# The Representation and Analysis of Dynamic Networks

THESIS

Presented in Partial Fulfillment of the Requirements for Graduation with Honors Research Distinction in the College of Engineering of The Ohio State University

By

Michael Griscom

Undergraduate Program in the Department of Computer Science and Engineering

The Ohio State University

2013

Thesis Committee:

Prof. Srinivasan Parthasarathy, Advisor

Prof. Arnab Nandi

Abstract

Many real-world phenomena, such as article citations and social interactions, can be viewed in terms of a set of entities connected by relationships. Utilizing this abstraction, the system can be represented as a network. This universal nature of networks, combined with the rapid growth in scope of data collection, has caused significant focus to be placed on techniques for mining these networks of high-level information. However, despite the strong temporal dependencies present in many of these systems, such as social networks, substantially less is understood about the evolution of their structures. A dynamic network representation captures this additional dimension by containing a series of static network "snapshots." The complexity and scale of such a representation poses several challenges regarding storage and analysis. This research explores a novel bit-vector representation of node interactions, which offers advantages in its ability to be compressed and manipulated through established methods from the fields of digital signal processing and information theory. The results have demonstrated high-level similarity between the considered datasets, giving insights into efficient representations. By way of the discrete Fourier transform, this research has also revealed underlying behavioral patterns, particularly in the social network realm. These approaches offer improved characterization and predictive capacity over that gained from analyzing the network as a static system, and the extent of this descriptive power obtainable through the bit-vector representation is a question which this research aims to address.

Acknowledgments

I would like to thank my advisor, Prof. Parthasarathy, for his immense help and insight throughout the duration of the project. Acknowledgement is also due to Farhang Zia for his assistance in this project, and to David Fuhry, Yiye Ruan, Yu-Keng Shih, and Ye Wang for their help in the research lab. I would also like to thank Prof. Nandi for serving on the thesis committee.

This research is funded in part by a generous research scholarship from the Undergraduate Honors Committee in the College of Engineering. This material is also based upon work supported by the National Science Foundation under Grant No. IIS 0917070 and an associated REU supplement. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and or his collaborators and advisor and do not necessarily reflect the views of the National Science Foundation.

Vita

May 2009 ........................................................Reynoldsburg High School

June 2010 – June 2011..................................Undergraduate Researcher, Department of Materials Science & Engineering, The Ohio State University

June 2011 – Aug 2011 ...................................Undergraduate Researcher, Tyndall National Institute, Cork, Ireland

Nov 2011 – June 2012 ...................................Undergraduate Researcher, Department of Biomedical Informatics, The Ohio State University

Sept 2012 – Current .......................................Undergraduate Researcher, Department of Computer Science & Engineering, The Ohio State University

Fields of Study

Major Field:  Computer Science and Engineering

Table of Contents

List of Tables

List of Figures

## Introduction

The ubiquity of digital communications, along with the advancement of storage technology, has introduced the ability to gather and store information on a scale exponentially greater than before. Social networks, search engines, and e-retailers are all examples of new innovations which have capitalized upon the ability to collect and utilize data. These data contain a wealth of knowledge that it not immediately evident, but by leveraging methods found in fields such as statistics and artificial intelligence, one is able to sift out patterns and rules. The resulting information provides the descriptive and predictive insight that is the aim of data mining [1] [2]. In the case of social networks, for example, this knowledge can then be used to suggest recommended connections or provide relevant advertising.

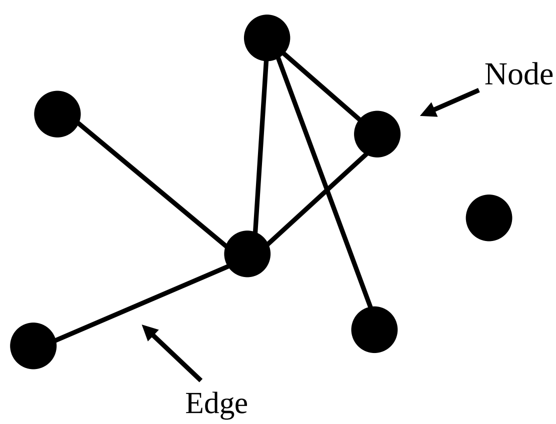


Figure 1: A simple network

For a subset of applications within data mining, representing the system as a network can prove useful. These networks consist of a set of points with connections between them, which are referred to as nodes and edges respectively. For the case of a social network, the people can be treated as nodes and the relationships between them as edges. A simple network is depicted in Figure 1; these networks correspond to a well-studied concept in mathematics called graphs [3].

Whereas many networks can be considered as fairly static, e.g., report structure hierarchies within businesses, others possess inherent flux. These dynamic networks are therefore represented as dynamic graphs, and are the subject of this research. A static graph can be compared to a freeze-frame snapshot of the state of a system, while dynamic graphs would be analogous to a movie composed of many static graph frames. By combining the knowledge gained from these static states with information from the time-based view, the structure and behavior of these dynamic graphs can be described and predicted.

The vast size and complexity of these graphs pose several problems related to computer representation and analysis methods, which are the subject of this research. Due to these complications, the study of dynamic graphs has not yet reached the maturity of their static counterparts. This research proposes a novel representation which treats edge evolutions as sequences of bits, and explores the storage and analysis of the resulting data.

An improved understanding of dynamic graphs would be immediately applicable across many domains. For example, protein-protein interactions are often modeled as a network, the description of which can improve medical capabilities. Further, the structure and evolution of social networks impacts the spread of phenomena such as information and disease. Capturing the temporal aspects of the graph can also improve the prediction of ebbs and flows in activity, along with the development of communities.

In the social realm, individuals follow general daily and weekly schedules. It is thus natural to expect these cycles to imprint themselves upon the behavior of the corresponding nodes in a social network, such as in email activity patterns. Knowledge of such tendencies could aid in server load prediction or in the detection of erratic behavior. Additionally, improved knowledge of edge formations can also allow for improved connection recommendations, for example, in suggesting additional email accounts to carbon-copy.

