HAL open science

Two types of densification scaling in the evolution of temporal networks

Teruyoshi Kobayashi, Mathieu Génois

To cite this version:

Teruyoshi Kobayashi, Mathieu Génois. Two types of densification scaling in the evolution of temporal networks. Physical Review E , 2020, 102 (5), pp.052302. 10.1103/PhysRevE.102.052302 . hal-03094946

HAL Id: hal-03094946

https://hal.science/hal-03094946

Submitted on 4 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Two types of densification scaling in the evolution of temporal networks

Teruyoshi Kobayashi*
Department of Economics, Kobe University, Kobe, Japan
and Center for Computational Social Science, Kobe University, Kobe, Japan

Mathieu Génois
CNRS, CPT, Aix-Marseille Université, Université de Toulon, Marseille, France
and GESIS, Leibniz Institute for the Social Sciences, Köln/Mannheim, Germany

(Received 11 June 2020; accepted 18 October 2020; published 9 November 2020)

Many real-world social networks constantly change their global properties over time, such as the number of edges, size, and density. While temporal and local properties of social networks have been extensively studied, the origin of their dynamical nature is not yet well understood. Networks may grow or shrink if (a) the total population of nodes changes and/or (b) the chance of two nodes being connected varies over time. Here, we develop a method that allows us to classify the source of time-varying nature of temporal networks. In doing so, we first show empirical evidence that real-world dynamical systems could be categorized into two classes, the difference of which is characterized by the way the number of edges grows with the number of active nodes, i.e., densification scaling. We develop a dynamic hidden-variable model to formally characterize the two dynamical classes. The model is fitted to the empirical data to identify whether the origin of scaling comes from a changing population in the system or shifts in the connecting probabilities.

DOI: 10.1103/PhysRevE.102.052302

I. INTRODUCTION

Along with the increasing availability of high-resolution data sets, dynamics of human social communication have been extensively studied over the past decades [1–6]. Many of these studies are based on data sets of online interactions, such as emails [7], text messages [8,9], and mobile phones [2,3,10,11], but the recent development of sensor devices has also enabled us to collect time-stamped data from face-to-face interactions in physical space [1,12–15]. Those data therefore cover a wide range of social contexts in which dynamic interactions among individuals form temporal social networks [6,16].

These real-world social networks exhibit very often nonstationarity: their structure constantly changes over time not only in shape but also in size. Generally, these dynamics are present because the studied system is not closed: it is in fact a common property of real-world social and economic networks that agents are free to enter and exit. In online social networks like Twitter and Facebook, anyone can basically join or quit the existing communication at any time. In financial markets, a bank becomes a part of an interbank network if it borrows from or lends to other banks and exits the network when the loan is repaid [17,18]. Another nonconservative aspect of real-world networks is the fact that, even if the population is constant, the networking activity might vary due to external factors such as schedule and diurnal rhythm; coffee breaks in a conference [13], pauses between classes in a school [5,19], lunch breaks in a company [15,20], etc.

In the present work, we focus on the evolution of two fundamental quantities that condition the global property of networks: the number of active nodes N and the number of edges M. The scaling relationship $M \propto N^{\gamma}$, known as the "densification power law," has been found in many real-world systems [21,22], where the scaling exponent γ is constant and $1 < \gamma < 2$. In this work, we present further empirical evidence that there exist two types of scaling in the evolution of networks. In addition to the well-known densification power law, we also show that some systems exhibit an accelerating growth of M in which the scaling exponent itself is increasing.

We consider two key factors that would lead networks to be time varying: N and M will vary over time if (a) the size of population (i.e., potential number of active nodes) changes and/or (b) the chance of two nodes being connected changes. Clearly, the size of the population in a system constrains the number of active nodes that form a network: at constant probability for links to appear, more nodes implies more links. Similarly, bilateral matching probability determines the number of edges in the network and thereby its density: at constant population size, the higher the probability for a link to exist, the higher the number of links. The question is then: given an empirical temporal network exhibiting time-varying global quantities, is it possible to identify the source of the dynamics?

Here, we develop a method to perform such a task by exploiting a scaling relationship between the numbers of active nodes and edges. To model the behavior of nodes, we use a dynamical version of a hidden-variable model in which the

*Corresponding author: kobayashi@econ.kobe-u.ac.jp

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

temporal probability of two nodes being connected is given by a product of "fitness" parameters [23,24]. The fitness parameters are considered to be intrinsic and constant features of the nodes. In the present model, the time-evolving aspect arises from two distinct channels. First, we introduce a parameter that modulates the average activity level of nodes. This modulation parameter allows the size of generated networks to vary through a change in the connecting probabilities while keeping the population size, including resting nodes, constant. Second, we allow the population size to vary with time. In the original fitness model [23,24], there is no distinction between population and the number of active nodes, because the population size is assumed to be large enough so that virtually all nodes in the system are active [18]. However, if the population in a system is not sufficiently large, a certain fraction of existing nodes may not be active [18], and thereby a change in the population size affects the rate at which the number of edges grows with the number of active nodes.

In the following, we first expose empirical evidence for the existence of two different types of scaling relationships between N and M. We then present a dynamical hidden-variable model with which we investigate the emergence of the two types of scaling patterns. Specifically, we define two classes of theoretical equations that connect N and M under different specifications on the average activity of nodes and the population size. By identifying a class of equations that better fits the observed data, the proposed method allows us to estimate the actual average activity and the (unobservable) population size. From this we are then able to identify for each empirical data which key factor drives the dynamics of the temporal network. We also briefly mention a variation of the model for cases where the population is fixed and known, allowing us to fit the empirical distribution of node fitnesses to a beta distribution. We conclude by a discussion of our results and the limitations of the model.

II. SCALING RELATIONSHIP BETWEEN N AND M

A. Data

We consider six data sets of social and economic interest, taken from contexts of very different nature (see Appendix A for a full description of the data sets):

(1) Interbank (bilateral transactions in the online interbank market in Italy);

(2) Enron (email communication network from the Enron Corporation [7,25]);

(3) CollegeMsg (online social network at the University of California, Irvine [8,9]);

(4) RealityMining (phone call data from the Reality Commons project [2]);

(5) LondonBike (bike trips from the London Bicycle Sharing Scheme [26]);

(6) Highschool (face-to-face contacts network in a French high school [19]).

All data sets are converted to temporal networks with undirected and unweighted edges. Bidirectional edges (i.e., edges in both directions) are regarded as undirected edges with weight 1. From each dataset, we construct a sequence of network snapshots by defining particular time intervals in each of which all the interactions between nodes are regarded as the edges of the corresponding network. We define N to be the number of nodes that have at least one edge in a given snapshot, and M denotes the corresponding number of edges. This suggests that $N/2 \leqslant M \leqslant N(N-1)/2$, where $N/2$ corresponds to the minimum number of edges that can exist between N active nodes (when all the nodes are connected to exactly one edge), while $N(N-1)/2$ corresponds to the maximum number of edges that can exist between N active nodes (i.e., a complete graph).

B. Evidence from empirical data

We investigate the dynamical relationship between N and M. Figure 1 shows scatter plots of M against N for each social context. Two important features appear. First and foremost, there is a strong positive correlation between N and M in all the data sets we examine. In particular, we observe superlinear scaling, i.e., the rate at which M rises with N is larger than that expected by a linear growth, as is occasionally reported for many real-world systems [21,22]. This phenomenon is also known as the "densification power law" or "densification scaling" [21,22,27].

Second, there are two different patterns as to how M grows with N. One is the densification scaling we mentioned, in which the scaling exponent is constant (>1), showing as a straight line on a log-log scale plot. In Fig. 1, Interbank, Enron, CollegeMsg, and RealityMining appear to belong to this category. Contrarily, for LondonBike and Highschool the growth of M for large values of N is accelerating: the slope itself increases in log-log space as N grows [28].

Both behaviors are striking, as they suggest the existence of simple mechanisms for the dynamics of global activity in temporal networks. However, the empirical dynamical relationship we observe in Fig. 1 cannot be reproduced by a class of common growing network models in which a new node joins the network with a given number of edges [29–31]. While these models are intended to explain the emergence of scaling in empirical degree distributions [29], the number of edges has a linear correlation with the number of active nodes asymptotically, i.e., $M \propto N$, which is not consistent with our finding. In the following, we present a model which explains how these different types of global behaviors can emerge from temporal social interactions.

