Community Detection in Uncertain Networks by Ensemble approaches

Johan Dahlin[a,b] and Pontus Svenson[a]
[a] FOI – Swedish Defence Research Agency, Sweden
[b] Department of Electrical Engineering, Linköping University, Sweden

Abstract—**Social network analysis can be an important tool to investigate networks of, e.g., criminals, terrorists, electrical power stations, or biological processes. In real world applications, there is seldom complete knowledge about the network of interest – we only have partial and incomplete information about the nodes and networks present. Community detection in networks is an important area of current research in social network analysis with many applications. Finding community structures is however a challenging task and despite significant effort no satisfactory method has been found. Here we study the problem of community detection in noisy and uncertain networks with missing and false edges and propose methods for detecting community structures in them. The method is based on sampling from an ensemble of certain networks that are consistent with the available information about the uncertain networks.**

I. Introduction And Background

The first applications of social network analysis were by sociologists who collected data by questionnaires or direct observations and used graph visualization to gain insight into the communication behavior of small groups. The field has expanded considerably since then, and has also been influenced by the advance in computer technology which has enabled collection of network data in enormous quantities. With the increased availability of data, however, come new challenges: algorithms are needed to handle much larger networks than previously, and some way has to be devised to take account of the inherent uncertainty of the data. In recent years, there has been much development in the social network analysis field of new, more efficient algorithms for handling ever larger amounts of data. So far, however, there has been very little done taking account of the uncertainty of the data. Uncertainty can arise in many different ways, depending on the way that the network data has been collected. Different observation models are needed for different kinds of data collection.

In previous work [1,2,3], we introduced and studied methods for computing community structures of networks where we do not have complete knowledge of nodes and edges. The methods are based on generating an ensemble of certain networks that are consistent with the information available about the real network. Community structures are then computed for each such certain network, and the results merged. The methods can be used not only when we have knowledge about edge probabilities, but also if there is information about more complicated network substructures and their probabilities. The methods for merging the results of the community detection methods can also be used to merge the results of several different community detection algorithms applied to the same certain network. This is especially useful for varying parameters determining the resolution limit in networks, i.e. the property of community detection algorithm to find communities of a certain size.

II. Results

We have used the method to compute community structure for both real-world networks and simulated (random) networks with community structures. The latter is accomplished using newly developed random network models and allows for varying e.g. the size of the network in a more controlled manner. The performance of the community detection method is evaluated by artificially generated imperfections and uncertainties in certain network structures. The resulting community structure is compared using Normalized Mutual Information (NMI) with a given external labeling of the community structure in the certain version of the network.

The results indicate that at least one of the three methods introduced in [1] generates good results and is able to recover the original community structure despite quite severe added imperfections. This is indicated in the figure below where the number of nodes is varied in random networks with different mixing parameters (describing the fraction of internal versus external edges in communities, i.e. higher mixing parameter results in more diffuse communities).

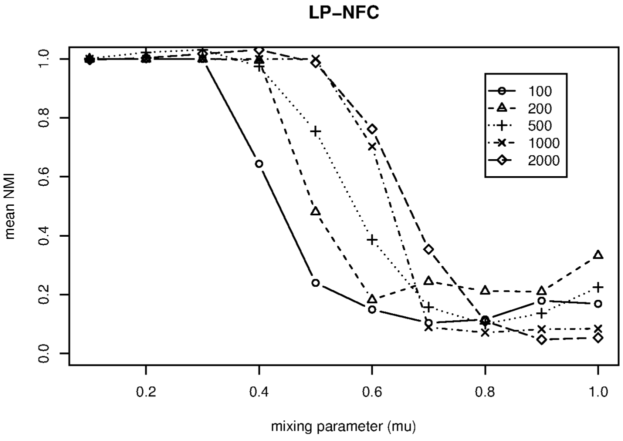

Figure: The mean NMI for random networks with a varying number of nodes and mixing parameters for the method: Label Propagation-Node-based Fusion of Communities.

Acknowledgements: This research was funded by the R&D Programme of the Swedish Armed Forces.

References

[1] Johan Dahlin, "Community Detection in Imperfect Networks", Master's thesis Umeå University 2011, FOI report no FOI-S-3710-SE

[2] Johan Dahlin, Pontus Svenson, "A Method for Community Detection in Uncertain Networks", to appear in Proceedings of the European Intelligence and Security Informatics Conference 2011

[3] Johan Dahlin, Pontus Svenson, "Ensemble approaches for improving community detection", manuscript in preparation for submission to Physical Review (2011)