

Non-Altering Time Scales for Aggregation of Dynamic Networks into Series of Graphs*

YANNICK LÉO,

Univ Lyon, ENS de Lyon, CNRS, Inria, UCB Lyon 1, LIP UMR 5668,
15 parvis René Descartes, F-69342, Lyon, FRANCE
`yannick.leo@inria.fr`

CHRISTOPHE CRESPELLE

Univ Lyon, UCB Lyon 1, CNRS, ENS de Lyon, Inria, LIP UMR 5668,
15 parvis René Descartes, F-69342, Lyon, FRANCE and
Institute of Mathematics, Vietnam Academy of Science and Technology,
18 Hoang Quoc Viet, Hanoi, Vietnam
`christophe.crespelle@inria.fr`

AND

ERIC FLEURY

Univ Lyon, ENS de Lyon, Inria, CNRS, UCB Lyon 1, LIP UMR 5668,
15 parvis René Descartes, F-69342, Lyon, FRANCE
`eric.fleury@inria.fr`

Abstract

Many dynamic networks coming from real-world contexts are *link streams*, i.e. a finite collection of triplets (u, v, t) where u and v are two nodes having a link between them at time t . A very large number of studies on these objects start by aggregating the data in disjoint time windows of length Δ in order to obtain a series of graphs on which are made all subsequent analyses. Here we are concerned with the impact of the chosen Δ on the obtained graph series. We address the fundamental question of knowing whether a series of graphs formed using a given Δ faithfully describes the original link stream. We answer the question by showing that such dynamic networks exhibit a threshold for Δ , which we call the *saturation scale*, beyond which the properties of propagation of the link stream are altered, while they are mostly preserved before. We design an automatic method to determine the saturation scale of any link stream, which we apply and validate on several real-world datasets.

*This is a complete version of the extended abstract appeared in CoNEXT 2015 [22].

1 Introduction

Many real world dynamic networks are naturally given in the form of a finite collection \mathcal{L} of triplets (u, v, t) , which we call a *link stream*, where $u, v \in V$ are two nodes of the network and t is a timestamp¹, with the meaning that nodes u and v have a link between them at time t . Depending on the context², these links can represent physical contacts between individuals, exchanges of emails between people, commercial interactions between companies, etc. When one wants to study such dynamic networks, a very common approach [11, 19, 25, 15, 6, 42, 9, 34, 30, 46, 44, 38, 43, 33, 4, 24, 13, 2, 23] is to transform them into series of graphs. The process used to do so is called *aggregation*. It consists in choosing a time window $[a, b] \subseteq [0, T]$ in the initial series, where T is the length of the period of study, and forming the graph $G_{[a, b]}$ with all edges u, v such that there exists a triplet $(u, v, t) \in \mathcal{L}$ with $t \in [a, b]$. Doing so for a collection of windows that covers the entire period of study, one obtains a representation of the dynamic network as a graph series: the graphs formed for each window, called *snapshots*. Very often, as in this paper, the windows are disjoint and all have the same length, but in some studies, they may also overlap [20, 1, 29, 40, 5, 37] or have different lengths [35, 39] or all start at the beginning of the period of study [21, 31, 14, 37]. In all cases, once the series is obtained, all analyses are conducted on it instead of the raw data, which is in the form of a link stream, referred to as the *original link stream* in the rest of the paper, see Figure 1.

There are two main reasons to use aggregation for studying dynamic networks. First, in many cases, it does not make sense to study the network at the scale of the time resolution of the timestamps of the given link stream. For example, in an email dataset, the timestamps of the events (sending of email) are often given with a 1-second resolution. However, studying the dynamic network at this time scale does not give a general and comprehensive view of its organization, like someone watching a painting with a microscope. Hence, aggregation allows to study the network at a scale which is relevant compared to its activity. The second reason for using aggregation is that it produces graphs. They give an instantaneous view of the network (snapshot) which is practical in itself to get a view of what the object under study looks like and one can use the rich set of notions developed in graph theory to analyze the considered dynamic network.

If the benefits of aggregation are clear, on the other hand, it also raises some important concerns. Indeed, the length chosen for the aggregation window usually has a strong impact on the properties of the aggregated graph series [18, 36, 37]. This raises the question of which time scale should be chosen to study a given dynamic network and how much the properties studied, based on which conclusions are derived, are sensitive to the length of the aggregation period used [34, 30, 45]. It points out that in any case, this period should not be chosen without well established evidence, as it is currently done in most of the studies cited above. Pushing further, it is not even clear whether an aggregated series faithfully describes the original link stream. Indeed, the

¹Time can be continuous or discrete. The method we design works in both frameworks. Though, the sample datasets on which we illustrate it all use discrete timestamps.

²See [16] for a survey on dynamic networks and the vocabulary used to refer to them in the various contexts in which they are encountered and studied.

aggregation process goes along with a loss of information: in each aggregation window, the information on the exact times at which links occur in this window is lost. In particular, in a given time window, it is impossible to know whether a given link (a, b) has occurred before or after another one (b, c) . This question, which determines whether it is possible to go from node a to node c , via b , within this time window (only if ab has occurred before bc), is crucial for many phenomena taking place on the dynamic network, such as epidemic spreads, possibilities of communications and cascade of influence for example. The wider the aggregation period, the greater the amount of information lost. At the limit, aggregating a link stream over the whole period of study yields one single static network which misses all the information on the order of occurrences of links and which therefore very poorly captures the structure of the original dynamic network, see e.g. [27]. Then, more generally, for a given aggregation period, one can ask whether the obtained graph series is a faithful representation of the original link stream. This is precisely the question we address here, through the prism of propagation properties.

1.1 Our results

We show that for many dynamic networks, the length Δ of the window chosen for aggregating the network into a graph series exhibits a threshold, which is proper to each network. Beyond this threshold, the propagation properties of the graph series obtained from aggregation show evidence of alteration, while they are mostly preserved below it. We design a method, called the *occupancy method*, in order to determine this threshold, which we call the *saturation scale* and denote γ . We apply and validate the occupancy method on various real-world datasets, as well as on synthetic dynamic networks.

This answers the fundamental question of deciding whether a given aggregation period gives rise to a graph series that faithfully describes the original dynamic network. The aggregation periods beyond the saturation scale alters the properties of propagation of the dynamics. This range of scale must then be avoided or used only for analyzing properties of the series that do not suffer this alteration.

Moreover, the saturation scale, which is the larger non-altering aggregation period, is a characteristic time scale of the network. It can then be used to compare the properties of different dynamic networks at a same level of aggregation, which is very interesting in itself. Finally, let us emphasize that our method is fully automatic and does not require any parameter as input. Therefore, it can easily be incorporated into any automatic tool for analyzing dynamic networks.

1.2 Related work

It is paradoxical to note that, while the question of the influence of aggregation on the properties of the formed graph series is largely ignored in most of the studies on dynamic networks, this question actually already received a lot of specific attention [18, 36, 37, 41, 7, 39, 12, 3].

In [18], the authors lead a systematic analysis of what is visible from the structure of a dynamic phone-call network when it is aggregated at different time scales. They show that significant characteristics of the dynamics of the

network appear at different scales of analysis, which implies that one should use the broad spectrum of possible scales in order to reveal these different properties of the dynamics. Though we are also concerned by the impact of the aggregation period on the properties of the formed graph series, our motivation and goal are clearly different from those of [18]. Here, we do not intend to find time scales that are able to reveal the key properties of the dynamic network or that are relevant to study phenomena taking place over the network. Instead, we aim at determining the range of aggregation scales such that the formed graph series faithfully describe the original network. Making statistics on the network out of this range of scale may still reveal interesting facts that are invisible at other scales. Nevertheless, for such aggregation scales greater than γ , one should consider only statistics that are not sensitive to the loss of information induced by aggregation (like those used in [18] for example), excluding all statistics based on propagation properties of the dynamics.

This is also the point of view developed in [36], where the authors study the impact of aggregation over the properties of random walks in a dynamic network. They show that the probability of occupation of nodes of the network by such random walks is deeply impacted by aggregation, implying that it should be used with great caution when dealing with phenomenon that depends on propagation properties of the dynamics. The key contribution of the work of [36] is to emphasize and analytically explain the impact of aggregation on random walks, but it does not provide any way of determining a maximum aggregation period that can be used safely, which is precisely our goal here.

In [37], the authors study the impact of the length of the aggregation window, as well as the impact of the type of windows used (disjoint or overlapping or starting at the beginning of the period of study), on the output of a dynamic community tracking algorithm taking as input a series of graphs. Their results show that both the length and the type of the windows used have a strong impact on the dynamic communities outputted by the algorithm. As [36], the purpose of their work is to provide a deeper understanding of the effect of aggregation, but it is not intended to choose a suitable aggregation period.

