Journal of Machine Learning Research 25 (2024) 1-40 Submitted 11/22; Revised 8/23; Published 1/24

Modeling Random Networks with Heterogeneous Reciprocity

Daniel Cirkovic CIRKOVD@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843, USA

Tiandong Wang* TD_WANG@FUDAN.EDU.CN
Shanghai Center for Mathematical Sciences
Fudan University
Shanghai 200438, China

Editor: Tina Eliassi-Rad

Abstract

Reciprocity, or the tendency of individuals to mirror behavior, is a key measure that describes information exchange in a social network. Users in social networks tend to engage in different levels of reciprocal behavior. Differences in such behavior may indicate the existence of communities that reciprocate links at varying rates. In this paper, we develop methodology to model the diverse reciprocal behavior in growing social networks. In particular, we present a preferential attachment model with heterogeneous reciprocity that imitates the attraction users have for popular users, plus the heterogeneous nature by which they reciprocate links. We compare Bayesian and frequentist model fitting techniques for large networks, as well as computationally efficient variational alternatives. Cases where the number of communities is known and unknown are both considered. We apply the presented methods to the analysis of Facebook and Reddit networks where users have non-uniform reciprocal behavior patterns. The fitted model captures the heavy-tailed nature of the empirical degree distributions in the datasets and identifies multiple groups of users that differ in their tendency to reply to and receive responses to wallposts and comments.

Keywords: Variational inference, community detection, preferential attachment, Bayesian methods

1 Introduction

A frequent goal in the statistical inference of social networks is to develop models that adequately capture and quantify common types of user interaction. One such feature is the propensity of users to generate links with other users that already have attracted a large number of links (Newman, 2001; Jeong et al., 2003). To model this "rich get richer" self-organizing feature of nodes in a growing network, Barabási and Albert (1999) developed the preferential attachment (PA) model. The classical preferential attachment model posits that as users enter a growing network, they connect with other users with probability proportional to their degree. This simple mechanism produces power-law degree distributions, yet another feature of many real-world networks (Mislove et al., 2007). Since its inception, many generalizations of the preferential attachment model have been developed to capture

*. T. Wang is the corresponding author.

more features of growing networks (Bhamidi et al., 2015; Hajek and Sankagiri, 2019; Wang and Zhang, 2022; Wang and Resnick, 2023a).

Another common feature of online social networks is a significant degree of reciprocity (see Newman et al., 2002; Zlatić and Štefančić, 2011, for example). Reciprocity describes the tendency of users to reply to links and is typically measured by the proportion of reciprocal links in a network (Jiang et al., 2015). A recent study by Wang and Resnick (2022a) found that the traditional directed preferential attachment model often produces a negligible proportion of reciprocal links. Motivated by this finding, Wang and Resnick (2022b) and Cirkovic et al. (2023a) developed a preferential attachment model with reciprocity that is a more realistic choice for fitting to social networks. The model assumes that upon the generation of a link between nodes through the typical preferential attachment scheme, the users reciprocate the link with a probability $\rho \in (0, 1)$ that is common to all users in the network. The model was used to analyze a Facebook wallpost network.

Although an improvement, the model of Cirkovic et al. (2023a) fails to account for the heterogeneity of reciprocal behavior in a social network. In reality, it is naïve to assume all users in a large network engage in similar levels of reciprocity. Such an assumption has caused Cirkovic et al. (2023a) to remove a subset of nodes that apparently engaged in dissimilar reciprocal behavior from their analysis of the Facebook wallpost network. Further, when a link is made between two nodes u and v, the decision of whether or not to reciprocate the link depends on the direction of the original link, (u, v) or (v, u). For example, a celebrity in a social network may be less likely to reply to a message sent by a fan, whereas a fan is very likely to respond to a message sent by the celebrity. Recently, Wang and Resnick (2023b) relax the assumption of having only one reciprocity parameter ρ to the case where reciprocity probabilities are different for users belonging to different communication classes. Theoretical results in Wang and Resnick (2022b) are obtained by assuming no new edge is added between existing nodes.

In this paper, we consider a further generalization of the model presented in Wang and Resnick (2023b) to allow for more realistic assumptions, i.e. heterogeneous, asymmetric reciprocity as well as edges between existing nodes. We assume that each user in the network is equipped with a communication class that governs its tendency to reciprocate edges. In the network generation process, initial edges between nodes are generated via preferential attachment, while the decision to reciprocate the edge is decided by a stochastic blockmodel-like scheme. We describe three methods to fit such a model to observed networks, both when the number of communication classes is known and unknown. Specifically, we propose a fully Bayesian approach, along with variationally Bayesian and frequentist approaches. The approaches and their performance on synthetic networks are then compared through simulation studies. Finally, we reconsider the Facebook wall post network as in Cirkovic et al. (2023a), as well as a newly analyzed Reddit network, and use the heterogeneous reciprocal preferential attachment model to glean new insights into communication patterns on Facebook and Reddit.

2 The PA Model with Heterogeneous Reciprocity

In this section, we formulate the preferential attachment model with heterogeneous reciprocity. Based on the model definition, we present the likelihood of a graph observed from

the proposed preferential attachment model. Said likelihood that will form the basis of all of our proposed inferential procedures.

2.1 The model

Let $G(n)$ be the graph after n steps and $V(n)$ be the set of nodes in $G(n)$. Attach to each node v a communication type W_v, where $\{W_v, v \geq 1\}$ are iid random variables with

$$\mathbb{P}(W_v = r) = \pi_r, \qquad \text{for} \quad \sum_{r=1}^{K} \pi_r = 1, \tag{1}$$

where K represents the number of communication classes. Define the vector $\boldsymbol{\pi} \equiv (\pi_r)_r$. Let $W(n) := \{W_v : v \in V(n)\}$ denote the set of group types for all nodes in $G(n)$. Throughout we assume that the communication group of node v is generated upon creation and remains unchanged throughout the graph evolution. Also, denote the set of directed edges in $G(n)$ by

$$E(n) := \{(u, v) : u, v \in V(n)\}.$$

Throughout this paper, we always assume $G(n) = (V(n), E(n), W(n))$, for $n \geq 0$.

We initialize the model with seed graph $G(0)$. $G(0)$ consists of $|V(0)|$ nodes, each of which is also endowed with its own communication class randomly according to (1). The edges $E(0)$ will have no impact on inference other than setting the initial degree distribution. For each new edge (u, v) with $W_u = r, W_v = m$, the reciprocity mechanism adds its reciprocal counterpart (v, u) instantaneously with probability $\rho_{m,r} \in [0, 1]$, for $m, r \in \{1, 2, \ldots, K\}$. Here $\rho_{m,r}$ measures the probability of adding a reciprocal edge from a node in group m to a node in group r. Note that the matrix $\boldsymbol{\rho} := (\rho_{m,r})_{m,r}$ is not necessarily a stochastic matrix, but can be an arbitrary matrix in $M_{K\times K}([0, 1])$, the set of all $K \times K$ matrices with entries belonging to $[0, 1]$.

We now describe the evolution of the network $G(n+1)$ from $G(n)$. Let $\left(D_v^{\text{in}}(n), D_v^{\text{out}}(n)\right)$ be the in- and out-degrees of node $v \in V(n)$, and we use the convention that $D_v^{\text{in}}(n) = D_v^{\text{out}}(n) = 0$ if $v \notin V(n)$.

1. With probability $\alpha \in [0, 1]$, add a new node $|V(n)| + 1$ with a directed edge $(|V(n)| + 1, v)$, where $v \in V(n)$ is chosen with probability

$$\frac{D_v^{\text{in}}(n) + \delta_{\text{in}}}{\sum_{v\in V(n)}(D_v^{\text{in}}(n) + \delta_{\text{in}})} = \frac{D_v^{\text{in}}(n) + \delta_{\text{in}}}{|E(n)| + \delta_{\text{in}}|V(n)|}, \tag{2}$$

   where $\delta_{\text{in}} > 0$ is an offset parameter, and update the node set $V(n + 1) = V(n) \cup \{|V(n)| + 1\}$ and $W(n + 1) = W(n) \cup \{W_{|V(n)|+1}\}$. The new node $|V(n)| + 1$ belongs to group r with probability π_r. If node v belongs to group m, then a reciprocal edge $(v, |V(n)| + 1)$ is added with probability $\rho_{m,r}$. Update the edge set as $E(n + 1) = E(n) \cup \{(|V(n)| + 1, v), (v, |V(n)| + 1)\}$. If the reciprocal edge is not created, set $E(n + 1) = E(n) \cup \{(|V(n)| + 1, v)\}$.

2. With probability $\beta \in [0, 1-\alpha]$, generate a directed edge (u, v) between two existing nodes $u, v \in V(n)$ with probability

$$\frac{D_v^{\text{in}}(n) + \delta_{\text{in}}}{\sum_{v \in V(n)} (D_v^{\text{in}}(n) + \delta_{\text{in}})} \frac{D_v^{\text{out}}(n) + \delta_{\text{out}}}{\sum_{v \in V(n)} (D_v^{\text{out}}(n) + \delta_{\text{out}})} = \frac{D_v^{\text{in}}(n) + \delta_{\text{in}}}{|E(n)| + \delta_{\text{in}}|V(n)|} \frac{D_v^{\text{out}}(n) + \delta_{\text{out}}}{|E(n)| + \delta_{\text{out}}|V(n)|}, \tag{3}$$

where $\delta_{\text{out}} > 0$ is also an offset parameter. If node u belongs to group r and node v belongs to group m, then a reciprocal edge (v, u) is added with probability $\rho_{m,r}$. Update the edge set as $E(n+1) = E(n) \cup \{(u, v), (v, u)\}$. If the reciprocal edge is not created, set $E(n+1) = E(n) \cup \{(u, v)\}$. Finally, update $V(n+1) = V(n)$ and $W(n+1) = W(n)$.

3. With probability $\gamma \equiv 1 - \alpha - \beta$, add a new node $|V(n)| + 1$ with a directed edge $(v, |V(n)| + 1)$, where $v \in V(n)$ is chosen with probability

$$\frac{D_v^{\text{out}}(n) + \delta_{\text{out}}}{\sum_{v \in V(n)} (D_v^{\text{out}}(n) + \delta_{\text{out}})} = \frac{D_v^{\text{out}}(n) + \delta_{\text{out}}}{|E(n)| + \delta_{\text{out}}|V(n)|}, \tag{4}$$

and update the node set $V(n+1) = V(n) \cup \{|V(n)| + 1\}$, $W(n+1) = W(n) \cup \{W_{|V(n)|+1}\}$. The new node $|V(n)|+1$ belongs to group r with probability π_r. If node v belongs to group m, then a reciprocal edge $(|V(n)|+1, v)$ is added with probability $\rho_{r,m}$. Update the edge set as $E(n+1) = E(n) \cup \{(v, |V(n)|+1, v), (|V(n)|+1, v)\}$. If the reciprocal edge is not created, set $E(n+1) = E(n) \cup \{(v, |V(n)| + 1)\}$.

Let $\{J_k\}$ be iid Categorical random variables that indicate under which scenario the transition from $G(k)$ to $G(k+1)$ has occurred. That is, $\mathbb{P}(J_k = 1) = \alpha$, $\mathbb{P}(J_k = 2) = \beta$ and $\mathbb{P}(J_k = 3) = 1 - \alpha - \beta$. At each step k, we denote the outcome of the reciprocal event via R_k where $R_k = 1$ if a reciprocal edge is added and $R_k = 0$ otherwise.

2.2 Likelihood inference

Suppose we observe the evolution of the graph sequence $\{G(k)\}_{k=0}^n$ so that we have the edges $e_k = E(k) \setminus E(k-1)$ added at each step according to the description in Section 2.1. Here,

$$e_k = \begin{cases} \{(s_k, t_k), (t_k, s_k)\} & \text{if } R_k = 1 \\ \{(s_k, t_k)\} & \text{if } R_k = 0, \end{cases} \tag{5}$$

where $s_k, t_k \in V(k-1) \cup \{|V(k-1)|+1\}$. Let $\boldsymbol{\theta} = (\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}})$. With these ingredients, the likelihood associated with the graph sequence $\{G(k)\}_{k=0}^n$ is given by

$$
\begin{aligned}
&p\left((e_k)_{k=1}^n, W(n) \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\rho}\right) \\
&= \alpha^{\sum_{k=1}^n 1_{\{J_k=1\}}} \beta^{\sum_{k=1}^n 1_{\{J_k=2\}}} (1-\alpha-\beta)^{\sum_{k=1}^n 1_{\{J_k=3\}}} \\
&\quad \times \prod_{k=1}^n \left(\frac{D_{t_k}^{\text{in}}(k-1) + \delta_{\text{in}}}{|E(k-1)| + \delta_{\text{in}}|V(k-1)|} \right)^{1_{\{J_k \in \{1,2\}\}}} \left(\frac{D_{s_k}^{\text{out}}(k-1) + \delta_{\text{out}}}{|E(k-1)| + \delta_{\text{out}}|V(k-1)|} \right)^{1_{\{J_k \in \{2,3\}\}}} \\
&\qquad \times \prod_{r=1}^K \pi_r^{\sum_{k=1}^n 1_{\{J_k=1\}} 1_{\{W_{s_k}=r\}} + \sum_{k=1}^n 1_{\{J_k=3\}} 1_{\{W_{t_k}=r\}}} \\
&\qquad\quad \times \prod_{r=1}^K \prod_{m=1}^K \rho_{m,r}^{\sum_{k=1}^n 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} 1_{\{R_k=1\}}} (1-\rho_{m,r})^{\sum_{k=1}^n 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} 1_{\{R_k=0\}}} \\
&\equiv p((e_k)_{k=1}^n \mid \boldsymbol{\theta}) \times p((e_k)_{k=1}^n, W(n) \mid \boldsymbol{\pi}, \boldsymbol{\rho}).
\end{aligned}
$$

The function $p(\cdot \mid \boldsymbol{\theta})$ collects the likelihood terms dependent on $\boldsymbol{\theta}$ and likewise $p(\cdot \mid \boldsymbol{\pi}, \boldsymbol{\rho})$ collects the terms dependent on $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. Such factorization implies that the estimation of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}, \boldsymbol{\rho}$ can be conducted independently. The frequentist estimation of $\boldsymbol{\theta}$ in homogeneous reciprocal PA models has already been considered in Cirkovic et al. (2023a). These estimators are unchanged in the heterogeneous case. Naturally, the maximum likelihood estimators (MLE) for α and β are given by $\hat{\alpha} = n^{-1} \sum_{k=0}^n 1_{\{J_k=1\}}$ and $\hat{\beta} = n^{-1} \sum_{k=0}^n 1_{\{J_k=2\}}$. The MLE for δ_{in} satisfies

$$
\sum_{k=1}^n 1_{\{J_k \in \{1,2\}\}} \frac{1}{D_{t_k}^{\text{in}}(k-1) + \hat{\delta}_{\text{in}}} - \sum_{k=1}^n 1_{\{J_k \in \{1,2\}\}} \frac{|V(k-1)|}{|E(k-1)| + \hat{\delta}_{\text{in}} N(k-1)} = 0, \tag{6}
$$

where (6) is obtained by setting $\frac{\partial}{\partial \delta_{\text{in}}} \log p((e_k)_{k=1}^n \mid \boldsymbol{\theta}) = 0$. The MLE for δ_{out} is obtained similarly. The estimators $\hat{\alpha}$ and $\hat{\beta}$ are strongly consistent for α and β, while consistency for $\hat{\delta}_{\text{in}}$ and $\hat{\delta}_{\text{out}}$ has not yet been verified since the reciprocal component of the model interferes with traditional techniques to analyze consistency in non-reciprocal preferential attachment models as in Wan et al. (2017). Estimation of $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$ is considerably more involved, and will be the main focus of this paper.

