COMPLESSITÀ NEI SISTEMI SOCIALI

LORENZO DALL'AMICO
ISI Foundation

Note per il corso in Fisica dei sistemi complessi

Università di Torino
March 26, 2024

Lorenzo Dall'Amico: *Complessità nei sistemi sociali*

CONTENTS

ACRONYMS

F2F *Face-to-face*

ER *Erdős-Rényi*

NMF *naïve mean field*

BP *belief propagation*

SIR *Susceptible-Infected-Recovered model*

GW *Galton Watson tree*

GFT *Graph Fourier transform*

CD *Community detection*

AMI *Adjusted mutual information*

DCSBM *Degree corrected stochastic block model*

SC *Spectral clustering*

SYMBOLS

- $\delta_{a,b}$ is the Kroeneker delta equal to 1 if $a = b$ and equal to 0 otherwise.
- $\Lambda(M)$ is the set of eigenvalues of a matrix M. $\lambda_i(M)$ is the i-th (smallest or largest, according to the context) eigenvalue of M
- The spectral radius of M (largest eigenvalue) is denoted with $\rho(M)$.
- With the notation $\mathbf{1}_n$ we denote the all-ones vector of size n.
- The entry-wise Hadamard product is denoted with $\circ$.
- The set of the first n integers is denoted with $[n]$.
- We adopt the Landau notation for the asymptotic behavior of variables. In particular $x = O_n(y)$ is equivalent to $\lim_{n\to\infty} \frac{x}{y} = c$ for some finite c. The notation $x = o_n(y)$ instead means $\lim_{n\to\infty} \frac{x}{y} = 0$.
- The set of neighbors of a node i on a graph is $\partial i = \{j \in \mathcal{V} \,:\, A_{ij} = 1\}$

1

TEMPORAL GRAPHS

1.1 WHY TEMPORAL GRAPHS?

Graphs are an essential mathematical tool to represent interacting systems. When using static graphs, the interactions between two nodes is often represented as a single Boolean variable determining whether or not those two nodes interacted with one another. Yet, we can think of many examples in which this variable should depend on time. Think of a graph in which an interaction between two people is a phone call. In most instants, for most people, no interactions are recorded at all, and, in all cases at most one interaction per time may occur. So, how would we determine the Boolean interaction variable? One approach would be to aggregate time as determine that two people interacted if they had enough phone calls during a specific time window. In this way, connections are created between people that frequently call each other, but we loose an important piece of information: the order of events. Imagine an event like the fire of Notre Dame de Paris: in few moments people witnessing the fire spread the news to their contacts who, themselves reported to others in chain. If we had no idea of what happened and what is the content of the calls or messages, we could actually retrieve the geographical location from which the burst of information was initiated by retracing backwards the chain of events. If instead we use a static representation of the graph as we did earlier, all this information would be lost. A temporal graph is then an object capable of representing the interactions between the elements of a system together with a time stamp. Besides communication graphs, other notable examples are face-to-face proximity graphs, biological graphs, ecological graphs and many others.

A pictorial representation of a chain of calls

As we will show in the remainder, in some cases it is necessary to keep the temporal dimension into account if one wants to understand the process happening on top of a graph. This is due to the fact that links may have

