This paper was presented as part of the Workshop on Network Science for Communication Networks (NetSciCom)

Measuring User Activity on an Online Location-based Social Network

Salvatore Scellato
Computer Laboratory, University of Cambridge
salvatore.scellato@cl.cam.ac.uk

Cecilia Mascolo
Computer Laboratory, University of Cambridge
cecilia.mascolo@cl.cam.ac.uk

***Abstract*—While in recent years online social networks have been largely shaping user experience on the Web, only in the last year there has been a soaring increase of novel location-based social applications. By allowing users to check-in at places and share their location with friends, these social platforms provide a new facet of user online behavior.**

In this work we present a measurement study of user activity on a popular online location-based social network with hundreds of thousands of users. We study not only how users connect with friends but also how they check-in at different places. We describe how, while the number of friends appears distributed according to a Double-Pareto law, both the number of check-ins and the number of places per user are better described by log-normal distributions. Moreover, we report how user activity spans decay faster than exponentially and how, over time, users add friends more quickly than they accumulate check-ins and places.

Our findings suggest that the difference in the distribution of friends and check-ins/places may be motivated by physical constraints that do not allow users to steadily visit very large numbers of new places, while online friends can be added at virtually no cost. These results shed new light on how users engage with location-based online social networks and prompt many more research questions.

I. Introduction

In the recent months location-based online social networks have experienced growingly higher levels of popularity, accumulating millions of users and attracting the attention of the mass media [1]. Among the biggest providers there are Foursquare and Gowalla, while other hugely popular social networking services such as Facebook and Twitter also introduced location-based features. Location is increasingly becoming a crucial facet of many online services: people appear more willing to share information about their geographic position with friends, while companies can customize their services by taking into account where the user is located. It appears more and more likely that location-based capabilities will gradually be offered as a feature by many online services.

In particular, location-based online social networks were among the first providers to explicitly build services tightly connected to physical places. They are often grounded on the *check-in* concept: users willingly share their own location with their friends by broadcasting the place where they are, usually through a location-sensing mobile device. Sometimes it might also be possible to unlock particular benefits or commercial deals by virtue of the check-in itself, attracting not only more users but also more investors to these services.

The check-in dynamics gives rise to a new dimension of user activity which was not present in previous online social networking tools. Data about the interplay between users and locations are for the first time available to researchers, providing unprecedented chances to understand how users actively engage with places and online friends. Among the research questions that are triggered by this new category of services, *a primary and fundamental issue is to measure and study how users take advantage of social and location-based features.*

This work gives some answers on this topic by presenting the first characterization of a popular location-based online social network with hundreds of thousands of users, Gowalla. We collect and analyze a complete and novel data set, focusing both on the social ties among users and on the check-ins made by users, which constitute the main user activity on the service. Since temporal information about user accounts has resulted fundamental to understand the distribution of friends [2], we further study how the age of user accounts and their activity spans influence the number of friends, check-ins and places users accumulate over time on Gowalla.

Our main contributions can be summarized as follows:

- We investigate user activity by analyzing not only the number of friends users have, but also the number of check-ins made and places visited by Gowalla users: we find that while the distribution of the number of friends can be described by a Double Pareto-like distribution, both check-ins and places distribution are better fitted by log-normal distributions.
- We study the age and the activity span of user accounts, highlighting that while the age of an account appears uniformly distributed over a wide range of values, user accounts remain active for much shorter time periods, exhibiting a decay faster than exponential.
- We suggest how the differences observed in the probability distributions of user activity can be motivated by differences in the rate at which users accumulate friends, check-ins and places over time.

Our findings uncover for the first time how user activity on location-based online social networks may differ from standard online social services and confirm that user account activity spans represent a key factor to study and analyze online social services.

978-1-4244-9920-5/11/$26.00 ©2011 IEEE

II. Data Acquisition

Studying large and active online social networks is often difficult because of the massive amount of user accounts combined with the limitations put in place by service providers against indiscriminate and unregulated data downloads. Nonetheless, random or sequential crawling is often the only method to acquire data and quantitatively measure and study these systems. In this section we describe our measurement methodology, highlighting how we were able to acquire a complete dataset.

