Community Recovery in a Preferential Attachment Graph

Bruce Hajek and Suryanarayana Sankagiri Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois
Email: {b-hajek,ss19}@illinois.edu

Abstract

A message passing algorithm is derived for recovering communities within a graph generated by a variation of the Barabási-Albert preferential attachment model. The estimator is assumed to know the arrival times, or order of attachment, of the vertices. The derivation of the algorithm is based on belief propagation under an independence assumption. Two precursors to the message passing algorithm are analyzed: the first is a degree thresholding (DT) algorithm and the second is an algorithm based on the arrival times of the children (C) of a given vertex, where the children of a given vertex are the vertices that attached to it. Comparison of the performance of the algorithms shows it is beneficial to know the arrival times, not just the number, of the children. The probability of correct classification of a vertex is asymptotically determined by the fraction of vertices arriving before it. Two extensions of Algorithm C are given: the first is based on joint likelihood of the children of a fixed set of vertices; it can sometimes be used to seed the message passing algorithm. The second is the message passing algorithm. Simulation results are given.[1]

Index terms: preferential attachment graph, message passing algorithm, graphical inference, clustering, community recovery

I. Introduction

Community detection, a form of unsupervised learning, is the task of identifying dense subgraphs within a large graph. For surveys of recent work, see [1]–[3]. Community detection is

[1]This paper was presented in part at the *2018 IEEE International Symposium on Information Theory*

often studied in the context of a generative random graph model, of which the stochastic block model is the most popular. The model specifies how the labels of the vertices are chosen, and how the edges are placed, given the labels. The task of community detection then becomes an inference problem; the vertex labels are the parameters to be inferred, and the graph structure is the data. The advantage of a generative model is that it helps in the design of algorithms for community detection.

The stochastic block model fails to capture two basic properties of networks that are seen in practice. Firstly, it does not model networks that grow over time, such as citation networks or social networks. Secondly, it does not model graphs with heavy-tailed degree distributions, such as the political blog network [4]. The Barabási-Albert model [5], a.k.a. the preferential attachment model, is a popular random graph model that addresses both the above shortcomings. We use the variation of the model introduced by Jordan [6] that includes community structure. The paper [6] considers labels coming from a metric space, though a section of the paper focuses on the case the label space is finite. We consider only a finite label set–the model is described in Section II-A. In recent years there has been substantial study of a variation of preferential attachment model introduced in [7] such that different vertices can have different *fitness*. For example, in a citation network, some papers attract more citations than others published at the same time. There has also been work done on recovering clusters from graphs with different fitness (see Chapter 9 of [8] and references therein). Our work departs from previous work by considering community detection for the model in which the affinity for attachment between an arriving vertex and an existing vertex depends on the labels of both vertices (i.e. for the model of [6]).

The algorithm we focus on is message passing. Algorithms that are precursors to message passing, in which the membership of a vertex is estimated from its radius one neighborhood in the graph, are also discussed. The algorithm is closest in spirit to that in the papers [9], [10]. Message passing algorithms are *local* algorithms; vertices in the graph pass messages to each of their neighbors, in an iterative fashion. The messages in every iteration are computed on the basis of messages in the previous iteration. The degree growth rates for vertices in different communities are different (unless there happens to be a tie) so the neighborhood of a vertex conveys some information about its label. A quantitative estimate of this information is the belief (a posteriori probability) of belonging to a particular community. A much better estimate of a vertex's label could potentially be obtained if the labels of all other vertices were known. Since this information is not known, the idea of message passing algorithms is to have vertices

simultaneously update their beliefs.

The main similarity between the preferential attachment model with communities and the stochastic block model is that both produce locally tree-like graphs. However, the probabilities of edges existing are more complicated for preferential attachment models. To proceed to develop the message passing algorithm, we invoke an independence assumption that is suggested by an analysis of the joint degree evolution of multiple vertices. This approach is tantamount to constructing a belief propagation algorithm for a graphical model that captures the asymptotic distribution of neighborhood structure for the preferential attachment graphs.

a) Organization of the paper: Section II lays the groundwork for the problem formulation and analysis of the community detection problem. It begins by presenting a model for a graph with preferential attachment and community structure, following [6]. The section then presents some key properties of the graphical model in the limit of a large number of vertices. In particular, the empirical distribution of degree, and the evolution of degree of a finite number of vertices, are examined. Stochastic coupling and total variation distance are used extensively. In addition, it is shown that the growth rate parameter for a given fixed vertex can be consistently estimated as the size of the graph converges to infinity. Section III formulates the community recovery problem as a Bayesian hypothesis testing problem, and focuses on two precursors to the message passing algorithm. The first, Algorithm C, estimates the community membership of a vertex based on the children of the vertex (i.e. vertices that attached to the vertex). The second, Algorithm DT, estimates the community membership of a vertex based on the number of children. Section IV investigates an asymptotically equivalent recovery problem, based on a continuous-time random process Z that approximates the evolution of degree of a vertex in a large graph. A key conclusion of that section is that, for the purpose of estimating the community membership of a single vertex, knowing the neighborhood of the vertex in the graph is significantly more informative than knowing the degree of the vertex. Section V presents our main results about how the performance of the recovery Algorithms C and DT scale in the large graph limit. Section VI presents an extension of Algorithm C whereby the labels of a fixed small set of vertices are jointly estimated based on the likelihood of their joint children sets. This algorithm has exponential complexity in the number of labels estimated, but can be used to seed the message passing algorithm. Since the vertices that arrive early have large degree, it can greatly help to correctly estimate the labels of a small number of such vertices. The message passing algorithm is presented in Section VII. Simulation results are given for a variety of examples in Section VIII. Various proofs, and the

derivation of the message passing algorithm, can be found in the appendices.

b) Related work: A different extension of preferential attachment to include communities is given in [11]. In [11], the community membership of a new vertex is determined based on the membership of the vertices to which the new vertex is attached. The paper focuses on the question of whether large communities coexist as the number of vertices converges to infinity. However, the dynamics of the graph itself is the same as in the original Barabási-Albert model. In contrast, our model assumes that community membership of a vertex is determined randomly before the vertex arrives, and the distribution of attachments made depends on the community membership. It might be interesting to consider a combination of the two models, in which some vertices determine community membership exogenously, and others determine membership based on the memberships of their neighbors.

Another model of graphs with community structure and possibly heavy-tailed degree distribution is the degree corrected stochastic block model – see [12] for recent work and references.

There is an extensive literature on degree distributions and related properties of preferential attachment graphs, and an even larger literature on the closely related theory of Polya urn schemes. However, the addition of planted community structure breaks the elegant exact analysis methods, such as the matching equivalence formulated in [13], or methods such as in [14] or [15]. Still, the convergence of the empirical distribution of the induced labels of half edges (see Proposition 2 below) makes the analysis tractable without the exact formulas. A sequence of models evolved from preferential attachment with fitness [7], towards the case examined in [6], such that the attachment probability is weighted by a factor depending on the labels of both the new vertex and a potential target vertex. The model of [16] is a special case, for which attachment is possible if the labels are sufficiently close. See [6], [8], [16] for additional background literature.

II. Preliminaries and some asymptotics

A. Barabási - Albert preferential attachment model with community structure

The model consists of a sequence of directed graphs, $(G_t = (V_t, E_t) : t \geqslant t_o)$ and vertex labels $(\ell_t : t \geqslant 1)$ with distribution determined by the following parameters:[2]

[2]The model is the same as the finite metric space case of [6] except for differences in notation. $\alpha, S, X, \mu, \nu, Y, \phi$ in [6] are $\beta^T, [r], \ell, \rho, \eta, C, 2\theta$ here. Also, [6] denotes the initial graph as G_0 while we denote it by G_{t_o}, we assume it has mt_0 edges, and we suppose the random evolution begins with the addition of vertex $t_o + 1$.

- $m \geqslant 1$: out degree of each added vertex
- $r \geqslant 1$: number of possible labels; labels are selected from $[r] \triangleq \{1, \ldots, r\}$
- $\rho = (\rho_1, \ldots, \rho_r)$: a priori label probability distribution
- $\boldsymbol{\beta} \in \mathbb{R}^{r \times r}$: matrix of strictly positive affinities for vertices of different labels; β_{uv} is the affinity of a new vertex with label u for attachment to a vertex of label v.
- $t_o \geqslant 1$: initial time
- $G_{t_o} = (V_{t_o}, E_{t_o})$: initial directed graph with $V_{t_o} = [t_o]$ and mt_o directed edges
- $(\ell_t : t \in [t_o]) \in [r]^{t_o}$: labels assigned to vertices in G_{t_o}.

For each $t \geqslant t_o$, G_t has t vertices given by $V_t = [t]$ and mt edges. The graphs can contain parallel edges. No self loops are added during the evolution, so if G_{t_o} has no self loops, none of the graphs will have self loops. Of course, by ignoring the orientation of edges, we could obtain undirected graphs.

Given the labeled graph G_t, the graph G_{t+1} is constructed as follows. First vertex $t+1$ is added and its label ℓ_{t+1} is randomly selected from $[r]$ using distribution ρ, independently of G_t. Then m outgoing edges are attached to the new vertex, and the head ends of those edges are selected from among the vertices in $V_t = [t]$ using sampling with replacement, and probability distribution given by preferential attachment, weighted based on labels according to the affinity matrix.

The probabilities are calculated as follows. Note that E_t has mt edges, and thus $2mt$ *half edges*, where we view each edge as the union of two half edges. For any edge, its two half edges are each incident to a vertex; the vertices the two half edges are incident to are the two vertices the edge is incident to. Suppose each half edge inherits the label from the vertex it is incident to. If $\ell_{t+1} = u$, meaning the new vertex has label u, and if one of the existing half edges has label v, then the half edge is assigned weight β_{uv} for the purpose of adding edges outgoing from vertex $t+1$. For each one of the new edges outgoing from vertex $t+1$, an existing half edge is chosen at random from among the $2mt$ possibilities, with probabilities proportional to such weights. The selection is done simultaneously for all m of the new edges, or equivalently, sampling with replacement is used. Then the vertices of the respective selected half edges become the head ends of the m new edges.

B. Empirical degree distribution for large T

For a vertex in G_t, where $t \geqslant t_o$, the distribution of the number of edges incident on the vertex from vertex $t+1$ depends on the label of the vertex, the degree of the vertex, and the labels on all the half edges incident to the existing vertices in G_t. The empirical distribution of labels of half edges in G_t converges almost surely as $t \to \infty$, as explained next. Let $C_t = (C_{t,u} : u \in [r])$ for $t \geqslant t_o$, where $C_{t,u}$ denotes the number of half edges with label u in G_t. It is easy to see that $(C_t : t \geqslant t_o)$ is a discrete-time Markov process, with initial state determined by the labels of vertices in G_{t_o}. Let $\eta_t = \frac{C_t}{2mt}$. Thus, $\eta_{t,u}$ is the fraction of half edges that have label u at time t. Let $h = (h_1, \ldots, h_r)$ where

$$h_v(\eta) = \rho_v + \sum_u \rho_u \left(\frac{\beta_{uv}\eta_v}{\sum_{v'} \beta_{uv'}\eta_{v'}} \right) - 2\eta_v. \tag{1}$$

The following is proved in [6], by appealing to the theory of stochastic approximation. For convenience we give essentially the same proof, using our notation, in Appendix A.

