Community Detection in Uncertain Networks Using an Ensemble Approach

Johan Dahlin and Pontus Svenson

Swedish Defence Research Agency

Introduction

Social Network Analysis (SNA) is an umbrella term for a large number of methods and algorithms for both visualization and study of relational data sets. These methods are used in investigating networks of e.g. criminals, terrorists, electrical networks, as well as epidemiology. A large effort has been devoted to develop efficient algorithms for estimating centrality measures and community structures during the last decade. This is essential as more and more network information is collected and the need for analysis in e.g. business and criminal intelligence increases.

Community detection is a collection of methods to find groupings (or clusters) in network data, in which member entities are more densely connected between themselves than to other entities outside the group. A large number of algorithms have been developed for weighted, directed, and dynamical networks. Little or no effort has so far been devoted to the practical problem of incomplete information, i.e. missing and imperfect network data. Earlier applications either removes these uncertain connections or wrongly includes them. This work aims to utilize the imperfect information is a more general and structured manner.

Fusion of Communities

The proposed framework to detect communities in uncertain and imperfect networks is named Fusion of Communities. The approach generates a large number of weak estimates of the community structure, called candidate communities. These are merged into a single structure by one of three methods:

- Two-step Fusion of Communities (TFC)
- Node-based Fusion of Communities (NFC)
- Community-based Fusion of Communities (CFC)

These methods are based on the idea of ensemble classification from the field of machine learning. Where a group of weak classifiers have been shown to often perform better than a single more advanced classifier. The proposed framework also enables more complicated uncertainties (triads, chains, or other structures) to be utilized in community detection.

The results indicate good convergence properties on both real-world and randomly generated networks. This framework is thus a viable method to detect communities in imperfect networks, as well as a methods to improve simple fast detection algorithms by merging several runs.

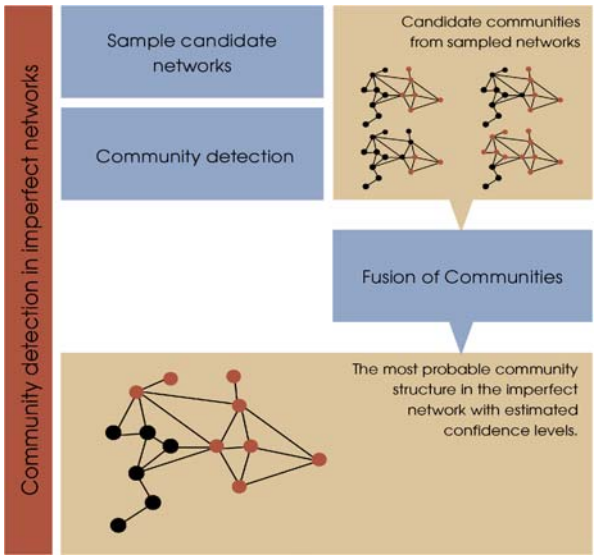

Figure 1: A simple cartoon of the framework to find communities in networks with uncertain and imperfect information. (i) the candidate networks are sampled from the ensemble of network consistent with the given information. (ii) candidate communities are found using standard techniques from SNA. (iii) the candidate communities are fusied into a single (most probable) community structure using one of three proposed algorithms, focusing on either nodes or clusters.

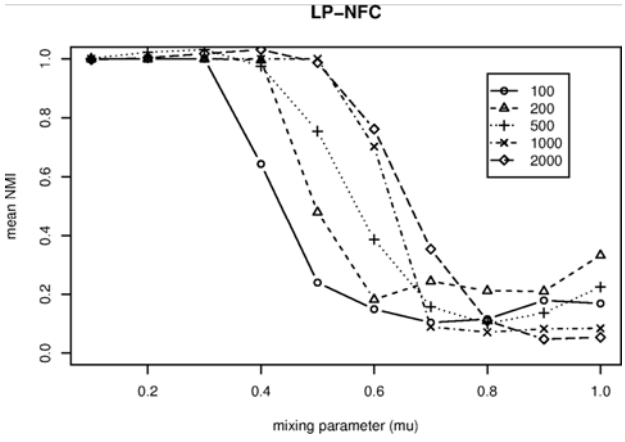

Figure 2: The mean Normalized Mutual Information (NMI) for the Label Propagation algorithm used in conjuncture with the Node-based Fusion of Communities approach. The algorithms are applied on scrambled randomly generated networks with community structures. The mixing parameter determines the fraction of edges placed between communities (lower values give more separated and distinct groupings.

Contact: pontus.svenson@foi.se Web: www.foi.se