Supporting Information

Onnela and Reed-Tsochas 10.1073/pnas.0914572107

SI Text

SI Section 1: Background to Fluctuation Scaling. A recent article by Eisler, Bartos, and Kertesz (1) provides a good introduction to fluctuation scaling (FS). In temporal fluctuation scaling, we start from a multitude of M time series measured in the interval $[0,T]$ and assume that the constituents, that is, the random variables making up the signal, are additive. The signals are divided into blocks of duration Δt, and for any block in the interval $[t, t + \Delta t)$ the signal can be decomposed as

$$f_i^{\Delta t}(t) = \sum_{n=1}^{N_i^{\Delta t}(t)} V_{i,n}^{\Delta t}(t), \qquad [S1]$$

where $N_i^{\Delta t}(t)$ is the number of constituents within the block, that is, the number of random variables $V_{i,n}^{\Delta t}(t)$ to be summed together, of signal i during $[t, t + \Delta t)$. We assume that $V_{i,n}^{\Delta t}(t) \geq 0$, so that the time average of $f_i^{\Delta t}$, denoted by $\langle f_i^{\Delta t} \rangle$, does not vanish. It is defined as

$$\langle f_i^{\Delta t} \rangle = \frac{1}{Q} \sum_{q=0}^{Q-1} f_i^{\Delta t}(q\Delta t) = \frac{1}{Q} \sum_{q=0}^{Q-1} \sum_{n=1}^{N_i^{\Delta t}(q\Delta t)} V_{i,n}^{\Delta t}(q\Delta t), \qquad [S2]$$

where $Q = T/\Delta t$. For any Δt, the variance can be obtained as

$$\sigma_i^2(\Delta t) = \left\langle \left[f_i^{\Delta t} \right]^2 \right\rangle - \left\langle f_i^{\Delta t} \right\rangle^2. \qquad [S3]$$

This quantity characterizes the fluctuations of the activity of signal i from block to block. When f is positive and additive, it is often observed that the relationship between the SD σ_i (Δt) and the mean $\langle f_i^{\Delta t} \rangle$ is given by a power law

$$\sigma_i(\Delta t) \propto \langle f_i \rangle^{\alpha_T}, \qquad [S4]$$

where one varies i keeping Δt fixed. Note that the value of Δt does not affect the scaling, as it can be absorbed in the proportionality constant. The exponent α_T is in the range [1/2, 1], and the subscript T indicates that the statistical quantities are defined as temporal averages to distinguish them from ensemble fluctuation scaling (1).

In the main text, we discuss a more system-specific form of fluctuation scaling using spin variables $S_{i,j}(t)$ as constituent variables. Further, instead of having access to signals in continuous time, we consider, as a starting point, data sampled at discrete time intervals such that two consecutive time points t and $t + 1$ are separated by δt in physical time. The corresponding events in real physical time may have an arbitrary time resolution but, due to finite temporal sampling resolution, all events within one block may be considered concurrent.

SI Section 2: Example of Fluctuation Scaling. Let us consider a set of state or spin variables $S_{i,j}(t) \in \{-1,0,1\}$, one for each application i of every user. Here $S_{i,j}(t) = 1$ corresponds to user n adopting application i at time t, $S_{i,j}(t) = 0$ corresponds to there being no activity from user j regarding application i at time t, and $S_{i,j}(t) = -1$ corresponds to user j dropping application i at time t. The FS exponent α can be interpreted in terms of correlations between the constituent variables, in this case the spin variables $S_{i,j}(t)$. This leads to two limiting cases. If the constituent variables are uncorrelated, one obtains square-root scaling with $\alpha = 1/2$, whereas if the constituent variables are fully correlated, one obtains a linear scaling with $\alpha = 1$.

