Scaling and correlation of human movements in cyberspace and physical space

Zhi-Dan Zhao,[1,2] Zi-Gang Huang,[2,3] Liang Huang,[2,3] Huan Liu,[4] and Ying-Cheng Lai[2,5,*]

[1]*Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2]*School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA*
[3]*Institute of Computational Physics and Complex Systems and Key Laboratory for Magnetism and Magnetic Materials of MOE, Lanzhou University, Lanzhou, Gansu 730000, China*
[4]*School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, Arizona 85287, USA*
[5]*Department of Physics, Arizona State University, Tempe, Arizona 85287, USA*

(Received 1 August 2014; published 12 November 2014)

Understanding the dynamics of human movements is key to issues of significant current interest such as behavioral prediction, recommendation, and control of epidemic spreading. We collect and analyze big data sets of human movements in both cyberspace (through browsing of websites) and physical space (through mobile towers) and find a superlinear scaling relation between the mean frequency of visit $\langle f \rangle$ and its fluctuation σ: $\sigma \sim \langle f \rangle^{\beta}$ with $\beta \approx 1.2$. The probability distribution of the visiting frequency is found to be a stretched exponential function. We develop a model incorporating two essential ingredients, preferential return and exploration, and show that these are necessary for generating the scaling relation extracted from real data. A striking finding is that human movements in cyberspace and physical space are strongly correlated, indicating a distinctive behavioral identifying characteristic and implying that the behaviors in one space can be used to predict those in the other.

DOI: 10.1103/PhysRevE.90.050802 PACS number(s): 89.75.Da, 05.40.Fb, 89.75.Hc

Traditionally, human movements are restricted to the real physical space (or geospace). Pioneering works demonstrated that there are intrinsic patterns underlying human mobility in physical space [1–3], which are key to deciphering the dynamics of human behaviors with wide applications ranging from traffic forecasting [4] to epidemic prevention [5]. Triggered by the tremendous advances in modern information and communication technologies, at present as well as in the future, human movements occur not only in physical space but also in virtual or cyberspace. Here movements in cyberspace are defined broadly as changes in online activities, typically corresponding to switchings in the websites of exploration. Examples of cyberspace movements include World Wide Web surfing along hyperlinks and continuous shopping from commercial websites in a single online session. Do human movements in cyberspace and physical space share common features? Are there general scaling relations underlying human movements in both spaces?

Studies of human behaviors have been greatly facilitated by the ubiquity of massive empirical data sets (big data sets) that typically record individuals' movements on various temporal and spatial scales [6,7]. For example, great insights into the dynamics of human movements in physical space were gained by tracking and analyzing the dispersal of dollar bills [1] and through mobile phone [2] and GPS [8] data. There were also efforts to uncover human movements in cyberspace during web surfing [9–11] and to probe into human interests dynamics unfolded during cyberspace shopping and browsing [12].

In this paper we analyze data sets that record mobile phone users' visits to websites in cyberspace and to mobile towers in the physical space *simultaneously* and search for a correlation between the movements and general scaling relations. Distinguished from existing approaches to human-mobility analysis [1–3], we focus on the relationship between flux and fluctuations [13–19]. In particular, from time to time an individual would visit various sites in both spaces. For any given site, the number of visits to it, or the frequency of visits, denoted by f, can be counted. Suppose the individual visits a large number of distinct sites. The frequency f can then be regarded as a random variable with a certain probability distribution, from which the mean frequency of visits and its variance, denoted by $\langle f \rangle$ and σ^2, respectively, can be defined. For a large number of individuals, each with a definite pair of coordinates in the $\langle f \rangle$-σ plane, we examine the dependence of σ on $\langle f \rangle$. In biological physics, the scaling relationship between the two quantities, also called Taylor's law, was originally discovered in the densities of different species of organisms [20], where the scaling exponent is a crucial quantity in characterizing or classifying the underlying dynamics of the system. In complex transportation systems such as rivers, highways, the Internet, and microchips, the scaling relationship between the nodal flux fluctuation σ and the mean flux $\langle f \rangle$ has also attracted much attention [13–19], with values of the scaling exponent typically in the range [0.5,1.0]. For human movements in cyberspace and physical space, we find the scaling relation between the two quantities as $\sigma \sim \langle f \rangle^{\beta}$ and that the scaling exponent β assumes a value greater than unity, hence the term superlinear scaling. Surprisingly, we find that the scaling exponents in cyberspace and physical space cannot be distinguished, indicating a remarkable similarity between the dynamics of human movements in virtual and real spaces. Indeed, we find a strong correlation between the movements in the two spaces, indicating the existence of a distinctive behavioral identifying characteristic associated with each user. The intriguing implication is that human behaviors in physical space may be predicted in terms of those in cyberspace and vice versa. To place these findings on a firm ground, we develop an analyzable model with two essential dynamical ingredients derived from real data analysis to understand the flux-fluctuation relations governing human movements in both cyberspace and physical space.

