Learning latent structure in complex networks

Lars Kai Hansen
www.imm.dtu.dk/~lkh

Current network research issues:

- Social Media
- Neuroinformatics
- Machine learning

Joint work with
Morten Mørup, Sune Lehmann

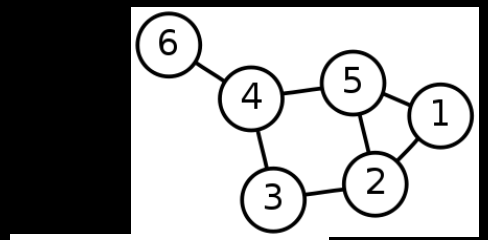
DTU Informatics
Department of Informatics and Mathematical Modeling

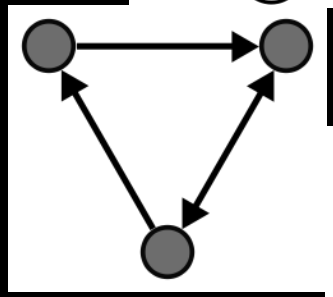
Network models

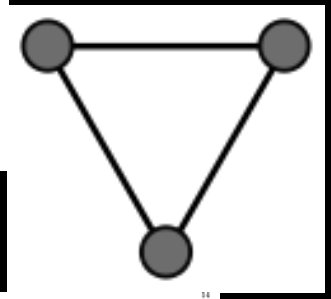
- N nodes/vertices and links/edges
  - Directed / undirected
  - Weighted / un-weighted
  - Here A_{ij} is symmetric matrix of 1/0's

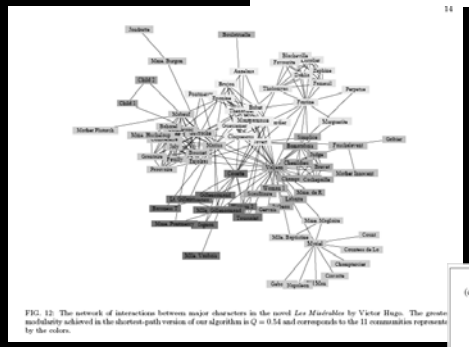
- Link distributions
  - Random
  - Long tail
  - Hubs and authorities
  - Friends of friends are friends
  - Assortative mixing "The rich club"

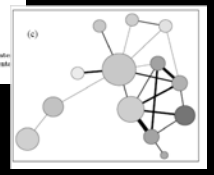
- Communities
  - A community is a set of densely linked nodes
  - Typically community structure is "hidden" or "latent"











DTU Informatics
Department of Informatics and Mathematical Modeling

The main points…

- Community detection can be formulated as an inference problem
- The success of the inference depends on the link sampling process. There is a phase transition like detection threshold. The location can be estimated with mean field analysis
- The phase transition shifts (sharpens?) if we simultaneously learn the parameters of a generative model
- For good link prediction we need more complex latent structures: Simple community models do not beat basic heuristics

Why look for latent community structure?

Communities may represent different mechanisms, hence different statistics … the network is non-stationarity

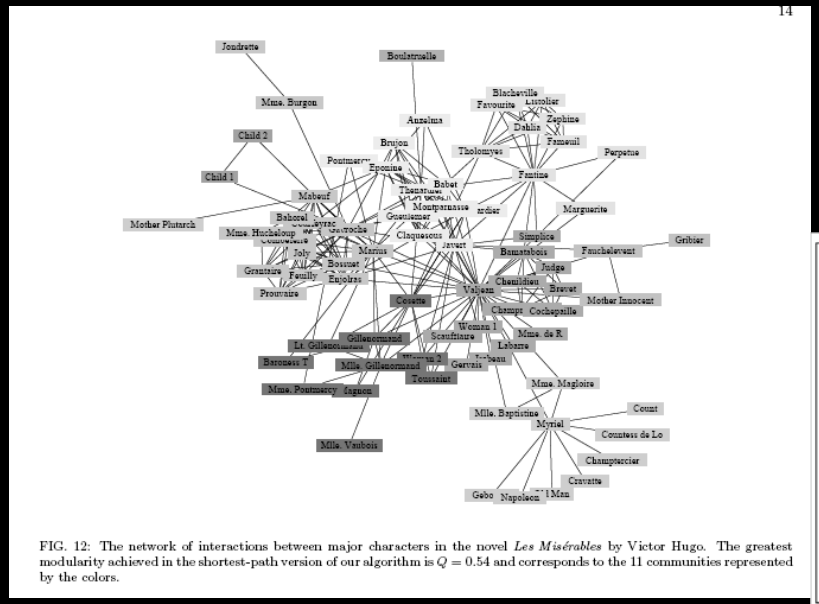
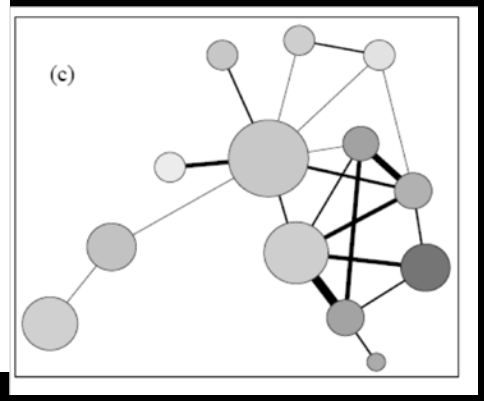


FIG. 12: The network of interactions between major characters in the novel *Les Misérables* by Victor Hugo. The greatest modularity achieved in the shortest-path version of our algorithm is $Q = 0.54$ and corresponds to the 11 communities represented by the colors.

Communities detected by spectral clustering



M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113 (2004)

Why look for latent community structure?

Communities may be predictive of dynamics and structural (in-)stability, e.g., Palla et al. (2007):

Small communities depend on stable core of membership, large communities can persist longer if they renew membership

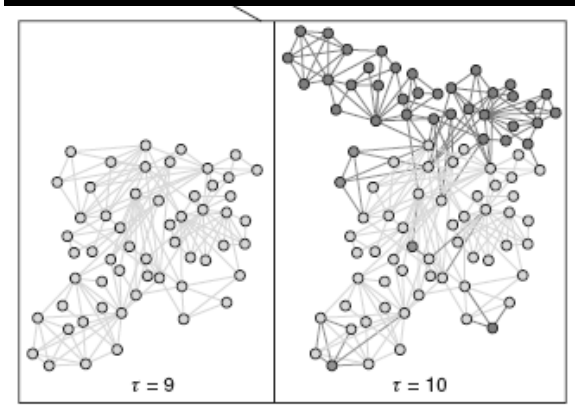
Communities found by the clique percolation method (CPM) for detection of overlapping communities

nature Vol 446|5 April 2007|doi:10.1038/nature05670

LETTERS

Quantifying social group evolution

Gergely Palla[1], Albert-László Barabási[2] & Tamás Vicsek[1,3]



Why look for latent community structure?

