Modeling homophily and stochastic equivalence in symmetric relational data

Peter D. Hoff

September 17, 2014

Symmetric relational data

This article discusses a latent variable model for inference and prediction of symmetric relational data.

$$\{y_{i,j} : 1 \leq i < j \leq n\}$$

Symmetric relational data

This article discusses a latent variable model for inference and prediction of symmetric relational data.

$$\{y_{i,j} : 1 \leq i < j \leq n\}$$

Data measured on pairs of a set of n objects or nodes, such as

- friendships among people
- associations among words
- interactions among proteins

Symmetric relational data

This article discusses a latent variable model for inference and prediction of symmetric relational data.

$$\{y_{i,j} : 1 \leq i < j \leq n\}$$

Data measured on pairs of a set of n objects or nodes, such as

- friendships among people
- associations among words
- interactions among proteins

Sociomatrix $Y = [y_{i,j}]_{n \times n}$ is a symmetric matrix with an undefined diagonal.
Observed explanatory variables is represented similarly

$$X = \{x_{i,j}, 1 \leq i < j \leq n\}$$

Symmetric relational data

This article discusses a latent variable model for inference and prediction of symmetric relational data.

$$\{y_{i,j} : 1 \leq i < j \leq n\}$$

Data measured on pairs of a set of n objects or nodes, such as

- friendships among people
- associations among words
- interactions among proteins

Sociomatrix $Y = [y_{i,j}]_{n \times n}$ is a symmetric matrix with an undefined diagonal.
Observed explanatory variables is represented similarly

$$X = \{x_{i,j}, 1 \leq i < j \leq n\}$$

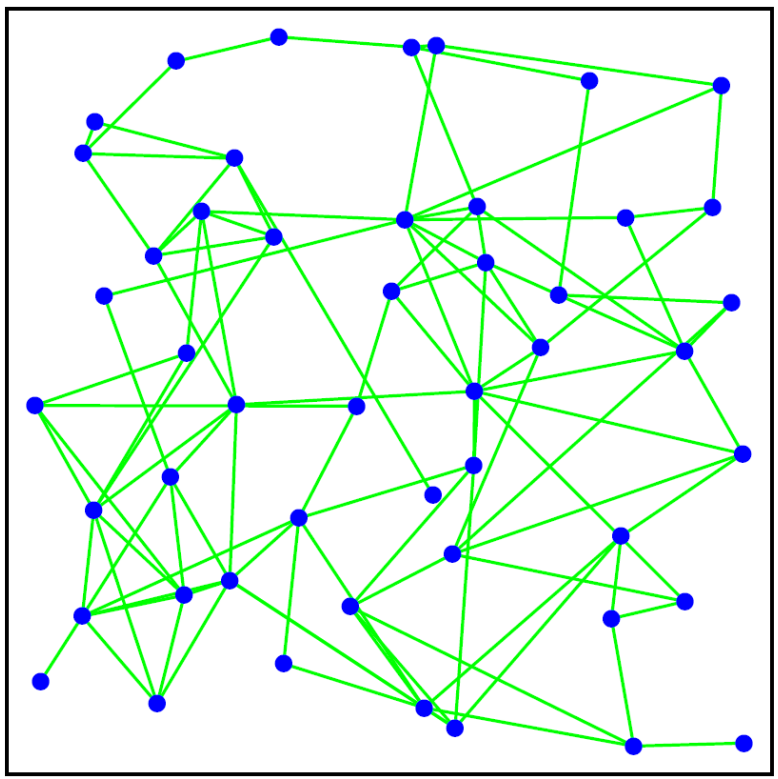
To discover the covariation of Y and X

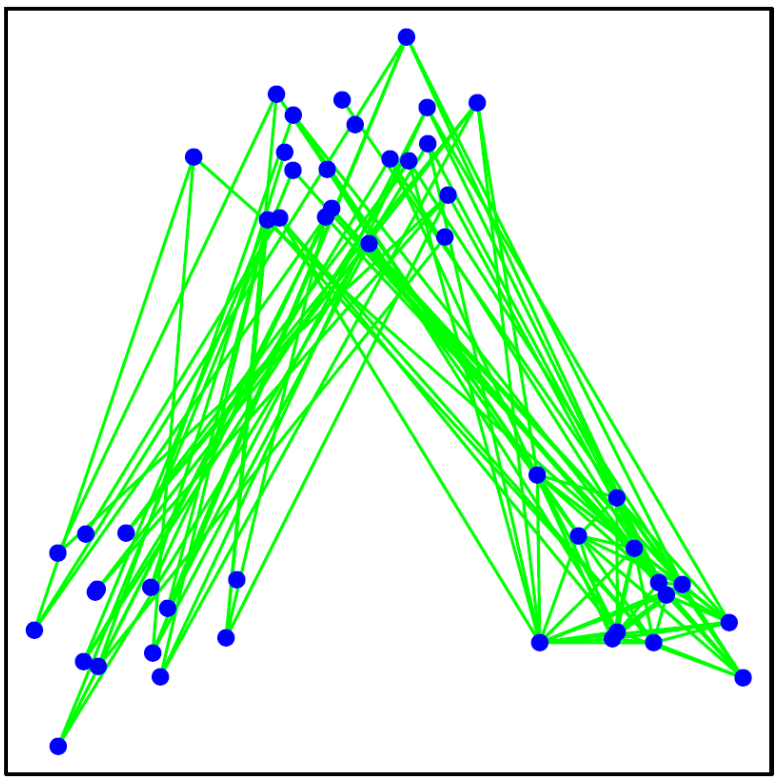
- linear predictor $\beta^T x_{i,j}$,
- node-specific latent variables $u_i, u_j \in \{u_1, \ldots, u_n\}$,