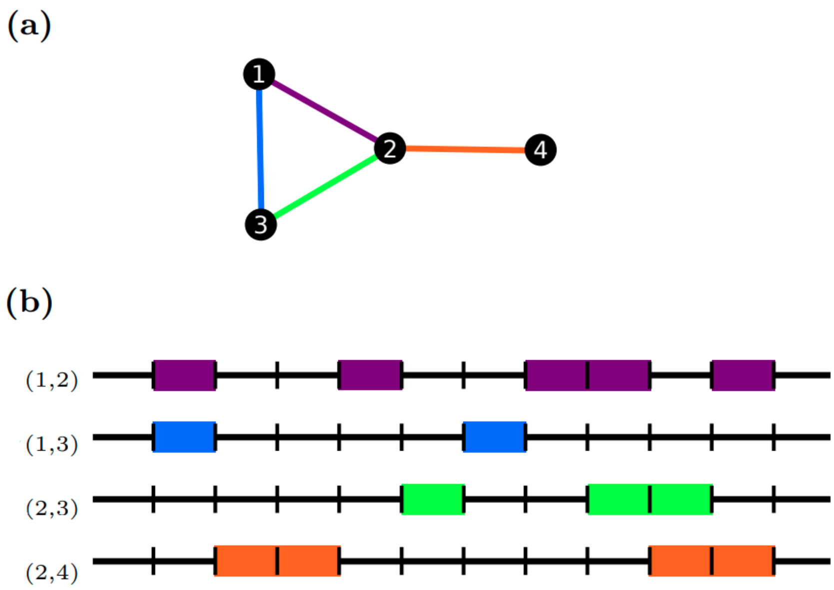


Figure 1.1: **A pictorial representation of a temporal graph**. (a) A graph with 4 nodes with $\mathcal{V} = \{1,2,3,4\}$ and $\mathcal{E} = \{(1,2),(1,3),(2,3),(2,4)\}$. (b) the temporal activation patterns of each edge (with color code). The x axis represents time. Picture taken from *Gauvin et al., Randomized reference models for temporal graphs.*

a causal relation (like the phone calls in the Notre Dame example) or simply because the aggregated static graph may be non representative of the interactions at any time stamp. We will then discuss how to mathematically define and represent temporal graphs, a relevant example of how to measure them and show some peculiar properties of temporal graphs.

1.2 REPRESENTING TEMPORAL GRAPHS

Let us consider a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$ connecting the nodes. Given a suited definition of interaction for our problem and a time window, $\mathcal{E}$ is the set of all pairs of nodes (i,j) that interacted at least once in the considered time window. We now want to add the notion of *when* these interactions occurred. Each edge (i,j) can appear multiple times and for each interaction we can consider a time t at which the interaction begun and a time duration τ.[1] We can then represent a temporal graph as a sequence of *temporal edges* in the form (i,j,t,τ). Figure 1.1 gives a pictorial representation of the temporal edges of a graph with 6 nodes. We now provide a more formal definition of a temporal graph.

1 We could equivalently replace τ with the time t_e at which the interaction ended.

Temporal graphs

A temporal graph is a tuple $\mathcal{G}(\mathcal{V}, \mathcal{E}_t)$, where $\mathcal{V}$ denotes the set of n nodes and $\mathcal{E}_t$ of temporal edges. Each $e \in \mathcal{E}_t$ can be written as (i, j, t, τ) where $i, j \in \mathcal{V}$, t is a time-stamp and $\tau \in \mathbb{R}^+$ is a positive interaction duration, implying that the link between i and j was active from t to $t + \tau$. If a node has at least one connection at time t we say it is *active* at time t and it is inactive otherwise.

Temporal graphs

One can generalize the concept of adjacency matrix to the temporal setting by letting

Temporal adjacency matrix

$$\tilde{A}_{ij}^{(t)} = \begin{cases} 1 & \text{if } \exists\, e = (i, j, t_0, \tau) \in \mathcal{E}_t \text{ s.t. } t \in [t_0, t_0 + \tau] \\ 0 & \text{else.} \end{cases}$$

In this representation, however, time is kept as a continuous variable and, even for a finite observation time, we obtain an infinite number of adjacency matrices. For this reason, the snapshot representation – that uses a discrete notion of time – may be more suited. In particular, we assume that the interaction duration τ is a multiple of a unit Δt that sets the temporal resolution of the graph. Let us define the concept of snapshot graphs, pictorially visualized in Figure 1.2.

