Evolution of a Location-based Online Social Network: Analysis and Models

Miltiadis Allamanis
Computer Laboratory
University of Cambridge
ma536@cam.ac.uk

Salvatore Scellato
Computer Laboratory
University of Cambridge
ss824@cam.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cm542@cam.ac.uk

ABSTRACT

Connections established by users of online social networks are influenced by mechanisms such as preferential attachment and triadic closure. Yet, recent research has found that geographic factors also constrain users: spatial proximity fosters the creation of online social ties. While the effect of space might need to be incorporated to these social mechanisms, it is not clear to which extent this is true and in which way this is best achieved.

To address these questions, we present a measurement study of the temporal evolution of an online location-based social network. We have collected longitudinal traces over 4 months, including information about when social links are created and which places are visited by users, as revealed by their mobile *check-ins*. Thanks to this fine-grained temporal information, we test and compare whether different probabilistic models can explain the observed data adopting an approach based on likelihood estimation, quantitatively comparing their statistical power to reproduce real events. We demonstrate that geographic distance plays an important role in the creation of new social connections: node degree and spatial distance can be combined in a *gravitational attachment* process that reproduces real traces. Instead, we find that links arising because of *triadic closure*, where users form new ties with friends of existing friends, and because of *common focus*, where connections arise among users visiting the same place, appear to be mainly driven by social factors.

We exploit our findings to describe a new model of network growth that combines spatial and social factors. We extensively evaluate our model and its variations, demonstrating that it is able to reproduce the social and spatial properties observed in our traces. Our results offer useful insights for systems that take advantage of the spatial properties of online social services.

IMC'12, October 29–30, 2012, Austin, Texas, USA.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Management**]: Database applications—*data mining*

Keywords

social network, graph evolution, gravity models

1. INTRODUCTION

Measurement studies of popular online social services have greatly improved the understanding of how users create social connections online [22]. Research efforts have taken advantage of the availability of large-scale datasets to study the temporal evolution of online social ties [15]. Models of network growth have been proposed to reproduce the global properties observed in real online social networks, such as power-law degree distributions and high clustering coefficient [17].

The fundamental importance of such models is due to the fact that they explain properties observed in measured traces in terms of the actions of individual users: for instance, mechanisms such as preferential attachment [2] and triadic closure [14] are inspired by the actions of individuals creating their social connections. Thus, by offering insights about how users behave, measurements and models of network evolution provide practical applications to link prediction systems [19, 21], but also suggest solutions to large-scale engineering problems faced by online service providers [24]. However, researchers have often neglected factors that are not inherently connected to the structure of the social network itself. In this work we aim to fill this gap, studying the influence of *spatial factors* on connections created by users of a location-based social service.

Spatial properties of social services. Recently, online social networks have integrated location-based features: services such as Foursquare and Gowalla have pioneered the idea of sharing one's geographic location with friends, attracting millions of users over a short period of time. These services offer an additional source of information about user behavior: the geographic mobility of individuals.

Recent works have taken advantage of this opportunity to shed some light on the relationship between spatial factors and online social interactions. For instance, the probability of seeing a social connection between users of online social services decreases with spatial distance [20, 1]. Similar but quantitatively different spatial constraints have been also found in mobile phone communication networks [16, 23].

Online social ties can also be inferred by mining geographic coincidences [7], suggesting that spatial encounters have an effect on the creation of new online connections. The places that online users visit offer even more accurate predictive power about future social connections [8, 27].

The importance of space. The importance of space for online social network goes beyond the definition of more accurate models. In fact, a better understanding of the spatial aspects of the evolution of social networks would greatly benefit engineering approaches based on the spatial constraints of online social ties.

For instance, online interactions tend to be spatially clustered: this geographic locality of interest has been exploited in Facebook interactions, to improve service responsiveness with distributed proxies [31], and in a company's email network, to partition email traffic across storage locations [13]. Spatial differences in content requests arising from online social sharing have been used to reduce latency and bandwidth costs associated to content delivery [29]. The spatial patterns observed in Facebook social connections have been exploited to predict the geographic location of users given their friends' locations [1]. Such engineering efforts confirm that understanding the effect that space has on online social services remains of crucial importance for modern large-scale online platforms.

However, the effect of space on the mechanisms that drive how online users create their links is still largely unknown. In more detail, there has been no investigation of the spatial aspects of the temporal evolution of an online social network: a better understanding of these mechanisms would pave the way to predictive models including geographic factors.

Our work. In this work we study the temporal evolution of a location-based social network: over a period of 16 weeks we have collected daily snapshots of a location-based social network with hundred of thousands of users, Gowalla, including the places visited by users and their social connections (Section 2). Thanks to this fine-grained temporal information about network evolution and users' mobility, we test and compare different edge attachment models that can explain the observed data, adopting an approach based on *likelihood estimation* [32]. This allows us to quantitatively compare these models according to their statistical ability to reproduce the real traces.

In more detail, we analyze these core facets of temporal network evolution:

- **how edges are created**: we test different edge attachment models based on the social and spatial properties of nodes: we show that node degree and spatial distance are simultaneously influencing edge creation, demonstrating that *a gravitational attachment model captures real network evolution better than purely social or spatial models* (Section 3);

- **how social triangles are created**: since social networks tend to have a dominant fraction of new edges closing triangles, we test several different models of triadic closure, some of them involving also spatial distance: we find that *social factors are more important than spatial constraints when an edge closes a triangle* (Section 4);

- **how users' mobility affects new edges**: because social connections might arise among users visiting the same place, we study models of edge creation that exploit the properties of shared places to connect users; we discover that both *the popularity of a place and the popularity of users visiting that place* help to predict which social connections are established (Section 5);

In addition, we study the temporal patterns of user behavior. We focus on the lifetime of a node, that is, the amount of time a user is actively creating new edges, and the inter-edge waiting time, which governs the amount of time elapsed before a node will create a new edge (Section 6).

Based on our findings, *we describe a new family of models of network growth which are able to reproduce both the social and spatial properties observed in the real data* (Section 7). Such models combine a *global gravitational attachment* process with a *local triadic closure* mechanism based on shared friends and shared places; the result is an evolutionary random process that grows a spatial network edge-by-edge. We demonstrate that the resulting synthetic networks exhibit social and spatial properties similar to the real network, while a similar model that considers preferential rather than gravitational attachment, effectively ignoring the effect of geographic distance, fails to reproduce real properties.

This work sheds light on the effect of geographic constraints on the evolution of online social networks. Our results offer useful insights for researchers and practitioners, with promising implications for the wide range of applications that already take advantage of the spatial properties of online social services.

2. MEASUREMENT METHODOLOGY

In this section we illustrate the measurement methodology used in our work to acquire data on a large-scale location-based service, Gowalla. We describe this service, our data collection procedure and we present the basic properties of the resulting dataset. We also introduce the likelihood estimation technique we adopt to quantify how edge attachment models explain the real traces.

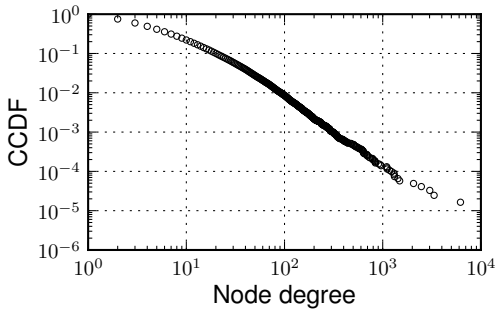
2.1 Gowalla

Gowalla is a location-based social networking service created in 2009 that allows users to add friends and share their location with them. It allows users to "check-in" at places through a dedicated mobile application, publicly disclosing their location on the service. These check-ins can then be pushed to friends. As a consequence, friends can see where a user is or has been. Users can create mutual friendship relationships, requiring each user to accept friendship requests. Gowalla was discontinued at the end of 2011 as the company was acquired by Facebook.

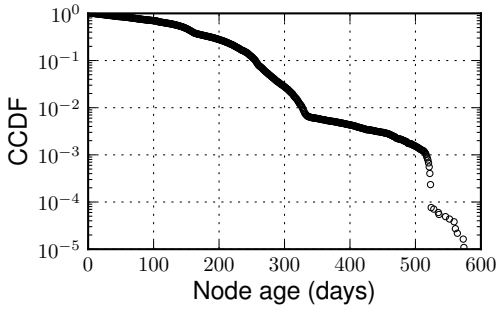
2.2 Data collection

Using the public API provided by Gowalla to allow other applications and services to access their content, we have downloaded daily snapshots of Gowalla data between May, 4^{th} and August, 19^{th} 2010.

