Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Network Inference Using the Hub Model and Variants

Zhibing He, Yunpeng Zhao, Peter Bickel, Charles Weko, Dan Cheng & Jirui Wang

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 546, 1264–1273: Theory and Methods
https://doi.org/10.1080/01621459.2023.2183133

Network Inference Using the Hub Model and Variants

Zhibing He[a], Yunpeng Zhao[a], Peter Bickel[b], Charles Weko[c], Dan Cheng[a], and Jirui Wang[d]

[a]Arizona State University, Tempe, AZ; [b]University of California, Berkeley, Berkeley, CA; [c]U.S. Army, Arlington, TX; [d]Medpace, Cincinnati, OH

ABSTRACT

Statistical network analysis primarily focuses on inferring the parameters of an observed network. In many applications, especially in the social sciences, the observed data is the groups formed by individual subjects. In these applications, the network is itself a parameter of a statistical model. Zhao and Weko propose a model-based approach, called the *hub model*, to infer implicit networks from grouping behavior. The hub model assumes that each member of the group is brought together by a member of the group called the *hub*. The set of members which can serve as a hub is called the *hub set*. The hub model belongs to the family of Bernoulli mixture models. Identifiability of Bernoulli mixture model parameters is a notoriously difficult problem. This article proves identifiability of the hub model parameters and estimation consistency under mild conditions. Furthermore, this article generalizes the hub model by introducing a model component that allows hubless groups in which individual nodes spontaneously appear independent of any other individual. We refer to this additional component as the *null component*. The new model bridges the gap between the hub model and the degenerate case of the mixture model—the Bernoulli product. Identifiability and consistency are also proved for the new model. In addition, a penalized likelihood approach is proposed to estimate the hub set when it is unknown. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2020
Accepted February 2023

KEYWORDS
Asymptotic properties; Bernoulli mixture models; Identifiability; Model selection; Network inference

1. Introduction

In recent decades, network analysis has been applied in science and engineering fields including mathematics, physics, biology, computer science, social sciences, and statistics (see Getoor and Diehl 2005; Goldenberg et al. 2010; Newman 2010 for reviews). Traditionally, statistical network analysis deals with parameter estimation of an observed network, that, an observed adjacency matrix. For example, community detection, a topic of broad interest, studies how to partition the node set of an observed network into cohesive overlapping or nonoverlapping communities (see Zhao 2017; Abbe 2018 for recent reviews). Other well-studied statistical network models include the preferential attachment model (Barabási and Albert 1999), exponential random graph models (Frank and Strauss 1986; Robins et al. 2007), latent space models (Hoff, Raftery, and Handcock 2002; Hoff 2007), and the graphon model (Diaconis and Janson 2007; Gao, Lu, and Zhou 2015; Zhang, Levina, and Zhu 2017).

In contrast to traditional statistical network analysis, this article focuses on inferring a latent network structure. Specifically, we model data with the following format: each observation in the dataset is a subset of nodes that are observed simultaneously. An observation is called a *group* and a full dataset is called *grouped data*. Wasserman and Faust (1994) introduced this format using the toy example of a children's birthday party. In their simple example, children are treated as nodes and each party represents a group—that is, a subset of children who attended the same party is a group. The reader is referred to Zhao and Weko (2019) and Weko and Zhao (2017) for applications of such data to the social sciences and animal behavior.

The observed grouping behavior presumably results from a latent social structure that can be interpreted as a network structure of associated individuals (Moreno 1934). The task is therefore to infer a latent network structure from grouped data. Existing methods mainly focus on ad-hoc descriptive approaches from the social sciences literature, such as the co-occurrence matrix (Wasserman and Faust 1994) or the half weight index (Cairns and Schwager 1987). Zhao and Weko (2019) propose the first model-based approach, called the *hub model*, which assumes that every observed group has a *hub* that brings together the other members of the group. When the hub nodes of grouped data are known, estimating the model parameters is a trivial task. In most research situations, hub nodes are unknown and need to be modeled as latent variables. Under this setup, estimating the model parameters becomes a more difficult task.

This article has three aims: first, to prove the identifiability of the canonical parameters and the asymptotic consistency for the estimators of those parameters *when hubs are unobserved*. The canonical parameters refer to the probabilities of being a hub node of a group and the probabilities of being included in a group formed by a particular hub node. The hub model is a restricted class from the family of finite mixtures of multivariate Bernoulli (Zhao and Weko 2019). Gyllenberg et al. (1994) showed that in general the parameters of finite mixture

CONTACT Yunpeng Zhao Yunpeng.Zhao@asu.edu Arizona State University, Tempe, AZ.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

models of multivariate Bernoulli are not identifiable. Zhao and Weko (2019) showed that the parameters are identifiable under two assumptions: the hub node of each group always appears in the group it forms and relationships are reciprocal. That is, the adjacency matrix is symmetric with diagonal entries as one. This article considers identifiability when adjacency matrices are asymmetric. The model is therefore referred as to the *asymmetric hub model*. We prove that when the hub set (i.e., the set of possible hubs) contains at least one fewer member than the node set, the parameters are identifiable under mild conditions. The new setup is practical and less restrictive than the symmetry assumption. Moreover, allowing the hub set to be smaller than the node set can reduce model complexity as pointed out by Weko and Zhao (2017). When proving the consistency of the estimators, we first prove the consistency of the hub estimates and then show that the estimators of model parameters are consistent as a corollary. Our proofs accommodate the most general setup in which the number of groups (i.e., sample size), the size of the node set, and the size of the hub set are all allowed to grow.

The second aim is to generalize the hub model to accommodate hubless groups and then prove identifiability and consistency of this generalized model. The classical hub model requires each group to have a hub. As observed in Weko and Zhao (2017), when fitting the hub model to data, one sometimes has to choose an unnecessarily large hub set due to this requirement. For example, a node that appears infrequently in general but appears once as a singleton must be included in the hub set. To relax the *one-hub* restriction, we add a component to the hub model that allows hubless groups in which nodes appear independently. We call this additional component the *null component* and call the new model the *hub model with the null component*. The proofs of identifiability and consistency for the new model do not parallel the first set of proofs and are more challenging.

Since the new models assume the hub set is a subset of the nodes, this raises a natural question: how to estimate the hub set from data, which is the third aim of the article. We formulate this problem as model selection for Bernoulli mixture models. We borrow the log penalty in Huang, Peng, and Zhang (2017), originally designed for Gaussian mixture models, to propose a penalized likelihood approach to select the hub set for the hub model with the null component. Instead of penalizing the mixing probability of every component as in Huang, Peng, and Zhang (2017), we modify the penalty function such that the probability of the null component is not penalized. The null component does not exist in the setup of Gaussian mixture models, but it creates a natural connection between the hub model and a null model in our scenario. That is, when all other mixing probabilities are shrunken to zero, the model naturally degenerates to the model in which nodes appear independently in a group—in other words, each group is modeled by independent Bernoulli trials.