The reciprocal component of the preferential attachment model with heterogeneous reciprocity is reminiscent of a stochastic block model. Nodes first attach via the preferential attachment rules in (2), (3) and (4), then a stochastic-block-model type mechanism dictates the reciprocal behavior. A large portion of the literature on stochastic block modeling is concerned with community detection Bickel and Chen (2009); Holland et al. (1983); Karrer and Newman (2011); Zhao et al. (2011). Here we are primarily concerned with the estimation of $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$, and consider the recovery of $W(n)$ as a secondary goal. The optimal recovery of $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$ hinges on the correct specification of K, the number of reciprocal classes. We will thus examine cases when K is known a priori, as well as cases where it must be inferred from the data.

We also note that a minor nuisance of modeling reciprocal PA models is the observation of the random variable R_k. Upon serial observation of the edges $\{(u,v),(v,u)\}$, it is not

possible to identify whether the second edge was generated under $R_k = 1$ or the events $R_k = 0$ and $J_{k+1} = 2$. Assuming that (u, v) was added according to one of the attachment rules (2)-(4), (v, u) is either added by reciprocation, which occurs with probability $\rho_{W_v W_u}$, or without reciprocation and with $J_{k+1} = 2$, which occurs with probability

$$(1 - \rho_{W_v W_u})\beta \left(\frac{D_u^{\text{in}}(k-1) + 1_{\{u=v\}} + \delta_{\text{in}}}{|E(k-1)| + 1 + \delta_{\text{in}}(|V(k-1)| + 1_{\{J_k \in \{1,3\}\}})} \right) \left(\frac{D_v^{\text{out}}(k-1) + 1_{\{u=v\}} + \delta_{\text{out}}}{|E(k-1)| + 1 + \delta_{\text{out}}(|V(k-1)| + 1_{\{J_k \in \{1,3\}\}})} \right).$$

The latter probability is generally of smaller order due to the attachment rule (3), and hence it is reasonable to assume that all reciprocated edges are generated under $R_k = 1$. As demonstrated by our simulation studies in Section 5, this assumption has seemingly no effect on model estimation. In real-world networks time will often pass between message replies. For such networks, we will thus employ window estimators from Cirkovic et al. (2023a). We defer further discussion of window estimators to Section 6.

We will continue to consider the estimation of $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$ based on $p((e_k)_{k=1}^n, W(n) \mid \boldsymbol{\pi}, \boldsymbol{\rho})$. Since $W(n)$ is unobservable, a natural probabilistic approach would marginalize over the unobservable communication types, and form a complete-data likelihood $p((e_k)_{k=1}^n \mid \boldsymbol{\pi}, \boldsymbol{\rho})$. This, however, involves a sum over all latent configurations of $W(n)$ which is analytically intractable, as well as computationally infeasible for large networks. Such difficulties encourage attempts to learn $W(n)$ from the conditional distribution of $W(n)$ given $(e_k)_{k=1}^n$ (á la an EM Algorithm Dempster et al. (1977)) and jointly estimate $W(n), \boldsymbol{\pi}$ and $\boldsymbol{\rho}$. Often, these attempts are computationally infeasible due to the lack of factorization in the conditional distribution. In the following section, we will consider both Bayesian and frequentist estimation methods for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ where K is known. We will first present an "ideal" fully Bayesian approach, and then move on to variationally Bayesian and frequentist approximations to that ideal. Afterwards, we will discuss how to perform model selection when K is unknown for each of these methods.

3 Inference for a known number of communication types

In this section, we propose three ways of fitting the preferential attachment model with heterogeneous reciprocity to observed networks when the number of communication types is known. The first is a fully Bayesian procedure. The latter two are computationally efficient variational alternatives, one Bayesian and one frequentist.

3.1 Bayesian inference

For Bayesian inference of the heterogeneous reciprocal PA model we follow Nowicki and Snijders (2001) and employ independent and conditionally conjugate priors

$$\begin{aligned} &\rho_{m,r} \overset{\text{i.i.d.}}{\sim} \text{Beta}(a, b), \ m, r = 1, \ldots, K, \\ &\boldsymbol{\pi} \sim \text{Dirichlet}(\eta, \ldots, \eta). \end{aligned} \tag{7}$$

The prior specification (7) leads to a simple Gibbs sampler that draws approximate samples from the posterior $p(\boldsymbol{\rho}, \boldsymbol{\pi}, W(n) \mid (e_k)_{k=1}^n)$. We present the Gibbs sampler as Algorithm

1. Here, $\boldsymbol{\rho}$ and $\boldsymbol{\pi}$ are initialized from prior draws and $W(n)$ is initialized by drawing from $p(W_v \mid \boldsymbol{\pi})$ for $v = 1, \ldots |V(n)|$. Although the sampler is standard, many samples are required to sufficiently explore the posterior distribution. For large networks, this can be computationally onerous, and hence we appeal to variational alternatives.

Algorithm 1 Gibbs sampling for heterogeneous reciprocal PA with known K

Input: Graph $G(n)$, # communication types K, prior parameters a, b, η, # MCMC iterations M

Output: Approximate samples from the posterior $p(\boldsymbol{\rho}, \boldsymbol{\pi}, W(n) \mid (e_k)_{k=1}^n)$

Initialize: Draw $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ from (7), draw $W_v \sim \text{Multinomial}(\boldsymbol{\pi})$ for $v \in V(n)$

for $i = 1$ to M **do**

1. Sample $W(n)$ from its conditional posterior

for all $v \in V(n)$ **do**

Sample W_v according to

$$P(W_v = \ell \mid \boldsymbol{\pi}, \boldsymbol{\rho}, (W_u)_{u \neq v}, (e_k)_{k=1}^n) \propto \pi_r \prod_{m=1}^{K} \rho_{m,\ell}^{\sum_{k:s_k=v} 1_{\{W_{t_k}=m\}} 1_{\{R_k=1\}}} (1-\rho_{m,\ell})^{\sum_{k:s_k=v} 1_{\{W_{t_k}=m\}} 1_{\{R_k=0\}}} \times \prod_{r=1}^{K} \rho_{\ell,r}^{\sum_{k:t_k=v} 1_{\{W_{s_k}=r\}} 1_{\{R_k=1\}}} (1-\rho_{\ell,r})^{\sum_{k:t_k=v} 1_{\{W_{s_k}=r\}} 1_{\{R_k=0\}}}$$

for $\ell = 1, \ldots, K$

end for

2. Sample $\boldsymbol{\rho}$ from its conditional posterior

for $m = 1$ to K **do**

for $r = 1$ to K **do**

Sample $\rho_{m,r}$ from

$$\rho_{m,r} \mid \boldsymbol{\pi}, W(n), (e_k)_{k=1}^n \sim \text{Beta}\left(a + \sum_{k=1}^{n} 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} 1_{\{R_k=1\}}, b + \sum_{k=1}^{n} 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} 1_{\{R_k=0\}}\right)$$

end for

end for

3. Sample $\boldsymbol{\pi}$ from its conditional posterior

$$\boldsymbol{\pi} \mid \boldsymbol{\rho}, W(n), (e_k)_{k=1}^n \sim \text{Dirichlet}\left(\eta + \sum_{v \in V(n)} 1_{\{W_v=1\}}, \ldots, \eta + \sum_{v \in V(n)} 1_{\{W_v=K\}}\right)$$

end for

3.2 Variational inference

In this section, we present variational alternatives for approximating posteriors associated with the heterogeneous reciprocal PA model. Variational inference aims to approximate the conditional distribution of latent variables $\mathbf{z}$ given data $\mathbf{x}$ via a class of densities $\mathcal{Q}$ typically chosen to circumvent computational inconveniences. If Bayesian inference is being performed, the latent variables $\mathbf{z}$ can also encompass the model parameters ($\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ in our setting). The variational inference procedure aims to find the density $q^\star \in \mathcal{Q}$ that minimizes the Kullback-Leibler (KL) divergence from $p(\cdot \mid \mathbf{x})$, i.e.

$$q^\star = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q(\cdot) \,||\, p(\cdot \mid \mathbf{x})\right). \tag{8}$$

We will restrict $\mathcal{Q}$ to the mean-field family, that is, the family of densities where components $\mathbf{z}$ are mutually independent. Naturally, such restriction will prevent $q^\star$ from capturing the dependence structure between the latent variables. Recently, however, some more structured, expressive families have been proposed that may improve the approximation; see for instance Yin et al. (2020). Conveniently, using the definition of the conditional density, the objective (8) can be expressed as

$$\mathrm{KL}\left(q(\cdot) \,||\, p(\cdot \mid \mathbf{x})\right) = E_q[\log q(\mathbf{z})] - E_q[\log p(\mathbf{z}, \mathbf{x})] + \log p(\mathbf{x}) \equiv -\mathrm{ELBO}(q) + \log p(\mathbf{x}), \tag{9}$$

so that minimizing the KL divergence from $p(\cdot \mid \mathbf{x})$ to $q(\cdot)$ is equivalent to maximizing the evidence lower bound ($\mathrm{ELBO}(q)$) since $\log p(\mathbf{x})$ does not depend on q. For more on variational inference, see Blei et al. (2017).

3.2.1 Bayesian Variational Inference

Now we consider solving the variational problem (8) for the probabilistic model presented in Section 3.1. Although we have presented a sampler in Algorithm 1 that draws approximate samples from the posterior, we aim for an estimate that sacrifices modeling the dependence in the posterior distribution in favor of computation time. Variational inference for stochastic blockmodels in the Bayesian setting was studied in Latouche et al. (2012). Following their strategy, we posit a mean-field variational family:

$$q(\boldsymbol{\pi}, \boldsymbol{\rho}, W(n)) = q(\boldsymbol{\pi})q(\boldsymbol{\rho})q(W(n)) = q(\boldsymbol{\pi}) \prod_{m=1}^{K} \prod_{r=1}^{K} q(\rho_{m,r}) \prod_{v \in V(n)} q_v(W_v).$$

We further assume that the variational densities have the following forms:

$$\begin{aligned}
q(\boldsymbol{\pi}) &\propto \prod_{r=1}^{K} \pi_r^{d_r},\ d_1, \ldots, d_K \geq 0, \\
q(\rho_{m,r}) &\propto \rho_{m,r}^{\omega_{m,r}} (1-\rho_{m,r})^{\xi_{m,r}},\ \omega_{m,r}, \xi_{m,r} \geq 0,\ m, r = 1, \ldots, K, \\
q_v(W_v) &= \prod_{r=1}^{K} \tau_{v,r}^{1_{\{W_v = r\}}}, \tau_{v,r} \geq 0,\ r = 1, \ldots, K,\ v = 1, \ldots |V(n)|,
\end{aligned}$$

and additionally $\sum_{r=1}^{K} \tau_{v,r} = 1$ for all $v \in V(n)$. In other words, the posterior of $\boldsymbol{\pi}$ is approximated by a Dirichlet$(d_1, \ldots, d_K)$ distribution, and the component-wise posteriors of $\boldsymbol{\rho}$ and $W(n)$ are approximated by Beta$(\omega_{m,r}, \xi_{m,r})$ and Multinomial$(1, (\tau_{v,r})_{r=1}^{K})$ distributions, respectively. These choices are made to obtain a tractable algorithm; independence among parameters allows for a simpler calculation of (8) and the distributional families mimic the functional forms of the prior and likelihood which helps to facilitate computation. In Algorithm 2 we present a coordinate ascent variational inference (CAVI) algorithm for optimizing the ELBO. Here, $\psi(\cdot)$ is the digamma function. Note that in step 3 of algorithm, we write $\sum_{k:s_k=v} \equiv \sum_{k:s_k=v, s_k \neq t_k}$ for brevity of notation. The inclusion of self-loops makes the optimization of the ELBO much more difficult, hence their exclusion. Here, the class probabilities, $\tau_{v,r}$, are initialized uniformly at random. We omit the calculations for the derivation of this algorithm, as they are very similar to Latouche et al. (2012).

To monitor the convergence of Algorithm 2, we recommend computing the ELBO after each iteration of the CAVI algorithm and terminating the algorithm once the increase in the ELBO is less than some predetermined threshold ϵ. Specifically, if the ELBO is computed after step 2, it has the simplified form:

$$\begin{aligned} \text{ELBO}(q) = &\log\left(\frac{\Gamma(K\eta)\prod_{r=1}^{K}\Gamma(d_r)}{\Gamma(\sum_{r=1}^{K} d_r)\Gamma(\eta)^K}\right) + \sum_{r=1}^{K}\sum_{m=1}^{K}\log\left(\frac{\Gamma(a+b)\Gamma(\omega_{m,r})\Gamma(\xi_{m,r})}{\Gamma(\omega_{m,r}+\xi_{m,r})\Gamma(a)\Gamma(b)}\right) \\ &- \sum_{v\in V(n)}\sum_{r=1}^{K}\tau_{v,r}\log\tau_{v,r}. \end{aligned} \tag{10}$$

Here, recall that a, b and η are prior parameters defined in (7).

3.2.2 Variational Expectation Maximization

In this section, we consider frequentist estimation of the PA model with heterogeneous reciprocity through a variational expectation maximization algorithm (VEM). VEM for stochastic blockmodel data was first considered in Daudin et al. (2008) which further inspired many interesting generalizations that could enhance the reciprocal PA model (see Matias and Miele, 2017, for example). The VEM algorithm augments the traditional EM algorithm by approximating the E-step for models in which the conditional distribution of the latent variables given the observed data is computationally intractable. The VEM estimates thus serve as a computationally efficient approximation to the maximum likelihood estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. Although a frequentist procedure, the VEM algorithm may enhance Bayesian inference of stochastic blockstructure data. For example, since the dimension of the posterior $p\left(\boldsymbol{\pi}, \boldsymbol{\rho} \mid (e_k)_{k=1}^{n}\right)$ does not grow with the size of the data, one might expect a Bernstein-von-Mises phenomena to occur. The VEM estimates may thus approximate the posterior mean or even be leveraged to enhance posterior sampling as in Donnet and Robin (2021).

