Scaling laws of human interaction activity

Diego Rybski[a], Sergey V. Buldyrev[b], Shlomo Havlin[c], Fredrik Liljeros[d], and Hernán A. Makse[a,1]

[a]Levich Institute and Physics Department, City College of New York, New York, NY 10031; [b]Department of Physics, Yeshiva University, New York, NY 10033; [c]Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel; and [d]Department of Sociology, Stockholm University, S-10691 Stockholm, Sweden

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved June 2, 2009 (received for review March 13, 2009)

Even though people in our contemporary technological society are depending on communication, our understanding of the underlying laws of human communicational behavior continues to be poorly understood. Here we investigate the communication patterns in 2 social Internet communities in search of statistical laws in human interaction activity. This research reveals that human communication networks dynamically follow scaling laws that may also explain the observed trends in economic growth. Specifically, we identify a generalized version of Gibrat's law of social activity expressed as a scaling law between the fluctuations in the number of messages sent by members and their level of activity. Gibrat's law has been essential in understanding economic growth patterns, yet without an underlying general principle for its origin. We attribute this scaling law to long-term correlation patterns in human activity, which surprisingly span from days to the entire period of the available data of more than 1 year. Further, we provide a mathematical framework that relates the generalized version of Gibrat's law to the long-term correlated dynamics, which suggests that the same underlying mechanism could be the source of Gibrat's law in economics, ranging from large firms, research and development expenditures, gross domestic product of countries, to city population growth. These findings are also of importance for designing communication networks and for the understanding of the dynamics of social systems in which communication plays a role, such as economic markets and political systems.

growth | Gibrat's law | long-term correlations | memory | network growth

The question of whether unforeseen outcomes of social activity follow emergent statistical laws has been an acknowledged problem in the social sciences since at least the last decade of the 19th century (1–4). Earlier discoveries include Pareto's law for income distributions (5), Zipf's law initially applied to word frequency in texts and later extended to firms, cities and others (6), and Gibrat's law of proportionate growth in economics (7–9).

Social networks are permanently evolving and Internet communities are growing more each day. Having access to the communication patterns of Internet users opens the possibility to unveil the origins of statistical laws that may lead us to the better understanding of human behavior as a whole. In this paper, we analyze the dynamics of sending messages in 2 Internet communities in search of statistical laws of human communication activity. The first online community (OC1) is mainly used by the group of men who have sex with men (MSM).* The data consists of over 80,000 members and more than 12.5 million messages sent over the course of 63 days. The target group of the second online community (OC2) is teenagers (10). The data covers 492 days of activity with more than 500,000 messages sent among almost 30,000 members. Both web sites are also used for social interaction in general. All data are completely anonymous, lack any message content, and consist only of the times at which the messages are sent and the identification numbers of the senders and receivers.

The act of writing and sending messages is an example of an intentional social action. In contrast to routinized behavior, the actants are aware of the purpose of their actions (2, 3). Nevertheless, the emergent properties of the collective behavior of the actants are unintended. In Fig. 1*A*, we show a typical example of the activity of a member of OC1 depicting the times when the member sends messages. Figure 1*B* provides the cumulative number of messages sent (green curve) compared with a random surrogate dataset (brown curve) obtained by shuffling the data, as discussed in *Materials and Methods*. As would be expected, there are large fluctuations in the members' activity when compared with a random signal (11–13, 15). The messages sent at random display small temporal fluctuations, whereas the OC1 member, sends many more messages in the beginning and a lot fewer at the end of the period of data acquisition (as also seen in Fig. 1*C*, displaying the number of messages sent per day). Although such extreme events or bursts have been documented for many systems, including e-mail and letter post communication, instant messaging, web browsing and movie watching (11–15), their origin is still an open question.

Results

Growth in the Number of Messages

The cumulative number $m_j(t)$ expresses how many messages have been sent by a certain member j up to a given time t [for better readability, we will not write the index j explicitly, $m(t)$; see details on the notation in the *SI Appendix*, Sec. I]. The dynamics of $m(t)$ between times t_0 and t_1 within the period of data acquisition T ($t_0 < t_1 \leq T$) can be considered as a growth process, where each member exhibits a specific growth rate r_j (r for short notation):

$$r \equiv \ln \frac{m_1}{m_0}, \qquad [1]$$

where $m_0 \equiv m(t_0)$ and $m_1 \equiv m(t_1)$ are the number of messages sent until t_0 and t_1, respectively, by every member. To characterize the dynamics of the activity, we consider 2 measures. (*i*) The conditional average growth rate, $\langle r(m_0)\rangle$, quantifies the average growth of the number of messages sent by the members between t_0 and t_1 depending on the initial number of messages, m_0. In other words, we consider the average growth rate of only those members who have sent m_0 messages until t_0 (see *Materials and Methods* for more details). (*ii*) The conditional standard deviation of the growth rate for those members who have sent m_0 messages until t_0, $\sigma(m_0) \equiv \sqrt{\langle (r(m_0) - \langle r(m_0)\rangle)^2\rangle}$, expresses the statistical spread or fluctuation of growth among the members depending on m_0. Both quantities are relevant in the context of Gibrat's law in economics (7–9) which proposes a proportionate growth process, entailing the assumption that the average and the standard deviation of the growth rate of a given economic indicator are constant and independent of the specific indicator value. That is, both $\langle r(m_0)\rangle$ and $\sigma(m_0)$ are independent of m_0 (9).

