Meaningful Selection of Temporal Resolution for Dynamic Networks

Rajmonda Sulo Tanya Berger-Wolf Robert Grossman

University of Illinois at Chicago
{rsulo1, tanyabw, grossman} @ uic.edu

May 28, 2010

Abstract

The understanding of dynamics of data streams is greatly affected by the choice of temporal resolution at which the data are discretized, aggregated, and analyzed. Our paper focuses explicitly on data streams represented as dynamic networks. We propose a framework for identifying meaningful resolution levels that best reveal critical changes in the network structure, by balancing the reduction of noise with the loss of information. We demonstrate the applicability of our approach by analyzing various network statistics of both synthetic and real dynamic networks and using those to detect important events and changes in dynamic network structure.

1. Introduction

Massive streams of complex data are collected in various domains such as sociology (social interactions, opinions), communication networks (IP data, cell phone and email communication), and biology (protein data). The high volume and temporal detail at which the data are collected poses a challenge in terms of both manageability and meaningful interpretation.

The need for determining the temporal resolution appropriate for meaningful analysis of changing data is even more evident in data streams represented as networks. Networks are graphs with nodes representing entities, such as people, computers, or proteins, and the edges representing interactions between pairs of entities, such as, sending an email or meeting a person, routing a packet between IP addresses, or proteins participating in the same regulatory process. A further abstraction of the network concept is the dynamic network where each edge has a time label. The network can therefore represent complex temporal structures. At the same time, it is very sensitive to the temporal resolution of the underlying data, that is, the time window the interactions are aggregated into a network. As our analysis will illustrate, too fine or too coarse of a temporal resolution will either disguise or smooth out important temporal dynamics of the network and the structure of the underlying interactions. A typical way dynamic networks are analyzed is to observe over time different network statistics and to develop insights from the corresponding time series. The first step of such analysis is typically the discretization of the dataset. There is a rich body of literature that recognizes the sensitivity of good analytical tools to the size of the discretization step [6, 9, 11, 15, 18]. If a measure is computed over a small window size the corresponding time series will have a lot of noise, meaningful connections in time might be lost and the data are not easy to manage. For example, Figure 1 illustrates amount of noise present when a network statistic on IP stream data is aggregated at very fine temporal resolution.

On the other extreme, if we compute the statistic over the entire duration of the dataset, while we remove a lot of noise from data, we also smooth out a lot of useful and critical information needed to understand the temporal

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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MLG '10, July 24-25, 2010 Washington, DC, USA

Copyright ©2010 ACM 978-1-4503-0214-2/10/07... $10.00

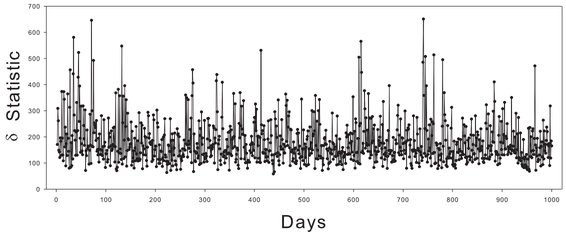

Figure 1. A time series of a network statistic on IP data aggregated on 5 minute intervals.

nature of the dataset.

Even though the discretization step seems to be a critical step to ensure the success of the analytical tools that we develop, too often this step is ignored or there is no systematic process that justifies the choice of a particular window size [1, 6].

This paper proposes an empirical framework for identifying the meaningful temporal resolution for dynamic network analysis. We propose a set of balancing criteria to automatically infer the appropriate level of data aggregation for analyzing different statistics of a network that changes over time. To quantify the terms ”appropriate” and ”meaningful”, we empirically evaluate our framework in some typical dynamic network analysis scenaria: event detection, community inference, and change detection. Although, currently there is no theoretical framework that addresses this problem, our paper's results offer strong empirical evidence of its usefulness as a first step in analysing a range of complex dynamic networks. The paper is organized as follows: Section 2 defines the basic concepts on our proposed framework. Section 3 summarizes related work. Section 4 describes the methodology. Section 5 presents the main results of our analysis followed by a discussion of their implications. Finally, Section 6 gives the conclusions.

