

Statistica Sinica Preprint No: SS-2016-0397.R1

Title	Network Inference From Grouped Observations Using Hub Models
Manuscript ID	SS-2016-0397.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0397
Complete List of Authors	Yunpeng Zhao and Charles Weko
Corresponding Author	Yunpeng Zhao
E-mail	yzhao15@gmu.edu
Notice: Accepted version subject to English editing.	

Network Inference From Grouped Observations Using Hub Models

YUNPENG ZHAO¹ AND CHARLES WEKO²

George Mason University¹ and United States Army²

2 *Abstract:* In medical research, economics, and the social sciences data frequently
3 appear as subsets of a set of objects. Over the past century a number of descriptive
4 statistics have been developed to infer network structure from such data. However,
5 these measures lack a generating mechanism that links the inferred network struc-
6 ture to the observed groups. To address this issue, we propose a model-based
7 approach called the *Hub Model* which assumes that every observed group has a
8 leader and that the leader has brought together the other members of the group.
9 The performance of Hub Models is demonstrated by simulation studies. We apply
10 this model to the characters in a famous 18th century Chinese novel.

11 *Key words and phrases:* Social network analysis, affiliation network, expectation-
12 maximization algorithm, half weight index, Dream of the Red Chamber.

13 1 INTRODUCTION

14 A network can be denoted by $N = (V, E)$, where $V = \{v_1, v_2, \dots, v_n\}$ is the set
15 of n nodes, and E is the set of edges between nodes. In this article, we focus
16 on symmetric weighted networks represented by an $n \times n$ adjacency matrix,
17 A , where the element A_{ij} measures the relationship strength between nodes
18 v_i and v_j .

19 Traditionally, statistical network analysis focuses on modeling *observed*
20 network structure (e.g., highway systems or electrical transmission grids). In
21 this situation, nodes are well defined and the physical links between nodes is
22 observable (Hiller and Lieberman, 2001; Newman, 2011). However, in some
23 fields of research (e.g., the social sciences) network structure is not explicit.
24 In these fields, the observable data are groups of individuals and a model is
25 presumed to produce the groups. The fundamental task is to estimate model
26 parameters from such data.

27 Wasserman and Faust (1994) introduce inference of relationships with the
28 example of children attending birthday parties. In their example, the children
29 act as nodes in the network and the birthday parties represent subsets of
30 children.

31 In this paper, a collection of nodes observed in the same sample is called

32 a *group* and a dataset is called *grouped data*. In Wasserman and Faust's
33 example, each party defines a group and the set of all parties is the grouped
34 data. Two individuals are said to *co-occur* if they appear in the same group.

35 One common technique used to estimate an adjacency matrix from grouped
36 data is to count the number of times that a pair of nodes appears in the same
37 group (Zachary, 1977; Freeman et al., 1989; Wasserman and Faust, 1994; Ko-
38 laczyk, 2009; Brent et al., 2011). Frequently, a threshold is applied to this
39 count to create an unweighted adjacency matrix; however, Choudhury et al.
40 (2010) show that the characteristics of networks inferred by this technique
41 are sensitive to the threshold. We adopt a generalized version of the inter-
42 citation frequency (Kolaczyk, 2009) which measures the number of times a
43 pair of nodes is observed to co-occur in the dataset. We refer to this measure
44 as the *co-occurrence matrix*.

45 An alternative technique, called the *half weight index* (Cairns and Schwa-
46 ger, 1987), estimates an adjacency matrix by the frequency that two nodes
47 co-occur given that one of them is observed. This addresses a shortcoming
48 of the co-occurrence matrix in which nodes that appear rarely can be es-
49 timated to have a weak relationship even though the relationship is quite
50 strong (Voelkl et al., 2011).

51 The co-occurrence matrix and half weight index both have probabilistic

52 interpretations. The co-occurrence matrix estimates the probability that
53 two nodes will be observed together. The half weight index estimates the
54 probability that two nodes will be observed together given that one of them
55 is observed. However, these are not equivalent to the probability of an active
56 relationship between nodes. In fact, neither of these techniques describe the
57 process which leads to the generation of the observed groups. It is unclear
58 how these descriptive statistics relate to the grouped data in these methods.

59 We propose a model-based approach for grouped data generation which
60 we refer to as the *Hub Model* because each observed group is assumed to be
61 brought together by a hub node (see Figure 1).

62 The Hub Model is fundamentally different from classical network mod-
63 els such as the stochastic blockmodel and its variants (Holland et al., 1983;
64 Airoldi et al., 2008), the exponential random graph models (Frank and Strauss,
65 1986; Robins et al., 2007), the latent space model and its variants (Hoff et al.,
66 2002; Handcock et al., 2007), among others (see Goldenberg et al. (2010) for
67 a comprehensive review). These models focus on modeling the statistical
68 behavior of the network, that is, they treat the network as the observed
69 data. By contrast, the Hub Model treats the network as latent governing the
70 grouping behavior of a population. Our task is to estimate the latent network
71 (i.e., the adjacency matrix) from the observed group data. In this article,

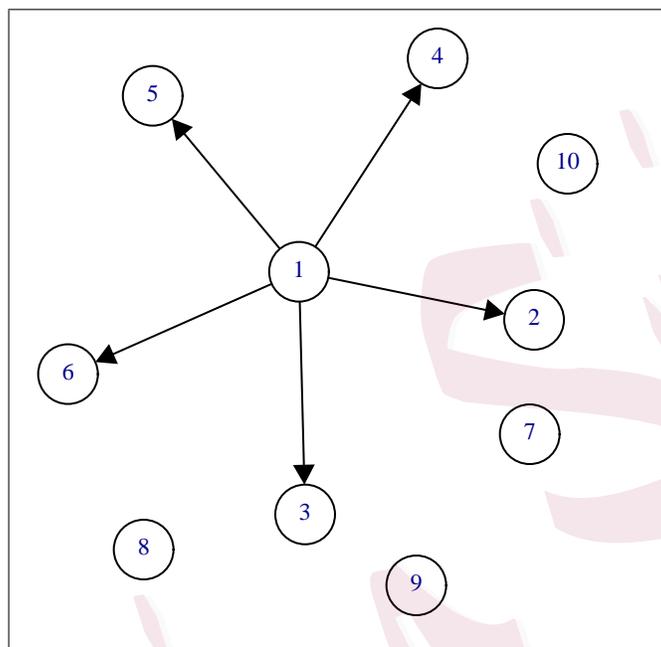


Figure 1: The generating mechanism of the Hub Model is demonstrated on a group of 10 nodes. In the observed sample, nodes v_1, \dots, v_6 are members of the group while nodes v_7, \dots, v_{10} are not members of the group. The observed group is the result of the hub node, v_1 , bringing together nodes v_2, \dots, v_6 .