[*] ying-cheng.lai@asu.edu

Typical data sets used in previous research of human mobility are real-time tracking of mobile phone users solely in physical space through various towers [2,3]. The big data sets that we analyze, however, record simultaneous movements of a large number of individuals in both cyberspace and physical space. In particular, our data sets are randomly sampled from millions of mobile phone users for approximately one month in a major city in China. For each individual, the data set recorded the mobile tower location each time some websites had been visited by mobile phone. The data thus recorded the individual's activities with respect to two types of sites—websites in cyberspace and mobile towers in physical space—from which the individual's trajectories in both spaces can be constructed. More specifically, the cyberspace activities are characterized through web surfing. The relevant events are thus those that involve simultaneous web surfing (in cyberspace) and phone or SMS communications (in physical space). Web surfing through the mobile devices provides a convenient platform to collect the required data, with the advantage that the physical locations of the users (corresponding to mobile towers) can be recorded at the same time. The raw data set used in our work recorded information of 20 000 users, who were randomly sampled from millions of mobile phone users with at least 100 actions. To ensure statistical significance, we impose the additional criterion that the numbers of both mobile towers and distinct websites visited during the one month observational period exceed 50. This results in a database of 3174 users with pronounced activities in both cyberspace and physical space.

We measure the mean visiting frequency $\langle f \rangle$ per site for a given individual and the corresponding standard deviation σ of the frequency distribution for the visited sites. Specifically, for each individual, the mean frequency of visits is defined as $\langle f \rangle = n/S$, where n is the user's total number of visits (actions) and S is the number of distinct sites visited by the user. The standard deviation σ characterizes the degree of heterogeneity for the user to distribute the n actions among the S sites. Figure 1(a) shows, in both cyberspace and physical space for a large number of users, σ versus $\langle f \rangle$, with σ averaged over the users with approximately equal values of $\langle f \rangle$. We observe a power-law scaling relation between the two quantities with the exponent $\beta \approx 1.2 > 1$. This superlinear behavior implies that, on average, the users with higher mean visiting frequency $\langle f \rangle$ per site are likely to distribute their visits more heterogeneously among the sites. The remarkable phenomenon is that the scaling relations in the two spaces are essentially indistinguishable.

An issue of interest is whether there is correlation between the user activities in cyberspace and physical space, i.e., do people possess distinct behavioral characters in different spaces or are there common features? As shown in Fig. 1(b), where the x and y coordinates of each circle represent the average visiting frequencies of an individual user in the respective spaces, we find a high concentration of circles near the diagonal line, indicating a strong correlation between the activities in cyberspace and physical space. This implies the existence of a distinctive behavioral identifying characteristic of a user, which is shared in both spaces. The phenomenon may have potential applications. For example, it may be possible to predict the behaviors of individuals in physical space based on their activities in cyberspace and vice versa.

