Network Inference From Time-varying Grouped Observations

Yunpeng Zhao*

Department of Statistics, George Mason University

August 31, 2017

Abstract

In social network analysis, the observed data is usually some social behavior, such as the formation of groups, rather than explicit network structure. Zhao and Weko (2017) proposed a model-based approach called the hub model to infer implicit networks from grouped observations. The hub model assumes independence between groups, which sometimes is not valid in practice. In this article, we generalize the idea of the hub model into the case of time-varying grouped observations. As in the hub model, we assume that the group at each time point is gathered by one leader. Unlike in the hub model, the group leaders are not sampled independently but follow a Markov chain, and other members in adjacent groups are also correlated.

An expectation-maximization (EM) algorithm is developed for this model, and a polynomial-time algorithm is proposed for the E-step. The performance of the new model is evaluated under different simulation settings. We apply this model to a data set of the Kibale Chimpanzee Project.

Keywords: Grouping behavior; Social networks; Forward-backward algorithm

1 Introduction

A network is a data structure consisting of nodes (vertices) connected by links (edges). A network with n nodes can be represented by an $n \times n$ adjacency matrix $A = [A_{ij}]$, where $A_{ij} > 0$ if there is an edge between nodes i and j, and $A_{ij} = 0$ otherwise. A network A can be weighted where A_{ij} measures the link strength between node i and j.

Increasing attention has been drawn to network analysis by a number of fields such as social sciences, physics, computer science, biology, and statistics (see Getoor and Diehl (2005); Goldenberg et al. (2010); Newman (2010); Zhao (2017) for reviews of this area).

In many cases, social networks of humans or animals are implicit, and sometimes they are even conceptual. The raw data that can be observed is usually social behavior. For example, we may not directly observe "friendships" between individuals. On the contrary, what we may see is some social behavior. For example, two people called each other more than twice a week on average, or two dolphins frequently co-occurred (Bejder et al., 1998). From the so-called social network perspective (Moreno, 1934), social behavior is generally presumed to be governed by latent social networks.

In this article, we focus on a special type of social behavior – grouping behavior. In such data, individuals that appear together form a *group*. A data set consisting of such groups is referred to as *grouped data* by Zhao and Weko (2017). To better explain the structure of grouped data, we need some notation. For a set of n individuals, $V = \{v_1, ..., v_n\}$, we observe T subsets $V^1, ..., V^T$ at times $1, ..., T$, called groups. As in Zhao and Weko (2017), each observed subset V^t can be represented as an n length row vector G^t where

$$G_i^t = \begin{cases} 1 & \text{if } v_i \in V^t, \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, we will slightly abuse the notation: we will also call the indicator vector G^t as a group from now on. The goal of this paper is to infer the adjacency matrix A with the groups being observed.

Wasserman and Faust (1994) introduced such a data set, which is shown in Table 1. By the use of the notation above, $G_2^2 = 1$ since Drew attended Party 2, but $G_2^3 = 0$ since Drew did not attend Party 3.

Table 1: Dataset for six children and three birthday parties, Adapted from Wasserman and Faust (1994).

Party	Allison	Drew	Eliot	Keith	Ross	Sarah
1	1	0	0	0	1	1
2	0	1	1	0	1	1
3	1	0	1	1	1	0

As another example, Zhao and Weko (2017) treated the characters of a novel appearing in the same paragraph as a group and use the inferred network structure to interpret the relationships between characters.

Existing methods for network inference from grouped data are mainly ad-hoc approaches from the social sciences literature. A simple technique is to count the number of times that a pair of nodes appears in the same group. This measure has different names from different places, e.g., the *co-citation* matrix in Section 6.4 of Newman (2010) or the *sociomatrix* in Section 8.4 of Wasserman and Faust (1994). Zhao and Weko (2017) refers to this measure as the *co-occurrence matrix*. Half weight index (Cairns and Schwager, 1987) is an alternative approach which uses the conditional frequencies of co-occurrences as estimates. A common difficulty of such methods is that they provide no statistical model to connect these descriptive statistics with the latent network.

Zhao and Weko (2017) recently proposed a model-based approach for grouped observations. In the so-called *hub model*, groups at different time points are assumed to form independently and there is a central node called *hub* or group leader in each group, who gathers other members into the group. For example, the hub is the child who hosted the party in the example above.

A crucial assumption made in Zhao and Weko (2017) is the independence between groups. In some cases, this assumption is reasonable if each group forms spontaneously. The assumption can also be approximately satisfied if researchers collect grouped data with sufficiently long time intervals between observations (see Bejder et al. (1998) for discussion).

However, the independence assumption may not be valid in other situations. In most

practical cases, the grouped observations are time-varying in nature. For example, in the study of animal behavior, researchers may observe the behavior of animals on an hourly or daily basis. In Section 5, we will analyze such a data set consisting of groups of wild chimpanzees studied by the Kibale Chimpanzee Project (https://kibalechimpanzees.wordpress.com/). It is inappropriate to assume that every group is independent with the previous group. A more plausible point of view is that the group at a particular time is a transformation of the previous one. That is, some new members may join the group and some may leave, but the group maintains a certain level of stability. Also, we focus on the case of only one group being observed at a time point in this article.

We generalize the idea of the hub model to the above time-varying setting. We will call the new model as the time-varying hub model or the time-varying model in short. This new model allows for dependency between group leaders as well as between others group members. We will explain both assumptions of dependency in the next two paragraph, respectively.

As in the classical hub model, we assume that there is one leader for each group. But leaders are not sampled independently in the time-varying model, but follow a Markov chain. That is, the probability of the current leader being a certain node depends on the leader in the previous group.

For other group members, we consider the following two cases to make the model flexible enough. If the current leader is inside the previous group, then we treat this group as a transformation of the previous one, as mentioned above. If the new leader is outside the previous group (e.g., some event could happen and thus completely break the previous group), then we treat this group as the start of a new segment. In this case, the leader will select the group members as in the classical hub model, i.e., independently on whether they are members of the previous group.

As will be seen in Section 3, the time-varying hub model can be viewed as a generalization of the hidden Markov model (HMM) when the group leaders are latent. An efficient algorithm is thus developed for model fitting.

Finally, we discuss some related work. First, the time-varying hub model is fundamentally different from many existing models for dynamic networks, such as the preferential

attachment model (Barabási and Albert, 1999), discrete/continuous time Markov models (Snijders, 2001; Hanneke and Xing, 2007), etc. In these works, the observed data are the snapshots of the network at different time points. In this article, the unknown parameters are a single latent network, and the observations are groups with time-varying structure.

Second, there are recent studies on estimating latent networks or related latent structures in the dynamic setting, but from data structures different from groups. Guo et al. (2015) proposed a Bayesian model to infer latent relationships between people from a special type of data – the evolution of people's language over time. Robinson and Priebe (2013) proposed a latent process model for dynamic relational network data. Such a data set consists of binary interactions at different times. Blundell et al. (2012) developed a nonparametric Bayesian approach for estimating latent communities from a similar type of data. The grouped data we consider in this article is more complicated than binary interactions in the sense that unlike a linked pair, the links within a group consisting of more than two members are unknown.

2 Model

2.1 The classical hub model

We briefly state the generating mechanism of the classical hub model proposed by Zhao and Weko (2017). The hub model assumes one leader for each group. The leader of G^t is denoted by z^t.

Under the hub model, each group G^t is independently generated by the following two steps.

1. The group leader is sampled from a multinomial distribution with parameter $\rho = (\rho_1, ..., \rho_n)$, i.e., $\mathbb{P}(z^t = i) = \rho_i$, with $\sum_i \rho_i = 1$.