2. Problem Definition

While the problem of finding the right temporal aggregation level is common for different kinds of stream data, we focus on streams of dynamic interactions or dynamic networks. Dynamic networks have a rich temporal structure [3, 10, 13, 14]. Given oversampled observations from an underlying dynamic network, the goal in this paper is to find the right temporal resolution at which meaningful information about the structure of the network is revealed, while the oversampling noise inherent in the data is removed. We propose to achieve the goal of identifying the right temporal resolution by finding the temporal window that optimizes the trade-off between a measure of noise and a measure of information content in the data. We now formally define dynamic networks and propose computational measures of noise and information.

2.1 Dynamic Network

Given a list of time labeled interactions sampled at regular intervals and a fixed window size w, we define a dynamic network the following way:

Definition 1 *A dynamic network at resolution w is a time-series of labeled graphs $\mathcal{G}(w) = [G_1, G_2, ..., G_t, ..., G_T]$, where $G_t = (V_t, E_t)$ is the graph of edges taking place over time interval $[t, t + w]$. $V_t \subseteq V$ is the set of vertices observed over time interval $[t, t + w]$, and an edge $e = (v_1, v_2)$ exists in E_t if v_1 and v_2 were observed interacting in that time period.*

Figure 2 illustrates how temporal edges are aggregated to form a dynamic network.

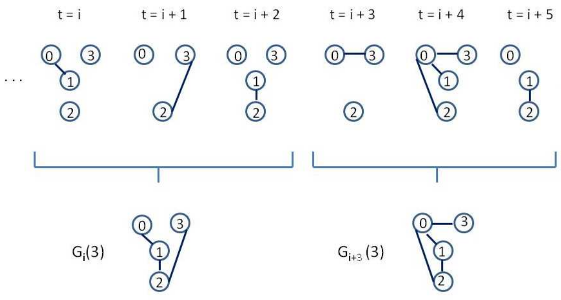

Figure 2. Example of a dynamic network aggregated over window size $w = 3$

2.2 Trade-off measures

Given a fixed window of aggregation w and the corresponding $\mathcal{G}(w)$, we compute different statistics on the dynamic graph and monitor their time series. Let f represent such a statistic and $\mathcal{F}_w$ its corresponding time-series:

$$\mathcal{F}_w(\mathcal{G}) = [f(G_1), f(G_2), ..., f(G_t), ..., f(G_T)]$$

If w is too large, the aggregated network over w will not capture a lot of the critical temporal information such as edge concurrency and time propagating paths [3, 10]. As a consequence, time series $\mathcal{F}_w$ will not correctly represent structural variation on the network.

On the other hand, if w is too small, the dynamic network is aggregated over insufficient time, over which interesting phenomena such as the formation of a giant component or the disappearance of a cluster might not be evident. It is therefore of great interest to determine what is the right window of aggregation that will balance between the loss of temporal structural information and fluctuations that clutter and obscure what is relevant in a structural change. We propose the use of the following measures to capture this compromise:

Variance: Let $V(\mathcal{F}_w)$ be the variance of $\mathcal{F}_w$:

$$V(\mathcal{F}_w) = \frac{1}{T}\sum_{i=1}^{T}[f_w(G_i) - \mu(f_w)]^2,$$

where $\mu(\mathcal{F}_w) = \frac{1}{T}\sum_{i=1}^{T} f_w(G_i)$.
We can think of the $V(\mathcal{F}_w)$ as a measure of noise present in $\mathcal{F}_w$. As we vary w, the value of $V(\mathcal{F}_w)$ changes. Large values of variance indicate $\mathcal{F}_w$ changes drastically in time making it hard to distinguish between the occurrence of a meaningful change and a noise effect. On the other hand, small values of variance indicate $\mathcal{F}_w$ is smooth and a lot of the noise is removed.

Compression Ratio: Let u represent the length of the string representation of $\mathcal{F}_w(\mathcal{G})$, and c represent the length of the compressed representation of $\mathcal{F}_w(\mathcal{G})$ produced by a data compression algorithm. Let $R(\mathcal{F}_w)$ be the compression ratio of $\mathcal{F}_w(\mathcal{G})$ defined as:

$$R(\mathcal{F}_w) = \frac{u}{c}$$

$R(\mathcal{F}_w)$ is one of the ways to represents information encoded in $\mathcal{F}_w$. A small value of $R(\mathcal{F}_w)$ represents a lot of randomness or noise in signal $\mathcal{F}_w$, while a large value of $R(\mathcal{F}_w)$ comes as a result of redundancies in $\mathcal{F}_w$. In the information theoretic sense, redundancies correspond to low entropy and low entropy corresponds to high information.

