Distance Matters: Geo-social Metrics for Online Social Networks

Salvatore Scellato
Computer Laboratory
University of Cambridge

Cecilia Mascolo
Computer Laboratory
University of Cambridge

Mirco Musolesi[*]
School of Computer Science
University of St. Andrews

Vito Latora
Dipartimento di Fisica
Università di Catania and INFN

Abstract

Online Social Networks (OSNs) are increasingly becoming one of the key media of communication over the Internet. The potential of these services as the basis to gather statistics and exploit information about user behavior is appealing and, as a consequence, the number of applications developed for these purposes has been soaring. At the same time, users are now willing to share information about their location, allowing for the study of the role of geographic distance in social ties.

In this paper we present a graph analysis based approach to study social networks with geographic information and new metrics to characterize how geographic distance affects social structure. We apply our analysis to four large-scale OSN datasets: our results show that there is a vast portion of users with short-distance links and that clusters of friends are often geographically close. In addition, we demonstrate that different social networking services exhibit different geo-social properties: OSNs based mainly on location-advertising largely foster local ties and clusters, while services used mainly for news and content sharing present more connections and clusters on longer distances. The results of this work can be exploited to improve many classes of systems and a potential vast number of applications, as we illustrate by means of some practical examples.

1 Introduction

In the recent years massive Online Social Networks (OSNs) such as Facebook, MySpace, LinkedIn, Flickr and Twitter have become increasingly popular, gathering millions of users and engaging them in the production, sharing and consumption of information over social links. The numbers are overwhelming: for example, Facebook has more than 400 million active users, which share about 3.5 billion pieces of content each week and upload about 2.5 billion photos each month [4].

The growth of these services has attracted the interest of the academic community [11, 15, 10]: even if "offline" social networks have been under investigation for many decades [19], the availability of such large OSNs provides fascinating opportunities to understand how social structures develop and arise at a large scale.

OSNs are increasingly becoming location-aware: they offer the opportunity to share geographic location in order to generate location-tagged information and to search for it. For instance, there are social networking services grounded on the idea of advertising your exact location to all your friends in real-time, such as Gowalla. Similarly, Twitter has always provided its users with the option of sharing some information about their location, with an increasing proportion of them already doing so. These new features open novel research directions which are largely unexplored, such as the design of new social applications and the improvement of existing large scale systems. Hence, it becomes important to investigate how geographic distance between individuals affects OSNs in order to deepen our understanding of these networks.

In this paper we firstly present a new approach for the analysis of networks with geographic information: then, we define new geo-social metrics which are able to quantify if an individual has short-range or long-distance social ties and to discern if its neighbors form clusters at a large or small scale. We show how these metrics provide a better understanding of these networks, since they take into account geographic properties of the social ties across people. We apply our metrics to four different OSNs which provide location information for their users: we study two purely location-based social networking services (BrightKite and FourSquare), one blogging community (LiveJournal) and a social microblogging platform (Twitter). We study their geo-social properties and we find that their users exhibit a tendency to have short-range social connections and that clusters of friends tend to be geographically confined. Moreover, we show how location-based OSNs exhibit different

[*] This work was carried out in part when Mirco Musolesi was at the Computer Laboratory, University of Cambridge.

geo-social properties than services mainly based on content sharing and broadcasting, for they seem to foster more local, short-range interactions among users.

The main contributions of this paper can be summarized as follows:

- We describe an analytical framework where network nodes are embedded in a metric space, in order to study the relationship between social connections and geographic distance. We define two new geo-social measures: a *node locality* metric, which quantifies how much a node is engaged with a local rather than global set of individuals, and a *geographic clustering coefficient*, which extends the standard notion of clustering by taking into account how much clusters of people are connected by short-range ties.
- We describe and analyze the social, geographic and geo-social properties of four different datasets of real OSNs with geographic information. We characterize some recurrent properties of the networks: in particular, we observe that there is a vast portion of users with short-range friendship links and that, at the same time, clusters of friends are often geographically close.
- We show how different OSNs present contrasting characteristics, which may be explained by a varying attitude of their users towards the social and geographic aspects of online friendship: location-based OSNs engage their users in short-range social connections more than sharing-based services, which exhibit social ties on a wider scale.

2 Motivation

Augmenting social structure with geographic information adds a new dimension to social network analysis and a large number of theoretical investigations and practical applications can be pursued on socio-geographic systems.

