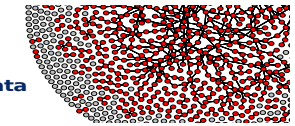
Fast Variational Algorithms for Statistical Network Modeling

and other network modeling advances

David Hunter
Michael Schweinberger
Duy Vu
Ruth Hummel

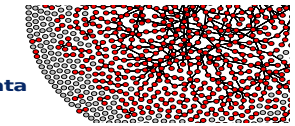
Department of Statistics, Penn State University

MURI meeting, Nov 12, 2010



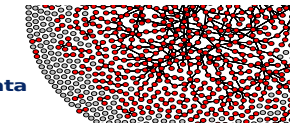
Outline

1. Variational EM
2. Maximum Likelihood Estimation for ERGMs
3. Hierarchical ERG models
4. On the horizon: Relational event models and degeneracy theory



Outline

1. Variational EM
2. Maximum Likelihood Estimation for ERGMs
3. Hierarchical ERG models
4. On the horizon: Relational event models and degeneracy theory



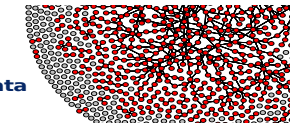
Variational EM algorithms

Goal: Scalable algorithm for clustering of nodes and simultaneous estimation of network parameters of interest (e.g., reciprocity, propensity to form edges) that:

- assumes *dyadic* (not *edgewise*) independence;
- assumes the nodes are partitioned in (latent) categories;
- allows for categorical (not merely 0/1) edge values;
- **is scalable to large ($\geq 1e+5$ nodes) networks**;
- allows for statatistical inference (e.g., confidence intervals).

	dyadic indep.	latent cat.	scalable alg.	cat. edges	stat. inf.
N & S (2001)	yes	yes	no	yes	no
D, P & R (2008)	no	yes	yes	no	no

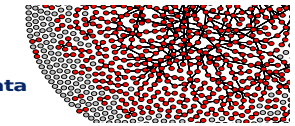
Nowicki & Snijders (2001); Daudin, Picard, & Robin (2008)



Dyadic independence ERGM with reciprocity

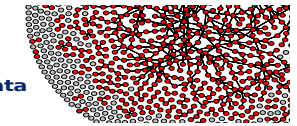
Work with Duy Vu, graduate student at PSU:

- Assume edges are directed, taking three values: $-1, +1, 0$
- There are five different types of dyads.
- Assuming homogeneity for now, Let π_i denote the probability of each type:
  - $\pi_1 = P_\theta(Y_{ij} = -1, Y_{ji} = 0)$
  - $\pi_2 = P_\theta(Y_{ij} = 1, Y_{ji} = 0)$
  - $\pi_3 = P_\theta(Y_{ij} = -1, Y_{ji} = 1)$
  - $\pi_4 = P_\theta(Y_{ij} = -1, Y_{ji} = -1)$
  - $\pi_5 = P_\theta(Y_{ij} = 1, Y_{ji} = 1)$
- Because we assume independent dyads, these parameters give the full model.