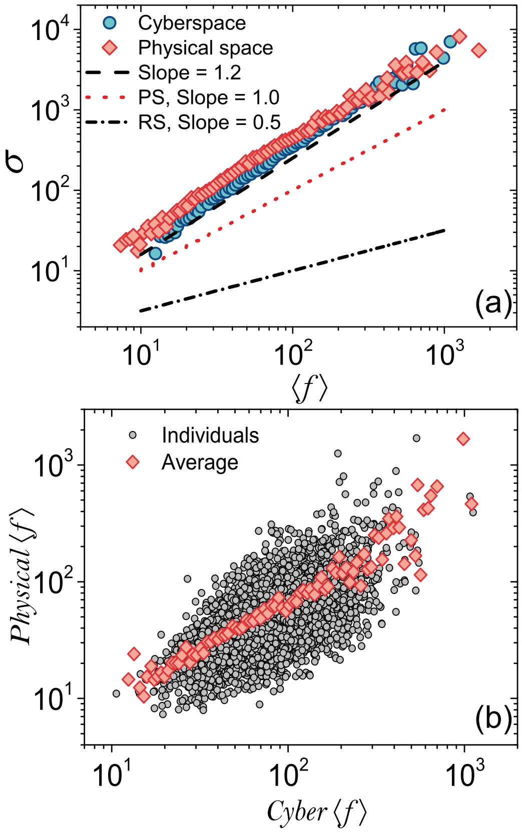

FIG. 1. (Color online) (a) Flux-fluctuation scaling of human movements, where the blue circles and red diamonds correspond to cyberspace (websites) and physical space (towers), respectively. The black dashed line has the slope 1.2. The red dotted and black dash-dotted lines correspond to random selection (RS) and preferential selection (PS) of sites with slope 0.5 and 1.0, respectively (see model construction). The scaling relations in the two spaces cannot be distinguished. (b) Average flux of each individual in the cyberspace space versus that in the physical space and the means over individuals with common values of the average cyberspace visiting frequency $\langle f \rangle$. There is a strong correlation between the user activities in both spaces.

To gain insights into the mechanisms that lead to the superlinear scaling relationship as exemplified in Fig. 1, we consider two scenarios: (i) homogeneous random selection (RS) and (ii) heterogeneous preferential selection (PS) of sites. For the first scenario, a given user randomly visits S distinct sites with identical probability $1/S$. The process continues until all S sites have been visited at least once. In this stochastic scenario, the frequency of visits to the S sites obeys a binomial distribution, so we have $\sigma \sim \langle f \rangle^{1/2}$ (see note 1 in [21]). Either an increase in the total number n of visits or a decrease in the number of sites S can cause $\langle f \rangle$ to increase, while the plots of all these cases collapse into a single curve in the σ-$\langle f \rangle$ plane with the scaling exponent $\beta = 1/2$. For the second scenario, the user visits sites with heterogeneous probabilities $\{p_i\}$, which are time independent over the S sites [14]. Under this mechanism, we obtain $\sigma \sim \langle f \rangle$ in note 2 in [21]. Simulation

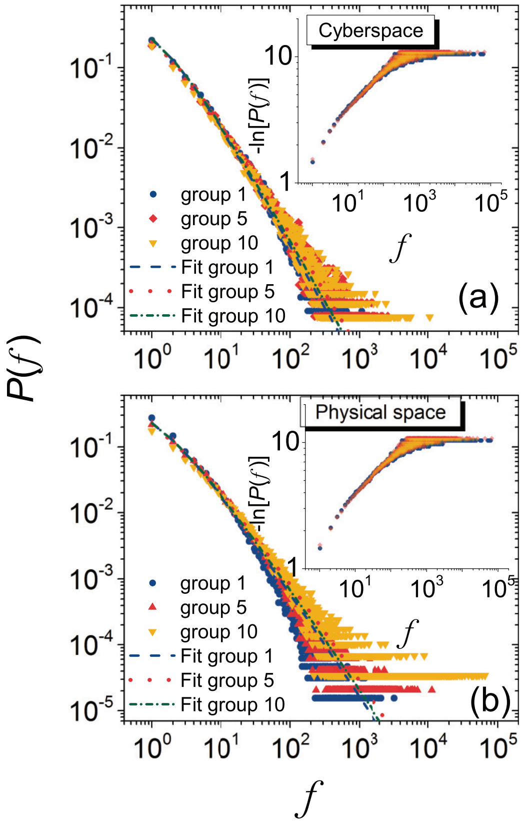

FIG. 2. (Color online) Distributions of the frequency of visit to (a) websites in the cyberspace and (b) mobile towers in physical space. The distributions shown are from three groups of users (out of ten groups). The insets show the plots on a logarithmic scale, in which the linear behaviors imply a stretched exponential distribution.

results from the RS and PS scenarios are shown in Fig. 1, where the scaling exponents are 0.5 and 1, respectively. Neither of the two mechanisms, however, is capable of explaining the superlinear behavior with $\beta \approx 1.2$ extracted from real data sets.