Contrastingly, the goal of [41] is precisely to determine an ideal aggregation period. In their method, this period is obtained as a trade-off between two metrics that vary monotonically and oppositely with regard to aggregation: one describing the loss of information (increasing with aggregation) and one describing the noise contained in the series of snapshots (decreasing with aggregation). Compared to them, here, we are concerned only with the loss of information. This allows us to avoid some drawbacks and limitations inherent to the approaches based on a trade-off: i) the value selected for the aggregation period strongly depends on the importance given to each metrics and ii) the selected value does not reveal any particular behavior of the properties of the network used in the trade-off, as each of them varies smoothly and monotonically from one extremal value to another one. On the contrary, our method does not depend on any arbitrary choice of ponderation and reveals a natural change in the way the network responds to aggregation at a certain aggregation scale that we determine.

[7] also aims at determining an appropriate time scale for aggregating a link stream into a graph series. Their method does not take into account the loss of information but is instead based on the modes of periodicity and on the self-similarity of the time series of some properties of the snapshots. They observe

that the offset time for which the self similarity of these time series is zero is close to half of the period of the highest frequency visible in their spectra, which is the aggregation period suggested as a result of their method. Though this provides a very relevant time scale for analyzing dynamic networks, its meaning is different from the meaning of the saturation scale we are looking for in this paper. Indeed, an important part of the activity of dynamic networks takes place at time scales much smaller than their modes of periodicity. Therefore, using such periods for aggregation usually induces an important loss of information, which we aim at avoiding here. Let us mention that a similar approach based on modes of periodicity of some time series associated to the network was previously used in [10, 11].

The approaches of [39, 8] are noticeable in that they develop a method to aggregate a link stream on variable length windows. To this purpose, they fix the beginning time of the current aggregation window and they observe the evolution of some statistics of events of the aggregated graph, like the density [39] or the recurrence of ties [8], as the ending time of the window increases. When the observed statistics is optimized, they end the current window and start a new one. The idea is to form a series of so-called "mature" graphs, meaning that these graphs have been aggregated on a time window long enough so that the properties of the formed graph would not change much if it was aggregated on a longer period of time. This motivation is clearly different from ours and the loss of information due to aggregation may occur before the convergence of the properties of the formed graph.

[12] and [3] consider the aggregation of a particular class of link streams: those that are obtained as the result of the oversampling of a dynamic network where links do not occur punctually but instead last over a time interval. Such dynamic networks are often measured by sampling processes that repeatedly check (often periodically) what are the links existing in the network at different times along the period of study, e.g. using sensor devices to measure contacts between individuals [11, 5]. These sampling processes introduce some noise in the data, for example due to failure to measure some links that do exist. The aim of [12, 3] is to find an aggregation period that removes the noise introduced by the sampling process and allows to retrieve the original signal. Here, both our purpose and the kind of dynamic networks we consider are different. We deal only with link streams where links are punctual and do not last over time. The approach of [12, 3] is not intended and not directly applicable to this kind of link streams. Conversely, it must be clear that applying our method to lasting links would require some adaptation and is one key perspective of our work.

For sake of completeness, let us mention two other works that address in different ways the problem of aggregation of link streams into graph series. [1] design a tool for visualizing a dynamic network as a series of snapshots, which takes as parameter the length of the aggregation window. One of the interest of this tool is that it helps to visually choose an aggregation period that gives a comprehensive view of the evolution of the network. Finally, [28] provides alternatives to aggregation by designing two representations of a dynamic network that encode both time and links in the form of a static graph structure.

2 Preliminaries

We describe our methodology in discrete time and with non-directed links, but actually, it applies the same if the time t is continuous and if the links are directed (as in the real-world datasets we consider in Section 5). The only restriction which is meaningful here is that links are punctual events and therefore have no duration. The case where links exist during one interval of time instead of one instant requires some adaptation, both for continuous and discrete time. We now formally define some of the concepts we use in the paper, starting with the process of aggregation of a link stream into a graph series, see example in Figure 1.

Definition 1 (Aggregation) *The aggregation, on disjoint time windows of equal length, of a link stream \mathcal{L} on the period of study $[0, T]$ consists in choosing a constant time period Δ such that $\Delta = T/K$ for some integer $K \geq 1$ and forming the graph series $\mathcal{G}_\Delta = (G_k)_{1 \leq k \leq K}$ defined by $G_k = (V, E_k)$ with*

$$E_k = \{uv \mid \exists(u, v, t) \in \mathcal{L}, (k-1)\Delta \leq t < k\Delta\}$$

where V is the set of nodes involved in the link stream \mathcal{L} .

Note that with this definition, the set of nodes V of each graph of the series \mathcal{G}_Δ is the same. This is a convention adopted for convenience of description. But our methodology applies the same if one keeps, in each snapshot, only the set of nodes that have at least one link during the time window used for aggregation. Doing so, one obtains a series where the set of nodes of the graphs in the series is not fixed but depends on the time. This kind of graph series may suit better in some contexts of study. The methodological tools developed here, though described for series of graphs on a fixed set of nodes, apply also on series of graphs with a variable set of nodes.

A temporal path, in a link stream or a graph series, is a sequence of edges defining a path and occurring at strictly increasing time along the path (see examples given in Figure 1).

Definition 2 (Temporal path in a link stream) *In a link stream \mathcal{L} , a temporal path P is a sequence (u_i, v_i, t_i) of triplets, with $1 \leq i \leq l$ and $l > 0$, such that $\forall i, (u_i, v_i, t_i) \in \mathcal{L}$ and $\forall i > 1, u_i = v_{i-1}$ and $\forall i, j$, if $i < j$ then $t_i < t_j$.*

Definition 3 (Temporal path in a series of graphs) *In a series of graphs $\mathcal{G} = (G_k)_{1 \leq k \leq K}$, a temporal path P is a sequence (u_i, v_i, k_i) of triplets, with $1 \leq i \leq l$ and $l > 0$, such that $\forall i, u_i v_i \in E(G_{k_i})$ and $\forall i > 1, u_i = v_{i-1}$ and $\forall i, j$, if $i < j$ then $k_i < k_j$.*

Remark 1 *Note that, in Definitions 2 and 3, the inequalities are strict. This implies that a temporal path cannot use two links belonging to the same graph of the series or occurring at the same time in the link stream.*

Temporal paths are an essential notion as they capture the propagation properties of the dynamic network. Indeed, all diffusion phenomena in the network, such as communication of information, spreading of epidemics and cascades of influence for example, respect time causality: a node needs to be

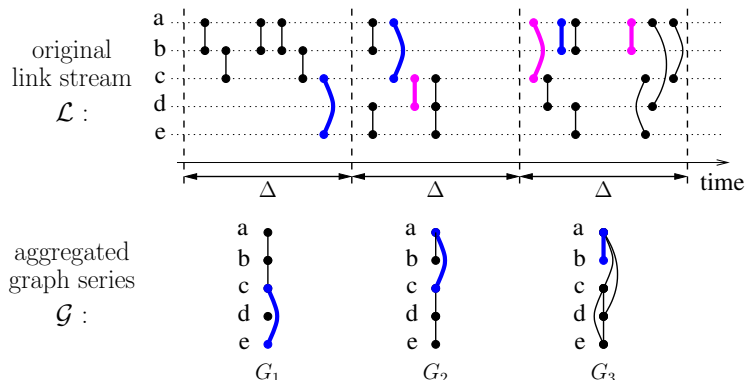


Figure 1: A link stream \mathcal{L} and the graph series $\mathcal{G} = (G_1, G_2, G_3)$ obtained by aggregating \mathcal{L} using a period Δ . The bold dark-blue links depict a temporal path, from e to b , in the link stream and its corresponding temporal path in the graph series. The bold light-pink links also form a temporal path in the link stream, from d to b , but there is no temporal path from d to b in the graph series, because it would require to use two links of graph G_3 , which is not allowed (see Remark 1).

reached by the diffusion before it can propagate it further. Therefore, all these phenomena occur on and follow temporal paths of the dynamic network.

There are two notions of length associated to a temporal path: the topological length, which is the classical one for static graphs, and the duration of the path.

Definition 4 (hops(P) and time(P)) *In a link stream or a graph series, the topological length of a temporal path $P = ((u_i, v_i, t_i))_{1 \leq i \leq l}$ is the number l of edges in the path. In the rest of the paper, we call it the number of hops of P and denote it $\text{hops}(P)$.*

The duration of path P , denoted $\text{time}(P)$, is $t_l - t_1$ in a link stream and $t_l - t_1 + 1$ in a graph series (because each t_i is not an instant as in a link stream but an interval of time which has a duration).