Proposition 1. *[6] (Limiting fractions of half edges with given labels) $\eta_t \to \eta^*$ a.s. as $t \to \infty$, where η^* is the unique probability vector such that $h(\eta^*) = 0$.*

A second limit result we restate from [6] concerns the empirical degree distribution for the vertices with a given label. For $v \in [r]$ and integers $n \geqslant m$ and T, let:

- $H^v(T)$ denote the number of vertices with label v in G_T
- $N_n^v(T)$ denote the number of vertices with label v and with degree n in G_T
- $P_n^v(T) = \frac{N_n^v(T)}{H^v(T)}$ denote the fraction of vertices with label v that have degree n in G_T.

Let

$$\theta^*_{u,v} = \frac{\beta_{uv}}{2\sum_{v'} \beta_{uv'}\eta^*_{v'}} \quad \text{for } u, v \in [r],$$

and

$$\theta^*_v = \sum_u \rho_u \theta^*_{u,v} \quad \text{for } v \in [r]. \tag{2}$$

Proposition 2. *[6] (Limiting empirical distribution of degree for a given label) Let $n \geqslant m$ and $v \in [r]$ be fixed. Then $\lim_{T\to\infty} P_n^v(T) = p_n(\theta^*_v, m)$ almost surely, where*

$$\begin{aligned} p_n(\theta, m) &= \frac{\Gamma\left(\frac{1}{\theta} + m\right)\Gamma(n)}{\theta\Gamma(m)\Gamma\left(n + \frac{1}{\theta} + 1\right)} \\ &\asymp \left[\frac{\Gamma\left(\frac{1}{\theta} + m\right)}{\theta\Gamma(m)}\right] \frac{1}{n^{\frac{1}{\theta}+1}} \end{aligned} \tag{3}$$

The asymptotic equivalence in (3) as $n \to \infty$ follows from Sterling's formula for the Gamma function. The proposition shows that the limiting degree distribution of a vertex with label v selected uniformly at random from among the vertices with label v in G_T has probability mass function with tail decreasing like $n^{-\left(\frac{1}{\theta_v^*}+1\right)}$. If $\beta_{u,v}$ is the same for all u, v then $\theta_v^* = 1/2$ for all v and we see the classical tail exponent -3 for the Barabási-Albert model.

The proof of Proposition 2 given in [6] is based on examining the evolution of the fraction of vertices with a given label and given degree n. Using the convergence analysis of stochastic approximation theory, this yields limiting difference equations for p_n that can be solved to find p_n. However, since all vertices with a given label are grouped together, the analysis does not identify the limiting degree distribution of a vertex as a function of the arrival time of the vertex.

The following section investigates the evolution of the degree of a single vertex, or finite set of vertices, conditioned on their labels. As a preliminary application, we produce an alternative proof of Proposition 2 in Appendix D. The main motivation for this alternative approach is that it can also be applied to analyze the probability of label error as a function of time of arrival of a vertex, for two of the recovery algorithms we consider.

C. Evolution of vertex degree–the processes $Y, \widetilde{Y}, \breve{Y}$, and Z

Consider the preferential attachment model defined in Section II-A. Given a vertex τ with $\tau \geqslant t_o + 1$, consider the process $(Y_t : t \geqslant \tau)$, where Y_t is the degree of vertex τ at time t. So $Y_\tau = m$. The conditional distribution (i.e. probability law) of $Y_{t+1} - Y_t$ given $(Y_t, \eta_t, \ell_\tau = v, \ell_{t+1} = u)$ is given by:

$$\mathcal{L}(Y_{t+1} - Y_t | Y_t, \eta_t, \ell_\tau = v, \ell_{t+1} = u) = \mathsf{binom}\left(m, \frac{\theta_{u,v,t} Y_t}{mt}\right),$$

where

$$\theta_{u,v,t} = \frac{\beta_{uv}}{2\sum_{v'} \beta_{uv'} \eta_{tv'}}.$$

It follows that, given $(Y_t, \eta_t, \ell_\tau = v)$, the conditional distribution of $Y_{t+1} - Y_t$ is a mixture of binomial distributions with selection probability distribution ρ, which we write as:

$$\mathcal{L}(Y_{t+1} - Y_t | Y_t, \eta_t, \ell_\tau = v) = \sum_{u \in [r]} \rho_u \mathsf{binom}\left(m, \frac{\theta_{u,v,t} Y_t}{mt}\right).$$

Proposition 1 implies, given any $\epsilon > 0$, if τ is sufficiently large, $\mathbb{P}\left\{\|\eta_t - \eta^*\| \leqslant \epsilon \text{ for all } t \geqslant \tau\right\} \geqslant 1 - \epsilon$. Therefore, $\theta_{u,v.t} \approx \theta_{u,v}^*$ for $v \in [r]$. A mixture of binomial distributions, all with small

means, can be well approximated by a Bernoulli distribution with the same mean. Thus, we expect $\mathcal{L}(Y_{t+1} - Y_t | Y_t, \ell_\tau = v) \approx \mathsf{Ber}\left(\frac{\theta^*_v Y_t}{t}\right)$.

Based on these observations, we define a random process that is an idealized variation of Y obtained by replacing η_t by the constant vector η^*, and allowing jumps of size one only. The process $\widetilde{Y}$ has parameters $\tau, m,$ and ϑ, where τ is the activation time, m is the state at the activation time, and $\vartheta > 0$ is a rate parameter. The process $\widetilde{Y}$ is a time-inhomogeneous Markov process with initial value $Y_\tau = m$. For $t \geqslant \tau$ and y such that $\frac{\vartheta y}{t} \leqslant 1$, we require:

$$\mathcal{L}(\widetilde{Y}_{t+1} - \widetilde{Y}_t | \widetilde{Y}_t = y) = \mathsf{Ber}\left(\frac{\vartheta y}{t}\right). \tag{4}$$

By induction, starting at time τ, we find that $\widetilde{Y}_t \leqslant m + t - \tau$ for $t \geqslant \tau$. If $\tau \geqslant m$ and $\vartheta \leqslant 1$, then $\frac{\vartheta \widetilde{Y}_t}{t} \leqslant 1$ for all $t \geqslant \tau$ with probability one, in which case (4) and the initial condition completely specify the distribution of $(\widetilde{Y} : t \geqslant \tau)$. However, for added generality we allow $\vartheta > 1$, in which case the above construction can break down. To address such situation, we define ζ such that ζ is the stopping time $\zeta \triangleq \inf\{t : \vartheta \widetilde{Y}_t > t\}$ and we define $\widetilde{Y}_t = +\infty$ for $t > \zeta$.

The process Y can be thought of as a (non Markovian) discrete time birth process with activation time τ and birth probability at a time t proportional to the number of individuals. However, the birth probability (or birth rate) per individual, $\frac{\theta^*_v}{t}$, has a factor $\frac{1}{t}$, which tends to decrease the birth rate per individual. To obtain a process with constant birth rate per individual we introduce a time change by using the process $(Y_{e^s} : s \geqslant 0)$. In other words, we use t for the original time variable and $s = \ln t$ as a new time variable. We will define a process Z such that $(Z_{\ln(t/\tau)} : t \geqslant \tau) \approx (Y_t : t \geqslant \tau)$, or equivalently, $(Z_s : s \geqslant 0) \approx (Y_{\tau \mathrm{e}^s} : s \geqslant 0)$, in a sense to be made precise.

The process $Z = (Z_s : s \geqslant 0)$ is a continuous time pure birth Markov process with initial state $Z_0 = m$ and birth rate ϑk in state k, for some $\vartheta > 0$. (It is a simple example of a Bellman-Harris process, and is related to busy periods in Markov queueing systems.) The process Z represents the total number of individuals in a continuous time branching process beginning with m individuals activated at time 0, such that each individual spawns another at rate ϑ. For fixed s, Z_s has the negative binomial distribution $\mathsf{negbinom}(m, \mathrm{e}^{-s\vartheta})$. In other words, its marginal probability distribution $(\pi_n(s, \vartheta, m) : n \in \mathbb{Z}_+)$ is given by

$$\pi_n(s, \vartheta, m) = \binom{n-1}{m-1} \mathrm{e}^{-m\vartheta s} (1 - \mathrm{e}^{-\vartheta s})^{n-m} \quad \text{for } n \geqslant m. \tag{5}$$

In particular, taking $m = 1$ shows $\pi(s, \vartheta, 1)$ is the geometric distribution with parameter $\mathrm{e}^{-\vartheta s}$, and hence, mean $\mathrm{e}^{\vartheta s}$. The expression (5) can be easily derived for $m = 1$ by solving the Kolmogorov

forward equations recursively in n: $\dot{\pi}_n = -\vartheta n \pi_n + \vartheta(n-1)\pi_{n-1}$ for $n \geqslant 1$, with the convention and base case, $\pi_0 \equiv 0$. For $m \geqslant 2$, the process Z has the same distribution as the sum of m independent copies of Z with $m = 1$, proving the validity of (5) by the same property for the negative binomial distribution.

Let $\breve{Y}_t = Z_{\ln(t/\tau)}$ for integers $t \geqslant \tau$. The mapping from Z to $\breve{Y}$ does not depend on the parameter ϑ, so a hypothesis testing problem for Z maps to a hypothesis testing problem for $\breve{Y}$. There is loss of information because the mapping is not invertible, but the loss tends to zero as $\tau \to \infty$, because the rate of sampling of Z increases without bound.

The following proposition, proven in Appendix B, shows that $Y, \widetilde{Y}$ and $\breve{Y}$ are asymptotically equivalent in the sense of total variation distance. Since the processes $Y, \widetilde{Y}$ and $\breve{Y}$ are integer valued, discrete time processes, their trajectories over a finite time interval $[\tau, T]$ have discrete probability distributions. See the beginning of Appendix B for a review of the definition of total variation distance and its significance for coupling. Sometimes we write $\widetilde{Y}(\vartheta)$ instead of $\widetilde{Y}$, and $\breve{Y}(\vartheta)$ instead of $\breve{Y}$, to denote the dependence on the parameter ϑ.

Proposition 3. *Suppose $\tau, T \to \infty$ such that $T > \tau$ and T/τ is bounded. Fix $v \in [r]$. Then*

$$d_{TV}((Y_{[\tau,T]}|\ell_\tau = v), \widetilde{Y}_{[\tau,T]}(\theta^*_v)) \to 0, \tag{6}$$

and for any $\vartheta > 0$,

$$d_{TV}\left(\widetilde{Y}_{[\tau,T]}(\vartheta), \breve{Y}_{[\tau,T]}(\vartheta)\right) \to 0. \tag{7}$$

The first part of Proposition 3 can be strengthened as follows. The labels in $\ell_{[1,T]}$ are mutually independent, each with distribution ρ. We can define a joint probability distribution over $(\widetilde{Y}_{[\tau,T]}, \ell_{[1,T]})$ by specifying the conditional probability distribution of $\widetilde{Y}_{[\tau,T]}$ given $\ell_{[1,T]}$ as follows. Given $\ell_{[1,T]}$, $\widetilde{Y}_{[\tau,T]}$ is a Markov sequence with $\widetilde{Y}_\tau = m$ and:

$$\mathcal{L}(\widetilde{Y}_t - \widetilde{Y}_{t-1}|\ell_t = u, \ell_\tau = v, \widetilde{Y}_{t-1} = y) = \mathsf{Ber}\left(\frac{y\theta^*_{u,v}}{t-1}\right). \tag{8}$$

By the law of total probability, this gives the same marginal distribution for $\mathcal{L}(\widetilde{Y}_{[\tau,T]}|\ell_{\tau=v})$ as (4) with $\vartheta = \theta^*_v$, as long as $\max_{u,v}\{\theta^*_{u,v}\}y \leqslant t$.

