Model selection for Stochastic Block Models

Xiaoran Yan

May 10, 2013

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

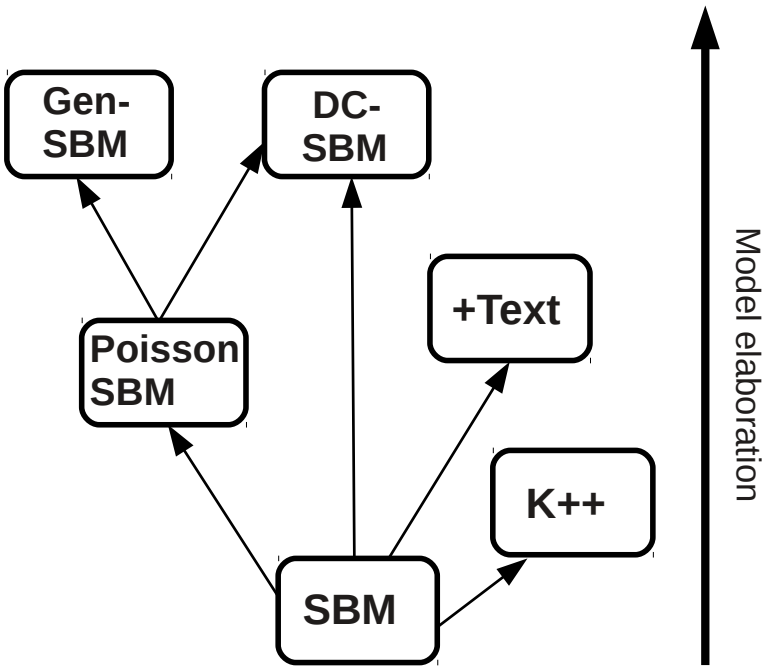
Stochastic Block Models

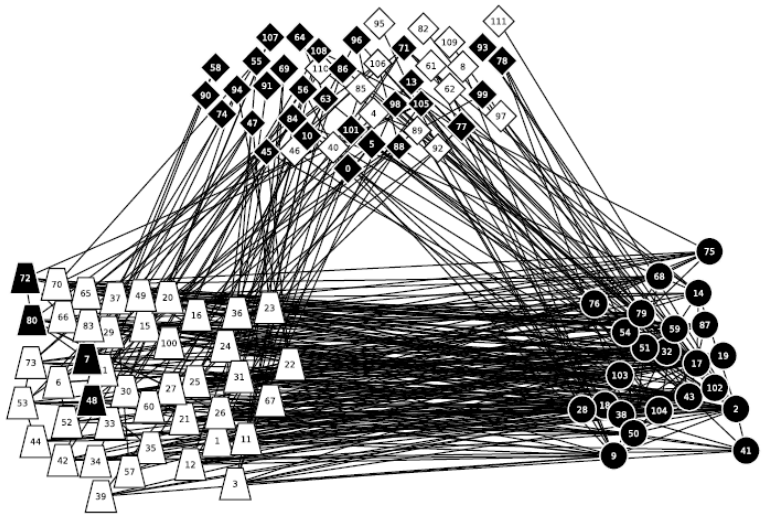
The good

- Has some good statistical properties
- General enough to capture different structures
- Flexible extensions for rich data

