Title:

Understanding social structures behind the six degrees of separation

Authors:

Carlos Herrera-Yagüe1,2,3, Christian M Schneider1, Thomas Couronne4, Zbigniew Smoreda4, Rosa M Benito1,5, Pedro J Zufiria2,3, Marta C Gonzalez1,6*

Affiliations:

1 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America
2 Depto. Matemática Aplicada a las Tecnologías de la Información, ETSI Teleco

municación, Universidad Politécnica de Madrid (UPM), Spain
3 Cátedra Orange, Universidad Politécnica de Madrid (UPM), Spain
4 Sociology and Economics of Networks and Services department, Orange Labs, Issy les Moulineaux, France
5 Grupo de Sistemas Complejos and Departamento de Física y Mecánica, ETSI Agrónomos, Universidad Politécnica de Madrid (UPM), Spain
6 Engineering Systems Division, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

*martag@mit.edu

One Sentence Summary:

We analyze large social networks from three countries, discovering that not the distance among acquaintances but their community structure makes them searchable within cities.

Abstract:

Empirical evidence has proved that social networks are searchable, meaning that a short path from one person to any other person can be collectively found, despite that each individual has only local information of the entire network. The structure of empirical social networks that allows these phenomena has not yet been completely uncovered, and this knowledge is important for communication systems and models of information spreading. In this work we evaluate diverse decentralized search strategies in social networks of over 25 million phone users from three countries. While the city of the target person is reached within a few hops almost independently of the origin and the routing strategy, we show, for the first time, that within cities, the strategies based on geographic information completely fail to reach the target. This failure occurs because the relationship between the social networks and their geographical space does fundamentally change within cities: social groups are geographically scattered across the city, as opposed to what happens at country scale where communities nearly preserve geo-political borders. We present structural evidence for the success of routing strategies based on the network's community structure within cities.

Main Text:

In the 1960's, social psychologist Stanley Milgram carried out a series of experiments where participants in Omaha, Nebraska, were given a letter whose addressee was a person in Boston, Massachusetts (1). The rules of the experiment stated that the participant was restricted to mail the letter to any acquaintance (who would become a new participant) with the aim that the letter would finally reach the addressee. Surprisingly, one third of these letters reached their intended target in a median number of only six hops, giving rise to the popular expression “six degrees
of separation”. Not only did short paths exist between individuals, but also people were able to find them without information of the global structure. While the former phenomenon fundamentally changed the way social networks were modeled (2-3), the latter remains under debate despite numerous experiments (4-8), simulations (9-12) and theoretical models (13-16).

Milgram-like routing experiments that require real participants suffer from low completion rates (< 30%), do not provide enough information to map the entire networks, and rely on self-reporting, which is prone to errors (8). Theoretical models of social networks that warrant searchability were proposed in the seminal work by Kleinberg et al (13-16), known as the group-model framework, which determines the conditions of decentralized searching depending of the distribution of a social distance $d(u, v)$ between all pairs of connected nodes $u$ and $v$. The social distance is defined as the size of the smallest group containing $u$ and $v$. Given the distance distribution of the form $P(u \text{ and } v \text{ directly linked}) \sim d(u, v)^{-\gamma}$, when $\gamma < 1$ the social network is not searchable; if $\gamma = 1$ the social network is always searchable, and if $\gamma > 1$ the network can be searchable. While routing simulations to date have relied either on social media networks (e.g. Facebook or LiveJournal) or corporate networks that may not generalize to other contexts (9-12, 17) there is no conclusive empirical evidence of the structural properties behind the success of large-scale decentralized routing strategies.

To fill this gap, we simulate message-forwarding strategies on three real world social networks in country scale. Note that actual participant behavior is not examined, as this has been the subject of extensive work (4-8). Rather, we systematically test, at unprecedented scales, different combinations of strategies reported by participants in real experiments (e.g. routing strategies based on geography, number of friends, and profession) and show whether the underlying social structure permits their success. Previous work has found that: (i) social connections form a small world network where geographic neighbors are highly clustered (2), (ii) the attrition rate is constant (8), and (iii) geographic routing dominates initial hops, while selection based on social features, such as profession and family, increases with path length (8). The results presented here uncover previously unknown geographical features of social networks, revealing that the empirically found routing strategies are a direct consequence of these features.

