Variational Inference of Community Models using Belief Propagation

A Case Study: Model Selection for Stochastic Block Models

Xiaoran Yan

June 4, 2013

Joint work with Cris Moore, Cosma Shalizi, Aurelien Decelle, Lenka Zdeborová, Florent Krzakala, Pan Zhang, Yaojia Zhu

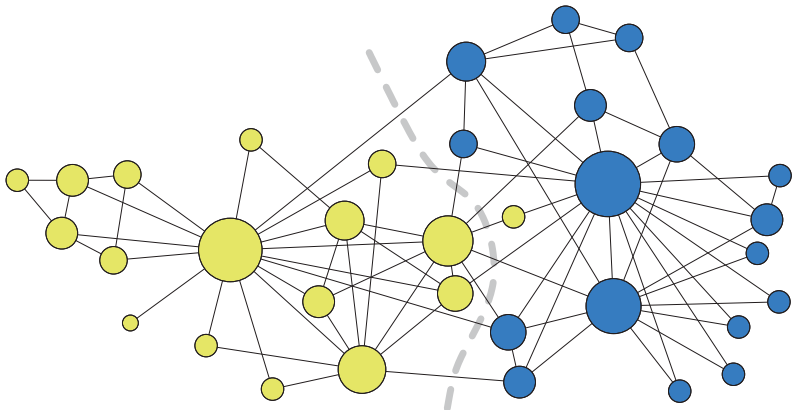
Community structures in networks

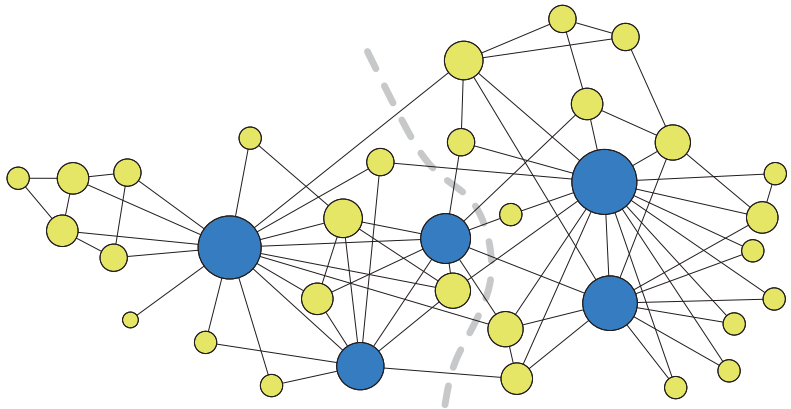
Assortative communities

- Dense connection within groups
- Sparse connection between groups
- Min cut, spectral partitioning, modularity, etc.

Functional communities

- Structure equivalence
- Disassortative communities
- Mixed structures, satellite structures
- Food webs, leaders and followers





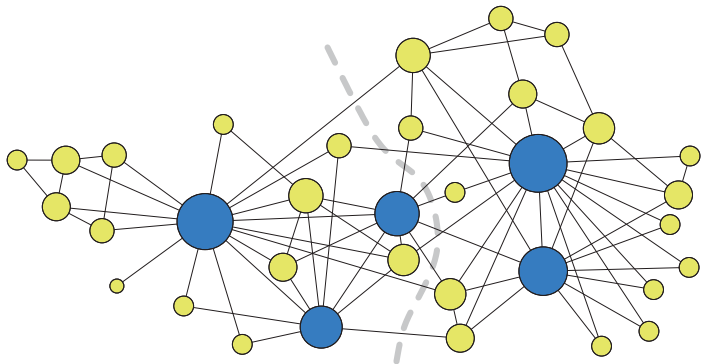
Stochastic block model

Assumptions:

- We represent our network as an undirected simple graph $G = (V, E)$ with the adjacency matrix A.
- Each node $u \in V$ has a hidden block label $g(u) \in \{1, \ldots, k\}$.
- Each node's block label $g(u)$ is first generated according to $q_{g(u)}$. Let n_s be the number of nodes of type s, with $n = \sum_s n_s$.
- Between each pair of nodes $\{u, v\}$, an edge is generated independently with probability $p_{g(u)g(v)}$. Let m_{st} be the number of edges from block s to block t, with $\sum_{st} m_{st} = m$.

