The objective of my research is to design computational systems that are capable of analyzing and processing big data. The term “big data” usually refers to 3V’s, i.e., Volume (the amount of data), Velocity (the speed of data generated), and Variety (the kinds of data available). My thesis research mainly focuses on addressing the variety issues of big data. Data variety is about the increasing number of data types that need to be handled differently from simple logs and conventional data records, and also many different data sources that need to be fused together. These include data collected from scientific studies, health care records, social networks and social media: user activities, events, locations, etc. Although technology advances (e.g., Apache Hadoop) have helped us enormously in dealing with the first two V’s (volume and velocity), data variety remains a challenging problem to solve programmatically, and once it succeeds, it will represent a fundamental advance in big data research with great benefits to diverse fields, such as biomedical research, social computing, business, etc.

The data variety issues are difficult to solve because the data usually have complex structures and involve many different types of information interrelated in a complex way. For example, in computational system biology, if we want to predict the efficacy of a molecular drug for a certain disease, there are many challenging issues we need to address. Chemical compounds have complex molecular structures that standard computational methods cannot handle directly. The biology system is also extremely complex, involving multiple types of components, such as genes, pathways, tissues, that are interrelated through multiple types of relationships like PPI (Protein-Protein-Interaction) and drug binding. One challenge when it comes to big data applications like this is that it is extremely hard to join all of these different data types together in one unified analytic. Furthermore, these data are quite often separated in different sources, such as DrugBank and PubChem. Unifying fractured data can compound big data complexity problems immensely.

A successful computational system for data variety must be able to process data with complex structures, of heterogeneous types, and from multiple sources in a holistic perspective. This challenge raises three fundamental questions: 1) How do we discover and extract important patterns from data with complex structures? 2) How do we join different data types together in one unified analytic? 3) How do we merge data from one source with other sources? To answer these questions, I performed research at three levels: 1) pattern/feature extraction for structural data (e.g., identifying subgraph patterns for molecular drugs and brain functional networks); 2) mining heterogeneous information networks (e.g., mining the complex biological system for drug discovery, analyzing social networks and social media); 3) aligning multiple data sources (e.g., aligning the user accounts across different social networks). In the following three sections I will outline my work in these three areas.
**Feature extraction for structural data in drug discovery and neuroimaging**

In many big data applications, data with complex structures can usually be modeled as graph objects. Examples include chemical compounds, brain functional networks and social networks. These data are quite different from traditional data objects, which have flat features and thus can be processed easily using conventional methods. A graph object (e.g., a chemical compound) usually has a complex structure, represented as a set of nodes (e.g., atoms) interconnected through a set of edges (e.g., chemical bonds). There is no feature readily available. I have done significant work on modeling various structural data as graph objects, and extracting important subgraph patterns in many different scenarios: drug discovery tasks where the labeling costs are very high or when we want to explore multiple gene targets; neuroimaging tasks where the structures of brain networks are highly noisy and uncertain. In such scenarios, significant, discriminative and reliable patterns for structural data would be of great utility and the prerequisite for any true analytics.

I have developed a new semi-supervised method for graph pattern mining to identify important subgraph features[4], that are suitable for small molecular drug discovery. Graphs are becoming increasingly important in modeling real-world data with complex structures. Different from widely used supervised machine learning settings, in graph data, the labeling process is very expensive and time consuming for automatic graph classification. I studied the problem of how to minimize the labeling efforts on graphs by using both labeled and unlabeled graphs to improve the classification performances. I then developed the first method for active graph pattern mining to identify promising drug candidates and important drug features simultaneously [2]. In this work, we studied the active learning setting for graph data. Based on the importance estimation of unlabeled graph objects, we developed an active sampler on graph data that can help reduce the labeling cost in graph classification tasks by only requesting labels of most important graphs in the dataset. I developed a novel multi-label method for extracting subgraph features targeting multiple genes in drug discovery [7][5]. One major problem with analyzing structural data lies in the lack of useful features, especially when the graphs are associated with multiple concepts simultaneously. I studied the problem of how to exploit the multiple concepts in graph data to extract useful features for multi-label classification. My approach can effectively boost multi-label graph classification performances and is more efficient by pruning the subgraph search space using multiple labels.

I further modeled brain functional connections as uncertain graphs and developed a method to extract important subgraph patterns for neuroimaging tasks [8], such as extracting structural patterns for brain diseases, such as Alzheimer’s disease, ADHD (Attention Deficit Hyperactivity Disorder) and brain injury. A key issue in analyzing the network of brain functional connectivities is modeling the structural uncertainty within the network structure. The functional connections among different brain regions are often uncertain, where we are not sure about the exact structure of the network. In this work, I modeled each brain as an uncertain graph to capture a variety of different possible outcomes of the network structures. I then developed a method that can take such uncertainties into consideration and extract important and reliable patterns from the brain network data.