Community structure may assist link prediction

	Degr. Prod.	Short. Path	Com. Neigh.	Jaccard	Hamilt.	Modularity	H&W	MMSB	LD
H&W model	0.627(5)	0.788(7)	0.827(8)	0.828(8)	0.869(2)	0.867(3)	**0.873(2)**	0.868(2)	0.872(1)
MMSB model	0.621(5)	0.786(5)	0.840(5)	0.845(5)	0.866(2)	0.856(6)	0.861(3)	**0.892(1)**	0.888(1)
LD model	0.690(2)	0.688(2)	0.834(3)	0.825(3)	0.866(2)	0.866(2)	0.870(1)	0.904(1)	**0.905(1)**
Yeast Network	0.783(6)	0.795(9)	0.675(9)	0.675(9)	0.640(13)	0.599(14)	0.573(9)	**0.836(7)**	0.794(9)
US Power	0.449(10)	**0.795(6)**	0.474(13)	0.474(13)	0.479(12)	0.489(13)	0.544(13)	0.407(8)	0.506(9)
Erdos02	0.586(9)	0.580(9)	0.584(5)	0.559(8)	0.530(15)	0.460(11)	0.377(12)	**0.954(3)**	0.873(22)
Free Assoc.	0.845(6)	0.877(5)	0.856(6)	0.854(6)	0.766(11)	0.632(8)	0.609(6)	**0.902(5)**	0.872(4)
Reuters911	0.928(3)	0.892(4)	0.929(3)	0.910(3)	0.728(8)	0.601(5)	0.766(4)	**0.942(2)**	0.935(2)
Wordnet3	0.602(6)	0.613(6)	0.356(4)	0.356(4)	0.507(8)	0.463(7)	0.481(6)	**0.795(8)**	0.658(8)
Dictionary28	0.808(5)	**0.917(4)**	0.776(4)	0.744(5)	0.644(6)	0.633(4)	0.678(6)	0.865(6)	**0.894(3)**
CondPhys2005	0.790(3)	0.963(2)	0.964(1)	**0.965(1)**	0.737(8)	0.689(5)	0.463(12)	0.909(2)	**0.924(1)**
Internet	0.604(5)	**0.745(4)**	0.501(4)	0.501(4)	0.662(4)	0.672(5)	0.607(8)	0.695(5)	0.663(5)
IMDB	0.918(1)	0.980(5)	**0.998(0)**	0.997(1)	0.864(2)	0.857(2)	0.861(2)	0.965(2)	0.974
Patents	0.766(1)	**0.946(3)**	0.743(5)	0.743(5)	0.770(1)	0.768(1)	0.696(14)	0.889(1)	-

Table 2: Link prediction performance of the various methods on the 14 networks. Given are the mean AUC values as well as their standard deviations on the last digit given in parenthesis across 10 data splits with randomly chosen links and non-links treated as missing. In bold black is given the best performing approach and in underline the best performing community detection approach. The Hamiltonian was based on average link density as the imposed null hypothesis, i.e. $B = \frac{m}{n^2}E$ all clusters were modelled with $c = 50$, as such a more dramatic difference in link performance was observed for the H&W, MMSB and LD models when initialized with the true number of clusters, however, in general the true number of clusters is unknown hence these results are not shown.

M. Mørup, L.K. Hansen: Learning latent structure in complex networks NIPS workshop "Analyzing Networks and Learning With Graphs" (2009)

Outline

- Community detection
  - Detection as a combinatorial optimization problem
  - Mean field approx and Gibbs sampling
  - Phase transition leads to sharp detection threshold
- Learning community parameters
  - The Hofman-Wiggins model
  - Detection threshold with learning
  - Stochastic block models
- Link prediction
  - Robustness to link dilution

DTU Informatics
Department of Informatics and Mathematical Modeling

Formal community detection .. Newman's Modularity

The modularity is expressed as a sum over links, such that we reward excess links in communities relative to a baseline measure P_{ij}

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - P_{ij} \right] \delta(c_i, c_j)$$

With $c_i = k$ if node i is in community k, *a total of* m links … $2m = \Sigma_{ij} A_{ij,}$

The baseline assumes independence $P_{ij} = k_i k_j / 2m$, *with* $k_i = \Sigma_j A_{ij,}$

Combinatorial optimization problem

M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69:026113, 2004, cond-mat/0308217.

DTU Informatics
Department of Informatics and Mathematical Modeling

Potts representation

Introduce *C* x *N* binary matrices S encoding the community assignment

$$\delta(c_i, c_j) = \sum_k S_{ki} S_{kj}$$

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - P_{ij} \right] \sum_k S_{ki} S_{kj}$$

$$Q = \frac{1}{2m} \sum_{ijk} B_{ij} S_{ki} S_{kj} = \frac{Tr(S'BS)}{2m}$$

Spectral heuristic

Newman makes a relaxation of the optimization problem to the unit sphere

$$Q = \frac{1}{2m} \sum_{ijk} B_{ij} S_{ki} S_{kj} = \frac{Tr(S'BS)}{2m}$$

$$BS = S\Lambda$$

Procedure: solve the eigenvalue problem to get a Fiedler vector (can be repeated), convert it to assignments and improve the resulting pre-community structure by local Lin-Kernighan post-processing

M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69:026113, 2004, cond-mat/0308217.