2. The group leader, v_i, will choose to include v_j in the group with probability A_{ij}, i.e., $\mathbb{P}(G_j^t = 1 | S_i^t = 1) = A_{ij}$.

2.2 Generating mechanism of the time-varying hub model

The hub model assumes that all the groups are generated independently across time. In practice, it is more natural to model the groups as time-varying observations.

We first explain the idea of the generating mechanism of time-varying groups and then give the formal definition. We generalize the idea of the hub model into the time-varying setting. Specifically, we keep assuming that there is only one leader z^t at each time who brought the group together, but the group at time t depends on previous group, which is different from the classical hub model.

At time $t = 1$, the group is generated from the classical hub model. For $t = 2, ..., T$, the group leader z^t can remain the same as the previous leader or change to a new one. We assume that the leader z^t will remain as z^{t-1} with a higher probability than the probability of changing to any other node.

If the new leader is outside the previous group, then the current group is considered as the start of a new segment and is generated by the classical hub model. It is worth noting that technically, the generation of the new group however still depends on the previous group. This will become more clear after we introduce the likelihood function. For the case that the new leader is inside the previous group – that is, if the leader remains unchanged, or the leader changes but is still a member of the previous group – we propose the following *In-and-Out procedure*. For any node v_j being in the previous group, it will remain in G^t with a probability higher than $A_{z^t,j}$ – the probability in the classical hub model. On the contrary, for any node v_k not being in the previous group, it will enter G^t with a probability lower than $A_{z^t,k}$. Intuitively, this *In-and-Out procedure* assumes that when a group forms, it will maintain a certain level of stability.

We now give the formal definition of the generating mechanism as follows.

- Step 1: (Classical hub model.) When $t = 1$, G^t is generated by the following two sub-steps.

    1) The leader is sampled from a multinomial distribution with parameter $\rho = (\rho_1, ..., \rho_n)$, i.e.,
    $$\mathbb{P}(z^t = i) = \rho_i \triangleq \frac{\exp(u_i)}{\sum_{k=1}^n \exp(u_k)}.$$

2) The leader v_i will choose to include v_j in the group with probability A_{ij}, i.e., $\mathbb{P}(G_j^{(t)} = 1 | z^t = i) = A_{ij}$, where $A_{ii} \equiv 1$, and

$$A_{ij} = A_{ji} \triangleq \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})}. \tag{2.1}$$

- Step 2: (Leader change.) For $t = 2, ..., T$,

$$\mathbb{P}(z^t = i | z^{t-1}) = \frac{\exp(u_i + \alpha I(z^{t-1} = i))}{\sum_{k=1}^{n} \exp(u_k + \alpha I(z^{t-1} = k))}.$$

- Step 3: (In-and-Out procedure.) For $t = 2, ..., T$, given v_i being the leader, G^t is generated by the following mechanism.

  If v_i is not within G^{t-1}, then it will include each v_j in the group with probability A_{ij}, otherwise, see below:

  1) If $G_j^{t-1} = 1$, v_i will include v_j in the group with probability

$$B_{ij} = B_{ji} \triangleq \frac{\exp(\theta_{ij} + \beta)}{1 + \exp(\theta_{ij} + \beta)}. \tag{2.2}$$

  2) If $G_j^{t-1} = 0$, v_i will include v_j in the group with probability

$$C_{ij} = C_{ji} \triangleq \frac{\exp(\theta_{ij} + \gamma)}{1 + \exp(\theta_{ij} + \gamma)}. \tag{2.3}$$

Remark

1. In the definition above, ρ_i's are re-parameterized as a exponential form and so are A_{ij}'s. This is for the convenience of optimization since log-likelihood is convex under this parametrization.

2. The parameters $(u_1, ..., u_n)$ are non-identifiable under this parametrization since $(u_1 + \delta, ..., u_n + \delta)$ gives the same likelihood. We will discuss the solution to this problem in Section 3 after introducing the algorithm.

3. We will not enforce $\alpha > 0$, $\beta > 0$ and $\gamma < 0$ in the model fitting. On the contrary, we will test these assumptions for the data example in Section 5.

2.3 Likelihood

For notational convenience, the leader in group G^t can also be indicated by an n length vector, S^t, where

$$S_i^t = \begin{cases} 1 & \text{if } z^t = i, \\ 0 & \text{otherwise.} \end{cases}$$

Only one element of S^t is allowed to be 1.

Clearly, $\{z^1, ..., z^T\}$ is a Markov chain according to the generating mechanism. Let $\Phi_{ij} = \mathbb{P}(z^t = i | z^{t-1} = j)$ be the transition probability, and $\Phi = [\Phi_{ij}]_{n \times n}$.

For the clarity of notation, we now give the vector/matrix form of the other parameters. Define $z = (z^1, ..., z^T)$, $u = (u_1, ..., u_n)$ and $\rho = (\rho_1, ..., \rho_n)$. Define $\theta = [\theta_{ij}]_{1 \leq i \leq n, 1 \leq j \leq n}$, $A = [A_{ij}]_{1 \leq i \leq n, 1 \leq j \leq n}$, $B = [B_{ij}]_{1 \leq i \leq n, 1 \leq j \leq n}$, and $C = [C_{ij}]_{1 \leq i \leq n, 1 \leq j \leq n}$. We assume θ, A, B and C to be symmetric in the article, in order to avoid any issues of identifiability (see the discussion in Zhao and Weko (2017)). And define G and S as $T \times n$ matrices, with G^t and S^t being their rows, respectively. We summarize notation in Table 2.

We now give the joint log-likelihood of S and G for the model defined in the previous sub-section:

$$
\begin{aligned}
&\log \mathbb{P}(S, G | \alpha, \beta, \gamma, \theta, u) \\
&= \sum_{i=1}^{n} S_i^1 \log \rho_i + \sum_{t=2}^{T} \sum_{i=1}^{n} \sum_{j=1}^{n} S_i^t S_j^{t-1} \log \Phi_{ij} \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ S_i^1 G_j^1 \log A_{ij} + S_i^1 (1 - G_j^1) \log(1 - A_{ij}) \right\} \\
&\quad + \sum_{t=2}^{T} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ S_i^t (1 - G_i^{t-1}) G_j^t \log A_{ij} + S_i^t (1 - G_i^{t-1})(1 - G_j^t) \log(1 - A_{ij}) \right\} \\
&\quad + \sum_{t=2}^{T} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ S_i^t G_i^{t-1} G_j^{t-1} G_j^t \log B_{ij} + S_i^t G_i^{t-1} G_j^{t-1} (1 - G_j^t) \log(1 - B_{ij}) \right\} \\
&\quad + \sum_{t=2}^{T} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ S_i^t G_i^{t-1} (1 - G_j^{t-1}) G_j^t \log C_{ij} + S_i^t G_i^{t-1} (1 - G_j^{t-1})(1 - G_j^t) \log(1 - C_{ij}) \right\}.
\end{aligned}
$$

$$(2.4)$$

Note that $\alpha, \beta, \gamma, \theta$ and u are essentially the parameters of this model and ρ, Φ, A, B and C are the functions of them. Despite its length, equation (2.4) has a clear structure. The

Table 2: Summary of Notation

	Notation	Remark
Parameter	ρ_i	Probability of v_i being the leader of G^1
	u_i	$\rho_i = \frac{\exp(u_i)}{\sum_{k=1}^{n} \exp(u_k)}$
	α	Adjustment factor for remaining leaders
	Φ_{ij}	$\Phi_{ij} = \mathbb{P}(z^t = i \mid z^{t-1} = j)$
	A_{ij}	Probability of v_j being inside the group when v_i is the leader in a newly formed group
	θ_{ij}	$\theta_{ij} = \log \frac{A_{ij}}{1 - A_{ij}}$
	β	Adjustment factor for nodes being inside the previous group
	B_{ij}	Adjusted probability of v_j being inside the group when inside the previous group
	γ	Adjustment factor for nodes being outside the previous group
	C_{ij}	Adjusted probability of v_j being inside the group when outside the previous group
Data	G^t	Group at time t
	z^t	Leader at time t
	S^t	Indicator of the leader at time t, with only one element being 1
Index	n	Size of the network
	T	Number of groups (sample size)

1st line gives the log-likelihood of S. The 2nd line gives the log-likelihood of G^1 given S^1. The 3rd line gives the log-likelihood of G^t given that the current leader z^t is outside the previous group G^{t-1}. The 4th and 5th lines give the log-likelihood of G^t given that z^t is inside G^{t-1}, based on the *In-and-Out procedure*.