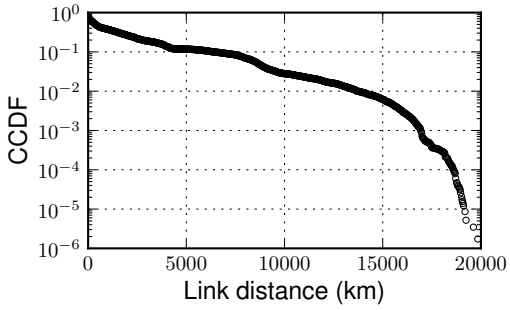
We built a multi-threaded self-limiting crawler to access their API without incurring into rate limitation. Since users were identified by consecutive numeric IDs, each day we were able to exhaustively gather profiles of all the registered user accounts on the service. Each user profile included information about the number of social connections of that user and the timestamp and the place of his/her last check-in. This



(a) Node degree



(b) Node age



(c) Link length

Figure 1: Complementary Cumulative Distribution Function (CCDF) of node degree (a), node age (b) and link geographic length (c) at the end of the measurement period.

allowed us to additionally download the friend list or the timestamped list of check-ins for all those users which either had new friends or new check-ins with respect to the previous day; we did the same for all the new users that were not registered before. As a result, we have a sequence of daily snapshots, each one including social connections and check-ins for each registered user. We also acquired the geographic coordinates of each place where users had checked in.

This dataset represents a sequence of *complete* snapshots of a large-scale location-based service, offering a chance to study how a social network grows over time and also over space. In particular, we have temporal information about all the social links created during our measurement process. This will allow us to study which social and spatial factors influence how links are created at the microscopic level. Finally, even though we have no temporal information about social connections created before our measurement started,

N	122,030
K	577,014
$\langle k \rangle$	9.28
$\langle C \rangle$	0.254
N_{GC}	116,910 (95.8%)
D_{EFF}	5.43
$\langle l \rangle$	1,792 km
$\langle D \rangle$	5,479 km

Table 1: Properties of the spatial social network at the end of the measurement period: number of nodes N, number of edges K, average node degree $\langle k \rangle$, average clustering coefficient $\langle C \rangle$, number of nodes in the giant component N_{GC}, 90-percentile effective network diameter D_{EFF} average geographic distance between nodes $\langle D \rangle$, average link length $\langle l \rangle$.

we have a reasonable estimation of the first time each user joined the network by observing the list of previous check-ins.

At the end of our measurement period we find about 400,000 registered users, with a total of more than 10 million check-ins across about 1,400,000 distinct places. More precisely, there are only 183,709 users with at least one check-in and only 162,239 with at least one friend. We focus our analysis on 122,030 users that have both friends and check-ins.

2.3 Notation

Formally, we represent the social network of Gowalla users as an undirected graph. We denote by N and K the total number of nodes and edges, while $G_t = (N_t, K_t)$ is the graph composed of the earliest t edges $(e_1, \ldots, e_t)$, with G_T being the final network at the end of the measurement process. The time when edge e was created is $t(e)$ and $t(u)$ is the time when node u joined the network. The degree of node u at time t is $k_u(t)$, while the number of nodes with degree k at time t is denoted as $n_k(t)$.

Every node of this network is embedded in a metric space: in our case, the metric space is the 2-dimensional surface of the planet and we adopt the great-circle distance over the Earth as distance metric. Rank-distance has been suggested as an alternative density-aware distance measure [20]: however, in our study the growing social graph would cause the measure to change as new nodes are added. Thus, we choose to adopt the simpler great-circle distance. We define the location of each user as the geographic location of the place where he/she has more check-ins overall at the end of the measurement period. We denote by D_{ij} the distance between nodes i and j. We assign a length l_{ij} to each social link so that $l_{ij} = D_{ij}$: since node positions do not change over time, link lengths and distances between nodes do not change either.

2.4 Basic properties

The number of nodes and the number of links grows approximately linearly over time, with the number of links growing at a faster pace. on average, the network gains about 375 new nodes and about 1,900 new edges per day. As a result, the average degree of the social network slowly increases: we find that a relationship $K_t \propto N_t^{\rho}$ holds with an exponent $\rho \approx 1.11$, which denotes superlinear growth of the edges with respect to the number of nodes, a sign of densification of the network as time passes by [18]. The main

properties of the spatial social network at the end of the measurement period are reported in Table 1. At the end of the measurement period the social network under analysis contains 122,030 nodes and 577,014 edges, with an average degree of 9.28 and an average clustering coefficient of 0.254. The giant component includes 96% of all observed nodes: the 90-th percentile of nodes network distances is 5.43 hops. There is evidence of the small-world effect, as found in other offline and online systems [22, 17]: while the shortest path lengths tend to be only a few hops, the average clustering coefficient is high, suggesting that strong local structures tend to be connected by occasional shortcuts.

The degree distribution exhibits a heavy tail, as depicted in Figure 1(a), with some nodes accumulating thousands of friends. In contrast, both the distribution of node age and link geographic distance, in Figures 1(b)-1(c), do not exhibit heavy tails but instead an almost exponential decay (notice the linear x-axis). We should note that whenever distance is calculated, we have used logarithmic binning with a minimum distance of 1 km. This allows a more robust analysis of the distribution of distances but preserves the distance characteristics of the dataset. There is a large fraction of short-range geographic connections: about 50% of social links span less than 100 km, with only a small fraction being longer than 4,000 km. The distribution of node age shows how users have joined the service with irregular temporal bursts; overall the trend can be approximately described as exponential, particularly for lower values of node age.

2.5 Limitations

Although our dataset represents a complete snapshot of Gowalla, this service was relatively small compared to other massive online social services. Furthermore, we underline a potential demographic bias: typical users of location-based services may have different mobility patterns and social habits than other Web users. In addition, some properties observed in our traces could be attributed to user engagement and Gowalla marketing and would not reflect user behavior in other online geosocial networks. Notwithstanding these limitations potentially present in our dataset, our analysis sheds new light on the spatial and social properties of Gowalla users and our findings can pave the way for further investigation on other services.

2.6 Model likelihood estimation

We take advantage of the fine-grained temporal information of our traces and we adopt a quantitative approach to compare how different attachment models describe the empirical data. We compute the likelihood that a model has to generate the observed events in our sequence of traces. The Maximum Likelihood Principle can then be applied: this principle is used to compare a family of models numerically and, as a result, pick the "best" model (and parameters) to explain the data.

Studying networks with likelihood methods requires a probabilistic model describing the evolution of the graph itself. In other words, the network is considered the result of an evolutionary stochastic process which drives its growth, both in terms of new nodes and new edges [32]. Given real data about the evolution of a network, one can test the extent to which the assumptions of a model are supported by the data.

In our case, estimating the likelihood of a model M involves considering each individual edge $e_t = (i, j)$ created during our measurement period and computing the likelihood $P_M(e_t)$ that the source i selects the actual destination j according to the model M. Thus, the likelihood $P_M(G)$ that model M reproduces graph G is given by the product of the individual likelihoods according to model M:

$$P_M(G) = \prod_t P_M(e_t) \tag{1}$$

We use log-likelihood for better numerical accuracy, obtaining

$$\log(P_M(G)) = \log(\prod_t P_M(e_t)) = \sum_t \log(P_M(e_t)) \tag{2}$$

Equation (2) suggests a simple algorithm to compute the log-likelihood of a given model M: for each new edge created during the graph evolution, we compute the probability that it would be created according to model M, we take the logarithm of this probability and we sum all the values obtained for each edge. When this procedure is repeated for several models, we can choose the model with the highest likelihood to explain the data.

Since every edge is undirected and we do not have information about which user initiated the social contact, we consider every new edge $e_t = (i, j)$ in both directions in the rest of our analysis, to avoid any bias. This methodology can be extended easily to handle directed graphs.

3. EDGE CREATION

In this section we study how the creation of individual edges is influenced by social and spatial properties of the nodes, exploring the effect of node degree, node age and spatial distance on the the edge attachment process.

