Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction

Mingyuan Zhou
McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA

Abstract

A hierarchical gamma process infinite edge partition model is proposed to factorize the binary adjacency matrix of an unweighted undirected relational network under a Bernoulli-Poisson link. The model describes both homophily and stochastic equivalence, and is scalable to big sparse networks by focusing its computation on pairs of linked nodes. It can not only discover overlapping communities and inter-community interactions, but also predict missing edges. A simplified version omitting inter-community interactions is also provided and we reveal its interesting connections to existing models. The number of communities is automatically inferred in a nonparametric Bayesian manner, and efficient inference via Gibbs sampling is derived using novel data augmentation techniques. Experimental results on four real networks demonstrate the models' scalability and state-of-the-art performance.

1 INTRODUCTION

Community detection and link prediction are two important problems in network analysis. A vast number of community detection algorithms based on various useful heuristics, such as modularity maximization (Newman and Girvan, 2004) and clique percolation (Palla et al., 2005), have been proposed. See Fortunato (2010) for a comprehensive review. These algorithms, however, are not based on generative models and hence usually cannot be used to generate networks and predict missing edges (links). Moreover, how to set the number of communities is a critical issue that has not been well addressed by them. In this paper, we will fit unweighted undirected relational networks using nonparametric Bayesian generative models, which can be used to simulate random networks, detect latent overlapping communities and community-community interactions, and predict missing edges, with the number of communities automatically inferred from the data.

For a relational network, a community can be considered as a subset of nodes (vertices) that are densely connected to each other but sparsely to the others, such as those in a social network, or it can be considered as a subset of nodes that are sparsely connected to each other but densely connected to the nodes belonging to another community, such as those in a network consisting of carnivores and herbivores: tigers and bears both hunt deers but rarely prey on each other. The former phenomenon is usually described as assortativity or homophily, while the latter is known as dissortativity or stochastic equivalence (Hoff, 2008). As a relational network may exhibit both homophily and stochastic equivalence, an algorithm capable of modeling both phenomena would usually be preferred if no prior information on assortativity is available. If analyzing assortative networks with dense intra-community connections is the main goal, then one may consider an assortative algorithm that models homophily but not necessarily stochastic equivalence.

The stochastic blockmodel (SBM) is a popular latent class model to detect latent communities (Holland et al., 1983; Nowicki and Snijders, 2001). It partitions the nodes into disjoint communities, and models the probability for an edge to exist between two nodes solely based on which two communities that they belong to. It is simple and scalable, and models both homophily and stochastic equivalence. In addition, the infinite relational model, a nonparametric Bayesian extension of the SBM based on the Chinese restaurant process (Aldous, 1985), allows the number of communities to be automatically inferred from the data (Kemp et al., 2006). Despite these attractive properties, the SBM is restrictive in that communities are not allowed to overlap. In practice, however, it is

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

common for a node to belong to multiple communities, motivating the development of more advanced latent class models, such as the mixed-membership stochastic blockmodel (MMSB) of Airoldi et al. (2008) and its various extensions (Gopalan et al., 2012; Kim et al., 2013). The MMSB generalizes the SBM to allow a node to participate in multiple communities, yet since it has to infer two community indicators for each pair of nodes, regardless of whether an edge exists in that pair, its computation grows quadratically as a function of the number of nodes N. Moreover, the number of communities in the MMSB is a model parameter that needs to be carefully selected.

In this paper, instead of clustering nodes, as in the SBM, or clustering all possible edges, as in the MMSB, we propose the edge partition model (EPM) to partition only the observed edges, which readily leads to the partition of nodes: if the edges linked to a node are partitioned into multiple communities, then the node is naturally affiliated with all these communities, and could be hard assigned to a single community that has the strongest presence in its edges. In contrast to the SBM, the EPM allows communities to overlap; and in contrast to the MMSB that spends $O(N^2)$ computation clustering all possible edges, the EPM spends $O(\bar{d}N)$ computation partitioning only observed edges, where $\bar{d}$ is the average degree (number of edges) per node, leading to notable computational savings as $\bar{d}$ is often much smaller than N in a big sparse network commonly observed in practice.

To support a potentially infinite number of communities and to model both homophily and stochastic equivalence in an unweighted undirected relational network, we propose a hierarchical gamma process (HGP) EPM, which links each observed edge to a latent count using a Bernoulli-Poisson link, and then factorizes the latent $N \times N$ random count matrix. The HGP supports the EPM to have an infinite dimensional feature vector for each node to describe its affiliations with communities, and an infinite dimensional square rate matrix, whose diagonal and off-diagonal elements describe the intra- and inter- community interactions, respectively. We also propose a gamma process EPM as a simplified version of the HGP-EPM, which omits inter-community interactions to gain simpler inference and faster computation at the expense of reduced ability to model stochastic equivalence.

Conceptually, our idea of directly partitioning edges and implicitly partitioning nodes into communities is related to the one in Ahn et al. (2010) and Evans and Lambiotte (2009). In terms of construction, our EPMs are related to the Poisson factor models of Ball et al. (2011) and the Eigenmodel of Hoff (2008). In terms of supporting an infinite number of features, our EPMs are related to the models in Miller et al. (2009) and Morup et al. (2011) that use the Indian buffet process of Griffiths and Ghahramani (2005) to support an infinite binary feature matrix. The proposed models depart from existing ones with several distinctions: 1) a Bernoulli-Poisson link connects each edge to a latent count that is further partitioned; 2) a hierarchical gamma process is constructed to support an infinite number of communities and an infinite-dimensional square matrix to describe community-community interactions; 3) two nonparametric Bayesian EPMs are constructed to factorize the $N \times N$ binary adjacency matrix under the Bernoulli-Poisson link, supporting a nonnegative feature matrix with an unbounded number of columns, and at the same time assign each edge and hence each node to one or multiple latent communities; and 4) efficient and scalable Bayesian inference via Gibbs sampling is provided.

2 FACTOR ANALYSIS AND BERNOULLI-POISSON LINK

Our basic idea is to factorize the BINARY network adjacency matrix using tools developed for COUNT data analysis, and to discover overlapping communities and their interactions by examining how the latent count for each edge is partitioned. This section will primarily discuss individual model components and their properties, with hierarchical Bayesian models presented later.

2.1 Poisson Factor Analysis

We propose a Poisson factor model for a weighted undirected $N \times N$ relational network as

$$m_{ij} \sim \text{Po}\left(\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\right), \quad (1)$$

where $m_{ij} \equiv m_{ji}$ is the integer-valued weight (observed or latent) that links nodes i and j, $(\phi_{i1}, \ldots, \phi_{iK})$ is the positive feature vector for node i, $\lambda_{k_1k_2} \equiv \lambda_{k_2k_1}$ is a positive rate, and the symbol $\equiv$ denotes "equal by definition." This model is conceptually simple: with ϕ_{ik_1} measuring how strongly node i is affiliated with community k_1 and $\lambda_{k_1k_2}$ measuring how strongly communities k_1 and k_2 interact with each other, the product $\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}$ measures how strongly nodes i and j are connected due to their affiliations with communities k_1 and k_2, respectively, and a weighted combination of all intra-community weights $\{\lambda_{kk}\}_{1\le k\le K}$ and inter-community ones $\{\lambda_{k_1k_2}\}_{1\le k_1\ne k_2\le K}$ is the expected value of m_{ij}.