III. A DYNAMIC HIDDEN-VARIABLE MODEL

A. Model

To explain the two types of scaling in a unified model, we consider a dynamical version of the hidden variable model in which the probability of two nodes being connected at time interval t is given by

$$p_{ij,t} = \kappa_t a_i a_j, \quad i, j = 1, \ldots, N_{\mathrm{p},t}, \quad t = 1, \ldots, T. \tag{1}$$

where a_i is the "fitness" of node i that represents the activity level of the node [23,24,32,33]. In the baseline model we assume that a_i is uniformly distributed on [0, 1] because in general we do not have any prior information about the distribution of activity levels. We will also consider a beta distribution as an alternative case in Sec. IV D.

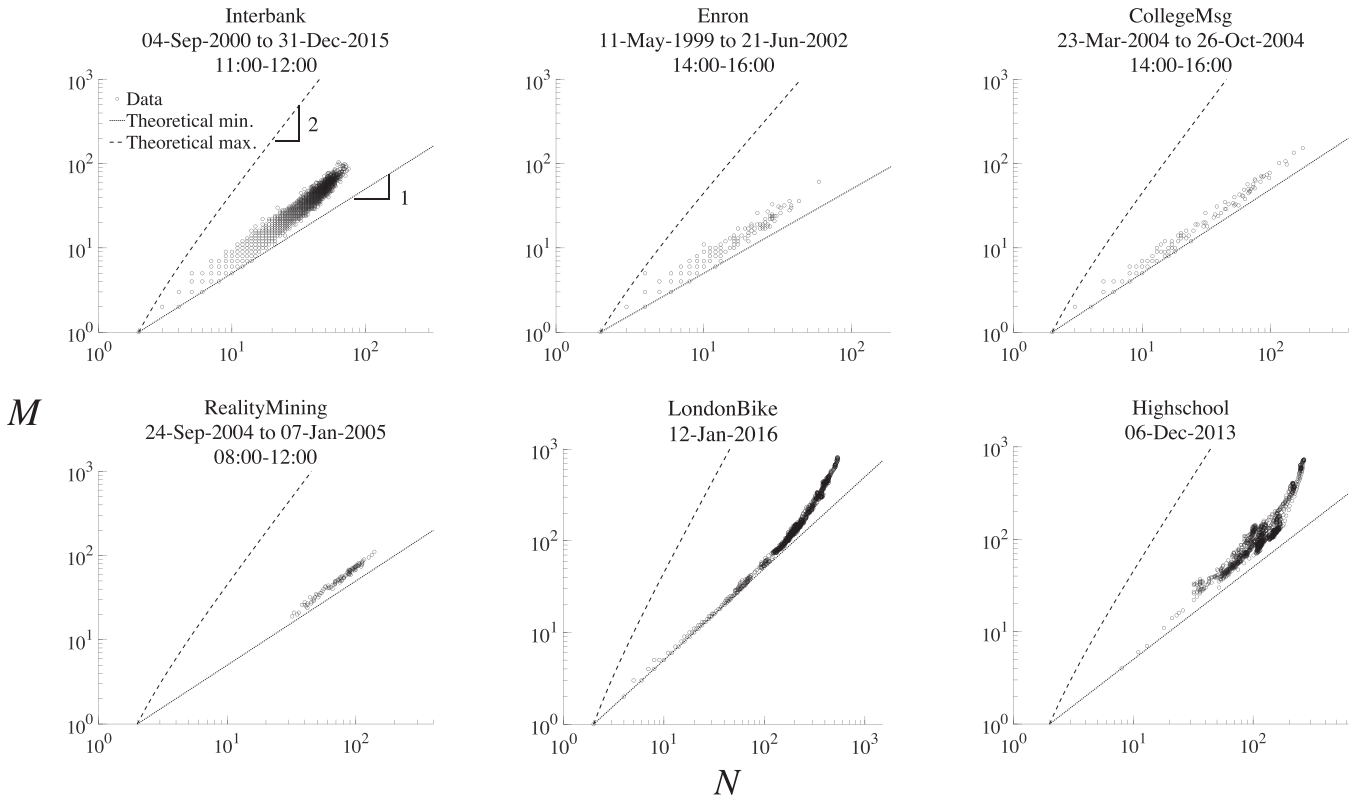

FIG. 1. Relationship between the number of active nodes N and the number of edges M in each snapshot for different social contexts. For Interbank, Enron, CollegeMsg, and RealityMining data, each dot represents the realization of (N, M) in a particular time window (annotated at the top) of a day. For LondonBike and Highschool data, each dot represents the realization of (N, M) in a 10-minute time window of a day (0:00–24:00). Black dotted and dashed lines denote the theoretical upper [$M = N(N-1)/2$] and lower ($M = N/2$) bounds, respectively.

There are two time-varying parameters in the model. One is $N_{\mathrm{p},t}$, which represents the potential number of active nodes in the system at time t, i.e., the total of active and inactive nodes that are in the system at time t. The number of active nodes having at least one edge at time t is denoted by N_t. We note that the number of active nodes N_t is always observable, but the potential number of nodes $N_{\mathrm{p},t}$ is not. In social networks, for instance, we do not usually know how many people are ready to interact with other people and what fraction of them actually created at least one edge. In many cases, what we can observe from data is the number of active nodes that appear in the record of interaction history, while there is no record of nodes without interactions. Since the observed active nodes may account for only a fraction of the potential nodes, it is generally written as $N_t = (1 - q_{0,t})N_{\mathrm{p},t}$, where $q_{0,t}$ denotes the fraction of inactive nodes that have no edge at time t, or equivalently, the probability of a randomly chosen node being isolated. To take an example of social networks, changes in N_{p} may represent a situation in which the number of students in the classroom changes over time according to the class schedule, leading to a variation in the maximum possible size of face-to-face contact networks. The potential number of nodes that are ready to interact with others is the first key parameter of the model, as it physically constrains the size of networks to be observed.

The second time-varying parameter of the model is $\kappa_t > 0$, which modulates the global activity level of nodes. In the financial system, for instance, the chance that two banks trade during the lunch time would be intrinsically lower than that in the morning [18], in which case the banks' activity levels may have a certain diurnal pattern. In social networks where individuals communicate with each other, κ would vary according to the time schedule of the school, workplace, academic conferences, or the circadian rhythm of humans [4,34–36].

With this specification, the observed network size N and the number of edges M coevolve as either N_{p} or κ or both change over time. One can see a change in N due to a shift in N_{p} represents an *extensive margin effect*, while a shift in κ leads to an *intensive margin effect*. Parameters κ and N_{p} can thus explain two different origins of the time-varying nature of networks.

B. Analytical expression for N and M

In this section we show an analytical solution for the dynamic hidden-variable model when the network size is not necessarily large enough [18]. Suppose that node $i\,(1 \leqslant i \leqslant N_{\mathrm{p}})$ is assigned activity $a_i \in [0, 1]$ which is drawn from density $\rho(a)$ [23]. The numbers of active nodes N and edges M can be expressed as functions of parameters κ and N_{p} (we drop time subscript t for notational convenience):

$$N = [1 - q_0(\kappa, N_{\mathrm{p}})]N_{\mathrm{p}},$$

$$M = \frac{\overline{k}(\kappa, N_{\mathrm{p}})N_{\mathrm{p}}}{2}, \tag{2}$$

where $\bar{k}(\kappa, N_p)$ denotes the average degree over all the existing nodes including isolated ones. To obtain the functional forms of N and M, we need to find the functional forms of $q_0(\kappa, N_p)$ and $\bar{k}(\kappa, N_p)$.

