Contents

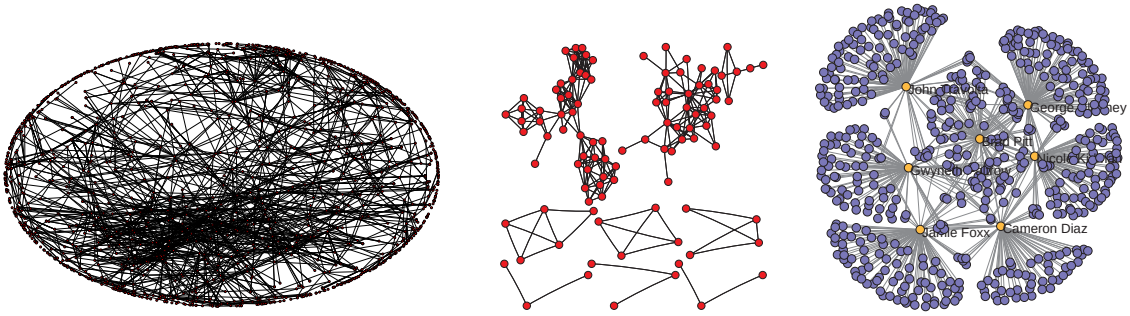
Chapter 1

Introduction

This thesis is about *dynamic network data*, with an individual emphasis on each word. In the most general sense, a dynamic network is a very powerful mathematical representation for time-varying systems that consist of many interacting entities. In particular, dynamic networks are most useful when the system exhibits extremely complex behavior and is continually changing over time, and when only a record of the interactions between entities are observed. One of the best known instances of network representation applied to a real-world system is the *social network* [Wasserman and Faust, 1994], where the relationships between people are mapped into a web-like network. The structure of this web of human connections has been used for decades to gain insight into the structure of human societies and behavior, and more recently animal societies and behavior as well. A dynamic network is a much more recent and powerful representation than a canonical social network that allows one to explicitly map the changing structure of such a web over time. It is a generic, graph-theoretic representation that can be applied to systems much more diverse than human and animal social connections. In this thesis, we will describe correspondingly generic methods for analyzing any dynamic network dataset in successively more complex ways.

The focus of this thesis is on developing generic analytical methods, rather than focusing on the analysis of networks from a specific domain. Since the diversity of sources of dynamic network data has been steadily increasing [Newman, 2003], the problems in this thesis are driven by the need for analytical tools that are powerful under a minimal set of assumptions, and as agnostic as possible to the characteristics of any single class of domains. As examples of this data diversity, dynamic network datasets can be produced by massive industrial logging systems attached to communications switches, or by researchers scraping the World Wide Web from a laptop; by social scientists painstakingly interviewing human subjects over many years, or by automatically processing the log files from a web server. Where this data might have earlier been condensed into simpler formats, new techniques for handling very large dynamic networks can now enable more powerful network analysis. In particular, this thesis examines and builds tools for analyzing the *dynamics* of networks, *i.e.,* the way in which its structure is changing over time, to extract information about the underlying system. The tools developed here fit between the statistician John Tukey's original notion of exploratory data analysis [Tukey, 1980], where the focus is on probing data to generate questions and hypotheses for subsequent investigation[1], and modern data mining, where the focus is on extracting the

[1] As opposed to *confirmatory data analysis* methods like hypothesis testing, *etc.*, which presume the existence of a hypothesis.



(a) Co-authorships of academic papers between researchers in theoretical computer science.

(b) Social associations between wild horses on Shackleford Island, North Carolina.

(c) Social associations of the top 7 Hollywood celebrities by PageRank, determined by co-appearance in publicly available photographs.

Figure 1.1: Three examples of physical systems represented as networks.

statistically 'interesting' parts of massive datasets under various definitions of what constitutes interesting.

Returning to the three emphasized words in the opening statement, a *network* is a graph-theoretic representation of a system where individually identifiable entities, either physical or abstract, interact with each other. Almost any interconnected system can be represented as a network, the most common representation for which is a graph with labeled vertices. Figure 1.1 shows an example of a graph drawing of three different network datasets. Figure 1.1a is a scientific co-authorship network, in which vertices represent individual researchers in theoretical computer science, and an edge between any pair of researchers indicates that they have published an academic paper together. Collaboration and citation networks can be easily extracted from computerized bibliographic databases. Figure 1.1b is a network representation of a type of animal association data that is routinely collected by ecologists and field biologists. Ecologists recorded social associations between wild horses on Shackleford Island, North Carolina over a period of three months.[2] Ordinarily, this association data might have been summarized into simpler statistics such as the mean group size, but networks offer a powerful new way to visualize and analyze all the information present in association data. Finally, Figure 1.1c illustrates the fact that network datasets can emerge from unusual places. The Internet Movie Database maintains a large online repository of publicly available photographs of actors, musicians, movie directors, and other individuals associated with the entertainment industry in the United States, taken at various times when the individuals in question appear in public. These photographs lead to a natural approximation of the professional association network between these individuals, and how it changes over time.[3]

Moving on to the second emphasized word in the opening statement of this chapter, we focus on *dynamic* networks in this thesis. It should be clear that the systems shown in Figure 1.1 are not frozen in time, but

[2]Data courtesy of Prof. Daniel I. Rubenstein of the Princeton Equid Research Group.