The factor model in (1) makes intuitive sense. For example, suppose persons i and j are both residents of City Avatar and active members of the Avatar anglers Meetup group that organizes fishing trips regularly. In addition, persons i is an active member of the Avatar

artificial intelligence (AI) Meetup group while person j is an active member of the Avatar statistics Meetup group. Denoting m_{ij} as the number of times that i and j attend the same group meeting in 2015, then due to their strong affiliations with the anglers Meetup group, m_{ij} would have a large expected value, which is likely to be further increased if the AI and statistics Meetup groups hold joint events regularly.

To model an assortative relational network exhibiting homophily but not necessarily stochastic equivalence, we may omit the inter-community interactions by letting $\lambda_{k_1k_2} \equiv 0$ for $k_1 \neq k_2$ and simplify (1) as

$$m_{ij} \sim \text{Po}\left(\sum_{k=1}^{K} r_k \phi_{ik} \phi_{jk}\right), \tag{2}$$

where $r_k \equiv \lambda_{kk}$ indicates the prevalence of community k, and two nodes with similar latent features are encouraged to be linked by an edge with a large weight.

We note Ball et al. (2011) had examined a model related to (2) and briefly mentioned a model related to (1). However, they used a heuristic approach to model binary data under the Poisson distribution, did not provide a principal way to set the number of communities K, and had to create possibly nonexistent self-edges in order to derive tractable expectation-maximization (EM) inference. This paper will address all these issues rigorously, in a nonparametric Bayesian manner, and carefully examine the models in both (2) and (1) and provide efficient Bayesian inference.

2.2 Bernoulli-Poisson Link

To use the Poisson factor models in (1) and (2) for an unweighted network with a binary adjacency matrix, we introduce a Bernoulli-Poisson (BerPo) link function that thresholds a random count at one to obtain a random variable in $\{0, 1\}$ as

$$b = \mathbf{1}(m \geq 1),\ m \sim \text{Po}(\lambda), \tag{3}$$

where $b = 1$ if $m \geq 1$ and $b = 0$ if $m = 0$. The intuition is that two nodes are connected if they interact at least once. The mathematical motivation is after transforming a binary-modeling problem into a count-modeling one, one is readily equipped with a rich set of statistical tools developed for count data analysis using the Poisson and negative binomial distributions.

If m is marginalized out from (3), then given λ, one obtains a Bernoulli random variable as

$$b \sim \text{Ber}\left(1 - e^{-\lambda}\right).$$

The conditional posterior of m can be expressed as

$$(m|b, \lambda) \sim b \cdot \text{Po}_+(\lambda),$$

where $x \sim \text{Po}_+(\lambda)$ follows a truncated Poisson distribution, with $P(x = k) = (1 - e^{-\lambda})^{-1}\lambda^k e^{-\lambda}/k!$ for $k \in \{1, 2, \ldots\}$. Thus if $b = 0$, then $m = 0$ almost surely (a.s.), and if $b = 1$, then $m \sim \text{Po}_+(\lambda)$, which can be simulated with rejection sampling: if $\lambda \geq 1$, we draw $m \sim \text{Po}(\lambda)$ till $m \geq 1$; and if $\lambda < 1$, we draw both $n \sim \text{Po}(\lambda)$ and $u \sim \text{Unif}(0, 1)$ till $u < 1/(n + 1)$, and then let $m = n + 1$. The acceptance rate is $1 - e^{-\lambda}$ if $\lambda \geq 1$ and $\lambda^{-1}(1 - e^{-\lambda})$ if $\lambda < 1$, and reaches its minimum, 63.2%, when $\lambda = 1$.

The BerPo link shares some similarities with the probit link that thresholds a normal random variable at zero, and the logit link that lets $b \sim \text{Ber}[e^x/(1 + e^x)]$. We advocate the BerPo link as an alternative to the probit and logit links since if $b = 0$, then $m = 0$ a.s., which could lead to significant computational savings if a considerable proportion of the data are equal to zero. In addition, the additive property of the Poisson allows us to model the link strength between any two nodes by aggregating the contributions of all possible intra- and inter- community interactions, and the conjugacy between the Poisson and gamma distributions makes it convenient to construct hierarchical Bayesian models amenable to posterior simulation.

2.3 Overlapping Community Structures

Note that (1) can be augmented as

$$m_{ij} = \sum_{k_1} \sum_{k_2} m_{ik_1k_2j},\ m_{ik_1k_2j} \sim \text{Po}\left(\phi_{ik_1} \lambda_{k_1k_2} \phi_{jk_2}\right). \tag{4}$$

where $m_{ik_1k_2j}$ represents how often nodes i and j interact due to their affiliations with communities k_1 and k_2, respectively. We may consider that the model is partitioning the count m_{ij} into $\{m_{ik_1k_2j}\}_{1 \leq k_1, k_2 \leq K}$, and hence we call the Poisson factor model in (1) together with the BerPo link in (3) as an edge partition model (EPM), in which each edge is partitioned according to all possible K^2 community-community interactions, and how strongly node i is affiliated with community k can be measured with $\phi_{ik}\omega_{ik}$, where

$$\omega_{ik} := \textstyle\sum_{j \neq i} \sum_{k'} \phi_{jk'} \lambda_{kk'} \tag{5}$$

represents how strongly node i would interact with all the other nodes through its affiliation with community k. We further introduce the latent count

$$m_{ik\cdot\cdot} := \textstyle\sum_{j>i} \sum_{k_2} m_{ikk_2j} + \sum_{j<i} \sum_{k_1} m_{jk_1ki}, \tag{6}$$

to represent how often node i is connected to the other nodes due to its affiliation with community k. We can then assign node i to multiple communities in $\{k : m_{ik\cdot\cdot} \geq 1\}$, or (hard) assign it to a single community using either $\underset{k}{\text{argmax}}(\phi_{ik}\omega_{ik})$ or $\underset{k}{\text{argmax}}(m_{ik\cdot\cdot})$. Similar analysis applies to a simpler EPM built on (2).

By hard assigning each node to a single community and ordering the nodes from the same community to be adjacent to each other, we expect the ordered adjacency matrix to exhibit a block structure, where the blocks along and off the diagonal represent the intra- and inter- community connections, respectively.

2.4 Scalability for Big Sparse Networks

We are motivated to construct the EPMs because they not only allow each edge and hence each node to participate in multiple communities, but also readily scale to a big sparse network whose average degree per node is much smaller than N. A key observation for scalable computation is that (1) can be augmented as (4), where $m_{ik_1k_2j} = 0$ a.s. for any k_1 and k_2 if no edge exists between nodes i and j (i.e., $m_{ij} = 0$). On a sparse network, where the edges constitute only a small portion of all possible N^2 edges, this property makes the EPMs computationally appealing. By contrast, conceptually related models, including the MMSB of Airoldi et al. (2008), Eigenmodel of Hoff (2008) and latent feature relational model of Miller et al. (2009), spend computation indiscriminately on all pairs of nodes (i, j) no matter whether an edge exists between nodes i and j, and hence they have $O(N^2)$ computation and do not scale well as N increases.

