Georgetown University Institutional Repository
http://www.library.georgetown.edu/digitalgeorgetown

The author made this article openly available online. Please tell us how this access affects you. Your story matters.

Challenges for dynamic, heterogeneous networks in observational sciences

L. Singh. "*Challenges for dynamic, heterogeneous networks in observational sciences.*" Next Generation of Data Mining. Ed. Hillol Kargupta. Florida: CRS Press, 2008: 395-414.

Collection Permanent Link: http://hdl.handle.net/10822/761593

© CRC Press

Copyright 2008 From Next Generation of Data Mining by Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, Vipin Kumar, editors. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

This material is made available online with the permission of the author, and in accordance with publisher policies. No further reproduction or distribution of this copy is permitted by electronic transmission or any other means.

Chapter 20

Challenges for Dynamic Heterogeneous Networks in Observational Sciences

Lisa Singh

Contents

20.1 Introduction

In corporate and scientific domains, there exist much data containing interrelated entities that are linked together. While this data can be analyzed by assuming independent instances and ignoring the relationships, the connectivity, relational structure, and associated dependencies can provide valuable insights, leading to potentially interesting data mining results with clearer semantics and higher degrees of accuracy. Different representations exist for domains with large, feature-rich entities and relationships. One natural representation that we consider in this chapter for this interconnected data is a graph. In an age of information overload, graphs give us a tangible, interpretable construct that allows us to more readily incorporate relationships into our analysis. We use graphs as the basis for describing models and abstractions that are useful for state-of-the-art visual mining of public and private graph data.

But graphs are still a very general construct. What should these graphs look like? How much detail is useful in the graph structure for data mining, more specifically visual mining? How much detail should be perturbed or removed in order to maintain privacy of individuals in a network? Even though simple graphs are easier to analyze and interpret, more complex graphs are sometimes more beneficial for sophisticated graph-mining tasks and domain-specific analysis. For example, using multiple node or relationship types allows for clearer semantic interpretations of clusters and associations, incorporates both node features and graph structural properties when building predictive models, and enables multiple abstractions of the data to help preserve privacy of individuals and support visual analytics of large graphs.

In this chapter, we consider graph representations and abstractions for dynamic, heterogeneous networks. A large number of domains contain dynamic, heterogeneous networks with multiple edge types and relationship types over time, for example, communication networks, protein interaction networks, social networks, transportation systems, and observational scientific networks. Here, we focus our discussion on observational scientific networks. These networks have a number of challenges for future researchers developing data mining algorithms and visual analytic tools, including high dimensionality, varying degrees of observational certainty, incomplete data, and highly fluid, dynamic network structures.

We begin by describing the semantics of the data associated with observational scientific data sets in the next section. We then discuss a generic graph model that can incorporate constraints useful for visual and graph mining of data generated by observational scientists. Section 20.3 presents the current state of visual analytics and mining as it relates to graph structures and describes both algorithmic and visualization challenges. In Section 20.4, we switch gears and consider situations were the observed graph data need to remain private. This is a difficult problem since there are many unique features of a graph. We formulate the problem and constraints for privacy of graphs and identify some of the early work in that arena. Finally, Section 20.5 presents some concluding thoughts.

20.2 Observational Science Motivation

20.2.1 Background Scenario

Observational data is prevalent in many fields including biology, sociology, medicine, and psychology. We begin by considering a simple observational science data set where researchers monitor a subject for a specified period of time. Example subjects include wild animals, humans, and planets. Each monitoring period can be viewed as an event consisting of a number of observations. Events include tracking a group of animals for a 30 min period, conducting a 30 min psychological evaluation of a person, and taking a snapshot every minute of the interaction between a planet and its moons. During a single event, researchers watch the subject or group of subjects, taking notes throughout the process. Sometimes, computers are used to record observations, but hand notes are common. These observational data sets tend to contain a large number of events/observations (thousands to millions) and features (hundreds to thousands) for a small number of subjects (tens to thousands).

It is also not unusual for photos and videos to accompany the more structured data. These data are used for identification of subjects in the wild. For example, if researchers know that certain monkeys are seen regularly together and suddenly the group composition changes, photos can be used to see if bites or injuries on a subject may have caused misclassification or if the community structure actually changed. Using network connectivity in conjunction with image data can help increase the quality of observational results and even help correct errors related to potential duplication. This example also highlights an important feature of observational data—many measurements in the data contain a time element. This time dimension is important for exploring questions regarding community stability and group formation: (1) How stable are these complex graph structures over time? (2) Which modes or relationships change most frequently? (3) What is the topological difference between a multi-featured community at time t_1 and t_2? (4) Why do some communities grow while others get smaller? For more details on long-term animal studies, we refer you to studies on dolphins and whales [1], pronghorn [2], and chimpanzees [3].

20.2.2 Data Representation

The majority of graph-mining algorithms (see Refs. [4,5] for surveys) and visual mining tools (see Ref. [6] for an overview) designed for graphs are developed for unimode, unirelation networks, where each node represents an object of a single type (e.g., an actor/subject, a Web page, or an observation) and each edge represents a relationship of a single type between two nodes in the network (e.g., friendship, kinship, or coauthorship) [7]. More formally, for a graph G containing n nodes or vertices, $G = (V,E)$, where $V = \{v_1,v_2,\ldots,v_n\}$ and $E = \{(v_i,v_j)\,|\,v_i, v_j \in V, i \neq j, 1 \leq i,j \leq n\}$. Here, V represents a set of nodes of a single type and E represents the set of relationships or edges of a single type, where each edge is defined as a pair of nodes from the set V.