Combinatorial optimization

- Alternatively, use Gibbs sampling with simulated annealing (Kirkpatrick et al. 1983, Geman, Geman 1984)

$$P(S \mid A,T) = \exp\left(\frac{Q(S)}{T}\right) = \exp\left(\frac{Tr(SBS')}{2mT}\right)$$

- Monte Carlo realization of a Markov process in which each variable is randomly assigned according to its marginal distribution

$$P(S_j \mid S_{-j}, A,T) = \frac{P(S \mid A,T)}{\sum_{S_j} P(S \mid A,T)}$$

S Geman,D Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (6): 721–741 (1984)

Gibbs sampling

$$\varphi_{ki} = \sum_j \frac{B_{ij}}{2m} S_{kj} = \sum_j \frac{A_{ij}}{2m} S_{kj} - \sum_j \frac{k_i}{2m} \frac{k_j}{2m} S_{kj}$$

$$\mu_{ki} = \frac{\exp(\varphi_{ki} / T)}{\sum_{k'} \exp(\varphi_{k'i} / T)}$$

$$S_{ki} = \text{discrete}(\mu_{ki})$$

Potts model: single node

- Discrete probability distribution on states k = 1,…,C

$$P(S \mid \varphi, T) \propto \exp\left(\frac{\sum_{k=1}^{C} S_k \varphi_k}{T}\right)$$

$$P(S \mid \varphi, T) = \prod_k (\mu_k)^{S_k}$$

$$\mu_k = \frac{\exp\left(\frac{\varphi_k}{T}\right)}{\sum_{k'} \exp\left(\frac{\varphi_{k'}}{T}\right)}$$

Mean Field method:

Approximate the posterior by product of discrete distributions

Minimize the KL distance between P(S|μ) and P(S|A,p,q)

$$P(S \mid \mu) = \prod_{ki} (\mu_{ki})^{S_{ki}}$$

$$\mu_{ki} = \frac{\exp(\varphi_{ki} / T)}{\sum_{k'} \exp(\varphi_{k'i} / T)}$$

$$\varphi_{ki} = \sum_j \frac{B_{ij}}{2m} \mu_{kj} = \sum_j \frac{A_{ij}}{2m} \mu_{kj} - \sum_j \frac{k_i}{2m} \frac{k_j}{2m} \mu_{kj}$$

DTU Informatics
Department of Informatics and Mathematical Modeling

S. Lehmann, L.K. Hansen: Deterministic modularity optimization
European Physical Journal B 60(1) 83-88 (2007).

Deterministic annealing

Iterative solution with a decreasing sequence of temperatures to reach the ground state = MAP solution

$$\mu_{ki}^{(t+1)} = \frac{\exp(\varphi_{ki}^{(t)} / T)}{\sum_{k'} \exp(\varphi_{k'i}^{(t)} / T)}$$

$$\varphi_{ki}^{(t)} = \sum_j \frac{B_{ij}}{2m} \mu_{kj}^{(t)} = \sum_j \frac{A_{ij}}{2m} \mu_{kj}^{(t)} - \sum_j \frac{k_i}{2m} \frac{k_j}{2m} \mu_{kj}^{(t)}$$

Experimental evaluation

Create a simple testbed with within link probability p and between "noise" links q

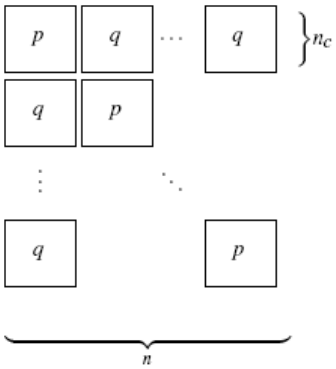


FIG. 1: A sketch of the simple network model. The figure displays the structure of the adjacency matrix with nodes arranged according to community. Inside each community (the blocks) along the diagonal, the probability of a link between two nodes is p and between communities, the probability of a link is q.

$q = f \cdot p$

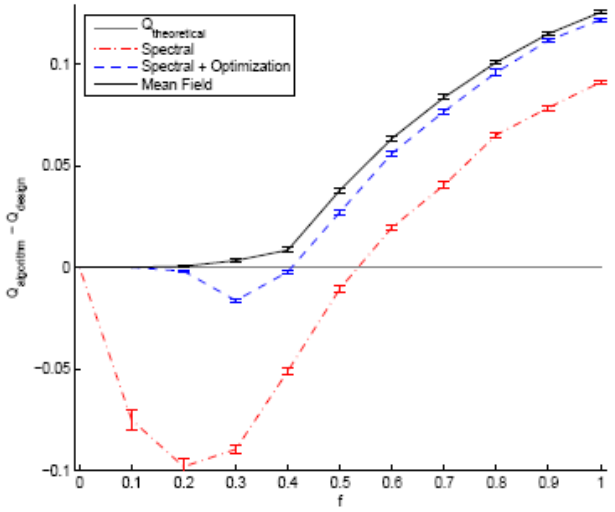


FIG. 3: Comparing spectral methods with the mean field solution. The networks were created according to the simple model, using parameters $n_c = 100$, $C = 5$, $p = 0.1$ and $f \in [0, 1]$. All data points display the point-wise differences between the value of $Q_{\text{algorithm}}$ found by the algorithm in question and Q_{design}. The error-bars are calculated as in Figure 2. The dash-dotted red line shows the results for the spectral method. The dashed blue line shows the results for the spectral optimization followed by KLN post-processing. The solid black curve shows the results for the mean field optimization. The grey, horizontal line corresponds to the theoretical prediction (Eq. (22)) for the designed communities.