Let $u(a, a')$ be the probability that there is an edge between two nodes having activity levels a and a', respectively. As is shown in Appendix B, the average degree $\bar{k}(\kappa, N_p)$ is given by

$$\bar{k}(\kappa, N_p) = (N_p - 1) \iint da\, da' \rho(a)\rho(a')u(a, a'), \tag{3}$$

which simply states that the average degree is equal to the number of nodes (excluding the focal node itself) times the expected connecting probability. It should be noted that Eq. (3) is equivalent to Eq. (21) of Ref. [24] if $N_p - 1$ is replaced with N, which is asymptotically true as will be shown below. From (B11) in Appendix B, the probability of a randomly chosen node being isolated is given by

$$q_0(\kappa, N_p) = \int da' \rho(a') \left[1 - \int u(a', a)\rho(a) da\right]^{N_p - 1}. \tag{4}$$

Substituting $\rho(a) = 1$ (i.e., uniform distribution on [0, 1]) and $u(a, a') = \kappa a a'$ into Eq. (3) gives

$$\bar{k}(\kappa, N_p) = \frac{\kappa}{4}(N_p - 1). \tag{5}$$

Similarly, q_0 is given by

$$q_0(\kappa, N_p) = \int_0^1 \left(1 - \frac{\kappa a'}{2}\right)^{N_p - 1} da'. \tag{6}$$

By defining a variable $x \equiv 1 - \frac{\kappa a'}{2}$, we have

$$\begin{aligned} q_0(\kappa, N_p) &= \frac{2}{\kappa} \int_{1-\frac{\kappa}{2}}^{1} x^{N_p - 1} dx \\ &= \frac{2}{\kappa N_p}\left[1 - \left(1 - \frac{\kappa}{2}\right)^{N_p}\right]. \end{aligned} \tag{7}$$

Combining these results with Eq. (2), we have

$$N = N_p\left[1 - \frac{2}{\kappa N_p}\left(1 - \left(1 - \frac{\kappa}{2}\right)^{N_p}\right)\right], \tag{8}$$

$$M = \frac{1}{8}\kappa N_p(N_p - 1). \tag{9}$$

This leads to interesting limit behaviors: if $|1 - \kappa/2| < 1$ and N_p is sufficiently large, then $q_0(\kappa, N_p) \simeq 0$ and thereby $N \simeq N_p$ and $M \propto N^2$, as is shown in the study of the static fitness model [23,24,32]. In contrast, if N_p is not large enough, then $q_0(\kappa, N_p) > 0$ and $N < N_p$, in which case M is not of order N^2 and the scaling exponent will take a value between 1 and 2 as is observed in empirical data (Fig. 1) [18]. Note that κ is not *per se* a probability, and that its value does not have any *a priori* upper bound (as it depends on the activity distribution). Clearly, the larger the population N_p and the overall activity κ, the lower the share of resting nodes q_0 in the population [Fig. 2(a)].

C. Role of κ and N_p in the emergence of scaling

Using Eqs. (8) and (9), we are now able to analyze numerically how M scales with N. First, we observe that if the value of κ is kept constant while N_p varies, the dynamical relationship between N and M is close to a straight line in a log-log plot [gray solid in Fig. 2(b)], as seen in some empirical data. If κ is small enough, the scaling is close to linear, approaching the lower bound of M indicated by the dashed line in Fig. 2(b). However, as κ increases, the scaling becomes more and more superlinear, which can be seen in Fig. 2(b) by following the same symbols in different colors.

By contrast, if we vary κ for a given value of N_p, the slope will bend upward. This can be seen in Fig. 2(b) by following different symbols in the same color (colored line). This reproduces the accelerating growth behavior observed in the empirical data, namely LondonBike and Highschool. We note that although the scaling relationships appear to be

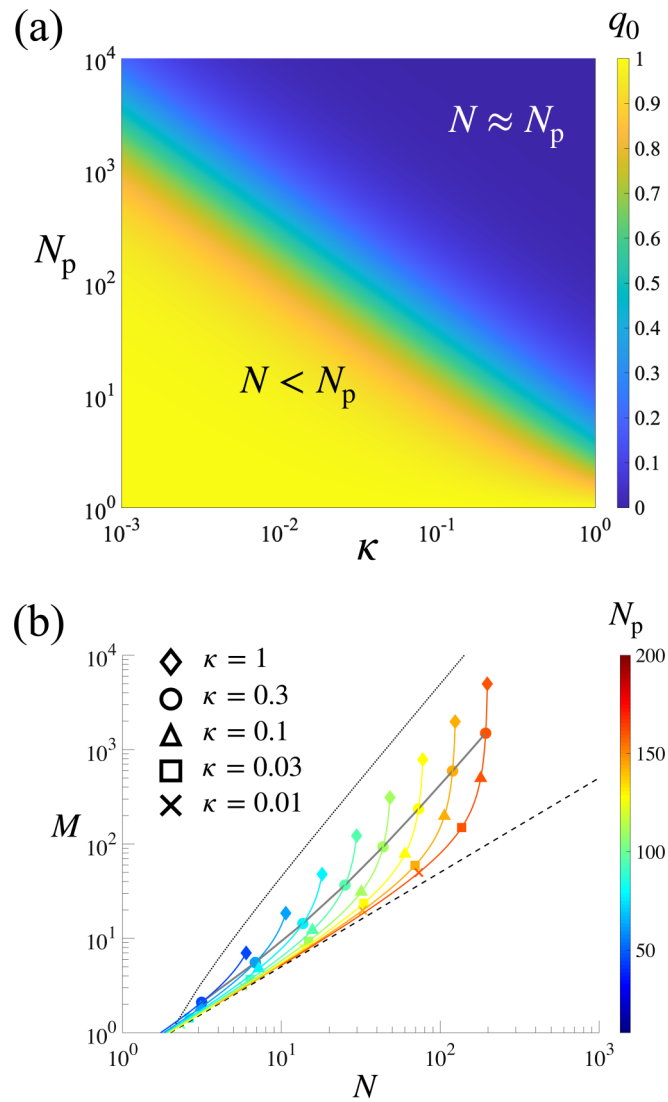

FIG. 2. Scaling in the dynamic hidden variable model. (a) Fraction of inactive nodes in the population, q_0. (b) Two types of scaling relationships between N and M. Each color represents a particular value of N_p, while different symbols denote different values of κ. Each colored line represents a scaling relation when N_p is fixed at a given value, i.e., scaling driven by time-varying κ. The gray solid line represents a scaling relation when κ is fixed (= 0.3), i.e., scaling driven by time-varying N_p.

quite regular, it proves to be very difficult (if not impossible) to extract from Eqs. (8) and (9) an analytical expression for them, because of the complicated dependencies of N on N_p.

IV. IDENTIFYING THE SOURCE OF NETWORK DYNAMICS

A. Estimation of model parameters from empirical data

We now propose a method to identify the dynamical class of a system and at the same time estimate κ and N_p from the empirical data. In fact, the two model parameters may be estimated in two ways. One is to directly solve the two nonlinear equations Eqs. (8) and (9) with respect to (κ, N_p) for a given observation of (N, M). This direct calculation gives us a one-to-one mapping of (N, M) to (κ^*, N_p^*), where an asterisk denotes the solution of the system of two equations. However, such a method proves to be unable to accurately estimate the parameters when the network is small, where there is a large degree of overlap of (N, M) generated under multiple combinations of (κ, N_p) [Figs. 2(b) and 4(a)].

B. Exploiting the dynamical relationship between N and M

1. Two classes of models for the two types of scaling

The other method is to use the dynamical relationship between N and M. This method is based on the idea that the estimation bias due to the overlap of (N, M) could be avoided if we exploit the dynamical relationship between N and M rather than a particular observation of (N, M) in a given snapshot. In this method, we fit the empirical N-M relationship to theoretical equations, which will give us nonlinear least squares estimators of κ and N_p.

Since the observable variables N and M appear separately in Eqs. (8) and (9), respectively, we formulate a regression equation by relating N with M through the substitution of κ or N_p. By doing so, we essentially categorize the empirical dynamic networks into two classes. In the regression equation for the first class, we express N as a function of M and parameter κ to endogenize the time variation of N_p. Hereafter we call this type of formulation "Model I." This corresponds to a situation in which the connecting probability $p_{ij}, \forall i \neq j$ is constant while the potential size of networks N_p is time varying.