Remark 2 *By definition, in a graph series, we always have $\text{hops}(P) \leq \text{time}(P)$ for any temporal path P . Note that this does not hold for link streams in general, since time is not necessarily indexed by an integer.*

In the rest of the paper, we also use three notions of distance at time t between two nodes u, v of a link stream or a graph series. These notions are based on the minimal arrival time t_{arr} , if any (otherwise t_{arr} is undefined), among all paths from u to v whose departure time is not before t . The *distance in time*, denoted $d_{time}(u, v, t)$, is simply defined as $t_{arr} - t$ in a link stream and $t_{arr} - t + 1$ in a graph series, with the convention $d_{time}(u, v, t) = +\infty$ if t_{arr} is undefined. The *distance in hops*, denoted $d_{hops}(u, v, t)$, is the minimum number of hops among all paths realizing the distance in time $d_{time}(u, v, t)$. By convention, $d_{hops}(u, v, t) = +\infty$ when $d_{time}(u, v, t) = +\infty$. Finally, the distance in absolute time, which is dedicated to aggregated graph series only, is denoted

$d_{time}^{abs}(u, v, t)$ and defined by $d_{time}^{abs}(u, v, t) = \Delta \cdot d_{time}(u, v, t)$. It is the absolute time needed to go from node u to node v in the aggregated graph series, with a departure time not before t , taking into account the fact that each graph of the series represent a time interval of length Δ . Interestingly, this last notion contains in itself the imprecision of the timestamps of the aggregated series.

3 Difficulty of the problem

As previously explained, the larger the length of the aggregation window, the greater the loss of temporal information due to aggregation, as the exact times of occurrence of the links within one given window are lost. Then, we can reformulate the problem as follows: what is the maximum aggregation period that induces no significant loss of information in the graph series compared to the original link stream? A natural way to proceed to answer this question is to make the aggregation period vary from its minimal value to its maximal value and to observe the variations of the properties of the obtained series of graphs in the meanwhile. Then, one would hope to find a time scale beyond which the variation of these properties exhibit a qualitative change. Unfortunately, this does not happen for the classical properties of interest of the graph series. On the opposite, when the aggregation period varies, these properties varies smoothly from one extremal value to another one. Figure 2 shows the results for several properties of the Irvine network, which we use as an example to describe our method in the first sections of this article (see Section 5 for a description of the Irvine dataset).

Figure 2 top-left shows the variations of the mean density of the snapshots of the series, which is also equivalent, up to a multiplicative factor $n - 1$ (where n is the number of vertices), to the mean degree of the nodes in all the snapshots. The plot shows that when the aggregation goes from the minimal temporal resolution of the timestamps (1s) to the total length of the period of study (~ 1175 h), these two properties linearly varies from a very small value (5.7×10^{-7}) to their maximal value (7.2×10^{-3}), which is the one obtained by aggregating the whole dynamic network into one single graph. In this case, the mean density of the graphs in the series is therefore equal to the density of the totally aggregated graph.

The plot in Figure 2 top-right shows that in the meanwhile the mean size of the largest connected component in each snapshot as well as the mean number of non isolated vertices per snapshot (which are very close) exhibit the same behavior: their values go increasingly from a minimal one (2.3 nodes) to a maximal one (1509 nodes), which is the total number of nodes in the network, without exhibiting any non-smooth behavior at any time scale.

Let us now examine the variations of distance properties according to the aggregation period. Figure 2 bottom-left gives the variation of the mean distance in time $d_{time}(u, v, t)$ (see Section 2) for all couples (u, v) of nodes and all time t (such that $d_{time}(u, v, t)$ is finite), in logarithmic scale. As one can see, the curve is almost a straight line, indicating that the mean distance in time depends on Δ following a power law. This comes from the fact that the number of graphs formed in the aggregated series varies as $1/\Delta$: the plot shows that the mean distance in time varies accordingly. This does not help much to detect a time scale at which the properties of the aggregated series significantly change their

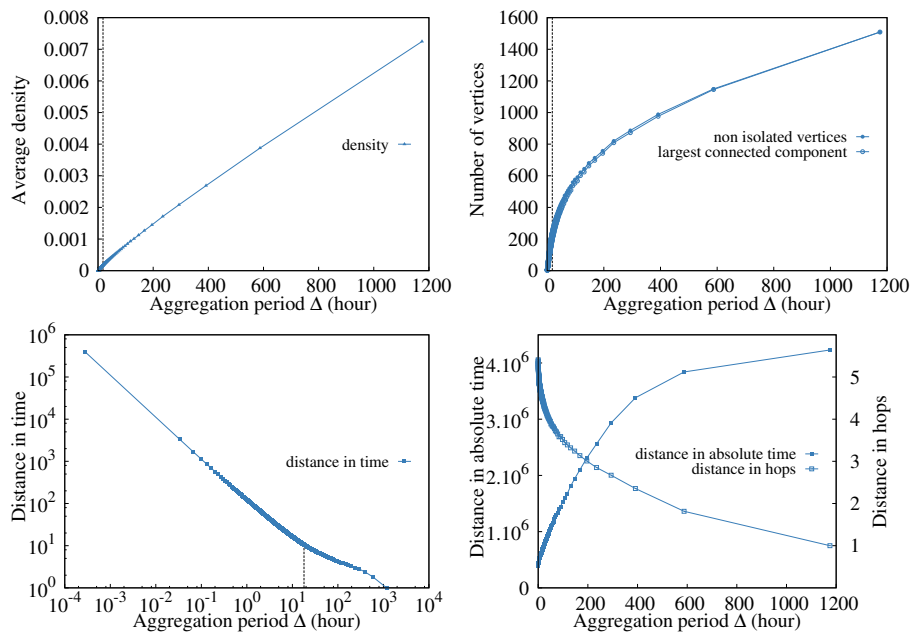


Figure 2: Variation of some classical parameters of the aggregated series of graphs (y-axis) according to the aggregation period Δ (x-axis), for the Irvine network (see Section 5). Top-left: density. Top-right: connectedness properties. Bottom-left: distance in time (log-log scale). Bottom-right: other distance properties (linear scale). The dotted line shows the aggregation period returned by the occupancy method: 18h.

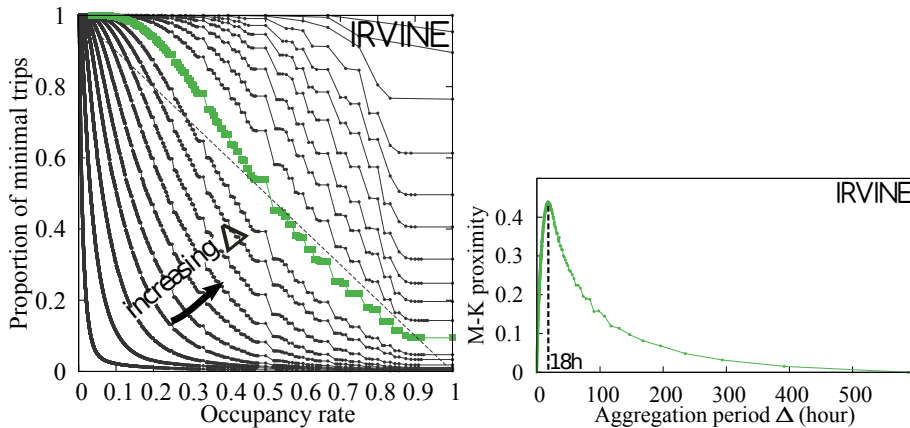


Figure 3: Left: Inverse Cumulative Distributions (ICD) of the occupancy rates (x-axis) of the minimal trips of the aggregated series \mathcal{G}_Δ for several values of the aggregation period Δ in the range $[1, T]$, for the Irvine network. Right: M-K proximity (y-axis) of these distributions with the uniform density distribution according to Δ (x-axis).

behavior.