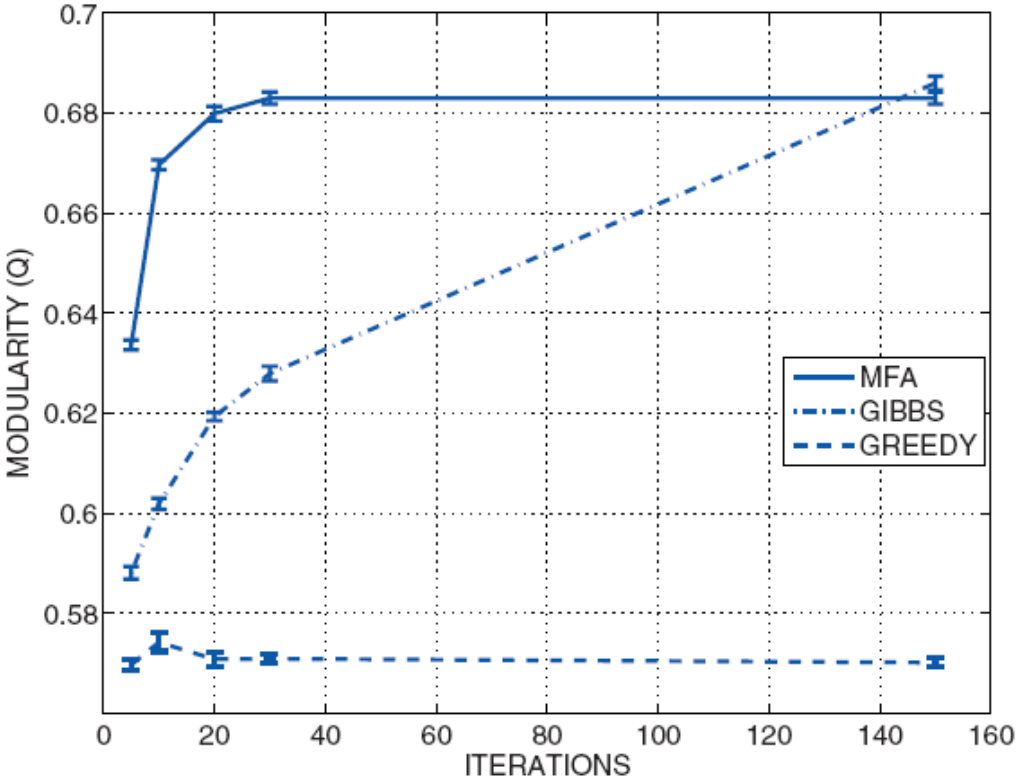


Fig. 5. Comparison of the modularity performance for mean field annealing, Gibbs sampling, and a 'greedy' optimizer. The greedy algorithm corresponds to Gibbs sampling at $T = 0$. The graph describes the co-authorship network of the Los Alamos condensed matter preprint archive, considering articles published between April 1998 and February 2004 [26], it has $n = 30\,561$ nodes, and $m = 125\,959$ links. The mean field method provides good modularity solutions for very few iterations, for the present graph the Gibbs sampling scheme outperforms mean field annealing at 150 iterations. The best modularity solutions we found in this network after extensive Gibbs sampling have $Q \equiv 0.71$. The zero temperature greedy search does not produce useful modularity solutions here.

S. Lehmann, L.K. Hansen:
Deterministic modularity optimization
European Physical Journal B 60(1) 83-88 (2007).

The Hofman-Wiggins model (2008)

- The H&W model is generalization of the Modularity heuristic to a proper statistical model with Bayesian inference

$$P(A \mid S, p, q) = \prod_{i>j} \left(p^{d_{ij}} q^{(1-d_{ij})} \right)^{A_{ij}} \left((1-p)^{d_{ij}} (1-q)^{(1-d_{ij})} \right)^{1-A_{ij}}$$

$$d_{ij} = \sum_k S_{ki} S_{kj}$$

- Consider the link probability parameters within p and between q unknown

J.M. Hofman and C.H. Wiggins.
A Bayesian approach to network modularity.
Phys. Rev. Lett. 100:258701, 2008.

The Hofman-Wiggins model (2008)

- Critical behavior for fixed parameters,

$$P(S \mid A, p, q) \propto \prod_{i>j} \left(p^{d_{ij}} q^{(1-d_{ij})} \right)^{A_{ij}} \left((1-p)^{d_{ij}} (1-q)^{(1-d_{ij})} \right)^{1-A_{ij}}$$

$$P(S \mid A, p, q) = Z^{-1} \exp\left(\log\left[\frac{p}{1-p} \frac{1-q}{q} \right] \sum_{ij} \sum_{k} S_{ki} S_{kj} A_{ij} \right)$$

Effective inverse temperature

The Hofman-Wiggins model (2008)

- Mean field critical behavior for fixed parameters, as function of p, q and A

$$\mu_{ki}^{(t+1)} = \frac{\exp(\varphi_{ki}^{(t)} / T(p,q))}{\sum_{k'} \exp(\varphi_{k'i}^{(t)} / T(p,q))}$$

$$\varphi_{ki}^{(t)} = \sum_j \mu_{kj}^{(t)} A_{ji}$$

Converges to $\mu = 1/C$ for $T > T_c$

$$T(p,q)^{-1} = \log\left(\frac{p}{1-p}\frac{1-q}{q}\right) \approx \log\left(\frac{p}{q}\right) \equiv \log \mathrm{SNR}$$

DTU Informatics
Department of Informatics and Mathematical Modeling

The community detection threshold

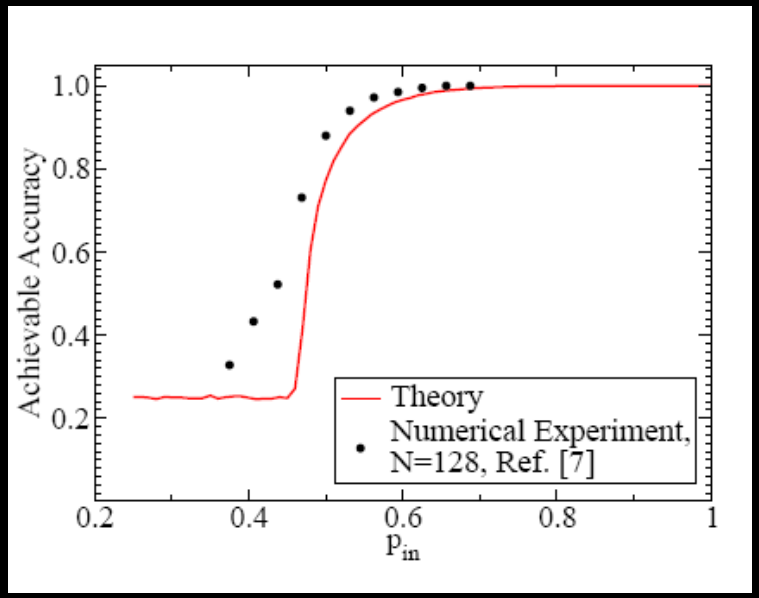
how many links are needed to detect the structure?

(Un)detectable cluster structure in sparse networks

Jörg Reichardt[1] and Michele Leone[2]

[1]*Institute for Theoretical Physics, University of Würzburg, 97074 Würzburg, Germany*
[2]*ISI Foundation, Viale S. Severo 65,I-10133 Torino, Italy*

(Dated: February 2, 2008)



DTU Informatics
Department of Informatics and Mathematical Modeling

Jorg Reichardt and Michele Leone,
Un)detectable Cluster Structure in Sparse Networks
Phys. Rev. Lett. 101, 078701 (2008),

The Hofman-Wiggins model
Mean field critical link density

$$1 = \frac{\lambda_{\max}(A)}{T(p_c, q_c)} = \frac{N}{C}(p_c - q_c)\cdot \log\left(\frac{p_c}{1-p_c}\frac{1-q_c}{q_c}\right)$$

Assume that A is indeed drawn with parameters *p, q=fp=p/SNR*

The iteration scheme converges to uniform random solution below the critical density