Intuitively, $V(\mathcal{F}_w)$ and $R(\mathcal{F}_w)$ have opposite behaviours with respect to window size which allows us to formulate the process of finding the range of appropriate discretization windows as a minimization problem.

3. Related Work

The problem of identifying the right resolution for analysis of data streams is a very broad problem and covers many research areas such as signal processing [7, 16], discretization of continuous variables [11], time series discretization [15, 18], model granularity [9]. Usually the approach involves a trade-off between loss of information and reduction of noise. While this literature offers a solid foundation on discretization analysis, it does not explicitly address datasets that are represented as networks and, furthermore, it does not address the dynamic nature of these networks. The focus of our method is to find the right aggregation levels for different graph theoretic metrics computed on a dynamic network. Also, it is important to note that in our analysis, aggregation happens at the level of the dynamic network rather than at the level of time series. This preserves a lot of the rich structure encoded in the network

While dynamic networks and their rich temporal structure have gained a lot of interest and motivated a series of informative papers [3, 5, 10, 12, 13, 22], there is no clear framework on how to aggregate temporal graphs in a meaningful way.

Analysis presented by Eagle [17] and Sun et. al. [20] deal with this problem in a more explicit way. Eagle illustrates the effect of the aggregation window in understanding the periodic dynamics of the Reality Mining dataset [6]. He recommends Fourier Transform analysis and auto-correlation analysis of time series. While these techniques have been successfully used to understand stationary time series, their application to time series originating from highly dynamic and complex networks might not be appropriate.

The approach in [20] is to group similar graphs into one time segment based on the Minimum Description Length principle. The contribution of [20] is to use drastic changes in the time series of compression levels to segment the timeline of the temporal network. Our approach builds on this idea, but instead of analyzing the compression levels of the graph stream, we analyze compression levels of the time series of different metrics computed on the graph. Also, we use the variance of such time series as an explicit trade off measure to ensure a balanced selection of window size.

4. Methodology

We propose algorithm TWIN : **T**emporal **W**indow **I**n **N**etworks, to empirically answer the question posed in Section 2.

Given a list of temporal edges, and a graph metric, TWIN generates time series of graphs at different resolutions. It then computes the time series of the given metric, together with its compression ratio and its variance. Finally, the algorithm analyzes the compression ratio and variance as functions of window size and selects the window size for which compression ratio and variance are close or equal to each other.

TWIN Algorithm:
Input: List of time labeled edges.
Fixed measure $f : G \rightarrow \mathbf{R}$, where G represents a graph
User define "goodness measure" $\gamma \geq 0$
Maximum window size analyzed $w_{max} \geq 1$
Output: List of appropriate window sizes w

```
for $w = 1$ to $w_{max}$ do
    Compute the time series of graphs $\mathcal{G}(w) : [G_1, G_2, ..., G_t, ..., G_T]$
    Compute the time series $\mathcal{F}_w : [f(G_1), f(G_2), ..., f(G_t), ..., f(G_T)]$
    if $V(\mathcal{F}_w) - R(\mathcal{F}_w) < \gamma$ then
        Output w
    end if
end for
```

We apply TWIN to a range of metrics and dynamic networks both real and synthetic. Following is the description of the measures and datasets used in our analysis.

4.1 Network Measures

Density: the proportion of the number of edges $|E|$ present in a graph relative to the possible number of edges $\binom{|V|}{2}$. Graphs with few edges (typically linear in the number of nodes) are called *sparse* and those with almost all possible edges present (or quadratic in the number of nodes) are called *dense*.

Number of Connected Components: a connected component is a set of nodes mutually reachable by paths in the graph.

Size of Giant Component: the size of the largest connected component.

Geodesic between a pair of nodes: the path with the smallest number of edges.

Eccentricity of a node: the greatest geodesic between the node and any other node in the graph.

Diameter: the maximum eccentricity of any node in the graph.

Radius: the minimum eccentricity of any node in the graph.

Average Path Length:the length of the average geodesic between any pair of nodes.

Clustering Coefficient: the number of triangles over the number of possible triangles in the graph [14].

Clique Number: the size of the largest clique.