Given the parameters p_{st} and a block assignment, i.e., a function $g : V \to \{1, \ldots, k\}$ assigning a label to each node, the likelihood of generating a given graph G in this model is:

$$
\begin{aligned}
& P(G, g|q, p) \\
&= \prod_u q_{g(u)} \prod_{u<v} (p_{g(u)g(v)})^{A_{uv}} (1 - p_{g(u)g(v)})^{1-A_{uv}} \\
&= \prod_{s=1}^{k} q_s^{n_s} \prod_{s,t=1}^{k} (p_{st})^{m_{st}} (1 - p_{st})^{n_s n_t - m_{st}} .
\end{aligned}
$$



The total likelihood

Summing over latent block assignment

- Overcoming discreteness
- A natural requirement in Bayesian approaches
- Avoid over-fitting by averaging over of latent states of different fit
- k^n number of terms in the sum

$$P(G \mid q, p) = \sum_g P(G, g \mid q, p)\,,$$

In statistical physics terms

- Assuming temperature $T = 1/\beta = 1$
- State energy $E(g) = -\log P(G, g \mid q, p)$ —likelihood
- Ground state $\arg\min_g E(g)$ —Maximum likelihood state
- The Boltzmann distribution $P^*(g) = \dfrac{e^{-E(g)}}{\sum_g e^{-E(g)}}$
- The free energy $\sum_g e^{-E(g)} = -\log P(G \mid q, p)$ —the total likelihood
- Calorimetry Tricks for estimating the free energy
  - simulated annealing, population annealing, parallel tempering

Variational approximation for the total likelihood

The variational trick

$$\begin{aligned}
\log P(G \mid q, p) &= \log \sum_g Q(g) \frac{P(G, g \mid q, p)}{Q(g)} \\
&\geq \sum_g \log \left[Q(g) \frac{P(G, g \mid q, p)}{Q(g)} \right] \\
&= \mathbb{E}_{Q(g)} \left[\log \frac{P(G, g \mid q, p)}{Q(g)} \right] \\
&= \mathbb{E}_{Q(g)} \left[\log P(G, g \mid q, p) \right] + \mathbb{S}[Q(g)] \\
&= - \langle \mathbb{E} \rangle + \mathbb{S}
\end{aligned}$$

The variational distribution

- If $Q(g) = P^*(g)$ the approximation is exact
- Use simpler forms of $Q(g)$ the approximation becomes a optimization

$$\log P(G \mid q, p) \approx \sup_Q \left[\mathbb{E}_{Q(g)} [\log P(G, g \mid q, p)] + \mathbb{S}[Q(g)] \right]$$

The choice of $Q(g)$

$$P(G, g|q, p) = \prod_u q_{g(u)} \prod_{u<v} (p_{g(u)g(v)})^{A_{uv}} (1 - p_{g(u)g(v)})^{1-A_{uv}} .$$

$$\log P(G \mid q, p) \geq \sum_g \log \left[Q(g) \frac{P(G, g \mid q, p)}{Q(g)} \right]$$

Wish list

- Be able to factor $Q(g)$ into local terms
- Each individual local factor can be efficiently solved
- Can be optimized to achieve good approximation

The mean field free energy

$$Q(g) = \prod_u b^u_{g_u}$$

- Easy to solve
- Total independence
- Poor approximation for almost any graphs

The Bethe free energy

Estimating the average energy

$$
\begin{aligned}
-\langle \mathbb{E} \rangle =& \mathbb{E}_{Q(g)}[\log P(G, g \mid q, p)] \\
=& \mathbb{E}_{Q(g)}[\sum_u \log q_{g(u)} + \sum_{(u,v)\in E} \log p_{g(u)g(v)} + \sum_{(u,v)\notin E} \log(1 - p_{g(u)g(v)})] \\
=& \sum_u \sum_s b_s^u \log q_s + \sum_{(u,v)\in E} \sum_{st} b_{st}^{uv} \log p_{st} + \sum_{(u,v)\notin E} \sum_{st} b_{st}^{uv} \log(1 - p_{st})
\end{aligned}
$$

The cluster variational distribution

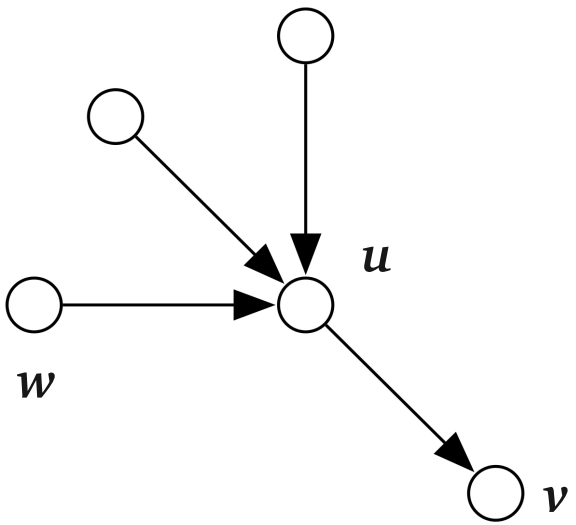
$$
Q(g) = \frac{\prod_{u<v} b_{g_u g_v}^{uv}}{\prod_u (b_{g_u}^u)^{d_u - 1}}, \quad \textit{with } b_{g_u}^u = \sum_{t,\forall v} b_{g_u t}^{uv}
$$