In Fig. 2 *A* and *B*, we show the results of $\langle r(m_0)\rangle$ and $\sigma(m_0)$ versus m_0 for both online communities. We find that the

Author contributions: D.R., S.V.B., S.H., F.L., and H.A.M. designed research; D.R. performed research; D.R., S.V.B., S.H., F.L., and H.A.M. analyzed data; and D.R., S.V.B., S.H., F.L., and H.A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. E-mail: hmakse@lev.ccny.cuny.edu.

*The study of the de-identified MSM dating site network data was approved by the Regional Ethical Review Board in Stockholm, record 2005/5:3.

This article contains supporting information online at www.pnas.org/cgi/content/full/0902667106/DCSupplemental.

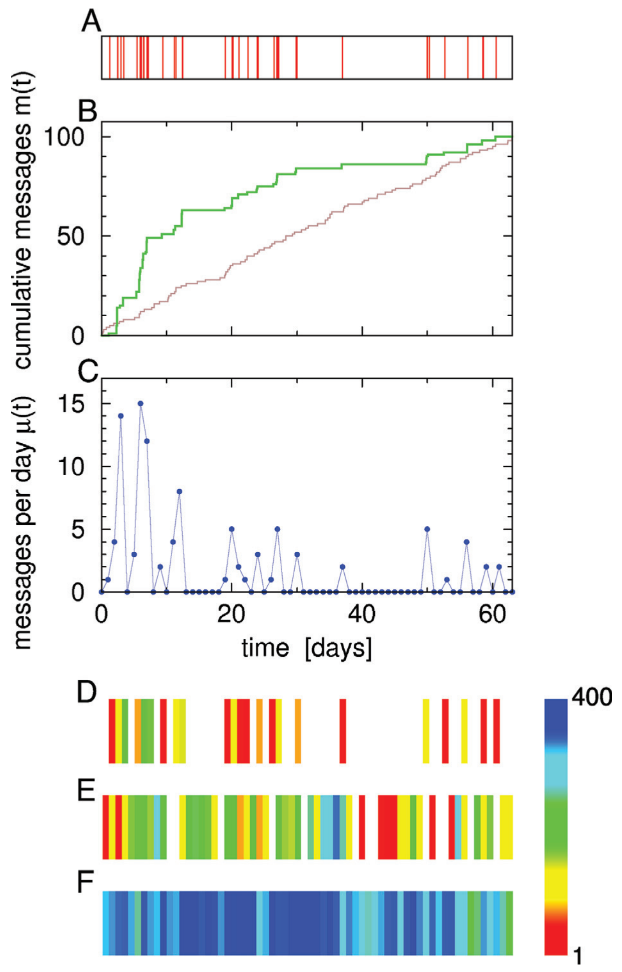

Fig. 1. A typical example of an individuals' message activity. (*A*) Instants at which messages were sent by a member belonging to OC1. (*B*) Cumulative number of messages $m(t)$ (green) and the same but with the messages placed at random (brown). (*C*) Sequence of number of messages sent per day, $\mu(t)$, for the same individual. (*D–F*) Color-coded sequences $\mu(t)$ for members sending $M = 100$; 1,000; or 10,000 messages overall, respectively. The color is proportional to the logarithm of the number of messages per day (red, 1 message; blue, 400 messages; white, no message).

conditional average growth rate is fairly independent of m_0. On the other hand, the standard deviation decreases as a power-law of the form:

$$\sigma(m_0) \sim m_0^{-\beta}. \qquad [2]$$

We obtain by least-square fitting the exponents $\beta_{\mathrm{OC1}} = 0.22 \pm 0.01$ for OC1 and $\beta_{\mathrm{OC2}} = 0.17 \pm 0.03$ for OC2 (the values deviate slightly for large m_0 due to low statistics). Although the web sites are used by different member populations, the power law and the obtained exponents are quite similar. The exponents are also close to those reported for growth in economic systems such as firms and countries ($0.15 - 0.18$) (16), research and development expenditures at universities (0.25) (17), scientific output ($0.28 - 0.4$) (18), and city population growth ($0.19 - 0.27$) (19). The approximate agreement between the exponents obtained for very different systems (social or of human origin) can be considered as a generalization of Gibrat's law, suggesting that the mechanisms behind the growth properties in different systems may originate in the human activity represented by Eq. **2**.

Figure 2*C* and *D* depicts the results when we randomize the data of OC1 and OC2, respectively (see *Materials and Methods* for details of the randomization procedure), such that any temporal correlations are removed. The typical dynamics for such surrogate data set are shown in Fig. 1*B* (the brown curve), which displays a clear random pattern of small fluctuations in comparison with the original data of larger fluctuations (green curve). We find that the random signal displays a close to constant average growth rate $\langle r(m_0)\rangle$ and that the fluctuations behave as in Eq. **2** but with an exponent $\beta_{\mathrm{rnd}} = 1/2$ (Fig. 2 *C* and *D*). The origin of this value has a simple explanation: If an isolated individual randomly flips an ideal coin with no memory of the previous attempt, then the fluctuations from the expected value of the fraction of obtained heads decay as a square-root of the number of throws, implying $\beta_{\mathrm{rnd}} = 1/2$. In contrast to randomness, here we hypothesize that the origin of the generalized version of Gibrat's law with $\beta < 1/2$ in Eq. **2** is a nontrivial long-term correlation in communication activity. These correlations possibly arise from internal and external stimuli from other members transmitted through the highly connected network of individuals, an effect that is absent in the randomized data. The exponent value of $\beta \approx 0.2$ for OC1 and OC2 implies that the fluctuations of very active members are smaller than the ones of less active members, but they are significantly larger compared with the random case (compare Fig. 2 *A* and *B* with Fig. 2 *C* and *D*).