Geographic analysis of social networks By taking into account geographic location we can understand which role distance plays in social phenomena such as the creation of friendship ties, the development of personal tastes and the spreading of information. Even though the structure and the dynamics of social networks have been under scrutiny for many years [3, 19], we still need to investigate which influence geography has on both. Previous research has studied how geographic distance affects social ties [14] and how location-based OSNs are changing our attitude towards the perception of space [9]. By analyzing large scale OSNs it becomes possible to understand how users choose their connections over space, how their interactions are affected by distance and whether social influence and trust fade away with remoteness. Similarly, standard social network models do not take into account geographic distance between nodes. As an example, users could be characterized by their preference towards global, long-range interactions rather than towards local, short-distance ties, in order to classify their behavior and understand what factors drive their social interactions.

Design of geo-social applications Geolocation availability on OSNs opens new alluring directions for novel applications and systems. Applications such as social search, social recommendation and advertising would greatly benefit from geographic information about users: search queries about local content could be directed to nearby users with many social links in the area of interest, while both advertising and recommender systems could better profile users by knowing how their social ties stretch over space. Moreover, information about social links and geographic placement can tell us a great deal about how culture and taste disseminate on an OSN. Some potential applications of this idea include targeted advertisement and more effective local content spreading (e.g., shop promotions, local news, job openings).

Improving large-scale systems Finally, large-scale systems would greatly profit from a better knowledge of how users are connected over space and how information spreading over these links creates demand for content and services around the planet. More specifically, with the recent rising interest in cloud services [8] and content delivery networks [13], it has become extremely important to understand how content and service requests are arising from all over the world and whether the design of such systems can be improved by exploiting the geographic properties of social processes. For instance, popularity of content can be geographically and temporally characterized, devising new strategies for replica placement and caching where distribution servers can be pre-loaded with content depending on the sharing interactions, their location and their temporal patterns.

3 Geographic Social Networks

OSNs encourage users to indicate their home location: for example, they can provide details about their hometown, neighborhood or, more recently, their exact location by uploading their mobile device's GPS readings.

In this section we present an approach to study these social networks which takes advantage of the geographic information about nodes to assign lengths to their links. We then define two new geo-social network metrics which are able to quantify, respectively, i) users' likelihood of exhibiting long (or short)-distance social interactions and ii) locality of social clusters (i.e., how users in the same cluster are close to each other).

A *Geographic Social Network* is represented as a graph G with N nodes and K links: nodes represent users and a link among two nodes exists if there is a social tie between them (e.g., a person lists another user as one of his/her friends). A link may be undirected or directed: in the latter case, the existence of a link from node i to node j does not imply the existence of the reverse link from j to i. Given a fixed location on the Earth for each user, nodes are embedded 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 Then, we define a metric to quantify the geographic closeness (i.e., the locality) of the neighbors of a certain node to the node itself. 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:

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

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 does not impact the relative values of node locality: a node with higher locality than another one will always have a higher value, regardless of the value of β. At the same time, β can be chosen so that networks with different geographic size can still be compared with each other, as we will discuss later. Finally, we adopt an exponential decay for node locality to highlight social ties which span over short geographic distances and to reduce, at the same time, the impact of longer ties.

In a similar fashion, in the case of directed graphs the *node in-locality* can be defined considering only the incoming connections of a node and the *node out-locality* is defined considering only outgoing links. A node without in-connections will have, by definition, node in-locality equal to 0; the same applies to out-locality.

Geographic clustering coefficient While node locality captures how close the neighbors of a node are, another measure is needed to quantify how tightly connected the neighborhood of a node is. Thus, the *geographic clustering coefficient* can be defined as an extension of the clustering coefficient used for complex networks [2]. The clustering coefficient measures the proportion of triangles among the neighbors of a given node: this geographic adaptation attempts to weigh differently triangles formed by nodes that are close to each other and triangles where nodes are at longer distance. 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:

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

where Δ_{ijk} is the maximum length among the three links, that is $\Delta_{ijk} = max(l_{ij}, l_{ik}, l_{jk})$. We define $w_{ijk} = 0$ if there is no link between j and k. 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.

In the case of directed graphs, as in the case of the standard clustering coefficient, we consider triangles containing undirected links joining node i to its neighbors and directed links for the remaining side. Thus, if we consider Γ_i as the set of all the neighbors of node i (considering both incoming and outgoing links), with $k_i = |\Gamma_i|$, the geographic clustering coefficient is defined as:

$$GC_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k \in \Gamma_i} w_{ijk} \qquad (3)$$

where the sum is extended only to existing triangles. Since there are exactly $k_i(k_i - 1)$ different ordered couples of neighbors in Γ_i, GC_i is normalized between 0 and 1 by definition.

4 Dataset Acquisition

We now describe the collection methodology we have used to extract data about social links and user geographic information of OSNs and, then, we present the four datasets we analyze in this work.

