Detecting community structures in uncertain social networks

Johan Dahlin

Swedish Defence Research Agency,
Department of Physics, Umeå University.

johan.dahlin@foi.se

March 30, 2011

Outline

1. Introduction
2. Uncertain social networks
3. Merging candidate communities
4. Examples
5. Concluding remarks

Summary

In this talk, I will present these two results from work in progress:

- Merging results from community detection methods can be used to improve performance of some methods.
- Merging can be used to find and analyze the community structure in uncertain (social) networks.

Uncertain network information

Why should we be concerned with uncertain networks? What is the problem?

- Difficult or impossible to find the entires network.
- Uncertain or contradicting information (measurement errors, lack of information etc.).
- Previous work have ignored the problem by only allowing binary edge structures.

Uncertain network information

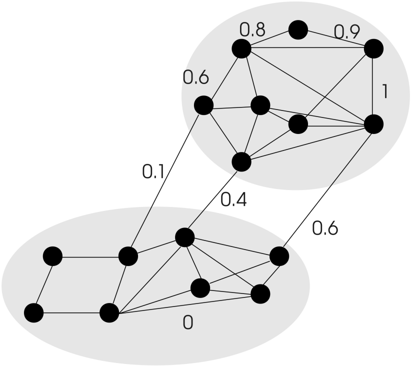


Figure: A small uncertain network.

Uncertain network information

- Assume that observation of the real network g is not possible.
- Only some proxy to the network f is observable.
- The different cases are contained in a observation model,

$$P(g_{ij}) = FP + TP = P(g_{ij}|\neg f_{ij}) + P(g_{ij}|f_{ij}), \quad (1)$$

$$P(\neg g_{ij}) = TN + FN = P(\neg g_{ij}|\neg f_{ij}) + P(\neg g_{ij}|f_{ij}), \quad (2)$$

  where $P(g_{ij})$ is the probability that an edge, $e(i,j)$, exist.
- FP is *false links* and FN is *missing links*.

Community detection

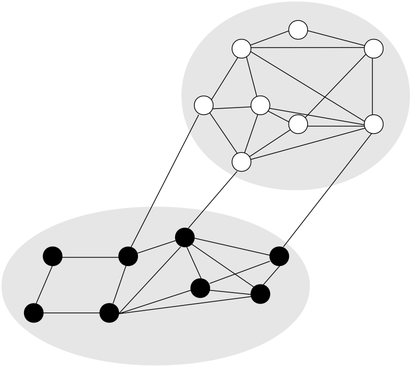


Figure: Two communities in a small network.

Community detection

A community is *a subset of nodes within a graph such that connections between nodes are denser than connections with the rest of the network* (Radicchi et. al., 2004).

Modularity (how unlikely that the structure is the result of randomnes) is a quality measure often used. Higher modularity is better.

Community detection

- Detecting communities is a non-trivial and important problem.
- Applications in many different fields and problems.
- No completely satisfactory method has been developed so far.
- What about merging the results from different methods?

Detecting communities in uncertain networks

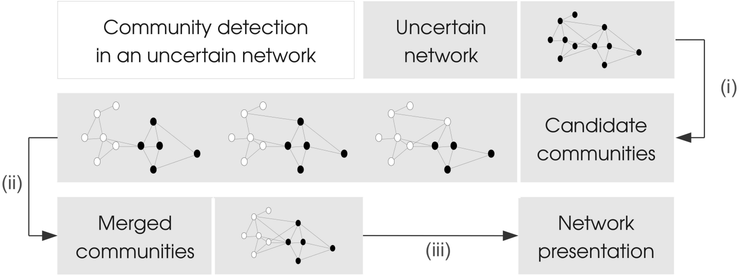


Figure: The proposed steps to detect community structures in social networks. (i) sample from the ensemble of consistent networks and find the community structure in each of them. (ii) merge the candidate communities. (iii) Validate the community structure found and evaluate some diagnostic measures.

