How Do Online Social Networks Grow?

Konglin Zhu1, Wenzhong Li1,2, Xiaoming Fu1, Jan Nagler3,4

1 Institute of Computer Science, Georg-August-Universität Göttingen, Göttingen, Germany, 2 State Key Laboratory for Novel Software and Technology, Nanjing University, Nanjing, China, 3 Computational Physics, IfB, ETH Zurich, Zurich, Switzerland, 4 Max Planck Institute for Dynamics and Self-Organization (MPI DS), Göttingen, Germany

Abstract

Online social networks such as Facebook, Twitter and Gowalla allow people to communicate and interact across borders. In past years online social networks have become increasingly important for studying the behavior of individuals, group formation, and the emergence of online societies. Here we focus on the characterization of the average growth of online social networks and try to understand which are possible processes behind seemingly long-range temporal correlated collective behavior. In agreement with recent findings, but in contrast to Gibrat’s law of proportionate growth, we find scaling in the average growth rate and its standard deviation. In contrast, Renren and Twitter deviate, however, in certain important aspects significantly from those found in many social and economic systems. Whereas independent methods suggest no significance for temporally long-range correlated behavior for Renren and Twitter, a scaling analysis of the standard deviation does suggest long-range temporal correlated growth in Gowalla. However, we demonstrate that seemingly long-range temporal correlations in the growth of online social networks, such as in Gowalla, can be explained by a decomposition into temporally and spatially independent growth processes with a large variety of entry rates. Our analysis thus suggests that temporally or spatially correlated behavior does not play a major role in the growth of online social networks.

Introduction

Online social networks (OSNs) have become increasingly important as they allow us to interact across any geographical scale. Communication networks, transport networks and OSNs are often interconnected and interdependent. This opens up great economic and social opportunities but can also involve considerable risks such as cascading breakdowns [1]. The study of OSNs is of importance for understanding the behavior of individuals, groups and societies. Hence, particular types of growth in social, economic and other networked systems have attracted a lot of attraction in the past years [2–8].

Gibrat’s law states that both the average growth rate and the standard deviation of the growth rate of a given observable are constant and independent of the specific value of the observable [9]. However, this empirical law, originally observed in economic systems, has been challenged by many socio-economic studies [10,11], notably very recently [4,5,8].

Any social growth dynamics is expected to depend on social factors such as gender, age, social status and so forth. Unfortunately, available datasets that comprise such information are typically too small to investigate emergent scaling or large-scale collective behavior. In this paper, we focus on the population growth dynamics of three large OSNs. Our datasets do not resolve individual social factors but their size allows for studying scaling and long-range correlations, both temporally and spatially.

We find evidence for certain scaling laws in the growth rate and the variance, although for Renren and Twitter the exponents characterizing fluctuations are found to deviate from those that have been reported previously for social and economic systems.

These deviations carry important information about the growth of online social systems. In particular, we find that the relative number of registered users increases almost temporally and spatially independently of each other. This contrasts the behavior of offline growth in many social and economic systems where growth is a long-range correlated process and thus a collective phenomenon. Even for Gowalla where scaling indicates long-range correlated growth a decomposition into independent growth processes unravels the seemingly long-range collective behavior to be a mere artifact of the large variability of entry rates [12].

Data

We analyze three OSN datasets. The first OSN Renren (rr), often referred to as the “Chinese Facebook”, is one of the largest online social networks in China. The dataset covers about 1,000,000 users in the time period of January 2006 to December 2010 (60 months) with online interactions from 10,000 registered locations.

The second OSN data set, comes from a subset of Twitter (tt), a microblogging online social service sited in the United States. It covers more than 250,000 members between August 2006 and September 2010 (50 months) from about 9,000 locations.

The third OSN, Gowalla (gw), was an online check in social service launched in 2007 and closed in 2012 in the United States. Users were able to check in at certain locations, referred to as Spots, either through a mobile phone application or Gowalla’s mobile website. Among other things, checking-in allowed for the dropping or swapping of virtual items. The dataset covers 21 months from
February 2009 to October 2010) with around 200,000 members from about 5,000 locations.