Collection methodology To extract and collect information about the social ties among users and their geographic location we have crawled a sample of each OSN employing snowball sampling [12]: the data extraction starts from a set of seed users and then it expands the extraction by following the outgoing links of these users to reach new users and so on. However, the snowball sampling procedure is known to be biased [12], since there is a higher probability of sampling nodes with more links. This aspect will be taken into consideration during the evaluation of the analytical results.

Location-aware OSNs The OSNs under analysis were created with different goals and, hence, they exhibit different characteristics. Nonetheless, they all provide static geographic information about their users,

Dataset	N	K	$\langle k \rangle$	$\langle C \rangle$	$\langle L \rangle$	$\langle D_{ij} \rangle$	$\langle l_{ij} \rangle$	$\langle NL \rangle$	$\langle GC \rangle$	ρ
BrightKite	54,190	213,668	7.88	0.181	4.71	5,683	2,041	0.82	0.165	1
FourSquare	58,424	351,216	12.02	0.256	4.60	4,312	1,296	0.85	0.237	1
LiveJournal	992,886	29,645,952	29.85	0.185	4.89	6,142	2,727	0.73/0.71	0.146	0.69
Twitter	409,093	182,986,353	447.29	0.207	2.77	6,087	5,117	0.57/0.49	0.108	0.79

Table 1: Properties of the datasets: number of nodes N and edges K, average node degree $\langle k \rangle$, average clustering coefficient $\langle C \rangle$, average shortest path length $\langle L \rangle$, average distance between nodes $\langle D_{ij} \rangle$ [km], average link length $\langle l_{ij} \rangle$ [km], average node locality $\langle NL \rangle$ (in/out), average geographic clustering coefficient $\langle GC \rangle$ and reciprocity ρ [6].

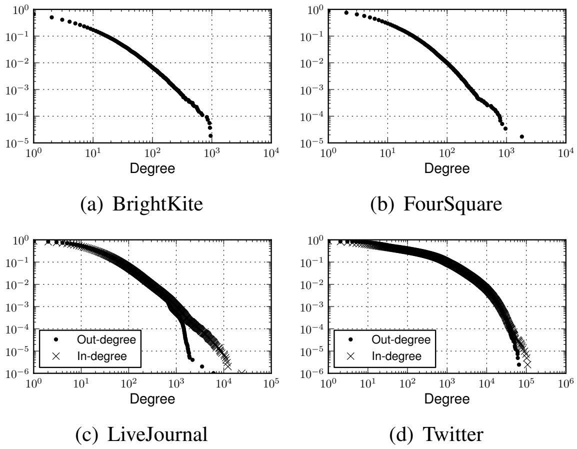
in explicit or implicit form (e.g., geographic coordinates or a city name).

BrightKite was founded in 2007 and it is a social networking website which allows users to share their location, to post notes and to upload photos through different interfaces. It is based on the idea of making "check-ins" at places, where users can see who is nearby and who has been there before. BrightKite users can establish bidirectional friendship links and send public and private messages to each other. It offers a public API which provides geographic coordinates of user home locations and lists of friends. The duration of our crawl was 2 days from September 20 to September 21, 2009: we seeded the crawl collecting 1,000 users from the public timeline and then exhausting the extraction. The dataset contains information about 54,190 users.

FourSquare is a location-based social networking website launched in 2008 which engages its users in a game competition. Users "check-in" at venues in order to be awarded points which contribute to their chart position. This fosters the engagement of users, which are encouraged to check-in as many times as possible. Initially, FourSquare allowed users to check-in only from about 100 cities in the US and in Europe; they have only recently removed this limitation [5]. Furthermore, the service enables the creation of bidirectional friendship links. The website audience has recently grown steadily, reaching about 100,000 members at the end of 2009. We used a public API to retrieve user friend lists and home locations with geographic coordinates. The duration of the crawl was 7 days from November 22 to November 28, 2009 and it was seeded with 1,000 randomly selected user identifiers. Due to a limit on the number of API requests that could be issued, we retrieved a subset of the entire network which contains information about 58,424 users.

LiveJournal is a community of bloggers with around 14 millions active users as the end of 2009. Users can keep a blog or a journal and establish friendship connections among them. Each user provides a personal profile which often includes home location, personal interests and a list of other bloggers considered as friends. Friendship links may not be reciprocal. There is a public API to explore the social network, but it does not expose any method to get user profiles, where location information may be obtained. Thus, the crawling process involved both crawling the social network links through the API and downloading the HTML profile pages of the visited users. Seed users were acquired by accessing the public timeline over 24 hours and then 1,000 users were randomly selected among all the users retrieved. The duration of the crawl was 9 days, from November 2 to November 9, 2009, obtaining a sample of 1,502,684 users. Given the 1,226,412 users which provide location information, we successfully obtained a meaningful geographic location for only 992,886 users.