Two simple examples will illustrate this interpretation. Consider a variable $S_{i,j}(t)$ with the mean and variance given by $\langle S_i \rangle$ and $\Sigma_{S_i}^2$, respectively. If the random variables $S_{i,j}(t)$ are independent and identically distributed for all j and t, we obtain by the linearity of the expectation operator $E[\cdot]$ taken over time that

$$\mu_i = E[f_i(t)] = E\Big[\sum_{j=1}^{N} S_{i,j}(t) \Big] = NE\big[S_{i,j}(t)\big] = N\langle S_i \rangle. \qquad [S5]$$

The variance is given by

$$\sigma_i^2 = \mathrm{Var}[f_i(t)] = \mathrm{Var}\Big[\sum_{j=1}^{N} S_{ij}(t) \Big] = N\mathrm{Var}\big[S_{i,j}(t)\big] = N\Sigma_{S_i}^2, \qquad [S6]$$

because the variance of the sum of uncorrelated random variables (as follows from their independence) is the sum of their variances. Combining the expression for the mean and the variance gives $\sigma_i^2 = (\Sigma_{S_i}^2/\langle S_i \rangle)\mu_i$, so that $\alpha = 1/2$. The exponent $\alpha = 1/2$ is then a consequence of the central limit theorem and is reminiscent of the $1/\sqrt{N}$ fluctuations of extensive quantities, such as energy, in equilibrium statistical mechanics (1). On the other hand, if the random variables $S_{i,j}(t)$ are completely correlated, i.e. $S_{i,1}(t) = \cdots = S_{i,N}(t)$, we can write $\sum_{j=1}^{N} S_{ij}(t) = NS_{i,1}(t)$ which, as before, gives

$$\mu_i = NE\big[S_{i,1}(t)\big] = N\langle S_i \rangle \qquad [S7]$$

but now

$$\sigma_i^2 = \mathrm{Var}\big[NS_{i,1}(t)\big] = N^2\mathrm{Var}\big[S_{i,1}(t)\big] = N^2\Sigma_{S_i}^2, \qquad [S8]$$

resulting in $\sigma_i = (\Sigma_{S_i}^2/\langle S_i \rangle)\mu_i$, so that $\alpha = 1$. One way to produce $\alpha = 1$ is by a global driving force that imposes strong fluctuations that dominate over the local dynamics of the system (1).

SI Section 3: Stationarity of Time Series. The fact that for most applications $n_i(t)$ is an increasing function of time suggests that the system is not stationary and, consequently, violates the assumption on stationarity. The question then becomes whether the system is sufficiently close to stationarity so that the fluctuation scaling exponents can be interpreted in terms of correlations among the constituent variables. We can write

$$\mu_i \equiv \langle f_i(t) \rangle = \frac{1}{T_i} \sum_{t=1}^{T_i} f_i(t) = \frac{1}{T_i} \sum_{t=1}^{T_i} [n_i(t) - n_i(t-1)] = \frac{1}{T_i} \sum_{t=1}^{T_i} \sum_{j=1}^{N} S_{i,j}(t), \qquad [S9]$$

where the latter sum is taken over all N Facebook users and we have used $\sum_{j=1}^{N} S_{i,j}(t) = n_i(t) - n_i(t-1)$. Let us now assume that only irreversible $S_{i,j}(t) = 0 \rightarrow S_{i,j}(t + 1) = 1$ changes are possible. The validity of this assumption has mostly to do with the choice of the investigated time period. Facebook applications had just recently been introduced, there was less choice of and less competition between applications and, hence, dropping of applications was conceivably rather rare. Quantifying the extent of uninstallation of applications would, however, require access to microlevel data.

Instead of letting the sum indexed by j in the equation run over the entire system (over all users), we construct a restricted sum consisting of those users only who have not adopted application i earlier. This yields

$$\mu_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\sum_{j=1}^{N} S_{ij}(t) = \frac{1}{T_i}\sum_{t=1}^{T_i}\sum_{k=1}^{N-n_i(t)} S_{ij_k}(t), \qquad \text{[S10]}$$

where the subset of indices $j_1, j_2, \ldots, j_{N-n_i(t)} \in \{1,2, \ldots, N\}$ such that $S_{ij_k}(\tau) = 0$ for $\tau < t$.

The nonstationarity of $f_i(t)$ is reflected in the fact that the number of terms in the above sum, $N - n_i(t)$, depends on (typically decreases with) time. Although this is true for almost every application, it may be a problem only for highly popular applications, namely, in the high-density regime. Let us impose the stringent condition that the system is within the low-density regime, corresponding to the set of applications for which $f_i(t)$ are sufficiently close to stationarity, when at most 1% of users have the application. Within this regime, the number of terms in the last sum of Eq. S10 is always between $0.99N$ and N and, consequently, it decreases only marginally and the time series can be taken to be sufficiently stationary.