Equivalently to (2.4), we can write the likelihood as a product of conditional probabilities:

$$\mathbb{P}(S, G) = \mathbb{P}(S^1)\mathbb{P}(G^1|S^1) \prod_{t=2}^{T} \mathbb{P}(S^t|S^{t-1}) \prod_{t=2}^{T} \mathbb{P}(G^t|S^t, G^{t-1}).$$

This factorization can be represented by a Bayesian network in Figure 1, where a node represents a variable and a directed arc is drawn from node X to node Y if Y is conditioned on X in the factorization. The reader may refer to Jordan et al. (1999) for a comprehensive introduction to Bayesian networks. This Bayesian network is not to be confused with the latent network A – the former is a representation of the dependency structure between variables while the latter reflects the relations between the group members.

Furthermore, the group leaders $z^1, ..., z^T$ are assumed to be latent (and so are $S^1, ..., S^T$) since in many applications only the groups themselves are observable.
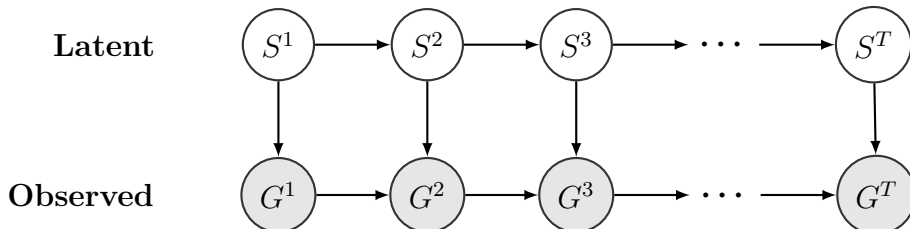

Figure 1: A Bayesian network representing the time-varying hub model. Nodes with dark colors indicate the observed variables.

3   Model fitting

In this section, we propose an algorithm to find the maximum likelihood estimators (MLEs) for α, β, γ, u and θ. With S being the latent variables, an expectation-maximization (EM)

algorithm will be used for this problem. The EM algorithm maximizes the marginal likelihood of the observed data, which is G in our case, by iteratively applying an E-step and an M-step.

Let $\alpha^{\text{old}}, \beta^{\text{old}}, \gamma^{\text{old}}, u^{\text{old}}$ and θ^{old} be the estimates at the current iteration. In the E-step, we calculate the conditional expectation of the complete log-likelihood given G under the current estimate. That is,

$$Q \triangleq Q(\alpha, \beta, \gamma, u, \theta | \alpha^{\text{old}}, \beta^{\text{old}}, \gamma^{\text{old}}, u^{\text{old}}, \theta^{\text{old}}) = \mathbb{E}_{\alpha^{\text{old}}, \beta^{\text{old}}, \gamma^{\text{old}}, u^{\text{old}}, \theta^{\text{old}}} \left[\log \mathbb{P}(S, G) | G \right].$$

In the M-step, we maximize this conditional expectation with respect to the unknown parameters. That is,

$$(\alpha^{\text{new}}, \beta^{\text{new}}, \gamma^{\text{new}}, u^{\text{new}}, \theta^{\text{new}}) = \underset{\alpha, \beta, \gamma, u, \theta}{\arg \max} \; Q(\alpha, \beta, \gamma, u, \theta | \alpha^{\text{old}}, \beta^{\text{old}}, \gamma^{\text{old}}, u^{\text{old}}, \theta^{\text{old}}).$$

It has been proved by Wu (1983) that the EM algorithm converges to a local maximizer of the marginal likelihood. The reader can refer to McLachlan and Krishnan (2008) for a comprehensive introduction to this algorithm. We now give details of the two steps in our context.

3.1 E-step

Since the complete log-likelihood $\log \mathbb{P}(S, G)$ is a linear function of S_i^t ($t = 1, ..., T; i = 1, ..., n$) and $S_i^t S_j^{t-1}$ ($t = 2, ..., T; i = 1, ..., n; j = 1, ..., n$), the computation of its conditional expectation is equivalent to calculating $\mathbb{P}(S_i^t = 1 | G)$ and $\mathbb{P}(S_i^t = 1, S_j^{t-1} = 1 | G)$. From now on, all the conditional probabilities are defined under the current estimates.

A brute-force calculation of these probabilities, such as

$$\mathbb{P}(S_i^t = 1 | G) = \mathbb{P}(z^t = i | G) = \frac{\sum_{z^1} \cdots \sum_{z^{t-1}} \sum_{z^{t+1}} \cdots \sum_{z^t} \mathbb{P}(z^1, ..., z^{t-1}, z^t = i, z^{t+1}, ..., z^T, G)}{\mathbb{P}(G)},$$

is infeasible since the numerator involves a sum of n^{T-1} terms. This is because $G^1, ..., G^T$ are not independent according to our model. An efficient algorithm is needed for all practical purposes.

From Figure 1, it can be seen that the time-varying hub model is similar to the hidden Markov model (HMM). A polynomial-time algorithm for this model, called the forward-backward algorithm, was developed for computing the conditional probabilities. See Smyth et al. (1997); Ghahramani (2001) for tutorials on HMMs and this algorithm.

In the HMM, the observed variable at time t only depends on the corresponding hidden state. But in our model, G^t depends on both the current leader z^t and the previous group G^{t-1}. We develop a new forward-backward algorithm for our model, which which has more steps than the classical algorithm but is also polynomial-time. We describe the algorithm here and leave the detailed derivation and justification to the Appendix.

Define $a = [a_i^t], b = [b_i^t]$ and $c = [c_i^t]$ as $T \times n$ matrices. These matrices are computed by the following recursive procedures.

$$a_i^1 = \mathbb{P}(z^1 = i, G^1) \quad (i = 1, ..., n).$$

$$a_i^t = \sum_{k=1}^{n} a_k^{t-1} \Phi_{ik} \mathbb{P}(G^t | z^t = i, G^{t-1}) \quad (t = 2, ..., T; i = 1, ..., n).$$

$$b_i^T = 1 \quad (i = 1, ..., n).$$

$$b_i^t = \sum_{k=1}^{n} b_k^{t+1} \Phi_{ki} \mathbb{P}(G^{t+1} | z^{t+1} = k, G^t) \quad (t = T - 1, ..., 1; i = 1, ..., n).$$

$$c_i^T = \mathbb{P}(G^T | z^T = i, G^{T-1}) \quad (i = 1, ..., n).$$

$$c_i^t = \sum_{k=1}^{n} c_k^{t+1} \Phi_{ki} \mathbb{P}(G^t | z^t = i, G^{t-1}) \quad (t = T - 1, ..., 2; i = 1, ..., n).$$

With these quantities,

$$\mathbb{P}(S_i^t = 1 | G) = \frac{a_i^t b_i^t}{\sum_k a_k^t b_k^t} \quad (t = 2, ..., T; i = 1, ..., n).$$

$$\mathbb{P}(S_i^t = 1, S_j^{t-1} = 1 | G) = \frac{a_j^{t-1} \Phi_{ij} c_i^t}{\sum_{kl} a_l^{t-1} \Phi_{kl} c_k^t} \quad (t = 2, ..., T; i = 1, ..., n; j = 1, ..., n).$$

The complexity of this algorithm is $O(Tn^2)$.

