Model-Based Clustering of Large Networks

and some data science-related comments

David R. Hunter
Joint work with Duy Q. Vu and Michael Schweinberger

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

Penn State
Jan. 2020
Prologue: Data Science and Prediction
this litigation. The OUA considers a number of factors in making admissions decisions, including high school grades, standardized test scores, high school quality, curriculum strength, geography, alumni relationships, leadership, and race. During all relevant periods, the University has considered African-Americans, Hispanics, and Native Americans to be “underrepresented minorities,” and it is undisputed that the University admits virtually every qualified applicant from these groups. The current guidelines use a selection method under which every applicant from an underrepresented racial or ethnic minority group is automatically awarded 20 points of the 100 needed to guarantee admission.

Prediction done well is both prediction and explanation.
Epinions.com: Example of large network dataset

- 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 ($\pm 1$)
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.

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

Unbiased Reviews by Real People
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
  \[ \psi(\theta) = \sum_{\text{all possible } y'} \exp\{\theta^\top g(x, y')\}. \]

For this 34-node, symmetric, binary network, the sum has about \( 7.54 \times 10^{168} \) terms.
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 limit flexibility but they

- facilitate estimation;
- facilitate simulation;
To model dependence, add $K$-component mixture structure

Let $Z_i$ denote the class membership of the $i$th node.

We assume

- $Z_i \overset{iid}{\sim} \text{Multinomial}(1; \gamma_1, \ldots, \gamma_K);$  

- $P_{\gamma, \theta}(Y = y \mid x) = \sum_z \prod_{i<j} P_\theta(D_{ij} = d_{ij} \mid x, Z = z) P_\gamma(Z = z).$

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

*Clustering using a statistical model reveals more than cluster assignments.*
Consider two examples of conditional dyadic independence for the Epinions dataset

1. “Full” stochastic block model (Nowicki & Snijders, 2001):

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

2. A more parsimonious model:

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

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

- 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. “Full” stochastic block model (Nowicki & Snijders, 2001):

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

2. A more parsimonious model:

\[ P_\theta(D_{ij} = d_{ij} \mid Z_i = k, Z_j = l) \propto \exp\left\{ \theta^- (y^-_{ij} + y^-_{ji}) + \theta^\Delta y_{ji} + \theta^\Delta y_{ij} + \theta^-^- y^-_{ij} y^-_{ji} + \theta^+^+ y^+_{ij} y^+_{ji} \right\} \]

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.
Goal: Approximate maximum likelihood estimation

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

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

after augmenting parameters by adding $\alpha$.
- 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 $\ell(\gamma, \theta)$.

Far better an approximate answer to the ‘right’ question, which is often vague, than an ‘exact’ answer to the wrong question, which can always be made precise.

— John W. Tukey
Variational EM: Create then maximize the lower bound

▶ Clever variational idea: Augment the parameter set, letting

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

▶ Let \( A_\alpha(Z) = \prod_i \text{Mult}(z_i; \alpha_i) \) denote the joint dist. of \( Z \).

▶ Direct calculation gives

\[
J(\gamma, \theta, \alpha) \overset{\text{def}}{=} \ell(\gamma, \theta) - \text{KL} \{ A_\alpha(Z), P_{\gamma,\theta}(Z \mid Y) \}
\]

\[
= \ldots
\]

\[
= E_\alpha [\log P_{\gamma,\theta}(Y, Z)] - H[A_\alpha(Z)].
\]

▶ Thus, an EM-like algorithm consists of alternately:

▶ maximizing \( J(\gamma, \theta, \alpha) \) with respect to \( \alpha \) (“E-step”)

▶ maximizing \( E_\alpha [\log P_{\gamma,\theta}(Y, Z)] \) with respect to \( \gamma, \theta \) (“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 $\alpha$, 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

$$J(\gamma, \theta, \alpha) = \sum_{i<j}^{K} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \alpha_{ik} \alpha_{j\ell} \log \pi_{d_{ij}; kl}(\theta)$$

$$+ \sum_{i=1}^{n} \sum_{C}^{C} \alpha_{ik} (\log \gamma_k - \log \alpha_{ik}).$$