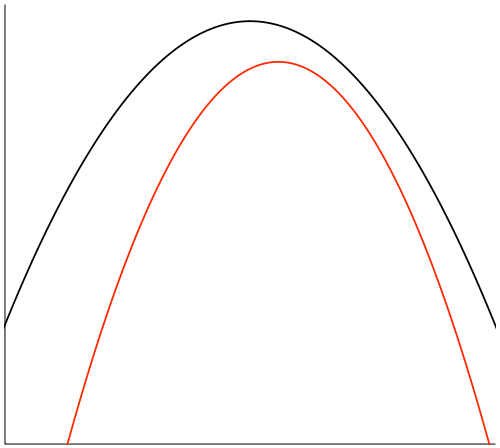


Mixture structure

- Assume each node comes from one of C latent classes.
- Instead of five parameters $\pi_1, \ldots, \pi_5$, we introduce $\pi^1_{k\ell}, \ldots, \pi^5_{k\ell}$, where k and ℓ range from 1 to C.
- Therefore, conditional on $Z_i = k$ and $Z_j = \ell$,
  - $\pi^1_{k\ell} = P_\theta(Y_{ij} = -1, Y_{ji} = 0)$
  - $\pi^2_{k\ell} = P_\theta(Y_{ij} = 1, Y_{ji} = 0)$
  - $\pi^3_{k\ell} = P_\theta(Y_{ij} = -1, Y_{ji} = 1)$
  - $\pi^4_{k\ell} = P_\theta(Y_{ij} = -1, Y_{ji} = -1)$
  - $\pi^5_{k\ell} = P_\theta(Y_{ij} = 1, Y_{ji} = 1)$
- Note: We assume $\pi^4_{k\ell} = \pi^4_{\ell k}$ and $\pi^5_{k\ell} = \pi^5_{\ell k}$.
- Conditional on all the Z_i, we have a closed-form loglikelihood (from earlier development).
- Marginally, let $\lambda_k = P(Z_i = k)$.



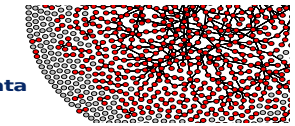
Variational approach



- For MLE, goal is to maximize the loglikelihood $\ell(\pi, \lambda)$.
- Basic idea: Establish lower bound

$$J(\pi, \lambda, \tau) \leq \ell(\pi, \lambda) \tag{1}$$

- Create an EM-like algorithm guaranteed to increase $J(\pi, \lambda, \tau)$ at each iteration.
- If we maximize the lower bound, then we're hoping that the inequality (1) will be tight enough to put us close to a maximum of $\ell(\pi, \lambda)$.

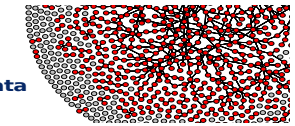


The eOpinion dataset (Richardson et al, 2003)

- General consumer review site Epinions.com.
- Members of the site can decide whether to "trust" each other.
- "Web of Trust" combined with review ratings to determine which reviews are shown to the user.
- 131,828 nodes, 841,372 signed edges

- To choose number of clusters, we use an Integrated Completed Likelihood (ICL) criterion as in Daudin et al (2008):

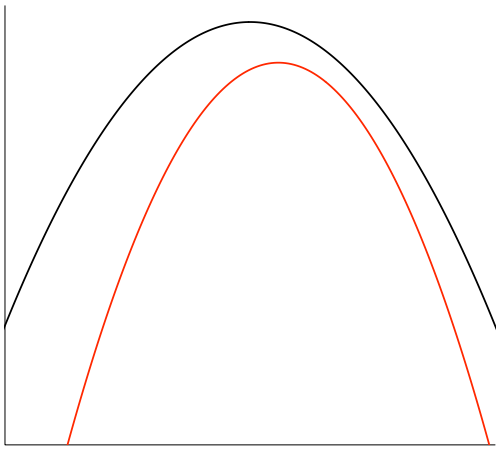
2	3	4	5	6	7	8	9	10
−1.29	−1.23	−1.19	−1.17	**−1.147**	−1.25	−1.32	−1.44	−1.45



Standard Error Estimates

Earlier, we established a lower bound

$$J(\pi, \lambda, \tau) \leq \ell(\pi, \lambda).$$



- Standard procedure: Find Hessian matrix $\nabla^2 \ell(\hat{\pi}, \hat{\lambda})$
- Flawed alternative: Use $\nabla^2 J(\hat{\pi}, \hat{\lambda}, \hat{\tau})$
- Better: Parametric bootstrap idea, which Duy has made scalable

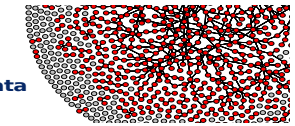
Outline

1 Variational EM

2 Maximum Likelihood Estimation for ERGMs

3 Hierarchical ERG models

4 On the horizon: Relational event models and degeneracy theory



Motivation: The likelihood function and MLE

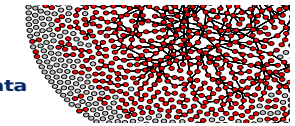
The ERG model class:

$$P_\theta(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } \kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}$$