A. Gowalla

Gowalla is a location-based social networking service created in early 2009 that allows users to add friends, to share their location and to post notes or photos. Users check-in at places through a dedicated mobile application, available for mobile platforms such as Google Android, Apple iPhone and BlackBerry, as well as through a mobile Web browser. These devices use GPS and other sensing technologies to automatically detect their location, though users are still required to willingly check-in.

Check-ins can be pushed via notifications to mobile applications and, by linking accounts, to Twitter and Facebook. As a consequence, friends can check where a user is or has been; conversely, it is possible to see all the users that have recently been in a given place. The friendship relationships is mutual, requiring each user to accept friendship requests to allow location sharing.

B. Data collection

Gowalla provides a public API to let other applications integrate with their service: in particular, they provide information about user profiles, friend lists, user check-ins and place details. We have collected a complete snapshot of Gowalla data in August 2010. For every user we have gathered the user profile, the friends list and the list of all the places where the user has checked-in. The API does not provide unauthorized access to fine-grained temporal information about user check-ins. However, the API presents the dates of the earliest and latest check-ins for each place where a user has checked-in at.

Since users are identified by consecutive numeric IDs, we were able to exhaustively query all user accounts, downloading all the aforementioned information. We were able to download an entire corpus of data within 6 hours, employing a multi-threaded crawler. As a result our dataset might be considered as an almost instantaneous and complete snapshot of the service itself, which allows us to analyze and investigate the properties of Gowalla users without incurring in any sampling bias or in major temporal heterogeneities arising because of the collection methodology.

C. General properties

Our dataset contains more than 400,000 users (see Table I), which constitute the entire user base of Gowalla at the date of our measurement. However, not all these accounts provide data about user activity: the fraction of *active* users, or users with at least one check-in, is only 45.7% of the total, while the percentage of *social* users, or users with at least one friendship connection, is even smaller at just 41.0%. A high fraction of inactive users has been found also in other online social networking services [3], and Gowalla does not seem to deviate from this behavior. In addition, active users are more numerous than social users.

Total users	402,030
Active users	183,709
Social users	165,016
Active & Social users	125,568
Places	1,470,727
Check-ins	13,770,652
Average check-ins per place	9.36
Average check-ins per active user	74.96

TABLE I
Properties of the Gowalla dataset under analysis

We extract a social network from all the friends lists, where the set of nodes V contains all the social users and E is the set of friendship connections. We represent it as an undirected graph $G = (V, E)$ with $|V| = N = 165,016$ and $|E| = K = 759,770$, with the average number of friends per user being about 9.21. The social network of Gowalla users exhibits properties found in many other online and offline social networks. The shortest network path length between users is distributed according to a normal distribution with average 4.5, with only a minor fraction of couples being more than 8 hops away from each other and a maximum network diameter of 14 hops. These findings, together with an average clustering coefficient of 0.228, suggest that the social network among Gowalla users presents highly clustered social ties yet short social distances, usual indicators of small-world behavior found in many other social systems [4].

III. User Activity Analysis

In this section we analyze our data on Gowalla users and the aggregated properties of their activity on the service, highlighting some differences between social and location-based features. Then, we investigate the age and the activity span of user accounts and, finally, we study how different user activities relate to account lifetimes.

A. User activity

Differently than other online social networks, where user activity mainly involves interacting with friends by means of posts, shared web links and messages, Gowalla users check-in at places and share these events with their connections. Hence, we focus on three important indicators of user acivity: *adding online friends*, *making check-ins* and *visiting new places*.

The probability distribution of the number of friends of Gowalla users is shown in Figure 1: there is a significant heavy tail in the distribution, with some accounts accumulating large number of friends as observed in many other online social networks [5], [6]. We adopt maximum-likelihood estimation to find the best fit for a power-law and for a log-normal distribution [7]. However, the empirical distribution can not

