

Beyond the Social Web: the Geo-Social Revolution

Salvatore Scellato
Computer Laboratory
University of Cambridge, United Kingdom

While in the last years massive online social networks such as Facebook and Twitter have become extremely popular, gathering and engaging millions of users, only recently the widespread adoption of mobile devices has led to a large portion of users continuously using these social services during their daily lives. These devices offer geolocation capabilities: the ability to share your location, to generate location-tagged information and to search for it adds a crucial spatial dimension to online social networking. As a result, online social services are increasingly becoming location-aware, allowing users to create and access information about their geographic whereabouts: the trend is progressively going from specialized providers offering *location-as-a-service* to a widespread new concept of *location-as-a-feature*, where every online social platform integrates geographic information into their services. This provides broad and plentiful data to investigate how spatial and social structure blend together, opening exciting research directions with promising scientific and practical applications.

In this article I will present how the socio-spatial properties of online social networks can be studied and how novel *geo-social* network measures can be defined to blend together social and spatial properties. The availability of geographic data, together with the rising importance of mobile devices, is likely to transform the Web and how it is used by billions of people every day.

DOI: 10.1145/2020936.2020941

<http://doi.acm.org/10.1145/2020936.2020941>

1. INTRODUCTION

Online location-based social networks have recently attracted millions of users, experiencing a huge popularity increase over a short period of time. Thanks to the widespread adoption of location-sensing mobile devices, users can share information about their location with their friends. As a consequence, these services offer a groundbreaking opportunity to expand our knowledge of the social Web and include an additional and crucial dimension in our investigation: *the places where people live*. This means that while online social interactions were often studied from a purely structural point of view, that is, who interacts with whom, it is now possible to study also the geographic distance that these interactions bridge.

Among the many interesting questions arising from the availability of this new kind of information a fundamental one is to understand whether, and how, distance is affecting online social interactions. This involves studying the social network arising among online users from a spatial perspective, embedding users in a metric space and then associating a geographic length to their links. Such spatial networks have been extensively studied in the past years, particularly in the case of transportation networks, power grids, urban road net-

Dataset	N	$\langle k \rangle$	$\langle C \rangle$	$\langle D \rangle$	$\langle l \rangle$
Brightkite	54,190	7.88	0.181	5,651	2,041
Foursquare	258,706	22.07	0.191	8,494	1,442
Gowalla	122,414	9.48	0.254	5,663	1,792

Table I. Properties of the mobile datasets: number of nodes N in the social network, average node degree $\langle k \rangle$, average clustering coefficient $\langle C \rangle$, average geographic distance between nodes $\langle D \rangle$ [km], average social link length $\langle l \rangle$ [km].

works and other systems where nodes are embedded in a metric space [Barthélemy 2011]. In general, metric distance directly influences the network structure of such systems by imposing higher costs on the connections between distant entities. Social networks, instead, have been largely studied from a purely topological perspective, focusing on the structural position of their nodes and on structural mechanisms that describe their evolution. Indeed, the connection cost that heavily affects other types of spatial networks may not be as important in social systems, particularly when focusing on online interactions. It has also been proposed that distance may cease to play a role because of the increasing availability of affordable long-distance travel and new communication media, resulting in the inevitable “Death of Distance” [Cairncross 2001]. However, some recent results have put forward the idea that distance still matters [Liben-Nowell et al. 2005]: online users tend to connect more with other individuals living nearby, giving rise to predictable patterns that can be studied and exploited to build systems and applications such as predicting where users really live [Backstrom et al. 2010]. Even though spatial distance might greatly influence online social networks, the socio-spatial properties of such systems are still largely unknown.

In this article I will discuss how the socio-spatial properties of online social networks can be studied and I will present some results on three different popular services: Brightkite, Foursquare and Gowalla. I will discuss how online users are affected in a heterogeneous way by geographic distance, with some individuals exhibiting mainly short-range rather than long-distance social ties and clusters. Then, I will present two novel geo-social measures that can capture and highlight these socio-spatial heterogeneities.

2. SPATIAL PROPERTIES OF ONLINE SOCIAL NETWORKS

The methodology involves studying a social network as a spatial network: every user is assigned to a “home location” and then social ties are stretched across space by considering the geographic distance they span. I will first address the global properties of the three social networks under analysis, focusing on their main social and geographic measures, also reported in Table I. For each service I extract the social network arising among users and their *check-ins*, that is the messages they post to share the place where they are. The *geographic home location* for each user is defined as the place where he/she has more check-ins overall.