- θ is a parameter vector to be estimated.
- $g(y)$ is a user-defined vector of graph statistics.
- The loglikelihood function is

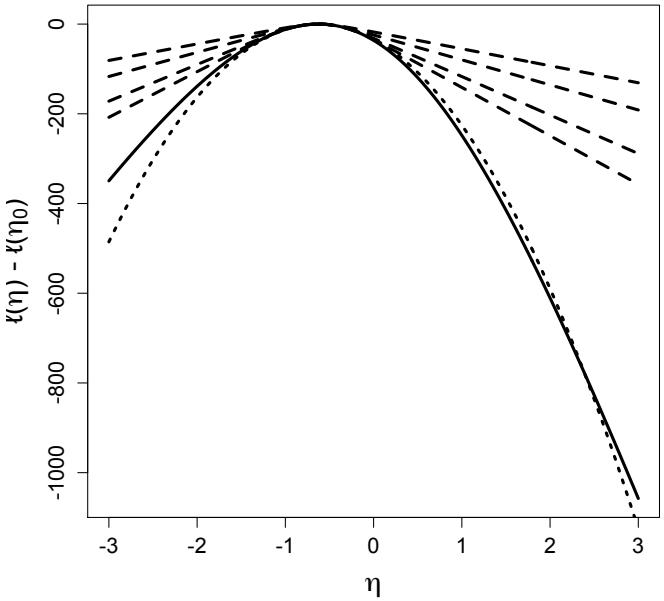
$$\ell(\theta) = \theta^t g(y^{\text{obs}}) - \log \kappa(\theta).$$

- The MLE is the maximizer $\hat{\theta}$ of the likelihood; finding it is very hard.



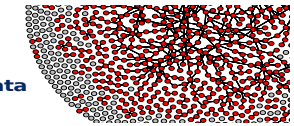
MCMC MLE, a new problem, and new solutions

Fix θ_0. By randomly simulating networks from the θ_0 model using MCMC, we can approximate the MLE.



Solid: Truth
Dashed: Approximations for samples of sizes up to 10^{15}
Dotted: Lognormal approximation

- Unfortunately, the quality of the approximation gets very poor as we move away from θ_0.
- Solution #1: Use a different (lognormal) approximation
- Solution #2: Use a "stepping" algorithm that tricks the estimation into staying close to θ_0.
- These solutions (Ruth Hummel's work) are now part of publicly available software
- More work to be done here!



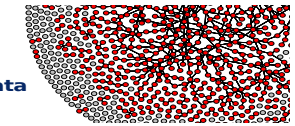
Outline

1 Variational EM

2 Maximum Likelihood Estimation for ERGMs

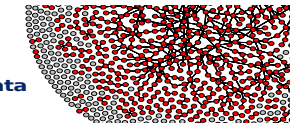
3 Hierarchical ERG models

4 On the horizon: Relational event models and degeneracy theory



Theory and Applications of hierarchical ERG models

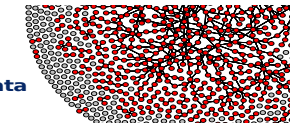
- A typical ERG model makes a *nodal homogeneity* assumption: All nodes have similar network-forming characteristics.
- Some of this is correctible by describing observable features (age, sex, job, etc.)
- Problem remains. For instance, consider degree heterogeneity:
  - Some nodes may be qualitatively different in their relationship-forming propensity
  - This quality may not be captured by an observable nodal trait.
- Michael Schweinberger has developed the `hergm` package to:
  - Impose a latent (unobserved) "edge-formation" attribute on the nodes;
  - use Bayesian methodology to perform inference for the result mixture model.



hergm: Application to disaster networks

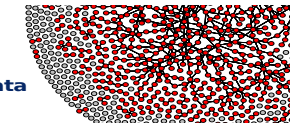
Michael Schweinberger (PSU) and Miruna Petrescu-Prahova (UW) studied the emergent multiorganizational networks (EMONs) formed during the first 12 days following the 9/11 attacks in New York.