Homophily and stochastic equivalence

Homophily: the relationships between nodes with similar characteristics are stronger than the relationships between nodes having different characteristics.

Stochastic equivalence: the nodes can be divided into groups such that members of the same group have similar patterns of relationships.





Cont'd

Hopmophily: by latent distance model

- transitivity: a friend of a friend is a friend
- balance: the enemy of my friend is an enemy
- existence of cohesive subgroups of nodes
- conditional means of $y_{i,j}$ is a function of $\beta' x_{i,j} - |u_i - u_j|$
- strong relationship between i and j suggests small $|u_i - u_j|$
- further implies $|u_i - u_k| \approx |u_j - u_k|$, i.e., nodes i and j are assumed to have similar relationships to other nodes

Cont'd

Hopmophily: by latent distance model

- transitivity: a friend of a friend is a friend
- balance: the enemy of my friend is an enemy
- existence of cohesive subgroups of nodes
- conditional means of $y_{i,j}$ is a function of $\beta' x_{i,j} - |u_i - u_j|$
- strong relationship between i and j suggests small $|u_i - u_j|$
- further implies $|u_i - u_k| \approx |u_j - u_k|$, i.e., nodes i and j are assumed to have similar relationships to other nodes

Stochastic Equivalence: by latent class model

- would require a large number of classes,
- none of which would be particularly cohesive or distinguishable from others

Issue

Two primary features of interest in social network and relational data analysis

- **classes of nodes with similar roles**: identified by latent class model
- **locational properties of the nodes**: identified by latent distance model

Issue

Two primary features of interest in social network and relational data analysis

- **classes of nodes with similar roles**: identified by latent class model
- **locational properties of the nodes**: identified by latent distance model

However,

- many real netwroks exhibit combinatins of structural equivalence and homophily in varying degrees
- use of either the latent class or distance model would only be representing part of the network structure

Issue

Two primary features of interest in social network and relational data analysis

- **classes of nodes with similar roles**: identified by latent class model
- **locational properties of the nodes**: identified by latent distance model

However,

- many real netwroks exhibit combinatins of structural equivalence and homophily in varying degrees
- use of either the latent class or distance model would only be representing part of the network structure

Latent eigenmodel:

$$\beta' x_{i,j} + u_i^T \Lambda u_j$$

$\{u_1, \ldots, u_n\}$ are node-specific factors and Λ is a diagonal matrix

Justification of latent variable modeling

- For undirected data, nodes as suggested exchangeable
- For any permulation π of the integers $\{1, \ldots, n\}$ and any set of sociomatrices A

$$Pr(\{y_{i,j} : 1 \leq i < j \leq n\} \in A) = Pr(\{y_{\pi i, \pi j} : 1 \leq i < j \leq n\} \in A)$$

Justification of latent variable modeling

- For undirected data, nodes as suggested exchangeable
- For any permulation π of the integers $\{1, \ldots, n\}$ and any set of sociomatrices A

$$Pr(\{y_{i,j} : 1 \leq i < j \leq n\} \in A) = Pr(\{y_{\pi i, \pi j} : 1 \leq i < j \leq n\} \in A)$$

- if a model satisfies the above exchangeability condition for each integer n,
- a latent variable model can be written in the form of

$$y_{i,j} = h(\mu, u_i, u_j, \epsilon_{i,j})$$

i.i.d. latent variables $\{u_1, \ldots, u_n\}$,
i.i.d. pair-specific effects $\{\epsilon_{i,j} : 1 \leq i < j \leq n\}$,
h that is symmetric in its second and third arguments

A general Probit model for binary network data

Different choice of function h lead to different models for y

$$\{\epsilon_{i,j} : 1 \leq i < j \leq n\} \sim i.i.d.\ N(0,1)$$
$$\{u_1, \ldots, u_n\} \sim i.i.d.\ f(u|\psi)$$
$$y_{i,j} = h(\mu, u_i, u_j, \epsilon_{i,j}) = \delta_{(0,\infty)}(\mu + \alpha(u_i, u_j) + \epsilon_{i,j})$$

μ and ψ are parameters to be estimated
α is a symmetric function

A general Probit model for binary network data

Different choice of function h lead to different models for y

$$\{\epsilon_{i,j} : 1 \leq i < j \leq n\} \sim i.i.d.\ N(0,1)$$
$$\{u_1, \ldots, u_n\} \sim i.i.d.\ f(u|\psi)$$
$$y_{i,j} = h(\mu, u_i, u_j, \epsilon_{i,j}) = \delta_{(0,\infty)}(\mu + \alpha(u_i, u_j) + \epsilon_{i,j})$$