We acquired the first two datasets by crawling user profiles in the web sites from Renren.com [13] and Twitter.com [14] through their APIs. We only crawled the user profiles which are publicly available. The Gowalla dataset is obtained from a shared data source [15] by other researchers. Due to the tremendous size of OSNs, we only acquired a sampled subset of each OSN. To eliminate sample bias we deployed the Breadth First Search (BFS) bias correction procedure by Kurant et al. [16].

For these three datasets we define a population at a location $l$ (an integer number ID) at time $t$ as the set of all users with home location $l$. The spatial resolution of the location refers to as a city code, associated with the administrative area (i.e. the city name) of the user’s home location. For Renren and Twitter we assume the registered location of the user as the user’s home location. For

Figure 1. Scaling in average growth rate and standard deviation. Both $\langle r(S_0) \rangle$ and $\sigma(r(S_0))$ as a function of the initial population size $S_0$ exhibit a power law, $\langle r(S_0) \rangle \sim S_0^{-\alpha}$, $\sigma(r(S_0)) \sim S_0^{-\beta}$. Renren: $\langle r(S_0) \rangle \sim S_0^{-0.07}$ ($R^2 = 0.326$), $\sigma(r(S_0)) \sim S_0^{-0.45}$ ($R^2 = 0.526$), that is, $\alpha_{ren} = 0.07$, $\beta_{ren} = 0.45$, Twitter: $\alpha_{tw} = 0.04$ ($R^2 = 0.142$), $\beta_{tw} = 0.37$ ($R^2 = 0.531$), Gowalla: $\alpha_{gw} = 0.09$ ($R^2 = 0.343$), $\beta_{gw} = 0.35$ ($R^2 = 0.625$). All values are obtained from MLE. Bootstrapping suggests 95% confidence for $\alpha > 0$ (violation of Gibrat’s law), and for $\beta_{gw} < 0.5$ (suggesting long-rang correlations). No statistical significance is found for $\beta < 0.5$ for Renren and Twitter. Vertical lines indicate 5% marks (insets).

doi:10.1371/journal.pone.0100023.g001

How Do Online Social Networks Grow?
Gowalla we assume the most visited location as the home location. For the spatial analysis we used an assignment of GPS coordinates (via Google Maps Application Programming Interfaces (API)) to the location \( l \), and calculated the distance between two locations via their GPS coordinates. The estimated GPS coordinates of a user’s home location may thus be incorrect for a certain fraction of a given population. This, however, may not alter any of the conclusions made in this article.

**Results**

Here, we investigate the mean growth rate and its fluctuation in OSN populations and ask the question how these observables depend on the initial population size.

We denote the population size, i.e. the number of users with home location index \( 1 \leq l \leq l_{\text{max}} \) at time \( 0 < t \leq T \), by \( S(t) \). Following Refs. [17–19] we define the logarithmic growth rate \( r \) between time \( t_0 \) and \( t_1 \) \( (t_0 < t_1 \leq T) \) as

\[
    r(t_0,t_1) = \ln \frac{S(t_1)}{S(t_0)}.
\]

where \( S_0 = S(t_0) \) and \( S_1 = S(t_1) \) are the population size at a location \( l \) but at different time points \( t_0 \) and \( t_1 \) [5].

To characterize fluctuations, we study the average growth rate \( \langle r(S_0) \rangle \) and the standard deviation

\[
    \sigma(r(S_0)) = \sqrt{\langle (r(S_0) - \langle r(S_0) \rangle)^2 \rangle}
\]

as a function of the initial population size \( S_0 \), see Figure 1. In other words, the average growth rate \( \langle r(S_0) \rangle \) corresponds to only those online populations with size at least \( S_0 \) until time \( t_0 \). The conditional standard deviation of the growth rate \( \sigma(r(S_0)) \) for those populations expresses the statistical spread or fluctuation of growth among populations with \( S_0 \). Both quantities show a power law dependence on the initial population size.
with positive exponents $\alpha, \beta > 0$ ($P < 0.05$), which suggests a deviation from the independence of Gibrat’s law that would imply $\sigma_{\text{Gibrat}} = \beta_{\text{Gibrat}} = 0$. Scale-invariant growth instead of Gibrat’s proportionate growth has been reported for economic systems such as firms [10] and countries ($\beta = 0.15 - 0.18$) [20], research and development expenditures at universities ($\beta \approx 0.25$) [21], scientific output ($\beta = 0.28 - 0.40$) [22], and more recently for city population growth ($\beta = 0.19 - 0.27$) [4] and online communities ($\beta = 0.15 - 0.22$) [5].