- EMONs characterized by a small number of high-degree nodes and a large number of low-degree nodes
- Employed hierarchical ERGM methodology
- Goal: Consider organizational attributes such as type (government, non-profit, profit, collective) and scale (local to federal) to identify the processes that have given rise to the observed structure of the networks.
- Possible implications for disaster planning and emergency management result.



Outline

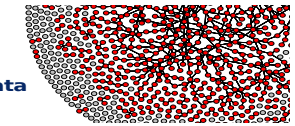
1. Variational EM
2. Maximum Likelihood Estimation for ERGMs
3. Hierarchical ERG models
4. On the horizon: Relational event models and degeneracy theory



Relational event models

Idea: When timing data are available on a network, do not merely treat the time-aggregated network;
Instead, consider each edge as an instantaneous event; model the stochastic process that produces these events.

- Duy Vu has begun to look at extending the ideas in Carter Butts' 2008 article on relational events.
- One can model "nodal intensity" processes or "dyadic intensity" processes. Scalable algoirthms possible for the former; more difficult for the latter.
- Applications to very large datasets such as citation networks; Duy has ideas for incorporating textual information (from, say, abstracts) to learn about citation networks
- Numerous collaborations with other MURI team members will be possible in this area: Carter's group, Padhraic's group, possibly others.



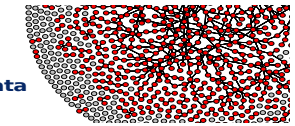
Contrastive Divergence (CD)

Consider the idea of MCMC MLE:

- Suppose we fix η_0. A bit of algebra shows that

$$-\log E_{\eta_0}\left[\exp\left\{(\eta-\eta_0)^t g(Y)\right\}\right] = \ell(\eta)-\ell(\eta_0). \qquad (2)$$

- The Law of Large Numbers suggests obtaining a sample of Y from the model using θ_0 as the parameter, then approximating the expectation by a sample mean.
- Q: How do we sample from $g(Y)$ using θ_0 as the parameter?
  A: Run MCMC infinitely long.
- But what if we only run MCMC for a single step (starting at y^{obs}), for a randomly chosen Y_{ij}?
- For this Y_{ij}, we're sampling from the conditional distribution given $(y^{\text{obs}})^c_{ij}$.

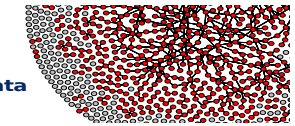


Contrastive Divergence (CD)

To summarize:

- Running an infinitely long Markov chain leads to the loglikelihood.
- Running a 1-step Markov chain leads to the pseudolikelihood.

Thus, if we alternately sample and then optimize the resulting "likelihood-like" function, we can view MLE and MPLE as two ends of a spectrum, the "contrastive divergence" spectrum. (MLE is CD-∞ and MPLE is CD-1.)



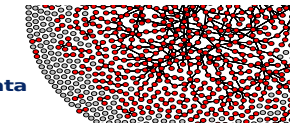
Contrastive Divergence (CD)

Considering CD-1...
Q: Is it better to

1. Repeatedly pick $i \neq j$ at random, or
2. Cycle through all possible $i \neq j$ in some systematic fashion?

A: The latter.

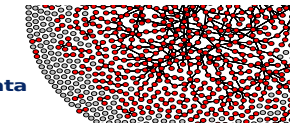
Considering CD-n...
Q: What is a good choice of n? How to optimize the tradeoff between n and sample size?
A: We don't yet know, but Arthur has been working on an answer.



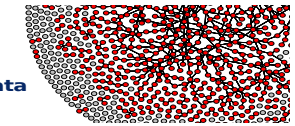
Instability and near-degeneracy of ERGMs

- ERGMs with interaction terms (e.g., stars, triangles) and strong node-homogeneity assumptions tend to be near-degenerate.
- Work begun by Michael Schweinberger sheds light on the near-degeneracy of ERGMs by introducing the notion of instability
- Can prove that unstable ERGMs tend to be asymptotically degenerate in a certain sense. sense of Strauss.
- Applications: ERGMs with Markov dependence and curved ERGMs
- Conclusion: Interaction in ERGM terms must be sufficiently weak.