We simulate different routing experiments on social networks obtained from more than seven billion reciprocal mobile phone calls in three European countries, France, Portugal, and Spain with 18.7, 1.2, and 5.9 million users, respectively. First task before simulating routing strategies is confirming that the network exhibits the “small world” property, which we demonstrate by showing that the average number of people in the shortest path between a sender and a recipient $<l>$ is 6.5, 6.4, and 8.4 in the different countries, similar to the observed in previous works (18-21). Furthermore, we can actually extract the spatial distribution of the most central people in the network. In Fig. 1 we
show the distribution of the average number of people in the shortest path between a sender and a recipient \( p(<l>) \) among the population for each country. This value is also known as the inverse of the closeness centrality (22) and it ranges from 3.8 to 11, so everyone in the country is on average within 4 hops from the most central people and within 11 from the less central ones. Each dot represents a mobile phone tower, which is our smallest spatial resolution. In order to expose the backbone of the social network, the color intensity of each mobile phone tower represents the closeness centrality of the most central person in that tower. While main cities appear brighter, centrality is not only determined by population density: Barcelona area (Spain NE) is highly populated but it seems to be socially less central than Alicante (SE) even though the latter has half the population. Notice also, that the most geographically central city is not necessarily the most socially central one. Additionally, the links highlight the social connections only among the 50 most central people in each country, showing significant differences in the social network analyzed in the three countries. Whereas in France, Paris hosts the most central people, in Portugal two cities, Lisbon and Porto, seem to be equally important. In contrast, the most central people in Spain are spread over the entire country including even the Gran Canaria and Mallorca islands.

Once established that the short paths exist all across the network, we explore the success of routing strategies at two levels: intercity and intracity. Intercity routing seeks to reach the correct municipality while intracity routing searches for the individual target within a city. Cities are defined by their administrative borders (S8). On both levels, we test different decentralized routing strategies which employ only information of neighbor nodes (also called contacts or friends). In a random search (run), individuals route the message by randomly selecting a neighbor node that did not have had the message previously. Geographical routing (geo) passes the message to the contact that is geographically closest to the final target, whereas degree routing (deg) selects the friend with the highest number of friends. Finally, community routing (com) forwards the message to a friend such that he/she belongs to the smallest community containing the target (see details in SI) (23-28). To detect communities in social networks, we use the well-established Louvain method (24). This method is a greedy optimization method that attempts to optimize the network modularity by aggregating nodes belonging to the same community and building a new network whose nodes are these communities. This method assigns to each person a set of communities at different hierarchical levels. Thus if two nodes share a certain community, these two nodes share all their larger communities. The number of hierarchical levels depends on the network itself and it is automatically detected by the algorithm. Note that these communities can be used as a proxy for the social groups of a person, since when individuals are presented algorithm-detected
communities in their social networks, they can identify the common social features in each community, such as family, workplace, or college attended (29).

First we evaluate intercity routing. We simulate routing processes using an upper limit of 100 hops and evaluate the performance of the different strategies if they reach the target municipality. We quantify the performance of a routing strategy by the success rate, $R$ (the ratio of delivered messages), and the average number of hops of the strategy, $<l_{tg}>$, taken by delivered messages. In Fig. 2a we show how these two quantities increase as the as the upper limit in the number of hops increases. These results demonstrate that decentralized routing to a given municipality (i.e. city) has high success rates (> 90%) and requires only a few hops when using either com or geo routing. The fact that both strategies perform similarly is explained when we explore communities in the social network of the entire country and verify that these are geographically determined (Fig 3a, S7) which is consistent with previous literature (27). As expected, the success rate depends only logarithmically on the population size of the destination city (Fig. S11), confirming that both strategies are equally efficient. The entire routing experiment can be replicated in our homepage (30). Geographic strategies had already been reported successful using a half million bloggers’ network across the US (10). However, intracity routing has not been previously explored because the low sample size of the network (0.15% of US population) and lack of information of the coordinates of individuals within cities obliged to relax the modeled network structure: namely, nodes were allowed to forward messages to anyone else within the target city, even if they were not directly connected. In contrast, our larger sample (12% - 40%) and higher spatial resolution (mobile phone tower scale) allows us to explore routing inside cities using strict routing among connected individuals.

