HAL open science

The switching mechanisms of social network densification

Teruyoshi Kobayashi, Mathieu Génois

To cite this version:

Teruyoshi Kobayashi, Mathieu Génois. The switching mechanisms of social network densification. Scientific Reports, 2021, 11 (1), pp.3160. 10.1038/s41598-021-82432-9 . hal-03228432

HAL Id: hal-03228432

https://hal.science/hal-03228432

Submitted on 3 Jun 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.

Distributed under a Creative Commons Attribution 4.0 International License

scientific reports

OPEN

The switching mechanisms of social network densification

Teruyoshi Kobayashi[1,2]✉ & Mathieu Génois[3,4]✉

Densification and sparsification of social networks are attributed to two fundamental mechanisms: a change in the population in the system, and/or a change in the chances that people in the system are connected. In theory, each of these mechanisms generates a distinctive type of densification scaling, but in reality both types are generally mixed. Here, we develop a Bayesian statistical method to identify the extent to which each of these mechanisms is at play at a given point in time, taking the mixed densification scaling as input. We apply the method to networks of face-to-face interactions of individuals and reveal that the main mechanism that causes densification and sparsification occasionally switches, the frequency of which depending on the social context. The proposed method uncovers an inherent regime-switching property of network dynamics, which will provide a new insight into the mechanics behind evolving social interactions.

Network representation of complex interactions among elements is an overarching framework heavily used in many fields of science[1–3]. For social systems, the dynamics of interactions between individuals (whether electronic, online or face-to-face) can be represented as time-varying networks, often called temporal networks, in which nodes come and go and edges are activated or deactivated as time goes on[4,5]. Many essential features of human behaviour encoded in the representation of temporal networks have been revealed over the past decade, such as burstiness[6–8], circadian/diurnal rhythms[9], temporal communities[10], higher-order interactions[11,12], etc.

While the studies of temporal networks shed light on the time-varying nature of interactions between nodes, dynamics in social systems emerge not only at the local level[13], but also at the global level. In a wide variety of social contexts, network size (i.e., the number of active nodes) and the number of edges observed at a given point in time are very often not constant, and accordingly the average degree increases or decreases[14,15]. In fact, the numbers of aggregate nodes and edges have been shown to have a scaling relationship known as the densification power law or densification scaling[14]. In temporal networks (i.e., a sequence of snapshot networks), any variation in the number of active nodes N and the number of edges M can be *a priori* attributed to changes in (i) the population in the system (e.g., the number of students present in a school, the number of attendees in a conference, etc); (ii) the probability of two nodes being connected; or (iii) both. With a constant probability of edge creation, N and M will increase if more nodes enter the system, since each node will have a higher chance of finding partners. Likewise, for a given population, if the probability of two nodes being connected increases, M will surely increase, and N will rise as well as isolated nodes, if they exist, will be more likely to get connected.

These two mechanisms are fundamental factors that bring about the dynamics of N and M, yet separating their contributions based on the dynamical behaviour of N and M is a challenging problem. In a wide variety of social and economic systems, network dynamics are likely to be driven by a mixture of these two mechanisms, and moreover their relative importance may occasionally change as the network evolves[15]. In theory, each of these two mechanisms leads to a distinctive type of densification scaling; The first one, generated by the evolution of population, is a scaling behaviour similar to the typical densification scaling in which the number of edges M scales with the number of active nodes N with a constant exponent α, i.e., $M \propto N^{\alpha}$[14]. The second one is an accelerating growth of M, which is caused by the evolution of the probability of edge creation[15]. In fact, for the human contact networks we study, neither of these two types of scaling is observed in their original form. Rather, we observe a "mixed" scaling behaviour which appears to be a composite of the two types and therefore cannot be explained by a single scaling law.

Here, we develop a Bayesian statistical method to identify the source of dynamics generating network densification and sparsification based on the sequence of N and M. To take into account possible changes in the source of dynamics, we derive two specifications (i.e., "regimes") for the solution of a simple generative model, namely a dynamic hidden variable model, each of which capturing one of the two fundamental mechanisms. By

[1]Department of Economics, Kobe University, Kobe, Japan. [2]Center for Computational Social Science, Kobe University, Kobe, Japan. [3]CNRS, CPT, Aix Marseille Univ, Université de Toulon, Marseille, France. [4]GESIS, Leibniz Institute for the Social Sciences, Köln/Mannheim, Germany. ✉email: kobayashi@econ.kobe-u.ac.jp; mathieu.genois@cpt.univ-mrs.fr

fitting the two specifications simultaneously to the observed mixed scaling relationship using a unified estimation framework, known as the Markov regime-switching model[16,17], we are able to estimate the probability that the dynamical source of densification or sparsification at a given point in time is attributed to a particular mechanism. At the same time, the Bayesian inference also allows us to trace the paths of the time-varying parameters directly related to the dynamical source, i.e., the population in the system and the activity level of nodes. An important advantage of the regime-switching model is that it allows the "true" model specification to occasionally switch, possibly depending on the social context.

In this work we analyse networks of face-to-face human interactions collected by the SocioPatterns collaboration[18]. We focus on four datasets: contact networks in two scientific conferences, a hospital and a workplace. Such networks can indeed be affected by the two fundamental mechanisms at the same time, because (i) individuals can always enter and exit the system, and (ii) presence of a time schedule could facilitate or inhibit face-to-face interactions (e.g., attendees of a conference are more likely to have interactions during coffee breaks than during keynote talks). In particular, using data on academic conferences has an important advantage, as it allows us to compare the dynamical regimes detected by the proposed method with the "ground-truth" conference time schedules. We find indeed that during keynote talks, parallel sessions and coffee breaks, the temporal densification and sparsification in the contact networks formed by conference attendees are mainly related to shifts in the chance of contacts being made between attendees present at the venue. On the other hand, shifts in the population are the main driving force of densification and sparsification during registration and poster sessions. This result is consistent with our intuition that the number of attendees in the middle of the program would be mostly constant, while it may be more likely to change during registration, which is held in the morning, and poster sessions in which not all of the attendees participate. For contact networks in a hospital and a workplace, this kind of comparison with a prespecified time schedule is not possible because there is no such rigorous time constraints to follow. Nevertheless, in all the systems we examined, the proposed method reveals that the main driving force of network densification and sparsification is occasionally switching, suggesting that the formation of social ties in physical space generally involves multiple dynamical sources.

Results

Empirical evidence on mixed densification scaling. We focus our analysis on temporal contact networks taken from the following four datasets:

- **WS-16**: Contacts between participants of the Computational Social Science Winter Symposium 2016 at GESIS in Cologne on November 30, 2016[19].
- **IC2S2-17**: Contacts between participants of the International Conference on Computational Social Science 2017 at GESIS in Cologne on July 12, 2017[19].
- **Hospital**: Contacts among patients, nurses, doctors and staffs in a Hospital in Lyon on December 8, 2010[20].
- **Workplace**: Contacts between workers in a office building in France on June 27, 2015[21].

These data consist of contacts between individuals collected every 20 seconds using RFID sensors[18,22]. A "contact" is here defined as a physical, face-to-face proximity event. The datasets thus give us temporal networks in which nodes are individuals and edges encode the contacts occurring between them. All datasets exhibit large and abrupt fluctuations of the number of edges that are typical in these non-stationary systems (see Fig. 1, lower panels). In these particular contexts of social interactions, these transitions between high an low activity periods are often related to specified schedules: from talk sessions to coffee breaks in the conferences, changes in shifts in the hospital, from desk work to meetings in the workplace.

