# Network Inference From Grouped Observations Using Hub Models

YUNPENG ZHAO[1] AND CHARLES WEKO[2]

*George Mason University*[1] *and United States Army*[2]

**Supplementary Material**

## S1 Identifiability

In Section 3.2, we introduced the notion of identifiability. For Hub Models, this means for any two sets of parameters $\{A, \rho\}$ and $\{A^*, \rho^*\}$:

$$\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*) \ \forall g \implies A = A^*, \rho = \rho^*. \tag{S1.1}$$

In this section, we provide a simple counterexample to show that Hub Models are not identifiable if the adjacency matrix is unconstrained and prove Theorem 1.

## S1.1 Counterexample

Consider a network of size $n = 4$ with the parameters defined in Table 1. Note that for the asymmetric case, $n = 4$ is the smallest population size where the number of possible groups exceeds the number of parameters to estimate. In this example, nodes $v_1$ and $v_2$ both always produce the same group while nodes $v_3$ and $v_4$ produce a different group.

| | | $A_{ij}$ | | | |
|---|---|---|---|---|---|
| | | $j$ | | | |
| $\rho_i$ | $i$ | 1 | 2 | 3 | 4 |
| 0.25 | 1 | 1 | 1 | 1 | 0 |
| 0.25 | 2 | 1 | 1 | 1 | 0 |
| 0.25 | 3 | 0 | 1 | 1 | 1 |
| 0.25 | 4 | 0 | 1 | 1 | 1 |

Table 1: Example of a Set of Parameters Which are Not Identifiable

The probability of $G$ has the form:

$$\mathbb{P}(G = g|A, \rho) = \begin{cases} \frac{1}{2} & g = \{1, 1, 1, 0\}, \\ \frac{1}{2} & g = \{0, 1, 1, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

There are an infinite number of parameters yielding the same distribution, but a simple alternative is as follows. Let $\rho^* = (0.5, 0, 0.5, 0)$, leave the first and third rows of $A$ unchanged but let all other components of $A^*$ assume arbitrary values. Obviously, we have $\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*)$

for all $g$. This counterexample demonstrates that the model requires an additional condition to be identifiable.

## S1.2 Theorem Proof

Theorem 1 states that symmetry of the adjacency matrix is a sufficient condition for identifiability.

**Theorem 1.** Let $A$ and $A^*$ be symmetric adjacency matrices with $A_{ii} = A^*_{ii} = 1$ for all $i$, $A_{ij} < 1$ and $A^*_{ij} < 1$ for all $i \neq j$. If $\mathbb{P}(g|A, \rho) = \mathbb{P}(g|A^*, \rho^*)$ for all $g$, then $\{A, \rho\} = \{A^*, \rho^*\}$.

Let $g^x$ and $g^y$ denote the singleton groups which consist only of nodes $v_x$ and $v_y$, respectively. Further, let $g^{xy}$ denote the group representing the pair of $v_x$ and $v_y$.

From (3.1) the probability of the singletons is:

$$\mathbb{P}(g^x|A, \rho) = \rho_x(1 - A_{xy}) \prod_{j \neq \{x,y\}} (1 - A_{xj}) \tag{S1.2}$$

$$\mathbb{P}(g^y|A, \rho) = \rho_y(1 - A_{xy}) \prod_{j \neq \{x,y\}} (1 - A_{yj}). \tag{S1.3}$$

In (S1.3) we have taken advantage of the symmetry of $A$ to replace $A_{yx}$ with $A_{xy}$.