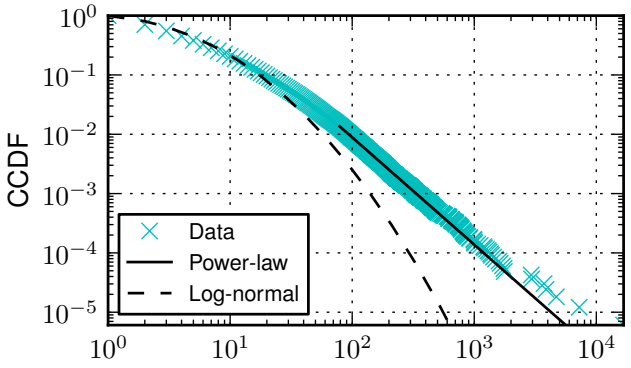
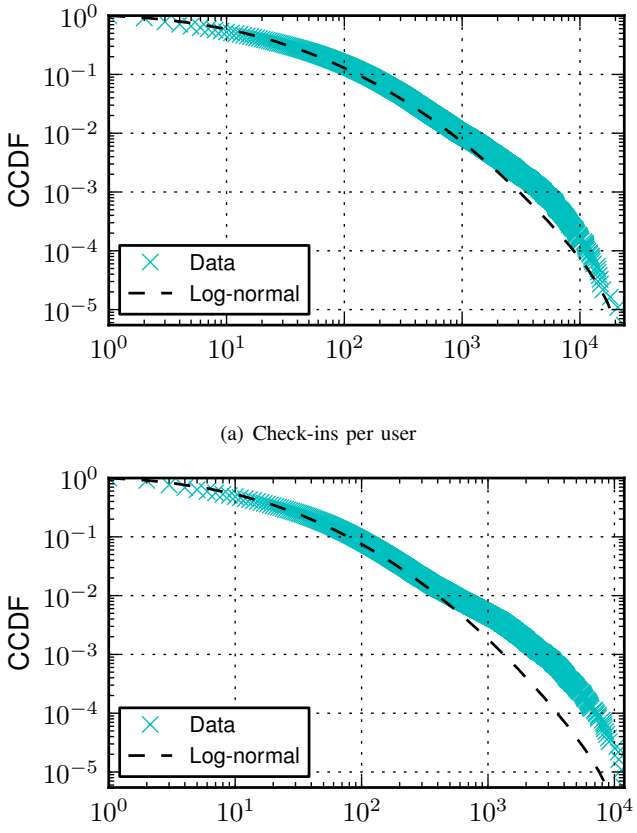


Fig. 1. Empirical Complementary Cumulative Distribution (CCDF) of the number of friends of Gowalla social users. The data approximately exhibit a log-normal body with a power-law tail, consistent with a Double-Pareto law.

be fitted by a single power-law distribution, nor by a single log-normal distribution. Instead, the distribution presents a log-normal body and a power-law tail, which is in agreement with a Double-Pareto distribution [8], [9]. A standard Double-Pareto law appears similar to both a log-normal and a power-law distribution: in particular, the main difference between a log-normal and a Double-Pareto is how the tail does not die as quickly as expected, following instead a power-law behavior. On the other hand, the probability distributions of the number of check-ins per user and of the number of different places each user has checked-in at show a different behavior. As observed in Figure 2, both distributions follow log-normal distributions across the entire range of values, instead of showing power-law tails.

While log-normal and power-law distributions seem similar, they tend to arise from different generative processes. For instance, log-normal distributions are known to appear under some weak conditions whenever a random process X_t grows over time through multiplicative steps ruled by random variables A_i, as in $X_t = A_t X_{t-1}$. As a result, $\ln(X_t) = \sum \ln(A_i) + X_0$: in particular, if all $\ln(A_i)$ are independent and identically distributed random variables with finite mean and variance, then $\ln(X_t)$ will asymptotically approach a normal distribution according to the Central Limit Theorem and, hence, X_t will approach a log-normal distribution. At the same time, there is a minor variation of this multiplicative generative model that results in a different behavior. Suppose that the random variable X_t is measured only at certain time intervals T, also randomly distributed. This is indeed our case, since each user is observed after a certain period of time since joining Gowalla: the aggregated distribution is given by mixing together users which might have undergone a different number of multiplicative steps. In particular, Reed considered a process X_t which at every time step t appears as a log-normal distribution with variance t and zero mean: if this process is sampled at exponentially distributed times T, then the resulting sampled variable X_T follows a Double-Pareto distribution with a power-law tail [8].



(a) Check-ins per user

(b) Places per user

Fig. 2. Empirical Complementary Cumulative Distribution (CCDF) of the number of check-ins (a) and of the number of different places (b) per active user. Both sets of values can be described by log-normal distributions.

