Geospatial Structure of a Planetary-Scale Social Network

Jure Leskovec and Eric Horvitz

***Abstract*—Little is known about geographic properties of large-scale social networks. In this paper, we examine the geospatial attributes of a planetary-scale social network of 240 million people and 1.3 billion edges. We study the interplay among topological, geographical, and algorithmically generated paths connecting pairs of nodes in a social network. Starting in the realm of cyberspace, we find that topologically shortest paths of average length of 6.6 exist between pairs of nodes in the network and that the average degree of separation among nodes is robust to removal of hub nodes. Moving to the realm of locations and distances in geographic space, we find that topologically shortest paths in the social graph grow with increasing geographic distance between path's endpoint nodes. We discover that shortest topological paths are geographically inefficient, but that geography provides an important cue for local algorithmic policies for navigating between source and target nodes. Local algorithmic strategies for navigating the larger network structure in the absence of global navigation procedures have varying success. At the early stages of the navigation, navigating to a hub node helps, while in the middle stage, geography provides the most important clue. While local algorithms for navigating have trouble reaching the target node, they are successful in reaching nodes that are geographically close to the target. Taken together, our results demonstrate a complex interplay between topological and geographical properties of social networks and explain the success of local strategies for navigating such networks.**

***Index Terms*—Decentralized search, network navigation, networks, small-world experiment, social search.**

I. Introduction

UNTIL recently, it was impossible to directly analyze the structure of the global human social network. Nevertheless, it has been asserted commonly that any individual in the world can reach any other individual through a chain of only a few social connections [1], [2]. Experimental evidence for short chains in the global human social network has been limited [3]–[5]. Subject nonparticipation and noncompletion of chains have posed significant problems [3], [6]. Travers and Milgram reported their results on six degrees of separation among people based on 64 completed chains [7], and most recently the email-based experiment of more than 60 000 participants, recorded only 384 completed chains [8]. The advent of widely used Web-based communication platforms has provided electronic forms of large-scale social and communication networks that span large regions of the Earth—and that can serve as laboratories for research on the structure of global social networks. Online human experiments [8]–[11] have probed the structure of the global human social network and suggested that the average number of steps of such chains is roughly six, and further theoretical [12] and empirical [13]–[16] studies have further validated the claim to a wide range of networks. In summary, studies of large-scale networks have stimulated new questions and avenues of research about the "small-world" hypothesis and the broader structure of the global social network [6], [10], [15].

Here, we follow on the above line of work and explore questions about the relationships between graph-theoretic and geographic distances in social networks. We explore three different notions of path length between pairs of nodes in the social network: *topologically shortest path* that traverses the minimum number of links to navigate from the source to the destination node, *geographic length of a path*, which is simply the sum of geographical distances of edges along the nodes of the path, and *algorithmic path*, which is a path identified by an algorithmic policy that only relies on local knowledge about the network.

We find that the average topologically shortest path length among pairs of nodes grows with increasing geographic distance between the nodes. Topologically shortest paths are generally quite inefficient geographically as they traverse long geographic distances. We focus on the value of using geographic cues in local algorithmic strategies for navigating from one node to another. Individuals in social networks have only limited, local views of the network yet are often able, without global knowledge of the network, to navigate the network in a decentralized way to find short chains of connections [17]–[25]. We experiment with local navigation policies and show the value of harnessing geographic information in local navigation actions. We find that, at the early stages of the navigation, navigating to a high-degree hub node helps, while in the middle stage of the navigation, geography provides the best cue. Generally, early steps of navigation are "easy" as there are many nodes leading topologically closer to the target, however approaching the target is hard, as in later stages less than 5% of node's neighbors lead topologically toward the target.

Manuscript received March 18, 2014; revised November 29, 2014; accepted November 29, 2014. Date of publication January 07, 2015; date of current version January 22, 2015. The work of J. Leskovec was supported in part by NSF IIS-1016909, by CNS-1010921, by IIS-1149837, by DARPA SMISC, by DARPA GRAPHS, by ARL AHPCRC, by the Okawa Foundation, and in part by the Alfred P. Sloan Fellowship.

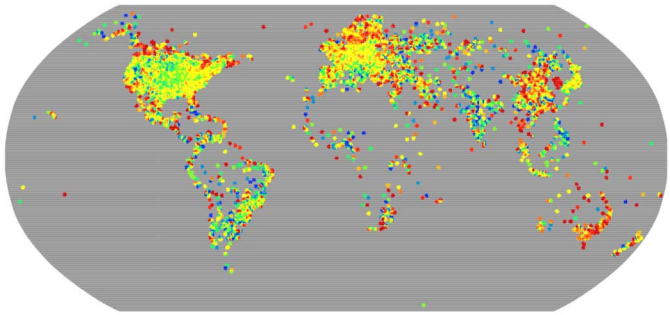
J. Leskovec is with the Department of Computer Science, Stanford Univeristy, Stanford, CA 94305 USA (e-mail: jure@cs.stanford.edu).

