An Experimental Study of Search in Global Social Networks

Peter Sheridan Dodds,[1] Roby Muhamad,[2] Duncan J. Watts[1,2]*

We report on a global social-search experiment in which more than 60,000 e-mail users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. We find that successful social search is conducted primarily through intermediate to weak strength ties, does not require highly connected "hubs" to succeed, and, in contrast to unsuccessful social search, disproportionately relies on professional relationships. By accounting for the attrition of message chains, we estimate that social searches can reach their targets in a median of five to seven steps, depending on the separation of source and target, although small variations in chain lengths and participation rates generate large differences in target reachability. We conclude that although global social networks are, in principle, searchable, actual success depends sensitively on individual incentives.

It has become commonplace to assert that any individual in the world can reach any other individual through a short chain of social ties (*1*, *2*). Early experimental work by Travers and Milgram (*3*) suggested that the average length of such chains is roughly six, and recent theoretical (*4*) and empirical (*4–9*) work has generalized the claim to a wide range of nonsocial networks. However, much about this "small world" hypothesis is poorly understood and empirically unsubstantiated. In particular, individuals in real social networks have only limited, local information about the global social network and, therefore, finding short paths represents a nontrivial search effort (*10–12*). Moreover, and contrary to accepted wisdom, experimental evidence for short global chain lengths is extremely limited (*13–15*). For example, Travers and Milgram report 96 message chains (of which 18 were completed) initiated by randomly selected individuals from a city other than the target's (*3*). Almost all other empirical studies of large-scale networks (*4–9*, *16–19*) have focused either on nonsocial networks or on crude proxies of social interaction such as scientific collaboration, and studies specific to e-mail networks have so far been limited to within single institutions (*20*).

We have addressed these issues by conducting a global, Internet-based social search experiment (*21*). Participants registered online (http://smallworld.sociology.columbia.edu) and were randomly allocated one of 18 target persons from 13 countries (table S1). Targets included a professor at an Ivy League university, an archival inspector in Estonia, a technology consultant in India, a policeman in Australia, and a veterinarian in the Norwegian army. Participants were informed that their task was to help relay a message to their allocated target by passing the message to a social acquaintance whom they considered "closer" than themselves to the target. Of the 98,847 individuals who registered, about 25% provided their personal information and initiated message chains. Because subsequent senders were effectively recruited by their own acquaintances, the participation rate after the first step increased to an average of 37%. Including initial and subsequent senders, data were recorded on 61,168 individuals from 166 countries, constituting 24,163 distinct message chains (table S2). More than half of all participants resided in North America and were middle class, professional, college educated, and Christian, reflecting commonly held notions of the Internet-using population (*22*).

In addition to providing his or her chosen contact's name and e-mail address, each sender was also required to describe how he or she had come to know the person, along with the type and strength of the resulting relationship. Table 1 lists the frequencies with which different types of relationships—classified by type, origin, and strength—were invoked by our population of 61,168 active senders. When passing messages, senders typically used friendships in preference to business or family ties; however, almost half of these friendships were formed through either work or school affiliations. Furthermore, successful chains in comparison with incomplete chains disproportionately involved professional ties (33.9 versus 13.2%) rather than friendship and familial relationships (59.8 versus 83.4%) (table S3). Successful chains were also more likely to entail links that originated through work or higher education (65.1 versus 39.6%) (table S4). Men passed messages more frequently to other men (57%), and women to other women (61%), and this tendency to pass to a same-sex contact was strengthened by about 3% if the target was the same gender as the sender and similarly weakened in the opposite case. Individuals in both successful and unsuccessful chains typically used ties to acquaintances they deemed to be "fairly close." However, in successful chains "casual" and "not close" ties were chosen 15.7 and 5.9% more frequently than in unsuccessful chains (table S5), thus adding support, and some resolution, to the longstanding claim that "weak" ties are disproportionately responsible for social connectivity (*23*).