Spectral Gap: the difference between the firs and the second eigenvalue of the Laplacian of the graph [4]. *The Laplacian* of graph $G(V, E)$ is define as

$$L = D - A$$

where $D = diag(d_1, ..., d_n)$ is the degree matrix of G and A is the adjacency matrix. Spectral gap of L is known to capture the connectivity properties of the graph [4] .

Compression Ratio of the Graph: the ratio of the length of string representation of the network and the length of the compressed representation of network. Instead of computing the time series of a graph theoretic measure and then the compression ratio of that time series, we were interested in analyzing directly the raw compression of the graph in time. This gives us the opportunity to evaluate any loss of information when we move our analysis from the actual graph to a measure on the graph.

The above list is not exhaustive of all measures used to analyze network structural properties. Rather, the goal is to illustrate the effect of aggregation window on the behaviour of a wide range of measures each revealing unique and interesting properties of the network. Analysis in this paper can easily be applied to other network metrics not mentioned here.

4.2 Datasets

Barabasi and Albert Model is a random graph generating algorithm designed to simulate a scale free network [2]. At each iteration **m** new nodes are added by following the preferential attachment property. For our analysis, we created a dataset of 1000 nodes and used a range of values for parameter m.

Enron Email is a publicly available database of e-mails sent between employees of the Enron corporation [1]. Each email address represents a vertex and an email exchange represents an edge. Timestamps were extracted from message headers for each day of e-mail activities. We are using a cleaner version of the dataset covering email exchanges from October 1998 to February 2003.

Reality Mining network consists of social interactions among 90 MIT students and faculty over a nine month period [6]. The dataset is designed based on the idea that spatial proximity between people implies a social interaction. Participants are equipped with Nokia 6600 smart phones and an edge between two participants exists if a bluetooth connection is recorded.

Haggle Infocomm network consists of social interactions among attendees at an IEEE Infocomm conference [19]. There were 41 participants and the duration of the conference was 4 days.

Grevy's Zebra The Grevy's dataset consists of social interactions among Grevy's zebra (Equus Grevyi) recorded by biologists over the period of June through August of 2002 in the Laikipia region of Kenya [8]. Predetermined census loops were driven approximately twice per week and individual zebra were identified by unique stripe patterns. Upon a sighting, the zebra's GPS location was taken. In the resulting dynamic network, each node represents an individual zebra and two animals are interacting if thir GPS locations are in close proximity. The dataset consists of 28 zebras.

Plains Zebra Plains zebra (Equus Burchelli) are another species of zebra. The data are collected in a similar fashion to that of Grevy's dataset. The data were collected through visual scans (approximately once per day) over a period of several months [21]. Each entity is a Plains zebra and the interactions represent spatial proximity as determined by ecologist based on GPS locations. The dataset represents data from observations of 282 individuals from July 2003 to September 2006.

5. Results

In this section, we first demonstrate how crucial the choice of the temporal resolution is in the analysis of complex dynamic networks. We then evaluate the results of our heuristic algorithm across datasets coming from different domains (simulations, sociology, communication, biology) in order to show the breadth of the applicability of the results. Through simulation and ground truth that comes from domain knowledge experts, we show that our algorithm produces meaningful and consistent levels of temporal aggregation. Finally we compare our method with GraphScope, in the context of event detection, and FFT analysis for identifying the scale of temporal dynamics and demonstrate that our results are equally robust and some times better when detecting events and patterns in dynamic networks.

Figure 3 shows the plot of the compression ratio R and variance V of each time series as functions of w generated for the radius of the Enron network.

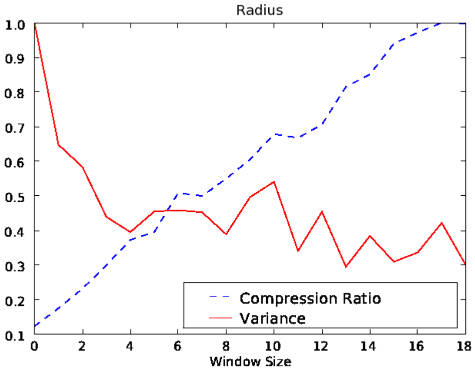

Figure 3. Compression ratio and variance of Radius with respect to window size for the Enron dataset.

Note that R increases as w increases, while V (overall) decreases. The plot suggests that an appropriate window for analyzing the radius of the Enron network is in the range 4-7 days, where variance is relatively small and compression is relatively high.