[3]The IMDB data is obtained using a method (photographing public sightings) that is curiously similar to the one used to collect social association data on the Shackleford horses and other wild animals [Lahiri et al., 2011], and indicative of the increasing use of technology to collect association and interaction data using technologies such as short-range wireless links and GPS positioning [Juang et al., 2002].

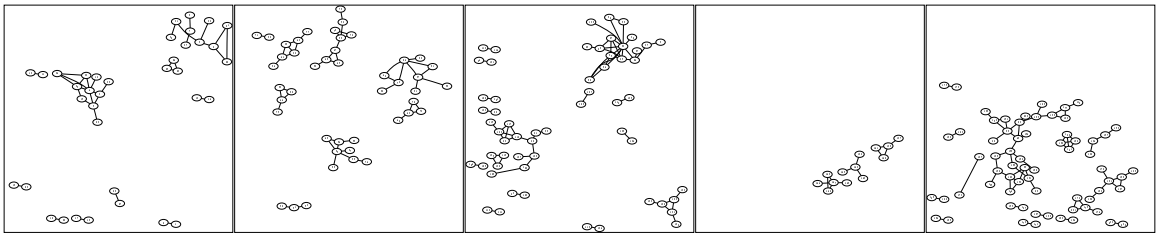


Figure 1.2: Individual network snapshots from five consecutive days of the IMDB photo-based dynamic network. A different portion of the larger, aggregate network is shown in Figure 1.1c.

are in essence 'network snapshots' of continually changing systems.[4] New researchers are continually born, and existing researchers form new collaborations, so the co-authorship network in Figure 1.1a is certainly evolving over time, as are the other two networks depicted. There are therefore two components to most networks: its structure at any point in time, such as the visualizations in Figure 1.1, and the dynamics of the underlying system that drive the formation and evolution of this structure through time. These dynamics have generally been difficult to study because of the limited availability of data with a temporal component. However, this has changed with recent technological advances, resulting in large and detailed datasets that depict the structures shown in Figure 1.1 through the time dimension.

Given a dynamic network dataset, we first address the most basic of analytical tasks:

- *How are network properties changing over time?* Graph theoretic measures summarize the structure of a graph into numeric scalar or vector values. Some examples are the basic counts of vertices and edges in a graph, the empirical distribution of shortest-path lengths between all pairs of connected vertices, and the spectrum of a graph. Since a dynamic network is a graph that changes over time, a fundamental question to ask is if and how these measures are changing as the network changes. For example, are scientific co-authorship networks getting denser over time, possibly indicating that collaboration between scientists is on the rise? Is the average number of hops between routers on the Internet decreasing over time, even at the rate that the Internet is currently growing? And could a trend in its clustering properties over time suggest, for example, that simpler decentralized packet routing algorithms could be used for efficient routing of packets [Kleinberg, 2000]?

In Chapter 2, we start with a methodological description of how basic structural properties of dynamic networks are measured from a collected dataset. Among other uses, these measurements form the basis of a very large class of *generative stochastic models* for the dynamics of the underlying physical system [Chakrabarti et al., 2010], and which are also used to generate realistic synthetic data to test network-based algorithms [Bilgic and Getoor, 2008]. We note that dynamic network datasets are almost always susceptible to various kinds of non-trivial measurement errors and missing data issues, so it is important to select graph measures that are robust under such circumstances. While the effects of sampling biases and missing data have been extensively studied in the context of a single static network [Ghani et al., 1998, Costenbader and Valente, 2003, Kossinets, 2006, Achlioptas et al., 2009], this has not been the case with dynamic measurements of networks. We show that the temporal properties of dynamic networks as reported in the literature use one

[4]The reason why these static snapshots allow us to draw any inferences about the system at all is a network analogue of the concept of statistical consistency: under certain assumptions, the larger our snapshot gets, the more likely it is to be an accurate representation of the underlying system.

of two basic aggregation methods, and a small set of simple graph theoretic measures. We also show that almost all of these measures are sensitive to incomplete data under one of the two aggregation methods, and can erroneously suggest extremely prominent trends where none, or contrary ones, truly exist.

