Machine Learning in Simple Networks

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

DTU Informatics
Department of Informatics and Mathematical Modeling

Outline

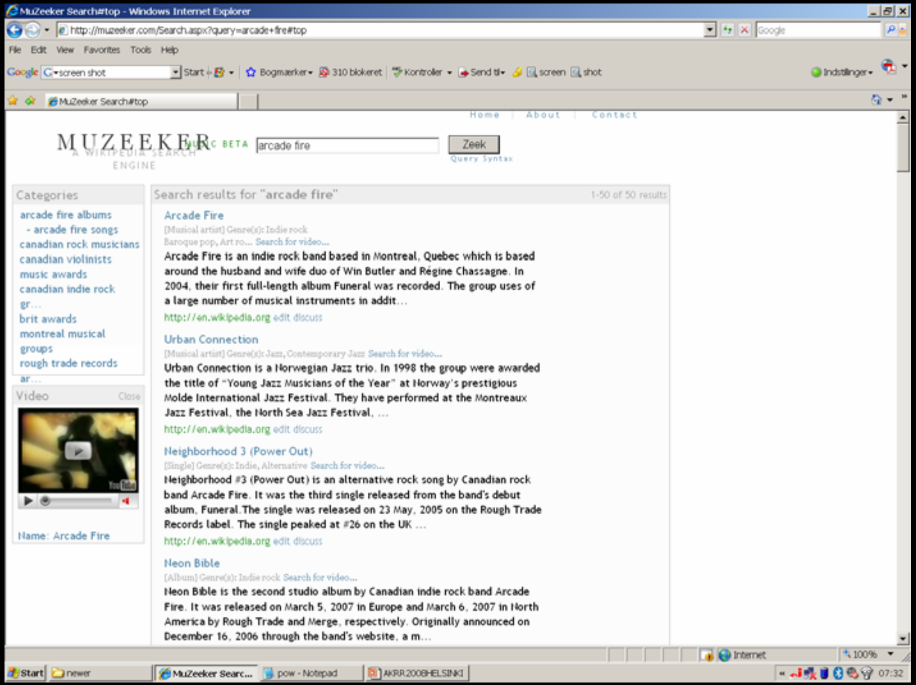
- Communities and link prediction
- Modularity
  - Modularity as a combinatorial optimization problem
  - Gibbs sampling
- Detection threshold – a phase transition?
- Learning community parameters
  - The Hofman-Wiggins generative model
  - Is there a threshold for detection when you learn the parameters and complexity

Muzeeker

- Wikipedia based common sense
- Wikipedia used as a proxy for the music users mental model
- Implementation: Filter retrieval using Wikipedia's article/ categories

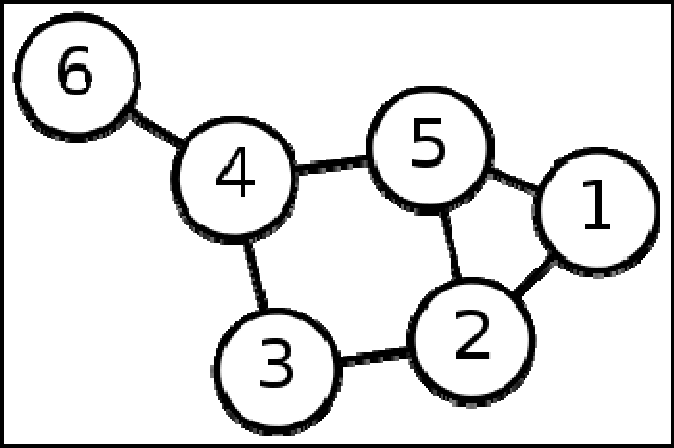
- Muzeeker.com

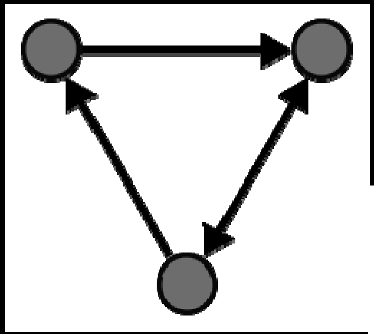
- LINK PREDICTION to complete the ontological quality of Wikipedia