In "Model II," on the other hand, we specify N as a function of M and parameter N_p to endogenize the time variation of κ. This type of model would be appropriate when the set of nodes is fixed while the connecting probabilities are affected by diurnal or circadian rhythms.

The regression equations in the two models are respectively given as follows:

Model I: *N_p is time varying and κ is constant.*

$$N = G(M;\kappa) \equiv N_p(M,\kappa)\left[1 - \frac{2}{\kappa N_p(M,\kappa)}\left(1 - \left(1 - \frac{\kappa}{2}\right)^{N_p(M,\kappa)}\right)\right], \tag{10}$$

where N_p is expressed as a function of M and κ: $N_p(M,\kappa) \equiv \frac{1+\sqrt{1+32M/\kappa}}{2}$ [Eq. (9)]. We obtain the estimator of κ, denoted by $\widehat{\kappa}$, by regressing N on M using a method of nonlinear least squares, where $\boldsymbol{N} = G(\boldsymbol{M};\widehat{\kappa}) + \boldsymbol{\varepsilon}_{\mathrm{I}}$, and $\boldsymbol{\varepsilon}_{\mathrm{I}}$ denotes the $T \times 1$ vector of residuals. Estimates of time-varying N_p are then given by

$$\widehat{N}_{p,t} = \frac{1 + \sqrt{1 + 32M_t/\widehat{\kappa}}}{2}. \tag{11}$$

Model II: *N_p is constant and κ is time-varying.*

$$N = F(M;N_p) \equiv N_p\left[1 - \frac{2}{\kappa(M,N_p)N_p}\left(1 - \left(1 - \frac{\kappa(M,N_p)}{2}\right)^{N_p}\right)\right], \tag{12}$$

where κ is expressed as a function of M and N_p: $\kappa(M,N_p) \equiv \frac{8M}{N_p(N_p-1)}$ [Eq. (9)]. We estimate $\widehat{N}_p$ based on a nonlinear regression equation $\boldsymbol{N} = F(\boldsymbol{M};\widehat{N}_p) + \boldsymbol{\varepsilon}_{\mathrm{II}}$. Estimates of time-varying κ are given by

$$\widehat{\kappa}_t = \frac{8M_t}{\widehat{N}_p(\widehat{N}_p - 1)}. \tag{13}$$

After estimating the parameters in both specifications, we select one that attains the lower sum of squared errors:

Model I is selected if $\boldsymbol{\varepsilon}_{\mathrm{I}}^{\top}\boldsymbol{\varepsilon}_{\mathrm{I}} < \boldsymbol{\varepsilon}_{\mathrm{II}}^{\top}\boldsymbol{\varepsilon}_{\mathrm{II}}$,
Model II is selected otherwise.

A schematic of the model selection is illustrated in Fig. 3. Note that the criterion of model selection is effectively the same as that of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) because we have only one parameter in both models.

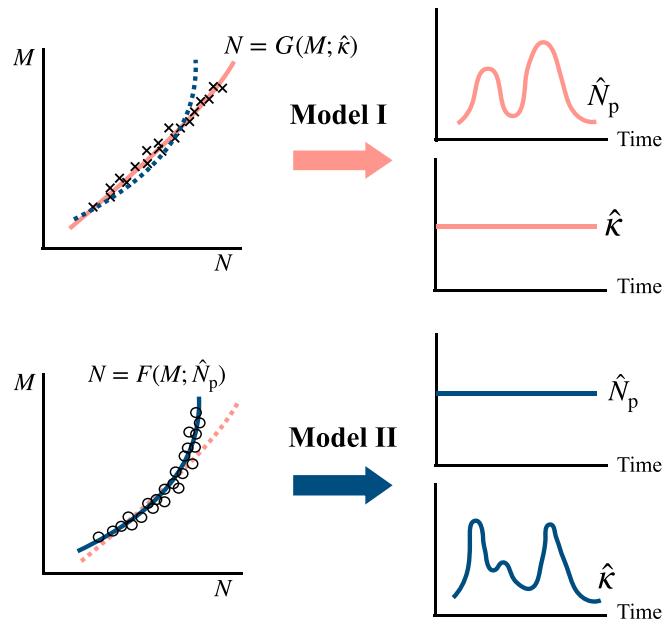

FIG. 3. Schematic of model selection. If Model I (resp. Model II) is selected, each bilateral connection probability $\hat{\kappa}$ (resp. population size $\widehat{N}_p$) is constant while $\widehat{N}_p$ (resp. $\hat{\kappa}$) is time varying.

2. Validation

We check the accuracy of the proposed estimation method by using synthetic networks. For the estimation of Model I

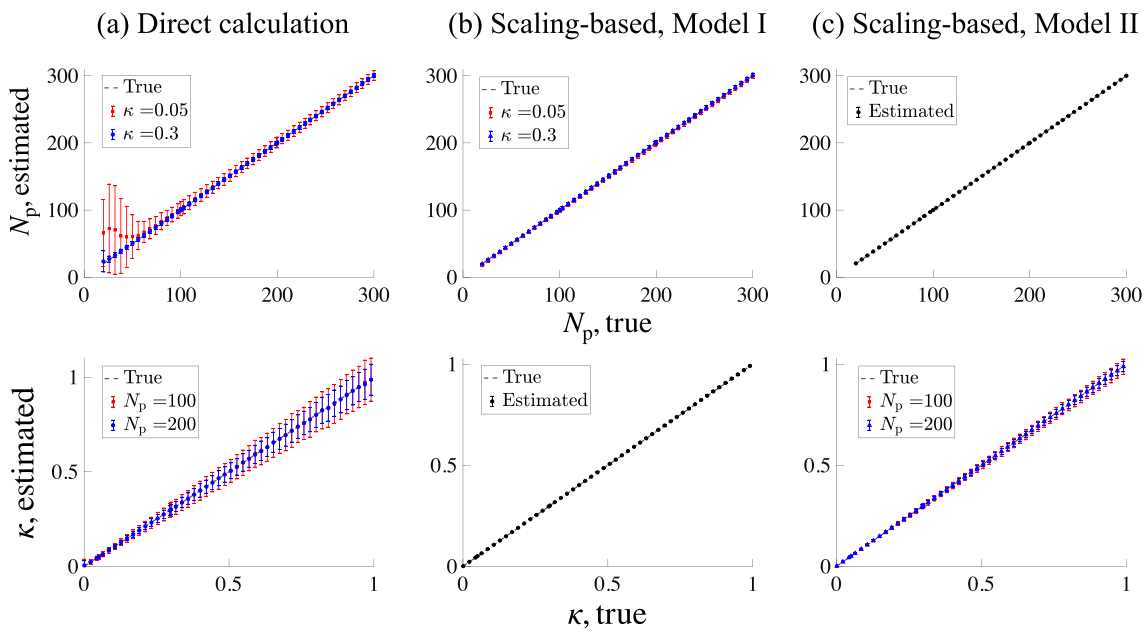

FIG. 4. Validation of estimation methods. Error bars are calculated over 1000 runs. (a) Solutions of Eqs. (8) and (9). (b) Estimation for Model I, in which N_p is time varying and κ is constant. (c) Estimation for Model II, in which N_p is constant and κ is time varying.

(resp. Model II), we generate 500 synthetic networks under various N_p ranging from 20 to 300 (resp. κ ranging from 0.001 to 0.99) for a given κ (resp. N_p).

While solving the system of two nonlinear equations is straightforward in principle, the question is whether the obtained solution matches the true values of κ and N_p. Obviously, the network generating mechanism is in reality not deterministic but stochastic, which means the same parameter combination (κ, N_p) may yield different observations of (N, M). Using a particular pair of (N, M) is therefore not sufficient to infer the true model parameters. Indeed, the solution of Eqs. (8) and (9) leads to a biased estimate of N_p especially when the true values of κ and N_p are small [Fig. 4(a)]. This is expected from Fig. 2(b) in which there is a large amount of data overlap in the lower left area of the corn. In fact, κ tends to take small values (e.g., <0.1) in real-world networks, in which case the biased estimation can become a serious problem.