We therefore demonstrate that even the task of determining how basic network measures like diameter are changing over time is not necessarily a settled issue. We suggest some alternatives, but conjecture that there is no generally applicable way to correct or even estimate the measurement bias without making additional assumptions about network dynamics and the size and gross structure of the missing data. This suggests the need to look laterally at a network through the time dimension directly to analyze dynamics. Noting that a dynamic network is analogous to a time series of changing graphs, we choose approaches that are inspired by methods for dealing with numeric time series. In signal analysis, the Fourier transform is a standard tool that is used to decompose a time-varying signal into a sum of sinusoidal components; we develop a similar tool for dynamic networks in Chapter 3 to answer the following question.

- *What are the periodicities and periodic patterns present in a dynamic network?* In systems like communications networks, there are likely to be patterns of interactions between specific individuals that recur or a periodic basis. If we can extract all such patterns in a tractable, succinct representation, we can determine a spectrum of periodicities in the interactions within a physical system similar to the frequency domain representation of time-varying numerical signals. For example, at what frequency do research groups publish in different fields, and do animal associations exhibit cyclical behavior? Alternatively, the periodic patterns themselves could be used for algorithmic tasks like clustering individuals into communities of individuals that interact periodically.

Our formulation of the problem above incorporates the notion of *local periodicity*, where a pattern of connections between individuals may exhibit periodic behavior only temporarily relative to the time extent of the dataset. Although there have been a number of approaches to mining periodic patterns in related types of data, we develop an asymptotically more succinct formulation of the problem that does not lose any information present in other formulation. Furthermore, our formulation of the problem naturally yields a polynomial time and space algorithm in the size of the input dataset, stemming from a worst-case output size that is provably polynomial in the size of the input. Our algorithm efficiently mines periodic patterns at all possible periodicities (a number of other mining algorithms require the user to input 'likely' periodicities), performing orders of magnitude faster on real datasets than its theoretical worst-case bound, and the frequency domain spectra we obtain reveal very plausible principle periodicities in various physical systems.

Since we were able to extract an appreciable amount of information from our periodic pattern mining formulation, a natural extension would be to attempt to detect more sophisticated forms of temporal relationships than periodicity. In doing so, we have to abandon the frequency-domain spectrum of periodicities, and focus solely on structural patterns instead:

- *Which interactions in a dynamic network are strongly correlated in time?* In Figure 1.2, it is difficult to tell whether any edges are temporally correlated with each other. In networks with hundreds of thousands or millions of edges that may appear entirely chaotic, is it possible to wean out the few that are temporally correlated? For example, can we look through the plethora of interactions and tell that a particular pair of scientists publishing a paper in the current year is a good predictor of a different

pair of scientists publishing a paper in the following year, or that an e-mail from a student to their Ph.D. advisor at any time is a reliable predictor of a reply within an hour?

Given the sheer size of typical dynamic networks, this problem can easily become intractable. Although the underlying physical system being represented by a dynamic network can sometimes arise from applying simple local rules, going in the other direction without knowing those rules is a formidable problem. We propose a formulation of this problem in Chapter 4 to mine a rich but limited set of predictably coupled interactions in a dynamic network. When posed as a data mining problem, the patterns of interest are specific interactions that reliably predict future occurrences of themselves or other interactions. Our contribution consists of a novel method of modeling and evaluating dependencies between interactions, in order to yield data mining results with a degree of generalization. In the later part of Chapter 4, we demonstrate some practical applications of mining strong relationships.

In the next section, we briefly survey the historical development of dynamic network data for context, as well as other related representations. Each subsequent chapter deals with one of the three problems described earlier, and more detailed surveys of the literature specific to each problem are contained within individual chapters. Section 1.2 contains a summary of the technical contributions of this thesis, and Section 1.3 briefly describes some of the datasets used throughout the thesis.

1.1 A brief history of network data

In the introductory paragraph of this thesis, we mentioned that there would be an individual emphasis on data, since all the methods presented in this thesis deal with inferring aspects of the underlying system from a dynamic network dataset. A number of developments in network analysis have been directed by the availability of large, comprehensive datasets, and so we briefly summarize the history of dynamic network data in this section, as well as the theoretical developments that accompanied the availability of datasets with greater detail.