As in Section 3.2.1, we approximate the distribution of the communication types given the observed network, $p\left(W(n) \mid \boldsymbol{\pi}, \boldsymbol{\rho}, (e_k)_{k=1}^{n}\right)$, via the mean-field approximation

$$q(W(n)) = \prod_{v\in V(n)} q_v(W_v).$$

Algorithm 2 CAVI for heterogeneous reciprocal PA with known K

Input: Graph $G(n)$, # communication types K, prior parameters a, b, η, tolerance $\epsilon > 0$
Output: Variational approximation to the posterior $q^\star$

Initialize: Draw $\tau_{v,r}$, $r = 1, \ldots, K$ uniformly at random from the K-simplex for every $v \in V(n)$

while the increase in ELBO(q) is greater than ϵ **do**

1. Update $q(\boldsymbol{\pi})$

for $r = 1$ to K **do**

$$d_r = \eta + \sum_{v \in V(n)} \tau_{u,r}$$

end for

2. Update $q(\boldsymbol{\rho})$

for $m = 1$ to K **do**

for $r = 1$ to K **do**

$$\omega_{m,r} = a + \sum_{k=1}^{n} \tau_{s_k,r} \tau_{t_k,m} 1_{\{R_k=1\}}$$

$$\xi_{m,r} = b + \sum_{k=1}^{n} \tau_{s_k,r} \tau_{t_k,m} 1_{\{R_k=0\}}$$

end for

end for

3. Update ELBO(q) according to (10)

4. Update $q(W(n))$

for all $v \in V(n)$ **do**

for $\ell = 1$ to K **do**

$$\begin{aligned}
\tau_{v,\ell} \propto \exp &\left\{ \psi\left(d_\ell\right) \; - \psi\left(\sum_{r=1}^{K} d_r \right) \right\} \\
&\times \prod_{m=1}^{K} \exp \left\{ \psi\left(\omega_{m,\ell}\right) \sum_{k:s_k=v} \tau_{t_k,m} 1_{\{R_k=1\}} + \psi\left(\xi_{m,\ell}\right) \sum_{k:s_k=v} \tau_{t_k,m} 1_{\{R_k=0\}} \right. \\
&\qquad \left. - \psi(\omega_{m,\ell} + \xi_{m,\ell}) \sum_{k:s_k=v} \tau_{t_k,m} \right\} \\
&\times \prod_{r=1}^{K} \exp \left\{ \psi\left(\omega_{\ell,r}\right) \sum_{k:t_k=v} \tau_{s_k,r} 1_{\{R_k=1\}} + \psi\left(\xi_{\ell,r}\right) \sum_{k:t_k=v} \tau_{s_k,r} 1_{\{R_k=0\}} \right. \\
&\qquad \left. - \psi(\omega_{\ell,r} + \xi_{\ell,r}) \sum_{k:t_k=v} \tau_{s_k,r} \right\}
\end{aligned}$$

end for

end for

end while

Via the mean-field family assumption, the ELBO is given by

$$\begin{aligned}
&\text{ELBO}(q, \boldsymbol{\pi}, \boldsymbol{\rho}) \\
&\quad = E_q \left[\log p\left(W(n), (e_k)_{k=1}^n \mid \boldsymbol{\pi}, \boldsymbol{\rho}\right)\right] - E_q\left[\log q(W(n))\right] \\
&\quad = \sum_{k=1}^{n}\sum_{r=1}^{K}\left(1_{\{J_k=1\}}\tau_{s_k,r} + 1_{\{J_k=3\}}\tau_{t_k,r}\right)\log \pi_r - \sum_{v\in V(n)}\sum_{r=1}^{K}\tau_{v,r}\log\tau_{v,r} \\
&\qquad + \sum_{k=1}^{n}\sum_{r=1}^{K}\sum_{m=1}^{K}\tau_{s_k,r}\tau_{t_k,m}\left(1_{\{R_k=1\}}\log\rho_{m,r} + 1_{\{R_k=0\}}\log(1-\rho_{m,r})\right).
\end{aligned} \tag{11}$$

Note that from (9), maximizing (11) concerning q (the E-step) is equivalent to minimizing the KL divergence from $p\left(\cdot \mid \boldsymbol{\pi}, \boldsymbol{\rho}, (e_k)_{k=1}^n\right)$ to $q(\cdot)$ and maximizing (11) for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ is equivalent to the M-step in the usual EM algorithm. Thus, the E-step is equivalent to performing variational inference for $p\left(\cdot \mid \boldsymbol{\pi}, \boldsymbol{\rho}, (e_k)_{k=1}^n\right)$ where $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are evaluated at their current estimates $\hat{\boldsymbol{\pi}}_{\text{VEM}}$ and $\hat{\boldsymbol{\rho}}_{\text{VEM}}$.

The VEM algorithm for the heterogeneous reciprocal PA model is given in Algorithm 3. As in Algorithm 2, we write $\sum_{k:s_k=v} \equiv \sum_{k:s_k=v, s_k\neq t_k}$ for ease of notation. We describe the initialization of the algorithm at the end of Appendix A in Algorithm 5. We further provide some derivations of the VEM algorithm in Appendix B. Similar types of computations can be employed to derive Algorithm 2. As in Algorithm 2, we recommend cycling through the updates of $\hat{\tau}_{v,\ell}$ in the E-step until the ELBO no longer increases beyond a pre-specified threshold $\epsilon > 0$.

4 Model selection for an unknown number of communication types

In this section we extend the methods discussed in Section 3 to the case where the number of communication types is not known a priori. This can be viewed as a model selection problem, where the Bayesian solution places a prior on K while the variationally Bayesian and EM algorithms aim to imitate marginal likelihood-based procedures.

4.1 A prior on K

This section extends the Bayesian solution in Section 3.1 to making inference on the unknown number of communication classes K. In a fully Bayesian framework, K is assigned a prior and inference is made on the posterior of K given the observed data. This, however, often requires the use of complicated reversible jump MCMC (RJMCMC) algorithms to make valid posterior inference on K. Generically, mixture models with a prior on the number of mixture components are known as mixture of finite mixture (MFM) models. For Bayesian MFMs, Miller and Harrison (2018) derived the Dirichlet process-like properties of MFMs and proposed a collapsed Gibbs sampler that circumvented the need for RJMCMC. Geng et al. (2019) used a similar collapsed Gibbs sampler for learning the number of components in a stochastic block model. Unfortunately, such collapsed Gibbs samplers require analytically marginalizing over K, restricting our ability to make inference on $\boldsymbol{\pi}$ without some ad-hoc post-processing of the posterior samples. Recently, a telescoping sampler has been developed by Frühwirth-Schnatter et al. (2021) for MFMs that obviates the

Algorithm 3 VEM for heterogeneous reciprocal PA with known K

Input: Graph $G(n)$, # communication types K, tolerances $\epsilon, \kappa > 0$

Output: Variational EM estimates $\hat{\boldsymbol{\pi}}_{\text{VEM}}$ and $\hat{\boldsymbol{\rho}}_{\text{VEM}}$

Initialize: Draw $\hat{\tau}_{v,r}$, $r = 1, \ldots, K$ uniformly at random from the K-simplex for every $v \in V(n)$, run Algorithm 5 to initialize $\hat{\boldsymbol{\pi}}_{\text{VEM}}$ and $\hat{\boldsymbol{\rho}}_{\text{VEM}}$

while at least one of the elements of $\hat{\boldsymbol{\pi}}_{\text{VEM}}$ and $\hat{\boldsymbol{\rho}}_{\text{VEM}}$ change by more than κ in absolute value **do**

1. **E-step**: Update $\hat{q}$ via

while the increase in ELBO(q) is greater than ϵ **do**

for all $v \in V(n)$ **do**

for $\ell = 1$ to K **do**

$$\hat{\tau}_{v,\ell} \propto \hat{\pi}_\ell \prod_{m=1}^{K} \hat{\rho}_{m,\ell}^{\sum_{k:s_k=v} \hat{\tau}_{t_k,m} 1_{\{R_k=1\}}} (1-\hat{\rho}_{m,\ell})^{\sum_{k:s_k=v} \hat{\tau}_{t_k,m} 1_{\{R_k=0\}}} \times \prod_{r=1}^{K} \hat{\rho}_{\ell,r}^{\sum_{k:t_k=v} \hat{\tau}_{s_k,r} 1_{\{R_k=1\}}} (1-\hat{\rho}_{\ell,r})^{\sum_{k:t_k=v} \hat{\tau}_{t_k,r} 1_{\{R_k=0\}}}$$

end for

end for

Update ELBO(q) according to (11)

end while

2. **M-step**: Update $\hat{\boldsymbol{\pi}}_{\text{VEM}}$ and $\hat{\boldsymbol{\rho}}_{\text{VEM}}$ via

for $m = 1$ to K **do**

$$\hat{\pi}_m = \sum_{v \in V(n)} \hat{\tau}_{v,m}$$

for $r = 1$ to K **do**

$$\hat{\rho}_{m,r} = \frac{\sum_{k=1}^{n} \hat{\tau}_{s_k,r} \hat{\tau}_{t_k,m} 1_{\{R_k=1\}}}{\sum_{k=1}^{n} \hat{\tau}_{s_k,r} \hat{\tau}_{t_k,m}}$$

end for

end for

end while

need to marginalize over K. Rather, K is explicitly sampled in the scheme by distinguishing between K, the number of mixture components, and K_+, the number of *filled* mixture components.

For the heterogeneous reciprocal PA model, we adopt the prior specification in (7) and additionally let $K-1$ follow a beta-negative-binomial (BNB) distribution with parameters c_1, c_2 and c_3 as recommended by Frühwirth-Schnatter et al. (2021). The BNB distribution is a hierarchical generalization of the Poisson, geometric, and negative-binomial distribution. If $K-1 \sim \text{BNB}(c_1, c_2, c_3)$ then the probability mass function on K is given by

$$p(K) = \frac{\Gamma(c_1 + K - 1)B(c_1 + c_2, K - 1 + c_3)}{\Gamma(c_1)\Gamma(K)B(c_2, c_3)},\ K = 1, 2, \ldots,$$

where B denotes the beta function. As discussed in Frühwirth-Schnatter et al. (2021), the BNB distribution allows the user to specify a heavier tail on the number of mixture components which is essential for the telescoping sampler to mix well. Previous analyses in Geng et al. (2019) and Miller and Harrison (2018) specify that $K-1 \sim \text{Poisson}(1)$, which is a highly informative choice with a light tail.

We present the telescoping sampler for heterogeneous reciprocal PA models in Algorithm 4. For ease of notation, we do not distinguish between $W(n)$, the communication types, and the random partition of the $|V(n)|$ nodes into K_+ clusters induced by $W(n)$. However, the alternating between sampling on the parameter space of the mixture distribution and the set partition space is a key aspect that allows K to be directly sampled from the conditional posterior of K given the partition induced by $W(n)$ (Step 3 in Algorithm 4). We refer to Frühwirth-Schnatter et al. (2021) for more details on the telescoping sampler. Note that within the sampler, K only decreases if one of the K_+ filled components loses all of its membership in Step 1. Thus, for the sampler to mix well, K must occasionally exceed K_+, emphasizing the need for a heavier-tailed prior on K.

Frühwirth-Schnatter et al. (2021) also present a dynamic mixture of finite mixture model where the prior on $\boldsymbol{\pi}$ is taken to be $\text{Dirichlet}(\varphi/K, \varphi/K, \ldots, \varphi/K)$ for some $\varphi > 0$. This specification would induce a sparse mixture model where a large number of mixture components K would be fit, but a majority of them would be unfilled (Frühwirth-Schnatter and Malsiner-Walli, 2019; Malsiner-Walli et al., 2016). In this sense, the posterior distributions on K and K_+ would differ greatly. Though this is undesirable for learning the parameters of a mixture model, it may be useful for analyses more focused on partitioning nodes into a small number of classes with similar reciprocal behavior.

4.2 Imitations of the marginal likelihood

In this section, we review criteria for choosing the number of communication types K for the variational methods proposed in Section 3.2. A typical strategy for Bayesian model selection is choosing the model that maximizes the marginal likelihood, or the probability distribution that is obtained by integrating the likelihood over the prior distribution of the parameters. For many of the same reasons presented in Section 2.2, the marginal likelihood is not available for stochastic blockmodel data. Instead, for the Bayesian Variational Inference method presented in Section 3.2.1, Latouche et al. (2012) recommend employing the ELBO

Algorithm 4 Telescoping sampler for heterogeneous reciprocal PA with known K

Input: Graph $G(n)$, parameters a, b, η, c_1, c_2, c_3, K initial/max values K_{init}, $K_{\max}$, # MCMC iterations M

Output: Approximate samples from the posterior $p\left(\boldsymbol{\rho}, \boldsymbol{\pi}, W(n), K \mid (e_k)_{k=1}^n\right)$

Initialize: Set $K = K_{\text{init}}$, draw $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ from (7), draw $W_v \sim \text{Multinomial}(\boldsymbol{\pi})$ for $v \in V(n)$

for $i = 1$ to M **do**

1. Sample $W(n)$ from its conditional posterior

for all $v \in V(n)$ **do**

Sample W_v according to

$$\begin{aligned} &P\left(W_v = \ell \mid \boldsymbol{\pi}, \boldsymbol{\rho}, (W_u)_{u \neq v}, (e_k)_{k=1}^n\right) \\ &\quad \propto \pi_\ell \prod_{m=1}^{K} \rho_{m,\ell}^{\sum_{k:s_k=v} 1_{\{W_{t_k}=m\}} 1_{\{R_k=1\}}} (1-\rho_{m,\ell})^{\sum_{k:s_k=v} 1_{\{W_{t_k}=m\}} 1_{\{R_k=0\}}} \\ &\qquad \times \prod_{r=1}^{K} \rho_{\ell,r}^{\sum_{k:t_k=v} 1_{\{W_{s_k}=r\}} 1_{\{R_k=1\}}} (1-\rho_{\ell,r})^{\sum_{k:t_k=v} 1_{\{W_{s_k}=r\}} 1_{\{R_k=0\}}}, \end{aligned}$$

for $\ell = 1, \ldots, K$

end for

and determine the number of filled components K_+. Relabel the communication classes such that the first K_+ components are filled and the rest are empty.