However, in the simple example described in Section 20.2.1, four types of objects exist (observer/scientist, subject, events, and observations within an event), each of which can be viewed as a different vertex or node type. Also, both observations and events may link to more than one subject at a particular time. Therefore, an extended network is necessary to capture multiple node types and multiple edge types. Singh et al. introduced the M^*3 model as the basic data model for Invenio, a visual mining tool for social networks with multiple node types and multiple edge types [8]. In this model, a relation exists for each node type and each edge type. The relations are semantically organized based on actors and events, similar to a traditional affiliation network studied by sociologists. While there are many benefits to the M^*3, there are still some limitations. It does not allow for complex data types (e.g., texts, photos, and videos). Nodes are simple objects and features of nodes and relationships are well structured. It also does not incorporate semantics for time-varying networks. While time can easily be represented as an attribute in the M^*3 model, the model does not support special semantics for aggregation and sophisticated time-varying analysis of dynamic features.

Therefore, we advocate a generic model that captures multiple node and object types (multi-mode), multiple edge types (multirelation), and multiple static and time-varying descriptive features (multi-feature) associated with each. Formally, a more generic graph $G^{'}$ that captures any number of node types and edge types can be represented as $G^{'} = \{\mathbf{V},\mathbf{E}\}$, where $\mathbf{V} = \{V_1,\ldots,V_{n_{\mathrm{VS}}}\}$, $\mathbf{E} = \{E_1,\ldots,E_{n_{\mathrm{ES}}}\}$, and n_{VS} and n_{ES} are the number of node sets and edge sets in $G^{'}$, respectively. The base case remains a unimode, unirelation graph.

A multi-feature graph has features or attributes associated with each node and each edge in the graph. Some nodes such as the observation event nodes are temporal and will also have a time stamp and a duration time associated with them. This will be crucial for supporting longitudinal graph-mining analysis.

Figure 20.1 shows an example observation network. Each node type is shown as a different shape. Researchers are circles, observation events are squares, observations within an event are diamonds, and subjects are triangles. Other possible node types for these data include photos, geographic locations, and audio clips. For clarity, we show only one link type between different node types and no links between nodes of the same type. In reality, many different link types may exist between any two nodes in the graph. Each node and each edge in Figure 20.1 also has a set of features or attributes associated with it. For example, the animal subject nodes may have the following attributes: birth date, name, gender, mother, father, and pregnancy date for females.

There are a number of advantages to use a graph model with numerous entity and relationship types. First, users can represent a rich feature set for nodes and edges. Next, underlying relational theory supports complex graph structures, allowing users to translate between different graph topologies and abstraction levels. This means

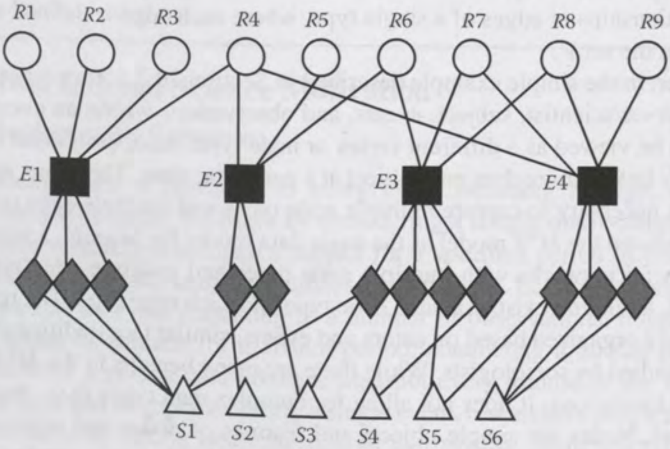

FIGURE 20.1: Observational network with four node types: researchers (circles), events (squares), observations within events (diamonds), and animal subjects (triangles).

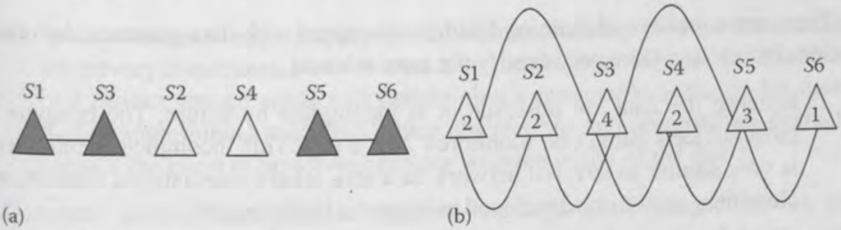

FIGURE 20.2: Unimode projections of Figure 20.1: (a) connections between animals observed exhibiting the same behavior when observed together and (b) connections between animals seen together in the same event.

that we can use relational algebra to easily create a subgraph that is projected on a single node type. For our example, we may want to create a unimode network of animal subjects. In this case, each node is a particular animal subject and each link represents animals that are observed together. Figure 20.2 shows two example projections. Figure 20.2a shows connections between animal subjects that have been observed together exhibiting the same behavior. Nodes that have a connection are highlighted in gray. Notice that there are few relationships in this unimode, unirelation representation. Figure 20.2b shows the connections between animal subjects that have been observed together in the same event. Here, all the animals have been seen with at least one other animal. Inside each node is the number of connections an animal has to other animals in the network. This network view highlights the link structure of the animal subject network, without displaying the structure of events or observations.