E. Horvitz is with Microsoft Research, Redmond, WA 98052 USA (e-mail: horvitz@microsoft.com).

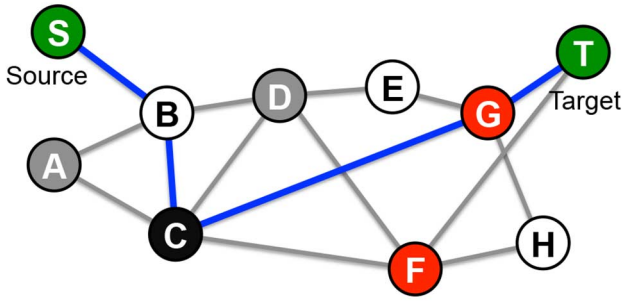
Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSS.2014.2377789

2329-924X © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

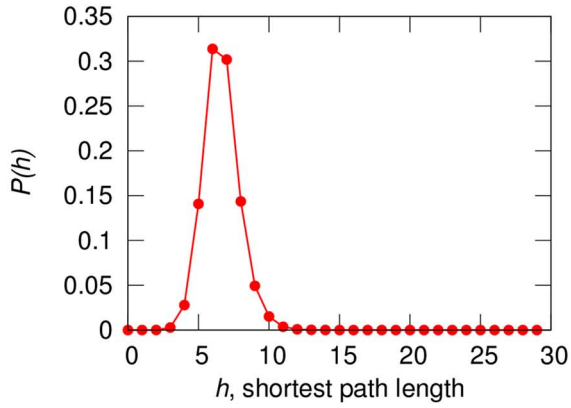


(a)

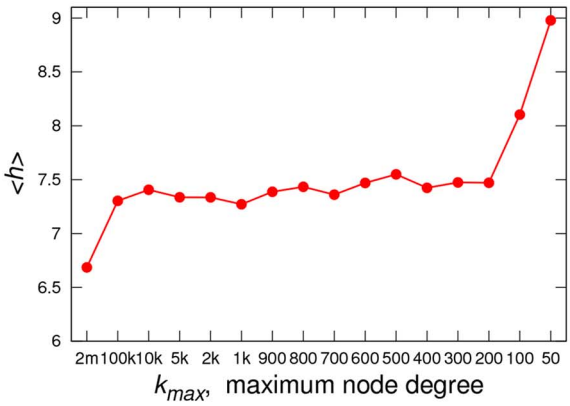


(b)

Fig. 1. Messenger network. (a) Geographic locations of 240 million Messenger users. Color (blue to red) of dots represents the logarithm of the number of users at a given location. (b) Messenger users are represented as nodes of the network and users are connected via undirected edges if they exchanged at least one message in the observation period of 1 month.



(a)



(b)

Fig. 2. Degree of separation on the Messenger network. (a) Distribution of the topologically shortest path lengths, measured as the minimum number of links h traversed by the path between nodes S and T in the largest connected component of the Messenger network which contains 180 million nodes and 1.342 billion edges. The average path length is $\langle h \rangle = 6.6$. The 90th percentile of the distribution is 7.8, 48% of nodes can be reached within six steps and 78% within seven steps. (b) Average topologically shortest path length $\langle h \rangle$ when the nodes of degree higher than $k_{\max}$ are removed from the network. 1.3% of nodes have degree $k > 100$ and they account for 25% of all edges. Shortest path lengths remain stable even if the high-degree hub nodes are removed from the network.

II. Seven-Degrees of Messaging

We conducted a computational experiment of a global communication network comprised 240 million users [Fig. 1(a)] who exchanged 255 billion messages during the 1 month observation period (refer to Appendix for further details) [10]. The users of the network cover a large portion of the populated areas of the earth and represent a nontrivial fraction of world's population. The communication network represents all people who have communicated with one another on the Microsoft Instant Messenger communication network. We represent people as nodes and focus only on the active communication ties by connecting via undirected edges pairs of users who exchanged at least one message during the observation period. We limit our study to the largest connected component of the network which contains 180 million nodes and 1.342 billion undirected edges [Fig. 1(b)]. We also determine the geographical location of each user by using the reverse geo-lookup based on the user's IP address (Appendix) [10]. We note that, in contrast to recent studies of large social networks like Facebook [11] and Twitter [26], where users have hundreds or even thousands of weak "friends," our study considers a network of strong active ties that are actually being used for communication.

The experimental setup enables a thorough examination of the small-world hypothesis. We studied the distribution of lengths of the topologically shortest paths [Fig. 2(a)]. We randomly sampled pairs of nodes and calculated the minimum number of links separating a given source–destination node pair. We found that the distribution of path lengths reaches the mode at 6 and has a median length of 7 [10]. A random pair of nodes in the Messenger network is separated by 6.6 degrees on average, which is close to the length measured by Travers and Milgram [7]. Via the lens provided on the world of social relationships by Messenger, we find that there are about "seven degrees of separation" among people. Long paths—pairs of nodes with high degree of separation—also exist in the network. We found pairs of nodes that are separated by shortest paths of length of up to 29 steps [10].

To put our findings in perspective, recent studies of Facebook [15] and Twitter [26] social networks found that degree of separation within these networks is around 4. The difference likely arises because of the way that the networks were constructed. We study a network of "strong" ties actively used for communication (and not, e.g., a network of address book contacts). In contrast, studies on Twitter and Facebook use all the ties with the majority of them being inactive or "weak." For example, the average degree in our network is 15, while

average degrees in Twitter and Facebook are in the hundreds, with some users having thousands of contacts, all referred to as "friends." Comparing our findings to those of Travers and Milgram [7] is intriguing. However, Milgram studied the lengths of the paths found by humans, while we compute topologically shortest paths. Moreover, one also has to be cautious interpreting Milgram's results due to attrition in participation, which means that in his experiment longer paths have a higher probability of failing [6].