Twitter is a social networking service which allows users to send short messages known as *tweets*. Tweets are composed only of text, with a strict limit of 140 characters: they are displayed on the author's profile page and delivered to the author's subscribers, who are also known as *followers*. Since its launch in 2006 it has gained a global and vast audience of millions of users all around the world [18]. Twitter does not enforce reciprocity in social connections: a user may follow another one even though the latter is not following back. Hence, the resulting graph is directed. Another key characteristic is the presence of a heterogeneous network structure, where a user may have many more followers than the number of users he/she is following, or vice versa. Twitter provides a public API to gather details on user profiles and follower lists. Due to a rate limit on API requests, it was not possible to collect information about all the Twitter users. The crawling process was seeded collecting 1,000 seed users from the public timeline, which shows a list of the 20 most recent tweets posted by users with unrestricted privacy settings to the entire service. The duration of the data crawling was 6 days from December 3 to December 8, 2009, gathering information about profiles and follower lists for 814,902 different users. Among them, 535,653 reported some information about their home location. We have successfully geocoded 409,093 users, translating their location information into a point on the Earth.



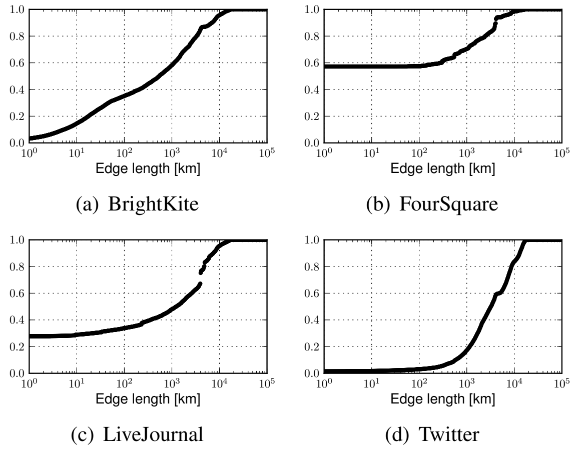
(a) BrightKite

(b) FourSquare

(c) LiveJournal

(d) Twitter

Figure 1: Complementary Cumulative Distribution Function (CCDF) of node degree.



(a) BrightKite

(b) FourSquare

(c) LiveJournal

(d) Twitter

Figure 2: Cumulative Distribution Function of edge length.

5 Geo-social Analysis

In this section we analyze the social structure of the OSNs under investigation. Then, we study their geographic structure with our novel geo-social metrics.

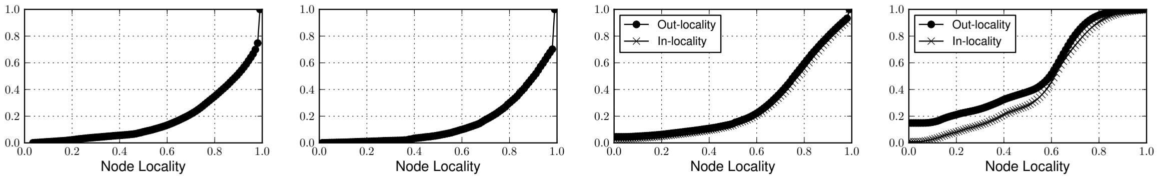
5.1 Social Structure

We have extracted a graph from each sampled dataset: BrightKite and FourSquare are modeled with undirected graphs, whereas LiveJournal and Twitter with directed graphs. In LiveJournal we have a link from user A to user B if A lists B as one of his/her friends; in Twitter there is the same link if A follows B. In Table 1 we report some basic properties of the datasets under analysis. The graphs extracted from these OSNs are quite different: BrightKite and FourSquare have an average node degree $\langle k \rangle$ of 7.88 and 12.02 respectively, whereas LiveJournal has an average degree of about 30 and Twitter shows a larger value of 447. These values give indication that while purely location-based OSNs such as BrightKite and FourSquare have not yet gathered a massive and tightly connected audience, LiveJournal users have built many more connections over the years. The case of Twitter is peculiar: this social network encourages users to follow a large number of other users and, since no reciprocation in link creation is needed, it is easier for a user to accumulate a large number of social connections. Our samples show a dominant giant component which contains almost all the nodes in the sample.

