Spatial Dissemination Metrics for Location-Based Social Networks

Antonio Lima
School of Computer Science
University of Birmingham
United Kingdom
a.lima@cs.bham.ac.uk

Mirco Musolesi
School of Computer Science
University of Birmingham
United Kingdom
m.musolesi@cs.bham.ac.uk

ABSTRACT

The importance of spatial information in Online Social Networks is increasing at a fast pace. The number of users regularly accessing services from their phones is rising and, therefore, local information is becoming more and more important, for example in targeted marketing and personalized services. In particular, news, from gossips to security alerts, are daily spread across cities through social networks. Content produced by users is consumed by their friends or followers, whose locations can be known or inferred. The spatial location of users' social connections strongly affects the areas where such information will be disseminated. As a consequence, some users can deliver content to a certain geographic area more easily and efficiently than others, for example because they have a larger number of friends in that area.

In this paper we present a set of metrics that quantitatively capture the effects of social links on the spreading of information in a given area. We discuss possible application scenarios and we present an initial critical evaluation by means of two datasets from Twitter and Foursquare by discussing a series of case studies.

1. INTRODUCTION

Location information is assuming increasing importance in Online Social Networks (OSNs). In particular, a very large number of users is now accessing these services using mobile devices [20, 14]. Certain social networks, such as Foursquare, are built around the very same concept of location [23]. Geo-tagging of posts and photos is becoming popular in Facebook and geographic information is often provided in user profiles and in the generated contents in Twitter. Through these services, information is disseminated across cities, regions, states and the entire planet. Contents of different types are propagated and are consumed by millions of people dispersed around the globe. Understanding the dynamics of the dissemination process is critical for a variety of purposes and to answer a set of fundamental questions that might have important implications for the design of the location-based online network services themselves. For example, to what extent does the geographic distribution of friendships in the network affect where the content will be potentially propagated? Are we able to determine which users are structurally central in delivering information to a specific spatial region? By identifying these users it will be possible to exploit them to deliver information efficiently to specific regions.

General structural properties of large-scale OSNs have already been explored in great detail in the past (see for example [16]). More recently, several works have focussed on geo-social properties of OSNs, focussing for example on the correlation between geography and social topology [19]. Others have investigated co-location and friendship [2, 18] and the possibility of predicting location using friendship information [18]. Indeed, these networks can be considered as a particular class of spatial networks [4]. In the context of complex networks a significant effort has been made to try to give an answer to the question "Which are the most important (i.e., the most central) nodes in a network?". Finding an answer is important because it has strong implications on the processes taking places in networks, such as information diffusion in a social network. The problem has been answered by defining various centrality metrics [17]. All these centrality measures are defined in different ways, by taking into account only social ties (i.e., *topological* information). However, the problem of finding the most important nodes person with respect to the people that are in a specific location (i.e. by using the *spatial* information of the social links) remains open and largely unexplored.

In this paper, we propose information diffusion metrics that capture and quantify geographic importance and centrality of users in geo-social networks. We evaluate these metrics by associating users to one or more locations, using datasets extracted from Twitter and Foursquare. Our metrics focus on the structural properties of the geo-social networks and not on the processes happening over them, such as information cascading and retweeting. Moreover, by separating structure and dynamics, they can be used as quantitative generic tools for evaluating the *potential* role of each node in disseminating information in the geographic space.

The need for modeling spatial social networks and finding measures for quantifying geographic centrality and influence comes not only from the ambition to study the complex interactions between the social and spatial dimensions more comprehensively, but also from a variety of potential

practical applications which could benefit from this analysis. These include:

Targeted information spreading. Being able to measure geographic centrality allows us to rank users according to the number of contacts they have in a certain area. Consequently, they can be used to select individuals to be targeted for spreading information. Applications include not only support for advertisement campaigns of certain products or promotions restricted to given areas, but also the design of systems for dissemination of emergency alerts in natural or man-made disaster situations, where information should be disseminated in a spatially-limited area (for example in case of security alerts in parts of a city or for weather alerts in a certain region).

Models of cultural influence. OSNs are an invaluable source of data for studies in social sciences that were simply not possible in the past [10, 12]. In particular, estimating social and political influence can be very important and relevant for analyzing and interpreting several cultural phenomena. For example, a person tweeting in London might have influence also outside it, for example in his/her hometowns, and in case of recent immigrants, in his/her country of origin. Other possible fields include health studies [7] and economics [21]: until now research in these fields has focussed mainly on the structure of the social networks without considering geographic aspects.

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

- Starting from some well-known metrics of centrality and clustering in location-agnostic networks, we define new measures of centrality for quantifying spatial influence, spatial closeness, and spatial efficiency for geo-social networks. We also propose a definition of spatial local clustering coefficient to quantify the presence of *social triangles* in a given location.

- We present a preliminary evaluation of the effectiveness of these metrics by means of two datasets obtained from real world OSNs, namely Twitter and Foursquare, and we discuss the application of these metrics to some realistic application scenarios.