72 we treat the adjacency matrix as fixed parameters and make no structural
73 assumption about it. If there were *a priori* information about the latent
74 network, such as that it follows the stochastic blockmodel or the exponential
75 random graph model, then one could take a Bayesian approach and use this
76 model as *a priori*. For more discussion, refer to Section 7.

77 The Hub Model belongs to the family of finite mixture models which
78 has been applied in many different situations including text classification

79 (Carreira-Perpinan and Renals, 2000), topic models (Anandkumar et al.,
80 2015), fingerprint identification (Vretos et al., 2012), and product recom-
81 mendation (Colace et al., 2015).

82 Hub Models have the advantage that relationship strength is both math-
83 ematically well defined and practical to researchers. In the Hub Model, A_{ij} ,
84 is defined as the probability that node v_i will include node v_j when v_i is the
85 hub node of a group. The formal definition of the Hub Model will be given
86 in Section 3.

87 As an introduction to Hub Models, consider the hypothetical relationships
88 in Figure 2a. In this example there is a pair of nodes, v_1 and v_2 , which never
89 directly pair to each other; however, they have an 80% chance of interacting
90 with five nodes. That is, $A_{ij} = 0.8$ for all $i \leq 2$ and $j \geq 3$ while $A_{ij} =$
91 0 otherwise. In Figure 2b, the co-occurrence matrix mistakenly assigns a
92 relatively strong relationship to nodes v_1 and v_2 because they often co-occur.
93 In Figure 2c, the half weight index arrives at a similar conclusion. In both
94 Figures 2b and 2c, the non-existent relationship between nodes v_1 and v_2 is
95 estimated to be stronger than all other relationships. By contrast, the Hub
96 Model in Figure 2d clearly captures the relationships of the population.

97 To the best of our knowledge, there have been limited attempts to apply
98 model-based approaches to these data. Rabbat et al. (2008) provide an

99 application for telecommunication networks. They modeled group formation
100 as a random walk from a source node to a terminal node. This model assumed
101 a distinctly different process of group formation from Hub Models. The nodes
102 along the path were subjected to an unknown permutation to account for the
103 lack of order information. Treating permutations as missing data, Rabbat et
104 al. employed a *Monte Carlo EM* algorithm based on importance sampling
105 to estimate the parameters of the model.

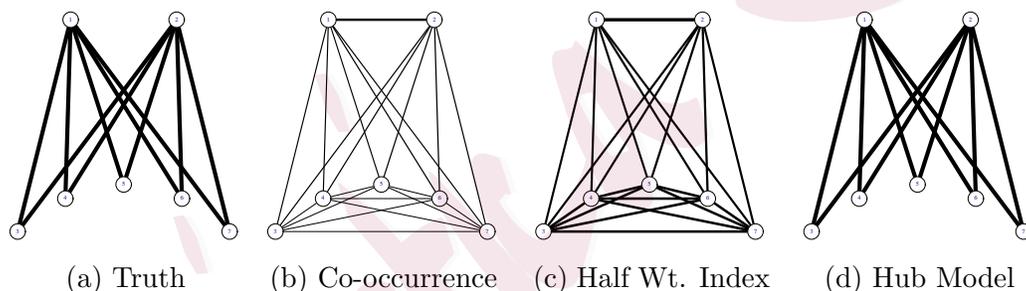


Figure 2: Comparison of Estimation Techniques

106 In the following sections we present a formal description of the grouped
107 data structure, review existing techniques, and define Hub Models. Then we
108 address Hub Model identifiability and provide a theorem that proves that a
109 symmetric adjacency matrix is a sufficient condition for identifiability. We
110 propose an EM algorithm to solve the maximum likelihood estimator of the
111 Hub Model. We then evaluate the model performance by simulation studies.
112 We apply the Hub Model to infer the relationships among the characters of

113 the 18th century Chinese novel, *Dream of the Red Chamber*. We close with
114 a discussion of how the size of the population impacts model efficiency and
115 ways to incorporate network structure assumptions to simplify the model.

116 **2 GROUPED DATA**

117 **2.1 Data Structure**

118 For a population of n individuals, $V = \{v_1, \dots, v_n\}$, we observe T subsets
119 of the global population, $\{V^{(t)} | V^{(t)} \subseteq V, t = 1, \dots, T\}$. Each observed subset
120 can be coded as an n length row vector $G^{(t)}$ where:

$$G_i^{(t)} = \begin{cases} 1 & \text{if } v_i \in V^{(t)} \\ 0 & \text{if } v_i \notin V^{(t)} \end{cases}$$

121 The full set of observations is denoted by a $T \times n$ matrix, G . The t^{th} row
122 of G is $G^{(t)}$.

123 **2.2 Existing Methods**

124 Inferring relationships from grouped data relies on descriptive statistics which
125 count the number of times that two nodes are observed together. We focus on

126 two popular techniques which estimate probabilities of individual behavior.

127 A simple measure of grouped data is the *co-occurrence matrix*. Versions
128 of this technique appear throughout the literature under many names and
129 notations including: *capacity matrix* (Zachary, 1977), *sociomatrix* (Wasser-
130 man and Faust, 1994), *inter-citation frequency* (Kolaczyk, 2009), *cocitation*
131 *matrix* (Newman, 2011), and *strength* (Brent et al., 2011).

132 A co-occurrence matrix, O , is an $n \times n$ symmetric matrix, defined by:

$$O = \frac{G'G}{T}, \quad (2.1)$$

133 which estimates the frequency that the nodes v_i and v_j are observed in
134 the same group.

135 One shortcoming of the co-occurrence matrix is that it estimates the
136 probability that two nodes *will be observed* to co-occur in a given observation.
137 That is, if two nodes have a strong relationship, but appear in the dataset
138 infrequently, the co-occurrence matrix will estimate a low probability that
139 the two nodes *will be observed* to co-occur.