Prior to the late 1990s, the analysis of real networks was an endeavor that was largely restricted to the fields of sociology, bibliometrics, and computer networks. In sociology, social network analysis had precedents as far back as the 1930s [Moreno, 1934]. However, the only practical way to collect network data was interviews with subjects. This severely limited the size of datasets and the ability to make inferences at a large scale. For example, a benchmark network dataset in social network analysis, called the *Southern Women* dataset, consists of just 18 individuals and can be printed in its entirety in a publishable table. A meta-analysis of this dataset identified 21 published attempts at analyzing it, each with different methods [Freeman, 2003]. Another sociological dataset, called the *Zachary karate club*, consists of approximately 100 individuals [Zachary, 1977], and is still used to test modern analytical methods [Newman and Leicht, 2007, Tong et al., 2010]. By modern standards, these would be woefully inadequate datasets, but have nonetheless helped develop a comprehensive literature on analytical techniques that use graph theory to gain sociological insights [Wasserman and Faust, 1994].

The current interest in network analysis, both static and dynamic, can probably be largely attributed to two developments in the late 1990s. The first is that physicists discovered that real networks look quite similar in certain ways, but at the same time quite different from purely random graphs. The best-known precursor to modern dynamic graph models is the 1960 publication of Erdős and Rényi [Erdős and Rényi,

1960], which analytically describes the structural properties of randomly generated graphs as the number of vertices (and edges) is increased. The so-called Erdős-Rényi (ER) random graph model was not intended to be realistic models, with the authors noting that "if one aims at describing ... a real situation, one should replace the hypothesis of equiprobability of all [edges] by some more realistic hypothesis" [Erdős and Rényi, 1960]. In 1998, Watts and Strogatz [Watts and Strogatz, 1998] showed that many types of real networks are highly clustered, unlike ER graphs, with short average path lengths relative to comparable ER graphs. This lead to the first realizations of Erdős and Rényi's "more realistic hypothesis" for networks, in the form of the *small-world hypothesis* [Watts and Strogatz, 1998] and the *preferential attachment* model [Barabási and Albert, 1999]. Bibliographic databases were some of the earliest testbeds for these hypothesis. The research questions, however, were rooted in statistical mechanics, and could broadly be classified into the development of generative models to explain observed network characteristics [Newman, 2003], and the analysis of 'critical points' in parameters governing the formation of network, at which networks exhibit sudden drastic changes in properties [Dorogovtsev et al., 2008].

The second development in the late 1990s was the wide-scale growth of the World Wide Web and the adoption of search engines. Commercial web crawlers were indexing the Web at continually increasing scales, leading to possibly the largest dynamic network dataset in existence. In 1998 and 1999, two of the most successful algorithms for ranking webpages were based on network analysis: the HITS algorithm [Kleinberg et al., 1999], and the PageRank algorithm [Brin and Page, 1998]. As a result, a number of algorithmic questions dealing with either static or dynamic network came to the forefront. In addition to the ranking algorithms powering search engines, some of the focus in computer science was on algorithmic issues such as routing in decentralized networks [Kleinberg, 2000, Kumar et al., 2005], targeting influential nodes in a network for marketing or immunization [Kempe et al., 2003, Aspnes et al., 2007, Domingos and Richardson, 2001], and characterizing specific computer networks such as the Internet [Faloutsos et al., 1999], the Web [Kleinberg et al., 1999], and recently, online social networks [Kumar et al., 2006,Hu and Wang, 2009,Ahn et al., 2007].

Finally, we note that at approximately the same time in the mid 1990s, the field of data mining was maturing rapidly, and although it would not be acknowledged till later, a number of techniques developed in data mining would be applicable to dynamic network data. Mining for frequent patterns in large *transactional databases*, also known as 'market basket data', was one of the earliest problems in data mining [Agrawal and Srikant, 1994]. A transactional database records co-occurrences of items, with the standard example being a supermarket retail database that records the items in each customer's 'basket'. The database can then be thought of as a set of subsets drawn from a universal set of 'items'. Although there is generally no notion of order between the transactions in such a database, some algorithms that operate on transactional databases do require an ordering [Özden et al., 1998, Tung et al., 1999].

Dynamic networks are graphs with unique node labels, a property that can be mapped to transactional databases for certain tasks [Lahiri and Berger-Wolf, 2008, Lahiri and Berger-Wolf, 2007]. Since each node within the graph of a single timestep is unique, an edge between any pair of nodes can be identified uniquely. Therefore, all nodes and edges can be uniquely mapped to the set of integers. Each timestep in the dynamic network becomes a transaction in the database, with each vertex and edge in the timestep being converted to an item. The entire dynamic network can then be treated as a transactional database, with a direct mapping between algorithmic tasks such as maximal common subgraph and set intersection [Dickinson

et al., 2003, Lahiri and Berger-Wolf, 2008, Lahiri and Berger-Wolf, 2007]. Note that the mapping is not valid for tasks where graph properties that have no equivalent mapping in set notation, such as connectivity, are required. However, as a benefit, frequent subgraphs can be mined using current tools for frequent itemset mining.