Proposition 4. *Suppose $\tau, T \to \infty$ such that $T > \tau$ and T/τ is bounded. Fix $v \in [r]$. Then*

$$d_{TV}\left(\left(Y_{[\tau,T]}, \ell_{[1,T]}\right), \left(\widetilde{Y}_{[\tau,T]}, \ell_{[1,T]}\right)\right) \to 0, \tag{9}$$

The proof is a minor variation of the proof of Proposition 3 because the estimates on total variation distance are uniform for θ^*_v or $\theta^*_{u,v}$ bounded. Details are left to the reader.

D. Joint evolution of vertex degrees

Instead of considering the evolution of degree of a single vertex we consider the evolution of degree for a finite set of vertices, still for the preferential attachment model with communities, $(G_t = (V_t, E_t) : t \geqslant t_o)$, defined in Section II-A. Given integers $\tau_1, \ldots, \tau_J$ with $t_o < \tau_1 < \cdots < \tau_J$, let $Y^j_t = 0$ if $1 \leqslant t < \tau_j$ and let Y^j_t denote the degree of vertex τ_j at time t if $t \geqslant \tau_j$. Let $Y^{[J]}_t = (Y^j_t : j \in [J])$. Let $(v_1, \ldots, v_J) \in [r]^J$. We consider the evolution of $(Y^{[J]}_t : t \geqslant 1)$ given $(\ell_{\tau_1}, \ldots, \ell_{\tau_J}) = (v_1, \ldots, v_J)$. Let $\vartheta_j = \theta^*_{v_j}$ for $j \in [J]$. About the notation θ^* vs. ϑ: The vector $\theta^* = (\theta^*_v : v \in [r])$ denotes the limiting rate parameters for the r possible vertex labels defined in (2), whereas $\vartheta = (\vartheta_j : j \in [J])$ denotes the limiting rate parameters for the specific set of J vertices being focused on, conditioned on their labels being $v_1, \ldots, v_J$.

The process $\widetilde{Y}^{[J]}$ is defined similarly. Fix $J \geqslant 1$, integers $\tau_1, \ldots, \tau_J$ with $1 \leqslant \tau_1 < \ldots < \tau_J$, and $\underline{\vartheta} \in (\mathbb{R}_{>0})^J$. Suppose for each $j \in [J]$ that $\widetilde{Y}^j$ is a version of the process $\widetilde{Y}$ defined in Section II-C, with parameters τ_j, m, and ϑ_j, with the extension $\widetilde{Y}^j_t = 0$ for $1 \leqslant t \leqslant \tau_j - 1$. Furthermore, suppose the J processes $(\widetilde{Y}^j)_{j \in [J]}$ are mutually independent. Finally, let $\widetilde{Y}^{[J]} = (\widetilde{Y}^{[J]}_t : t \geqslant 1)$ where $\widetilde{Y}^{[J]}_t = (\widetilde{Y}^j_t : j \in [J])$. Note that $\widetilde{Y}^{[J]}$ is itself a time-inhomogeneous Markov process. In what follows we write $\widetilde{Y}^{[J]}(\underline{\vartheta})$ instead of $\widetilde{Y}^{[J]}$ when we wish to emphasize the dependence on the parameter vector $\underline{\vartheta}$. Let $\breve{Y}^{[J]}$ be defined analogously, based on $\breve{Y}$.

Proposition 5. *Fix the parameters of the preferential attachment model, $m, r, \beta, \rho, t_o, G_{t_o}, \ell_{[1,t_o]}$. Fix $J \geqslant 1$ and $v_1, \ldots, v_J \in [r]$, and let $\vartheta_j = \theta^*_{v_j}$ for $j \in [J]$. Let $\tau_0 \to \infty$ and let $\tau_1, \ldots, \tau_J$ and T vary such that $\tau_0 \leqslant \tau_1 < \ldots < \tau_J$, and T/τ_0 is bounded. Then*

$$d_{TV}\left(\widetilde{Y}^{[J]}_{[1,T]}(\underline{\vartheta}), \left(Y^{[J]}_{[1,T]} \middle| \ell_{\tau_j} = v_j \text{ for } \ j \in [J]\right)\right) \to 0$$
$$d_{TV}\left(\widetilde{Y}^{[J]}_{[1,T]}(\underline{\vartheta}), \breve{Y}^{[J]}_{[1,T]}(\underline{\vartheta})\right) \to 0$$

The proposition is proved in Appendix C. A key implication of the proposition is that the degree evolution processes for a finite number of vertices are asymptotically independent in the assumed asymptotic regime. In particular, the following corollary is an immediate consequence of the proposition. It shows that the degrees of J vertices at a fixed time T are asymptotically independent with marginal distributions given by (5).

Corollary 1. *(Convergence of joint distribution of degrees of J vertices at a given time) Under the conditions of Proposition 5, for a vector $\underline{n} = (n_1, \ldots, n_J)$ with $n_j \geqslant m$,*

$$\lim_{\tau_0 \to \infty} \sup_{\tau_1, \ldots, \tau_J, T} \left| \mathbb{P}\left\{ Y_T^{[J]} = \underline{n} \middle| (\ell_{\tau_1}, \ldots, \ell_{\tau_J}) = (v_1, \ldots, v_J) \right\} - \prod_{j \in [J]} \pi_{n_j} \left(\ln(T/\tau_j), \vartheta_j, m \right) \right| = 0.$$

Remark 1. *Corollary 1 implies, given $\ell_\tau = v$, the limiting distribution of the degree of τ in G_T is* $\mathsf{negbinom}(m, (\tau/T)^{\theta_v^*})$*, as $\tau, T \to \infty$ with $\tau \leqslant T$ and T/τ bounded. This generalizes the result known in the classical case $\beta_{u,v} \equiv 1$ where $\theta_v^* = 1/2$, shown on p. 286 of [13].*

E. Large time evolution of degree of a fixed vertex and consistent estimation of the rate parameter of a vertex

Consider the Barabási-Albert model with communities. Fix $\tau \geqslant 1$ and let Y_t denote the degree of τ in G_t for $t \geqslant t_o$. To avoid triviality, assume τ is not an isolated vertex in the initial graph G_{t_o}. The following proposition offers a way to consistently estimate the rate parameter $\theta_{\ell_\tau}^*$. If the parameters θ_v^* of the Barabási-Albert model are distinct, it follows that any fixed finite set of vertices could be consistently classified in the limit as $T \to \infty$, without knowledge of the model parameters.

Proposition 6. *(Large time behavior of degree evolution) For τ fixed,*

$$\lim_{T \to \infty} \frac{\ln Y_T}{\ln(T/\tau)} = \theta_{\ell_\tau}^* \quad a.s. \tag{10}$$

Here, "a.s." means almost surely, or in other words, with probability one.

The following strengthening of Proposition 6 is conjectured.

Conjecture 1. *(Sharp large time behavior of degree evolution) For τ fixed,*

$$\lim_{T \to \infty} \frac{Y_T}{(T/\tau)^{\theta_{\ell_\tau}^*}} = W \ a.s. \tag{11}$$

for a random variable W with $\mathbb{P}\{W > 0\} = 1$.

See Appendix E for a proof of the proposition and a proof that (11) holds with Y replaced by $\breve{Y}$.

III. Community recovery based on children

Given vertices τ and τ_0, we say τ is a child of τ_0, and τ_0 is a parent of τ, if $\tau \geqslant \max\{\tau_0, t_o\}+1$, and there is an edge from τ to τ_0. It is assumed that the known initial graph G_{t_o} is arbitrary and carries no information about vertex labels. Thus, for the purpose of inferring the vertex labels, the edges in G_{t_o} are not relevant beyond the degrees that they imply for the vertices in G_{t_o}. Assuming T is an integer with $1 \leqslant \tau < T$, let $\partial\tau$ denote the children of τ in G_T and $\wp\tau$ the parents of τ. So $\wp\tau = \varnothing$ if $\tau \leqslant t_o$ and $\partial\tau \subset \{t_o + 1, \ldots, T\}$.

Consider the problem of estimating ℓ_τ given observation of a random object $\mathcal{O}$. For instance, the object could be the degree of vertex τ in G_T, or it could be the set of children of τ in G_T, or it could be the entire graph. This is an r-ary hypothesis testing problem. It is assumed a priori that the label ℓ_τ has probability distribution ρ, so it makes sense to try to minimize the probability of error in the Bayesian framework. Let Λ_τ denote the log-likelihood vector defined by $\Lambda_\tau(\mathcal{O}|i) = \ln p(\mathcal{O}|\ell_\tau = i)$ for $i \in [r]$. By a central result of Bayesian decision theory, the optimal decision rule is the MAP estimator, given by

$$\widehat{\ell}_{\tau,MAP} = \arg\max_i \left(\ln \rho_i + \Lambda_\tau(\mathcal{O}|i)\right)$$

Remark 2. *(i) Knowing G_T is equivalent to knowing the indices of the vertices and the undirected graph induced by dropping the orientations of the edges of G_T.*

(ii) The estimators considered in this paper are assumed to know the order of arrival of the vertices (which we take to be specified by the indices of the vertices for brevity) and the parameters m, β and ρ. It is clear that in some cases the parameters can be estimated from a realization of the graph for sufficiently large T. In particular, the parameter m is directly observable. By Proposition 6, if the order of arrival is known, the set of growth rates $\{\theta^*_v : v \in [r]\}$ can be estimated. So if the θ^*_v's are distinct, the distribution ρ can also be consistently estimated.

(iii) If the indices of the vertices are not known and only the undirected version of the graph is given, it may be possible to estimate the indices if m is sufficiently large. Such problem has been explored recently for the classical Barabási-Albert model [17], but we don't pursue it here for the variation with a planted community.