- Exact for trees
- Empirical results show that it works pretty well for loopy graphs
- Conditional independence
- Corresponds to the 2nd order Kikuchi free energy
- Belief Propagation leads to the same fixed points (Yedidia, 2001)

Belief Propagation

The message passing algorithm

- Equivalent to the cavity method in statistical physics
- Each node u send a message to each neighbor v about u's 1-point marginal
- The message $b_s^{u\to v}$ is based on all the other neighbors of u, as if v is absent
- Pass the messages around until convergence



Updating the messages

$$b_s^{u\to v} = \frac{1}{Z^{u\to v}} q_s \prod_{w\neq u,v} \sum_{t=1}^{k} b_t^{w\to u} (p_{st})^{A_{uv}} (1-p_{st})^{1-A_{uv}}$$

- Since we take non-edges into account, there are $O(n^2)$ messages to update
- In sparse networks, the messages along non-edges is of order $O(1/n)$, if we apply a mean field approximation for $\forall v, (u,v) \notin E$

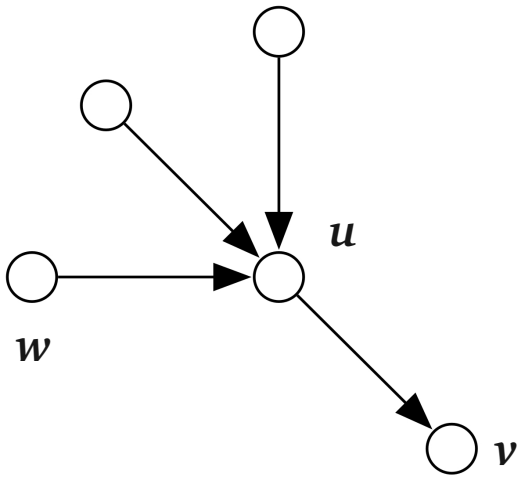
$$b_s^{u\to v} = b_s^u = \frac{1}{Z^u} q_s \prod_{w\neq u} \sum_{t=1}^{k} b_t^{w\to u} (p_{st})^{A_{uv}} (1-p_{st})^{1-A_{uv}}$$

Belief Propagation (continued)

The messages with a mean field on non-edges

$$b_s^{u\to v} = \frac{1}{Z^{u\to v}} q_s \prod_{w\neq u,v}^{(w,u)\in E} \sum_{t=1}^{k} b_t^{w\to u} p_{st} \prod_{w\neq u,v}^{(w,u)\notin E} \sum_{t=1}^{k} b_s^u (1-p_{st}) \text{ if } (u,v)\in E$$

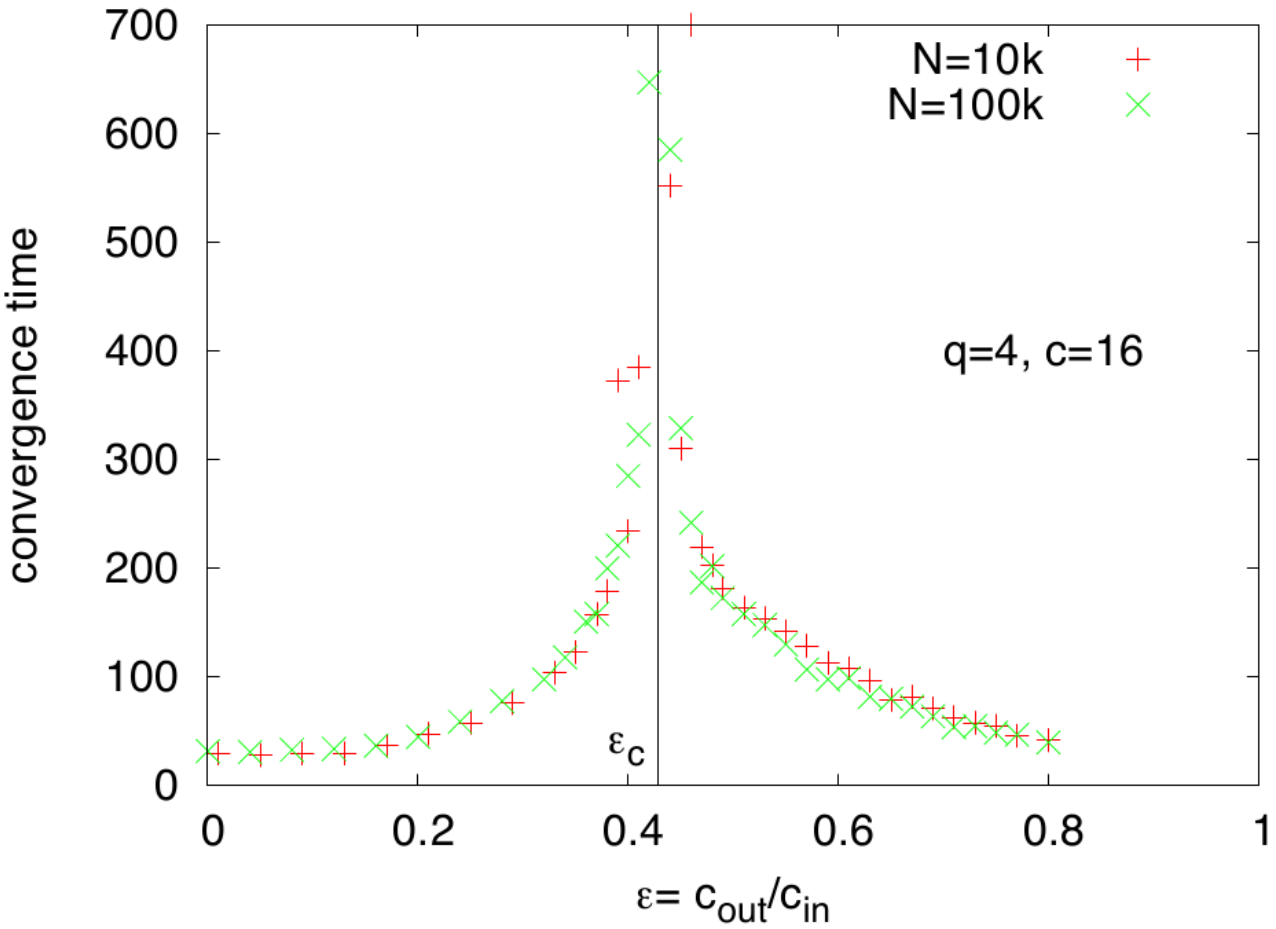
$$b_s^u = \frac{1}{Z^u} q_s \prod_{w\neq u}^{(w,u)\in E} \sum_{t=1}^{k} b_t^{w\to u} p_{st} \prod_{w\neq u}^{(w,u)\notin E} \sum_{t=1}^{k} b_s^u (1-p_{st})$$