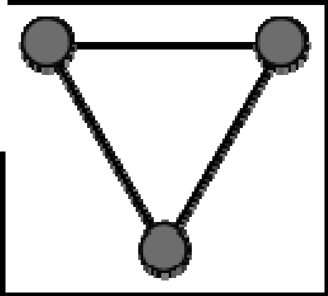


Network models

- Nodes/vertices and links/edges
  - Directed / undirected
  - Weighted / un-weighted
- Link distributions
  - Random
  - Long tail
  - Hubs and authorities
- Link induced correlations
  - The Rich club
- Communities
  - Link prediction

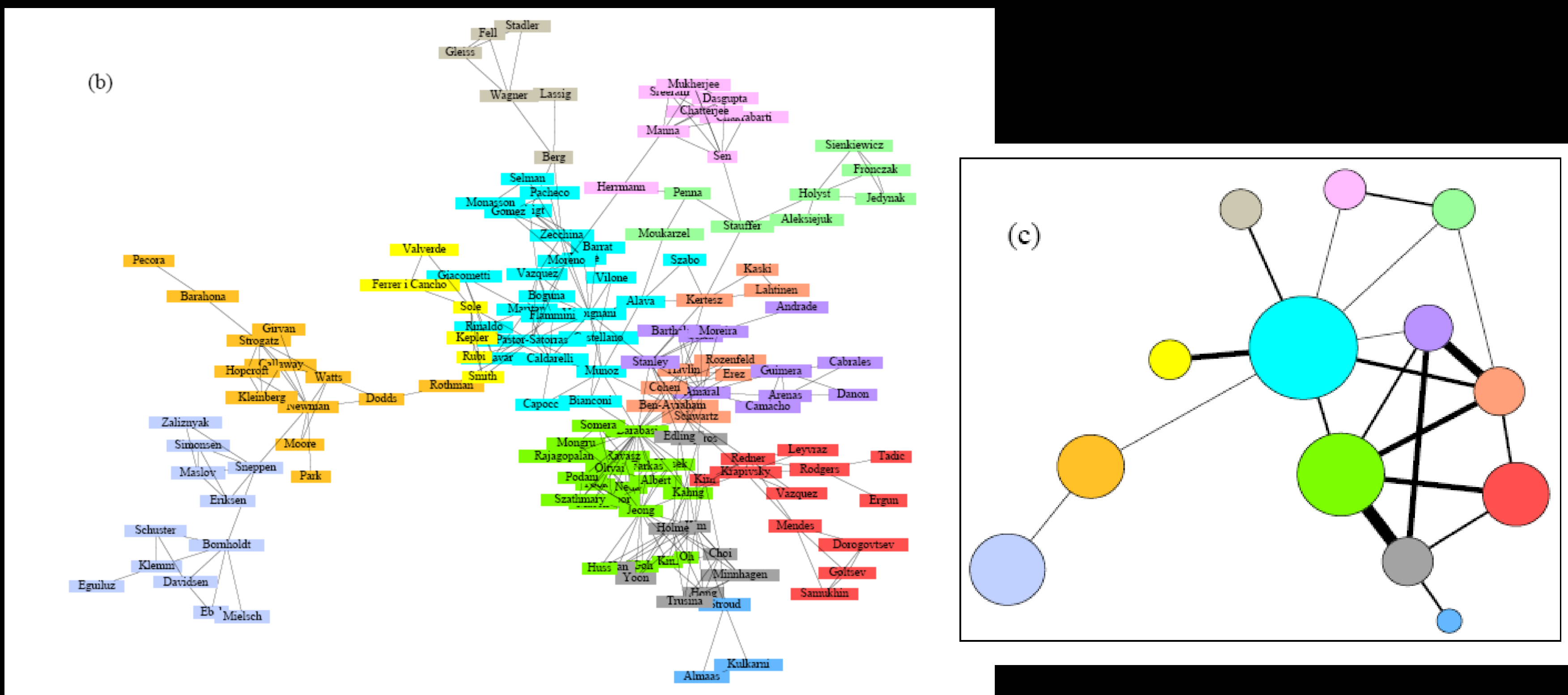






Motivation for community detection

- Community structure may mark a non-stationary link distribution with “high and low density” sub-networks, hence summarizing with a single “model” could be misleading