Small-world effect We study the average path length $\langle L \rangle$ of these networks by sampling random pairs of nodes in the network and evaluating their path distance. The average path length is obtained by considering the average shortest path among the sampled pairs. The sampled average path length is above 4.5 hops in FourSquare, BrightKite and LiveJournal and only 2.77 hops in Twitter. Differences among the OSNs are present also regarding the average clustering coefficient $\langle C \rangle$: Twitter and FourSquare have a higher coefficient of 0.207 and 0.256 respectively, while LiveJournal scores 0.185 and BrightKite 0.181. These results confirm the existence of the small-world effect also in this type of social networks, as found in other offline systems [15]: while the average shortest path length is only a few hops, their clustering coefficient is still higher than in a randomized network of the same size.

Degree distribution As shown in Figure 1, the complementary cumulative probability distributions of node degree present a heavy tail, with a vast proportion of nodes with lower degrees and only few nodes with significantly larger degrees. Similar degree distributions have also been found in many other complex networks, such as those that exhibit power-law degree distributions [2]. However, our distributions present a flat head and a fast decay in the tail, which is in contrast with power-law models: this may be due to the fact that snowball sampling underestimates the proportion of nodes with lower degrees, thus lowering the initial values of the distribution.

In the case of LiveJournal the out-degree distribution decays much faster than the in-degree one: this may be related to the fact that LiveJournal had temporarily limited the number of outgoing connections of every user to 750. As expected, this results in a sharp cut-off right after this value. Twitter exhibits degree distributions with a flat head and then a rapid decay, as found in other studies [10]. Since these two OSNs are represented with directed graphs we report their value of reciprocity ρ [6]: this metric measures how likely each link is present in both directions and spans from $\rho = 1$ for perfect reciprocity to $\rho = -1$ if each link is present only in one di-



(a) BrightKite (b) FourSquare (c) LiveJournal (d) Twitter

Figure 3: Cumulative Distribution Function of node locality.

rection. We have $\rho = 0.69$ for LiveJournal, while Twitter has $\rho = 0.79$. Hence, both networks exhibit high values of reciprocity, albeit Twitter appears more symmetric: this property might be related to the fact that it encourages more reciprocal interactions than LiveJournal.

5.2 Geographic Structure

After investigating the social structure of the OSNs we now analyze their geo-social properties.

5.2.1 Geographic Distance

One of the most important characteristics is the geographic distance that social connections span: even if a link between two users denotes some sort of social relationship, it is also important to take into account how it stretches across space. First of all, the OSNs under analysis present different values of the average distance $\langle D_{ij} \rangle$ between users: FourSquare users exhibit an average distance of only 4,312 km, while in BrightKite this value goes up to 5,683 km and in LiveJournal and Twitter it is above 6,000 km. Thus, regardless of social connections, FourSquare has a more limited geographic user base, while the other datasets are more widely spread and comparable among them. In Figure 2 we report the cumulative probability distribution of edge length for the four different social networks. FourSquare has the smallest average, only 1,296 km, and it contains more than 50% of links shorter than 1 km. It has also the shortest average distance between nodes. These values can be explained by the fact that FourSquare was available only in 100 different cities when this dataset was sampled, so users were not distributed all over the world and several social links were among friends colocated in the same city. Indeed, the link length distribution appears flat below 100 km. A similar phenomenon appears in LiveJournal, with around 30% of links shorter than 1 km, albeit the average link length is 2,727 km. On the contrary, BrightKite contains only about 4% of extremely short links, with a global average length of 2,041 km. Nonetheless, about 60% of links are shorter than 1,000 km. Finally, Twitter links have an average length of 5,117 km: there are only less than 5% of links shorter than 100 km, while there are more than 80% of links longer than 1,000 km. This is a clear indication that Twitter users are likely to be engaged with a global audience of followers, even though there are also short-range social connections. These properties may also be affected by the fact that Twitter users tend to have a high number of connections, which makes the network less sparse than the other ones.

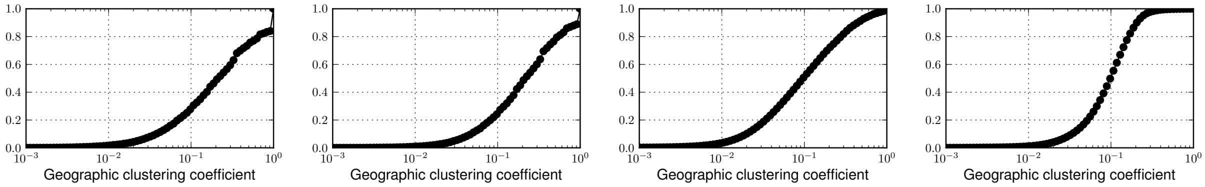
5.2.2 Geo-social Properties

Here we present some results on the novel geo-social metrics we have previously defined, the *node locality* and the *geographic clustering coefficient*.