Running time analysis

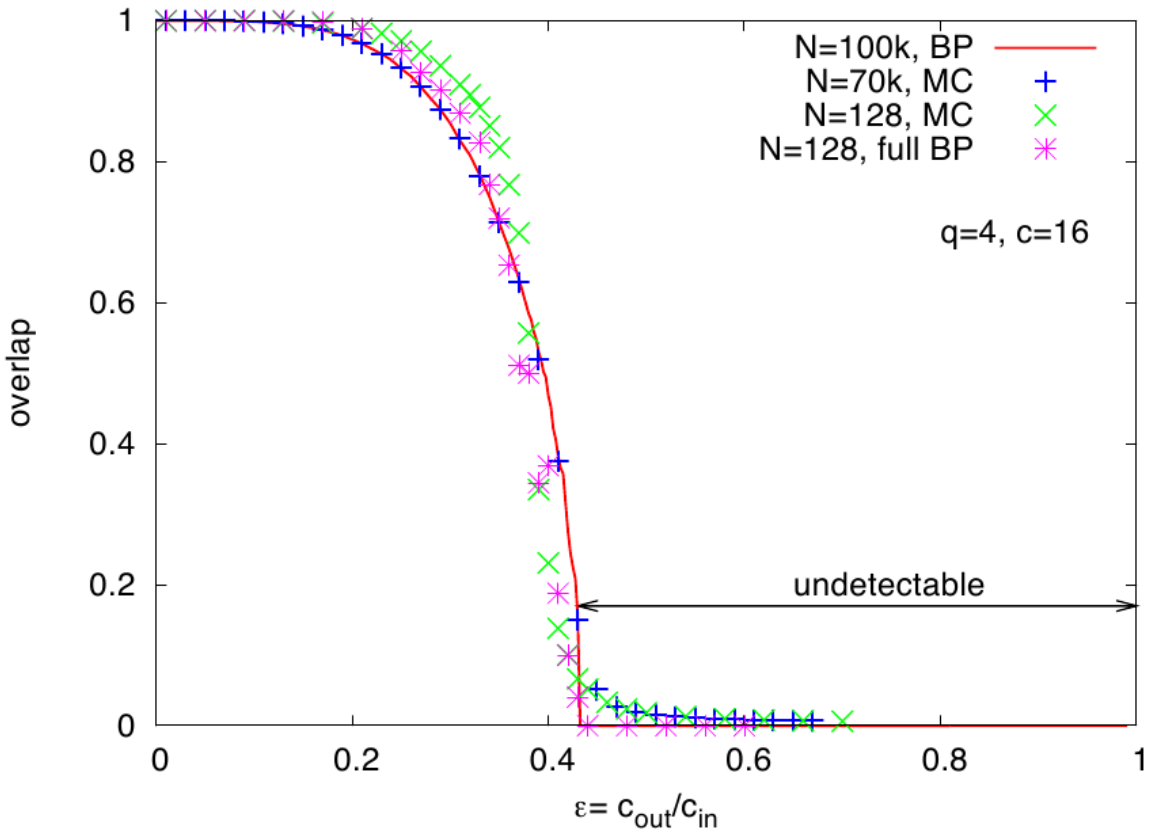
- $O(m+n)$ messages to update each sweep
- The messages converge within constant sweeps
- Linear time E-step
- The EM outer loop converges even faster
- Constant number of initial restarts
- $O(m+n)$ linear total time