Now, we consider the probability of $g^{xy}$.

$$\begin{aligned}\mathbb{P}(g^{xy}|A,\rho) &= \rho_x A_{xy}\prod_{j\neq\{x,y\}}(1-A_{xj}) + \rho_y A_{xy}\prod_{j\neq\{x,y\}}(1-A_{yj}) \\ &= A_{xy}\Big[\rho_x\prod_{j\neq\{x,y\}}(1-A_{xj}) + \rho_y\prod_{j\neq\{x,y\}}(1-A_{yj})\Big] \\ &= A_{xy}\Big[\frac{\mathbb{P}(G=g^x|A,\rho)}{(1-A_{xy})} + \frac{\mathbb{P}(g^y|A,\rho)}{(1-A_{xy})}\Big] \\ &= \frac{A_{xy}}{(1-A_{xy})}\Big[\mathbb{P}(g^x|A,\rho)+\mathbb{P}(g^y|A,\rho)\Big], \end{aligned} \tag{S1.4}$$

which implies that:

$$A_{xy} = \frac{\mathbb{P}(g^{xy}|A,\rho)}{\mathbb{P}(g^x|A,\rho)+\mathbb{P}(g^y|A,\rho)+\mathbb{P}(g^{xy}|A,\rho)}. \tag{S1.5}$$

Therefore, $A_{xy} = A^*_{xy}$ for all $x$ and $y$.

To complete the proof, consider an arbitrary node $v_x$ which appears as a singleton represented by $g^x$:

$$\mathbb{P}(g^x|A,\rho) = \rho_x\prod_{j\neq x}(1-A_{xj}). \tag{S1.6}$$

If $A_{xy} = A^*_{xy}$ for all $x$ and $y$ and $\mathbb{P}(g|A,\rho) = \mathbb{P}(g|A^*,\rho^*)$ for all $g$, then:

$$\rho_x\prod_{j\neq x}(1-A_{xj}) = \rho^*_x\prod_{j\neq x}(1-A_{xj}) \tag{S1.7}$$

and it is easy to see that $\rho_x = \rho_x^*$ for all $x$.

□

Equation (S1.5) suggests a method of moments estimator for the adjacency matrix based on frequencies of doubletons and singletons. However, this estimator requires that the probability of doubletons and singletons be estimated accurately. In practice, this technique would be very inefficient, because small groups appear infrequently in many real-world datasets. Thus, we continue to focus on the MLE which presumably uses all available information.

# S2 Estimating Equations

In Section 3.3, we presented estimating equations for Hub Model parameters ((3.5) and (3.6)). Here we derive those equations.

We begin by taking the derivative of (3.2) with respect to $A_{xy}$ and $A_{yx}$.

$$\frac{\partial \Lambda(G|A,\rho)}{\partial A_{xy}} = \frac{\partial \mathcal{L}(G|A,\rho)}{\partial A_{xy}} - \lambda_{xy} = 0 \text{ if } x < y, \tag{S2.1}$$

$$\frac{\partial \Lambda(G|A,\rho)}{\partial A_{yx}} = \frac{\partial \mathcal{L}(G|A,\rho)}{\partial A_{yx}} + \lambda_{xy} = 0 \text{ if } x > y. \tag{S2.2}$$

Therefore,

$$\frac{\partial \mathcal{L}(G|A,\rho)}{\partial A_{xy}} = -\frac{\partial \mathcal{L}(G|A,\rho)}{\partial A_{yx}}. \tag{S2.3}$$

We now focus on the derivative of the log likelihood function of (3.2):

$$\sum_t \frac{\rho_x G_x^{(t)} \Big[ \prod_{j \neq y} A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1 - G_j^{(t)}} \Big] \frac{\partial}{\partial A_{xy}} \Big( A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{1 - G_y^{(t)}} \Big)}{\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1 - G_j^{(t)}}}. \tag{S2.4}$$

Note that the derivative in the numerator of (S2.4) is equal to 1 if node $v_y$ is in observation $G^{(t)}$, and $-1$ if $v_y$ is not in the observation. We represent this by the function:

$$\gamma(G_y^{(t)}) = \begin{cases} 1 & \text{if } G_y^{(t)} = 1, \\ -1 & \text{if } G_y^{(t)} = 0. \end{cases}$$

Therefore,

$$\frac{\partial}{\partial A_{xy}} \mathcal{L}(G|A,\rho) = \sum_t \frac{\Big[ \rho_x G_x^{(t)} \prod_{j \neq y} A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1 - G_j^{(t)}} \Big] \gamma(G_y^{(t)})}{\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1 - G_j^{(t)}}}. \tag{S2.5}$$

