Research Challenges for Knowledge Discovery

We categorize the research of knowledge discovery in information networks into three main categories: knowledge mining, integration, and understanding.

- **Mining Uncertain Heterogeneous Information Networks**
  Scalable and intelligent methods to handle the challenges of uncertainty, heterogeneity, and large volumes and/or low bandwidth, focused on three major data sources in military: network
  - spatial-temporal data + streams
- **Domain Integration for Decision Support**
  - Integrate knowledge learned from networks, spatial-temporal data and streams all together for better decision support.
- **Human-centric Network Analysis**
  - Consider human factors for better understanding and applying knowledge. Interface to Human Network Proposal

Task 1. Mining Uncertain Heterogenous Information Networks

Scalable methods for situation awareness and network mining

Methods for mining network data lie at the core of network analysis. Mining not only enables direct discovery of new interesting patterns and knowledge buried in the multidimensional data of heterogeneous networks, but also provides necessary support for many other tasks of network analysis (e.g., supporting network search through mining frequent substructures for indexing). We will develop scalable methods for several key network mining tasks with a focus on noisy network information and missing data: (1) substructure correlation mining, (2) classification across multiple heterogeneous networks, (3) clustering subnetworks, and (4) outlier analysis.

Much of Han's previous work can be leveraged and extended in this direction. For example, gSpan and CloseGraph 5 can be extended to integrate constraint-based graph mining, "virtual" graph cutting, and graph pattern growth for substructure correlation mining. His recently developed CrossMine 6 and DDPMine 2 can be integrated for cross-network classification. Clustering work (LinkClus 7 and CrossClus) 8 can be extended to exploit the power-law distribution of many networks and perform user-guided clustering. Frequent patterns and rare-event classification techniques can be exploited for outlier analysis. Furthermore, all the proposed methods will be explored in the context of a variety of different networks types, as well as their heterogeneous superpositions. Examples include Web-based information networks (about people and their related public content) and financial networks (about organizations, prices, and monetary dependencies). This will break the limitation of current graph mining algorithms that the nodes and edges are all assumed to be of the same type (e.g., humans for nodes and phone calls for edges) and develop new mining methods that can consider attributes of nodes (e.g., job-code for humans), and attributes of edges (e.g., phone-call, versus e-mail, versus text message, etc.).

RankClus: Ranking and Clustering Information Networks

For network mining, an essential task is to cluster and rank nodes in networks based on user- or application-oriented similarity or ranking criteria. Although ranking and clustering each can provide overall views on information network data, ranking objects globally without considering which clusters they belong to often leads to dumb results, e.g., ranking database and computer architecture conferences together may not make much sense. Similarly, clustering a huge number of objects (e.g., thousands of authors) in one huge cluster without distinction is very dull as well.

We propose to develop an effective RankClus method by integration of ranking and clustering, which mutually enhances the quality of clustering and ranking results. The general idea is to iteratively refine the clusters and ranks generated based on the tightness of the links as well as the authority-criteria proposed by experts. In particular, we use a mixture model to decompose each object into a K-dimensional vector, where each dimension is a component coefficient with respect to a cluster, which is measured by rank distribution. Objects then are reassigned to the nearest cluster under the new measure space. As a result, quality of clustering and ranking are mutually enhanced, which means that the clusters are getting more accurate and the ranking is getting more meaningful. Our initial experiments show that RankClus can generate more accurate clusters and in a more efficient way than the state-of-the-art link-based clustering methods. Moreover, the clustering results with ranks can provide more informative views of data compared with traditional clustering. This method will be further developed to find multilevel clusters and the object ranking in the corresponding clusters.

2.2 Real-time methods for spatiotemporal context analysis

Information networks operate in physical (e.g., spatio-temporal) environmental context that has impact on network analysis techniques. An example is a network that models the evolving state of a battlefield. Observations made in such a network are subject to resource constraints, noise, uncertainty, and other limitations of the physical environment. Hence, understanding, representing, and accounting for the spatiotemporal context of a network (e.g., its physical constraints and uncertainty) becomes key. To optimize monitoring performance, one may need to actively control physical resource allocation in the environment, which itself may be modeled by a network of assets that acquire and filter data streams, or a network that models resource allocation of ground, air and satellite communication resources. We make a distinction at this point between the evolving information network itself and its physical monitoring networks.

Physical (e.g., monitoring) networks and information networks have traditionally been studied in their respective communities separately, focusing on performance analysis methods for the former and search and mining methods for the latter. A significant innovation of the proposed work lies in a multidisciplinary approach that puts them under a common analytic foundation and accounts for their interdependencies, when present. For
example, in a dynamic scenario, a resource-constrained or partially damaged monitoring network might throttle the information needed for maintaining and updating parameters of an evolving battlefield model. In turn, real-time analysis of battlefield information streams may require additional information from the monitoring network to reduce uncertainty (in much the same way as initial medical diagnosis may motivate further tests). The efficacy of battlefield analysis in detecting distributed anomalies or predicting an opponent’s move may thus depend on proper understanding and exploitation of the resulting feedback loop and the imposed environmental constraints. Of particular interest is developing an understanding of the sensitivity of such loops to both damage (e.g., to the physical network) and uncertainty (e.g., in the state of the information network).

To facilitate analysis of evolving networks, we propose to extend our multidimensional mining to stream data. We develop real-time methods for discovering and evolving stream patterns. Moreover, we develop methods to detect anomalies in data streams by statistical analysis, clustering-based outlier detection, and rare-event classification. We will further develop our recently developed methods on stream data mining, including stream data cube 4, clustering evolving data streams 1, stream classification under rare events 3 and extend them to handle mining data streams in heterogeneous physical and information networks.