Note that the first row of c is undefined but is also unused. Also note that the elements of a, b and c will quickly vanish as the recursions progress. Therefore, we renormalize each row to sum to one at each step. One can easily check that this normalization will not affect the conditional probabilities. Finally, it is worth emphasizing that this algorithm gives the exact values of the conditional probabilities in a fixed number of steps – that is, it is not an approximate or iterative method.

3.2 M-step

The M-step is somewhat routine compared with the E-step. First, it is clear that $\{\alpha, u\}$ and $\{\beta, \gamma, \theta\}$ can be handled separately.

We apply the coordinate ascent method (see Boyd and Vandenberghe (2004) for a comprehensive introduction) to iteratively update α and u, as well as β, γ and θ. Since the complete log-likelihood is concave and so is Q, coordinate ascent can guarantee a global maximizer.

At each step, we optimize the log-likelihood over parameter one by one with others being fixed. The procedure is repeated until convergence. At each step, we use the standard Newton-Raphson method to solve each individual optimization problem. Specifically, for a parameter ϕ (here ϕ can represent α, β, γ, u_i or θ_{ij} $(i < j)$), the estimate at $(m+1)$-th iteration is updated by the following formula given its estimate at m-th iteration:

$$\hat{\phi}_{m+1} = \hat{\phi}_m - \left(\frac{\partial^2 Q}{\partial \phi^2} \bigg|_{\phi=\hat{\phi}_m} \right)^{-1} \left(\frac{\partial Q}{\partial \phi} \bigg|_{\phi=\hat{\phi}_m} \right).$$

The calculation of these derivatives is straightforward but tedious, so we will provide the details in the Appendix. We set $\hat{\theta}_{ii} \equiv \infty$ since by definition $A_{ii} \equiv 1$.

As remarked in Section 2.2, the model is not identifiable with respect to u. A standard solution to this problem is to set some $u_i \equiv 0$. But it does not work for our case. This is because for small data sets, some $\hat{\rho}_i$ estimated by the EM algorithm may be zero implying that v_i never became the leader. Furthermore, these zero ρ_i cannot be predetermined since the leaders are unobserved. We observe that without constraint on u_i, the algorithm converges to different $\hat{u}$ with different initial values, but the corresponding $\hat{\rho}$ will be the same. Therefore, identifiability is not the real issue for model fitting.

Finally, it is obvious that the algorithm for the time-varying hub model with known leaders is the same as the M-step with Q being replaced by the original complete log-likelihood.

3.3 Initial value

As many optimization algorithms, the EM algorithm cannot guarantee to find the global maximizer. Ideally, one should use multiple random initial values and find the best solution

by comparing the marginal likelihoods $\mathbb{P}(G)$ under the corresponding estimates.

In principle, $\mathbb{P}(G)$ can be computed by $\sum_k a_k^t b_k^t$ as shown in Section 3.1. But the marginal likelihood vanishes quickly even for a moderate T. Note that we cannot renormalize a and b for the purpose of computing $\mathbb{P}(G)$.

We use half weight index (Dice, 1945; Cairns and Schwager, 1987) as the initial value of A, which is defined by

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

This measure estimates the conditional probability that two nodes co-occur given that one of them is observed, which is a reasonable initial guess of the strength of links. Furthermore, we use zero as the initial values of α, β and γ, and use $\sum_t G_i^t / T$ as the initial value of ρ_i.

4   Simulation studies

In all simulation studies, we fix the size of the network to be $n = 50$ and set $\beta = 3$ and $\gamma = -1$. We generate u_i as independently and identically distributed variables with $N(0, 2)$ and $\rho_i = u_i / \sum_k u_k$. And θ_{ij} $(i < j)$ are generated independently with $N(-2, 1)$. We generate θ_{ij} in this way to control the average link density of the network (≈ 0.12), which is more realistic than a symmetric setting, i.e., $\theta_{ij} \sim N(0, 1)$. For clarification, we will not use the prior information on u and θ in our estimating procedure. That is, we still treat u and θ as unknown fixed parameters in the algorithm. We generate them as random variables for the whole purpose of adding more variations to the parameter setup in our study.

We consider three levels of $\alpha = \log((n-1)/2), \log(n-1), \log(2(n-1))$, which correspond to that a leader from the previous group remains unchanged in the current group with probability $1/3, 1/2, 2/3$ on average, respectively. For each α we try five different sample sizes, $T = 1000, 1500, 2000, 2500, 3000$.

Table 3-5 show the average root of mean squared errors (RMSEs) for A, ρ, α, β and γ over 100 replicates. For each simulation, we compare three different methods, the classical hub model without time-varying structure (HM), the time-varying hub model (TVHM) and the time-varying hub model with known leaders (KTVHM). Note that the classical

Table 3: RMSE for $A, \rho, \alpha, \beta, \gamma$. HM: Hub Model; TVHM: Time-varying Hub Model; KTVHM: Known leader Time-varying Hub Model.

	$n = 50, \alpha = \log((n-1)/2)$								
	RMSE for A			RMSE for ρ					
T	HM	TVHM	KTVHM	HM	TVHM	KTVHM			
1000	0.1243	0.1037	0.0712	0.0085	0.0069	0.0049			
1500	0.1011	0.0837	0.0573	0.0073	0.0058	0.0039			
2000	0.0871	0.0705	0.0493	0.0067	0.0051	0.0034			
2500	0.0772	0.0620	0.0439	0.0062	0.0045	0.0030			
3000	0.0697	0.0557	0.0399	0.0059	0.0040	0.0028			
	RMSE for α			RMSE for β			RMSE for γ		
T	HM	TVHM	KTVHM	HM	TVHM	KTVHM	HM	TVHM	KTVHM
1000	-	0.0776	0.0745	-	0.0601	0.0470	-	0.0337	0.0327
1500	-	0.0584	0.0516	-	0.0406	0.0326	-	0.0265	0.0250
2000	-	0.0541	0.0442	-	0.0321	0.0287	-	0.0221	0.0200
2500	-	0.0482	0.0395	-	0.0267	0.0229	-	0.0201	0.0181
3000	-	0.0452	0.0381	-	0.0219	0.0201	-	0.0166	0.0155

hub model does not contain α, β and γ so we only list the RMSEs of A and ρ for it.

From the tables, our first observation is simply that the RMSEs decrease as sample sizes increase, which is consistent with common sense in statistics.

Secondly, the RMSEs for all the parameters increase as α increases. This phenomenon can be interpreted as follows. With a larger value of α, the correlation between adjacent groups becomes stronger and hence the effective sample size becomes smaller. The ratio of the sample size to the number of parameters decreases with α, which makes the inference more difficult.

Moreover, the discrepancy between the time-varying hub model estimates and the corresponding hub model estimates becomes larger as α increase. This is because the behavior of the time-varying model deviates more from the classical hub model as α increases.

Finally, the estimates from the time-varying hub model with known leaders in general

15

Table 4: RMSE for $A, \rho, \alpha, \beta, \gamma$. HM: Hub Model; TVHM: Time-varying Hub Model; KTVHM: Known leader Time-varying Hub Model.