The denominator of (S2.5) is simply the probability of $G^{(t)}$ (see (3.1)). In addition, the term in brackets can be made equal to $\mathbb{P}(G^{(t)}, S_x = 1)$ by multiplying $A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{(1 - G_y^{(t)})}$. To conserve space, we suppress $\{A, \rho\}$

going forward. This gives:

$$\frac{\partial}{\partial A_{xy}}\mathcal{L}(G) = \sum_t \frac{\gamma(G_y^{(t)})\mathbb{P}(G^{(t)}, S_x^{(t)} = 1)}{A_{xy}^{G_y^{(t)}}(1 - A_{xy})^{(1-G_y^{(t)})}\mathbb{P}(G^{(t)})}. \tag{S2.6}$$

This equation can be further simplified by observing that $\frac{\mathbb{P}(G^{(t)}, S_x^{(t)}=1)}{\mathbb{P}(G^{(t)})}$ is equivalent to $\mathbb{P}(S_x^{(t)} = 1|G^{(t)})$:

$$\frac{\partial}{\partial A_{xy}}\mathcal{L}(G) = \sum_t \frac{\gamma(G_y^{(t)})\mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{A_{xy}^{G_y^{(t)}}(1 - A_{xy})^{(1-G_y^{(t)})}}. \tag{S2.7}$$

Plugging (S2.7) into (S2.3), we get:

$$\sum_t \frac{\gamma(G_y^{(t)})\mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{A_{xy}^{G_y^{(t)}}(1 - A_{xy})^{(1-G_y^{(t)})}} = -\sum_t \frac{\gamma(G_x^{(t)})\mathbb{P}(S_y^{(t)} = 1|G^{(t)})}{A_{yx}^{G_x^{(t)}}(1 - A_{yx})^{(1-G_x^{(t)})}}. \tag{S2.8}$$

By applying symmetry and breaking the summations, this becomes:

$$\begin{aligned} &\sum_{t:G_y^{(t)}=1} \frac{\mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{A_{xy}} - \sum_{t:G_y^{(t)}=0} \frac{\mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{1 - A_{xy}} \\ &= -\sum_{t:G_x^{(t)}=1} \frac{\mathbb{P}(S_y^{(t)} = 1|G^{(t)})}{A_{xy}} + \sum_{t:G_x^{(t)}=0} \frac{\mathbb{P}(S_y^{(t)} = 1|G^{(t)})}{1 - A_{xy}}. \end{aligned} \tag{S2.9}$$

With some simple algebra, it is easy to see that:

$$\hat{A}_{xy} = \frac{\sum_t G_y^{(t)}\mathbb{P}(S_x = 1|G^{(t)}) + \sum_t G_x^{(t)}\mathbb{P}(S_y = 1|G^{(t)})}{\sum_t \left[\mathbb{P}(S_x = 1|G^{(t)}) + \mathbb{P}(S_y = 1|G^{(t)})\right]}. \tag{S2.10}$$

It is worth repeating that (S2.10) is not a closed form solution for $\hat{A}_{xy}$. This is because the right hand side of the equation depends on $\hat{A}_{xy}$.

We next derive the estimating equation for $\hat{\rho}$. By taking the derivative of (3.3) with respect to $\rho_x$, we get the following:

$$\begin{aligned}\frac{\partial}{\partial \rho_x}\Lambda(G) &= \sum_t \frac{G_x^{(t)} \prod_j A_{xj}^{G_j^{(t)}} (1-A_{xj})^{1-G_j^{(t)}}}{\mathbb{P}(G^{(t)})} - \lambda_o \\ &= \sum_t \frac{\mathbb{P}(G^{(t)}, S_x^{(t)} = 1)}{\rho_x \mathbb{P}(G^{(t)})} - \lambda_o \\ &= \frac{1}{\rho_x} \sum_t \mathbb{P}(S_x^{(t)} = 1 | G^{(t)}) - \lambda_o. \end{aligned} \tag{S2.11}$$