Long-Term Correlations

The exceptional quality of the data (more than 10 million messages spanning several effective decades of magnitude in terms of both activity and time) allows to test the above hypothesis by investigating the presence of temporal correlations in the individuals' activity. We aggregate the data to records of messages per day (an example is shown in Fig. 1*C*) to avoid the daily cycle in the activity and analyze the number of messages sent by individuals per day, $\mu(t)$, where t denotes the day [$m(t) \equiv \sum_{t'=1}^{t} \mu(t')$, Fig. 1 *D–F* shows the color coded daily activity of 3 members in OC1]. For every member, we obtain a record of a length of 63 days (OC1) or 492 days (OC2). We note that former studies reporting Eq. **2** (such as refs. 16–19) typically were not based on data with temporal resolution as we use it here and therefore were not able to investigate its origin in terms of temporal correlations.

We quantify the temporal correlations in the members' activity by mapping the problem to a 1-dimensional random walk. The quantity $Y(t) \equiv \sum_{t'=1}^{t}(\mu(t') - \langle\mu(t)\rangle)$, where $\langle\mu(t)\rangle$ is the average of the corresponding record $\mu(t)$, represents the position of the random walker that performs an up or down step given by $\mu(t') - \langle\mu(t)\rangle$ at time step t'. The correlations after Δt steps are reflected in the behavior of the root-mean-square displacement $F(\Delta t) \equiv \sqrt{\langle[Y(t+\Delta t) - Y(t)]^2\rangle}$ (20), where $\langle\cdot\rangle$ is the average over t and members. If the activity $\mu(t)$ is uncorrelated or short-term correlated, then one obtains $F(\Delta t) \sim (\Delta t)^{1/2}$, Fick's law of diffusion, after some cross-over time. In the case of long-term correlations, the result is a power-law increase

$$F(\Delta t) \sim (\Delta t)^{H}, \qquad [3]$$

where $H > 1/2$ is the fluctuation exponent [also known as Hurst exponent (20)]. In statistical physics, long-term correlation or persistence is also referred to as long-term "memory". Because, in general, the records might be affected by trends, we use the standard detrended fluctuation analysis (DFA) (21) to calculate H (see *SI Appendix*, section III for a detailed description).

The results for OC1 are shown in Fig. 3 *A* and *B*, where we calculate Eq. **3** by separating the members in groups with different total number of messages sent by the members, M. We find that $F(\Delta t)$ asymptotically follows a power law with $H \approx 1/2$ for the less active members who sent fewer than 10 messages in the entire period ($M < 10$). The dynamics of the more active members display clear long-term correlations. We find that the fluctuation exponent increases to $H \approx 0.75$ for members with $M > 10^3$ (see Fig. 3*B*). The smaller value of H for less active members could be due to the small amount of information that these members

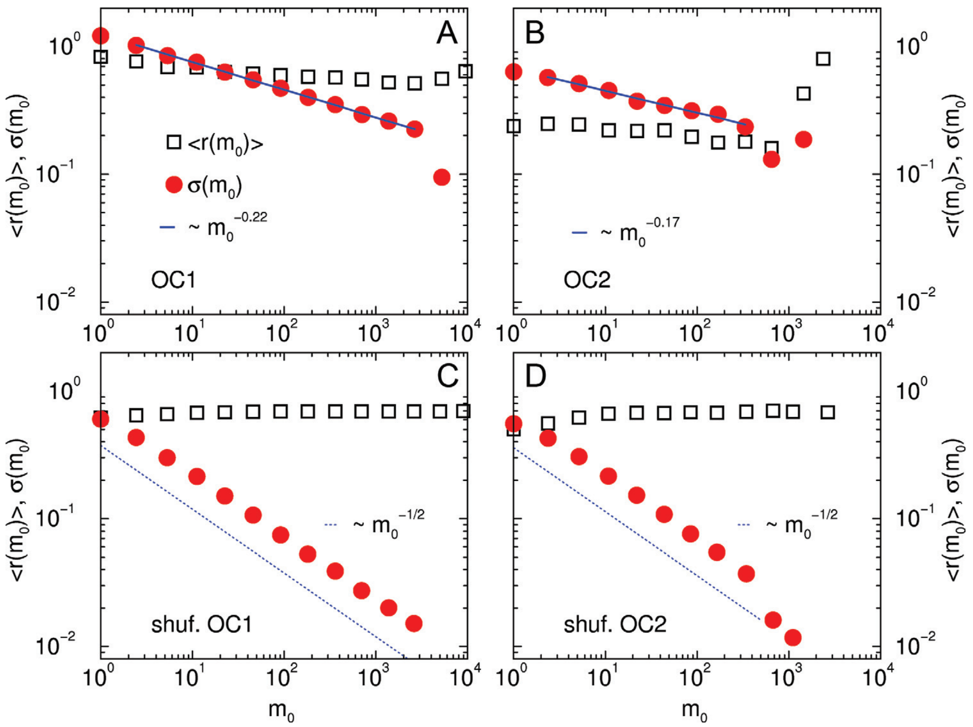

Fig. 2. Average and standard deviation of the growth rate versus number of messages. (*A*) Results for OC1. The average growth rate of messages conditional to m_0 is almost constant and the standard deviation decays with an exponent $\beta_{\rm OC1} = 0.22 \pm 0.01$. (*B*) Results for OC2. The standard deviation conditional to m_0 decays with an exponent $\beta_{\rm OC2} = 0.17 \pm 0.03$. (*C*) Results for OC1, when the messages are shuffled, displaying $\beta_{\rm rnd} = 1/2$. (*D*) Results for OC2, when the messages are shuffled. In all cases, t_0 corresponds to half of the period of data acquisition and t_1 to the end, which we found to provide optimal statistics (see *SI Appendix*, Fig. 1).

provide in the available time of data acquisition. When we shuffle the data to remove any temporal correlations, we obtain the random exponent $H_{\rm rnd} = 1/2$ (as seen in Fig. 3*B*), confirming that the correlations in the data are due to temporal structure.