3 EDGE PARTITION MODELS

3.1 Hierarchical Gamma Process

The EPM takes a weighted combination of all possible intra- and inter- community weights to explain each pair of node, however, the number of communities K is still a model parameter that needs to be set appropriately. To allow K to be inferred from the data and potentially grow to infinity, we need to introduce a stochastic process that can generate a countably infinite number of atoms $\{\boldsymbol{\phi}_k\}_{1,\infty}$, where $\boldsymbol{\phi}_k = (\phi_{1k}, \ldots, \phi_{NK})^T$ measures how strongly the N nodes are affiliated with community k, and an infinite dimensional square matrix $\{\lambda_{k_1k_2}\}_{1\le k_1,k_2\le\infty}$, where $\lambda_{k_2k_1} = \lambda_{k_1k_2}$ measures how strongly communities k_1 and k_2 interact with each other. Moreover, we need to ensure $\sum_{k_1=1}^{\infty}\sum_{k_2=1}^{\infty}\lambda_{k_1k_2}$ to be finite a.s. and we may wish to impose some structural regularization on the infinite square matrix.

To satisfy all these needs, we first define

$$G \sim \Gamma\mathrm{P}(G_0, 1/c_0) \tag{7}$$

as a gamma process on a product space $\mathbb{R}^+ \times \Omega$, where $\mathbb{R}^+ = \{x : x > 0\}$, Ω is a complete separable metric space, $1/c_0$ is a positive scale parameter, and G_0 is a finite and continuous base measure, such that $G(A) \sim \mathrm{Gam}(G_0(A), 1/c_0)$ for each Borel set $A \subset \Omega$ (Ferguson, 1973; Kingman, 1993). The Lévy measure of the gamma process can be expressed as $\nu(drd\boldsymbol{\phi}) = r^{-1}e^{-c_0r}drG_0(d\boldsymbol{\phi})$, and a draw from the gamma process, consisting of countably infinite atoms, can be expressed as $G = \sum_{k=1}^{\infty} r_k\delta_{\boldsymbol{\phi}_k}$, where $\boldsymbol{\phi}_k \overset{iid}{\sim} g_0$, $G_0 = \gamma_0 g_0$, $g_0(d\boldsymbol{\phi}) = G_0(d\boldsymbol{\phi})/\gamma_0$ is the base distribution, and $\gamma_0 = G_0(\Omega)$ is the mass parameter. A gamma process based model has an inherent shrinkage mechanism, as in the prior the number of atoms with r_k greater than $\varepsilon \in \mathbb{R}^+$ follows $\mathrm{Po}(\gamma_0\int_{\varepsilon}^{\infty} r^{-1}e^{-cr}dr)$, whose Poisson rate decreases as ε increases.

Given $G = \sum_{k=1}^{\infty} r_k\delta_{\boldsymbol{\phi}_k}$, we further define a relational gamma process (rΓP) as

$$\Lambda|G \sim \mathrm{r}\Gamma\mathrm{P}(G, \xi, 1/\beta), \tag{8}$$

a draw from which is defined as

$$\Lambda = \textstyle\sum_{k_1=1}^{\infty}\sum_{k_2=1}^{\infty}\lambda_{k_1k_2}\delta_{(\boldsymbol{\phi}_{k_1}, \boldsymbol{\phi}_{k_2})},$$

where ξ and β are both in $\mathbb{R}^+$, $\lambda_{k_2k_1} \equiv \lambda_{k_1k_2}$, and

$$\lambda_{k_1k_2} \sim \begin{cases} \mathrm{Gam}(\xi r_{k_1}, 1/\beta)\ , & \text{if } k_2 = k_1, \\ \mathrm{Gam}(r_{k_1}r_{k_2}, 1/\beta), & \text{if } k_2 > k_1. \end{cases}$$

Given a relational gamma process draw Λ, we generate a binary adjacency matrix $\mathbf{B} \in \{0, 1\}^{N\times N}$ as

$$\mathbf{B}|\Lambda \sim \mathrm{Ber}\left[1 - \prod_{k_1=1}^{\infty}\prod_{k_2=1}^{\infty}\exp\left(-\boldsymbol{\phi}_{k_1}\lambda_{k_1k_2}\boldsymbol{\phi}_{k_2}^T\right)\right]. \tag{9}$$

Equations (9), (8) and (7) constitute an HGP-EPM that supports countably infinite atoms and a countably infinite square matrix, the total sum of whose elements has a finite expectation, as shown in the following Lemma, with proof provided in the Appendix.

Lemma 1. *The expectation of* $\sum_{k_1=1}^{\infty}\sum_{k_2=1}^{\infty}\lambda_{k_1k_2}$ *is finite and can be expressed as*

$$\mathbb{E}\left[\sum_{k_1}\sum_{k_2}\lambda_{k_1k_2}\right] = \frac{\xi}{c_0\beta}\gamma_0 + \frac{1}{c_0^2\beta}\gamma_0^2.$$

The usual scenario to consider an HGP construction is when one models grouped data and wishes to share statistical strengths across groups. For example, the gamma-negative binomial process of Zhou and Carin (2012), related to the hierarchical Dirichlet process of Teh et al. (2006), is considered for topic modeling, where each document is associated with a gamma process, and these gamma processes are coupled by sharing a lower-level (*i.e.*, further from the data) gamma process as their atomic base measure. The proposed HGP is distinct in that the product of the weights of any two atoms of the lower-level gamma process is used to parameterize the shape parameter of a gamma random variable higher in the hierarchy.

The proposed HGP also helps express our prior belief that an atom with a small weight tends to represent a small community, which also tends to interact with the others less frequently. Note that if we let $\lambda_{kk} \sim \text{Gam}(r_k^2, 1/\beta)$, then the expectation of the matrix $\{\lambda_{k_1k_2}\}_{1\le k_1,k_2\le\infty}$ given $\{r_k\}_{1,\infty}$ has a rank of one. We use ξr_k instead of r_k^2 as the shape parameter of λ_{kk} to allow r_k to be inferred with Gibbs sampling and to prevent overly shrinking λ_{kk} for small communities. Note that Palla et al. (2014) proposed a reversible infinite hidden Markov model using a related HGP infinite square rate matrix, the normalization of whose each row represents a state transition probability vector. Our HGP serves a distinct modeling purpose; no normalization is required for the infinite square rate matrix, and our model allows exploiting unique data augmentation techniques to infer both $\lambda_{k_1k_2}$ and r_k with closed-form Gibbs sampling update equations, as discussed in Section 3.4 and the Appendix.