2. Hub Model and Variants

2.1. Model Setup

First, we review the grouped data structure and then propose a modified version of the hub model, called the *asymmetric hub model*. For a set of n individuals, $V = \{1, \ldots, n\}$, we observe T subsets, called *groups*.

In this article, groups are treated as a random sample of size T with each group being an observation. Each group is represented by an n length row vector $G^{(t)}$, where

$$G_i^{(t)} = \begin{cases} 1 & \text{if node } i \text{ appears in group } t, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The full dataset is a $T \times n$ matrix $\boldsymbol{G}$ with $G^{(t)}$ being its rows.

Let V_0 be the set of all nodes which can serve as a hub and let $n_L = |V_0|$. We refer to V_0 as the *hub set* and call the nodes in this set *hub set member*. In contrast to the setup in Zhao and Weko (2019) where the hub set contains all nodes, we assume that the hub set contains fewer members than the whole set of nodes, that is, $n_L < n$. We assume in this section that V_0 is known and consider the problem of estimating V_0 in Section 3. For simplicity of notation, we further assume $V_0 = \{1, \ldots, n_L\}$ in this section. We refer to nodes from $n_L + 1$ to n as *followers*. Given this notation, the true hub of $G^{(t)}$ is represented by $z_*^{(t)}$ which takes on values from $1, \ldots, n_L$.

Under the hub model, each group $G^{(t)}$ is independently generated by the following two-step process:

(i) The hub is sampled from a multinomial trial with parameter $\rho = (\rho_1, \ldots, \rho_{n_L})$, that is, $\mathbb{P}(z_*^{(t)} = i) = \rho_i$, with $\sum_{i=1}^{n_L} \rho_i = 1$.
(ii) Given the hub node i, each node j appears in the group independently with probability A_{ij}, that is, $\mathbb{P}(G_j^{(t)} = 1 | z_*^{(t)} = i) = A_{ij}$.

Note that multiple hub set members may appear in the same group although only one of them will be the hub of that group.

A key assumption from Zhao and Weko (2019) which we adopt in this article is that a hub node must appear in any group that it forms (i.e., $A_{ii} \equiv 1$, for $i = 1, \ldots, n_L$). The parameters for the hub model are thus

$$\rho = (\rho_1, \ldots, \rho_{n_L}),$$

$$A_{n_L \times n} = \begin{pmatrix} 1 & A_{12} & \cdots & A_{1,n_L} & A_{1,n_L+1} & \cdots & A_{1,n} \\ A_{21} & 1 & \cdots & A_{2,n_L} & A_{2,n_L+1} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ A_{n_L,1} & A_{n_L,2} & \cdots & 1 & A_{n_L,n_L+1} & \cdots & A_{n_L,n} \end{pmatrix}.$$

As in Zhao and Weko (2019), we interpret A_{ij} as the strength of the relationship between node i and j. We differ from Zhao and Weko (2019) in that A is a non-square matrix and A_{ij} is not necessarily equal to A_{ji}. The setting in this article is more natural. Social relationships are usually nonreciprocal and in most organizations there are members who do not have the authority or willingness to initiate groups.

We begin with the case where both $\boldsymbol{G}$ and $z_* = (z_*^{(1)}, \ldots, z_*^{(T)})$ are observed. The likelihood function is

$$\mathbb{P}(\boldsymbol{G}, z_* | A, \rho) = \prod_{t=1}^{T} \prod_{i=1}^{n_L} \prod_{j=1}^{n} \left[A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{(1-G_j^{(t)})} \right]^{1(z_*^{(t)}=i)} \prod_{i=1}^{n_L} \rho_i^{1(z_*^{(t)}=i)},$$

where $1(\cdot)$ is the indicator function. With both $\boldsymbol{G}$ and z_* being observed, it is straightforward to estimate A and ρ by their respective Maximum Likelihood Estimators (MLEs):

$$\hat{A}_{ij}^{z_*} = \frac{\sum_t G_j^{(t)} 1(z_*^{(t)} = i)}{\sum_t 1(z_*^{(t)} = i)}, \quad i = 1, \dots, n_L, j = 1, \dots, n,$$
$$\hat{\rho}_i^{z_*} = \frac{\sum_t 1(z_*^{(t)} = i)}{T}, \quad i = 1, \dots, n_L.$$

When the hub node of each group is latent, that is, when z_* is unobserved, the estimation problem becomes challenging. Integrating out z_*, the marginal likelihood of $\boldsymbol{G}$ is

$$\mathbb{P}(\boldsymbol{G}|A, \rho) = \prod_{t=1}^{T} \sum_{i=1}^{n_L} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}, \qquad (1)$$

which has the form of a Bernoulli mixture model. Hereafter the term hub model refers to the case where z_* is unobserved, unless otherwise specified.

Although less stringent than the original symmetric hub model, the asymmetric hub model has a significant limitation: it cannot naturally transition to a null model. In general, a null model generates data that match the basic features of the observed data, but which is otherwise a random process without structured patterns. In other words, a null model is the degenerate case of the model class being studied. The null model for grouped data, naturally, generates each group by independent Bernoulli trials. That is, if the grouping behavior is not governed by a network structure then every node is assumed to appear independently in a group. The likelihood of $G^{(t)}$ under the null model is

$$\mathbb{P}(G^{(t)}) = \prod_{j=1}^{n} \pi_j^{G_j^{(t)}} (1 - \pi_j)^{1-G_j^{(t)}},$$

where π_j is the probability that node j appears in a group.

The asymmetric hub model needs generalization to accommodate the null model because if there is only one component in (1), say, node i is the only hub set member, the likelihood of $G^{(t)}$ becomes

$$\mathbb{P}(G^{(t)}) = \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}},$$

which is not a proper null model because the assumption $A_{ii} \equiv 1$ forces node i to appear in every group.

To allow the hub model to degenerate to the null model, we add the null component. This null component allows groups without hubs where nodes independently appear in such groups. We call this model the *hub model with the null component*. We use $z_*^{(t)} = 0$ to represent a hubless group.

The parameters for the hub model with the null component are $\rho = (\rho_0, \rho_1, \dots, \rho_{n_L}), A_{(n_L+1)\times n} = [A_{ij}]_{i=0,1,\dots,n_L, j=1,\dots,n}$. Here the row indices of A start from 0, that is, $A_{0j} \equiv \pi_j$ for $j = 1, \dots, n$. We will use A_{0j} and π_j interchangeably below. As before we assume $A_{ii} \equiv 1$ for $i = 1, \dots, n_L$. The marginal likelihood of $\boldsymbol{G}$ under the new model is

$$\mathbb{P}(\boldsymbol{G}|A, \rho) = \prod_{t=1}^{T} \sum_{i=0}^{n_L} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}. \qquad (2)$$

The above model degenerates to the null model when $\rho_0 = 1$. For simplicity of notation, we use the same notation such as ρ and A for both the hub model with and without the null component when the meaning is clear from context.