The range of $\beta$ for Renren and Twitter is in agreement with those previously reported exponents for $\beta$. However, in contrast to the previous work mentioned above our analysis (employing maximal likelihood estimation (MLE) and bootstrapping) do not suggest significant deviations from $\beta = 0.5$ which would indicate uncorrelated growth. In contrast, for Gowalla we find $\beta$ significantly smaller than $\beta = 0.5$ ($P < 0.05$).

Second, we find the range of exponents for the average growth rate for all studied online social networks significantly above $\alpha = 0$ ($P < 0.05$) indicating a violation of Gibrat’s proportionate growth, which is in agreement with social and economic systems [4,5,10,20–22].

The average growth rate $\langle r(S_0) \rangle$ and the conditional standard deviation of the growth rate $\sigma(r(S_0))$ allow for direct comparison with the literature for other social systems and Gibrat’s law but are only averages. As suggested by studies of certain assets in economical systems the distribution of the variance can often exposes important information that cannot be seen in averages [23].

Since for a given $S_0$ there is only a single value of $\sigma(r(S_0))$ (see Fig. 1), we ask what is the relative variation of $\sigma(r(S_0))$ across all values of $\sigma(r(S_0))$ that occur in a given dataset. We thus focus on the relative fluctuation function (rff),

$$\text{rff}(S_0) = \frac{\sigma(r(S_0))}{\langle r(S_0) \rangle} \sum_{S_0} \sigma(r(S_0'))$$

as a function of $S_0$. Specifically, we study the complementary cumulative relative fluctuation function (ccrff), which is given by the complement of the integrated rff,

$$\text{ccrff}(S_0) = 1 - \sum_{S_0 \leq S_0} \text{rff}(S_0').$$

We chose the ccrff representation because it shows (if exists) a clearer scaling than the rff and thus better exposes different (scaling) regimes. The ccrff is obtained by collecting all locations with a given value of $S_0$ using exponential binning (see Fig. 2 and Methods).

In contrast to Renren and Twitter where we find no significant bimodality, for Gowalla the ccrff as a function of $S_0$ exhibits a remarkable bimodal behavior.

Figure 3. Gowalla: Temporal short- and long-term correlations in the population growth rate. Short-term correlations for $S_0 < S^*$ (log-lin plots), $C(t) \sim e^{-\gamma t}$, and long-term correlations for $S_0 \geq S^*$ (log-log plots), $C(t) \sim t^{-\delta}$. Fits using MLE suggest $\gamma = 0.13$ ($R^2 = 0.992$), $\delta = 0.73$ ($R^2 = 0.955$); log-log-scaling for determining the coefficient of determination for the power law, and log-linear-scaling for the exponential. doi:10.1371/journal.pone.0100023.g003

Figure 4. Gowalla: Decomposition of the growth into independent short-term correlated population growth processes. ACF for the three data sets according to the superposition scheme explained in the text. The power law exponents from fitting $C(t) \sim t^{-\delta}$ obtained from the decomposition via surrogate data (sur). Best fits from MLE: $\delta \approx 0.59$ ($R^2 = 0.986$). doi:10.1371/journal.pone.0100023.g004
For Gowalla, Figure 2C suggests a bimodal distribution of standard deviations, characterized by an exponential decay that is followed by a power law

\[ c_{\text{ccfrf}}(S_0) = \begin{cases} e^{-\mu S_0} & (S_0 < S^*) \\ S_0^{-\delta} & (S_0 \geq S^*) \end{cases}, \tag{6} \]

where \( S^* = 93 \) for Gowalla marks the crossover point (determined from MLE, see Methods). MLE suggests that the power law decay is characterized by the exponent \( n_{gw} \sim 0.13 \) \((R^2 = 0.966)\).