Algorithm C: The first recovery algorithm we describe, Algorithm C ("C" for "children"), is to let $\mathcal{O}$ denote the set of children, $\partial t = \{t_1, \ldots, t_n\}$, of vertex τ in G_T. Equivalently, $\mathcal{O}$ could be observation of $Y_{[\tau \vee t_o, T]}$, with parameters m and θ^*_v, where $\tau \vee t_o = \max\{\tau, t_o\}$. However,

motivated by Proposition 3, we consider instead observation of $\widetilde{Y}_{[\tau\vee t_o,T]}$, which has a distribution asymptotically equivalent to the distribution of $Y_{[\tau\vee t_o,T]}$. Let $d_0(\tau)$ denote the initial degree of vertex τ, defined to be the degree of τ in G_{t_o} if $\tau \leqslant t_o$ and $d_0(\tau) = m$ otherwise. Given a possible children set $\partial t = \{t_1, \ldots, t_n\}$, let $y^{\partial\tau}_{[\tau,T]}$, denote the corresponding degree evolution sample path: $y^{\partial\tau}_t = d_o(\tau) + |\partial\tau \cap [\tau, t]|$ for $\tau \vee t_o \leqslant t \leqslant T$, The probability $\widetilde{Y}_{[\tau\vee t_o,T]}$ corresponds to children set $\partial t = \{t_1, \ldots, t_n\}$ is given by

$$P(\partial t = \{t_1, \ldots, t_n\}) = \prod_{t\in[\tau\vee t_o+1,T]\setminus\partial\tau} \left(1 - \frac{y^{\partial\tau}_{sot-1}\theta^*_v}{t-1}\right) \prod_{t\in\partial\tau} \frac{y^{\partial\tau}_{t-1}\theta^*_v}{t-1},$$

so the log likelihood for observation $\widetilde{Y}_{[\tau\vee t_o,T]} = y^{\partial\tau}_{[\tau\vee t_o,T]}$ is:

$$\Lambda^C_\tau = |\partial\tau| \ln\theta^*_v + \sum_{t\in[\tau\vee t_o+1,T]\setminus\partial\tau} \ln\left(1 - \frac{y^{\partial\tau}_{t-1}\theta^*_v}{t-1}\right)$$

Algorithm C for estimating ℓ_τ is to use the MAP estimator based on ρ and Λ^C_τ. Using the approximation $\ln(1+s) = s$ and approximating the sum by an integral we find $\Lambda^C_\tau \approx \lambda_\tau$, where

$$\begin{aligned}\lambda^C_\tau(v) &\triangleq |\partial\tau| \ln\theta^*_v - \theta^*_v \int_{\tau\vee t_o}^T \frac{y^{\partial\tau}_t}{t} dt \\ &= |\partial\tau| \ln\theta^*_v + \theta^*_v \left(d_o(\tau) \ln\frac{\tau\vee t_o}{T} + \sum_{t\in\partial\tau} \ln\frac{t}{T}\right). \end{aligned} \tag{12}$$

Algorithm DT: The second recovery algorithm we describe, Algorithm DT ("DT" for "degree thresholding"), is to let $\mathcal{O}$ denote the number of children of vertex τ in G_T, or, equivalently, the degree of τ at time T minus the initial degree of τ. Equivalently, $\mathcal{O}$ could be observation of $Y_T - d_o(\tau)$. However, motivated by Proposition 3, we consider instead consider observation of $\breve{Y}_T - d_0(\tau)$, which has the $\mathsf{negbinom}\left(d_o(\tau), (\tau/T)^{\theta^*_v}\right)$ distribution given $\ell_\tau = v$, for $v \in [r]$. The log likelihood vector in this case, given the number of children, $|\partial\tau|$, is:

$$\Lambda^{DT}_\tau(v) = -d_o(\tau)\theta^*_v \ln(T/\tau) + |\partial\tau| \ln\left(1 - (\tau/T)^{\theta^*_v}\right),$$

where we have dropped a term (log of binomial coefficient) not depending on v. Algorithm DT for estimating ℓ_τ is to use the MAP decision rule based on ρ and Λ^{DT}, or in other words, the MAP decision rule based on $\mathcal{O} = \breve{Y}_T$, or equivalently, based on $\mathcal{O} = Z_{\bar{s}}$, where $\bar{s} = \ln(T/\tau)$ (because $\breve{Y}_T = Z_{\bar{s}}$). Let $f^{DT}_Z(\rho, \theta^*, m, \bar{s})$ denote the resulting average error probability p_e.

IV. HYPOTHESIS TESTING FOR Z

Proposition 3 gives an asymptotic equivalence of $Y_{[\tau,T]}, \widetilde{Y}_{[\tau,T]}$, and $\breve{Y}_{[\tau,T]}$. Recall that $\breve{Y}_{[\tau,T]}$ is obtained by sampling the continuous time process $Z_{\ln(t/\tau)}$ at integers $t \in [\tau, T]$. Thus, the continuous time process Z is not observable. However, as $\tau \to \infty$, the rate that Z is sampled increases without bound, so asymptotically $Z_{[0,\ln(T/\tau)]}$ is observed. We consider here the hypothesis testing problem based on observation of $Z_{[0,\ln(T/\tau)]}$ such that under H_v it has rate parameter $\vartheta = \theta_v^*$ for $v \in [r]$. This is sensible in case the parameter values θ_v^*, $v \in [r]$, are distinct. To this end, we derive the log likelihood vector.

Suppose $\{s_1, \ldots, s_n\} \subset (0, \bar{s}]$ such that $0 < s_1 < \cdots < s_n$ and $\bar{s} = \ln T/\tau$. Since the inter-jump periods are independent (exponential) random variables, the likelihood of $s_1, \ldots, s_n$ being the jump times during $[0, \bar{s}]$ under hypothesis H_v, is the product of the likelihoods of the observed inter-jump periods, with an additional factor of the likelihood of not seeing a jump in the last interval:

$$\left(\prod_{i=0}^{n-1} \theta_v^*(m+i) e^{-\theta_v^*(m+i)(s_{i+1}-s_i)} \right) e^{-\theta_v^*(n+m)(\bar{s}-s_n)}$$

Thus, the log likelihood for observing this is (letting $s_0 = 0$):

$$\Lambda^Z = n \ln \theta_v^* - \theta_v^* \left(m\bar{s} + \sum_{i=1}^{n} (\bar{s} - s_i) \right) \tag{13}$$

(With $s_i = \ln(t_i/\tau)$, (13) is the same as (12), although in (12) the variables t_i are supposed to be integer valued.) Let $A_{\bar{s}} \triangleq (m\bar{s} + \sum_{i=1}^{n} (\bar{s} - s_i))$. Note that $A_{\bar{s}}$ is the area under the trajectory of $Z_{[0,\bar{s}]}$. Moreover, $n + m$ is the value of $Z_{\bar{s}}$. So the log-likelihood vector is given by:

$$\Lambda^Z = (Z_{\bar{s}} - m) \ln \theta_v^* - (A_{\bar{s}})\theta_v^*, \tag{14}$$

which is a linear combination of $Z_{\bar{s}} - m$ and $A_{\bar{s}}$. Thus, the MAP decision rule has a simple form. Let $f_Z^C(\rho, \theta^*, m, \bar{s})$ denote the average error probability p_e for the MAP decision rule based on observation of $Z_{[0,\bar{s}]}$.

There is apparently no closed form expression for the distribution of Λ^Z, so computation of $f_Z^C(\rho, \theta^*, m, \bar{s})$ apparently requires Monte Carlo simulation or some other numerical method. A closed form expression for the moment generating function of Λ^Z is given in the following proposition, proved in Appendix F, and it can be used to either bound the probability of error or to accelerate its estimation by importance sampling.

Proposition 7. *The joint moment generating function of Z_s and A_s is given as follows, where $\mathbb{E}_{\lambda,m}[\cdot]$ denotes expectation assuming the parameters of Z are λ, m :*

$$\begin{aligned}\psi_{\lambda,m}(u,v,s) &\triangleq \mathbb{E}_{\lambda,m}\left[e^{uZ_s+vA_s}\right] \\ &= \left(\frac{e^{(v-\lambda)s+u}}{1+\frac{\lambda e^u}{v-\lambda}\left(1-e^{(v-\lambda)s}\right)}\right)^m. \end{aligned} \tag{15}$$

Proposition 7 can be used to bound p_e for the special case of two possible labels, $r = 2$, in which estimating ℓ_τ is a binary hypothesis testing problem: $H_1 : \vartheta = \theta_1^*$, vs. $H_2 : \vartheta = \theta_2^*$. For such a problem the likelihood vector Λ^Z can be replaced by the log likelihood ratio, $\Lambda = \Lambda^Z(1) - \Lambda^Z(2)$. By a standard result in the theory of binary hypothesis testing (due to [18], stated without proof in [19], proved in special case $\pi_1 = \pi_2 = 0.5$ in [20], and same proof easily extends to general case), the probability of error for the MAP decision rule is bounded by

$$\pi_1\pi_2\rho_B^2 \leqslant p_e \leqslant \sqrt{\pi_1\pi_2}\rho_B, \tag{16}$$

where the Bhattacharyya coefficient (or Hellinger integral) ρ_B is defined by $\rho_B = \mathbb{E}\left[\mathrm{e}^{\Lambda/2}\middle|H_2\right]$, and π_1 and π_2 are the prior probabilities on the hypotheses. The proposition with $\lambda = \theta_2^*$, $u = \frac{1}{2}\ln(\theta_1^*/\theta_2^*)$, $v = -\frac{\theta_1^*-\theta_2^*}{2}$, and $s = \bar{s}$ yields

$$\begin{aligned}\rho_{B,C} &= \mathbb{E}_{\lambda,m}\left[e^{u(Z_s-m)+vA_s}\right] = \psi_{\lambda,m}(u,v,s)e^{-mu} \\ &= \left(\frac{\mathrm{e}^{-(\theta_1^*+\theta_2^*)\bar{s}/2}}{1-\frac{2\sqrt{\theta_1^*\theta_2^*}}{\theta_1^*+\theta_2^*}\left(1-\mathrm{e}^{-(\theta_1^*+\theta_2^*)\bar{s}/2}\right)}\right)^m.\end{aligned}$$

Here we wrote $\rho_{B,C}$ to denote it as the Bhattacharyya coefficient for Algorithm C (for the large T limit). Using this expression in (16) provides upper and lower bounds on $p_e = f_Z^C(\rho, \theta^*, m, \bar{s})$ in case $r = 2$.

For the sake of comparison, we note that the Bhattacharyya coefficient for the hypothesis testing problem based on $\check{Y}_T$ alone, i.e. Algorithm DT, is easily found to be:

$$\rho_{B,DT} = \left(\frac{\mathrm{e}^{-(\theta_1^*+\theta_2^*)\bar{s}/2}}{1-\sqrt{(1-\mathrm{e}^{-\theta_1^*\bar{s}})(1-\mathrm{e}^{-\theta_2^*\bar{s}})}}\right)^m.$$

V. Performance scaling for Algorithms C and DT

Consider the community recovery problem for m, r, ρ, and β fixed, and large T, such that the rate parameters $\theta_v^* : v \in [r]$ are distinct. Let δ be an arbitrarily small positive constant.

The problem of recovering ℓ_τ for some vertex τ with $\delta T \leqslant \tau \leqslant T$ from G_T using children (C) (respectively, degree thresholding (DT)) is asymptotically equivalent to the r-ary hypothesis testing problem for observation $Z_{[0,\ln(T/\tau)]}$ (respectively, $Z_{\ln(T/\tau)}$) with the same parameters m and $\theta_v^* : v \in [r]$. This leads to the following proposition, based on the results on coupling of Y, $\widetilde{Y}$ and $\breve{Y}$ and the connection of $\breve{Y}$ to Z.

Proposition 8. *(Performance scaling for Algorithms C and DT) (a) Let $p_{e,\tau,T}^{(C)}$ denote the probability of error for recovery of the label ℓ_τ using Algorithm C. For any $\delta \in (0,1)$, as $T \to \infty$,*

$$\max_{\tau:\delta T \leqslant \tau \leqslant T} |p_{e,\tau,T}^{(C)} - f_Z^C(\rho, \theta^*, m, \ln(T/\tau))| \to 0.$$

(b) Let $\widehat{p}_{e,T}^{(C)}$ denote the fraction of errors for recovery of the labels of G_T using Algorithm C for each vertex. Then,

$$\widehat{p}_{e,T}^{(C)} \stackrel{T\to\infty}{\longrightarrow} \int_0^1 f_Z^C(\rho, \theta^*, m, \ln(1/\delta))d\delta,$$

where the convergence is in probability.
(c) Parts (a) and (b) hold with C replaced by DT.