B. Account age and activity span

In order to understand whether the differences among user activity distributions can be described by such generative mechanisms, we further investigate the age and the activity span of user accounts. Information about when a user account was created is not directly available in our dataset: however, for every place where a user has checked-in there are the dates of both the earliest and the latest check-in. Thus, it is possible to estimate when each active user account was created and when it was last accessed by the user.

Given user i we define S_i as the date of the earliest check-in and L_i as the date of the latest check-in: then, the *age* of a user account is computed by taking the difference between our measurement date and S_i, while the *activity span* of a user account is the difference $L_i - S_i$. These estimates let us understand how groups of users with similar age or activity span can be statistically described. The probability distribution of user account age and activity span show different patterns, as seen in Figure 3: while account age is approximately uniformly distributed in the range 0-300 days, accounts are usually active for much shorter values. The median account age is about 147 days, but the median activity span is only

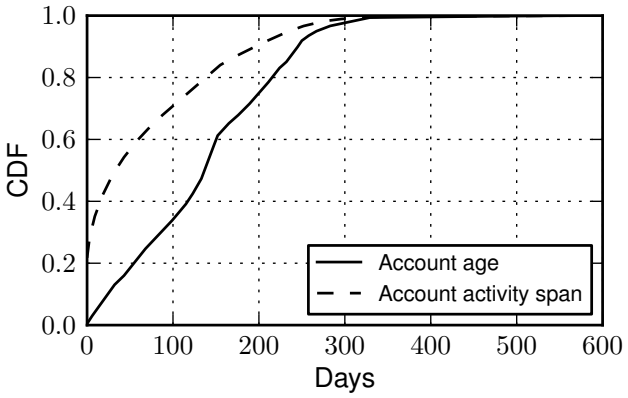


Fig. 3. Empirical Cumulative Distribution (CDF) of the user account age and user activity span for Gowalla active users.

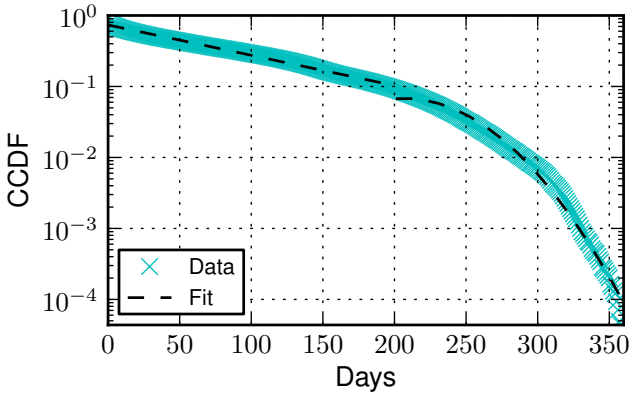


Fig. 4. Empirical Complementary Cumulative Distribution (CCDF) of user activity span. The activity span distribution follows an exponential law up until 200 days, then it decays faster than an exponential.

33 days. Although Gowalla was about 600-day old when our dataset was collected, almost all values are below 350 days: this is because in the first months after its launch Gowalla did not accumulate a significant number of users.

We further investigate the distribution of user activity span in Figure 4. The CCDF of activity span shows two different behaviors: the majority of user account activity spans are distributed exponentially up until 200 days, when the distribution starts to decay faster than an exponential following a quadratic law $e^{-\lambda t^2}$. A comparable pattern has been found in other online services, suggesting that this might be a recurrent property of these platforms [2]. While exponential user activity spans may arise from an exponential increment of the number of users in Gowalla, this conclusion can be already refuted by comparing the two cumulative distributions of account age and activity span in Figure 3.