Low distance between friends The average geographic distance between users $\langle D \rangle$ is consistently larger than the distance between friends $\langle l \rangle$: while the first value ranges between 5,600 and 8,500 km, the latter has much lower values, between 1,400 and 2,000 km. This provides evidence that the probability of having a social link between two users might

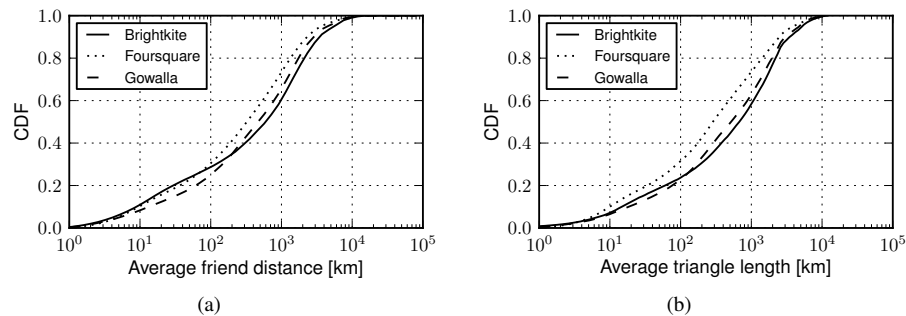


Fig. 1. Empirical Cumulative Distribution (CDF) of the average friend distance (a) and average triangle length (b) for each user in the social networks.

be higher at lower distances, as found in other recent studies [Backstrom et al. 2010].

Spatial user heterogeneity of social ties An interesting question is how the social ties of individual users stretch across space. A basic and intuitive measure of the socio-spatial properties of a given user is the average geographic distance of his/her friends: the probability distribution of this quantity across the users of the three different services is presented in Figure 1(a), from [Scellato et al. 2011]. The existence of values over all geographic scales is due to the existence of users with different characteristic lengths of interaction. For instance, about 10% of users have connections with an average length of just 10 km, whereas about 20% of users have average friend distances above 2,000 km. In particular, links with different geographic lengths do not appear homogeneously across all users. Instead, there is heterogeneity between users, with some of them with only short-range connections and others with long-distance ties.

Spatial user heterogeneity of social triangles After analyzing individual social ties, another interesting point to look at regards the geographic properties of social triangles. In fact, users tend to belong to several triads, resulting in high values of clustering coefficient: the networks under analysis exhibit clustering values between 0.18 and 0.26. To assess whether geographic heterogeneity arises also for social triangles, it is useful to compute the geographic average triangle length of the three links of each triangle and then to compute the average triangle geographic length for each user by considering all the triangles he/she belongs to. The aim is to assess the geographic span of a user's social triangles, whatever their number might be. Figure 1(b) from [Scellato et al. 2011] displays the distribution of the average triangle length for each user is shown: triangles with different geographic span are not equally arising among all users, but instead there are users with smaller triads and users with wider ones. For example, there are at least 20% of users with an average triangle length less than 100 km, while the top 20% have values above 2,000 km.

3. GEO-SOCIAL MEASURES

Given the properties observed in the previous section, my aim is now to present new geo-social network measures which are able to capture, respectively, the geographic heterogeneity observed in social ties and social triangles [Scellato et al. 2010]. Nodes are em-

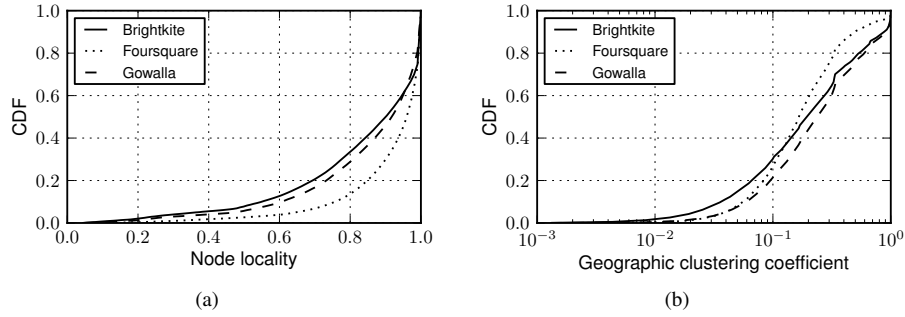


Fig. 2. Empirical Cumulative Distribution (CDF) of node locality (a) and geographic clustering coefficient (b) for each user in the social networks.

bedded in a 2-dimensional metric space where the distance between two nodes i and j is given by the geographic distance D_{ij} between their locations on the planet. This distance is used as the length of the link l_{ij} between nodes i and j .