Figures 4(b) and 4(c) shows the error bars of the estimated parameters for the second method over 1000 runs. The estimated values of N_p and κ nicely match the true values even when the network size is fairly small and thereby multiple combinations of (κ, N_p) can yield the same (N, M). This is an advantage of this method with which we do not rely on a particular realization of (N, M), but rather we exploit the whole dynamical relationship. Furthermore, in the case where N_p is fixed and κ varies in time, Model II also gives a better estimate than the direct calculation.

It should be noted that the observed N can be much lower than its potential value N_p, which suggests that the potential number of active nodes cannot necessarily be inferred directly from the observed number of nodes. This is particularly true when κ is so small that the network is fairly sparse (see Fig. S1 in the Supplemental Material [37]).

3. Empirical results

Figure 5 shows the empirical results for the CollegeMsg and LondonBike data sets (see Figs. S2 and S3 in the Supplemental Material [37] for the other data sets). Our results illustrate the fact that scaling relations in social and economic temporal networks may be driven by the two previously described factors. For Interbank, Enron, CollegeMsg, and RealityMining, Model I is selected, which means the time-varying nature of the global network properties comes from shifts in the potential number of nodes, i.e., the population in the system changes over time. On the other hand, for LondonBike and Highschool, Model II is selected, which means the population remains almost unchanged, and the changes in the numbers of edges and active nodes are due to time-varying connecting probabilities.

Since all we need for the model classification is a variety of combinations of (N, M), one can implement the method for any timescale. For instance, if we see intraday activity in the Interbank data set, the scaling behaviors on the vast majority of days are still better explained by Model I (Fig. S4). For the LondonBike dataset, the scaling relationship over different days is identified as being driven by a time-varying κ for each time interval (Fig. S5), again indicating that the population (i.e., the number of bike stations) is essentially fixed throughout the data period.

For the data sets for which Model I is selected, we note that the estimated values of κ are fairly small, ranging from 0.011 (RealityMining) to 0.078 (Interbank). $\widehat{\kappa}$ is time varying for LondonBike and Highschool, but their values are still small with the maximum value being no larger than 0.02. This suggests that the direct calculation discussed above would not work well for empirical networks [Fig. 4(a)].

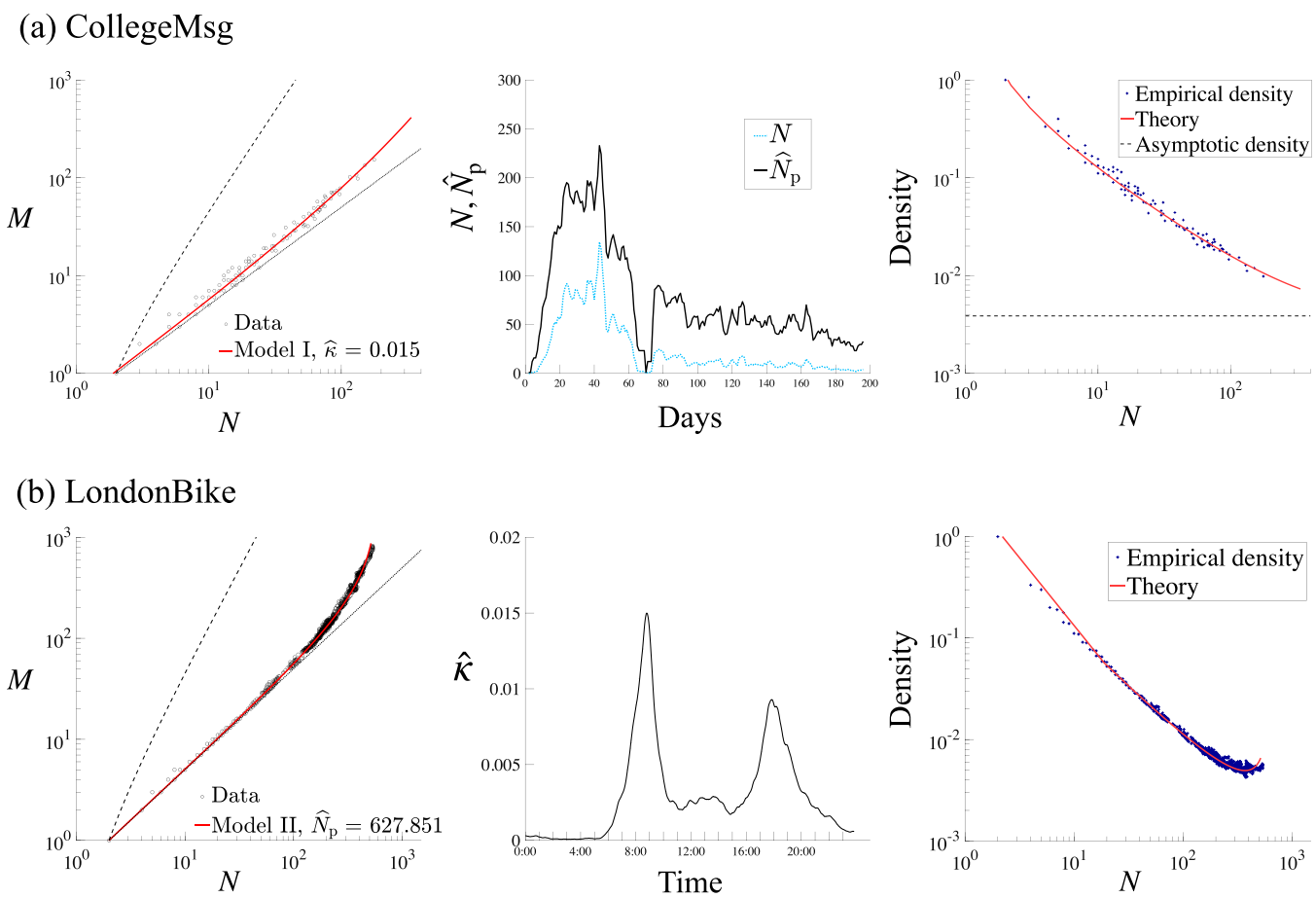

FIG. 5. Estimation results for (a) CollegeMsg and (b) LondonBike data sets. Model I is selected for CollegeMsg and Model II is selected for LondonBike. In each panel, the left column shows the fitted N-M curves. Black dotted and dashed lines denote the theoretical upper $[M = N(N-1)/2]$ and lower $(M = N/2)$ bounds, respectively. The middle column shows in (a) the estimated $\widehat{N}_p$ and in (b) the estimated $\hat{\kappa}$. The right column shows the fitted theoretical density. In (a), the asymptotic density derived from Eq. (14) is marked by the dotted line.

C. Network density

Another global quantity that might be of interest is network density. From the estimates of κ and N_p we can write the theoretical network density as

$$\frac{2M}{N(N-1)} = \frac{\widehat{\kappa}}{4}\left(\frac{1}{1-q_0(\widehat{\kappa},\widehat{N}_p)}\right)^2\left(1+\frac{q_0(\widehat{\kappa},\widehat{N}_p)}{N-1}\right). \quad (14)$$

As discussed above, the parameter q_0 approaches 0 as $N_p \to \infty$. This suggests that in Model I, in which κ is constant, the network density converges to $\kappa/4$ as N_p (and N) grows.

We compare the theoretical and the empirical network density in Fig. 5 (right panels) for CollegeMsg and LondonBike (see Fig. S6 for the other data sets). For the data sets for which Model I is selected (Interbank, Enron, CollegeMsg, and RealityMining), the density monotonically decreases as N increases, approaching the asymptotic value $\widehat{\kappa}/4$ (dashed line).

For the other data sets (LondonBike and Highschool), on the other hand, the relationship is nonmonotonic; density increases with N when N is sufficiently large. In Model II, where N_p is constant, the density can be regarded as a function of κ, and a shift in κ has two effects on the density. First, an increase in κ leads the network to be denser because it has a positive impact on the probability of two nodes being connected. Second, an increase in κ would cause the number of active nodes N to rise, which has a negative impact on the density. Since there is a finite fraction of inactive nodes when the network is not large enough (i.e., $q_0 > 0$), the number of active nodes can increase in accordance with a rise in κ. This increases the denominator of the density by definition, which would lead to a reduction in the theoretical density. Indeed, we find that there exists a threshold of N above which the former effect dominates the latter [Fig. 5(b), right].