The new model has an advantage in data analysis in addition to the theoretical benefit. Grouped data usually contain a number of tiny groups such as singletons and doubletons. When fitting the asymmetric hub model to such a dataset, one sometimes has to include these nodes in the hub set due to the one-hub restriction. Doing so may result in an unnecessarily large hub set (see Section 4 in the supplemental materials). In the hub model with the null component, these small groups can be treated as hubless groups and the corresponding nodes may be removed from the hub set. Therefore, the model complexity is significantly reduced.

2.2. Model Identifiability

Before considering estimation of ρ and A under (1) and (2), we need to establish the identifiability of parameters ρ and A. Zhao and Weko (2019) proved model identifiability under the symmetry condition. We seek a new set of identifiability conditions as the new models do not assume symmetry of A.

To precisely define identifiability, let $\mathcal{P}$ be the parameter space of the hub model with the null component, where $\mathcal{P} = \{(\rho, A) | 0 < \rho_i < 1, i = 0, \dots, n_L; A_{ii} = 1, i = 1, \dots, n_L; 0 \le A_{ij} \le 1, i = 0, \dots, n_L, j = 1, \dots, n, i \neq j\}$. The parameter space of the hub model without the null component is similar except that the index i always begins with 1. Let $\boldsymbol{g} = (g_j^{(t)})_{t=1,\dots,T, j=1,\dots,n}$ be any realization of $\boldsymbol{G}$ under the hub model.

Definition 1. The parameters (ρ, A) within the parameter space $\mathcal{P}$ are identifiable (under the hub model with or without the null component) if the following holds:

$$\forall \boldsymbol{g}, \forall (\tilde{\rho}, \tilde{A}) \in \mathbb{P}(\boldsymbol{G} = \boldsymbol{g}|\rho, A) = \mathbb{P}(\boldsymbol{G} = \boldsymbol{g}|\tilde{\rho}, \tilde{A})$$
$$\iff (\rho, A) = (\tilde{\rho}, \tilde{A}).$$

We define identifiability in the strictest sense and the above definition does not allow label swapping of latent classes. In cluster analysis label swapping refers to the fact that nodes can be successfully partitioned into latent classes, but individual classes cannot be uniquely identified. For example, community detection may correctly partition voters into communities based on their political preferences, but cannot identify which political party each community prefers. This is not an issue in the hub model due to the constraint $A_{ii} = 1$. In addition, note that we only need to consider identifiability for the distribution of a single observation, that is, $T = 1$ because the data are independently and identically distributed. Let g be a realization of a single observation hereafter.

We now give the identifiability result for the asymmetric hub model.

Theorem 1. The parameters (ρ, A) of the asymmetric hub model are identifiable under the following conditions:

(i) $A_{ij} < 1$, for $i = 1, \dots, n_L, j = 1, \dots, n, i \neq j$;

(ii) for all $i = 1, \ldots, n_L$, $i' = 1, \ldots, n_L, i \neq i'$, the vectors $(A_{i,n_L+1}, A_{i,n_L+2}, \ldots, A_{i,n})$ and $(A_{i',n_L+1}, A_{i',n_L+2}, \ldots, A_{i',n})$ are not identical.

Condition (ii) implies that for any pair of nodes in the hub set, there exists a follower with different probability of being included in groups formed by the two hubs, respectively. All proofs are given in the supplementary materials.

Identifiability under the model with the null component is more difficult to prove than the case of the asymmetric hub model due to the extra null component in the model. In particular, there is no constraint such as $\pi_i = 1$ on parameters of the null component. The conditions for identifiability in the following theorem are; however, as natural as those in Theorem 1.

Theorem 2. The parameters (ρ, A) of the hub model with the null component are identifiable under conditions (1) and (ii) in Theorem 1 (index i begins with 0 in (1)), and

(iii) for any $i = 1, \ldots, n_L$, the vectors $(A_{i,n_L+1}, A_{i,n_L+2}, \ldots, A_{i,n})$ and $(\pi_{n_L+1}, \pi_{n_L+2}, \ldots, \pi_n)$ are different by at least two entries.

Condition (3) adds the requirement that for any hub i, there exist two followers which each has different probabilities of appearing in a group led by hub i than of appearing in a hubless group. This condition implies that there should exist at least two more nodes in the node set than in the hub set. This condition is natural if one compares it to condition (ii), as both imply that there exists at least one more column than rows in A.

2.3. Consistency of the Maximum Profile Likelihood Estimator

We consider the asymptotic consistency for the hub model in the most general setting. That is, we allow the number of groups (T), the size of the node set (n), and the size of the hub set (n_L) to grow. As mentioned in Section 1, we reformulate the problem as a clustering problem where a cluster is defined as the groups formed by the same hub node. We borrow the techniques from the community detection literature to prove the consistency of class labels, that is, the consistency of hub labels. The consistency of parameter estimation then holds as a corollary. Note that n is necessarily to go to infinity for proving the consistency of hub labels because when n is fixed, the posterior probability of the hub label of a group given the data cannot concentrate on a single node. If one is only interested in the consistency of parameter estimation, it is possible to allow n fixed. The problem degenerates to the classical case, that is, estimating a nongrowing number of parameters, and the classical theory of MLE is expected to be applicable.

We first consider the asymmetric hub model without the null component. Let $z = (z^{(t)})_{t=1,\ldots,T}$ be an assignment of hub labels. Given z, the log-likelihood of the full dataset $\boldsymbol{G}$ is

$$L_G(A|z) = \sum_{t=1}^{T} \sum_{j=1}^{n} G_j^{(t)} \log A_{z^{(t)},j} + (1 - G_j^{(t)}) \log(1 - A_{z^{(t)},j}). \quad (3)$$

For $i = 1, \ldots, n_L$, let $t_i = \sum_t 1(z^{(t)} = i)$ be the number of groups with hub i. Given z, the MLE of A is

$$\hat{A}_{ij}^z = \frac{\sum_t G_j^{(t)} 1(z^{(t)} = i)}{t_i}, \text{ for } t_i > 0.$$

If $t_i = 0$, define $\hat{A}_{ij}^z = 0$. We will omit the upper index z when it is clear from the context. Plugging $\hat{A}_{ij}$ back into (3), we obtain the profile log-likelihood

$$L_G(z) = \max_A L_G(A|z) = \sum_t \sum_j G_j^{(t)} \log \hat{A}_{z^{(t)},j} + (1 - G_j^{(t)}) \log(1 - \hat{A}_{z^{(t)},j}).$$

Furthermore, let

$$\hat{z} = \arg\max_z L_G(z).$$

The framework of profile likelihoods are adopted from the community detection literature (Bickel and Chen 2009; Choi, Wolfe, and Airoldi 2012), where z is treated as an unknown parameter and we search for the z that optimizes the profile likelihood.