Choice of scaling factor To compare the results for different geographic social networks a value for the scaling factor β used in Equations 1 and 2 needs to be chosen. By using the same value for every OSN, the network whose nodes are at shorter distances from each other will have higher values of geo-social metrics. Instead, we want to be able to compare the geo-social structure of two networks even if it arises at different geographic scales, i.e., city-wide or nation-wide. Thus, for each OSN we chose to adopt a scaling factor β equal to the mean distance between all its users, which is reported in Table 1. This choice is dependent only on the positions of the nodes of the social graph, not on their links: in this way we can understand how the same set of nodes might have different geo-social properties according to the links among them.

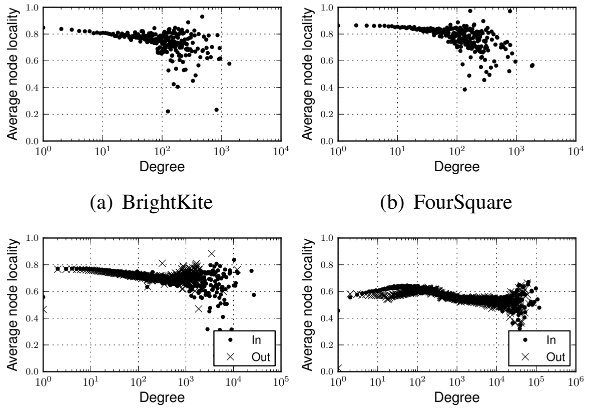
Node locality The probability distributions of node locality for the four datasets are shown in Figure 3. In the BrightKite, FourSquare and LiveJournal networks there is a non-negligible fraction of nodes with node locality close to 1. Hence, there are some users who have social connections only with other individuals within a close geographic distance. Furthermore, in the BrightKite network about 40% of users have a node locality higher than 0.90, whereas in the FourSquare dataset this phenomenon is even more evident, with 1 out of 4 users with node locality close to 1. Users of these location-based Owns exhibit an overall high average node locality: BrightKite has an average value of 0.82, while in FourSquare this value goes up to 0.85.

In the LiveJournal network this effect is weaker: only



(a) BrightKite (b) FourSquare (c) LiveJournal (d) Twitter

Figure 5: Cumulative Distribution Function of geographic clustering coefficient.



(c) LiveJournal (d) Twitter

Figure 4: Average node locality as a function of node degree. For directed networks the relationship is shown both for incoming and outgoing links.

10% of users have a node locality close to 1 and the mean values are 0.73 for in-locality and 0.71 for out-locality. The node locality distribution appears similar both for in- and out-locality. Instead, in Twitter the distribution of node locality shows less nodes with high values. This may provide evidence that Twitter users are more likely to engage with a geographically spread set of individuals rather than only with users at closer distances. Moreover, in- and out-locality show different patterns, since there are more than 15% of nodes with an out-locality of 0, probably nodes without outgoing connections. The average values are lower than in the other networks: 0.57 for in-locality and 0.49 for out-locality.

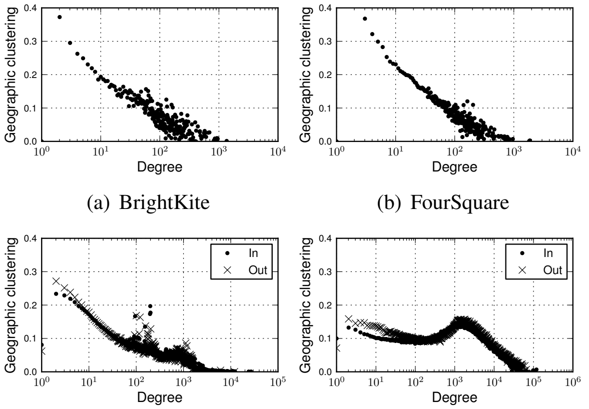
These results show how the new generation of location-based services services, such as BrightKite and FourSquare, is characterized by short-range friendship links among users, resulting in a vast proportion of them with high values of node locality. Thus focusing merely on user location, rather than on what users share and post, may give more opportunities to discover potential friends that live nearby. On the contrary, these patterns are not present in social networks which are less centered on user location: in LiveJournal users have connections with heterogeneous length and this effect is even greater in Twitter. Their users may be more interested in becoming friends with individuals which post and share interesting content rather than simply with people at close distance.

Node locality and node degree We now analyze the correlation between node degree and node locality to understand the geo-social properties of users with different numbers of connections. The average node locality as a function of node degree for BrightKite and LiveJournal is shown in Figure 4: node locality is slowly decreasing with node degree and only users with many connections have lower values of node locality. In FourSquare this is less evident, as the trend is fairly constant. Since LiveJournal and Twitter are modeled as directed graphs we investigate the same correlation in both directions. In LiveJournal the decreasing trend is evident for both in- and out-locality. Instead, Twitter users show a maximum value of out-locality as their number of outgoing links grows larger than 100, whereas in-locality shows a maximum just before 100 incoming connections. Both relationships then decrease until they reach a plateau.