140 As an example, consider four nodes v_1, \dots, v_4 and the grouped data repre-
141 sented in Table 1. For this dataset, both $O_{1,2} = \frac{2}{5}$ and $O_{3,4} = \frac{2}{5}$. However,
142 notice that every time node v_3 is present node v_4 is also present. A researcher

	Node			
Event	v_1	v_2	v_3	v_4
1	1	0	0	0
2	1	1	0	0
3	1	1	0	0
4	1	0	1	1
5	0	1	1	1

Table 1: Notional Grouped Data

143 may conclude that there is some aspect of the relationship between nodes v_3
144 and v_4 which has been understated.

145 As an alternative, the *half weight index* estimates the probability that
146 two nodes will be observed to co-occur given that one of them is observed
147 (Cairns and Schwager, 1987).

148 The half weight index has been introduced in a number of equivalent
149 forms (Dice, 1945). Computationally, the most direct form is:

$$H_{ij} = \frac{2 \sum_t G_i^{(t)} G_j^{(t)}}{\sum_t G_i^{(t)} + \sum_t G_j^{(t)}}. \quad (2.2)$$

150 Returning to the example in Table 1, we can see that $H_{1,2} = \frac{4}{7}$ while
151 $H_{3,4} = \frac{4}{4}$. Therefore, the half weight index infers a different network than
152 the co-occurrence matrix.

153 3 HUB MODELS

154 3.1 Generating Mechanism

155 Hub Models (HM) assume that each group is a star subgraph on the global
156 population. The hub node connecting the observed group is represented by
157 an n length row vector, $S^{(t)}$, where

$$S_i^{(t)} = \begin{cases} 1 & \text{if } v_i \text{ the hub node of sample } t, \\ 0 & \text{otherwise.} \end{cases}$$

158 There is one and only one element of $S^{(t)}$ that is equal to 1.

159 Each group is independently generated by a two step process.

160 1. The hub node is drawn from a multinomial distribution with parameter

161 $\rho = (\rho_1, \dots, \rho_n)$, i.e., $\rho_i = \mathbb{P}(S_i^{(t)} = 1)$. The following constraint applies:

162

$$\sum_i \rho_i = 1. \quad (3.3)$$

163 2. The hub node, v_i , chooses to include v_j in the group with probability

164 A_{ij} , i.e., $A_{ij} = \mathbb{P}(G_j^{(t)} = 1 | S_i^{(t)} = 1)$.

165 In most practical applications, the hub node of each group is unknown.

166 This article focuses on this case. We refer to the model where leaders are

3.2 Likelihood of the Hub Model

167 known as the Known Hub Model (KHM).

168 Since the co-occurrence matrix and half weight index produce a symmetric
169 adjacency matrix, we assume the Hub Model adjacency matrix is symmetric.
170 The symmetry condition will be shown to ensure the identifiability of the Hub
171 Model when group leaders are unobserved (Supplemental Material S1.2).

172 Further, we assume that the hub node will always include itself in the
173 group, i.e. $A_{ii} = 1$ for all i .

174 This generating mechanism implies that each observed group is indepen-
175 dent of every other observed group. In particular, $G^{(t)}$ is not a transformation
176 of $G^{(t-1)}$ and the order in which groups are observed contains no information
177 about the relationships between group members. Researchers often collect
178 data in such a way to ensure this property (Bejder et al., 1998).

179 **3.2 Likelihood of the Hub Model**

180 Under the HM, the probability of an observation has the form of a finite
181 mixture model with n components:

$$\mathbb{P}(G^{(t)}|A, \rho) = \sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}. \quad (3.4)$$

182 By taking the log of the product of individual observed groups, the log

3.2 Likelihood of the Hub Model

183 likelihood function for the full set of observations is:

$$\mathcal{L}(G|A, \rho) = \sum_t \log \left[\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1 - G_j^{(t)}} \right]. \quad (3.5)$$

184 Solving the MLE of HM is an optimization problem with the equality
185 constraints $\sum_i \rho_i = 1$, and $A_{ij} = A_{ji}$ for all i and j . From (3.5), we denote
186 the log likelihood function as $\mathcal{L}(G|A, \rho)$. This gives the following Lagrange
187 function:

$$\Lambda(G|A, \rho) = \mathcal{L}(G|A, \rho) - \lambda_o \left[\left(\sum_i \rho_i \right) - 1 \right] - \sum_{i < j} \lambda_{ij} (A_{ij} - A_{ji}). \quad (3.6)$$

188 The log likelihood does not have a closed-form solution for the MLE.
189 Instead we will derive estimating equations which can be incorporated into
190 an Expectation Maximization algorithm. Before doing so we investigate the
191 identifiability of the Hub Model.

A basic requirement for any model is *identifiability*. For Hub Models, this means for any two sets of parameters $\{A, \rho\}$ and $\{A^*, \rho^*\}$:

$$\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*) \quad \forall g \implies A = A^*, \rho = \rho^*. \quad (3.7)$$

192 The generating mechanism for Hub Models is equivalent to a finite mix-

3.2 Likelihood of the Hub Model

193 ture model of multivariate Bernoulli random variables. In general, such a
194 model is not identifiable (Teicher, 1961). This shortcoming does not prevent
195 such models from being useful in many applications. For example, when
196 dealing with classification problems where the researcher only has to identify
197 which component density an observation came from, this type of mixture can
198 be effectively used (Carreira-Perpinan and Renals, 2000). In such a situation,
199 the individual parameters of the multivariate Bernoulli random variables are
200 not of interest. However, the issue of identifiability presents a challenge in
201 our application because we are specifically interested in the individual pa-
202 rameters of the adjacency matrix.

203 If no constraint is put on the adjacency matrix, the model is unidentifi-
204 able. The following theorem establishes a sufficient condition for identifiabil-
205 ity. See Supplemental Material S1 for more details.

206 **Theorem 1.** Let A and A^* be symmetric adjacency matrices with $A_{ii} =$
207 $A_{ii}^* = 1$ for all i , $A_{ij} < 1$ and $A_{ij}^* < 1$ for all $i \neq j$. If $\mathbb{P}(g|A, \rho) = \mathbb{P}(g|A^*, \rho^*)$
208 for all g , then $\{A, \rho\} = \{A^*, \rho^*\}$.

209 It is worth noticing that even though symmetry of the adjacency matrix
210 is a natural assumption, it is only a sufficient condition for identifiability ac-
211 cording to Theorem 1. For future work, we will explore other assumptions to

212 ensure identifiability and ultimately find a necessary and sufficient condition.