**Mining heterogeneous information networks in bioinformatics and social computing**

A key issue in data variety is handling heterogeneous information, i.e., joining many different data types and extracting their complex interrelations in one unified analytic. Social, biomedical information systems usually consist of a large number of interacting components. Examples of such systems include online social networks, biological networks, transportation systems, epidemic networks. All the above systems share one common feature: they are network systems, consisting of multiple types of individual components (i.e., nodes) interrelated with each other through various kinds of relationships (i.e., edges), forming large heterogeneous networks. We call such systems as heterogeneous information networks. Heterogeneous information networks are ubiquitous and of great utility in many big data applications.

My earlier work focused on classification problems in social networks, where the networks have complex interactions among the individual nodes and the classification tasks involve multiple overlapping categories
Conventional predictive models have been quite successful in modeling flat data, such as document classification. However, in social network data, users are closely connected with complex structures and each user’s interests may involve multiple different categories simultaneously, such as movies, sports and cars. I studied the problem of how to extract the complex dependencies among different users on multiple labels to improve the classification performances on network data. To take advantage of structured data of social networks, I designed a new relational learning method for collective classification of multiple label concepts. The model is able to capture the dependencies among the preferences for a group of related individuals and the dependencies among the multiple concepts simultaneously.

Based upon the relational model, I developed advanced methods to analyze heterogeneous information in social networks [6] and biomedical systems for drug discovery [1]. Heterogeneous information networks involve many different types of nodes (e.g., genes, pathways, chemical compounds) interacting with each other through multiple types of relationships (e.g., PPI, gene binding). As shown in Figure 2, I researched on a heterogeneous information network in drug discovery that involves 10 types of nodes interconnected through 11 types of relationships. The key issues in heterogeneous information networks are extracting the complex and heterogeneous relationships between various types of nodes and exploiting such relationships in data analysis, such as classification tasks. I developed a novel classification model on heterogeneous information networks which can capture the semantic differences among different types of paths in the networks, and use such paths to collectively predict the labels of a set of nodes based upon their interactions with other types of nodes. This work may open new avenues for many tasks in analyzing the biological systems in drug discovery, such as drug-target binding prediction, gene-disease association prediction.

Aligning multiple sources of heterogeneous information networks

Another step towards data variety is handling multiple data sources. In many big data applications, the data we are often separated in different sources. Leveraging data from multiple sources allows a computer system to accomplish tasks that cannot be accomplished by processing each data source separately.

I focus on leveraging information from multiple social networks to improve the social computing analysis, such as node classification, link prediction and content/location recommendation. To do so, one needs an alignment (mapping) between the data nodes (e.g., user accounts) in multiple sources. For example, in online social networks, people can get involved in many different kinds of social networks simultaneously, such as Facebook, Twitter and Foursquare, where each of the network contains abundant information about who, where, when and what. Thus, each user usually has multiple separate accounts in different social networks. I developed a novel approach to map the user accounts in one social network with those in another social network [9]. This approach can extract important structural features from multiple social networks and find a mapping using constrained stable matching.

Future Agenda

My future research interests are directed towards a long-term goal of making sense of big data in a holistic perspective. In the future, I will actively apply for research grants from the NSF and other funding agencies. I will also seek collaborations with big data industries (e.g., Yahoo Labs) and researchers from
other fields (e.g., biomedical research and business). This section outlines some future opportunities that I am excited to pursue.

**Mining medical and neuroimaging data using heterogeneous models.** Medical research and neuroscience have been facing big data challenges for a long time, and the data are becoming larger and more complex by each year. A typical dataset for a functional MRI scan can take up to Gigabytes per person, providing the measurements on subregions of the brain at a spatial resolution of 1 to 5 mm per voxel, and a temporal resolution of one scan per seconds or so. In addition to fMRI, there are also other types of medical images (e.g., EEG, PET, MRI, DTI, CT, CAT, MEG) providing multiple views of the patient. Moreover, doctors may also have measures on thousands of other bio-markers of the patient, such as blood markers, anti-bodies, virus-levels, RNAs etc. Clinicians are interested in using all these measures (imaging and bio-markers) to map the brain as well as the blood system to detect the effects of stroke, brain injury, or diseases such as Alzheimer’s and ADHD. I am interested in investigating the heterogeneous sources of medical data, figuring out the relationships among different components and automatically extract most relevant features from the data.

**Mining heterogeneous social media.** With the recent advances of smart phones and mobile apps, social media (such as Twitter and Facebook) are providing more and more information and various types of data in real time. For example, Twitter nowadays is playing a role as the world’s largest free “sensor network”, where the users are the “sensors”, providing abundant information about who, where, when and what. These data could help us detect events, monitor natural disasters, disease outbreaks, etc. A large part of the challenge in mining heterogeneous social media lies in putting the different types of data into the right context. Nowadays most of the social media data available for analysis is linked to outside entities and organizations. I am interested in investigating the problem of data variety in social media by making sense of and adding context to the diverse data types and sources.

**Mining large-scale, dynamic information networks.** In the future, I will try to investigate the data variety issues in large-scale and dynamic systems. The first two V’s (Volume and Velocity) in big data are usually mixed with data variety issues. For example, how can we discover the structural relationships among a set of entities in heterogeneous information networks when the linkage structures are highly dynamic. These studies can help improve the predictive analysis on heterogeneous networks by utilizing the temporal dynamics of the network.

**References**