μ and ψ are parameters to be estimated
α is a symmetric function

- By adding a linear predictor $\beta^T x_{i,j}$ to μ to represent covariation between Y and X
- Integrating over $\epsilon_{i,j}$,

$$Pr(y_{i,j} = 1|x_{i,j}, u_i, u_j) = \Phi[\mu + \beta^T x_{i,j} + \alpha(u_i, u_j)]$$

cont'd

Assume $\epsilon_{i,j}$'s independent,

- In the case of binary relational datasets,

$$Pr(y_{i,j} = 1|x_{i,j}, u_i, u_j) \equiv \theta_{i,j} = \Phi[\mu + \beta^T x_{i,j} + \alpha(u_i, u_j)]$$

$$Pr(Y|X, u_1, \ldots, u_n) = \prod_{i<j} \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1-y_{i,j}}$$

cont'd

Assume $\epsilon_{i,j}$'s independent,

- In the case of binary relational datasets,

$$Pr(y_{i,j} = 1|x_{i,j}, u_i, u_j) \equiv \theta_{i,j} = \Phi[\mu + \beta^T x_{i,j} + \alpha(u_i, u_j)]$$

$$Pr(Y|X, u_1, \ldots, u_n) = \prod_{i<j} \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1-y_{i,j}}$$

- In the case of relational datasets have ordinal, non-binary measurements,

$$Pr(y_{i,j} = y|x_{i,j}, u_i, u_j) \equiv \theta_{i,j}^{(y)}$$

$$= \Phi[\mu_y + \beta^T x_{i,j} + \alpha(u_i, u_j)] - \Phi[\mu_{y-1} + \beta^T x_{i,j} + \alpha(u_i, u_j)$$

$$Pr(Y|X, u_1, \ldots, u_n) = \prod_{i<j} \theta_{i,j}^{(y_{i,j})}$$

where $\{\mu_y\}$ are parameters to be estimated for all but the lowest value y in the sample space.

Effects of nodal variation

Latent calss model:

$$\alpha(u_i, u_j) = m_{u_i,u_j}$$
$$u_i \in \{1, \ldots, K\}, i \in \{1, \ldots, n\}$$
$$M \text{ a } K \times K \text{ symmetric matrix}$$

Latent distance model:

$$\alpha(u_i, u_j) = -|u_i - u_j|$$
$$u_i \in \mathbb{R}^K, i \in \{1, \ldots, n\}$$

Latent eigenmodel:

$$\alpha(u_i, u_j) = u_i^T \Lambda u_j$$
$$u_i \in \mathbb{R}^K, i \in \{1, \ldots, n\}$$
$$\Lambda \text{ a } K \times K \text{ diagonal matrix}$$

Interpredation of the latent eigenmodel

Latent eigenmodel:

$$\beta' x_{i,j} + u_i^T \Lambda u_j$$

$\{u_1, \ldots, u_n\}$ are node-specific factors and Λ is a diagonal matrix

- Each node i has a vector of unobserved characteristics $u_i = \{u_{i,1}, \ldots, u_{i,K}\}$,

Interpredation of the latent eigenmodel

Latent eigenmodel:

$$\beta' x_{i,j} + u_i^T \Lambda u_j$$

$\{u_1, \ldots, u_n\}$ are node-specific factors and Λ is a diagonal matrix

- Each node i has a vector of unobserved characteristics $u_i = \{u_{i,1}, \ldots, u_{i,K}\}$,
- Similar values of $u_{i,k}$ and $u_{j,k}$ will contribute positively or negatively to the relationship between i and j, depending on whether $\lambda_k > 0$ or $\lambda_k < 0$.

Interpredation of the latent eigenmodel

Latent eigenmodel:

$$\beta' x_{i,j} + u_i^T \Lambda u_j$$

$\{u_1, \ldots, u_n\}$ are node-specific factors and Λ is a diagonal matrix

- Each node i has a vector of unobserved characteristics $u_i = \{u_{i,1}, \ldots, u_{i,K}\}$,
- Similar values of $u_{i,k}$ and $u_{j,k}$ will contribute positively or negatively to the relationship between i and j, depending on whether $\lambda_k > 0$ or $\lambda_k < 0$.
- the model can represent both positive or negative homophily in varying degrees,

Interpredation of the latent eigenmodel

Latent eigenmodel:

$$\beta' x_{i,j} + u_i^T \Lambda u_j$$

$\{u_1, \ldots, u_n\}$ are node-specific factors and Λ is a diagonal matrix

- Each node i has a vector of unobserved characteristics $u_i = \{u_{i,1}, \ldots, u_{i,K}\}$,
- Similar values of $u_{i,k}$ and $u_{j,k}$ will contribute positively or negatively to the relationship between i and j, depending on whether $\lambda_k > 0$ or $\lambda_k < 0$.
- the model can represent both positive or negative homophily in varying degrees,
- can also represent that stochastically equivalent nodes may or may not have strong relationship with one another.