The representation and modes of analysis for a dynamic graph mutually inform each other. By introducing a new way of expressing this network data, this research invites the exploration of alternate perspectives of knowledge discovery in their behavior through time.

## Related Work

In the context of dynamic graph representation, Sulo et al. detailed a method for representing a dynamic graph data stream with a resolution which balances noise reduction and information loss for a given graph statistic [4]. Barbay et al. developed a labeled graph representation that approaches the information-theoretic lower bound of storage cost [5]. The TimeArr storage manager developed by Soroush and Balazinska handles queries of past versions of an array database, using a backward-delta storage method in combination with array tilting and variable-length encoding [6]. A similar platform, DeltaGraph, was developed by Khurana and Deshpande which allows for retrieval of historical graph data and auxiliary information, such as in subgraph matching [7].

With respect to graph analysis, Eagle represented the cell phone usage of a group of participants in terms of a dynamic graph, and applied probabilistic analysis in addition to a Fourier Transform on a specific graph point characteristic, identifying so-called eigenbehavioral classes of nodes [8]. In [9], Leskovec et al. explore the pattern of activity of blog graphs, in addition to common cascade patterns of influence spread. Asur et al. explored event detection in dynamic graphs and the analysis of resulting behavior-based measurements, with applications to link prediction [10]. Sun et al. explored the identification of communities and their evolution in the context of dynamic graphs with the development of the GraphScope framework [11], a compressed storage representation which also enables the detection of discontinuities and anomalous activity. Berlingerio et al. examined the recurrence of small-scale patterns in dynamic graphs, and used these to reason about the evolution of the whole through the creation of graph evolution rules [12].

## Methodology

### *Bit Stream Representation*

Without loss of generality, a dynamic graph can be considered as a discrete sequence of graph states, where each state is separated by a variable time interval. If the node set is kept constant, the graph at time $t$ can be represented as $G_t(N, E_t)$ where $N$ represents the set of nodes and $E_t$ represents the set of edges at time $t$. For a given $t$, it can be determined whether $e \in E_t$. Thus, $e$ can be represented as a bit stream where the state of each bit is determined by the presence of $e$ in the graph at the given time. In this manner, the temporal aspect of the graph can be captured by storing a set of bit streams, one for each edge. An example of this representation is given in Figure 2, along with the corresponding graph snapshots.

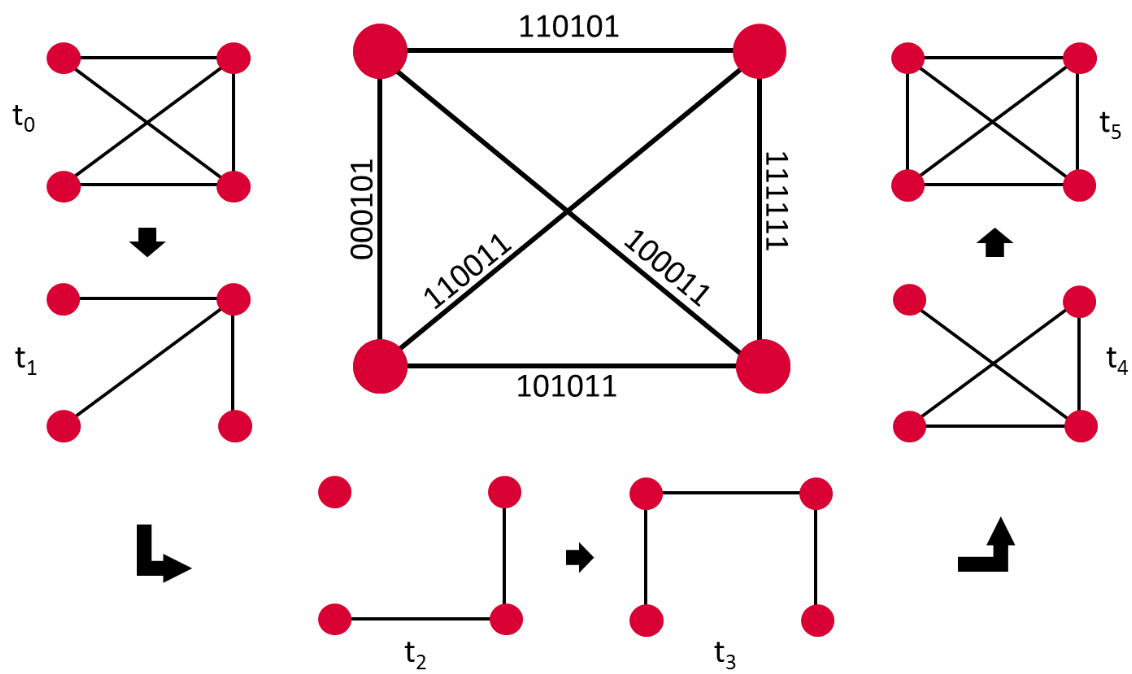


Figure 2: Sample bit stream edge representation and corresponding graph snapshots

In the framework that was created during this research to process these graphs, the length between successive snapshots, $\tau$, is treated as a constant parameter throughout the graph. As such, increasing $\tau$ results in a temporal coarsening of the graph, which enables improved storage and processing performance at the expense of a loss of information. As with the original graph, within the coarsened graph edge existence remains binary; that is, for a given snapshot $E_t'$, $e \in E_t'$ implies that for at least one snapshot $E_x$ in the original graph, $e \in E_x$ where $t \le x < t + \tau$ .

*Datasets*

Four datasets were used during the course of this research, statistics of which are given in Table 1. Being that the representation and analysis performed during this research is intended to apply across a variety of networks, these datasets were chosen due to their difference in size, time scale, and nature. The DBLP dataset is a citation network from articles in the field of computer science dated 1938-2006. This network is represented as a set of authors (nodes) connected by citations (edges). For the purposes of this research, the edges from all datasets are treated as undirected, e.g., a citation in the DBLP dataset from author $X$ to author $Y$ at time $t$ is treated in the same manner as would be a citation from author $Y$ to author $X$ at time $t$.