Modularity can be predictive for dynamics

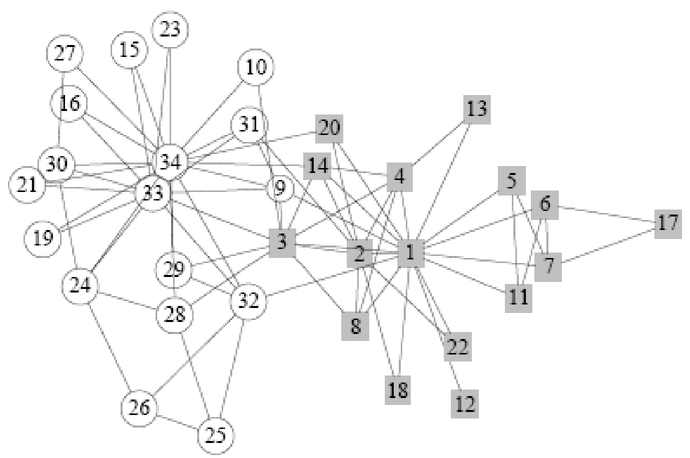


FIG. 8: The network of friendships between individuals in the karate club study of Zachary [35]. The administrator and the instructor are represented by nodes 1 and 33 respectively. Shaded squares represent individuals to who ended up aligning with the club's administrator after the fission of the club, open circles those who aligned with the instructor.

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

Modularity objective function

The modularity is expressed as a sum over links, such that we penalize missing links in communities - missing is measured relative to a null distribution P^0_{ij}.

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

C_i is the community assignment of node j
and $2m = \Sigma_{ij} A_{ij}$, $k_i = \Sigma_j A_{ij}$

The null is a baseline distribution $P_{ij} = k_i k_j / (2m)^2$

The value of the modularity lies in the range [−1,1]. It is positive if the number of edges within groups exceeds the number expected on the basis of chance

DTU Informatics
Department of Informatics and Mathematical Modeling

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

Potts representation

Introduce 0,1 binary variables S_{kj} coding the community assignment: “node j is member of community k”

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

$$P(j,i) = \frac{A_{ij}}{2m}$$

$$Q = \sum_{ij} \left[\frac{A_{ij}}{2m} - P_i P_j \right] \delta(c_i, c_j) = \sum_{ij} \left[\frac{A_{ij}}{2m} - P_i P_j \right] \sum_k S_{ki} S_{kj}$$

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

Spectral optimization

- Newman relaxes the optimization problem to the simplex

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

$$L = \frac{Tr(SBS')}{2m} + Tr(\tilde{\Lambda} S)$$

$$BS = S\Lambda$$

Combinatorial optimization

- We can use a physics analogy Simulated Annealing (Kirkpatrick et al. 1983)

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

- Gibbs sampling is a 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)

Potts model 1-node

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

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

$$P(S \mid A,T) = \prod_k (r_k)^{S_k}$$

$$\langle S_k \rangle = r_k = \frac{\exp\left(\frac{\varphi_k}{T}\right)}{\sum_{k'} \exp\left(\frac{\varphi_{k'}}{T}\right)}$$

Gibbs sampling

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

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

$$S_i = \mathrm{potts}(r_i)$$

Deterministic annealing

- Instead of drawing Gibbs samples according to the marginals we can **average** instead, this provides a set of self-consistent equations for the means (for 0,1 Bernoulli variables the mean is the probability μ_{ki} =P(S_{ki}))

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

$$\varphi_{ki} = \sum_j \frac{B_{ij}}{2m} r_{kj} = \sum_j \frac{A_{ij}}{2m} r_{kj} - \sum_j P_i P_j r_{kj}$$

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

Experimental evaluation

- Create a simple testbed with link probability and “noise”