This paper is organized as follows: in Section 2 we introduce the influence metrics; then, in Section 3, we evaluate these metrics by means of the two datasets. We discuss the potential use of these geo-structural metrics for studying dynamic processes in Section 4. Finally, in Section 5 we conclude the paper by discussing future work.

2. SPATIAL INFORMATION DISSEMINATION METRICS

We can represent a social network as a graph $\mathcal{G} = (V, E)$ with N nodes and K links, where nodes are users and links are the social connections between them[1]. We define a *spatial social network* as a social network where each user i is assigned a set of n_i points on Earth $\mathcal{P}_i = \{p_0^{(i)}, p_1^{(i)}, \dots p_{n_i}^{(i)}\}$ including locations that are significant for him/her (e.g., hometown, workplace, favorite restaurant, etc.)[2]. We will firstly introduce a set of accessory definitions that will be used in the remainder of this paper.

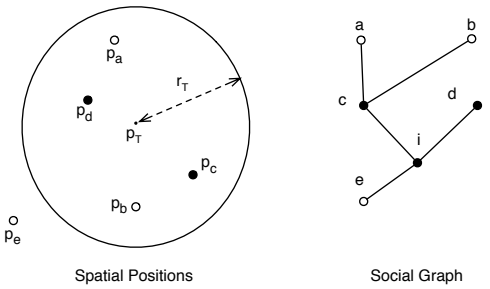


Figure 1: Example of social graph (on the right) and spatial dimension (on the left). In this example the social neighborhood $\mathcal{N}_i$ of i is composed by the nodes c, d, and e; the points of interest of a, b, c and d are inside the spatial neighborhood $\mathcal{S}$, which is the circle of center p_T and radius r_T. As a consequence, the socio-spatial neighborhood $\mathcal{N}_{i,\mathcal{S}}$ indicated with full black dots, does not include node e because it falls outside the spatial neighborhood and also excludes nodes a and b because their positions are located outside the social neighborhood.

- As far the social graph is concerned, we define the neighbors (or connections) of node i as the set of nodes j that are reachable from i through the out-link e_{ij} (content can flow from i to j). The *social neighborhood* $\mathcal{N}_i$ of a node i is the set of all the k_i neighboring nodes of i (e.g., all the followers of user i in Twitter); k_i is often referred to as the degree of node i. This set is defined only on social ties and does not take into consideration any geographic information.

- As far as the spatial dimension is concerned, we use the notation $d_G(p_1, p_2)$ to indicate the geodesic distance between two points on Earth p_1 and p_2. We then define the *spatial neighborhood* $\mathcal{S}$ as an arbitrarily shaped part of the geographic surface; this is a continuous set of geographic points. For simplicity, in this work we will often consider circular regions specified by their center and radius but the definitions presented here can be applied to regions of any shape.

- Given a node j and a geographic region $\mathcal{S}$, the intersection $\mathcal{P}_j \cap \mathcal{S}$ contains all the significant points of j falling inside the region. We define the *socio-spatial neighborhood* $\mathcal{N}_{i,\mathcal{S}}$ of the node i with respect to $\mathcal{S}$ as the set of neighbors j who have at least one significant point inside $\mathcal{S}$:

$$\mathcal{N}_{i,\mathcal{S}} = \{j \in \mathcal{N}_i : \mathcal{P}_j \cap \mathcal{S} \neq \emptyset\}. \quad (1)$$

  With $k_{i,\mathcal{S}}$ we denote the number of users in this set. An example is presented in Figure 1.

[1]This representation can be considered as a snapshot of the graph at a given time t. A treatment considering the time-varying nature of the social graphs is outside the scope of this work.

[2]In the simpler case each user can be assigned a single significant location. In the evaluation section we will present two examples covering both cases.

2.1 Spatial Degree Centrality

In general, in a social graph degree centrality is used to rank users according to the number of ties they have within the network [17]; its value is a simple indicator of *influence* and prestige [22]. Methods based on degree centrality are generally used to select the best nodes for spreading information [9]. We extend the concept of degree centrality to spatial social networks with respect to a given spatial neighborhood $\mathcal{S}$ by introducing the concept of *spatial degree centrality*:

$$C_{i,\mathcal{S}} = \sum_{j \in \mathcal{N}_i} |\mathcal{P}_j \cap \mathcal{S}| \,. \tag{2}$$

This value indicates how many significant points the social neighborhood of user i has got inside the considered spatial neighborhood $\mathcal{S}$. If every user is associated only one significant point, this value indicates the size of the audience of user i in the region. In the general case of many significant points for each user, this also takes into account the strength of the potential audience in the region (i.e. social connections with many significant places inside the region give a larger contribution than those with fewer).