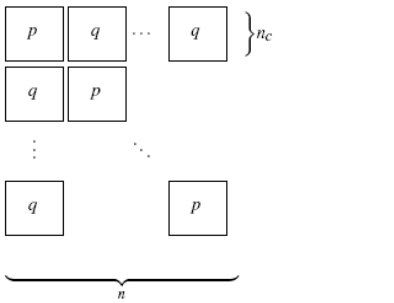
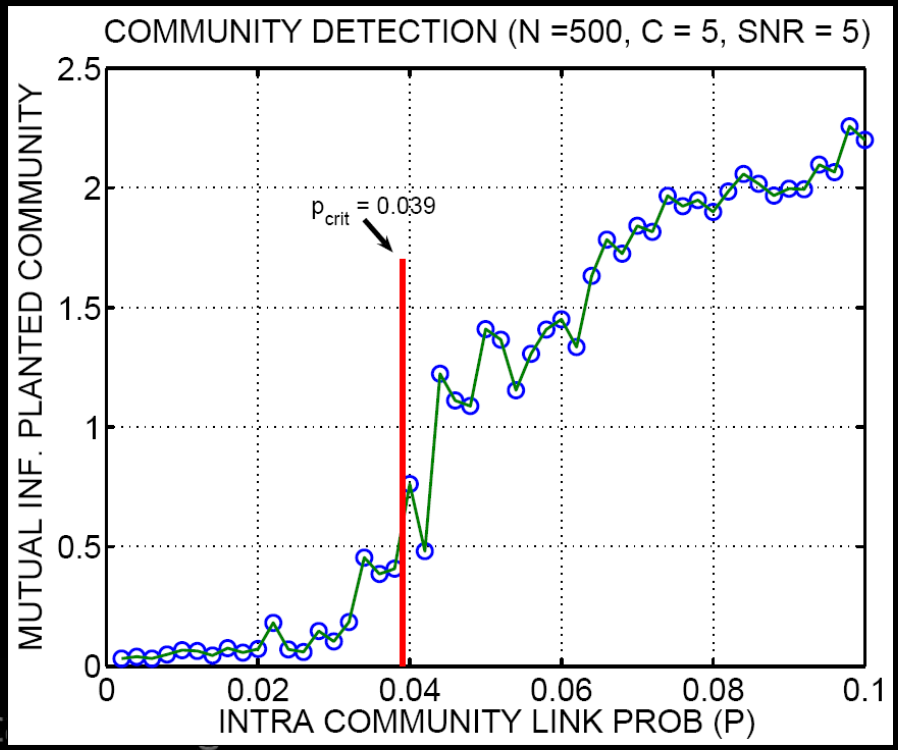


FIG. 1: A sketch of the simple network model. The figure displays the structure of the adjacency matrix with nodes arranged according to community. Inside each community (the blocks) along the diagonal, the probability of a link between two nodes is p and between communities, the probability of a link is q.



Learning the parameters of the generative model

- Hofman & Wiggins (2008)
  - “Variational Bayes”
  - Dirichlets/beta prior and posterior distributions for the probabilities
  - Independent binomials for the assignment variables
- Here
  - Maximum likelihood for the parameters
  - Gibbs sampling for the assignments

Jake M. Hofman and Chris H. Wiggins,
Bayesian Approach to Network Modularity
Phys. Rev. Lett. 100, 258701 (2008),

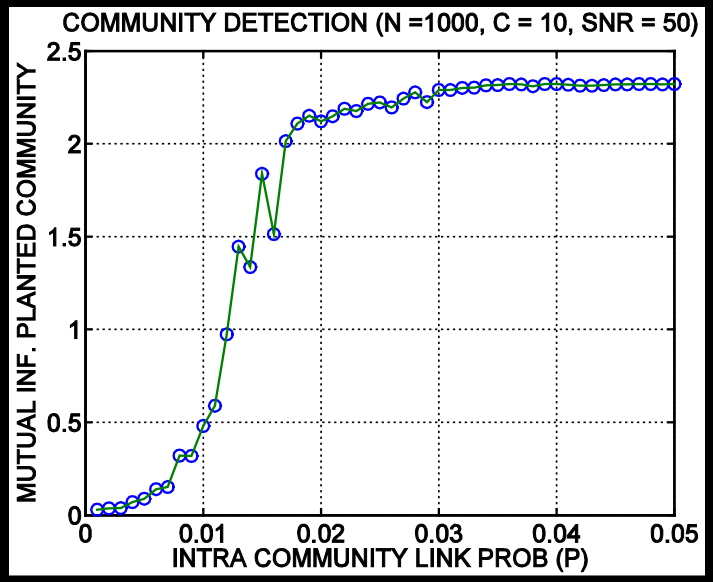
Experimental design

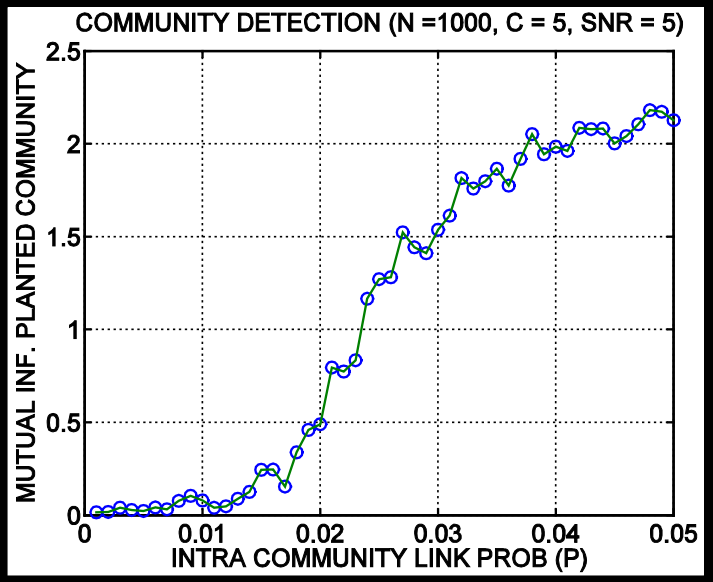
- Planted solution
  - N = 1000 nodes
  - C_{true} = 5
  - Quality: Mutual information between
    - planted assignments and the best identified