20.2.3 Challenging Data Features

While one can focus on any single node type, integrating the data can help explore methodological questions, data-quality issues, and graph-mining research questions that are obscured in graphs containing only one node type and one edge type. For example, are certain subjects oversampled by certain researchers? Answering this question requires investigation of three node types: researchers, events, and subjects.

Researchers within disciplines involving observational sciences are interested in individual behaviors as well as group dynamics. Because of the large number of features collected by observational researchers, targeted data exploration is necessary. We have observed that visualization can enhance the process considerably. Researchers can then visually see the results of sophisticated graph-mining algorithms and determine whether or not clusters and predictions coincide with their domain knowledge. They can use visual analytics to enhance their understanding of community structure and alliance creation, information transmission or behavior propagation through the community, and synergies between genetic relationships and community substructure. Further, observational scientists can use advanced statistical analysis on robust graph structures to visually discover patterns that may be difficult to interpret when looking at the statistical results alone.

There are a number of defining features associated with data generated by observational scientists. Here, we identify the most relevant.

1. Because the data are observed, it is incomplete by nature. The behavior of most subjects cannot be monitored 24 h a day. This incomplete ground truth is very similar to terrorist network data sets, where associations, clusters, and classifiers need to be developed using partial information.

2. Many of the observed attributes have an element of uncertainty associated with them. It can occur that a researcher monitoring animals in the field may be uncertain about a behavior that has occurred (e.g., Was animal Sl petting or hitting?) or the individuals involved in the behavior (e.g., Was that S2 or S3 that was jumping?). Are the observations reliable? Using the observational certainty information, is there a relationship between the field condition or the animal subject and the quality of the observation? These uncertainties need to be considered so that confidence in associations, clusters, and predictive models are tempered and new ways to understand and interpret the results are considered.

3. The data are dynamic. Many longitudinal studies of people or animals already have decades of data. Therefore, they are well suited for investigating changing community structures or behaviors over time. Information transmission can also be analyzed in this context. Are behaviors taught based on social relationships or are they self-learned? How are community structures of prominent subjects, exhibiting given behaviors, changing over time? How does their community structure compare to the norm of the entire population? Do certain patterns of behavior occur from generation to generation?

4. These data are feature rich and heterogeneous. The dimensionality is large compared to other social network data sets (e.g., blogs and e-mail). Which attributes and relationships are most relevant for pattern discovery, clustering, and classification? Which ones are noisy? How do we integrate knowledge from diverse types of data, for example, minute by minute focal data and snapshot survey data, for mining applications?

5. Researchers have observational bias. When building predictive models from real data, the sample used is very important. Machine-learning researchers use different techniques to reduce bias in samples that are used to train classifiers. If there are a large number of researchers monitoring subjects, observational bias may be very subtle and difficult to detect. Do some observers have favorites that are oversampled in the population? When there are teams of observers, is the terminology consistent across the team, for example, does "large" mean the same size to all the observers?

6. As more data are collected, graphs will continue to grow. While the number of subjects in the network may be small, the number of observations, relationships, and attributes tend to be larger than blog, e-mail, or communication networks. How do we reduce the graph size prior to exploration while maintaining the properties necessary for accurate analysis?

7. Privacy concerns may not apply to wild animals, but they do to patients. If there are privacy constraints, how must we alter the graph to decrease the likelihood of a privacy breach while still maintaining a reasonable accuracy for meaningful graph-mining analysis? Which abstractions and perturbation strategies balance the goals of preserving privacy with the utility of the data?

Computer science researchers in machine learning, data mining, statistics, and graph theory have begun developing approaches that consider some of these issues. What is missing is an integrated environment that facilitates data exploration as a dynamic, iterative process. An important component of this environment is a sophisticated visual-mining component that incorporates interactive visual exploration of complex data mining results.

20.3 Visual Mining of Complex Graphs

One reason why we like graphs is because they are naturally visual. We can look at them and begin interpretation before any formal analysis begins. While the visualization community has come up with some exceptional visual representations of networks (see Freeman survey [9] and Keim overview [10] for more details), visualization is different from visual analytics and visual mining. Visualization shows a view of the data and provides certain insight resulting from the interface design and layout decisions. The benefits are related to the old cliche, "a picture is worth a thousand words." In contrast, visual mining combines visualization, data mining, and other analytic techniques to support advanced data-exploration tasks [11]. The goal is to help users sift through and manipulate the data more effectively, particularly large or complex data sets, to better understand the data space. Visual analytic and mining tools may serve to help identify and interactively explore clusters or groups, find common structures in the network, compare roles of different classes of people, and visually analyze changes to community structures over time. These capabilities parallel the needs identified in the last section for observational scientists.

20.3.1 Visual Analytic Tools

Over the last few years, numerous toolkits and tools have been developed that use visualization to help represent patterns and results from mining algorithms graphically. Tools fall into two categories, those that have sophisticated statistical analysis using matrix operations [12–14] and those that focus on interactive visualization of unimode networks [15–19] and multi-mode, heterogeneous networks [8,20–22].