Testing the method

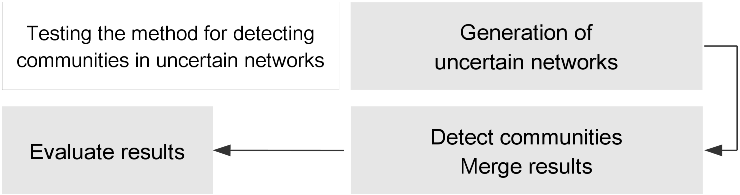

Testing the method for detecting communities in uncertain networks
Generation of uncertain networks
Detect communities
Merge results
Evaluate results

Figure: The proposed steps to test the method.

Generating test networks and candidate communities

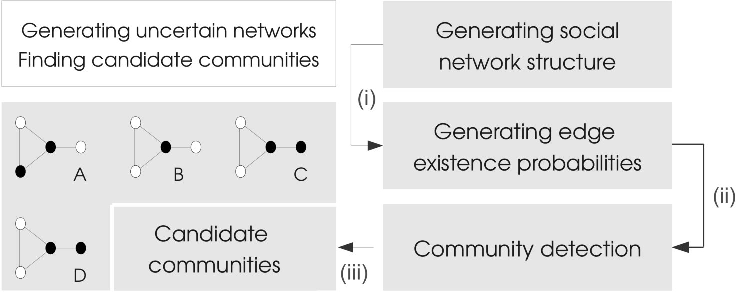

Generating uncertain networks
Finding candidate communities
Generating social network structure
(i)
Generating edge existence probabilities
(ii)
Community detection
(iii)
Candidate communities
A
B
C
D

Figure: The proposed steps to (i) generate uncertain social networks, (ii) sample candidate networks and (iii) find the community structure of each candidate network.

Generating test networks and candidate communities

- In uncertain social networks edges are not completely known.
- To test the method, some uncertain social networks are generated.
- Candidate networks are sampled randomly from an ensemble of consistent networks.
- For each candidate network a community detection is made.

Generating edge existence probabilities

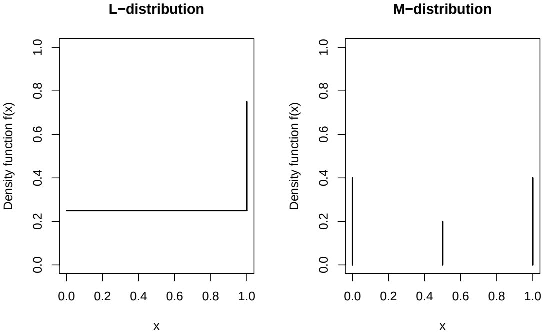
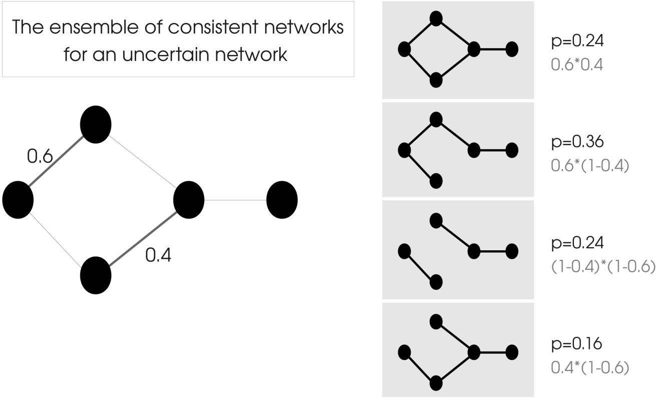


Figure: Two probability density functions for generating edge existence probabilities.

Generating edge existence probabilities

- L-distribution varies the amount of edges that are certain versus uncertain.
- In the L-distribution, when $p_t = 0$, all edges are uncertain and larger p_t increases the amount of certain edges.

Sampling candidate communities



The ensemble of consistent networks for an uncertain network

Figure: An example of the generation of networks consistent with the uncertain network.

Sampling candidate communities (details)

- Idea: Need certain networks for detecting communities. Create a ensemble consistent with the uncertain network.

Sampling candidate communities (details)

- Idea: Need certain networks for detecting communities. Create a ensemble consistent with the uncertain network.
- A uniformly distributed random matrix, $\mathbf{R} = [R_{ij}]$, is generated with $R_{ij} \sim \mathcal{U}[0, 1]$.
- An edge is included in the graph if, $R_{ij} \leq E_{ij}$, where E_{ij} is the edge existence probability.
- Repeating this creates an ensemble of candidate networks consistent with the uncertain information.