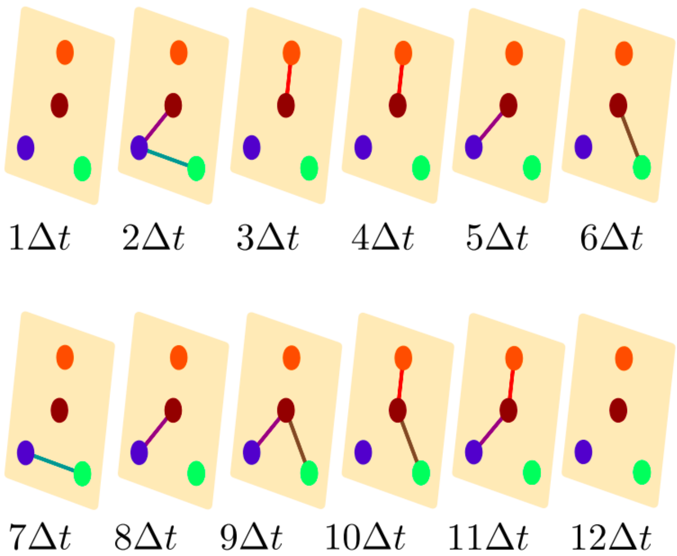


Figure 1.2: **A pictorial representation of a snapshot graph**. This plot represents the same graph of Figure 1.1 in which time was already discretized for convenience. Each *slice* corresponds to a different time step in which the edges progressively are activated. Picture taken from *Gauvin et al., Randomized reference models for temporal graphs.*

Snapshot graphs

Snapshot graphs

A snapshot graph is a tuple $\mathcal{G}(\mathcal{V}, \mathcal{E}_t)$, where $\mathcal{V}$ denotes the set of n nodes and $\mathcal{E}_t$ of temporal edges. Each $e \in \mathcal{E}_t$ can be written as (i, j, t) where $i, j \in \mathcal{V}$, t is a *discrete* time-stamp and models an instantaneous interaction between i, j at time t.

With this representation at hand, one can obtain the adjacency matrix representation of a temporal graph as follows

The snapshot adjacency matrix

$$A_{ij}^{(t)} = \begin{cases} 1 & \text{if } (i, j, t) \in \mathcal{E}_t \\ 0 & \text{else.} \end{cases}$$

In this way we obtain a sequence of T adjacency matrices $\{A^{(t)}\}_{t \in 1, \ldots, T}$, where T is the total number of snapshots. A natural question that poses is how which of the two representations is more appropriate and how much information is lost when choosing to discretize time. We address this point in the following remark.

Discretizing time

Discretizing time

If we want to choose a discrete representation of time, the natural questions that arise are how to choose Δt, the minimal interaction duration and how much do we loose in performing this simplification. The first point we want to raise is that any measured quantity (including time) is not truly continuous and it is bounded by a resolution that makes time discrete by design. So, in all cases, one can say Δt is the measurement instrument resolution and the snapshot representation is always appropriate. Yet, if Δt is much smaller than the total observation time, the number of time frames T – even if finite – will tend to be very large, hence untractable. So, setting a longer Δt might be more appropriate in some cases and the choice of a good Δt is necessarily problem-dependent because Δt determines the scale at which we consider interactions to be *simultaneous*.

The coupling between the process and graph dynamics

Let us make two examples to make this point clearer. Suppose we have two spreading phenomena: in one case the propagation of a piece of information (such as the fire of Notre Dame) and in the other a flu-like illness transmission. In the former case, the propagation of the information from one person to the other moves very fast and so Δt must be small, in the order of seconds/minutes to capture the rapid dynamics of the news propagation. If we consider the flu-like propagation, instead, we know that a person, after being infected, is not typically able to infect someone for a couple of days, hence we can set Δt of the order of one day, assuming that one cannot change its own infectious state in the course of a day. So, summarizing, the proper time aggregation depends on the time-scale of the dynamic process happening on top of the graph. If this process if much slower than the temporal evolution then we can simply aggregate the graph. If instead the two time scales (of the process and of the graph evolution) are similar, then we have an interesting coupling that must be taken into account.

To conclude this remark, there is still a quantity we want to preserve when we aggregate time, that is the cumulative interaction duration: what do we do with all interactions so that $\tau < \Delta t$? Also in this case the answer depends on the problem under consideration. Take for instance the flu propagation with a time aggregation of 24 hours. If an infectious individual has a interaction with a susceptible one, the interaction duration is key to determine the probability of infection: an hour-long interaction is much more likely to propagate the disease than a minute-long interaction and all interactions will be shorter than Δt in this case. To preserve this piece of information, we might want to associate a weight to each edge, representing the cumulative interaction duration. We relate the continuous time and snapshot adjacency matrices as follows

$$W_{ij}^{(t)} = \int_{t}^{t+\tau_0} dt' \, \tilde{A}_{ij}^{(t')}.$$

The weighted aggregated graph

If this representation may seem very reasonable, it must be noted that is not the only admissible one: in the case of the propagation of sexual diseases, for instance, the interaction duration is not relevant and may simply want to keep a Boolean representation of the edges, without attributing any weight.