213 **3.3 Estimating Equations**

214 In Supplemental Materials S2, we derive (3.8) and (3.9) which are estimating
215 equations that the MLE must satisfy. The maximum likelihood estimator of
216 HM does not have a closed-form solution for the parameters because the right
217 hand side of the estimating equations includes the estimated parameters.
218 Next we will show that solving these equations iteratively is equivalent to an
219 EM algorithm. The details of the EM algorithm will be given in the next
220 section.

$$\hat{A}_{xy} = \frac{\sum_t G_y^{(t)} \mathbb{P}(S_x = 1|G^{(t)}) + \sum_t G_x^{(t)} \mathbb{P}(S_y = 1|G^{(t)})}{\sum_t [\mathbb{P}(S_x = 1|G^{(t)}) + \mathbb{P}(S_y = 1|G^{(t)})]}. \quad (3.8)$$

$$\hat{\rho}_x = \frac{\sum_{t=1}^T \mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{T}. \quad (3.9)$$

221 **4 EM ALGORITHM**

222 The estimating equations shown above depend on the probability $\mathbb{P}(S_x^{(t)} =$
223 $1|G^{(t)})$. This implies an algorithm updating $\{\hat{A}, \hat{\rho}\}$ and $\mathbb{P}(S_x^{(t)} = 1|G^{(t)})$ iter-

224 atively, which can be fitted into the general framework of an EM algorithm.

225 The key technique of any EM algorithm is to formulate a complete data
226 model then solve the model as if some data is observed and other data is
227 missing. In this case, the Known Hub Model serves as the complete data
228 model, G is the observed data, and S is the missing data. Each iteration of
229 the EM algorithm consists of an expectation step followed by a maximization
230 step (McLachlan and Krishnan, 2008).

231 E-Step

Since the log likelihood function of the complete data model is linear in the unobserved data, the E-Step (on the $(m + 1)^{th}$ iteration) simply requires calculating the current conditional expectation of $S_i^{(t)}$ given the observed data (see McLachlan and Krishnan (2008) for detailed explanation).

$$\begin{aligned} E[S_x^{(t)} | G^{(t)}] &= \mathbb{P}(S_x^{(t)} = 1 | G^{(t)}) \\ &= \frac{\rho_x G_x^{(t)} \prod_j A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1-G_j^{(t)}}}{\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}} \end{aligned} \quad (4.10)$$

232 M-Step

233 The M-Step replaces $\mathbb{P}(S_x^{(t)} = 1 | G^{(t)})$ on the right hand side of (3.8) and

234 (3.9) with $E[S_x^{(t)}|G^{(t)}]$ from (4.10).

235 **Algorithm**

236 Algorithm 1 illustrates the details of the Hub Model.

237 Several standard techniques are used to improve the performance of the
238 EM algorithm. Firstly, we run the EM algorithm 10 times with different
239 starting points and choose the solution with the highest likelihood. Secondly,
240 we limit the number of iterations applied to a starting point. This second
241 treatment is based in part on the observation that when this algorithm has
242 a bad starting point, it will take a very long time to converge and the point
243 that it converges to is not close to the maximum. As a final step, we treat
244 any $\hat{A}_{xy} \leq 10^{-4}$ as $\hat{A}_{xy} = 0$. We apply this finishing step to remove clutter
245 from the returned solutions.

246 **5 SIMULATION**

247 In order to perform simulations, we generate parameters $\{A, \rho\}$ using the
248 following techniques.

249 For ρ , we select n random numbers, X_i , uniformly and divide each random

Data: G
Result: $\hat{A}, \hat{\rho}$
 Initialize:
 $\mathcal{L}(G|\hat{A}) = -\infty$
for rep=1 to 10 **do**
 Initialize:
 $\hat{A}_{ij}^{(0)} = \text{unif}(0, 1) \quad \forall \{i, j\}$
 $X_i = \text{unif}(0, 1) \quad \forall i$
 $\hat{\rho}_i^{(0)} = \frac{X_i}{\sum_k X_k}$
 $\Delta\mathcal{L}(G|A^{(0)}) = 10^4$
 counter=1
 while $\left| \frac{\Delta\mathcal{L}(G|A^{(m+1)})}{\mathcal{L}(G|A^{(m)})} \right| > 10^{-4}$ and counter < 100 **do**
 E-Step
 Update $\mathbb{P}(S_k^{(t)} = 1|G^{(t)})$ by Equation 4.10
 M-Step
 Update $A^{(m+1)}$ by Equation S2.10
 Update $\rho^{(m+1)}$ by Equation S2.13
 $\Delta\mathcal{L}(G|A^{(m+1)}) = \mathcal{L}(G|A^{(m+1)}) - \mathcal{L}(G|A^{(m)})$
 counter=counter+1
 end
 if $\mathcal{L}(G|A^{(m+1)}) > \mathcal{L}(G|\hat{A})$ **then**
 if $\hat{A}_{ij} \leq 10^{-4}$ **then**
 $\hat{A}_{ij} = 0$
 else
 $\hat{A}_{ij} = A_{ij}^{(m+1)}$
 end
 end
end

Algorithm 1: Expectation Maximization Algorithm for the Hub Model

250 number by the sum of all X_i 's. That is, $\rho_i = \frac{X_i}{\sum_i X_i}$.

251 We use a two step process to generate the adjacency matrix. First, we
 252 create a symmetric unweighted undirected random graph on n nodes using

253 the configuration model (Jackson, 2010) with a power law degree distribution
254 $\mathbb{P}(k) \propto k^{-\eta}$, where k is the possible value of the node degree. We assume a
255 power law degree distribution because it is commonly believed that many real
256 world social networks have this property Newman (2011). In all simulation
257 studies, we choose $\eta = 2$, because many networks are reported to have power
258 between 2 and 3 and a power of 2 generates the most dense networks, which is
259 a more challenging setup. We refer to this unweighted graph as the *structure*
260 of the network.

261 Each edge in the graph is then assigned a relationship strength with a
262 beta distribution,

$$A_{ij} = \begin{cases} \text{Beta}(\alpha, \beta) & \text{if there is an edge between } v_i \text{ and } v_j \\ 0 & \text{otherwise} \end{cases}$$

263 We simply let $A_{ji} = A_{ij}$ to ensure symmetry. We set $\alpha = 1$ and $\beta = 4$ in the
264 beta distribution so that the average relationship strength is less than 0.5,
265 which we believe is realistic.