Solving this equation for zero, we obtain:

$$\rho_x = \frac{1}{\lambda_o} \sum_t \mathbb{P}(S_x^{(t)} = 1 | G^{(t)}). \tag{S2.12}$$

Using the constraint on $\rho$, we get:

$$\hat{\rho}_x = \frac{\sum_{t=1}^T \mathbb{P}(S_x^{(t)} = 1 | G^{(t)})}{T}. \tag{S2.13}$$

# S3 Data Analysis

In Section 6, we performed analysis on a dataset from *Dream of the Red Chamber* to show how Hub Models can provide sharper contrast between

the relationships with a population. Here we present analysis of two additional datasets.

The first dataset records co-sponsorship of legislation in the Senate of the $110^{th}$ United States Congress. The rules of the Senate require that each piece of legislation have a unique sponsor; however, other members may co-sponsor the bill (Fowler, 2006a). These rules mean that the data conform to the assumption of the Hub Model.

The second dataset has been extracted from the USDA plant database. Unlike the other two datasets, this one does not deal with "social" data, but with "spatial" data. Each observation represents a single species of plant along with each North American state or territory in which the plant is observed to grow. For this analysis, states and territories represent nodes. As with the *Dream of the Red Chamber* dataset, we find that the Hub Models return meaningful information about underlying structure even when the assumption of a single hub node is not valid.

## S3.1 Senate of the $110^{th}$ United States Congress

The first dataset records co-sponsorship of legislation in the Senate of the $110^{th}$ United States Congress. The rules of the Senate require that each piece of legislation have a single, unique sponsor; however, other members

may co-sponsor the bill (Fowler, 2006a). These rules mean that the data conform to the assumption of the Hub Model.

The United States Senate is a chamber in the bicameral legislature of the United States, and together with the U.S. House of Representatives makes up the U.S. Congress. A key function of both chambers of Congress is to originate legislation. Each piece of legislation can have only one originating sponsor; however, since the mid-1930s, Senators have had an opportunity to express support for a piece of legislation by signing it as a co-sponsor (Fowler, 2006a).

The $110^{th}$ United States Congress occurred between January 3, 2007 and January 3, 2009. The Democratic Party controlled a majority in both chambers for the first time since 1995 with a voting share of 50.5 % of the Senate membership.

The United States Senate consists of 100 members with each state represented by 2 Senators at any time. During this session of Congress, there were a total of 102 individuals who served in the Senate. One original member died and a second resigned to become a lobbyist. Both members were replaced by appointed state representatives.

Data for legislative co-sponsorship are available in the Library of Congress Thomas legislative database. This database includes more than 280,000

pieces of legislation proposed in the U.S. House and Senate with over 2.1 million co-sponsorship signatures. Most bills do not pass, and cosponsors need not invest time and resources crafting legislation; so co-sponsorship is a relatively costless way to signal one's position on issues important to constituents and fellow legislators. For the purposes of this study, we include all forms of legislation including all available resolutions, public and private bills, and amendments (Fowler, 2006b). During the $110^{th}$ Congress, the Senate initiated 10,327 pieces of legislation.

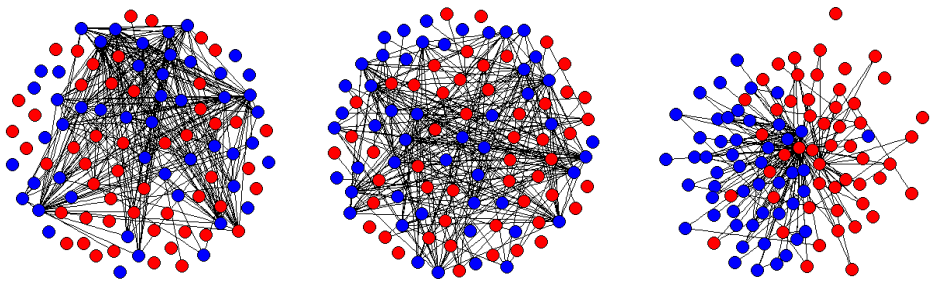
It is a trivial task to apply the KHM to this dataset when we treat the sponsor as known; therefore, we focus on the case where the sponsor is unknown. That is, we intentionally confound sponsors with co-sponsors so that the only data that we have is $G$. We would like to investigate whether the HM can provide a meaningful estimate of the latent social structure even when the information of hub nodes is missing. The average difference between edges estimated by KHM and HM is 0.03, which suggests that the HM estimate is very accurate even when we confound the hub nodes.