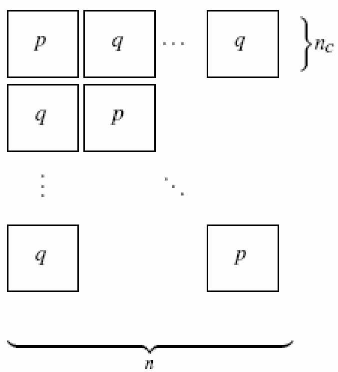


FIG. 1: A sketch of the simple network model. The figure displays the structure of the adjacency matrix with nodes arranged according to community. Inside each community (the blocks) along the diagonal, the probability of a link between two nodes is p and between communities, the probability of a link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

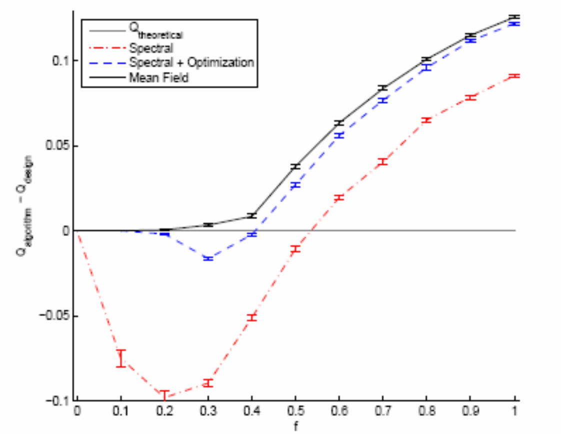


FIG. 3: Comparing spectral methods with the mean field solution. The networks were created according to the simple model, using parameters $n_c = 100$, $C = 5$, $p = 0.1$ and $f \in [0,1]$. All data points display the point-wise differences between the value of $Q_{\text{algorithm}}$ found by the algorithm 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.

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

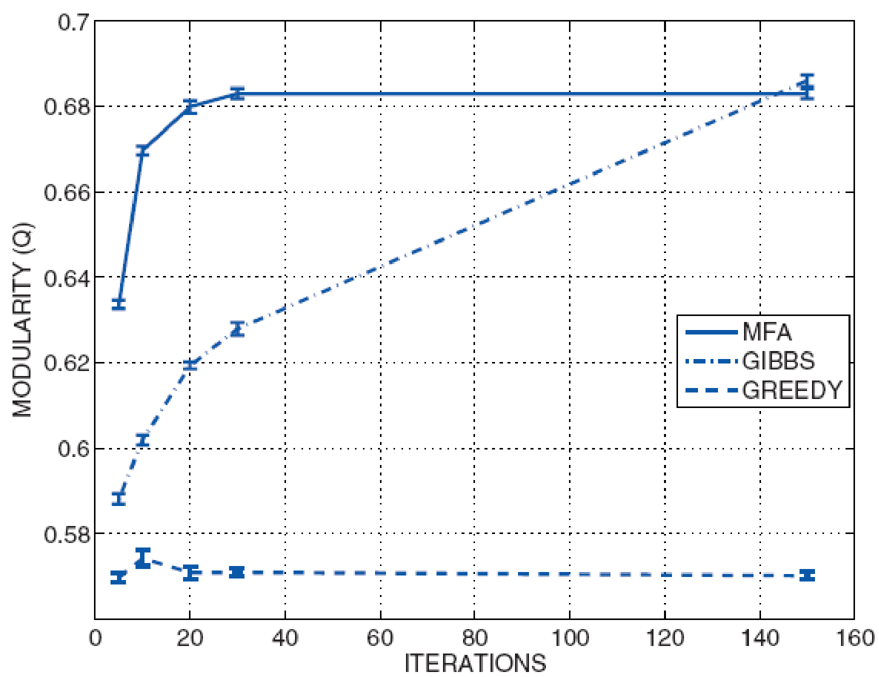


Fig. 5. Comparison of the modularity performance for mean field annealing, Gibbs sampling, and a 'greedy' optimizer. The greedy algorithm corresponds to Gibbs sampling at $T = 0$. The graph describes the co-authorship network of the Los Alamos condensed matter preprint archive, considering articles published 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).