2. Sample the filled components of $\boldsymbol{\rho}$ from its conditional posterior

for $m = 1$ to K_+ **do**

for $r = 1$ to K_+ **do**

$$\rho_{m,r} \mid \boldsymbol{\pi}, W(n), (e_k)_{k=1}^n \sim \text{Beta}\Bigg(a + \sum_{k=1}^{n} 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} 1_{\{R_k=1\}}, \\ b + \sum_{k=1}^{n} 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} 1_{\{R_k=0\}}\Bigg),$$

end for

end for

3. Sample K from

$$p(K|W(n)) \propto p(K) \frac{K!}{(K-K_+)!} \frac{\Gamma(\eta K)}{\Gamma(|V(n)| + \eta K)\Gamma(\eta)^{K_+}} \prod_{r=1}^{K_+} \Gamma\left(\sum_{v \in V(n)} 1_{\{W_v=r\}} + \eta\right),$$

where $K = K_+, K_+ + 1, \ldots, K_{\max}$. If $K > K_+$, generate $K - K_+$ empty components and fill the corresponding $\boldsymbol{\rho}$ components with draws from the prior $\text{Beta}(a, b)$.

4. Sample $\boldsymbol{\pi}$ from its conditional posterior

$$\boldsymbol{\pi} \mid \boldsymbol{\rho}, W(n), (e_k)_{k=1}^n \sim \text{Dirichlet}\left(\eta + \sum_{v \in V(n)} 1_{\{W_v=1\}}, \ldots, \eta + \sum_{v \in V(n)} 1_{\{W_v=K\}}\right)$$

end for

as the model selection criterion. From (9), it can be seen that

$$\text{ELBO}(q) = -\text{KL}\left(q(\cdot) \,||\, p(\cdot|(e_k)_{k=1}^n)\right)) + \log p\left((e_k)_{k=1}^n\right) \leq \log p\left((e_k)_{k=1}^n\right).$$

That is, the ELBO lower bounds the marginal likelihood, and if the variational approximation to the posterior is good, the ELBO should approximate it. Though, there is no evidence that the variational approximation results in a sufficiently small KL divergence such that the ELBO accurately estimates the marginal likleihood. Regardless, this criterion is often used in practice (Blei et al., 2017).

For the VEM algorithm, Daudin et al. (2008) recommend employing the Integrated Classification Likelihood (ICL). Although the VEM algorithm is a frequentist procedure, the ICL criterion is derived by assuming a Jeffrey's prior on $\boldsymbol{\pi}$ ($\eta = 1/2$) and further employs a BIC approximation to the distribution of $(e_k)_{k=1}^n$ given $W(n)$. The ICL for reciprocal PA models is given by

$$\text{ICL}(K) = \log p((e_k)_{k=1}^n, \hat{W}(n) \mid \hat{\boldsymbol{\pi}}_{\text{VEM}}, \hat{\boldsymbol{\rho}}_{\text{VEM}}) - \frac{K^2}{2}\log n - \frac{K-1}{2}\log|V(n)|,$$

where $\hat{W}(n)$ is the modal approximation of $W(n)$ given by $\hat{W}_v = \arg\max_{\ell=1,\ldots,K} \hat{\tau}_{v,\ell}$.

5 Simulation Studies

In this section, we evaluate the performance of the estimation procedures presented in Sections 3 and 4 on synthetic datasets. We evaluate the performance of estimation methods for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ when K is known, as well as the accuracy of the model selection criteria presented in Section 4 when K is unknown. When K is known, we employ the Monte Carlo averages of the approximate posterior samples, the posterior means of the variational densities and the variational EM estimates as point estimators of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ for the fully Bayesian (B), variational Bayes (VB) and variational EM (VEM) methods, respectively. Since the B and VB methods produce approximate posteriors, we also provide marginal coverage rates of credible intervals constructed using the element-wise 2.5% and 97.5% quantiles of the respective posteriors for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. In the case of known K, we further provide the average Rand index for estimating $(W_v)_{v\in V(n)}$ for each method. When K is unknown, we record the frequencies of the estimated K under each model selection criteria. We employ the posterior mode as the estimated K for the fully Bayesian method.

In each simulation, we assume non-informative priors Dirichlet$(1/2,\ldots,1/2)$ on $\boldsymbol{\pi}$ and Beta$(1/2,1/2)$ on $\boldsymbol{\rho}$ for the VB and B methods. Although a prior is not explicitly assumed for the VEM method, the ICL model selection criterion implicitly assumes the same prior on $\boldsymbol{\pi}$, hence these choices are consistent. For the VEM algorithm, we terminate the E-step once either the ELBO has increased by less than $\epsilon = 0.01$ or the total number of iterations exceeds 500, and terminate the entire algorithm once the element-wise differences in the parameters fall below $\kappa = 0.01$. We also terminate the VB algorithm via the same conditions as in the E-step of the VEM algorithm. We run the fully Bayesian method for $M = 5{,}000$ MCMC samples, and discard the first half as burn-in. Further, when K is unknown, we assume a BNB$(1,4,3)$ prior on K as recommended by Frühwirth-Schnatter et al. (2021) and set $K_{\max} = 20$. For the variational methods, we search over $K = 1,2,3,4$.

Method	$\pi_1 = 0.8$	$\rho_{11} = 0.5$	$\rho_{12} = 0.9$	$\rho_{21} = 0.05$	$\rho_{22} = 0.2$
			Mean(SE)		
B	0.803(0.003)	0.500(0.002)	0.900(0.004)	0.050(0.002)	0.198(0.010)
VB	0.805(0.003)	0.500(0.002)	0.896(0.004)	0.050(0.002)	0.198(0.010)
VEM	0.788(0.004)	0.501(0.002)	0.889(0.005)	0.052(0.002)	0.196(0.010)
			% Coverage		
B	98	93	94	92	94
VB	50	90	70	87	92

Table 1: Average point estimates and standard errors for 100 networks generated from a PA model with $\boldsymbol{\theta} = (0.15, 0.8, 1, 1)$. Coverage rates for equal-tailed credible intervals produced by the B and VB methods are also provided.

5.1 Simulations on Pragmatic Networks

In this section, we evaluate the presented model estimation and selection criteria on networks that are generated to reflect those found in real-world applications. In particular, we generate a PA network with heterogeneous reciprocity such that $\boldsymbol{\theta} = (\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}}) = (0.15, 0.8, 1, 1)$,

$$\boldsymbol{\pi} = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\rho} = \begin{bmatrix} 0.5 & 0.9 \\ 0.05 & 0.2 \end{bmatrix}.$$

This network generating process contains two groups, the first of which can be thought of as typical users and the other can be thought of as celebrities. Here typical users will often reciprocate the messages from celebrities, but a celebrity is far less likely to respond to a typical user. As one might expect, there are far more typical users than celebrities on this network.

Table 1 displays the means and standard errors of the element-wise point estimators across the simulations, as well as the coverage of credible intervals produced from the B and VB methods. Here, the estimation procedures have virtually identical performance in terms of point estimation. Further, the coverage rates for the fully Bayesian method hover around the expected 95%, while the coverage rates for the variational Bayes method vary across the parameters. The VB method seems to have difficulty capturing the larger reciprocity $\rho_{12} = 0.9$, as well as $\pi_1 = 0.8$. The methods also perform similarly in terms of classification, as the average Rand index for the communication types is given by 0.767, 0.767, and 0.766 for the B, VB, and VEM methods respectively.

Table 2 displays the performance of the model selection criteria on the same preferential attachment model but for unknown K. For the fully Bayesian method, we initialize at $K_{\text{init}} = 4$ to exhibit the insensitivity of the telescoping sampler to initialization. Note that the ELBO and ICL select the correct class for every simulated data set, while the fully Bayesian method has a slight tendency to over-select the number of classes. However, analysis of such networks results in variational methods that perform comparably to the fully Bayesian method, at less computational cost.

We continue our simulations by evaluating the performance of the estimation procedures on 100 synthetic networks generating from a PA network with heterogeneous reciprocity

Method	$\hat{K}$			
	1	2	3	4
B	0	68	31	1
VB	0	100	0	0
VEM	0	100	0	0

Table 2: Estimated K from 100 networks generated from a PA model with $\boldsymbol{\theta} = (0.15, 0.8, 1, 1)$ and $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ as in Table 1.

such that $\boldsymbol{\theta} = (0.15, 0.8, 1, 1)$ but now

$$\boldsymbol{\pi} = \begin{bmatrix} 0.8 \\ 0.2 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\rho} = \begin{bmatrix} 0.5 & 0 \\ 0.05 & 0.2 \end{bmatrix}.$$

Note that the only difference between this simulation setup and the previous one is that ρ_{12} has decreased from 0.9 to 0. The inclusion of 0 into the $\boldsymbol{\rho}$ matrix is motivated by the data example in Section 6.1, where we find a group of users that do not receive reciprocal edges. This set-up is analogous to a diagonally-dominant stochastic block model where users are likely to communicate within groups but not across groups.

Table 3 displays the point estimates for all three methods, along with the coverage probabilities for the B and VB methods. With the decrease in ρ_{12}, the variational methods struggle to recover ρ_{22}. This is sensible since class 2 communicating with class 2 should be the least common communication type according to $\boldsymbol{\pi}$ and, unlike the case when $\rho_{12} = 0.9$, the difference between the communication classes is not obvious. Otherwise, the estimation accuracy of the other parameters is relatively consistent across all the methods. Although coverage rates are similar to Table 1, we also observe a reduction in the coverage of ρ_{22}. Equal-tailed credible intervals are a poor choice for capturing ρ_{12} and if one had prior knowledge of the behavior of $\boldsymbol{\rho}$, the highest posterior density interval would be a sensible choice. The average Rand index for the B, VB, and VEM methods are given by 0.762, 0.762, and 0.760, respectively, again indicating that the methods classify similarly when the number of edges far exceeds the number of nodes. Additionally, we verify the computational efficiency of the variational methods by tracking the average completion time for each algorithm across the 100 network realizations. Indeed, the average time required to draw 5,000 samples from Algorithm 1 is 229.80 seconds with a standard error of 3.61 seconds, while Algorithms 2 and 3 require on average 5.62 and 3.55 seconds to run with standard errors of 0.440 and 0.641 seconds, respectively. Naturally, completion times will vary with the choices of tolerances and other algorithmic parameters; such choices were not necessarily equated across the methods and the reported times serve as a rough comparison.

Table 4 displays the performance of the model selection criteria presented in Section 4 for the preferential attachment model as in 3. Again, K is initialized at $K_{\text{init}} = 4$ for the fully Bayesian method. Note that, again, the variational methods select the correct number of classes in each simulation, while the telescoping sampler has a slight tendency to overfit. The average completion time for Algorithm 4 was 245.12 seconds with a standard deviation of 16.66 seconds, while the cycles over K through Algorithms 2 and Algorithms 3 terminated at 33.76 and 19.90 seconds with standard deviations of 7.08 and 11.36 seconds,

Method	$\pi_1 = 0.8$	$\rho_{11} = 0.5$	$\rho_{12} = 0.00$	$\rho_{21} = 0.05$	$\rho_{22} = 0.2$
			Mean(SE)		
B	0.802(0.003)	0.501(0.002)	0.001(0.001)	0.051(0.003)	0.199(0.017)
VB	0.805(0.003)	0.501(0.002)	0.001(0.001)	0.052(0.003)	0.172(0.016)
VEM	0.791(0.011)	0.503(0.003)	0.006(0.003)	0.057(0.005)	0.152(0.017)
			% Coverage		
B	96	89	0	94	97
VB	55	88	0	85	45

Table 3: Average point estimates and standard errors for 100 networks generated from a PA model with $\boldsymbol{\theta} = (0.15, 0.8, 1, 1)$. Coverage rates for equal-tailed credible intervals produced by the B and VB methods are also provided.

Method	$\hat{K}$		
	1	2	3
B	0	77	23
VB	0	100	0
VEM	1	99	0

Table 4: Estimated K from 100 networks generated from a PA model with $\boldsymbol{\theta} = (0.15, 0.8, 1, 1)$ and $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ as in Table 4.

respectively. Despite Algorithm 4 evaluating multiple models over a single Gibbs sampler, the variational methods are still significantly faster than the fully Bayesian method.

5.2 Comparisons to the SBM

In this section, we evaluate the same estimation and model selection procedures on synthetic networks with a comparably low number of edges relative to the number of nodes. Such networks serve to highlight the additional difficulties faced by estimating reciprocal PA models compared to stochastic block models. We simulate 100 preferential attachment networks of size $n = 30{,}000$ from a PA model with $\boldsymbol{\theta} = (\alpha, \beta, \delta_{\text{in}}, \delta_{\text{out}}) = (0.75, 0, 0.8, 0.8)$ and the reciprocal component governed by

$$\boldsymbol{\pi} = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\rho} = \begin{bmatrix} 0.1 & 0.4 \\ 0.5 & 0.8 \end{bmatrix}.$$

Wang and Resnick (2023b) have shown that, under suitable conditions, such heterogeneous reciprocal PA models with $\beta = 0$ exhibit networks with out/in-degrees that exhibit a complex extremal dependence structure (see Appendix A for more details). Additionally, since $\beta = 0$, such models allow for the complete observation of the reciprocal edge events as there are no $J_k = 2$ edges that could be mistaken as reciprocal edges.

Here we assume K is known. Table 5 displays the average value of the point estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ for each method, as well as their associated standard errors. Clearly, the fully

Bayesian method outperforms both the VB and VEM methods by producing accurate point estimates with lower standard errors. Additionally, the coverage rates for the fully Bayesian method are near the expected 95% level, while the VB method produces posteriors that do not reliably capture the true $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. The fully Bayesian method also dominates in terms of classification, as the average Rand index for the communication types is given by 0.590, 0.583, and 0.552 for the B, VB, and VEM methods, respectively.

The superiority of the fully Bayesian method compared to the variational methods is unsurprising in this setting. Although variational methods exhibit strong point estimation for stochastic block models, estimation for PA models with heterogeneous reciprocity is an inherently harder problem. Namely, in a directed stochastic block model, each node has the opportunity to connect to every other node in the network. This results in $m(m-1)$ many potential edges for m many nodes in the network. For the PA model, one expects the number of potential edges to scale linearly with the number of nodes. Thus, there is inherently less observed information that can be leveraged to learn the latent communication classes. Such lack of information induces a multimodal ELBO, and therefore the variational methods struggle to find a global optimum. The fully Bayesian method is better able to incorporate this uncertainty since it is sampling from, not optimizing, a multimodal posterior.