We may define a minorizing function as follows:

$$Q(\alpha, \gamma, \theta, \alpha(t)) = \sum_{i<j}^{K} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \left( \alpha_{ik}^{2} \frac{\alpha_{j\ell}(t)}{2\alpha_{ik}(t)} + \alpha_{i\ell}^{2} \frac{\alpha_{ik}(t)}{2\alpha_{ik}(t)} \right) \log \pi_{d_{ij}; kl}(\theta)$$

$$+ \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{ik} \left( \log \gamma_k - \log \alpha_{ik}(t) - \frac{\alpha_{ik}}{\alpha_{ik}(t)} + 1 \right).$$

- Can be maximized (in $\alpha$) using quadratic programming.
The parsimonious model for the Epinions dataset

\[ P_\theta(D_{ij} = d_{ij} \mid Z_i = k, Z_j = l) \propto \exp\{\theta^- (y^-_{ij} + y^-_{ji}) + \theta^\Delta_k y_{ji} + \theta^\Delta_l y_{ij} + \theta^-^- y^-_{ij} y^-_{ji} + \theta^++ y^+_ij y^+_ji\} \]

where \( y^-_{ij} = I\{Y_{ij} = -1\} \) and \( y^+_ij = I\{Y_{ij} = +1\} \).

Dyad \( D_{ij} \), directed case:

- \( \theta^- \): Overall tendency toward distrust
- \( \theta^-^\Delta \): Category-specific trustedness
- \( \theta^-^- \): lex talionis tendency (eye for an eye)
- \( \theta^++ \): quid pro quo tendency (one good turn...)
Parameter estimates themselves are of interest

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Estimate</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative edges ($\theta^-$)</td>
<td>$-24.020$</td>
<td>($-24.029$, $-24.012$)</td>
</tr>
<tr>
<td>Positive edges ($\theta^+$)</td>
<td>$0$</td>
<td>—</td>
</tr>
<tr>
<td>Negative reciprocity ($\theta^{--}$)</td>
<td>$8.660$</td>
<td>($8.614$, $8.699$)</td>
</tr>
<tr>
<td>Positive reciprocity ($\theta^{++}$)</td>
<td>$9.899$</td>
<td>($9.891$, $9.907$)</td>
</tr>
<tr>
<td>Cluster 1 Trustworthiness ($\theta_{1}^\Delta$)</td>
<td>$-6.256$</td>
<td>($-6.260$, $-6.251$)</td>
</tr>
<tr>
<td>Cluster 2 Trustworthiness ($\theta_{2}^\Delta$)</td>
<td>$-7.658$</td>
<td>($-7.662$, $-7.653$)</td>
</tr>
<tr>
<td>Cluster 3 Trustworthiness ($\theta_{3}^\Delta$)</td>
<td>$-9.343$</td>
<td>($-9.348$, $-9.337$)</td>
</tr>
<tr>
<td>Cluster 4 Trustworthiness ($\theta_{4}^\Delta$)</td>
<td>$-11.914$</td>
<td>($-11.919$, $-11.908$)</td>
</tr>
<tr>
<td>Cluster 5 Trustworthiness ($\theta_{5}^\Delta$)</td>
<td>$-15.212$</td>
<td>($-15.225$, $-15.200$)</td>
</tr>
</tbody>
</table>

95% Confidence intervals based on parametric bootstrap using 500 simulated networks.
Multiple starting points converge to the same solution

Trace plots from 100 different randomly selected starting parameter values:

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.)
Different algorithms can give different results

Values of $J(\gamma, \theta, \alpha)$ as a function of algorithm time for fixed-point (FP) and MM implementations of variational EM using 100 randomly selected starting points:
This work extends the clustering on networks 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.

Recent extensions are myriad (e.g., netclust package, time-varying models).
A few big-picture thoughts

- A statistical point of view on some prediction problems brings more than ability to predict.
- “An approximate answer to the right question is better than an exact answer to the wrong question.”
- Theoretical results are not the same as computed results.
- Algorithms (e.g., for optimization) should be in a statistician’s toolkit.
- Statisticians should advocate for both theoretical and practical reproducibility.
References