Senders were also asked why they considered their nominated acquaintance a suitable recipient (Table 2). Two reasons—geographical proximity of the acquaintance to the target and similarity of occupation—accounted for at least half of all choices, in general agreement with previous findings (*24*, *25*). Geography clearly dominated the early stages of a chain (when senders were geographically distant) but after the third step was cited less frequently than other characteristics, of which occupation was the most often cited. In contrast with previous claims (*3*, *12*), the presence of highly connected individuals (hubs) appears to have limited relevance to the kind of social search embodied by our experiment (social search with large associated costs/rewards or otherwise modified individual incentives may behave differently). Participants relatively rarely nominated an acquaintance primarily because he or she had many friends (Table 2, "Friends"), and individuals in successful

Table 1. Type, origin, and strength of social ties used to direct messages. Only the top five categories in the first two columns have been listed. The most useful category of social tie is medium-strength friendships that originate in the workplace.

Type of relationship	%	Origin of relationship	%	Strength of relationship	%
Friend	67	Work	25	Extremely close	18
Relatives	10	School/university	22	Very close	23
Co-worker	9	Family/relation	19	Fairly close	33
Sibling	5	Mutual friend	9	Casual	22
Significant other	3	Internet	6	Not close	4

[1]Institute for Social and Economic Research and Policy, Columbia University, 420 West 118th Street, New York, NY 10027, USA. [2]Department of Sociology, Columbia University, 1180 Amsterdam Avenue, New York, NY 10027, USA.

*To whom correspondence should be addressed. E-mail: djw24@columbia.edu

chains were far less likely than those in incomplete chains to send messages to hubs (1.6 versus 8.2%) (table S6). We also find no evidence of message "funneling" (*3*, *9*) through a single acquaintance of the target: At most 5% of messages passed through a single acquaintance of any target, and 95% of all chains were completed through individuals who delivered at most three messages. We conclude that social search appears to be largely an egalitarian exercise, not one whose success depends on a small minority of exceptional individuals.

Although the average participation rate (about 37%) was high relative to those reported in most e-mail–based surveys (*26*), the compounding effects of attrition over multiple links resulted in exponential attenuation of chains as a function of their length and therefore an extremely low chain completion rate (384 of 24,163 chains reached their targets). Chains may have terminated (i) randomly, because of individual apathy or disinclination to participate (*3*, *27*); (ii) preferentially at longer chain lengths, corresponding to the claim that chains get "lost" or are otherwise unable to reach their targets (*13*); or (iii) preferentially at short chain lengths, because, for example, individuals nearer the target are more likely to continue the chain.

Our findings support the random-failure hypothesis for two reasons. First, with the exception of the first step (which is special because senders register rather than receive a message from an acquaintance), the attrition rate remains almost constant for all chain lengths at which we have a sufficiently large N; hence small confidence intervals (Fig. 1A). Second, senders who did not forward their messages after one week were asked why they had not participated. Less than 0.3% of those contacted claimed that they could not think of an appropriate recipient, suggesting that lack of interest or incentive, not difficulty, was the main reason for chain termination.

To estimate the reachability of all targets, we first aggregate the 384 completed chains across targets (Fig. 1B), finding the average chain length to be $<L> = 4.05$. However, this number is misleading because it represents an average only over the completed chains, and shorter chains are more likely to be completed. An "ideal" frequency distribution of chain lengths $n'(L)$ (i.e., the chain lengths that would be observed in the hypothetical limit of zero attrition) may be estimated by accounting for observed attrition as follows: $n'(L) = n(L)/\Pi_{i=0}^{L-1}(1-r_i)$ (Fig. 1C, bars), where $n(L)$ is the observed number of chains completed after L steps (Fig. 1B) and r_L is the maximum-likelihood attrition rate from step L to step $L + 1$ (Fig. 1A, circles). Using the observed values of r_L, we have reconstructed the most likely ideal distribution $n'(L)$ (Fig. 1C, bars) under our assumption of random attrition. Because the tail of the distribution is poorly specified (owing to the small number of observed chains at large, L), we measure its median L_* rather than its mean. We find $L_* = 7$, and this can be thought of as the typical ideal chain length for a hypothetical average individual. By repeating the above procedure for chains that started and ended in the same country ($L_* = 5$) or in different countries ($L_* = 7$), we can disentangle to some extent the different underlying distributions of chains, yielding an estimated range of typical chain lengths $5 \leq L_* \leq 7$, depending on the geographical separation of source and target.