To see how far the low-density regime extends, we set $N - n^* = 0.99N$, giving an upper limit $n^* = N/100$. The number of users at the end of the time period is $n_i(T) = n_i(0) + \mu_i T \approx \mu_i T$, the approximation being rather good in the low-density regime, and we can assume that the approximate stationarity holds throughout the time horizon for applications with $n_i(T) \leq n*$, and setting $n^* = \mu^* T$ defines the low-density regime as $0 < \mu < \mu^*$ with $\mu^* = N/(100T)$. The stationarity can be expected to break down for applications with $\mu_i > \mu^* \approx 414$ so that $\log(\mu^*) \approx 2.6$. This means that, even under this relatively strict interpretation of stationarity, 97.8% of time series are stationary. This also means that the scaling in Fig. 2*C* holds for over two orders of magnitude *above* the cross-over point μ_x. We conclude that the system is sufficiently stationary so that the fluctuation scaling exponents for temporal fluctuations may be interpreted in terms of correlations between the constituent variables.

We can also relax the assumption about having only irreversible $S_{i,j}(t) = 0 \rightarrow S_{i,j}(t+1) = 1$ changes. Let $S_{i,j}(t) = 1$ correspond to user j adopting application i at time t, $S_{i,j}(t) = 0$ correspond to there being no activity from user j regarding application i at time t, and $S_{i,j}(t) = -1$ correspond to user j dropping application i at time t. Allowing $S_{i,j}(t) = -1$ means that the value of $\langle f_i \rangle$ may vanish or become negative. Of the $M = 2{,}705$ applications analyzed, 2,562 have positive $\mu_i > 0$, 5 have $\mu_i = 0$, and for 138 applications $\mu_i < 0$. Combining these numbers, we can see that 95% of the temporal averages μ_i are, in fact, positive and, consequently, nonnegativity does not pose a problem.

SI Section 4: Breakpoint Analysis for Linear Regression. Consider the linear regression model

$$y_i = x_i^T \beta_i + u_i, i = 1, \ldots, n \qquad \text{[S11]}$$

where y_i is observation i of the dependent variable, x_i is a $k \times 1$ vector of regressors with the first component set equal to unity, and β_i is a $k \times 1$ vector of regression coefficients that may vary over time. The null hypothesis is that the regression coefficients remain constant,

$$H_0 : \beta_i = \beta_0, i = 1, \ldots, n, \qquad \text{[S12]}$$

against the alternative hypothesis H_1 that at least one of the coefficients changes. In general, if there are m breakpoints, the regression coefficients are constant within the resulting $m + 1$ segments. The model can be rewritten to incorporate the breakpoints as

$$y_i = x_i^T \beta_j + u_i, i = i_{j-1} + 1, \ldots, i_j, j = 1, \ldots, m + 1, \qquad \text{[S13]}$$

where $\{i_1, \ldots, i_m\}$ are the set of breakpoints and j is the segment index. Conventionally, $i_0 = 0$ and $i_{m+1} = n$. Breakpoints are typically not given exogenously but need to be estimated from the data. Finding breakpoints in data is also known as testing for structural change in data, and there are two frameworks for doing that: F statistics and generalized fluctuation tests (2). Here we follow the F-statistics test that can be used to test against a single breakpoint, corresponding to the case with $m = 1$ in the above framework, at an unknown observation i_1 with segment $j = 1$ covering observations $i = 1, \ldots, i_1$ and segment $j = 2$ covering observations $i = i_1 + 1, \ldots, n$. To identify the breakpoint i_1, we compute a sequence of F statistics for a change at observation i given by

$$F_i = \frac{\hat{u}^T\hat{u} - \hat{u}(i)^T\,\hat{u}(i)}{\hat{u}(i)^T\,\hat{u}(i)/(n-2k)}, \qquad \text{[S14]}$$

where $\hat{u}$ are the ordinary least-squares residuals from the unsegmented (no breakpoint) model and $\hat{u}(i)$ are the ordinary least-squares residuals from a segmented model with a breakpoint at observation i, and the regression is carried out separately for each segment (2).