The dynamics of the message activity in OC2 is similar to OC1 (see Fig. 3*C*). On large time scales, we measure the fluctuation exponent increasing from $H \approx 1/2$ to $H \approx 0.9$, with increasing M (the exponents for very active members are based on poor statistics and therefore carry large error bars). Analogous to the results obtained for OC1, there are no correlations in the shuffled records ($H_{\rm rnd} = 1/2$ in Fig. 3*D*). The fact that $H > 1/2$ means that a sudden burst in activity of a member persists on times scales ranging from days to years. The distribution of activity is self-similar over time. Similar correlation results have been found in traded values of stocks and e-mail data (22).

Relation Between β and H

Next, we elaborate the mathematical framework that relates the growth process Eq. **2** to the long-term correlations, Eq. **3**. To relate the exponent from Eq. **2**, β, to the temporal correlation exponent γ, from Eq. **4**, and therefore to H, one can first rewrite Eq. **1** as

$$r = \ln\frac{m_1}{m_0} = \ln\frac{m_0+\Delta m}{m_0} \qquad \text{with } \Delta m = m_1 - m_0$$
$$= \ln\left(\frac{\Delta m}{m_0}+1\right) \approx \frac{\Delta m}{m_0} \qquad \text{for small } \frac{\Delta m}{m_0}.$$

Next, the total increment of messages Δm is expressed in terms of smaller increments $\mu(t)$, such as messages per day:

$$\Delta m = \sum_{t=t_0+1}^{t_0+\Delta t} \mu(t),$$

which is (assuming stationarity) statistically equivalent to $\Delta m = \sum_{t=1}^{\Delta t}\mu(t)$, and one can write $r \approx \frac{1}{m_0}\sum_{t=1}^{\Delta t}\mu(t)$ for the growth rate. The conditional average growth is then

$$\langle r(m_0)\rangle = \left\langle \frac{1}{m_0}\sum_{t=1}^{\Delta t}\mu(t)\right\rangle \approx \frac{1}{m_0}\sum_{t=1}^{\Delta t}\langle\mu(t)\rangle.$$

Then the conditional standard deviation $\sigma(m_0) = \sqrt{\langle[r(m_0)-\langle r(m_0)\rangle]^2\rangle}$ can be written in terms of the autocorrelation function as follows:

$$r(m_0) - \langle r(m_0)\rangle = \frac{1}{m_0}\left(\sum_{t=1}^{\Delta t}\mu(t) - \sum_{t=1}^{\Delta t}\langle\mu(t)\rangle\right)$$
$$[r(m_0) - \langle r(m_0)\rangle]^2 = \frac{1}{m_0^2}\left(\sum_{t=1}^{\Delta t}(\mu(t) - \langle\mu(t)\rangle)\right)^2$$
$$\langle[r(m_0) - \langle r(m_0)\rangle]^2\rangle \approx \frac{1}{m_0^2}\sum_{i}^{\Delta t}\sum_{j}^{\Delta t}\sigma_\mu^2 C(j-i),$$

where $C(\Delta t) = \frac{1}{\sigma_\mu^2}\langle[\mu(t) - \langle\mu(t)\rangle][\mu(t+\Delta t) - \langle\mu(t)\rangle]\rangle$ is the autocorrelation function of $\mu(t)$ and σ_μ is the standard deviation of $\mu(t)$. The autocorrelation function $C(\Delta t)$ measures the interdependencies between the values of the record $\mu(t)$. For uncorrelated values, $C(\Delta t)$ is zero for $\Delta t > 0$ because, on average, positive and negative products of the record will cancel out each other. In the case of short-term correlations, $C(\Delta t)$ has a characteristic decay time, $\Delta t_\times$. A prominent example is the exponential decay $C(\Delta t) \sim \exp(-\Delta t/\Delta t_\times)$. Long-term correlations are described by a slower decay, namely a power-law