	$n = 50, \alpha = \log(n-1)$								
	RMSE for A			RMSE for ρ					
T	HM	TVHM	KTVHM	HM	TVHM	KTVHM			
1000	0.1397	0.1091	0.0793	0.0099	0.0073	0.0055			
1500	0.1144	0.0870	0.0631	0.0088	0.0060	0.0044			
2000	0.0995	0.0733	0.0539	0.0082	0.0052	0.0038			
2500	0.0896	0.0646	0.0479	0.0079	0.0047	0.0034			
3000	0.0826	0.0587	0.0437	0.0076	0.0043	0.0031			
	RMSE for α			RMSE for β			RMSE for γ		
T	HM	TVHM	KTVHM	HM	TVHM	KTVHM	HM	TVHM	KTVHM
1000	-	0.0856	0.0864	-	0.0598	0.0651	-	0.0371	0.0373
1500	-	0.0654	0.0640	-	0.0480	0.0439	-	0.0297	0.0279
2000	-	0.0535	0.0518	-	0.0335	0.0297	-	0.0242	0.0237
2500	-	0.0499	0.0450	-	0.0293	0.0253	-	0.0220	0.0213
3000	-	0.0466	0.0417	-	0.0272	0.0229	-	0.0194	0.0184

Table 5: RMSE for $A, \rho, \alpha, \beta, \gamma$. HM: Hub Model; TVHM: Time-varying Hub Model; KTVHM: Known leader Time-varying Hub Model.

	$n = 50, \alpha = \log(2(n-1))$						
	RMSE for A			RMSE for ρ			
T	HM	TVHM	KTVHM	HM	TVHM	KTVHM	
1000	0.1631	0.1225	0.0945	0.0124	0.0081	0.0068	
1500	0.1366	0.0956	0.0730	0.0113	0.0067	0.0055	
2000	0.1199	0.0813	0.0617	0.0106	0.0060	0.0048	
2500	0.1087	0.0712	0.0541	0.0102	0.0054	0.0043	
3000	0.1015	0.0636	0.0491	0.0099	0.0049	0.0039	

	RMSE for α			RMSE for β			RMSE for γ		
T	HM	TVHM	KTVHM	HM	TVHM	KTVHM	HM	TVHM	KTVHM
1000	-	0.1382	0.1092	-	0.0845	0.1059	-	0.0400	0.0460
1500	-	0.0871	0.0783	-	0.0655	0.0659	-	0.0321	0.0343
2000	-	0.0679	0.0648	-	0.0556	0.0540	-	0.0247	0.0256
2500	-	0.0586	0.0536	-	0.0452	0.0422	-	0.0217	0.0225
3000	-	0.0504	0.0471	-	0.0382	0.0359	-	0.0193	0.0193

outperform the estimates from the other models because the model is correctly specified and has the most information. When $\alpha = \log(2(n-1))$, the average RMSEs of $\hat{\beta}$ and $\hat{\gamma}$ by the time-varying model with known leaders are slightly higher than the estimates by the time-varying model without known leaders when T is small. This unusual phenomenon requires further investigation. But the discrepancy is very small relative to the magnitude of the true parameters. This phenomenon may be due to the bias of $\hat{\beta}$ and $\hat{\gamma}$: The parameters θ_{ij} are asymmetric, i.e., $\theta_{ij} \sim N(-2,1)$. The relatively small link density of the graph will cause that a number of pairs v_i and v_j never co-occur in any groups, even though the true link strength is non-zero. But we have to estimate A_{ij} with $\hat{A}_{ij} = 0$ and θ_{ij} with $\hat{\theta}_{ij} = -\infty$ in this case. This also introduces bias on the estimations of β and γ.

5　Data example

In this section, we study a data set of groups formed by chimpanzees. This data set is compiled from the research results of the Kibale Chimpanzee Project, which is a long-term field study of the behavior, ecology, and physiology of wild chimpanzees (https://kibalechimpanzees.wordpress.com/).

We focus on grouping behavior in our analysis. We analyze the grouped data collected from January 1, 2009 to June 30, 2009 (Kibale Chimpanzee Project, 2011). The group identification was taken at 1pm daily during this time period. If there is no group observed at 1pm for a given day, it is not included in the data. Only one group is observed at 1pm in 75.29% of the remaining days over this period of six months. In the other days, multiple groups (usually two) are observed at 1pm. For these cases, we keep the group which has the most overlap with the previous group in our analysis. We use the Jaccard index to measure the overlap between two groups G^{t-1} and G^t,

$$J(G^{t-1}, G^t) = \frac{\sum_{j=1}^n G_j^{t-1} G_j^t}{\sum_{j=1}^n [n - (1 - G_j^{t-1})(1 - G_j^t)]},$$

where the numerator is the size of the intersection of two groups and the denominator is the size of their union. One may refer to Liben-Nowell and Kleinberg (2007) for an introduction to this measure.

Moreover, five chimpanzees never appear in any group and thus are removed. After the preprocessing, the data set contains 170 groups on 40 chimpanzees.

Figure 2 illustrates the data set by grayscale with rows representing the groups over time and columns representing the chimpanzees. Black indicates $G_i^t = 1$ at location (t, i) while white indicates $G_i^t = 0$. The pattern in Figure 2 clearly demonstrates the existence of dependency between groups.

Figure 2 also shows the inferred grouped leaders indicated in red and the inferred segments separated by blue lines. By the inferred grouped leader for G^t, we mean that the chimpanzee with the highest posterior probability being the leader given G^t. We can see from the figure that the leaders remain a certain level of stability, which is consistent with the estimates of α ($= 1.7291$). Also, recall that by our definition, a new segment starts if the current leader is not within the previous group. From Figure 2, the inferred segments are coincident with the visualization of the data set.

Figure 3 shows the result of estimated adjacency matrices by the classical hub model and the time-varying hub model, respectively. As in the previous figure, darker color indicates higher value of $\hat{A}_{ij}$. The red blocks indicate clusters of chimpanzees in a biological sense. The first cluster consists of 12 adult males and each of the rest nine clusters consists of an adult female and its children. From the estimates from both the classical hub model and the time-varying hub model, there are strong connections within these biological clusters. This suggests in this community of the chimpanzees, adult males usually do activities together but females usually stay with their children.

The graph density ($= 0.2286$) of the estimated network by the classical hub model is larger than the corresponding value ($= 0.1973$) of the time-varying hub model. This fact is consistent with the estimated values of the adjustment factors, $\hat{\beta} = 2.5703$ and $\hat{\gamma} = -0.1922$. The magnitude of $\hat{\beta}$ is larger than the magnitude of $\hat{\gamma}$, which suggests that individuals have a stronger tendency to join a group than leave a group. In other words, the groups may start with small size and grow larger over time. This phenomenon can be visualized in Figure 2. But the classical hub model does not take this time-varying effect into account and thus has to give a denser estimate of the adjacency matrix.

The significance of α, β and γ is tested by the parametric bootstrap method (Efron and

Tibshirani, 1994). This method was applied to HMMs and showed a good performance (Visser et al., 2000). We generate 5000 independent data sets from the fitted time-varying hub model and compute the MLEs for each simulated data set. Figure 4 shows the histograms of the MLEs for α, β and γ. The 95% bootstrap confidence intervals for α, β and γ are (1.2177, 1.9774), (2.0710, 2.8944) and (-0.5208, 0.0410), respectively, which shows that the effects of α and β are significant while γ is not at the significance level of 0.05. This further supports the remark in the previous paragraph – chimpanzees have a stronger tendency to join a group than leave a group in this data set.

6 Summary and discussion

In this article, we generalize the idea of the hub model and propose a novel model for time-varying grouped data. This new model allows for dependency between groups. Specifically, the group leaders follow a Markov chain, and a group is either a transformation of the previous group or a new start, depending on whether the current leader is within the previous group. An EM algorithm is applied to this model with a polynomial-time algorithm being developed for the E-step.