In Figure 1, we plot the co-occurrence matrix, half-weight index, and the adjacency matrix of the Hub Model using the force directed graph drawing technique of Fruchterman-Reingold. Each Senator is represented as a node where the color of the node represents the Senator's official political party.

Red nodes represent Republicans while blue nodes represent Democrats.

First, notice that each estimated adjacency matrix produces a different layout because of the different weights estimated for each relationship. For the co-occurrence matrix and the half weight index, the estimated adjacency matrices do not separate Republican and Democrat Senators while the Hub Model produces an adjacency matrix which clusters Senators by political party.

Second, in Figure 1, we plot only the 5% strongest relationships in each estimate. Again, the Hub Models provide insight into the individuals in the population which have key relationships. In Figure 1c, individuals with many strong connections are closer to the center of the layout while in Figures 1a and 1b a number of these individuals are actually present at the perimeter of the layout. We have not presented all relationships because this results in a nearly complete graph.



(a) Co-occurrence (b) Half Weight Index (c) Hub Model

Figure 1: Comparison of Estimation Techniques for the $110^{th}$ Senate

This implies that the HM provides more meaningful information about the community structure of this dataset than classical measures. This is further supported by the normalized cut value corresponding to the Senator's official party membership (Shi and Malik, 2000) .

$$\frac{\sum\limits_{i\in C_1, j\in C_2} A_{ij}}{\sum\limits_{i,j\in C_1} A_{ij}} + \frac{\sum\limits_{i\in C_1, j\in C_2} A_{ij}}{\sum\limits_{i,j\in C_2} A_{ij}}. \tag{S3.14}$$

Equation (S3.14) gives the normalized cut value for two communities $C_1$ and $C_2$. Lower normalized cut values indicate stronger community differentiation. For the Senate data, Table 2 presents the values for each inferred network. The normalized cut value for HM is lower than the co-occurrence matrix and half weight index, which strengthens the visual intuition that the estimates from HM provide better distinction between communities.

| | Normalized Cut Value |
|---|---|
| Co-Occurrence Matrix | 0.837 |
| Half Weight Index | 0.823 |
| Hub Model | 0.757 |

Table 2: Normalized Cut Ratios of $110^{th}$ Senate for Different Inference Techniques

## S3.2 North American Flora

The final dataset has been extracted from the USDA plant database. In the previous examples, we have worked with datasets which are essentially

from the social sciences. However, we believe that Hub Models are useful in other situations where observations are the result of nodes coalescing around a single node or observations are the result of some resource dispersing outward from a single node to multiple nodes.

As a demonstration of how this kind of data can be used to estimate the relationship between different regions, we use a dataset from the University of California Irvine Machine Learning Repository which had been extracted and encoded from the USDA plants database (Hamalainen and Nykanen, 2008). 34,781 plant species or geneses are included in the dataset. For each plant, the dataset indicates in which of 68 areas the plant is found. These areas include all United States states, Canadian provinces and territories, along with the Virgin Islands, Puerto Rico, Greenland, and St. Pierre and Miquelon (islands off the northeast coast of Canada). For simplicity, we will refer to all of these areas as "states".

We would expect that contiguous states would tend to have many flora in common while states which are far apart would be less likely to share common flora. For example, Connecticut and Massachusetts are small states which share a common boarder; therefore, we would expect them to appear together in the regions of many plants. Conversely, California and Greenland are very far apart and at different latitudes; therefore, we would expect

a weak relationship.

Of course, the map of North America is well known and our objective here is not to compare the Hub Model to spacial modeling. Instead, we are using the regions of North America as a proxy for a system of distribution.