This is also confirmed by the probability distribution of account age in Figure 5, which gives us insight on when user accounts were created. After an initial brief phase of expansion there is no sign of exponential increase, while, instead, growth remains more or less constant. The peak at the age of 150 days can be explained by the huge media exposure that Gowalla

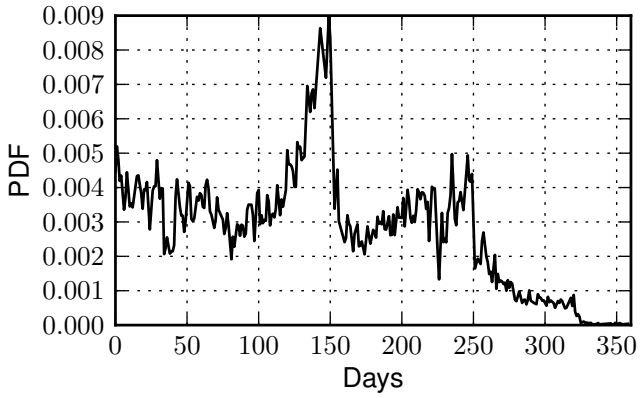


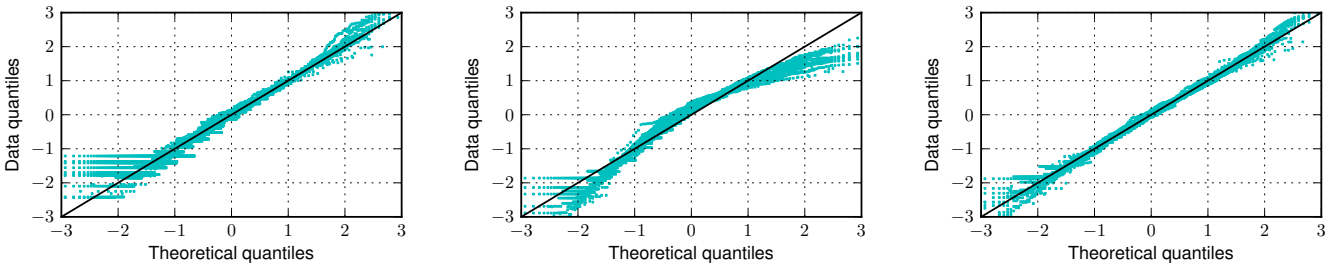
Fig. 5. Empirical Probability Distribution (PDF) of the user account age in Gowalla. User account ages show no sign of steady exponential growth, while it is possible to notice peaks occuring simultaneously with external known events.

had during the XIII South by Southwest Interactive (SXSW) conference, where Gowalla won the award as best company in the Mobile category in March 2010. This award unleashed a large number of new user accounts as a result of the hype surrounding the company in the following days [10], [11].

C. User activity over time

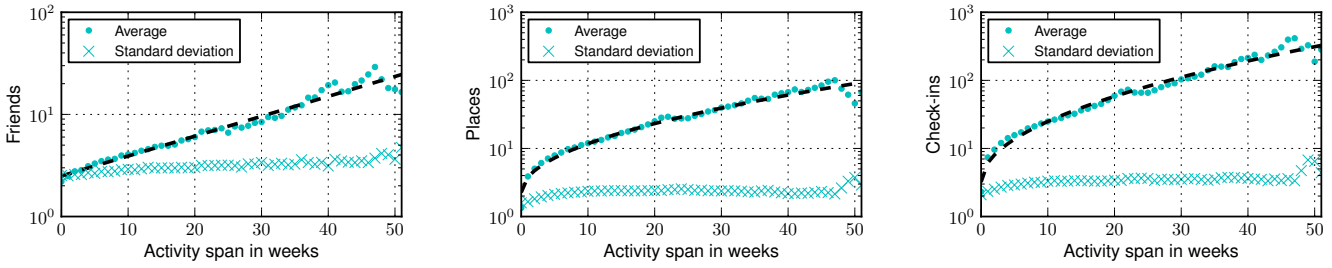
Since user activity span is distributed exponentially, albeit with a tail which decays much faster than predicted, user accounts might be considered as sampled at randomly distributed times of their evolution. However, an account which has been active for a longer period is more likely to accumulate more friends, check-ins and places than an account only active for a shorter amount of time. The distributions observed in Figures 1-2 are thus arising by mixing together the distributions of groups of users with different activity spans. Since those distributions exhibit different patterns, *it is crucial to understand whether users activity shows also varying behavior as a function of user account activity span*. In particular, given the models behind the Double-Pareto and log-normal distributions, we want to investigate what type of user activity distribution is observed in user accounts with similar activity span.