Figures 4, 5, and 6 display the time series of radius for the Enron dataset at $w = 1$ (high resolution), $w = 5$ (resolution within the appropriate range 4-7 days) and $w = 12$ (coarse resolution), correspondingly.

As seen in Figure 4, the drastic variations of the radius time series at 1 day aggregation make it impractical to understand any pattern on the dynamics of email exchanges of the Enron employees.

As we increase the aggregation window to 5 days (Figure 5), some peaks corresponding to important events in the lifetime of the Enron company become clear. For example, the peak at timestep 950 (Event 1) represents the time when Karl Rove sold off his energy stocks, the peak at timestep 1100 (Event 2) represents the unsuccessful attempt of Dynegy to acquire the bankrupt Enron, while the peak at timestep 1150 (Event 3) represents the resignation of Enron's CEO in January 2002, and the beginning of FBI investigation.

When we aggregate the dynamic network beyond the 4-7 day range as in Figure 6, we notice that the time series looks smoother, but at the same time, some critical temporal events are lost. For example, the collapse of the Dynegy deal represented by a sharp peak at aggregation level 5 is not identifiable anymore.

Similar behaviours for measures computed on the Barabasi-Albert, Reality Mining and Haggle datasets are illustrated on Figures 7, 8, and 9. Also, summaries of the appropriate window ranges for the Barabasi-

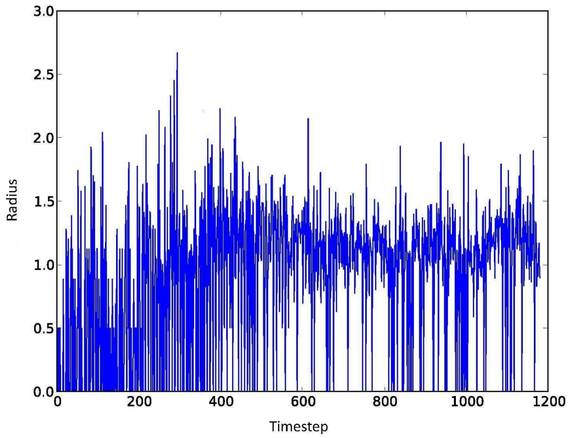

Figure 4. Radius time series for the Enron dataset, w = 1 day.

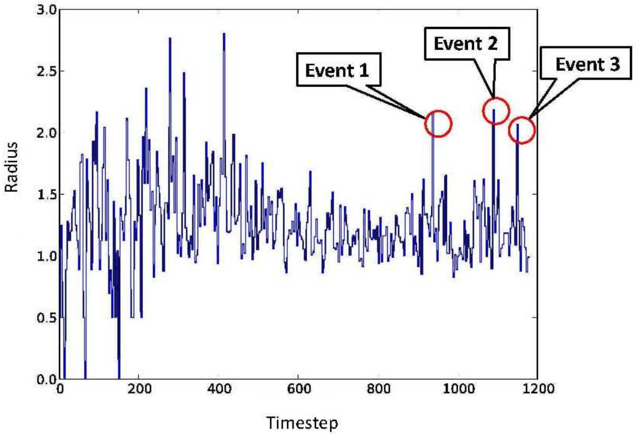

Figure 5. Radius time series for the Enron dataset, w = 5 days.

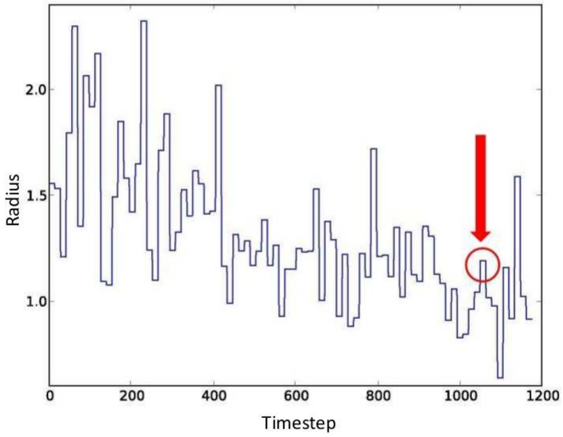

Figure 6. Radius time series for the Enron dataset, w = 12 days.