Generative community model (Hofman & Wiggins, 2008)

DTU

A Bayesian Approach to Network Modularity

Jake M. Hofman*
Department of Physics, Columbia University, New York, NY 10027

Chris H. Wiggins†
Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027
(Dated: June 23, 2008)

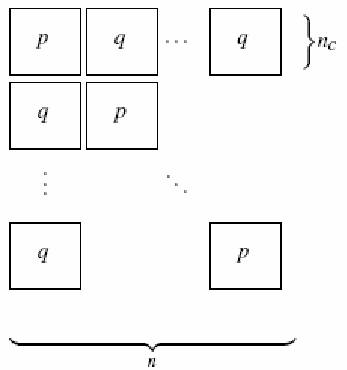


FIG. 1: A sketch of the simple network model. The figure displays the structure of the adjacency matrix with nodes arranged according to community. Inside each community (the blocks) along the diagonal, the probability of a link between two nodes is p and between communities, the probability of a link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

$$P(A|S,p,q) = p^{c}(1-p)^{d}q^{e}(1-q)^{f}$$

$$c = \frac{1}{2}\sum_{j\neq i,k} A_{ij}S_{kj}S_{ki}$$

$$d = \frac{1}{2}\sum_{j\neq i,k} (1-A_{ij})S_{kj}S_{ki}$$

$$e = \frac{1}{2}\sum_{j\neq i} A_{ij}\left(1-\sum_{k} S_{kj}S_{ki}\right)$$

$$f = \frac{1}{2}\sum_{j\neq i} (1-A_{ij})\left(1-\sum_{k} S_{kj}S_{ki}\right)$$

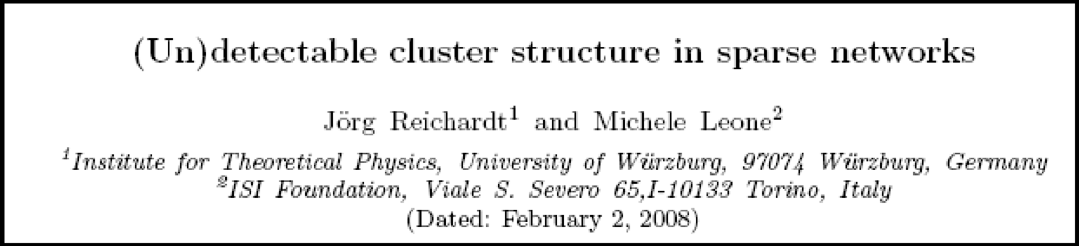
Learning parameters of the generative model

- Hofman & Wiggins (2008)
  - “Variational Bayes”
  - Dirichlets/beta prior and posterior distributions for the probabilities
  - Very well determined (over kill)
  - Independent binomials for the assignment variables (misses correlation)
- 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),

The community detection threshold

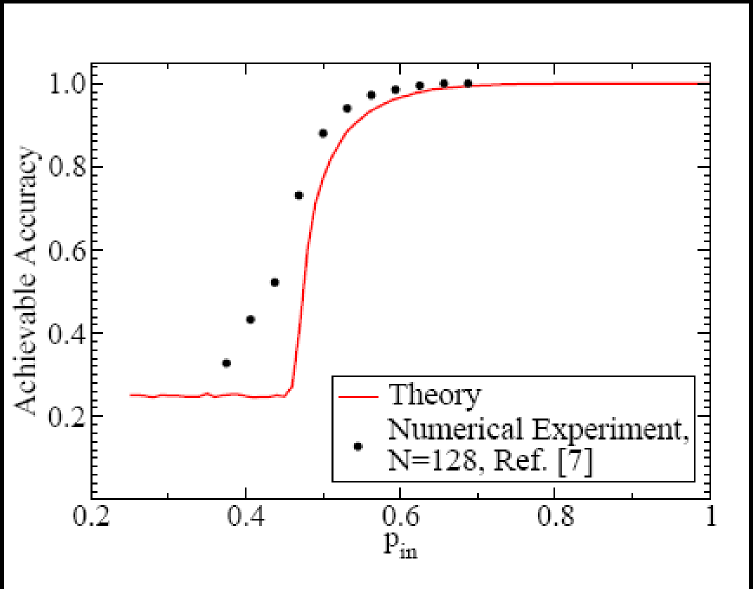
how many links are needed to detect the structure?