From the above definition it is clear that F_i is proportional to the residuals of the unsegmented model, $\hat{u}^T\hat{u}$, and inversely proportional to the residuals of the segmented model, $\hat{u}(i)^T\hat{u}(i)$. To ensure that each regression model can be estimated with a sufficient number of data points, we need to introduce a trimming parameter h such that we compute F_i for a subset of $i = h, h+1, \ldots, n-h$ observations. In practice, we can compute F_i for all $i = 1, \ldots, n$ and simply ignore the resulting values of F_i for very small and very large values of i, where a suitable value of h is chosen by the practitioner. The null hypothesis H_0 is rejected if the maximum value of F is "large" (2). What precisely it means for F to be large depends on the context. In any case, what matters is the relative height and narrowness of the maximum value of F with respect to all of the other values: A peak that is high and narrow is stronger evidence of a structural change in data than a peak that is low and wide.

The results are shown in Fig. S1. The data have been sorted in ascending order based on the x variable such that $\mu_{(1)} \leq \mu_{(2)} \leq \ldots \leq \mu_{(M)}$. In the case of empirical data, the F statistic behaves smoothly and develops a clear maximum. This is strong evidence of there being a structural change in the data such that the two regimes to the left and right of the breakpoint are governed by different exponents, $\alpha_I \approx 0.55$ and $\alpha_C \approx 0.85$, respectively.

The behavior of the F statistic for the synthetic data, however, is qualitatively very different. Instead of a smooth, single maximum, the error landscape is more rugged, and the maximum appears to be degenerate. Strictly speaking, there is a single maximum at $F_{(k)} \approx 186$ for observation $k = 562$, corresponding to $\log(\mu_{(562)}) \approx -0.38$, but there is also a secondary maximum for $k \approx 1{,}800$. The lack of a clearly defined maximum suggests that there is no sufficient statistical evidence to introduce a breakpoint in the data. Note that the above framework does not allow introducing multiple breakpoints. Although this could be done in principle by adding more degrees of freedom (more parameters), it becomes exceedingly difficult to justify them, especially if the differences in the slopes are very small. To demonstrate this, consider accepting the view that there is, in fact, a legitimate breakpoint at $k = 562$ in the synthetic data. This results in two exponents, $\alpha \approx 0.84$ and $\alpha \approx 0.87$, which are so close to one another that it is difficult to justify theoretically their slightly different values. We conclude, given these considerations, that the behavior of the synthetic data is governed by just a single exponent $\alpha_S \approx 0.84$.

SI Section 5: Supporting Data Analysis. Let us define the *total activity* as $F(t) = \sum f_i(t)$, where the sum runs over all applications that are in existence at time t. The total activity $F(t)$, which is not to be confused with the F statistic in *SI Section 4*, corresponds to the total number of applications installed in the 1-hour interval

between t and $t-1$. We show $F(t)$ in Fig. S2, where the daily 24-hour period of activity is clearly visible.

It is possible that, for a given application, the mean and SD of activity f_i result from the application being at a certain stage of its lifetime. Consequently, given that we have a mixture of old and new applications, if the scaling of SD of f_i with the mean of f_i were dependent in a discontinuous manner on the age of the application, this could in principle contribute to the cross-over reported in the main text. To test this hypothesis, we define the time-shifted activity for application i as $g_i(\tau) = f_i(t_i + \tau)$ with $\tau \geq 0$, where t_i is the (approximate) introduction time of application i. The time-shifted aggregated numbers $n_i(t_i + \tau)$ are shown in the upper panel of Fig. S3, and the time-shifted activities $g_i(\tau)$ are in the lower panel. We can now compute the mean and SD of the time-shifted activity $g_i(\tau)$ by truncating the time series at various values of τ, that is, by taking the first τ points of the time series since birth. We define an ensemble average of the time-shifted activities taken over all $N(\tau)$ applications that have a lifetime of at least τ as

$$g(\tau) = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} \frac{1}{\tau} \sum_{t=1}^{\tau} g_i(t). \qquad \text{[S15]}$$