$$C(\Delta t) \sim (\Delta t)^{-\gamma}, \qquad \textbf{[4]}$$

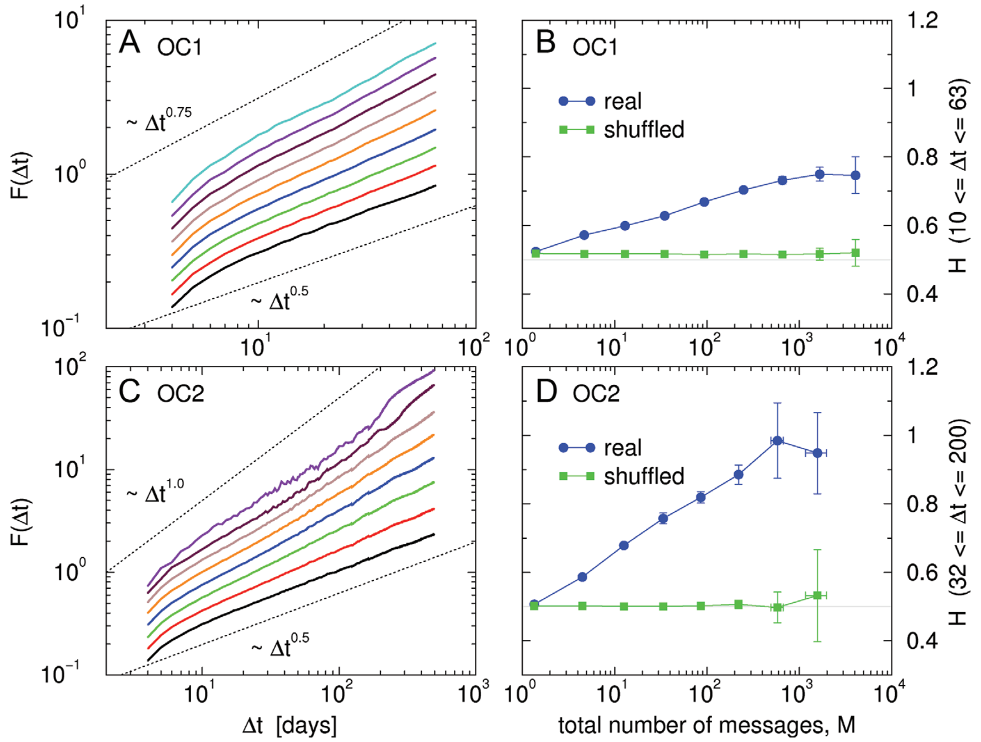

Fig. 3. Long-term correlations in the message activity of OC1 (*A* and *B*) and OC2 (*C* and *D*). (*A*) DFA fluctuation functions averaged conditional to *M*, the total number of messages sent by each member (black, 1-2; red, 3-7; green, 8-20; blue, 21-54; orange, 55-148; brown, 149-403; maroon, 404-1096; violet, 1097-2980; turquoise, 2981-8103). The dotted lines serve as guides, the one in the bottom corresponds to the uncorrelated case, whereas the one in the top corresponds to the exponent 0.75. (*B*) Fluctuation exponent *H* measured from *A* on the scales 10 days $\leq \Delta t \leq$ 63 days as a function of the total number of messages sent, *M*, for real (blue) and individually shuffled (green) records. (*C*) DFA fluctuation functions averaged conditional to *M* (colors as in *A*). The dotted lines correspond to the uncorrelated case (bottom) and to the exponent 1 (top). (*D*) Fluctuation exponents obtained from *C* on the scales 32 days $\leq \Delta t \leq$ 200 days as a function of the total number of messages sent, *M*. Due to weak statistics causing large error bars, we do not consider the last 2 values for $M > 500$ as reliable. For clarity, the fluctuation functions in *A* and *C* are shifted vertically.

with the correlation exponent $0 < \gamma < 1$, which is related to the fluctuation exponent H from Eq. **3** by $\gamma = 2 - 2H$ (20). We note that $\gamma = 1$ (or $\gamma > 1$) corresponds to an uncorrelated record with $H = 1/2$. A key property of long-term correlations is a pronounced mountain–valley structure in the records (20). Statistically, large values of $\mu(t)$ are likely to be followed by large values and small values by small values. Ideally, this feature holds on all time scales, which means a sequence in daily, weekly, or monthly resolution is correlated in the same way as the original sequence.

Assuming long-term correlations asymptotically decaying as in Eq. **4**, we approximate the double sum with integrals and obtain

$$\langle [r(m_0) - \langle r(m_0)\rangle]^2\rangle \approx \frac{1}{m_0^2}\sigma_\mu^2 \iint_1^{\Delta t} (j-i)^{-\gamma} \mathrm{d}j\mathrm{d}i \sim \frac{1}{m_0^2}\sigma_\mu^2(\Delta t)^{2-\gamma}.$$

In order to relate Δt and m_0, one can use $\Delta t = x t_0$ [where x is an arbitrary (small) constant that simply states how large Δt is compared with t_0], and $m_0 \sim t_0$, which states that the number of messages is proportional to time assuming stationary activity. By using these 2 arguments we obtain

$$\langle [r(m_0) - \langle r(m_0)\rangle]^2\rangle \approx \frac{1}{m_0^2}\sigma_\mu^2(x)^{2-\gamma}(t_0)^{2-\gamma} \sim \sigma_\mu^2 m_0^{-\gamma},$$

$$\sigma(m_0) \sim \sigma_\mu m_0^{-\gamma/2}.$$

Comparing this equation with Eq. **2**, we finally obtain $\beta = \gamma/2$, and, with $\gamma = 2 - 2H$,

$$\beta = 1 - H. \qquad \textbf{[5]}$$

Eq. **5** is a scaling law formalizing the relation between growth and long-term correlations in the activity and is confirmed by our data. For OC1, we measured $\beta_{\mathrm{OC1}} \approx 0.22$ yielding $H_{\mathrm{OC1}} \approx 0.78$ from Eq. **5**, which is in approximate agreement with the (maximum) exponent we obtained by direct measurements for OC1 ($H = 0.75 \pm 0.05$ from Fig. 3*B*). For OC2, we obtained $\beta_{\mathrm{OC2}} \approx 0.17$ and therefore $H_{\mathrm{OC2}} \approx 0.83$ through Eq. **5**, which is not too far from the (maximum) exponent found by direct measurements for OC2 ($H = 0.88 \pm 0.03$). According to Eq. **5**, the original Gibrat's law ($\beta_{\mathrm{G}} = 0$) corresponds to very strong long-term correlations with $H_{\mathrm{G}} = 1$. This is the case when the activity on all time scales exhibits equally strong correlations. In contrast, $\beta_{\mathrm{rnd}} = 1/2$ represents completely random activity ($H_{\mathrm{rnd}} = 1/2$), as obtained for the randomized data in Fig. 3 *B* and *D*.