(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)

$$P_{in} = \frac{p}{q(C-1)} = \frac{SNR}{C-1}$$

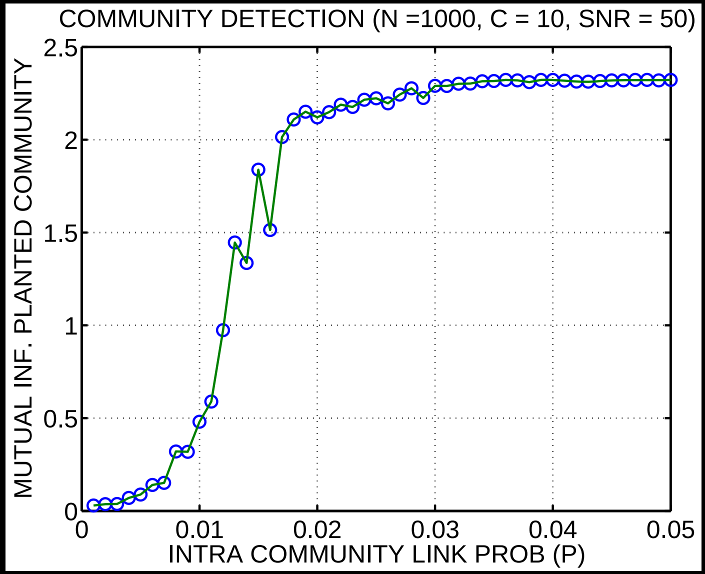


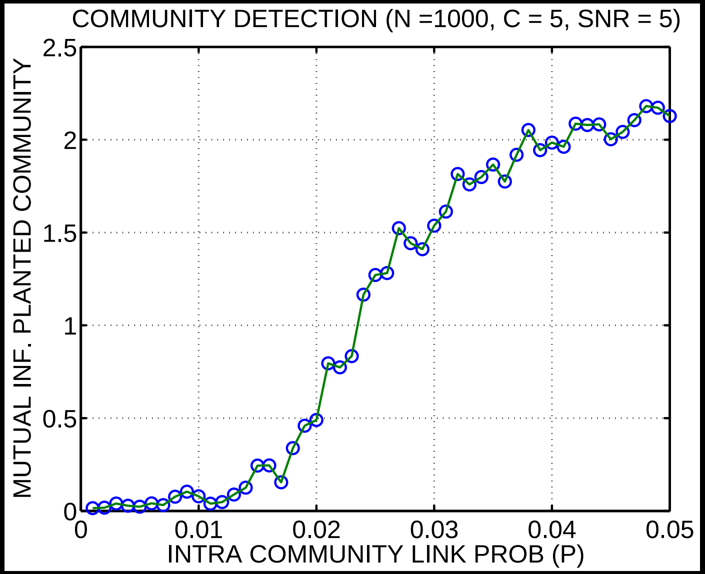
Jorg Reichardt and Michele Leone,
Un)detectable Cluster Structure in Sparse Networks
Phys. Rev. Lett. 101, 078701 (2008),