Merging candidate communities

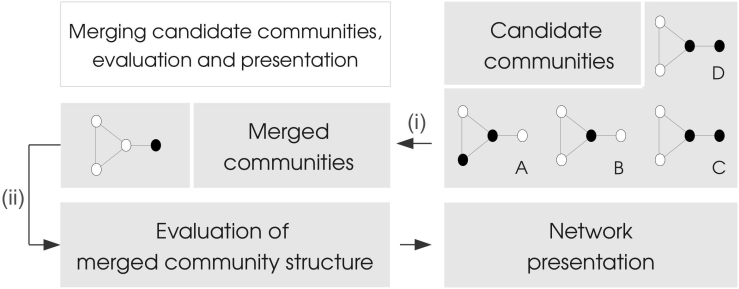

Merging candidate communities, evaluation and presentation
Candidate communities
D
Merged communities
(i)
A
B
C
(ii)
Evaluation of merged community structure
Network presentation

Figure: The proposed steps to (i) merge candidate communities, (ii) evaluate merged communities and (iii) present the network.

Community merging methods

- Two-Step Community Merging (TSCM).
- Instance-Based Community Merging (IBCM).
- Cluster-Based Community Merging (CBCM).

Two-step Community Merging

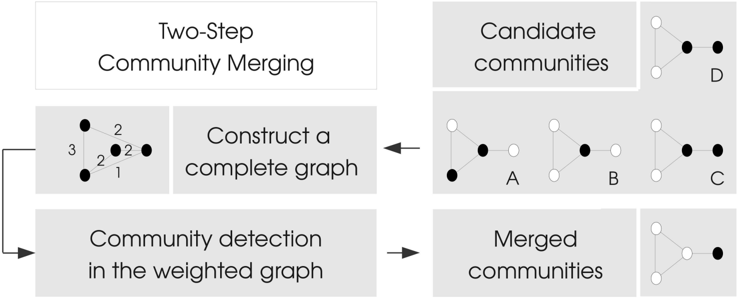


Figure: Two-step Community Merging using a community detection method, that can handle a weighted network, twice.

Two-step Community Merging (details)

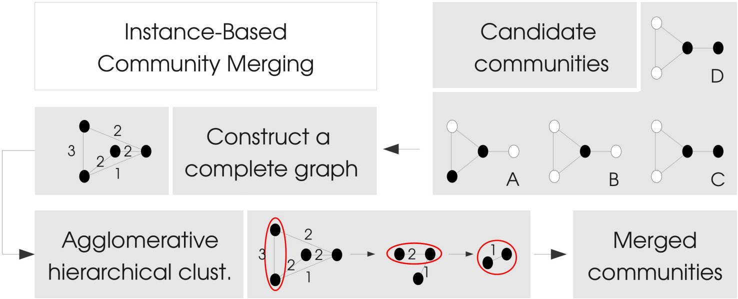
- Create a new (complete) graph with the same nodes and edge weight as the frequency that nodes have been clustered together.
- Run community detection on this new weighted graph.
- The presented communities from the merge graph are taken as the communities found in the candidate communities.

Ensemble clustering methods

- Problem: How to combine a given ensemble of clusterings to produce a single clustering?
- Often occurs in re-sampling methods (Dudiot and Fridlyand, 2003) or random projections (Fern and Brodley, 2004).

- Strehl and Ghosh (2003) propose two approaches to the problem, using *Instance-based Ensemble Clustering* (IBEC) and *Cluster-based Ensemble Clustering* (CBEC).

Instance-Based Community Merging



Instance-Based Community Merging

Candidate communities

Construct a complete graph

Agglomerative hierarchical clust.

Merged communities

Figure: Instance-based Community merging using agglomerative hierarchical clustering with a special linkage calculation.

Instance-Based Ensemble Clustering (details)

- Create a new graph of instances and use some similarity measure to describe the connection between instances.
- Similarity measure used is the fraction of instances that two nodes are in the same cluster.

