reported median of 8 [39].

Next we consider the Dodds et al. [6] data, again under a homogeneous attrition assumption. Here, however, we consider a more nuanced attrition model in which the

[12]We apply the same estimator (3) to each "experiment" but because the data given to the estimator changes with each bootstrap sample, we obtain different estimates.

[13]The agreement is even closer than it seems, on account of additional assumption made by White that senders at the last remove, being personally acquainted with the target (by definition) would always (i.e. with 100% likelihood) pass on the letter. As noted earlier, by effectively increasing the probability that chains would complete, this assumption actually increases the estimated chain length by one, versus the alternative assumption (made by Dodds et al. [6]) that all links were equally susceptible to attrition. Thus White's estimate of 8 is essentially the same as our estimate of 7.

first individual in a chain has attrition $r_0 = .41$ and all the remaining individuals in the chain have attrition $r = .70$. These values of attrition were empirically determined from the data, and the substantial difference found between r_0 and r motivates our decision to treat them separately[14]. To reach the target, the chain initiator must first forward the message to the second participant in the chain (which happens with probability $1 - r_0$), and then each of the next participants must also forward the message (which occurs with probability $(1-r)^{L(\omega)-1}$). Consequently, in this model, the probability $Q(\omega)$ that a chain reaches its target (i.e., is observed as complete) is given by

$$Q(\omega) = (1 - r_0)(1 - r)^{L(\omega)-1} = (1 - .41)(1 - .70)^{L(\omega)-1}.$$

With this choice of $Q(\omega)$, the estimator (3) for mean chain length and the estimator (4) for the full chain length distribution yield a "true" mean length of 41.5 (95% CI: 20-68) and a robust median of 6 (95% CI: 6-6). This estimate of the median is in agreement with Dodds et al.'s estimated range of 5-7 [6].

Heterogeneous Attrition. We now proceed with the main analysis, estimating the chain length distribution while allowing for heterogeneous attrition, as modeled in Section 3. Recall that due to limits in the experimental design, we estimate the "next-step continuance" probability: Given an individual A who forwards the message to B (and assuming B is not the ultimate target), we estimate the probability that B continues the chain based on A's attributes. In particular, the next-step continuance probability is a function of A's gender, race, age, education, work field, work position, income, strength of A's relationship with B, A's reason for selecting B, origin of A's relationship with B, and the final target.

We first define the probability $Q(\omega)$ of observing a completed chain in this heterogeneous attrition model. For a chain ω of length $L(\omega)$, enumerate the links in the chain by $\omega_0 \to \omega_1, \omega_1 \to \omega_2, \ldots, \omega_{L(\omega)-1} \to \omega_{L(\omega)}$, and let $r_{\omega_i \to \omega_{i+1}}$ denote the probability that the i^{th} participant in the chain fails to forwards the message. We assume the chain initiator fails to forwards the message with fixed probability $r_{\omega_0 \to \omega_1} = .41$, found empirically from the data. For $i > 1$, $r_{\omega_i \to \omega_{i+1}}$ is estimated from the next-step continuance model of Section 3 based on the attributes of the $(i-1)^{st}$ participant in the chain ω. For example, if the $(i-1)^{st}$ participant in ω is a 20-29 year-old white male with a graduate degree, etc., we may estimate $r_{\omega_i \to \omega_{i+1}} = .67$. Combining these continuance probabilities from each step, we define the probability of observing a complete chain (i.e., a chain that makes it to its target) by

$$Q(\omega) = (1 - r_{\omega_0 \to \omega_1})\,(1 - r_{\omega_1 \to \omega_2}) \cdots \left(1 - r_{\omega_{L(\omega)-1} \to \omega_{L(\omega)}}\right).$$