Similarly, we can define the ensemble SD of the time-shifted activities as

$$h(\tau) = \frac{1}{N(\tau)} \sum_{i=1}^{N(\tau)} \left[\frac{1}{\tau - 1} \sum_{t=1}^{\tau} (g_i(t) - \langle g_i(\tau) \rangle)^2 \right]^{1/2}, \qquad \text{[S16]}$$

where $\langle g_i(\tau) \rangle = (1/\tau) \sum_{t=1}^{\tau} g_i(t)$. We plot $h(\tau)$ versus $g(\tau)$ for a number of different truncation points $\tau \in 50,60,\ldots,1{,}000$ in Fig. S4. A linear fit describes their dependence very well, and demonstrates that the relationship between the mean and the SD for the ensemble of applications does not depend in a discontinuous manner on the age of the application, that is, the stage of the application in its lifetime. The fact that the dependence of $h(\tau)$ on $g(\tau)$ holds throughout the measured lifetime of applications demonstrates that the cross-over in the fluctuation scaling plot in the main text cannot be explained by having a mixture of applications that are at different stages of their lifetime. Finally, we repeat the fluctuation scaling plot in Fig. S5, this time using only applications that have lifetimes $50 \leq \tau_i \leq T$ such that they were introduced during the first $T - 50$ time steps, corresponding to $t_i \in [0, T - 50]$, such that for each application we have at least 50 points for estimating the first and second moments. The result is essentially identical to the one presented in the main text. In particular, the high-μ applications are still present, as is the cross-over (fits not shown). This demonstrates explicitly that the high-μ regime is not simply produced by applications that have a large number of installations for $t < 0$, that is, before the start of data collection.

SI Section 6: Facebook and Facebook Applications. Here we provide a brief description of the platform (Facebook) and the studied cultural products (Facebook applications).

Facebook. Facebook is a social networking website operated by Facebook. At the time of data collection (June 25 to August 14, 2007), Facebook had approximately 50 million active users worldwide. However, at the time of writing, the site had more than 500 million users worldwide, reflecting the quickly growing popularity of the site. Since September 2006, anyone age 13 and over with a valid email address has been able to become a Facebook user. Facebook users, in line with other social networking sites, can construct a public or semipublic profile within a bounded system, articulate a list of other users, "Facebook friends," with whom they share a connection, and view and traverse their list of connections and those made by others within the system. Users can add friends and send them messages, and update their personal profiles to notify friends about themselves. Facebook has changed its interface several times, and the functionality and behavior of the site have also changed throughout its lifetime, and there is every reason to believe that this evolution will also continue in the future.

Facebook applications. Facebook launched a framework for software developers to create applications that interact with core Facebook features on May 24, 2007. Facebook introduced several applications, allowing users to send virtual gifts to each other, post free classified ads, inform their friends about upcoming events, and more. Some applications include interactivity, which allows users to play games with their friends. The moves made during the game are saved on the website, allowing the next move to be made at any later time.

At the time of the data collection, Facebook users could access an applications page from their profile which, among other things, provided a list of all available applications rank-ordered by their popularity. This allowed Facebook users, at any time, to access an exhaustive list of existing applications. In addition to this "global signal," reflecting the aggregate decisions of the user population, Facebook users could visit the profiles of their friends, which gave them an unobstructed view of the applications installed by their friends. In addition, when a user logged onto Facebook, the system would first present him or her with a "news feed," which includes status updates from friends as well as information about their application installations (but not uninstallations), a practice that was subsequently discontinued. These two factors, profile browsing and news feeds, make up the "local signal."

Adoption of Facebook applications. In our study, we view Facebook applications as cultural products or technological innovations. We emphasize that we had no data on the behavior of individuals. Instead, we had records on the cumulative number of application installations for each application at multiple points in time. These data enabled us to monitor every single application installation in the system with 50 million potential adopters; that is, no application installation went unobserved, except for 1% of applications which were discarded due to data-quality issues. Any Facebook user could install applications, and although they were free of charge, it seems that most users avoided installing too many. One plausible reason for this behavior is that having too many applications would easily clutter the profile, making some of the other information less visible to friends viewing the profile.

To give a better idea of the spread of Facebook applications, we show additional examples of cumulative application adoption curves in Fig. S6. We have divided the examples into different panels by hand based on the qualitative nature of their behavior. For example, applications in the upper left panel seem to have saturated their growth, and perhaps most closely resemble the classic S-shaped adoption curves as described, for example, by the logistic function. The other curves either show distinct "bumps" in their growth (upper right panel), or continue growing at different rates (lower panels). Note that pure exponential growth would appear as a straight upward-sloping line in semilogarithmic plots like these. The bumpiness of the curves is most likely related to the structure of the social network, in particular its community structure (3, 4), underlying the adoption behavior. The shapes of the adoption curves underscore the importance of taking the structures of social connections into account in the study of diffusion and influence processes.

