Model-Based Clustering of Large Networks

Duy Q. Vu, Uni. Melbourne
David R. Hunter, Penn State Univ.
Michael Schweinberger, Rice Univ.

Supported by Office of Naval Research
MURI Award No. N00014-08-1-1015

Interface Conference on Applied Statistics
October 23, 2014

Epinions.com: Example of large network dataset

Unbiased Reviews by Real People

- Members of Epinions.com can decide whether to ”trust” each other.
- “Web of Trust” combined with review ratings to determine which reviews are shown to the user.
- Dataset of Massa and Avesani (2007):
  - $n = 131{,}828$ nodes
  - $n(n-1) = 17.4$ billion observations
  - 841,372 of these are nonzero (± 1)

The Goal: Cluster 131,828 users

- Basis for clustering: Patterns of trusts and distrusts in the network
- If possible: understand the features of the clusters by examining parameter estimates.

Unbiased Reviews by Real People

Notation: Throughout, we let y_{ij} be rating of j by i and $y = (y_{ij})$.

Next, a few words about how we might model observed (y_{ij}) *data. . .*

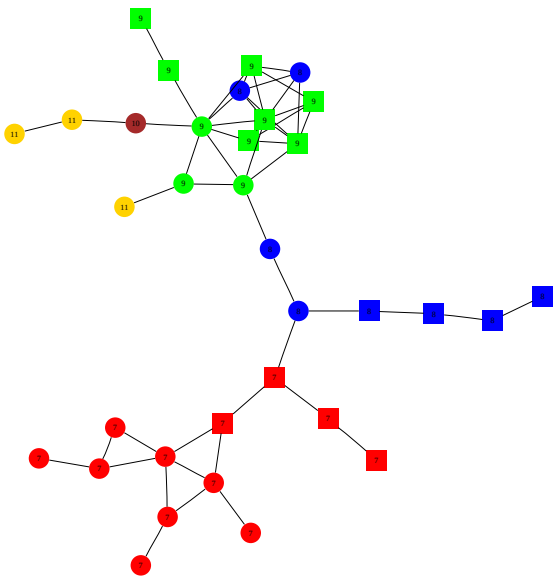
Estimation for exponential-family models can be hard

- General exponential-family random graph model (ERGM):

$$P_\theta(Y = y \mid x) = \exp\{\theta^\top g(x, y) - \psi(\theta)\},$$

  where y is a particular realization of the random network Y and x represents any covariates.
- The normalizing function is given by

$$\psi(\theta) = \sum_{\text{all possible } y'} \exp\{\theta^\top g(x, y')\}.$$



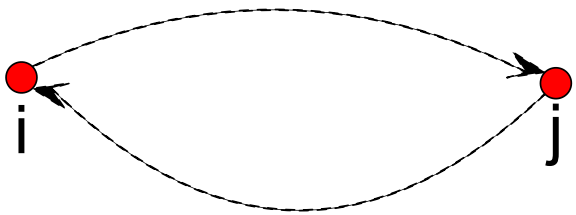
Pop quiz: How large is the set of all possible y' for this 34-node, symmetric, zero-one network?

We restrict attention to a more tractable model class

Special case of ERGMs called *dyadic independence*:

$$P_\theta(Y = y \mid x) = \prod_{i<j} P_\theta(D_{ij} = d_{ij} \mid x)$$

Dyad D_{ij}, directed case:



Dyad D_{ij}, undirected case:

Dyadic independence models have drawbacks but they

- facilitate estimation;
- facilitate simulation;
- avoid degeneracy issue (cf. Schweinberger, 2011).

To model dependence, add K-component mixture structure

Let Z_i denote the class membership of the ith node.

We assume

- $Z_i \overset{\text{iid}}{\sim} \text{Multinomial}(1; \gamma_1, \ldots, \gamma_K)$;
- $P_{\gamma,\theta}(Y = y \mid x) = \sum_z \prod_{i<j} P_\theta(D_{ij} = d_{ij} \mid x, Z = z) P_\gamma(Z = z)$.