The time-varying hub model can be further extended in the following directions which we will explore for future study. First, a group may contain zero or multiple hubs. Second, multiple groups may exist at the same time (some of these groups may be unobserved). It is worth mentioning that these generalizations will significantly increase model complexity. Therefore, the total number of possible leader needs to be limited. A method by the author and a collaborator (Weko and Zhao, 2017) was proposed to reduce this upper bound. More test-based and penalization methods are under development.

Furthermore, we also plan to investigate the theoretical properties of the proposed model. When the size of the network is being fixed and the number of observed groups goes to infinity, the theoretical properties of the MLE may be studied via a standard theory of the Markov Chain. The case that the size of the network also goes to infinity is more intriguing but more complicated since the number of parameters diverges.

Acknowledgements

A   Forward-backward algorithm in the E-steps

We derive the forward-backward algorithm for the time-varying hub model introduced in Section 3.1. Before proceeding, we state two propositions of Bayesian networks. These results (or the equivalent forms) can be found in a standard textbook or tutorial on Bayesian networks, for example, Jordan et al. (1999). Here we follow Ghahramani (2001).

Proposition A.1. *Each node is conditionally independent from its non-descendents given its parents. Here the node X is a parent of another node Y if there is a directed arc from X to Y and if so, Y is a child of X. The descendents of a node are its children, children's children, etc.*

Proposition A.2. *Two disjoint sets of nodes $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given another set $\mathcal{C}$, if on every undirected path between a node in $\mathcal{A}$ and a node in $\mathcal{B}$, there is a node X in $\mathcal{C}$ which is not a child of both the previous and following nodes in the path.*

Define $G^{s:t}$ as a collection of groups from time s to time t.

Let $a_i^t = \mathbb{P}(z^t = i, G^{1:t})$. Then

$$
\begin{aligned}
a_i^t &= \sum_{k=1}^{n} \mathbb{P}(z^t = i, z^{t-1} = k, G^{1:t-1}, G^t) \\
&= \sum_{k=1}^{n} \mathbb{P}(z^{t-1} = k, G^{1:t-1})\mathbb{P}(z^t = i | z^{t-1} = k, G^{1:t-1})\mathbb{P}(G^t | z^t = i, z^{t-1} = k, G^{1:t-1}) \\
&= \sum_{k=1}^{n} a_k^{t-1} \Phi_{ik} \mathbb{P}(G^t | z^t = i, G^{t-1}).
\end{aligned}
$$

The last equation holds by Proposition A.1.

Similarly, let $b_i^t = \mathbb{P}(G^{t+1:T}|z^t = i, G^t)$. Then

$$
\begin{aligned}
b_i^t &= \sum_{k=1}^n \mathbb{P}(z^{t+1} = k, G^{t+1}, G^{t+2:T}|z^t = i, G^t) \\
&= \sum_{k=1}^n \mathbb{P}(G^{t+2:T}|z^{t+1} = k, G^{t+1}, z^t = i, G^t)\mathbb{P}(z^{t+1} = k|z^t = i, G^t)\mathbb{P}(G^{t+1}|z^{t+1} = k, z^t = i, G^t) \\
&= \sum_{k=1}^n b_k^{t+1}\Phi_{ki}\mathbb{P}(G^{t+1}|z^{t+1} = k, G^t).
\end{aligned}
$$

In the last equation, $\mathbb{P}(G^{t+2:T}|z^{t+1} = k, G^{t+1}, z^t = i, G^t)$ holds by Proposition A.2. This is because a path from $\{G^{t+2:T}\}$ to $\{z^t, G^t\}$ must pass z^{t+1} or G^{t+1}. If it only passes one of these two variables, then we can take that variable as X in Proposition A.2. If it passes both, then take z^{t+1} as X. The rest part of the last equation holds by A.1.

The computation of a and b is essentially the same as in the classical forward-backward algorithm for the HMM with minor modifications. Different from the HMM, the dependence between the current and the previous groups requires another quantity c.

Let $c_i^t = \mathbb{P}(G^{t:T}|z^t = i, G^{t-1})$.

$$
\begin{aligned}
c_i^t &= \sum_{k=1}^n \mathbb{P}(z^{t+1} = k, G^{t+1:T}, G^t|z^t = i, G^{t-1}) \\
&= \sum_{k=1}^n \mathbb{P}(G^{t+1:T}|z^{t+1} = k, G^t, z^t = i, G^{t-1})\mathbb{P}(z^{t+1} = k|z^t = i, G^{t-1})\mathbb{P}(G^t|z^{t+1} = k, z^t = i, G^{t-1}) \\
&= \sum_{k=1}^n c_k^{t+1}\Phi_{ki}\mathbb{P}(G^t|z^t = i, G^{t-1}).
\end{aligned}
$$

The last equation can be justified by a similar argument as before.

Since

$$
\begin{aligned}
&\mathbb{P}(z^t = i, G^{1:T}) \\
=&\mathbb{P}(z^t = i, G^{1:t})\mathbb{P}(G^{t+1:T}|z^t = i, G^{1:t}) \\
=&\mathbb{P}(z^t = i, G^{1:t})\mathbb{P}(G^{t+1:T}|z^t = i, G^t),
\end{aligned}
$$

$$
\mathbb{P}(S_i^t = 1|G) = \frac{a_i^t b_i^t}{\sum_k a_k^t b_k^t}.
$$

Similarly,

$$\mathbb{P}(z^t = i, z^{t-1} = j, G^{1:T})$$

$$=\mathbb{P}(z^{t-1} = j, G^{1:t-1})\mathbb{P}(z^t = i|z^{t-1} = j, G^{1:t-1})\mathbb{P}(G^{t:T}|z^t = i, z^{t-1} = j, G^{1:t-1})$$

$$=a_j^{t-1}\Phi_{ij}\mathbb{P}(G^{t:T}|z^t = i, G^{t-1}),$$

which implies

$$\mathbb{P}(S_i^t = 1, S_j^{t-1} = 1|G) = \frac{a_j^{t-1}\Phi_{ij}c_i^t}{\sum_{kl} a_l^{t-1}\Phi_{kl}c_k^t}.$$

B   Derivatives of Q

We give the first and second derivatives of Q with respect to $\alpha, \beta, \gamma, u_i$ and θ_{ij}, which are used in the coordinate ascent method introduced in Section 3.2.

Define,

$$R_i^t = \mathbb{P}(S_i^t = 1|G),$$

$$V_{ij} = \sum_{t=2}^{T} \mathbb{P}(S_i^t = 1, S_j^{t-1} = 1|G),$$

$$D_{ij}^1 = R_i^1 G_j^1 + \sum_{t=2}^{T} R_i^t (1 - G_i^{t-1}) G_j^t,$$

$$D_{ij}^2 = R_i^1 (1 - G_j^1) + \sum_{t=2}^{T} R_i^t (1 - G_i^{t-1})(1 - G_j^t),$$