3.1 Edge attachment by node degree

The preferential attachment model [2] posits that the probability of creating a new connection with a node is proportional to the number of its existing connections. This cumulative advantage held by high-degree nodes results in a degree distribution with heavy tail, as some nodes accumulate a large number of connections. We test if a similar mechanism is governing our data by computing the probability $P_{deg}(k)$ that a new link will be created with a node with degree k:

$$P_{deg}(k) = \frac{|\{e_t : e_t = (i, j) \wedge k_j(t-1) = k\}|}{\sum_t n_k(t-1)} \tag{3}$$

where the normalization factor considers all nodes with degree k just before the edge creation. If preferential attachment is not governing the growth $P_{deg}(k)$ should not depend on k: instead, we see in Figure 2(a) that $P_{deg}(k) \propto k^{0.74}$, denoting how nodes with higher degrees are more likely to attract new edges than nodes with fewer connections. Although the trend is not exactly linear as in the original preferential attachment model, node degree is related to the creation of new edges.

3.2 Edge attachment by node age

The amount of time a node has been part of the network could also be a factor which drives the creation of edges.

Older nodes might have more visibility on the service; at the same time, when new users join the network they might experience intense activity as they create their first connections. We compute $E(a)$, the number of edges created by nodes of age a normalized by the number of nodes that ever achieved age a [17]:

$$E(a) = \frac{|\{e_t : e_t = (i,j) \wedge t(e) - t(i) = a\}|}{\sum_t |\{n : T - t(n) \geq a\}|} \quad (4)$$

where T is the time when the last node joined the network during the measurement period. As reported in Figure 2(b), there is a spike at age 0: this represents nodes that join the network, create some links and then never come back. The number of created edges then quickly goes down with age a but grows again for higher values of a. This denotes that older nodes might benefit from receiving incoming links. The overall effect suggests that there is an abnormal spike of links created when a node joins the network followed by lower levels of edge creation: older nodes tend then to establish further links.

3.3 Edge attachment by node distance

The probability of having a social connection between two individuals decreases with their distance, although the exact functional form of this relationship is still under debate and appears slightly different in different systems [20, 1, 26]. We compute the probability $P_{geo}(d)$ that a new edge spans geographic distance d, normalized by the number of nodes at distance d from the source:

$$P_{geo}(d) = \frac{|\{e_t : e_t = (i,j) \wedge D_{ij} = d\}|}{\sum_t |\{n : D_{in} = d\}|} \quad (5)$$

Our data show how $P_{geo}(d)$ decreases with distance d, as reported in Figure 2(c), even though the trend appears noisy: in particular, the data roughly follow a trend $P_{geo}(d) \approx d^{-\alpha}$ with $\alpha = 0.6$ (depicted). While a similar functional form has been found in other spatial social networks, but with different exponents α, this is the first time it is measured at microscopic level on individual edge creation events. The main difference is that in this case there is a lower value of α, while other systems exhibit values closer to 1 [26]. Nonetheless, geographic distance affects the edge creation process in a straightforward way: longer links have a lower probability of appearance than short-range ones.

3.4 Evidence of gravitational attachment

The effect of node degree on network evolution is well captured by the preferential attachment model, where the probability of connection between nodes i and j, P_{ij}, is proportional to the degree of node j, $P_{ij} \propto k_j$. This model generates networks with degree distribution exhibiting a heavy tail, as there are a few nodes, the so called "hubs", that accumulate an extremely high number of connections. Real-world examples such as transportation and communication networks can be described by a preferential attachment model, but geographic distance is an important parameter as well. In fact, long-range connections tend to exist mainly between well-connected hubs [5].

The effect of geographic distance can be included in the attachment probability, $P_{ij} \propto k_i k_j f(D_{ij})$, where f is a decreasing *deterrence function* of the geographic distance D_{ij} between the nodes. Thus, long distances tend to be covered only to connect to important hubs, while nodes with less connections become attractive when they can be reached over a short distance. When the deterrence function has a simple functional form such as $f(d) \sim d^{-\alpha}$, then the probability of a connection between two nodes becomes similar to the gravitational attraction between celestial bodies, $P_{ij} \propto \frac{k_i k_j}{D_{ij}^{\alpha}}$. Hence, this family of attachment models has been known as *gravity models* [6]. We want now to understand whether there is any evidence that similar factors are shaping the evolution of our real spatial social network.

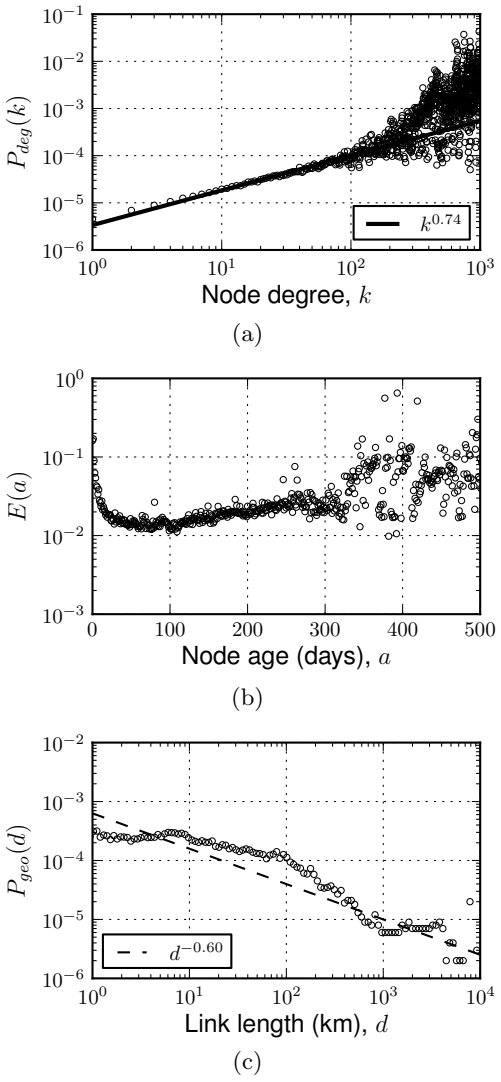


Figure 2: Probability of creating a new social link as a function of node degree (a), age of the node (b) and geographic distance of the node (c).

A consequence of the gravity model is that nodes with higher degrees tend to attract longer links: thus, we define $\lambda_i(k)$ as the geographic length of the k-th edge created by user i and we study $\lambda(k)$ for different values of k. The influence of degree k on the geographic properties of social links appears strong: as described in Figure 3, both the average and the median value of the geographic length $\langle\lambda(k)\rangle$ of the k-th edge increase with k: while the average length of the first edge is about 1,100 km, the 100th edge is about 2,400

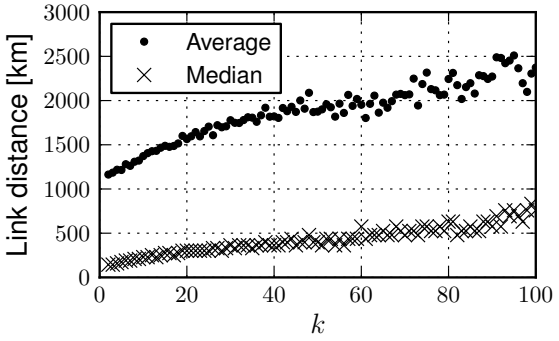


Figure 3: Average and median geographic span gap of the k-th edge created by a node as a function of k.

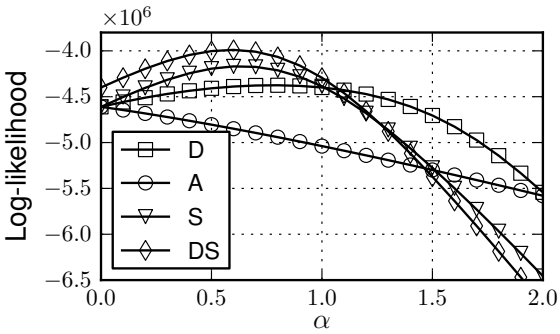


Figure 4: Log-likelihood of each edge attachment model as a function of their parameter α. The gravity model DS outperforms all the others.

km. The median value shifts in accordance with k, increasing from 150 km to more than 900 km for higher degrees. These findings are compatible with a gravity model where node degree and geographic distance simultaneously influence social connections created over space, as we will see in the next section.