While it is expected that nodes with larger degrees exhibit smaller locality values, since it is statistically more likely that they are connected to distant users, this behavior is not observed in Twitter: users with about 10 outgoing connections have lower values of out-locality, but as the out-degree grows there is a maximum at 100.

One possible explanation is that users with a small number of links are probably mainly connected to popular accounts, i.e., institutions, media and commercial entities, that are usually not geographically close to them. Indeed, when users join Twitter, the website suggests a list of 20 popular people and organizations which are unlikely to be located close to the joining user. As a consequence, people that just join the service and abandon it after a short period of time end up with a small number of connections which are not close from both social and geographic point of view.

Geographic clustering coefficient The other geo-social metric that we have studied is the geographic clustering coefficient. Since social networks are widely known to be characterized by the presence of triangles,



(a) BrightKite

(b) FourSquare

(c) LiveJournal

(d) Twitter

Figure 6: Average geographic clustering coefficient as a function of node degree. For directed networks the relationship is shown both for in- and out-degree.

the aim of this metric is to understand whether triplets of mutually connected users are more likely to be geographically close or, instead, distant from each other. Thus, a user with high geographic clustering coefficient has neighbors which are tightly interconnected and close to the user itself and to each other. The four datasets exhibit different values of geographic clustering coefficient: while BrightKite has an average value of 0.165 and FourSquare of 0.237, the average of LiveJournal is 0.146 and Twitter scores 0.108. Also, the first two datasets exhibit a geographic clustering coefficient which is closer to their standard clustering coefficient, while LiveJournal and Twitter present lower values when geographic distance is taken into account. Thus, in the former two OSNs clusters form at shorter distances than in the latter.

The probability distributions of the geographic clustering coefficient are shown in Figure 5. The higher mean values of BrightKite and FourSquare are explained by the fact that a non-negligible portion of users have a coefficient of 1.0: about 20% in BrightKite and about 10% in FourSquare. On the other hand, in LiveJournal and Twitter higher values are less likely to be observed and there is no discernible proportion of users with a coefficient of 1. These results show how location-based OSNs such as FourSquare and BrightKite tend to have more geographically confined triangles than social networks more focused on content production and sharing such as LiveJournal and Twitter.

Geographic clustering coefficient and node degree We now investigate how the geographic clustering coefficient is related to node degree. As reported in Figure 6, in BrightKite, FourSquare and LiveJournal the geographic clustering coefficient steadily decreases as the number of neighbors grows: thus, if a user has only few friends they are more likely to create connections with people nearby. On the other hand, Twitter shows a different behavior: the geographic clustering coefficient is slowly decreasing as the degree increases, but then it grows again until reaching a local maximum around the value of 1,000, while it decreases again for larger degrees.

This particular property of the Twitter network may be explained by the existence of users which are popular only in a particular region: they have both incoming and outgoing links with a large audience which has, however, several interconnections on a confined scale. Indeed, a user which is locally popular might have lower values of node locality because of his/her large audience, as shown in Figure 4, but users which are following him/her are also likely to share the same interests (since they follow the same popular user) and to become connected with each other. Instead, when a user reaches a wider popularity, his/her followers will be both more geographically spread and less interconnected. Further investigation on this point may unravel interesting findings.

5.2.3 Discussion

We have found common characteristics of these networks, namely the existence of a small-world effect, heavy-tailed degree distributions, a tendency to exhibit users with high node locality and social triangles on a local geographic scale.

However, from a geo-social point of view we have also seen some differences across the OSNs. Whereas purely location-based social networking services such BrightKite and FourSquare, which mainly focus on the geographic dimension of social interaction, have high node locality and geographic clustering close to standard clustering, OSNs based more on the idea of sharing information and content result in users with lower node locality and geographic clustering coefficient values.

It is important to note that the standard clustering coefficient is not affected by this distinction: indeed, BrightKite has the lowest clustering coefficient among all the datasets but the second largest geographic clustering coefficient, whereas Twitter has the second largest standard clustering coefficient but the lowest geographic one. Thus, it seems that geographic distance influences only the geographic properties of the triangles, not their likelihood of appearance. Indeed, Twitter and LiveJournal show a larger difference between standard and geographic clustering coefficients, while FourSquare and BrightKite present closer values. This result indicates how taking into account geographic distance in these metrics provides insightful information for the design of systems and applications that could potentially exploit the underlying geographic and social structure of OSNs.