A. Fragility of Topologically Shortest Paths

While topologically short paths exist in the network, they can be "fragile" in the sense that removing a small number of edges (or nodes) could lead to large changes in the connectivity of the network structure [27]. In particular, for scale-free networks, it has been shown that removing a small number of nodes or edges can lead to large changes in the network connectivity structure [27].

We investigate the fragility of topologically shortest paths, and the effect of high-degree hub nodes by examining how the average topologically shortest path length $\langle h \rangle$ changes as a function of the maximum degree $k_{\max}$ of any node in the network [27]. As soon as the highest degree node is removed, the average path length jumps from 6.6 to 7.4 and remains stable up to the point when all nodes of degree greater than 100 are removed. Overall, there are 2.3 million (1.3%) nodes of degree $k > 100$ and they account for 25% of all edges in the network. This indicates that the human social network represented by the Messenger data is relatively robust and maintains connectivity even if high-degree nodes (25% of all edges) are removed, which is somewhat surprising given the findings on scale-free networks [27]. In contrast, our experiment gives evidence for the hypothesis that highly connected hubs are not required for short chains to exist in the network [24], [25].

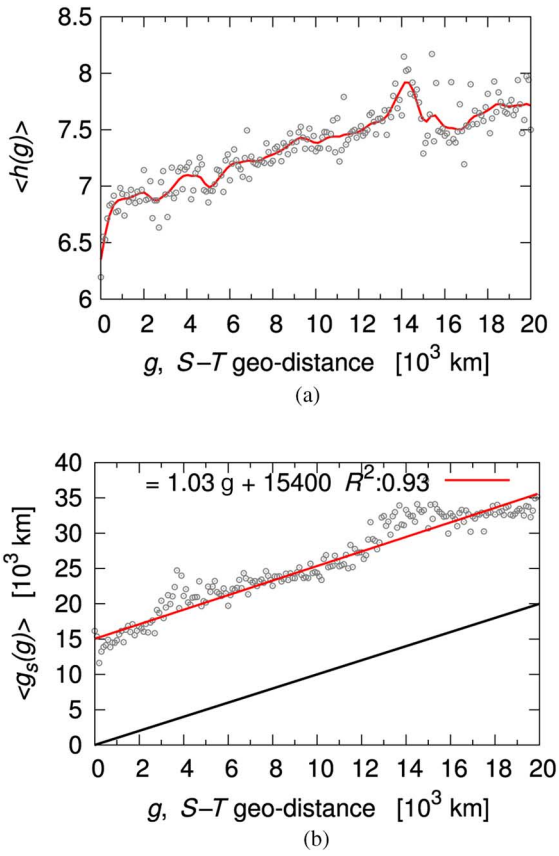


Fig. 3. Geographical properties of topologically shortest paths. (a) Average topologically shortest path length $\langle h(g) \rangle$ between pairs of nodes $S{-}T$ that are at geographical distance g. (b) Average geographical distance $\langle g_s(g) \rangle$ of the topologically shortest path s among pairs of nodes $S{-}T$ that are at geographical distance g. Straight line $y = x$ illustrates the lower bound on the geo-distance of the shortest path. For nodes S and T that are more than 500 km apart we observe that the topologically shortest paths s tend to be 15 400 km longer than the geographical distance g between the path endpoints. In other words, $\langle g_s(g) \rangle \approx g + 15\,400$.

B. Geographical Properties of Topologically Shortest Paths

We now investigate the connection between geographic g and the graph-theoretic h notion of distance. More precisely, we distinguish between geographic and topological measures of distance as illustrated in Fig. 1(b).

1) *Length $h(S,T)$ of the topologically shortest path*: The minimal number of links we need to traverse to get from source node S to the target node T. For example, $h(S,T) = 4$ in Fig. 1(b).
2) *Geographic distance $g(U,V)$* between nodes U and V is the distance between geographic locations of nodes U and V measured along the surface of the Earth.
3) *Geographic path length g_p* of the path p is defined as the sum of the geographic distances of links $g(U,V)$ along the path p: $g_p = \sum_{(U,V)\in p} g(U,V)$, where $g(U,V)$ is the geographical distance between nodes U and V.
4) *Geographic length g_s of the topologically shortest path s* is the geographic length of path s, where s is the shortest (in a network sense) path between nodes S and T.

We compare the geographic distance $g(S,T)$ and the topologically shortest path distance $h(S,T)$ between the source node S and target node T. Fig. 1(b) illustrates the topologically shortest path from S to T with blue edges [$h(S,T) = 4$] and the geographic distance of the path as the sum of path edge lengths. When analyzing the geographic structure of shortest paths, it is often the case that multiple shortest paths of different geographic lengths exist between a pair of nodes. In such cases, we choose a random path among all topologically shortest paths. Thus, for each pair of nodes S and T, we obtain a single unique topologically shortest path and our figures then show the averages of the shortest paths over many random $S{-}T$ pairs.

We find [Fig. 3(a)] that, for nodes $S{-}T$ that are geographically close ($g < 1000$ km), the average topologically shortest path length $\langle h(g) \rangle$ is 6.5, while for nodes that are far apart ($g > 10\,000$ km) the path length $\langle h(g) \rangle$ increases to 7.6. Thus, we conclude that there exists a connection between geographical and topological distance between a pair of nodes. Geographically, closer nodes tend to have shorter topologically shortest paths. Increasing the distance between the nodes from 0 to 20 000 km on average increases the degree of separation for about 1.2 hops, from 6.4 to 7.6.