As previously mentioned, we use administrative borders to define cities: provinces as upper limits (usually containing large cities plus surroundings) and municipalities as lower limits (see SI for details). Thus, we analyze the three different routing strategies in social networks of 155 large municipalities and all 150 provinces in the three countries. In contrast to intercity routing, routing inside municipalities is significantly more successful if the strategy uses community information (Figs. S13 – S18 show additional strategies). For different routing strategies Fig. 2b shows the success rate for municipalities (filled circles) and provinces (open circles) in each country as a function of the population size $N$; an upper limit of 100 hops was employed and Fig. S26 shows qualitatively similar results with a smaller upper limit. We find that at both scales the community based routing is efficient because of the slow decay in success rate $R \sim \beta - \eta \ln N$ ($\beta = 2 \pm 0.03$ and $\eta = 0.133 \pm 0.003$) and in contrast to the random strategy, which as expected decays almost reversely linear as $R \sim N^{-\alpha}$ ($\alpha = 0.95 \pm 0.03$). Interestingly, the geographically based routing
presents a crossover behavior between municipalities (only intracity routing) and provinces (including an initial intercity stage). This behavior is due to the fact that a province consists of several municipalities. Although the geographically based routing reaches the correct municipality, within the municipality this strategy fails. This explains the different scaling observed for geographic routing in municipalities and provinces: while within municipalities the routing success rate scales similarly to the random routing \( R \sim N^{-\alpha} \) (\( \alpha = 0.66 \pm 0.03 \)), the province routing success rate scales similarly to community routing \( R \sim \beta - \eta \ln N \) (\( \beta = 0.82 \pm 0.05 \) and \( \eta = 0.056 \pm 0.004 \)), but with a lower success rate as a consequence of its inefficacy within municipalities.

To explain the failure of geographic routing within cities, we investigate further the spatial properties of social networks within cities. On the country scale the identified communities are highly spatial correlated and even redraw the administrative borders as shown in Fig. 3a for France; the colors indicate the dominant community of each mobile phone tower. However, in the city scale the communities are scattered over space and within the downtown area they are nearly randomly distributed. This shows for the first time that communities within cities are not geographically determined. Interestingly, despite the fact that social ties are locally driven, i.e. the probability of connection is known to be a distance decaying function, the network communities are not. We find this ubiquitous social network structure causes the georouting strategies to fail. The same results are presented for Portugal and Spain in Fig S7.

Next, we show how this network structure confirms the performance of geo and com strategies within cities in agreement with the theoretical conjectures. To do so, we examine the theoretical searchability conditions provided by the group-model framework (14), which generalizes previous results on hierarchical networks (13) and spatial lattices (15). To apply it to our data, we divide the network into groups of size \( S_X \) using either geographic (\( X = r \)) information or existing communities (\( X = c \)) as illustrated in the insets of Figure 3b. Then we calculate the probability that two nodes that belong to the same group (being that group the smallest they both belong to) share a link and how this probability depends on the group size. We observe that both functions have the exponent close to \( \gamma = 1 \), but in the groups based on geography these exponents are mainly below 1, while the exponents are consistently above 1 for communities as shown in Fig. 3b for France. Although the group-model framework does not capture all of our network properties (heterogeneous degree distribution and clustering coefficient) we find that our results confirm theoretical statements regarding the conditions for searchability of social networks.

Interestingly, the failure of geo routing is rooted in a structural feature: the connectivity of social networks within small radius in the city vanishes as a consequence of \( \gamma < 1 \). This is equivalent to saying that if a message headed to
target user B has reached a user A, then there is no path between A and B only through the people who are closer to B than A is (this is illustrated by the lower inset of Fig 3b).