D. A more general activity distribution

The estimation methods we proposed above assume that activity parameters $\{a_i\}$ are distributed uniformly, because in many real-world systems we have no prior knowledge about the activity level of (unobservable) resting nodes. Nevertheless, if we could have further information about the system (in addition to N and M), we could also obtain an estimate of the empirical activity distribution that covers the entire set of nodes.

In this section, we propose a method to estimate activity distribution when the total number of potentially active nodes

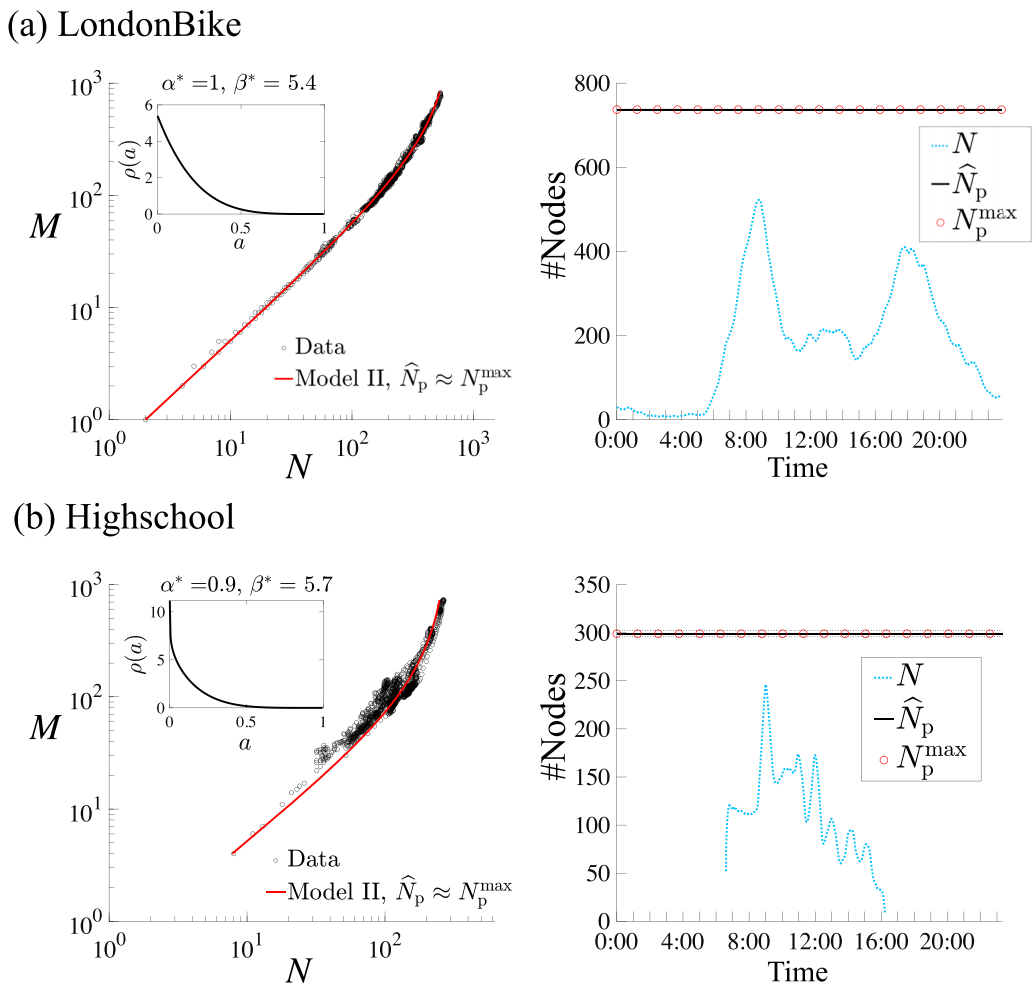

FIG. 6. Estimates based on generalized regression equations. In the left panels, the red line shows the theoretical equation [Eq. (15)] with optimized parameters (α^*, β^*). *Inset*: Estimated activity distribution $\rho(a)$. In the right panels, the number of active nodes N (dotted blue), estimated $\widehat{N}_p$ (black solid), and the empirical counterpart of N_p, denoted by N_p^{max} (red circle), are shown. The 5% confidence interval for $\widehat{N}_p$ is depicted by black dotted lines.

in the system (i.e., N_p) is known. We focus on the systems in which N_p is considered to be constant (i.e., systems for which Model II is selected), namely LondonBike and Highschool, and assume that the true N_p is given by the total number of active nodes of a day. The implicit assumption here is that nodes that are ready to be active would have at least one temporal edge during a day. We choose a beta distribution, $\rho(a) = f(a;\alpha,\beta) \equiv \frac{a^{\alpha-1}(1-a)^{\beta-1}}{B(\alpha,\beta)}$ for $a \in [0,1]$, as a general form for the activity distribution. Parameters α and β are estimated such that the estimated $\widehat{N}_p$ matches the empirical counterpart.

A generalized version of the nonlinear regression equation [Eq. (12)] is given by [see Eq. (B11) in Appendix B)

$$N = F(M;N_p,\alpha,\beta) \equiv N_p\left[1 - \int da f(a,\alpha,\beta) \times \left(1 - \left(\frac{\alpha+\beta}{\alpha}\right)\frac{2Ma}{N_p(N_p-1)}\right)^{N_p-1}\right]. \quad (15)$$

Note that endogenous variable κ is now expressed as a function of M, taking parameters N_p, α, and β as given: $\kappa(M;N_p,\alpha,\beta) \equiv (\frac{\alpha+\beta}{\alpha})^2 \frac{2M}{N_p(N_p-1)}$.

The estimation procedure under a generalized activity distribution is then given by the following four steps:

(1) For a given combination of (α,β), obtain the estimate of N_p, denoted by $\widehat{N}_p(\alpha,\beta)$, by implementing the nonlinear least squares on Eq. (15).

(2) Repeat step 1 for various combinations of (α,β).

(3) Find a combination of (α^*,β^*) such that $(\alpha^*,\beta^*) = \arg\min_{\alpha,\beta}|\widehat{N}_p(\alpha,\beta) - N_p^{max}|$, where N_p^{max} denotes the empirical counterpart of the total number of nodes in the system including temporally resting nodes.

(4) The estimator of N_p is given by $\widehat{N}_p(\alpha^*,\beta^*)$.

The estimation results suggest that the activity distribution is skewed to the left in both data sets (Fig. 6, insets), and the generalized regression equation still well fits the empirical N-M curve (Fig. 6). We note that while the goodness of fit generally improves due to the introduction of additional parameters (i.e., α and β), the fitted curve is little affected by the specification of activity distribution [see Figs. 5(b) and S2].

This suggests that the dynamic hidden variable model well explains the macroscopic fluctuations of empirical networks for alternative specifications of activity distribution.

V. DISCUSSION

We proposed a method to identify the source of scaling in temporal networks, namely the dynamical relationship between the numbers of active nodes and edges. Building on a model including both population and activity dynamics, we showed that these two mechanisms are sufficient to explain the two types of scaling observed in real-world systems. The estimating method we developed enables us to compute the parameters for the activity rhythm κ and the population size N_{p} (and thereby the number of resting nodes $N_{\mathrm{p}} - N$). While an observation of (N, M) in a particular snapshot is not sufficient to identify the source of dynamics, a sequence of N and M allows for such an estimation. We apply the method to six empirical data sets, and identify for each the main driving factor responsible for the emergence of scaling.

It should be noted that our proposed framework does not depend on whether the network under study is growing or shrinking. As we already pointed out, the only information needed for the method is a time variation of N and M. Indeed, in many real networks such as the six networks we examined, the size of networks does not exhibit a monotonic behavior but rather nonmonotonic shifts, depending on external factors that affect the activity rhythm and/or the population. Thus, the method can reveal the key factor that may lead a network to grow or shrink.