To generate confidence intervals for estimates based on heterogeneous attrition, we require a slightly more complicated bootstrap sampling method than for the homogeneous case above, as now we also need to account for uncertainty in our estimates of attrition, which in turn translates to uncertainty in $Q(\omega)$. As in the case of homogenous attrition, bootstrap samples $S_1, \ldots, S_k$ are generated by resampling $n = 162,328$ chains from the original set of n complete and incomplete chains. Each bootstrap sample is a simulated dataset that mimics what could have occurred had the entire small-world experiment been repeated. Next, we simulate vectors of regression coefficients $\vec{\beta}_1, \ldots, \vec{\beta}_k$ for the attrition model of Section 3, where each vector of parameters $\vec{\beta}_i$ is a complete set of coefficients for the model. In particular, $\vec{\beta}_i$ consists of parameter values for each group-level effect (e.g., "graduate school" and "20-29"), and allows one to estimate next-step continuance probabilities for given attribute data. These vectors of coefficients are generated by taking into account both the uncertainty in individual parameters, and the correlation between parameters[15]. In short, for a given chain ω, each coefficient vector $\vec{\beta}_i$ produces different estimates for the probability $r_{\omega_i \to \omega_{i+1}}$ of failing to forwarding a message across each link in the chain, and hence generates different estimates of the probability $Q_{\vec{\beta}_i}(\omega)$ that the chain ω is observed, reflecting uncertainty in the attrition model. The mean for each bootstrap sample S_i is computed by way of the observation probability $Q_{\vec{\beta}_i}(\cdot)$, as derived from the attrition model with parameter vector $\vec{\beta}_i$. As before, this procedure results in $k = 10,000$ different estimates of the mean chain length, and the reported confidence interval is the range of the middle 95% of these estimates.

Appealing as before to the estimators (3) and (4), we find the mean under this heterogeneous attrition model is 22 (95% CI: 4.5-57.5), and the median is 7 (95% CI: 6-8.5). Estimation of mean chain length is clearly sensitive to uncertainty in estimates of the probability of observation $Q(\omega)$, resulting in a wide confidence interval. The median, however, appears to be much more stable. The entire estimated cumulative distribution function (CDF) for chain length is shown in Figure 2. Since there are relatively few long chains, the variance in the estimated CDF grows with chain length.

Randomized Attrition. One possible objection to these estimates is that since individuals in completed chains by

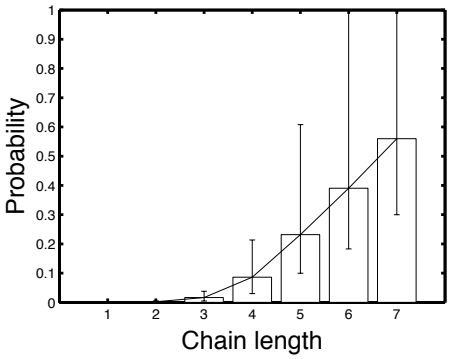

Figure 2: The estimated CDF of chain length under the heterogeneous attrition model. Error bars indicate 95% confidence intervals.

[14]The reason for the much lower attrition at the first step derives from the nature of the experiment, which relied on initial senders signing up at the web site to participate. Naturally, subsequent senders in a given chain did not volunteer, but were instead recruited by previous senders; thus were presumably not as motivated to complete the exercise as those who volunteered to initiate the chains.

[15]In a Bayesian interpretation, the coefficient vectors are generated from a "non-informative" prior, and represent configurations of parameters that are consistent with the data (see Gelman & Hill [7], Section 7.2).

definition must have passed on messages, we are at risk of over-fitting the attrition probabilities, mistakenly inferring that attributes of participants who happen to be in completed chains predict lower attrition rates. Indeed, the average estimated attrition for individuals in completed chains is 3% lower than the average estimated attrition for individuals in incomplete chains.

To address this possibility of over-fitting, we consider a second heterogeneous attrition model, in which attrition probabilities R_i are randomly generated from the distribution of estimated attrition rates shown in Figure 1. That is, we assume individuals have attrition rates that are randomly drawn from this estimated population distribution, and define the probability of observing a completed chain ω of length $L(\omega)$ to be

$$Q(\omega) = (1 - R_{\omega_0})(1 - R_{\omega_1} \cdots (1 - R_{\omega_{L(\omega)-1}}).$$

Each individual in each chain is independently assigned an attrition rate chosen from the population distribution. Again invoking the estimators (3) and (4), we find that under this modified attrition model, the "true" mean chain length is now 49 (95% CI: 37-63), where confidence intervals are generated in analog to the previous heterogeneous attrition case. Also as before, we estimate the median, which we find to be 6 (95% CI: 6-6).