3.2 Hierarchical Gamma Process EPM

We choose the base distribution of the gamma process $G \sim \Gamma\text{P}(G_0, 1/c_0)$ as $g_0(\boldsymbol{\phi}) = \prod_{i=1}^N \text{Gam}(a_i, 1/c_i)$. For implementation convenience, we consider a discrete base measure as $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K}\delta_{\boldsymbol{\phi}_k}$, where K is a truncation level that is set large enough to ensure a good approximation to the truly infinite model. We express the (truncated) HGP-EPM as

$$
\begin{aligned}
&b_{ij} = \mathbf{1}(m_{ij} \ge 1),\ m_{ij} = \sum_{k_1=1}^K \sum_{k_2=1}^K m_{ik_1k_2j},\\
&m_{ik_1k_2j} \sim \text{Po}\left(\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\right),\\
&\phi_{ik} \sim \text{Gam}(a_i, 1/c_i),\ a_i \sim \text{Gam}(e_0, 1/f_0),\\
&\lambda_{k_1k_2} \sim \begin{cases} \text{Gam}(\xi r_{k_1}, 1/\beta), & \text{if } k_2 = k_1,\\ \text{Gam}(r_{k_1}r_{k_2}, 1/\beta), & \text{if } k_2 > k_1,\end{cases}\\
&r_k \sim \text{Gam}(\gamma_0/K, 1/c_0),
\end{aligned} \tag{10}
$$

where $\lambda_{k_2k_1} = \lambda_{k_1k_2}$ a.s., and conjugate gamma priors are imposed on γ_0, ξ, c_0, c_i and β. Note that marginalizing out both m_{ij} and $m_{ik_1k_2j}$ from (10) leads to

$$
b_{ij} \sim \text{Ber}\left[1 - \prod_{k_1=1}^K \prod_{k_2=1}^K \exp(-\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2})\right]. \tag{11}
$$

A noticeable advantage of the augmented representation in (10) over (11) is that (10) is amenable to posterior simulation, as discussed in Section 3.4.

Note that similar to Hoff (2008) and Lloyd et al. (2012), we assume that the nodes are exchangeable and hence the discussions of Hoover (1982) and Aldous (1985) on exchangeability also apply to our EPMs.

3.3 Gamma Process EPM

If we omit inter-community interactions by letting $\lambda_{k_1k_2} \equiv 0$ for $k_1 \ne k_2$ and $\lambda_{kk} \equiv r_k$, then the HGP-EPM reduces to a gamma process EPM (GP-EPM), which is likely to well fit assortative networks but not necessarily disassortative ones. We notice an interesting connection to the community-affiliation graph model (AGM) of Yang and Leskovec (2012, 2014): the GP-EPM generates an edge with probability

$$
P(b_{ij}=1) = 1 - \textstyle\prod_k \{1 - [1 - \exp(-r_k\phi_{ik}\phi_{jk})]\}; \tag{12}
$$

if we define $p_k = 1 - e^{-r_k}$ and further impose the restriction that $\phi_{ik} \in \{0, 1\}$, then (12) reduces to

$$
P(b_{ij} = 1) = 1 - \textstyle\prod_{k\in C_{ij}} (1 - p_k), \tag{13}
$$

where $C_{ij} = \{k : \phi_{ik} = 1 \text{ and } \phi_{jk} = 1\} \subset \{1, \ldots, K\}$ is a set of communities that nodes i and j share; note that (13) is almost the same as the AGM of Yang and Leskovec (2012, 2014). In fact, one may consider the GP-EPM with $b_{ij} \sim \text{Ber}[1 - e^{-\epsilon}\prod_k \exp(-r_k\phi_{ik}\phi_{jk})]$, where $\epsilon \in \mathbb{R}^+$ and $\phi_{ik} \in \{0, 1\}$, as a nonparametric Bayesian AGM. Similarly, we also notice that (11) of the HGP-EPM is related to the model of Morup et al. (2011) if we restrict $\phi_{ik} \in \{0, 1\}$.

Yang and Leskovec (2012, 2014) argue that all previous community detection methods, including clique percolation and MMSB, would fail to detect communities with dense overlaps, because they all had a hidden assumption that a community's overlapping parts are less densely connected than its non-overlapping ones. The same as the AGM, both the GP-EPM and HGP-EPM do not make such a restrictive assumption, and they both allow overlaps of communities to be denser than communities themselves; Beyond the AGM, we do not restrict ϕ_{ik} to be either zero or one, and our generative models are built under a rigorous nonparametric Bayesian framework with efficient Bayesian inference, as presented below.

3.4 MCMC Inference

In this paper, we consider an unweighted undirected network, where $b_{ji} \equiv b_{ij}$ and self-links b_{ii} are not defined. Thus we only consider b_{ij} for $j > i$ in (10). Let $m_{ik\cdot\cdot}$ be defined as in (6) and $m_{\cdot k_1k_2\cdot}$ as

$$
m_{\cdot k_1k_2\cdot} := 2^{-\delta_{k_1k_2}} \textstyle\sum_i \sum_{j>i} (m_{ik_1k_2j} + m_{ik_2k_1j}),
$$

where $\delta_{k_1k_2} = 1$ if $k_1 = k_2$ and $\delta_{k_1k_2} = 0$ otherwise. Using (5) and the Poisson additive property, we have

$$
m_{ik\cdot\cdot} \sim \text{Po}(\phi_{ik}\omega_{ik}), \tag{14}
$$

$$
m_{\cdot k_1k_2\cdot} \sim \text{Po}\left(\lambda_{k_1k_2}\theta_{k_1k_2}\right), \tag{15}
$$

where $\theta_{k_1k_2} := 2^{-\delta_{k_1k_2}} \sum_i \sum_{j\ne i} \phi_{ik_1}\phi_{jk_2}$ represents how strongly the nodes interact through communities k_1 and k_2. Marginalizing out ϕ_{ik} from (14)

and $\lambda_{k_1k_2}$ from (15), with $p'_{ik} := \omega_{ik}/(c_i + \omega_{ik})$ and $\tilde{p}_{k_1k_2} := \theta_{k_1k_2}/(\beta + \theta_{k_1k_2})$, we have

$$m_{ik\cdot\cdot} \sim \mathrm{NB}\left(a_i, p'_{ik}\right), \tag{16}$$

$$m_{\cdot k_1k_2\cdot} \sim \mathrm{NB}\left[r_{k_1}\xi^{\delta_{k_1k_2}}(r_{k_2})^{1-\delta_{k_1k_2}},\ \tilde{p}_{k_1k_2}\right]. \tag{17}$$

Using the BerPo link, the gamma-Poisson conjugacy, and the augment-and-conquer techniques to infer the negative binomial dispersion parameters (Zhou and Carin, 2012, 2014), we exploit (14)-(17) to derive closed-form Gibbs sampling update equations for all model parameters except γ_0, and construct an excellent proposal distribution to sample γ_0 using an independence chain Metropolis-Hastings algorithm. We present in the Appendix the details of MCMC inference for the HGP-EPM, and the hierarchical model and closed-form Gibbs sampling update equations for the GP-EPM. The inference of the nonparametric Bayesian AGM would be almost the same as that of the GP-EPM, with the only difference that the $(\phi_{ik}|-)$ would be sampled from Bernoulli distributions.

4 EXPERIMENTAL RESULTS

For comparison, we consider the infinite relational model (IRM) of Kemp et al. (2006), the Eigenmodel of Hoff (2008), the infinite latent attribute (ILA) model of Palla et al. (2012), the AGM of Yang and Leskovec (2012, 2014), and our GP- and HGP-EPMs. We use the R package provided for the Eigenmodel. We use the ILA code[1] provided for Palla et al. (2012), in which it is shown that the ILA outperforms the related nonparametric latent feature relational model of Miller et al. (2009). We implement a nonparametric Bayesian version of the AGM as a special case of the GP-EPM, as discussed in Section 3.3. Matlab code for the EPMs is available on the author's website.