In many social and economic dynamical networks, the numbers of aggregate edges and nodes have a superlinear scaling relationship called the "densification power law"[14,23,24], in which the average degree is increasing with the number of nodes, i.e., "densification". For temporal networks, where there is a sequence of network snapshots, a similar type of scaling emerges from the dynamics of the population, in which nodes enter and leave the system, keeping the chance of two nodes being connected constant[15,25]. However, another type of scaling emerges in real-world systems for which the population is fixed. In such systems densification is "explosive", with the scaling exponent increasing with N[15]. While these two classes of scaling could be differentiated and identified from data if we observe a specific type of scaling[15], in general there may exist a mixture of them that cannot be easily classified as one of the two classes. Indeed, in the four datasets we study, no clear scaling relationship appears (Fig. 1, upper panels). In the following we show that the mixed shape of empirical densification behavior reflects a mixing of both classes of scaling.

Two dynamical regimes in the dynamic hidden-variable model. To explore the temporal dynamics of densification and sparsification, we consider a dynamic version of the hidden variable model. The probability that two nodes i and j are in contact within a given time window t is:

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

where a_i is the "fitness" that represents the intrinsic activity level of node i[26–28], and T denotes the last time window in the data. There are two time-varying parameters in the model. The first one is $\kappa_t > 0$, which modulates the overall activity rhythm of nodes. A variation in κ would reflect the time-schedule of a conference or a school, working hours in an office or a hospital, or the circadian rhythm of individuals[9,22,29,30]. The second time-varying parameter $N_{\mathrm{p},t}$ denotes the potential number of active nodes at time t, i.e., the total of active and inactive nodes that are in the room or the building. It should be noted that although the number of active nodes (i.e., nodes

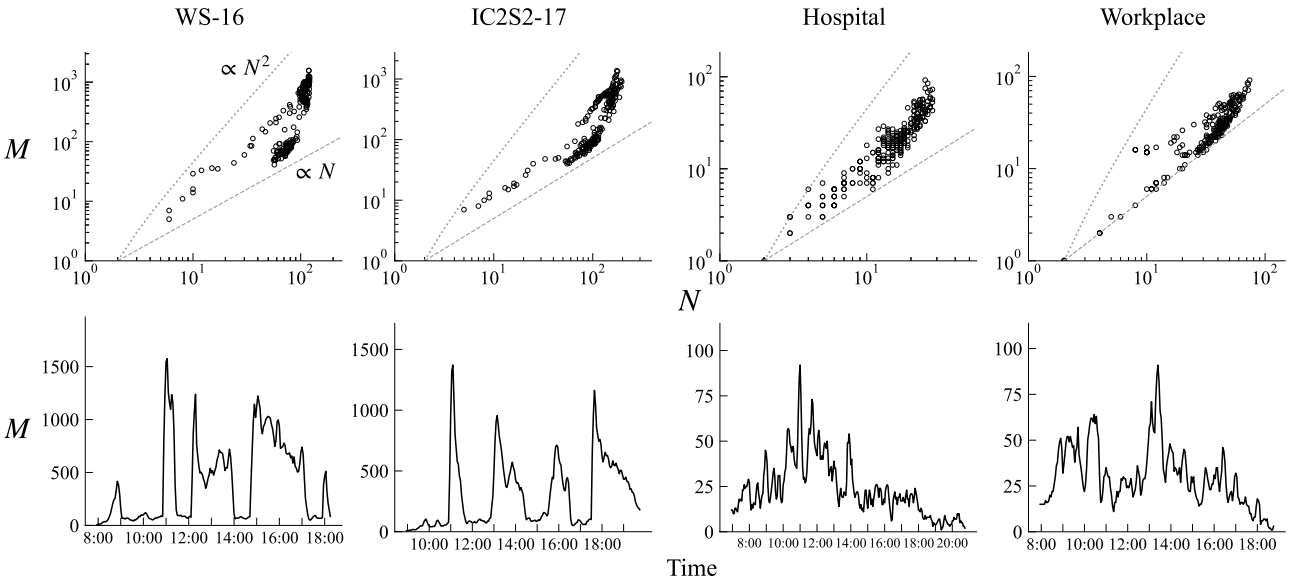

Figure 1. Dynamical behaviour of number of active nodes N and number of active edges M. In upper panels, dynamical relationship between N and M is shown. Each dot represents a snapshot network created over a 10-min time window. Gray dashed and dotted lines respectively denote $N/2$ (i.e., the lower bound for M) and $N(N-1)/2$ (i.e., the upper bound for M). Lower panels show the behaviour of M over time.

having at least one edge) N_t is always observable from the data, the potential number of nodes $N_{p,t}$ is not. We do not usually know how many people were actually in the room at a given time because people could enter and exit the room at any time without being interacting with any other individual. We can observe the number of active nodes that appear in the record of contacts, but in many cases there is no record of nodes without any interaction. We assume that activity a_i is uniformly distributed on [0, 1], because (i) we do not have any prior information about the full distribution of the activity levels of all nodes including isolated ones, and (ii) introducing a more general distribution prohibits us from obtaining an analytical solution, which makes it difficult to implement parameter estimation.

The average numbers of active nodes N and edges M are analytically given as (see "Analytical expression for N and M" section in Methods for derivation):

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

$$M = \frac{1}{8}\kappa N_p(N_p - 1), \tag{3}$$

where we drop time subscript t for brevity. From these expressions, it is clear that the two parameters κ and N_p play different roles in the determination of N and M, but it is not clear how N and M correlate. To see the direct relationship between N and M, we eliminate one of the two parameters in Eq. (2), using Eq. (3). By doing this, we can effectively endogenise either κ or N_p. Depending on whether we endogenise κ or N_p, we obtain different functional forms that connect N and M.

Regime 1: N_p-driven dynamics. First, let us consider the case of time-varying N_p. This is a situation in which the dynamics of N and M are fully driven by changes in the population. We call this system as being in "Regime 1" or "state 1":

Definition 1. A system is in Regime 1 if N_p is time-varying and κ is constant, in which case the dynamical relationship between N and M is given by:

$$\begin{aligned} N_t &= N_p(M_t,\kappa)\left[1 - \frac{2}{\kappa N_p(M_t,\kappa)}\left(1 - \left(1 - \frac{\kappa}{2}\right)^{N_p(M_t,\kappa)}\right)\right] \\ &\equiv h^1(M_t;\kappa), \end{aligned} \tag{4}$$

where the time-varying N_p value is expressed as a function of M_t and κ: $N_p(M_t,\kappa) \equiv \frac{1+\sqrt{1+32M_t/\kappa}}{2}$ (see, Eq. 3).