Recall that z_* is the true class assignment. We will treat z_* as a random vector to maintain continuity with the previous section.

Let $P_j^{(t)} = \mathbb{P}(G_j^{(t)} = 1 | z_*^{(t)}) = A_{z_*^{(t)},j}$. Then by replacing $G_j^{(t)}$ by $P_j^{(t)}$, we obtain a "population version" of $L_G(z)$:

$$L_P(z) = \sum_t \sum_j P_j^{(t)} \log \bar{A}_{z^{(t)},j} + (1 - P_j^{(t)}) \log(1 - \bar{A}_{z^{(t)},j}),$$

where

$$\bar{A}_{ij} = \frac{\sum_t P_j^{(t)} 1(z^{(t)} = i)}{t_i}, \text{ for } t_i > 0. \quad (4)$$

Otherwise, define $\bar{A}_{ij} = 0$. Let $T_e = \sum_t 1(z_*^{(t)} \neq \hat{z}^{(t)})$ be the number of groups with incorrect hub labels. As discussed previously, we do not allow label swapping in the definition of T_e. Our aim is to prove

$$T_e/T = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We make the following assumptions throughout the proof of consistency under the asymmetric hub model:

H_1: $Tc_{\min}/n_L \leq t_{i*} \leq Tc_{\max}/n_L$ for $i = 1, \ldots, n_L$, where $t_{i*} = \sum_t 1(z_*^{(t)} = i)$ and $c_{\min}$ and $c_{\max}$ are constants.

H_2: $A_{ij} = s_{ij} d$ for $i = 1, \ldots, n_L, j = 1, \ldots, n$ and $i \neq j$ where s_{ij} are unknown constants satisfying $0 < s_{\min} \leq s_{ij} \leq s_{\max} < \infty$ while d goes to zero as n goes to infinity.

H_3: There exists a set $V_i \subset \{n_L + 1, \ldots, n\}$ for $i = 1, \ldots, n_L$ with[1] $|V_i| \geq \nu n/n_L$ such that $\tau = \min_{i,i'=1,\ldots,n_L, i \neq i', j \in V_i} (s_{ij} - s_{i'j})$ is bounded away from 0.

H_4: $A_{ii'} \leq c_0/n_L$ for $i = 1, \ldots, n_L$, $i' = 1, \ldots, n_L$, $i \neq i'$, where c_0 is a positive constant.

[1] $|\cdot|$ is the cardinality of a set.

H_1 ensures that no hub set members appear too infrequently. The assumption in fact automatically holds with high probability if $(n_L^2 \log n_L)/T = o(1)$, which can be proved by applying Hoeffding's inequality. Here we directly assume the condition for simplicity. H_2 allows the expected density of A to shrink as n grows, which is a common setup in the community literature. H_3 implies that for every hub set member there exists a set of nodes that are more likely to join groups initiated by this particular hub set member than others. The size of this set is influenced by v and the magnitude of this preference is influenced by d (since $A_{ij} = ds_{ij}$). The decay rates of d and v, as well as the growth rates of n_L, n and T, will be specified in the following consistency results. H_4 is a technical assumption that prevents label swapping from influencing the consistency results.

Now we state a lemma that T_e/T is bounded by $L_P(z_*) - L_P(\hat{z})$. That is, z_* is a *well-separated* point of maximum of L_P. The reader is referred to Section 5.2 in Van der Vaart (2000) for the classical case of this concept.

Lemma 1. Under $H_1 - H_4$, for some positive constant δ,

$$\mathbb{P}\left(\frac{\delta n_L}{dvnT}(L_P(z_*) - L_P(\hat{z})) \geq \frac{T_e}{T}\right) \to 1, \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We consider the most general setup in which n_L, n, and T all go to infinity in the main text. For the easier case of n_L being fixed, we give the corresponding results (Theorem 3′ and 4′ for the asymmetric hub model and Theorem 5′ and 6′ for the hub model with the null component) in the supplementary materials. Based on Lemma 1, we establish label consistency:

Theorem 3. Under $H_1 - H_4$, if $n_L^2 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^2 \log n_L/(dnv^2) = o(1)$, and $(\log T)^2 n_L^2 \log n_L/(dnv^2) = o(1)$, then

$$T_e/T = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

The next result addresses the consistency for parameter estimation of A, which is based upon a faster decay rate of T_e/T than Theorem 3 (see the proof of Theorem 4 in the supplemental materials for details).

Theorem 4. Under $H_1 - H_4$, if $n_L \log n/T = o(1)$, $n_L^3 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^4 \log n_L/(dnv^2) = o(1)$, and $(\log T)^2 n_L^4 \log n_L/(dnv^2) = o(1)$, then

$$\max_{i\in\{1,\ldots,n_L\}, j\in\{1,\ldots,n\}} \left|\hat{A}_{ij}^{\hat{z}} - A_{ij}\right| = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We now establish the consistency for the hub model with the null component. The proofs are more challenging due to the extra null component. We make the following assumptions throughout the proofs, parallel to $H_1 - H_4$:

H_1^*: $Tc_{\min}/n_L \leq t_{i*} \leq Tc_{\max}/n_L$ for $i = 0, \ldots, n_L$, where $t_{i*} = \sum_t 1(z_*^{(t)} = i)$ and $c_{\min}$ and $c_{\max}$ are constants.
H_2^*: $A_{ij} = s_{ij}d$ for $i = 0, \ldots, n_L, j = 1, \ldots, n$ and $i \neq j$ where s_{ij} are unknown constants satisfying $0 < s_{\min} \leq s_{ij} \leq s_{\max} < \infty$ while d goes to zero as n goes to infinity.
H_3^*: There exists a set $V_i \subset \{n_L+1, \ldots, n\}$ for $i = 1, \ldots, n_L$ with $|V_i| \geq vn/n_L$ such that $\tau = \min_{i=1\ldots,n_L, i'=0,\ldots,n_L, i\neq i', j\in V_i}(s_{ij} - s_{i'j})$ is bounded away from 0.
H_4^*: $A_{ii'} \leq c_0/n_L$ for $i = 0, \ldots, n_L$, $i' = 1, \ldots, n_L$, $i \neq i'$, where c_0 is a positive constant.

The main difference between the two sets of assumptions is on the range of the indices. For example, index i is from 0 to n_L in H_1^*. In particular, t_{0*} is the true number of hubless groups. Index i starts from 1 in H_3^* because we only define the set V_i for each hub set member i but not for the hubless case.