3.5 Evaluation of attachment models

With our analysis we have discovered that individual node properties and geographic distance affect how edges are created. Our aim is now to understand what type of edge attachment mechanisms better explain the temporal evolution of the network.

We deliberately choose simple models, since our goal is not to accurately reproduce the temporal evolution of the network but rather to understand which factors mainly drive its growth. We consider 4 different edge attachment models, each one with a single parameter α:

- D: the probability of creating an edge with node n is proportional to a power α of its degree: $k_n(t)^\alpha$
- A: the probability of creating an edge with node n is proportional to a power α of its age: $a_t(n)^\alpha$
- S: the probability of creating an edge with node n is inversely proportional to a power α of its spatial distance from source i: $D_{in}^{-\alpha}$
- DS: the probability of creating an edge with node n is proportional to its degree and inversely proportional to a power α of its spatial distance from i: $k_n(t)D_{in}^{-\alpha}$

Figure 4 displays the log-likelihood values obtained by each model as a function of the parameter α. First, we note that the models S and DS, which incorporate geographic distance, have higher log-likelihood than the other two models D and A, with the maximum log-likelihood achieved by DS. The maximum log-likelihood for DS is achieved for $\alpha \approx 0.6$, which is in agreement with the results obtained measuring $P_{geo}(d)$. Node age does not seem a key factor for edge attachment, as the model A shows decreasing values of log-likelihood for values of α between 0 and 2, with its maximum log-likelihood of -4.4×10^6 reached instead only for $\alpha = -0.8$, failing to outperform S and DS. Indeed, we have tested models which also combine node age with geographic distance and node degree, but they do not exhibit significant improvements with respect to the models without node age. Hence, it seems that *the main driving factors, of those examined, in edge attachment are node degree and geographic distance* and that a gravity model which combines them is the most suitable option.

4. SOCIAL TRIADIC CLOSURE

The edge attachment mechanisms previously investigated only take into account the influence of global network properties on new edge creation. However, local network properties can be equally or more important: for instance, new links tend to connect users that already share friends, creating social triangles that are extremely common on social networks [19]. This mechanism, where a node just copies a connection from a node it is already connected to, has turned out to be essential to reproduce the structure observed in many networks [25]. Hence, in this section our aim is to study the extent to which new links generate social triangles and whether different models based on local network properties can reproduce the patterns observed in the data.

4.1 Measuring triangle creation

Social connections tend to link together individuals that are already at close social distance: the vast majority of new links tend to be between nodes that already share at least a connection, thus only 2 hops away from each other, with larger social distances exponentially less likely [17]. We notice a similar pattern in our data: Figure 5(a) shows that the number of edges E_h that connect nodes h hops away exponentially decays with h. Furthermore, many edges also connect nodes that were not in the same connected component, as when a new node joins the network and creates its first link.

A better understanding of this process can be achieved by considering not only how many new links connect nodes h hops away, but also considering the number of nodes at that social distance. In fact, since E_h exponentially decreases with h and the number of available nodes increases with h, the probability P_h that a new link spans h hops must be decreasing much faster than exponentially. More precisely, we compute P_h as

$$P_h = \frac{|\{e_t : e_t = (i,j) \wedge d_{t-1}(i,j) = h\}|}{\sum_t |\{n : d_{t-1}(i,n) = h\}|} \qquad (6)$$

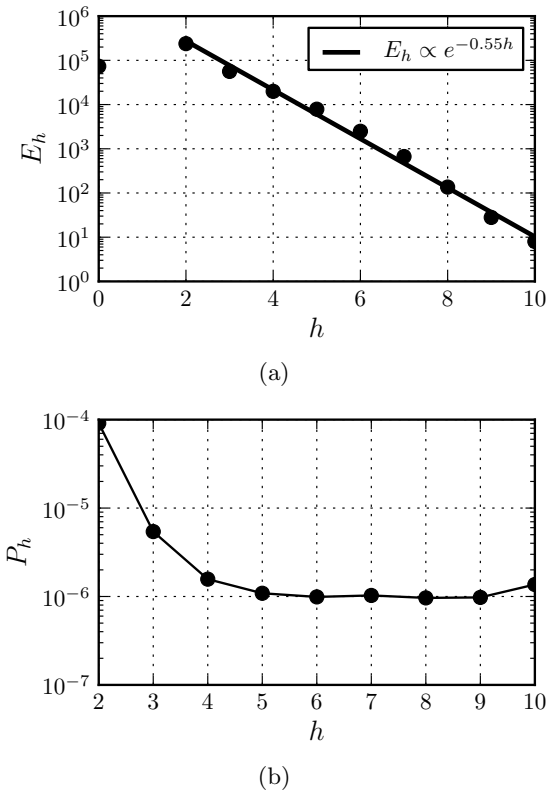


Figure 5: Number of new links E_h created between nodes h hops away (a) and probability P_h that a new link connects nodes h hops away. The single E_h value at $h = 0$ denotes the number of edges connecting nodes previously in separate disconnected components.

where $d_t(i, j)$ is the number of hops between nodes i and j at time t. Figure 5(b) plots P_h as a function of h: the probability quickly decays and finally reaches a constant value. Triadic closure seems to be the predominant factor shaping network growth over time: new edges are most likely to connect people who already share at least one friend.

In summary, our analysis of triadic closure provides evidence that two users sharing at least one friend are much more likely to create direct connections than two users without friends in common.

4.2 Triangle-closing models

Since a vast majority of new edges close social triangles, our aim is now to understand what factors influence which node to choose when an edge is closing a triangle. Again, we make use of the maximum likelihood principle to test and compare whether different triangle-closing models would be able to generate the triangles created during the real network evolution.

We consider the case when a source node s has to choose another target node t 2 hops away to create a new link. A simple model would be for node s to choose t uniformly at random from all the nodes at a distance of 2 hops, which will be our baseline model. We then take into account more complex models where a source node s first chooses according to a given strategy an intermediate node i among its neighbors

	random	shared	degree	distance	gravity
random	**12.34**	9.48	-3.47	-28.17	-35.26
shared	**14.54**	11.47	-0.95	-24.74	-34.46
degree	7.33	5.16	-6.79	-25.17	-41.98
distance	-0.92	-3.70	-16.94	-39.32	-41.53
gravity	2.71	0.25	-12.11	-33.01	-43.18

Table 2: Performance of different triangle closing models: on each row there is the model to pick the intermediate node and on each column the model to then pick the target node. The value in each cell gives the percentage improvement over the baseline, which is the log-likelihood of choosing a random node two hops away from the source.

and then picks a target t among i's neighbors with, potentially, a different strategy. The edge (s, t) is then created, closing the triangle (s, i, t). Since every strategy involves only choosing a node among the neighbors of a given node, we consider 5 different strategies to choose a neighbor v of a given node u:

- random: uniformly at random;
- shared: proportional to the number of shared friends between u and v;
- degree: proportional to the degree of the neighbor v;
- distance: inversely proportional to the geographic distance between u and v;
- gravity: proportional to the degree of v and inversely proportional to the geographic distance between u and v.

Since there are 5 different triangle-closing models, there are 25 combinations: we compute the log-likelihood of each combination and we measure the percentage improvement over the log-likelihood of the baseline model. The results are presented in Table 2: the general trend is that random and shared offer the largest improvements over the baseline, with a maximum improvement of 14.54% in the combination shared-random and 12.34% for random-random. Instead, models based on degree or on distance have performance much lower than the baseline, with degradation up to 40% when the gravity model is adopted. In particular, the random-random model works surprisingly well, as it favors connections between nodes that have multiple 2-hop paths between them and that have higher degrees, while being extremely simple and computationally fast.

These results show that *triadic closure is mainly driven by social processes*. Nonetheless, it only reproduces some aspects of network evolution, as edges that do not close triangles are also arising in the network. As we will see, we can exploit users' mobility information to understand how other online social connections are created, adopting closure models based on the *places* visited by online users.