Although the range of L_* and the variation in attrition rates across targets do not appear great, the compounding effects of attrition over the length of a message chain can nevertheless generate large differences in message completion rates. For example, a decrease of 15% in attrition rates, when compounded over the same ideal distribution with $L_* = 6$, can generate an 800% increase in completion rate. The same attrition rates [e.g., $r_0 = 0.75$, $r_L = 0.63$ ($L \geq 1$)], when applied over chains with $L_* = 5$ and 7, respectively, can lead to completion rates that vary by up to a factor of three.

Taken together, this evidence suggests a mixed picture of search in global social networks. On the one hand, all targets may in fact be reachable from random initial senders in only a few steps, with surprisingly little variation across targets in different countries and professions. On the other hand, small differences in either participation rates or the underlying chain lengths can have a dramatic impact on the apparent reachability of different targets. Target 5 (a professor at a prominent U.S. university) stands out in this respect. Because 85% of senders were college educated and more than half were American, participants may have anticipated little difficulty in reaching him, thus accounting for his chains' attrition rate (54%) being much lower than that of any other target (60 to 68%). Target 5 received a notable 44% of all completed chains, yet this result is consistent with his "true" reachability being little different from that of other targets; his allocated senders may simply have been more confident of success.

Our results therefore suggest that if individuals searching for remote targets do not have sufficient incentives to proceed, the small-world hypothesis will not appear to hold (*13*), but that even a slight increase in incentives can render social searches success-

Table 2. Reason for choosing next recipient. All quantities are percentages. Location, recipient is geographically closer; Travel, recipient has traveled to target's region; Family, recipient's family originates from target's region; Work, recipient has occupation similar to target; Education, recipient has similar educational background to target; Friends, recipient has many friends; Cooperative, recipient is considered likely to continue the chain; Other, includes recipient as the target.

L	N	Location	Travel	Family	Work	Education	Friends	Cooperative	Other
1	19,718	33	16	11	16	3	9	9	3
2	7,414	40	11	11	19	4	6	7	2
3	2,834	37	8	10	26	6	6	4	3
4	1,014	33	6	7	31	8	5	5	5
5	349	27	3	6	38	12	6	3	5
6	117	21	3	5	42	15	4	5	5
7	37	16	3	3	46	19	8	5	0