The size of the considered region $\mathcal{S}$ affects the calculation of the values of the metrics. For this reason, the size should be set according to the characteristics of the dataset (measurement granularity and precision) and the goal of the analysis itself (for example, researchers might be interested in an analysis at city level). Since the degree of each node also affects this value, a normalization of this metric might also be necessary. The normalization is particularly convenient when comparing users who have a number of followers that differs by orders of magnitude. This might be the case that happens when comparing accounts of news agencies and celebrities, often followed by hundreds of thousands users, with users who have dozens or hundreds of followers. We call the normalized version *spatial degree ratio*, formally defined as:

$$\rho_{i,\mathcal{S}} = \frac{1}{\sum_{j \in \mathcal{N}_i} n_i} C_{i,\mathcal{S}} \tag{3}$$

where n_i is the number of significant places of the user i; this is equivalent, for the one-place case, to:

$$\rho_{i,\mathcal{S}} = \frac{1}{k_i} C_{i,\mathcal{S}}. \tag{4}$$

This metric has values in the range $[0, 1]$. It represents the ratio of connections of i that are inside the area $\mathcal{S}$, therefore it allows to compare nodes that have different degrees in the graph.

These centralities might be considered as simple measures of spatial influence, which can be used in the selection of a user for spreading information to a certain geographic region. However, as they are based on the concepts of geographic membership and social membership, they might not be entirely sufficient to describe the geographic distribution of the neighbors of users. For this reason, in the next subsection we will introduce metrics that take also into account geographic distances.

2.2 Spatial Closeness Centrality

We have defined spatial degree centrality relating to a region. Now we will define a measure of centrality concerning a *punctual* location. Given a target point $p^\star$ on Earth, we define the *spatial closeness centrality* for a user i towards this point as its average geographic distance from all the significant places of his/her connections, formally:

$$C^C_{i,p^\star} = \frac{1}{\sum_{j \in \mathcal{N}_i} n_i} \sum_{j \in \mathcal{N}_i} d_G(p_j, p^\star). \tag{5}$$

This definition is an indicator of how the influenced audience of a user is geographically close to the target point. It can be considered as the spatial counterpart of closeness centrality, which for complex networks is defined as the average distance of the shortest path from the node to all the other nodes [17] it is used as a heuristic when selecting nodes in information diffusion processes [9]. However, this metric might have some drawbacks in specific scenarios given the fact it is calculated as an average of all the distances. This metric can be generalized to the case of multiple locations.

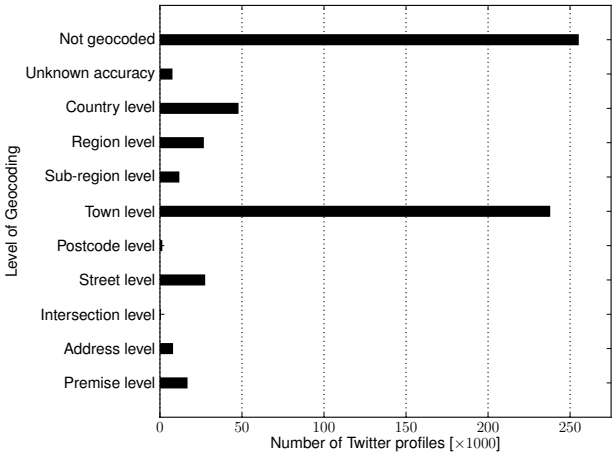


Figure 2: Accuracy of geocoding for the Twitter dataset.

2.3 Spatial Efficiency

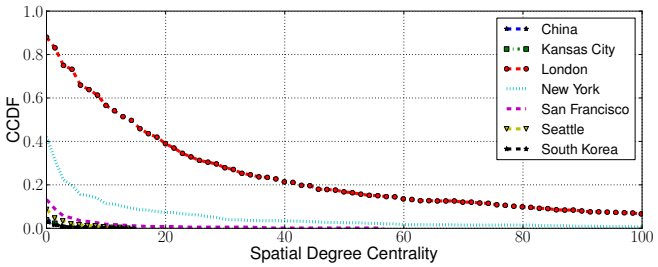
In order to deal with the problem of very large distances which might skew the value of spatial closeness centrality, we define *spatial efficiency* of user i with respect to a point $p^\star$ as follows:

$$C^E_{i,p^\star} = \frac{1}{k_i} \sum_{j \in \mathcal{N}_i} \frac{1}{d_G(p_j, p^\star)}. \tag{6}$$