For the purpose of parameter estimation, we introduce an error term as $N_t = h^1(M_t;\widehat{\kappa}) + \varepsilon_{1,t}$, where $\widehat{\kappa}$ denotes the estimated value of κ, and ε_t^1 is a residual term following a normal distribution with mean zero and standard deviation σ_1. Estimated value of $N_{p,t}$ when the system is in Regime 1 leads to:

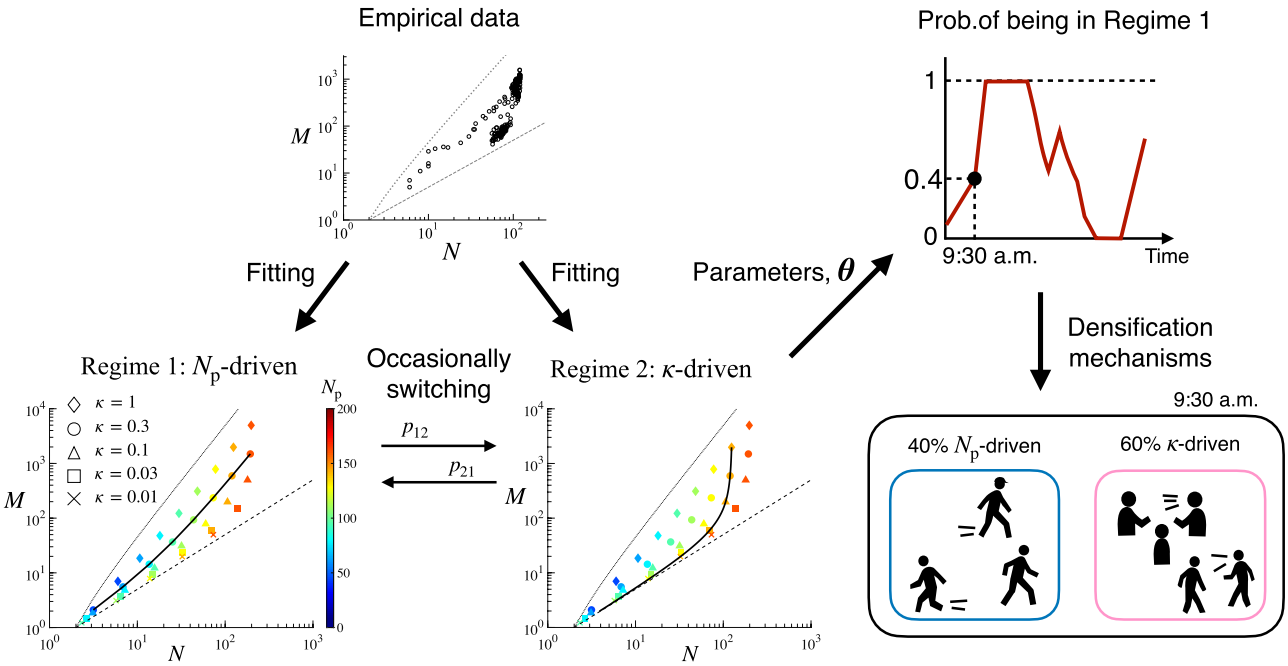

Figure 2. Schematic of the identification method. Empirical densification is fitted to the regime switching model in which the model switches from Regime 1 to Regime 2 (resp. from Regime 2 to Regime 1) with probability p_{12} (resp. p_{21}). Then, the estimated parameters are used to infer the probability of the system being in Regime 1 at a given time t. For panels at lower left and lower middle, different colours denote different N_p, and different symbols denote different κ (see Eqs. 2 and 3). Solid line in the bottom left (resp. bottom middle) denotes the N_p-driven (resp. κ-driven) scaling with $\kappa = 0.3$ (resp. $N_p = 126$). If the scaling is N_p-driven (resp. κ -driven), the time variation of N and M is fully caused by shifts in N_p (resp. κ).

$$\widehat{N}_{p,t}|_{S_t=1} = \frac{1+\sqrt{1+32M_t/\widehat{\kappa}}}{2}, \tag{5}$$

where $S_t = 1$ denotes the fact that the system is in Regime 1 at time t. In Regime 1, network dynamics is totally driven by the time-varying nature of the population, what we call "N_p-driven" dynamics. For a given κ, the slope of densification scaling is close to constant, while different κ yield different slopes (Fig. 2, lower left).

Regime 2: κ-driven dynamics. Next, let us consider the case of time-varying κ. This corresponds to a situation in which the dynamics of the system is fully driven by changes in the overall activity of individuals. We call this system as being in "Regime 2" or "state 2":

Definition 2. A system is in Regime 2 if κ is time-varying and N_p is constant, in which case the dynamical relationship between N and M is given by

$$\begin{aligned} N &= N_p\left[1-\frac{2}{\kappa(M,N_p)N_p}\left(1-\left(1-\frac{\kappa(M,N_p)}{2}\right)^{N_p}\right)\right] \\ &\equiv h^2(M;N_p), \end{aligned} \tag{6}$$

where the time-varying value of κ is expressed as a function of M_t and N_p: $\kappa(M_t,N_p) \equiv \frac{8M_t}{N_p(N_p-1)}$ (see, Eq. 3).

For estimating, we add an error term as $N_t = h^2(M_t;\widehat{N}_p) + \varepsilon_{2,t}$, where $\widehat{N}_p$ denotes the estimated value of N_p, and $\varepsilon_{2,t}$ is a residual term following a normal distribution with mean zero and standard deviation σ_2. Estimated value of κ at time t when the system is in Regime 2 leads to:

$$\widehat{\kappa}_t|_{S_t=2} = \frac{8M_t}{\widehat{N}_p(\widehat{N}_p-1)}. \tag{7}$$

In Regime 2, network dynamics is fully driven by the individuals' time-varying activity levels, what we call "κ -driven" dynamics, and the slope of densification scaling in fact increases with N (Fig. 2, lower middle). This kind of accelerating growth of M naturally happens when edges are created in a fixed-population system, in which case the network tends to be denser as the number of inactive nodes vanishes.

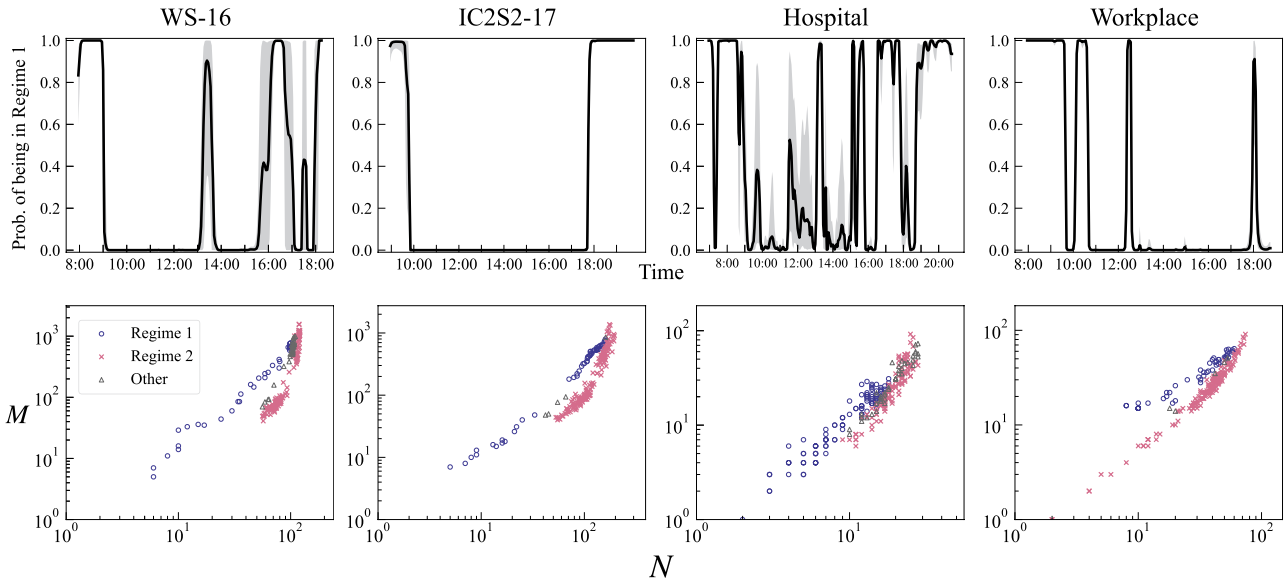

Figure 3. Identification of dynamical regime. Upper panels show the smoothed probability of being in Regime 1 (i.e., N_p-driven dynamics) at each time window. 95 % credible interval is indicated by shading. Lower panels show *N*-*M* plots with classified regimes being denoted by different colours and symbols. We identify a snapshot network as being in Regime 1 (resp. Regime 2) if the estimated probability of being in Regime 1 (resp. Regime 2) is greater than 0.5 in more than 95 % of MCMC sampling. Otherwise, a network is considered as being in an undetermined "gray area".