We need a result on the separation of $L_P(z_*)$ from $L_P(\hat{z})$ which is similar to Lemma 1. However, the technique in the original proof cannot be directly applied to the new model. A key step in the proof of Lemma 1 relies on the fact that we can obtain a nonzero lower bound for the number of correctly classified groups with node i as the hub node in the asymmetric hub model. Specifically, let $t_{ii'} = \sum_t 1(z_*^{(t)} = i, \hat{z}^{(t)} = i')$ for $i = 0, \ldots, n_L$, $i' = 0, \ldots, n_L$. Thus, t_{ii} is the number of correctly classified groups where node i is the hub node. For the asymmetric hub model, we obtain a lower bound for t_{ii}/t_{i*} ($i = 1, \ldots, n_L$) from the fact that a node cannot be labeled as the hub of a particular group if the node does not appear in the group. This is due to the assumption $A_{ii} \equiv 1$ for $i = 1, \ldots, n_L$. For the hub model with the null component, the lower bound for t_{ii}/t_{i*} cannot be proved by the same technique because all groups can be classified as hubless groups without violating the assumption $A_{ii} \equiv 1$.

We take a different path in the proof to overcome this issue and other technical difficulties due to the null component. We first bound t_{i0}/t_{i*} for $i = 1, \ldots, n$.

Lemma 2. Under $H_1^* - H_4^*$, if $n_L^4 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^6 \log n_L/(dnv^2) = o(1)$ and $(\log T)^2 n_L^6 \log n_L/(dnv^2) = o(1)$, then for all $\eta > 0$,

$$\frac{t_{i0}}{t_{i*}} \leq \eta, \quad i = 1, \ldots, n_L,$$

with probability approaching 1.

Based on the result in Lemma 2, we establish the label consistency for the hub model with the null component.

Theorem 5. Under the conditions of Lemma 2,

$$\frac{T_e}{T} = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

We conclude this section by the result on consistency for parameter estimation of A under the hub model with the null component.

Theorem 6. Under $H_1^* - H_4^*$, if $n_L \log n/T = o(1)$, $n_L^5 \log T/(dTv) = o(1)$, $(\log d)^2 n_L^8 \log n_L/(dnv^2) = o(1)$ and $(\log T)^2 n_L^8 \log n_L/(dnv^2) = o(1)$, then

$$\max_{i\in\{0,\ldots,n_L\}, j\in\{1,\ldots,n\}} \left|\hat{A}_{ij}^{\hat{z}} - A_{ij}\right| = o_p(1), \quad \text{as } n_L \to \infty, n \to \infty, T \to \infty.$$

3. The Hub Model with the Null Component and Unknown Hub Set

3.1. Model Setup

The asymmetric hub model (with or without the null component) assumes that the hub set is a subset of the nodes. The previous section addressed the estimation problem when the hub set is known, but in practice, the hub set is usually not known a priori. In this section, we study the selection of the hub set under the hub model with the null component.

Recall that V_0 denotes the hub set with $|V_0| = n_L$. In the following, we no longer assume $V_0 = \{1, \ldots, n_L\}$ and the goal is to estimate V_0. We begin with a known *potential hub set*, denoted by $\bar{V}_0$, which is subset containing all nodes that can potentially serve as hub set members. One might assume that the ideal $\bar{V}_0$ would be the same as the entire node set V; however, to prove identifiability of parameters when the hub set is unknown (see Theorem S1 in the supplemental materials), we require the potential hub set $\bar{V}_0$ to be smaller than V. In practice, this means we have prior knowledge that certain nodes do not play an important role in group formation and are therefore not included in the hub set. Let $M = |\bar{V}_0|$ with $n_L < M < n$. Without loss of generality, assume $\bar{V}_0 = \{1, \ldots, M\}$.

The data generation mechanism is the same as the hub model with the null component. The parameters are $\rho = (\rho_0, \rho_1, \ldots, \rho_M)$, $A_{(M+1)\times n} = [A_{ij}]_{i=0,1,\ldots,M, j=1,\ldots,n}$. For $i = 1, \ldots, M$, $\rho_i = 0$ if $i \notin V_0$. The corresponding $\{A_{ij}\}_{j=1,\ldots,n}$ therefore do not play a role in the model and will not be estimated. If all $\rho_i = 0$, $i = 1, \ldots, M$, the model degenerates to the null model in which nodes appear independently in all groups. The marginal likelihood of $\boldsymbol{G}$ is

$$\mathbb{P}(\boldsymbol{G}|A, \rho) = \prod_{t=1}^{T} \sum_{i=0}^{M} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}.$$

3.2. Penalized Likelihood

We propose to maximize the following penalized log-likelihood function to estimate V_0:

$$L(A, \rho) - T\lambda \sum_{i=1}^{M} [\log(\epsilon + \rho_i) - \log \epsilon], \tag{5}$$

$$\text{subject to } \rho_i \geq 0,\ i = 0, 1, \ldots, M,\ \sum_{i=0}^{M} \rho_i = 1,$$

where

$$L(A, \rho) = \log \mathbb{P}(\boldsymbol{G}|A, \rho) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{M} \rho_i \prod_{j=1}^{n} A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}} \right].$$

λ is the tuning parameter which controls the penalty on the mixing weights. ϵ is a small positive number. We use $\epsilon = 10^{-8}$ in all numerical studies. The estimated hub set V_0 includes node i $(i = 1, \ldots, M)$ if and only if $\hat{\rho}_i \neq 0$ in the maximizer of (5).

The penalty function in (5) was inspired by a similar penalty function proposed by Huang, Peng, and Zhang (2017) for selecting the number of components in Gaussian mixture models. However, our penalty function has a subtle but substantial difference: the hub node index m in the penalty function begins with 1 instead of 0—that is, we do not penalize the coefficient of the null component ρ_0. The model is therefore penalized toward the null model, that is, the independent Bernoulli model, when λ is sufficiently large. The penalty function uses $\log(\epsilon + \rho_i)$ instead of $\log \rho_i$ as in Huang, Peng, and Zhang (2017), because $\log(\epsilon + \rho_i)$ will not go to infinity when ρ_i goes to zero, which makes it possible for $\hat{\rho}_i$ to reach exactly zero.

Gu and Xu (2019a) studied model selection under another constrained class of Bernoulli mixture models—structured latent attribute models (SLAMs). Gu and Xu (2019a) proposed a penalty function similar to Huang, Peng, and Zhang (2017) but with a hard threshold. Huang, Peng, and Zhang (2017) and Gu and Xu (2019a) proved the selection consistency under their respective assumed models which we will study for our model in future work. That is, in the context of hub models, whether the selected hub set is identical to the true hub set with high probability when the size of the potential hub set (M) diverges.

3.3. Algorithm

We propose a modified expectation-maximization (EM) algorithm for optimizing (5).

***Algorithm** 1 (Modified EM).*
Iteratively update $\hat{A}$ and $\hat{z}$ by the following E-step and M-step until convergence.
Define $h_{ti} = \mathbb{P}(z^{(t)} = i|\boldsymbol{G}, A)$ for $t = 1, \ldots, T$ and $i = 0, \ldots, M$.