The bad

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

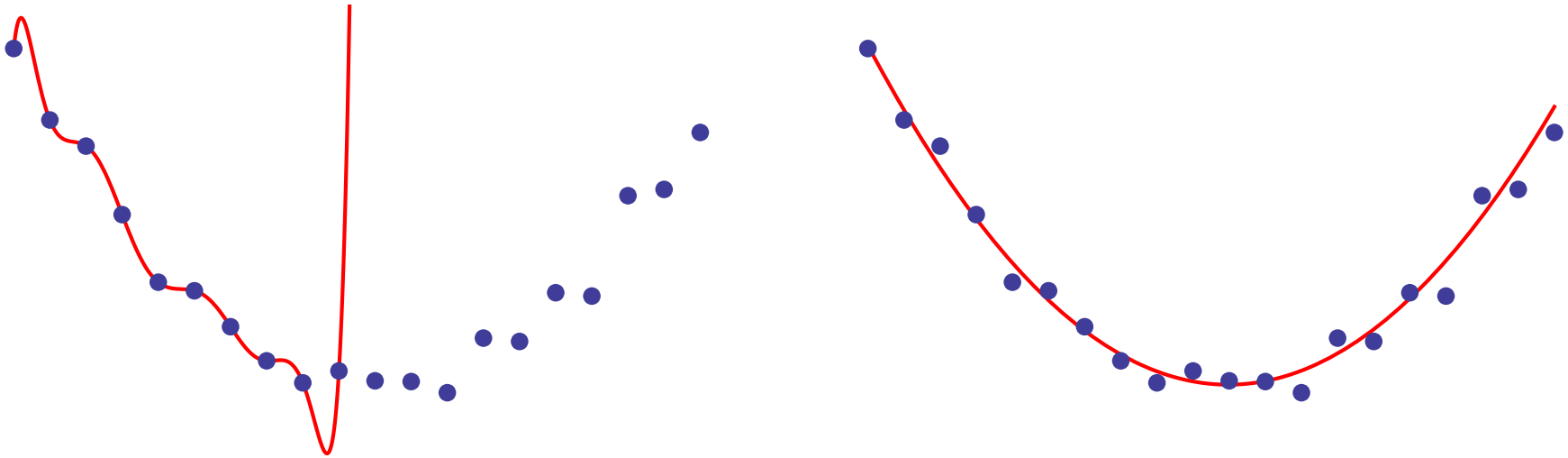




Model selection

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 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.

Generalization test (cross validation)

- Link prediction
- Node label prediction
- Network feature check
- The good
  - can compare any model
  - generalization performance focused
- The bad
  - require multiple data samples
  - relational data lead to biased Subsamples
  - multiple runs lead to inefficiency

Bayesian model selection

- Integrating over parameters of different fit
- The posterior is proportional to total likelihood
- The good
  - compare any model with proper posterior
  - combine domain prior with data
  - conjugate priors lead to tractability
- The bad
  - conjugate priors often not realistic
  - realistic priors often not conjugate
- BIC has close relation with Minimum Description Length

Likelihood Ratio Test for block models

Frequentist model selection

- 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

Constructing a LRT

- Null model H_0,
- The more general, nesting alternative H_1
- $$\Lambda(G) = \log \frac{\sup_{H_1} P(G \mid H_1)}{\sup_{H_0} P(G \mid H_1)},$$
- 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
  - parametric bootstrapping
  - analytic prediction

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
- Sparse networks far from large data limit
- Degenerated nesting for order selection

Likelihood Ratio Test for Poisson-SBM vs DC-SBM

Constructing a LRT

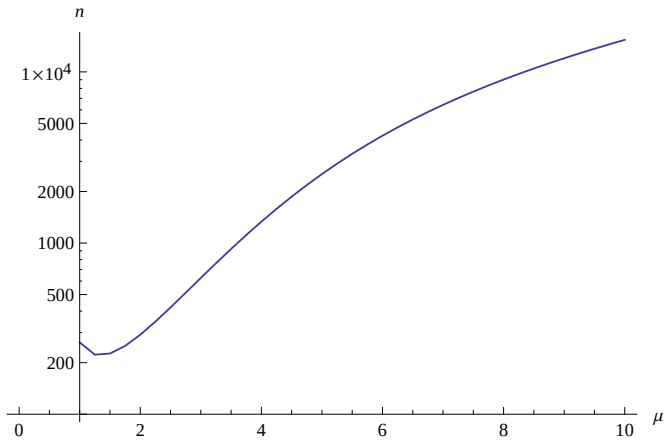
- H_0 : Graph is generated according to the Poisson-SBM
- H_1 : Graph is generated according to the DC-SBM

$$\Lambda_{DC}(G) = \log \frac{\sup_{H_1} \sum_g \prod_u \theta_u^{d_u} \prod_r q_s^{n_s} \prod_{st} \omega_{st}^{m_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st})}{\sup_{H_0} \sum_g \prod_r q_s^{n_s} \prod_{st} \omega_{st}^{m_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st})}$$

Parametric bootstrapping

- Bethe free energy
- Weak non-edge messages
- Further approximation of $O(n+m)$ complexity
- 1k samples for networks of size $n = 10^5$
- Results show that the classic $\frac{1}{2}\chi^2_\ell$ $(\ell = n - k)$ requires correction.

$$\mu_r^{u\to v} = \frac{1}{Z^{u\to v}} \gamma_r \prod_{w\neq u,v} \sum_{s=1}^{k} \mu_s^{w\to u} f(\theta_w, \theta_u, \omega_{rs}, A_{wu})$$



The size n, as a function of the average degree μ, above which naive χ^2 testing commits a type I error with 95% confidence.