Similarly, sequences of symbols can be treated as discrete time series if they are scanned in a temporal direction. Although sequence data is almost ubiquitous in bioinformatics, it is also used to represent event logs, such as hardware and networks alert logs [Domeniconi et al., 2002, Vilalta and Ma, 2002]. Event logs (or to be precise, *multievent* logs) are slightly more general than sequences, because event logs can often contain more than one symbol or event at each position [Oates and Cohen, 1996, Oates et al., 1997]. With this property, multievent logs can be considered equivalent to an ordered transactional database. An advantage of this mapping is that it also forms a link between dynamic networks, transactional databases, and the well-studied area of machine learning in sequences [Dietterich, 2002]. Although some assumptions made for learning in sequences might not hold for dynamic networks[5], the general techniques are still be applicable to networks. This connection also allows one to capitalize, if needed, on well established algorithms for mining frequently occurring sequential patterns [Pei et al., 2004].

1.2 Summary of contributions

The following is a summary of contributions in this thesis.

1. *(Chapter 2)* A survey of how dynamic networks are measured over time to yield time series of various graph theoretic properties. Specifically, we find that almost all literature on measuring dynamic networks uses one of two aggregation methods, and a handful of simple graph theoretic measures.

   (a) When the commonly *growing network* aggregation method is used, many common trends observed in dynamic networks can be explained by a doubly stochastic sampling process involved in the collection of data, and not as an intrinsic feature of the network itself.

   (b) Using simulations on several network models, we show that networks that exhibit certain temporal properties, when subjected to even a small amount of missing temporal data, erroneously display either a contrary trend, or no trend at all.

   (c) Given a dynamic network dataset, we propose a method using statistical randomization (permutation) tests to determine how likely it is that the properties of the underlying network are in flux under various assumptions.

   (d) We note that measuring dynamic network datasets in the presence of noise and missing data is a difficult issue, and suggest some alternatives. If networks must be measured over time in the presence of noise and missing data, then it is important to either pick a measurement method that is less prone to biases, or to choose measures or methods that are more robust.

2. *(Chapter 3)* The development of a Fourier-like decomposition for detecting periodicity and periodic patterns in dynamic networks.

[5] In particular, the scale and dimensionality (*i.e.,* of the adjacency matrix) of dynamic networks is several orders of magnitude larger than typical multi-event sequences, and the dimensionality may not even be known in advance.

(a) We propose a new mining problem for dynamic networks that involves periodicity detection and is well grounded in theory.

(b) We prove a polynomial upper bound in the size of the input on the number of patterns in a dynamic network that can satisfy our mining criteria. This is in contrast to related periodic pattern mining formulations that are intractable in the worst case and include redundant information in the output.

(c) We describe an efficient polynomial-time algorithm that makes a single pass over the data.

(d) We show how our algorithm extracts both a spectrum of periodicities from the network, as well as the basis patterns that comprise the spectrum.

3. *(Chapter 4)* The development of a technique for finding strong temporal correlations between edges in dynamic networks. A correlation is defined as strong if it holds a degree of predictive power on unseen data.

   (a) We approach the intractability of the general problem by capitalizing on the graph-theoretic properties of real networks. Specifically, the skewed degree distribution of real networks is used to build a tractable dependency structure.

   (b) We describe a problem formulation that works on a continuous time stream of interaction data, without requiring it to be quantized into discrete timesteps. We also describe an evaluation framework for our continuous-time formulation.

   (c) We describe a novel Hidden Markov Model (HMM) formulation that models the time delay between any pair of edges. Using the dependency structure we described earlier, we mine pairs of edges that are best modeled by this HMM.

1.3 Datasets used in this thesis

Dynamic networks can often appear in different guises. For example, 'call graphs' are collected by telecommunications companies in real time, even though the number of customers can number in the millions [Nanavati et al., 2006]. GPS and radio collars allow ecologists to tag wild animals and the social interactions between them [Juang et al., 2002, Fischhoff et al., 2007], resulting in continuous streams of proximity data. In humans, a similar effect is achieved by monitoring connections between Bluetooth-equipped cellphones [Eagle and Pentland, 2006], manually annotated photographs [Lahiri and Berger-Wolf, 2008] or the headers of email traffic [Chapanond et al., 2005, Diesner and Carley, 2005]. In computer networks research, a time-series of labeled graphs representing network traffic has been used as the basis for network intrusion detection [Bunke, 2003, Bunke et al., 2005]. The following is a description of the datasets that are used in various parts of this thesis.