Under suitable conditions on the data-generating parameters, which are satisfied by the current choices of $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$, Wang and Resnick (2023b) show that, for large enough networks, total degrees drawn from a heterogeneous reciprocal PA model have a power law tail. More specifically, the empirical total degree distribution tends towards a mass function which is regularly varying (see Bingham et al., 1989, for more on regular variation). Wang and Resnick (2023b) find that the tail index is given by

$$\iota \equiv \frac{1+\rho^\star+\delta}{\lambda_1},$$

where λ_1 is the largest eigenvalue among the matrices

$$A_m = \begin{bmatrix} \alpha & \alpha\rho_{m\bullet} \\ \gamma\rho_{\bullet m} & \gamma \end{bmatrix},$$

and $\rho_{m\bullet} = \sum_{r=1}^{K} \rho_{m,r}\pi_r$ and $\rho_{\bullet m} = \sum_{r=1}^{K} \rho_{r,m}\pi_r$ for $m = 1, \ldots, K$. Additionally, as $n \to \infty$,

$$\frac{|E(n)|}{n} - 1 \xrightarrow{a.s.} \rho^\star.$$

Since the tail index ι is an important parameter in the study of scale-free networks, we employ it to examine the efficiency of the Bayesian sampler. Suppose that $\iota^{(1)}, \ldots, \iota^{(2500)}$ are draws from the posterior constructed by using the draws of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ from Algorithm 1 after discarding the burn-in and assuming that $\boldsymbol{\theta}$ is known. We evaluate the effective sample size of the Monte Carlo estimator $2500^{-1}\sum_{i=1}^{2500} \iota^{(i)}$ of the posterior mean of ι. The effective sample size quantifies the extent to which correlation results in a loss of efficiency when estimating the posterior mean of ι (Roy, 2020). For each PA network realization used in Table 5, we estimate the effective sample size using the `mcse()` function in R's `mcmcse` package (Flegal et al., 2017). In particular, we employ the overlapping batch means estimator with batch size $2500^{1/3}$. The average effective sample size across all simulations

Method	$\pi_1 = 0.6$	$\rho_{11} = 0.1$	$\rho_{12} = 0.5$	$\rho_{21} = 0.4$	$\rho_{22} = 0.8$
			Mean(SE)		
B	0.604(0.015)	0.102(0.011)	0.500(0.018)	0.400(0.016)	0.800(0.020)
VB	0.587(0.028)	0.168(0.058)	0.523(0.035)	0.331(0.0216)	0.718(0.082)
VEM	0.599(0.073)	0.121(0.005)	0.441(0.091)	0.286(0.077)	0.686(0.074)
			% Coverage		
B	95	93	98	93	92
VB	19	0	6	11	0

Table 5: Average point estimates and standard errors for 100 networks generated from a PA model with $\boldsymbol{\theta} = (0.75, 0, 0.8, 0.8)$. Coverage rates for equal-tailed credible intervals produced by the B and VB methods are also provided.

Method	$\hat{K}$			
	1	2	3	4
B	0	86	11	3
VB	0	94	6	0
VEM	100	0	0	0

Table 6: Estimated K from 100 networks generated from a PA model with $\boldsymbol{\theta} = (0.75, 0, 0.8, 0.8)$ and $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ as in Table 5.

is 225.38 with a standard error of 6.587. Hence we may determine that the autocorrelations in the Markov chains drawn from Algorithm 1 are not overly onerous and that we may efficiently estimate the functionals of interest.

Table 6 displays the performance of the model selection criteria for 100 networks generated under the same PA model. For the fully Bayesian method, we initialize at $K_{\text{init}} = 1$. Despite the poor performance of the VB method at the parameter estimation, it captures the true $K = 2$ the most often, indicating that the ELBO is a good model selection criterion. The VEM algorithm always chooses $K = 1$, though we expect that this is again due to the lack of information in the data. The likelihood associated with $\boldsymbol{\pi}$ has a much larger role in the ICL for PA models than in stochastic block models. This, combined with the poor estimation of the classes for known K, results in the poor performance of the ICL criteria.

6 Data Examples

In this section we fit the heterogeneous reciprocal PA model to two real-world networks: the Facebook wall post data and the Reddit comment data. Although Facebook wallposts are now defunct, the network was also analyzed in Cirkovic et al. (2023a) using a homogeneous reciprocal PA model and allows us to highlight the remarkable improvement in the fit provided by modeling heterogeneous reciprocity. The Reddit network, on the other hand, provides user interaction dynamics that are reflective of those seen on present-day social media platforms.

We will use the VEM, VB, and fully Bayesian methods to fit the PA model with heterogeneous reciprocity to both real networks. Before introducing each network and evaluating the model fits, we state the inputs chosen for each algorithm; both are analyzed using the same inputs. For the VEM algorithm, we terminate the variational E-step when the increase in the ELBO is less than $\epsilon = 0.1$ and terminate the overall algorithm once the largest absolute difference in the estimated components of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ between M-steps falls below $\kappa = 0.001$. For the Bayesian methods, we again assume non-informative priors on $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. Analogous to the VEM algorithm, we terminate the VB procedure once the change in the ELBO falls below $\epsilon = 0.1$. Both the VEM and VB methods are fit for $K = 1, \ldots, 10$. The telescoping sampler for the fully Bayesian method is run for $M = 100{,}000$ MCMC iterates, where the first 90,000 iterates are discarded as burn-in. Within the telescoping sampler, we set $K_{\max} = 20$.

6.1 Facebook Wallposts

Now we apply the heterogeneous reciprocal PA model to the Facebook wall post data from KONECT analyzed in Viswanath et al. (2009) and Cirkovic et al. (2023a). The Facebook wall post data tracks a group of users in New Orleans and their wall posts from September 9th, 2004 to January 22nd, 2009. The network is temporal: when user u posts to user v's wall, a directed edge (u, v) is generated, and the timestamp of the post is recorded. The full dataset consists of 876,933 wall posts and 46,952 users. In Figure 1, we display the out/in-degree of each user in a trimmed version of the network; we postpone the discussion of the data cleaning procedure to the following paragraph. Note that upon first observation, the degree distribution indicates the existence of two populations that exhibit differing reciprocal behavior. The first group, concentrated on the out-degree axis, mostly posted on other users' walls while not receiving any posts on their own. The second group both sends and receives wall posts at a commensurate rate. Further, the marginal out/in-degree distributions exhibit power law tails as indicated by Figure 2 where, on the log-log scale, the empirical tail functions seem to scale linearly with large degrees.

In Cirkovic et al. (2023a), the Facebook wall post data was analyzed assuming that each user exhibited homogeneous reciprocal behavior. In the wording of Section 2.1, it was assumed that $\boldsymbol{\pi} \equiv 1$ and $\boldsymbol{\rho} \equiv \rho \in \mathbb{R}$. In doing so, the users concentrated on the out-degree axis in Figure 1 were excluded from the analysis as the homogeneous model could not model the observed heterogeneous reciprocal behavior. Additionally, by extreme value-based methods being sensitive to the choice of seed graph, Cirkovic et al. (2023a) also removed nodes that became inactive as the graph evolved, a phenomenon not modeled by the proposed PA model. The likelihood-based methodology in Cirkovic et al. (2023a) returned a homogeneous reciprocity estimate of $\hat{\rho} = 0.28$. The flexibility provided by the heterogeneous reciprocal PA model aims to capture the additional, intricate dynamics underlying the Facebook wall post data not previously considered in Cirkovic et al. (2023a).

According to the analysis of the Facebook wallpost data in Viswanath et al. (2009), there is a sudden uptick in the number of wall posts from July 2008 and onwards. They conjecture that this uptick is likely due to a Facebook redesign, introduced in July, that allowed users to interact with more wall posts through friend feeds. This likely results in a distributional shift in the network's evolution, and thus we discard the portion of the network

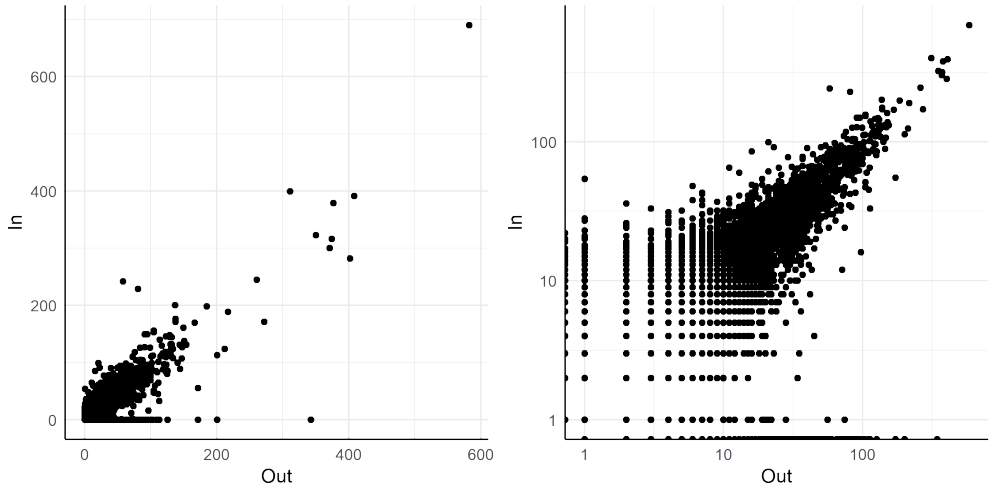

Figure 1: Out/in-degree plot for the Facebook wallpost data

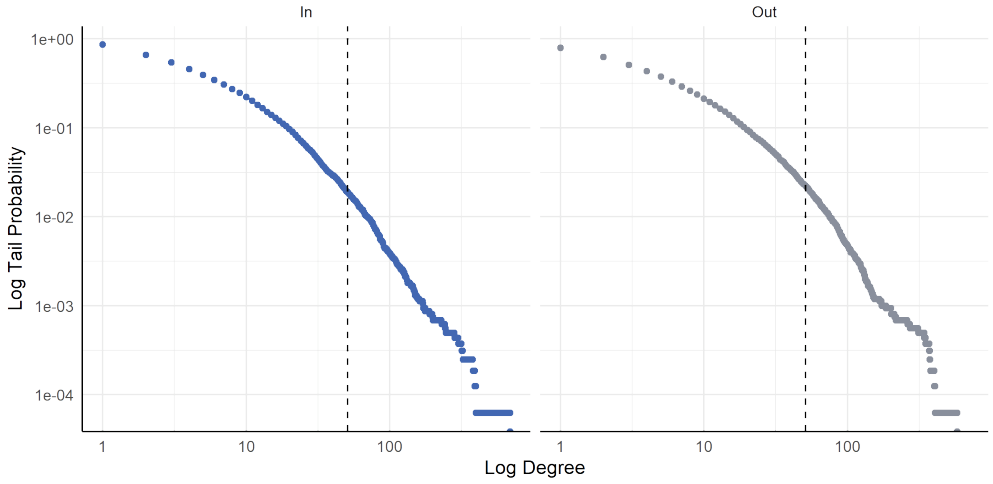

Figure 2: Plot of empirical tail probability function for the Facebook wallpost degrees on a log base 10 scale

observed beyond June 2008, resulting in a network with 22,286 nodes and 165,776 edges. This observation, however, may lead to additional analyses via changepoint detection (see Banerjee et al., 2023; Bhamidi et al., 2018; Cirkovic et al., 2023b, for example). Additionally, the evolution of the PA network specified in Section 2.1 posits that every new edge must attach to at least one node that was previously observed in the network evolution. To better adhere to this assumption we define a sequence of networks by first selecting the node with the largest total degree and pairing it with the first node it makes a connection with to create a seed graph $G(0)$. Then, we only retain the edges (u, v) that are (i) observed after the introduction of the seed graph and (ii) $u \in V(k-1)$ or $v \in V(k-1)$. This trimming procedure results in a connected network of 16,099 nodes and 123,920 edges that could have realistically been generated by a heterogeneous reciprocal PA model.

The reciprocal PA model assumes that reciprocal edges (t_k, s_k) are generated instantaneously with their parent edge (s_k, t_k). However, in the Facebook wall post network, it is likely that in the time between reciprocated wall posts, wall posts between other users have been generated. Thus, similar to Cirkovic et al. (2023a), we employ window estimators to identify reciprocal edges. That is, if $e_k = (s_k, t_k)$ has a reciprocal counterpart (t_k, s_k) appear in 24 hours, we attribute the event $R_k = 1$ to the edge e_k, redefine $e_k := e_k \cup (t_k, s_k)$ and drop (t_k, s_k) from the edgelist. This results in an edgelist that is in alignment with Section 2.1. We note that this is only one way to define reciprocal pairs. More complex models would account for reciprocation triplets, quadruplets and so on that could potentially be modeled via a geometric distribution.

To conclude our exploratory data analysis, we study the tail behavior of the out/in-degrees for the trimmed Facebook network. We employ the minimum distance procedure (Clauset et al., 2009) on the total degrees to obtain a threshold beyond which a power-law tail for the in/out-degree can be safely assumed. The minimum distance procedure computes a tail threshold of 51. Note that computing the tail threshold on the total degree implicitly assumes that the out/in-degree tails have the same power-law index. We find this to be a reasonable assumption as indicated by the similarity of the empirical tail functions in Figure 2. In fact, using a threshold of 51, the tail index estimates for the out/in-degrees are 2.212 and 2.231, respectively. Further observation of Figure 1 indicates that, beyond this threshold, there is an extremal dependence structure in the out/in-degree distribution; nodes with a total degree larger than 51 tend to cluster around multiple lines through the origin. This extremal dependence structure is further analyzed in Appendix A.

The global PA parameters $\boldsymbol{\theta}$ are estimated by maximizing the likelihood $p(\cdot \mid \boldsymbol{\theta})$. Maximum likelihood returns $(\hat{\alpha}, \hat{\beta}, \hat{\delta}_{\text{in}}, \hat{\delta}_{\text{out}}) = (0.071, 0.829, 1.756, 1.571)$. The small values of $\hat{\delta}_{\text{in}}$ and $\hat{\delta}_{\text{out}}$ indicate that preferential attachment is indeed a viable mechanism to describe how users send and receive wall posts. Analyzing the reciprocal component of the model, the VEM algorithm identifies 3 classes, while the VB and fully Bayesian algorithms identify 6 and 11 clusters, respectively. Figure 3 displays the ICL, ELBO, and posterior of K for the VEM, VB, and fully Bayesian methods. The ICL criterion clearly identifies $K = 3$ as the choice that optimally balances model parsimony with fidelity to the data. Though the VB method chooses $K = 6$, we note that the ELBO for the VB method becomes very flat at $K = 4$, indicating that perhaps a simpler model may fit the data nearly as well as the model with $K = 6$ mixture components. Further, the telescoping sampler seems to overfit the number of classes by producing classes whose mixture weights have small pos-

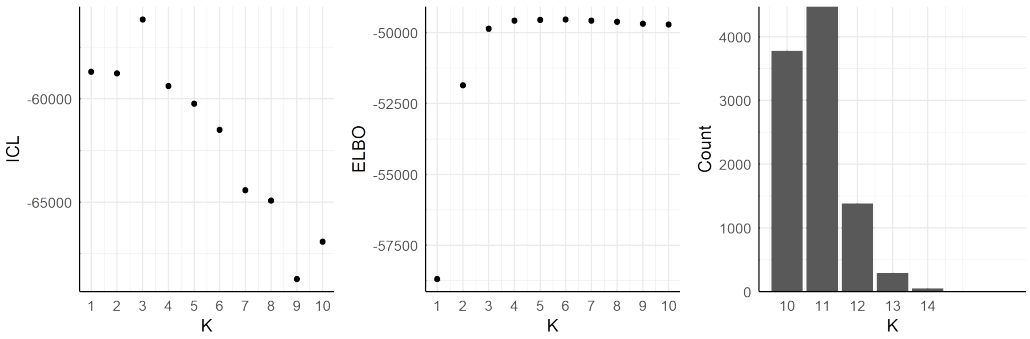

Figure 3: ICL, ELBO, and posterior on K from the VEM, VB, and fully Bayesian methods. For the VEM and VB algorithms, we consider $K = 1, \ldots, 10$.

terior means. We suspect that the fully Bayesian method overfits the number of mixture components due to model misspecification. It is unlikely that the Facebook wall post data exactly follows the specification in Section 2.1. For example, there is empirical evidence that the degree of each node may influence reciprocal behavior (Cheng et al., 2011). There is strong evidence that mixtures of finite mixtures do not reliably learn the number of mixture components under model misspecification (Cai et al., 2021; Miller and Dunson, 2018).

The estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ from the VEM and VB algorithms are

$$\hat{\boldsymbol{\pi}}_{\text{VEM}} = \begin{bmatrix} 0.538 \\ 0.251 \\ 0.211 \end{bmatrix}, \qquad \hat{\boldsymbol{\rho}}_{\text{VEM}} = \begin{bmatrix} 0.242 & 0.273 & 0.001 \\ 0.597 & 0.650 & 0.001 \\ 0.0701 & 0.053 & 0.001 \end{bmatrix}$$

$$\hat{\boldsymbol{\pi}}_{\text{VB}} = \begin{bmatrix} 0.122 \\ 0.285 \\ 0.153 \\ 0.060 \\ 0.197 \\ 0.184 \end{bmatrix}, \qquad \hat{\boldsymbol{\rho}}_{\text{VB}} = \begin{bmatrix} 0.088 & 0.094 & 0.084 & 0.038 & 0.001 & 0.083 \\ 0.383 & 0.427 & 0.431 & 0.182 & 0.001 & 0.375 \\ 0.670 & 0.699 & 0.718 & 0.433 & 0.001 & 0.641 \\ 0.467 & 0.464 & 0.499 & 0.214 & 0.002 & 0.437 \\ 0.089 & 0.089 & 0.059 & 0.036 & 0.005 & 0.082 \\ 0.206 & 0.225 & 0.230 & 0.098 & 0.001 & 0.201 \end{bmatrix}$$

Note that both the VEM and VB methods identify a group of nodes that receives nearly no reciprocal edges as indicated by a column of near-zero estimates in $\boldsymbol{\rho}$. The fully Bayesian methods agree; class 1 has class probability 0.195 and, on average, receives a reciprocal edge with probability 0.002 as estimated by the posterior means of π_1 and $\rho_{\bullet,1} = \sum_{m=1}^{11} \pi_m \rho_{m,1}$, respectively. Additionally, the VEM algorithm seems to indicate that reciprocity on Facebook is receiver-dependent; we estimate that $\rho_{m,1} \approx \rho_{m,2}$ for $m = 1, 2$, indicating reciprocity does not tend vary depending on who sends the initial message.

Figure 4 displays the degree distribution of the trimmed Facebook wall post network, grouped by the VEM class estimates. The VEM algorithm clearly identifies class 3 as nodes that do not receive reciprocal edges. This class may very well correspond to bot users who regularly post spam messages that most users ignore. Despite class 2 having a heavier tail, classes 1 and 2 tend to concentrate in similar regions of $\mathbb{R}_+$. Further, the similarity of the estimates $\hat{\boldsymbol{\rho}}_{\text{VEM}}$ indicate that classes 1 and 2 engage in similar reciprocal behavior.

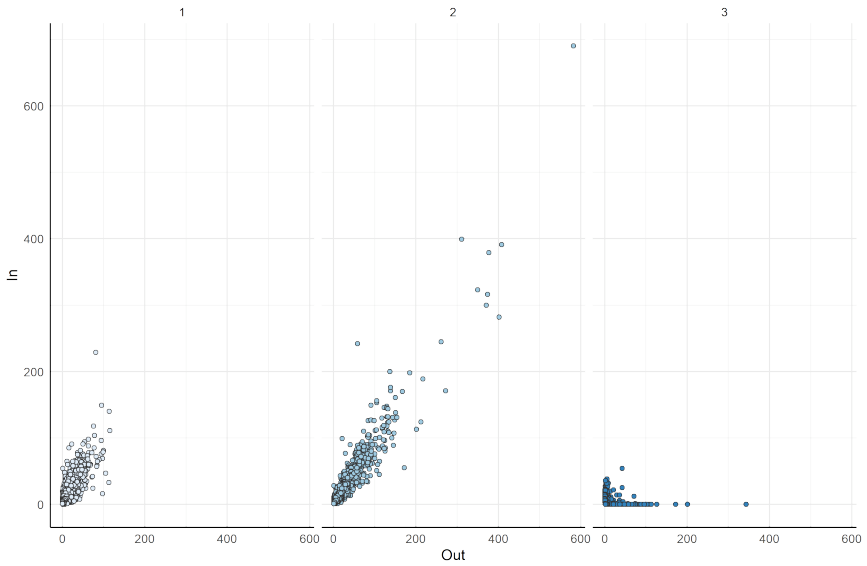

Figure 4: Reciprocal components identified by the VEM algorithm

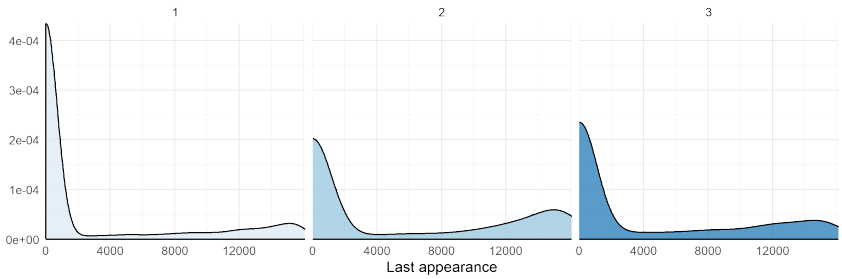

Figure 5: Density plots for last appearance time (by class) of each node that posted more than once in the network

These visual measures warrant further inspection of the differences between classes 1 and 2. Figure 5 displays the discrete time of the last post made by each node in the network that posts more than once. Note that nodes in class 1 are more likely to become inactive in the early period of the network evolution. These inactive nodes were noted by Cirkovic et al. (2023a) and Viswanath et al. (2009) as well. The lighter tails of class 1 thus can be explained by the relatively short lifetimes of the nodes, as such nodes do not have as long enough time to send and receive wall posts. The VEM algorithm may have picked up on this inactivity by proxy. However, such observations warrant extension to a preferential attachment model that incorporates nodes that become inactive over time.

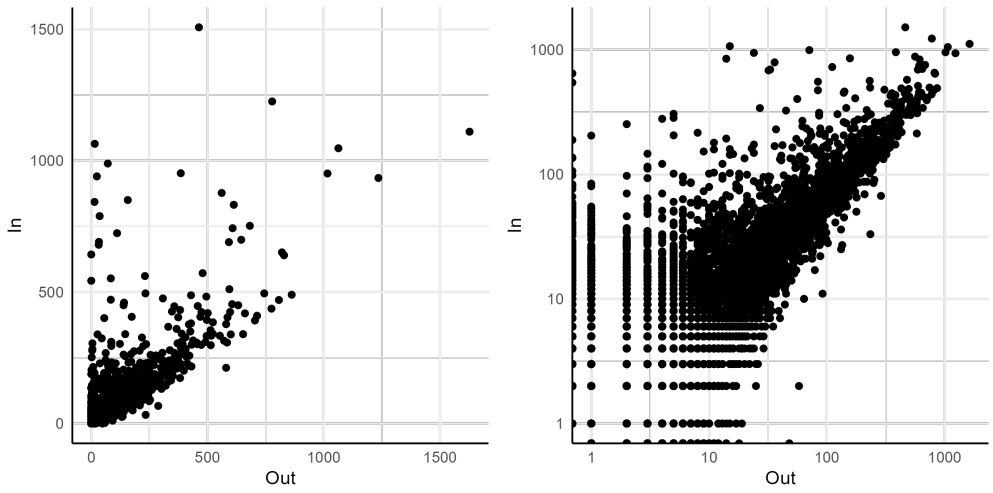

Figure 6: Out/in-degree plot for the Reddit data

6.2 Reddit replies

We additionally analyze Reddit user replies from December 11th, 2005 to December 31st, 2006 (Hessel et al., 2016; Liu et al., 2019). Here an edge (u, v) indicates that user u commented on user v's comment or post. The network is temporal; each edge has an associated timestamp. Following the analysis in Section 6.1, we trim the Reddit network by defining the first comment containing the user with maximum total degree as the seed graph and retaining any newer comments that have at least one user that has already commented in the trimmed network. By trimming the network in such a fashion, we obtain a graph that mimics the evolution of the reciprocal PA process. The resulting network contains 14,080 users and 197,719 comments.

As displayed by Figure 6, the degree distribution for the trimmed Reddit is heterogeneous. A significant segment of users post and receive comments at a common rate, while a smaller portion of users rarely comment but generate an exceptional amount of interaction. Additionally, Figure 7 indicates that the marginal out/in-degree distributions can be suitably modeled via power-law tails since, on the log-log scale, the empirical tail function and large degrees are approximately linearly related. Here, large degrees refer to the degrees that surpass the threshold chosen by the minimum distance procedure applied to total degrees. The threshold, 231, is computed using the total degree since marginal the out/in-degree tail indices are close; using a threshold of 231, the estimated tail indices for the out/in-degrees are 3.01 and 2.87, respectively. Thus the heterogeneous, power-law tails of the degree distribution lead us to expect that the reciprocal PA model would provide a good fit to the Reddit data.

We fit the heterogeneous reciprocal PA model to the Reddit network using the VEM, VB, and fully Bayesian methods. Each method maximizes the likelihood $p(\cdot \mid \boldsymbol{\theta})$ in order estimate $\boldsymbol{\theta}$. Maximum likelihood returns global parameter estimates of $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}, \hat{\delta}_{\text{in}}, \hat{\delta}_{\text{out}}) = (0.048, 0.919, 1.291, 0.717)$. Relative to $\hat{\delta}_{\text{in}}$, the smaller value of $\hat{\delta}_{\text{out}}$ indicates that the extent to which users reply to comments more heavily depends on their tendency to have already

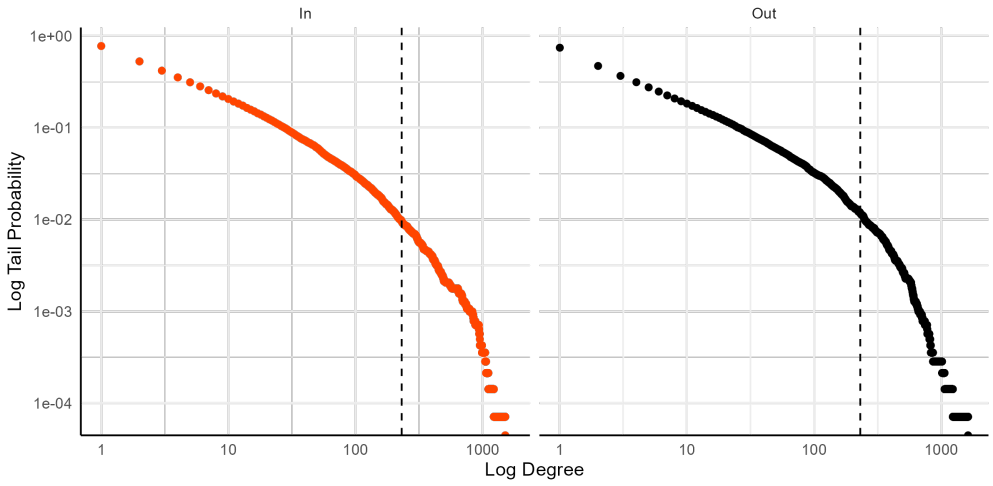

Figure 7: Plot of empirical tail probability function for the Reddit degrees on a log base 10 scale

replied to comments in the past. In comparison, receiving many replies is less impactful on the ability of users to continue receiving replies in the future. Independently, however, the small values of $\hat{\delta}_{\text{in}}$ and $\hat{\delta}_{\text{out}}$ indicate that preferential attachment tends to govern comment behavior on Reddit.

We now analyze the reciprocity component of the model. Figure 8 displays the ICL, ELBO, and posterior of K for the VEM, VB, and fully Bayesian methods. The ICL criterion chooses $K = 2$ as the optimal number of reciprocal components, though $K = 3$ results in a similar ICL. The maximal ELBO occurs for $K = 8$, though marginal gains are made in the ELBO for $K > 3$. For the fully Bayesian method, we estimate that the posterior places maximal probability mass on $K = 17$. As in Section 6.1, we suspect that the large number of communication classes inferred by the fully Bayesian method is a result of its sensitivity to model misspecification.

We report the estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ for VEM and VB algorithms below:

$$\hat{\boldsymbol{\pi}}_{\text{VEM}} = \begin{bmatrix} 0.210 \\ 0.790 \end{bmatrix}, \qquad \hat{\boldsymbol{\rho}}_{\text{VEM}} = \begin{bmatrix} 0.318 & 0.205 \\ 0.052 & 0.040 \end{bmatrix}$$

$$\hat{\boldsymbol{\pi}}_{\text{VB}} = \begin{bmatrix} 0.602 \\ 0.020 \\ 0.040 \\ 0.021 \\ 0.044 \\ 0.015 \\ 0.219 \\ 0.040 \end{bmatrix}, \qquad \hat{\boldsymbol{\rho}}_{\text{VB}} = \begin{bmatrix} 0.055 & 0.020 & 0.026 & 0.017 & 0.017 & 0.061 & 0.030 & 0.030 \\ 0.395 & 0.534 & 0.561 & 0.214 & 0.397 & 0.731 & 0.470 & 0.552 \\ 0.224 & 0.402 & 0.328 & 0.150 & 0.226 & 0.321 & 0.292 & 0.340 \\ 0.094 & 0.152 & 0.125 & 0.483 & 0.085 & 0.435 & 0.120 & 0.128 \\ 0.188 & 0.308 & 0.287 & 0.184 & 0.218 & 0.408 & 0.225 & 0.312 \\ 0.092 & 0.129 & 0.149 & 0.082 & 0.098 & 0.561 & 0.100 & 0.210 \\ 0.092 & 0.178 & 0.127 & 0.092 & 0.097 & 0.177 & 0.120 & 0.130 \\ 0.126 & 0.203 & 0.266 & 0.125 & 0.162 & 0.399 & 0.170 & 0.210 \end{bmatrix}.$$

As evidenced by class 2 for the VEM algorithm and class 1 for the VB algorithm, both methods identify a large contingent of individuals who do not tend to reply to comments.