Pushing further, in Figure 2 bottom-right, we also plotted the mean distance in hops (empty squares) and the mean distance in absolute time (filled squares), both in linear scale. The rationale for using the distance in absolute time, defined as $d_{time}^{abs}(u, v, t) = \Delta \cdot d_{time}(u, v, t)$, is that it does not suffer from the dependence on $1/\Delta$ previously highlighted for the distance in time, since it is canceled by the multiplication by Δ . Then, it gives a clearer insight into the variations of the distance in time with the aggregation period. Unfortunately, as one can see, the situation is the same as for the other parameters previously studied. When the aggregation period Δ increases from the minimal temporal resolution of the timestamps (1s) to the total length of the period of study (~ 1175 h), the mean distance in absolute time increases as well, going monotonically from its minimal value (~ 110 h) until its maximal value (~ 1175 h), which is by definition equal to the total length of the period of study, as there is only one graph in the series formed using the maximum value of Δ . In the meanwhile the mean distance in hops (empty squares in Figure 2 bottom-right) decreases, from 5.4 to 1, without exhibiting any remarkable change at any value of the aggregation period.

Thus, the observation of the variations of the classical properties of the graph series with the aggregation period does not point out scale at which some qualitative changes occur in the way the dynamic network responds to aggregation. Instead, one finds a regular drift from one extreme value to another one³. This constitutes the main difficulty of the problem we consider.

³Note that we present results for only one dataset but they hold similarly for all the four datasets we consider in this paper, cf. Section 5.

4 The occupancy method

We now give the definitions necessary to describe our method and we illustrate it on a sample real-world network, the Irvine network (cf. Section 5).

Definition 5 (Trip and minimal trip) *A trip is a quadruplet (u, v, t_{dep}, t_{arr}) such that there exists a temporal path from u to v whose starting time from u and arriving time at v are both in the interval $[t_{dep}, t_{arr}]$. A trip (u, v, t_{dep}, t_{arr}) is minimal if there exists no trip from u to v in an interval $[t'_{dep}, t'_{arr}]$ strictly included in $[t_{dep}, t_{arr}]$ (i.e. $[t'_{dep}, t'_{arr}] \subsetneq [t_{dep}, t_{arr}]$).*

Definition 6 (Transition and shortest transition) *A temporal path P on two hops, i.e. $P = ((a, b, t_1), (b, c, t_2))$, is called a transition, and P is a shortest transition if (a, c, t_1, t_2) is a minimal trip.*

Definition 7 (Occupancy rate) *For a graph series \mathcal{G} and a temporal path P in \mathcal{G} , the occupancy rate of path P , denoted $occ(P)$, is defined as $occ(P) = hops(P)/time(P)$. The occupancy rate of a minimal trip (u, v, t_{dep}, t_{arr}) is the occupancy rate of a temporal path starting from u at t_{dep} and arriving at v at t_{arr} and having the minimum number of hops among such paths. Note that we always have $0 < occ(P) \leq 1$, since from Remark 2, $0 < hops(P) \leq time(P)$.*

The rationale behind the occupancy rate $occ(P)$, which is comprised between 0 and 1, is to count the proportion of time steps between t_{dep} and t_{arr} that are effectively used by path P to move from one node of the dynamic network to another one. Indeed, only some of the graphs G_t with $t_{dep} \leq t \leq t_{arr}$ contain a link of path P , but not all of them. Therefore, a path P uses some time steps to move to the next node on the path and spends the rest of the time steps simply waiting on the node reached so far, until the next hop to be performed on path P occurs. Then, in other words, the occupancy rate quantifies how much the path P is busy moving from one node to the next one, taking into account that P also spends some proportion of the time without moving. As we explain later in this section, the occupancy rate of paths in the aggregated graph series is strongly impacted by the aggregation period used to form the series, and it is at the core of our method to determine the saturation scale γ .

To this purpose, we make the aggregation period Δ vary from its minimal value, the resolution of the timestamps, until the whole length T of study of the network. For each value of Δ we form the aggregated graph series \mathcal{G}_Δ for which we compute the set of minimal trips and their occupancy rates. Then, for each Δ , we plot the distribution of occupancy rates of all the minimal trips in \mathcal{G}_Δ (considering all pairs of nodes and all time intervals), see Figure 3 left.

Necessarily, when Δ is close to its minimal value, provided that the resolution of the timestamps is fine enough, the distribution of occupancy rates must be concentrated on values close to 0. The reason is that the aggregation windows contain only few data and the shortest paths therefore need to wait several slot of times before finding one opportunity to perform the next hop. On the opposite, when the aggregation period reaches its maximum value, by definition, all the minimal trips are made of one single link (because there is only one graph in the aggregated series) and their occupation rate is 1. Then, the distribution is again concentrated, this time on the value 1. What is remarkable here (Figure 3 left) is that the distribution changes from values concentrated near 0 to values

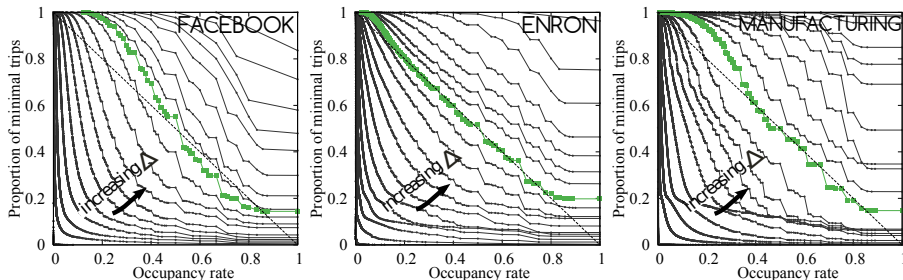


Figure 4: Inverse Cumulative Distributions (ICD) of the occupancy rates (x -axis) of the minimal trips of the aggregated series \mathcal{G}_Δ for several values of the aggregation period Δ in the range $[1, T]$, for the Facebook, Enron and Manufacturing networks.

concentrated on 1 in a very specific manner: it first progressively stretches toward 1 until it almost equally occupies all the values on the range from 0 to 1 and then it contracts again, leaving the low values to progressively concentrate on the values close to 1.

The saturation scale γ is precisely the value of Δ for which the distribution is maximally stretched on the interval $[0, 1]$ (curve marked with green squares on Figure 3 left). In order to detect it, we compute for each value of Δ in the total range of variation, the M-K distance $d(\Delta)$ (see Section 7 for a definition) between the distribution obtained for Δ and the uniform density distribution on $[0, 1]$, i.e. the distribution whose inverse cumulative is the straight line $y = 1 - x$. We then plot the M-K proximity, defined as $1/2 - d(\Delta)$, in Figure 3 right. This confirms the observation made above on the way the distribution first stretches and then concentrate again: accordingly, the M-K proximity first increases and then decreases. As a consequence, the value γ returned by our method is the value of Δ that realizes the maximum of the M-K proximity. Of course, one may think of many other ways to determine which Δ gives the maximum stretch of the distribution. We actually tried several of them (see Section 7) and we chose to use the M-K distance with the uniform density distribution because it gives results that are visually satisfying and it is conceptually simple.

Let us now explain the meaning of γ . As pointed out in the introduction, the most significant loss of information due to aggregation is the loss of the order in which the links involving one given node u occur in a given aggregation window. This loss makes it impossible to know whether there exists a transition from node v_1 to node v_2 going through node u within an aggregation window where both links (v_1, u) and (u, v_2) occur: this depends on whether (v_1, u) occurs before (u, v_2) or not, which is lost with aggregation. Then, the propagation properties of the aggregated series do not faithfully reflect those of the original link stream.

To this regard, a very low occupancy rate for most minimal trips of the series denotes that most paths spend a long time waiting between two consecutive hops. This means that there are only very few links in each aggregation window, which implies that only very few information is lost in the whole graph series about the order of occurrences of links. On the opposite, a very high occupancy rate for most of the minimal trips reveals that at each time in the graph series,

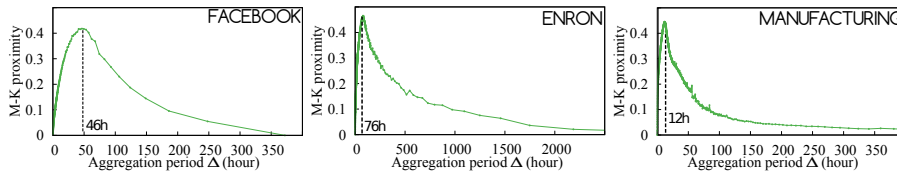


Figure 5: M-K proximity (y-axis) of the distribution of occupancy rates of minimal trips of the aggregated series \mathcal{G}_Δ according to the aggregation period Δ (x-axis), for the Facebook, Enron and Manufacturing networks.

there is a high probability to find a next hop to perform on any given shortest path, meaning that, in each snapshot, a high proportion of nodes are involved in a high number of edges. Then, at the same time, the information on the existence or the non existence of a transition, in the original link stream, using a couple of these edges incident to one same node is lost, which constitutes the essential loss resulting from the aggregation process.