Table 1: Summary statistics for the datasets considered

| Dataset | Number of nodes (approx.) | Number of connections (approx.) |
|---|---|---|
| DBLP [13] | 400,000 | 600,000 |
| Enron [14] | 9,500 | 115,000 |
| Facebook [15] | 60,000 | 840,000 |
| Infection [16] | 14,000 | 230,000 |

The Enron dataset represents emails involving Enron Corporation accounts during the span of January 1999-July 2002, for which each node represents an email account and each edge an email. The Facebook dataset consists of wall posts between users in the New Orleans network from September 2006-January 2009. The infection dataset consists of contact during a simulation carried out over a span of three months as an exhibit at the Science Gallery in Dublin, Ireland, in collaboration with SocioPatterns [17]. Each participant wore an RFID emitter which would record any other emitter in range (approx. 1.5m) along with a timestamp window of 20s. All contacts were recorded regardless of the infected status of the participants involved; node infection status is also disregarded for the purposes of this research.

Timestamps of the DBLP dataset correspond to the year of publication, thus the value of $\tau = 1\ yr$ was used in processing. Likewise, the infection simulation dataset was treated using the resolution of the dataset, $\tau = 20s$ For the Enron and Facebook datasets the timestamps have an accuracy of $1s$, but except where otherwise noted, $\tau = 30\ min$ was used during processing, which was selected based on the time-scale of the interactions within the datasets.

## Results

### *Encoding Scheme*

The strength of the compression scheme used for a dynamic graph is dependent upon the properties of the individual bit-streams; for example, the space efficiency gains from a run-length encoding strategy are highly dependent upon the frequency of repeating values in a string. In order to determine the general nature of the bit-streams, the number of ‘true’ bits was counted for each stream in a given graph, resulting in the Hamming weight of the stream. Bins were then created for these weights, and the resulting bin counts calculated. For each dataset, these values are plotted on a log-log scale in Figure 3.

Aside from the infection dataset, the bin distributions appear to follow a power-law distribution. It is likely that the more rapid drop-off in the infection plot is due to the restricted duration of a given node's activity, as the time required for participants to form excessive connections would exceed the normal amount of time to traverse the exhibition. The form for a power-law distribution follows $Cx^{-\alpha}$, where $\alpha$ is a scaling factor; an increase in this factor thus implies an increase in the ratio of sparse to dense bit streams within the graph. For the three plots that followed this distribution, the scaling factor for a best-fit approximation is given in Table 2, along with the corresponding values for the coefficient of determination.

Table 2: Best-fit exponential scaling factor for edge activity magnitudes

| Dataset | $\alpha$ | $R^2$ |
|---|---|---|
| DBLP | 3.37 | 0.97 |
| Enron | 1.65 | 0.93 |
| Facebook | 2.06 | 0.94 |

One class of possible compression methods involves the encoding of bit streams. Decompression of a given graph state thus involves decoding the data for each bit-stream. In [18], the compression of sparse binary data is analyzed for a memory-less signal model described by the probability $p$ of a 'true' bit occurring, which was found to give similar performance as a linear predictive coding method. For bit stream encoding, the average percentage of ‘true’ bits for the nonzero streams of each dataset for a given granularity can be considered an estimate of $p$.

Three encoding schemes were tested: run length encoding [19], coordinate encoding, and a block-encoding method developed by Zeng and Ahmed [18] within the image processing domain. This block-encoding method resembles that developed by Kunt [20], with the distinction that Kunt’s method stores the entire sequence of a nonzero block, whereas Zeng’s stores the location of the nonzero values within the block. This method outperforms Kunt’s within the range of $p$ values that the datasets in this experiment fall [18], thus Kunt’s method was not tested. Representative space savings for each bit stream of the Enron dataset are given in Figure 4. It can be seen that the storage space required is correlated with the number of ‘true’ bits of the bit stream. Overall compression statistics for the datasets are given in Table 3; the DBLP dataset was omitted from this analysis because the small number of snapshots, 68, allows for only marginal bit stream compression.

Table 3: Bit stream encoding results

| | | | Post-encoding Space Savings (%) | | |
|---|---|---|---|---|---|
| Dataset | $\tau$ | Average ‘true’ bits (%) | Run Length | Coordinate | Block |
| Enron | 30 min | 0.27 | 91.83 | 95.61 | 97.14 |
| Infection | 20 s | 1.87 | 45.24 | 71.43 | 83.33 |
| Facebook | 30 min | 0.11 | 96.93 | 98.22 | 98.76 |