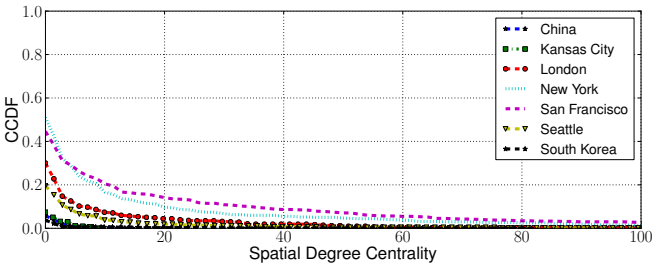
This measure can be thought of as a spatial version of efficiency of traditional graphs [11]. However, this definition has also a potential drawback: if the the neighbor location p_j coincides with $p^\star$ this formula is not defined. For this reason, we modify the above formula by introducing a smoothing decay term as follows:

$$C^E_i(p) = \frac{1}{k_i} \sum_{j \in \mathcal{N}_i} e^{-d_G(p_j, p^\star)/\gamma)} \tag{7}$$

where γ is a scaling factor that can be used to give different weights to the distance $d_G(p_j, p^\star)$. In this formula, the contribution for every neighbor j is at most 1. It is equal to 1 if the neighbor location p_j coincides with $p^\star$, whereas it is negligible if the point is very distant (asymptotically zero if the distance is infinite). This definition can be generalized to multiple locations in a similar way to the formulae presented above.

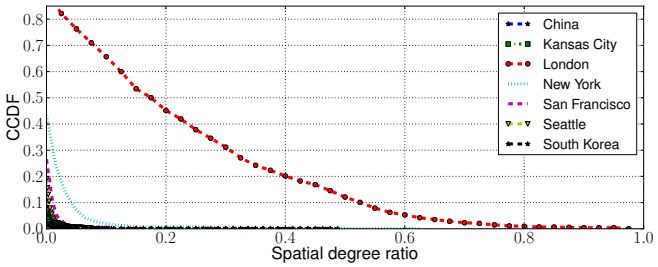


(a) Spatial degree centrality from London.

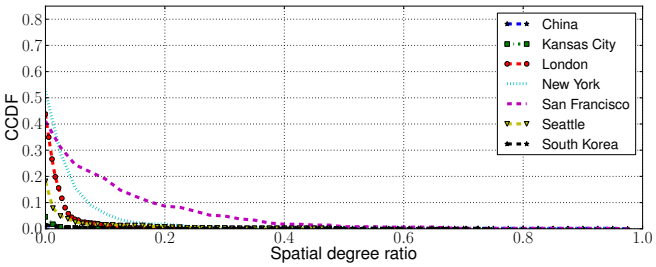


(b) Spatial degree centrality from San Francisco.

Figure 3: Spatial degree centrality of Twitter users in London and San Francisco.



(a) Spatial degree ratio from London.



(b) Spatial degree ratio from San Francisco.

Figure 4: Spatial degree ratio of Twitter users in London and San Francisco.

2.4 Local Spatial Clustering Coefficient

All the definitions presented up to here concern links between pairs of nodes. An interesting measure often used in social network analysis, which deals with triplets of people, is the local clustering coefficient, also called transitivity [22]. It is a local measure quantifying the fraction of triangles among a node and its neighbors. Its spatial version, the geographic clustering coefficient, weights every triangle depending on the geographic distances between the nodes of the triangle [19]. However, this geographic version does not give insights on how neighbors of neighbors might interact in a *specific* geographic region. For this reason, we define the *local spatial clustering coefficient* as the number of triangles present in the socio-spatial neighborhood taken into analysis, formally:

$$C_{i,\mathcal{S}} = \frac{|\{e_{jk} \in E : j,k \in \mathcal{N}_{i,\mathcal{S}}\}|}{k_{i,\mathcal{S}}(k_{i,\mathcal{S}}-1)} \quad (8)$$

where the numerator counts how many links in the social graph are present between users in the socio-spatial neighborhood and the denominator counts how many there could be at most, if they were all connected between each other. The local clustering coefficient measures to which extent neighbors of a node are connected to each other. This metric acquires a special meaning in its spatial version. Nodes scoring high values are part of "social circles", making them potentially highly influential[3]. Social circles defined in this way can be considered a simple example of spatial network motifs, i.e., patterns of interactions on which the network is built [15]. The investigation of the role of spatial network motifs in information dissemination is outside the scope of this work.

[3]At the same time, it is worth noting that, according to some existing theories such as Burt's structural holes [5], nodes scoring low values might also be considered very influential but in a different way as they are able to bring information to users who are not connected between each other, therefore controlling information flows for these users.

3. EVALUATION

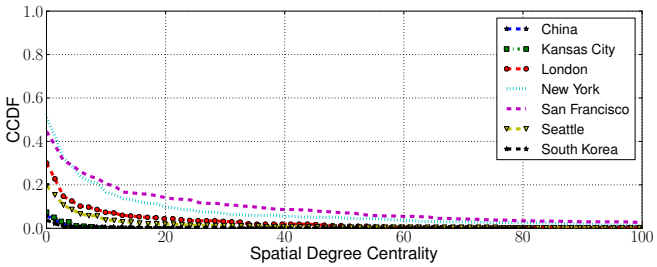
In this section we provide a preliminary evaluation of the proposed metrics. We first present the datasets and we analyze the results deriving from the application of the metrics to different case studies.