All four estimates—one based on Travers and Milgram's data; and three based on the experiments described above under assumptions of (a) homogenous attrition; (b) empirically observed heterogeneous attrition; and (c) randomized heterogeneous attrition—are presented in Table 3, which prompts three observations. First, estimates of the median are extremely stable, with all four procedures predicting a "true" median chain length in the range of 6-7 steps; thus, for approximately half the population, the claim that "everyone is connected by six degrees of separation" appears to be valid not only in the topological sense, but also in the algorithmic sense. Second, in contrast with the median, estimates of the mean are extremely unstable, ranging from 11.8 to 49 steps, and with 95% confidence intervals ranging from 4.5 to 68. That estimated means can vary so wildly as a function of different assumptions about the heterogeneity of attrition is perhaps not surprising given the known sensitivity of chain completion to attrition, but it does suggest that any conclusions regarding the mean should be treated with caution. In spite of this, however, a third conclusion appears warranted—namely that at least some, and possibly many, chains are much, much longer than the median. Thus although the algorithmic small-world phenomenon appears to be satisfied for at least half the population, it also appears not to be satisfied for at least some fraction.

Table 3: Summary of "true" average algorithmic distance under homogenous and heterogeneous attrition models.

Model	Mean (95% CI)	Median (95% CI)
Homogeneous attrition (Travers/Milgram)	11.8 (8.5-15)	7 (6-7)
Homogeneous attrition (Dodds et. al.[3])	41.5 (20-68)	6 (6-6)
Heterogeneous attrition (Dodds et al.[3])	22 (4.5-57.5)	7 (6-8.5)
Randomized attrition (Dodds et al.[3])	49 (37-63)	6 (6-6)

6. CONCLUSION

In concluding, we return to the distinction posed in the Introduction between the topological and algorithmic versions of the small-world hypothesis. Empirically, the most striking contrast between the two is that whereas in our analysis, we find very large differences between the mean and the median path lengths, studies of topological path lengths typically find that the two measures are almost interchangeable: in the largest such study to date, for example, Leskovec and Horvitz [20] found that the mean shortest path length was 6.6 while the median was 7. Precisely why topological and algorithmic path lengths differ in this manner is ultimately unclear; however, it probably derives from the observation that the number of steps in a search chain depends not only on the actual (i.e. topological) distance between the source and target, but also on the search strategies of the intermediaries. Two individuals, in other words, may be the same distance from a given target, but if one has a better search strategy, the resulting chain will be shorter. Moreover, if one sender merely perceives the task to be easier, the two chains may experience different outcomes—for example, one may terminate while the other continues—thus leading to the appearance of a difference in difficulty when in fact the "real" difficulty is the same. Given only a set of complete and incomplete chains, therefore, it is arguably impossible to determine how much of the observed differences in chain length arise from (a) differences in topological distance, (b) differences in search strategies, and (c) differences in beliefs and motivations. In simulation exercises, these ambiguities can be eliminated by assuming a search rule [13, 37], but of course there is no guarantee that these rules resemble those being used by the participants in real-world experiments; thus simulations alone cannot resolve the problem either.

In principle, the problem might be resolved with an alternative experimental design, in which one could sample pairs of individuals according to their known topological distances, and also to represent a wide range of individual-level attributes. One could then run a large number of small-world type experiments, where participants would be drawn from the known sample, and would be required to pass messages exclusively through the known network. By comparing the lengths of completed chains with known topological shortest paths, one could place a measure of the "true" difficulty of a search between pairs with given attributes. And by varying incentives to participants as well as performing follow-up surveys, search difficulty could be further parsed in terms of differences in search strategies, perceived difficulty, and apathy. Although non-trivial to implement, the recent explosion of large social networking sites like Facebook, as well as email [18] and instant messaging networks [20], at least render such a design feasible. The results of such a study, moreover, would not only help to settle a decades-long debate over the effective connectivity of the social world, but would also shed new light on the relation between individual social capital, topological network structure, and the algorithmic properties of social search.