E-step: Given $\hat{A}$ and $\hat{\rho}$,

$$\hat{h}_{ti} = \frac{\hat{\rho}_i \mathbb{P}(G^{(t)}|z^{(t)} = i, \hat{A})}{\sum_{i=0}^{M} \hat{\rho}_i \mathbb{P}(G^{(t)}|z^{(t)} = i, \hat{A})}, \quad \text{for } i = 0, \ldots, M.$$

M-step: For i such that $\hat{\rho}_i \neq 0$, given $\hat{h}_{ti}$,

$$\hat{A}_{ij} = \frac{\sum_{t=1}^{T} \hat{h}_{ti} G_j^{(t)}}{\sum_{t=1}^{T} \hat{h}_{ti}}, \quad \text{for } j = 1, \ldots, n.$$

Update $\hat{\rho}$ by solving the following optimization problem:

$$\hat{\rho} = \arg\max_{\rho} L(\hat{A}, \rho) - T\lambda \sum_{i=1}^{M} \log(\epsilon + \rho_i), \tag{6}$$

$$\text{subject to } \rho_i \geq 0,\ i = 0, \ldots, M,\ \sum_{i=0}^{M} \rho_i = 1.$$

The only difference between Algorithm 1 and the standard EM algorithm is the update of $\hat{\rho}$ in the M-step. In the standard EM algorithm for the likelihood without the penalty term, $\hat{\rho}_i$ has a closed-form solution, that is, $\hat{\rho}_i = \sum_{t=1}^{T} \hat{h}_{ti}/T,\ i = 0, \ldots, M$. By contrast, (6) is a nonlinear optimization problem with inequality constraints, which we use a numerical technique—the augmented Lagrange multiplier (Ghalanos and Theussl 2015) method to solve the problem. In addition, since (5) is a nonconvex optimization problem, we use multiple different initial values (20 random initial values are used in this article) to help guard against local maxima.

4. Numerical Studies

4.1. Numerical Studies When the Hub Set is Known

In this section, we examine the performance of the estimators for the asymmetric hub model and the hub model with the null component when the hub set is known, under varying n_L, n and T. Hub set selection will be considered in the next section. The parameters are estimated by the standard EM algorithm and the estimated hub labels are determined according to the largest posterior probabilities.

For the asymmetric hub model, let ρ_i be generated independently from $U(0, 1)$ and renormalize ρ_i such that $\sum_{i=1}^{n_L} \rho_i = 1$. Let the size of the node set, n, be 100 or 500. We partition the follower set $\{n_L + 1, \ldots, n\}$ into n_L nonoverlapping sets $V_1, \ldots, V_{n_L}$. Each set V_i is the set of followers with a preference for hub set member i over other hub set members. As in Theorem 1, we assume different ranges of probabilities of joining a group for followers that prefer a specific hub set member than for followers which do not prefer that member. Specifically, for $j \in V_i$, the parameters A_{ij} are generated independently from $U(0.2, 0.4)$, and for $j \notin V_i$, the parameters A_{ij} are generated independently from $U(0, 0.2)$. The numerical results for sparser A will be given in Section 4 of the supplemental materials. For clarification, we will not use prior information about how A was generated in the estimating procedure. That is, we still treat A as unknown fixed parameters in the estimation. We generate these probabilities from uniform distributions for the sole purpose of adding more variations to the parameter setup. In each setup, we consider four different sample sizes, $T = 500, 1000, 1500$, and 2000, and two different values of the size of hub set, $n_L = 10$ and 20.

For the hub model with the null component, let the probability of hubless groups $\rho_0 = 0.2$, and let ρ_i be generated independently from $U(0, 1)$ and renormalize ρ_i such that $\sum_{i=1}^{n_L} \rho_i = 0.8$ for $i = 1, \ldots, n_L$. For a hubless group, each node will independently join the group with probability $\pi_j \equiv 0.05$ for $j = 1, \ldots, n$. The setups on n_L, n, $\{V_1, \ldots, V_{n_L}\}$, A, n_L and T are identical to the asymmetric hub model case.

Tables 1 and 2 show the performance of the estimators for the asymmetric hub model and the hub model with the null component, respectively. The first measure of performance we are interested in is the proportion of mislabeled groups, T_e/T. As the proportion of mislabeled groups approaches zero, we expect the parameter estimates to approach the accuracy achievable if the hub nodes are known. The second measure of performance is the RMSE($\hat{A}_{ij}$). As a reference point, we also provide the RMSE achieved when we treat the hub nodes as known, RMSE*. All results are averaged by 1000 replicates.

From the tables, the estimators for the asymmetric hub model generally outperform those for the hub model with the null component as the latter is a more complex model. The patterns within the two tables are, however, similar. First, the performance becomes better as the sample size T grows, which is in line with common sense in statistics. Second, the performance becomes worse as n_L grows because n_L is the number of components in the mixture model, and thus a larger n_L indicates a more complex model. Third, the effect of n is more complicated: the RMSE* for the case that hub labels are known slightly increases as n grows because the model contains more parameters. What we are interested in is the case where hub labels are unknown, and this is what our theoretical studies focused on. In this case, the RMSE($\hat{A}_{ij}$) significantly improves as n grows. This is because the clustered pattern becomes clearer as the number of followers increases, which is in line with the label consistency results in Section 2.3.

Table 1. Asymmetric hub model results.

$n_L = 10$	$n = 100$			$n = 500$		
	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.0479	0.0501	0.0475	0.0011	0.0483	0.0483
$T = 1000$	0.0335	0.0344	0.0332	0.0000	0.0337	0.0337
$T = 1500$	0.0295	0.0280	0.0272	0.0000	0.0274	0.0274
$T = 2000$	0.0262	0.0243	0.0236	0.0000	0.0235	0.0235
$n_L = 20$	$n = 100$			$n = 500$		
	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.2396	0.0791	0.0662	0.0605	0.0686	0.0673
$T = 1000$	0.1528	0.0548	0.0463	0.0096	0.0466	0.0463
$T = 1500$	0.1186	0.0433	0.0375	0.0029	0.0380	0.0379
$T = 2000$	0.0998	0.0366	0.0325	0.0013	0.0328	0.0328

NOTE: Mislabels: the fraction of groups with incorrect hub labels. RMSE($\hat{A}_{ij}$): average RMSEs when the hub labels are unknown. RMSE*: average RMSEs when the hub labels are known.

Table 2. Hub model with null component results.