The first category of tools is less interactive; their strength is the sophisticated statistical calculations. For example, UCINet calculates social network centrality measures, permutation-based statistics, matrix algebra, and multivariate statistics [13]. StOCNET calculates some similar metrics, but focuses on using stochastic methods to analyze longitudinal data [14]. Finally, Pajek contains sophisticated blockmodeling to support analysis of large networks [12].

The more interactive tools lack the sophisticated statistical analysis, but allow users to switch between different visual layouts and levels of data detail with ease. Some tools look at the entire network, while others focus on a piece of the network. Views of graph data range from the more traditional node–edge views similar to Figure 20.1 [8,15,17,18,20,21] to tree-structure representations of subgraphs in the network [19], maps, histograms and nested rectangles [16], and hybrids [22].

Various toolkits have also been developed to help programmers create interactive visualizations and visual-mining tools themselves. The most robust include JUNG [23], Prefuse [24], Piccolo [25], and GUESS [26]. These toolkits support the rapid development of graphic-based applications. JUNG has the largest support for graph mining and path algorithms. Prefuse contains a large number of visual layouts for graphs. Piccolo allows developers to create applications that incorporate different visual and graphic features beyond network-based visualizations with ease. Finally, GUESS contains database support and a simple query language that can be used to focus on analysis and easily change the perspective of a graph.

Even though the tools we have described identify patterns in the graph, zoom in on interesting groups or clusters in the graph, and allow for interactive exploration of the graph, we are missing visual analytic tools that allow us to answer sensitivity or what-if questions: What if we remove this node from the network? What if we add a new member to this clique? What nodes need to be added or removed so the distribution of node degree approximates a power law distribution? These questions are very relevant to group and network stability, information flow, and network characterizations.

While a wealth of tools exist for visualization of graph structures, approaches that integrate measures and algorithms from graph theory with interactive visualizations are still in its infancy. Therefore, the next step is to tightly integrate data mining and visualization so that user-selected subsets and abstractions of the graph can be used as inputs into different data mining algorithms and intermediate results can be visually explored and manipulated, iteratively. Accomplishing this using a flexible framework with a high degree of interactivity that supports a range of analytic tasks is still an open question. One reason is that a platform of this magnitude transcends cutting edge research in many areas of computer science, requiring sophisticated data management and indexing schemes, robust software engineering design, alternative visual paradigms, and scalable statistical and graph-mining procedures.

20.3.2 Developing Metrics for Understanding Complex Structures

Many measures have been developed for characterizing structural or topological features of a graph. Centrality measures are those that are node specific. Some of the more well-known ones include degree (the number of connections an arbitrary node v_i has to other nodes in the network), clustering coefficient (an indicator of the number of neighbors of node v_i that are connected to each other), betweenness (a metric based on the number of shortest paths going through node v_i), and eigenvector (a metric that determines v_i's relative importance in the network based on the importance of its neighbors). More general graph invariants include the number of nodes, the number of edges, the graph density, and the diameter of the graph (the longest shortest path in the graph) [7].

While all these metrics and properties are very important for understanding certain abstraction levels of the graph, not all nodes and edges have the same intrinsic properties in heterogeneous graphs. Understanding the relationship between different node types (modes) and relationship types is necessary to understand the detailed dynamics of the network and the effects of specific graph invariants and properties within and across modes. Therefore, in order to mine heterogeneous networks, we need to extend the unimode, unirelation measures to meaningful multi-mode, multirelation measures. As an example, Singh et al. developed a measure for multi-mode hop expansion [27]. This centrality metric identifies neighborhood size of different modes at different distances from each node v_i. This information can be used to see which nodes have similar positions in the network in terms of distances to other nodes in different modes.

Building on this work, we suggest developing measures that take varying edge types, node types, and feature distributions into consideration. Possible examples include multiedge path length (the number of edges traversed between two nodes in a multi-relational path), strength of connections (frequency and duration of different types of interactions), multi-mode density (a measure of the number of edges connecting nodes in different modes as a fraction of the total number of edges in the graph), transmission rate (path lengths of feature value expansion through different modes), and network turnover (a longitudinal measure that captures affiliation changes over time).

With these topological metrics, we can use the structure to understand the growth distribution of temporal networks. We can analyze the effect of growth rates in one mode on growth rates of others and consider the relationship between attribute features and structural properties of the network. While physicists have been studying the dynamics of network formation and growth [28,29], only the simplest of models are understood. Those findings need to be extended to more complex structures and specialized metrics for measuring individual local community and global network statistics will be vital for exploration of heterogeneous graphs. As we describe in the next section, for large graphs, these metrics can also be useful for eliminating irrelevant or noisy parts of the graph prior to execution of graph-mining algorithms. We now consider possible preprocessing of large graphs.