3.1 Description of the Datasets

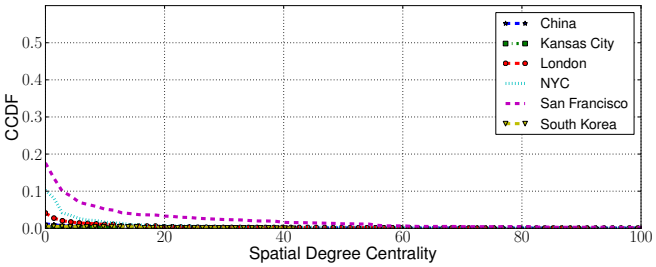
In order to evaluate our metrics, we analyze two popular real-world OSNs, Twitter and Foursquare. In general, datasets were acquired using 2-hop snowball sampling, seeded with random users chosen in well-defined geographic areas. Due to different properties of the two social networking services taken into consideration, the two datasets were obtained following different methodologies, as explained below.

With respect to Twitter, we crawled a dataset containing information about 657,777 users, starting from two evenly distributed sets of 1375 seed users. These were chosen randomly among users that were tweeting from two urban areas, London, UK and San Francisco, California[4]. This location bias was necessary given the nature of our investigation, which requires to have a statistically significant sample of users in the area. It is also worth noting that this can be

[4]The locations for the "seeds" were retrieved from geotags, i.e., spatial tags which are associated to tweets either by automatic geographic sensors as GPS or manually by the user.

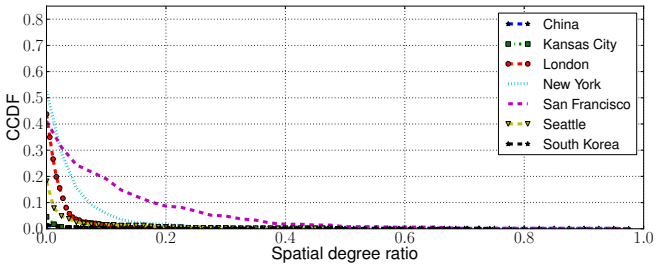


(a) Spatial degree centrality from London.

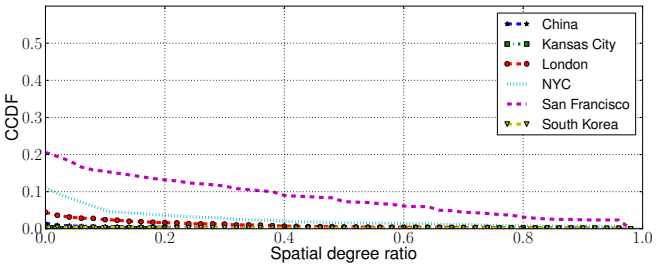


(b) Spatial degree centrality from San Francisco.

Figure 5: Spatial degree centrality of Foursquare users in London and San Francisco.



(a) Spatial degree ratio from London.



(b) Spatial degree ratio from San Francisco.

Figure 6: Spatial degree ratio of Foursquare users in London and San Francisco.

considered as a practical way of retrieving these users for a potential deployment of the algorithms for the calculation of the proposed metrics. We assigned a single significant place to each user, by fetching the information in the "location" field of their personal profile, and converting it to geographic coordinates through the Google Geocoding API. The geocoder was able to identify location for 378,829 users, with different levels of precision; the majority of the identified locations were at town level, according to the distribution shown in Fig. 2, similar to that shown in [8]. We are indeed aware of the fact that locations are not precise and the data are noisy[5].

In Foursquare, a location-based online social network, users "check-in" at venues to let their friends know about their whereabouts, to keep track of their habits and to explore places related to the interests they have in common with other people. The user with the highest number of check-ins over the last 60 days is called the "mayor" of the venue in Foursquare jargon. For this reason, mayorship provides information of potentially strong spatial significance of a certain place for that user. This is also a fine-grained information, as venues are commonly specified at premises level. For this reason, we used the collection of mayorships locations to build the set of significant places. We crawled a dataset of 177,809 users. Since the number of connections in Foursquare is typically smaller than the number of followers in Twitter and the former tend to link with spatially close people [19], our sampling strategy followed a different approach in order to avoid geographically sparse data. We selected a group of interesting urban areas and we crawled venues in the area using the Foursquare API. It is worth noting that these considerations are of great importance for a practical implementation of systems for calculating these metrics in (quasi) real-time, also considering the crawling limitations of the APIs[6].

Finally, we make the simplifying assumption that the rate of change of the network topology is negligible with respect to the information dissemination process taking place over it. This assumption seems reasonable in networks such as Twitter or Foursquare where the rate of change of links is usually very low at the scale of 1 day for example. In fact, the number of new added and removed followers and friends is quite low for a given user after an initial period where a large number of users is added.

3.2 Results

In this section we will present a selection of measurements for each metric. More specifically, in this preliminary study, we choose to compare areas that are heterogeneous from a cultural point of view and different in size.