266 In Tables 2 and 3, we consider five different network sizes $n = 10, 20, 50, 100, 150$.

267 For the first two cases, we set the minimum node degree to be 1 in the power
268 law distribution. And for the last three cases, we set the minimum degree to

269 be 5 in order to make sure the networks are not too sparse. For each size, we
270 generate 100 sets of parameters (A, ρ) using the setup described above. For
271 each (A, ρ) , we generate a dataset with T groups. Each average and standard
272 deviation are calculated over this 100 datasets. We use 9 different values of
273 $T = 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000$.

274 We first measure the ability of the estimated adjacency matrix \hat{A} to cor-
275 rectly identify the structure. To do this we define true positives and true
276 negatives as follows:

$$TP = \sum_{i < j} \mathbb{1}_{(A_{ij} > 0)} \mathbb{1}_{(\hat{A}_{ij} > 10^{-4})},$$

$$TN = \sum_{i < j} \mathbb{1}_{(A_{ij} = 0)} \mathbb{1}_{(\hat{A}_{ij} \leq 10^{-4})}.$$

277 Here, v_i and v_j are considered to have no relationship if the estimated
278 link strength is below 10^{-4} . False positives and false negatives are calculated
279 similarly. We use the Matthews correlation coefficient (MCC) to measure
280 the identification of the structure because it is a binary classification measure
281 that accounts for situations where the number of ones is significantly different
282 than the number of zeros (Liu et al., 2015). Based on our setup, our simulated
283 structures will have many more zeros than ones.

284 For the non-zero elements A_{ij} , we further evaluate the difference between
285 the numerical values of A_{ij} and \hat{A}_{ij} by calculating the mean absolute error
286 (MAE) of non-zero A_{ij} ,

$$MAE(A) = \frac{\sum_{i < j} |\hat{A}_{ij} - A_{ij}| \mathbf{1}_{(A_{ij} > 0)}}{\sum_{i < j} \mathbf{1}_{(A_{ij} > 0)}}.$$

287 We also report the average run time and the average number of iterations
288 for the EM algorithm when the simulation is run on an Intel Pentium CPU
289 G2030 at 3.00 GHz with 4.00GB of RAM.

290 The first observation from Tables 2 and 3 is that for a fixed value of n
291 the average error of both the MCC and the MAE decline as the number of
292 observations increases. By contrast, for a fixed number of observations, the
293 average error increases as the number of nodes increases.

294 The standard deviation of estimates generally improves once the number
295 of observations exceeds the number of parameters in the model. For example,
296 with 100 nodes there are roughly 10,000 parameters to estimate, thus samples
297 of only 2,000 or 5,000 observations demonstrate high standard deviations.

298 Finally, average run time generally increases as the number of obser-

299 vations increases and the number of nodes increases. An important factor
300 affecting the run time is the number of iterations the EM algorithm performs
301 before converging. In Table 2 the number of iterations declines as observa-
302 tions increase until it appears to approach a minimum number of iterations.
303 Table 3 provides further insight as the number of iterations generally in-
304 creases until the number of observations is roughly equal to the number of
305 parameters in the model after which the iterations declines. Up to that point,
306 the algorithm quickly converges to an adjacency matrix which is sparser than
307 the true adjacency matrix due to the insufficient sample size. The implica-
308 tion of these declining iterations is that run time is not strictly a function
309 of the size of the dataset, but the relationship between the number of nodes
310 and the number of observations.

	$n = 10$					
Obs	Avg MCC	StDev MCC	Avg MAE(A)	StDev MAE(A)	Avg Run Time (sec)	Avg Iterations
100	0.8010	0.0977	0.0533	0.0219	0.0472	20.258
200	0.8929	0.0903	0.0349	0.0128	0.0431	16.670
500	0.9487	0.0530	0.0212	0.0071	0.0411	13.618
1000	0.9770	0.0364	0.0147	0.0047	0.0369	12.011
2000	0.9865	0.0279	0.0102	0.0030	0.0353	10.613
5000	0.9984	0.0115	0.0067	0.0019	0.0298	9.604
10000	0.9988	0.0086	0.0045	0.0014	0.0295	9.416
20000	0.9994	0.0060	0.0035	0.0009	0.0305	9.327
50000	1	0	0.0020	0.0006	0.0316	9.210
	$n = 20$					
100	0.6727	0.0972	0.0833	0.0210	0.1005	21.007
200	0.7984	0.0756	0.0599	0.0154	0.0992	19.961
500	0.8781	0.0576	0.0340	0.0079	0.1039	17.793
1000	0.9147	0.0594	0.0225	0.0056	0.1131	15.418
2000	0.9360	0.0612	0.0150	0.0033	0.1473	13.803
5000	0.9734	0.0367	0.0099	0.0024	0.1653	11.571
10000	0.9842	0.0393	0.0069	0.0019	0.1806	10.662
20000	0.9937	0.0187	0.0048	0.0013	0.2052	10.260
50000	0.9989	0.0070	0.0031	0.0006	0.2320	9.888

Table 2: Average and Standard Deviation of Mean Absolute Error as Observations Increase

	$n = 50$					
Obs	Avg MCC	StDev MCC	Avg MAE(A)	StDev MAE(A)	Avg Run Time (sec)	Avg Iterations
100	0.3454	0.0503	0.1680	0.0139	0.2272	5.261
200	0.3987	0.0622	0.1368	0.0081	0.9216	16.237
500	0.5815	0.0668	0.0936	0.0085	2.7233	36.148
1000	0.8499	0.0302	0.0526	0.0049	2.6903	38.222
2000	0.9013	0.0176	0.0345	0.0030	2.3761	24.713
5000	0.9127	0.0193	0.0212	0.0017	2.8953	17.802
10000	0.9074	0.0259	0.0145	0.0012	5.1788	15.343
20000	0.9080	0.0327	0.0104	0.0008	7.1548	13.932
50000	0.9142	0.0383	0.0065	0.0006	12.190	12.866
	$n = 100$					
100	0.2620	0.0352	0.1955	0.0096	0.2058	2.040
200	0.3187	0.0346	0.1756	0.0109	0.2922	2.533
500	0.3495	0.0519	0.1359	0.0070	1.8683	9.151
1000	0.3857	0.0498	0.1109	0.0074	6.9431	25.852
2000	0.5343	0.1055	0.0748	0.0100	14.6644	44.035
5000	0.8236	0.1469	0.0351	0.0080	17.5031	34.544
10000	0.9128	0.0826	0.0219	0.0028	19.4031	23.370
20000	0.9355	0.0579	0.0148	0.0015	22.4366	17.494
50000	0.9484	0.0282	0.0092	0.0006	33.8123	13.905
	$n = 150$					
100	0.2247	0.0366	0.1994	0.0105	0.3373	1.536
200	0.2674	0.0316	0.1909	0.0081	0.3705	1.547
500	0.2965	0.0431	0.1632	0.0091	0.8822	2.623
1000	0.2625	0.0600	0.1363	0.0067	7.4969	11.65
2000	0.2354	0.0628	0.1247	0.0089	42.4597	47.525
5000	0.2700	0.1402	0.1075	0.0144	98.8080	75.973
10000	0.4276	0.2247	0.0822	0.0252	150.6061	72.416
20000	0.6025	0.2601	0.0532	0.0280	184.3534	60.144
50000	0.7602	0.2441	0.0275	0.0230	217.9005	41.975