Another important theme for handling temporal data in information networks is to discover evolution regularities and anomalies in dynamic information networks. We propose a density-based subnetwork clustering method for evolutionary analysis of dynamic networks. The general philosophy is an information network consists of many interconnected subnetworks or subgraphs. Such subgraphs can be clustered using a density-based approach. Evolution regularity can be viewed as such connected graphs will maintain certain stability along with time and such a transition from one state to another should be somewhat incremental and density-based subgraph clustering along with time could be an interesting model for such analysis. This will lead to analysis techniques that, for example, can uncover dominant structures and dynamics in complex networks to infer information such as an adversaries command structure, the sequence of clearances they need to engage specific resources, or vulnerabilities involving interactions between their organizational and physical assets.

**Actionable Information and Decision Support from Data Streams**

A key aspect of information processing is to convert the large volume of the incoming data stream into actionable information. The definition of actionable information may vary considerably with the scenario at hand. Therefore, it is important to focus our research on the design of tools which are reusable across different situations. In order to achieve this goal, we will design a number of “data mining primitives” for probabilistic and deterministic data streams, which are reusable across many scenarios. Many standard problems in data mining such as event detection, abnormality detection, clustering and classification have very wide applicability and can be used as components to a larger analytical scenario. While a number of mining and management techniques have been developed from data streams in the literature, it is still an emerging area of research on how the unique challenges of network data can be handled. In particular, the heterogenous and uncertain nature of the underlying data pose a serious challenge to a variety of data mining and management applications.

**Task 2. Domain Integration for Decision Support**

Currently, many of the decision support and data mining techniques in the literature assume "pure data availability", in the sense that the underlying data has a specific structure (e.g. all fields are quantitative). This is typically not the case for network data, since the data arrives from multiple data sources such as text, video, images, or other conventional data such as multi-dimensional records. Therefore a critical focus will be on the development of tools which can convert the information from multiple data sources into generic summaries which can be used in conjunction with a variety of tools. This requires the development of data mining and management tools which assume heterogenous input in order to determine useful and actionable information.

**Task 3. Human-centric Analysis of Information Networks**

**User-Friendly Methods for Network Analytics**

Graphs and networks are observed universally in military, physical, and social worlds. For instance, a large number of information networks can be learned from the full range of structured and unstructured data sources; yet it is unknown what is hidden inside and how to find them. It is critical to (1) invent novel concepts and methodologies in network analytics that could facilitate the understanding of networks for better decision making, and (2) provide efficient data-parallel and approximation algorithms so that real-time online knowledge discovery can be performed quickly in information networks.

1. Improve interpretability of discovered knowledge and patterns.

The exponential pattern sets or rules generated by many mining processes have undermined the potential of knowledge discovery in complex networks. We are drowning in interwound networks; ironically, we are also drowning in structured and unstructured knowledge discovered by machines. To overcome this issue, we have successfully developed statistical and combinatorial models to summarize itemset patterns. In this project, we are going to explore the same issue in the context of network domain.

2. Data-parallel algorithms for network analytics

Network models adopted by existing network algorithms are often limited in scale, and assume that all nodes and edges can be accommodated into main memory. This assumption does not hold in large-scale information networks. We must develop divide-and-conquer style mining
algorithms that can run on network partitions in parallel, and output local results as soon as possible, which could be aggregated later into a global solution, if needed. In this project, we propose a series of data-parallel algorithms for the most important network mining tasks such as network centrality, association patterns, structural patterns, clustering and grouping, classification and link analysis.

3. Approximation algorithms for knowledge discovery

When the computation cost of mining full set of knowledge and patterns becomes prohibitive, a user-friendly method should instead turn to mine only (or at least generate first) the most significant knowledge and distinctive patterns from the underlying information networks. Dr. Yan's recent research has developed a set of principles for such analysis, including (1) pattern summarization, (2) colossal pattern mining, and (3) mining significant patterns directly by leap search.

### 4.2 Human-centric Awareness Analysis

In this task, we propose human-centric analysis of information networks, which will systematically evaluate agents' role and interactions from a information network point of view. This task might also provide detailed input to human network analysis. Given an information network recording information exchange between agents or information access by agents, it will allow a high-quality assessment of agents' expertise and the awareness of expertise between agents. Such knowledge will help coordinators/commanders successfully allocate the right human resources and react to emergent events appropriately in the first place.

1. Search experts in information networks

This project studies an approach to search experts through the analysis of organizational information from the full range of data sources. Existing expert search systems are often designed to explicitly integrate heterogeneous information sources with identical data types. Instead, our search framework will focus on an autonomous approach to integrate text, sequential, and network information all together for better expertise analysis. This project will also explore the potential of implicit models to recommend experts without revealing their expertise. At IBM, we have successfully applied sequential patterns discovered from multiple decision chains to quickly locating right experts for the right problems.

2. Awareness analysis for accurate decision making

Existing expert recommendation or more general, problem solving systems often ignore the potential of collective intelligence for collaborative work. In order to improve decision making process, it is critical to assess the awareness of expertise between agents. Upon receipt of a problem, an agent tries to resolve the problem using his/her own expertise and if fails, will forward the problem to a successor agent that is considered to be of the right expertise to resolve the problem. Obviously, the accurate understanding of successors' expertise becomes the key to the fast and accurate decision making. In this project, we propose developing autonomous data mining techniques to evaluate expertise awareness so that the weak components in a decision making process could be identified.

### References