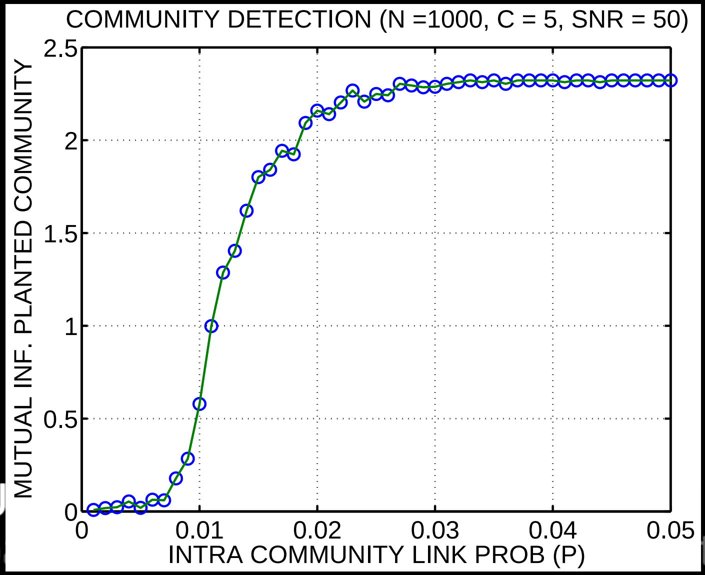
Experimental design

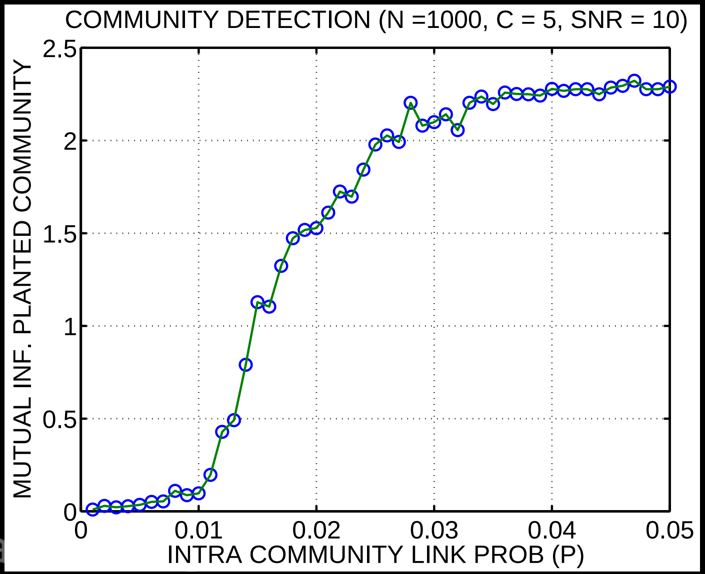
- Planted solution
  - N = 1000 nodes
  - $C_{true} = 5$
  - Quality: Mutual information between
    - planted assignments and the best identified
- Gibbs sampling
  - No annealing
  - Burn-in 200 iterations
  - Averaging 800 iterations
- Parameter learning
  - Q = 10 iterations