We illustrate this fact further by calculating the size of the largest cluster in the part of the network within different radius $r$ and group sizes $S_r$, performing the analysis in different initial locations from the capital municipality or a province of the three countries. Fig. 3c shows that the fraction of nodes in the giant component is much smaller within cities than within provinces. Surprisingly, we find that this lack of connectivity is not caused by not having enough short-distance links (actually between 18% and 40% of links are within resolution limit), in fact, when we zoom into a region of the city we see highly clustered small groups which only connect to each other through people living far away.

Although it might be natural to think of a critical distance, group size or area, which changes the regime from intracity to intercity, our data show no evidence of such threshold. In contrast, we find that the fact of being in the same city, often defined as an area more densely populated than its surroundings (31-33), is what matters to measure this structure in the social network. For example, we show consistently that people living in a small cities form a network with a giant component much larger than the same number of people living within the same radius in a large city (Fig S24). One possible explanation for this is that people living in large cities are more willing to travel a bit further to meet acquaintances, so a larger geographic ball is needed to get a giant component.

The presented results lead to the following discoveries: (i) Communities within cities are not geographically determined in contrast to bigger geographical regions. Consistently, while geo and com strategies are both efficient at the intercity stage, they do show different performance at the intracity level (ii) While people living within geographic radius including several cities form a connected network, the same radius within cities leads to highly clustered components only connected through people in distant parts of the city. This behavior occurs across different city and region sizes, highlighting cities as functional entities of the social networks (iii) Geography does not provide an efficient routing criteria within cities, however the existence of a community structure which allows routing indicates people can still find short paths in the urban scenario. This finding is consistent with experimental results that suggest people do use the profession or name of the target in the final steps to make inferences about his/her education or ethnicity, as a hint to help routing (8).

In closing, our results uncover an unknown feature of social networks: while traditional models of social networks consist of highly connected and geographically close communities, we find that geography plays only a minor role when forming communities within cities. In fact, social small world networks consist of highly connected
communities within cities, but having geographically long-range links between them. This structure explains why people are able to successfully route in Milgram-like experiments, provided they correctly identify the community of the target. Our results support the theoretical hypothesis of Kleinberg: the likelihood to find friendships within communities decays as a power-law with increasing community size (14), confirming that among all possible network configurations, humans have favored those such that a message can reach anyone even if delivered using only local information. This is a remarkable example of a self-organized structure that allows a small group of individuals to solve a complex problem by cooperating to take collective knowledge (34, 35).

References and Notes:


**Acknowledgements:**

This work was funded by New England UTC Year 23 grant, awards from NEC Corporation Fund, the Solomon Buchsbaum Research Fund. CHM acknowledges support from Orange Spain (France Telecom Group). PIZ from MTM2010-15102 of Ministerio de Ciencia e Innovacion and Q10 0930-144 of the UPM, and RMB from Fundación Caja Madrid (Spain) and project grant MINECO-Spain MTM2012-39101.
Figure 1: Visualization of central places in France, Spain and Portugal. Each circle represents a mobile phone tower and its color (the brighter the more central) corresponds to the inverse of closeness centrality $<l>$ (average number of hops to any other person) of the most central persons in this tower. A person is always assigned either to its billing zip code in Spain or to its most used tower in France and Portugal. White lines highlight the social network between the 50 most central persons of each country. In the three insets the distribution of the $<l>$ of all persons and the relation to the used color are also shown.
Figure 2: a) Dependence of the number of hops $\langle l_{rg} \rangle$ on the success rate $R$ for intercity routing (results for completing the delivery within 15 and 50 hops are highlighted by circles). b) Success rate $R$ versus population size $N$ for three strategies in municipalities and provinces. All logarithmic and power-law functions are guides to the eye.
Figure 3: a) Communities are not mainly geographically based within cities. Each mobile phone tower is colored according to its dominant social community structure. b) Comparison of the exponent $\gamma$ for the probability of finding a link between two people as a function of smallest common group size: $p(S) \sim S^{-\gamma}$ for 96 cities in France. Groups are constructed either based on geography ($S_r$ - black) or on community ($S_c$ - red). c) Fraction of nodes in the giant component $G_r$ for various geographical groups with size $S_r$ for the three capitals (closed circles) and for the entire country (open circles).