To demonstrate the ability of the Hub Model to capture the connections between states, we split them into 13 different regions. The United States is identified by the 9 divisions of the US Census Bureau. The Canadian provinces are identified by three regions. The final region includes islands in the Atlantic ocean which are not included in any other region.

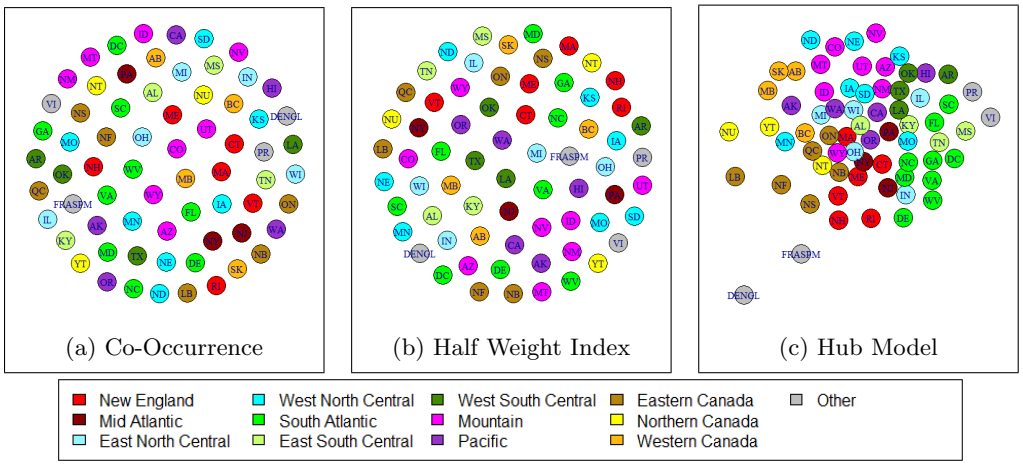



Figure 2: Adjacency Matrix Estimates for North American Flora Data

In Figure 2, we apply a force directed graph drawing technique to layout the states according to the estimated adjacency matrices of each approach. The HM graph is striking in how closely states in the same region are

grouped. Southern states are generally on the right side of Figure 2c while northern states are generally to the left. Eastern states are generally at the bottom of the figure while western states are generally at the top. Atlantic islands are on the outer edge of the plot.

In Figures 2a and 2b, there is almost no distinction between the organization of the states. This suggests that even in situations where the data does not clearly conform to the Hub Model assumption that valuable information about the relationship between nodes can be identified.

# S4 Discussion

## S4.1 Theoretical Curve

We gave a simple counterexample in Section S1.1 to demonstrate that without any constraints, the model is not identifiable. Here we will randomly select parameters to explore the issue in general. We randomly generate an asymmetric adjacency matrix with $n = 4$ (see Table 3). The number of observed groups is set high ($T = 100,000$) to ensure good performance of the algorithm. We ran Algorithm 1 100 times, and obtained 100 different estimators with the same or similar likelihoods. Figure 3 gives a scatterplot of the one hundred pairs of $\{\hat{A}_{1,2}, \hat{A}_{2,1}\}$ indicated by blue circles. This plot clearly demonstrates a non-linear relationship between these two values where increases in one are associated with decreases in the other.

| | | $A_{ij}$ | | | |
|---|---|---|---|---|---|
| | | j | | | |
| $\rho_i$ | i | 1 | 2 | 3 | 4 |
| 0.5499 | 1 | 1.0000 | 0.7854 | 0.9063 | 0.7957 |
| 0.3269 | 2 | 0.7032 | 1.0000 | 0.8324 | 0.5885 |
| 0.1016 | 3 | 0.9464 | 0.8817 | 1.0000 | 0.9334 |
| 0.0216 | 4 | 0.7452 | 0.8594 | 0.9478 | 1.0000 |