$$D_{ij}^3 = \sum_{t=2}^{T} R_i^t G_i^{t-1} G_j^{t-1} G_j^t,$$

$$D_{ij}^4 = \sum_{t=2}^{T} R_i^t G_i^{t-1} G_j^{t-1}(1 - G_j^t),$$

$$D_{ij}^5 = \sum_{t=2}^{T} R_i^t G_i^{t-1}(1 - G_j^{t-1}) G_j^t,$$

$$D_{ij}^6 = \sum_{t=2}^{T} R_i^t G_i^{t-1}(1 - G_j^{t-1})(1 - G_j^t).$$

Therefore,

$$
\begin{aligned}
Q = \ &\sum_{i=1}^{n} R_i^1 \left[u_i - \log \left\{ \sum_{k=1}^{n} \exp(u_k) \right\} \right] \\
&+ \sum_{i=1}^{n} V_{ii} \left[u_i + \alpha - \log \left\{ \sum_{k=1}^{n} \exp(u_k + \alpha I(k=i)) \right\} \right] \\
&+ \sum_{i=1}^{n} \sum_{j \neq i} V_{ij} \left[u_i - \log \left\{ \sum_{k=1}^{n} \exp(u_k + \alpha I(k=j)) \right\} \right] \\
&+ \sum_{ij} \left[D_{ij}^1 \log \frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}} + D_{ij}^2 \log \frac{1}{1 + e^{\theta_{ij}}} \right. \\
&\qquad + D_{ij}^3 \log \frac{e^{\theta_{ij} + \beta}}{1 + e^{\theta_{ij} + \beta}} + D_{ij}^4 \log \frac{1}{1 + e^{\theta_{ij} + \beta}} \\
&\qquad \left. + D_{ij}^5 \log \frac{e^{\theta_{ij} + \gamma}}{1 + e^{\theta_{ij} + \gamma}} + D_{ij}^6 \log \frac{1}{1 + e^{\theta_{ij} + \gamma}} \right].
\end{aligned}
$$

The first and second order derivatives are given as follows,

$$\frac{\partial Q}{\partial \alpha} = \sum_{i=1}^{n} V_{ii} + \sum_{j=1}^{n} \left[\sum_{i=1}^{n} V_{ij}\right] \left[-\frac{\exp(u_j + \alpha)}{\sum_{k=1}^{n} \exp(u_k + \alpha I(k=j))}\right],$$

$$\frac{\partial^2 Q}{\partial \alpha^2} = \sum_{j=1}^{n} \left[\sum_{i=1}^{n} V_{ij}\right] \left[-\frac{\exp(u_j + \alpha)\sum_{k=1}^{n} \exp(u_k + \alpha I(k=j)) - \exp(u_j + \alpha)\exp(u_j + \alpha)}{(\sum_{k=1}^{n} \exp(u_k + \alpha I(k=j)))^2}\right],$$

$$\frac{\partial Q}{\partial u_r} = R_r^1 + \left[\sum_{i=1}^{n} R_i^1\right] \left[-\frac{\exp(u_r)}{\sum_{k=1}^{n} \exp(u_k)}\right],$$

$$\frac{\partial^2 Q}{\partial u_r^2} = \left[\sum_{i=1}^{n} R_i^1\right] \left[-\frac{\exp(u_r)\sum_{k=1}^{n} \exp(u_k) - \exp(u_r)\exp(u_r)}{(\sum_{k=1}^{n} \exp(u_k))^2}\right]$$

$$+ \sum_{j=1}^{n} \left[\sum_{i=1}^{n} V_{ij}\right] \left[-\frac{\exp(u_r + \alpha I(r=j))\sum_{k=1}^{n} \exp(u_k + \alpha I(k=j)) - (\exp(u_r + \alpha I(r=j)))^2}{(\sum_{k=1}^{n} \exp(u_k + \alpha I(k=j)))^2}\right]$$

$$+ \sum_{j=1}^{n} V_{rj} + \sum_{j=1}^{n} \left[\sum_{i=1}^{n} B_{ij}\right] \left[-\frac{\exp(u_r + \alpha I(r=j))}{\sum_{k=1}^{n} \exp(u_k + \alpha I(k=j))}\right],$$

$$\frac{\partial Q}{\partial \beta} = \sum_{i \neq j} D_{ij}^3 - (D_{ij}^3 + D_{ij}^4)\frac{e^{\theta_{ij}+\beta}}{1 + e^{\theta_{ij}+\beta}},$$

$$\frac{\partial^2 Q}{\partial \beta^2} = -\sum_{i \neq j} (D_{ij}^3 + D_{ij}^4)\frac{e^{\theta_{ij}+\beta}}{(1 + e^{\theta_{ij}+\beta})^2},$$

$$\frac{\partial Q}{\partial \gamma} = \sum_{i \neq j} D_{ij}^5 - (D_{ij}^5 + D_{ij}^6)\frac{e^{\theta_{ij}+\gamma}}{1 + e^{\theta_{ij}+\gamma}},$$

$$\frac{\partial^2 Q}{\partial \gamma^2} = -\sum_{i \neq j} (D_{ij}^5 + D_{ij}^6)\frac{e^{\theta_{ij}+\gamma}}{(1 + e^{\theta_{ij}+\gamma})^2},$$

$$\frac{\partial Q}{\partial \theta_{ij}} = (D_{ij}^1 + D_{ji}^1) - (D_{ij}^1 + D_{ji}^1 + D_{ij}^2 + D_{ji}^2)\frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}}$$

$$+ (D_{ij}^3 + D_{ji}^3) - (D_{ij}^3 + D_{ji}^3 + D_{ij}^4 + D_{ji}^4)\frac{e^{\theta_{ij}+\beta}}{1 + e^{\theta_{ij}+\beta}}$$

$$+ (D_{ij}^5 + D_{ji}^5) - (D_{ij}^5 + D_{ji}^5 + D_{ij}^6 + D_{ji}^6)\frac{e^{\theta_{ij}+\gamma}}{1 + e^{\theta_{ij}+\gamma}},$$

$$\frac{\partial^2 Q}{\partial \theta_{ij}^2} = -(D_{ij}^1 + D_{ji}^1 + D_{ij}^2 + D_{ji}^2)\frac{e^{\theta_{ij}}}{(1 + e^{\theta_{ij}})^2}$$

$$- (D_{ij}^3 + D_{ji}^3 + D_{ij}^4 + D_{ji}^4)\frac{e^{\theta_{ij}+\beta}}{(1 + e^{\theta_{ij}+\beta})^2}$$

$$- (D_{ij}^5 + D_{ji}^5 + D_{ij}^6 + D_{ji}^6)\frac{e^{\theta_{ij}+\gamma}}{(1 + e^{\theta_{ij}+\gamma})^2}.$$

References

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

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

Blundell, C., Beck, J., and Heller, K. A. (2012). Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization.* Cambridge University Press, New York.

Cairns, S. J. and Schwager, S. J. (1987). A comparison of association indices. *Animal Behavior*, 35.

Dice, L. R. (1945). Measures of the amount of ecological association between species. *Ecology*, 26:297–302.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press.

Getoor, L. and Diehl, C. P. (2005). Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International journal of pattern recognition and artificial intelligence*, 15(01):9–42.

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.

Guo, F., Blundell, C., Wallach, H., and Heller, K. (2015). The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Artificial Intelligence and Statistics*, pages 315–323.

Hanneke, S. and Xing, E. P. (2007). Discrete temporal models of social networks. *Lecture Notes in Computer Science*, 4503:115.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kibale Chimpanzee Project (2011). Kpc january-june 2009.

Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley and Sons, Inc.

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

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Robinson, L. F. and Priebe, C. E. (2013). Detecting time-dependent structure in network data via a new class of latent process models. arXiv:1212.3587.

Smyth, P., Heckerman, D., and Jordan, M. I. (1997). Probabilistic independence networks for hidden markov probability models. *Neural computation*, 9(2):227–269.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395.

Visser, I., Raijmakers, M. E., and Molenaar, P. (2000). Confidence intervals for hidden markov model parameters. *British journal of mathematical and statistical psychology*, 53(2):317–327.

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

Weko, C. and Zhao, Y. (2017). Penalized component hub models. *Social Networks*, 49:27–36.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

Zhao, Y. (2017). A survey on theoretical advances of community detection in networks. *Wiley Interdisciplinary Reviews: Computational Statistics*.

