Statement of Research Accomplishments

Purnamrita Sarkar

June 1, 2009

1 Overview

A broad range of graph-based real world applications, collaborative filtering in recommender networks, link prediction in social networks, fraud detection, personalized graph search techniques and graph visualization over time require a deep understanding of the underlying graph structure. These networks can consist of millions of entities, and hence scalability is a critical concern for designing algorithms in these settings. The focus of my thesis is design and analysis of scalable algorithms for proximity search in graphs, which is a core primitive at the heart of many graph-based learning problems.

2 Proximity search in graphs

Random walks on graphs is an extremely well studied domain of mathematics. However, most random walk based proximity measures are computationally impractical for large graphs. I have devised local algorithms which quickly compute approximate nearest neighbors of a node in these proximity measures. The main idea is to adaptively expand a neighborhood around the node in order to prune away nodes which are *provably* not potential nearest neighbors. We also proved some surprising compactness properties of these measures, which justifies our neighbor-expansion algorithm. This algorithmic framework can be generalized to other related settings, as we will show in the following sections.

GRANCH: All pairs of nearest neighbors: We have developed an algorithm (Sarkar,Moore,UAI'07) to compute all *interesting* pairs of ϵ-approximate nearest neighbors in truncated hitting and commute time in large graphs. On link prediction tasks on $100,000$ node real graphs and a single CPU machine, our algorithm consistently outperformed existing algorithms in the social networks/link prediction domain.

Fast Incremental Proximity Search: In this project (Sarkar et al, ICML'08) we combined the deterministic neighborhood-expansion algorithm GRANCH with sampling techniques to compute approximately correct rankings with high probability *at query time*. The proposed algorithm can answer queries *on the fly without caching any information about the entire graph*. On a $600,000$ node Citeseer network this algorithm can process queries in around 4 seconds on a single CPU machine.

Fast Dynamic Reranking in Large Graphs: So far we have considered ranking problems. In (Sarkar,Moore,WWW'09)) we design algorithms for quickly *reranking* search results based on a small set of results labeled as relevant or irrelevant by the user. Since the labeled set of nodes vary from user to user we want an algorithm which will be extremely fast in order to be able to incorporate user-feedback on the fly. We successfully extended our algorithmic framework for computing top ranking nodes under *harmonic measures* in graphs. On the entire DBLP citation corpus (1.4 M nodes and 2.2 M edges) our branch and bound algorithm takes about 1.5 seconds to retrieve the top 10 nodes w.r.t. this measure.

3 DSNL: Probabilistic Modeling of Dynamic Social Networks

I have also worked on tractable generative models for large dynamic networks (Sarkar,Moore,NIPS'05). In this paper we generalized a successful static model of relationships into a dynamic model that accounts for friendships drifting over time. We showed used an interesting array of statistical and computational tools to tractably learn such models from data. Our algorithm scaled easily to a real world graph of $11,000$ nodes, which is two orders of magnitude larger than the graphs that could be analyzed by prior statistical network modeling algorithms. We also applied DSNL on the co-occurrence data between USDA-controlled food processing establishments and various strains of Salmonella (serotypes) as a network which evolves over time (Sarkar et al, BioSecure'08). Experimental results indicate predictive utility of analyzing establishments as a network of interconnected entities as opposed to modeling their risk independently of each other.

DSNL efficiently obtains point estimates for latent coordinates. In (Sarkar et al,AISTATS'07) we developed the first algorithm for dynamic embedding of co-occurrence data which provides *full distributional information* for its coordinate estimates.