$n_L = 10$	$n = 100$			$n = 500$		
	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.0842	0.0542	0.0511	0.0058	0.0516	0.0516
$T = 1000$	0.0595	0.0376	0.0357	0.0006	0.0362	0.0362
$T = 1500$	0.0512	0.0308	0.0294	0.0001	0.0292	0.0292
$T = 2000$	0.0489	0.0264	0.0253	0.0001	0.0253	0.0253
$n_L = 20$	$n = 100$			$n = 500$		
	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*	Mislabels	RMSE($\hat{A}_{ij}$)	RMSE*
$T = 500$	0.3206	0.0839	0.0734	0.1146	0.0732	0.0719
$T = 1000$	0.2102	0.0607	0.0506	0.0229	0.0510	0.0509
$T = 1500$	0.1598	0.0488	0.0411	0.0076	0.0418	0.0416
$T = 2000$	0.1419	0.0414	0.0355	0.0022	0.0359	0.0359

NOTE: Mis-labels: the fraction of groups with incorrect hub labels. RMSE($\hat{A}_{ij}$): average RMSEs when the hub labels are unknown. RMSE*: average RMSEs when the hub labels are known.

4.2. Numerical Results for Hub Set Selection

We study the performance of hub set selection by the penalized log-likelihood (5), which is optimized by the modified EM algorithm (Algorithm 1). We use the same settings as the hub model with the null component in the previous section. The only difference is we need to specify the potential hub set $\bar{V}_0 = \{1, \ldots, M\}$: we consider $M = 80$ for $n = 100$ and $M = 80, 200$ and 300 for $n = 500$. In each setup, AIC and BIC are used to select the tunning parameter, λ. Let $\widehat{V}_0$ be the estimate of V_0. The performance of hub set selection is evaluated by the True Positive Rate (TPR) and the False Positive Rate (FPR), where

$$\mathrm{TPR} = \frac{\sum_{i=1}^{M} 1(i \in V_0, i \in \widehat{V}_0)}{n_L},$$

$$\mathrm{FPR} = \frac{\sum_{i=1}^{M} 1(i \notin V_0, i \in \widehat{V}_0)}{M - n_L}.$$

Table 3. TPR and FPR for hub set selection.

n_L	T	Parameter tuning	$n = 100$		$n = 500$					
			$M = 80$		$M = 80$		$M = 200$		$M = 300$	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
10	1000	AIC	0.6438	0.0719	0.9460	0.0128	0.7338	0.0081	0.6986	0.0128
		BIC	0.5787	0.0283	0.9381	0.0127	0.6831	0.0042	0.6472	0.0081
20	1000	AIC	0.5140	0.1410	0.6972	0.0249	0.4831	0.0229	0.4780	0.0370
		BIC	0.5100	0.1350	0.6859	0.0239	0.4494	0.0132	0.4673	0.0318
10	2000	AIC	0.8613	0.0187	0.9909	0.0010	0.9130	0.0018	0.8585	0.0015
		BIC	0.7675	0.0043	0.9883	0.0005	0.8956	0.0007	0.8400	0.0004
20	2000	AIC	0.6560	0.1050	0.8551	0.0074	0.6770	0.0155	0.6250	0.0140
		BIC	0.4438	0.0344	0.7884	0.0034	0.5848	0.0058	0.5519	0.0056

Table 3 shows the TPR and FPR for hub set selection under various settings. The patterns in the table with respect to n_L, n and T are similar to Tables 1 and 2. That is, the performance of hub set selection is better for smaller n_L, larger n, and/or larger T. Among all settings, the model with $n_L = 10, T = 2000$ and $n = 500$ is the simplest for hub set selection purpose, which has the largest TPR and smallest FPR with λ selected by either AIC or BIC. Furthermore, the selection performance becomes worse as M grows because a larger M corresponds to a larger potential hub set and hence a larger candidate set of models.

5. Analysis of Passerine Data

We apply the hub model with the null component to analyze a dataset on grouping behavior of passerines (Shizuka and Farine 2016). The dataset includes 63 color-marked passerines in Australia for daily observations, which are 2 scarlet robins (*Petroica boodang*), 13 striated thornbills (*Acanthiza lineata*), 26 buff-rumped thornbills (*Acanthiza reguloides*), 14 yellow-rumped thornbills (*Acanthiza chrysorrhoa*), 4 speckled warblers (*Chthonicola sagittatus*), 2 white-throated treecreepers (*Cormobates leucophaea*), one white-eared honeyeater (*Lichenostomous leucotis*), and one unkown bird. A group is defined as individuals observed together in a flock, and in total there are 109 groups, that is, $T = 109$. Species information is summarized in Table 4.

In the following analysis, we set the potential hub set $\bar{V}_0$ with $M = 55$ as the collection of birds in the first four species (Table 4) and the other eight birds belonging to small-scale species as followers.[2] Table 5 shows the estimated hub set under various λ values where a gray block indicates that a node is included in the hub set. As λ increases, nodes are removed gradually from the hub set and at $\lambda = 0.065$, the hub model degenerates to the null model where the hub set is empty. The BIC selects $\lambda = 0.055$, where the estimated hub set includes v_9, v_{30}, and v_{42}, each belonging to one of the three large-scale species.

Table 4. Summary of passerine species.

Species	Binomial nomenclature	Number	Label
Scarlet robin	*Petroica boodang*	2	$v_1 - v_2$
Striated thornbill	*Acanthiza lineata*	13	$v_3 - v_{15}$
Buff-rumped thornbill	*Acanthiza reguloides*	26	$v_{16} - v_{41}$
Yellow-rumped thornbill	*Acanthiza chrysorrhoa*	14	$v_{42} - v_{55}$
Speckled warbler	*Chthonicola sagittatus*	4	$v_{56} - v_{59}$
White-throated treecreeper	*Cormobates leucophaea*	2	$v_{60} - v_{61}$
White-eared honeyeater	*Lichenostomus leucotis*	1	v_{62}
Unknown	unknown	1	v_{63}

Table 5. Estimated hub set for passerine data.

λ	v_7	v_9	v_{10}	v_{20}	v_{30}	v_{33}	v_{37}	v_{42}	v_{46}
0.045	■	■	■	■	■	■	■	■	■
0.050		■	■	■		■		■	
0.055		■			■			■	
0.060		■			■			■	
0.065									

6. Summary and Discussion

In this article we studied the theoretical properties of the hub model and its variants from the perspective of Bernoulli mixture models. The contributions of the article are four-fold. First, we proved the model identifiability of the hub model. Bernoulli mixture models are a notoriously difficult model to prove identifiability on, especially under mild conditions. Second, we proved the label consistency and estimation consistency of the hub model. Third, we generalized the hub model by adding the null component that allows nodes to independently appear in hubless groups. The new model can naturally degenerate to the null model—the Bernoulli product. We also proved identifiability and consistency of the newly proposed model. Finally, we proposed a penalized likelihood method to select the hub set, which estimates not only the size of the hub set, n_L, but also which nodes belong to the set. The new method can handle data with no prior knowledge of the hub set and hence greatly expands the domain of the applicability of the hub model.

A natural constraint from Zhao and Weko (2019) that we apply in this article is $A_{ii} = 1$ $(i = 1, \dots, n_L)$, which turns out to be a key condition for ensuring model identifiability and avoiding the label swapping issue in the proof of consistency. On the other hand, this constraint prevents the asymmetric hub model from naturally degenerating to the null model because one node always appear in every group when there is only one component in the hub model, which motivated adding the null component to the model.