We also consider two practical case studies. The first is related to the London riots that took place in August 2011: we measure the centrality of Londoners on Croydon, which was one of the theaters of the most violent acts in the British capital. This scenario is an example of usage of this technique in case of emergency. In other words, we are able to answer the following question: *what is the best set of people to target in order to have localized influence through social media in case of natural and man-made emergency and disasters?*

[5]The problem of dealing with noisy data is part of ongoing work.

[6]The Foursquare API returns at most 50 venues per call and does not allow to paginate over all venues in a given large area. Therefore, we queried for venues in categories in small-radius areas (i.e., with a 50 m radius) randomly selected inside the larger areas considered.

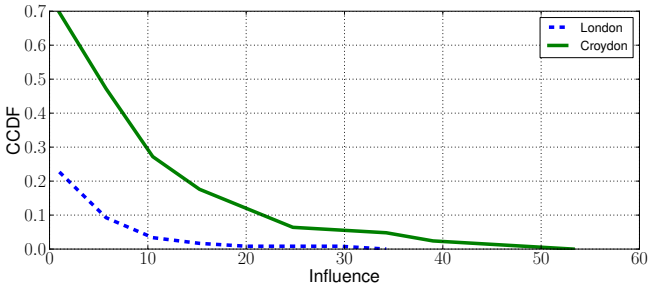


Figure 7: Spatial degree centrality of Foursquare users in Croydon and London towards Croydon.

The second consists in quantifying the centrality of both San Franciscans on people living in Chinatown, and people living in Chinatown on people living in China. It can be seen as an application to the area of geo-demographics [1], aimed at quantifying the potential cultural influence of the inhabitants of certain areas of the city over other areas.

Spatial Degree Centrality and Ratio In Figure 3 we report the Complementary Cumulative Distribution Function (CCDF) of spatial degree centrality of the users located in London and in San Francisco towards four cities (New York, Kansas City, London and Seattle) and two countries (China and South Korea) using the Twitter dataset. We selected the countries by considering the presence of a non-negligible percentage of their population belonging to these ethnic groups. It is possible to observe that both cities have a high degree centrality with respect to themselves, as expected. It is surprising, though, that the degree centrality of Londoners on themselves is very high; in comparison, San Franciscans are not significantly central with respect to their fellow-citizens, and have a self-centrality similar to the centrality shown towards New Yorkers. In our opinion, a possible cause might be that many people who spend most of the day in San Francisco (for example, because their workplace is based there), actually live in the neighboring areas and commute everyday. While 9 users out of 10 in London have at least 1 followers from their own city, only 1 user out of 2 in San Francisco has at least a fellow-citizen reading his content. San Franciscans have some limited potential influence on Londoners, though not as much as on New Yorkers. Users from London and Seattle are also potentially influenced in a substantial way, though not as much as New Yorkers. The countries, China and South Korea, score very low centrality measures in both scenarios, and their curves overlap with those related to Kansas City, the city on which both Londoners and San Franciscans influence the least.

It is worth noting that these results could be influenced by a culture-related tendency to include location information: users from some locations might be keener to include the real personal location, compared to users from other places, due to a different sensibility about privacy issues. Unfortunately, we do not have hard evidence about this fact.

Similar observations can be made for the CCDF for Spatial Degree Ratio in Figure 4. We can notice how the high degree centrality of London with respect to itself is actually connected to a low spatial heterogeneity of followers: nearly one Londoners out of two has *at least* 20 followers living in the same city, while in San Francisco only one out of ten satisfies this property. This peculiar characteristic might be explained both with the tendency of Londoners to follow people from London and with a low interest shown by non-Londoners for the content shared by Londoners. We can also observe how the two highest curves of ratio show a more linear progress, compared to their spatial degree centrality counterparts.

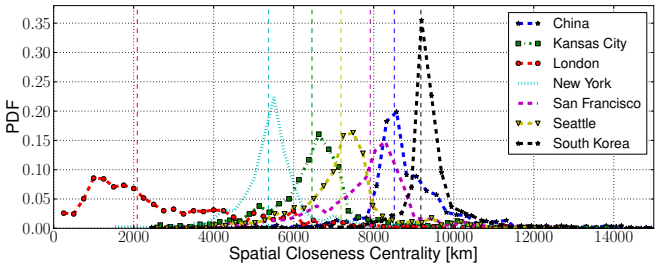
We also perform a similar analysis using the Foursquare dataset. The lower penetration of Foursquare leads to a lower average degree (i.e., on average users in Foursquare have a smaller number of connections than in Twitter) and consequently to smaller centrality values, which include many zeros. However, results shown in Fig. 5 and Fig. 6 are still in accordance with those observed for Twitter. Considering the characteristics of the users in the city, London is again a place of high centrality with respect to itself.