Convergence time: finite correlation length



[Decelle, Krzakala, Moore, Zdeborová, PRL 2011]

Optimal detectability of Belief Propagation



Detectability

- BP fails when the p_{st} are too similar
- MCMC with extended running time also fails
- The same bound from spectral method in dense graphs
- Better bound than spectral method in sparse graphs
- Theoretical proof that BP is optimal in trees (Mossel, Neeman and Sly, 2001)
- No algorithm can do better!

The variational expectation maximization framework

Variational E-step

$$\log P(G \mid q, p) \approx \sup_{Q} \left[\mathbb{E}_{Q(g)}[\log P(G, g \mid q, p)] + \mathbb{S}[Q(g)]\right]$$

- Choose your favorite $Q(g)$
- Optimize $Q(g)$ while fixing the parameters q, p
- For SBMs: Bethe approximation with linear Belief Propagation

M-step

$$\widehat{q}_s = \frac{\bar{n}_s}{n} = \frac{\sum_u b_s^u}{n}, \quad \widehat{\omega}_{st} = \frac{\overline{m}_{st}}{n_s n_t} = \frac{\sum_{u \neq v} A_{uv} b_{st}^{uv}}{(\sum_u b_s^u)(\sum_u b_t^u)},$$

- Solving for the MLEs of the parameters q, p while fixing $Q(g)$
- Go back to E-step, rinse and repeat until a fixed point is reached
- Gradient ascent in the joint space, maximizing the total likelihood
- Linear approximation for an exponential size problem

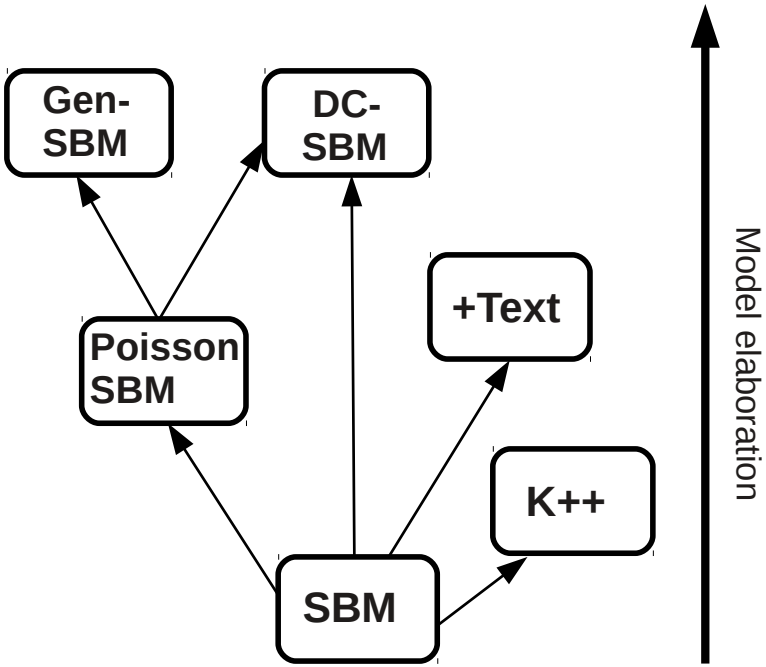
A case study: building the right stochastic block model

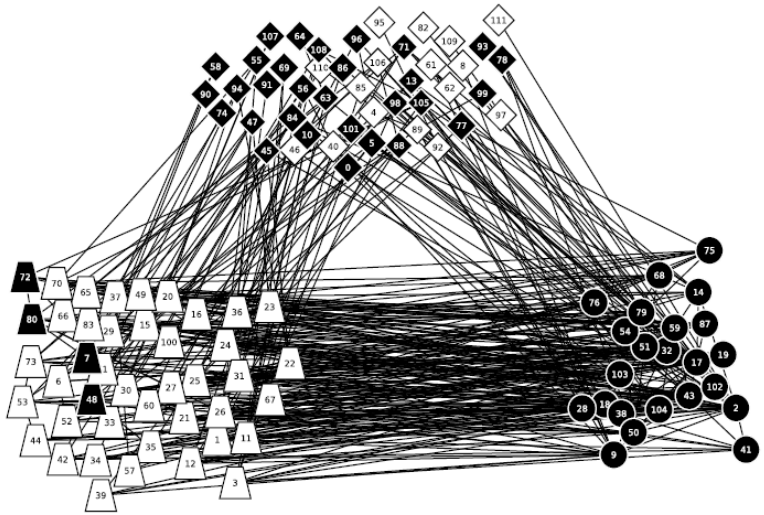
Variants and elaborations

- Degree corrected SBM
- Extensions for rich data
- More blocks!