To develop a model that gives rise to the superlinear flux-fluctuation scaling behavior, we calculate the probability distribution $P(f)$ of the visiting frequency from real data sets, as shown in Fig. 2. In particular, for both cyberspace and physical space, we divide the users into ten groups in terms of their values of $\langle f \rangle$. The distributions associated with all the groups exhibit a stretched exponential form [22] $P(f) \sim f^{\alpha-1} e^{\kappa f^{\alpha}}$, where the exponents $\alpha < 0$ and $\kappa < 0$ can be calculated using the standard maximum-likelihood estimation method [23]. We observe that the distributions are nearly identical for both cyberspace and physical space, in accordance with Fig. 1(b). The stretched exponential distribution is consistent across different user groups (see note 3 in [21]). The long-tail and exponential cutoff features in $P(f)$, common to both cyberspace and physical space, indicate the existence of some particular sites that receive more frequent visits. The strong correlation between the activities in the two spaces [Fig. 1(b)] and the common frequency distributions suggest that a single mechanism is responsible for the dynamical behaviors in both spaces.

Based on results from real data (Figs. 1 and 2) and previous works on human mobility [3] and human interest [12] dynamics, we hypothesize two fundamental ingredients underlying human activities in cyberspace and physical space: *exploration* and *preferential return*, as shown schematically in Fig. 3. To initiate a trajectory either in cyberspace or in physical space, an individual has two options [3,24]: to explore a new site with probability p_{new} or to return to a previously visited site with probability $1 - p_{\text{new}}$. Based on a combination of numerical calculation and physical reasoning, we show below that the model successfully predicts the superlinear scaling relation between σ and $\langle f \rangle$.

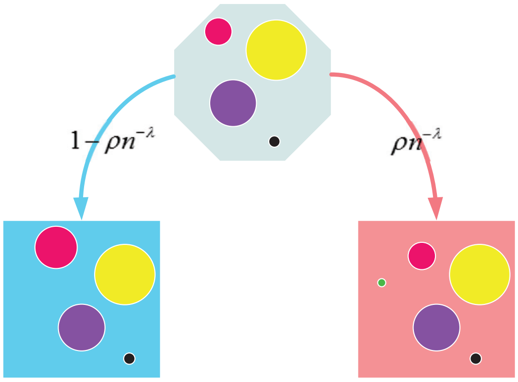

FIG. 3. (Color online) Schematic illustration of our model, where an individual can perform one of the two complementary processes at each step: exploring a new site with the probability $\rho n^{-\lambda}$ (exploration) or returning preferentially to one of the previously explored sites with the probability $1 - \rho n^{-\lambda}$ (preferential return).

Extensive analysis of real data sets revealed that the probability for a user to explore a new site is algebraically related to the total number of visits n (see Fig. S1 in note 4 in [21]). In particular, the number of sites already visited, denoted by S, increases by one at the nth visit of the user with probability $p_{\text{new}} = \rho n^{-\lambda}$. The form of p_{new} implies that the *growth rate* of S decreases as the number of actions n is increased. Approximating the dynamical process as continuous in n, we have

$$\frac{dS}{dn} = p_{\text{new}} = \rho n^{-\lambda}, \tag{1}$$

which gives the dependence of S on the number of visits n as

$$S \sim n^{-\lambda+1}. \tag{2}$$

The preferential return process occurring with probability $1 - p_{\text{new}}$ is the complementary event to exploration. Data analysis indicates that users revisit sites preferentially based on the corresponding frequencies of previous visits. A further indication of preferential return is the long tail in the distribution of the visiting frequency [2,3,12], as exemplified in Fig. 2. The probability for a user to select a particular already visited site i is $p_i = f_i/n$, where f_i is the accumulated times visiting site i and $n = \sum_{j=1}^{S} f_j$ is the total times visiting all sites.

Let n_i be the time of visit in the trajectory when site i is visited for the first time. The initial frequency of visiting site i is thus $f_i = 1$. When site i is selected again in the preferential return process, the frequency f_i will increase by one. The time evolution of f_i is then governed by $df_i/dn = (1 - p_{\text{new}})p_i$, where p_i is the probability for site i to be selected from all the