Generalization

let S_n be the set of $n \times n$ sociomatrices, and let

$$\mathcal{C}_K = \{C \in S_n : c_{i,j} = m_{u_i,u_j}, u_i \in \{1, \ldots, K\}, M_{K\times K}\}$$
$$\mathcal{D}_K = \{D \in S_n : d_{i,j} = -|u_i - u_j|, u_i \in \mathbb{R}^K\}$$
$$\mathcal{E}_K = \{E \in S_n : e_{i,j} = u_i^T \Lambda u_j, u_i \in \mathbb{R}^K, \Lambda_{K\times K}\}$$

- Latent eigenmodel generalizes the latent class model
- Latent eigenmodel weakly generalizes the latent distance model
- Latent distance model cannot generalizes the latent eigenmodel

Parameter estimation by MCMC

- Include an additional latent variable $z_{i,j}$,

$$z_{i,j} \sim \mathcal{N}(\beta' x_{i,j} + \alpha(u_i, u_j))$$

$y_{i,j} = y$, if $\mu_y < z_{i,j} < \mu_{y+1}$

- Using conjugate prior distributions where possible, the MCMC algorithms proceed by generating a new state

$$\phi^{(s+1)} = \{Z^{(s+1)}, \mu^{(s+1)}, \beta^{(s+1)}, u_1^{(s+1)}, \ldots, u_n^{(s+1)}\}$$

from a current state $\phi^{(s)}$

Model comparison on three different datasets

Model comparison on three different datasets

Three data sets:

- Adolescent health social network - homophily
- Word neighbors in Genesis - stochastic equivalence
- Protein-protein interaction data - both

Model comparison on three different datasets

Three data sets:

- Adolescent health social network - homophily
- Word neighbors in Genesis - stochastic equivalence
- Protein-protein interaction data - both

Five-fold cross validation for each combination of data, dimension ($K \in \{3, 5, 10\}$) and model

Model comparison on three different datasets

Three data sets:

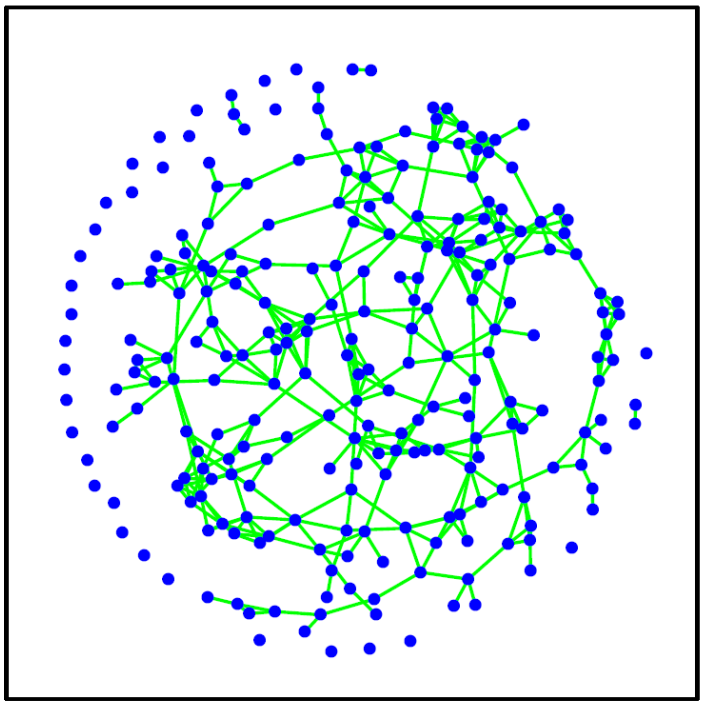
- Adolescent health social network - homophily
- Word neighbors in Genesis - stochastic equivalence
- Protein-protein interaction data - both

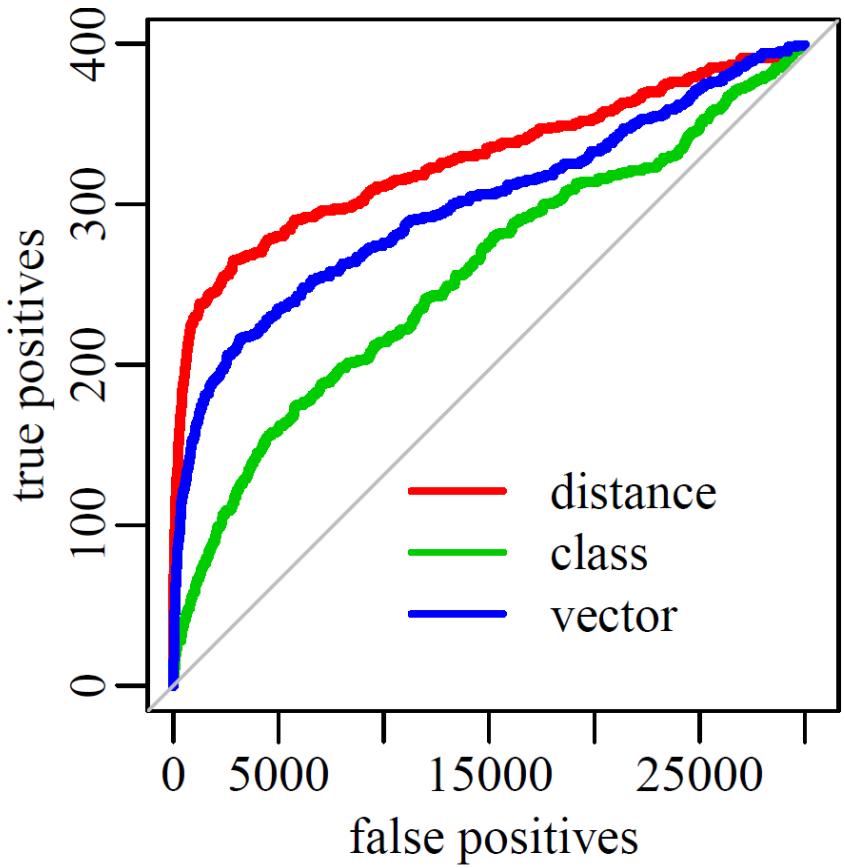
Five-fold cross validation for each combination of data, dimension ($K \in \{3, 5, 10\}$) and model