Proof. Observing the children of vertex τ in G_T is equivalent to observing $Y_{[\tau,T]}$. In view of Proposition 5, the binary hypothesis testing problem based on observation of $Y_{[\tau,T]}$ is asymptotically equivalent to the binary hypothesis testing problem based on observation of $\widetilde{Y}_{[\tau,T]}$ or on $\breve{Y}_{[\tau,T]}$. The upper bound on total variation distance is uniform for T/τ bounded. In particular, the minimum average probabilities of error for the problems become arbitrarily close as $T \to \infty$. To complete the proof of (a), we next compare the probability of recovery error based on observation of $\breve{Y}$ vs. observation based on the continuous time process Z.

The process $\breve{Y}_{[\tau,T]}$ is obtained by sampling the process $Z_{\ln(t/\tau)}$ at integer times $t \in [\tau, T]$. The mapping from Z to $\breve{Y}$ does not depend on the parameter ϑ, which could equal θ_v^* for any $v \in [r]$. In other words, observing $\breve{Y}_{[\tau,T]}$ is equivalent to observing Z_s for all $s \in [0, \ln(T/\tau)]$ such that τe^s is an integer, where Z has rate parameter θ_v^* under the hypothesis $\ell_\tau = v$. Thus, in the terminology of source coding, $\breve{Y}_{[\tau,T]}$ is a quantized version of $Z_{[0,\ln(T/\tau)]}$, with the quantizer becoming arbitrarily fine as $\tau \to \infty$. Therefore, the minimum probability of error for recovering ℓ_τ based on the children of τ in G_T, in the limit as $\tau, T \to \infty$ with $1 \leqslant T/\tau \leqslant 1/\delta$ is uniformly arbitrarily close to $f_Z^{(C)}(\rho, \theta^*, m, \ln(T/\tau))$. This completes the proof of part (a). Therefore, by

the bounded convergence theorem and the fact δ can be taken arbitrarily small, convergence of the expected fraction of label errors follows:

$$\mathbb{E}\left[\widehat{p}_{e,T}^{(C)}\right] \stackrel{T\to\infty}{\longrightarrow} \int_0^T f_Z^C(\rho,\theta^*,m,\ln(T/\tau))d\tau$$
$$= \int_0^1 f_Z^C(\rho,\theta^*,m,\ln(1/\delta))d\delta.$$

The last part of the proof is to show that the convergence is true not only in mean, but also in probability. That follows by the same method used for the alternative proof of Proposition 2, about the empirical degree distribution, given in Appendix D. The key step is a proof that the joint degree evolution processes $(\widetilde{Y}^j)$ for a finite number J of vertices (we only need to consider $J = 2$ here) are asymptotically independent in the sense that the total variation distance to a process with independent degree evolution converges to zero. That implies the error events for different labels are asymptotically uncorrelated, so convergence in probability to the mean follows by the Chebychev inequality. The same proof works for C replaced by DT. □

We conjecture that a result similar to Proposition 8 exists for label recovery using the message passing (MP) algorithm described in the next section.

The following proposition, proved in Appendix G, addresses the case that $\tau = o(T)$, including the possibility that τ is a constant. The estimation procedure is a modification of Algorithm C.

Proposition 9. *Suppose $T \to \infty$, with $\tau^o \geqslant 1$ being a function of T such that $\tau^o/T \to 0$. Then ℓ_{τ^o} can be recovered from knowledge of the children of τ^o in G_T with probability converging to one.*

Example 1 (Numerical comparison for a single community plus outliers)**.** *Numerical results are shown in Figure 1 for $m = 5, r = 2$, $\rho = (0.5, 0.5)$ and $\beta = \begin{pmatrix} b & 1 \\ 1 & 1 \end{pmatrix}$, with $b = 4$, corresponding to a graph with a single community of vertices and outlier vertices. For these parameters, $\eta^* = (0.622839, 0.377121)$ and $\theta^* = (0.598612, 0.337153)$. There is little difference between the error probabilities of Algorithms DT and C for $t/T \geqslant 10^{-1}$ but the difference is quite large for $t/T \leqslant 10^{-2}$. Thus, for the vertices arriving in the top one percent of time, Algorithm C, which uses the identity of children of a vertex, substantially outperforms Algorithm DT, which uses only the number of children. The Bhattacharyya upper bounds are not very tight but the ratio of upper bounds for DT and C is similar to the ratio f_Z^{DT}/f_Z^C. The derivative of $f_Z^{DT}(\rho,\theta^*,m,\ln(T/t))$ with respect to t/T has jump discontinuities at values of t/T such that*

the threshold in the MAP test changes from one integer to the next, which is noticeable in the plot for t/T close to 1, where the thresholds are small.

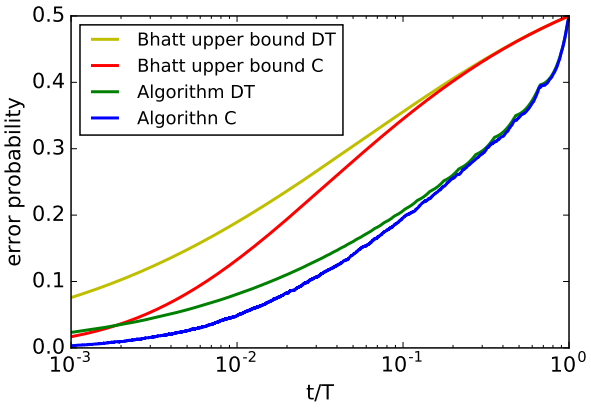

Fig. 1. Semilog plot of Bhattacharyya upper bounds $\frac{1}{2}\rho_{B,DT}$ and $\frac{1}{2}\rho_{B,C}$, and functions f_Z^{DT} and f_Z^C, for an example with a single community of vertices and outlier vertices.

VI. Joint estimation of labels of a fixed set of vertices

The idea of algorithm C is to estimate the label of a single vertex based on the likelihood of the observed set of children of the vertex, given the possible labels of the vertex. A natural extension, described in this section, is to jointly estimate the labels of a small fixed set of vertices from the joint likelihood of the children sets of the fixed set of vertices. Given a vector of possible labels of the vertices in the set, under the approximation $\eta_t \equiv \eta^*$ for all t, it is possible to compute the joint likelihood of the children sets for the vertices. Maximizing over all label vectors gives an approximate maximum likelihood estimate of the label vector. We use the following notation.

- $V \subset \mathbb{N}$, a finite set of vertices to be jointly classified
- $b \in [r]^V$, an assignment of labels for the vertices in V
- Y_t^τ is the degree of vertex τ in G_t.
- A_t^τ is the number of edges from vertex t to vertex τ
- $A_t^{V^c} = m - \sum_{\tau \in V} A_t^\tau$
- Attachment of vertices in $[\bar{t}+1, T]$ is observed, for some $\bar{t}$ and T with $\max\{\tau : \tau \in V\} \leqslant \bar{t} < T$.

Joint estimation algorithm: The joint estimation algorithm for estimating $(\ell_t : t \in V)$ is to calculate

$$\widehat{b}_{\text{ML}} = \arg\max_b \ln P\left(\left(A^\tau_{[\bar{t}+1,T]} : \tau \in V\right) \middle| b\right),$$

using the the following approximate expression for the log likelihoods:

$$\ln P\left(\left(A^\tau_{[\bar{t}+1,T]} : \tau \in V\right)|b\right) \approx const+$$
$$\sum_{t=\bar{t}}^{T-1} \ln \sum_{u\in[r]} \rho_u \left(\prod_{\tau\in V}\left(\frac{Y^\tau_t \theta^*_{u,b_\tau}}{mt}\right)^{A^\tau_{t+1}}\right)\left(1-\sum_{\tau'\in V}\frac{Y^{\tau'}_t \theta^*_{u,b_{\tau'}}}{mt}\right)^{A^{V^c}_{t+1}},$$

where $const$ represents a constant not depending on b (it is the sum of logarithms of multinomial coefficients) and the approximation stems entirely from approximating η_t by η^*. We could calculate either the approximate ML estimator, $\widehat{b}_{\text{ML}}$ by finding the arg max of the approximate log likelihood with respect to b, or $\widehat{b}_{\text{MAP}}$ in the same way but first adding the log of the prior probability of b. The complexity of the algorithm is $\Theta(r^n nT)$, which is feasible for small values of n.

Remark 3. *By Proposition 5, if the set V were to have a fixed number of vertices, but the vertices depended on T in such a way that $V \subset [\delta T, T]$ for some fixed $\delta > 0$, then the sets of children of the vertices would be asymptotically independent in the sense of total variation distance. Hence, in that limit, the joint estimation algorithm of this section would have no better performance than Algorithm C. That is why we envision using the joint estimation algorithm for a fixed set of vertices as $T \to \infty$.*

To see why joint estimation can help, consider two fixed vertices, τ and τ' with $\ell_\tau = v$ and $\ell_{\tau'} = v'$. By Proposition 6 we expect the degrees of the two vertices at time t to be on the order of $m(t/\tau)^{\theta^*_v}$ and $m(t/\tau')^{\theta^*_{v'}}$. Thus, if $m \geqslant 2$, the probability of the two vertices having a common child at time t to be proportional to the product of their degrees divided by t^2, or on the order of $(const)t^{\theta^*_v+\theta^*_{v'}-2}$. Thus, if $\theta^*_v + \theta^*_{v'} \geqslant 1$ we expect the number of common children of vertices τ and τ' in G_T to converge to infinity as $T \to \infty$, with a constant multiplier that can thus be consistently estimated as $T \to \infty$. In particular, if $\theta^*_v = \theta^*_{v'} \geqslant 0.5$, the rate of growth of joint children would typically depend on whether the two vertices are in the same community, providing consistent estimation whereas Algorithm C would fail.

VII. THE MESSAGE PASSING ALGORITHM

In this section, we describe how Alorithm C (the MAP rule given children) can be extended to a message passing algorithm. We describe the algorithm for the case of $r \geqslant 2$ possible labels for a general $r \times r$ matrix β with positive entries, and fixed $m \geqslant 1$. *Throughout the remainder of this section, let* (V, E) *be a fixed instance of the random graph,* (V_T, E_T)*, with known parameters* $m, r, \beta, \rho, t_o, G_{t_o}$*, and* T. The message passing algorithm is run on this graph, with the aim of calculating Λ_τ for $1 \leqslant \tau \leqslant T$, where for each τ, Λ_τ is a log-likelihood vector:

$$\Lambda_\tau(v) \triangleq \ln \mathbb{P}\left\{E_T = E | \ell_\tau = v\right\} + const, \quad v \in [r]$$

where $const$ represents a constant that can depend on the graph but does not depend on the vertex label v. Then we can calculate the maximum likelihood (ML) and maximum a posteriori probability (MAP) estimators of the label of a vertex τ by $\widehat{\ell}_{\tau,ML} = \arg\max_{v\in[r]} \Lambda_\tau(v)$ and $\widehat{\ell}_{\tau,MAP} = \arg\max_{v\in[r]} \rho_v \Lambda_\tau(v)$.

The messages in the message passing algorithm given below are also log likelihood vectors, so two values, $\nu, \nu' \in \mathbb{R}^r$, of such a message are considered to be equivalent if $\nu - \nu'$ is proportional to the all ones vector in $\mathbb{R}^r$. For example, given a log likelihood vector ν there is a *canonical equivalent log likelihood vector* ν' such that $\max_{u\in[r]} \nu'(u) = 0$, namely, ν' defined by $\nu'(u) = \nu(u) - \max_{u'\in[r]} \nu(u')$. This fact is useful for numerical computation; in our computer code we stored all log likelihood vectors in their equivalent canonical forms. A log likelihood vector is said to be a *null log likelihood vector* if it is a constant multiple of the all one vector. In other words, a null log likelihood vector is equivalent to the zero vector. In the special case $r = 2$, $\Lambda_\tau(1) - \Lambda_\tau(2)$ and $\nu(1) - \nu(2)$ represent log likelihood ratios, and the algorithm below can easily be restated using real valued messages that have interpretations as log likelihood ratios instead of using length two log likelihood vectors.