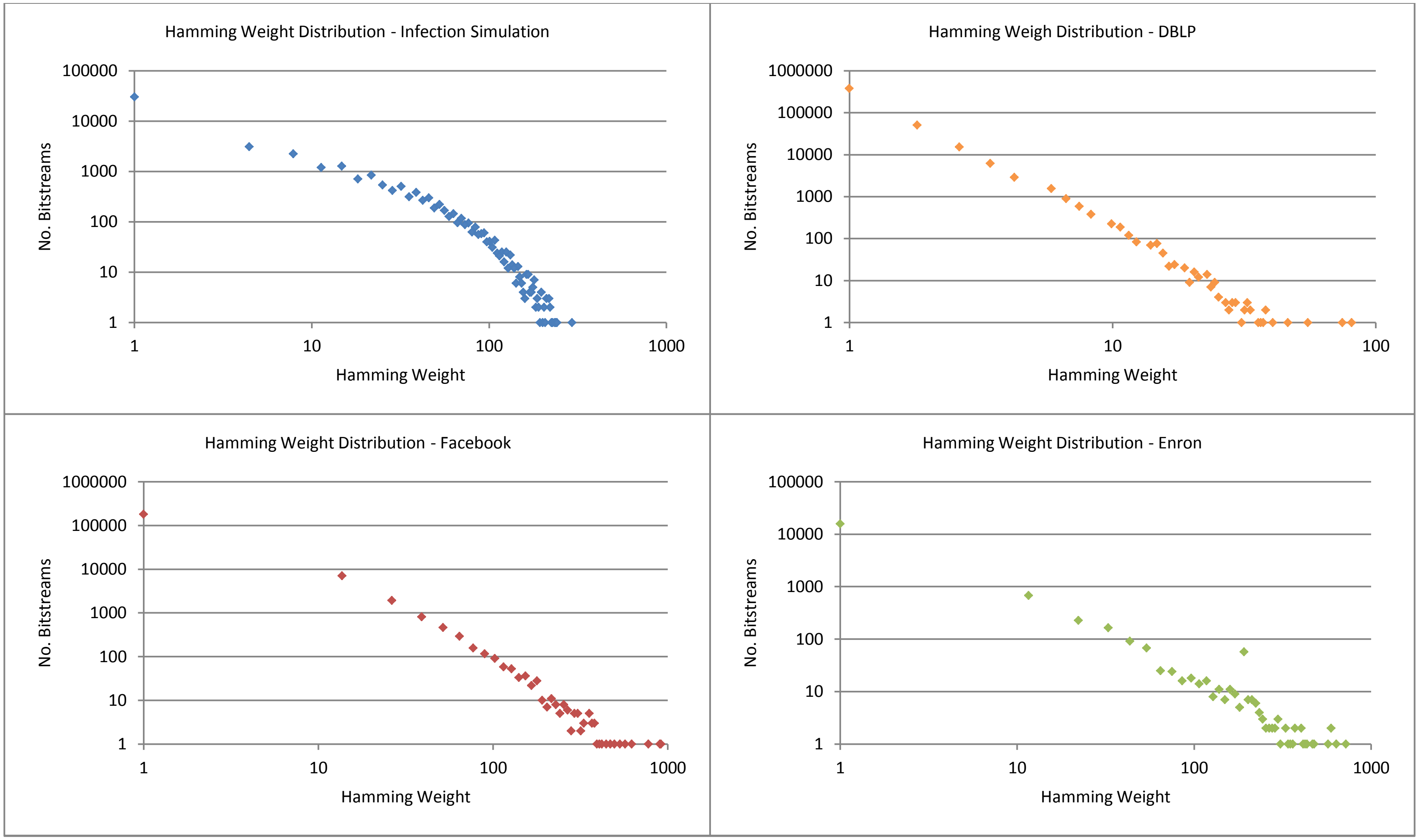


Figure 3: Bit stream Hamming weight distributions for each dataset

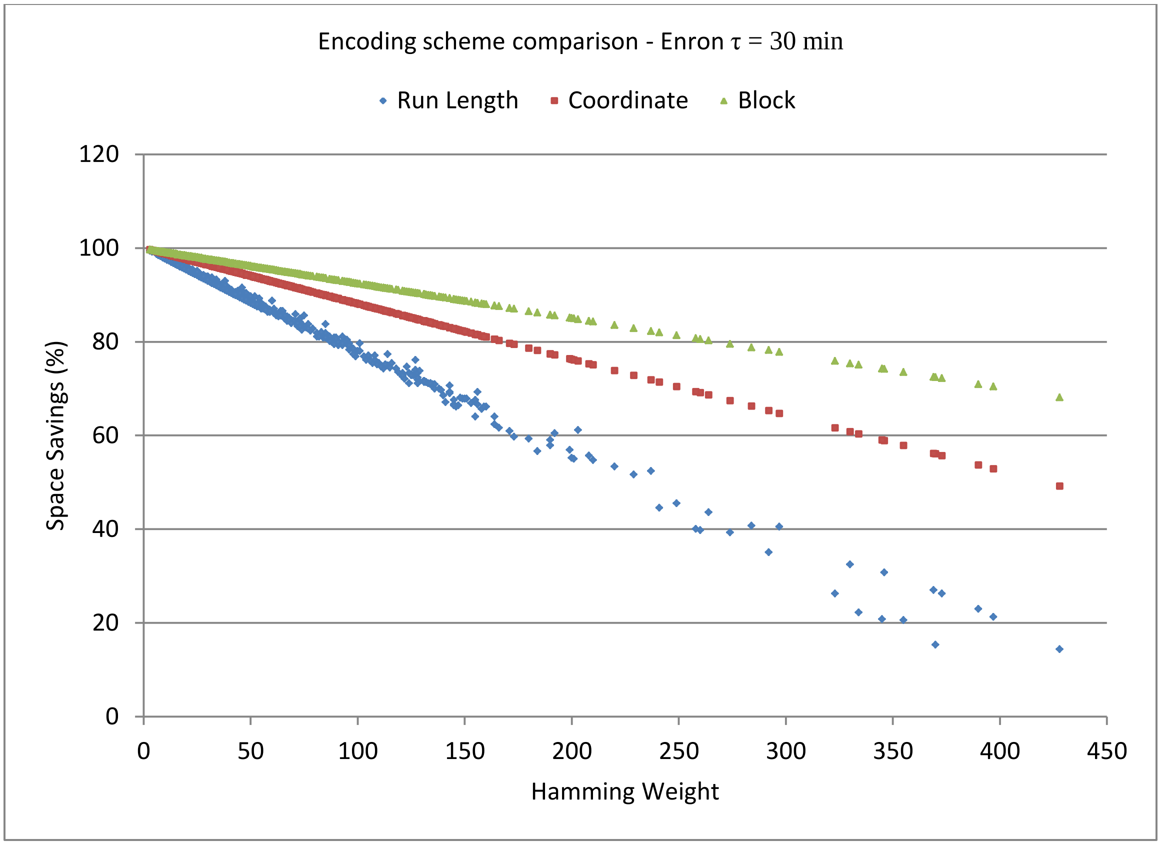


Figure 4: Space savings across encoding schemes for bit streams of Enron dataset

*Fourier Analysis*

To obtain a holistic characterization of the activity frequencies of the graph, a discrete Fourier transform (DFT) was calculated for each bit stream. This algorithm, popular in the digital signal processing community, decomposes a signal into its constituent frequencies, each with an amplitude and phase. The magnitudes of each frequency component were totaled across the bit streams. The frequency values were then inverted to allow for direct interpretation of time spans. In the subsequent DFT plots, the magnitudes have been normalized such that each magnitude is the percentage of the maximum value achieved by the DFT.