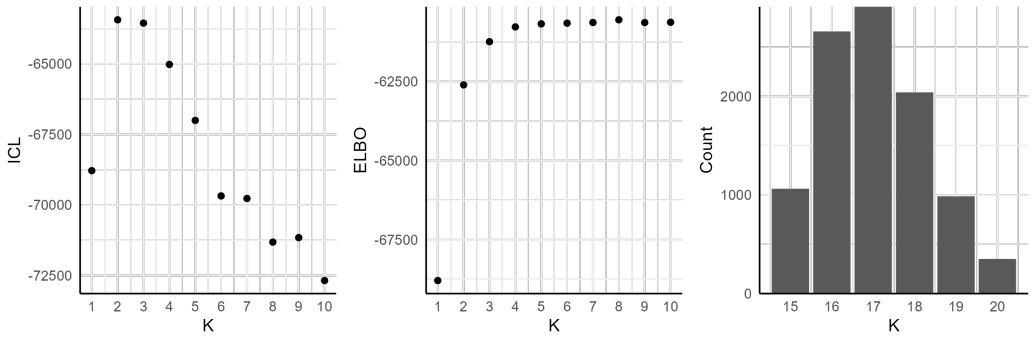

Figure 8: ICL, ELBO, and posterior on K from the VEM, VB, and fully Bayesian methods for the Reddit network. For the VEM and VB algorithms, we consider $K = 1, \ldots, 10$.

The fully Bayesian also arrives at the same conclusion as it reports that class 7, a class with estimated class probability 0.201, tends to reply to comments with probability 0.032 as estimated by the posterior mean of $\rho_{7,\bullet} = \sum_{m=1}^{17} \pi_m \rho_{7,m}$. Further, both the VB and fully Bayesian methods identify classes with very low-class probabilities, indicating that the methods may be overfitting the number of communication classes.

Given that the ICL for $K = 2, 3$ are comparable and the ELBO becomes relatively flat after $K = 3$, we set $K = 3$ as a consensus choice and employ Algorithm 1 to refit the heterogeneous reciprocal PA model. As displayed in Section 5, the fully Bayesian method allows superior uncertainty quantification of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ while the ICL and ELBO criteria more consistently capture the correct number of communication classes. Thus, we recommend this melding of methods as a computationally efficient, robust alternative for real-world network analysis. As before, we run Algorithm 1 for $M = 100{,}000$ MCMC iterates, where the first 90,000 iterates are discarded as burn-in.

Estimates of the marginal posteriors of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ from Algorithm 1 are provided in Figure 9. Note that Algorithm 1 identifies a large contingent of nodes that do not tend to reciprocate incoming edges. This is supported by the fact that the estimated posterior mean of class 1's average reply probability, $\rho_{1,\bullet} \equiv \sum_{m=1}^{K} \pi_m \rho_{1,m}$, is given by 0.022. It also identifies a small group of users that engage in high levels of reciprocity, most of which occurs within the same communication class. We additionally observe that the reciprocal behavior between classes is asymmetric. For example, a 95% credible interval for $\rho_{2,3}/\rho_{3,2}$ is given by $(1.956, 2.157)$, indicating that class 2 is approximately twice as likely to reply to a comment from class 3 than class 3 is to reply to a comment from class 2.

Lastly, we evaluate the ability of the consensus method to capture the heterogeneity of the Reddit degree distribution. The degree distribution of the Reddit data is plotted by estimated class in Figure 10. Here, the estimated class corresponds to the maximum a posteriori class as estimated by the posterior samples from the Gibbs sampler. Algorithm 1 clearly isolates the nodes that concentrate along the in-degree axis as class 1. Additionally, when compared to class 3, class 2 consists of nodes with out/in-degree pairs that concentrate more heavily on a line through the origin. For higher levels of reciprocity, we expect that

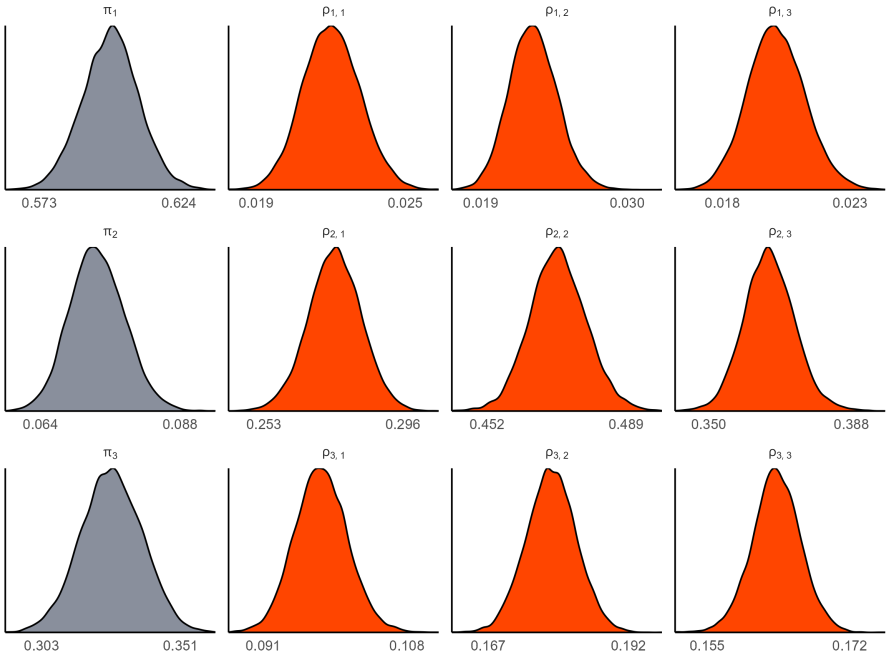

Figure 9: Marginal posterior density estimates for $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ from the Reddit network using Algorithm 1 with $K = 3$.

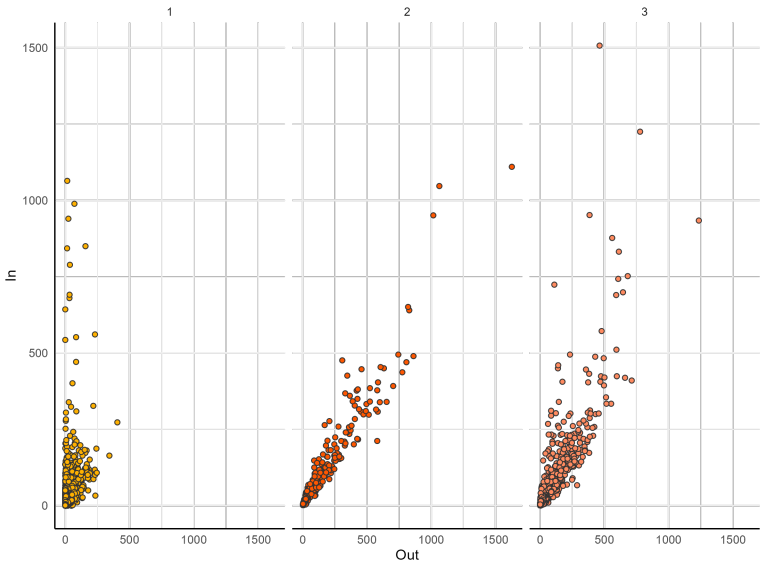

Figure 10: Reciprocal components as identified by Algorithm 1 with $K = 3$.

the out- and in-degree of a given node will be highly correlated, whereas for lower levels of reciprocity, the degree distribution will diffuse as in class 3.

7 Conclusion

In this paper, we outline a preferential attachment model with heterogeneous reciprocity and offer three methods for fitting the model to both simulated and real-world networks. Through simulations, we find that when analyzing networks that have many edges compared to the number of nodes, the variational alternatives offer similar performance to the fully Bayesian method in terms of point estimation at less computational cost. However, the credible intervals generated by the fully Bayesian method more reliably capture the true data-generating parameters. We also compare the ability of each method to select the number of communication classes in heterogeneous reciprocal PA networks. Generally speaking, when the number of edges is again large compared to the number of nodes, all three methods consistently choose the true number of classes, with the fully Bayesian method having a slight tendency to overfit. We then showcase the ability of the heterogeneous reciprocal PA model to capture non-uniform reciprocal behavior across users in the Facebook wallpost and Reddit comment networks. The proposed model offers the additional flexibility needed to model such data.

Upon analyzing the Facebook wallpost network, we find that the VEM algorithm uncovered two reciprocal classes that engage in somewhat similar reciprocal behavior, though one of the classes consisted of more inactive users. The propensity of some users to become inactive in a network as it evolves is a common feature of many networks. It warrants the extension of the preferential attachment model to account for such behavior. In future work, we will also consider models allowing users to become inactive as the network grows over time.

Additionally, we believe that a powerful way to model both the Facebook and Reddit data is to assume that each user is embedded with their own reciprocity parameter which is drawn from some prior distribution. Such a model would account for individual heterogeneity in reciprocal behavior. For the Facebook and Reddit networks in particular, a spike-and-slab prior with a point mass at zero would capture the users that do not reciprocate edges while also allotting a sufficient amount of flexibility to take into account individual effects. We intend to explore such a model in future work.

Acknowledgments and Disclosure of Funding

D. Cirkovic is supported by NSF Grant DMS-2210735. T. Wang is supported by the National Natural Science Foundation of China Grant 12301660 and the Science and Technology Commission of Shanghai Municipality Grant 23JC1400700.

Appendix A: Statistical Tools for Multivariate Extremes

Here we detail, in a non-technical fashion, some tools used to analyze data subject to extremal observations. For more rigorous treatments, we refer to the works of Beirlant et al. (2004) and Resnick (2007). A central goal in the study of multivariate extremes is to identify how extremes cluster. In other words, if one or more components of a random vector are large, how likely is it that the other components of the random vector will also be large? For PA models with homogeneous reciprocity, Cirkovic et al. (2023a) proved that the extremal out/in-degrees tend to cluster on a line through the origin. With heterogeneous reciprocity, Wang and Resnick (2023b) proved that the model with $\beta = 0$ generates extreme out/in-degrees that concentrate on multiple lines through the origin.

An exploratory tool used to identify where such extremes cluster in $\mathbb{R}_+^2$ is the *angular density*, a plot of the angles

$$\Theta_r \equiv \left\{ D_v^{\text{out}}(n)/(D_v^{\text{out}}(n) + D_v^{\text{in}}(n)) : v \in V(n), D_v^{\text{out}}(n) + D_v^{\text{in}}(n) > r \right\}$$

for some large threshold r. Intuitively, if the angular density concentrates mass around some point in $(0, 1)$, then one would expect extremes to cluster on a line through the origin. On the other hand, if the angular density only places mass on the set $\{0, 1\}$, then the out/in-degrees are *asymptotically independent*; a large in-degree does not necessarily imply a large out-degree, and vice versa. Figure 11 displays the angular density for the Facebook wallpost and Reddit comment data analyzed in Section 6.1.

When the angular density concentrates mass on the set $\{0.5, 1\}$, it indicates the existence of two extremal populations: one that has approximately equal in/out-degree, and another that has high out-degree but small in-degree. The threshold for Θ_r was chosen as $r = 51$ by the minimum distance method applied to the total degrees (Clauset et al., 2009). The minimum distance method chooses a threshold that minimizes the Kolmogorov-Smirnov distance between a power-law tail and the empirical tail of the observations above the threshold. Note that the angular density is naturally sensitive to the choice of r. If r is chosen too large, some extremal features of the data may be passed over, while if r is chosen too small, the extremal behavior will be corrupted by non-extremal observations.

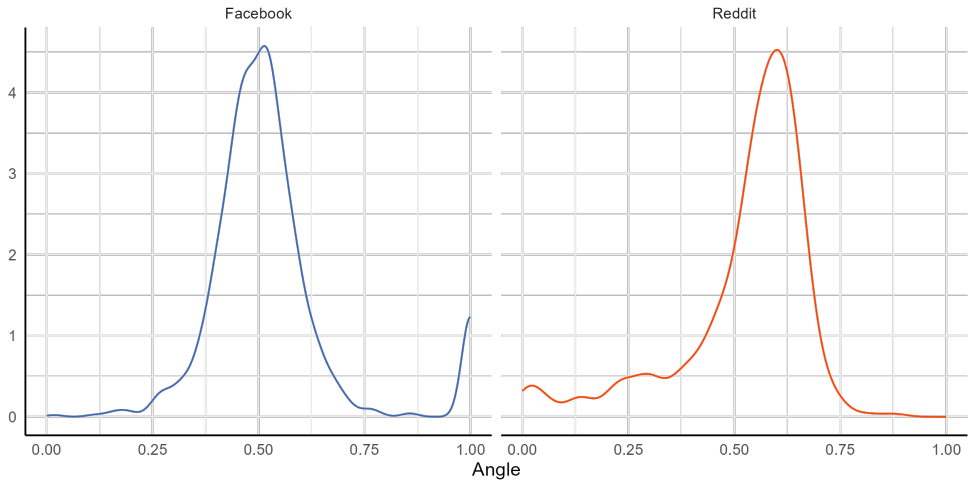

Figure 11: Angular density for the Facebook wallpost and Reddit comment data with thresholds chosen via the minimum distance method.

We now have the tools to describe the initialization of the VEM algorithm for a fixed K presented in Section 3.2.2. First, the set Θ_r is constructed via threshold r chosen by the minimum distance method available in R package `igraph` (Csardi et al., 2006). We then employ K-means on the set Θ_r to determine an initial clustering of nodes. Note that this only clusters nodes with a total degree larger than r. This clustering is then used to compute empirical class probabilities $(\hat{\pi}_r)_{r=1}^K$ and empirical reciprocities $(\hat{\rho}_{m,r})_{m,r=1}^K$. Note that $(\hat{\rho}_{m,r})_{m,r=1}^K$ is computed only on edges that connect nodes which both have total degree larger than r. $(\hat{\pi}_r)_{r=1}^K$ and $(\hat{\rho}_{m,r})_{m,r=1}^K$ are thus used as initial parameter values and the initial $(\tau_{w,\ell})_{w\in V(n),\ell\in\{1,\ldots,K\}}$ are chosen according to a uniform distribution on the K-simplex. The full initialization algorithm is given in Algorithm 5.

Appendix B: Sample Derivations for the VEM Algorithm

In this appendix, we present some sample derivations for the variational EM algorithm presented in Section 3.2.2. We note that the derivations are very similar to those of Daudin et al. (2008) and Latouche et al. (2012), though we reformulate them in our setting for convenience. The same type of calculations can be employed to derive the variational Bayes algorithm.