Model selection

- Which model to choose given the data?
- Number of blocks (order selection)?
- Over-fitting?

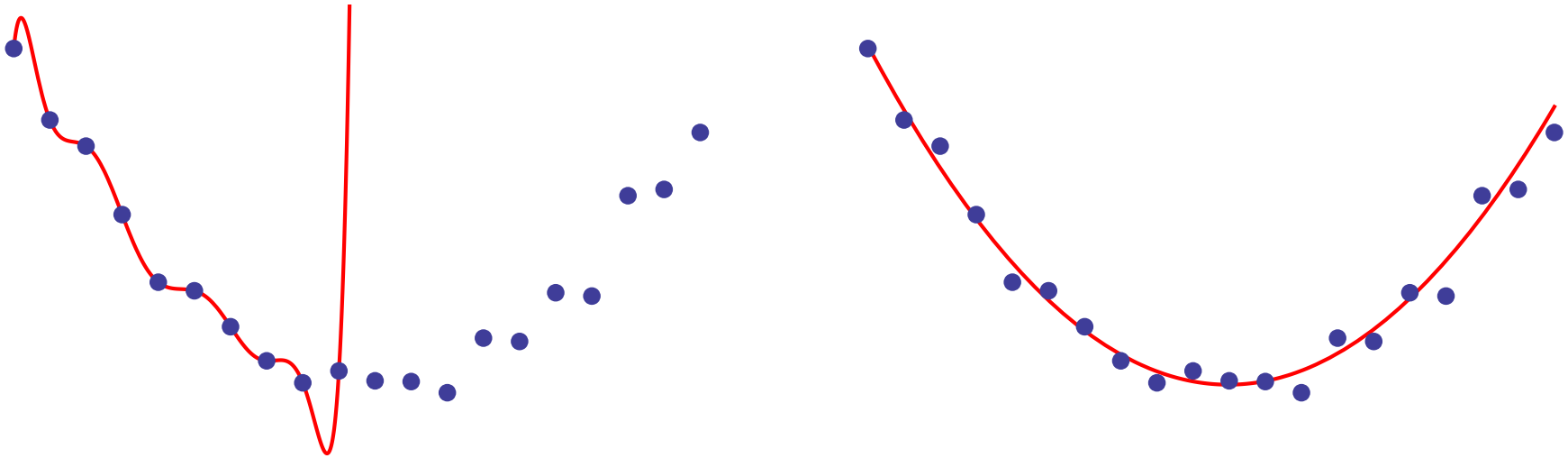




Model selection

Occam's razor

- Complex models with more parameters have a natural advantage at fitting data.
- Simpler models have lower variability, thus less sensitive to noise in the data.
- Balance the trade-off between bias and variance.
- Excessive complexity not only increases the cost of the model, but also hurts the generalization performance.



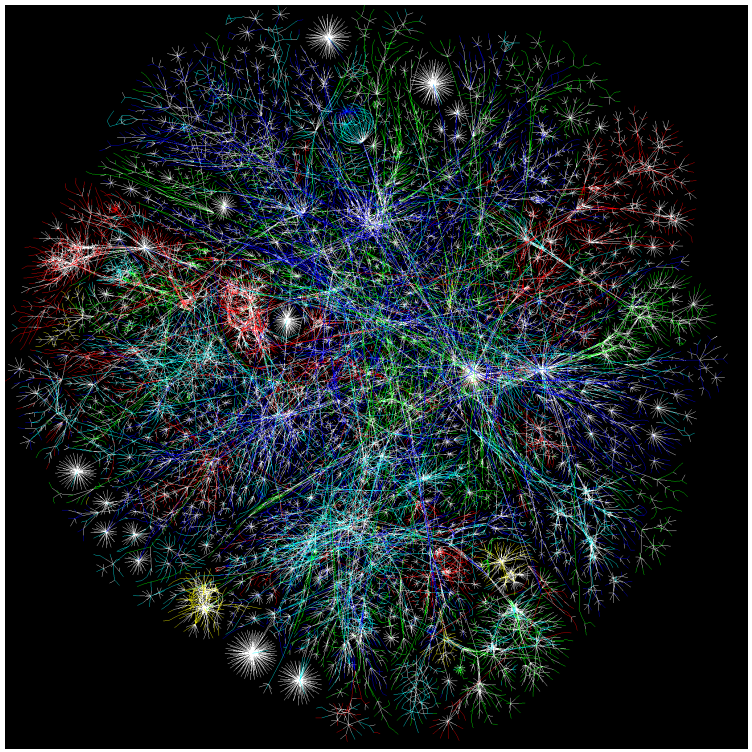
Model selection for stochastic block models