We test the hypothesis that user activity follows a log-normal distribution across different activity spans, as suggested by previous findings on MySpace [2]. We separate users with different activity spans with a temporal granularity of one month (30 days) and then we study the probability distribution of the number of friends, check-ins and places for each set of users. Each set of values is log-transformed, then the average and the standard deviation are computed in order to rescale the values and obtain a distribution with zero mean and unitary standard deviation. If the original values follow a log-normal distribution, then the transformed values should follow a t-Student distribution [2]. The two distributions can then be directly compared by means of a QQ-plot, which shows how the quantiles of two distributions correspond: if the QQ-plot follows the $y = x$ relationship, with a clear correlation of the



(a) Friends (b) Places (c) Check-ins

Fig. 6. QQ-plots of the transformed distributions of friends (a), places (b) and check-ins (c) against a theoretical t-Student distribution, for users with the same activity span with a temporal granularity of 30 days. Each graph presents 20 diffent plots, one for each set of users with the same account lifetime in months.



(a) Friends (b) Places (c) Check-ins

Fig. 7. Average number of friends (a), of places (b) and of check-ins (c) as a function of the account activity span T. While the number of friends grows according to e^T, both the number of places and of check-ins grow more slowly as $e^{\sqrt{T}}$ (note the logarithmic scale of the y-axis). The standard deviation is instead largely unaffected by T.

quantiles, then the original distribution can be described as a log-normal [12]. The resulting QQ-plots for the number of friends, check-ins and places can be found in Figure 6: while the number of friends and the number of check-ins match a log-normal distribution across different user activity spans, albeit with some deviations mainly at the extreme quantiles, the number of places presents stronger deviations both in the body of the distribution and in the extreme quantiles. This discrepancy might explain why the distribution of the number of places in Figure 2(b), instead, does not exactly follow a log-normal distribution as the number of check-ins nicely does.

However, even if both the number of friends and the number of check-ins appear equally in agreement with the log-normal distribution, the aggregated distributions are different: the former shows a Double-Pareto law, the latter a log-normal. Hence, we study user activity as a function of user account temporal span in order to investigate how quickly accounts accumulate friends, places and check-ins. We report in Figure 7 the average and the standard deviation of the number of friends, check-ins and places as a function of user activity span. While all variables grow over time, they also follow different behaviors: the average number of friends of an account with span T grows according to e^T, but the average number of check-ins and places grows more slowly as $e^{\sqrt{T}}$. At the same time, the standard deviation of these variables is largely unaffected by the activity span, since it remains fairly constant as T increases.

D. Discussion

Our analysis and our results highlight some key differences in the relationship between the amount of time an account is active and the activity related to that account. In particular, *it appears easier and quicker to accumulate friends than to accumulate new places and check-ins*. While users quickly visit more places and make many check-ins in the first weeks, this trend then slows down as it takes more time to add new places and new check-ins. This may be motivated by the fact that creating new online friendship connections does not involve a cost as large as visiting new places or making new check-ins, which is likely to need either time or money, as physical movement becomes necessary. While finding new places may seem difficult, making additional check-ins should be easier: nonetheless, both variables experience a similar growth pattern. Even though these findings may support the claim that user engagement slows down as users spend more time using the service, this relationship deserves further investigation.

At the same time, the differences found in this temporal behavior may be the cause of the different distributions arising at the aggregated levels: since friends are accumulated more quickly over time, it might be more likely that high values can be reached by accounts that are active long enough, thus resulting in a heavy tail in the distribution. Instead, when temporal growth is slower the highest values are less likely to constitute a heavy tail. In fact, the assumptions behind the Double-Pareto generative model put forward by Reed are not

entirely met by Gowalla users: user activity spans are not strictly exponentially distributed and at each time step the distribution does not exactly follow a log-normal distribution. Furthermore, user activity values show increasing average values with constant standard deviation, instead of constant average with increasing standard deviation. Nonetheless, we see that the resulting distributions nicely fit the properties of Double-Pareto and log-normal distributions. Understanding how these differences in the temporal evolution affect the final distribution represents another interesting research question.