A complete specification of a message passing algorithm includes specification of the following elements:

1) initial messages
2) mappings from messages received at a vertex to messages sent by the vertex
3) timing of message passing and termination criterion
4) mappings from messages received at a vertex to the output log likelihood vector of the vertex

About element 3). A natural choice for the timing of message passing is synchronous. For synchronous timing, all messages to be sent along each edge of the graph G_T (excluding edges in the initial graph G_{t_o}) are computed. Based on those, log likelihood vectors are computed for each vertex and the next round of messages to be sent is computed. An alternative timing of messages is to alternate between updating only messages from children to parents and updating only messages from parents to children. For termination, we stopped the message passing when the sum of Euclidean norms of differences in the canonical log likelihood vectors was below a threshold.

In this section we specify the equations for elements 1), 2), and 4).

Given vertices τ and τ_0, we say τ is a child of τ_0, and τ_0 is a parent of τ, if $\tau \geqslant \max\{\tau_0, t_o\}+1$, and there is an edge from τ to τ_0. It is assumed that the known initial graph G_{t_o} is arbitrary and carries no information about vertex labels. Thus, for the inference problem at hand, the edges in G_{t_o} are not relevant beyond the degrees that they imply for the vertices in G_{t_o}. Let $\partial\tau$ denote the children of τ in G_T and $\wp\tau$ the parents of τ. So $\wp\tau = \varnothing$ if $\tau \leqslant t_o$ and $\partial\tau \subset \{t_o+1, \ldots, T\}$. Let $\nu_{\tau\to\tau_0}$ denote a message passed from child to parent, and $\mu_{\tau_0\to\tau}$ denote a message passed from parent to child.

Let $g^{cp} : \mathbb{R}^r \mapsto \mathbb{R}^r$ and $g^{pc} : \mathbb{R}^r \mapsto \mathbb{R}^r$ be defined as follows (here "cp" denotes child to parent, and "pc" denotes parent to child)

$$g^{cp}(\nu)(v) = \ln\left(\sum_{u\in[r]} \mathrm{e}^{\nu(u)}\rho_u\theta^*_{u,v}/\theta^*_v\right) \quad \text{for } \nu \in \mathbb{R}^r$$

$$g^{pc}(\mu)(v) = \ln\left(\sum_{v'\in[r]} \theta^*_{v,v'}\mathrm{e}^{\mu(v')}\rho_{v'}/\theta^*_{v'}\right) \quad \text{for } \mu \in \mathbb{R}^r,$$

where $\theta^*_{u,v}$ and θ^*_u are defined in Section II-B. For convenience, we repeat the expression in (12) for the approximate log likelood vector based on observation of children:

$$\lambda^C_\tau(v) = |\partial\tau| \ln\theta^*_v + \theta^*_v\left(d_0(\tau)\ln\frac{\tau\vee t_o}{T} + \sum_{t\in\partial\tau}\ln\frac{t}{T}\right), \tag{17}$$

where $\tau \vee t_o = \max\{\tau, t_o\}$ and $d_0(\tau)$ is the initial degree of vertex τ, defined to be the degree of τ in G_{t_o} if $\tau \leqslant t_o$ and $d_0(\tau) = m$ otherwise. The message passing equations are given as follows. See Appendix H for a derivation.

$$\nu_{\tau\to\tau_0} = \lambda_\tau^C + \sum_{t\in\partial\tau} \widetilde{\nu}_{t\to\tau} + \sum_{\tau_1\in\wp\tau\setminus\{\tau_0\}} \widetilde{\mu}_{\tau_1\to\tau} \tag{18}$$

$$\mu_{\tau_0\to\tau} = \lambda_{\tau_0}^C + \sum_{t\in\partial\tau_0\setminus\{\tau\}} \widetilde{\nu}_{t\to\tau_0} + \sum_{\tau_1\in\wp\tau_0} \widetilde{\mu}_{\tau_1\to\tau_0} \tag{19}$$

$$\widetilde{\nu}_{\tau\to\tau_0} = g^{cp}(\nu_{\tau\to\tau_0}) \tag{20}$$

$$\widetilde{\mu}_{\tau_0\to\tau} = g^{pc}(\mu_{\tau_0\to\tau}) \tag{21}$$

$$\Lambda_\tau = \lambda_\tau^C + \sum_{t\in\partial\tau} \widetilde{\nu}_{t\to\tau} + \sum_{\tau_0\in\wp\tau} \widetilde{\mu}_{\tau_0\to\tau}, \tag{22}$$

with the initial conditions:

$$\widetilde{\nu}_{\tau\to\tau_0} = 0 \qquad \widetilde{\mu}_{\tau_0\to\tau} = 0, \tag{23}$$

or equivalently

$$\nu_{\tau\to\tau_0} = \lambda_\tau^C \quad \mu_{\tau_0\to\tau} = \lambda_{\tau_0}^C. \tag{24}$$

In (18) - (22) messages with the letter ν are sent from child to parent, and messages with letter μ are sent from parent to child. The r coordinates of a message without a tilde represent likelihoods given possible labels of the sending vertex, while the r coordinates of a message with a tilde represent likelihoods given possible labels of the receiving vertex. The equations could be written entirely using only the ν's and μ's by applying (20) and (21) within (18) and (19). Or the equations could be written entirely using only the $\widetilde{\nu}$'s and $\widetilde{\mu}$'s by applying (18) and (19) within (20) and (21).

The edges in the initial graph G_{t_o} are not relevant in the algorithm beyond the fact they determine the degrees of the vertices in G_{t_o}. The message passing equations are written as if there are no parallel edges in (V, E). While the fraction of edges that are parallel to other edges will be small for large T, they are permitted. The convention used in the message passing algorithm is that $\partial\tau$ and $\wp\tau$ are to be considered as multisets, so that if a vertex appears with some multiplicity in one of those sets, then the corresponding term in the summations will be appearing the corresponding number of times.

Remark 4. *The* fitness only *case of the preferential attachment model with communities occurs if either of the following two equivalent conditions hold:*

1) β *has rank one*
2) $\theta^*_{u,v} = \theta^*_v$ *for all* u.

Since the distribution of the preferential attachment model with communities does not change if a row of β *is multiplied by a positive constant, for the fitness only case of the model it could be assumed that the rows of* β *are identical.*

In the fitness only case of the model, both g^{pc} *and* g^{cp} *map to null log likelihood vectors for any choice of their arguments, so all messages generated in the message passing algorithm are null log likelihood vectors. Consequently, if* β *has rank one then the message passing algorithm converges in one iteration and it coincides with algorithm* C.

VIII. Monte Carlo simulation results

The simulation results reported in this paper were computed for random graphs with $m = 5$, $\rho_u = 1/r$ for $u \in [r]$, and two vertices in the initial graph (i.e. $t_o = 2$) with degree $2m$ each. The specific choice of initial edges is not relevant, but there could for example be $2m$ parallel edges between the two initial vertices, or for example each of the two vertices could have m self loops.

A. Single community

The performance of the message passing algorithm is described for the case of a single community plus outliers, described in Example 1. Through numerical experimentation, we found the following timing of message passing works well. We take the initial values of all $\widetilde{\mu}$ and $\widetilde{\nu}$ messages to be zero. For the timing of message passing we run two phases. In the first phase the messages from children to parents (i.e. the $\widetilde{\nu}$'s) are repeatedly updated, while messages from parents to children are held fixed. In the second phase the messages $\widetilde{\nu}$ are held fixed and the messages from parents to children are repeatedly updated until the messages converge. In both phases the messages converge in a finite number of iterations. After both phases are completed, the (approximate) likelihood ratios are computed. Numerical results are shown in Figure 2. The message passing algorithm significantly outperforms the other two algorithms. Another version of algorithm with about the same performance is to use synchronous scheduling of all messages, while applying the message balancing method described in Section VIII-B.

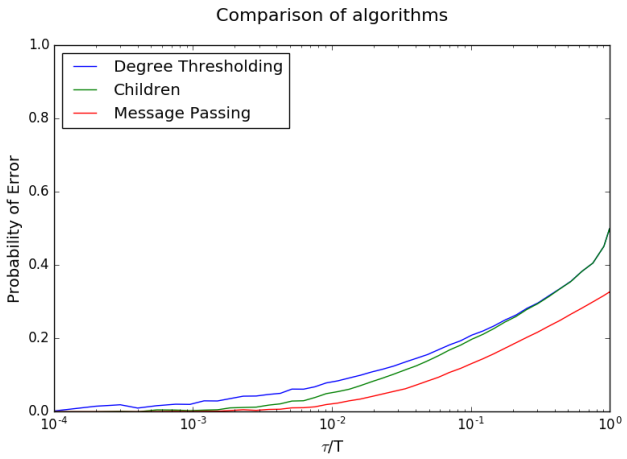

Fig. 2. Semilog plot of error probability vs. vertex index for algorithms DT,C, and MP for single community example with $m = 5$, $\rho = (0.5, 0.5)$, and $b = 4$. The average over 1000 runs of MP is shown.

B. Symmetric multiple community graphs

To model the situation that each vertex is in one of r communities with equal probability, with equal affinities within each community, let $\rho_v = 1/r$ for $v \in [r]$ and, for some $b > 1$,

$$\beta_{u,v} = \begin{cases} b & \text{if } u = v \\ 1 & \text{else} \end{cases}.$$

Then $\eta^* = \rho$, $\theta^*_{u,u} = \frac{br}{2(b+r-1)}$ and, for $u \neq v$, $\theta^*_{u,v} = \frac{r}{2(b+r-1)}$. Also, $\theta^*_v = 0.5$ for all v. Note that λ^C_τ is a null log likelihood vector for all τ. Up to equivalence of log likelihood vectors (i.e. ignoring addition of constant multiples of the all one vector) $g^{cp}(\cdot) = g^{pc}(\cdot) = g(\cdot)$, where

$$g(\nu)(v) = \ln\left(b\mathrm{e}^{\nu(v)} + \sum_{v' \in [r]\setminus\{v\}} \mathrm{e}^{\nu(v')}\right).$$

In the special case $r = 2$, the messages can be taken to be scalars representing log likelihood ratios, with g taking the form $g(\mu) \triangleq \ln \frac{b\mathrm{e}^{\mu}+1}{\mathrm{e}^{\mu}+b}$.

The functions g^{cp} and g^{pc} map null log likelihood vectors to null log likelihood vectors, so all messages equal to null log likelihood vectors is a fixed point of the message passing equations (18) - (21). Community detection is apparently rather difficult for this model in case $m = 1$ because G_T is a tree and for the symmetric two or more community graphs the local neighborhood of a vertex does not indicate which community the vertex is in, at least under the idealized assumption $\eta_t \equiv \eta$. We restrict attention to the case $m \geqslant 2$. In that case, we can

apply the joint estimation algorithm given in Section VI to identify the labels of a small number of vertices, which we call *seeds* to help initialize the message passing algorithm. Accordingly, for the message passing algorithm, we assume that the labels of the seed vertices are correctly revealed to the algorithm. Accordingly, the μ and ν messages sent by a seed vertex τ with $\ell_\tau = u$ would all be the same, and be given by:

$$\nu_{\tau\to\tau_0}(v) = \begin{cases} 0 & \text{if } v = u \\ -\infty & \text{else} \end{cases}$$

All other messages are initially set to zero. At every iteration, all the messages (both μ and ν) are updated synchronously.