Poisson stochastic block model (Poisson-SBM)

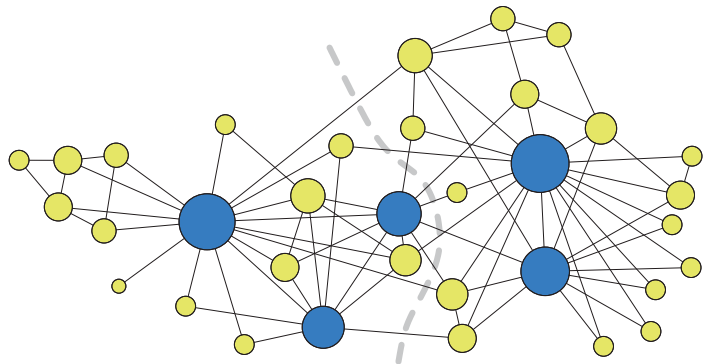
Assumptions (Karrer and Newman)

- We represent our network as an undirected multi-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 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 $\{u, v\}$, the number of edges follow a Poisson distribution with mean $\omega_{g(u)g(v)}$, and they are independent. Let m_{st} be the number of edges from type s to type t, with $\sum_{st} m_{st} = m$.

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

$$P(G, g|\omega, q) = \prod_u q_{g_u} \prod_{u<v} \frac{\omega_{g_u g_v}^{A_{uv}} e^{-\omega_{g_u g_v}}}{A_{uv}!}$$

$$= \prod_{s=1}^{k} q_s^{n_s} \prod_{s,t=1}^{k} \omega_{st}^{e_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st}) \prod_{u<v} \frac{1}{A_{uv}!} . \quad (1)$$



Degree corrected stochastic block model (DC-SBM)

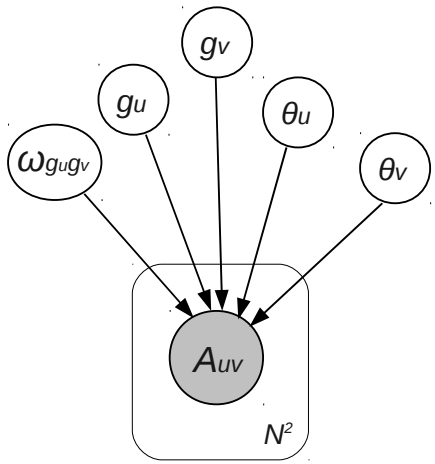
Motivations:

- For the vanilla SBMs, any two nodes in the same block follow the same Poisson degree distribution.
- As a consequence, vanilla SBMs "resists" putting nodes with very different degrees in the same block.
- DC-SBM introduces an additional parameter θ_u for each node, which scales the number of edges connecting it to other nodes(Karrer and Newman). We force θ_u to sum to the total number of nodes within each block: $\sum_{u:g_u=s} \theta_u = n_s$

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

$$P(G, g \mid \theta, \omega, q) = \prod_u q_{g_u} \prod_{u<v} \frac{(\theta_u \, \theta_v \, \omega_{g_u g_v})^{A_{uv}}}{A_{uv}!} \exp(-\theta_u \, \theta_v \, \omega_{g_u g_v})$$

$$= \prod_u \theta_u^{d_u} \prod_{s=1}^{k} q_s^{n_s} \prod_{s,t=1}^{k} \omega_{st}^{e_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st}) \prod_{u<v} \frac{1}{A_{uv}!} \qquad (2)$$



Likelihood Ratio Test for Poisson-SBM vs DC-SBM

Constructing a LRT

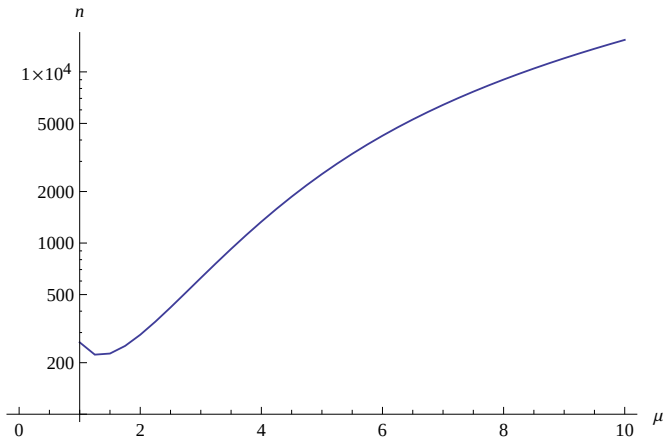
- H_0 : Graph is generated according to the Poisson-SBM
- H_1 : Graph is generated according to the DC-SBM