In Figure 5, a DFT plot is given of the DBLP dataset with $\tau = 1\ yr$, which results in no coarsening of the graph. The Nyquist frequency thus lies at 2 $yr^{-1}$, frequencies within the data above this value are aliased, and thus their magnitudes cannot be determined. The graph demonstrates a generally consistent magnitude until approximately 6 years, at which point there is an increasing trend that continues into the long time-span (low frequency) values. The phenomenon of high magnitude values at low frequencies is mirrored by the other datasets, and is a reflection of the large number of sparse edges within the graph, each of which contributes low-frequency terms.

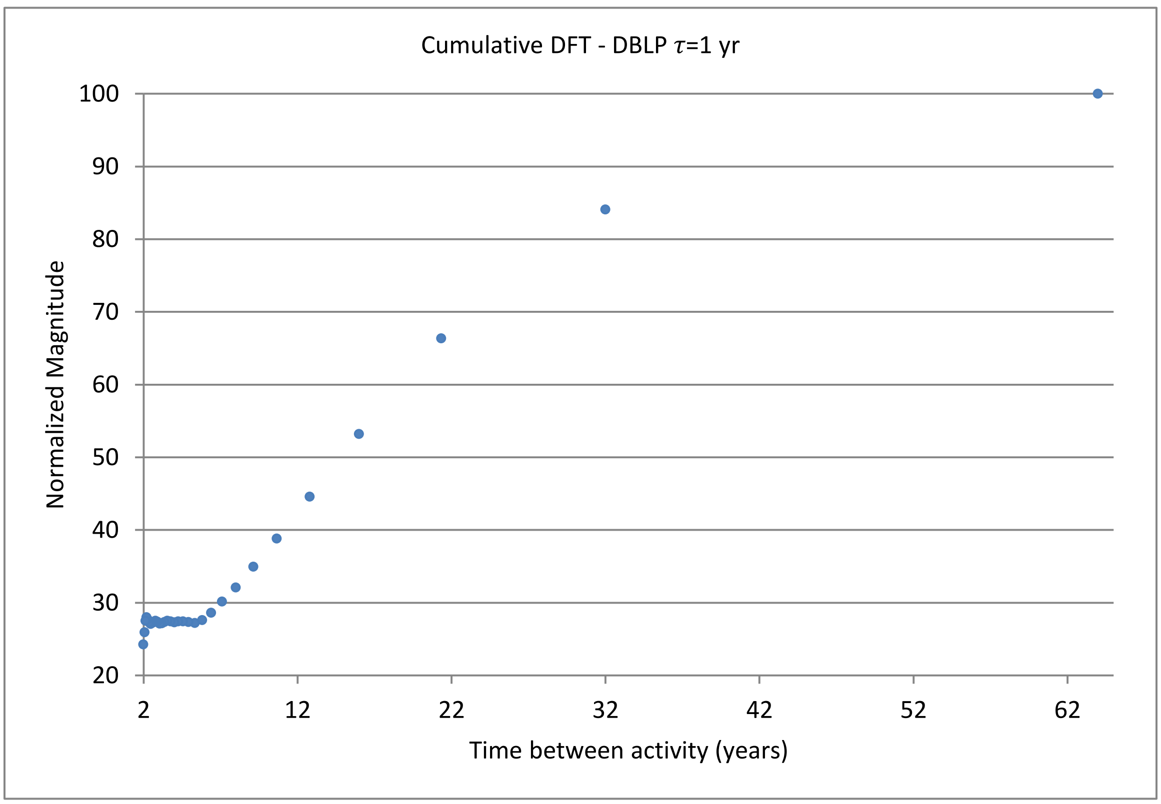


Figure 5: Cumulative DFT values for DBLP dataset with $\tau = 1$ yr

In Figure 6 the cumulative DFT values for the Facebook and Enron datasets with $\tau = 12\ hr$ are plotted. The frequencies were then normalized to 1 day$^{-1}$; Figure 7 depicts cumulative DFT results for the Facebook and Enron datasets with $\tau = 30\ min$.