Analysis of switching dynamics behind temporal densification and sparsification. *A Markov regime switching model.* In real-world networks, the mechanism of densification and sparsification may occasionally change depending on the context, such as working schedule, coffee breaks, lunch time, etc. To incorporate such a possibility, we propose a unified framework based on the Markov regime switching model in which the hidden state of a system can switch from Regime 1 to Regime 2 (respectively from Regime 2 to Regime 1) with probability p_{12} (resp. p_{21})[16,17]. An important advantage of the regime switching model is that it allows us to calculate the probability of a system being in Regime $s \in \{1, 2\}$ at time t for a given parameter set $\boldsymbol{\theta} = \{N_p, \kappa, \sigma_1, \sigma_2, p_{11}, p_{22}\}$. This probability of the system being in Regime s can then be interpreted as the relevancy of each mechanism in explaining the densification dynamics at a given time (Fig. 2). We employ a Bayesian approach for the estimation of the parameters, using the Markov chain Monte Carlo (MCMC) to obtain posterior distributions (see, Methods "Bayesian estimation" for the estimation method). It should be noted that $\boldsymbol{\theta}$ does not contain $N_{p,t}|_{S_t=1}$ and $\kappa_t|_{S_t=2}$ since $\boldsymbol{\theta}$ only contains constant parameters to be directly estimated by the Bayesian method.

In the following, we use the smoothed probability $\Pr(S_t = s|\psi_T; \boldsymbol{\theta})$ which is calculated conditional on all the information available at time T, denoted by ψ_T (see, Methods "Smoothed probability" for full derivation) [31]. Validation analyses using synthetic networks show that the proposed method correctly detects the switching of regimes and estimates the model parameters quite accurately (Table S1, Figs. S1 and S2 in Supporting Information (SI)). Given the probability of being in Regime $s \in \{1, 2\}$, we can estimate the dynamical parameters $N_{p,t}$ and κ_t as:

$$\widehat{N}_{p,t} = \Pr(S_t = 1|\psi_T; \widehat{\boldsymbol{\theta}}) \cdot \widehat{N}_{p,t}|_{S_t=1} \; + \; \Pr(S_t = 2|\psi_T; \widehat{\boldsymbol{\theta}}) \cdot \widehat{N}_p, \tag{8}$$

$$\widehat{\kappa}_t = \Pr(S_t = 1|\psi_T; \widehat{\boldsymbol{\theta}}) \cdot \widehat{\kappa} \; + \; \Pr(S_t = 2|\psi_T; \widehat{\boldsymbol{\theta}}) \cdot \widehat{\kappa}_t|_{S_t=2}, \tag{9}$$

where $\widehat{\boldsymbol{\theta}}$ denotes the set of estimated parameters, which is summarised in Table 1 in Methods.

Classification of network dynamics. The Bayesian estimation of the parameters suggests that the empirical systems' dynamics are indeed occasionally switching between N_p-driven and κ-driven (Fig. 3, upper panels). For the conference data, a common feature is that the probability of being in Regime 1 is almost 1 prior to the first session and after the last keynote session of the day, and mostly zero in between (see Fig. 4 for the correspondence between the dynamics and the schedule of the conferences). For WS-16, we see further fluctuations between the two regimes, one linked to the lunch break, the other to the poster session which closed the day. This suggests that the dynamics during the oral sessions, keynote talks and breaks are mainly driven by changes in the activity level of participants, while in the "opened" time slots, such as registration, closing and poster session, their dynamics are explained by time-varying population. The same patterns linked to the schedule are found on the other days of the conferences (see S3a–c).

For the Workplace data we see a roughly similar pattern (Fig. 3, top right). The dynamics in the early morning and evening are driven by a variation in N_p, as well as around lunch time and coffee break, and changes in

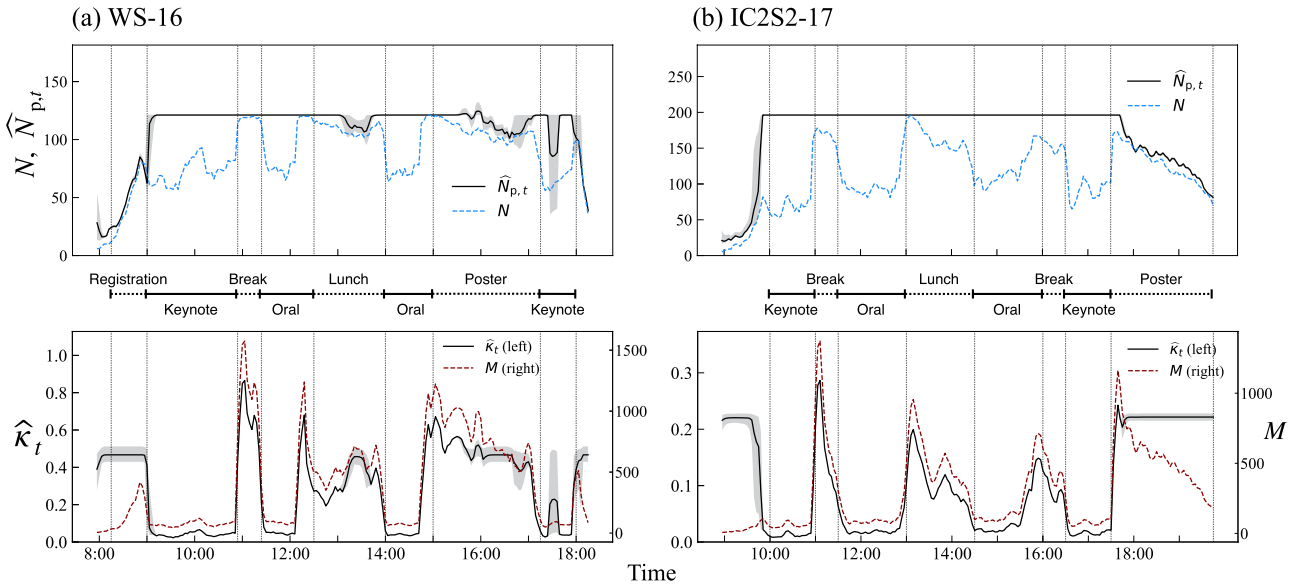

Figure 4. Estimation of $N_{p,t}$ and κ_t for (**a**) WS-16 and (**b**) IC2S2-17. $\widehat{N}_{p,t}$ and $\widehat{\kappa}_t$ are shown in the upper and the lower panels, respectively, and the 95 % credible interval is indicated by shading. Upper panels show the number of active nodes (dashed blue line) at each time, thus the difference between the two lines represents the number of isolated nodes. Lower panels also show the number of edges at each time (dashed red line). Vertical dotted lines indicate the time windows of the scheduled sessions, with the labels in the middle.

activity level are the main source of dynamics in between. This is, of course, not necessarily a general property of contact networks in physical space. We also see that the regime remains almost constant in most of the day (Fig. S3e), or there might be days in which the regime constantly changes throughout the day (Fig. S3f). In the case of the Hospital data, there is no clear tendency for the regime-switching pattern (Fig. 3, third column and S3d), which seems natural for such an open environment with visitors and medical workers coming and going, and no general, fixed schedule for working hours.