20.3.3 Preprocessing Prior to Visualization

It is difficult to interpret a graph with a large number of nodes and edges. If the network is well connected, the initial visualization using a traditional layout may appear to be a large black ball containing hundreds if not thousands of nodes. Statistics of node distributions and connectivity structure are useful, but visual analysis of such a large graph is limited. Therefore, initial preprocessing of the graph to identify important components of the graph, hiding irrelevant data, for the specific task of interest makes the problem more tractable [30]. Many techniques have been used to find the important nodes including the following: using structural metrics that capture important attributes of the graph [31,32], using clustering algorithms and blockmodeling to decompose large structures on attribute features or graph-link information [33], applying known compression techniques to graph structures [34], and using graph-matching techniques to identify substructures of interest [35].

Several proposed methods for classification of network objects consider the link structure of the network (see Getoor survey [36]). When considering link-based approaches for prediction in heterogeneous networks, the logical relationship between objects and the probabilistic dependencies between attributes may cause a huge search space for subgraph mining. By removing some of the less relevant components of the data, we can improve predictive accuracy of classifiers and extract smaller, more meaningful clusters and graph substructures from the data, enabling analysts to focus on the most meaningful set of patterns.

In previous work, we showed that predictive accuracy can be maintained on affiliation network (two-mode network) objects if instead of random pruning, the network under consideration is pruned or reduced in size based on attribute values or structural properties like degree and betweenness [32,37]. The goal here was to maximize predictive accuracy on attributes of event nodes in affiliation networks. Standard classifiers were built using pruned networks, where edges were removed based on centrality measures of nodes in the network, feature values, and random sampling. Two affiliation-network data sets were analyzed, an executive corporation network containing board of directors information for a subset of companies traded on the NASDAQ and the NYSE, and an author publication network based on the ACM SIGMOD anthology. For both data sets, pruning on descriptive attributes and graph invariants outperformed random pruning. Further, the underlying networks created using these pruning strategies had significant differences, meaning that each strategy was optimized differently. This finding is also consistent with Airoldi and Carley [38]. They found that pure network topologies are sensitive to random sampling.

Still, further investigation is necessary to determine the role topological structure plays in dynamic, heterogeneous networks in terms of graph-mining accuracy. Is structure less of a predictive indicator for multi-mode networks? Given that no one pruning approach will work across data sets, we suggest selecting a pruning approach prior to subgraph extraction and classification based on local, structural graph invariants (hop expansion, clique structure, clustering coefficient, etc.) and node-specific feature measures (behaviors, gender, lineage, etc.). We can then compare the structural similarities and the predictive accuracies of these pruned networks to the full network to better understand the strengths and weaknesses of the different pruning strategies on networks with specific topological structures.

20.3.4 Graph Mining Applicability to Observational Sciences

Many algorithms have been developed to uncover community substructure, flow of information, and prominent node identification on unimode networks with few, if any, features associated with each node. In the case of unimode networks, various local and global network statistics are used to help interpret network relationships, influence, clusters, or flows. For a recent survey, see Newman [30].

There are a number of problems being explored in the graph-mining community that can benefit from interactive visual analytics and are applicable to research conducted by observational scientists. Here, we describe the most relevant and explain their significance to this domain.

Hidden community identification or group clustering. This problem involves identifying groups of individuals or clusters of individuals that interact frequently together or share some common properties [39–44]. For example, if a scientist is observing a community of monkeys, he may identify a few pockets of 5–10 monkeys that play together regularly. Other significant features may be how often they play or the location where they play. Different unsupervised algorithms have been proposed for identifying hidden communities, including approaches based on subgraph identification, exhaustive search, and greedy heuristics. This problem also has similarity to the graph cut problem [45]. A tangential line of work investigates changes in community structure over time [46–48] and longitudinal network analysis that examines changes to network structural properties over time [7,49]. The networks being used for the analysis contain a single node and a single edge type. We are also interested in extracting unknown or hidden substructures across features and relations. To this end, Cai and colleagues [50] used a linear combination of weighted matrices for each relation to extract unknown community substructures. Because of the volume of data that needs to be analyzed when multiple relations are combined, the preprocessing options discussed in Section 20.3.3 are very relevant here.

Information diffusion and transmission. Here researchers investigate how information spreads through a network [29,51–53]. These papers attempt to find the most influential nodes in the network. How fast will information disseminate if we tell the right people? Who are the right people? Two well-known applications are disease transmission and viral marketing. For animals, behavior transmission is an important application. Approaches included using a global, probabilistic model [51], using a diffusion process that begins with an initial set of active nodes and adds neighbors based on different weighting schemes [52], and using graph invariants [53].

Group formation. The growth of communities in a network and the likelihood that individuals will join a particular community has dependencies to the underlying network structure. Given members in groups, what are the structural features that influence whether an individual will join a particular group? Can we use topological information to determine whether a group will grow or whether the group focus will change? The groups and communities can be viewed as subgraphs of the network, growing and overlapping in complex ways [48,54,55].

Detecting and matching subgraph patterns. In observational sciences that monitor groups of animals, researchers are interested in matching patterns of animal groups based on subgraph structure and feature distributions of nodes and links in a network. For example, does the calf's network emulate that of her mother and is it dependent on the sex of the offspring? The subgraph isomorphism problem looks at matching exact graph structures between two different graphs [56]. Many algorithms have been developed for finding subgraph patterns in massive unimode graphs. Approaches include greedy algorithms [57] and indicative logic programming approaches [58]. For visual exploration, it is important to integrate these approaches with meaningful visualization that can help detect stable subgraphs for more dynamic networks.