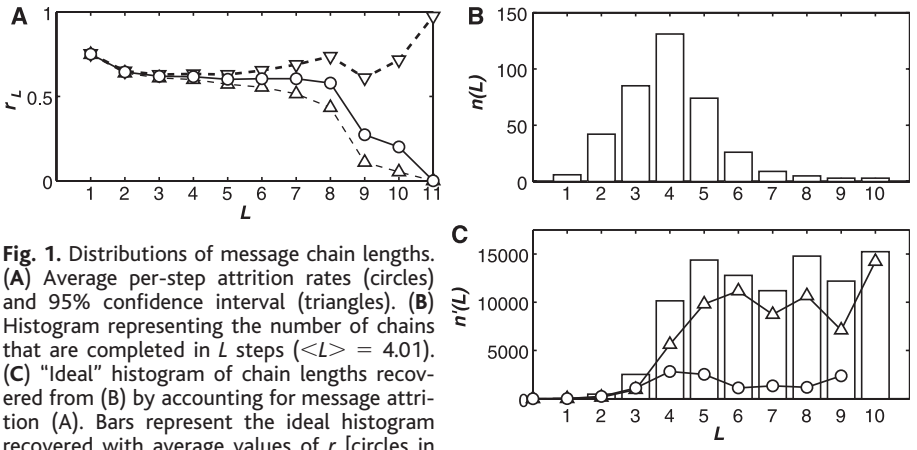

Fig. 1. Distributions of message chain lengths. (**A**) Average per-step attrition rates (circles) and 95% confidence interval (triangles). (**B**) Histogram representing the number of chains that are completed in L steps ($<L> = 4.01$). (**C**) "Ideal" histogram of chain lengths recovered from (B) by accounting for message attrition (A). Bars represent the ideal histogram recovered with average values of r [circles in (A)] for the histogram in (B); lines represent a decomposition of the complete data into chains that start in the same country as the target (circles) and those that start in a different country (triangles).

ful under broad conditions. More generally, the experimental approach adopted here suggests that empirically observed network structure can only be meaningfully interpreted in light of the actions, strategies, and even perceptions of the individuals embedded in the network: Network structure alone is not everything.

References and Notes

1. I. de Sola Pool, M. Kochen, *Soc. Networks* **1**, 1 (1978).
2. S. H. Strogatz, *Nature* **410**, 268 (2001).
3. J. Travers, S. Milgram, *Sociometry* **32**, 425 (1969).
4. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
5. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130 (1999).
6. L. A. Adamic, in *Lecture Notes in Computer Science 1696*, S. Abiteboul, A. Vercoustre, Eds. (Springer, Heidelberg, 1999), pp. 443–454.
7. L. A. N. Amaral, A. Scala, M. Barthelemy, H. E. Stanley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11149 (2000).
8. A. Wagner, D. Fell, *Proc. R. Soc. London, B* **268**, 1803 (2001).
9. M. E. J. Newman, *Phys. Rev. E* **64**, 016131 (2001).
10. J. Kleinberg, *Nature* **406**, 845 (2000).
11. D. J. Watts, P. S. Dodds, M. E. J. Newman, *Science* **296**, 1302 (2002).
12. L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, *Phys. Rev. E* **64**, 046135 (2001).
13. J. S. Kleinfeld, *Society* **39**, 61 (2002).
14. C. Korte, S. Milgram, *J. Pers. Soc. Psychol.* **15**, 101 (1970).
15. N. Lin, P. Dayton, P. Greenwald, in *Communication Yearbook: Vol. 1*, B. D. Ruben, Ed. (Transaction Books, New Brunswick, NJ, 1977), pp. 107–119.
16. A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).
17. M. Faloutsos, P. Faloutsos, C. Faloutsos, *Comp. Comm. Rev.* **29**, 251 (1999).
18. L. A. Adamic, B. A. Huberman, *Science* **287**, 2115a (2000).
19. H. Jeong, B. Tombor, R. Albert, Z. N. Oltavi, A.-L. Barabási, *Nature* **407**, 651 (2000).
20. H. Ebel, L.-I. Mielsch, S. Bornholdt, *Phys. Rev. E* **66**, 035103 (2002).
21. Materials and methods are available as supporting material on *Science* Online.
22. W. Chen, J. Boase, B. Wellman, in *The Internet in Everyday Life*, B. Wellman, C. Haythornthwaite, Eds. (Blackwell, Oxford, 2002), pp. 74–113.
23. M. S. Granovetter, *Am. J. Sociol.* **78**, 1360 (1973).
24. P. D. Killworth, H. R. Bernard, *Soc. Networks* **1**, 159 (1978).
25. H. R. Bernard, P. D. Killworth, M. J. Evans, C. McCarty, G. A. Shelly, *Ethnology* **27**, 155 (1988).
26. K. Sheehan, *J. Comput. Mediated Commun.* **6**(2). Available online at www.ascusc.org/jcmc/vol6/issue2/sheehan.html (2001).
27. H. C. White, *Soc. Forces* **49**(2), 259 (1970).

28. This research was supported in part by the National Science Foundation, Intel Corporation, and Office of Naval Research.