5. MOBILITY-DRIVEN CLOSURE

The edge attachment mechanisms discussed so far do not include any information on users' mobility. However, the places where users check in could help explain how new social ties are created. According to the *common focus* theory [10],

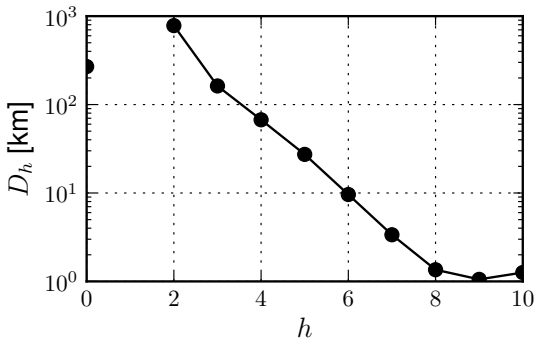


Figure 6: Average geographic distance D_h of new links created between nodes h hops away. The single value at $h = 0$ denotes the average geographic distance of links connecting nodes previously in separate disconnected components.

individuals who visit the same places tend to establish new social connections. In this section we measure the impact that users' mobility has on network evolution. In agreement with the common focus theory, we study edge attachment mechanisms that connect users that visit the same places.

5.1 Measuring mobility-based attachment

In our Gowalla traces, 32.28% of all new edges are established between users that share at least one common place. In particular, about 10% of new links are created between users that do share common places, but no common friends. This means that adopting only social triadic closure mechanisms would fail to reproduce that users create new social connections beyond their 2-hop neighborhood.

The importance of social ties that connect users without friends in common is confirmed when we examine the average geographic distance D_h of all new edges that connect nodes previously h hops away, shown in Figure 6. There is an evident trend: *social connections at shorter social distance tend to have higher geographic distances, while links spanning more hops have lower spatial distance.* A potential explanation is that both social and spatial factors tend to affect the edge creation process: a new link is created either between users sharing friends, even if they are far from each other, or between spatially close users, even if they have no friends in common. In particular, *geographic proximity becomes complementary to social closeness*: both factors are shaping the network, but in different ways. The challenge is to go beyond geographic distance when modeling the evolution of the network: mobility-based attachment provides the additional source of information, based on the places visited by users. Such information may be important to model network evolution, since it provides more accurate geographic information than users' home locations.

5.2 Mobility-driven closure models

We consider mobility-driven closure models to be two-step processes. A source node s first selects a place $p \in P_s$, where P_s is the set of all places where node s has checked in; then, given place p a target note $t \in Q_p$ is selected, where Q_p is the set of all nodes that have checked in at place p. We consider different strategies that a node u adopts to select a place p from the set P_u:

- random: uniformly at random;
- friends: proportional to the number of user u's friends that have visited p;
- user-checkins: proportional to the number of check-ins made by user u at p;
- tot-checkins: proportional to the total number of check-ins made at p by all users;
- tot-users: proportional to the total number of users who have checked in at p;
- place-distance: inversely proportional to the distance between user u's home location and p;
- place-gravity: proportional to the total number of check-ins made by all users at place p and inversely proportional to the distance between user u's home location and p.

Given a selected place p, we then consider another set of strategies to select a target user t from Q_p:

- random: uniformly at random;
- degree: proportional to user t's degree;
- deg-diffusion: proportional to user t's degree and inversely proportional the logarithm of user t's total number of visited places;
- user-checkins: proportional to user t's number of check-ins at p;
- tot-checkins: proportional to user t's total number of check-ins;
- inv-tot-checkins: inversely proportional to user t's total number of check-ins;
- distance: inversely proportional to the distance between user t's home location and p;
- gravity: proportional to user t's degree and inversely proportional to the distance between user t's home location and p.

To test and evaluate mobility-driven models we use again the maximum likelihood principle; we only evaluate the likelihood that a model has to reproduce real edge attachments where the source and target nodes share at least one place. We adopt a baseline model that selects at random target users from the set of all users that share places with the source. Table 3 presents the results for all the possible combinations.

In general, in the first step the best improvement is given by selecting a popular place that has already been visited by many users, friends or not. For the second step, node degree plays an important role, akin to a local preferential attachment. The greatest improvement over the baseline is provided by first selecting a place that has been visited many times (tot-checkins) and then choosing a node proportionally to its degree "diffused" over the number of visited places (deg-diffusion). This mobility measure corrects for the fact that popular users that visit only a few places might be more related to that place, thus enticing other visitors to

	random	degree	deg-diffusion	user-checkins	tot-checkins	inv-tot-checkins	distance	gravity
random	0.28	6.88	9.24	0.16	-17.02	-4.51	-19.36	-7.04
friends	4.70	**11.60**	**13.63**	4.74	-10.63	-1.56	-14.88	-1.71
user-checkins	0.05	6.59	8.94	-0.03	-17.27	-4.80	-19.69	-7.41
tot-checkins	6.09	**13.13**	**15.18**	6.14	-9.29	0.04	-13.15	-0.02
tot-users	5.10	**12.33**	**14.33**	5.16	-9.96	-1.08	-14.19	-0.84
place-distance	-23.41	-15.57	-13.21	-23.56	-40.82	-28.27	-43.67	-30.17
place-gravity	0.37	7.22	9.46	0.32	-16.26	-5.29	-19.60	-6.81

Table 3: Performance of mobility-driven closure models: on each row there is a model to pick the intermediate place and on each column a model to then pick the target node. The value in each cell gives the percentage improvement over the baseline, which is the log-likelihood of choosing a node at random among all the nodes that share at least one place with the source.

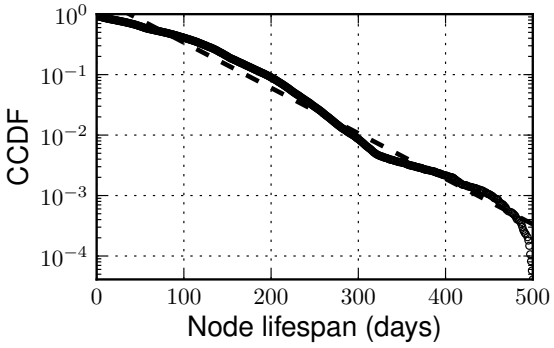


Figure 7: Complementary Cumulative Distribution Function (CCDF) of node lifespan and exponential fit.

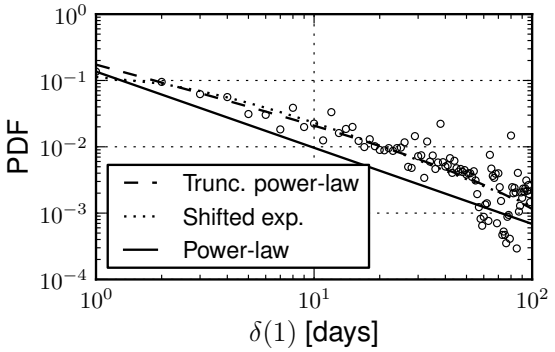


Figure 8: Probability Distribution Function (PDF) of $\delta(1)$, the temporal gap elapsing between the time when the first and the second edge are created by a user. The fits show a power law, an exponentially truncated power law and a shifted exponential.

connect. The tot-checkins-degree model has a similar but slightly inferior performance, yet it is simpler and computationally faster.

In addition to the models presented in Table 3, we experimented with variations of tot-users and tot-checkins where we use a probability of attachment *inversely* proportional to the total number of users or check-ins. All these models provided inferior performance compared to the baseline.

6. TEMPORAL EVOLUTION

In this section we study how users create new connections as they spend more time on the network. We study the amount of time users remain active for, their lifespan; then, we investigate the inter-edge temporal gap between the creation of consecutive edges. In this section we consider only users that joined the service after our measurement process started, in order to observe their behavior from the very first moment.

6.1 Node lifespan

We define the *lifespan* of a node as the difference between the time the node created the last and the first edge. Figure 7 plots the distribution of lifespan for all users: the distribution shows an approximately exponential behavior, with a deviation only at longer lifespans for few users who were early adopters and started using the service from the very first days. The fit is reasonably accurate for a wide range of lifespan values.

6.2 Inter-edge temporal gap

Different users can show significant differences in the pace they add new edges: users with higher degree create new ties at a faster rate. Thus, we study $\delta_i(k)$, the temporal gap between the k-th and $k+1$-th edges of user i, for different values of k.