Common approaches

- Use the model you like.
- Make a choice based on domain expertise.
- Use off-the-shelf Information Criteria for independent data.
  - Akaike information criterion (AIC), Bayesian information criterion (BIC), etc.
- Non-parametric methods

Generalization test (cross validation)

- Node classification
- Link prediction
- The good
  - can compare any model
  - generalization performance focused
- The bad
  - require multiple data instances
  - or the ability to divide data into i.i.d. subsamples
  - multiple runs lead to inefficiency



Frequentist model selection

Likelihood Ratio Test for block models

- Model selection between a pair of nested models as a hypothesis test
- Test results have proper confidence intervals
- The likelihood ratio test (LRT) is the uniformly most powerful test
- Basis for many off-the-shelf statistical tools (AIC)

Constructing a LRT

- Null model H_0, nesting alternative H_1
- $$\Lambda(G) = \log \frac{\sup_{H \in H_1} P(G \mid H)}{\sup_{H \in H_0} P(G \mid H)},$$
- Reject the null model when Λ exceeds some threshold, which is based on
  - our desired error rate
  - Null distribution of Λ
- To get the Null distribution of Λ, we can
  - analytic prediction
  - parametric bootstrapping

LRT for block models

- Classic $\frac{1}{2}\chi^2_\ell$ result
- Key assumptions:
  - parameter estimates have Gaussian distributions
  - central limit theorems for IID data
  - large data limit
- Networks data is relational
- The latent block assignment variables are discrete
- Sparse networks far from large data limit

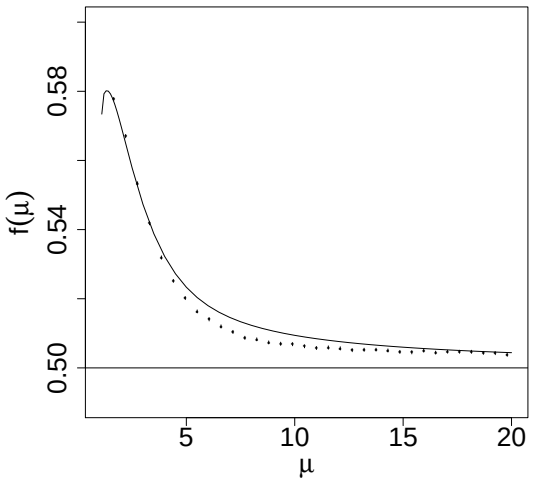
Bootstrapping results using BP

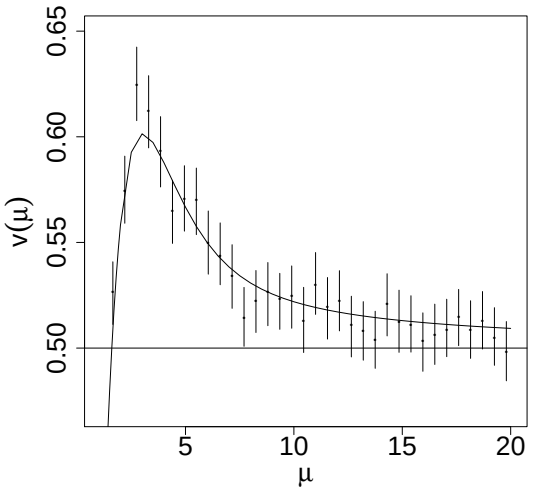
Likelihood Ratio Test for SBM vs DC-SBM

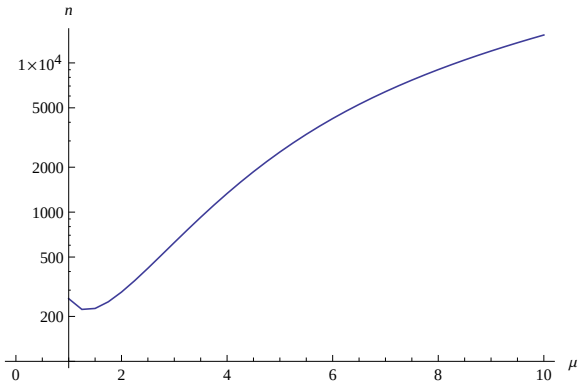
- H_0 : Graph is generated according to the Vanilla-SBM
- H_1 : Graph is generated according to the DC-SBM, where an edge is generated with probability $\theta_u \theta_v p_{g(u)g(v)}$.

$$\Lambda(G) = \log \frac{\sup_{H \in H_1} P(G \mid H)}{\sup_{H \in H_0} P(G \mid H)},$$

- According to classic χ^2 test, $\Lambda(G) \sim \frac{1}{2}\chi^2_\ell$ with $\ell = n - k$







Bayesian model selection

The Bayesian approach

- Integrating over parameters of different fit
- The posterior is proportional to total likelihood
- BIC has close relation with Minimum Description Length