However, we find that topologically shortest paths are very inefficient geographically [Fig. 3(b)]. We measure the

geographical length g_s of the topologically shortest path s as the sum of geographical lengths of the links on the path s and find that the average geographical length $\langle g_s \rangle$ of the topologically shortest path is 15 400 km longer than the geographical distance between path endpoints *S*–*T*. Except for the nodes that are within a few kilometers from each other, it seems that topologically shortest paths show a constant penalty in the geographical distance that they traverse and the penalty seems to be independent of the geographical distance between the start and the end point of the path. Thus, if one wishes to travel from S to T in a minimal number of steps, then, on average, the distance that must be traveled is 15 400 km longer than the $S{-}T$ distance.

We find this result surprising as one would expect that the paths between nodes that are spatially very close to one another would be short topologically and geographically. In fact, we find this to be the case for $S{-}T$ pairs that are less than 500 km apart. However, for $S{-}T$ pairs above roughly 500 km apart, we observe a clear linear trend in the relation between the geographical distance between nodes $S{-}T$ and the geographic length taken by the topologically shortest path.

III. Geographic Navigation

We now explore the value of using geographic information in local algorithmic policies to find short paths between source and target nodes. Computer scientists can compute shortest paths among any two nodes in a graph using Dijkstra's algorithm. However, when humans navigate social networks and search for short chains of connections, they are limited to using local, decentralized policies that do not have access to global knowledge of the larger network structure [17], [25]. Studies [4], [8], [9], [28] have observed that geography tends to be one of the main cues people use in navigating the global social network. To gain insight into geographic navigation in networks, we next examine how topologically shortest paths geographically "navigate" through the network.

A. Geographical Anatomy of Topologically Shortest Paths

We compute topologically shortest paths between pairs of nodes and then examine paths' geographical properties. We discover that, even though topologically shortest paths in the network are geographically very long, they eventually zoom-in on the target node T.

Fig. 4(a) demonstrates that as the shortest path navigates topologically closer to T in the network, it also approaches T geographically. However, it is in the steps $i = 5$, 4, and 3 (counting backwards from target node T) when the path makes the largest geographical strides toward T. An average path starts geographically far away from T ($g_T = 8500$ km), dwells at approximately this distance for a few steps, and then in just three hops zooms-in on T from 8000 km away to 500 km from the target T [Fig. 4(a)].

To gain further insights into the structure of topologically shortest paths, we quantify two properties: The degree of nodes along the path [Fig. 4(b) circles] and the number of *topologically closer* neighbors [Fig. 4(b) squares]. We define node V to be topologically closer to the target T than node U if $h(V,T) < h(U,T)$. The intuition behind the definition is that

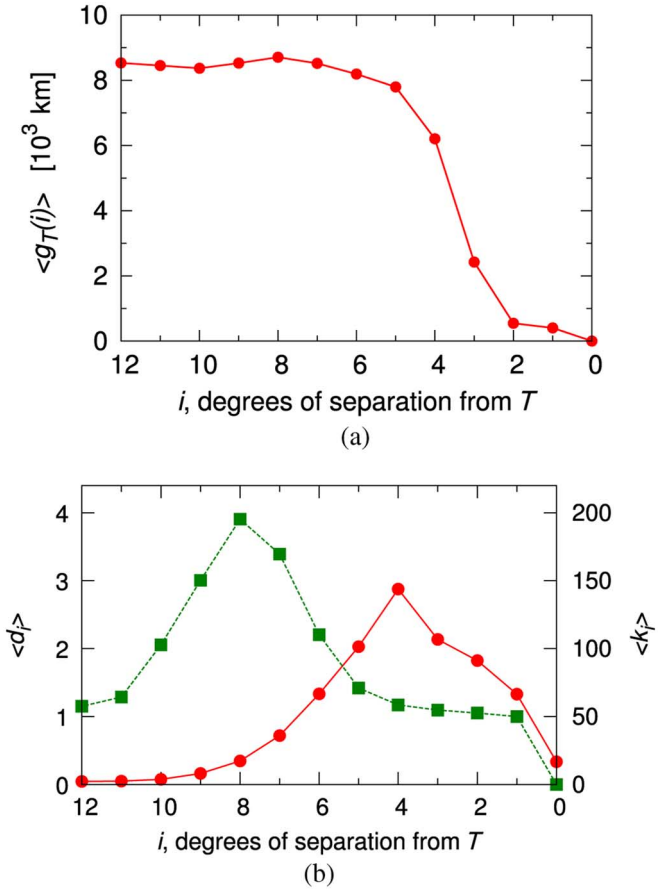


Fig. 4. Geographical properties of shortest paths. (a) Average geographical distance $\langle g_T(g) \rangle$ to target node T as the shortest path is i steps away from target T. The shortest path makes largest jumps geographically toward the target T at steps $i = 5$, 4, and 3 (counting backwards from the end T of the path). (b) Squares: Average number of topologically closer neighbors $\langle d_i \rangle$ of a node with i degrees of separation from T. Circles: Average degree $\langle k_i \rangle$ of a node on the shortest path toward T as a function of the degrees of separation i of the node from target T (shown on a different scale, drawn in the left side of the frame).

if we are at some node U then navigating to node V decreases the topological distance to target T and thus gets us closer to the target.