Table 3: True Adjacency Matrix in Identifiability Example

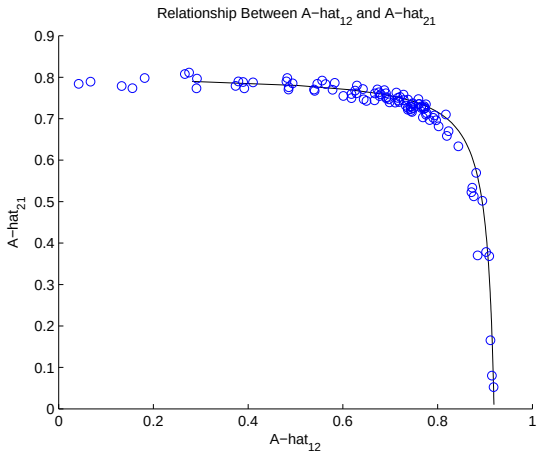



Figure 3: Nonlinear Relationship of Symmetric Elements for $\hat{A}$

It is possible to derive the theoretical relationship between symmetric elements of the adjacency matrix. As in Section S1.2, let $g^x$ and $g^y$ represent the observed groups which contain only node $v_x$ and $v_y$ respectively. Further, let $g^{xy}$ represent the group that is observed to contain only the pair $v_x$ and $v_y$.

Under the Hub Models $\mathbb{P}(G = g^x)$ and $\mathbb{P}(G = g^x)$ are given by (S1.2) and (S1.3). In the asymmetric case, the probability of the pair is:

$$\mathbb{P}(G = g^{xy}) = \rho_x A_{xy} \prod_{j \neq \{x,y\}} (1 - A_{xj}) + \rho_y A_{yx} \prod_{j \neq \{x,y\}} (1 - A_{yj}). \qquad \text{(S4.15)}$$

By simply reordering the terms of Equations S1.2 and S1.3, they can

be plugged back into Equation S4.15 to give:

$$\mathbb{P}(G = g^{xy}) = \frac{A_{xy}}{(1 - A_{xy})}\mathbb{P}(G = g^{x}) + \frac{A_{yx}}{(1 - A_{yx})}\mathbb{P}(G = g^{y}) \qquad \text{(S4.16)}$$

By some simple algebra, we find the following relationship between the elements of the adjacency matrix:

$$A_{xy} = \frac{\mathbb{P}(G = g^{xy}) - A_{yx}\big[\mathbb{P}(G = g^{y}) + \mathbb{P}(G = g^{xy})\big]}{\big[\mathbb{P}(G = g^{x}) + \mathbb{P}(G = g^{xy})\big] - A_{yx}\big[\mathbb{P}(G = g^{x}) + \mathbb{P}(G = g^{y}) + \mathbb{P}(G = g^{xy})\big]}. \qquad \text{(S4.17)}$$

Using (S4.17), we can calculate the relationship between $A_{1,2}$ and $A_{2,1}$ from the example above. This is represented in Figure 3 by the solid line. Clearly the observed solutions are falling along this theoretical curve.

## S4.2 Self-sparsity

In Section 6, we introduced a property of the Hub Model estimators which we refer to as *self-sparsity*. When $T$ is small relative to $n$, the model tends to produce a sparse adjacency matrix. Rabbat et al. (2008) observed similar behavior in their research. Sparsity in the adjacency matrix is achieved without any penalty in the log-likelihood, hence the name.

To begin, observe that the true probability of co-occurrence is related

to the parameters by the following equation:

$$\mathbb{P}(v_i \text{ and } v_j \text{ co-occur}) = \sum_{k=1}^{n} \rho_k A_{ki} A_{kj}. \tag{S4.1}$$

Suppose that there is a pair of nodes, $\{v_i, v_j\}$, for which the probability of co-occurrence is exactly zero. Equation (S4.1) implies that:

$$\rho_k A_{ki} A_{kj} = 0 \quad \forall k.$$

Hence, for every $k$, at least one of the following is true: $\rho_k = 0$, $A_{ki} = 0$, or $A_{kj} = 0$. At a minimum, this requires that there be $n$ elements of the parameters which are exactly equal to zero for every pair of nodes which fails to co-occur.