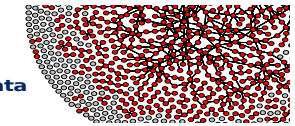
Michael's preliminary characterizations of these instabilities are the simplest I've seen. With further development, these ideas could be extremely useful in guiding effective modeling of networks.



Thank you!



Extra slides on variational EM

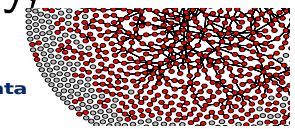


Dyadic independence ERGM with reciprocity

- Assume edges are directed, taking three values: $-1, +1, 0$
- Extension of p_1 model of Holland and Leinhardt (*JASA*, 1981):

$$P_\theta(Y = y) \propto \exp\left\{\sum_{i=1}^{5} \theta_i g_i(y)\right\}, \qquad \text{where}$$

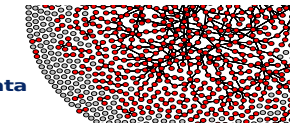
  - $g_1(y) =$ total # of -1 edges
  - $g_2(y) =$ total # of $+1$ edges
  - $g_3(y) =$ total # of discordant $-1, +1$ dyads
  - $g_4(y) =$ total # of concordant $-1, -1$ dyads
  - $g_5(y) =$ total # of concordant $+1, +1$ dyads
- NB: In principle, it would be easy to add more terms (say, nodal covariate terms)



Reparameterize for simplicity

$$P_\theta(Y = y) = \frac{\exp\left\{\sum_{i=1}^{5} \theta_i g_i(y)\right\}}{\kappa(\theta)}$$

- There are five different types of dyads.
- Assuming homogeneity for now, Let π_i denote the probability of each type:
  - $\pi_1 = P_\theta(Y_{ij} = -1, Y_{ji} = 0)$
  - $\pi_2 = P_\theta(Y_{ij} = 1, Y_{ji} = 0)$
  - $\pi_3 = P_\theta(Y_{ij} = -1, Y_{ji} = 1)$
  - $\pi_4 = P_\theta(Y_{ij} = -1, Y_{ji} = -1)$
  - $\pi_5 = P_\theta(Y_{ij} = 1, Y_{ji} = 1)$
- Because we assume independent dyads, these parameters give the full model.



Another reparameterization

- Recall

$$P_\theta(Y = y) = \frac{\exp\left\{\sum_{i=1}^{5} \theta_i g_i(y)\right\}}{\kappa(\theta)},$$

  where $g_1(y) =$ total # of -1 edges, etc.

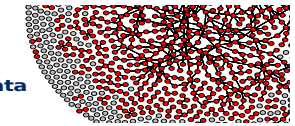
- Therefore, the *mean value parameterization* is

$$\mu_1 \stackrel{\text{def}}{=} E_\theta[g_1(Y)] = \sum_i \sum_{j \neq i} P_\theta(Y_{ij} = -1) = \sum_i \sum_{j<i} 2\left[\pi_4 + \pi_3 + \pi_1\right]$$

  because

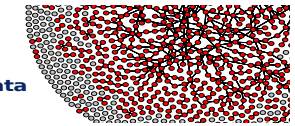
  - $\pi_1 = P_\theta(Y_{ij} = -1, Y_{ji} = 0)$
  - $\pi_3 = P_\theta(Y_{ij} = -1, Y_{ji} = 1)$
  - $\pi_4 = P_\theta(Y_{ij} = -1, Y_{ji} = -1)$

- Similarly, each μ_i is easily written in terms of π



Why Three Parameterizations?