What makes the occupancy rate so remarkable with regard to aggregation, compared to the classical parameters such as density for example (see Figure 2), is that the evolution of the distribution of occupancy rates when the aggregation period varies clearly shows two distinct phases. In the first phase of variation, below γ , only the low values of the distribution increase, while the proportion of high occupancy rates almost does not change. This means that during this phase, the effect of increasing the aggregation period is mainly to fill the lack of links in the aggregation windows without inducing a significant loss of information. On the opposite, in the second phase, beyond γ , there is a strong increase of the proportion of minimal trips having a very high occupancy rate, 1 or close to 1, indicating that the loss of information due to aggregation becomes non-negligible. Therefore, the saturation scale γ appears as a separation between the range of values, below γ , where the aggregated graph series still faithfully describes the original link stream and the range of values, beyond γ , where aggregation alters the properties of propagation of the original link stream.

5 Results on real-world datasets

In this section we apply our methodology and discuss the results obtained on four link streams, whose timestamps have a resolution of 1s. The *UC Irvine messages* network [32], which is the one used for presentation of the method in the previous section, is made of 48 000 messages sent between 1 509 users of an online community of students from the University of California, Irvine, over a period of 48 days. The *Facebook wall posts* network [46] is made of 11 991 wall posts between a group of 3 387 Facebook users over a period of 1 month. The *Enron emails* network [17] contains 15 951 individual emails sent between a group of 150 employees of the Enron company during year 2001. Finally, the *Manufacturing emails* network [26] contains 82 894 internal emails between 153 employees of a mid-sized manufacturing company over a period of 8 months.

From a computational point of view, let us mention that the algorithm we use for computing the distribution of occupancy rates in one given graph series

\mathcal{G} has time complexity $O(nM)$, where n is the number of vertices of one graph in the series (which is the same for all graphs) and M is the total number of edges in all the graphs of the series. It is a dynamic programming scheme going backward in time: at one step, knowing all the minimal trips of the series starting not before time $k+1$, the algorithm computes the minimal trips starting exactly at time k , their duration and their minimum number of hops. In the occupancy method, this algorithm has to be run as many times as the number of values of Δ used for the aggregation period. But the most costly computations are the ones made for small values of Δ , as M is then large.

We applied the occupancy method on each of the four datasets mentioned above. The distributions of occupancy rates of the minimal trips in the aggregated graph series are given on Figure 3 left for Irvine and Figure 4 for the three other networks, their M-K proximity with the uniform density distribution is given on Figure 3 right and Figure 5. One can see that the observations made on the Irvine network in Section 4, hold for all the four datasets. When the aggregation period Δ increases, the distribution of occupancy rates, initially concentrated near 0, stretches until it occupies almost equitably all the range of values between 0 and 1, and then concentrates again on the values close to 1. Consequently, the proximity with the uniform density distribution first increases, until it reaches a maximum for $\Delta = \gamma$, which is the saturation scale returned by the occupancy method, and then decreases until the aggregation period reaches its maximum value T . This shows that the way the distribution of occupancy rates evolves with the aggregation period is a fundamental phenomenon common to many dynamic networks, therefore guaranteeing that our method is sound and that it can be used for a wide range of dynamic networks.

The values returned for γ in each of the four cases are: 18 hours for the Irvine message network, 46 hours for the Facebook wall-post network, 78 hours for the Enron email network and 12 hours for the Manufacturing email network. These values, between half a day and three days, are in accordance with the fact that both emails and on-line social network messages are generally not dedicated to live discussions. In the case of email networks for example, most of people only send some emails a day and frequently wait for some hours or some days before getting a reply. Therefore, this range of values seems appropriate for the largest aggregation scales providing accurate views of the original link streams.

The aggregation periods returned by our method also appear to be in accordance with the level of activity of these 4 networks. The two greater values, 46h for Facebook and 78h for Enron, are obtained for the two networks that have the lower activity, 0.12 and 0.29 messages sent in average per person per day for Facebook and Enron respectively. The two other networks have higher activities, 0.66 messages per person per day in the Irvine network and 2.22 in the Manufacturing network, and have smaller saturation scales, 18 hours and 12 hours respectively. As one can see, the average activity has a strong influence on the saturation scale, even though this is not the only parameter affecting it. We further investigate the relationship between the level of activity and the saturation scale in the next section.

Finally, let us emphasize that the aggregation period γ returned by our method should not be interpreted as the best possible one but instead as an upper-bound on the aggregation periods that are suitable for studying the network. For many practical studies, one may prefer to choose an aggregation period slightly lower than γ , which will preserve more carefully the properties

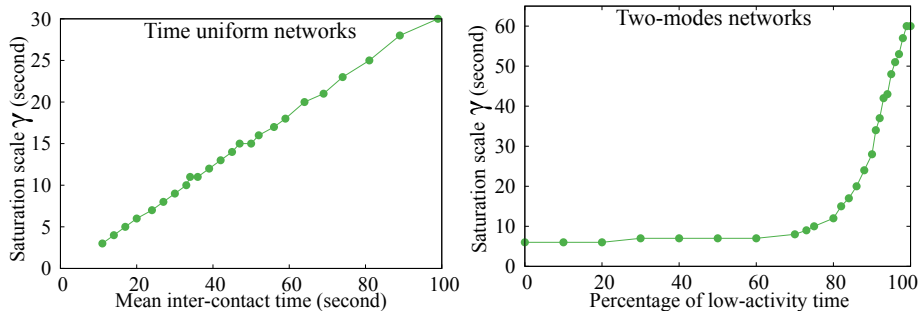


Figure 6: Left: for time uniform networks, saturation scale (y-axis) in function of the mean inter-contact time of nodes (x-axis). Right: for two-mode networks, saturation scale (y-axis) in function of the percentage of low-activity time (x-axis).

of the network. For example, in the case of the four networks we study here, one can note that the proportion of minimal trips having occupancy rate 1 started to increase just before the distribution of occupancy rates reaches its maximal stretched position (the one selected by our method). Then, one could prefer to use an aggregation period smaller than γ in order to get a finer grain representation of the dynamic network. In Section 8, we give some ways to directly estimate the loss of information in the aggregated graph series that can be used to choose more accurately the aggregation period in the range of scales immediately preceding γ .

6 Results on synthetic networks

We now investigate how the aggregation period returned by our method depends on the level of activity of the link streams considered, i.e. the number of links per node and per unit of time, and on the temporal heterogeneity of this activity. To this purpose, we use two kinds of synthetic dynamic networks, where the activity is uniformly distributed between all pairs of nodes. The first kind, called *time uniform networks*, is generated by assigning N links ($N \ll T$) to each pair of the $n = 100$ nodes of the network and uniformly randomly choosing each of their timestamps between 0 and $T = 100\,000$ s. We make the value of N vary from 10 to 100 and for each of these values, we compute the aggregation period γ returned by the occupancy method. Results are given in Figure 6 left, which shows γ as a function of the average inter-contact time of one node, that is $T/(N(n-1))$. For these time uniform networks, the aggregation period returned by the occupancy method is perfectly proportional to the average inter-contact time, showing that our method correctly takes into account the level of activity of the link stream.

However, most of the dynamic networks encountered in practice are far from being uniformly active over time. Many of them instead alternate periods of intense activity with periods of lower activity. In particular, this is the case for networks coming from human activities, such as the ones considered in Section 5, which often exhibit circadian rhythms. Then the question naturally arises to

know how the saturation scale behaves according to this temporal heterogeneity. Does it simply make the average between the different levels of activity? Or does it favor one of them? To answer these questions we generate *two-mode networks* that are built by 10 alternations of one period of high activity and one period of low activity, which are time uniform networks with parameters N_1, T_1 and N_2, T_2 respectively. N_1, N_2 and the whole length $T = 10(T_1 + T_2)$ of study are fixed and we vary the ratio between T_1 and T_2 .

Figure 6 right gives the saturation scale γ as a function of the percentage $\rho = T_2/(T_1 + T_2)$ of low-activity time in the network. The curve goes from the value of γ for the high-activity mode (for $\rho = 0\%$) to the one, much larger, for the low-activity mode (for $\rho = 100\%$). The plot shows that when the proportion of low activity varies from 0% to 70-80%, the saturation scale almost does not increase: it remains very close to the smaller value of the high-activity network, which preserves better the information contained in the original link stream. This is surprising as one would rather expect the saturation scale to be a compromise between its value for the low-activity periods and its value for the high-activity periods. This shows that in presence of heterogeneity of the activity along time, even with high-activity periods occupying only 30% to 20% of the time, the saturation scale returned by the occupancy method is respectful of this important part of the dynamics. Moreover, and importantly, the fact that the saturation scale does not linearly vary with respect to the percentage of low-activity time in the network shows that, for networks that are not time uniform (which is in particular the case of real-world networks), the saturation scale returned by the occupancy method does not only depend on the mean inter-contact time of nodes in the network (or equivalently on the frequency of links in the network).