1. Eisler Z, Bartos I, Kertesz J (2008) Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv Phys* 57:89–142.
2. Zeileis A, Kleiber C, Krämer W, Hornik K (2003) Testing and dating of structural changes in practice. *Comput Stat Data Anal* 44:109–123.
3. Fortunato S (2010) Community detection in graphs. *Phys Rep* 486:75–174.
4. Porter MA, Onnela JP, Mucha PJ (2009) Communities in networks. *Not Amer Math Soc* 56:1082–1164–10971166.

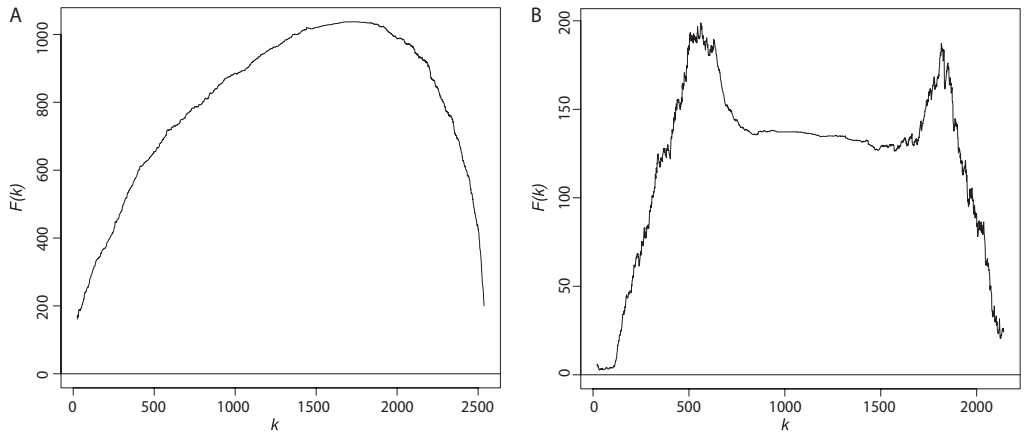

Fig. S1. *F* statistics for breakpoint analysis. (*A*) The *F* statistic is smooth and well-behaved for the empirical data and reaches a maximum of $F_{(k)} \approx 1{,}035$ for observation $k = 1{,}759$. This maximum corresponds to a breakpoint at log $(\mu_{(1,759)}) \approx 0.36$ and separates the data into two regimes characterized by exponents $\alpha_I \approx 0.55$ and $\alpha_C \approx 0.85$ to the left and right of the point, respectively. (*B*) The *F* statistic for the synthetic data are very rugged and the resulting maximum is effectively degenerate. This irregular behavior of the *F* statistic violates the underlying assumption of having a well-defined maximum and, consequently, does not provide sufficient statistical evidence for introducing a breakpoint in the data.

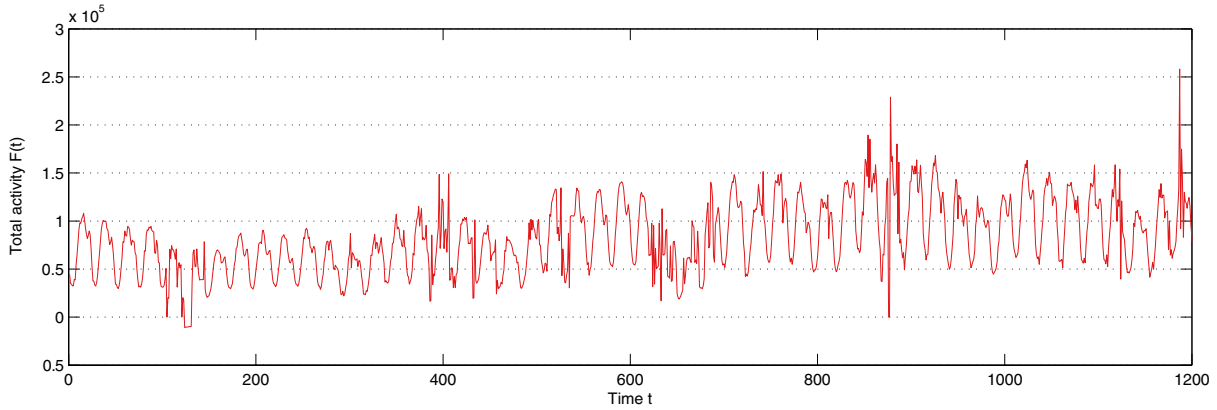

Fig. S2. Total activity $F(t)$ as a function of time t, where a unit of time is one observation, corresponding to calendar time from June 25 to August 14, 2007.

