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

Yannick Léo
ENS Lyon, DANTE/INRIA
LIP UMR CNRS 5668
Université de Lyon
yannick.leo@inria.fr

Christophe Crespelle
UCBL, DANTE/INRIA
LIP UMR CNRS 5668
Université de Lyon
christophe.crespelle@inria.fr

Eric Fleury
ENS Lyon, DANTE/INRIA
LIP UMR CNRS 5668
Université de Lyon
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 great number of studies on these objects start by aggregating the data on disjoint time windows of length \(\Delta\) 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 \(\Delta\) on the obtained graph series. We address the fundamental question of knowing whether a series of graphs formed using a given \(\Delta\) faithfully describes the original link stream. We answer the question by showing that such dynamic networks exhibit a threshold for \(\Delta\), 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.

CCS Concepts

- Networks → Network dynamics;

Keywords

aggregation; time scale; link stream

1. INTRODUCTION

*This work is supported in part by the French National Research Agency contract CODDDE ANR-13-CORD-0017-01.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CoNEXT’15 December 01 - 04, 2015, Heidelberg, Germany
© 2015 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-3412-9/15/12 ...$15.00
DOI: http://dx.doi.org/10.1145/2716281.2836114

Many real-world dynamic networks are naturally given in the form of a finite collection \(L\) of triplets \((u, v, t)\), which we call a link stream, where \(u, v \in V\) are two nodes of the network and \(t \in [0, T]\) is a timestamp, with the meaning that nodes \(u\) and \(v\) have a link between them at time \(t\). Depending on the context, this link 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 [1, 2, 3, 4, 7, 8, 9] 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 and forming the graph \(G_{[a, b]}\) with all edges \(u, v\) such that there exists a triplet \((u, v, t) \in L\) 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 and have different lengths. In both cases, once the series is obtained, all analyses are conducted on it instead of the original link stream.

Aggregation offers some clear benefits: i) changing the scale of study can give a more relevant view of the dynamics and ii) obtaining graphs allows to use the rich set of notions developed in graph theory to analyze the dynamic network. But it also raises some important concerns, as the length chosen for the aggregation window has a strong impact on the properties of the aggregated graph series and therefore on the conclusions derived from analysis [4, 7]. Pushing further, it is not even clear whether this series faithfully describes the original link stream. Indeed, the 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, 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, 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. Then, one can ask whether the graph series obtained for a given aggregation period is a faithful representation of the original dynamic network? This is precisely the question we address here. Our results. We show that for many dynamic networks, the length \(\Delta\) of the window chosen for aggregating the network into a graph series exhibits a threshold, which is proper to each network, and beyond which the propagation properties of the graph series obtained from aggregation show evidence of alteration, while these properties are mostly preserved below this threshold. We design a method, called the occupancy method, in order to determine this threshold, which we call the saturation scale and denote \(\gamma\). We apply and validate the occupancy method on various real-world datasets, as well as on synthetic dynamic networks.

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 been incorporated into any automatic tool for analyzing dynamic networks.

Related works. In [4], the authors show that significant characteristics of the dynamics of a phone-call network appear at different time scales of analysis, which implies that one should use the broad spectrum of possible scales in order to observe these properties. It is very interesting to note that they do so using only statistics that are meaningful regardless of the loss of information due to aggregation, which is precisely what we are interested in here. It should be clear that the aggregation period \(\gamma\) we determine is not intended to reveal the key properties of the dynamics. Making statistics at other time scales may reveal interesting facts that are invisible otherwise. Nevertheless, for aggregation scales greater than \(\gamma\), one should consider only statistics that are not sensitive to the loss of information induced by aggregation, excluding all statistics based on propagation properties of the dynamics.

The motivations of [7] are closer to ours since they are also concerned with the loss of information. But they appreciate it only on properties of the snapshots, which are by nature unable to capture the propagation properties of the dynamics. Moreover, compared to their approach, which relies on a trade-off between two oppositely varying metrics, a key benefit of our work is to exhibit a natural change in the way the dynamic network responds to aggregation beyond a certain time scale.

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 3). The only restriction which is meaningful here is that links are punctual events and therefore have no duration. The aggregation of link stream \(L\) consists in choosing a time period \(\Delta\) such that \(T = K \Delta\) for some integer \(K \geq 1\) and forming the graph series \(G_\Delta = (G_k)_{1 \leq k \leq K}\) defined by \(G_k = (V,E_k)\) with \(E_k = \{uv \mid \exists (u,v,t) \in L, (k-1)\Delta \leq t < k\Delta\}\). Note that \(V\) is the same for all graphs of all aggregated series: it is the set of nodes involved in the link stream \(L\). A temporal path \(P\), 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. More formally, in a link stream \(L\) (resp. a series of graphs \(G\)) 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 \(\forall i, (u_i, v_i, t_i) \in L\) (resp. \(\forall i, u_i,v_i \in E(G_k)\)) and \(\forall i > 1, u_i = v_{i-1}\) and \(\forall i,j, t_i < t_j\) (resp. \(k_i < k_j\)). 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 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, is the number \(l\) of edges in the path, which we call number of hops in the rest of the paper and denote \(hops(P)\). The second one is the duration, that is \(t_l - t_1\) for a link stream and \(k_l - k_1 + 1\) for a graph series\(^1\), we denote it \(time(P)\).