For the Eigenmodel, we find the best K in $\{5, 10, 25, 50\}$. For the ILA, we use its default parameter setting[2]. For the IRM, we choose $\mathrm{Beta}(0.1, 1)$ as the prior for each latent block and $\mathrm{Gam}(0.01, 1/0.01)$ as the prior for the Chinese restaurant process concentration parameter; for the nonparametric Bayesian AGM, we let $\phi_{ik} \sim \mathrm{Ber}(\pi_i)$, $\pi_i \sim \mathrm{Beta}(0.01, 0.01)$ and $\epsilon \sim \mathrm{Gam}(0.01, 1/0.01)$; these parameters are found to consistently provide good performance. For our models' hyper-parameters, we choose $e_0 = f_0 = 0.01$ and let γ_0, c_i, c_0 and β be all drawn from $\mathrm{Gam}(1, 1)$.

We consider 3000 MCMC iterations and collect the last 1500 samples, unless otherwise stated. We consider two small-scale benchmark networks, for which we test all algorithms and set the truncation level as $K_{\max} = 100$ for our algorithms, and another two networks with more than 2000 nodes, for which we set $K_{\max} = 256$.

To test a model's ability to predict missing edges of an unweighted undirected relational network, we randomly[3] hold out 20% pairs of nodes and use the the remaining 80% to predict the probability for an edge to exist in each of these held-out pairs. Letting $o_{ij} = 0$ if b_{ij} is held out and $o_{ij} = 1$ otherwise, we only need to slightly modify the inference by only considering $\{b_{ij} : o_{ij} = 1\}$ in the likelihood. For example, ω_{ik} in (5) would be redefined as $\omega_{ik} = \sum_{j:o_{ij}=1}\sum_{k'} \lambda_{kk'}\phi_{jk'}$. We consider exactly the same five random training-testing partitions for all algorithms and report the average area under the curve (AUC) of both the receiver operating characteristic (ROC) and precision-recall (PR) curves (Davis and Goadrich, 2006). For link prediction, the AUC-PR is more sensitive to the percentage of true edges among the top ranked ones. Note that in addition to link prediction, the HGP-EPM, GP-EPM, AGM and IRM all have easily interpretable latent representations that will be used to detect overlapping/disjoint communities.

4.1 Protein230 Network

We first consider the Protein230 dataset of Butland et al. (2005) that describes the interactions between 230 proteins, with 595 edges. This is a small-scale benchmark network that exhibits both homophily and stochastic equivalence, as shown in Hoff (2008) and also tested in Lloyd et al. (2012). We are able to run 3000 MCMC iterations quickly enough for all algorithms except for the ILA on this network.

As shown in Tab. 1, the HGP-EPM has the best overall performance. The Eigenmodel is the second best with $K = 10$ and the IRM is the third best. The AGM is not competitive as it restricts its features to be binary. In this and all future tables, we highlight in bold both the best result and the ones that are less than one standard error away from the best. Below we analyze why the HGP-EPM performs the best while the simpler GP-EPM is not that competitive on this dataset.

As shown in Figs. 1 (b)-(d), the HGP-EPM captures both homophily and stochastic equivalence by accurately modeling both diagonal and off-diagonal dense regions of the adjacency matrix; the GP-EPM captures homophily by accurately modeling diagonal dense regions that represent intra-community interactions, but at the expense of creating nonexistent blocks in order to fit dense off-diagonal regions that represent strong

[1] http://mlg.eng.cam.ac.uk/konstantina/ILA/ILA_code(v1).tar.gz

[2] The default training/testing partition of the ILA code sends self-edges into the testing set; whereas in this paper, we do not intend to predict self-edges and hence we do not allow them to appear in the testing set.

[3] If removing an edge disconnects a node to all the others, then the edge will be kept in the training set.

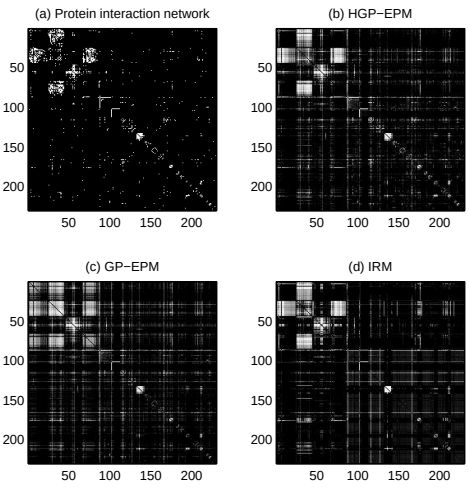

Figure 1: Comparison of three models on estimating the link probabilities for the Protein230 network using 80% of its node pairs. The nodes are reordered to make a node with a larger index belong to the same or a smaller-size community, where the disjoint community assignments are obtained by analyzing the results of the HGP-EPM. (a) The binary adjacency matrix. (b)-(c) Estimated link probabilities displayed on the log-10 scale from -2 to 1, with a brighter color representing a higher link probability.

Table 1: Comparison of six algorithms on predicting missing edges of the Protein230 network. The Eigenmodel achieves its best performance at $K = 10$.

Model	AUC-ROC	AUC-PR
IRM	0.9338 ± 0.0128	0.5026 ± 0.0676
Eigenmodel	0.9314 ± 0.0188	**0.5468** ± 0.0500
ILA	0.8971 ± 0.0297	0.3693 ± 0.0234
AGM	0.9145 ± 0.0160	0.3339 ± 0.0359
GP-EPM	0.9335 ± 0.0110	0.4011 ± 0.0452
HGP-EPM	**0.9519** ± 0.0100	**0.5655** ± 0.0505

inter-community interactions; and the IRM captures these large dense blocks, but produces a cartoonish estimation, which overlooks small communities that represent fine details along the diagonal.

Fig. 2 shows how the HGP-EPM works. First, each feature vector $\boldsymbol{\phi}_k$ shown in Fig. 2 (a) clearly describes how strongly the nodes are affiliated with the community it represents, and each node may have large weights on multiple community. Second, about 30 latent feature vectors are inferred and the remaining ones are essentially drawn from the prior $\prod_i \mathrm{Gam}(a_i, 1/c_i)$. Third, the inter- and intra-community interaction strengths in Fig. 2 (b) can be matched to the corresponding communities (subsets of nodes) in Figs. 1 (a) and (b). For example, Fig. 2 (a) suggests that the first and second largest communities have 24 and 22 nodes, respectively, and Fig. 2 (b) suggests that the first and second communities have sparse and dense intra-community connections, respectively, and have denser connections between them, as confirmed by examining the block structures within the top-left 46×46 corner of both Figs. 1 (a) and (b).

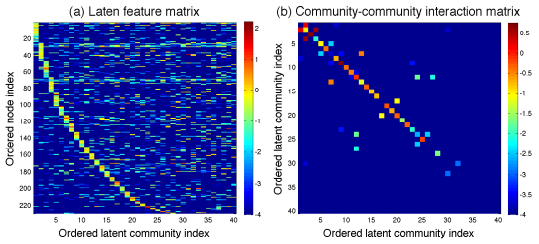

Figure 2: The inferred latent feature matrix $\{\boldsymbol{\phi}_k\}$ and community-community interaction rate matrix $\{\lambda_{k_1 k_2}\}$ for the HGP-EPM on Protein230. The nodes are reordered to make a node with a larger index belong to the same or a smaller-size community, and the communities are ordered to make a community with a larger index to have a smaller size. The pixel values are displayed on the log-10 scale.