- The dyad probabilities (the π_i) are convenient.
- The canonical (θ) and mean-value (μ) parameters are linked by *duality theory*:
  - Let $A(\theta) = \log \kappa(\theta)$ be the log normalizing constant.
  - Then
$$A^*(\mu) \stackrel{\text{def}}{=} \sup_{\theta} \left[\theta^\top \mu - A(\theta)\right]$$
    is the entropy of the model under $\theta(\mu)$.
  - Furthermore,
$$A(\theta) = \sup_{\mu} \left[\theta^\top \mu - A^*(\mu)\right]$$
- Since the entropy may be written explicitly in terms of π, we obtain an explicit formula for the loglikelihood (including the normalizing constant!) in terms of π.



A lower bound and variational EM

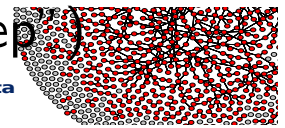
- Clever variational idea: Augment the parameter set, letting

$$\tau_{ik} = P(Z_i = k) \qquad \text{for all } 1 \le i \le n \text{ and } 1 \le k \le C.$$

- Let $R_\tau(Z) = \prod_i \text{Mult}(z_i; \tau_i)$ denote the joint dist. of Z.
- Direct calculation gives

$$\begin{aligned} J(\pi, \lambda, \tau) &\stackrel{\text{def}}{=} \ell(\pi, \lambda) - \text{KL}\left\{R_\tau(Z), P(Z \mid Y)\right\} \\ &= \ldots \\ &= E_\tau\left[\log P(Y, Z)\right] - H\left[R_\tau(Z)\right]. \end{aligned}$$

- Thus, an EM-like algorithm consists of alternately:
  - maximizing $J(\lambda, \pi, \tau)$ with respect to τ ("E-step")
  - maximizing $E_\tau\left[\log P(Y, Z)\right]$ with respect to π, λ ("M-step")

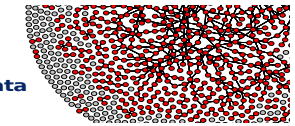


The Slashdot Zoo dataset (Kunegis et al, 2008)

- Technology-related news website known for its specific user community.
- In 2002 Slashdot introduced the Slashdot Zoo feature, which allows users to tag each other as friends or foes.
- The network was obtained in February 2009.
- 79,120 nodes, 515,581 signed edges

- To choose number of clusters, we use an Integrated Completed Likelihood (ICL) criterion as in Daudin et al (2008):

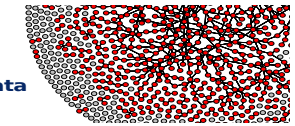
2	3	4	5	6	7	8	9	10
−8.76	−8.50	−8.12	−8.02	−7.93	−7.83	**−7.75**	−9.02	−9.16



Slashdot reciprocity parameters

- The ICL criterion chose 8 clusters.
- Here are the within-group estimates for discordant, -1-concordant, and $+1$-concordant parameters.
- Asymptotic standard errors in parentheses.

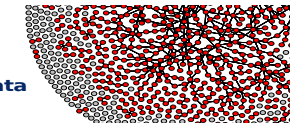
Parameters	discordant (θ_3)	-1-concordant (θ_4)	$+1$-concordant (θ_5)
Group 1 (269 actors)	0.2140 (0.4071)	2.7148 (0.2188)	3.1862 (0.0995)
Group 2 (55596 actors)	13.1101(0.0403)	13.8155(0.0570)	12.6018 (0.0427)
Group 3 (7206 actors)	8.6447 (0.2116)	10.0198(0.2456)	8.8422 (0.1164)
Group 4 (9257 actors)	9.0203 (0.1675)	10.9795(0.1759)	10.7145 (0.0705)
Group 5 (3492 actors)	4.7432 (0.2068)	8.0928 (0.1125)	7.5580 (0.0485)
Group 6 (1848 actors)	2.4766 (0.4102)	5.2717 (0.1968)	5.1187 (0.0793)
Group 7 (597 actors)	1.6174 (0.3697)	4.7837 (0.0903)	5.3132 (0.1339)
Group 8 (796 actors)	2.8147 (0.2064)	5.6858 (0.2303)	5.6640 (0.0402)



eOpinion reciprocity parameters

- The ICL criterion chose 6 clusters.
- Here are the within-group estimates for discordant, -1-concordant, and $+1$-concordant parameters.
- Asymptotic standard errors in parentheses