We next attempt to classify the snapshot networks into two groups based on their probability of being in a particular regime. We identify a snapshot network at t as being in Regime 1 (resp. Regime 2) if more than 95 % of samples for the value of $\Pr(S_t = 1|\psi_T; \boldsymbol{\theta})$ generated by MCMC are greater than 0.5 (resp. lower than 0.5), i.e., in more than 95 % of parameter sampling the dynamics at t is considered to be attributed to Regime 1 (resp. Regime 2). Otherwise, the system is considered as being in an undetermined "gray area". As seen in the lower panels of Fig. 3, the location of snapshot networks in the N-M space is strongly related to which regimes they belong to. As expected, the snapshots in Regime 1 exhibit a scaling whose slope is almost constant (i.e., N_p -driven scaling), while the snapshots in Regime 2 exhibit accelerating growth patterns (i.e., κ-driven scaling). Classifying each time window according to the underlying dynamical mechanism is essentially equivalent to identifying patterns in the N-M space.

Temporal dynamics of population and activity level. We also examine the evolution of the dynamical parameters for both regimes (Fig. 4). For the two conferences (WS-16 and IC2S2-17), the estimated population size $\widehat{N}_{p,t}$ increases at the beginning of the day and decreases at the end, consistent with the dynamics of participants entering and exiting the venue. The estimated activity parameter $\widehat{\kappa}_t$ is high during these periods, and the level is consistent with those seen in highly active windows during social breaks. During the main program, the population is virtually constant and the size is consistent with the number of attendants ($\sim$ 120 for WS-16, $\sim$ 200 for IC2S2-17). The variation of network size is thus mainly driven by the schedule, which constrains the participants' networking activity. In the case of WS-16, the fluctuation of $\widehat{N}_{p,t}$ during the lunch break and the poster session are worth noting since the variation of observed network size N seems to be driven by both mechanisms; we see slight reductions in the estimated population while the overall activity is still high in these time windows. This demonstrates the ability of the proposed method to extract mixed-regime periods in which both of the two mechanisms are at work (see Fig. 2, right, for schematic). Similar patterns are also found in the other days (see S4).

In the Hospital data, the regime-switching dynamics is much less periodic, with lots of transitions and mixed periods (Fig. 5a). This is however not surprising, because there is no fixed schedule regulating either the activity or the number of people present. For the Workplace data, we also do not expect *a priori* to see a clear segmentation of regimes because of the absence of a rigid schedule as in a hospital. However, the dynamics uncovered by our method indicates that the situation is much simpler than that for Hospital, as there seem to be less variation in population size, aside from the "opening" and "closing" effects and a reduction in population around the lunch time (Fig. 5b). The day that exhibited many regime switches presents however many episodes of small variations in population size (see S4), similar to the dynamics observed in a Hospital.

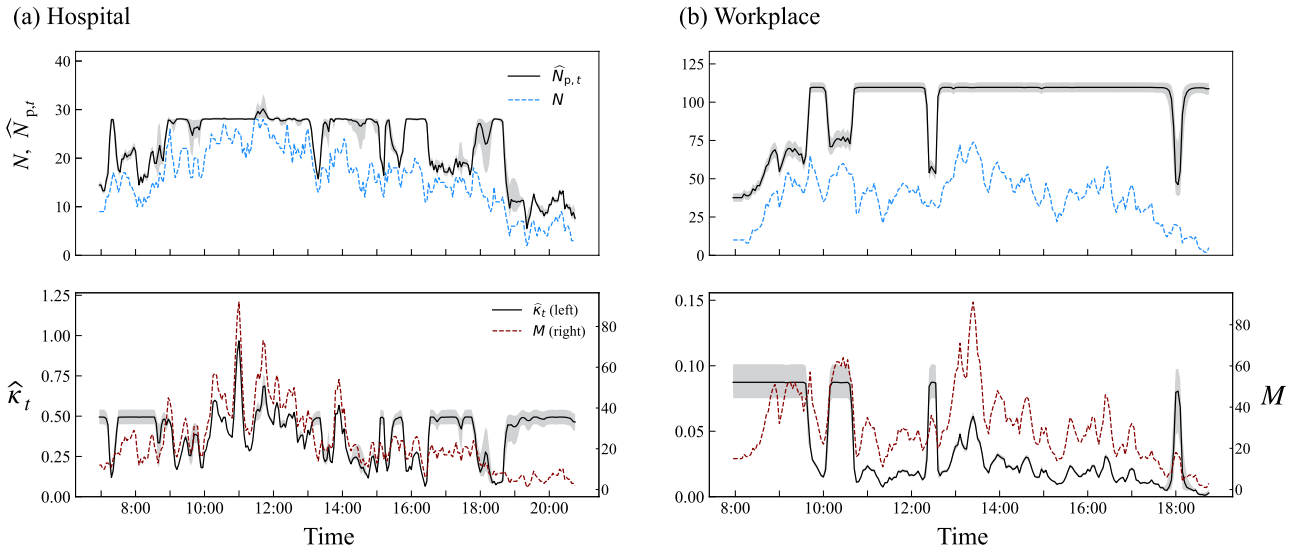

Figure 5. Estimation of $N_{p,t}$ and κ_t for (**a**) Hospital and (**b**) Workplace.

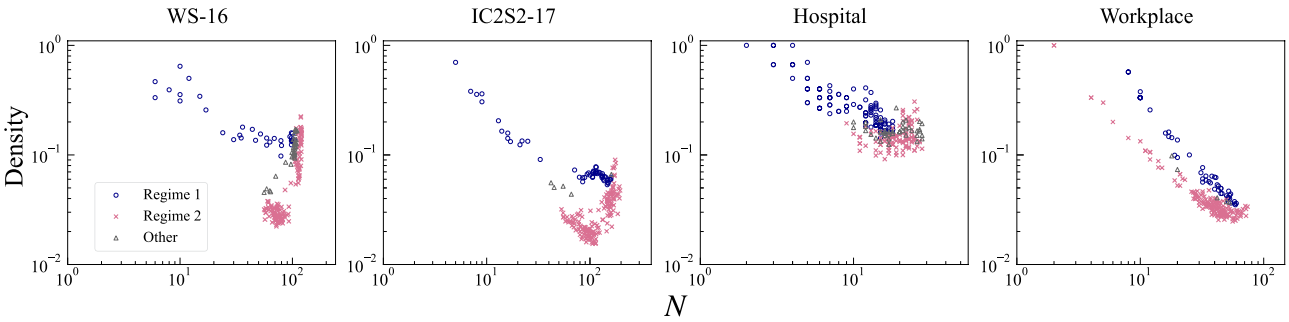

Figure 6. Density versus the number of active nodes. Classification of dynamical regimes is conducted in the same way as in Fig. 3.