For example, in Fig. 1(b), nodes F and G are topologically closer to T than node C since $h(F,T), h(G,T) = 1 < h(C,T) = 2$. Nodes A and D are not closer to T than C since $h(A,T) = 3$, while $h(C,T) = 2$.

Examining the two quantities in Fig. 4(b), we find that topologically shortest paths navigate through the high-degree nodes at steps 5–3. We further find that this fast-paced geographical approach occurs because topologically shortest paths traverse geographically longest links at steps 5–3 [Fig. 4(a)], which indicates the core-periphery structure [29], [30] of the Messenger network, where the shortest path lurks at the periphery of the network so as to find a quick way through the high-degree network core to reach the target.

However, navigating the network core is hard. At every step, there are only between 1 and 4 possible nodes for the path to proceed to, so that it progresses topologically toward the target [Fig. 4(b) squares]. To navigate from a given node i, steps from T to a node topologically closer to T, the path can choose only

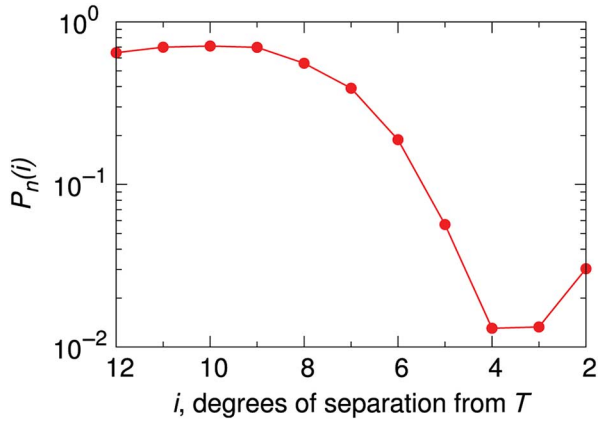


Fig. 5. Fraction of topologically closer neighbors. Average fraction of $P_n(i)$ of neighbors V of node U ($h(U,T)=i$) that reside on the shortest path ($h(V,T)=i-1$).

among 1.7 suitable nodes on the average. The average degree $\langle k_i \rangle$ of nodes on the topologically shortest paths peaks later ($i=4$) than the average number of neighbors $\langle d_i \rangle$ that lead closer to T ($i=8$).

The ratio $P_n(i)=\langle d_i \rangle/(\langle k_i \rangle - 1)$ indicates the probability that navigating to a random node would make a step topologically closer to the target node T (Fig. 5). When the path is very far from T ($i>8$), such navigation is easy; on average 70% of the neighbors of a node lead toward T. However, as the path gets closer to T ($i<6$), navigation becomes extremely difficult. When the path is ≤ 6 steps from T, less than 5% of nodes' neighbors lead toward T. On average, only 10% of node's neighbors lead topologically closer to T. This means that random navigation will successfully navigate toward the target with probability of 0.1 at every step. The overall probability of random navigation reaching the target node in the expected number of seven steps is thus extremely unlikely, $\approx 10^{-7}$.

B. Strategies for Network Navigation

Various hypotheses have been proposed about strategies that might be used by people to navigate the global social network in a local, decentralized manner [8], [28]. We examine decentralized search strategies by simulating the process of local navigation in a network between a starting node S and a target node T. In advance of finding a full path, a navigator is at some node U and tries to evolve the path to target node T by navigating to one of U's neighbors [9], [17], [19], [25]. We model this process as a greedy best–first search procedure, where node U evaluates each of its neighbors V and navigates to the node of the highest score.

Consider that the decentralized navigation strategy navigates the network and tries to find a path from starting node S to target node T. Consider further that the navigation strategy currently resides at node U and wants to move from U to some neighbor V such that the degree of separation to T is decreased. We call node V a topologically closer node as, by navigating to it, the navigation strategy makes a step toward to T.

We say that the navigation strategy makes a *successful* move if it moves to a topologically closer node, i.e., it navigates from node U, which resides at distance $h(U,T)=i$ from T, to node V, which is at distance $h(V,T)=i-1$. We use $P_n(i)$ to denote the *accuracy*, i.e., the probability that the navigation at node U chooses topologically closer neighbor V. Note that, by definition, the topological distance to the target T can decrease by at most one at each step of navigation.

We consider that U performs a greedy best-first search algorithm, i.e., we move from U to node V, where $V=\arg\max_{W\in\mathcal{N}(U)} f_T(W)$ and $\mathcal{N}(U)$ is a set of yet unvisited neighbors of node U. For example, one such heuristic $f_T(V)$ would simply measure the geographic distance between nodes V and target T, and this heuristic would generate greedy moves to a neighbor V that is geographically closest to T. For the network illustrated in Fig. 1(b) geographical navigation would take the path S–B–D–E–G–T of length five even though topologically shortest path of length four exists.

We consider the following heuristic procedures to navigate to a node $V\in\mathcal{N}(U)$. The procedure navigates to node V that maximizes $f_T(V)$:

1) RANDOM: Navigate to a random neighbor V of current node U, i.e., $f_T(V)$ is random;
2) GEO: Navigate to neighbor V that is geographically closest to the target T, i.e., $f_T(V)=-g(T,V)$;
3) DEG: Navigate to highest degree k_V neighbor V of U, $f_T(V)=\deg(V)$;
4) DEGGEO: Navigate to neighbor V with the highest degree-to-distance ratio, $f_T(V)=\deg(V)/g(T,V)^2$.