In other words:
Conditional on the Z_i, we have a dyadic independence model.

Consider two examples of conditional dyadic independence for the Epinions dataset

1. The "full model" of Nowicki and Snijders (2001):

$$P_\theta(D_{ij} = d \mid Z_i = k, Z_j = l) = \theta_{d;kl}$$

2. A more parsimonious model:

$$\begin{aligned} P_\theta(D_{ij} = d_{ij} \mid Z_i = k, Z_j = l) \quad \propto \quad & \exp\{\theta^-(y_{ij}^- + y_{ji}^-) \\ & +\theta_k^\Delta y_{ji} + \theta_l^\Delta y_{ij} \\ & +\theta^{--} y_{ij}^- y_{ji}^- + \theta^{++} y_{ij}^+ y_{ji}^+\} \end{aligned}$$

where $y_{ij}^- = I\{Y_{ij} = -1\}$ and $y_{ij}^+ = I\{Y_{ij} = +1\}$.

- The term $\theta^+(y_{ij}^+ + y_{ji}^+)$ is missing from the second model to avoid perfect collinearity.

Consider two examples of conditional dyadic independence for the Epinions dataset

1. The "full model" of Nowicki and Snijders (2001):

$$P_\theta(D_{ij} = d \mid Z_i = k, Z_j = l) = \theta_{d;kl}$$

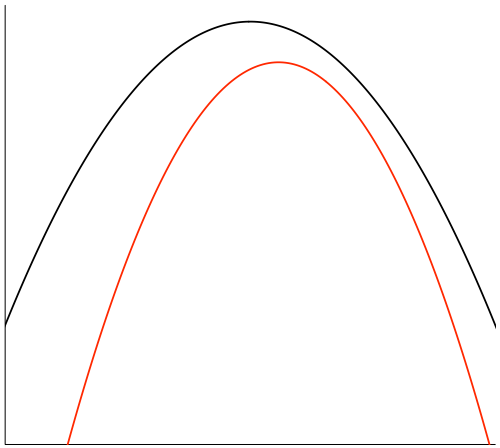
2. A more parsimonious model:

$$\begin{aligned} P_\theta(D_{ij} = d_{ij} \mid Z_i = k, Z_j = l) \quad \propto \quad & \exp\{\theta^-(y_{ij}^- + y_{ji}^-) \\ & +\theta_k^\Delta y_{ji} + \theta_l^\Delta y_{ij} \\ & +\theta^{--} y_{ij}^- y_{ji}^- + \theta^{++} y_{ij}^+ y_{ji}^+\} \end{aligned}$$

where $y_{ij}^- = I\{Y_{ij} = -1\}$ and $y_{ij}^+ = I\{Y_{ij} = +1\}$.

- When $K = 5$ components, these models have 109 and 12 parameters, respectively.

Approximate maximum likelihood estimation uses a variational EM algorithm



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

$$J(\gamma, \theta, \alpha) \leq \ell(\gamma, \theta) \tag{1}$$

  after augmenting parameters by adding α.
- Create an EM-like algorithm guaranteed to increase $J(\gamma, \theta, \alpha)$ 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 $\ell(\gamma, \theta)$.

We adapt the variational EM idea of Daudin, Picard, & Robin (2008).