The mathematical framework relating long-term correlations quantified by H and the growth fluctuations quantified by β could be relevant to other complex systems. While the generalized version of Gibrat's law has been reported for economic indicators displaying $\beta \approx 0.2$ (16–18), the origin of this scaling law is not clear and is still being investigated. Our results suggest that the value of β could be explained by the existence of long-term correlations in the activity of the corresponding system ranging from firms and markets to social and population dynamics. In turn, Eq. **5** establishes a missing link between studies of growth processes in economic or social systems (16–18) and studies of long-term correlations, such as in finance and the economy (23), Ethernet traffic (24), and human brain (25) or motor activity (26). Our results foreshadow the possibility that systems involving other types of human interaction, such as various Internet activities, communication via cell phones, trading activity, etc., may display similar growth and correlation properties as found here, offering the possibility of explaining their dynamics in terms of the long-term persistence of individuals' behavior.

Growth of the Degree in the Underlying Social Network

Communication among the members of a community represents a type of a social interaction that defines a network, whereas a message is sent either based on an existing relation between 2 members or establishing a new one. There is considerable interest in the origin of broad distributions of activity in social systems. Two paradigms have been invoked for various applications in social systems: the "rich-get-richer" idea used by Simon in 1955 (27) and the models based on optimization strategies as proposed by Mandelbrot (28). Regarding network models, the preferential attachment (PA) model has been introduced (29) to generate a type of stochastic scale-free network with a power-law degree distribution in the network topology. Considering the social network of members linked when they exchange at least 1 message (that has not been sent before), we examine the dynamic of the number of outgoing links of each member [the out-degree $k(t)$] in analogy to Eq. **2**.

We start from the empty set of nodes consisting of all the members in the community and chronologically add a directed link between 2 members when a messages is sent. In analogy to the growth in the number of messages $m(t)$ of each member, we study the growth of the members' out-degree $k(t)$, i.e. the number of links to others. We define the growth rate of every member as

$$r_k = \ln \frac{k_1}{k_0}, \qquad [6]$$

where $k_0 \equiv k(t_0)$ is the out-degree of a member at time t_0 and $k_1 \equiv k(t_1)$ is the out-degree at time t_1. Again, there is a growth rate for each member j, but for better readability we skip the index. In Fig. 4, we study $\langle r_k(k_0)\rangle$, the average growth rate conditional to the initial out-degree k_0, and $\sigma_k(k_0)$, the standard deviation of the growth rate conditional to the initial out-degree k_0 for OC1 and OC2. We obtain almost constant average growth $\langle r_k(k_0)\rangle$ as a function of k_0 as in the study of messages.

The conditional standard deviation of the network degree, $\sigma_k(k_0)$, is shown in Fig. 4 for both social communities. We obtain a power-law relation analogous to Eq. **2**:

$$\sigma_k(k_0) \sim k_0^{-\beta_k}, \qquad [7]$$

with β-exponents very similar to those found for the number of messages, namely $\beta_{k,\mathrm{OC1}} = 0.22 \pm 0.02$ for OC1 and $\beta_{k,\mathrm{OC2}} = 0.17 \pm 0.08$ for OC2. These values are consistent with those we obtained for the activity of sending messages.

Next, we consider the preferential attachment model, which has been introduced to generate scale-free networks (29) with power-law degree distribution $P(k)$ of the type investigated in the present study. Essentially, it consists of subsequently adding nodes to the network by linking them to existing nodes that are chosen randomly with a probability proportional to their degree. We consider the undirected network and study the degree growth properties using Eqs. **6** and **7** and calculate the conditional average growth rate $\langle r_{\mathrm{PA}}(k_0)\rangle$ and the conditional standard deviation $\sigma_{\mathrm{PA}}(k_0)$. The times t_0 and t_1 are defined by the number of nodes attached to the network. Fig. 2 in the *SI Appendix*, section IV shows the results where an average degree $\langle k\rangle = 20$; 50,000 nodes in t_0 and 100,000 nodes in t_1 were chosen. We find constant average growth rate that does not depend on the initial degree k_0. The conditional standard deviation is a function of k_0 and exhibits a power-law decay characterized by Eq. **7** with $\beta_{\mathrm{PA}} = 1/2$. The value $\beta_{\mathrm{PA}} = 1/2$ in Eq. **5** corresponds to $H = 1/2$, indicating complete randomness. There is no memory in the system. Because each addition of a new node is completely independent from precedent ones, there cannot be temporal correlations in the activity of adding links. Therefore, purely preferential attachment type of growth is not sufficient to describe the social network dynamics found in the present study and further temporal correlations have to be incorporated according to Eq. **3**.

For the PA model, it has been shown that the degree of each node grows in time as $k(t) \sim (\frac{t}{t^*})^b$, where t^* is the time when the corresponding node was introduced to the system and $b = 1/2$ is the dynamics exponent in growing network models (30). Accordingly, the growth rate is given by $r_{\mathrm{PA}} = b \ln \frac{t_1}{t_0}$, which is a constant independent of k_0, in accordance with our numerical findings. Furthermore, in the *SI Appendix*, section IV we analytically obtain the exponent $\beta_{\mathrm{PA}} = 1/2$ and confirm the numerical results as well. Interestingly, an extension of the standard PA model has been proposed (31) that takes into account different fitnesses of the nodes to acquiring links involving a distribution of b-exponents and therefore a distribution of growth rates. This model opens the possibility to relate the distribution of fitness values to the fluctuations in the growth rates, a point that requires further investigation.