Table 3: Average and Standard Deviation of Mean Absolute Error as Observations Increase (continued)

311 6 DATA ANALYSIS

312 In this section, we perform data analysis on the 18th century Chinese novel,
313 *Dream of the Red Chamber*. The observed groups in this dataset do not nec-
314 essarily conform to the Hub Model assumption. However, we will show that
315 even without this assumption being explicitly valid, important information
316 about the relationships can be estimated.

317 The Supplemental Materials S3 include two additional data sets estimat-
318 ing co-sponsorship of legislation in the Senate of the 110th United States
319 Congress and the dispersion of plant species across North America.

320 As noted by Kolaczyk (2009), a significant challenge with estimating the
321 parameters of implicit networks is that for a real world dataset there is usually
322 no way to verify the extent to which the estimate matches reality. That
323 is, there is no so-called “ground truth” or “golden standard” to compare
324 the estimated results against. Therefore, it is useful to analyze data about
325 which there is some qualitative knowledge of the relationships between nodes.

326 To this end, we construct a dataset of characters from *Dream of the Red*
327 *Chamber*. Since novels contain a qualitative social structure that is familiar to
328 readers, the results of quantitative analysis can be compared to this standard.

329 This novel is chosen for two reasons. Firstly, the relationships between

330 the characters are subtle and complex. Secondly, the novel has been carefully
331 studied by scholars. Therefore, the story presents a challenge to the task of
332 estimating relationships and there is a body of knowledge to compare the
333 estimates against.

334 Traditionally datasets are built from novels by carefully reading the text
335 and identifying dyadic interactions between characters based on criteria es-
336 tablished by the researchers, e.g., characters A and B have a conversation
337 (MacCarron and Kenna, 2013). This method may construct high quality
338 datasets; however, in order to identify interactions, it requires readers who
339 can read the novel and have time to build the datasets. Since *Dream of the*
340 *Red Chamber* is written in classical Chinese and the English translation runs
341 over 2,600 pages, directly generating the dataset would be excessively time
342 consuming.

343 Therefore, we built the dataset using text mining and define a group
344 as characters who co-occur in the same paragraph. Paragraphs with no
345 characters named in them are ignored. For a complete description of the
346 text mining protocol, see Supplemental Materials S5.

347 We analyze the relationships of 29 important characters. The character
348 names presented here are based on the original pinyin pronunciations and the
349 David Hawkes translation (Hawkes, 1974). A Chinese version of the novel

350 was used for text-mining. The complete novel contains 120 chapters, but we
351 focus on the first 80 because it is commonly believed that the last 40 chapters
352 are written by a different author and may not reflect the original themes of
353 the novel (Hsueh-Chin, 2016). The resulting dataset has 1,389 observations
354 of groups containing at least one of the 29 characters.

355 In Figure 3, the adjacency matrix is represented as an $n \times n$ grid where
356 the $i^{th} \times j^{th}$ cell represents the relationship between nodes v_i and v_j . The
357 relationship strength is represented by the cell's color. Nodes with weak
358 relationships have light cells while nodes with strong relationships have dark
359 cells. Cells representing relationships of intermediate strength are shaded
360 along the gray scale.

361 This visualization demonstrates another difference in the performance of
362 the techniques. The co-occurrence matrix estimates all relationships as being
363 very weak and it is difficult to differentiate strong relationships from the
364 absence of a relationship. The half-weight index presents a much stronger
365 set of relationships but there is evidence of relationships which have been
366 imputed transitively. In general, HM returns a much sparser network where
367 relationship strengths demonstrate higher contrast. This tendency towards
368 sparsity is discussed in more detail in the Supplemental Materials S4.2.

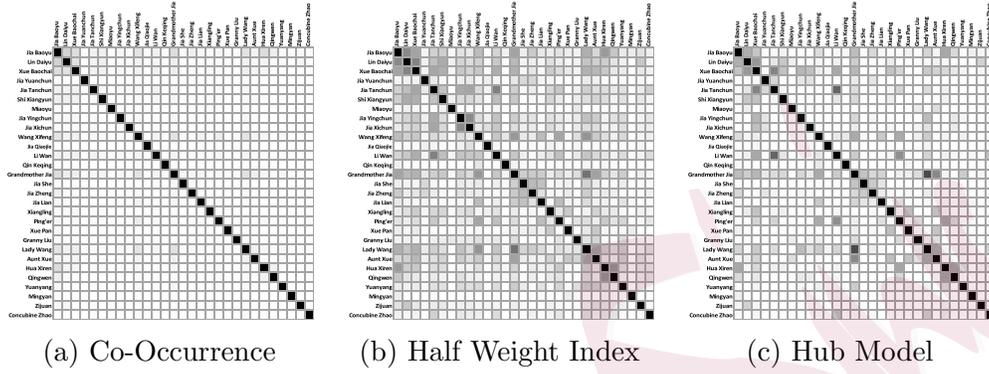


Figure 3: Comparison of Results for *Dream of the Red Chamber*

369 The EM algorithm of HM provides very stable solutions. By selecting
 370 multiple starting points, we find that the adjacency matrix (Figure 3c) is
 371 repeatedly returned as the most likely parameter of the observed data.