Gowalla: Correlations in the growth rate

The above findings suggest to consider two groups of locations: one group with initial population size \( S_0 < S^* \), and the other one with initial population size \( S_0 \geq S^* \). We study the monthly population growth rates for each location and calculate their autocorrelation function (ACF) [24–26]. For \( S_0 < S^* \) the ensemble averaged ACF exhibits an exponential decay, \( C(t) \sim e^{-\gamma t} \), indicating that the population growth is short-term correlated, see Figure 3A. We obtain the exponent \( \gamma = 0.13 \) \((R^2 = 0.992)\) from MLE, which is equivalent to a correlation time constant of about two weeks.

In contrast, for \( S_0 \geq S^* \) the ACF is well described by a power law, \( C(t) \sim t^{-\delta} \) with power law exponent \( \delta = 0.73 \) \((R^2 = 0.955)\), see Figure 3B.

This is consistent with long-term correlations characterized by \( \sum_{t=-\infty}^{\infty} C(t) = \infty \), see [27] and references therein.

Superposition model

Seemingly long-range correlation can often be explained by a finite set of independent processes whose superposition accounts for the algebraic decay in the ACF, and the divergence of its infinite sum. In 1979 van der Ziel established that any ensemble of uncoupled short-range correlated stochastic oscillators is sufficient for explaining long-range correlations in their superposition, if and only if the time constants of the mixed processes are sufficiently broadly distributed [12]. More recently, it has been shown that a superposition of Poisson processes, together with circadian activity, very likely account for many scaling laws of human activity patterns [28]. Here, as the growth rates are broadly distributed, we follow this spirit by considering a superposition of populations and surrogate time series from these, see Methods.

Gowalla’s population growth of the superposition ensemble obtained from a random selection of population with \( S_0 < S^* \) results in the occurrence of seemingly long-term correlations for

![Figure 5. Spatial independence of the population growth rates.](image-url)
locations with $S_0 \geq S^*$. The exponents $\delta^{\text{sur}}$ for the surrogate superpositions (sur) are obtained from fitting the superposition ensemble averaged ACF by MLE (see Methods) with $R^2 = 0.986$, $\delta^{\text{sur}} = 0.59$ ($R^2 = 0.986$, $\delta = 0.73$), see Figure 4.

This suggests that the seemingly long-term correlated population growth found for locations with $S_0 \geq S^*$ results from superpositions of short-term correlated growing populations.

Spatial dependence
To study geographical factors we investigate correlations of the populations growth rates $r_i$ and $r_j$ between different places [29,30].

We therefore study the Pearson's correlation coefficient

$$c = \frac{\langle r_{ij} - \langle r_i \rangle \rangle (r_j - \langle r_j \rangle) \rangle}{\sigma_i \sigma_j}. \quad (7)$$

where $\sigma_{ij} = \sqrt{\langle r_{ij} - \langle r_i \rangle \rangle ^2}$ is the standard deviation of $r_i$ and $r_j$, respectively.

We investigate the monthly population growth rates and Pearson’s correlation coefficient between a pair of locations as a function of the geographic distance of the users. Figure 5 shows the Pearson’s correlation coefficients for the three data sets. The average correlation $\langle c \rangle$ is found at a level of about $0.7-0.8$, effectively independent of the geographic distance. The high value of the cross correlation agrees well with the plausible assumption that individuals join online social networks collectively but independently of the geographic distance to each other.

**Discussion**

We find scaling in the population growth rate and variance in online social networks. Our results suggest that the population growth in online social networks is neither significantly determined by population size [31] nor by spatial factors. The results deviate from Gibrat’s law as previously found in many social and economic systems. The seemingly long-term correlated growth behavior for Gowalla suggested by scaling in the standard deviation is explained by a simple decomposition into short-term correlated population growth with broadly distributed growth rates. Our method may help interpreting (seemingly) long-range correlations in the growth of large heterogenous (online) social and economic systems. Seemingly collective behavior in online social systems may result from the high variability of loners’ actions and not from correlated collective behavior.