Figure 8 displays the probability distribution of $\delta(1)$, the amount of time between the first and the second edges created by a user. Even though many users add their second edge after a few days, some wait for several weeks. The distribution can be reproduced by different functional forms: an exponentially truncated power law $\delta(1)^{-\alpha_1} exp(-\delta(1)/\beta_1)$ yields a slightly higher log-likelihood than a pure power-law, a shifted exponential and an exponential; the average log-likelihood improvement over the exponential fit is about 5%. This result also holds for different values of k.

Then, we study the effect of current degree k: in particular, we are interested in how the probability distribution of $\delta(k)$ changes with k. A first indication is given in Figure 9(a), which plots the average temporal gap $\langle\delta(k)\rangle$ between the k-th and $k+1$-th edges for different values of k: users with higher degrees tend to wait, on average, for a shorter amount of time. In fact, users wait on average 20 days before adding their second edge but only 7 days when they have about 100 friends. While α_k tends to be unrelated to k, the exponential cut-off β_k becomes smaller as k grows larger, as seen in Figure 9(b). The final effect is that nodes with higher degrees are more likely to wait for

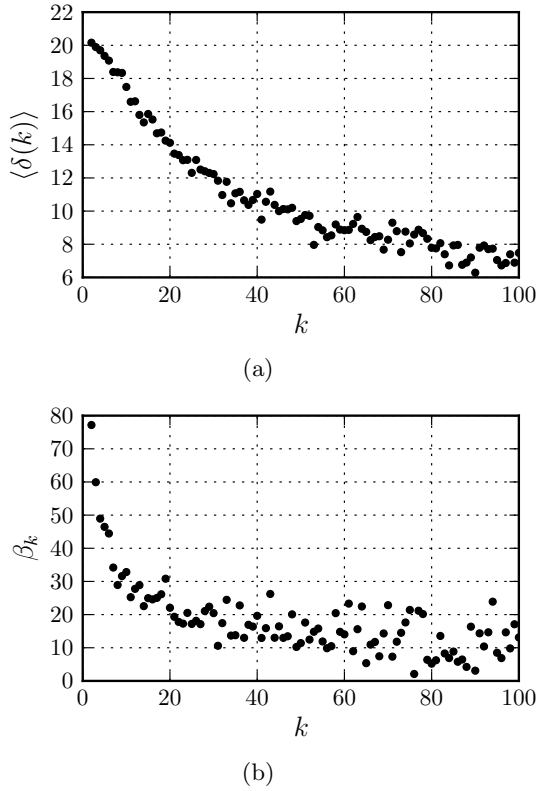


Figure 9: Average temporal gap $\langle\delta(k)\rangle$ between the k-th and $k+1$-th edge (a) and exponential cut-off β_k (b) in the truncated power law $p(\delta(k)) \propto \delta(k)^{-\alpha_k} exp(-\delta(k)/\beta_k)$, as a function of node degree k.

a shorter time span, as the truncated tail of the power law $P(\delta(k))$ increasingly constrains larger gap values.

It is not surprising that nodes with higher degree add links at a higher pace: given a fixed temporal period, as in our measurement, higher degree nodes add more links than lower degree ones, so their activity has to be faster in the same temporal period. Nonetheless, this heterogeneous temporal behavior is crucial to foster the heterogeneity observed in the degree distribution of social systems [17].

7. PUTTING IT ALL TOGETHER: NEW MODELS

We have seen that a gravity-based attachment, combining spatial distance and node degree, influences how new edges are created (in Section 3). At the same time, we have discussed that triadic and mobility-based closure are mainly shaped by social factors rather than geographic ones (in Section 4 and 5). These two mechanisms seem to be complementary: while the gravity attachment is responsible for edges connecting together different parts of the network, the closure mechanisms seem involved in the creation of local edges between nodes that already share either a friend or a place. Finally, we have analyzed how nodes tend to become faster and faster in creating new edges as they get more connections (in Section 6). Building on all these results, our aim is now to define network growth models which are able to reproduce the spatial and social properties observed in the real network. We stress that the goal of our models is not to accurately reproduce the network or predict edge creation events, but to describe the fundamental mechanisms affecting user behavior.

7.1 Model definition

Following the methodology presented in [17], we describe our model as a simple algorithm to grow a network one node, and one edge, at a time. Our model combines global attachment mechanisms and local closure mechanisms:

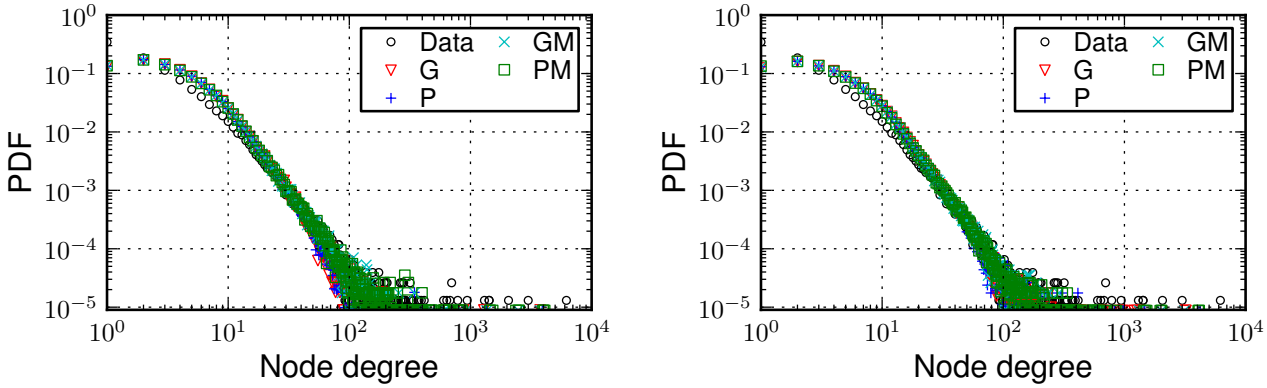
1. A new node u joins the network according to a certain arrival discipline and positions itself over the space;
2. A new node u samples its lifetime from an exponential distribution;
3. Node u adds its first edge to node v according to a global connection model (preferential or gravity-based attachment);
4. A node with degree k samples a time gap δ from a distribution $p(\delta(k)) \propto \delta(k)^{-\alpha_k} exp(-\delta(k)/\beta_k)$ and then goes to sleep for δ time steps;
5. When a node wakes up, if its lifetime has not expired yet it creates a new edge: with probability p the node uses the random−random social triangle-closing model, otherwise it uses the tot-checkins − degree mobility-based closure.
6. The node repeats step 4.

The probability p allows us to assess the impact that the local closure models have on overall accuracy. In particular, we adopt two variations: we use $p = 1$ so that the model only includes social triadic closure, and we adopt $p = 0.66$ to introduce also mobility-driven closure (this value is motivated by the observed frequency of edge attachments in the real data). This yields a total of 4 different combinations: gravity-based (G), gravity-based with mobility-driven closure (GM), preferential attachment (P) and preferential attachment with mobility-driven closure (PM).

Finally, we note that local closure models only account for about two-thirds of all new social links. The remaining fraction includes ties between users that do not share common friends and that do not visit the same places. Thus, we introduce a variation into our models whose aim is to adopt global attachment models also during a node's lifetime, and not only in step 3. In more detail, when a node wakes up in step 5, with probability $q = 0.33$ it creates an edge according to the global attachment mechanisms; otherwise, the model proceeds as defined. This yields 4 additional model combinations, for a total of 8 combinations. We will present our results separately for combinations that do and do not trigger global attachment mechanisms during a node's lifetime.