1. The **Enron email network** is inferred from the mailboxes of about 150 employees of the former Enron Corporation [Shetty and Adibi, 2004]. The contents of the mailboxes were publicly released by the Federal Energy Regulatory Commission in the course of investigations into the workings of the company, and consist of the full text and headers of emails, both sent and received. This allows us to

construct a partial, dynamic view of email communications within a large, complex organization like Enron, and has spurred much research in various areas [Diesner and Carley, 2005, Chapanond et al., 2005, Shetty and Adibi, 2004]. Each vertex represents an email address, with a directed edge from the sender of an email to all its recipients. The timestep quantization is one day, although arbitrarily smaller quantizations are also possible due to the presence of full email header information.

2. The **Plains Zebra** and **Grevys Zebra** networks are observations of social associations in two species of wild Zebra in Kenya [Fischhoff et al., 2007,Sundaresan et al., 2007]. They are currently collected by direct visual observations made by ecologists, although more advanced and accurate methodologies like radio and GPS collars are being investigated [Juang et al., 2002]. The manual collection of interaction data results in missing data, but for Grevys Zebra, the missing data rate is estimated to be under 50%. Both species of Zebra are *fission-fusion* species, which means that they come together in groups that subsequently dissolve to form new groupings. An analysis of dynamic communities has confirmed differences in the grouping habits of the two species [Tantipathananandh et al., 2007]. Each vertex represents an individual Zebra, identified by the pattern of stripes on specific parts of its body, and an edge represents a social association as determined by ecologists. The timestep quantization is one day, which corresponds to the frequency of the ecologists' observation rounds.

3. The **IMDB Photo Network** is collected from metadata about people tagged in photographs on the Internet Movie Database (IMDB) [Lahiri and Berger-Wolf, 2008]. The photographs are generally of actors, musicians, directors and other people associated with the entertainment industry, and may either be candid or professional shots. Since IMDB is not a general photo-sharing site and its pictures are labeled by staff members, one might reasonably assume that the people tagged in photos are 'celebrities' of some sort and that a degree of social association exists between them. This methodology is quite similar to the Zebra sighting datasets, and the observed structure is a partial view of the true set of interactions. Metadata on a total of 194,430 photographs were collected, with about 75,000 photographs containing more than one person. Each vertex corresponds to a manually identified and disambiguated person (conducted either by IMDB or professional photo agency staff), with an edge representing co-appearance in a photograph. The discretization timestep is one day.

4. **Reality Mining** was an experiment conducted at MIT to collect a variety of data related to movements and social dynamics in humans (specifically, students at MIT) [Eagle and Pentland, 2006]. It involved equipping volunteers with cellphones augmented with special tracking software. Among the recorded data was physical proximity data inferred from two cellphones being able to establish a direct Bluetooth connection, with the maximum range for such a connection being approximately 30 ft. [Bluetooth SIG, Inc., 2009] Each vertex represents a study participant with a Bluetooth-equipped cellphone, and an edge represents physical proximity of less than 30 ft. The quantization timestep is four hours.

5. **Call Detail Records-C** is a Call Detail Record (CDR) dataset that is collected whenever a phone subscriber makes a call to another telephone number. In the process, information such as the originating and destination number (encrypted), and the time and date of the call are logged. We obtained 4 months of CDRs from mobile phone subscribers in a dense urban area. For this dataset, we only considered subscribers that made at least 3 phone calls per day, and included *all* successful phone calls

made or received during the observation period. We treat a phone call between two subscribers as an undirected edge, because phone conversations, as opposed to phone calls, are inherently bi-directional. The number of nodes in the network and the number of phone calls recorded are on the order of 10^5 and 10^7 respectively. We are unable to disclose further details due to privacy considerations.

6. **Call Detail Records-J** is similar to the CDR-C dataset, but includes CDRs from all phone users of a large geographical region (an entire state) of a particular country, for a period of 5 months, without any sampling bias as in CDR-C. It is also unlikely that there is significant overlap between the customers in CDR-J and CDR-C due to the geographical separation between the regions. The number of nodes and interactions are on the order of 10^6 and 10^7 respectively.

1.4 Declaration of prior published work

Parts of Chapters 3 and 4 appear in the following publications:

1. M. Lahiri and T.Y. Berger-Wolf. *Periodic subgraph mining in dynamic networks.* Knowledge and Information Systems, Volume 24, Issue 3 (2010), p. 467.
2. M. Lahiri and T.Y. Berger-Wolf. Mining Periodic Behavior in Dynamic Social Networks. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, Pisa, Italy. December 2008.
3. M. Lahiri and T.Y. Berger-Wolf. Structure Prediction in Temporal Networks using Frequent Subgraphs. In *Proceedings of the IEEE Computational Intelligence and Data Mining Conference (CIDM 2007)*, Honolulu, Hawaii. April 2007.