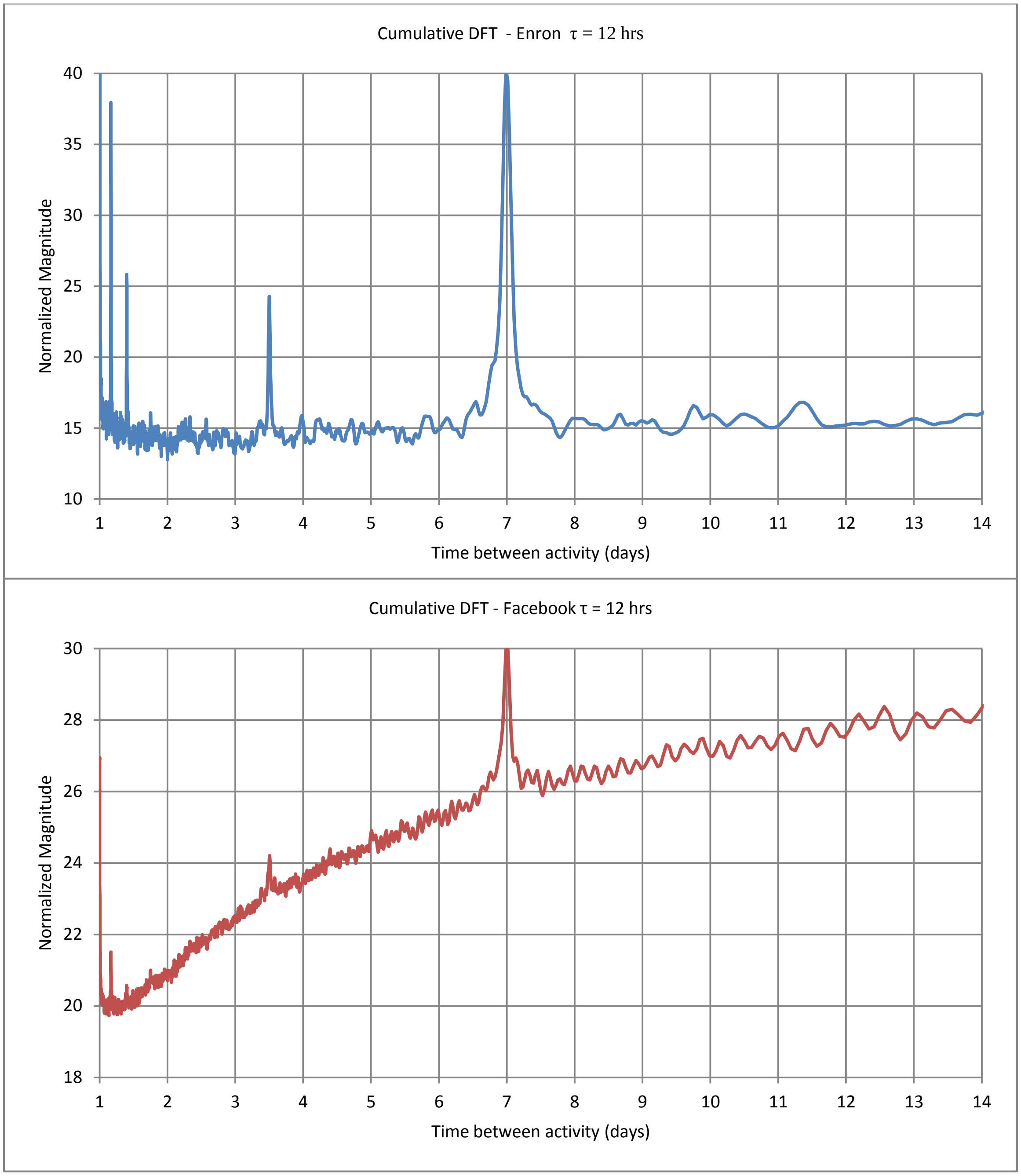


Figure 6: Cumulative week-scale DFT magnitudes for Facebook and Enron datasets

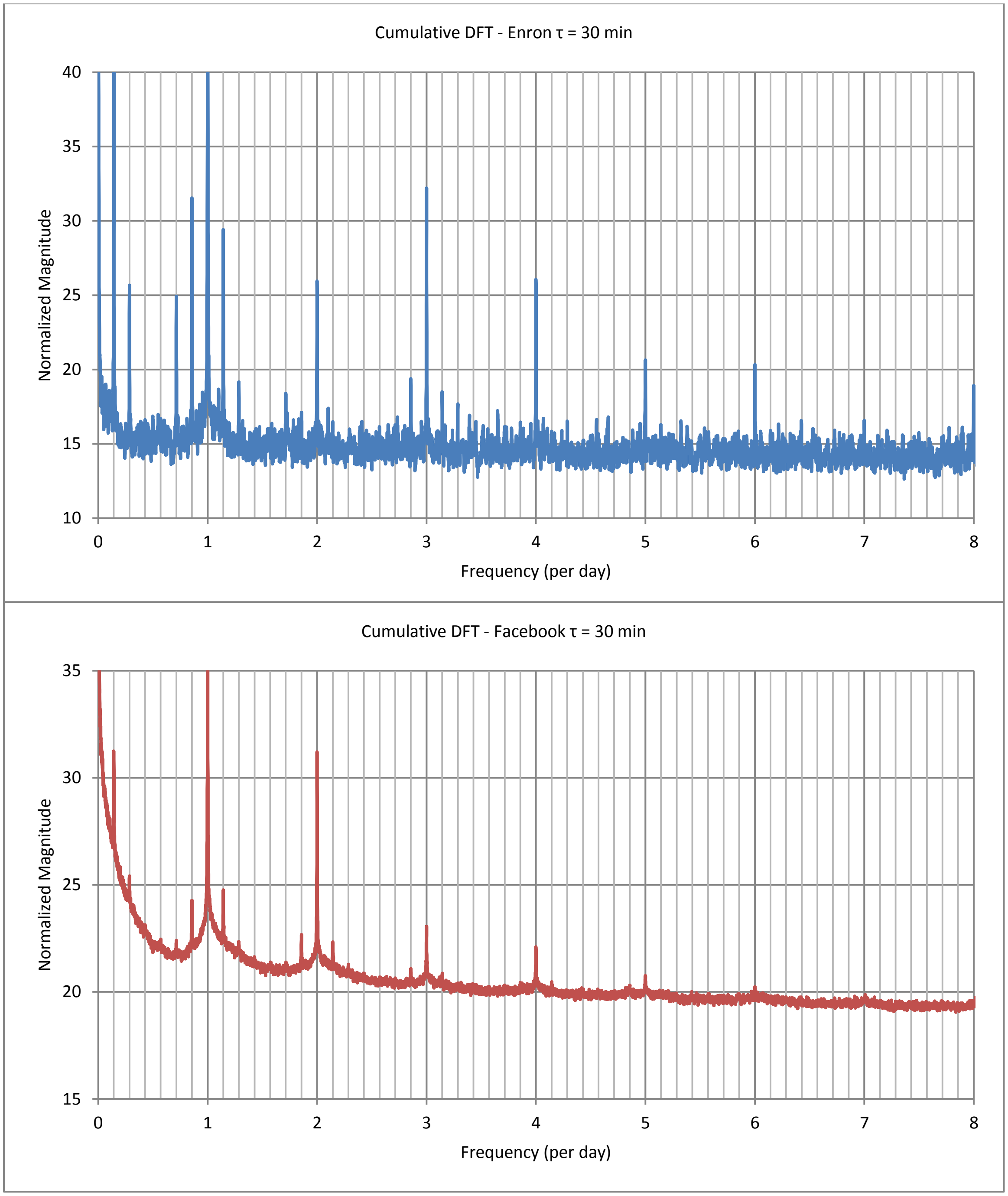


Figure 7: Cumulative DFT frequency magnitudes for Facebook and Enron datasets

To account for the possibility of the sub-peaks in the previous figures being byproducts of the spectral decomposition by way of spectral leakage [21], windowing functions were applied to each bit-stream of the Facebook dataset. In Figure 8, a subset of the cumulative DFT plot is given after processing without a windowing function, along with that resulting from processing with the application of a Hanning window and with a 4-term Blackman-Harris window [21]. It can be seen that the peaks survive both windowing procedures. Additionally, a value of $\tau$ – namely, 1751 – relatively prime to the number of seconds in a day was also used to perform the DFT, the magnitudes of which again displayed these peaks. This was done in order to account for the presence of resonance or aliasing effects of the sampling frequency with respect to the peak frequencies; the plots are omitted for brevity.