When the proportion of low-activity time goes beyond 80%, the aggregation period returned starts to increase until it reaches the value for the low-activity network when its proportion in time is 100%. This seems natural as when the low-activity part takes most of the time of the dynamic network, it does not make sense to continue to study it with a scale which is suitable only for a marginal part of the time. Nevertheless, we note that the increase of γ is progressive. For example for 90% of low activity, the returned value is close to the arithmetic mean between the values for the two modes of the network. This shows that the returned aggregation does not forget too quickly the high activity part of the dynamics, which once again is a desirable feature of such a method.

7 Detection of the more uniform distribution

In this section we consider several methods for selecting the aggregation period γ that gives the distribution of occupancy rates that is the more uniformly spread on $[0, 1]$, and we study the dependence of γ on the chosen selection method. Until now, all the results we gave were obtained by selecting the distribution which minimizes the M-K distance with the uniform density distribution on $[0, 1]$. Of course, one may think of many other ways to select γ . This includes plotting the sets of distributions obtained when the aggregation period spans its entire range of variation, like on Figure 3, and selecting one distribution by visual mean. This empirical method is likely to give the most satisfactory results

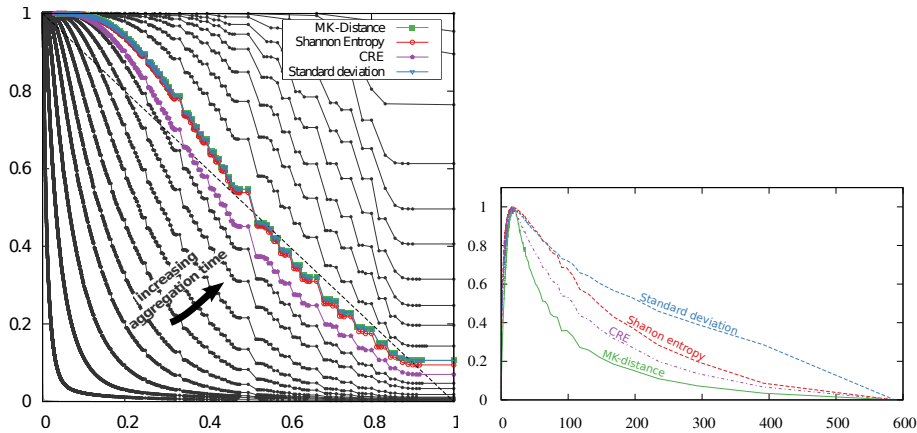


Figure 7: Results of four methods for selection of the more uniformly spread distribution: M-K proximity, standard deviation, Shannon entropy with 10 slots and cumulative residual entropy (CRE). The left plot shows the distributions selected by the maximum of each of the metrics and the right plot shows the variations (normalized to have maximum 1) of each metric (y-axis) depending on the aggregation period Δ (x-axis).

in practice and will therefore be preferred in many studies. However, here, we are interested only in quantitative methods of selection, for two reasons. Firstly, we want to provide a uniquely defined value which can be used as a reference for comparing the saturation scales of different dynamic networks. Secondly, we want our method to be fully automatic in order to be easily incorporable to any tool for analysis of dynamic networks.

In addition to the method based on the M-K distance, which we used until now, we now consider four other selection methods and compare their results. These four additional methods are based respectively on: standard deviation, variation coefficient, Shannon entropy and cumulative residual entropy. Figure 7 gives the results obtained when applying these methods on the Irvine data set. We now describe and analyze them one by one each before giving a global comparison of their respective results.

M-K distance with the uniform density distribution. The Monge-Kantorovich distance is a way to measure the distance between two distributions of probability on the same support, here $[0, 1]$. It is defined as the area comprised between the two inverse cumulative distributions of the probability distributions to be compared. Here, as we are looking for the distribution which is maximally spread over $[0, 1]$, we compare each distribution with the uniform density distribution, which gives $dist_{M-K}(X) = \int_{[0,1]} |P(X > \lambda) - (1 - \lambda)| d\lambda$, where X is the random variable defined by the occupancy rate. Then, the aggregation period we select is the one for which the distribution of occupancy rates X realizes the minimum of $dist_{M-K}(X)$. In order to get the desired distribution for the maximum of the measure, instead of the minimum, as for all the other measures we consider, we rather use the corresponding proximity measure defined as $1/2 - dist_{M-K}(X)$, as $dist_{M-K}(X)$ is always less than $1/2$. This metric is

the one we use throughout the article. It gives visually very satisfying results for all the data sets.

Standard deviation. This method selects the distribution having the maximum standard deviation $\sigma = \sqrt{E[(X - \mu)^2]}$, where X is the random variable defined by occupancy rate and μ is its mean value. This is one of the most direct measure one can think of in order to compare the spread of distributions on support $[0, 1]$. It gives very satisfactory results, comparable to the one obtained with the M-K distance. Nevertheless, it tends to select slightly higher aggregation period than the M-K distance, as the standard deviation is less penalized by the increasing of occupancy rates 1, which is the maximal value in the distribution. Then, usually, the aggregation period selected by the M-K distance is visually a bit more satisfying. This is the reason why we prefer to present our methodology using the M-K distance, but the two metrics actually give comparable results.

Variation coefficient. Another very natural method is the one that selects the distribution having the maximum variation coefficient $c_v = \sigma/\mu$, where μ is the mean and σ the standard deviation of the distribution. Moreover, it could possibly correct the slight drawback of the standard deviation pointed above, which tends to select a little bit higher value than would desire. Unfortunately, the method based on variation coefficient suffers from a much more severe limitation: it favors too much distributions having a small mean and therefore proposes only very short aggregation periods, or even not to aggregate at all. Among all the methods we tried, this is the only one which gives clearly unsatisfactory results to select the more spread distribution.

Shannon entropy. In information theory, the Shannon entropy $H(X) = E[-\ln(P(X))]$ of a random variable X (here the occupancy rate) allows to measure the spread of distributions on a given fixed finite support. This means that in order to compare different distributions, the set of possible values taken by the distributions (the support) must be the same and must be finite. As for the measure based on the M-K distance, the distribution which maximizes the Shannon entropy is the one with uniform density on the considered support. The difficulty we face here to use this measure is that the supports of the distributions we want to compare are not the same: the set of possible values of the occupancy rate is different for each aggregation period. There are different ways to deal with this issue in order to compare all the distributions we obtain when varying the aggregation period. The first one is to artificially take one common support for all the distributions, the minimal such support being the union of the supports of all obtained distributions. Unfortunately, when applied with this support, the method always selects the distributions that effectively use the larger part of the support, that is those obtained for very short aggregation periods. As noted for the variation coefficient, this does not give satisfactory results.

Another possibility to solve this issue is to discretize the segment $[0, 1]$ into k slots of equal length and to compute for each distribution what is the probability that the value of the occupancy rate belongs to each slot. For example, when applied with $k = 10$, this measures gives very satisfactory results. On the

other hand, the returned aggregation period depends on the number of slots chosen. The results are sensibly different using $k = 5$ or $k = 20$. With smaller number of slots, the method tends to select higher aggregation periods, while with greater number of slots, like previously, it favors the distributions having greater original supports, which are those obtained for short aggregation periods. For $k = 100$ this trend is already clearly marked: the value returned for γ is less than half of the one returned with $k = 10$ and visually, the distribution selected does not appear to be spread over $[0, 1]$ in the best possible way among all the distributions obtained. Despite of this, we note that in the range of variation we consider for k , namely $[2, 100]$, and for the data sets we use, the selection method based on the Shannon entropy properly determines the order of magnitude of the saturation scale γ and even gives visually very satisfying results for values of k between 5 and 20. Nevertheless, because of its sensitivity to the chosen k , we decided not to use this selection method. The next metric is another attempt to correct the difficulties arising from the use of the Shannon entropy.

Cumulative residual entropy (CRE). This is a variation of the Shannon entropy that is able to compare distributions with the same *infinite* support. It is suitable for our purpose as all the distributions we consider have support $[0, 1]$. In this case, the cumulative residual entropy is defined as $\varepsilon(X) = - \int_{[0,1]} P(X > \lambda) \log(P(X > \lambda)) d\lambda$. As for the Shannon entropy, the maximum value of the CRE is reached for the uniform density on the considered support. It turns out that this selection method performs well on all the data sets we used. It gives aggregation periods close to the one obtained by the M-K distance, usually shorter. Visually the results are satisfying, even if on some example this method appears to favor a bit too much distributions with large supports. But this is only a slight effect and this method seems quite suitable for our needs, and theoretically well funded. The reason why we preferred the M-K distance with the uniform density distribution is that it is conceptually much simpler and gives as good results.