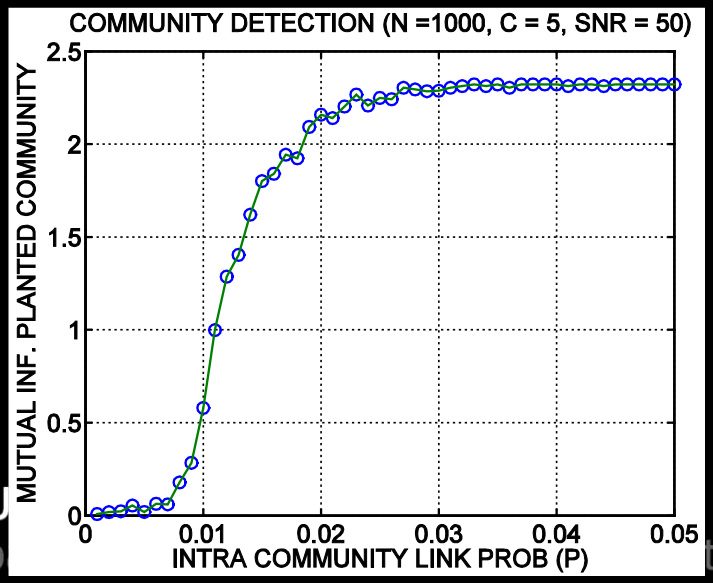
- Gibbs sampling
  - No annealing
  - Burn-in 200 iterations
  - Averaging 800 iterations

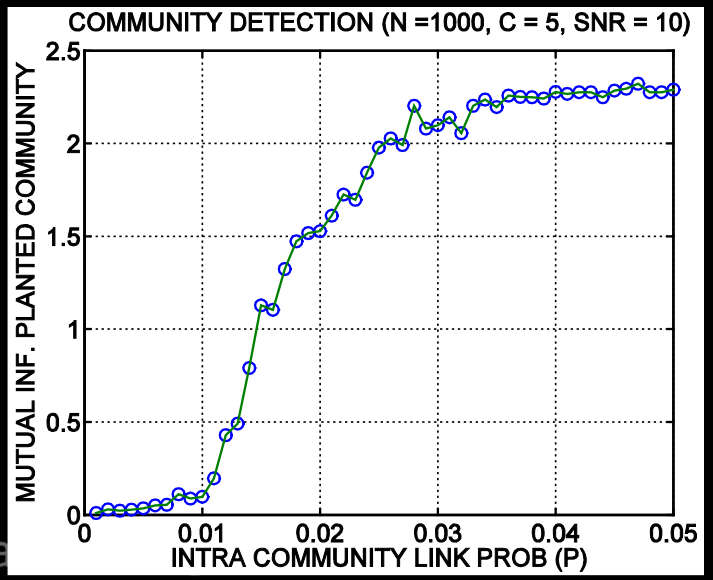
- Parameter learning
  - Q = 10 iterations

Community Detection – fully informed on number of communities and probabilities

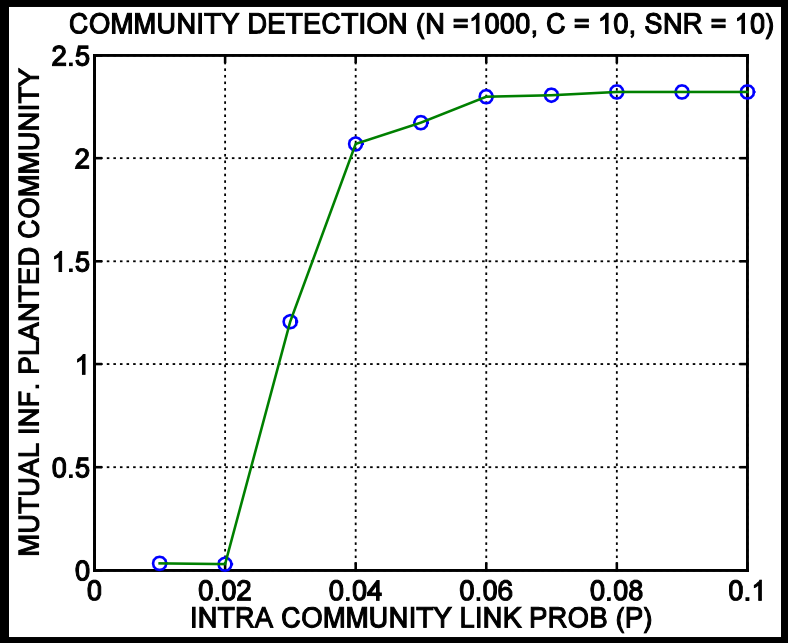


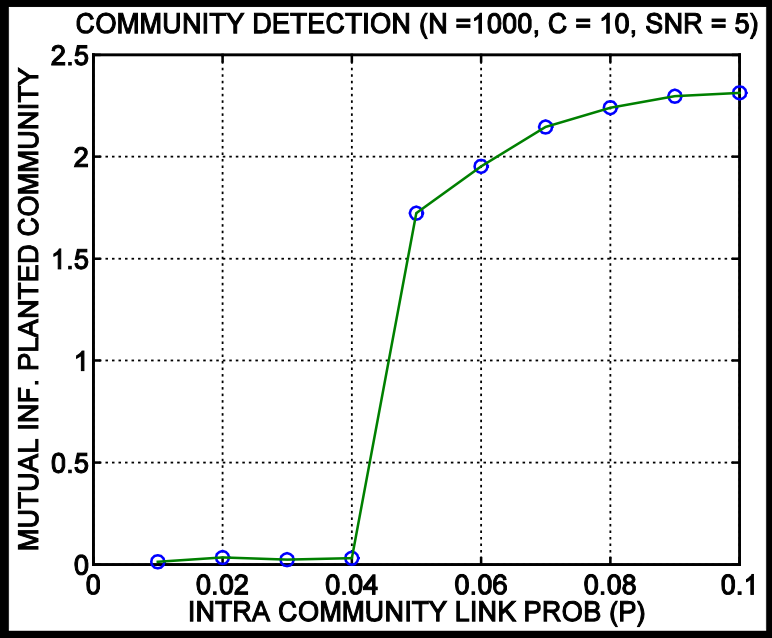


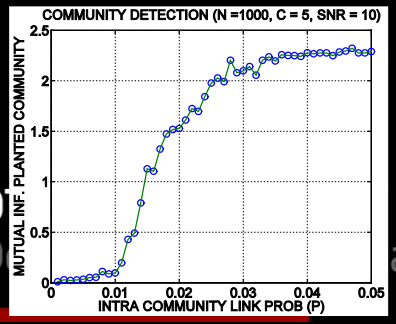


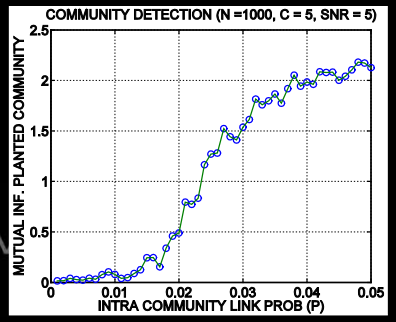


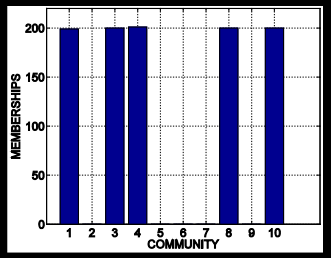
Now what happens to the phase transition if we learn the parameters … with a too complex model ($C > C_{true} = 5$) ?