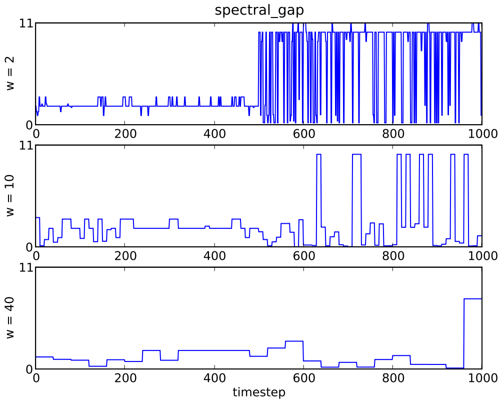

Figure 7. Spectral gap of the Barabasi-Albert dataset in three different resolutions, w=2, w=10, w=40. Identified appropriate window size is w=10.

Albert, Enron, Reality Mining, Haggle, Grevy's and Plains datasets are given in Tables 1 and 2.

Synthetic Dynamic Networks

We chose to use the Barabasi and Albert model to simulate the growth of a dynamic network because the preferential attachment model generates networks with properties found in many realistic networks (although, clearly, they do not capture the entire complexity of real networks).

The firs column in Table 1 represents the appropriate window sizes when the network growth rate is kept the same throughout the network evolution. The second and third columns represent the appropriate window sizes when the growth rate increases after a certain iteration by 5 times and 10 times, respectively. The reason we change the growth rate for the Barabasi-Albert dataset is to simulate a change in the synthetic dataset to ensure that our method detects the embedded change and responds consistently.

An important observation on the ranges of aggregation for the different network statistics is that they are not the same. This illustrates the essential property of dynamic networks where interesting behaviour happens

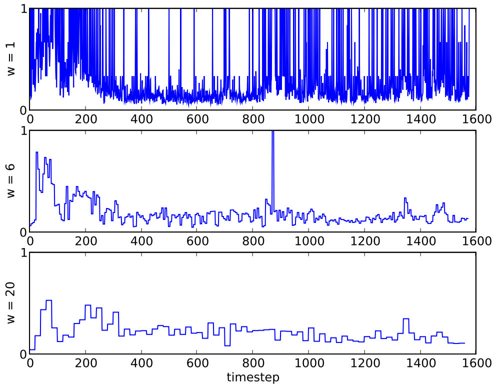

Figure 8. Density of the Reality Mining dataset in three different resolutions, w=1, w=6, w=20. Identified appropriate window size is w=6.

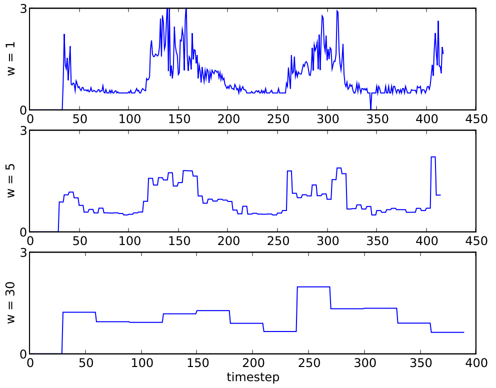

Figure 9. Average path length of the Haggle dataset in three different resolutions, w=1, w=5, w=30. Identified appropriate window size is w=5

at various resolutions. Moreover, it reveals the problem with techniques that choose only one aggregation level when analyzing complex dynamic datasets.

In addition, note that for some measure-network combinations, no range of appropriate aggregation windows was found by our method. These results suggest that for some measure-network combinations, the measure is not revealing of the change in the temporal structure of the network. Our approach allows us not only to identify the appropriate aggregation window, but at the same time, it allows us to identify the measures that are more suitable in revealing the essential dynamics of the network.

It is important to mention that identifying an appropriate range of temporal resolutions does not mean resolutions beyond such range are not necessarily suitable for network dynamic analysis. Drastic or substantial structural changes are not smoothed out even at very coarse resolution. Our approach, instead, offers a range of window sizes at which we can guarantee that the right balance between minimization of jitter and maximization of temporal information is achieved.

The results imply that different types of changes in the dynamic network affect the choice of the appropriate level of aggregation. In the case of the Barabasi-Albert network, we notice that when the rate of growth changes, so does the range of resolutions required to best detect this change.

Real Dynamic Networks where we have ground truth

The behaviour of the Grevy's and Plains zebras have been extensively studied by biologists at Princeton University. The two species have similar patterns of social organizations, yet the data collected about them was sampled at different rates. As seen in table 2, TWIN algorithm outputs 3 days as the appropriate window of aggregation for both zebra networks. By outputing the same resolution level, the algorithm correctly infers that the underlying social network of the two species of zebra is similar, regardles of the initial sampling rate.