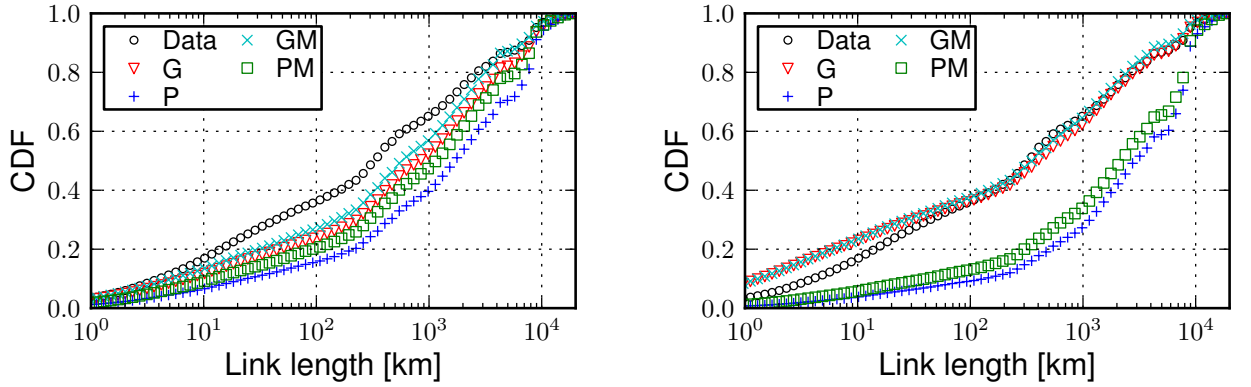
7.2 Evaluation

To test our model we take the real network at the beginning of our measurement, G_t, and simulate its growth by adding the missing nodes and check-ins, with their real geographic locations, and the check-ins according to when and where they happen in the real network. However, once they



(a) Without global attachment (b) With global attachment

Figure 10: Probability distribution function (PDF) of node degree for real data and different models: gravity-based (G), gravity-based with mobility-driven closure (GM), preferential attachment (P), preferential attachment with mobility-driven closure (PM).



(a) Without global attachment (b) With global attachment

Figure 11: Cumulative distribution function (CDF) of link geographic length for real data and different models: gravity-based (G), gravity-based with mobility-driven closure (GM), preferential attachment (P), preferential attachment with mobility-driven closure (PM).

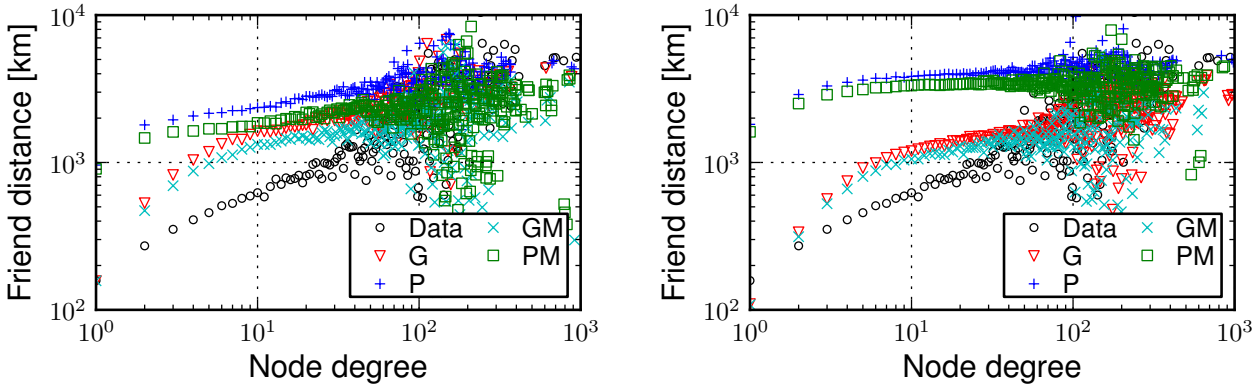
join the network they add new edges according to our algorithmic model. We stop the evolution when the simulated network has the same number of nodes as the real one, G_T.

All 8 model combinations are run 10 times with different random seeds and then their properties are averaged over all these realizations. When computing the properties shown in Figures 10, 11, 12 and 13 we only consider edges added after the start of our measurement period, both in the real network and in the simulated models, to avoid that the properties of the initial graph G_t dominate the final result.

The degree distribution observed in the real network and the two models are depicted in Figure 10: all models are able to reproduce the distribution, capturing the social properties of the real network. There is no noticeable difference between models with and without global attachment. However, as shown in Figure 11, the probability distribution of link geographic length is closer to the real one for gravity-based model, while models P and PM have links with longer geographic length. We note that models with global attachment have more similar characteristics than those without: the effect is that models G and GM reproduce better the real distribution, while social links created by models P and PM tend to span longer geographic distance. This suggests a positive effect of a gravity-based global attachment mechanism on the accuracy of the model.

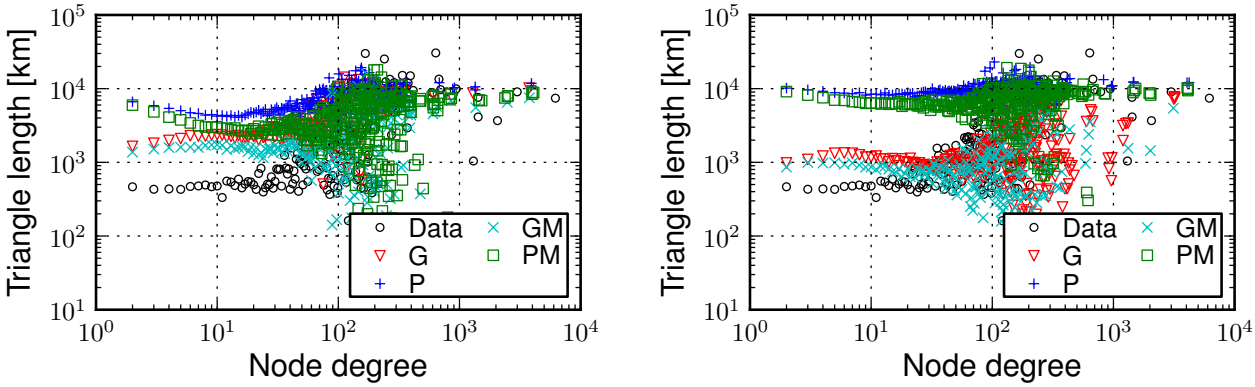
Another important difference between the gravity-based and the preferential attachment models is highlighted by considering the geometric average distance of the links of a user as a function of the node degree. As seen in Figure 12, models G and GM show an increasing trend as the original data, while, on the contrary, models P and PM have weaker correlation between node degree and geographic length of the links. Again, global attachment emphasizes the difference between gravity-based and preferential attachment models, with the first ones reproducing more accurately the real trend. A similar behavior is observed in Figure 13 the preferential attachment models create triangles on a much wider geographic scale than models G and GM, which are closer to the real data.

We note that when mobility-based closure is used results always marginally improve with respect to models with only social triadic closure. This performance increase can be



(a) Without global attachment (b) With global attachment

Figure 12: Average geographic friend distance as a function of node degree for real data and different models: gravity-based (G), gravity-based with mobility-driven closure (GM), preferential attachment (P), preferential attachment with mobility-driven closure (PM).



(a) Without global attachment (b) With global attachment

Figure 13: Average geographic triangle length as a function of node degree for real data and different models: gravity-based (G), gravity-based with mobility-driven closure (GM), preferential attachment (P), preferential attachment with mobility-driven closure (PM).

attributed to the latent geographic information embedded in user check-ins. The effect of global attachment is even stronger, as it enhances the accuracy of gravity-based models, while also reducing the validity of preferential attachment models. These results confirm that the effect of geographic distance can not be neglected when social networks are studied and modeled: preferential attachment mechanisms need to be modified into gravity-based mechanisms, which are able to correctly balance the effects of node attractiveness and the connection costs imposed by spatial distance. Furthermore, mobility-based closure improves model accuracy, offering additional information about the geographic whereabouts of online users.

7.3 Implications

The importance of our findings goes beyond the definition of accurate models of network evolution. Our results show that the effect of geographic distance cannot be neglected when online social networks are studied and modeled. In reality, preferential attachment and triadic closure together are already able to reproduce the global social properties observed in real social networks, namely the degree distribution and the level of clustering. However, neglecting spatial information about where users are located fails to account for the effect of distance. In real systems users preferentially connect over short distances, resulting in a considerable fraction of short-range ties; instead, ignoring spatial constraints would predict an unlikely majority of long-range connections. This goes against empirical evidence, both in offline and online social systems.

Our findings support the idea that distance has a simple effect on the creation of social ties: the probability of connection between two individuals decreases as a negative power of the spatial distance between them. Yet, this effect must be combined with a process based on "popularity" or "visibility" that introduces heterogeneity across users, such as attachment to the best connected nodes, in order to fully recreate the self-reinforcing mechanisms that lead to the scale-free degree distributions observed in social graphs.

