

Research Summary

Sitaram Asur (asur@cse.ohio-state.edu)

In the recent years, rapid advances in technology have led to an exponential growth in data, with billions and trillions of observations being generated constantly in numerous domains such as astronomy, sociology, computer science, biology, chemistry, metabolism and nutrition. It has been observed that real world data from these diverse domains can be modeled as complex interaction networks where nodes represent entities of interest and edges mimic the interactions or relationships among them. Specific examples range from Protein-Protein Interaction (PPI) networks to relationship (social) networks, from co-authorship networks to the World Wide Web and from metabolic networks to peer-to-peer networks. These networks have been recognized to be not only outcomes of complex interactions, but also key determinants of structure, function, and dynamics in systems that span the biological, physical, and social sciences. Hence, there is a need to design suitable algorithms to extract or infer meaningful information from these networks. However, the challenges involved are daunting.

First, most of these real-world networks have specific topological constraints (e.g. scale-free) that make the task of extracting useful patterns using traditional data mining techniques difficult. Additionally, these networks can be noisy (containing unreliable interactions), which makes the process of knowledge discovery difficult. Second, these networks are usually dynamic in nature. Earlier research has concentrated mostly on the static features of these networks. It is crucial to consider the evolutionary aspect of these networks to identify and model key structural and behavioral changes occurring in these networks over time. To address these challenges, as part of my dissertation research, I have developed a framework of algorithms designed to detect, analyze and reason about the structure, behavior and evolution of real-world interaction networks.

1 Dissertation Research

My dissertation explores the development of noise-resistant and topology-aware algorithms for the joint static and dynamic analysis of interaction graphs in order to capture the interplay among structure, behavior and evolution of these graphs. I will describe the key contributions in turn below.

Static Analysis for Motif Extraction

The objective here is to develop algorithms and techniques for mining interaction graphs to extract useful clusters, modules and motifs. In Protein-Protein interaction (PPI) networks, the discovery of key functional modules can help understand the functions of proteins and also aid in predicting the function of unknown (un-annotated) proteins. Traditional clustering/graph partitioning algorithms have not performed well in this task due to the presence of a) noisy false positive interactions b) scale-free topology, and c) multi-faceted hub nodes. To overcome these challenges, I have proposed an *ensemble clustering* strategy which enables one to combine multiple topological features and obtain meaningful cluster partitions. The ensemble strategy makes use of extrinsic and intrinsic topology-based similarity measures, with different base clustering algorithms to form informative base partitions. I have designed and evaluated a consensus method that relies on *Principal Component Analysis (PCA)* to reduce the dimensionality of the consensus determination problem. The ensemble solution on the reduced dimensional space can then be efficiently computed using traditional consensus methods. To handle the issue that most proteins are multi-faceted, I have designed an adaptation to the above approach that allows for *soft ensemble clustering of proteins* in interaction networks. This enables our method to model and account for

the different functions that proteins possess. I have performed a detailed experimental evaluation of the above methodology using topological, information theoretic and domain-specific cluster validation metrics to evaluate and modulate the improvements gained from each component of the proposed ensemble clustering methodology. Parts of this work have been published in BIBE 2005, PKDD 2006, Link-KDD 2006, ISMB 2007 and the Journal of Bioinformatics, July 2007. Although in these papers, I have focused on PPI networks, these challenges and techniques are applicable to several other real-world interaction networks.

Event Detection for Dynamic Analysis

The task of dynamic analysis involves characterizing the changes in structure and behavior that occur in interaction graphs over time. Previous research has almost entirely concentrated on the static analysis of interaction graphs, ignoring the fact that most of these real-world interaction graphs are constantly evolving over time, with structure and behavior changing. Examples of evolving networks include dynamic social networks, WWW network, epidemiology studies, examining changes in gene expression or protein interactions over time, and longitudinal clinical trials studies. Identifying the portions of the network that are changing, characterizing the type of change, predicting future events (link prediction), and developing generic models for evolving networks are critical challenges that I have looked to address.

Cluster-based Evolutionary Analysis: I have developed a general event-based framework studying the evolution of clusters of these networks, in particular their formation, transitions and dissolution, for effectively characterizing the corresponding changes to the network, quantifying behavior and performing temporal reasoning on these dynamic networks. The advantage of the general event-detection framework is that it can be used to derive custom behavioral measures as well, which is extremely useful in the context of social information management. I have shown how semantic content and category hierarchy information can be incorporated for temporal reasoning. This work has been published in the proceedings of SIGKDD 2007 and also secured the Best Paper Award in the Applications category. An extended version is to be published in the TKDD journal.