Node locality Let us consider an undirected geographic social network, a node i with a particular geographic position and the set Γ_i of its neighbors. The node degree k_i is the number of these neighbors, that is $k_i = |\Gamma_i|$. Then, the *node locality* of i can be defined as a measure of how much geographically close its neighbors are and it is computed as follows:

$$NL_i = \frac{1}{k_i} \sum_{j \in \Gamma_i} e^{-l_{ij}/\beta} \quad (1)$$

where β is a scaling factor which avoids extremely small values of node locality when links have large lengths. By definition, NL_i is always normalized between 0 and 1. The value of β can be chosen so that networks with different geographic size can still be compared with each other.

Users exhibit an overall high average node locality: Brightkite has an average value of 0.82 and Gowalla of 0.85, while in Foursquare this value goes up to 0.90. The distributions of node locality for the three networks are shown in Figure 2(a): node locality is able to capture how different users have heterogeneous spatial properties, with values spanning the entire range. For example, in Brightkite and Gowalla about 40% of users have a node locality higher than 0.90, whereas in the Foursquare dataset this phenomenon is more evident, with 70% of users above 0.90.

Geographic clustering coefficient Similarly, the *geographic clustering coefficient* is defined as an extension of the clustering coefficient used for complex networks. The clustering coefficient measures the proportion of triangles among the neighbors of a given node: the geographic clustering coefficient of node i is thus defined in the same way as the clustering coefficient, but each existing triangle between nodes i , j and k is assigned a weight w_{ijk} defined as:

$$w_{ijk} = e^{-\frac{\Delta_{ijk}}{\beta}} \quad (2)$$

where Δ_{ijk} is the maximum length among the three links, that is $\Delta_{ijk} = \max(l_{ij}, l_{ik}, l_{jk})$. If there is no link between j and k , then $w_{ijk} = 0$. Since this measure uses the maximum

weight among all the links of a triangle, it focuses on nodes which are all close to each other: when just one of the three nodes is not close to the other two, the weight will immediately decrease. This emphasizes social triangles where users are extremely close to each other. Again, the parameter β is used to scale the values of the measure.

The three networks exhibit different values of geographic clustering coefficient: while Brightkite has an average value of 0.165 and Gowalla of 0.171, Foursquare scores a much higher 0.209. The geographic clustering coefficient is close to the standard clustering coefficient, signaling how triangles tend to form at shorter distances. The probability distributions of the geographic clustering coefficient are shown in Figure 2(b): again, the three networks exhibit a non-negligible portion of users with a coefficient close to 1.

4. APPLICATIONS AND FUTURE WORK

The increasing availability of geographic and spatial information on the Web is already revolutionizing the services offered to users: these novel geo-social metrics can be effectively exploited to improve large-scale systems tightly coupled with online social networking services. For instance, since users exhibit geographic locality of interest, and since social networks are already driving a large fraction of Web traffic, many requests to services such as YouTube tend to be both geographically and socially correlated. As a result, distributed caching systems can prioritize items by tracking their spreading on online social networks such as Twitter and Facebook and predicting whether future accesses will remain geographically close or, instead, reach planetary scale [Scellato et al. 2011].

At the same time, the availability of geographic and social data might dramatically improve our understanding of online and offline social behavior: the riveting potential offered by this new generation of social Web platforms is likely to spawn innovative research efforts and applications.

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Cecilia Mascolo for her wise guidance and support to my research efforts, my co-authors for all their essential efforts and my colleagues at the Computer Laboratory for many interesting discussions about this topic.

REFERENCES

- BACKSTROM, L., SUN, E., AND MARLOW, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW '10*. 61–70.
- BARTHÉLEMY, M. 2011. Spatial Networks. *Physics Reports* 499, 1–101.
- CAIRNCROSS, F. 2001. *The Death of Distance: How the Communications Revolution Is Changing our Lives*. Harvard Business School Press, Cambridge, MA, USA.
- LIBEN-NOWELL, D., NOVAK, J., KUMAR, R., RAGHAVAN, P., AND TOMKINS, A. 2005. Geographic routing in social networks. *PNAS* 102, 33 (August), 11623–11628.
- SCELLATO, S., MASCOLO, C., MUSOLESI, M., AND CROWCROFT, J. 2011. Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades. In *Proceedings of WWW'11*.
- SCELLATO, S., MASCOLO, C., MUSOLESI, M., AND LATORA, V. 2010. Distance Matters: Geo-social Metrics for Online Social Networks. In *Proceedings of WOSN'10*.
- SCELLATO, S., NOULAS, A., LAMBIOTTE, R., AND MASCOLO, C. 2011. Socio-Spatial Properties of Online Location-Based Social Networks. In *Proceedings of ICWSM'11*.