Parameters	discordant (θ_3)	-1-concordant (θ_4)	$+1$-concordant (θ_5)
Group 1 (5102 actors)	9.0377 (0.3258)	11.1657 (0.4108)	10.2289 (0.1154)
Group 2 (13007 actors)	7.6427 (0.1167)	10.1946 (0.1383)	9.4737 (0.0271)
Group 3 (107668 actors)	13.1540 (0.0208)	13.8155 (0.0294)	12.4924 (0.0230)
Group 4 (4303 actors)	2.7610 (0.1662)	5.9872 (0.1311)	6.1862 (0.0174)
Group 5 (976 actors)	0.3536 (0.0723)	2.7972 (0.0472)	3.6937 (0.0188)
Group 6 (738 actors)	1.7554 (0.3495)	5.5189 (0.5427)	4.4886 (0.0270)



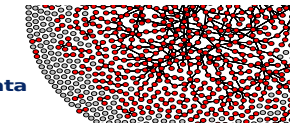
Clustering coefficients

As a check on the model, we compare the fitted model to the observed network based on four mixture-model clustering coefficients:

- Friend of friend is a friend

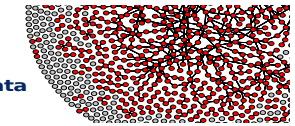
$$\text{i.e., } P(Y_{km} = 1 \mid Y_{kl} = Y_{lm} = 1)$$

- Enemy of Enemy is a friend
- Friend of enemy is an enemy
- Enemy of friend is an enemy

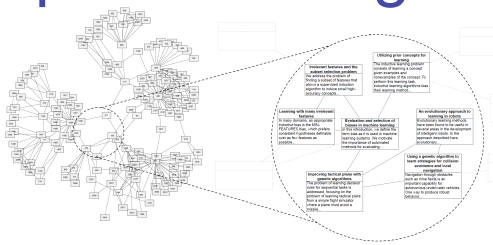


Cited References

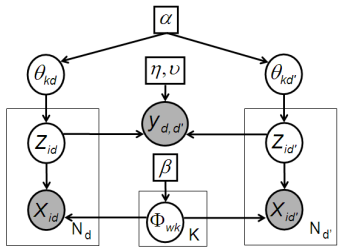
- Daudin JJ, Picard F, Robin S (2008, *Stat. & Comp.*), A Mixture Model for Random Graphs.
- Kunegis J, Lommatzsch A, and Bauckhage C (2009, *Proc. 18th Intl. Conf. on WWW*), The slashdot zoo: mining a social network with negative edges.
- Nowicki K and Snijders TAB (2001, *J. Am. Stat. Assoc*) Estimation and Prediction for Stochastic Blockstructures.
- M. Richardson M, Agrawal R, Domingos P (2003, *Intl. Sem. Web Conf.*), Trust Management for the Semantic Web.



FAST: **F**ast **A**nd **S**calable **T**opic-Modeling

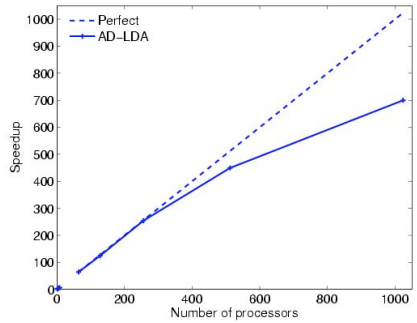


- Topic models are useful for analyzing data:
  - Text corpora, image databases, social networks
  - But many of these data sets are massive!



- This toolbox focuses on *efficient* & *scalable* inference:
  - Parallel/distributed inference (700x speedup on 1K procs)
  - Accelerated Gibbs sampling and variational inference ("real-time" learning)
  - Efficient inference for the Relational Topic Model

- MATLAB, C, & MPI code available at: http://www.ics.uci.edu/~asuncion/software/fast.htm



Scalable Methods for the Analysis of Network-Based Data