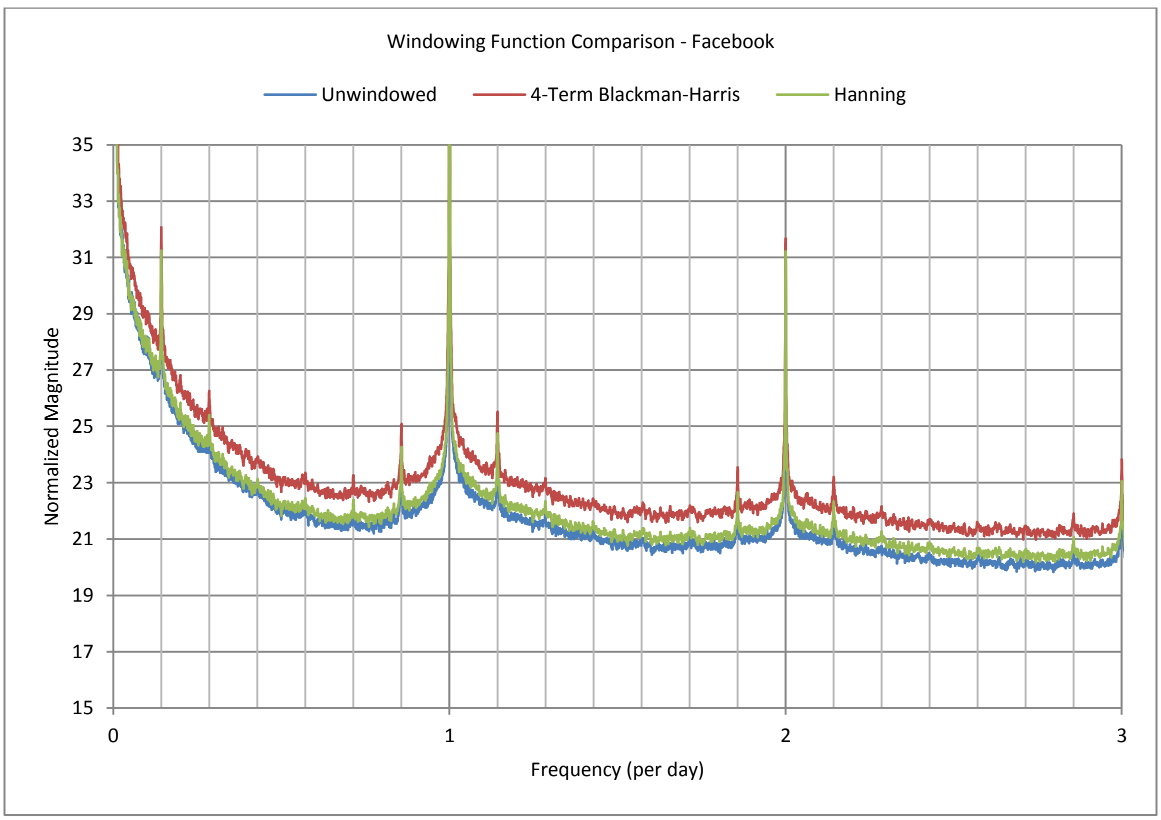


Figure 8: Effects of windowing on Facebook DFT

# Discussion

The power-law distributions seen in Figure 3 can be interpreted as a temporal analog of the established result that many real-world static graphs exhibit a scale-free property with respect to node degrees [22]. These distributions indicate that many of the node connections in the data occur a very few number of times. Their consistency also allows for prediction of unmeasured values; for example, knowledge of the prevalence of highly active pairs on Facebook would be sufficient in predicting the number of low-activity pairs.

The representation explored, which compresses individual bit streams, has several advantages over compressing the graph snapshots; firstly, bit streams are likely to show temporal locality and therefore lend themselves to direct compression. Additionally, this approach allows the bit stream data of individual edges to be stored contiguously, enabling more efficient analysis of an edge's behavior throughout time. This also enables straightforward real-time recording, which could be distributed across systems or delegated to the nodes themselves.

Although the data sets considered were small enough to be memory-resident during processing, systems that contain many interacting nodes or that are extended through time will require compression. A block-encoding scheme was shown in Table 3 to offer strong space-savings. The scheme also allows for efficient querying, as blocks can be quickly traversed, only decompressing the relevant locations.

In examining the frequency spectra, the transformations of the Facebook and Enron datasets reveal particularly interesting behavior. Figure 6 exhibits several peaks, including at 7 days and 3.5 days, along with multiple peaks between 1 and 2 days. This seems to indicate a rhythmic behavior having period of approximately one week present in both datasets. Additionally, the 3.5 day peak represents a frequency at the first harmonic of the 7 day peak. This harmonic pattern is more evident on the day-scale, as depicted in Figure 7. It can be seen that the major peaks lie at the harmonics, e.g., $\frac{1}{12}hr^{-1}$, $\frac{1}{8}hr^{-1}$, etc. with sub-peaks present at $\frac{1}{7}$ sub-harmonics of these values. The windowing results of Figure 8, along with the variations in $\tau$ outlined in the previous section, imply the existence of these sub-peaks within the data, rather than as artifacts of the DFT method.

## Conclusions and Future Work

This research introduced a novel bit stream representation of dynamic graphs, and explored resulting methods of storage and analysis. Despite the enormous range of possible behavior for a dynamic graph, the bit streams for each of the datasets considered show similar properties, allowing for an efficient encoding method for use across systems. The encoding method explored results in lossless compression, although one could also utilize lossy storage; for example, by storing the $k$ most dominant frequencies of the Fourier transform for each bit stream, a variable compression mechanism could be created. This representation would also enable analysis to be performed directly in the frequency domain which could, in turn, show similarities that result from phase-shifts of edge activity patterns, which would be non-apparent under alternate representations.