372 The Hub Model parameter’s standard deviation was estimated using the
 373 bootstrap technique. In general, the standard deviation was low. This was
 374 particularly true for $\hat{\rho}$ where the maximum standard deviation was 0.0173.
 375 Table 4 presents the standard deviation of the estimated adjacency matrix
 at different percentiles.

Percentile	Max	95 %	75 %	Med	25 %	5 %	Min
StDev	0.2696	0.1025	0.0374	0.0100	0.0000	0.0000	0.0000

Table 4: Percentiles of Standard Deviation in \hat{A} estimated by HM for *Dream of the Red Chamber*

376

377 One of the main themes of *Dream of the Red Chamber* is the love story

378 surrounding the protagonist Jia Baoyu (1st character in Figure 3c) and two
379 potential fiances, the sickly Lin Daiyu (2nd character) and the “ideal” Xue
380 Baochai (3rd character). Although Jia Baoyu shares a special bond with
381 Lin Daiyu and has no significant emotional connection to Xue Baochai, he
382 is ultimately tricked into marrying Xue Baochai (Hsueh-Chin, 2016). In
383 Table 5, we present the relationships between these two girls and the other
384 characters as estimated by the co-occurrence matrix, half weight index, and
385 HM.

386 From the novel, Lin Daiyu is a sensitive girl who prefers to be alone. By
387 contrast, Xue Baochai is a social and calculating girl. She is extremely good
388 at interpersonal communication especially with the protagonist’s mother
389 (Lady Wang) and grandmother (Grandmother Jia) (Hsueh-Chin, 2016). These
390 different personalities are clearly represented by the HM estimator while the
391 other estimators do not identify this difference.

	Co-Occurrence Matrix (O)		Half Weight Index (H)		Hub (\hat{A})	
	Lin Daiyu	Xue Baochai	Lin Daiyu	Xue Baochai	Lin Daiyu	Xue Baochai
Jia Baoyu	0.1728	0.1274	0.4563	0.3587	0.3113	0.2258
Lin Daiyu	1.0000	0.1109	1.0000	0.4866	1.0000	0.4072
Xue Baochai	0.1109	1.0000	0.4866	1.0000	0.4072	1.0000
Jia Yuanchun	0.0072	0.0050	0.0531	0.0449	0.0156	0.0228
Jia Tanchun	0.0439	0.0533	0.2490	0.3482	0.0915	0.4848
Shi Xiangyun	0.0590	0.0490	0.3273	0.3119	0.2194	0.2365
Miaoyu	0.0072	0.0036	0.0552	0.0337	0.0597	0
Jia Yingchun	0.0252	0.0274	0.1667	0.2141	0	0.2846
Jia Xichun	0.0187	0.0202	0.1313	0.1692	0.0102	0.2461
Wang Xifeng	0.0497	0.0526	0.1840	0.2131	0.0317	0.0697
Jia Qiaojie	0.0022	0.0022	0.0170	0.0208	0	0.0348
Li Wan	0.0367	0.0482	0.2086	0.3160	0.0580	0.3384
Qin Keqing	0.0007	0.0007	0.0052	0.0062	0	0
Grandmother Jia	0.0655	0.0648	0.2725	0.2985	0.1925	0.2820
Jia She	0.0065	0.0043	0.0449	0.0357	0	0
Jia Zheng	0.0122	0.0144	0.0701	0.0952	0.0143	0.0174
Jia Lian	0.0072	0.0036	0.0423	0.0245	0.0002	0.0073
Xiangling	0.0180	0.0252	0.1185	0.1961	0.0741	0.2344
Ping'er	0.0122	0.0209	0.0668	0.1306	0.0016	0.1643
Xue Pan	0.0043	0.0101	0.0292	0.0809	0	0
Granny Liu	0.0072	0.0050	0.0493	0.0411	0.0101	0.0113
Lady Wang	0.0490	0.0590	0.2248	0.3037	0.0224	0.2065
Aunt Xue	0.0302	0.0396	0.1806	0.2750	0.0479	0.1657
Hua Xiren	0.0403	0.0389	0.1938	0.2105	0.0283	0.1469
Qingwen	0.0166	0.0115	0.1020	0.0829	0.0155	0.0886
Yuanyang	0.0086	0.0101	0.0556	0.0763	0	0.0430
Mingyan	0.0007	0.0007	0.0053	0.0064	0	0
Zijuan	0.0317	0.0108	0.2184	0.0888	0.1775	0.0376
Concubine Zhao	0.0050	0.0058	0.0361	0.0495	0	0.0338

Table 5: Relationships of Lin Daiyu and Xue Baochai to other characters in *Dream of the Red Chamber*

392 7 CONCLUSION

393 To the best of our knowledge, Hub Models introduce an innovative approach
394 to the task of implicit network inference. By defining a model-based gener-
395 ating mechanism to link the latent network to observed grouped data and
396 applying an EM algorithm, we are able to estimate the network using this
397 model.

398 Not only are the estimators easy to calculate in a reasonable amount of
399 time, but they have a practical interpretation. The parameter ρ_i measures
400 the probability that node v_i will form a group. A_{ij} measures the probability
401 that a member of the population will be included in a group formed by node
402 v_i .

403 The Hub Models compare favorably against existing techniques. Since
404 the co-occurrence matrix and half weight index lack a generating mechanism
405 to connect them to the observed grouped data, these measures often cannot
406 detect important features of a network. By applying the Hub Model to the
407 18th century Chinese novel *Dream of the Red Chamber*, we demonstrate that
408 the HM is able to detect important features in the relationships between
409 nodes in complex situations.

410 By the standards of statistical network analysis, the size of the adjacency

411 matrices presented in this paper are small. An important question is how
412 the Hub Model would perform with 10,000 or even 1,000,000 nodes. While
413 it is computationally feasible to apply the Hub Model to populations of this
414 size, there is a practical challenge of collecting enough observations to have
415 sufficient statistical power.

416 We observe that how “small” or “large” a dataset is depends on the rela-
417 tionship between the number of nodes and the number of observed groups. In
418 principle, if there are n nodes, the Hub Model must estimate n^2 parameters.
419 If the number of observations is less than the number of nodes, multiple sets
420 of parameters have the same likelihood and parameter estimation is unstable.
421 In general, it is only when the number of observations exceeds the square of
422 the number of nodes, that we have stable estimates.

423 This means that to estimate the Hub Model parameters of a population
424 with hundreds of thousands of nodes, we would expect to have tens of billions
425 of observations. Therefore, applying Hub Models directly to text or even a
426 recommender system would be impractical.