We may derive a lower bound by simple algebra

- Clever variational idea: Augment the parameter set, letting

$$\alpha_{ik} = P(Z_i = k) \qquad \text{for all } 1 \leq i \leq n \text{ and } 1 \leq k \leq K.$$

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

$$\begin{aligned} J(\gamma, \theta, \alpha) &\stackrel{\text{def}}{=} \ell(\gamma, \theta) - \text{KL}\left\{A_\alpha(Z), P_{\gamma,\theta}(Z \mid Y)\right\} \\ &= \dots \\ &= E_\alpha\left[\log P_{\gamma.\theta}(Y, Z)\right] - H\left[A_\alpha(Z)\right]. \end{aligned}$$

- Thus, an EM-like algorithm consists of alternately:
  - maximizing $J(\gamma, \theta, \alpha)$ with respect to α ("E-step")
  - maximizing $E_\alpha\left[\log P_{\gamma,\theta}(Y, Z)\right]$ with respect to γ, θ ("M-step")

The variational E-step may be modified using a (non-variational) MM algorithm

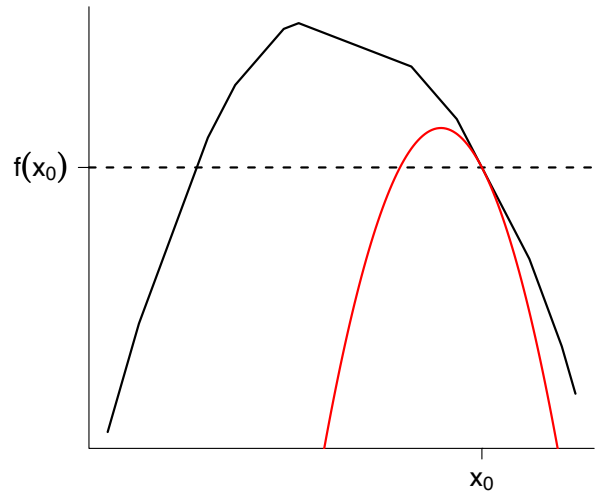
- Idea: Use a "generalized variational E-step" in which $J(\gamma, \theta, \alpha)$ is increased but not necessarily maximized.
- To this end, we create a surrogate function

$$Q(\alpha, \gamma^{(t)}, \theta^{(t)}, \alpha^{(t)})$$

of α, where t is the counter of the iteration number.

- The surrogate function is a *minorizer* of $J(\gamma, \theta, \alpha)$:
  It has the property that maximizing or increasing its value will guarantee an increase in the value of $J(\gamma, \theta, \alpha)$.

In the figure, the red curve minorizes $f(x)$ at x_0.



Construction of the minorizer of $J(\gamma, \theta, \alpha)$ uses standard MM algorithm methods

$$
\begin{aligned}
J(\gamma, \theta, \alpha) &= \sum_{i<j} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \alpha_{ik} \alpha_{j\ell} \log \pi_{d_{ij};k\ell}(\theta) \\
&\quad + \sum_{i=1}^{n} \sum_{k=1}^{C} \alpha_{ik} \left(\log \gamma_k - \log \alpha_{ik} \right).
\end{aligned}
$$

We may define a minorizing function as follows:

$$
\begin{aligned}
Q(\alpha, \gamma, \theta, \alpha^{(t)}) &= \sum_{i<j} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \left(\alpha_{ik}^2 \frac{\alpha_{j\ell}^{(t)}}{2\alpha_{ik}^{(t)}} + \alpha_{j\ell}^2 \frac{\alpha_{ik}^{(t)}}{2\alpha_{j\ell}^{(t)}} \right) \log \pi_{d_{ij};k\ell}(\theta) \\
&\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{ik} \left(\log \gamma_k - \log \alpha_{ik}^{(t)} - \frac{\alpha_{ik}}{\alpha_{ik}^{(t)}} + 1 \right).
\end{aligned}
$$

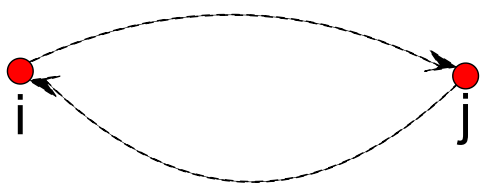
- Can be maximized (in α) using quadratic programming.