The analysis performed demonstrates high-level regularity in node behavior within social contexts. This could be used in server load prediction or in the detection of erratic behavior. Additionally, an understanding of edge formations can also allow for improved connection recommendations, for example, in suggesting additional email accounts to carbon-copy, or which walls for a user to post on. The location of peak frequency values on the harmonics of the day also suggest the ability to use sound as a novel representation method for dynamic networks, where an audible frequency of the standard musical scale could represent the $\frac{1}{24}hr^{-1}$ frequency, with harmonics of that tone scaled according to the data.

Due to the holistic nature of the characterizations performed, further analysis would likely benefit from a decrease in scaling, considering community and node-level behaviors. For example, a modified version of the oscillator model found in Arenas et al. [23] could be applied to edges, using nodes as couplers and edge 'true'-bits for oscillator excitations. By obtaining the amplitude and phase of dominant frequencies for individual bit streams, one could provide seed values for the oscillators. By monitoring the synchronization of these oscillators within subgraphs, community structures could be obtained in a manner similar to that developed by Arenas. Through the dynamic nature of this representation, one could extend such results by modeling information propagation through imparting a pulse to a given oscillator and examining the resulting impact to the system.

This representation can also be viewed as an ultra-high dimensional dynamical system [24], where each node has many degrees of freedom in the target and time of its connection formations. The corresponding state of the graph can be visualized as tracing a complex path through such a space, where the periodic patterns explored in this research correspond to identifying loops within projections of this path onto certain dimensions. There are likely more complex aspects of the system's evolution, e.g. attractors which the state trajectory tends to trace. The key step to finding these lies in the determination of a suitable dimensionality restriction to the graph. This could entail focusing upon individual communities or motif structures, such as Triangle K-Cores [25], which can be efficiently extracted to provide an indicator of the structural makeup of the graph.

Through this lower-dimensional description, it would then become possible to make quantitative claims as to the stability of substructures in the network, which would result in improved predictions of link formations and other aspects of the network's evolution. In particular, examining the Lyapunov stability [24] of community structures could enable the discrimination between growing and stable social groups. Notions of stability and synchronization such as these are only samples of a large class of analysis techniques that are supported through the novel stream-based representation described in this work.

Bibliography


[1] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison-Wesley, 2005.

[2] J. Han, M. Kamber and J. Pei, Data mining: Concepts and Techniques, Boston: Elsevier/Morgan Kaufmann, 2012.

[3] M. E. J. Newman, Networks: An Introduction, New York: Oxford University Press, 2010.

[4] R. Sulo, T. Berger-Wolf and R. Grossman, "Meaningful selection of temporal resolution for dynamic networks," in *Mining and Learning with Graphs*, New York, 2010.

[5] J. Barbay, L. Aleardi, M. He and J. I. Munro, "Succinct Representation of Labeled Graphs," in *Lecture Notes in Computer Science*, vol. 4835, Springer, 2007, pp. 316-328.

[6] E. Soroush and M. Balazinska, "Time Travel in a Scientific Array Database," in *IEEE International Conference on Data Engineering*, Brisbane, Australia, 2013.

[7] U. Khurana and A. Deshpande, "Efficient Snapshot Retrieval over Historical Graph Data," in *IEEE International Conference on Data Engineering*, Brisbane, Austrailia, 2013.

[8] N. Eagle, Machine perception and learning of complex social systems, Massachusetts Institute of Technology, Department of Media Arts and Sciences, 2005.

[9] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance and M. Hurst, "Cascading behavior in large blog graphs," in *SIAM International Conference on Data Mining*, Minneapolis, Minnesota, 2007.

[10] S. Asur, S. Parthasarathy and D. Ucar, "An Event-Based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 4, 2009.

[11] J. Sun, C. Faloutsos, S. Papadimitriou and P. S. Yu, "GraphScope: parameter-free mining of large time-evolving graphs," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, 2007.

[12] M. Berlingerio, F. Bonchi, B. Bringmann and A. Gionis, "Mining Graph Evolution Rules," in *Lecture Notes in Computer Science*, vol. 5781, Springer, 2009, pp. 115-130.

[13] "The DBLP Computer Science Bibliography," [Online]. Available: http://dblp.uni-trier.de/.

[14] B. Klimt and Y. Yang, "Introducing the Enron Corpus," in *Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2004.

[15] B. Viswanath, A. Mislove, M. Cha and K. P. Gummadi, "On the Evolution of User Interaction in Facebook," in *Proceedings of the 2nd ACM SIGCOMM Workshop on*

*Social Networks (WOSN'09)*, Barcelona, Spain, 2009.

[16] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton and W. Van den Broeck, "What's in a Crowd? Analysis of Face-to-Face Behavioral Networks," *Journal of Theoretical Biology,* vol. 271, pp. 166-180, 2011.

[17] "SocioPatterns," [Online]. Available: http://www.sociopatterns.org.

[18] G. Zeng and N. Ahmed, "A block coding technique for encoding sparse binary patterns," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 5, pp. 778-780, 1989.

[19] Y. Yasuda, "Overview of digital facsimile coding techniques in Japan," *Proceedings of the IEEE*, vol. 68, no. 7, pp. 830-845, 1980.

[20] M. Kunt and O. Johnsen, "Block coding of graphics: A tutorial review," *Proceedings of the IEEE*, vol. 68, no. 7, pp. 770-786, 1980.

[21] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE,* vol. 66, no. 1, pp. 51-83, 1978.

[22] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, p. 509–512, 1999.

[23] A. Arenas, A. Diaz-Guilera and C. J. Perez-Vicente, "Synchronization reveals topological scales in complex networks," *Physical Review Letters,* vol. 96, 2006.

[24] S. Strogatz, Nonlinear dynamics and chaos: with applications to physics, biology, chemistry and engineering, Westview Press, 2001.

[25] Y. Zhang and S. Parthasarathy, "Extracting Analyzing and Visualizing Triangle K-Core Motifs within Networks," in *IEEE International Conference on Data Engineering*, Washington DC, 2012.