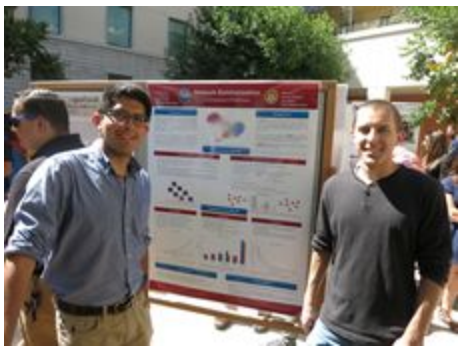
2014 Summer Undergraduate Research Experience Project Showcase
University of California, Santa Barbara

Funded by two grants from the National Science Foundation: an IGERT in Network Science (DGE-1258507), and a project on the querying and mining of dynamic graphs (IIS-1219254).
PI: Prof. Ambuj Singh, Computer Science

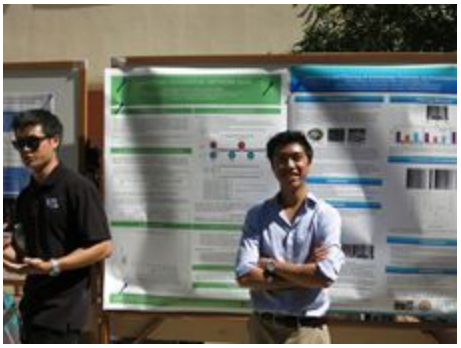
A Comparison of Methods for Network Summarization



Ali Hajimirza, Computer Science, University of Oklahoma, Norman
Jason White, Computer Science, CSU San Bernardino
Arlei Silva, Sourav Medya, Prof. Ambuj Singh, Department of Computer Science

Networks are ubiquitous – they model numerous complex structures and processes. A network is a set of nodes associated with dynamic attributes and a set of edges that represent relationships between nodes. Networks generated by real-world complex systems are extremely large. For example, Twitter consists of over 40 million users (or nodes) and over 1.4 billion connections (or edges) between users. In order to analyze such networks, the size of the network data must be reduced and, therefore, summarization becomes important. We experimented with three methods of compression: Slice Tree, Spectral Graph Fourier, and Spectral Graph Wavelets. Slice Tree partitions a network into smooth regions such that each region can be compactly represented by a single value. This single value summarizes the values of the nodes inside the region. Spectral Graph algorithms, in general, summarize a graph by performing a transformation from the graph spectral domain into the signal domain and filtering for important signal values. We examine each method for its scalability, accuracy, and sensitivity to input parameters on real and synthetic datasets. We discover that the Slice Tree algorithm is scalable and outperforms the Spectral Graph methods when the network values change smoothly across the topology.

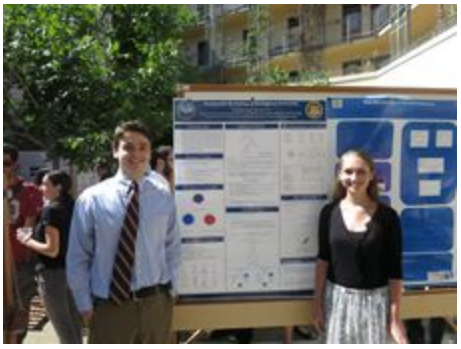
Interpreting Ecological Network Data



Sean Nguyen, Biology, UC Santa Barbara
Kyoungmin Roh, Sourav Medya, Prof. Ambuj Singh, Department of Computer Science, Prof. Hillary Young, Department of Ecology, Evolution, and Marine Biology

In recent years, there has been an increasing number of studies on complex networks within ecological systems. Those focusing on food web networks have produced simple models, allowed assessments of network robustness, and other properties. However, it is difficult to analyze a food web without a complete set of necessary data. To accurately describe such networks, either additional observational data or an effective and efficient way to infer missing network data is needed. Our current network includes three species of parasites and six host species; however, field data on one parasite species is missing. To find the missing data, we first validated the Expectation Maximization (EM) algorithm across sample data. We then applied the EM algorithm to our current network in order to find missing parasite parameters. Our findings provide a simple and fast approach that has time, labor, and monetary saving implications when describing incomplete ecological food web networks.

Multiscale Modeling of Biological Networks

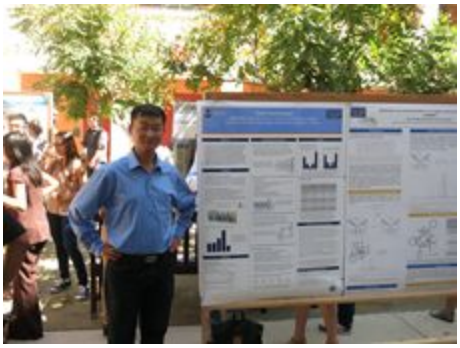


Kara Goodman, Computer Science, CSU San Bernardino
Austen Piers, Computer Science, UC Santa Barbara
Xuan Hong Dang, Sourav Medya, Hongyuan You, Kyoungmin Roh, Prof. Ambuj Singh, Department of Computer Science

A genetic network consists of gene expression levels and the genes' underlying PPI (protein-protein interaction) network. The project's goal is to identify a small number of sub-network biomarkers within three genetic networks that predict a phenotype. Our data consists of microarray data from breast and liver cancer patients, as well as cell proliferation in *Caenorhabditis elegans*. The collected microarray data has features in the low thousands; allowing for a large number of possible sub-networks, which, in turn, makes the search for discriminative sub-networks NP-hard. Our lab's machine learning algorithms MINDS (MINing Discriminative Subgraphs) and SNL (Sub-Network spectral Learning) are two methods that overcome this intractability. MINDS performs MH (Metropolis Hastings) sampling to discover discriminative sub-networks that are used to create NCDT (Network Constrained Decision Trees), which classify network snapshots. SNL uses regularized subspace learning under network topology constraints to discover discriminative sub-networks. Both SNL and MINDS reveal influential genetic biomarkers of the underlying phenotype with accuracies above 70 percent, respectively.

Tweet Classification

Yuanlin Xu, Computer Science, CSU Los Angeles
Michael Busch, Prof. Ambuj Singh, Department of Computer Science



An automatic and robust tweet topic classification technique plays an important role on many challenging research problems such as information retrieval, document categorization and social network analysis. The purpose of our project was to build an automatic tweet classification system with a domain-specific Wikipedia Knowledge Base. In this project, we used a 48.7GB Wikipedia Latest Articles data set, which was last modified on June 2014, to train our tweet classification system. We tested the system performance on 6000 tweets that were labeled from five topic labels in a Mechanical Turk survey. We constructed a Bag of Words model, a topic-specific vocabulary model, based on the Wikipedia Latest Articles data set and implemented a Naive Bayes' classifier to classify arbitrary text in a given tweet. We also provided a least-squares regression approach for mapping the scores of Wikipedia categories to a topic-specific tweet as well.