**Methods**

**Ethics statement**

We use the APIs that provided by Renren.com and Twitter.com for data collection from these two websites. The acquirement of Renren and Twitter datasets is in accordance with the websites’ terms of service.

**Data availability statement**

We use three datasets in this article. The Renren and Twitter datasets can be obtained upon the request, which is “data available on request”. The request can be send to the Computer Networks Group at University of Goettingen via email (net@cs.uni-goettingen.de). The Gowalla dataset is obtained from a shared data source [15] by other researchers. The requester can download it from snap.stanford.edu, which is “data available from online”.

**Exponential binning**

Fitting of average, standard deviation and the ccrff is performed by exponential binning, by which the bins are evenly distributed on a logarithmic scale. Specifically, the beginning of each bin is $b_j = \lfloor c R^j \rfloor$, exponentially increasing in $j$, with constants $c$ and $R > 1$, so that bins have size $b_{j+1} - b_j \approx b_j (R-1)$. We use exponential binning for both Figure 1 (with $c=1$, $R=2$) and Figure 2 (with $c=1$, $R=1.2$).

**Choice of $t_0$ and $t_1$**

The datasets are analyzed within a time window given by the time points $t_0$ and $t_1$. $t_1$ is chosen as the end point of the data set, $t_1 = T$. For the choice of $t_0$ we consider two factors: the number of populated locations in the time window and the size of the time window. A too small $t_0$ would lead to only a few populated locations whereas any large $t_0$ would reduce the width of the
window. Following the methodology of studies in human population growth in the real world [4] and the human interaction activities in OSNs [5], we determined $t_0$ as a result from time when the number of locations with growing populations reaches the peak. That is, $t_0 = 35$ for Renren, $t_0 = 36$ for Twitter and $t_0 = 14$ for Gowalla, respectively, see Figure 6.

**Determination of $S^*$**

To determine the best value for $S^*$, we fit the distribution of standard deviation with respect to $S_0$ ranging from 0 to 300 by using MLE. For each $S_0$, we calculate $R^2$ for exponential and power-law fitting, denoted as $R^2_{exp}$ and $R^2_{pow}$, respectively. To characterize the overall fitting quality ($FQ$) we use

$$FQ = R^2_{exp} R^2_{pow} \quad (8)$$

where we use log-log-scaling for determining the coefficient of determination for the power law, and log-linear-scaling for the exponential.

We choose $S^* = \text{argmax}(FQ(S_0))$ where $FQ$ takes its maximum at the value of $S^* = 93$ for Gowalla.

**Spatially resolved monthly growth rates**

For each location with integer ID $I$, we extract a time series from $t_0$ to $t_1$ of the monthly population growth rate according to equation (1) as $r_t = \ln \frac{X}{X_0}$, $0 \leq t \leq t_1$ being the $t$th month. We calculate the autocorrelation function (ACF) from the time series $r_t$ as

$$C(\tau) = \frac{\langle (r_t - \langle r_t \rangle)(r_{t+\tau} - \langle r_{t+\tau} \rangle) \rangle}{\sigma_r^2} \quad (9)$$

where $\tau$ is the time lag and $\sigma_r$ is the standard deviation of $r_t$.

**Superposition construction**

To study superpositions we select all populations at locations with $S_0 < S^*$. The randomized surrogate data set is created by shuffling these entries, and creating a time series from these shuffled entries as follows.

(1) From the set of populations with $S_0 < S^*$ we select randomly a population and add up its initial population size $S_0$ irrespective of its location. (2) We repeat (1) until the sum exceeds $S^*$. This results in a set of locations whose total initial population size equals or slightly exceeds $S^*$. We call this set of locations one realization of a superposition. (3) For each realization we study the temporal development with respect to total populations size of the thereafter fixed selected locations. For each superposition we construct a time series, that is, the population growth rates in monthly resolution, from $t_0$ to $t_1$. For this set of time series we obtain the ensemble averaged ACF.