Now that we have introduced some of the main concepts related to temporal graphs, let us consider the specific set of proximity graphs as a case study, first describing a method to measure these graphs and then using the open source, real data to describe some relevant properties.

1.3 MEASURING PROXIMITY GRAPHS

In proximity graphs the edges represent a *Face-to-face* (F2F) contact between two persons. The interest of this type of graphs resides in the fact that F2F interactions are the vehicle of human communication and of infectious diseases and, more generally, they quantify how humans interact with one another. Measuring *Face-to-face* (F2F) proximity graphs is, however, a very challenging task. Among the most used approaches to quantify them we have the use of questionnaires in which the interactions one has are self-reported. It was shown that this method is biased towards long interactions – in the sense that short interactions tend to be forgotten – and can achieve a low temporal resolution. A very important contribution to the field of measur-

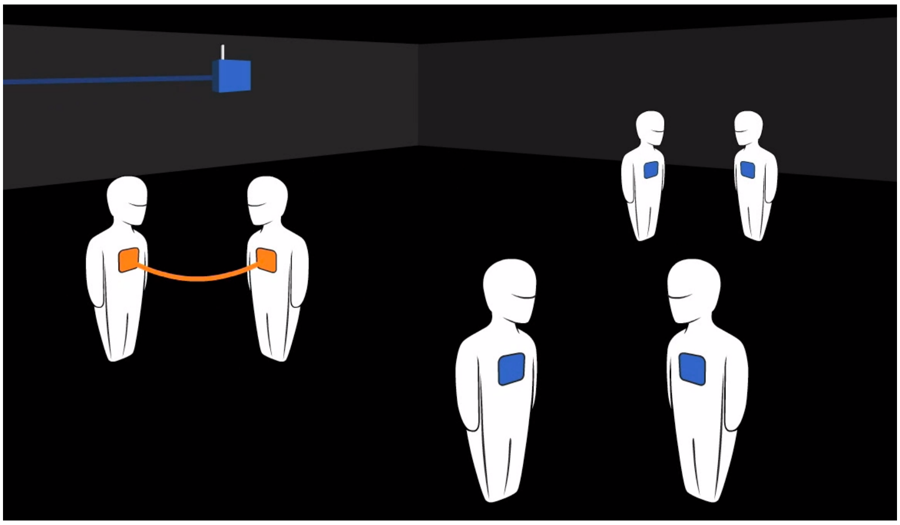
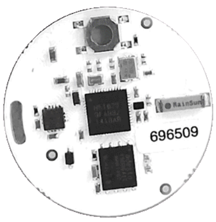


Figure 1.3: **Pictorial example of the use of SocioPatterns proximity sensors**. Six people in a room wearing a proximity sensor. The orange ones (with the line), indicate a recorded F2F interactions between two individuals. Picture taken from http://www.sociopatterns.org/.

ing F2F proximity graphs was made with the creation of the SocioPatterns collaboration by the ISI Foundation.

The SocioPatterns collaboration

SocioPatterns was formed in 2008 and developed wearable proximity sensors that are capable to *measure* temporal proximity graphs. Their functioning is based on the transmission and exchange of information packets using radio-frequency electromagnetic waves. Briefly speaking, each device is associated with a code and continuously switches from a *listener* to a *speaker* mode. When it is in the *speaker* mode it emits an information packet containing its own code and the power at which the signal was emitted. When it is in listener mode, instead, the device intercepts the packets emitted by the "speakers" and records on its memory the code and the power declared, the time stamp at which this interaction occurs as well as the power of the received signal. The devices have to be worn on the chest of the participants and, by design, record F2F proximity. Figure 1.3 shows a demo of the functioning of the SocioPatterns proximity sensors.