We consider these navigation heuristics because they are well motivated. DEG navigates through hubs and GEO navigates purely using geography. While these heuristics have been considered in the past, we propose DEGGEO which combines the two strategies based on our findings in the previous sections. In particular, Kleinberg's model [17] suggests that the probability that a node has a friend d km away decays as d^{-2} and assuming that edges are created independently, then the expected number of edges from a node V to the target T is exactly $\deg(V)/g(T,V)^2$. This means that the proposed DEGGEO policy navigates to a node V that is most likely to directly link to T. When far away from T, DEGGEO navigates using degrees, while when getting closer to T, geographic proximity plays a more important role.

We also experimented with other heuristics such as: $f_T(V)=\text{lang}(V,T)$ and $f_T(V)=\text{country}(V,T)$, where $\text{lang}(A,B)$ ($\text{country}(A,B)$) is the probability that a user of language (country) A links to user of language (country) B. We found that these heuristics do not perform as well as GEO.

For each of the four navigation heuristics, we consider 214 662 pairs of nodes S–T. For each of the S–T pairs, we simulate each navigation heuristic until it hits the target node or until the navigation reached 1000 steps, at which point we terminate the process. Also, if there is no better move, then the algorithm may be forced to move from node U to V which has a worse heuristic estimate, $f_T(V)<f_T(U)$.

C. Accuracy of Hitting the Target

Investigating the methods described above, we find that navigation strategies give different performance depending on the

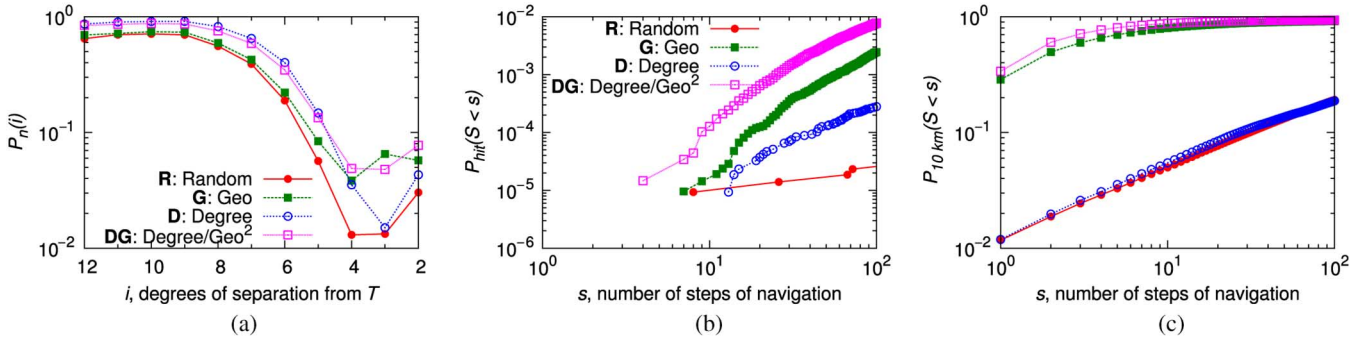


Fig. 6. Comparison of navigation strategies. Refer to the main text for the description of the strategies. (a) Probability $P_n(i)$ that a given navigation strategy that is currently i steps away from target T ($h(U,T) = i$) navigates to a topologically closer node V ($h(V,T) = i - 1$). (b) Probability $P_{\text{hit}}(S < s)$ of hitting the target T in less than s steps of the navigation procedure. (c) Probability $P_{10km}(S < s)$ of navigating inside the 10 km of the target T in less than s steps of the navigation procedure.

stage of the navigation procedure [Fig. 6(a)]. In particular, DEG navigation strategy has best accuracy when the search is topologically far ($i \geq 5$ steps) from T. When closer to T, DEGGEO provides highest accuracy.

For example, for $i \geq 8$ DEG outperforms RANDOM for 35% (DEGGEO improves RANDOM for 28%, and GEO improves RANDOM for 6%), for $i = 7 - 5$ DEG outperforms RANDOM for 120% (DEGGEO 90%, GEO 24%). But for $i = 4 - 2$ DEG gives 75% improvement over RANDOM, while DEGGEO gives 230% and GEO 225% improvement over RANDOM.

Generally, it is easy to navigate toward T when the navigator is far away from T while the probability of successfully navigating toward T is lowest in the last three steps. When the navigation is topologically far from T, it is important to navigate to and through high-degree hub nodes [24], [25]. As the search zooms, in on the target, geography becomes increasingly more important guidance for finding the target.

D. Getting Close to the Target

We discovered that hitting the target T exactly is a low probability event, but that it is much easier to navigate to the close vicinity of T [Fig. 6(b) and (c)]. In 1000 or less steps, the random navigation procedure hits T in only 0.02% of the cases. The best performing strategy is DEGGEO that in 4.92% of the cases hits T in less than 1000 steps (GEO 2.52% and DEG 0.21%). Overall, DEGGEO navigation strategy is 250 times more likely to hit the target than random navigation (DEG outperforms RANDOM for a factor of 10, GEO for 125). Directly hitting the target node T is unlikely, but navigating inside 10 km of T is much more likely. Even random and maximum-degree navigation approach within 10 km of the target in less than 100 steps 20% of the cases. On the other hand, GEO needs 11 steps to get close to T with probability 0.8 and 50 steps to get in 10 km of T with prob. 0.9. DEGGEO is even more successful. The procedure takes only five steps to zoom-in with probability 0.8 and 16 steps to navigate inside 10 km of T with probability 0.9 [Fig. 6(c)].