Understanding network growth. Many social networks, including the World Wide Web, follow a power law growth distribution [28,59,60]. In order to better understand the growth distribution of temporal heterogeneous networks created from observational scientific data, we need to analyze how growth rates of one relationship or mode affect the growth rates of others and consider how different attribute features relate to the structural properties of the network. How do growth rates of these more complex networks compare to growth rates of known unimode, unirelation networks?

20.4 Complex Social Networks and Privacy

As if the horizon was not complicated enough with the heterogeneity of the data, the dynamic nature of the data, and the graph mining and visual analytic complexities of working with large observational data, a need sometimes exists to keep the identities of the individuals in the data private. While wild animals do not have any well-established privacy rights, human subjects have varying levels of guaranteed privacy. In this scenario, the data need to be released for graph-mining analysis, but some level of privacy must be maintained.

While privacy preservation of data mining approaches has been an important topic for a number of years (see Verykios et al. [61] for an overview), privacy of multirelational medical and social network data is a relatively new area of interest [62–65]. One reason is its complexity. Social networks, human or animal, are not random. This is one reason that anonymization alone is not sufficient for hiding identity information on certain real-world data sets [62,63]. These networks contain topological structures that are identifying marks of the network. If we analyze an unlabeled graph of the Web link structure, finding the Google homepage node may be straightforward because of its dense incoming link structure. This is an example of a unimode network. If we consider complex networks containing more unique features, identification of individuals in the network becomes easier.

How do we combat this? To what degree is network topology a factor compared to node and edge features? Are relationships between nodes more apparent when local neighborhoods have certain topological structures? How can we use the topological structure of complex networks to measure the level of anonymity in the network? To study some of the behaviors associated with social networks, how accurate do the network measures need to be for data mining applications, for example, clustering, community discovery, prominent-node identification, etc. In other words, how much error in the released data is acceptable? While we anticipate many of these topics will be explored soon for unimode networks, a far-reaching goal is to consider privacy preservation in the context of dynamic, heterogeneous networks.

In order to begin to answer these questions, we must first define "What constitutes a privacy breach?" While this may seem straightforward on the surface, many of the authors that have written on this topic have defined different types of adversaries and breaches. In the remainder of this section, we explore different types of attack models, adversaries, and privacy breaches.

Graphs have a great deal of variation. This variation is particularly important in the context of privacy. One may release a single graph or multiple graphs at different time points, depending on the analysis task. Irrespective of the number of graphs, what are the properties of the nodes and edges being released? How many object or node types are there in the network? How many relationship or edge types are in the network? Do the objects or relationships have features?

As the complexity of the data increases, the data become more unique. These unique components can be exploited by adversaries. The goal of any adversary is to determine the identity of one or more individuals or relationships in the social network. In previous work, we investigate ways to determine the level of uniqueness of nodes and edges in a unimode graph by introducing a metric called topological anonymity [66]. This measure combines different structural components, variations of degree, and clustering coefficient to measure the hideability of nodes in the network. If the hideability is low, then perturbation strategies [62,63,65] or abstractions of the original graph [64] need to be investigated.

Since adversaries have varying degrees of information about the original social network, we need to define different adversaries based on their background knowledge. Example adversary-background models include (1) a single adversary that is a member of the network and knows his own degree; (2) a single adversary that is a member of the network, knows own degree, and degree of some neighbors; (3) a single adversary that is a member of the network, knows own degree, and knows if his neighbors are connected; (4) a single adversary that is a member of the network and has insight about a community within the network; (5) a single adversary that is not part of network, but has insight about a community within the network; and (6) two or more adversaries colluding and having varying degrees of information based on Models (1)–(5).

Any of the adversaries listed above may attack the network in different ways. Two possible types of attacks are passive and active [62]. A passive attack occurs when the adversary is trying to learn the identity of nodes after the data are released. In other words, the adversary does not have access to the data before they are released. An active attack occurs when the adversary adds an arbitrary set of nodes to the original data. Edges are placed in a unique structure to targeted users that the adversary wants to identify. Once the data are released, the adversary then looks for a pattern of connections that correspond to a subgraph created by the adversary.

Finally, how can we maintain some level of privacy if a subset of nodes and edges can be labeled correctly by adversaries? As previously mentioned, if the data have breached nodes or edges, then the graph either needs to be altered through perturbation or deletion or generalized, that is, abstracted to hide the individual details of the data. The better approach depends upon the mining task that needs to be executed on the released graph. If the strategy we choose alters the graph significantly, the results of the data mining algorithm will be inaccurate and useless. Further, because real-world graphs have many nodes with more unique centrality measures, random perturbation strategies have a large effect on the accuracy of the released graph [63].

Strategies that incorporate the distribution of the graph invariants into the perturbation strategy may perform better. More research is necessary in this area.

k-Anonymity was introduced for privacy preservation of independent, unlinked data records. Using this approach, each individual should not be distinguishable from $k - 1$ other individuals [67]. However, because our nodes are not independent and are linked together, we believe k-anonymity as identified in Ref. [67] is difficult to achieve in graphs where clear semantic dependencies exist in the data. We feel that it is even more difficult for the newer metrics of l-diversity [68] and t-closeness [69]. Two nodes that are indistinguishable across some node structural metrics or whose distribution is similar across different structural attributes do not guarantee that they are across other nodes, particularly path-related measures for nodes in the network. However, if we limit anonymity to local neighborhood structure of a node, k-anonymity, l-diversity, and t-closeness can be an important metrics for improving privacy of a graph.