While this city-level analysis can be carried out on the Twitter dataset, we cannot use it for a meaningful analysis at a finer scale, given the nature and quality of the data. Therefore, we use instead the Foursquare dataset, in particular to study the potential influence of Chinatown towards San Francisco and China. The metric is able to identify 8% of users that have a non-null centrality on China and to *rank* them according to their centrality (which quantifies their potential influence over China). When analyzing the average values of the measure, it is interesting to note that the centrality of San Francisco towards Chinatown and the centrality of Chinatown towards itself are comparable (3.2 vs 3.06). This might support the hypothesis that the district is considerably influenced by people living in other parts of the city and that choosing to deliver information to people in Chinatown instead of San Francisco might not have a significant impact on how the information is spread in Chinatown itself. Moreover, given its history and ethnic composition it is not surprising to discover that the average centrality of Chinatown towards China is almost 3 times bigger than the average centrality of the city of San Francisco on China (32.24 vs 11.87).

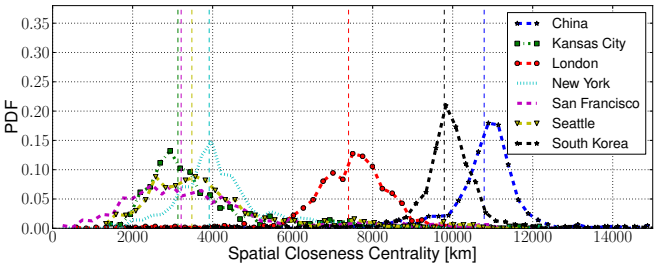
For the Croydon scenario, in Figure 7 we report the centrality of Croydon and London on the Croydon area itself. Users in Croydon appear to have substantially higher values of degree centrality than users of London. This suggests that when disseminating information, targeting people in the area of Croydon, instead of the whole London, might give an advantage in reaching the area of Croydon itself.

Spatial Closeness Centrality In Figure 8 we represent the probability distribution function for the seven distributions of spatial closeness centrality. For each curve, a dashed vertical line represents the median. We can firstly notice that for both cities taken into consideration, London and San Francisco, the closeness centrality curves are more spread out compared to the other curves, which are generally narrower and characterized by a series of peaks. London shows this behavior with stronger emphasis; this can be another evidence of the the high locality of London followers. By definition, geographic constraints have a strong impact on this metric; therefore, we would expect that the peak and the median are very close to the physical distance between the considered points. Indeed, this is the case for all the pairs we report in the figure.

Spatial Efficiency In order to characterize spatial efficiency, we set the value of γ equal to the maximum radius of the geographic area taken into consideration. Figure 9 shows the CCDF for the Twitter users in London and San

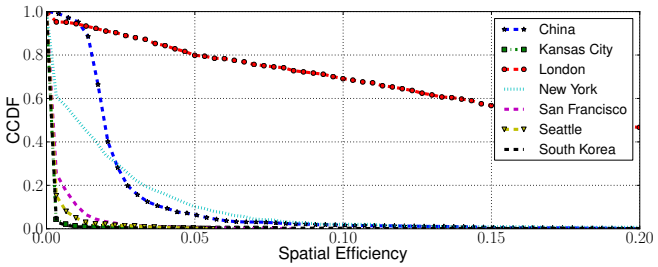


(a) Spatial closeness centrality from London.

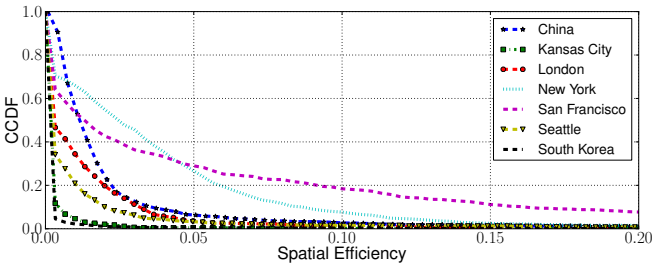


(b) Spatial closeness centrality from San Francisco.

Figure 8: Spatial closeness centrality of Twitter users in London and San Francisco.



(a) Spatial efficiency centrality from London.



(b) Spatial efficiency centrality from San Francisco.

Figure 9: Spatial efficiency centrality of Twitter users in London and San Francisco.

Francisco with respect to the areas considered for the other metrics. As this measure emphasizes the role of neighbors which are close to the location taken into consideration, we can see how the efficiency of London with respect to itself stands out from all the other curves.

Local Spatial Clustering Coefficient The local spatial clustering coefficient allows to identify how many "social triangles" are present in a specified area. As an example, we compute this metric for neighbors of Twitter users which indicated their location in the London area. The percentage of null values is quite high (88%) indicating that a small number of Londoners have social circles in their own city. In Figure 10 we show the CDF for the non-null values.