The parsimonious model for the Epinions dataset

$$P_\theta(D_{ij} = d_{ij} \mid Z_i = k, Z_j = l) \quad \propto \quad \exp\{\theta^-(y_{ij}^- + y_{ji}^-) + \theta_k^\Delta y_{ji} + \theta_l^\Delta y_{ij} + \theta^{--} y_{ij}^- y_{ji}^- + \theta^{++} y_{ij}^+ y_{ji}^+\}$$

where $y_{ij}^- = I\{Y_{ij} = -1\}$ and $y_{ij}^+ = I\{Y_{ij} = +1\}$.

Dyad D_{ij}, directed case:



- θ^-: Overall tendency toward distrust
- θ_k^Δ: Category-specific trustedness
- θ^{--}: *lex talionis* tendency (eye for an eye)
- θ^{++}: *quid pro quo* tendency (one good turn...)

Parameter estimates themselves are of interest

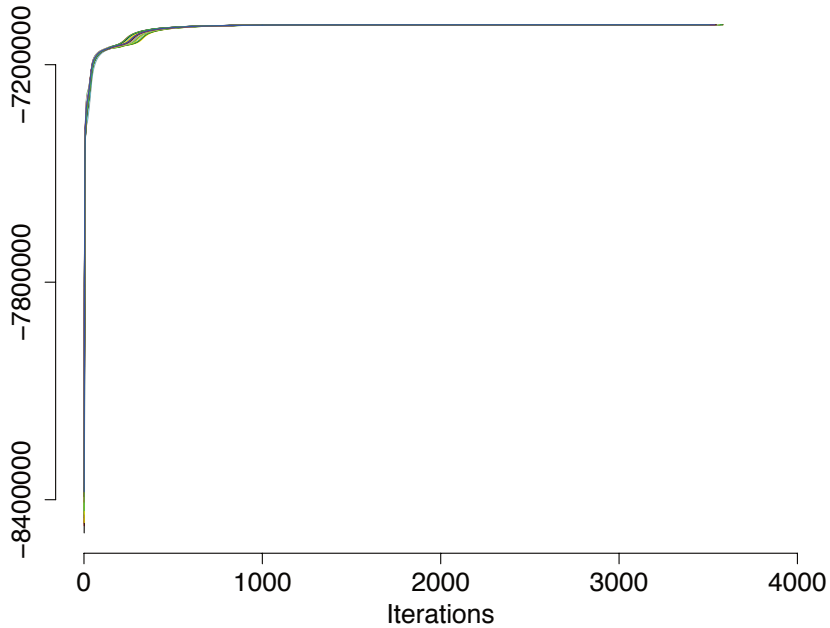
Parameter	Parameter Estimate	Confidence Interval
Negative edges (θ^{-})	-24.020	$(-24.029, -24.012)$
Positive edges (θ^{+})	0	—
Negative reciprocity (θ^{--})	8.660	$(8.614, 8.699)$
Positive reciprocity (θ^{++})	9.899	$(9.891, 9.907)$
Cluster 1 Trustworthiness (θ_1^{Δ})	-6.256	$(-6.260, -6.251)$
Cluster 2 Trustworthiness (θ_2^{Δ})	-7.658	$(-7.662, -7.653)$
Cluster 3 Trustworthiness (θ_3^{Δ})	-9.343	$(-9.348, -9.337)$
Cluster 4 Trustworthiness (θ_4^{Δ})	-11.914	$(-11.919, -11.908)$
Cluster 5 Trustworthiness (θ_5^{Δ})	-15.212	$(-15.225, -15.200)$

95% Confidence intervals based on parametric bootstrap using 500 simulated networks.