One other technique, we call *message balancing*, was employed to get the algorithm to give good performance. Intuitively, the idea is to balance the total amount of negativity about each community within the messages. The following description of message balancing assumes the messages are stored in their equivalent canonical form, described near the beginning of Section VII. At the beginning of each iteration, the $\widetilde{\mu}$ messages are scaled by a positive vector f: $\widetilde{\mu}_v \to f_v\widetilde{\mu}_v$. The scale vector f is chosen for the iteration so that the sum of all the scaled $\widetilde{\mu}$ messages is a null log likelihood vector (i.e. multiple of all ones vector) and the sum of the messages is preserved. The $\widetilde{\mu}$ messages are similarly scaled. Empirically we found similar performance if only the $\widetilde{\mu}$ messages were scaled, or if only the $\widetilde{\mu}$ messages sent by seeds were scaled.

We first present numerical results for an example with two communities for $T = 10,000, m = 5$, and $b = 4$. We first describe the performance of the joint estimation algorithm for estimating the labels of the first ten vertices, taken to be seed vertices, and then describe the performance of the message passing algorithm assuming the seed vertices are correctly classified. The performance of the joint estimation algorithm is shown in Figure 3. Two different methods of determining which ones of vertices 2 through 10 are in the same community as vertex 1 were used. The first method, called "partial data" in the figure, estimates the label of each vertex τ with $2 \leqslant \tau \leqslant 10$ by jointly estimating labels for the set of two vertices $V = \{1, \tau\}$, while the second method, called "complete data" in the figure, is to jointly infer the labels of vertices in $V = \{1, 2, \ldots, 10\}$. The value $\bar{t} = 20$ was used. It was observed that the last term in the likelihood expression is sometimes negative (a result of the approximation $\eta_t \equiv \eta$) for some values of t and b. That was only observed in the simulations for some values of t with $t \leqslant 30$. If for some t a negative likelihood was observed for some b, then the likelihood term for that t was dropped for all

vectors b. The performance gives good evidence that for n fixed, the labels of the first n vertices can be inferred with error probability converging to zero as $T \to \infty$, for the symmetric two community model.

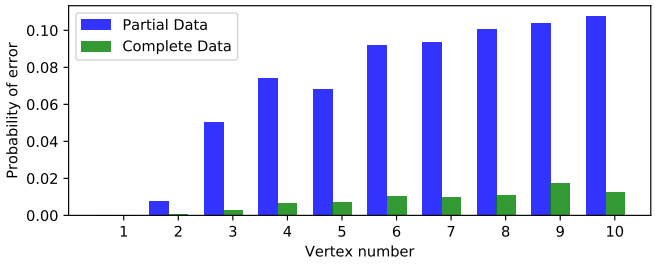

Fig. 3. Error probabilities for determining whether vertices 1 and τ are in the same community, for $2 \leqslant \tau \leqslant 10$, assuming symmetric two community model with parameters $b = 4$, $m = 5$, $T = 10,000$. Error probabilities are shown for (a) estimation based on joint likelihoods given labels for two vertices at a time (i.e. vertices 1 and τ with $2 \leqslant \tau \leqslant 10$), and (b) for estimation based on approximate maximum likelihood estimate of labels of vertices 1 through 10 simultaneously. Error probabilities are estimated by fraction of errors in 2048 simulations of graph, for estimation based on children with time of arrival t in the interval $[20, 10^4]$.

Next Figure 4 shows the performance of the message passing algorithm run on 100 graphs of size $T = 10,000$, with parameters $m = 5, b = 4$ with two communities with ten seed vertices. The message passing algorithm is run until the norm of the difference in the vector of log-likelihoods is less than 1. The probability of error curve plotted for each random graph is averaged over bins of width increasing with time. The ends of the bin intervals are chosen as a geometric progression with factor 1.2. Although there were only ten seed vertices, the algorithm nearly always correctly classified the first 100 vertices, and also most of the first 1000 vertices.

Performance of the message passing algorithm for four communities with 20 seed vertices is shown in Figure 5. The result of running on 100 sample graphs is shown. The algorithm had poor performance for one sample labeled graph, for which one of the communities was not represented among the seeds. In other simulations we have seen the algorithm fail occasionally even if all communities are represented among the seeds.

C. Three communities with symmetry between two of them

Consider three communities 1,2,3 such that each vertex is equally likely to be in any of the three communities. Vertices in community 1 have a growth rate distinct from the growth rates

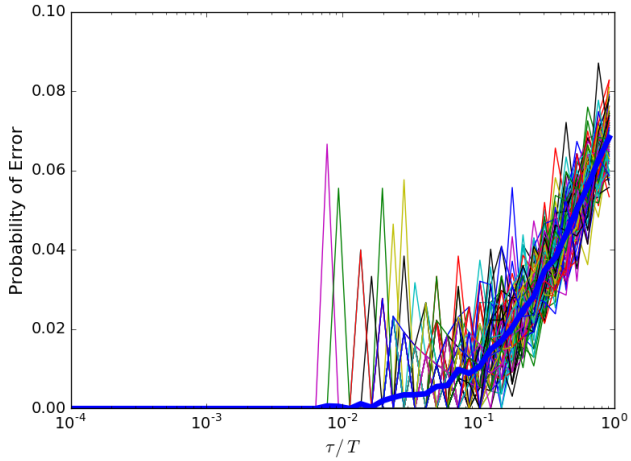

Fig. 4. Semilog plot of error probability vs. vertex index for algorithm MP for symmetric two ($r = 2$) community graphs with $m = 5$ and $b = 4$. The algorithm was given labels of the first ten vertices and message balancing was used. Smoothed results for 100 graphs are shown, with the average of them represented by the thicker blue curve.

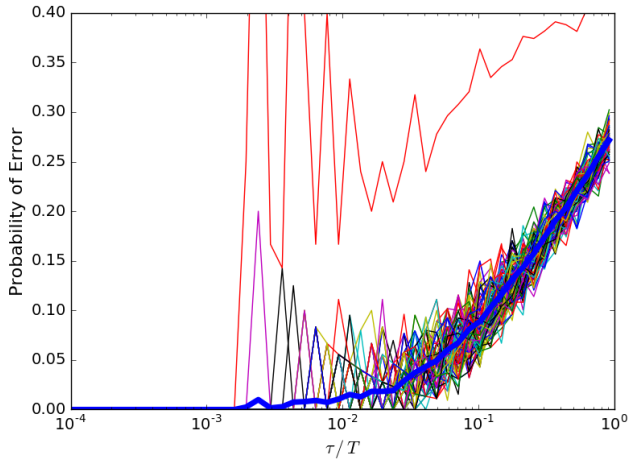

Fig. 5. Semilog plot of error probability vs. vertex index for algorithm MP for symmetric four ($r = 4$) community graphs with $m = 5$ and $b = 4$. The performance for MP run on 100 independently generated graphs is shown. The algorithm was given labels of the first twenty vertices and message balancing was used. Smoothed results for 100 graphs are shown, with the average of them represented by the thicker blue curve.

of the other two communities, and the other two communities are statistically identical. We again begin with the joint estimation algorithm, because identifying seed vertices can help the message passing algorithm distinguish vertices in the two statistically identical communities. To

display the performance of the joint estimation algorithm we need to adjust for the fact that the assignment of labels 2 vs. 3 to the two symmetric communities is arbitrary. Thus, before computing errors, we see whether swapping the 2's and 3's of the output label vector reduces the number of errors. If yes, the 2's and 3's of the output vector are swapped. If there is a tie, with probability 0.5, the 2's and 3's are all swapped. Then, for each seed vertex, we say a *big error* is made if the true label is 1 and the estimate is not 1, or vice versa. We say a *small error* is made if both the true label and estimated label are in $\{2, 3\}$ but they are unequal. The event that the label of a seed vertex is in error is the disjoint union of a big error event and small error event. The message passing algorithm was run using synchronous message timing with 15 seed vertices and message balancing.

Two different β matrices were tried, which we list with their corresponding vectors (θ_v^*)

$$\beta^I = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix} \qquad \beta^{II} = \begin{pmatrix} 4 & 1 & 1 \\ 2 & 4 & 1 \\ 2 & 1 & 4 \end{pmatrix}$$

$$(\theta^*)^I = (0.420, 0.532, 0.532]) \quad (\theta^*)^{II} = (0.590, 0.438, 0.438)$$

For version I of the model, Figure 6 displays the performance of the joint estimation algorithm and Figure 7 displays the performance of the message passing algorithm for 15 seed vertices. Proposition 6 implies that as $T \to \infty$ the probability of big errors converges to zero. The probability of small errors is apparently small for this model and algorithm.

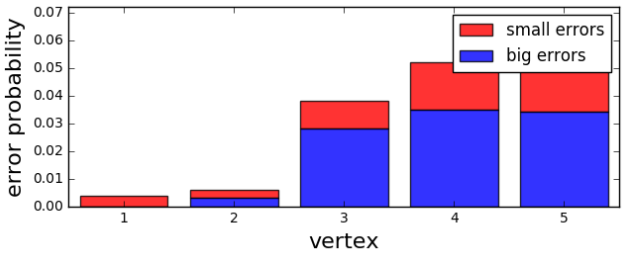

Fig. 6. Big errors and small errors for joint estimation of the labels of first five vertices for version I of the three communities example, estimated using 1000 sample graphs. At least one label is incorrect in 0.139 fraction of graphs.

For version II of the model, Figure 8 displays the performance of the joint estimation algorithm and Figure 9 displays the performance of the message passing algorithm for 15 seed vertices.

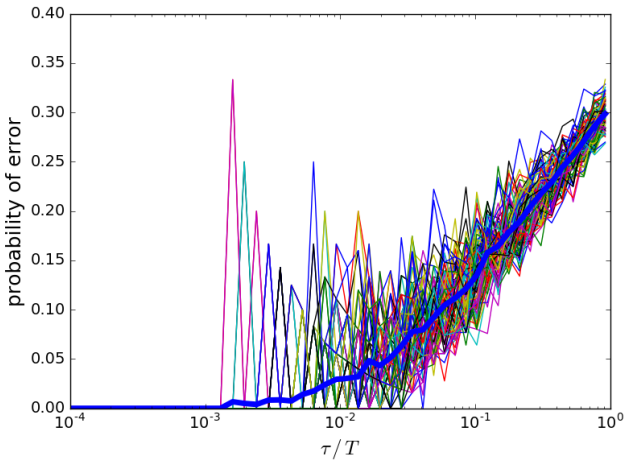

Fig. 7. Error probabilities by vertex for version I of the three communities example, for message passing with 15 seed vertices. Smoothed results for 100 graphs are shown, with the average of them represented by the thicker blue curve.

There are many more small errors for version II of the model than for version I, which is explained by the fact that for version II, the two equal sized communities that can't be distinguished by growth rates alone (because $\theta_2^* = \theta_3^*$) have much smaller degrees than vertices in the two equal sized communities of version I. In fact, we conjecture that the probability of small errors does not converge to zero for the joint estimation algorithm for version II. The reason is that the mean number of common children of two vertices that have labels in $\{2, 3\}$ is stochastically bounded above as $T \to \infty$, because $\theta_v^* + \theta_{v'}^* < 1$ for $v, v' \in \{2, 3\}$. See Remark 3.