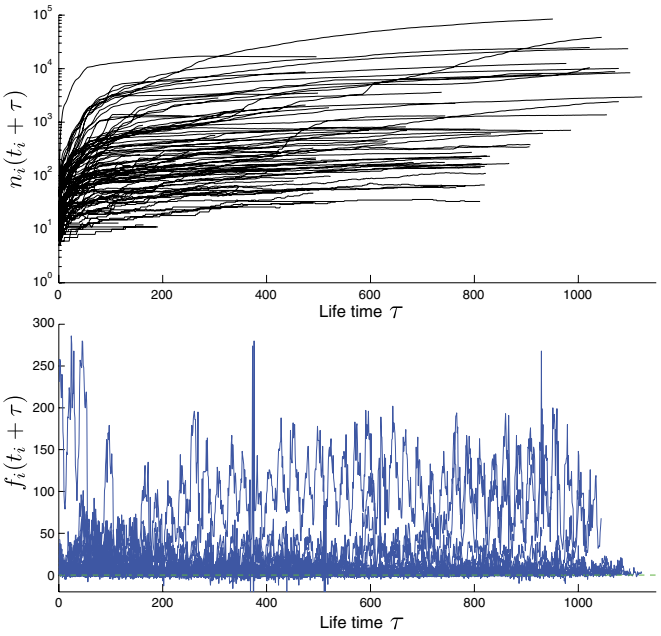

Fig. S3. Time-shifted aggregate numbers of application installations $n_i(t_i + \tau)$ as a function of application lifetime τ (*Upper*) and the related time-shifted activity values $g_i(\tau)$ (*Lower*). For purposes of visualization, in both plots are included a subset of 100 applications.

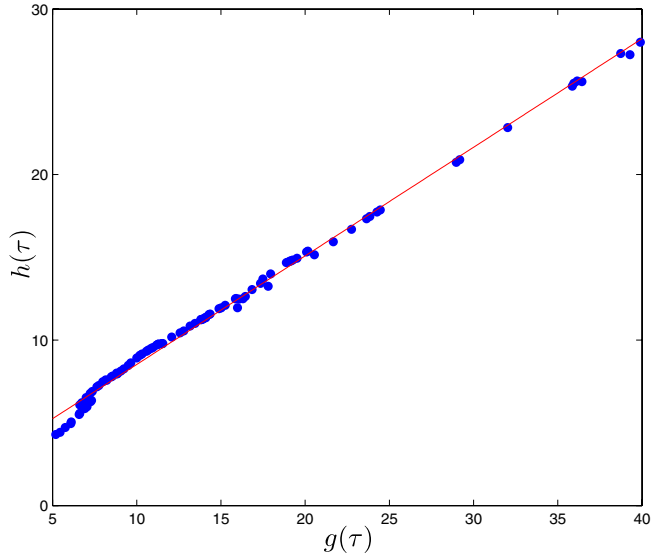

Fig. S4. Ensemble average of the SD of the time-shifted activity, $h(\tau)$, has a fixed dependence on the ensemble average of the time-shifted activity, $g(\tau)$, throughout the lifetime of applications. The different points correspond to different values of $\tau \in 50,60,70,\ldots,1{,}000$ such that increasing the value of τ leads to increasing values of $g(\tau)$. The values of τ start at 50 because we required that for each application there should be at least 50 points in the time series to estimate its first and second moments sufficiently accurately.