Posterior of the SBM

$$P(M_i \mid G) = \frac{P(M_i)}{P(G)} P(G \mid M_i)$$

$$\propto \sum_g \int\int\int_0^1 \mathrm{d}\{p_{st}\}\mathrm{d}\{q_s\}\, P(G, g \mid q, p)$$

- Uniform prior on $P(M_i)$
- Constant evidence $P(G)$
- Bayes factor

Bayesian model selection for SBMs

- The good
  - compare any model with proper posterior
  - combine domain prior with data
  - conjugate priors lead to tractability
- The bad
  - realistic priors often not conjugate
  - model selection is inherently NOT Bayesian
- Belief propagation for the sum over integrated likelihood

Conclusions

The variational expectation maximization framework

- A flexible framework for community based models
  - Mixed membership block model (Airoldi, Blei, Fienberg and Xing, 2008)
  - The Ball-Karrer-Newman model (2011)
- Choice of variational distribution balance between accuracy and scalability
- SBM: Linear Belief Propagation for the exponential sum
- The analogy between machine learning and statistical physics is powerful

Applications

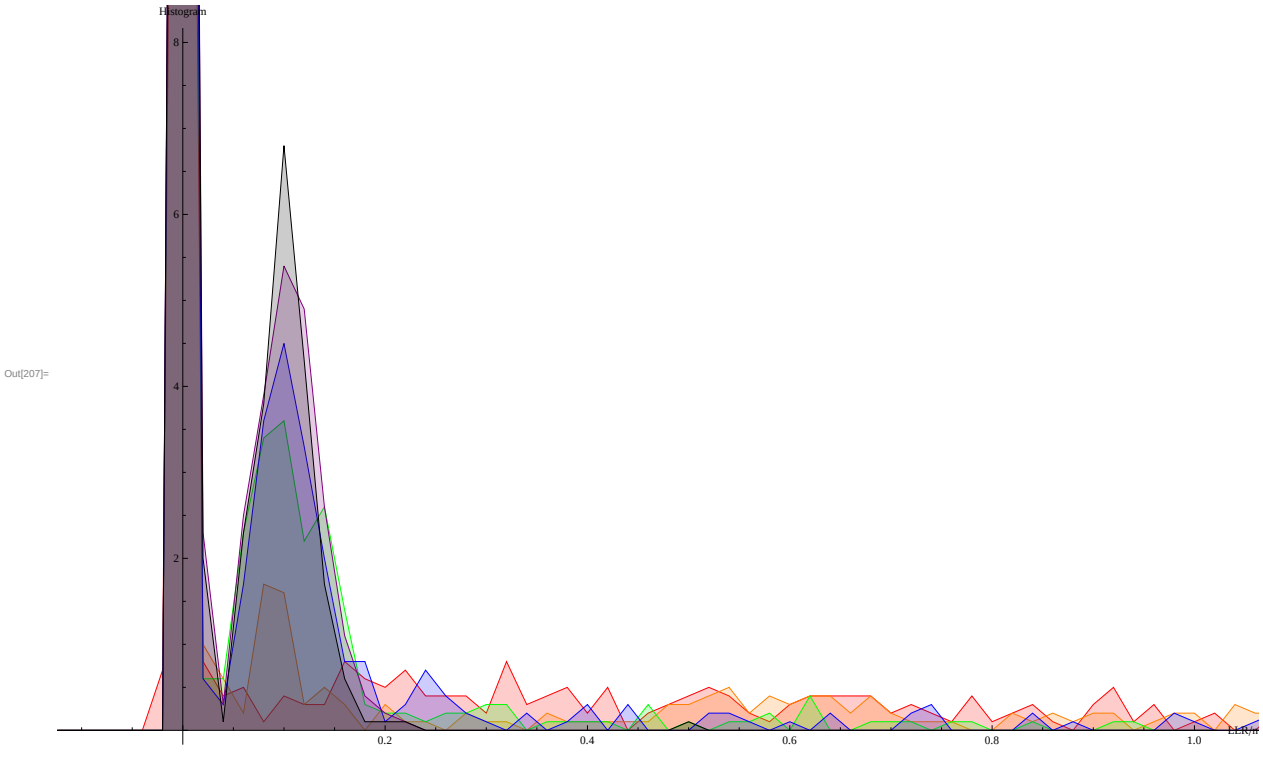
- Works for both Frequentist and Bayesian model selection
- The right scalability and accuracy lead to better theories
- Project: Bayesian recommendation system based on the SBM (Roger Guimerà)
- Even faster algorithm with message sub-sampling

Thank you

Looking for Postdoc opportunities

- Getting my Ph.D. this July
- Up for interesting projects in any discipline
- Preferably in the US, but open for other places
- everyxt@gmail.com

The secret last slide



Order selection for SBMs

- Even a challenge for classic i.i.d mixture models
- Degeneracy at zero
- The non-zero peak scale with the size of network
- AIC and BIC are bad for order selection