Table 1: Cross validation results and area under the ROC curves.

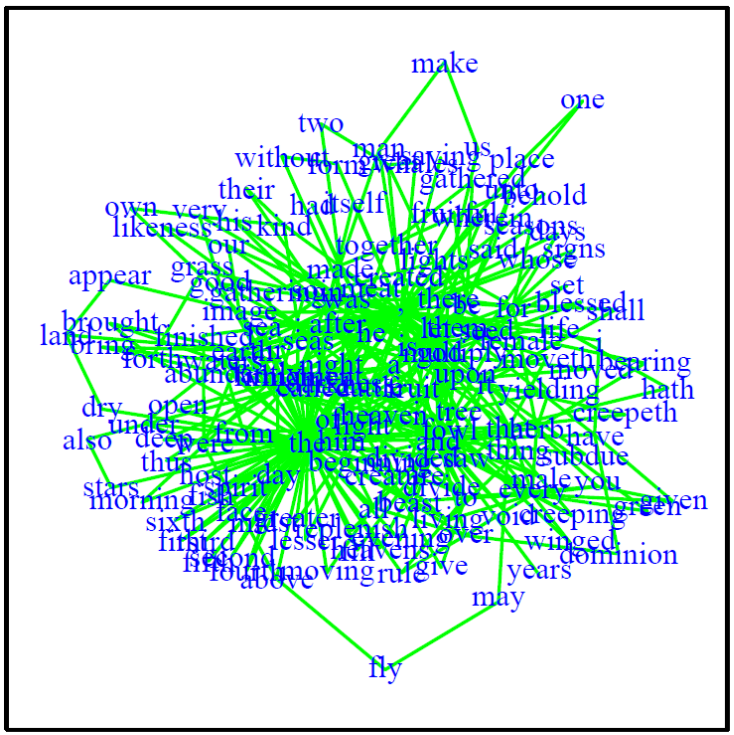
K	**Add health**			**Genesis**			**Protein interaction**		
	dist	class	eigen	dist	class	eigen	dist	class	eigen
3	0.82	0.64	0.75	0.62	0.82	0.82	0.83	0.79	0.88
5	0.81	0.70	0.78	0.66	0.82	0.82	0.84	0.84	0.90
10	0.76	0.69	0.80	0.74	0.82	0.82	0.85	0.86	0.90