Clearly, $O_{ij} = 0$ implies $A_{ij} = 0$. However, self-sparsity shows that the absence of co-occurrence contains even more information than just the relationship between two nodes. Absence of co-occurrence means that no member of the population chooses to simultaneously interact with both nodes.

However, the question remains as to why self-sparsity occurs in the estimation of the parameters when $T$ is small relative to $n$. We observe that sparsity in $\hat{A}$ or $\hat{\rho}$ is a consequence of the EM-algorithm of HM.

First, it is easy to see that zero is an absorbing state for $\hat{A}$ and $\hat{\rho}$ in the EM-algorithm. If, at the $m^{th}$ iteration, $\hat{\rho}_i^{(m)} = 0$, then by (4.1) $\mathbb{P}(S_i = 1|G^{(t)}) = 0$ for all $i$, and by (S2.13) $\hat{\rho}_i^{(m+1)} = 0$. Therefore, $\hat{\rho}_i = 0$ is an absorbing state. For a similar reason, $\hat{A}_{ij} = 0$ is also an absorbing state.

Recall that in the EM-algorithm we set $\hat{A}_{ij} = 0$ when it is below a certain threshold after estimation is complete. But why the estimate approaches zero is not fully understood. This aspect of the model will be explored in future work.

# S5 Text Mining Protocol

In Section 6, we alluded to a text mining protocol for creating grouped data from the Chinese novel *Dream of the Red Chamber*. Here we present an overview of this technique.

We first downloaded a Chinese text version of the novel from the public domain and pasted the text into a Microsoft Word document. Then we developed a list of character names in Chinese and assigned each character a distinguishing letter from the Latin alphabet. For each character name, we performed a find-and-replace search for the Chinese characters and replaced them with a highlighted version of their distinguishing Latin letter. We also performed a find-and-replace for paragraph breaks and represented them

with an identifying character. This resulted in text like that shown in Figure 4. Then we removed all the text in the Word document which was not highlighted to give us Figure 5.

□□·B·方进入房时，只见两个人搀着一位鬓发如银的老母迎上来，·B·便知是他外祖母．方欲拜见时，早被他外祖母一把搂入怀中，心肝儿肉叫着大哭起来．当下地下侍立之人，无不掩面涕泣，·B·也哭个不住．一时众人慢慢解劝住了，·B·方拜见了外祖母．___此即冷子兴所云之史氏太君，·O··P·之母也．当下·N·一一指与·B·："这是你大舅母，这是你二舅母，这是你先珠大哥的媳妇珠大嫂子。"·B·一一拜见过．·N·又说："请姑娘们来．今日远客才来，可以不必上学去了。"众人答应了一声，便去了两个．····¶

Figure 4: Example of text from *Dream of the Red Chamber* after the Chinese names of characters have been replaced with Latin alphabet code letters.


Figure 5: Result of deleting non-highlighted text from document.

We pasted the remaining text into the first column of an Excel document. This gave us a text string which indicates the individuals mentioned in each paragraph. We deleted any rows with no data (these rows represent paragraphs in which none of the characters were mentioned). Then in the Excel document, we created a column for each individual where the value of a cell is 1 if the individual is mentioned in the associated text string and 0 otherwise.

Please contact the corresponding author if you are interested in raw data or estimated parameters.

# Bibliography

Carreira-Perpinan, M. A. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate bernoulli distributions. Neural Computation, 12:141–152.

Fowler, J. H. (2006a). Connecting the congress: A study of cosponsorship networks. Political Analysis, 14(4):456–487.

Fowler, J. H. (2006b). Legislative cosponsorship networks in the u.s. house and senate. Social Networks, 28(4):454–465.

Hamalainen, W. and Nykanen, M. (2008). Efficient discovery of statistically significant association rules. Proceedings of the 8th IEEE International Conference on Data Mining, pages 203–212.

Rabbat, M., Figueiredo, M., and Nowak, R. (2008). Network inference from co-occurrences. IEEE Transactions on Information Technology, 54(9):4053–4068.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell., 22(8):888–905.

259 Teicher, H. (1961). Identifiability of mixtures. The Annals of Mathematical
260 Statistics, 32(1):244–248.