Bibliography

[Achlioptas et al., 2009] Achlioptas, D., Clauset, A., Kempe, D., and Moore, C. (2009). On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *J.ACM*, 56(4):1–28.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. In *Proc. of the 20th Intl. Conf. on Very Large Data Bases*, pages 487–499, San Francisco, CA. Morgan Kaufmann Publishers Inc.

[Ahn et al., 2007] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In *Proc. of the 16th Intl. Conf. on World Wide Web*, pages 835–844, New York, NY, USA. ACM.

[Aspnes et al., 2007] Aspnes, J., Rustagi, N., and Saia, J. (2007). Worm versus alert: Who wins in a battle for control of a large-scale network? *Principles of Distributed Systems*, pages 443–456.

[Barabási and Albert, 1999] Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509.

[Bilgic and Getoor, 2008] Bilgic, M. and Getoor, L. (2008). Effective label acquisition for collective classification. In *Proc. of the 14th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 43–51. ACM.

[Bluetooth SIG, Inc., 2009] Bluetooth SIG, Inc. (2009). Bluetooth: Technical comparison. `http://www.bluetooth.com/Bluetooth/Technology/Works/Compare/Technical/`.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proc. of the seventh Intl. Conf. on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

[Bunke, 2003] Bunke, H. (2003). Graph-based tools for data mining and machine learning. *Lecture Notes in Computer Science*, pages 7–19.

[Bunke et al., 2005] Bunke, H., Dickinson, P., Irniger, C., and Kraetzl, M. (2005). Analysis of time series of graphs: Prediction of node presence by means of decision tree learning. In *Proc. of the 4th Intl. Conf. on Machine Learning and Data Mining in Pattern Recognition*, volume 3587, pages 366–375. Springer.

[Chakrabarti et al., 2010] Chakrabarti, D., Faloutsos, C., and McGlohon, M. (2010). Graph mining: Laws and generators. *Managing and Mining Graph Data*, pages 69–123.

[Chapanond et al., 2005] Chapanond, A., Krishnamoorthy, M. S., and Yener, B. (2005). Graph theoretic and spectral analysis of Enron email data. *Comput. Math. Organ. Theory*, 11(3):265–281.

[Costenbader and Valente, 2003] Costenbader, E. and Valente, T. (2003). The stability of centrality measures when networks are sampled. *Social networks*, 25(4):283–307.

[Dickinson et al., 2003] Dickinson, P. J., Bunke, H., Dadej, A., and Kraetzl, M. (2003). *On Graphs with Unique Node Labels*, volume 2726 of *Lecture Notes in Computer Science*, pages 409–437. Springer Berlin.

[Diesner and Carley, 2005] Diesner, J. and Carley, K. M. (2005). Exploration of Communication Networks from the Enron Email Corpus. In *Proc. of the 2005 SIAM Wkshp. on Link Analysis, Counterterrorism and Security*, pages 3–14.

[Dietterich, 2002] Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Proc. of the Joint IAPR Intl. Wkshp. on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, London, UK. Springer-Verlag.

[Domeniconi et al., 2002] Domeniconi, C., Perng, C.-S., Vilalta, R., and Ma, S. (2002). A classification approach for prediction of target events in temporal sequences. In *Proc. of the 6th European Conf. on Principles of Data Mining and Knowl. Disc.*, pages 125–137, London, UK. Springer-Verlag.

[Domingos and Richardson, 2001] Domingos, P. and Richardson, M. (2001). Mining the network value of customers. In *Proc. 7th ACM SIGKDD*, pages 57–66.

[Dorogovtsev et al., 2008] Dorogovtsev, S., Goltsev, A., and Mendes, J. (2008). Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4):1275–1335.

[Eagle and Pentland, 2006] Eagle, N. and Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

[Erdős and Rényi, 1960] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *Proc. of the Conf. on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY. ACM.

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

[Freeman, 2003] Freeman, L. (2003). Finding social groups: A meta-analysis of the southern women data. In *Dynamic Social Network Modeling and Analysis: workshop summary and papers*, pages 39–77.

[Ghani et al., 1998] Ghani, A., Donnelly, C., and Garnett, G. (1998). Sampling biases and missing data in explorations of sexual partner networks for the spread of sexually transmitted diseases. *Statistics in medicine*, 17(18):2079–2097.

[Hu and Wang, 2009] Hu, H. and Wang, X. (2009). Evolution of a large online social network. *Physics Letters A*, 373:1105–1110.

[Juang et al., 2002] Juang, P., Oki, H., Wang, Y., Martonosi, M., Peh, L. S., and Rubenstein, D. I. (2002). Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet. *ACM SIGPLAN Notices*, 37(10):96–107.

[Kempe et al., 2003] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proc. 9th ACM SIGKDD*, pages 137–146.