Taken together, this evidence suggests that navigation strategies, based on geography and simple statistics about the network structure, are effective in guiding navigation on the network. When the search is far away from T, degree-centric navigation is most successful as the search needs to reach a good hub node. Thus, while hub nodes are not necessary for the existence of short paths in the network, they aid the navigation [24]. In the intermediate stage of the search, geography is the best strategy for zooming, in on the target. However, when the search is extremely close to T (less than 10 km), geography is not useful. We hypothesize that, when the navigator is proximal to the target, heuristics based on interests, profession, and other personal characteristics of T may prove more successful [8], [22].

IV. CONCLUSION

Web-based communication platforms have come to innervate large portions of the world and, as such, provide laboratories for studying the structure of connections among people. We used Messenger communications to investigate geographical properties of the social graph captured electronically by the planetary-scale service. We found that shortest paths among people are robust to removal of hub nodes and only increase slightly with large changes in geographic distance among people. Given the abilities of humans to navigate global social networks with local steps, we studied several local navigation procedures. We found that geography provides an important cue in navigating between arbitrary source and target nodes. Overall, we found that navigating through the periphery of the network is relatively easy, but that it is difficult to navigate through the core of the network composed of high-degree nodes.

APPENDIX A
NETWORK AND EXPERIMENTAL DETAILS

A. Network Preparation

We consider all the pairwise conversation activity during 30 days of June 2006. We observed 242 720 596 users log onto the Messenger system during the observation period. On a representative day of June 1, 2006, nearly 1 billion (982 005 323) different conversations (sessions) took place.

During this time, 180 million users sent *or* received messages, while 170 million sent *and* received messages. We

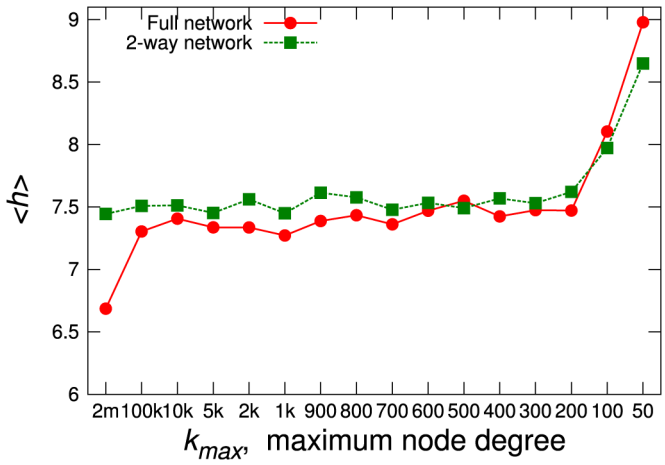


Fig. 7. Comparison of the degree of separation and the network robustness of the full and the 2-way networks.

introduce an arc joining a pair of nodes if the users represented by the nodes exchange at least one message in either direction in June 2006. This *full network* contains 179 792 538 nodes (of nonzero degree) and 1 342 246 427 undirected edges.

We also consider another version of the network where we connect a pair of nodes only if each of the users sends at least one message to the other. We refer to this network as *2-way network* as we only connect pairs of nodes that were engaged in bidirectional communications. The *2-way network* contains 169 415 316 (with nonzero degree) nodes and 1 155 190 935 undirected edges. For both networks, we limit our study to the largest connected component of the network which contains 99.9% of the nodes the particular network.

We note that degree of separation in the Messenger network is independent of the way we construct the network and that both networks exhibit the same level of robustness (Fig. 7). We investigate the fragility of topologically shortest paths and the effect of high degree hub nodes by examining how the average topologically shortest path length $\langle h \rangle$ changes as a function of the maximum degree of any node in the network. We start with the full network, remove all nodes with degree larger than $k_{\max}$, and measure the average shortest path length as a function of $k_{\max}$. We observe (Fig. 7) that the average shortest path length of the full network and the 2-way network are essentially equivalent. The only difference is that when a single highest degree node is kept in the full network, the average shortest path length is 6.6. As soon as this node is removed, the average path length jumps to 7.4 and remains stable up to the point when all nodes of degree greater than 100 are removed.

B. Determining User Geographic Location

We use the IP address of the computers users used to login to the Messenger service in order to decode the geographical coordinates, which we then use to position users on the globe and to calculate distances. The advantage of our approach is that it is not self-reported and cannot be easily gamed by the users. If a user logged in from multiple IP addresses, we use the location of the first login. The reported accuracy of the Geo-IP resolution is as follows: 99.8% accurate on a country level, 90% accurate on a state level, and 83% accurate on a city level with a 25 mile radius [31]. This means that the accuracy of such services is at the same level or better than using the US ZIP code as a proxy for the location. Overall, we consider such service to be accurate and robust enough for the purpose of our analyses.

C. Computing the Degree of Separation

For computations, we used the Stanford Network Analysis Platform (SNAP) [32]. We implemented a parallel version of Dijkstra's algorithm for finding shortest paths between pairs of nodes. The code we developed to perform the experiments is available at http://snap.stanford.edu.