Event Detection: Comparison with Graphscope Algorithm [20] and FFT analysis

GraphScope analysis on the Enron dataset discretizes the time line on segments that vary from 2 weeks to 6 weeks, during the eventful period of November 2001-May 2002. Some of the major events are captured using this segmentation. There are however, several important events that get smoothed out and can not be spotted when analyzing the time series aggregated at such coarse levels (Figure 10).

Since GraphScope focuses on variations of graph compression levels, it is the magnitude of change in the graph structure that drives the timeline segmentation. TWIN analyzes the regularity of compression levels of different metrics on the graph, and therefore, it is the rate of change, not the magnitude, that will have the most affect in the aggregation.

Furthermore, since the rate of change in a temporal graph does not follow a simple pattern, using periodicity to determine the right aggregation levels might not

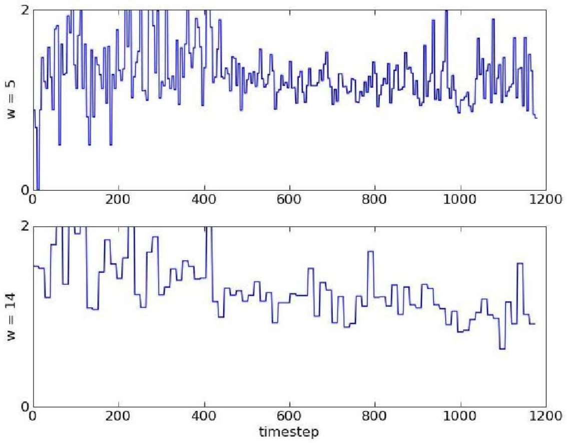

Figure 10. Average Path length for the Enron dataset at w=5 days and w=14 days

always be appropriate.

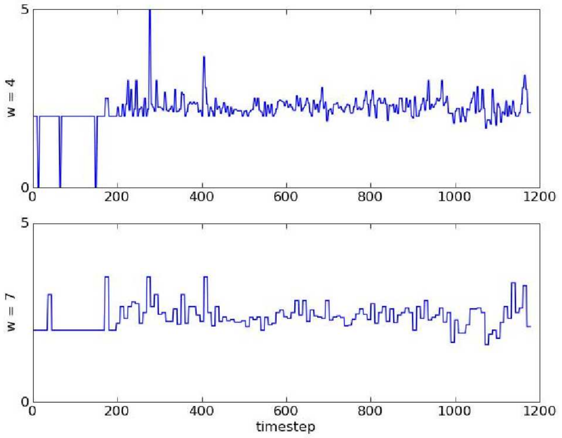

Figure 11. Clique number for the Enron dataset at w=4 days and w=7 days

Figure 11 shows the clique number for the Enron dataset when the underlying temporal graph is aggregated at 4 days, as recommended by our method, and 7 days, following the predominant cycle identified by the FFT analysis. We notice that at 7 days important peaks of the signal are not as easy to identify or completely disappear.

6. Conclusions

The dynamic network abstraction of data streams has allowed for rich structural representation, but at the same time has introduced many challenges that include understanding how the network structure changes over time. A critical aspect of this analysis is the level of resolution at which the network is aggregated and analyzed. There has been relatively little work done on systematically identifying a meaningful window size for aggregating and analyzing a dynamic network. This paper focuses precisely on this problem and offers the following contributions:

- We give a quantitative trade-off criterion defining the appropriate window size for discretizing a dynamic networks. By choosing windows of aggregation that balance between the minimization of noise and loss of temporal structural information, our approach offers a systematic framework to empirically discover interesting network dynamics that would otherwise be lost.

- The framework presented here does not restrict the analysis to one network statistic. We show that different aggregation levels are appropriate for different network measures. Not only this is not a drawback of our method, but it is a desirable feature , since each measure reveals distinct properties of the network. Furthermore, it is another illustration of the fact that interesting network behaviour happens at various temporal resolutions and our method automatically reveals those interesting temporal scales.

- We propose a simple algorithm that produces consistent and meaningful results for datasets arising from different domains and different underlying network dynamics.