Zhao, Y. and Weko, C. (2017). Network inference from grouped observations using hub models. *Statistica Sinica*. In press.
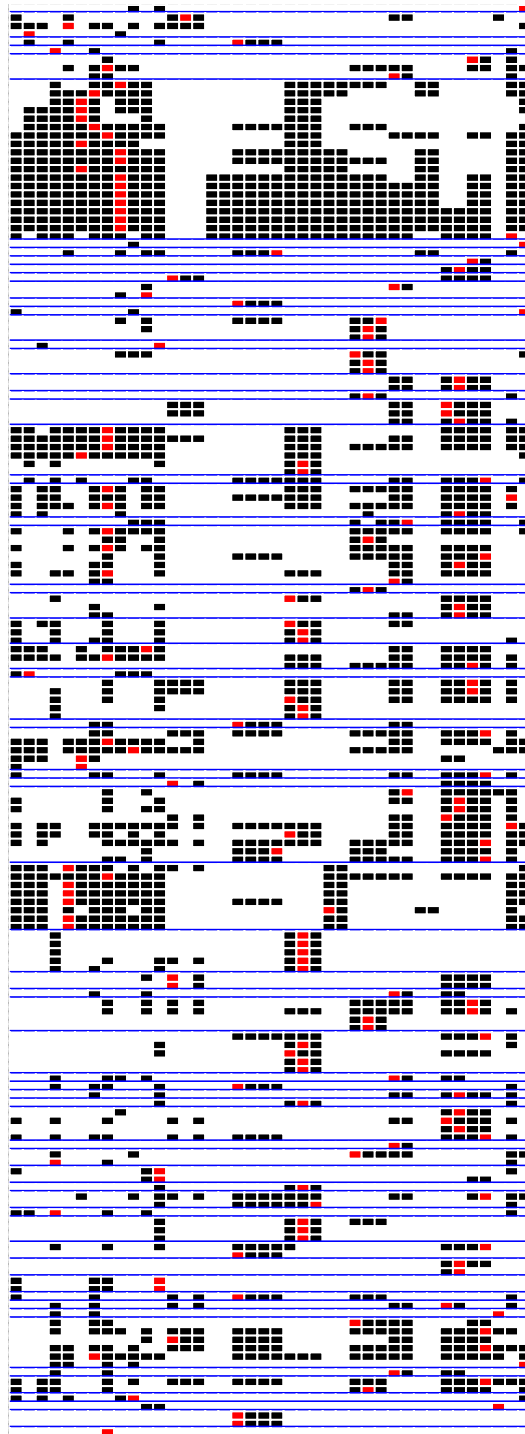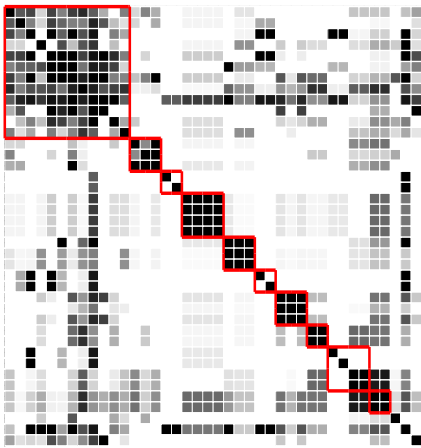
Figure 2: Grayscale plot for the chimpanzee data. The 170 rows represent the chimpanzees and the 40 columns represent the groups over time. Black indicates the presence of the group membership. Red indicates the inferred group leaders. The blue lines separate the inferred segments.

29

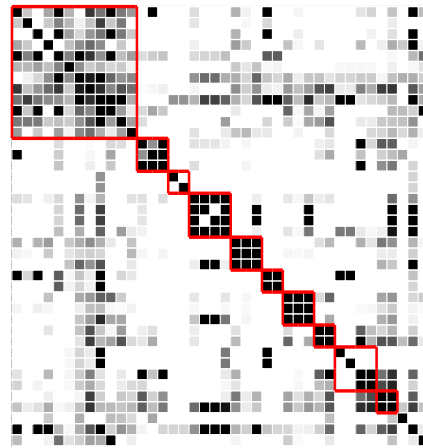(a) HM   (b) TVHM

Figure 3: Grayscale plot for the estimated adjacency matrices from the chimpanzee data by the classical and the time-varying hub model. The rows and the columns represent the 40 chimpanzees. Darker colors indicate stronger relationships. The red blocks indicate the biological clusters of chimpanzees.
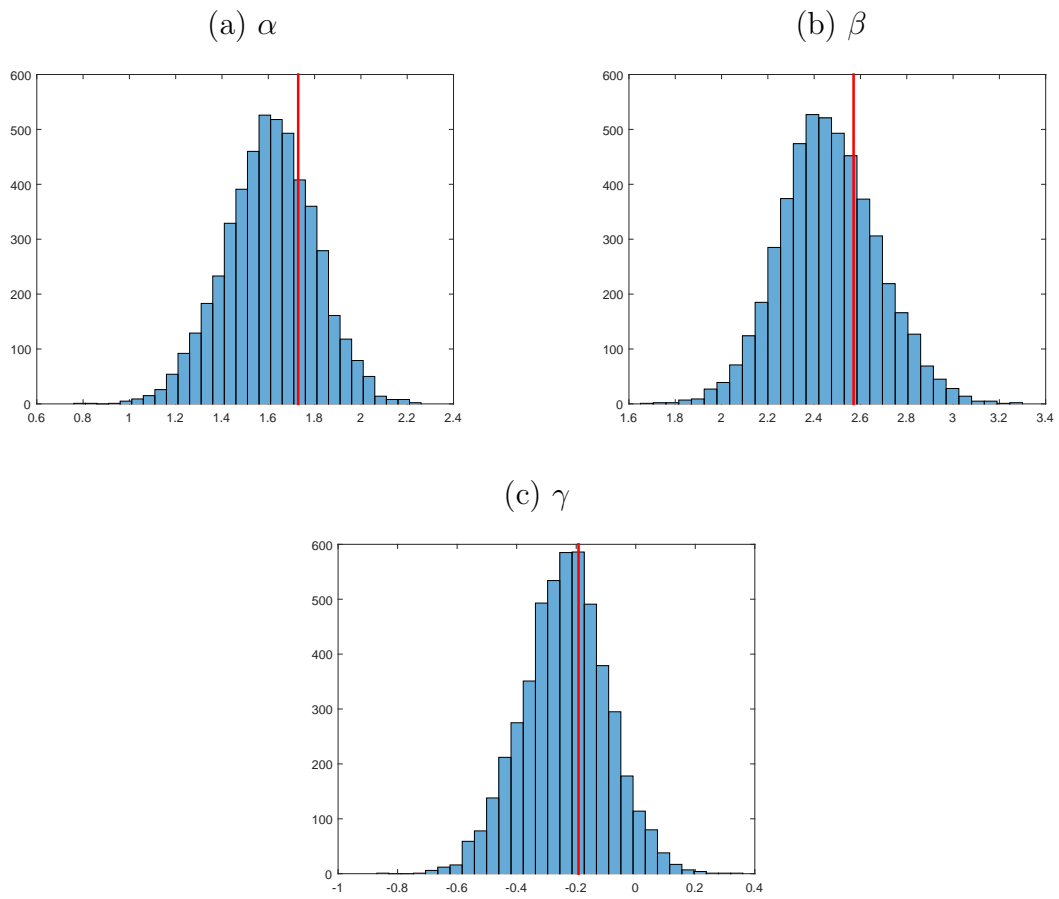
Figure 4: Histograms of estimates from parametric bootstrap samples. Red lines indicate the estimated values from the original data set.