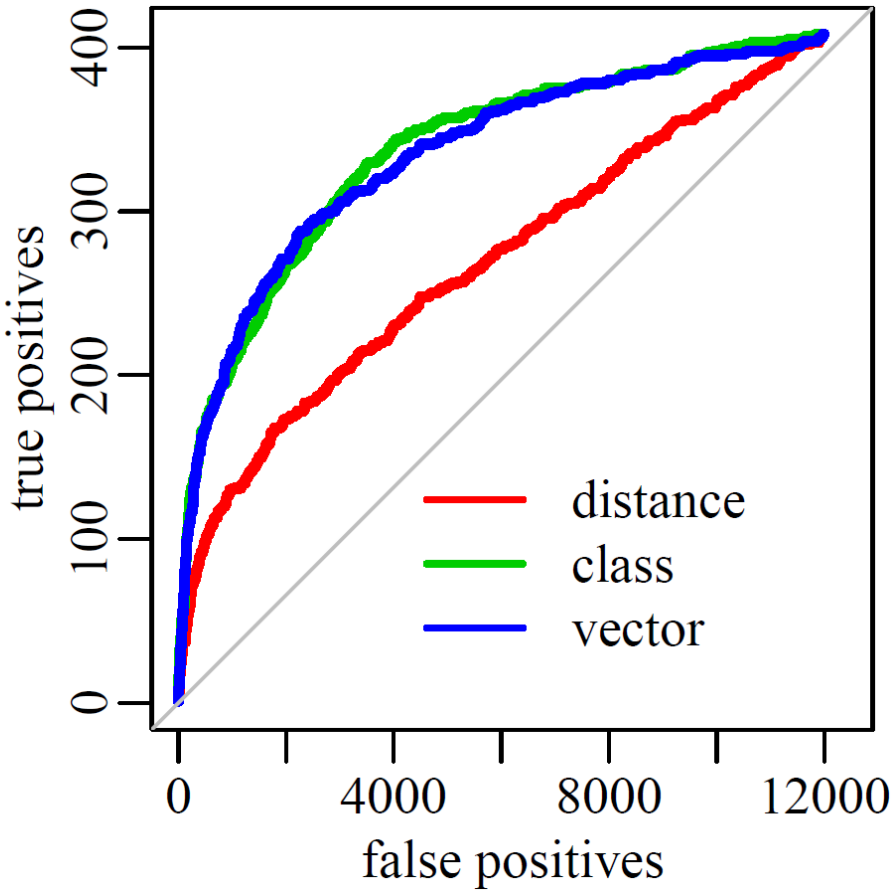
Adolescent Health social network



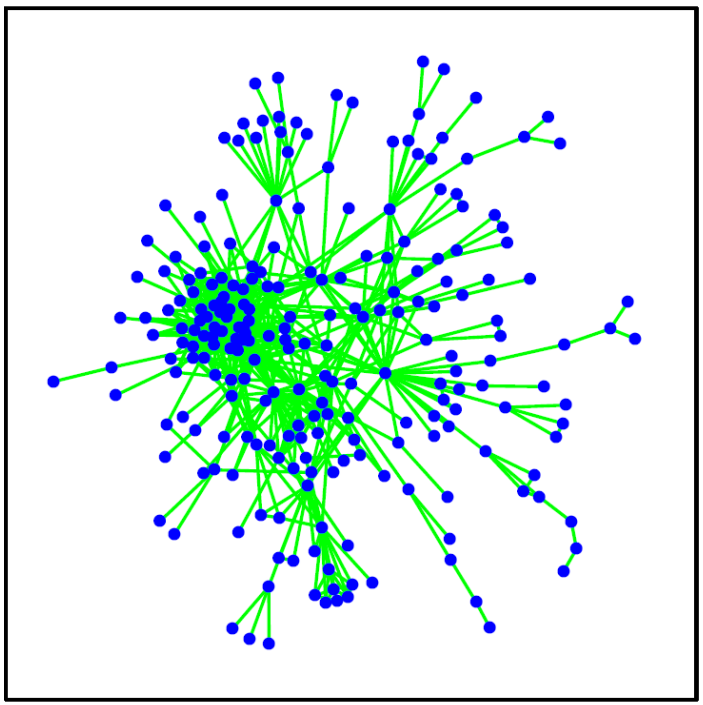


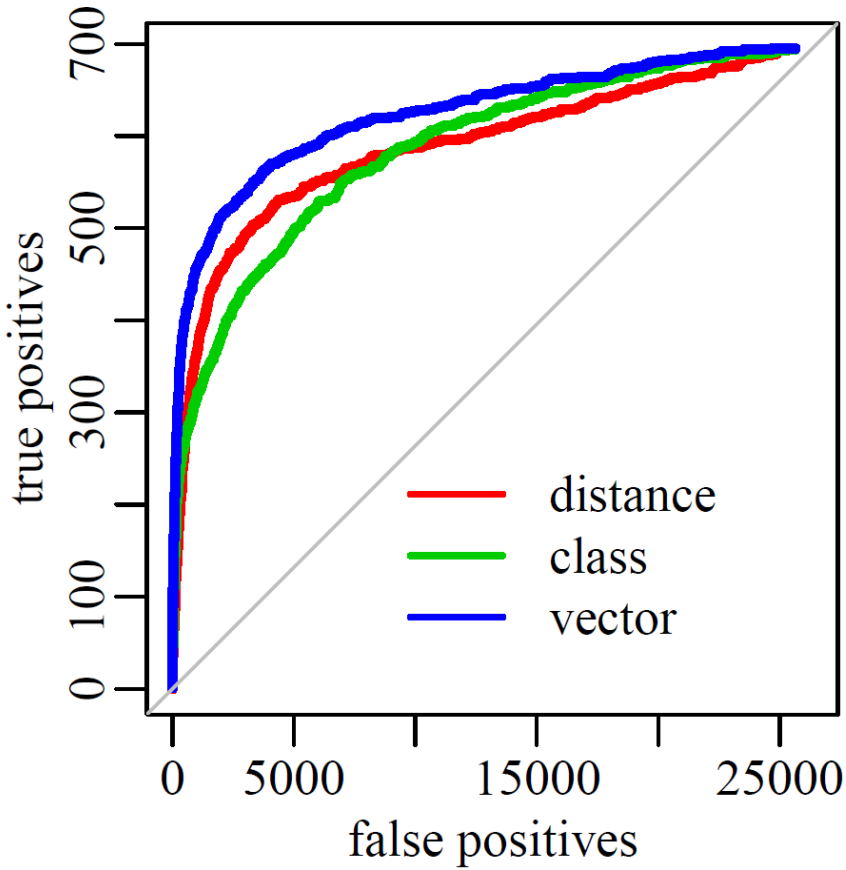
Word neighbors in Genesis





Protein-protein interaction data





Conclusion

- Latent distance and latent class models provide concise, easily interpreted descriptions of social networks and relational data