Acknowledgments

The authors wish to acknowledge helpful conversations with P. Bearman and A. Gelman. Research supported in part by NSF grants SES-0094162 and SES-0339023, and by the James S. McDonnell Foundation.

7. REFERENCES

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27:187–203, 2005.

[2] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman. Search in power-law networks. *Physical Review E*, 64, 2001.

[3] J. Best, B. Krueger, and C. Smith. An assessment of the generalizability of internet surveys. *Social Science Computer Review*, 19:131–145, 2001.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. W. 2000. Graph structure in the web. *Computer Networks*, pages 309–320, 2000.

[5] J. S. Coleman. Social capital in the creation of human capital. *The American Journal of Sociology*, 94:S95–S120, 1988.

[6] P. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, 2003.

[7] A. Gelman and J. Hill. *Data Analysis Using Regression and Multileve/Hierarchical Models*. Cambridge University Press, 2007.

[8] S. Golder, D. Wilkinson, and B. A. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *3rd International Conference on Communities and Technologies*, 2007.

[9] M. Granovetter. *Getting a Job: A Study of Contacts and Careers*. Harvard University Press, 2nd edition, 1995.

[10] J. Guare. *Six Degrees of Separation*. 1990.

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

[12] H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[13] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *32nd ACM Symposium on Theory of Computing*, 1999.

[14] J. Kleinberg. Navigation in a small world. *Nature*, 406, 2000.

[15] J. Kleinfeld. Could it be a big world after all? *Society*, 39:61–66, 2002.

[16] B. Kogut and G. Walker. The small world of germany and the durability of national networks. *American Sociological Review*, 66:317–335, 2001.

[17] C. Korte and S. Milgram. Acquaintance networks between racial groups: Application of the small world method. *Journal of Personality and Social Psychology*, 15:101–108, 1970.

[18] G. Kossinets and D. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

[19] N. H. Lee. *The Search for an Abortionist*. The University of Chicago Press, 1969.

[20] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *17th International World Wide Web Conference*, 2008.

[21] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA*, 102:11623–11628, 2005.

[22] N. Lin, P. Dayton, and P. Greenwald. The urban communication network and social stratification: A "small world experiment". In B. D. Ruben, editor, *Communication Yearbook*, pages 107–119. Transaction Books, 1978.

[23] C. Lundberg. Patterns of acquaintanceship in society and complex organization: A comparative study of the small world problem. *Pacific Sociological Review*, 18:206–222, 1975.

[24] J. W. Meshel. One phone call away: Secrets of a master networker, 2005. Portfolio.

[25] S. Milgram. The small world problem. *Psychology Today*, 1:60–67, 1967.

[26] A. E. Motter, T. Nishikawa, and Y. C. Lai. Large-scale structural organization of social networks|. *Physical Review E*, 68, 2003.

[27] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98:404–409, 2001.

[28] O. Pierson. *The Unwritten Rules of Highly Effective Job Search*. McGraw-Hill, 2006.

[29] A. Portes. Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology*, 24:1–24, 1998.

[30] D. Rezac. *Work the Pond!* Berkley Publishing Group, 2005.

[31] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the indian railway network. *Physical Review E*, 67, 2003.

[32] R. Shotland. *University Communication Networks: The Small World Method*. Wiley, 1976.

[33] O. Simsek and D. Jensen. Decentralized search in networks using homophily and degree disparity. In *Proceedings of the Ninenteenth International Joint Conference on Artificial Intelligence*, 2005.

[34] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[35] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 268:1803–1810, 2001.

[36] D. J. Watts. *Six Degrees: The Science of A Connected Age*. W. W. Norton, 2003.

[37] D. J. Watts, P. S. Dodds, and M. E. J. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.

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

[39] H. C. White. Search parameters for small world problem. *Social Forces*, 49:259–264, 1970.