4.2 NIPS234 Coauthor Network

We consider the small-scale NIPS234 network consists of the top 234 authors in NIPS 1-17 conferences[4] in terms of the number of publications, as studied in Miller et al. (2009). There are 598 edges. As shown in Tab. 2, the GP-EPM and HGP-EPM have the best overall performance, followed by the IRM. Comparing with the simpler GP-EPM, the extra flexibility to model stochastic equivalence does not bring the HGP-EPM additional advantages on this dataset, which is not surprising as Fig. 3 suggests that this coauthor network mainly exhibits homophily. Note that the IRM performs well measured by the AUC-ROC, but its AUC-PR is clearly worse than those of the EPMs. This may again be explained by its overly smoothed cartoonish estimation that overlooks small communities, as clearly shown in Fig. 3 (d).

Table 2: Comparison of six algorithms on predicting missing edges of the NIPS234 coauthor network. The Eigenmodel achieves its best performance at $K = 10$.

Model	AUC-ROC	AUC-PR
IRM	**0.9476** ± 0.0114	0.6677 ± 0.0201
Eigenmodel	0.9269 ± 0.0177	0.6784 ± 0.0364
ILA	0.9171 ± 0.0222	0.6793 ± 0.0295
AGM	0.8906 ± 0.0164	0.5842 ± 0.0357
GP-EPM	**0.9501** ± 0.0123	**0.7415** ± 0.0319
HGP-EPM	**0.9469** ± 0.0163	**0.7289** ± 0.0540

4.3 Yeast and NIPS12 Networks

We also consider the Yeast[5] protein interaction network of Bu et al. (2003), with 2361 nodes and 6646 non-self edges, and the NIPS12 coauthor network[6] that includes all the 2037 authors in NIPS papers vols 0-12, with 3134 edges. These two median-size networks are already too large for the Eigenmodel and ILA to produce reasonable results given our computational resources. The results in Tabs. 3 and 4 show

[4] http://chechiklab.biu.ac.il/∼gal/data.html
[5] http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm
[6] http://www.cs.nyu.edu/∼roweis/data.html

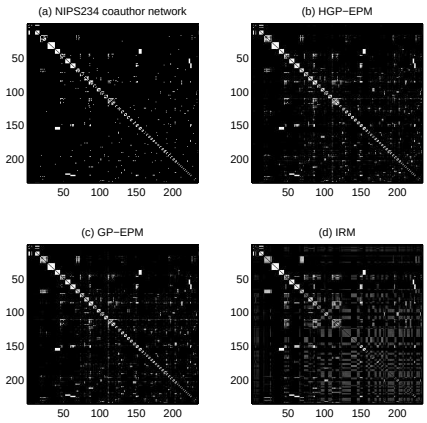

Figure 3: Comparison of three models on estimating the link probabilities for the NIPS234 coauthor network. (a)-(d) Analogous plots to Figures 1 (a)-(d).

Table 3: Comparison of four algorithms on predicting missing edges of the Yeast protein interaction network.

Model	AUC-ROC	AUC-PR
IRM	0.9093 ± 0.0059	0.1878 ± 0.0142
AGM	0.9009 ± 0.0025	0.1225 ± 0.0129
GP-EPM	0.9331 ± 0.0014	**0.2486** ± 0.0149
HGP-EPM	**0.9367** ± 0.0012	**0.2628** ± 0.0184

Table 4: Comparison of four algorithms on predicting missing edges of the NIPS12 coauthor network.

Model	AUC-ROC	AUC-PR
IRM	0.9427 ± 0.0121	0.2066 ± 0.0331
AGM	0.9328 ± 0.0049	0.2350 ± 0.0177
GP-EPM	**0.9768** ± 0.0079	**0.4705** ± 0.0362
HGP-EPM	**0.9762** ± 0.0081	**0.4493** ± 0.0229

that the HGP-EPM performs the best on the Yeast protein-protein interaction network, which is found to clearly exhibit stochastic equivalence by examining the plots corresponding to the ones in Figs. 1 and 3 (not shown for brevity), and the HGP-EPM and GP-EPM both perform well on the NIPS12 coauthor network, which is found to mainly exhibit homophily by examining related plots (not shown for brevity).

As discussed before, the HGP-EPM, GP-EPM, AGM and IRM can all be used to assign nodes to disjoint communities. In Fig. 4 we plot the size of an inferred latent community as a function of its rank (smaller ranks indicate larger sizes) on the log-10 scale, for the four scalable algorithms on the four tested real networks. It is clear that in contrast to the other three latent factor models, the IRM, a latent class model, infers a smaller number of communities, with more larger-size and fewer smaller-size ones. Examining the details we find that the IRM tends to place all the low-degree nodes into one or several large-size communities, whereas the other models are able to better preserve fine details involving small-size communities.

We mention that the HGP-EPM, GP-EPM and AGM have $O(N\bar{d} + NK)$ computation, whereas the Eigenmodel and ILA have at least $O(N^2 + NK)$ computation, where K is the number of latent features. With unoptimized Matlab on a 2.7 GHz CPU, for 1000 MCMC iterations, the HGP-EPM (GP-EPM) takes about 80 (20) seconds on Protein230, about 85 (28) seconds on NIPS234, about 50 (18) minutes on Yeast, and about 32 (12) minutes on NIPS12. The Eigenmodel with $K = 25$ takes about 200 seconds on NIPS234 to run 1000 MCMC iterations. For the ILA on NIPS234, we considered 1000 MCMC iterations that took over 18 hours to run; for Protein230, the ILA inferred about two times more features as it did on NIPS234, and we considered 500 MCMC iterations that took over 21 hours to run.

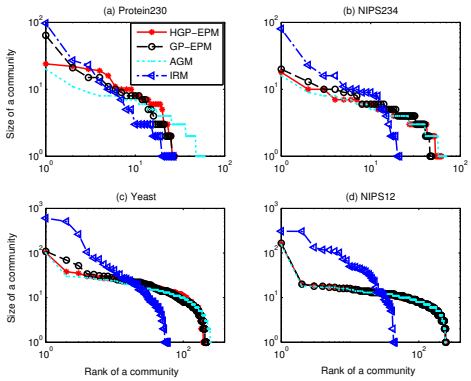

Figure 4: A community' size and its rank.

5 CONCLUSIONS

To model unweighted undirected relational networks characterized by both homophily and stochastic equivalence, we propose a hierarchical gamma process edge partition model (EPM) that supports an infinite number of communities and an infinite square rate matrix to describe community-community interactions. The EPM exploits a Bernoulli-Poisson link to assign a latent count to each binary edge, and further partitions that count according to the edge's affiliations with all pairs of communities, which naturally leads to the partition of each node into overlapping communities. We also provide a simpler gamma process EPM that omits inter-community interactions, which is found to perform well on assortative networks. Efficient MCMC inference with closed-form update equations is provided. Experimental results on four real networks illustrate the EPMs' working mechanisms and properties, as well as their state-of-the-art performance and interpretable latent representations. While previous latent feature relational models and their nonparametric Bayesian versions are often not scalable, our infinite EPMs are readily scalable to networks with thousands of nodes. It would be interesting to investigate strategies to make them scalable to relational networks with millions of nodes and edges.