Viewpoint Neighborhoods: Clusters provide global structural information, but do not retain local relationships. One important problem in keyword search is to extract subgraphs that correspond to a given search query. Here, one is interested in, not only nodes that satisfy the query but also their relationships. In advertising, one may be interested in identifying influential nodes as well as their sphere of influence i.e the nodes affected by this node and the degree of the effect. From a dynamic perspective, one can be interested in how changes that occur in the graph over time affect different nodes and their neighbors. Crucial to addressing these problems, is the notion of a local neighborhood of interest for a node. Such a neighborhood needs to reflect not only the local topology but also other relationships such as semantic similarity or global betweenness. I have developed algorithms using an activation function for identifying local neighborhoods of interest for a node and a group of nodes in interaction networks, while also quantifying local relationships within these neighborhoods. Different activation functions can be employed capturing different intrinsic and extrinsic properties of nodes in the graph. A crucial problem in this context is to characterize the changes occurring in these neighborhoods over time and show how they can be used to build models for dynamic behavior. For this purpose, I have defined certain temporal events that can characterize changes and measure important structural and behavioral patterns such as stability and popularity over time. This work is currently under review.

Visual Analysis of Dynamic Interaction Networks

One of the challenges in visualizing dynamic interaction networks is to identify and localize the portions of the

network that are changing to help characterize the type of change and its potential causes, *visually*. A related challenge is to facilitate interactive interrogation, i.e., the user needs to be able to interactively select and zoom down to clusters, entities of interest, as well as specific dynamic interactions and events that govern the evolution of interaction networks over time. To address these challenges, I have developed a visual toolkit specifically designed to analyze dynamic graphs. To facilitate visual analysis, the front-end of the toolkit presents the user with the option of multiple views - a *graph view* which is a cumulative snapshot representation of the graph at different points in time, a *community view* which represents the cluster arrangements of the snapshot graphs, an *event view* which demonstrates the transformations that have occurred over time, and a *node view* which details the evolutionary behavior of individual entities. The user can pick intervals of interest and drill down onto the corresponding events and behavioral measures within that time-frame. A weighting function is used to associate different behavioral characteristics such as influence and sociability with nodes and importance and recency (temporal stability) with edges. These weights are then mapped onto effective visual cues to localize features of interest. This work was published at SIGKDD 2008.

2 Other Research

Apart from my dissertation work, I have been involved in research projects on membrane protein crystallization and wireless sensor networks.

Mining Membrane Protein Crystallization Trials Data: Membrane proteins are integral to all cellular functions acting as mediators between the cell and its environment. However, there is still very little known about their function since many of their structures remain unknown. The science of crystallization is still quite preliminary and there is very limited knowledge on what actually causes crystallization to occur. The key challenges are a) The training data consists of mostly negative samples (corresponding to unsuccessful crystallization trials) b) The conditions have been sampled from only a few regions in the crystallization space. To overcome these challenges, I have developed a model-based approach using an ensemble of suitable supervised learning algorithms to examine relationships or correlations between the input parameters (protein properties, crystallization conditions) and model the response output (crystals, precipitates or no crystals) for existing trials and finally identify interesting ‘hot spots’ (areas with high potential for yielding good quality crystals) in the space for future trials. This work was published at ISMB 2006 and in the Journal of Bioinformatics, July 2006.

Intrusion Detection on Wireless Sensor Networks: Wireless sensor networks are becoming ubiquitous in their use in security, defense, monitoring and tracking applications. Intrusion detection for sensor networks involves: 1) continuous monitoring for threats and intrusions, 2) rapid detection, and possibly even classification and tracking, of intrusions, and 3) rapid decision making. Sensor networks are burdened by limited battery power, which dictates the need for energy-efficient classification models to address this issue. I have proposed a technique to build local classification models in clustered sensor networks to perform efficient detection of rare events, while also improving the lifetime of the network by reducing energy losses. I have detailed a correlation-based scheme to partition the features observed by the sensor nodes into disjoint mutually uncorrelated feature subsets. An ensemble of local classifiers are then trained on these subsets. An energy efficient routing scheme designed for the above model helps reduce energy losses in the system. This work was published in the proceedings of the Sensor-KDD Workshop at SIGKDD 2007.