427 In order to make the Hub Model useful for such large populations, some
428 technique must be applied to reduce the number of parameters in the model.
429 In this paper, we have placed no restrictions on the adjacency matrix. How-
430 ever, there are a number of restrictions which could be applied to enable us

431 to handle populations with “small” datasets.

432 One major way is to make an assumption about the structure of the
433 underlying network. For example, one might assume that the latent network
434 is itself the result of a block model or exponential random graph model. Such
435 an approach would create a hierarchical model for group formation.

436 A second way that assumptions about the structure of the underlying
437 network could be applied is to change the dimensions of the adjacency matrix.
438 In doing this, researchers may limit the number of nodes which can act as
439 leaders or treat some nodes as having the same behavior.

440 The Hub Model can be potentially useful to model the term-document
441 matrix in text mining. Such a matrix describes the frequency of terms that
442 occur in a collection of documents, which is similar to the format of group
443 data. So far many text mining techniques are based on a co-occurrence
444 matrix created from the term-document matrix. The Hub Model may provide
445 more meaningful estimates of the relations between terms.

446 **Supplementary Materials**

447 The supplemental materials contain additional details regarding the proof
448 of Theorem 1, calculation of the estimating equations 3.8 and 3.9. Addition-
449 ally, we provide data analysis for co-sponsorship of the 110th Congress and

REFERENCES

450 a dataset of North American flora. We conclude with a discussion of iden-
451 tifiability, self-sparsity, and the protocol for text mining *Dream of the Red*
452 *Chamber*.

453 Acknowledgements

454 This work is partially supported by NSF DMS 1513004.

455 Author's Statement

456 The views expressed in this paper are those of the authors and do not
457 reflect the official policy or position of the US Army, the Department of
458 Defense, or the US Government.

459 References

460 Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic
461 blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

462 Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S. M., and Liu, Y. (2015). A spectral algorithm
463 from latent dirichlet allocation. *Algorithmica*, 72(1):193–214.

464 Bejder, L., Fletcher, D., and Brager, S. (1998). A method for testing association patterns of social
465 animals. *Animal Behavior*, 56:719–725.

466 Brent, L. J. N., Lehmann, J., and Ramos-Fernandez, G. (2011). Social network analysis in the

REFERENCES

- 467 study of nonhuman primates: A historical perspective. *American Journal of Primatology*,
468 73:720–730.
- 469 Cairns, S. J. and Schwager, S. J. (1987). A comparison of association indices. *Animal Behavior*,
470 35.
- 471 Carreira-Perpinan, M. A. and Renals, S. (2000). Practical identifiability of finite mixtures of
472 multivariate bernoulli distributions. *Neural Computation*, 12:141–152.
- 473 Choudhury, M., M., W. A., Hofman, J. M., and Watts, D. J. (2010). Inferring relevant social
474 networks from interpersonal communication. *International World Wide Web Conference*
475 *Committee*.
- 476 Colace, F., De Santo, M., Greco, L., Moscato, V., and Picariello, A. (2015). A collaborative
477 user-centered framework for recommending items in online social networks. *Computers in*
478 *Human Behavior*.
- 479 Dice, L. R. (1945). Measures of the amount of ecological association between species. *Ecology*,
480 26:297–302.
- 481 Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Associa-*
482 *tion*, 81:832–842.
- 483 Freeman, L. C., White, D. R., and Romney, A. K. (1989). *Research Methods in Social Network*
484 *Analysis*. George Mason University Press.
- 485 Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical

REFERENCES

- 486 network models. *Foundations and Trends in Machine Learning*, 2:129–233.
- 487 Handcock, M. D., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social
488 networks. *J. R. Statist. Soc. A*, 170:301–354.
- 489 Hawkes, D. (1974). *The Story of the Stone, or The Dream of the Red Chamber, Vol. 1: The*
490 *Golden Days*. Penguin Classics.
- 491 Hiller, F. S. and Lieberman, G. L. (2001). *Introduction to Operations Research*. McGraw-Hill.
- 492 Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social
493 network analysis. *Journal of the American Statistical Association*, 97:1090–1098.
- 494 Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: first steps.
495 *Social Networks*, 5(2):109–137.
- 496 Hsueh-Chin, T. (2016). *CliffsNotes: Dream of the Red Chamber*. Houghton Mifflin Harcourt.
- 497 Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.
- 498 Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.
- 499 Liu, Y., Cheng, J., Yan, C., Wu, X., and Chen, F. (2015). Research on the matthews correla-
500 tion coefficients metrics of personalized recommendation algorithm evaluation. *International*
501 *Journal of Hybrid Information Technology*, 8(1):163–172.
- 502 MacCarron, P. and Kenna, R. (2013). Viking sagas: Six degrees of icelandic separation-social
503 networks from the viking era. *Significance*, pages 12–17.
- 504 McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley and

REFERENCES

- 505 Sons, Inc.
- 506 Newman, M. E. J. (2011). *Networks: An Introduction*. Oxford University Press.
- 507 Rabbat, M., Figueiredo, M., and Nowak, R. (2008). Network inference from co-occurrences. *IEEE*
508 *Transactions on Information Technology*, 54(9):4053–4068.
- 509 Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential
510 random graph (p^*) models for social networks. *Social networks*, 29(2):173–191.
- 511 Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–
512 248.
- 513 Voelkl, B., Kasper, C., and Schwab, C. (2011). Network measures for dyadic interactions: Stability
514 and reliability. *American Journal of Primatology*, 73:731–740.
- 515 Vretos, N., Nikolaidis, N., and Pitas, I. (2012). Video fingerprinting using latent dirichlet alloca-
516 tion and facial images. *Pattern Recognition*, 45(7):2489–2498.
- 517 Wasserman, S. and Faust, C. (1994). *Social Network Analysis: Methods and Applications*. Cam-
518 bridge University Press.
- 519 Zachary, W. W. (1977). An information flow model for conflicts and fission in small groups.
520 *Journal of Anthropological Research*, 33:452–473.

521 Department of Statistics

522 George Mason University

REFERENCES

- 523 4400 University Drive, MS 4A7
- 524 Fairfax, VA 22030-4444
- 525 E-mail: (yzhao15@gmu.edu)
- 526 United States Army
- 527 1400 Defense Pentagon
- 528 Washington, DC 20301
- 529 E-mail: (charles.w.weko.mil@mail.mil)