20.5 Final Thoughts

While we are still investigating ways to analyze simple networks with a single node type and a single edge type, the complexity of today's network data forces us to begin thinking of ways to handle and analyze more heterogeneous data. In order to mine the data, we need to develop robust models that capture the interconnected nature of the data, while allowing for the inclusion of complex features and time-varying attributes. Integrating longitudinal statistical analysis, graph-mining exploratory analysis, and visual analytic approaches to interpret complex, heterogeneous networks with incomplete, uncertain data and potentially, additional privacy constraints is an outstanding challenge. In 2006, a group of data mining researchers created a list of the top 10 data mining challenges [70]. The integration proposed here encompasses portions of six of the challenges mentioned. While researchers are working on these challenges, harnessing and integrating the advances in different areas of computer science in a meaningful, intuitive way are difficult. However, these advances in computer science are necessary to help researchers in other sciences advance their fields at a faster pace than they can today. As the last decade has shown, baby steps in computer science can translate to large strides in other disciplines.

References

[1] J. Mann, R. C. Connor, P. Tyack, and H. Whitehead (Eds.), *Cetacean Societies: Field Studies of Dolphins and Whales*. University of Chicago Press, Chicago, IL, 2000.

[2] J. A. Byers. *American Pronghorn: Social Adaptations and the Ghosts of Predators Past*. University of Chicago Press, Chicago, IL, 1997.

[3] J. Goodall. *The Chimpanzees of Gombe: Patterns of Behavior.* Harvard Univeristy Press, Cambridge, MA, 1986.

[4] D. Cook and L. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2): 32–41, 2000.

[5] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Exploration Newsletter*, 5(1): 59–68, 2003.

[6] M. C. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3): 378–394, 2003.

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

[8] L. Singh, M. Beard, L. Getoor, and M. B. Blake. Visual mining of multimodal social networks at different abstraction levels. In *Proceedings of the 11th International Conference Information Visualization*, Zurich, Switzerland, 2007. IEEE Computer Society.

[9] L. C. Freeman. Visualizing social networks. *Journal of Social Structure*, 1(1), 2000.

[10] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1): 1–8, 2002.

[11] M. Kreuseler and H. Schumann. A flexible approach for visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1): 39–51, 2002.

[12] V. Batagelj and A. Mrvar. Pajek—program for large network analysis. *Connections*, 21: 47–57, 1998.

[13] L. C. Freeman, S. P. Borgatto, and M. G. Everett. Ucinet for windows: Software for social network analysis, 2002. http://www.analytictech.com.

[14] M. Huisman and M. A. J. van Duijn. Stocnet: Software for the statistical analysis of social networks. *Connections*, 25(1): 7–26, 2003.

[15] M. Baur, M. Benkert, U. Brandes, S. Cornelsen, M. Gaertler, B. Köpf, J. Lerner, and D. Wagner. Visone software for visual social network analysis. In P. Mutzel, M. Jünger, and S. Leipert (Eds.), *Graph Drawing Software*, pp. 463–464. Springer, 2002.

[16] B. Bederson, B. Shneiderman, and M. Wattenberg. Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. *ACM Transactions on Graphics*, 21(4): 833–854, 2002.

[17] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization (INFOVIS'05)*, p. 5, Washington DC, 2005. IEEE Computer Society.

[18] G. Namata, B. Staats, L. Getoor, and B. Shneiderman. A dual-view approach to interactive network visualization. In *ACM Conference on Information and Knowledge Management*, Lisbon, Portugal, 2007. ACM Press.

[19] C. Plaisant, J. Grosjean, and B. B. Bederson. Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *INFO-VIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, pp. 57, Washington DC, 2002. IEEE Computer Society.

[20] H. Kang, L. Getoor, and L. Singh. Visual analysis of dynamic group membership in temporal social networks. *SIGKDD Explorations Newsletter*, 9(2): 13–21, 2007.

[21] Netminer ii: Social network mining software. Available at http://www.netminer.com/NetMiner/home_01.jsp.

[22] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5): 693–700, 2006.

[23] P. Smyth, S. White, J. O'Madadhain, D. Fisher, and Y. B. Boey. Analysis and visualization of network data using jung. *Journal of Statistical Software*, W(II) 2005.

[24] J. Heer, S. K. Card, and J. A. Landay. Prefuse: A toolkit for interactive information visualization. In *SIGCHI Conference on Human Factors in Computing Systems*, pp. 421–430, New York, 2005. ACM Press.

[25] B. Bederson, J. Grosjean, and J. Meyer. Toolkit design for interactive structured graphics. *IEEE Transactions on Software Engineering*, 30(8): 535–546, 2004.

[26] E. Adar. Guess: A language and interface for graph exploration. In *SIGCHI Conference on Human Factors in Computing Systems*, pp. 791–800, New York, 2006. ACM Press.

[27] L. Singh, M. Beard, B. Gopalan, and G. Nelson. Structure-based hierarchical transformations for interactive visual exploration of social networks. In *Pacific Asian Conference on Knowledge Discovery and Data Mining*, Osaka, Japan, 2008 Springer.