$$\Lambda_{DC}(G) = \log \frac{\sup_{H_1} \sum_g \prod_u \theta_u^{d_u} \prod_r q_s^{n_s} \prod_{st} \omega_{st}^{m_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st})}{\sup_{H_0} \sum_g \prod_r q_s^{n_s} \prod_{st} \omega_{st}^{m_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st})}$$

Parametric bootstrapping

- Bethe free energy
- Weak non-edge messages
- Further approximation of $O(n+m)$ complexity
- 1k samples for networks of size $n = 10^5$
- Results show that the classic $\frac{1}{2}\chi^2_\ell$ $(\ell = n - k)$ requires correction.

$$\mu_r^{u\to v} = \frac{1}{Z^{u\to v}} \gamma_r \prod_{w\neq u,v} \sum_{s=1}^{k} \mu_s^{w\to u} f(\theta_w, \theta_u, \omega_{rs}, A_{wu})$$



The size n, as a function of the average degree μ, above which naive χ^2 testing commits a type I error with 95% confidence.

Deriving the correct Null distribution

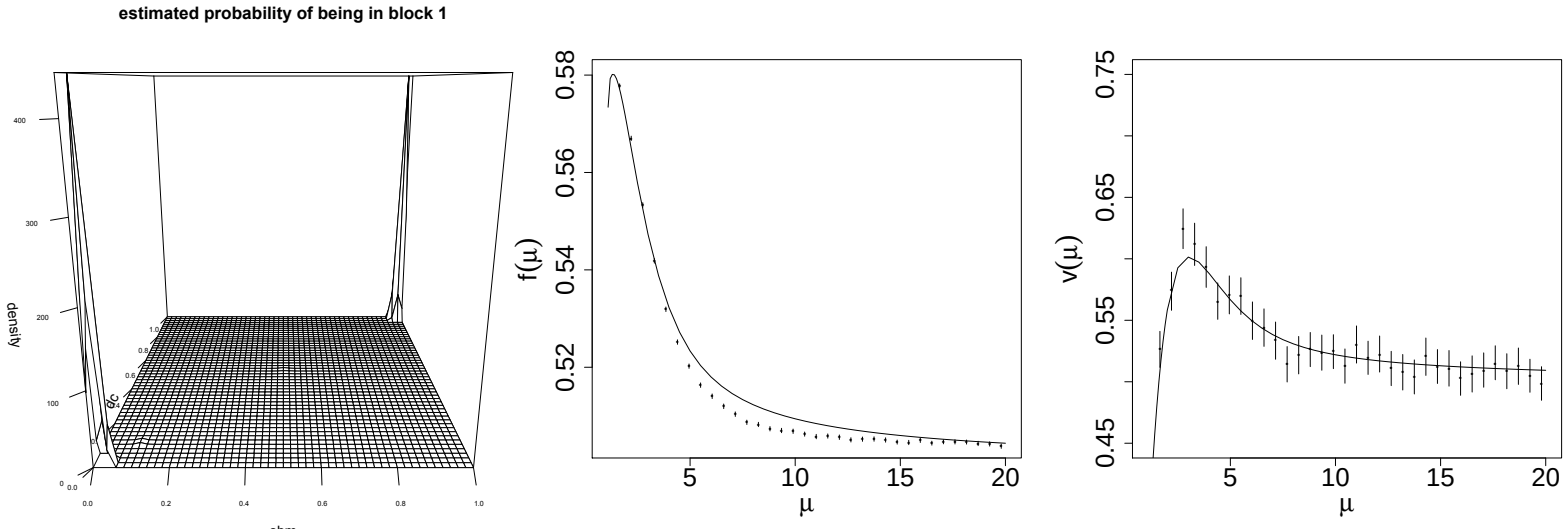
Theoretical derivation

MLEs of both model converge to:

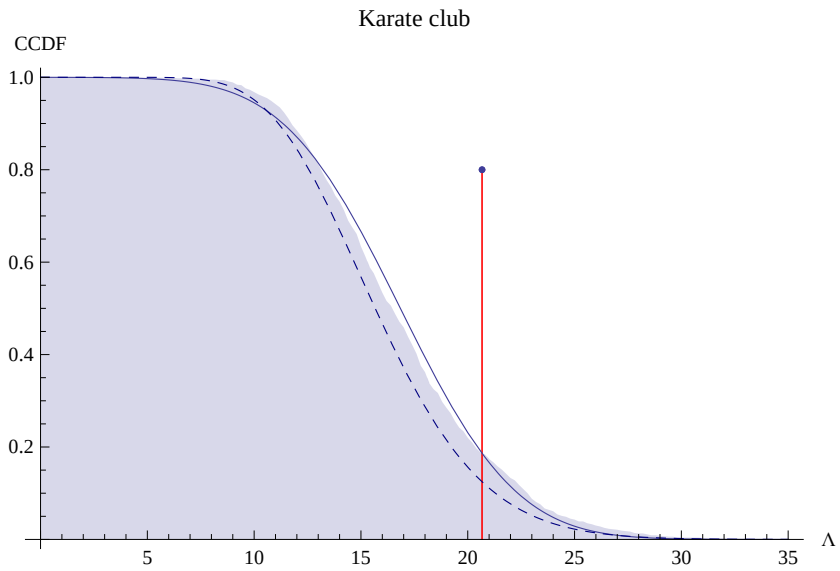
$$\hat{q}_s = \frac{\bar{n}_s}{n}\,, \quad \hat{\omega}_{st} = m_{st}/(n_s n_t)\,, \quad \hat{\theta}_u = \frac{d_u}{\bar{d}_{g_u}}\,.$$

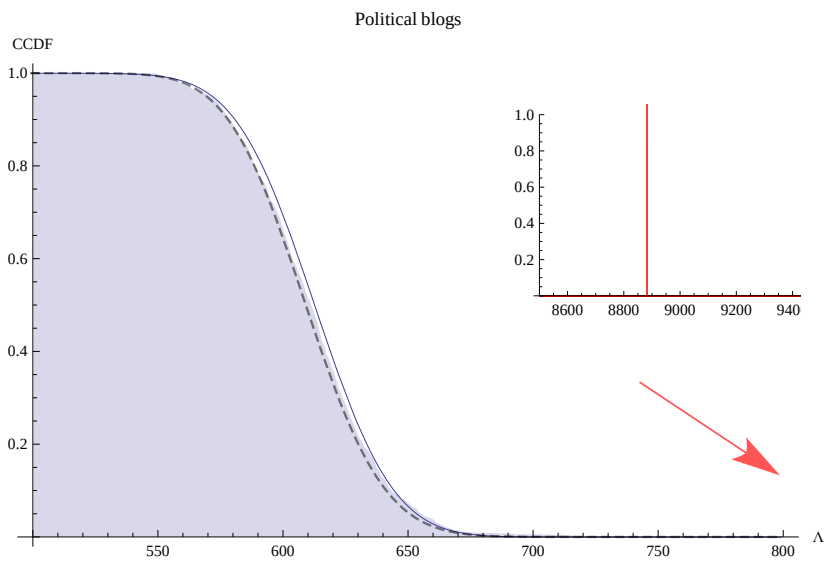
$$\Lambda_{DC}(G) = \log \frac{\sup_{H_1} \sum_g \prod_u \theta_u^{d_u} \prod_r q_s^{n_s} \prod_{st} \omega_{st}^{m_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st})}{\sup_{H_0} \sum_g \prod_r q_s^{n_s} \prod_{st} \omega_{st}^{m_{st}/2} \exp(-\frac{1}{2} n_s n_t \omega_{st})}$$

$$= \log \prod_u \left(\frac{d_u}{\bar{d}_{g_u}} \right)^{d_u} = \sum_u d_u \log \frac{d_u}{\bar{d}_{g_u}}\,.$$



Results on real world networks





- Evidence is not strong enough to reject Poisson-SBM
- The observed $\Lambda = 20.7$ has a p-value of 0.19 according to the theoretical Gaussian and the simulation (shade)
- 0.13 according to the χ^2 (dashed)

- Evidence is overwhelmingly strong to reject Poisson-SBM
- The observed log-likelihood ratio $\Lambda = 8883$ is 330 standard deviations above the mean.
- according all three distributions.

Thank you

Looking for Postdoc opportunities

- Graduating this July
- Up for any interesting projects
- everyxt@gmail.com