Non-monotonic behaviour of network density. Since both types of scaling emerging from two different dynamics exhibit superlinearity, the average degree is always increasing in N. However, the density of networks, defined by $2M/(N(N-1))$, is not always increasing with N (Fig. 6). In fact, when the dynamics is N_p-driven, the network density mostly decreases as the network size N increases (Fig. 6, blue circle). So, a rise in N causes the density to be smaller when the engine of dynamics is changes in population. In contrast, when changes in κ play a dominant role, the network density may increase when the network size is sufficiently large (Fig. 6, pale-red cross). This is because when the number of active nodes N is close to its upper bound N_p, at which the activity levels of remaining inactive nodes are fairly low, the overall activity κ needs to be large enough for those low-activity nodes to get at least one edge. This would necessarily increase the total number of edges in the network to a large extent, which leads to a "true" densification of networks.

These properties are also confirmed by the analytical equation for the average network density[15]

$$\frac{2M}{N(N-1)} = \frac{\kappa}{4}\left(\frac{1}{1-q_0(\kappa,N_p)}\right)^2\left(1+\frac{q_0(\kappa,N_p)}{N-1}\right), \tag{10}$$

where q_0 denotes the fraction of isolated nodes in the system (see Eq. 16 in Methods "Analytical expression for N and M"). If the system is in Regime 1, in which κ is constant, the density monotonically approaches $\kappa/4$ as $N_p \to \infty$ (i.e., $q_0 \to 0$ and $N \to \infty$). On the other hand, if the system is in Regime 2, in which N_p is constant, there is no a *priori* upper bound, and the density exhibits a non-monotonic behaviour. In Regime 2, a change in κ has two opposing effects on the network density. First, an increase in κ directly increases density through a rise in the probability of edges being created. Second, a shift in κ would also increase N, which reduces the density through the third term in Eq. 10. Since $q_0 \to 0$ as N becomes sufficiently large, the latter effect is vanishing, and therefore the density begins to rise with N for a sufficiently large N.

Discussion

Densification and sparsification of networks can occur for two reasons, namely a variation in the population N_p and a variation in the overall activity level κ. A key finding of this work is that the relative importance of each of these two dynamical factors occasionally change, depending on the social context under study. By fitting the model to the observed scaling relations, we can detect the main factor that is relevant at a given point in time. Shifts in N_p and/or κ affect the activity of all individuals equally, so κ could be considered as an effective "temperature" and N_p as an effective "chemical potential" of the system in a grand canonical approach. While in this work we studied face-to-face networks of individuals, by its versatility the proposed method could also be used for a wide variety of dynamical systems.

There are some remaining issues for future research. First, the baseline model, a dynamic hidden variable model, relies on a "homogeneous mixing" hypothesis, which implies that nodes are connected to each other at random, given their activity levels. If we look at the structural properties of networks, such as triadic closure and community structure, such a hypothesis—especially for social contexts—would be unrealistic. However, the fact that the proposed method works remarkably well indicates that, as long as we look at network dynamics at a sufficiently coarse scale, keeping local properties aside, this homogeneous mixing assumption is a good approximation. In fact, introducing a non-random structure would easily make it impossible to obtain analytical expressions that would be needed for identification.

Second, we assumed that the distribution of intrinsic node activities is uniform for simplicity. Ideally, one would need to set this distribution based on empirical evidence. However, measuring the empirical intrinsic activity levels of individuals is extremely difficult because one needs activity levels of totally inactive individuals as well. Furthermore, this parameter might very well have its own temporal evolution. If available, such a rich information would allow for a refinement of the method.

Third, while the current method works well for temporal networks whose dynamical regime is occasionally switching, for fixed-regime systems in which the whole dynamics could be explained by either a N_p-driven *or* a κ-driven regime, the proposed regime-switching model is unnecessary. In such cases, one would fit the empirical scaling to each of the two models separately, and then find out which model is better fitted[15].

Fourth, the parameters N_p and κ could also be estimated using a numerical joint distribution of (N_t, M_t) that would be obtained by simulation for a given (N_p, κ), which is considered as a likelihood function. However, since a particular combination of (N_t, M_t) can be generated by different combinations of (N_p, κ), the observed combination of (N_t, M_t), as opposed to the whole scaling relationship, could be insufficient to accurately uncover the true values of $N_{p,t}$ and κ_t especially when network size is small[15]. On the other hand, there will be an advantage of the simulated likelihood approach in that we could introduce realistic network properties such as clustering, community structure, homophily, etc.

In many cases, examining the source of network dynamics from the level of each individual would be prohibitively difficult because each individual has his/her own circumstance, and privacy issues often prohibit researchers from obtaining enough information to reveal particular individuals' behaviour. In contrast, global quantities, such as the total numbers of nodes and edges, are much more widely accessible, and therefore utilising these quantities will be inevitable when high-resolution data are difficult to collect. A contribution of this work is that the proposed model allows us to detect the role of the two fundamental dynamical factors just by using information on the global network dynamics. Any dynamical processes occurring *on* networks, regardless of whether they are micro- or macro-phenomena, would be largely affected by the underlying dynamics *of* networks. This is in particular the case for spreading processes such as epidemics. A better understanding of the dynamics of densification and sparsification could thus benefit public health policies, which are of central importance for modern social systems.

Methods

Analytical expression for *N* and *M*. In this section we derive Eqs. (2) and (3). 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 brevity):

$$\begin{cases} N &= (1 - q_0(\kappa, N_p))N_p, \\ M &= \frac{\bar{k}(\kappa, N_p)N_p}{2}, \end{cases} \tag{11}$$

where $\bar{k}(\kappa, N_p)$ denotes the average degree over all the existing nodes including isolated ones, and $q_0(\kappa, N_p)$ denotes the fraction of isolated nodes or equivalently the probability that a randomly chosen node being isolated.

Let $\rho(a)$ be the density of node activities, and define $u(a, a')$ as the probability that there is an edge between two nodes having activity levels a and a'. The average degree $\bar{k}(\kappa, N_p)$ is given by the number of possible partners times the average of $u(a, a')$ (see, section S1 in SI for a full derivation):

$$\bar{k}(\kappa, N_p) = (N_p - 1) \int \int dada' \rho(a)\rho(a')u(a, a'), \tag{12}$$

It should be noted that Eq. (12) is equivalent to the average degree in the standard fitness model[27] if $N_p - 1$ is replaced with N, which is only asymptotically true in our model.