For all of the experiments, we used a four processor server with 64 GB of main memory. Given a single source node, our implementation was able to find shortest path distances to all other nodes in the network in about 16 min. All of our analyses are performed with experiments using a set of 1000 source nodes, and thus we consider shortest paths between $\approx 180 \times 10^9$ pairs of nodes.

REFERENCES

[1] I. de Sola Pool and M. Kochen, "Contacts and influence," *Social Netw.*, vol. 1, no. 1, pp. 5–51, 1978.
[2] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.
[3] J. Kleinfeld, "Could it be a big world after all? 'The six degrees of separation' myth," *Society*, vol. 39, p. 61, 2002.
[4] P. Killworth and H. Bernard, "Reverse small world experiment," *Social Netw.*, vol. 1, no. 1, pp. 159–192, 1978.
[5] C. Korte and S. Milgram, "Acquaintance networks between racial groups: Application of the small world method," *J. Pers. Social Psychol.*, vol. 15, no. 2, pp. 101–108, 1970.
[6] S. Goel, R. Muhamad, and D. Watts, "Social search in "small-world" experiments," in *Proc. Int. Conf. World Wide Web*, 2009, pp. 701–710.
[7] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry*, vol. 32, no. 4, pp. 425–443, 1969.
[8] P. Dodds, R. Muhamad, and D. Watts, "An experimental study of search in global social networks," *Science*, vol. 301, no. 5634, pp. 827–829, 2003.
[9] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," in *Proc. Nat. Acad. Sci.*, 2005, vol. 102, pp. 11623–11628.
[10] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proc. Int. Conf. World Wide Web*, 2008, pp. 915–924.
[11] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint*, arXiv: 1111.4503, 2011.
[12] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.
[13] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, Sep. 1999.
[14] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *Proc. Nat. Acad. Sci.*, vol. 97, no. 21, pp. 11149–11152, Oct. 2000.
[15] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four degrees of separation," *arXiv e-prints*, arXiv: 1111.4570, 2011.
[16] A. Wagner and D. A. Fell, "The small world inside large metabolic networks," *Proc. Roy. Soc. London*, vol. 268, no. 1478, pp. 1803–1810, Sep. 2001.
[17] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, 2000.
[18] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, no. 5571, p. 1302, 2002.

[19] O. Simsek and D. Jensen, "Navigating networks by using homophily and degree," *Proc. Nat. Acad. Sci.*, vol. 105, no. 35, pp. 12758–12762, 2008.
[20] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Phys. Rev. E*, vol. 64, no. 4, p. 046135, Sep. 2001.
[21] F. Menczer, "Growing and navigating the small world web by local content," *Proc. Nat. Acad. Sci.*, vol. 99, no. 22, pp. 14014–14019, 2002.
[22] L. Adamic and E. Adar, "How to search a social network," *Social Netw.*, vol. 27, no. 3, pp. 187–203, 2005.
[23] P. Killworth, C. McCarthy, H. R. Bernard, and M. House, "The accuracy of small world chains in social networks," *SocNet*, vol. 28, no. 1, pp. 86–96, 2006.
[24] R. West and J. Leskovec, "Human wayfinding in information networks," in *Proc. Int. Conf. World Wide Web*, 2012, pp. 619–628.
[25] R. West and J. Leskovec, "Automatic versus human navigation in information networks," in *Proc. Int. Conf. Weblogs Soc. Med.*, 2012, pp. 362–369.
[26] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network? The structure of the twitter follow graph," in *Proc. Int. Conf. World Wide Web*, 2014, pp. 493–498.
[27] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, p. 378, 2000.
[28] S. Milgram, "The small-world problem," *Psychol. Today*, vol. 1, no. 1, pp. 60–67, 1967.
[29] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Math.*, vol. 6, no. 1, pp. 29–123, 2009.
[30] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Soc. Netw.*, vol. 21, no. 4, pp. 375–395, 2000.
[31] (2014). *Geoip2 City* [Online]. Available: http://www.maxmind.com/app/city
[32] J. Leskovec and R. Sosič. (2014, Jun.). *SNAP: A General Purpose Network Analysis and Hraph Mining Library in C++* [Online]. Available: http://snap.stanford.edu/snap

Jure Leskovec received the bachelor's degree in computer science from the University of Ljubljana, Ljubljana, Slovenia, and the Ph.D. degree in machine learning from Carnegie Mellon University, Pittsburgh, PA, USA, in 2008.

He received postdoctoral training at Cornell University, Ithaca, NY, USA. He is an Assistant Professor of Computer Science with Stanford University, Stanford, CA, USA. His research interests include mining large social and information networks.

Dr. Leskovec was the recipient of several awards including a Microsoft Research Faculty Fellowship, the Alfred P. Sloan Fellowship, and numerous Best Paper Awards.

Eric Horvitz received the Ph.D. degree in biomedical informatics and M.D. degree from Stanford University, Stanford, CA, USA, in 1991 and 1994, respectively.

He is a Distinguished Scientist with Microsoft Research, Redmond, WA, USA. His research interests include principles and applications of machine perception, learning, and inference. Beyond efforts on core challenges in artificial intelligence, he has pursued studies in search and retrieval, biomedical informatics, human–computer interaction, human computation and crowdsourcing, and analyses of social networks and behavioral data drawn from online systems.

Dr. Horvitz was elected as a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI), the Association for Computing Machinery (ACM), the American Association for the Advancement of Science (AAAS), and the National Academy of Engineering (NAE), and was inducted into the CHI Academy of ACM SIGCHI.