Instance-Based Ensemble Clustering (details)

- Create a new graph of instances and use some similarity measure to describe the connection between instances.
- Similarity measure used is the fraction of instances that two nodes are in the same cluster.
- Following Strehl and Ghosh (2003) by partitioning the generated graph by agglomerative hierarchal clustering.
- Using a special linkage rule to account for the network structure,

$$\mathrm{sim}(l, j) = |M_{jl} \cap M_{i_1 l} \cap M_{i_2 l} \cap \ldots \cap M_{i_{n_l} l}|,$$

for some cluster l containing nodes $v_1, v_2, \ldots, v_{n_l}$. This incur some information loss.

Cluster-Based Community Merging

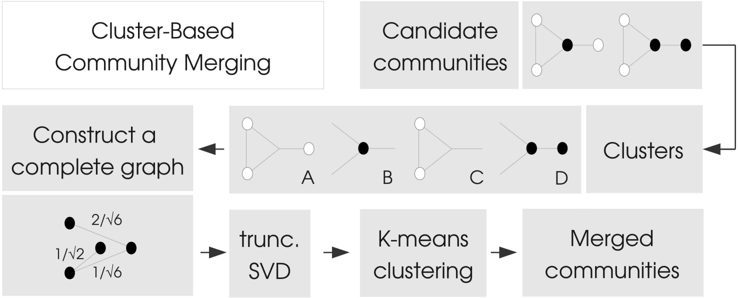


Figure: Cluster-based Community merging using k-means clustering on the truncated singular value decomposition of the adjacency matrix. The example uses the cosine measure to estimate the similarity between clusters.

Cluster-Based Ensemble Clustering (details)

- Clusters from different ensembles that have large overlap are similar to each other.
- Construct a graph of candidate clusters and some similarity between nodes.

Cluster-Based Ensemble Clustering (details)

- Clusters from different ensembles that have large overlap are similar to each other.
- Construct a graph of candidate clusters and some similarity between nodes.
- Which similarity measure should be used?

  $$\mathrm{sim}(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{\sqrt{|\mathbf{x}||\mathbf{y}|}}$$

  where **x** and **y** are two clusterings.
- Following Ng and Weiss (2001) by partitioning the generated graph by spectral partitioning methods.

Merging results from label propagation algorithm

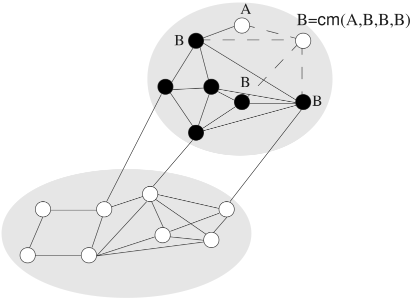


Figure: The label propagation algorithm (Raghavan et al., 2007) in action.

Merging results from label propagation algorithm

- The label propagation algorithm (Raghavan et al., 2007) is a fast stochastic community detection algorithm.
- Generates a different result for each run.
- A merger would probably be useful for creating a faster accurate community detection algorithm.

Merging results from label propagation algorithm

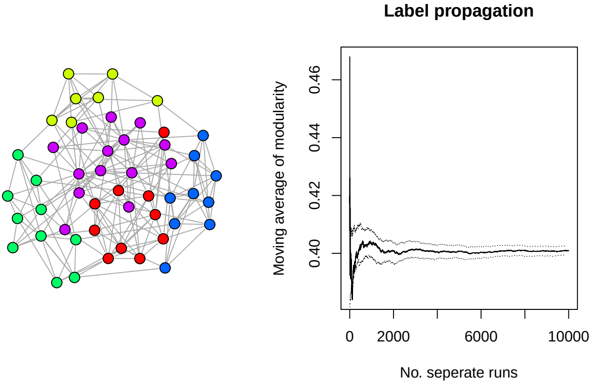


Figure: Random network generated using algorithms from Lancichinetti and Fortunato (2009).

Merging results from label propagation algorithm

	Modularity			Difference in mod.			No. com.		
N_s	T	I	C	T	I	C	T	I	C
2	0.468	0.468	0.468	0.000	0.000	0.000	5	5	5
5	0.348	0.468	0.437	-0.120	0.000	-0.031	3	5	4
10	0.348	0.428	0.468	-0.080	0.040	0.040	3	5	5
15	0.468	0.408	0.468	0.060	0.060	0.060	5	5	5
20	0.468	0.420	0.468	0.048	0.048	0.048	5	5	5
25	0.405	0.417	0.424	-0.013	0.051	0.007	4	5	4
50	0.468	0.425	0.393	0.043	0.043	-0.032	5	5	3
100	0.437	0.426	0.433	0.011	0.042	0.007	4	5	4

Table: Results from TSCM, IBCM and CBCM for the label propagation algorithm on a random network. Difference calculate in comparison the modularity from spin glass method, 0.468.