The fraction of isolated nodes in the system is given by (see, section S1 in SI):

$$q_0(\kappa, N_p) = \int da' \rho(a') \left[1 - \int u(a', a)\rho(a) da \right]^{N_p - 1}. \tag{13}$$

Substituting $\rho(a) = 1$ (i.e., uniform distribution on $[0, 1]$) and $u(a, a') = \kappa a a'$ into Eq. (12) leads to:

$$\bar{k}(\kappa, N_{\mathrm{p}}) = \frac{\kappa}{4}(N_{\mathrm{p}} - 1). \tag{14}$$

Similarly, q_0 is given by:

$$q_0(\kappa, N_{\mathrm{p}}) = \int_0^1 \left(1 - \frac{\kappa a'}{2}\right)^{N_{\mathrm{p}}-1} da'. \tag{15}$$

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

$$\begin{aligned} q_0(\kappa, N_{\mathrm{p}}) &= \frac{2}{\kappa}\int_{1-\frac{\kappa}{2}}^{1} x^{N_{\mathrm{p}}-1} dx \\ &= \frac{2}{\kappa N_{\mathrm{p}}}\left[1 - \left(1 - \frac{\kappa}{2}\right)^{N_{\mathrm{p}}}\right]. \end{aligned} \tag{16}$$

Combining these results with Eq. (11), we have:

$$N = N_{\mathrm{p}}\left[1 - \frac{2}{\kappa N_{\mathrm{p}}}\left(1 - \left(1 - \frac{\kappa}{2}\right)^{N_{\mathrm{p}}}\right)\right], \tag{17}$$

$$M = \frac{1}{8}\kappa N_{\mathrm{p}}(N_{\mathrm{p}} - 1). \tag{18}$$

It should be noted that if $|1 - \kappa/2| < 1$ and N_{p} is sufficiently large, then $q_0(\kappa, N_{\mathrm{p}}) \simeq 0$ and thereby $N \simeq N_{\mathrm{p}}$ and $M \propto N^2$, as is shown in the study of the static fitness model[26–28].

Bayesian estimation. This section describes how we can infer the model parameters and the dynamical regime at a given time interval t. Let $\Pr(S_t = s|\psi_{t-1}; \boldsymbol{\theta})$ be the probability that a network is in state s (i.e., in Regime s) conditional on information available at the end of time interval $t-1$, denoted by ψ_{t-1}, for a given set of constant parameters $\boldsymbol{\theta} = \{N_{\mathrm{p}}, \kappa, \sigma_1, \sigma_2, p_{11}, p_{22}\}$. More specifically, ψ_{t-1} is the set of variables observed from time 0 up to $t-1$, which represents the full history of the observed variables at the end of $t-1$. The likelihood function is then given by:

$$L(\{\boldsymbol{D}_t\}|\boldsymbol{\theta}) = \prod_{t=1}^{T}\sum_{s=1}^{2} f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}; \boldsymbol{\theta})\Pr(S_t = s|\psi_{t-1}; \boldsymbol{\theta}), \tag{19}$$

where $\{\boldsymbol{D}_t\}$ denotes the sequence of observations $\boldsymbol{D}_t = (N_t, M_t)$, and f is given by:

$$f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_s^2}}\exp\left(-\frac{(N_t - h^s)^2}{2\sigma_s^2}\right), \quad s = 1, 2. \tag{20}$$

The log-likelihood function leads to:

$$\begin{aligned} \log L(\{\boldsymbol{D}_t\}|\boldsymbol{\theta}) &= \sum_{t=1}^{T}\log\sum_{s=1}^{2} f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}; \boldsymbol{\theta})\Pr(S_t = s|\psi_{t-1}; \boldsymbol{\theta}), \\ &= \sum_{t=1}^{T}\log\sum_{s=1}^{2}\sum_{r=1}^{2} f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}; \boldsymbol{\theta})\Pr(S_{t-1} = r|\psi_{t-1}; \boldsymbol{\theta})p_{rs}. \end{aligned} \tag{21}$$

Bayesian inference is conducted based on the relationship $p(\boldsymbol{\theta}|\{\boldsymbol{D}_t\}) \propto L(\{\boldsymbol{D}_t\}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta}|\{\boldsymbol{D}_t\})$ and $p(\boldsymbol{\theta})$ are posterior and prior densities, respectively. For each parameter we collect 20,000 samples (four chains, 5,000 samples after 5,000 burn-in for each chain) generated from the posterior using Markov chain Monte Carlo (MCMC). We implement MCMC using Pystan ver. 2.19.0[32], which runs the No-U-Turn sampler (NUTS)[33]. The dataset and the Python code used in this work are available from Zenodo[34]. The mean parameter values are summarised in Table 1.

Now we describe how information is updated in each period. The probability of being in state s conditional on information at time t is written as:

$$\begin{aligned} \Pr(S_t = s|\psi_t; \boldsymbol{\theta}) &= \frac{f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}) \cdot \Pr(S_t = s|\psi_{t-1})}{\sum_s f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}) \cdot \Pr(S_t = s|\psi_{t-1})}, \\ &= \frac{\sum_r f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}) \cdot \Pr(S_{t-1} = r|\psi_{t-1})p_{rs}}{\sum_s\sum_r f(\boldsymbol{D}_t|S_t = s, \psi_{t-1}) \cdot \Pr(S_{t-1} = r|\psi_{t-1})p_{rs}}, \end{aligned} \tag{22}$$

where we drop argument $\boldsymbol{\theta}$ in f for brevity. Given the initial guess for $\Pr(S_0 = r|\psi_0)$, we can recursively update the probability of being in state s.

	WS-16	IC2S2-17	Hospital	Workplace	Prior distribution
N_p	121.152	196.236	28.087	109.694	Uniform(N_{max}, $2N_{max}$)
	[121.005, 121.555]	[196.005, 196.884]	[28.002, 28.306]	[106.449, 112.984]	
κ	0.467	0.222	0.495	0.087	Uniform(0, 1)
	[0.430, 0.512]	[0.215, 0.228]	[0.450, 0.541]	[0.075, 0.101]	
p_{11}	0.930	0.969	0.919	0.917	Beta(5, 1)
	[0.853, 0.980]	[0.914, 0.996]	[0.857, 0.965]	[0.836, 0.973]	
p_{22}	0.969	0.988	0.926	0.974	Beta(5, 1)
	[0.932, 0.992]	[0.966, 0.999]	[0.861, 0.971]	[0.944, 0.993]	
σ_1	8.587	5.358	1.617	5.815	Cauchy(0, 2)
	[4.186, 12.618]	[4.360, 6.598]	[1.395, 1.870]	[4.820, 7.064]	
σ_2	5.828	15.776	1.760	2.913	Cauchy(0, 2)
	[4.330, 7.520]	[14.096, 17.714]	[1.532, 2.007]	[2.602, 3.270]	

Table 1. Estimated parameters. For each parameter, mean and 95% credible interval obtained by MCMC are shown at the upper and lower rows, respectively. N_{max} denotes $\max_t\{N_t\}$.

Smoothed probability. The probability $\Pr(S_t = s|\psi_t;\boldsymbol{\theta})$ obtained in Eq. (22) is based on information available at time t for a given parameter set $\boldsymbol{\theta}$. We can also obtain the probability based on all information, represented by information set ψ_T. Let $\boldsymbol{\xi}_{t|T} \equiv [\Pr(S_t = 1|\psi_T;\boldsymbol{\theta}), \Pr(S_t = 2|\psi_T;\boldsymbol{\theta})]'$ be the vector of probabilities conditional on information at T. $\boldsymbol{\xi}_{t|T}$ can be calculated by conducting backward iteration from T[16]:

$$\begin{aligned}
\boldsymbol{\xi}_{T-1|T} &= \boldsymbol{\xi}_{T-1|T-1} \odot \{\boldsymbol{P}'[\boldsymbol{\xi}_{T|T}(\div)\boldsymbol{\xi}_{T|T-1}]\},\\
\boldsymbol{\xi}_{T-2|T} &= \boldsymbol{\xi}_{T-2|T-2} \odot \{\boldsymbol{P}'[\boldsymbol{\xi}_{T-1|T}(\div)\boldsymbol{\xi}_{T-1|T-2}]\},\\
&\vdots\\
\boldsymbol{\xi}_{t|T} &= \boldsymbol{\xi}_{t|T-1} \odot \{\boldsymbol{P}'[\boldsymbol{\xi}_{t+1|T}(\div)\boldsymbol{\xi}_{t+1|T-1}]\},
\end{aligned} \tag{23}$$

where $\odot$ and $(\div)$ denote element-by-element multiplication and element-by-element division, respectively, and $\boldsymbol{P} = (p_{ss})$ is the transition matrix. Note that all the terms in the RHS of the first equality are already known from the previous estimation procedure. After calculating $\boldsymbol{\xi}_{T-1|T}$, we use it to calculate the RHS of the second line. We repeat this until we obtain $\boldsymbol{\xi}_{t|T}$.