Discussion

From a statistical physics point of view, the finding of long-term correlations opens the question of the origin of such a persistence pattern in the communication. At this point, we speculate on 2 possible scenarios that require further studies. The question is whether the finding of an exponent $H > 0.5$ is due to a power-law (Levy-type) distribution (32, 33) in the time interval between 2 messages of the same person or just from pure correlations or long-term memory in the activity of people. In the first scenario, the intervals between the messages follow a power law (13, 34). Accordingly, the activity pattern comprises many short intervals and few long ones, implying persistent epochs of small and large activity. This fractal-like activity leads to long-term correlations with $H > 1/2$ (see the analogous problem of the origin of

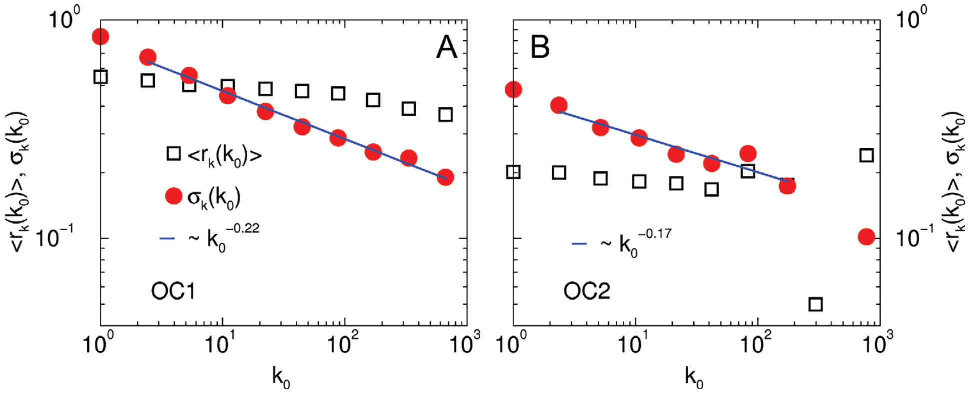

Fig. 4. Mean out-degree growth rate and standard deviation versus initial out-degree. (*A*) Results for OC1. The average growth of out-degree conditional to the out-degree at t_0 is almost constant. The standard deviation decays with an exponent $\beta_{k,\mathrm{OC1}} = 0.22 \pm 0.02$. (*B*) Results for OC2. The standard deviation conditional to the out-degree at t_0 decays with an exponent $\beta_{k,\mathrm{OC2}} = 0.17 \pm 0.08$. The quantities are analogous to those of Fig. 2 except that here the growth rate of the out-degree r_k is considered instead of the number of messages sent.

long-term correlations in DNA sequences as discussed in ref. 33). This scenario implies a direct link between the correlations and the distribution of interevent intervals which can be obtained analytically. In the second scenario, the intervals between the messages do not follow a Levy-type distribution, but the value of the time intervals are not independent of each other, again representing long-term persistence. For example, the distribution of interevent times could be stretched exponential [see recent work on the study of extreme events of climatological records exhibiting long-term correlations (35)]. Thus, deciding between these 2 possible scenarios for the origin of correlations in activity requires an extended analysis of interevent intervals as well as correlations to determine whether the behavior is Levy-like or pure memory-like. A careful statistical analysis is needed.

To some extent, the human nature of persistent interactions enables the prediction of the actants' activity. Our finding implies that traditional meanfield approximations based on the assumption that the particular type of human activity under study can be treated as a large number of independent random events (Poisson statistics) may result in faulty predictions. On the contrary, from the growth properties found here, one can estimate the probability for members of a certain activity level to send more than a given number of messages in the future. This result may help to improve the proper allocation of resources in communication-based systems ranging from economic markets to political systems. As a by-product, our finding that the activity of sending messages exhibits long-term persistence suggests the existence of an underlying long-term correlated process. This process can be understood as an unknown individual state driven by various internal and external stimuli (36, 37) providing the probability to send messages. In addition, the memory in activity found here could be the origin of the long-term persistence found in other records representing a superposition of the individuals' behavior, such as the Ethernet traffic (24), highway traffic, stock markets, and so forth.

Materials and Methods

Calculations of $\langle r(m_0)\rangle$, $\sigma(m_0)$ and Optimal Times t_0 and t_1. The average growth rate, $\langle r(m_0)\rangle$, and the standard deviation, $\sigma(m_0) = \sqrt{\langle r(m_0)^2\rangle - \langle r(m_0)\rangle^2}$, are defined as follows. Calling $P(r|m_0)$ the conditional probability density of finding a member with growth rate $r(m_0)$ with the condition of initial number of messages m_0, we obtain

$$\langle r(m_0)\rangle = \int r P(r|m_0)\,\mathrm{d}r \qquad [8]$$

and

$$\langle r(m_0)^2\rangle = \int r^2 P(r|m_0)\,\mathrm{d}r. \qquad [9]$$

In order to calculate the growth rate in Eq. **1**, one has to choose the times t_0 and t_1 in the period of data acquisition T. Naturally, it is best to use all data in order to have optimal statistics. Accordingly, t_1 is chosen best at the end of the available data ($t_1 = T$). We argue that if the choice of t_0 is too small, then $m(t_0)$ is zero for many members (those that send their messages later), which are then rejected in the calculation because of the division in Eq. **1**. Conversely, if the choice of t_0 is too large, then there is not enough time to observe the member's activity and $r = 0$ will occur frequently, indicating no change (members have sent their messages before). Thus, there must be an optimal time in between. In the *SI Appendix*, section II, Fig. 1, we plot, as a function of t_0, the number of members with at least 1 message at t_0 [$m_0 > 0$] and further exhibit at least some activity until $t_1 = T$ [$m_1 - m_0 > 0$]. For both online communities, we find an optimal t_0 in the middle of the period of observation $t_0 = T/2$, a value that is used for the analysis in the main text.

Shuffling of the Message Data. The raw data comprises 1 entry for each message consisting of the time when the message is sent, the sender identifier, and the receiver identifier. For example:

time	sender	receiver
1	a	b
2	a	c
4	b	a
6	c	d
7	a	b
…		

At $t = 1$ member a sends a message to member b, at $t = 2$ member a sends a message to member c, and so on.