[28] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physics Review*, 68, May 26, 2003.

[29] M. Boguna and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review*, E 66(4), 2002.

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

[31] D. Auber, Y. Chiricota, F. Jourdan, and G. Melancon. Multiscale visualization of small world networks. *IEEE Symposium on Information Visualization*, p. 10, Los Alamitos, CA, 2003. IEEE Computer Society.

[32] L. Singh, L. Getoor, and L. Licamele. Pruning social networks using structural properties and descriptive attributes. In *Proceedings of IEEE International Conference on Data Mining*, Houston, TX, 2005. IEEE Computer Society.

[33] V. Batagelj, P. Doreian, and A. Ferligoj. Generalized blockmodeling of two-mode network data. *Social Networks*, 26(1): 29–53, 2004.

[34] T. Suel and J. Yuan. Compressing the graph structure of the web. In *Data Compression Conference*, Snowbird, Utah, 2001. IEEE Computer Society.

[35] A. Baritchi, D. Cook, and L. Holder. Discovering structural patterns in telecommunications data. In *Proceedings of the 13th International Florida Artificial Intelligence Research Society Conference*, pp. 82–85, Orlando, FL, 2000. AAAI Press.

[36] L. Getoor. Link-based classification. In S. Bandyopadhyay, U. Maulik, L. Holder, and D. Cook (Eds.), *Advanced Methods for Knowledge Discovery from Complex Data*. Springer, 2005.

[37] L. Singh and L. Getoor. Increasing the predictive power of affiliation networks. *IEEE Data Engineering Bulletin*, 30(2): 41–50, 2007.

[38] K. M. Airoldi and K. M. Carley. Sampling algorithms for pure network topologies: A study on the stability and the separability of metric embeddings. *SIGKDD Explorations Newsletter*, 7(2): 13–22, 2005.

[39] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–160, Boston, MA, 2000. ACM Press.

[40] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In *Proceedings of the National Academy of Sciences*, pp. 7821–7826, 2002. Academy of Sciences.

[41] M. B. Hastings. Community detection as an inference problem. *Physics Review E*, 74(035102), 2006.

[42] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physics Review E*, 74(036104), 2006.

[43] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435: 814-818, 2005.

[44] S. White and P. Smyth. A spectral clustering approach to finding communities in graph. In *SIAM International Conference on Data Mining*, Newport Beach, CA, 2005. Society for Industrial and Applied Mathematics.

[45] Harary, Frank, *Graph Theory* (1969), Addison-Wesley, Reading, MA.

[46] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 717–726, New York, 2007. ACM Press.

[47] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 523–528, New York, 2006. ACM Press.

[48] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 44–54, New York, 2006. ACM Press.

[49] T. Snijders. The statistical evaluation of social network dynamics. *Sociology Methodology*, 31: 361–395, 2001.

[50] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Mining hidden community in heterogeneous social networks. In *LinkKDD '05: Proceedings of the 3rd International Workshop on Link Discovery*, pp. 58–65, New York, 2005. ACM Press.

[51] P. Domingos and M. Richadson. Mining the network value of customers. In *ACM International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2001. ACM Press.

[52] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *ACM International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 2003. ACM Press.

[53] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *ACM Conference on Information and Knowledge Management*, New Orleans, LA, 2003. ACM Press.

[54] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins. The Web as a graph: Measurements, models and methods. *Lecture Notes in Computer Science*, 1627: 1–17, 1999.

[55] X. Wang and G. Chen. Synchronization in scale-free dynamical networks: Robustness and fragility. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(1): 54–62, 2002.

[56] J. R. Ullmann. An algorithm for subgraph isomorphism. *Journal of ACM*, 23(1): 31–42, 1976.

[57] L. Holder, D. Cook, and S. Djoko. Substructure discovery in the subdue system. In *International Conference on Knowledge Discovery in Databases*, New York, 1994. ACM Press.

[58] L. Dehaspe and H. Toivonen. Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery*, 3(1): 7–36, 1999.

[59] R. Albert, A. Barabsi, and H. Jeong. Scale-free characteristics of random networks: The topology of the World Wide Web. *Physica A*, 281: 69–77, 2000.

[60] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pp. 251–262, New York, 1999. ACM Press.

[61] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Record*, 33(1): 50–57, 2004.

[62] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pp. 181–190, New York, 2007. ACM Press.

[63] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks, University of Massachusetts, Amherst, MA, Technical report no. 07–19, March 2007.

[64] M. E. Nergiz, C. Clifton, and A. E. Nergiz. Multirelational k-anonymity. In *the 23rd IEEE International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, 2007. IEEE Computer Society.

[65] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In *1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07)*, San Jose, CA, 2007. ACM Press.

[66] L. Singh and J. Zhan. Measuring topological anonymity in social networks. In *Proceedings of the International Conference on Granular Computing*, San Jose, CA, 2007. IEEE Computer Society.

[67] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty*, 10(5): 557–570, 2002.

[68] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In *IEEE International Conference on Data Engineering*, Atlanta, GA, 2006. IEEE Computer Society.

[69] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *IEEE International Conference on Data Engineering*, Istanbul, Turkey, 2007. IEEE Computer Society.

[70] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4): 597–604, 2006.