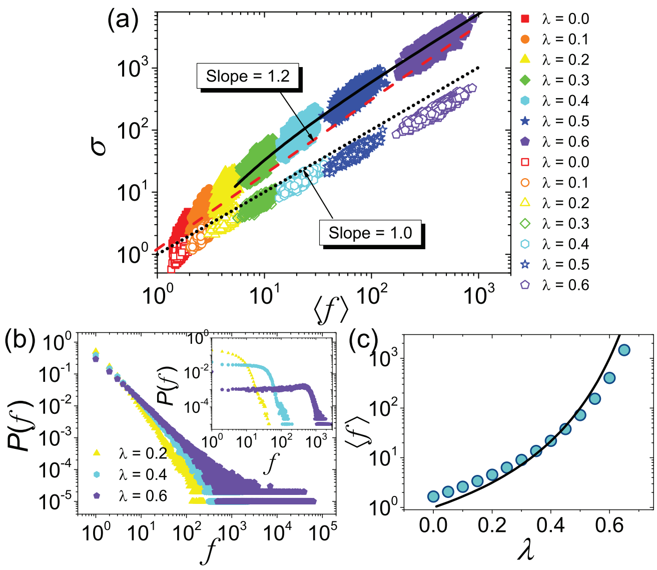

FIG. 4. (Color online) Simulation results from our model. (a) Model predicted relation between σ and $\langle f \rangle$ (closed symbols). The model can generate the value of the scaling exponent β (about 1.2) from data for a wide range of the parameter λ, e.g., $\lambda \in [0,0.6]$. For comparison, results from a variant of the model with exploration and random return are included (open symbols), which predicts the exponent $\beta = 1$ (linear relation between σ and $\langle f \rangle$). The solid black curve is the result from Eq. (3) and the red dashed and black dotted lines are to guide the eye. (b) Model predicted frequency distribution $P(f)$ for $\lambda = 0.2$ (yellow triangles), 0.4 (cyan hexagons), and 0.6 (purple pentagons). The inset shows the corresponding $P(f)$ from the model variant with random return. (c) Mean visiting frequency $\langle f \rangle$ versus λ obtained from simulation (closed circles) and from Eq. (4) (solid curve).

visited sites. We see from Eq. (1) that, for $\lambda > 0$, as n increases, we have $p_{\text{new}} \to 0$. Asymptotically, we have $df_i/dn = p_i$, with the initial condition $f_i(n_i) = 1$. We obtain $f_i = n/n_i$, which indicates that, the earlier one site was explored (smaller n_i), the higher the frequency f_i of visits to it. From Eq. (2) we have

$$f_i = \frac{n}{n_i} \sim \frac{S^{1/(1-\lambda)}}{n_i} \sim \frac{S^{1/(1-\lambda)}}{i^{1/(1-\lambda)}}, \tag{3}$$

which can be solved numerically to yield the frequency of visit to the ith site. For any given value of λ, the variance in the frequency of visit to all sites can also be calculated. The relation between σ and $\langle f \rangle$ can then be obtained.

Figure 4 shows the simulation results from our model. The visiting behaviors for different values of λ (varied from 0 to 0.6) are realized 10^3 times and the values of σ and $\langle f \rangle$ are plotted (closed symbols). We (somewhat arbitrarily) set the model parameters to be $S = 100$ and $\rho = 0.6$, as in previous works [3,12,24]. The maximum number of visit is $n_{\max} = 1 \times 10^5$, at which time the exploration probability p_{new} approaches zero. The simulation results shown in Figs. 4(a) and 4(b), respectively, are the superlinear scaling relation between $\langle f \rangle$ and σ and the stretched exponential distribution $P(f)$. There is good agreement with results from real data sets (Figs. 1 and 2). In addition, the mean frequency of visits is $\langle f_i \rangle = n/S$. From Eq. (2) we get the approximate relation between the mean frequency and the parameter λ as

$$\langle f \rangle \sim \frac{S^{1/(1-\lambda)}}{S} = S^{\lambda/(1-\lambda)}, \tag{4}$$

as shown in Fig. 4(c) (solid curve).

To further validate our model, we consider two model variants. First, we consider the situation where the return process is random. This model generates $\beta = 1$, thereby predicting a linear relation between $\langle f \rangle$ and σ [open symbols in Fig. 4(a)], which deviates markedly from the value of 1.2 obtained from real data sets. In addition, the predicted distribution $P(f)$ can no longer be approximated by a stretched exponential function, as shown in the inset in Fig. 4(b). These indicate strongly the necessity of preferential return in the model. Second, we consider a model variant in which the exploration process has a constant growth rate [instead of the decay rate defined in Eq. (1)]. The predictions from this model deviate significantly from the results from real data as well (Fig. S3 in note 5 in [21]). We also verify that different values of the site number S have little effect on the scaling relation (note 6 in [21]).