Derivation of the ELBO

In this section, we derive (11). Recall that we posit the mean-field variational family on the communication types $W(n)$ given by

$$q(W(n)) = \prod_{v\in V(n)} q_v(W_v).$$

Algorithm 5 Initialization of VEM for heterogeneous reciprocal PA

Input: Graph $G(n)$, # communication types K

Output: Initial variational EM estimates $\hat{\boldsymbol{\pi}}_{\text{VEM}}$ and $\hat{\boldsymbol{\rho}}_{\text{VEM}}$

1. Compute the tail threshold r according to the minimum distance procedure.
2. Construct the sets

$$\aleph_r = \left\{v \in V(n) : D_v^{\text{out}}(n) + D_v^{\text{in}}(n) > r\right\}$$
$$\Theta_r = \left\{D_v^{\text{out}}(n)/(D_v^{\text{out}}(n) + D_v^{\text{in}}(n)) : v \in \aleph_r\right\}$$

3. Employ K-means on Θ_r to form initial communication class estimates $\hat{W}_v$ for $v \in \aleph_r$
4. Form initial VEM estimates via

for $m = 1$ to K **do**

$$\hat{\pi}_m = \frac{1}{|\aleph_r|} \sum_{v \in \aleph_r} 1_{\{\hat{W}_v = m\}}$$

for $r = 1$ to K **do**

$$\hat{\rho}_{m,r} = \frac{\sum_{k:\hat{W}_{s_k}=r, \hat{W}_{t_k}=m} 1_{\{R_k=1\}}}{\left|\left\{k : \hat{W}_{s_k} = r, \hat{W}_{t_k} = m\right\}\right|}$$

end for

end for

Then, the ELBO is given by

$$\mathrm{ELBO}(q, \boldsymbol{\pi}, \boldsymbol{\rho}) = E_q \left[\log p \left(W(n), (e_k)_{k=1}^n \mid \boldsymbol{\pi}, \boldsymbol{\rho}\right)\right] - E_q \left[\log q(W(n))\right].$$

Focusing on the first term, recall that the log-likelihood is given by

$$\begin{aligned}
&\log p \left(W(n), (e_k)_{k=1}^n \mid \boldsymbol{\pi}, \boldsymbol{\rho}\right) \\
=&\sum_{k=1}^{n} \sum_{r=1}^{K} \left(1_{\{J_k=1\}} 1_{\{W_{s_k}=r\}} + 1_{\{J_k=3\}} 1_{\{W_{t_k}=r\}}\right) \\
&+ \sum_{k=1}^{n} \sum_{r=1}^{K} \sum_{m=1}^{K} 1_{\{R_k=1\}} 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} \log \rho_{m,r} \\
&+ \sum_{k=1}^{n} \sum_{r=1}^{K} \sum_{m=1}^{K} 1_{\{R_k=0\}} 1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}} \log(1-\rho_{m,r}).
\end{aligned}$$

Taking an expectation with respect to q gives that

$$\begin{aligned}
&E_q \left[\log p \left(W(n), (e_k)_{k=1}^n \mid \boldsymbol{\pi}, \boldsymbol{\rho}\right)\right] \\
=&\sum_{k=1}^{n} \sum_{r=1}^{K} \left(1_{\{J_k=1\}} E_q \left[1_{\{W_{s_k}=r\}}\right] + 1_{\{J_k=3\}} E_q \left[1_{\{W_{t_k}=r\}}\right]\right) \\
&+ \sum_{k=1}^{n} \sum_{r=1}^{K} \sum_{m=1}^{K} 1_{\{R_k=1\}} E_q \left[1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}}\right] \log \rho_{m,r} \\
&+ \sum_{k=1}^{n} \sum_{r=1}^{K} \sum_{m=1}^{K} 1_{\{R_k=0\}} E_q \left[1_{\{W_{s_k}=r\}} 1_{\{W_{t_k}=m\}}\right] \log(1-\rho_{m,r}),
\end{aligned}$$

and employing the mean-field family assumption,

$$
\begin{aligned}
&E_q\left[\log p\left(W(n),(e_k)_{k=1}^n \mid \boldsymbol{\pi},\boldsymbol{\rho}\right)\right] \\
&=\sum_{k=1}^{n}\sum_{r=1}^{K}\left(1_{\{J_k=1\}}E_q\left[1_{\{W_{s_k}=r\}}\right]+1_{\{J_k=3\}}E_q\left[1_{\{W_{t_k}=r\}}\right]\right) \\
&\quad+\sum_{k=1}^{n}\sum_{r=1}^{K}\sum_{m=1}^{K}1_{\{R_k=1\}}E_q\left[1_{\{W_{s_k}=r\}}\right]E_q\left[1_{\{W_{t_k}=m\}}\right]\log\rho_{m,r} \\
&\quad+\sum_{k=1}^{n}\sum_{r=1}^{K}\sum_{m=1}^{K}1_{\{R_k=0\}}E_q\left[1_{\{W_{s_k}=r\}}\right]E_q\left[1_{\{W_{t_k}=m\}}\right]\log(1-\rho_{m,r}) \\
&=\sum_{k=1}^{n}\sum_{r=1}^{K}\left(1_{\{J_k=1\}}\tau_{s_k,r}+1_{\{J_k=3\}}\tau_{t_k,r}\right) \\
&\quad+\sum_{k=1}^{n}\sum_{r=1}^{K}\sum_{m=1}^{K}1_{\{R_k=1\}}\tau_{s_k,r}\tau_{t_k,m}\log\rho_{m,r} \\
&\quad+\sum_{k=1}^{n}\sum_{r=1}^{K}\sum_{m=1}^{K}1_{\{R_k=0\}}\tau_{s_k,r}\tau_{t_k,m}\log(1-\rho_{m,r}).
\end{aligned}
$$

Finally, the entropy term is given by

$$
\begin{aligned}
E_q\left[\log q(W(n))\right] &= E_q\left[\sum_{v\in V(n)}\sum_{r=1}^{K}1_{\{W_v=r\}}\log\tau_{v,r}\right] \\
&= \sum_{v\in V(n)}\sum_{r=1}^{K}E_q\left[1_{\{W_v=r\}}\right]\log\tau_{v,r} \\
&= \sum_{v\in V(n)}\sum_{r=1}^{K}\tau_{v,r}\log\tau_{v,r}.
\end{aligned}
$$

Derivation of the E-Step

In this section, we derive the E-step of the variational EM algorithm (Step 1 of Algorithm 3). Recall that the E-step maximizes the ELBO for the variational density q. We perform this optimization with a coordinate ascent. From Blei et al. (2017), for every $\ell = 1,\ldots,K$, the optimal $q_w(W_w)$ satisfies

$$
\tau_{w,\ell} = q_w^\star\left(W_w=\ell\right) \propto \exp\left\{E_{q_{-w}}\left[\log p\left(W_w=\ell \mid (W_v)_{v\neq w},(e_k)_{k=1}^n,\boldsymbol{\pi},\boldsymbol{\rho}\right)\right]\right\},
$$

which, by definition of conditional density, is proportional to

$$
\propto \exp\left\{E_{q_{-w}}\left[\log p\left(W_w=\ell,(W_v)_{v\neq w},(e_k)_{k=1}^n \mid \boldsymbol{\pi},\boldsymbol{\rho}\right)\right]\right\}.
$$

Here, q_{-w} denotes the variational density on $(W_v)_{v\neq w}$. Up to some constant C not depending on w, the log-likelihood term is given by

$$\begin{aligned}\log p\left(W_w=\ell,(W_v)_{v\neq w},(e_k)_{k=1}^n \mid \boldsymbol{\pi},\boldsymbol{\rho}\right) = \log\pi_\ell &+ \sum_{m=1}^K \log\rho_{m,\ell} \sum_{k:s_k=w} 1_{\{W_{t_k}=m\}}1_{\{R_k=1\}} \\ &+ \sum_{m=1}^K \log(1-\rho_{m,\ell}) \sum_{k:s_k=w} 1_{\{W_{t_k}=m\}}1_{\{R_k=0\}} \\ &+ \sum_{r=1}^K \log\rho_{\ell,r} \sum_{k:t_k=w} 1_{\{W_{s_k}=r\}}1_{\{R_k=1\}} \\ &+ \sum_{r=1}^K \log(1-\rho_{\ell,r}) \sum_{k:t_k=w} 1_{\{W_{s_k}=r\}}1_{\{R_k=0\}} + C,\end{aligned}$$

and taking an expectation with respect to q_{-w} gives

$$\begin{aligned}\log p\left(W_w=\ell,(W_v)_{v\neq w},(e_k)_{k=1}^n \mid \boldsymbol{\pi},\boldsymbol{\rho}\right) = \log\pi_\ell &+ \sum_{m=1}^K \log\rho_{m,\ell} \sum_{k:s_k=w} \tau_{t_k,m}1_{\{R_k=1\}} \\ &+ \sum_{m=1}^K \log(1-\rho_{m,\ell}) \sum_{k:s_k=w} \tau_{t_k,m}1_{\{R_k=0\}} \\ &+ \sum_{r=1}^K \log\rho_{\ell,r} \sum_{k:t_k=w} \tau_{s_k,r}1_{\{R_k=1\}} \\ &+ \sum_{r=1}^K \log(1-\rho_{\ell,r}) \sum_{k:t_k=w} \tau_{s_k,r}1_{\{R_k=0\}} + C.\end{aligned}$$

Hence,

$$\begin{aligned}\tau_{w,\ell} = q_w^\star\left(W_w=\ell\right) &\propto \exp\left\{E_{q_{-w}}\left[\log p\left(W_w=\ell,(W_v)_{v\neq w},(e_k)_{k=1}^n \mid \boldsymbol{\pi},\boldsymbol{\rho}\right)\right]\right\} \\ &\propto \pi_\ell \prod_{m=1}^K \rho_{m,\ell}^{\sum_{k:s_k=w}\tau_{t_k,m}1_{\{R_k=1\}}}(1-\rho_{m,\ell})^{\sum_{k:s_k=w}\tau_{t_k,m}1_{\{R_k=0\}}} \\ &\quad\times \prod_{r=1}^K \rho_{\ell,r}^{\sum_{k:t_k=w}\tau_{s_k,r}1_{\{R_k=1\}}}(1-\rho_{\ell,r})^{\sum_{k:t_k=w}\tau_{s_k,r}1_{\{R_k=0\}}}.\end{aligned} \tag{12}$$

Thus, in the E-step, one cycles through (12) for each $w \in V(n)$ until the ELBO no longer meaningfully increases.

References

Sayan Banerjee, Shankar Bhamidi, and Iain Carmichael. Fluctuation bounds for continuous time branching processes and evolution of growing trees with a change point. *The Annals of Applied Probability*, 33(4):2919–2980, 2023.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons, 2004.

Shankar Bhamidi, J Michael Steele, and Tauhid Zaman. Twitter event networks and the superstar model. *The Annals of Applied Probability*, 25(5):2462–2502, 2015.

Shankar Bhamidi, Jimmy Jin, and Andrew Nobel. Change point detection in network models: Preferential attachment and long range dependence. *The Annals of Applied Probability*, 28(1):35–78, 2018.

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50): 21068–21073, 2009.

Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*. Number 27. Cambridge university press, 1989.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pages 1158–1169. PMLR, 2021.

Justin Cheng, Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Predicting reciprocity in social networks. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 49–56. IEEE, 2011.

Daniel Cirkovic, Tiandong Wang, and Sidney I Resnick. Preferential attachment with reciprocity: Properties and estimation. *Journal of Complex Networks*, 11(5):cnad031, 2023a.

Daniel Cirkovic, Tiandong Wang, and Xianyang Zhang. Likelihood-based inference for random networks with changepoints. *arXiv preprint arXiv:2206.01076*, 2023b.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

J-J Daudin, Franck Picard, and Stéphane Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Sophie Donnet and Stéphane Robin. Accelerating bayesian estimation for network poisson models using frequentist variational estimates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(4):858–885, 2021.

James M Flegal, John Hughes, Dootika Vats, and Ning Dai. mcmcse: Monte carlo standard errors for mcmc. *R package version*, 1(2), 2017.

Sylvia Frühwirth-Schnatter and Gertraud Malsiner-Walli. From here to infinity: sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in data analysis and classification*, 13(1):33–64, 2019.

Sylvia Frühwirth-Schnatter, Gertraud Malsiner-Walli, and Bettina Grün. Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307, 2021.

Junxian Geng, Anirban Bhattacharya, and Debdeep Pati. Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, 114(526):893–905, 2019.

Bruce Hajek and Suryanarayana Sankagiri. Community recovery in a preferential attachment graph. *IEEE Transactions on Information Theory*, 65(11):6853–6874, 2019.

Jack Hessel, Chenhao Tan, and Lillian Lee. Science, askscience, and badscience: On the coexistence of highly related communities. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 171–180, 2016.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003.

Bo Jiang, Zhi-Li Zhang, and Don Towsley. Reciprocity in social networks with capacity constraints. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–466, 2015.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Pierre Latouche, Etienne Birmele, and Christophe Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012.

Paul Liu, Austin R Benson, and Moses Charikar. Sampling methods for counting temporal motifs. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 294–302, 2019.

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Model-based clustering based on sparse finite gaussian mixtures. *Statistics and computing*, 26(1):303–324, 2016.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, 2007.

Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

Mark EJ Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

Vivekananda Roy. Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 7:387–412, 2020.

Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42, 2009.

Phyllis Wan, Tiandong Wang, Richard A Davis, and Sidney I Resnick. Fitting the linear preferential attachment model. *Electronic Journal of Statistics*, 11(2):3738–3780, 2017.

Tiandong Wang and Sidney Resnick. Measuring reciprocity in a directed preferential attachment network. *Advances in Applied Probability*, pages 1–25, 2022a. doi: https://doi.org/10.1017/apr.2021.52.

Tiandong Wang and Sidney Resnick. Poisson edge growth and preferential attachment networks. *Methodology and Computing in Applied Probability*, 25(1):8, 2023a.

Tiandong Wang and Sidney Resnick. Random networks with heterogeneous reciprocity. *Extremes*, pages 1–39, 2023b.

Tiandong Wang and Sidney I Resnick. Asymptotic dependence of in-and out-degrees in a preferential attachment model with reciprocity. *Extremes*, pages 1–34, 2022b.

Tiandong Wang and Panpan Zhang. Directed hybrid random networks mixing preferential attachment with uniform attachment mechanisms. *Annals of the Institute of Statistical Mathematics*, pages 1–30, 2022.

Mingzhang Yin, YX Rachel Wang, and Purnamrita Sarkar. A theoretical case study of structured variational inference for community detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3750–3761. PMLR, 2020.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326, 2011.

Vinko Zlatić and Hrvoje Štefančić. Model of wikipedia growth based on information exchange via reciprocal arcs. *EPL (Europhysics Letters)*, 93(5):58005, 2011.