While our framework is useful for understanding the evolution of temporal networks in any contexts, there remain some issues that need to be addressed in future research. First, our method assumes that there are two types of systems, which are described as Model I (i.e., activity rhythm κ is constant and population size N_{p} is time varying) and Model II (i.e., population size N_{p} is constant and activity rhythm κ is time varying). In real-world systems, there may exist an intermediate state in which both the activity rhythm and the population size are evolving with similar time scales. To study those systems, one would need to include additional information other than N and M to inform the model, in order to be able to separate the effects of both mechanisms. Second, one key parameter of the model is the distribution of node fitnesses. Currently, we specified this distribution to be either a uniform or a beta distribution, which gives satisfactory estimates of the dynamical parameters. The method would of course yield more accurate estimates if we could incorporate an empirical distribution of fitnesses. However, measuring those is a complicated task: to do so, one needs to observe the activity levels of totally inactive nodes (i.e., nodes without edges), which is paradoxical. The fitness of a node in the model is indeed a rather abstract property, which integrates many realistic characteristics that depend on the context. Such characteristics can also be time dependent. Third, in the hidden variable model, the structure of the generated network is basically the same as that in a configuration model. Therefore, the model is not sufficient to replicate the empirical structural properties while the aggregate properties are well explained by the model. Explaining the structural and local properties, however, is beyond the scope our paper and should be left for future research.

ACKNOWLEDGMENTS

T.K. acknowledges financial support from JSPS KAKENHI Grants No. 15H05729 and No. 19H01506. M.G. acknowledges that this work was partially supported by the ANR project DATAREDUX (ANR-19-CE46-0008).

T.K. conceived and directed the study. T.K. and M.G. defined the model. T.K. performed the analytical calculations and the data analyses. T.K. and M.G. drafted the final manuscript.

The authors declare no competing financial interests.

APPENDIX A: DATA SETS

The Interbank data set is constructed from bilateral transactions in the online interbank market in Italy between September 4, 2000 and December 31, 2015 (i.e., 3922 business days). The data are commercially available from e-MID SIM S.p.A. based in Milan, Italy [38] From the data we build a temporal network where nodes are banks, with one snapshot per day. For each day, two banks are connected by an edge if a loan is made from a bank to another between 11:00 and 12:00.

The Enron data set is an email-based communication network from the Enron Corporation [7,25] collected from May 11, 1999 to June 21, 2002. From the data we build a temporal network where nodes are employees, with one snapshot per day. For each day, two employees are connected by an edge if at least one e-mail has been sent from one employee to the other between 14:00 and 16:00. The data are taken from [39].

The CollegeMsg data set is an online social network at the University of California, Irvine collected from Mar 23, 2004 to October 26, 2004 [8,9]. From the data we build a temporal network where nodes are users, with one snapshot per day. For each day, two users are connected if one has sent a private message to the other between 14:00 and 16:00. The data are taken from [40].

The RealityMining data set is built from the call data from the Reality Commons project [2] collected from September 24, 2004 to January 7, 2005. From the data we build a temporal network where nodes are individuals, with one snapshot per day. For each day, two individuals are connected if there has been a phone call between them or a voicemail has been left, during the 8:00–12:00 time window. The data are taken from [39].

The LondonBike data set describes the trips taken by customers of London Bicycle Sharing Scheme [26] collected on January 12, 2016. From the data we build a temporal network where nodes are bike sharing stations, with snapshots every 20 seconds aggregating the data from a 10-minute sliding time window. For each 10-minute time interval, two stations are connected if there has been at least one trip between them.

The Highschool data set is a face-to-face contacts network recorded in a high school in France on December 6, 2013, using wearable sensors by the SocioPatterns collaboration [1,19]. As in LondonBike, from the data we build a temporal network where nodes are individuals, with snapshots constructed every 20 seconds with a 10 minutes sliding time

window. For each 10-minute time interval, two individuals are connected if they have been at least once in contact.

APPENDIX B: FULL DERIVATION OF EQS. (8) AND (9)

In this Appendix we show a full derivation of Eqs. (8) and (9). The numbers of active nodes N and edges M are expressed as functions of parameters κ and $N_{\rm p}$:

$$N = [1 - q_0(\kappa, N_{\rm p})]N_{\rm p},$$
$$M = \frac{\bar{k}(\kappa, N_{\rm p})N_{\rm p}}{2}, \tag{B1}$$

where $q_0(\kappa, N_{\rm p})$ is the probability of a randomly chosen node being isolated and $\bar{k}(\kappa, N_{\rm p})$ denotes the average degree over all the existing nodes including isolated ones.

Given the vector of each node's activity $\vec{a} = (a_1, a_2, \ldots, a_{N_{\rm p}})$, the probability that node i has degree k_i is written as

$$g(k_i|\vec{a}) = \sum_{\vec{c}_i}\left[\prod_{j\neq i} u(a_i, a_j)^{c_{ij}}[1 - u(a_i, a_j)]^{1-c_{ij}}\right] \times \delta\left(\sum_{j\neq i} c_{ij}, k_i\right), \tag{B2}$$

where $c_{ij} \in \{0, 1\}$ is the (i, j) element of the $N_{\rm p} \times N_{\rm p}$ adjacency matrix, whose ith column is given by $\vec{c}_i = (c_{1i}, c_{2i}, \ldots, c_{N_{\rm p}i})^\top$, and function $\delta(x, y)$ denotes the Kronecker delta.

Let us redefine a product term in the square brackets of (B2) as

$$f_j(c_{ij}; a_i, a_j) \equiv u(a_i, a_j)^{c_{ij}}[1 - u(a_i, a_j)]^{1-c_{ij}}. \tag{B3}$$

Since $g(k_i|\vec{a})$ is the convolution of $\{f_j(c_{ij}; a_i, a_j)\}_j$, its generating function

$$\hat{g}_i(z|\vec{a}) \equiv \sum_{k_i} z^{k_i} g(k_i|\vec{a}) \tag{B4}$$

is decomposed as

$$\hat{g}_i(z|\vec{a}) = \prod_{j\neq i} \hat{f}_j(z; a_i, a_j), \tag{B5}$$

where $\hat{f}_j$ is the generating function of $f_j(c_{ij}; a_i, a_j)$, given by

$$\hat{f}_j(z; a_i, a_j) \equiv \sum_{a_{ij}} z^{a_{ij}} f_j(a_{ij}; a_i, a_j). \tag{B6}$$

Degree distribution $p(k_i; \kappa, N_{\rm p})$ is defined by the probability that node i has degree k_i and is related to $g(k_i|\vec{a})$ so that

$$p(k_i; \kappa, N_{\rm p}) = \int g(k_i|\vec{a})\rho(\vec{a})d\vec{a}, \tag{B7}$$

where we define $\rho(\vec{a}) \equiv \prod_i \rho(a_i)$ and $d\vec{a} \equiv \prod_i da_i$. Therefore, differentiation of $\hat{g}_i(z|\vec{a})$ with respect to z gives the average degree $\bar{k}(\kappa, N_{\rm p})$:

$$\begin{aligned}
\bar{k}(\kappa, N_{\rm p}) &= \sum_{k_i} k_i p(k_i; N_{\rm p}) \\
&= \sum_{k_i} k_i \int g(k_i|\vec{a})\rho(\vec{a})d\vec{a} \\
&= \frac{d}{dz}\int \hat{g}_i(z|\vec{a})\rho(\vec{a})d\vec{a}\Big|_{z=1} \\
&= \frac{d}{dz}\int \rho(a_i)da_i \prod_{j\neq i}\int \hat{f}_j(z; a_i, a_j)\rho(a_j)da_j\Big|_{z=1} \\
&= \int \rho(a_i)da_i \frac{d}{dz}\left[\int \hat{f}(z; a_i, h)\rho(h)dh\right]^{N_{\rm p}-1}\Big|_{z=1} \\
&= (N_{\rm p} - 1)\int \rho(a_i)da_i\left[\int da\rho(a)\hat{f}(z; a_i, a)\right]^{N_{\rm p}-2} \\
&\quad \times \int da\rho(a)\frac{d}{dz}\hat{f}(z; a_i, a)\Big|_{z=1}.
\end{aligned} \tag{B8}$$