A proximity sensor

Inside the memory of each sensor we then have a list of entries of the type $(j, \text{pow}_{\text{tr}}, \text{pow}_{\text{rec}}, t)$, where j is the code of the sensor that emitted the signal, pow_{tr} is the transmission power, pow_{rec} is the received power and t is the time-stamp that has a temporal resolution of 20 seconds. Looking at the difference $\text{pow}_{\text{tr}} - \text{pow}_{\text{rec}}$ one can measure the attenuation of the signal and filter out the interactions that are too attenuated, thus keeping only those that happened at a distance within approximately 2 meters. From this we can create a snapshot graph with $\Delta t = 20\ s$ as described above.

The SocioPatterns sensors have been used in several contexts, including schools, hospitals, offices and rural African villages among others. They constitute a well known benchmark of temporal graph measurement that has

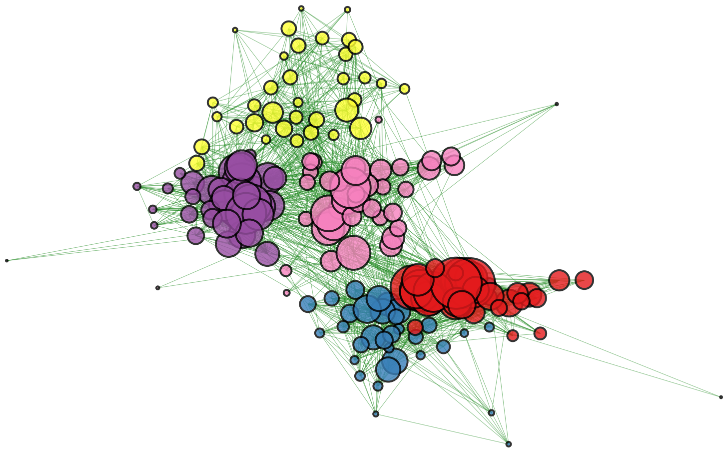


Figure 1.4: **The *School* dataset**. Pictorial representation of the *School* dataset (see Table 1.1) aggregated over all the observation time. The size of each node is the determined by the total interaction time of that node, while the color is determined by the class the student belongs to.

been used in many applications and many of the collected datasets are publicly available at http://www.sociopatterns.org/datasets/. We will make use of some of these data to study some relevant properties of temporal graphs. Table 1.1 summarizes some descriptive properties of the considered graphs, while Figure 1.4 shows the aggregated graph collected in a high school.

Name	***n***	**Observation time**	**Description**
School	180	from a Monday to the Tuesday of the following week in November 2012.	interactions between students in a high school in Marseilles, France belonging to 5 classes.
Office	92	June 24 to July 3, 2013	interactions between individuals measured in an office building in France
Village	86	between 16th December 2019 and 10th January 2020	interactions between the people of Mdoliro village in Dowa district in the Central Region of Malawi.
Conference	405	June 4-5, 2009	interactions at the SFHH conference in Nice

Table 1.1: **Summary statistics of the *SocioPatterns* temporal networks**. The first column indicates the name used in these notes. The column indexed by n indicates the number of nodes appearing in the graph. The column *Observation time* describes the experiment duration, while *Description* provides a few details on the context of the data collection. For more information, refer to the SocioPatterns website.

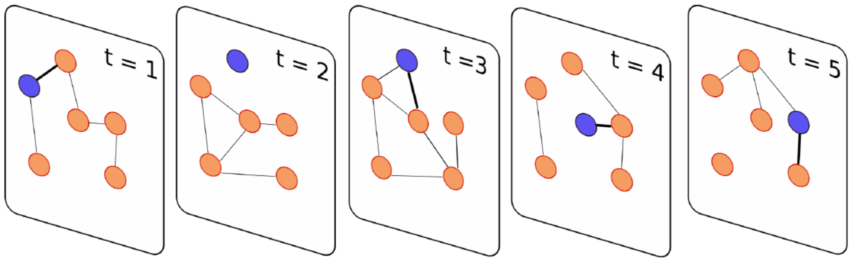


Figure 1.5: **Time respecting paths**. Five snapshots of a temporal graph in which unoccupied nodes are depicted in orange at each time, while the currently occupied node is in blue. A larger width is used to highlight the edge that causes the transition.