Validation. We check the accuracy of the inference method based on synthetic network data generated by the regime-switching hidden variable model. For given parameters N_p, κ, p_{11}, and p_{22}, and time-varying variables $\{N_{p,t}\}$ and $\{\kappa_t\}$, we generate sequences of $\{N_t\}$ and $\{M_t\}$ in a way prescribed in the model. When the network at t is in Regime 1 (Regime 2), the true $N_{p,t}$ (κ_t) is given by $N_{p,t} = 0.95N_{p,t-1}$ ($\kappa_t = 0.95\kappa_{t-1}$), and $N_{p,t} = N_p$ ($\kappa_t = \kappa$) otherwise. We assume that the initial probability of being in Regime 1 is set at 0.5, and $N_{p,0} = N_p$ and $\kappa_0 = \kappa$. For each parameter, we collect 20,000 samples by MCMC (5,000 samples from four chains after 5,000 burn-in iterations).

The estimated parameters under different sets of ground-truth $\boldsymbol{\theta}$ are summarised in Table S1 in SI. The estimated smoothed probabilities well match the true states of the generated networks (Fig. S1 in SI). We also group the generated networks based on the probability of being in Regime 1; For each time period, if more than 95% of the sampled values for $\Pr(S_t = 1|\psi_t;\widehat{\boldsymbol{\theta}})$ are higher (lower) than 0.5, then we classify the corresponding snapshot as being in Regime 1 (Regime 2). If it is not classified as Regime 1 or 2, the network is considered to be in a "gray area". As shown in the middle and the right columns of Fig. S1, the classification of generated networks based on estimated parameters is consistent with the ground truth, while there are some networks that are in gray zones especially when the observed pairs of (N_t, M_t) are overlapped between the two regimes. A comparison between the estimated and the true paths of $N_{p,t}$ and κ_t is also shown in Fig. S2.

Received: 18 November 2020; Accepted: 20 January 2021
Published online: 04 February 2021

References

1. Newman, M. E. J. *Networks – An Introduction* (Oxford University Press, Oxford, 2010).
2. Barabási, A. L. The network takeover. *Nat. Phys.* **8**, 14–16 (2012).
3. Barabási, A.-L. *Network Science* (Cambridge University Press, Cambridge, 2016).
4. Holme, P. & Saramäki, J. Temporal networks. *Phys. Reports* **519**, 97–125 (2012).
5. Holme, P. Modern temporal network theory: a colloquium. *Eur. Phys. J. B* **88**, 234 (2015).
6. Jo, H. H., Pan, R. K. & Kaski, K. Emergence of bursts and communities in evolving weighted networks. *PLOS ONE* **6**, e22687 (2011).
7. Karsai, M. *et al.* Small but slow world: How network topology and burstiness slow down spreading. *Phys. Rev. E* **83**, 025102(R) (2011).
8. Karsai, M., Kaski, K., Barabási, A. L. & Kertész, J. Universal features of correlated bursty behaviour. *Sci. Rep.* **2**, 397 (2012).

9. Jo, H. H., Karsai, M., Kertész, J. & Kaski, K. Circadian pattern and burstiness in mobile phone communication. *New J. Phys.* **14**, 013055 (2012).
10. Gauvin, L., Panisson, A. & Cattuto, C. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PLOS ONE* **9**, e86028 (2014).
11. Scholtes, I., Wider, N. & Garas, A. Higher-order aggregate networks in the analysis of temporal networks: path structures and centralities. *Eur. Phys. J. B* **89**, 1–15 (2016).
12. Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nat. Phys.* **15**, 313–320 (2019).
13. Gautreau, A., Barrat, A. & Barthélemy, M. Microdynamics in stationary complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8847–8852 (2009).
14. Leskovec, J., Kleinberg, J. & Faloutsos, C. (2007) Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data (TKDD)* 1:2.
15. Kobayashi, T. & Génois, M. Two types of densification scaling in the evolution of temporal networks. *Phys. Rev. E* **102**, 052302 (2020).
16. Hamilton, J. *Time Series Analysis* (Princeton University Press, Princeton, NJ, 1994).
17. Hamilton, J. D. Regime switching models. In *Macroeconometrics and Time Series Analysis* 202–209 (Springer, 2010).
18. http://www.sociopatterns.org/.
19. Génois, M., Zens, M., Lechner, C., Rammstedt, B. & Strohmaier, M. Building connections: How scientists meet each other during a conference. *arXiv:1901.01182* (2019).
20. Vanhems, P. *et al.* Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLOS ONE* **8**, e73970 (2013).
21. Génois, M. *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**, 326–347 (2015).
22. Cattuto, C. *et al.* Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLOS ONE* **5**, 1–9 (2010).
23. Leskovec, J., Kleinberg, J. & Faloutsos, C. 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* 177–187 (ACM, 2005).
24. Bettencourt, L. M., Kaiser, D. I. & Kaur, J. Scientific discovery and topological transitions in collaboration networks. *J. Informetrics* **3**, 210–221 (2009).
25. Kobayashi, T. & Takaguchi, T. Social dynamics of financial networks. *EPJ Data Sci.* **7**, 15 (2018).
26. Caldarelli, G., Capocci, A., De Los Rios, P. & Muñoz, M. A. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89**, 258702 (2002).
27. Boguñá, M. & Pastor-Satorras, R. Class of correlated random networks with hidden variables. *Phys. Rev. E* **68**, 036112 (2003).
28. De Masi, G., Iori, G. & Caldarelli, G. Fitness model for the Italian interbank money market. *Phys. Rev. E* **74**, 066112 (2006).
29. Aledavood, T., Lehmann, S. & Saramäki, J. Digital daily cycles of individuals. *Front. Phys.* **3**, 73 (2015).
30. Kobayashi, T., Takaguchi, T. & Barrat, A. The structured backbone of temporal social ties. *Nat. Commun.* **10**, 220 (2019).
31. Kim, C.-J. Dynamic linear models with Markov-switching. *J. Econom.* **60**, 1–22 (1994).
32. https://mc-stan.org/users/interfaces/pystan.html.
33. Hoffman, M. D. & Gelman, A. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014).
34. https://doi.org/10.5281/zenodo.4267207.

Acknowledgements
T.K. acknowledges financial support from JSPS KAKENHI Grant Nos. 15H05729 and 19H01506. This work was partially supported by the ANR Project DATAREDUX (ANR-19-CE46-0008) to M.G.

Author contributions
T.K. conceived the research. T.K. and M.G. defined the model. T.K. performed analytical calculation and data analysis. T.K. and M.G. discussed the results and wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material availlable at https://doi.org/10.1038/s41598-021-82432-9.

Correspondence and requests for materials should be addressed to T.K. or M.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021, corrected publication 2021