Gravity mechanisms provide an elegant and insightful way

of combining the effect of distance and the influence of social factors. Surprisingly, the influence of distance on the formation of social triads appears negligible, as other factors become more important at this level. The main implication of the gravity mechanism is that one user may be interested in another because the other user is hugely popular, regardless of their spatial distance, or because the other user is spatially close, *regardless of popularity and importance*. These mechanisms can be adopted in scenarios where the future evolution of an online social graph has to be estimated: some examples include security mechanisms for online services [1], the design of distributed storage solutions for massive social graphs [24] and the delivery of user-generated content [29].

The overall picture is that proximity both over physical and social dimensions fosters the creation of new social links; the result is that the likelihood of a new connection increases when two individuals share many other connections or when two individuals are close to each other. We also point out that no friend recommendation mechanism was in place on the online service under analysis during the measurement period.

This dual role of proximity has promising applications in a wide range of systems. In particular, despite the abundance of friend recommendation services, only few of them have included spatial closeness in their systems [8, 27]. Our work provides new insights for further research in this direction.

8. RELATED WORK

The temporal patterns of network evolution have been the focus of many studies and several models have been put forward to describe the basic mechanisms that drive network growth. A set of works studied the evolution of online social networks, discussing the densification and diameter reduction observed during the growth of the graph [11, 15]. Even though online social graphs tend to be have an heterogeneous degree distribution in agreement with the preferential attachment principle, these findings highlighted that, in social networks, different mechanisms seems to be in place. More in detail, Leskovec et al. [18] propose a "forest-fire" copying process: when a new node joins the network and connects to a first neighbor, it starts recursively creating new links among the neighbors of the first neighbor, effectively copying the connection already in place. This process mixes preferential attachment, as more connected nodes are more likely to be selected, and transitivity, which fosters new connections between nodes in social proximity. This confirms the importance that triadic closure holds for the evolution of social graphs, as we have seen in our results.

Other works have also been focusing on triadic closure: Simmel noted that people sharing many friends might be more likely to become connected [28]. This effect was then measured in online social networks [19, 14] and included in growth models. With respect to these results, our work explores, for the first time, the effect of spatial distance on network evolution: specifically, we study how distance influences preferential attachment and triadic closure.

Other works have been focusing on general models for spatial networks. One of the earliest examples is the Waxman model, where nodes are distributed at random over space and then connected with probability exponentially decreasing with distance [30]. The Waxman model has also been modified as a growth model, where new nodes join the network and connect with a similar rule [12]. Barthélemy proposed to combine preferential attachment with spatial distance, studying how the resulting graphs move away from being scale-free as the effect of spatial distance is increased [4], albeit this case only considered an exponential decay with distance as in the original Waxman model. Barrat et al. [3] also considered a similar model for weighted networks where preferential attachment is driven by the weight of the existing connections and hampered by spatial distance. While these works contain the initial ideas about including spatial influence in models of network growth, they were based on systems such as transportation networks that lacked social properties. Hence, they tend to focus on an exponential decay of the probability of connection as a function of distance, differently to what observed in social graphs, and they ignore properties arising from triadic closure. Our contribution builds on these findings and bridges together several different insights in order to obtain a suitable model for spatial social graphs.

Another set of works have investigated the spatial properties of social networks: the influence of geographic distance on social connections was firstly discussed in the LiveJournal community [20]. This influence appears so important that it can be exploited to predict where people live given their friends' locations [1]. Other studies on mobile phone communication networks have found that social triangles tend to extend over large geographic distances [16] and that community detection approaches should take spatial distance into account to achieve better results [9]. The fostering effect of geographic proximity on social ties has been demonstrated considering both purely spatial coincidences [7] and repeated visits to venues [27]. Our work extends these results by analyzing the effect of spatial distance not on the static structure of social networks but on their temporal evolution.

Finally, we adopt the maximum likelihood methodology from and we base our growth model on results presented in [17], where the evolution of four different online social networks was discussed. Again, our work differs as it addresses the effect of geographic distance on the temporal mechanisms that govern network evolution, providing a more complete understanding of the factors driving social behavior. Furthermore, we describe a model of network growth which successfully reproduces both social and spatial properties observed in online social graphs.

9. CONCLUSIONS

In this work we studied the evolution of the social graph of a location-based service. We collected data about social links created and places visited by users over a period of 4 months and we studied the effect of spatial factors on the growth of the network.

We tested different models of edge attachment and we found that on a *global scale* node degree and spatial distance simultaneously affect how individuals create social connections. On the other hand, on a *local scale* we studied triadic closure models based on shared friends and on shared places: in these cases social factors are more important than spatial ones. Finally, we explored the temporal properties of network evolution, studying how much time users remain active on the service and how the time elapsed between the creation of consecutive social connections becomes shorter as users have more friends.

Based on our findings, we defined and tested network growth models combining a global gravity-based attachment with local closure models based on shared friends and places. Our models are able to reproduce the structural and spatial properties observed in the traces. Our results highlight basic factors driving social network growth that could impact a range of research efforts and practical applications. Overall, this work builds up on previous results and provides further evidence that spatial factors should not be neglected when studying and modeling online social services.

Acknowledgments

This research was partly funded by a Google Research Award.

10. REFERENCES

[1] 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*, 2010.

[2] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439), 1999.

[3] A. Barrat, M. Barthélemy, and A. Vespignani. The effects of spatial constraints on the evolution of weighted complex networks. *Journal of Statistical Mechanics*, (05), 2005.

[4] M. Barthélemy. Crossover from scale-free to spatial networks. *Europhysics Letters*, 63, 2003.

[5] M. Barthélemy. Spatial Networks. *Physics Reports*, 499, 2011.

[6] V. Carrothers. A Historical Review of the Gravity and Potential Concepts of Human Interaction. *Journal of the American Institute of Planners*, 22, 1956.

[7] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. M. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436–22441, 2010.

[8] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of UbiComp'10*, Copenhagen, Denmark, 2010.

[9] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *PNAS*, 108(19):7663–7668, May 2011.

[10] S. L. Feld. The Focused Organization of Social Ties. *American Journal of Sociology*, 86(5):1015–1035, 1981.

[11] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of WWW'03*, 2003.

[12] M. Kaiser and C. C. Hilgetag. Spatial growth of real-world networks. *PRE*, 69:036103, Mar 2004.

[13] T. Karagiannis, C. Gkantsidis, D. Narayanan, and A. Rowstron. Hermes: clustering users in large-scale e-mail services. In *Proceedings of SoCC'10*, 2010.

[14] D. Krackhardt and M. S. Handcock. Heider vs Simmel: emergent features in dynamic structures. In *Proceedings of ICML'06*, 2006.

[15] R. Kumar, J. Novak, and A. Tomkins. Structure and Evolution of Online Social Networks. In *Proceedings of KDD'06*, 2006.

[16] R. Lambiotte, V. Blondel, C. Dekerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Vandooren. Geographical dispersal of mobile communication networks. *Physica A*, 387(21), 2008.

[17] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of KDD'08*, 2008.

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of KDD'05*, 2005.

[19] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of CIKM'03*, 2003.

[20] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.

[21] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *Proceedings of KDD'10*, 2010.

[22] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of IMC '07*, 2007.

[23] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis. Geographic constraints on social network groups. *PLoS ONE*, 6(4):e16939, 2011.

[24] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez. The little engine(s) that could: scaling online social networks. In *Proceedings of SIGCOMM'10*, 2010.

[25] D. M. Romero and J. Kleinberg. The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. In *Proceedings of ICWSM'11*, 2011.

[26] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial Properties of Online Location-based Social Networks. In *Proceedings of ICWSM'11*, 2011.

[27] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of KDD'11*, 2011.

[28] G. Simmel. *The Sociology of Georg Simmel*. The Free Press, 1908.

[29] S. Traverso, K. Huguenin, V. Trestian, I. Erramilli, N. Laoutaris, and K. Papagiannaki. TailGate: Handling Long-Tail Content with a Little Help from Friends. In *Proceedings of WWW'12*, 2012.

[30] B. M. Waxman. Routing of multipoint connections. *Selected Areas in Communications*, 6(9):1617–1622, 1988.

[31] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao. Exploiting locality of interest in online social networks. In *Proceedings of CONEXT'10*, 2010.

[32] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf. A likelihood approach to analysis of network data. *PNAS*, 103(20):7566–7570, 2006.