Supporting Online Material

www.sciencemag.org/cgi/content/full/301/5634/827/DC1
Methods
Tables S1 to S6

2 December 2002; accepted 23 May 2003

Phylogenetics and the Cohesion of Bacterial Genomes

Vincent Daubin,[1] Nancy A. Moran,[2] Howard Ochman[1]*

Gene acquisition is an ongoing process in many bacterial genomes, contributing to adaptation and ecological diversification. Lateral gene transfer is considered the primary explanation for discordance among gene phylogenies and as an obstacle to reconstructing the tree of life. We measured the extent of phylogenetic conflict and alien-gene acquisition within quartets of sequenced genomes. Although comparisons of complete gene inventories indicate appreciable gain and loss of genes, orthologs available for phylogenetic reconstruction are consistent with a single tree.

In all but the most reduced bacterial genomes, there is a substantial fraction of genes whose distributions and compositional features indicate that they originated by lateral gene transfer (LGT) (*1*). There is also clear evidence of LGT between distantly related organisms based on phylogenetic studies involving large taxonomic samples (*2*). Given these findings, incompatibility of phylogenies within and among bacterial phyla based on different genes has routinely been ascribed to LGT (*3–10*). However, building molecular phylogenies for distantly related species is often a difficult task, and choice of phylogenetic methods, genes, or taxa can yield different results. For example, there is still no consensus on the monophyly of rodents (*11*, *12*) or the branching order of amniotes (*13*, *14*), and these groups are young compared to bacterial phyla. In addition, distinguishing between orthologous genes (sequences that trace their divergence to the splitting of organismal lineages) and paralogous (duplicated) genes becomes increasingly difficult when considering more distantly related taxa.

The effects of LGT have been extended from the deepest to the shallowest levels of bacterial relationships. Indeed, the similarities in gene sequence and gene content that define widely accepted bacterial taxa have been proposed to reflect boundaries to gene transfer, rather than vertical transmission and common organismal ancestry (*10*). Thus, LGT may overwhelm attempts to reconstruct the relationships among bacterial taxa. The claim that the history of bacteria might be more faithfully depicted as a net than as a tree (*7*) relies upon the postulate that the substantial incidence of acquired DNA within genomes is the basis for findings of phylogenetic incongruence among genes. However, the genes detected as recently transferred are, by and large, different from those used to build species phylogenies. The former are disproportionately A+T-rich, have restricted phylogenetic distributions, and usually encode accessory functions. In contrast, species phylogenies are based on genes with wide taxonomic distributions and having key roles in cellular processes. However, such differences are often ignored when considering the impact of LGT on bacterial relationships. Although the incidence of recently acquired DNA in bacterial genomes is the most direct indication of extensive LGT among species (*1*), the question of whether the incongruence in gene phylogenies is linked to the amount of new DNA in a genome has not been addressed.

To investigate the relation between DNA acquisition and phylogenetic incongruence, we selected quartets of related, sequenced genomes whose phylogenetic relationships, based on small subunit ribosomal RNA (SSU rRNA) sequences, display the branching topology shown in Fig. 1. For each quartet, we inferred both the number of recently acquired and lost genes (based on their phylogenetic distributions) and the proportion of ortholog phylogenies supporting lateral transfers. We applied a conservative method for identifying orthologs by including only those genes having a single significant match per genome, thus minimizing the risks of including hidden paralogs descending from within-genome duplication events. This contrasts with the commonly used "reciprocal best-hit method" (*15*) to infer orthology, which can yield misleading results (*16*), especially when paralogs experience different evolutionary rates. We retained all quartets of species for which >25% of the genes from the smallest genome were recovered as orthologs. We then tested which of the three possible trees was significantly supported for each ortholog family, using the Shimodaira-Hasegawa (SH) (*17*) test implemented in Tree-puzzle 5.1 (*18*) at the 5% level of significance (*19*). This method tests if an alignment significantly supports a tree by estimating the confidence limits of the likelihood estimates of the topologies.

[1]Department of Biochemistry and Molecular Biophysics, [2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA.

*To whom correspondence should be addressed. E-mail: hochman@email.arizona.edu