Merging results from label propagation algorithm

- Merging several runs of the label propagation algorithm increases the modularity.
- Results in a faster and more accurate method for community detection.
- Difficult to identify the best choice on this network and algorithm.

Merging results from uncertain social network

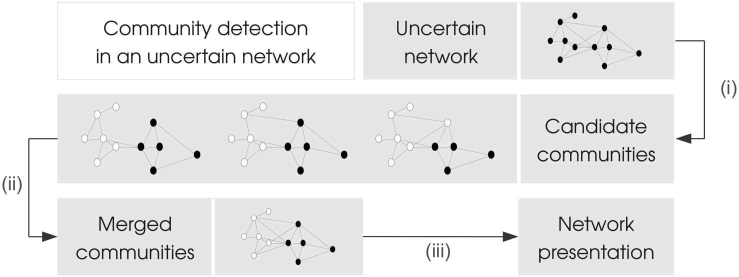


Figure: The proposed steps to detect community structures in social networks. (i) sample from the ensemble of consistent networks and find the community structure in each of them. (ii) merge the candidate communities. (iii) Validate the community structure found and evaluate some diagnostic measures.

Merging results from uncertain social network

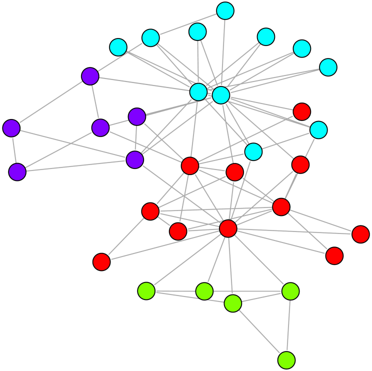


Figure: The famous Karate Network (Zachary, 1977) with the communities as found by the spin glass method (Reichardt and Bornholdt, 2006).

Merging results from uncertain social network

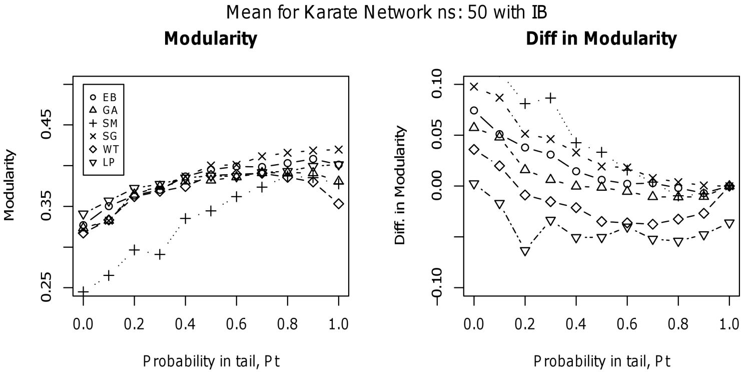


Figure: 20 runs (each with a different set of probabilities for edge existence) on the Karate network at $N_s = 50$ samplings and different P_t, using the Instance-Based Community Merging-method.

Merging results from uncertain social network

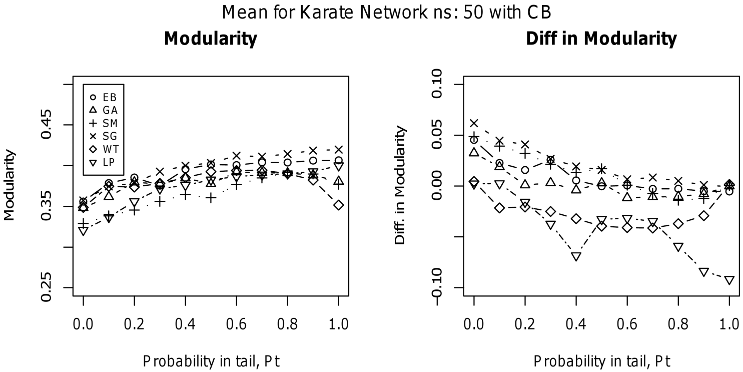


Figure: 20 runs (each with a different set of probabilities for edge existence) on the Karate network at $N_s = 50$ samplings and different P_t, using the Cluster-Based Community Merging-method.