**Author Contributions**

Conceived and designed the experiments: KZ WL XF JN. Performed the experiments: KZ. Analyzed the data: KZ. Contributed reagents/materials/analysis tools: KZ WL XF JN. Wrote the paper: KZ WL XF JN.

**References**

1. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464: 1025–1028.
2. Kossinets G, Watts DJ (2006) Empirical Analysis of an Evolving Social Network. Science 311: 88–90.
3. Palla G, lszl Barabsi A, Vicsek T, Hungary B (2007) Quantifying social group evolution. Nature 446: 2007.
4. Rozenfeld HD, Rybski D, Andrade JS Jr, Batty M, Stanley HE, et al. (2008) Laws of population growth. Proceedings of The National Academy of Sciences of the United States of America 105: 18762–18767.
5. Rybski D, Buldyrev SV, Havlin S, Liljeros F, Makse HA (2009b) Scaling laws of human interaction activity. Proceedings of the National Academy of Sciences of the United States of America 106: 12649–12655.
6. Rozenfeld HD, Rybski D, Gabaix X, Makse HA (2011) The area and population of cities: New insights from a different perspective on cities. American Economic Review 101: 2205–223.
7. Rybski D, Garcia Cantu Ros A, Kropp JP (2013) Distance-weighted city growth. Physical Review E 87: 042114.
8. Gallos LK, Rybski D, Liljeros F, Havlin S, Makse HA (2012) How people interact in evolving online affiliation networks. Physical Review X 2: 031014.
9. Gibrat R (1931) Les inégalités économiques. Libraire du Recueil Sirey, Paris.
10. Stanford MH, Amaral LA, 232 Buldyrev SV, Havlin S, Leschhorn H, et al. (1996) Scaling behaviour in the growth of companies. Nature 379: 804–806.
11. Santarelli E, Klopff L, Thurnik AR (2006) Gibrat’s law: An overview of the empirical literature. In: Entrepreneurship, growth, and innovation, Springer. pp. 41–73.
12. van der Ziel A (1988) Unified presentation of 1/f noise in electronic devices: fundamental 1/f noise sources. Proceedings of the IEEE 76: 253–258.
13. Renren net. Available: http://www.renren.com. Accessed 2014 May 27.
14. Twitter. Available: http://www.twitter.com. Accessed 2014 May 27.
15. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ’11). New York, NY, USA: ACM, pp. 1082–1090.
16. Kurant M, Markopoulou A, Thiran P (2011) Interactions in evolving online affiliation networks. Physical Review E 87: 042114.
17. Levy M (2009) Gibrat’s law for (all) cities: Comment. American Economic Review 99: 1672–75.
18. Ioannides Y, Overman H (2003) Zipf’s law for cities: an empirical examination. Regional Science and Urban Economics 33: 127–137.
19. Matias K, Nunes Amaral LA, Luwel M, Moed HF, Stanley HE (2005) Scaling phenomena in the growth dynamics of scientific output. Journal of the American Society for Information Science and Technology 56: 893–902.
20. Mantegna RN, Stanley HE (2000) Introduction to econophysics: correlations and complexity in finance. Cambridge university press.
21. Dunn P (2010) Measurement and Data Analysis for Engineering and Science. CRC Press/Taylor & Francis.
25. Beran J, Sherman R, Taqqu M, Willinger W (1995) Long-range dependence in variable-bit-rate video traffic. IEEE Transactions on Communications 43: 1566–1579.
26. Samorodnitsky G (2007) Long range dependence. Foundations and Trends in Stochastic Systems 1: 163–257.
27. Hennig H, Fleischmann R, Fredebohm A, Hagmayer Y, Nagler J, et al. (2011) The nature and perception of fluctuations in human musical rhythms. PloS one 6: e26457.
28. Malmgren R, Stouffer D, Motter A, Amaral L (2008) A poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences of the United States of America 105: 18153–18158.
29. Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. The American Statistician 42: pp. 59–66.
30. Nikolic D, Muresan RC, Feng W, Singer W (2012) Scaled correlation analysis: a better way to compute a cross-correlogram. European Journal of Neuroscience 35: 742–762.
31. Easley D, Kleinberg J (2010) Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.