From Eqs. (B3) and (B6), we have $\hat{f}(z; a_i, a) = \sum_{c_{ij}} z^{c_{ij}} f(c_{ij}; a_i, a) = (z - 1)u(a_i, a) + 1$. It follows that

$$\int da\, \rho(a)\hat{f}(z; a_i, a) = (z - 1)\int da\, \rho(a)u(a_i, a) + 1, \tag{B9}$$

$$\int da\, \rho(a)\frac{d}{dz}\hat{f}(z; a_i, a) = \int da\, \rho(a)u(a_i, a). \tag{B10}$$

From (B7), the probability of a node being isolated, $q_0(\kappa, N_{\rm p}) \equiv p(k_i = 0; \kappa, N_{\rm p})$, is given by

$$\begin{aligned}
q_0(\kappa, N_{\rm p}) &= \int g(k_i = 0|\vec{a})\rho(\vec{a})d\vec{a} \\
&= \int da_i\rho(a_i)\left[1 - \int u(a_i, a)\rho(a)da\right]^{N_{\rm p}-1}.
\end{aligned} \tag{B11}$$

Then, substituting $\rho(a) = 1$ (i.e., uniform distribution on $[0, 1]$) and $u(a, a') = \kappa aa'$ into Eq. (3) gives

$$\bar{k}(\kappa, N_{\rm p}) = \frac{\kappa}{4}(N_{\rm p} - 1). \tag{B12}$$

Similarly, substituting the same conditions into Eq. (B11) gives

$$q_0(\kappa, N_{\rm p}) = \int_0^1 \left(1 - \frac{\kappa a_i}{2}\right)^{N_{\rm p}-1} da_i. \tag{B13}$$

By defining a variable as $x \equiv 1 - \frac{\kappa a_i}{2}$, we have

$$\begin{aligned}
q_0(\kappa, N_{\rm p}) &= \frac{2}{\kappa}\int_{1-\frac{\kappa}{2}}^1 x^{N_{\rm p}-1}dx \\
&= \frac{2}{\kappa N_{\rm p}}\left[1 - \left(1 - \frac{\kappa}{2}\right)^{N_{\rm p}}\right].
\end{aligned} \tag{B14}$$

Note that $q_0(\kappa, 1) = 1$ and $\lim_{N_{\rm p}\to\infty} q_0(\kappa, N_{\rm p}) = 0$. Combining these results with Eq. (B1), we have

$$\begin{aligned}
N &= N_{\rm p}\left[1 - \frac{2}{\kappa N_{\rm p}}\left(1 - \left(1 - \frac{\kappa}{2}\right)^{N_{\rm p}}\right)\right], \\
M &= \frac{1}{8}\kappa N_{\rm p}(N_{\rm p} - 1).
\end{aligned} \tag{B15}$$

[1] http://www.sociopatterns.org/

[2] N. Eagle and A. Pentland, Reality mining: Sensing complex social systems, Pers. Ubiquitous Comput. **10**, 255 (2006).

[3] J.-P. Onnela *et al.*, Structure and tie strengths in mobile communication networks, Proc. Natl. Acad. Sci. USA **104**, 7332 (2007).

[4] C. Cattuto *et al.*, Dynamics of person-to-person interactions from distributed RFID sensor networks, PLoS ONE **5**, e11596 (2010).

[5] J. Stehlé *et al.*, High-resolution measurements of face-to-face contact patterns in a primary school, PLoS ONE **6**, e23176 (2011).

[6] P. Holme and J. Saramäki, *Temporal Networks* (Springer-Verlag, Berlin, 2013).

[7] B. Klimt and Y. Yang, The Enron corpus: A new dataset for email classification research, in *Machine Learning: ECML 2004* (Springer-Verlag, Berlin, 2004), pp. 217–226.

[8] T. Opsahl, V. Colizza, P. Panzarasa, and J. J. Ramasco, Prominence and Control: The Weighted Rich-Club effect, Phys. Rev. Lett. **101**, 168702 (2008).

[9] P. Panzarasa, T. Opsahl, and K. M. Carley, Patterns and dynamics of users' behavior and interaction: network analysis of an online community, J. Am. Soc. Inf. Sci. Technol. **60**, 911 (2009).

[10] M. Schläpfer *et al.*, The scaling of human interactions with city size, J. R. Soc. Interface **11**, 20130789 (2014).

[11] A. Ghosh, D. Monsivais, K. Bhattacharya, R. I. Dunbar, and K. Kaski, Quantifying gender preferences in human social interactions using a large cellphone dataset, EPJ Data Sci. **8**, 9 (2019).

[12] A. Barrat *et al.*, Empirical temporal networks of face-to-face human interactions, Eur. Phys. J.: Spec. Top. **222**, 1295 (2013).

[13] L. Isella *et al.*, What's in a crowd? Analysis of face-to-face behavioral networks, J. Theor. Biol. **271**, 166 (2011).

[14] M. Starnini, A. Baronchelli, and R. Pastor-Satorras, Modeling Human Dynamics of Face-to-Face Interaction Networks, Phys. Rev. Lett. **110**, 168701 (2013).

[15] M. Génois *et al.*, Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers, Netw. Sci. **3**, 326 (2015).

[16] N. Masuda and R. Lambiotte, *A Guide to Temporal Networks* (World Scientific, Singapore, 2016).

[17] G. Iori *et al.*, Networked relationships in the e-MID interbank market: A trading model with memory, J. Econ. Dyn. Control **50**, 98 (2015).

[18] T. Kobayashi and T. Takaguchi, Social dynamics of financial networks, EPJ Data Sci. **7**, 15 (2018).

[19] R. Mastrandrea, J. Fournet, and A. Barrat, Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys, PLoS ONE **10**, e0136497 (2015).

[20] M. Génois and A. Barrat, Can co-location be used as a proxy for face-to-face contacts? EPJ Data Sci. **7**, 11 (2018).

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (ACM Press, New Yotk, 2005), pp. 177–187.

[22] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).

[23] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Scale-Free Networks from Varying Vertex Intrinsic Fitness, Phys. Rev. Lett. **89**, 258702 (2002).

[24] M. Boguñá and R. Pastor-Satorras, Class of correlated random networks with hidden variables, Phys. Rev. E **68**, 036112 (2003).

[25] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters, Internet Math. **6**, 29 (2009).

[26] F. Munoz-Mendez, K. Han, K. Klemmer, and S. Jarvis, Community structures, interactions and dynamics in London's bicycle sharing network, in *Proceedings of the 2018 ACM International Joint Conference and Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (ACM Press, New York, NY, 2018), pp. 1015–1023.

[27] L. M. A. Bettencourt, D. I. Kaiser, and J. Kaur, Scientific discovery and topological transitions in collaboration networks J. Informetrics **3**, 210 (2009).

[28] S. N. Dorogovtsev and J. F. F. Mendes, Effect of the accelerating growth of communications networks on their structure, Phys. Rev. E **63**, 025101(R) (2001).

[29] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[30] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Structure of Growing Networks with Preferential Linking, Phys. Rev. Lett. **85**, 4633 (2000).

[31] P. L. Krapivsky and S. Redner, Organization of growing random networks, Phys. Rev. E **63**, 066123 (2001).

[32] G. De Masi, G. Iori, and G. Caldarelli, Fitness model for the Italian interbank money market, Phys. Rev. E **74**, 066112 (2006).

[33] M. Starnini and R. Pastor-Satorras, Topological properties of a time-integrated activity-driven network, Phys. Rev. E **87**, 062807 (2013).

[34] H. H. Jo, M. Karsai, J. Kertész, and K. Kaski, Circadian pattern and burstiness in mobile phone communication, New J. Phys. **14**, 013055 (2012).

[35] T. Aledavood, S. Lehmann, and J. Saramäki, Digital daily cycles of individuals, Front. Phys. **3**, 73 (2015).

[36] T. Kobayashi, T. Takaguchi, and A. Barrat, The structured backbone of temporal social ties, Nat. Commun. **10**, 220 (2019).

[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.102.052302 for details.

[38] http://www.e-mid.it

[39] http://networkrepository.com

[40] http://snap.stanford.edu