1.4 PROPERTIES OF TEMPORAL GRAPHS

We now proceed to describe some important concepts that characterize temporal graphs and use the four aforementioned datasets to show them on empirical data.

TIME-RESPECTING PATHS

When we consider a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, we define a path on it as an ordered sequence of nodes $\{i_1, i_2, \ldots, i_T\}$ for that, for all $p \in [T]$, $i_p \in \mathcal{V}$ and for all $p \in [T-1]$, $(i_p, i_{p+1}) \in \mathcal{E}$. In words, every step of a path allows one to only move from a node to one of its neighbors. When we consider a temporal graph, instead, we must generalize the concept of path, to encode the role played by time, introducing the *time-respecting paths*.

Time respecting paths

Time respecting paths

Given a temporal graph $\mathcal{G}(\mathcal{V},\mathcal{E}_t)$, we denote a time respecting path as $\{i_1(t_1), i_2(t_2), \ldots, i_T(t_T)\}$ if it satisfies the following conditions

- For all $p \in [T]$, $i_p \in \mathcal{V}$: the path is defined on the graph nodes.
- For all $p \in [T-1]$, $t_p < t_{p+1}$: these two times indicate the beginning of the residency on the respective nodes and time must be increasing.
- For all $p \in [T-1], \exists\, t_0, \tau$ s.t. $(i_p, i_{p+1}, t_0, \tau) \in \mathcal{E}_t$ and $t_{p+1} \in [t_0, t_0+\tau]$: the transition between one node and the other can only take place at a time at which the two nodes are connected.

This definition is given for a continuous time representation but it can simply be adapted to the discrete one. For this case, we give a simple representation of a time respecting path in Figure 1.5.

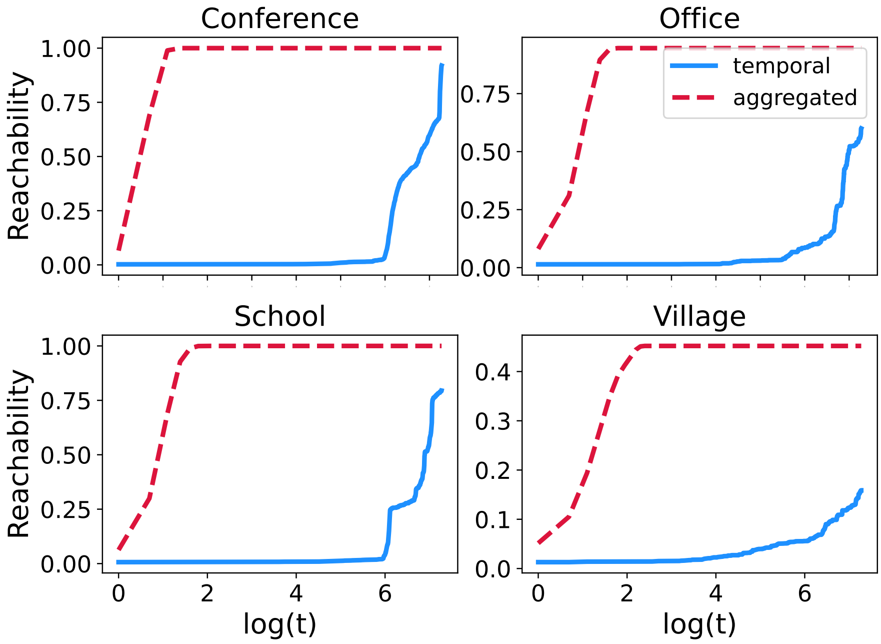


Figure 1.6: **Time-respecting vs aggregate reachability**. Each plot refers to one of the 4 *SocioPatterns* datasets described in Table 1.1, considering the first 8 hours of measurements. The red dashed line is the average of the reachability matrix R_t defined in Equation (1.1) as function of time, using for all t A_t the weighted aggregated matrix over all the observation period. The blue continuous line, instead, is obtained from the snapshot adjacency matrices and encodes time-respecting paths.