4. DISCUSSION

When defining metrics to determine how users are influential in a social network, or equivalently how central they are in the process of information diffusion, it seems natural to consider quantities related to the level of actual engagement of users (e.g., how many elements they share, how many reactions/retweets they receive in turn from their friends and so on) and about the semantics of the shared content (e.g., whether it is multimedia content, news links, games, etc.). Such measures can give information about the role of the user in the network and also about his/her topics of interests. For each topic he could either be a pure provider of content or a pure consumer of content or, as it happens more commonly, a combination of the two. The goal of this work is to explore spatio-social centralities relying only on structural properties of social networks, without considering data derived from processes taking place in the network, such as information diffusion [13].

This might be considered a limitation of the current metrics. However, our goal is to propose a generic set of metrics that can be used as the basis for an analysis of the dynamic processes happening over them [3].

Popular OSNs are used by millions of users and handle massive amounts of data. Given the fact that the proposed metrics are calculated on very large datasets, computational complexity is a key issue. First of all, we observe that the metrics we have defined are *local*: in order to perform the calculation we do not need global information about the entire graph. Given a specific location, described by a geographic point or surface, in order to determine the measures defined above for a set of n users, we need to know the coordinates of the neighbors' significant places. Spatial degree centrality, spatial closeness centrality and efficiency measures scale as $\mathcal{O}(nkt)$ where k is the expected number of neighbors of each node and t is the expected number of significant points for each neighbor. In order to determine local spatial clustering coefficient we also need to retrieve the neighbors of neighbors of the starting node (so that we can determine if two of his/her neighbors are neighbors in turn); the complexity of the calculation of this metric scales as $\mathcal{O}(nk^2t^2)$.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented metrics for quantifying potential information dissemination in social networks where geographic information is associated to each user. We have evaluated these metrics by means of two datasets extracted from Twitter and Foursquare by analyzing different realistic case studies, which might be relevant for emergency communications and social sciences. The applications of these metrics are many, including targeted location-aware mar-

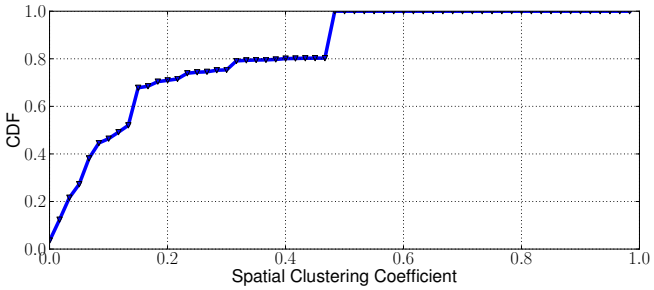


Figure 10: Spatial clustering coefficient of Twitter users in London towards London itself.

keting and efficient information spreading during emergency events.

We plan to extend this analysis by taking into consideration explicit actions (such as retweets or mentions, in Twitter [13, 6]). We also plan to explore the aspects related to the implementation of these algorithms to extract these indicators in real-time.

Acknowledgements

The authors would like to thank Manlio De Domenico for his insightful comments about an earlier draft of this paper. This work was supported through the EPSRC Grant "The Uncertainty of Identity: Linking Spatiotemporal Information Between Virtual and Real Worlds" (EP/J005266/1).

6. REFERENCES

[1] D. I. Ashby and P. A. Longley. Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS*, 9(1):53–72, 2005.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW'10*, pages 61–70, New York, NY, USA, 2010. ACM.

[3] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[4] M. Barthélemy. Spatial networks. *Physics Reports*, 499:1–101, 2011.

[5] R. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1994.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of ICWSM'10*. AAAI, 2010.

[7] N. A. Christakis and J. H. Fowler. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4):370–379, 2007.

[8] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of CHI'11*, pages 237–246, New York, NY, USA, 2011. ACM.

[9] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the Spread of Influence in a Social Network. In *Proceedings of KDD'03*, pages 137–146. ACM, 2003.

[10] J. Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51:66–72, Nov. 2008.

[11] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87:198701, Oct 2001.

[12] D. Lazer et al. Computational Social Science. *Science*, 323:721–723, February 2009.

[13] K. Lerman and R. Gosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of ICWSM'10*. AAAI, 2010.

[14] R. MacManus. Facebook mobile usage set to explode, October 2011. ReadWriteWeb.

[15] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[16] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of IMC'07*, pages 29–42, New York, NY, USA, 2007. ACM.

[17] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[18] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of WSDM'12*, pages 723–732, New York, NY, USA, 2012. ACM.

[19] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance Matters: Geo-social Metrics for Online Social Networks. In *Proceedings of WOSN'10*, Boston, MA, USA, June 2010.

[20] E. Schonfield. Meeker Says Majority of Pandora's and Twitter's Traffic is Mobile, October 2011. TechCrunch.

[21] O. Sorenson. Social networks and industrial geography. *Journal of Evolutionary Economics*, 13:513–527, 2003.

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

[23] Y. Zheng. Location-based social networks: Users. In *Computing with Spatial Trajectories*. Eds. Springer, 2011.