\(^1\)Because each \(k\) does not index an instant as in a link
The main difficulty in determining the maximum non-altering aggregation period $\gamma$ is that all the classical properties of the graph series, such as density, degrees, connectedness, distances in time and in hops, vary smoothly from one extremal value to another when the aggregation period spans its range of variations (see Fig. 1), therefore giving no hint on any scale at which a change occurs in the way the dynamic network responds to aggregation. Our main contribution is to exhibit a parameter able to reveal such a qualitative change. We now give the definitions necessary to describe our method and we illustrate it on a sample real-world network, the Irvine network introduced in Section 3.

**Definition 1** (Trip and minimal trip). A trip is a quadruplet $(u, v, t_{dep}, t_{arr})$ such that there exists a stream but an interval of time with a duration.

Note that, by definition, in a graph series, we always have $\text{hops}(P) \leq \text{time}(P)$ for any temporal path $P$.

In the following, we also use two notions of distance at time $t$ between two nodes $u, v$. Both definitions 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 $t_{arr} - t$, 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. By convention, $d_{hops}(u, v, t) = +\infty$ when $d_{time}(u, v, t) = +\infty$.

**2. THE OCCUPANCY METHOD**

In order to determine the saturation scale $\gamma$, we make the aggregation period $\Delta$ vary from its minimal value, the resolution of timestamps, until the whole length $T$ of study of the network. For each value of $\Delta$ we form the aggregated graph series $G_{\Delta}$ for which we compute the set of minimal trips and their occupancy rates. Then, for each $\Delta$, we plot the distribution of occupancy rates of all the minimal trips in $G_{\Delta}$ (considering all pairs of nodes and all time intervals), see Fig. 2.

Necessarily, when $\Delta$ is close to its minimal value, provided that the resolution of 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 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

**Figure 2:** Inverse Cumulative Distributions (ICD) of the occupancy rates (x-axis) of the minimal trips of the aggregated series $G_{\Delta}$ for several values of the aggregation period $\Delta$ in the range $[1, T]$.

**Figure 3:** $M - K$ proximity (y-axis) of the distributions of occupancy rates with the uniform density distribution according to $\Delta$ (x-axis).

A trip $(u, v, t_{dep}, t_{arr})$ is minimal if there exists no trip from $u$ to $v$ on an interval $[t'_{dep}, t'_{arr}]$ strictly included in $[t_{dep}, t_{arr}]$ ($[t'_{dep}, t'_{arr}] \subset [t_{dep}, t_{arr}]$).

In the following, a temporal path on two hops is called a transition, and a shortest transition if it is also a minimal trip.

**Definition 2** (Occupancy Rate). For a graph series $G$ and a temporal path $P$ in $G$, the occupancy rate of path $P$, denoted $\text{occ}(P)$, is defined as $\text{occ}(P) = \text{hops}(P) / \text{time}(P)$. 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.

The rational behind the occupancy rate $\text{occ}(P)$ 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. In particular, note that since $\text{hops}(P) \leq \text{time}(P)$ then we always have $\text{occ}(P) \leq 1$.

In order to determine the saturation scale $\gamma$, we make the aggregation period $\Delta$ vary from its minimal value, the resolution of timestamps, until the whole length $T$ of study of the network. For each value of $\Delta$ we form the aggregated graph series $G_{\Delta}$ for which we compute the set of minimal trips and their occupancy rates. Then, for each $\Delta$, we plot the distribution of occupancy rates of all the minimal trips in $G_{\Delta}$ (considering all pairs of nodes and all time intervals), see Fig. 2.