References

Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, pages 761–764, 2010.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, pages 1981–2014, 2008.

D. Aldous. Exchangeability and related topics. *École d'ete de probabilités de Saint-Flour XIII-1983*, pages 1–198, 1985.

B. Ball, B. Karrer, and M. E. J. Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 2011.

D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen. Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic acids research*, pages 2443–2450, 2003.

G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, pages 531–537, 2005.

J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *ICML*, 2006.

T. S. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, page 016105, 2009.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1973.

S. Fortunato. Community detection in graphs. *Physics Reports*, pages 75–174, 2010.

P. Gopalan, S. Gerrish, M. Freedman, D. M. Blei, and D. M. Mimno. Scalable inference of overlapping communities. In *NIPS*, pages 2249–2257, 2012.

T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

P. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *NIPS*, 2008.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, pages 109–137, 1983.

D. N. Hoover. Row-column exchangeability and a general model for exchangeability. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*. 1982.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, 2006.

D. I. Kim, P. Gopalan, D. M. Blei, and E. Sudderth. Efficient online inference for Bayesian nonparametric relational models. In *NIPS*, pages 962–970, 2013.

J. F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

J. Lloyd, P. Orbanz, Z. Ghahramani, and D. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *NIPS*, 2012.

K. Miller, M. I. Jordan, and T. L. Griffiths. Nonparametric latent feature models for link prediction. In *NIPS*, 2009.

M. Morup, M. N. Schmidt, and L. K. Hansen. Infinite multiple membership relational modeling for complex networks. In *IEEE MLSP*, 2011.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, page 026113, 2004.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *JASA*, pages 1077–1087, 2001.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, pages 814–818, 2005.

K. Palla, D. Knowles, and Z. Ghahramani. A reversible infinite HMM using normalised random measures. In *ICML*, 2014.

K. Palla, D. A. Knowles, and Z. Ghahramani. An infinite latent attribute model for network data. In *ICML*. 2012.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *JASA*, 2006.

J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In *ICDM*, 2012.

J. Yang and J. Leskovec. Structure and overlaps of ground-truth communities in networks. *ACM Trans. Intell. Syst. Technol.*, pages 26:1–26:35, 2014.

M. Zhou and L. Carin. Augment-and-conquer negative binomial processes. In *NIPS*, 2012.

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014.

M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.

Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction: Appendix

A Proof for Lemma 1

Using the law of total expectation, we have

$$\mathbb{E}\left[\sum_{k_1}\sum_{k_2}\lambda_{k_1k_2}\right]=\frac{1}{\beta}\mathbb{E}\left[\xi G(\Omega)+[G(\Omega)]^2-\sum_k r_k^2\right].$$

Using Campbell's theorem (Kingman, 1993), we have

$$\mathbb{E}\left[\sum_k r_k^2\right]=\int_\Omega\int_0^\infty r^2 r^{-1}e^{-c_0 r}drG_0(d\omega)=\frac{\gamma_0}{c_0^2}.$$

The proof is completed by further using $\mathbb{E}[G(\Omega)]=\gamma_0/c_0$ and $\mathbb{E}[G^2(\Omega)]=\gamma_0^2/c_0^2+\gamma_0/c_0^2$. □

B MCMC Inference for HGP-EPM

Sample m_{ij}**.** As in Section 2.2, we sample a latent count for each b_{ij} as

$$(m_{ij}|-)\sim b_{ij}\mathrm{Po}_+\left(\sum_{k_1=1}^K\sum_{k_2=1}^K\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\right). \quad (18)$$

Sample $m_{ik_1k_2j}$**.** Using the relationships between the Poisson and multinomial distributions, similar to the derivation in Zhou et al. (2012), we partition the latent count m_{ij} as

$$(\{m_{ik_1k_2j}\}|-)\sim\mathrm{Mult}\left(m_{ij};\frac{\{\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}\}}{\sum_{k_1}\sum_{k_2}\phi_{ik_1}\lambda_{k_1k_2}\phi_{jk_2}}\right). \quad (19)$$

Note that in each MCMC iteration we store $m_{ik\cdot\cdot}$ and $m_{\cdot k_1k_2\cdot}$ but not necessarily $m_{ik_1k_2j}$ in the memory.

Sample a_i**.** Using (16) and the data augmentation technique developed in Zhou and Carin (2012, 2014) for the negative binomial distribution, we sample a_i as

$$(\ell_{ik}|-)\sim\sum_{t=1}^{m_{ik\cdot\cdot}}\mathrm{Ber}\left(\frac{a_i}{a_i+t-1}\right),$$
$$(a_i|-)\sim\mathrm{Gam}\left(e_0+\sum_k\ell_{ik},\frac{1}{f_0-\sum_k\ln(1-p'_{ik})}\right), \quad (20)$$

where with a slight abuse of notation, but for added conciseness, we use $x\sim\sum_{t=1}^m\mathrm{Ber}[a/(a+t)]$ to represent $x=\sum_{t=1}^m u_t,\ u_t\sim\mathrm{Ber}[a/(a+t)]$.

Sample ϕ_{ik}**.** Using (14) and the gamma-Poisson conjugacy, we have

$$(\phi_{ik}|-)\sim\mathrm{Gam}\big[a_i+m_{ik\cdot\cdot},1/(c_i+\omega_{ik})\big]. \quad (21)$$

Sample r_k**.** Similar to the inference of a_i, using (17), we sample r_k as

$$(l_{kk_2}|-)\sim\sum_{t=1}^{m_{\cdot\cdot kk_2}}\mathrm{Ber}\left(\frac{r_k\xi^{\delta_{kk_2}}(r_{k_2})^{1-\delta_{kk_2}}}{r_k\xi^{\delta_{kk_2}}(r_{k_2})^{1-\delta_{kk_2}}+t-1}\right),$$
$$(r_k|-)\sim\mathrm{Gam}\Bigg[\frac{\gamma_0}{K}+\sum_{k_2}l_{kk_2},\frac{1}{c_0-\sum_{k_2}\xi^{\delta_{kk_2}}(r_{k_2})^{1-\delta_{kk_2}}\ln(1-\tilde{p}_{kk_2})}\Bigg]. \quad (22)$$

Sample ξ**.** We resample the auxiliary variables l_{kk} using the updated r_k and then sample ξ as

$$(\xi|-)\sim\mathrm{Gam}\left[e_0+\sum_k l_{kk},\frac{1}{f_0-\sum_k r_k\ln(1-\tilde{p}_{kk})}\right]. \quad (23)$$

Sample $\lambda_{k_1k_2}$**.** Using (15) and the gamma-Poisson conjugacy, we have

$$(\lambda_{k_1k_2}|-)\sim\mathrm{Gam}\Big[r_{k_1}\xi^{\delta_{k_1k_2}}(r_{k_2})^{1-\delta_{k_1k_2}}+m_{\cdot k_1k_2\cdot},1/(\beta+\theta_{k_1k_2})\Big]. \quad (24)$$

Sample β***,*** c_i ***and*** c_0**.** They can be sampled from gamma distributions using the conjugacy between gamma distributions, omitted here for brevity.