Conclusion

- Latent distance and latent class models provide concise, easily interpreted descriptions of social networks and relational data
- However, neither of these models will provide a complete picture of relational data that exhibit degrees of both homophily and stochastic equivalence

Conclusion

- Latent distance and latent class models provide concise, easily interpreted descriptions of social networks and relational data
- However, neither of these models will provide a complete picture of relational data that exhibit degrees of both homophily and stochastic equivalence
- In contrast, latent eigenmodel is able to represent datasets with either or both of these data patterns

Appendix

Eigenmodel generalizes the class and distance model

Eigenmodel generalizes the class and distance model

To show this statement, let S_n be the set of $n \times n$ sociomatrices, and let

$$\mathcal{C}_K = \{C \in S_n : c_{i,j} = m_{u_i,u_j}, u_i \in \{1, \ldots, K\}, \\ M \; is \; a \; K \times K \; symmetricmatrix\}$$

$$\mathcal{D}_K = \{D \in S_n : d_{i,j} = -|u_i - u_j|, u_i \in \mathbb{R}^K\}$$

$$\mathcal{E}_K = \{E \in S_n : e_{i,j} = u_i^T \Lambda u_j, u_i \in \mathbb{R}^K, \\ \Lambda \; is \; a \; K \times K \; diagonal \; matrix\}$$

Eigenmodel generalizes the class and distance model

To show this statement, let S_n be the set of $n \times n$ sociomatrices, and let

$$\mathcal{C}_K = \{C \in S_n : c_{i,j} = m_{u_i,u_j}, u_i \in \{1, \ldots, K\}, \\ M \text{ is a } K \times K \text{ symmetricmatrix}\}$$

$$\mathcal{D}_K = \{D \in S_n : d_{i,j} = -|u_i - u_j|, u_i \in \mathbb{R}^K\}$$

$$\mathcal{E}_K = \{E \in S_n : e_{i,j} = u_i^T \Lambda u_j, u_i \in \mathbb{R}^K, \\ \Lambda \text{ is a } K \times K \text{ diagonal matrix}\}$$

In other words,

- $\mathcal{C}_K$ is the set of possible values of $\{\alpha(u_i, u_j), 1 \leq i < j \leq n\}$ under a K-dimensional latent class model,
- similarly for $\mathcal{D}_K$ and $\mathcal{E}_K$

$\mathcal{E}_K$ generalizes $\mathcal{C}_K$:

$\mathcal{E}_K$ generalizes $\mathcal{C}_K$:

- Let $C \in \mathcal{C}_K$
- let $\tilde{C}$ be a completion of C obtained by setting $c_{i,j} = m_{u_i,u_i}$
- At most K unique rows of $\tilde{C}$ and so $\tilde{C}$ is of rank K at most

$\mathcal{E}_K$ generalizes $\mathcal{C}_K$:

- Let $C \in \mathcal{C}_K$
- let $\tilde{C}$ be a completion of C obtained by setting $c_{i,j} = m_{u_i,u_i}$
- At most K unique rows of $\tilde{C}$ and so $\tilde{C}$ is of rank K at most
- Since $\mathcal{E}_K$ contains all sociomatrices that can be completed as a rank-K matrix,
  we have $\mathcal{C}_K \subset \mathcal{E}_K$
- Since $\mathcal{E}_K$ includes matrices with n unique rows,
  $K \subset \mathcal{E}$ unless $K \geq n$ in which case the two sets are equal

$\mathcal{E}_{K+1}$ weakly generalizes $\mathcal{D}_K$:

- Let $D \in D_K$, generally be of full rank
  for $K < n$, it cannot be represented exactly by an $E \in \mathcal{E}_K$

$\mathcal{E}_{K+1}$ weakly generalizes $\mathcal{D}_K$:

- Let $D \in D_K$, generally be of full rank
  for $K < n$, it cannot be represented exactly by an $E \in \mathcal{E}_K$
- What is critical from a modeling perspective is whether or not the order of the entries of each D can be matched by the order of the entries of an E.
  This is becasue $\{\mu_y : y \in \mathcal{Y}\}$ which can be adjusted to accommodate monotone transformations of $\alpha(u_i, u_j)$

$\mathcal{E}_{K+1}$ weakly generalizes $\mathcal{D}_K$:

- Let $D \in D_K$, generally be of full rank
  for $K < n$, it cannot be represented exactly by an $E \in \mathcal{E}_K$
- What is critical from a modeling perspective is whether or not the order of the entries of each D can be matched by the order of the entries of an E.
  This is becasue $\{\mu_y : y \in \mathcal{Y}\}$ which can be adjusted to accommodate monotone transformations of $\alpha(u_i, u_j)$
- Note that the matrix of squared distances among a set of K-dimensional vectors $\{z_1, \ldots, z_n\}$ is a monotonic transformation of the distances,
  is a of rank $K + 2$ or less, (since
  $D^2 = [z_1'z_1, \ldots, z_n]^T 1^T + 1[z_1'z_1, \ldots, z_n'z_n] - 2ZZ^T$)
  so is in $\mathcal{E}_{K+2}$.