Given a sequence of adjacency matrices, we now define the *reachability matrix* R_t as follows

Reachability matrix

$$R_t = \text{sign}\left[\prod_{t'=1}^{t}(A_{t'} + I_n)\right], \tag{1.1}$$

where I_n is the identity matrix, the sign function has to be considered entry-wise, while the product has to be taken from right to left, *i.e.* $\prod_{t=1}^{3} A_t = A_3A_2A_1$. The entry $R_{t,ij}$ equals 1 if there exists a time-respecting path of length smaller or equal to t that allows one to go from i to j.

Time-respecting paths are not symmetric

An important fact related to this matrix is that it is not necessarily symmetric. This comes from the fact that the product of matrices (such as the A_t's) is symmetric only if the matrices commute. This is not the case in general and it implies that if there is a time-respecting path from i to j, that does not imply that there exists also a time-respecting path from j to i.

By taking the average of the reachability matrix, we also have a measure of how well its nodes are connected. Figure 1.6 compares the reachability on the 4 real temporal graphs described above with the one obtained on their aggregated version and clearly shows that temporal graphs have a lower reachability. This is because the valid time-respecting paths are constrained

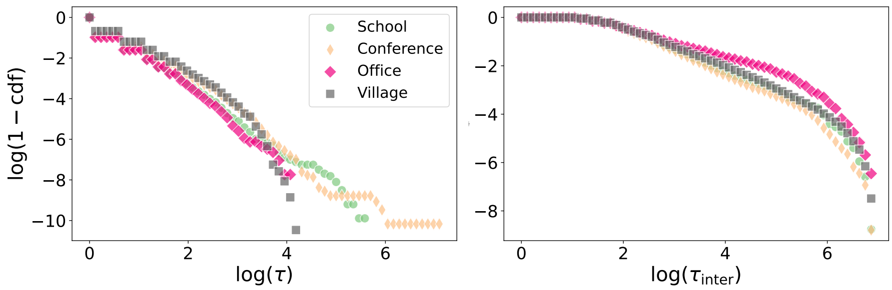


Figure 1.7: **Event and inter-event duration distributions**. The figures show the scatter plot in log-log scale of the interaction duration (left) and inter-event duration (right) distributions vs the $1 - cdf$, *i.e.* the complementary of the cumulative density function. Each line refers to one of the datasets described in Table 1.1 and is color and marker coded.

and are only a subset of all the possible paths that one can perform on the aggregate version of the graph. This is an important ingredient to consider when coupling a dynamic process with the graph, because only a fraction of all possible paths can actually take place.

DURATION DISTRIBUTION AND BURSTINESS

We now focus on a very peculiar aspect of contact graphs, that is the contact duration distribution. It has been observed in many instances (with no apparent exception) that this distribution is very broad and follows approximately a power law decay that appears to be a universal behavior. Figure 1.7 (left plot) shows in log-log scale 1 minus the cumulative density function vs the interaction duration and confirms this trend, since the relation is approximately linear in the logarithmic scale. This is an important observation, because it tells us that very long interactions are much more common than what we would expect for a thin tail distribution, such as the Poisson. The consequence is that, if we have a process that needs a minimal time of interaction to consider the interaction to be valid, then, in practice, even if the threshold is very large, there will be valid interaction edges with high probability. On the other hand, we also know that most of the distribution is concentrated around small values.

$1 - \mathrm{cdf}(x) = \mathbb{P}(\tau \geq x)$

A similar behavior is observed for the inter-event duration distribution. We define the inter-event duration as the time elapsed between two successive interactions of the same pair of nodes (ij). Since this distribution is broad, we say that the interaction dynamics is *bursty*, *i.e.* that typically we have an alternation of time intervals in which the activity is very low and some in which it is very high. To best understand the effect that a bursty dynamics may have on a process, let us consider the following example.