To summarize, through analysis of big data sets of mobile phone users in both cyberspace and physical space, we uncovered a superlinear scaling relation between the average visiting frequency and its fluctuation (a kind of flux-fluctuation relation). The underling mechanisms are (i) exploration of new sites with a probability that decays with the user's total number of actions and (ii) preferential return to highly visited sites. These two factors reveal the essential features in human movements. We developed a model incorporating these two factors, which generates robust superlinear behavior with the observed scaling exponent. The necessity of the two mechanisms was established by considering model variants, which lead to results that do not agree with those from real data. A striking finding is that there is a strong correlation between human movements in cyberspace and physical space. Although there are individuals that can be far more active in cyberspace than in physical space and vice versa, their behaviors in both spaces follow certain patterns and possess unique identifying characteristics. This suggests the possibility of predicting behaviors in one space based on those in the other. Our work provides insights into questions of significant current interest ranging from human-behavior prediction and design of searching algorithms [3,25,26] to controlling epidemic spreading processes [5].

This work was partially supported by ARO under Grant No. W911NF-14-1-0504, by ONR under Grant No. N000141410095, and by NSF of China under Grant No. 11275003.

[1] D. Brockmann, L. Hufnagel, and T. Geisel, Nature (London) **439**, 462 (2006).

[2] M. C. González, C. A. Hidalgo, and A.-L. Barabási, Nature (London) **453**, 779 (2008).

[3] C. Song, T. Koren, P. Wang, and A.-L. Barabási, Nat. Phys. **6**, 818 (2010).

[4] D. Helbing, Rev. Mod. Phys. **73**, 1067 (2001).

[5] D. Balcan and A. Vespignani, Nat. Phys. **7**, 581 (2011).

[6] M. Starnini, A. Baronchelli, and R. Pastor-Satorras, Phys. Rev. Lett. **110**, 168701 (2013).

[7] H.-J. Gao, J.-L. Tang, X. Hu, and H. Liu, in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (ACM, New York, 2013), pp. 1673–1678.

[8] I. Rhee, M. Shin, S. Hong, K. Lee, and C. Song, in *Proceedings of the 27th IEEE International Conference on Computer Communications (INFOCOM 2008)* (IEEE, Piscataway, 2008), pp. 924–932.

[9] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose, Science **280**, 95 (1998).

[10] A. Chmiel, K. Kowalska, and J. A. Hołyst, Phys. Rev. E **80**, 066122 (2009).

[11] S. Kumar, R. Zafarani, and H. Liu, *Special Track on AI and the Web (AIW) at the 25th AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, 2011).

[12] Z.-D. Zhao, Z. Yang, Z. Zhang, T. Zhou, Z.-G. Huang, and Y.-C. Lai, Sci. Rep. **3**, 3472 (2013).

[13] M. A. de Menezes and A.-L. Barabási, Phys. Rev. Lett. **92**, 028701 (2004).

[14] M. A. de Menezes and A.-L. Barabási, Phys. Rev. Lett. **93**, 068701 (2004).

[15] Z. Eisler and J. Kertész, Phys. Rev. E **71**, 057104 (2005).

[16] Z. Eisler, I. Bartos, and J. Kertész, Adv. Phys. **57**, 89 (2008).

[17] J. Duch and A. Arenas, Phys. Rev. Lett. **96**, 218702 (2006).

[18] B. Kujawski, B. Tadić, and G. J. Rodgers, New J. Phys. **9**, 154 (2007).

[19] Z. Zhou, Z.-G. Huang, L. Huang, Y.-C. Lai, L. Yang, and D.-S. Xue, Phys. Rev. E **87**, 012808 (2013); Z.-G. Huang, J.-Q. Dong, L. Huang, and Y.-C. Lai, Sci. Rep. **4**, 6787 (2014).

[20] L. R. Taylor, Nature (London) **189**, 732 (1961).

[21] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.90.050802 for additional empirical details and simulation results.

[22] J. Laherrère and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).

[23] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).

[24] M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora, Sci. Rep. **2**, 457 (2012).

[25] A. Vespignani, Science **325**, 425 (2009).

[26] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction* (Cambridge University Press, Cambridge, 2014).