More complex latent structures

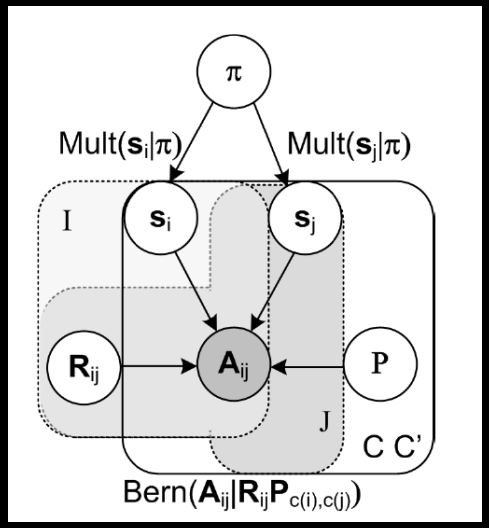
- **There is a very rich statistics literature on closely related models**
  - Review: Goldberg, Zeng, Fienberg, Airoldi (2010)

- The equivalent of the H&W model was analysed by Snijders & Nowicki (1997) using both EM and Gibbs sampling

- Stochastic block membership models parameterize the link density using a *C* x *C* matrix of parameters describing the potential different link probability between two given communities

- MMSB -the Mixed membership stochastic block model recently proposed and analyzed by Airoldi et al. (2008)

The general link density model



Stochastic block model with a (learned) node specific link probability ($R_{ij} = r_j r_i$) ala Modularity

$$A_{ji} \sim Bern(R_{ij} \otimes P_{c(i),c(j)})$$

Key research questions

- How to evaluate these representations?
  - Link prediction
- Can we speed-up the inference process to make large graphs feasible?
  - a new NMF-like relaxation to the simplex avoids annealing

DTU Informatics
Department of Informatics and Mathematical Modeling

M. Mørup, L.K. Hansen: Learning latent structure in complex networks
NIPS Workshop "Analyzing Networks and Learning With Graphs" (2009)

Link prediction

Inspired by Clauset et al. (2008) we use a cross-validation like procedure where we predict the presence of held-out links in a number of networks:

	♯ nodes	♯ links	r	c	L	ρ
H&W model	8,800	687,272	0.7027	0.0389	2.7408(0.0256)	8.9e-3
MMSB model	8,800	1,181,692	0.8612	0.0616	2.6226(0.0323)	15.3e-3
LD model	8,800	1,484,592	0.4589	0.0959	2.6228(0.0696)	19.2e-3
Yeast Network[?]	2,284	13,292	-0.0991	0.0958	4.3877(0.2443)	2.5e-3
US Power	4,941	13,188	0.0035	0.0500	19.8832(0.6681)	5.4e-4
Erdos02	5,534	16,944	-0.0399	0.0470	3.8807(0.0038)	5.5e-4
Free Assoc.	10,617	127,576	-0.0720	0.0938	3.9081(0.1317)	1.1e-3
Reuters911	13,314	296,076	-0.1090	0.2734	3.0600(0.0976)	1.7e-3
Wordnet3	31,867	240,798	-0.0911	0.0165	7.1956(0.3195)	2.4e-4
Dictionary28	39,327	178,076	0.7080	0.1597	7.8093(0.5067)	1.2e-4
CondPhys2005[7]	39,577	351,386	0.1863	0.4855	4.5589(0.5053)	2.2e-4
Internet	124,650	387,240	-0.0078	0.0382	11.4765(0.6301)	2.5e-5
IMDB	896,308	115,025,018	0.2002	-	3.5886(0.1252)	1.4e-4
Patents	3,774,768	29,941,533	0.1071	-	8.5683(0.2259)	2.1e-6

	$corr(\boldsymbol{k}, \boldsymbol{r}^{\mathrm{LD}})$
H&W model	0.713(3)
MMSB model	0.690(2)
LD model	0.762(5)
Yeast Network	0.399(10)
US Power	0.505(12)
Erdos02	0.401(21)
Free Assoc.	0.351(5)
Reuters911	0.309(9)
Wordnet3	0.140(11)
Dictionary28	0.510(8)
CondPhys2005	0.126(7)
Internet	0.171(11)
IMDB	0.0854
Patents	-

Table 1: **Left Table:** Properties of the analyzed networks. r denotes the networks assortativity [8], c denotes the clustering coefficient [12], L the average shortest path and ρ the density of the network. The average shortest path measure was calculated as the average of 10 samples of up to 10,000 links in the network disregarding non-existing paths between nodes (in parenthesis is given the standard deviation of this mean over the samples). **Right Table:** Correlation between node degree distribution and the estimated node specific parameter r of the LD model. Given are the average correlation over 10 model estimations as well as the standard deviation on the last digit. Clearly, there is a significant correlation for all the networks.

A. Clauset, C. Moore, and M.E.J. Newman. Hierarchical structure and the prediction of missing links in networks. Nature, 453:98–101 (2008).