6 Related Work

The effect of geography over complex networks has been studied mainly in communication and transportation networks [7]. For instance, it has been found that the spatial properties of the Internet topology are mainly determined by both preferential attachment and linear distance dependence [21], whereas Internet traffic is spatially bound to a spanning network which connects the most important centers around the globe [1]. However, these works do not investigate social networks, which do not have any spatial constraint.

Nonetheless, not many studies have addressed how geographic distance affects social interactions. While it seems still true that physically proximity fosters social interaction even on online channels as e-mail or instant-messaging [16], the rise of social networking services has been claimed to have caused the death of distance among social relationships. Certainly, mainly because of the latest technological changes, social communities have become ”glocalized” [20], with both extensive local links and significant long-distance relationships.

One of the first attempts to analyze how interactions on OSNs are affected by spatial distance is presented in [14], where the authors show that the probability of friendship decreases not only with distance but more precisely with the number of closer people. Some studies have investigated the structural properties of a location-based OSN and how social and geographic distance influences the creation of new connections among its users [17]. Instead, we describe novel geo-social metrics which offer a new perspective over these systems by combining both social structure with geographic distance.

7 Conclusion and Future Directions

In this paper we have presented a study of four geographical online social networks through the use of novel graph metrics able to capture geo-social relationships. A number of potential applications spark from the results presented in this work, including the ability of producing effective targeted advertising, efficient content placement and caching, and faster and more relevant information diffusion. We plan to investigate each of these applications in the near future. As location-aware OSNs become more and more popular, there will be even more data available about how people move and interact. Hence, we plan to extend this work by taking into account how users change their location over time and how this affects their behavior.

Acknowledgements

This work was supported through EPSRC grants EP/C544773, EP/E012914 and EP/D077273.

References

[1] Barthélemy, M., Gondran, B., and Guichard, E. Spatial structure of the internet traffic. *Physica A: Statistical Mechanics and its Applications 319* (March 2003), 633–642.

[2] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. U. Complex Networks: Structure and Dynamics. *Physics Reports 424*, 4-5 (February 2006), 175–308.

[3] Castellano, C., Fortunato, S., and Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys. 81*, 2 (2009).

[4] Facebook. Statistics. `http://www.facebook.com/press/info.php?statistics`.

[5] FourSquare. Foursquare. everywhere. `http://bit.ly/coJPSY`.

[6] Garlaschelli, D., and Loffredo, M. I. Patterns of link reciprocity in directed networks. *Phys. Rev. Lett. 93*, 26 (2004).

[7] Gastner, M. T., and Newman. The spatial structure of networks. *The European Physical Journal B - Condensed Matter and Complex Systems 49*, 2 (January 2006), 247–252.

[8] Hayes, B. Cloud computing. *Commun. ACM 51* (July 2008).

[9] Humphreys, L. Mobile Social Networks and Social Practice: A Case Study of Dodgeball. *Journal of Computer-Mediated Communication 13*, 1 (2007).

[10] Krishnamurthy, B., Gill, P., and Arlitt, M. A Few Chirps About Twitter. In *Proceedings of WOSN '08* (New York, NY, USA, 2008), ACM, pp. 19–24.

[11] Kumar, R., Novak, J., and Tomkins, A. Structure and Evolution of Online Social Networks. In *Proceedings of KDD '06* (New York, NY, USA, 2006), ACM, pp. 611–617.

[12] Lee, S. H., Kim, P. J., and Jeong, H. Statistical properties of sampled networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 73* (2006), 016102.

[13] Leighton, T. Improving Performance on the Internet. *Communications of the ACM 52* (February 2009), 44–51.

[14] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. *Proceedings of the National Academy of Sciences, USA*.

[15] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. Measurement and Analysis of Online Social Networks. In *Proceedings of IMC '07* (New York, NY, USA, 2007), ACM, pp. 29–42.

[16] Mok, D., Wellman, B., and Carrasco, J. A. Does Distance Still Matter in the Age of the Internet? *Urban Studies 46* (2009).

[17] Nan, L., and Guanling, C. Analysis of a location-based social network. In *IEEE International Conference on Computational Science and Engineering* (Los Alamitos, CA, USA, 2009), vol. 4, IEEE Computer Society, pp. 263–270.

[18] RJMetrics. New Data on Twitter Users and Engagement. `http://bit.ly/9JSNCf`.

[19] Wasserman, S., and Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[20] Wellman, B., and Hampton, K. Living Networked On and Offline. *Contemporary Sociology 28*, 6 (1999), 648–654.

[21] Yook, S.-H., Jeong, H., and Barabási, A.-L. Modeling the Internet's large-scale topology. *PNAS 99*, 21 (2002), 13382–13386.