cont'd

- $u_i = (z_i, \sqrt{r^2 - z_i^T z_i}) \in \mathbb{R}^{K+1} \forall i \in \{1, \ldots, n\}$,
  $u_i' u_j = z_i' z_j + \sqrt{(r^2 - |z_i|^2)(r^2 - |z_j|^2)}$
- For large r, this is approximately $r^2 - |z_i - z_j|^2/2$, which is an increasing function of the negative distance $d_{i,j}$
- For large enough r the numerical order of the entries of this $E \in \mathcal{E}_{K+1}$ is the same as that of $D \in \mathcal{D}_K$.

$\mathcal{D}_K$ does not weakly generalize $\mathcal{E}_1$:

- Consider $E \in \mathcal{E}_1$ generated by $\Lambda = 1, u_1 = 1$ and $u_i = r < 1$ for $i > 1$
  $r = e_{1,i_1} = e_{1,i_2} > e_{i_1,i_2} = r^2, \forall i_1, i_2 \neq 1.$
- For which K is such an ordering of the elements of $D \in \mathcal{D}_K$ possible?
  If $K = 1$ then such an ordering is possible only if $n = 3$.
  If $K = 2$ such an ordering is possible for $n \leq 6$. (kissing spheres)
- Increasing n increases the necessary dimension of the Euclidean space, and so for any K there are n and $E \in \mathcal{E}_1$ that have entry orderings that cannot be matched by those of any $D \in \mathcal{D}_K$

MCMC cont'd

1. $\forall \{i, j\}$, sample $z_{i,j}$
2. $\forall y \in \mathcal{Y}$, sample μ_y
3. Sample β
4. Sample $u_1, \ldots, u_n$ and their associated parameters

MCMC cont'd

1. $\forall\{i, j\}$, sample $z_{i,j}$
2. $\forall y \in \mathcal{Y}$, sample μ_y
3. Sample β
4. Sample $u_1, \ldots, u_n$ and their associated parameters
   - For latent distance model
     - propose and accept or reject new values of u_i with Metropolis algorithms
     - sample the population variances of u_i from (inverse-gamma) full conditional distributions

MCMC cont'd

1. $\forall\{i, j\}$, sample $z_{i,j}$
2. $\forall y \in \mathcal{Y}$, sample μ_y
3. Sample β
4. Sample $u_1, \ldots, u_n$ and their associated parameters
   - For latent distance model
     - propose and accept or reject new values of u_i with Metropolis algorithms
     - sample the population variances of u_i from (inverse-gamma) full conditional distributions
   - For latent class model
     - Update u_i from multinomial conditional distribution given current $Z, \{u_j : j \neq i\}$ and the variance of the elements of M
     - sample the elements of M from their normal full conditional distributions and the variance of the entries of M from its (inverse-gamma) full conditional distribution

MCMC cont'd

1. $\forall\{i, j\}$, sample $z_{i,j}$
2. $\forall y \in \mathcal{Y}$, sample μ_y
3. Sample β
4. Sample $u_1, \ldots, u_n$ and their associated parameters
   - For latent distance model
     - propose and accept or reject new values of u_i with Metropolis algorithms
     - sample the population variances of u_i from (inverse-gamma) full conditional distributions
   - For latent class model
     - Update u_i from multinomial conditional distribution given current $Z, \{u_j : j \neq i\}$ and the variance of the elements of M
     - sample the elements of M from their normal full conditional distributions and the variance of the entries of M from its (inverse-gamma) full conditional distribution
   - for latent vector model
     - sample u_i from its multivariate normal full conditional distribution
     - sample the mean of u_i's from their normal full conditional distributions

Model comparison on three different datasets

Model comparison on three different datasets

Three data sets:

- Adolescent health social network - homophily
- word neighbors in genesis - stochastic equivalence
- protein-protein interaction data - both

Model comparison on three different datasets

Three data sets:

- Adolescent health social network - homophily
- word neighbors in genesis - stochastic equivalence
- protein-protein interaction data - both

Five-fold cross validation for each combination of data, dimension ($K \in \{3, 5, 10\}$) and model

- Randomly divide the $\binom{n}{2}$ data values into 5 sets of roughly equal size, letting $s_{i,j}$ be the set to which pair $\{i, j\}$ is assigned.
- For each $s \in \{1, \ldots, 5\}$:
  - obtain posterior distributions of the model parameter codnitional on $\{y_{i,j} : s_{i,j} \neq s\}$, the data on pairs not in set s.
  - for pairs $\{k, l\}$ in set s, let $\hat{y}_{k,l} = E(y_{k,l} | \{y_{i,j} : s_{i,j} \neq s\})$, the posterior predictive mean of $y_{k,l}$ obtained using data not in set s.