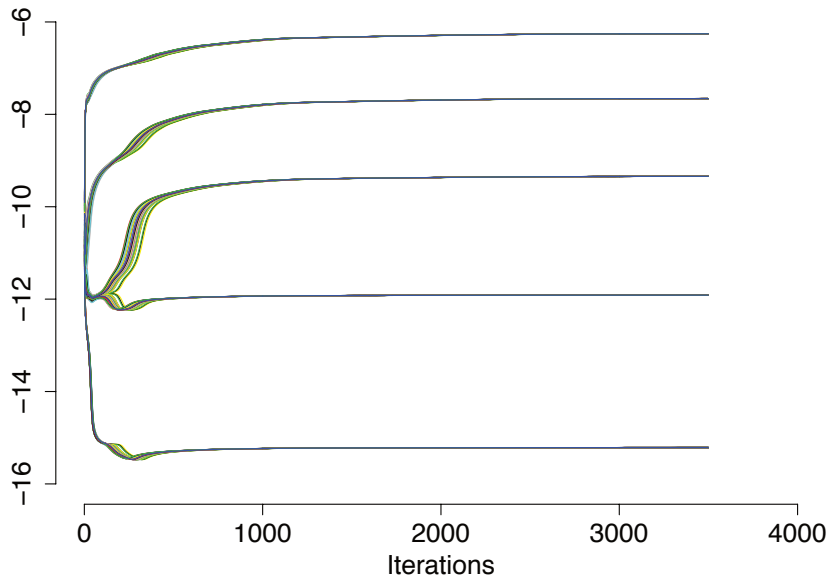
NB: There are some strange aspects of the bootstrap we cannot explain yet.

Multiple starting points converge to the same solution

Trace plots from 100 different randomly selected starting parameter values:



Loglikelihood values

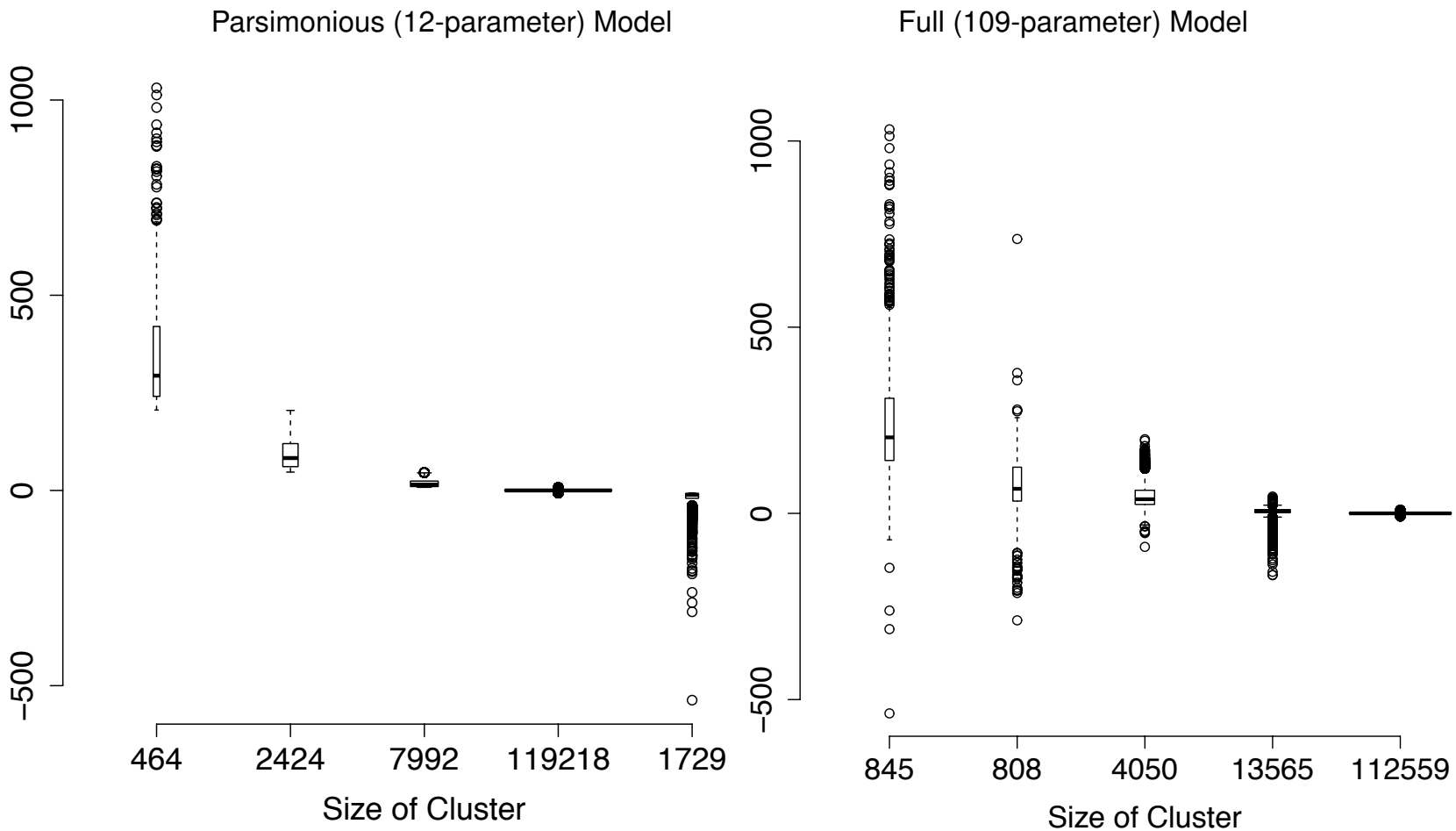


Trustedness parameters

Full (109-parameter) model results look nothing like this.

We may use average ratings of reviews by other users as a way to "ground-truth" the clustering solutions

659,290 articles categorized by author's highest-probability component. (Vertical axis is average article rating.)



Conclusion: This work extends the current state of the art in at least four ways

- Advances existing model-based clustering approaches via a simple and flexible modeling framework based on dyadic independence exponential family models.
- Introduces algorithmic improvements to the variational EM approach to approximate MLE.
- Considers bootstrap standard errors for parameter estimates
- Applies these methods to networks at least an order of magnitude larger than other networks previously considered.

Finally, we'd like to acknowledge the two giant mileposts in this developing body of work: Nowicki and Snijders (2001) and Daudin, Picard, and Robin (2008).

Cited References

- Daudin JJ, Picard F, Robin S (2008, *Stat. & Comp.*), A Mixture Model for Random Graphs.
- Massa P and Avesani P (2007, *Intl. J. on Semant. Web and Inf. Syst.*), Trust Metrics on Controversial Users.
- Nowicki K and Snijders TAB (2001, *J. Am. Stat. Assoc*), Estimation and Prediction for Stochastic Blockstructures.
- Schweinberger, M (2011, *J. Amer. Stat. Assoc.*), Instability, Sensitivity, and Degeneracy of Discrete Exponential Families.
- Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013, *Ann. Appl. Stat.*), Model-Based Clustering of Large Networks.

Simulating networks from the model is challenging

Let b denote the most common (baseline) value of a dyad in the network.
(Assuming sparsity, b is the empty dyad.)

1. Sample Z by sampling $\mathbf{M} \sim \text{Multinomial}(n; \gamma_1, \ldots, \gamma_K)$ and assigning nodes $1, \ldots, M_1$ to component 1, nodes $M_1 + 1, \ldots, M_1 + M_2$ to component 2, etc.
2. Sample $Y \mid Z$ as follows: For each $1 \leq k \leq l \leq K$,
   - 2.1 sample the number of dyads S_{kl} with non-baseline values: $S_{kl} \sim \text{Binomial}(N_{kl}, 1 - \pi_{b;kl})$, where N_{kl} is the number of pairs of nodes belonging to components k and l;
   - 2.2 sample S_{kl} out of N_{kl} pairs of nodes $i < j$ without replacement;
   - 2.3 for each of the S_{kl} sampled pairs of nodes $i < j$, sample the non-baseline value D_{ij} according to the probabilities $\pi_{d;kl}/(1 - \pi_{b;kl})$, $d \neq b$.