[Kleinberg et al., 1999] Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). The web as a graph: Measurements, models, and methods. In *Proc. of the 5th annual Intl. Conf. on Computing and combinatorics*, pages 1–17. Springer-Verlag.

[Kleinberg, 2000] Kleinberg, J. M. (2000). Navigation in a small world. *Nature*, 406(6798):845.

[Kossinets, 2006] Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28(3):247–268.

[Kumar et al., 2005] Kumar, A., Xu, J., and Zegura, E. (2005). Efficient and scalable query routing for unstructured peer-to-peer networks. In *INFOCOM 2005: Proc. of the 24th Annual Joint Conf. of the IEEE Computer and Communications Societies*, volume 2, pages 1162 – 1173 vol. 2.

[Kumar et al., 2006] Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *Proc. of the 12th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 611–617. ACM New York, NY, USA.

[Lahiri and Berger-Wolf, 2007] Lahiri, M. and Berger-Wolf, T. Y. (2007). Structure prediction in temporal networks using frequent subgraphs. In *Proc. of IEEE Symposium on Computational Intelligence and Data Mining*, pages 35–42.

[Lahiri and Berger-Wolf, 2008] Lahiri, M. and Berger-Wolf, T. Y. (2008). Mining periodic behavior in dynamic social networks. In *Proc. of the IEEE Intl. Conf. on Data Mining*, pages 373–382.

[Lahiri et al., 2011] Lahiri, M., Tantipathananandh, C., Warungu, R., Rubenstein, D., and Berger-Wolf, T. (2011). Biometric animal databases from field photographs: Identification of individual zebra in the wild. In *Proc. of the ACM Intl. Conf. on Multimedia Retr.* ACM Press.

[Moreno, 1934] Moreno, J. L. (1934). *Who Shall Survive?* Nervous and Mental Disease Publishing Company, Washington, D.C.

[Nanavati et al., 2006] Nanavati, A. A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., and Joshi, A. (2006). On the structural properties of massive telecom call graphs: findings and implications. In *Proc. of the 15th ACM Intl. Conf. on Information and knowledge management*, pages 435–444, New York, NY, USA. ACM.

[Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

[Newman and Leicht, 2007] Newman, M. E. J. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proc. of the National Academy of Sciences*, 104(23):9564.

[Oates and Cohen, 1996] Oates, T. and Cohen, P. R. (1996). Searching for structure in multiple streams of data. In *In Proc. of the Thirteenth Intl. Conf. on Machine Learning*, pages 346–354. Morgan Kaufmann.

[Oates et al., 1997] Oates, T., Schmill, M., Jensen, D., and Cohen, P. (1997). A family of algorithms for finding temporal structure in data. In *6th Intl. Wkshp. on A.I. and Statistics*.

[Özden et al., 1998] Özden, B., Ramaswamy, S., and Silberschatz, A. (1998). Cyclic association rules. In *Proc. of the Fourteenth Intl. Conf. on Data Engineering*, pages 412–421, Washington, DC, USA. IEEE Computer Society.

[Pei et al., 2004] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. (2004). Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440.

[Shetty and Adibi, 2004] Shetty, J. and Adibi, J. (2004). The Enron email dataset database schema and brief statistical report. Technical report, Information Sciences Institute, University of Southern California.

[Sundaresan et al., 2007] Sundaresan, S. R., Fischhoff, I. R., Dushoff, J., and Rubenstein, D. I. (2007). Network metrics reveal differences in social organization between two fission–fusion species, Grevy's zebra and onager. *Oecologia*, 151(1):140–149.

[Tantipathananandh et al., 2007] Tantipathananandh, C., Berger-Wolf, T. Y., and Kempe, D. (2007). A framework for community identification in dynamic social networks. In *Proc. of the 13th ACM SIGKDD Intl. Conf. on Knowl. disc. and data mining*, pages 717–726.

[Tong et al., 2010] Tong, H., Prakash, B., Tsourakakis, C., Eliassi-Rad, T., Faloutsos, C., and Chau, D. (2010). On the vulnerability of large graphs. In *2010 IEEE Intl. Conf. on Data Mining*, pages 1091–1096. IEEE.

[Tukey, 1980] Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25.

[Tung et al., 1999] Tung, A. K., Lu, H., Han, J., and Feng, L. (1999). Breaking the barrier of transactions: mining inter-transaction association rules. In *Proc. of the 5th ACM SIGKDD Intl. Conf. on Knowl. disc. and data mining*, pages 297–301, New York, NY, USA. ACM.

[Vilalta and Ma, 2002] Vilalta, R. and Ma, S. (2002). Predicting rare events in temporal domains. *2nd IEEE Intl. Conf. on Data Mining*, page 474.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[Zachary, 1977] Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473.