Merging results from uncertain social network

- The community structure is recovered from the original network, at quite low p_t.
- The best combination is given by Cluster-based merging of spin glass algorithm.

Summary

- Merging results from community detection methods can be used to improve performance of some methods.
- Merging can be used to find and analyze the community structure in uncertain (social) networks.

Further work

- Test the methods on networks with missing and false edges.
- Analyze the performance of merging on different networks.
- Bipartite graph partitioning to merge candidate communities.

The end

Thank you for your attention

Let the estimated (correct) community structure represented by a neighbor matrix $\hat{\mathbf{N}} = [\hat{N}_{ij}]$, where $\hat{N}ij = 1$ if nodes i and j are found in the same community and 0 otherwise. The matrix MSE is the mean of the square root of MSE for each row,

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\hat{N}_{ij} - N_{ij}\right)^{2}}. \tag{3}$$

The MSE of the matrix then corresponds to how many elements in the neighbor matrices that differ.

The correlation is also used to calculate a slightly different value, which indicates how the rows in the matrices tends to be similar. The mean correlation $\bar{\rho}$, between the two matrices, N and $\hat{N}$, is found as the mean of the *Pearson correlations* for each row, ρ_i,

$$\bar{\rho} = \frac{1}{n}\sum_{i=1}^{n}\rho_i = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{Cov}(N_i,\hat{N}_i)}{\mathbb{V}[N_i]\mathbb{V}[\hat{N}_i]}. \tag{4}$$

using the covariance between each element in each row and the expected value (mean) of the that specific row,

$$\mathrm{Cov}(N_i,\hat{N}_i) = \frac{1}{n}\sum_{j=1}^{n}(N_{ij} - \mathbb{E}[N_i])(\hat{N}_{ij} - \mathbb{E}[\hat{N}_i]). \tag{5}$$

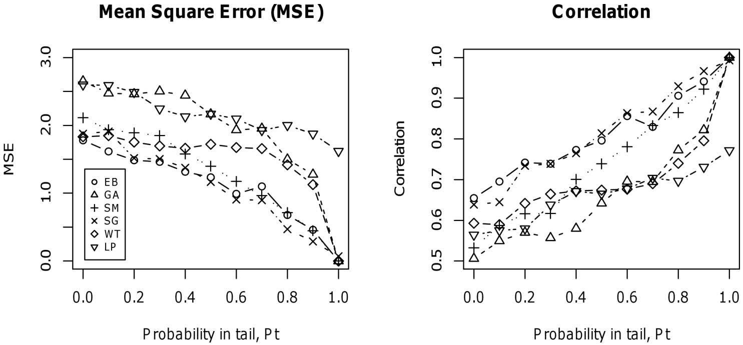


Figure: 20 runs (each with a different set of probabilities for edge existence) on the Karate network at $N_s = 50$ samplings and different P_t, using the Instance-Based Community Merging-method.

Additional remarks

Merging results from uncertain social network

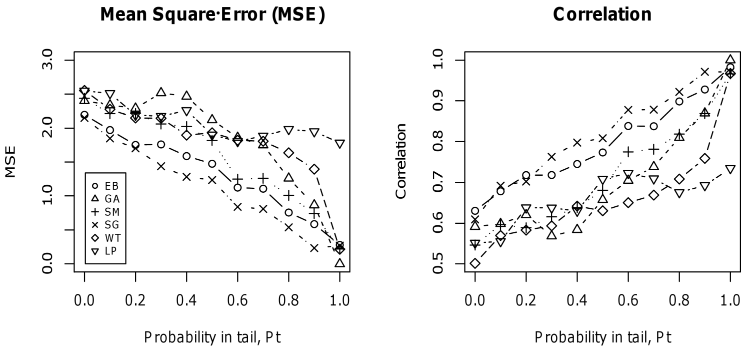


Figure: 20 runs (each with a different set of probabilities for edge existence) on the Karate network at $N_s = 50$ samplings and different P_t, using the Cluster-Based Community Merging-method.