Let us now compare the aggregation periods selected by the 5 methods above for the Irvine data set. First of all, let us note that all the selected aggregation periods are very close between 14.5h and 18.7h, except one of them, the one based on the variation coefficient. This method proposes an aggregation period of 1 second, which is the resolution of the timestamps, and the distribution it selects is very far from being uniformly spread on $[0, 1]$. The variation coefficient method therefore appears not to be suitable for our purpose. On this data set, the distributions selected by the M-K distance and the standard deviation methods are exactly the same. They are the distribution obtained with an aggregation period of 18.7h. The distribution selected by the method based on Shannon entropy with 10 slots of width 0.1 is almost indistinguishable from the previous one, it is the distribution obtained for an aggregation period of 18.1h. These two distributions are indeed visually quite well spread on $[0, 1]$ and therefore, these three methods give here very satisfying results, and they also did on other data-sets (not presented here). The method based on the cumulative residual entropy selects the distribution obtained with an aggregation period of 14.5h, which is slightly lower than the three previous method. In this case, this

distribution is also visually very well spread on $[0, 1]$ and then quite good for our purpose.

As a conclusion, except the method based on the variation coefficient, all the four other methods we considered appear to give satisfactory results on all the data sets we use. We chose the method based on the M-K distance because it is conceptually simple and it gives very satisfactory results. Beyond the slight differences between these methods, the fact that they all give very close values of γ shows that each of them is sound and is appropriate to detect the aggregation period that maximally stretches the distribution of the occupancy rates in the interval $[0, 1]$.

8 Validation

In this section we quantify the amount of information which is lost when one aggregates the network using a given period Δ . This allows us to validate our approach by evaluating the loss obtained for $\Delta = \gamma$. Moreover, this provides tools to select more accurately an aggregation period, in the range preceding γ , that is suitable for representing a given link stream as a graph series.

The first measure of loss we use is the proportion of shortest transitions (minimal trips with two hops, cf. Definition 6) that lay entirely in one aggregation window. These are exactly the shortest transitions of the original link stream that do not exist anymore in the aggregated series of graphs: all the other minimal trips having their two hops, say $(a, b, t_1), (b, c, t_2)$, in two different aggregation windows, say indexed t'_1 and t'_2 , still exist in the form $(a, b, t'_1), (b, c, t'_2)$ in the aggregated series. We chose this way of measuring the loss as the shortest transitions are the key units that capture the possibilities of propagation in the link streams. In other words, note that if all the shortest transitions of the link stream are conserved in the graph series (in the sense above), so are all the minimal trips, and therefore, the possibilities of propagation in the dynamic network are unchanged.

Figure 8 left depicts the proportion of lost transitions as a function of the aggregation period Δ , for the Irvine network. One can see that when the aggregation increases, starting from 1 second, the number of lost transitions first remains very low during several orders of magnitude, until an aggregation period of 0.5h where only 10% have been lost. The main part of the loss (80%) is concentrated on the range between 0.5h and 235h, i.e. a bit more than 2 orders of magnitude. The saturation scale $\gamma = 18h$ returned by the occupancy method is in the beginning of this range, and in the middle in terms of order of magnitude. This shows that the occupancy method successfully detects the order of magnitude of the time scale from which the loss of information starts to be visible. For $\Delta = \gamma$, 48% of the shortest transitions are lost. Therefore, one may prefer to limit further the range of aggregation periods used, for example one order of magnitude below γ .

On the other hand, it must be clear that the measure of the loss used above is rather pessimistic. Indeed, some of the shortest transitions of the original link stream that are lost can be replaced by some others slightly longer or occurring a bit later. This limits the actual impact of this loss on the possibilities of propagation in the aggregated series. As lost transitions can be replaced, the duration of a minimal trip that was using some of these transitions may be

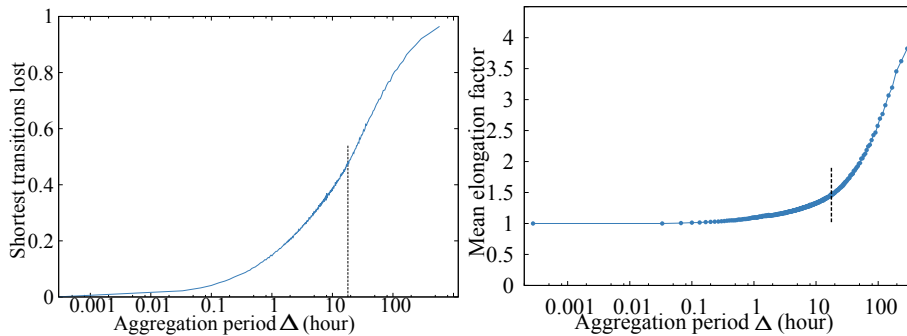


Figure 8: Left: proportion of shortest transitions lost (y-axis) in the aggregated series \mathcal{G}_Δ according to the aggregation period Δ (x-axis). Right: mean elongation factor of minimal trips of \mathcal{G}_Δ (y-axis) according to Δ (x-axis).

only slightly altered by their loss (or even not at all). For this reason, we also use a measure of loss which is based on the elongation of minimal trips in the aggregated series \mathcal{G}_Δ compared to the original link stream \mathcal{L} .

Definition 8 (Elongation factor) *The elongation factor e_P of a minimal trip $P = (u, v, t_u, t_v)$ of \mathcal{G}_Δ , with $t_u \neq t_v$, is defined as the ratio*

$$e_P = (t_v - t_u + 1) \cdot \Delta / \text{time}_{\mathcal{L}}(P)$$

where $\text{time}_{\mathcal{L}}(P) = \min\{t'_v - t'_u \mid (u, v, t'_u, t'_v) \text{ is a minimal trip of } \mathcal{L} \text{ and } t'_u, t'_v \in [(t_u - 1) \cdot \Delta, t_v \cdot \Delta]\}$

Note that when $t_u \neq t_v$, we necessarily have $\text{time}_{\mathcal{L}}(P) \neq 0$. Therefore, the elongation factor is properly defined. Figure 8 right gives the mean elongation factor (y-axis) of all minimal trips of the series aggregated with period Δ (x-axis), for the Irvine network. When Δ increases, the elongation factor of minimal trips first stays very close to 1 during several orders of magnitude, before it suddenly raises when the aggregation period reaches values around the saturation scale γ . This shows that our method properly determines the scale at which the properties of propagation of the link streams start to be altered by aggregation. For $\Delta = \gamma$, the mean elongation ratio of minimal trips is less than 1.5, showing that despite the 48% of shortest transitions lost, the propagation properties of the original link stream are not yet too drastically altered.

9 Conclusion

We showed that there exists a threshold, called the saturation scale γ , for the aggregation period of a link stream at which a qualitative change occurs in the way the network responds to aggregation. We showed that this change of behavior reveals an alteration of the properties of propagation of the dynamics, implying that dynamic networks should not be aggregated with a period larger than γ to perform analyses that depend on these properties. In addition, we designed a fully automatic and parameter-free method to determine the value of γ for an arbitrary link stream.

Our work opens several perspectives to improve the method and broaden its field of application. The first of these perspectives is to extend the occupancy method to the case where links have a duration. The method presented in this article applies to both discrete and continuous time, to both undirected links and directed links, but it is able to deal only with links that are punctual events. However, in some contexts, the links of the dynamic network last during an interval of time (e.g. phone calls and physical contacts between individuals). Adapting the occupancy method to this case would be highly desirable. One particularly interesting way to do so would be to develop a notion of minimal trip that is specifically adapted to links that have a duration.

In Section 6, we pointed out a nice behavior of the occupancy method in presence of temporal heterogeneity in the activity of the link stream processed: the aggregation scale γ returned in this case gives more importance to the parts of the dynamics that have a high level of activity, even if they do not occupy the majority of the time. Nevertheless, if these periods are really too short, they will have only a limited impact on the value of γ . As a consequence, these highly active parts of the link stream, which are likely to contain a valuable information for the whole dynamics, may be smoothed out by the aggregation process. Avoiding this phenomenon by better taking into account the temporal heterogeneity of the activity of the link stream would constitute a key improvement. To this end, one could enhance the method so that it is able to separate the high activity periods from the lower activity periods and to determine an appropriate aggregation scale for each of these parts independently. Then one could decide either to aggregate the whole link stream at the shortest aggregation scale detected, which is the one that better preserves the information contained in it, or to partition the period of study and aggregate each part with a different length of window.