The randomized surrogate dataset is created by randomly swapping the instants (time) at which the messages are sent between 2 events chosen at random. Thus, each message entry randomly obtains the time of another one, meaning that the total number of messages is preserved and the associations between them get shuffled. Temporal correlations are destroyed, but the set of instants at which the messages are sent remains unchanged. For instance, swapping events at $t = 1$ and $t = 6$ results in $t = 1$, c → d, and $t = 6$, a → b.

ACKNOWLEDGMENTS. We thank C. Briscoe, L.K. Gallos, and H.D. Rozenfeld for discussions. This work was supported by National Science Foundation Grant SES-0624116. F.L. acknowledges financial support from The Swedish Bank Tercentenary Foundation.

1. Merton RK (1936) The unanticipated consequences of purposive social action. *Am Sociol Rev* 1:894–904.
2. Weber M (1968) *Economy and Society*. (University of California Press, Berkeley), Vol. 1.
3. Giddens A (1993) *New Rules of Sociological Method*. (Stanford Univ Press, Stanford).
4. Durkheim É (1897) *Le suicide*. Étude de sociologie. (Les Presses Universitaires de France, Paris).
5. Pareto V (1896) *Cours d'Economie Politique*. (Droz, Geneva).
6. Zipf G (1932) *Selective Studies and the Principle of Relative Frequency in Language*. (Harvard Univ Press, Cambridge, MA).
7. Gibrat R (1931) *Les Inégalités Économiques*. (Libraire du Recueil Sierey, Paris).
8. Sutton J (1997) Gibrat's legacy. *J Econ Lit* 35:40–59.
9. Gabaix X (1999) Zipf's law for cities: An explanation. *Q J Econ* 114:739–767.
10. Holme P, Edling CR, Liljeros F (2004) Structure and time evolution of an Internet dating community. *Soc Networks* 26:155–174.
11. Paxson V, Floyd S (1995) Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Trans Networking* 3:226–244.
12. Dewes C, Wichmann A, Feldman A (2003) *Proceedings of the. 2003 ACM SIGCOMM Conference on Internet Measurement (IMC-03)*. (ACM Press, New York).
13. Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435:207–211.
14. Oliveira JG, Barabási AL (2005) Darwin and Einstein correspondence patterns. *Nature* 437:1251.
15. Zhou T, Kiet HAT, Kim BJ, Wang BH, Holme P (2008) Role of activity in human dynamics. *Europhys Lett* 82:28002.
16. Stanley MHR, et al. (1996) Scaling behaviour in the growth of companies. *Nature* 379:804–806.
17. Plerou V, Amaral LAN, Gopikrishnan P, Meyer M, Stanley HE (1999) Similarities between the growth dynamics of university research and of competitive economic activities. *Nature* 400:433–437.
18. Matia K, Amaral LAN, Luwel M, Moed HF, Stanley HE (2005) Scaling phenomena in the growth dynamics of scientific output. *J Am Soc Inf Sci Tec* 56:893–902.
19. Rozenfeld HD, et al. (2008) Laws of population growth. *Proc Nat Acad Sci USA* 105:18702–18707.
20. Feder J (1988) *Fractals, Physics of Solids and Liquids*. (Plenum Press, New York).
21. Peng CK, et al. (1994) Mosaic organization of DNA nucleotides. *Phys Rev E* 49:1685–1689.
22. Eisler Z, Bartos I, Kertész J (2008) Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv Phys* 57:89–142.
23. Mantegna RN, Stanley HE (1999) *An Introduction to Econophysics: Correlations and Complexity in Finance*. (Cambridge Univ Press, Cambridge, UK).
24. Leland WE, Taqqu MS, Willinger W, Wilson DV (1994) On the self-similar nature of ethernet traffic (extended version) *IEEE/ACM Trans Networking* 2:1–15.
25. Linkenkaer-Hansen K, Nikouline VV, Palva JM, Ilmoniemi RJ (2001) Long-range temporal correlations and scaling behavior in human brain oscillations. *J Neurosci* 21:1370–1377.
26. Ivanov PC, Hu K, Hilton MF, Shea SA, Stanley HE (2007) Endogenous circadian rhythm in human motor activity uncoupled from circadian influences on cardiac dynamics. *Proc Nat Acad Sci USA* 104:20702–20707.
27. Simon HA (1955) On a class of skew distribution functions. *Biometrika* 42:425–440.
28. Mandelbrot B (1953) *An Informational Theory of the Statistical Structure of Language*, ed Jackson W (Butterworth, London), pp 486–504.
29. Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512.
30. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97.
31. Bianconi G, Barabási, AL (2001) Competition and multiscaling in evolving networks. *Europhys Lett* 54:436–442.
32. Shlesinger MF, West BJ, Klafter J (1987) Lévy dynamics of enhanced diffusion: Application to turbulence. *Phys Rev Lett* 58:1100–1103.
33. Buldyrev SV, et al. (1993) Generalized Lévy-walk model for DNA nucleotide sequences. *Phys Rev E* 47:4514–4523.
34. Gerstein GL, Mandelbrot B (1964) Random walk models for spike activity of single neuron. *Biophys J* 4:41–68.
35. Bunde A, Eichner JF, Kantelhardt JW, Havlin S (2005) Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys Rev Lett* 94:048701.
36. Hedström P (2005) *Dissecting the Social: On the Principles of Analytical Sociology*. (Cambridge Univ Press, Cambridge, UK).
37. Kentsis A (2006) Mechanisms and models of human dynamics. *Nature* 441:E5–E6.