Sample γ_0**.** As show in Lemma 1, the mass parameter γ_0 plays an important role in determining the total sum of the infinite rate matrix $\{\lambda_{k_1k_2}\}$. Our experiments show that it could be used as a tuning parameter to impose one's prior preference on the number of active communities to be discovered. In this paper, we impose a gamma prior as $\gamma_0\sim\mathrm{Gam}(1,1)$ to let the data infer the posterior of γ_0. We employ an independence chain Metropolis-Hastings algorithm to sample γ_0, with the proposal distribution constructed as

$$Q(\gamma_0^*)=\mathrm{Gam}\left(1+\sum_k\tilde{l}_k,\frac{1}{1-\frac{1}{K}\sum_k\ln(1-\tilde{\tilde{p}}_k)}\right), \quad (25)$$

where $(\tilde{l}_k|-)\sim\mathrm{Gam}\left(\sum_{k_2}l_{kk_2},\gamma_0/K\right)$ and

$$\tilde{\tilde{p}}_k:=\frac{-\sum_{k_2}\xi^{\delta_{kk_2}}(r_{k_2})^{1-\delta_{kk_2}}\ln(1-\tilde{p}_{kk_2})}{c_0-\sum_{k_2}\xi^{\delta_{kk_2}}(r_{k_2})^{1-\delta_{kk_2}}\ln(1-\tilde{p}_{kk_2})}.$$

We accept γ_0^* with probability $\min\{1,\pi\}$, where π is

$$\frac{\prod_{k=1}^K\mathrm{Gam}(r_k;\gamma_0^*/K,1/c_0)\mathrm{Gam}(\gamma_0^*;1,1)Q(\gamma_0)}{\prod_{k=1}^K\mathrm{Gam}(r_k;\gamma_0/K,1/c_0)\mathrm{Gam}(\gamma_0;1,1)Q(\gamma_0^*)},$$

which is usually greater than 50% for the networks considered in this paper.

Each iteration of the MCMC for the HGP-EPM proceeds from (18) to (25).

C Gamma Process EPM

The gamma process EPM differs from the HGP-EPM in that it omits inter-community interactions, which leads to a simpler hierarchical model and faster computation at the expense of reduced ability to model stochastic equivalence. It is found to have good performance on assortative networks but not necessarily on disassortative ones.

C.1 Hierarchical Model

The (truncated) gamma process EPM is expressed as

$$\begin{aligned}
&b_{ij} = \mathbf{1}(m_{ij} \geq 1),\\
&m_{ij} = \sum_{k=1}^{K} m_{ijk},\ m_{ijk} \sim \mathrm{Po}\left(r_k \phi_{ik} \phi_{jk}\right),\\
&\phi_{ik} \sim \mathrm{Gam}(a_i, 1/c_i),\ a_i \sim \mathrm{Gam}(e_0, 1/f_0),\\
&r_k \sim \mathrm{Gam}(\gamma_0/K, 1/c_0),\ \gamma_0 \sim \mathrm{Gam}(e_1, 1/f_1),. \quad (26)
\end{aligned}$$

where the $\mathrm{Gam}(1,1)$ prior is also imposed on c_0 and c_i. As $K \to \infty$, we recover the (exact) gamma process with a finite and continuous base measure. We usually set K to be large enough to ensure a good approximation to the truly infinite model.

Note that if we marginalize out both m_{ij} and m_{ijk}, then we have

$$b_{ij} \sim \mathrm{Bernoulli}\left[1 - \prod_{k=1}^{K} \exp\left(-r_k \phi_{ik} \phi_{jk}\right)\right].$$

C.2 Gibbs Sampling

Let the latent counts $m_{i\cdot k}$ and $m_{\cdot\cdot k}$ be defined as

$$\begin{aligned}
m_{i\cdot k} &:= \sum_{j=i+1}^{N} m_{ijk} + \sum_{j=1}^{i-1} m_{jik},\\
m_{\cdot\cdot k} &:= \sum_{i=1}^{N} \sum_{j=i+1}^{N} m_{ijk} = \frac{1}{2} \sum_{i=1}^{N} m_{i\cdot k}\ .
\end{aligned}$$

Using the Poisson additive property, we have

$$m_{i\cdot k} \sim \mathrm{Po}\left(r_k \phi_{ik} \sum_{j \neq i} \phi_{jk}\right), \quad (27)$$

$$m_{\cdot\cdot k} \sim \mathrm{Po}\left(r_k \frac{\sum_i \sum_{j \neq i} \phi_{ik} \phi_{jk}}{2}\right). \quad (28)$$

Marginalizing out ϕ_{ik} from (27), we have

$$m_{ik\cdot} \sim \mathrm{NB}\left(a_i, p'_{ik}\right), \quad (29)$$

where

$$p'_{ik} := \frac{r_k \sum_{j \neq i} \phi_{jk}}{c_i + r_k \sum_{j \neq i} \phi_{jk}}.$$

Marginalizing out r_k from (28), we have

$$m_{\cdot\cdot k} \sim \mathrm{NB}\left(\gamma_0/K,\ \tilde{p}_k\right), \quad (30)$$

where

$$\tilde{p}_k := \frac{\sum_i \sum_{j \neq i} \phi_{ik} \phi_{jk}}{2c_0 + \sum_i \sum_{j \neq i} \phi_{ik} \phi_{jk}}.$$

Similar to the inference techniques used in Appendix B, one may exploit (27)-(30) to derive closed-form Gibbs sampling update equations for all model parameters, omitted here for brevity.

D Gamma Process AGM

Closely related to the gamma process EPM, the hierarchical model for the (truncated) gamma process AGM can be expressed as

$$\begin{aligned}
&b_{ij} = \mathbf{1}(m_{ij} \geq 1),\\
&m_{ij} = u_{ij} + \sum_{k=1}^{K} m_{ijk},\ m_{ijk} \sim \mathrm{Po}\left(r_k \phi_{ik} \phi_{jk}\right),\\
&u_{ij} \sim \mathrm{Po}(\epsilon),\ \epsilon \sim \mathrm{Gam}(a_0, 1/b_0),\\
&\phi_{ik} \sim \mathrm{Ber}(\pi_i),\ \pi_i \sim \mathrm{Beta}(a_1, b_1),\\
&r_k \sim \mathrm{Gam}(\gamma_0/K, 1/c_0),\ \gamma_0 \sim \mathrm{Gam}(e_1, 1/f_1). \quad (31)
\end{aligned}$$

We sample r_k, γ_0 and c_0 in the same way we sample them in the gamma process EPM. To sample ϕ_{ik}, one may use (27) as the likelihood, under which ϕ_{ik} is equal to one a.s. if $m_{i\cdot k} > 0$ and is drawn from a Bernoulli distribution if $m_{i\cdot k} = 0$. Gibbs sampling update equations for the other model parameters can be conviniently derived by exploiting conditional conjugacies, omitted here for brevity.