Necessarily, when $\Delta$ is close to its minimal value, provided that the resolution of 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 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 contact and their occupation rate is 1. Then, the distribution is again concentrated, this time on the value 1. What is remarkable here is that the distribution changes from values concentrated near 0 to values 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 \( \gamma \) returned by our method is precisely the value of \( \Delta \) for which the distribution is maximally stretched on the interval \([0,1]\) (curve marked with green squares on Fig. 2). In order to detect it, we compute for each value of \( \Delta \) in the total range of variation, the \( M-K \) distance \( d(\Delta) \) between the distribution obtained for \( \Delta \) 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 \( 1/2 - d(\Delta) \), on Fig. 3. This confirms the observation made above: the proximity first increases and then decreases. The saturation scale \( \gamma \) returned by our method is the value of \( \Delta \) realizing the maximum of the proximity. Of course, one may think of many other ways to determine which \( \Delta \) gives the maximum stretch of the distribution. We actually tried several of them (not presented here for lack of space) and we chose 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 \( \gamma \). A very low occupancy rate for most minimal trips of the series denotes a lack of data in each aggregation window, which means that the link stream is aggregated with a time scale much finer than the actual activity of the network. In this situation, the information contained in the link stream is mainly preserved in the graph series. On the opposite, a very high occupancy rate for most of the minimal trips goes along with a loss of information. Indeed, this indicates that, at each time in the graph series, there is a high probability to find a next hop to perform on any given shortest path, meaning that a high proportion of nodes are involved in a high number of edges in their snapshot. Then, at the same time, the information on the existence of a transition in the original link stream using a given couple of these edges is lost, which constitutes the more crucial loss of information in the aggregation process.

In the first phase of variation of the aggregation period, below \( \gamma \), 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 data in the aggregation windows without significant loss of information. On the opposite, in the second phase, beyond \( \gamma \), 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 \( \gamma \) appears as a separation between the range of values, below \( \gamma \), where the aggregated graph series still faithfully describes the original link stream and the range of values, beyond \( \gamma \), where aggregation alters the properties of propagation of the original link stream.

3. 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 [6] is made of 48 000 messages sent between the 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 [9] 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 [3] contains the 15 951 individual emails sent between a group of 150 employees of the Enron company during year 2001; and finally, the Manufacturing emails

![Figure 4: Inverse Cumulative Distributions (ICD) of the occupancy rates (x-axis) of the minimal trips of the aggregated series \( G_\Delta \) for several values of the aggregation period \( \Delta \) in the range \([1,T]\).](image-url)
network [5] contains the 82,894 internal emails between 153 employees of a mid-sized manufacturing company over a period of 8 months.

We applied our method on each of these four datasets. The distributions of occupancy rates of the minimal trips in the aggregated graph series are given on Fig. 2 for Irvine and Fig. 4 for the three other networks, their $M-K$ proximity with the uniform density distribution is given on Fig. 3 and 5. One can see that the observations made on the Irvine network in Section 2, on which our method is based, hold for all the four datasets. 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 $\gamma$ 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 online social networks 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, 12 hours and 18 hours respectively. As one can see, the average activity has a strong influence but is not the only parameter affecting the saturation scale. We further investigate this relationship in the next section.

4. 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 ($N << T$) to each pair of the $n = 100$ nodes of the network and randomly choosing each of their timestamps between 0 and $T = 100,000 \text{ s}$. We make the value of $N$ vary from 10 to 100 and for each of these values, we compute the aggregation period returned by the occupancy method. Results are given in Fig. 6 left, which shows $\gamma$ (y-axis) as a function of the average inter-contact time of one node, that is $T/(n-1)N$. 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 3, 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 activities? Or does it favor one of them? To answer this question we generated *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 $T_2$.

Fig. 6 right gives the saturation scale $\gamma$ 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
\(\gamma\) for the high-activity mode (for \(\rho = 0\%\)) to the one, much larger, of 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).

5. VALIDATION

In this section we quantify the amount of information which is lost when one aggregates the network with a given aggregation period \(\Delta\). 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 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) 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 with 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. 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.

Fig. 7 left depicts the proportion of lost transitions as a function of the aggregation period \(\Delta\), 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, 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 \(\gamma\).

On the other hand, some of the lost shortest transitions can be replaced by some others a little bit longer, therefore limiting the actual impact of this loss on the possibilities of propagation in the aggregated series. For this reason, we also use a measure of loss which is based on the elongation of minimal trips in the aggregated series \(G_\Delta\) compared to the original link stream \(L\). The elongation factor of a minimal trip \(P = (u, v, t_u, t_v)\) of \(G_\Delta\) is defined as the ratio \((t_v - t_u)\cdot\Delta/time\_L(P)\), where \(time\_L(P) = \min\{t'_v - t'_u | (u, v, t'_u, t'_v)\}\) is a minimal trip of \(L\) and \(t'_u, t'_v\) \((t_u - 1)\cdot\Delta, t_v, \Delta\)\.

Fig. 7 right gives the mean elongation factor (y-axis) of all minimal trips of the series aggregated with period \(\Delta\) (x-axis), for the Irvine network. When \(\Delta\) increases,
the elongation factor of minimal trips first stays very close to 1 before it suddenly raises when the aggregation period reaches values around the saturation scale $\gamma$. 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.

6. CONCLUSION

We showed that there exists a threshold, called the saturation scale $\gamma$, for the aggregation period of a dynamic network 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 $\gamma$ to perform analyses that depend on these properties. In addition, we designed a fully automatic and parameter-free method to determine the value of $\gamma$ for an arbitrary link stream.

7. REFERENCES