Our findings are only exploratory and generate a number of further questions: in particular, the real generative mechanisms behind user behavior are yet to be unraveled before user activity on location-based services can be accurately modeled and predicted. Our results may have an impact on how systems and applications built on top of social platforms perform: the presence of heavy tails in the distributions, or the lack thereof, denotes how much extreme conditions are likely to arise as a result of user activity. For instance, users with thousands of check-ins or with several thousands of friends which receive their real-time check-in notifications can impose a large stress on the storage and delivery infrastructure of the system: hence, understanding when such worst-case scenarios might take place becomes of paramount importance to service providers.

IV. Related Works

User activity on online social networks has already been under study [5], [6]. An investigation on MySpace user accounts showed that the number of friends also exhibits a Double-Pareto distribution, with exponential user account lifetimes and a temporal evolution consistent with the multiplicative generative model [2]. Our study builds on this result and extends it by considering for the first time a location-based online social network, which provides other types of user activity such as check-ins and visited places. Furthermore, while we find similar results with respect to user account spans, we also find different evolution patterns than those found in MySpace, showing how location-based services may exhibit different properties than standard online social networks.

Power-laws have been detected in several online and offline computer science scenarios. For instance, the relation between exponential growth and power-law behavior has been addressed by Huberman and Adamic [13], proposing that the power-law degree distribution of the WWW could be explained by the exponential growth of the number of web pages. Other generative models that explain how power-laws and log-normal distributions arise have been put forward by Reed [8], Mitzenmacher [9] and Clauset et al. [7]. We show how user behavior on Gowalla might obey similar mechanisms, discussing connections between these models and the observed distributions of user activity.

Location-based online social networks have been only recently under investigation: Scellato et al. have studied their spatial properties, describing how a considerable fraction of users exhibit a prevalence of short-distance spatially-clustered social ties and how such user heterogeneity appears driven by a mixture of social and geographic factors [14], [15]. In an earlier study, Liben-Nowell et al. demonstrated how geographic distance plays a key role in the probability of friendship between users of online social services and in the resulting navigability of such networks [16]. Our work, while also studying a popular online location-based social network, does not focus on geographic and spatial properties of social connections among online users, but instead presents the first investigation of their user activity on a broader sense.

V. Conclusions and Future Work

In this work we have presented a measurement study of user activity on a popular location-based online social network, Gowalla. We have highlighted differences in the distributions of friends, check-ins and places of Gowalla users and how user account lifetimes decay faster than exponentially. We have then investigated the temporal evolution of such distributions, noting that users add new friends at a faster rate than they accumulate new places and check-ins, discussing how these variations may influence the aggregated distributions.

While this work provides one of the first characterizations of user activity on the fastly rising location-based social services, there are many more interesting research questions that are elicited by our findings, such as the complete understanding of the generative mechanisms which rule user activity.

References

[1] GigaOM, "Foursquare Hits 4 Million Users," http://gigaom.com/2010/10/21/foursquare-hits-4-million-users/.

[2] B. Ribeiro, W. Gauvin, B. Liu, and D. Towsley, "On MySpace Account Spans and Double Pareto-Like Distribution of Friends," in *Proceedings of NetSciCom '10*, March 2010, pp. 1–6.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proceedings of ICWSM '10*, May 2010.

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, June 1998.

[5] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proceedings of SIGKDD '06*. New York, NY, USA: ACM, 2006, pp. 611–617.

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

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, no. 4, 2009.

[8] W. Reed, "The Pareto, Zipf and other power laws," *Economics Letters*, vol. 74, no. 1, pp. 15–19, December 2001.

[9] M. Mitzenmacher, "A Brief History of Generative Models for Power Law and Lognormal Distributions," *Internet Mathematics*, vol. 1, 2001.

[10] Gowalla, "SXSW 2010 Redux," http://blog.gowalla.com/post/977686051/sxsw-2010-redux.

[11] Wired, "Gowalla Tops Foursquare at SXSW Web Awards ," http://www.wired.com/underwire/2010/03/sxsw-web-awards/.

[12] M. B. Wilk and R. Gnanadesikan, "Probability plotting methods for the analysis of data," *Biometrika*, vol. 55, no. 1, pp. pp. 1–17, 1968.

[13] B. A. Huberman and L. A. Adamic, "Growth dynamics of the World-Wide Web," *Nature*, vol. 401, no. 6749, p. 131, 1999.

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

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

[16] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," *PNAS*, vol. 102, no. 33, pp. 11 623–11 628, August 2005.