Suppose we have a quantity $\mathcal{Q}$ that is increased by one unit every time there is a interaction and it is decreased by a factor α for each time step in which no interaction occurs. If $\mathcal{Q}$ exceeds a threshold value $\mathcal{Q}_{\text{th}}$, then some

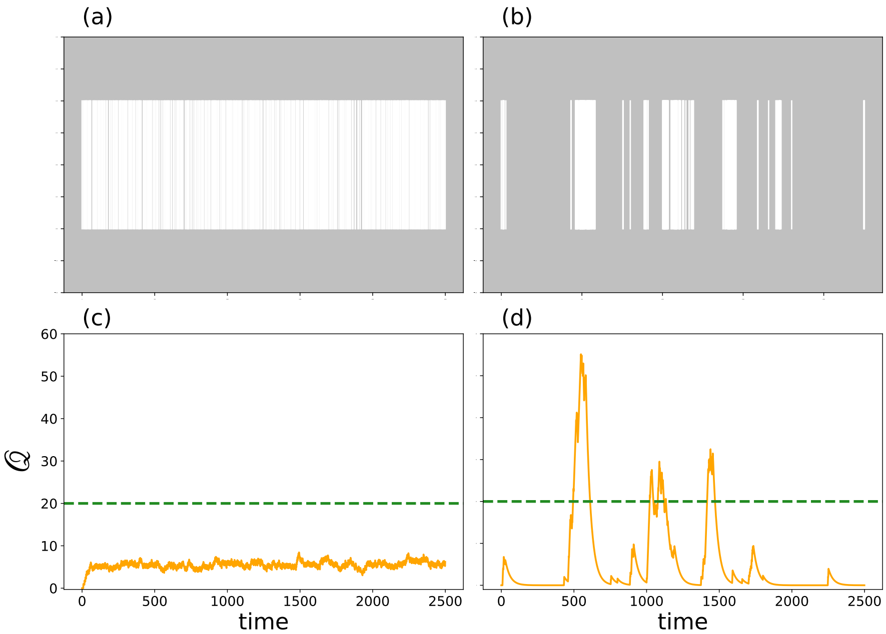


Figure 1.8: **The effect of bursty dynamics**. The first row represents a temporal time series for (a) a Poisson process and (b) the interaction times of a node of the *Conference* graph (see Table 1.1. Each horizontal line indicates an active time. The second line displays the dynamics of a quantity $\mathcal{Q}$ evolving according to the process described in the main text for (c) the Poisson dynamics of (a), and for (d) the bursty dynamics of (b). The horizontal dashed line indicates an arbitrary selected threshold value that triggers some process when $\mathcal{Q} > \mathcal{Q}_{\text{th}}$. Image adapted from *Holme, Saramaki, Temporal networks.*

process is triggered otherwise it is not. In Figure 1.8 we compare this dynamical process on a Poisson temporal series made of 362 interaction events with one extracted from an individual activity pattern of the *Conference* graph. The bottom plots clearly evidence that the bursty dynamics, being highly concentrated in some time regions, allows one to go beyond the threshold several times, while this does not happen to the Poisson distribution.

A method to measure the burstiness level of a time series is as follows

$$B = \frac{s - m}{s + m},$$

where s and m are the standard deviation and mean of the inter-event duration distribution, respectively. For a periodic process, $s = 0$ and $B = -1$, while for the burstiest of processes, $s \to \infty$ and $B \to 1$. In our case, the Poisson dynamics of Figure 1.8(a) has $B = -0.45$, while the one of Figure 1.8(b) has $B = 0.70$.

1.5 CONCLUSION

Temporal networks are a powerful tool to model complex dynamical systems. Empirical networks often show very broad distribution of the interaction duration as well as bursty dynamics. These features are of great importance to some dynamical processes that may unfold on networks and the temporal framework is a relevant generalization of static graphs. However, the dynamic component of graph evolution must always be compared with the typical time scale of the process unfolding over the graph in order to understand whether it is necessary to have an additional layer of complexity given by time or, more in general, to choose an appropriate time discretization to perform the analysis.

1.6 REFERENCES

- P. Holme, J. Saramäki, *Temporal networks*. Physics reports, 519(3), 97-125 2012.
  This is *the* reference for an introduction to temporal graphs.