- Finally, our results demonstrate that changes of different types and scales need to be analyzed at different resolutions. Whether those are implicit changes, in the case of the real datasets we analyzed, or synthetic changes, in the case of the Barabasi-Albert model, it is clear that one level of aggregation does not fit all.

References

[1] J. I. Adibi. Enron email dataset. Downloaded from http://www.isi.edu/ adibi/Enron/Enron.htm.

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[3] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *Proc. of the 12th ACM SIGKDD international conference on Knowledge*

Table 1. Appropriate window ranges for the Barabasi dataset

Measure	Growth Rate		
	Constant	5 x	10 x
Density	5-15	1-2	1-2
Number of connected components	-	-	-
Size of the giant component	-	-	-
Diameter	10-40	5-25	15 - 50
Radius	10-40	10-25	15 - 50
Average Path Length	10-60	35-40	40-50
Clustering Coefficien	2 - 25	10-15	5-10
Clique Number	-	-	-
Spectral Gap	-	-	-
Graph Compression Ratio	-	-	10-30

Table 2. Appropriate window ranges for the Enron, Reality Mining, Haggle, Grevy's and Plains datasets

Measure	Enron Dataset	Reality Mining	Haggle	Grevy's	Plains
Density	4-5	5-10	-	2-3	3-4
Number of connected components	3-5	10-14	35-50	-	-
Size of the giant component	-	45-55	38-58	-	1-4
Diameter	6-8	15-25	5-45	-	-
Radius	4-7	15-35	10-45	-	-
Average Path Length	6-10	20-30	5-20	2-3	3-4
Clustering Coefficien	-	15-25	38-55	2-3	3-4
Clique Number	2-4	-	-	2-3	3-4
Spectral Gap	-	-	5-10	-	-
Graph Compression Ratio	-	2-5	5-35	-	-

discovery and data mining, pages 523–528, New York, NY, 2006. ACM.

[4] F. Chung. *Spectral Graph Theory*. CBMS. AMS, 1997.

[5] P. Desikan and J. Srivastava. Mining Temporally Evolving Graphs. In *Proc. of WebKDD 2004*, pages 22–25, 2004.

[6] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, V10(4):255–268, May 2006.

[7] A. Feldmann, A. C. Gilbert, W. Willinger, and T. Kurtz. The changing nature of network traffic Scaling phenomena. *Computer Communication Review*, 28:5–29, 1998.

[8] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M.-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influenc leadership roles in movements of plains zebra, equus burchellii. *Animal Behaviour*, 73(5):825–831, May 2007.

[9] Q. Gao, M. Li, M. B, and P. Vitnyi. Applying mdl to learning best model granularity, 2000.

[10] D. Kempe, J. Kleinberg, and A. Kumar. Connectivity and inference problems for temporal networks. *J. Comput. Syst. Sci.*, 64(4):820–842, 2002.

[11] S. Kotssiantis and D. Kanellopoulos. Discretization techniques: A recent survey. In *GESTS International Transactions on Computer Science and Enginerting, 32(1)*, pages 47–58, 2006.

[12] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM.

[13] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densificatio laws, shrinking diameters and possible explanations. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, 2005. ACM.

[14] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[15] S. Papadimitriou, F. Li, G. Kollios, and P. S. Yu. Time series compressibility and privacy. In *In VLDB*, pages 459–470, 2007.

[16] C. Partridge, D. Cousins, A. W. Jackson, R. Krishnan, T. Saxena, and W. T. Strayer. Using signal processing to

analyze wireless data traffic In *Proc. ACM workshop on Wireless Security*, pages 67–76, 2002.

[17] A. P. Pentland, N. N. Eagle, and N. N. Eagle. Machine perception and learning of complex social systems. In *Ph.D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology*, 2005.

[18] M. H. Pesaran and A. Timmermann. Model instability and choice of observation window. University of California at San Diego, Economics Working Paper Series 99-19, Department of Economics, UC San Diego, Sep 1999.

[19] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD trace cambridge/haggle/imote/infocom (v. 2006-01-31), Jan 2006.

[20] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 687–696, New York, NY, USA, 2007. ACM.

[21] S. R. Sundaresan, I. R. Fischhoff, J. Dushoff, and D. I. Rubenstein. Network metrics reveal differences in social organization between two fission-fusio species, grevy's zebra and onager. *Oecologia*, September 2006.

[22] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identificatio in dynamic social networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726, New York, NY, USA, 2007. ACM.