Funding

This work was supported by the French National Research Agency contract CODDDE [ANR-13-CORD-0017-01]. It was performed within the framework of the LABEX MILYON [ANR-10-LABX-0070] of Université de Lyon, within the program "Investissements d'Avenir" [ANR-11-IDEX-0007] operated by the French National Research Agency (ANR). The second author gratefully acknowledges the support from a grant from Région Rhône-Alpes and from the delegation program of CNRS.

References

- [1] Skye Bender-deMoll and Daniel A. McFarland. The art and science of dynamic network visualization. *Journal of Social Structure*, 7, 2006.
- [2] D. Braha and Y. Bar-Yam. From centrality to temporary fame: Dynamic centrality in complex networks. *Complexity*, 12(2):59–63, 2006.
- [3] Rajmonda Sulo Caceres, Tanya Berger-Wolf, and Robert Grossman. Temporal scale of processes in dynamic networks. In *IEEE 11th International*

- Conference on Data Mining Workshops (ICDMW 2011)*, pages 925–932. IEEE, 2011.
- [4] J. Candia, M.C. González, P. Wang, T. Schoenharl, G. Madey, and A.L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A*, 41:224015+, 2008.
 - [5] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5:e11596, 2010.
 - [6] Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 163–172. ACM, 2007.
 - [7] A. Clauset and N. Eagle. Persistence and periodicity in a dynamic proximity network. In *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, 2007.
 - [8] Richard K Darst, Clara Granell, Alex Arenas, Sergio Gómez, Jari Saramäki, and Santo Fortunato. Detection of timescales in evolving complex systems. *Scientific Reports*, 6, 2016.
 - [9] Prasanna Desikan and Jaideep Srivastava. Mining temporally changing web usage graphs. In *Advances in Web Mining and Web Usage Analysis: 6th International Workshop on Knowledge Discovery on the Web (WebKDD 2004)*, number 3932 in LNAI, pages 1–17. Springer, 2006.
 - [10] Nathan Eagle. *Machine perception and learning of complex social systems*. PhD thesis, Massachusetts Institute of Technology, Department of Media Arts and Sciences, 2005.
 - [11] Nathan Eagle and Alex Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.
 - [12] Benjamin Fish and Rajmonda S. Caceres. Handling oversampling in dynamic networks using link prediction. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2015)*, volume 9285, part II of LNAI, pages 671–686. Springer, 2015.
 - [13] A. Gautreau, A. Barrat, and M. Barthélemy. Microdynamics in stationary complex networks. *PNAS*, 106:8847–8852, 2009.
 - [14] P. Holme. Network dynamics of ongoing social relationships. *Europhysics Letters*, 64(3):427, 2003.
 - [15] P. Holme, S. Min Park, B.J. Kim, and C.R. Edling. Korean university life in a network perspective: Dynamics of a large affiliation network. *Physica A: Statistical Mechanics and Its Applications*, 373:821–830, 2007.
 - [16] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

- [17] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *15th European Conference on Machine Learning (ECML 2004)*, pages 217–226. Springer, 2004.
- [18] Gautier Krings, Márton Karsai, Sebastian Bernharsson, Vincent D. Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of aggregated networks. *EPJ Data Science*, 1(4):1–16, 2012.
- [19] Mayank Lahiri, Arun S. Maiya, Rajmonda Sulo, Habiba, and Tanya Y. Berger-Wolf. The impact of structural changes on predictions of diffusion in networks. In *IEEE International Conference on Data Mining Workshops (ICDMW 2008)*, pages 939–948, 2008.
- [20] Matthieu Latapy, Assia Hamzaoui, and Clémence Magnien. Detecting events in the dynamics of ego-centered measurements of the internet topology. *Journal of Complex Networks*, 2013.
- [21] Matthieu Latapy and Clémence Magnien. Complex network measurements: Estimating the relevance of observed properties. In *27th IEEE Conference on Computer Communications (INFOCOM 2008)*, pages 1660–1668. IEEE, 2008.
- [22] Yannick Léo, Christophe Crespelle, and Eric Fleury. Non-altering time scales for aggregation of dynamic networks into series of graphs. In *11th International Conference on emerging Networking EXperiments and Technologies (CoNEXT 2015)*. ACM, 2015. To appear.
- [23] Kristina Lerman, Rumi Ghosh, and Jeon Hyung Kang. Centrality metric for dynamic networks. In *8th Workshop on Mining and Learning with Graphs (MLG 2010)*, pages 70–77. ACM, 2010.
- [24] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.
- [25] Lucie Martinet, Christophe Crespelle, and Eric Fleury. Dynamic contact network analysis in hospital wards. In *5th Workshop on Complex Networks (CompleNet 2014)*, number 549 in Studies in Computational Intelligence, pages 241–249. Springer, 2014.
- [26] Radosław Michalski, Sebastian Palus, and Przemysław Kazienko. Matching organizational structure and social network extracted from email communication. In *14th International Conference on Business Information Systems (BIS 2011)*, volume 87 of *LNBIP*, pages 197–206. Springer, 2011.
- [27] James Moody. The importance of relationship timing for diffusion. *Social Forces*, 81(1):25–56, 2002.
- [28] James Moody. Static representations of dynamic networks. Technical report, Duke Population Research Institute, Duke University, Durham, 2008.
- [29] James Moody, Daniel McFarland, and Skye Bender-deMoll. Dynamic network visualizations. *American Journal of Sociology*, 110(4):1206–1241, 2005.

- [30] Vincent Neiger, Christophe Crespelle, and Eric Fleury. On the structure of changes in dynamic contact networks. In *Workshop on Complex Networks and their Applications (Complex Networks 2012)*. In *8th International Conference on Signal Image Technology and Internet Based Systems (SITIS 2012)*, pages 731–738. IEEE, 2012.
- [31] Roberto N. Onody and Paulo A. de Castro. Complex network study of Brazilian soccer players. *Phys. Rev. E*, 70:037103, 2004.
- [32] Pietro Panzarasa, Tore Opsahl, and Kathleen M. Carley. Patterns and dynamics of users’ behavior and interaction: Network analysis of an online community. *J. Am. Soc. Inf. Sci. Technol.*, 60(5):911–932, 2009.
- [33] Leto Peel and Aaron Clauset. Detecting change points in the large-scale structure of evolving networks. In *29th AAAI Conference on Artificial Intelligence (AAAI 2015)*, pages 2914–2920. AAAI Press, 2015.
- [34] N. Perra, B. Goncalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Scientific Reports*, 2, 2012.
- [35] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80:056103+, 2009.
- [36] Bruno Ribeiro, Nicola Perra, and Andrea Baronchelli. Quantifying the effect of temporal resolution on time-varying networks. *Scientific Reports*, 3, 2013.
- [37] Stanislaw Saganowski, Piotr Brodka, and Przemyslaw Kazienko. Influence of the dynamic social network timeframe type and size on the group evolution discovery. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 679–683. IEEE, 2012.
- [38] Purnamrita Sarkar and Andrew W. Moore. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40, 2005.
- [39] Sucheta Soundarajan, Acar Tamersoy, Elias B. Khalil, Tina Eliassi-Rad, Duen Horng Chau, Brian Gallagher, and Kevin Roundy. Generating graph snapshots from streaming edge data. In *WWW’16 Companion*. ACM, 2016.
- [40] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176, 2011.
- [41] Rajmonda Sulo, Tanya Berger-Wolf, and Robert Grossman. Meaningful selection of temporal resolution for dynamic networks. In *8th Workshop on Mining and Learning with Graphs (MLG 2010)*, pages 127–136. ACM, 2010.

- [42] Jimeng Sun, Spiros Papadimitriou, Philip S. Yu, and Christos Faloutsos. Graphscope: Parameter-free mining of large time-evolving graphs. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*, pages 687–696. ACM, 2007.
- [43] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: Dynamic tensor analysis. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pages 374–383. ACM, 2006.
- [44] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8(9):e73970, 2013.
- [45] Jordan Viard and Matthieu Latapy. Identifying roles in an ip network with temporal and structural density. In *6th IEEE International Workshop on Network Science for Communication Networks (NetSciCom 2014)*. In *INFOCOM IEEE Conference on Computer Communications Workshops*, pages 801–806. IEEE, 2014.
- [46] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *2nd ACM SIGCOMM Workshop on Online Social Networks (WOSN 2009)*, pages 37–42. ACM, 2009.