We consider the profile likelihood estimator in the proofs of consistency. The marginal likelihood MLE could also be studied using a different framework. Bickel et al. (2013) and

[2] Nodes v_1 and v_2 appear frequently so we include them in the potential hub set.

Brault, Keribin, and Mariadassou (2020) proved the consistency of the Marginal Llikelihood MLE under the block models for undirected and directed networks, respectively. Their approach is to first prove the consistency of the MLE under the complete data likelihood and to further show that the marginal likelihood is asymptotically equivalent to the complete data likelihood, which implies the consistency of the MLE under the marginal likelihood. We plan to extend the above framework to the hub model for future works. Moreover, we plan to study the model selection consistency of the proposed hub set selection method, especially when n_L, n, and T are all allowed to grow. What we would also like to explore is to go beyond the independence assumption and to develop theories and model selection methodologies for correlated or temporally dependent groups (Zhao 2022).

Finally, we briefly review other work on Bernoulli mixture models. Gyllenberg et al. (1994) first showed that finite mixtures of Bernoulli products are not identifiable. Allman, Matias, and Rhodes (2009) introduced and studied the concept of generic identifiability, which means that the set of non-identifiable parameters has Lebesgue measure zero. Identifiability under another class of mixture Bernoulli models has been recently studied (Xu 2017; Gu and Xu 2019a, 2019b). This class of models, for example, Structured Latent Attribute Models (SLAMs), has applications in psychological and educational research. The motivation, the model setup, and the proof techniques presented in this article are all different from previous research, and the result of neither implies the other. Gu and Xu (2019a) further established the selection consistency in SLAMs when the number of potential latent patterns goes to infinity. It is intriguing to combine the techniques in the present paper and in Gu and Xu (2019a) to study the selection consistency in the hub model with the null component, especially for the case that both the size of the true hub set (n_L) and the size of the potential hub set (M) go to infinity.

Supplementary Materials

The supplementary materials contain proofs of all technical results in the article, additional numerical studies, and an analysis of extended bakery data.

Acknowledgments

The authors thank the editor, the associate editor, and two anonymous referees for their constructive feedback and suggestions.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Yunpeng Zhao acknowledges support from National Science Foundation grant DMS-1840203. Peter Bickel acknowledges support from National Science Foundation grant DMS-1713083. Dan Cheng and Zhibing He acknowledge support from National Science Foundation grant DMS-1902432 Simons Foundation Collaboration grant 854127.

References

Abbe, E. (2018), "Community Detection and Stochastic Block Models: Recent Developments," *Journal of Machine Learning Research*, 18, 1–86. [1264]

Allman, E. S., Matias, C., and Rhodes, J. A. (2009), "Identifiability of Parameters in Latent Structure Models with Many Observed Variables," *The Annals of Statistics*, 37, 3099–3132. [1272]

Barabási, A.-L., and Albert, R. (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509–512. [1264]

Bickel, P., and Chen, A. (2009), "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073. [1267]

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013), "Asymptotic Normality of Maximum Likelihood and its Variational Approximation for Stochastic Blockmodels," *The Annals of Statistics*, 41, 1922–1943. [1271]

Brault, V., Keribin, C., and Mariadassou, M. (2020), "Consistency and Asymptotic Normality of Latent Block Model Estimators," *Electronic Journal of Statistics*, 14, 1234–1268. [1272]

Cairns, S. J., and Schwager, S. J. (1987), "A Comparison of Association Indices," *Animal Behavior*, 35, 1454–1469. [1264]

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012), "Stochastic Blockmodels with Growing Number of Classes," *Biometrika*, 99, 273–284. [1267]

Diaconis, P., and Janson, S. (2007), "Graph Limits and Exchangeable Random Graphs," arXiv preprint arXiv:0712.2749. [1264]

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Asscociation*, 81, 832–842. [1264]

Gao, C., Lu, Y., and Zhou, H. H. (2015), "Rate-Optimal Graphon Estimation," *The Annals of Statistics*, 43, 2624–2652. [1264]

Getoor, L., and Diehl, C. P. (2005), "Link Mining: A Survey," *ACM SIGKDD Explorations Newsletter*, 7, 3–12. [1264]

Ghalanos, A., and Theussl, S. (2015), *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, R package version 1.16. [1269]

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning*, 2, 129–233. [1264]

Gu, Y., and Xu, G. (2019a), "Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models," *Journal of Machine Learning Research*, 20, 1–58. [1269,1272]

——— (2019b), "The Sufficient and Necessary Condition for the Identifiability and Estimability of the Dina Model," *Psychometrika*, 84, 468–483. [1272]

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994), "Non-Uniqueness in Probabilistic Numerical Identification of Bacteria," *Journal of Applied Probability*, 31, 542–548. [1264,1272]

Hoff, P. D. (2007), "Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data," in *Advances in Neural Information Processing Systems* (Vol. 20), Curran Associates, Inc. [1264]

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Asscociation*, 97, 1090–1098. [1264]

Huang, T., Peng, H., and Zhang, K. (2017), "Model Selection for Gaussian Mixture Models," *Statistica Sinica*, 27, 147–169. [1265,1269]

Moreno, J. L. (1934), *Who Shall Survive? A New Approach to the Problem of Human Interactions*, Washington DC: Nervous and Mental Disease Publishing Co. [1264]

Newman, M. (2010), *Networks: An Introduction*, Oxford: Oxford University Press. [1264]

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007), "An Introduction to Exponential Random Graph (p*) Models for Social Networks," *Social Networks*, 29, 173–191. [1264]

Shizuka, D., and Farine, D. R. (2016), "Measuring the Robustness of Network Community Structure Using Assortativity," *Animal Behaviour*, 112, 237–246. [1271]

Van der Vaart, A. W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [1268]

Wasserman, S., and Faust, C. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press. [1264]

Weko, C., and Zhao, Y. (2017), "Penalized Component Hub Models," *Social Networks*, 49, 27–36. [1264,1265]

Xu, G. (2017), "Identifiability of Restricted Latent Class Models with Binary Responses," *The Annals of Statistics*, 45, 675–707. [1272]

Zhang, Y., Levina, E., and Zhu, J. (2017), "Estimating Network Edge Probabilities by Neighbourhood Smoothing," *Biometrika*, 104, 771–783. [1264]

Zhao, Y. (2017), "A Survey on Theoretical Advances of Community Detection in Networks," *Wiley Interdisciplinary Reviews: Computational Statistics*, 9, e1403. [1264]

——— (2022), "Network Inference from Temporally Dependent Grouped Observations," *Computational Statistics & Data Analysis*, 171, 107470. [1272]

Zhao, Y., and Weko, C. (2019), "Network Inference from Grouped Observations using Hub Models," *Statistica Sinica*, 29, 225–244. [1264,1265,1266,1271]