Link prediction results (fixed community #)

	Degr. Prod.	Short. Path	Com. Neigh.	Jaccard	Hamilt.	Modularity	H&W	MMSB	LD
H&W model	0.627(5)	0.788(7)	0.827(8)	0.828(8)	0.869(2)	0.867(3)	**0.873(2)**	0.868(2)	0.872(1)
MMSB model	0.621(5)	0.786(5)	0.840(5)	0.845(5)	0.866(2)	0.856(6)	0.861(3)	**0.892(1)**	0.888(1)
LD model	0.690(2)	0.688(2)	0.834(3)	0.825(3)	0.866(2)	0.866(2)	0.870(1)	0.904(1)	**0.905(1)**
Yeast Network	0.783(6)	0.795(9)	0.675(9)	0.675(9)	0.640(13)	0.599(14)	0.573(9)	**0.836(7)**	0.794(9)
US Power	0.449(10)	**0.795(6)**	0.474(13)	0.474(13)	0.479(12)	0.489(13)	0.544(13)	0.407(8)	0.506(9)
Erdos02	0.586(9)	0.580(9)	0.584(5)	0.559(8)	0.530(15)	0.460(11)	0.377(12)	**0.954(3)**	0.873(22)
Free Assoc.	0.845(6)	0.877(5)	0.856(6)	0.854(6)	0.766(11)	0.632(8)	0.609(6)	**0.902(5)**	0.872(4)
Reuters911	0.928(3)	0.892(4)	0.929(3)	0.910(3)	0.728(8)	0.601(5)	0.766(4)	**0.942(2)**	0.935(2)
Wordnet3	0.602(6)	0.613(6)	0.356(4)	0.356(4)	0.507(8)	0.463(7)	0.481(6)	**0.795(8)**	0.658(8)
Dictionary28	0.808(5)	**0.917(4)**	0.776(4)	0.744(5)	0.644(6)	0.633(4)	0.678(6)	0.865(6)	**0.894(3)**
CondPhys2005	0.790(3)	0.963(2)	0.964(1)	**0.965(1)**	0.737(8)	0.689(5)	0.463(12)	0.909(2)	**0.924(1)**
Internet	0.604(5)	**0.745(4)**	0.501(4)	0.501(4)	0.662(4)	0.672(5)	0.607(8)	0.695(5)	0.663(5)
IMDB	0.918(1)	0.980(5)	**0.998(0)**	0.997(1)	0.864(2)	0.857(2)	0.861(2)	0.965(2)	0.974
Patents	0.766(1)	**0.946(3)**	0.743(5)	0.743(5)	0.770(1)	0.768(1)	0.696(14)	0.889(1)	-

Table 2: Link prediction performance of the various methods on the 14 networks. Given are the mean AUC values as well as their standard deviations on the last digit given in parenthesis across 10 data splits with randomly chosen links and non-links treated as missing. In bold black is given the best performing approach and in underline the best performing community detection approach. The Hamiltonian was based on average link density as the imposed null hypothesis, i.e. $B = \frac{m}{n^2}E$ all clusters were modelled with $c = 50$, as such a more dramatic difference in link performance was observed for the H&W, MMSB and LD models when initialized with the true number of clusters, however, in general the true number of clusters is unknown hence these results are not shown.

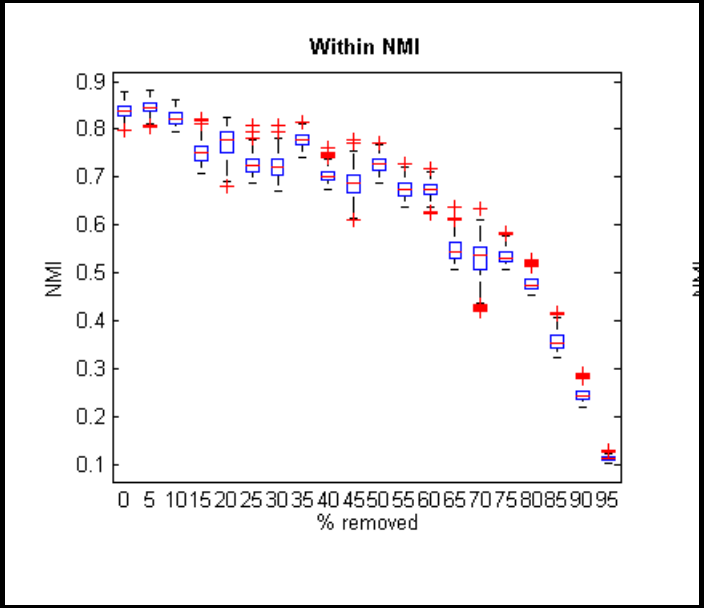
M. Mørup, L.K. Hansen: Learning latent structure in complex networks
NIPS Workshop "Analyzing Networks and Learning With Graphs" (2009)

Potential critique of link prediction

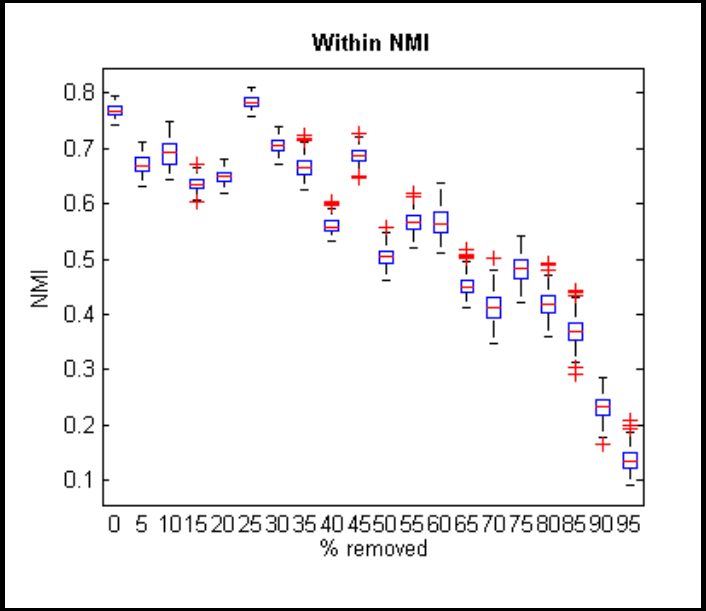
i) cross validation – is the structure robust to dilution?

ii) can we relax the fixed cap on number of communities?

Free word association



Yeast



Large communities seem very robust to link dilution. These runs use non-parametric Bayes Dirichlet process priors … the number of communities is on the order of C = 50 as in earlier results, drops to about 10-20 when community structure deteriorates

Conclusions

- Community detection can be formulated as an inference problem
- The sampling process for fixed SNR has a phase transition like detection threshold - we can estimate the threshold from MF analysis
- The phase transition remains (sharpens?) if we learn the parameters of a generative model with unknown complexity
- For link prediction more complex latent structures are necessary: Modularity and H&W do not beat simple non-parametric models

Acknowledgements

Morten Mørup
www.imm.dtu.dk/~mm

Sune Lehmann
sune.barabasilab.com

Funding

Danish Research Councils
The Lundbeck Foundation

DTU Informatics
Department of Informatics and Mathematical Modeling

DTU Informatics

Department of Informatics and Mathematical Modeling