Community Detection – fully informed on number of communities and probabilities

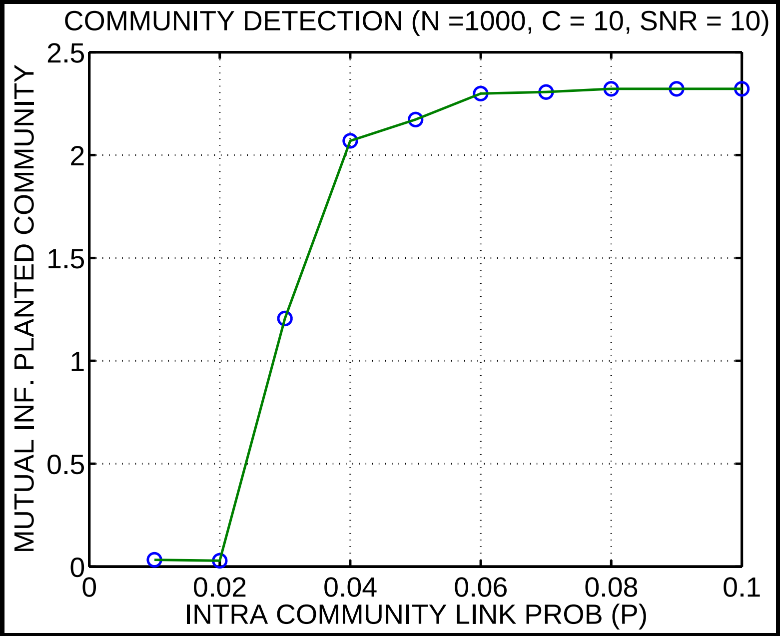


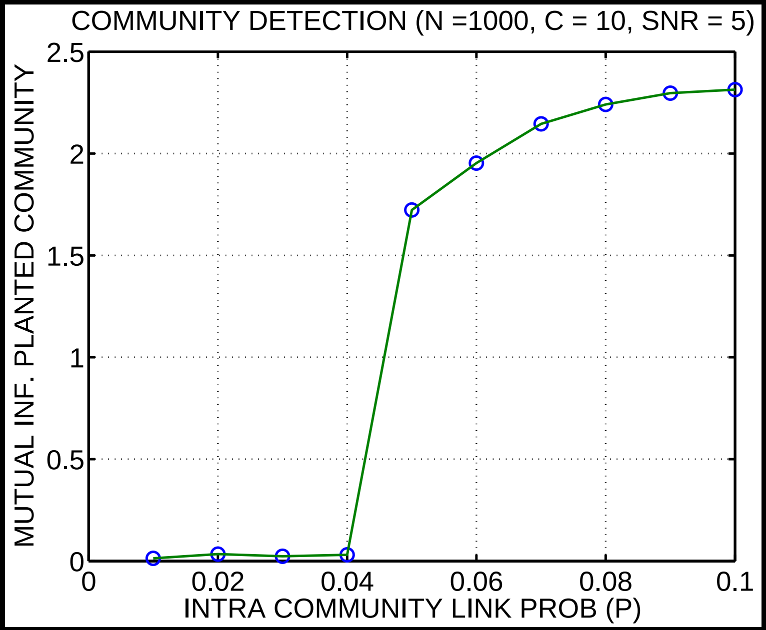


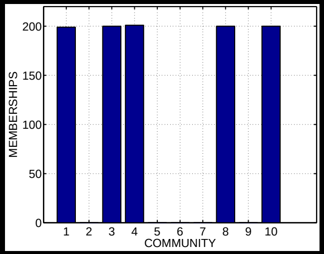




Now what happens to the phase transition if we learn the parameters … with a too complex model ($C > C_{true} = 5$) ?

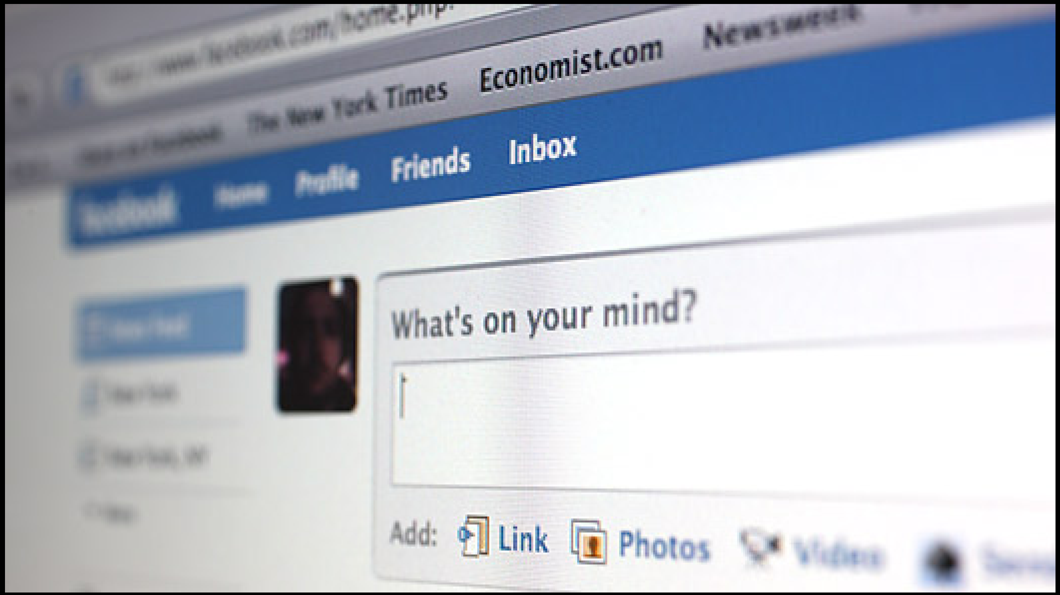






Conclusions

- Community detection can be formulated as an inference problem (Hofman & Wiggins, 2008)
- The sampling process for fixed SNR has a phase transition like detection threshold (Richard & Leone, 2008)
- The phase transition remains (sharpens?) if you learn the parameters of a generative model with unknown complexity



DTU Informatics
Department of Informatics and Mathematical Modeling