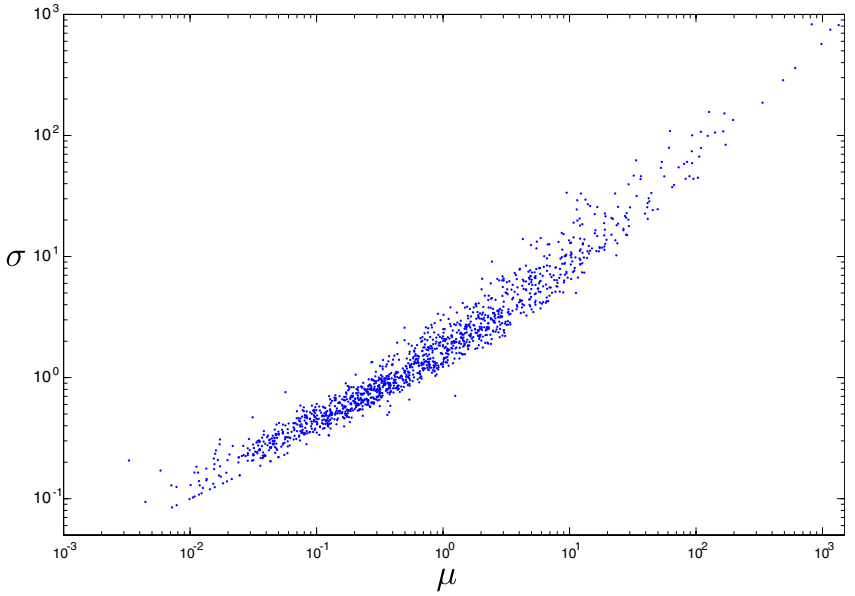

Fig. S5. Fluctuation scaling plot for activity f_i using only those applications i that were introduced during the studied time period and for which we had at least 50 time steps worth of data.

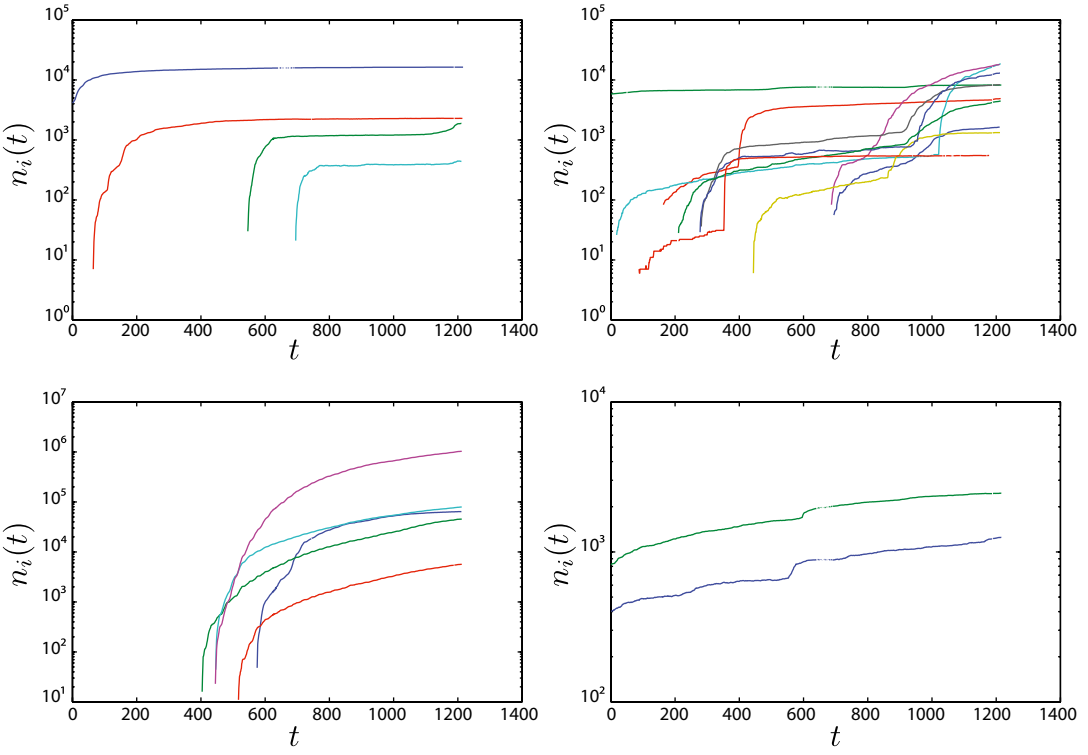

Fig. S6. Examples of adoption curves, that is, the cumulative number of application adoptions by a given time. The different panels group together applications exhibiting similar growth patterns.