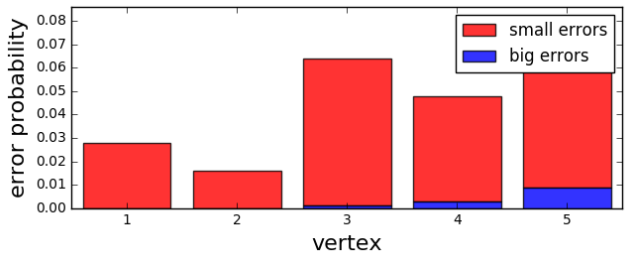

Fig. 8. Big errors and small errors for joint estimation of the labels of first five vertices for version II of the three communities example , estimated using 1000 sample graphs.

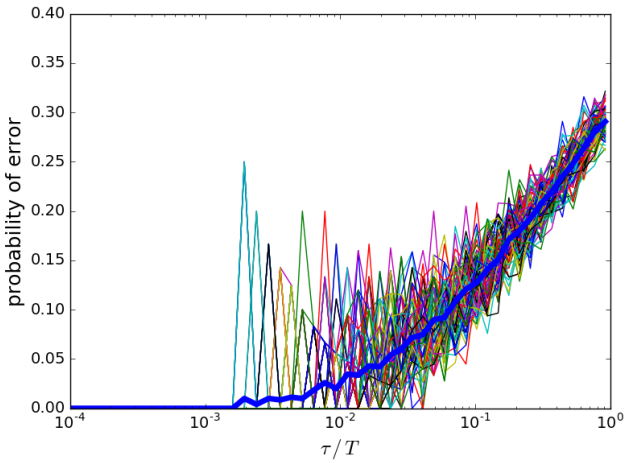

Fig. 9. Error probabilities by vertex for version II of the three communities example, for message passing with 15 seed vertices.

IX. CONCLUSION

The message passing algorithm, together with seeding by the joint inference algorithm and balancing method, appear to work well in Monte Carlo simulations. The use of seeds takes advantage of the large degrees of a few vertices. The performance of the joint inference algorithm is related to the large time degree evolution of one or more fixed vertices τ such that $T/\tau \to \infty$ as $T \to \infty$, whereas the derivation of the message passing algorithm is based on the joint degree evolution for one or more vertices τ such that $\tau \to \infty$ and T/τ remains bounded. As version II of the three community example points out, it may not always be possible to consistently recover a fixed set of vertex labels as $T \to \infty$, while it is possible if the parameters $\theta^*_v : v \in [r]$ are distinct.

APPENDIX A
PROOF OF PROPOSITION 1

Simple algebra yields

$$\eta_{t+1} - \eta_t = \frac{C_{t+1} - C_t - 2m\eta_t}{2m(t+1)}. \tag{25}$$

The conditional distribution of $C_{t+1} - C_t$ given C_t and given $\ell_{t+1} = u$ can be represented using a random variable with a multinomial distribution as

$$C_{t+1} - C_t \stackrel{d.}{=} me_u + \text{multinom}\left(m, \left(\frac{\beta_{uv}\eta_{tv}}{\sum_{v'}\beta_{uv'}\eta_{tv'}} : v \in [r]\right)\right),$$

where e_u is the unit length r vector with u^{th} coordinate equal to one. Therefore,

$$E[C_{t+1,v} - C_{t,v}|C_t] = m\rho_v + \sum_u m\rho_u \left(\frac{\beta_{uv}\eta_{tv}}{\sum_{v'}\beta_{uv'}\eta_{tv'}}\right) \tag{26}$$

Combining with (25) yields that

$$E[\eta_{t+1} - \eta_t|C_t] = \frac{1}{2(t+1)}h(\eta_t). \tag{27}$$

This gives the representation

$$\eta_{t+1} = \eta_t + \frac{1}{2(t+1)}\left[h(\eta_t) + M_t\right] \tag{28}$$

where

$$M_t = C_{t+1} - C_t - E[C_{t+1} - C_t|C_t]. \tag{29}$$

Note that M is a bounded martingale difference sequence; $\mathbb{P}\{\|M_t\|_1 \leqslant 4m\} = 1$ for all t. Also, the Jacobian matrix of h is uniformly bounded over the domain of probability vectors so h is Lipschitz continuous. In view of (28) and these properties, the theory of stochastic approximation implies the possible limit points of η_t is the set of stable equilibrium points of the ode $\dot{\eta} = h(\eta)$ [21, Chapter 2, Theorem 2] .

Since $\sum_v h_v(\eta) \equiv 0$, the ode $\dot{\eta} = h(\eta)$ can be restricted to the space of probability vectors. A Lyapunov function is used in [6] to show that the ode $\dot{\eta} = h(\eta)$ restricted to the space of probability vectors has a unique globally stable equilibrium point, which we denote by η^*.

APPENDIX B
PROOF OF PROPOSITION 3

Remark 5. *(i) We shall use extensively the connection between total variation distance and coupling. Given two discrete probability distributions a and b on the same discrete set, the total variation distance between a and b is defined by $d_{TV}(a,b) = \frac{1}{2}\sum_i |a_i - b_i|$. If A and B are random variables, not necessarily on the same probability space, we write $d_{TV}(A,B)$ to represent $d_{TV}(\mathcal{L}(A),\mathcal{L}(B))$, which is the total variation distance between the probability distributions of A and B. Clearly d_{TV} is a distance metric; in particular it satisfies the triangle inequality. An operational meaning is $d_{TV}(a,b) = \min \mathbb{P}\{A \neq B\}$, where the minimum is taken over all pairs of jointly distributed random variables (A,B) such that A has distribution a and B has distribution b. In other words, $d_{TV}(a,b)$ is the minimum failure probability when one attempts to couple a random variable with distribution a to a random variable with distribution b.*

(ii) The distance $d_{TV}(a,b)$ can be expressed as

$$d_{TV}(a,b) = \sum_i (b_i - a_i)_+ \tag{30}$$

Expression (30) *is especially useful if $b_i > a_i$ for only a small set of indices i. For example, if a and b are distributions on $\mathbb{Z}_+$ such that b is a Bernoulli probability distribution and $a_0 \geqslant b_0$, then $d_{TV}(a,b) = b_1 - a_1$.*

The proofs of (6) and (7) are similar. Since the proof of (6) depends slightly on (7), we prove (7) first.

Fix $t \geqslant \tau$ and $y \geqslant m$. The conditional distribution of the increment $\breve{Y}_{t+1} - \breve{Y}_t$ of the Markov process $\breve{Y}$ given $\breve{Y}_t = y$ can be identified as follows:

$$\begin{aligned}
&\mathcal{L}\left(\breve{Y}_{t+1} - \breve{Y}_t \middle| \breve{Y}_t = y\right) \\
&= \mathcal{L}\left(Z_{\ln((t+1)/\tau)} - Z_{\ln(t/\tau)} \middle| Z_{\ln t} = y\right) \\
&= \mathcal{L}\left(Z_{\ln\left(1+\frac{1}{t}\right)} \middle| Z_0 = y\right) \\
&= \mathcal{L}\left(\mathsf{negbinom}\left(y, \left(1+\frac{1}{t}\right)^{-\vartheta}\right) - y\right)
\end{aligned}$$

Hence, the following lemma is relevant, were ϵ represents $\frac{1}{t}$.

Lemma 1. *Let y be a positive integer and $\vartheta, \epsilon > 0$. Then*

$$\begin{aligned} &d_{TV}\left(\mathsf{negbinom}\left(y,(1+\epsilon)^{-\vartheta}\right)-y,\mathsf{Ber}(\vartheta y\epsilon)\right)\\ &\leqslant \frac{\epsilon^2}{2}\left(\vartheta y+\vartheta^2(2y+1)y\right). \end{aligned} \tag{31}$$

Proof. The shifted negative binomial distribution assigns more probability mass to 0 than the Bernoulli distribution:

$$\mathbb{P}\left\{\mathsf{negbinom}\left(y,(1+\epsilon)^{-\vartheta}\right)-y=0\right\}\geqslant 1-\vartheta y\epsilon,$$

or equivalently,

$$(1+\epsilon)^{-\vartheta y}\geqslant 1-\vartheta y\epsilon,$$

as is readily proved by considering the derivative of each side with respect to ϵ for $\epsilon > 0$. Therefore, by Remark 5(ii), the total variation distance to be bounded is given by the difference in probability mass at 1 for the two distributions. In other words, if δ denotes the variational distance on the lefthand side of (31), then

$$\delta=\vartheta y\epsilon-y(1+\epsilon)^{-\vartheta y}\left(1-(1+\epsilon)^{-\vartheta}\right).$$

Note that $\delta = 0$ for $\epsilon = 0$. Dividing through by y and differentiating with respect to ϵ we find

$$\frac{d\delta}{yd\epsilon}=\vartheta+\vartheta y(1+\epsilon)^{-\vartheta y-1}-\vartheta(y+1)(1+\epsilon)^{-\vartheta(y+1)-1},$$

and in particular the derivative at $\epsilon = 0$ is also zero. Differentiating again yields:

$$\begin{aligned} \frac{d^2\delta}{y(d\epsilon)^2} &= -\vartheta y(\vartheta y+1)(1+\epsilon)^{-\vartheta y-2}\\ &\qquad +\vartheta(y+1)(\vartheta(y+1)+1)(1+\epsilon)^{-\vartheta(y+1)-2}\\ &\overset{(a)}{\leqslant} -\vartheta y(\vartheta y+1)+\vartheta(y+1)(\vartheta(y+1)+1)\\ &= \vartheta+\vartheta^2(2y+1) \end{aligned}$$

where to get inequality (a) we first multiply the lefthand side by $(1+\epsilon)^{\theta y+2}$ (thus increasing it) and then multiplying the second term on the lefthand side by $(1+\epsilon)^{2\theta}$, thus increasing the positive term further. The lemma follows by twice integrating with respect to ϵ. □

Proof of (7). Let n be a positive integer with $n \geqslant m$. We appeal to Lemma 1 to show that $\widetilde{Y}_{[\tau,T]}$ and $\breve{Y}_{[\tau,T]}$ can be coupled (i.e. constructed on the same probability space) such that the probability of coupling failure before both processes reach state n is bounded as follows:

$$\begin{aligned}\mathbb{P}\left\{\widetilde{Y}_{[\tau,T]} \wedge n \neq \breve{Y}_{[\tau,T]} \wedge n\right\} &\leqslant \sum_{t=\tau}^{T} \frac{\vartheta n + \vartheta^2(2n+1)n}{2t^2} \\ &\leqslant \frac{\vartheta n + \vartheta^2(2n+1)n}{\tau - 1}.\end{aligned}$$

The construction is done sequentially in time, starting with the process $\widetilde{Y}$, letting $\breve{Y}_\tau = m$, and enlarging the probability space $\widetilde{Y}$ is defined on in order to construct $\breve{Y}_t$ for $\tau+1 \leqslant t \leqslant T$ on the same probability space. For each time t in the range $\tau \leqslant t \leqslant T-1$, once the random variable $\breve{Y}_t$ has been constructed, if the coupling has been successful so far (i.e. $\widetilde{Y}_{[\tau,t]} = \breve{Y}_{[\tau,t]}$) and if $\breve{Y}_t \leqslant n-1$, we appeal to Lemma 1 with $y = \breve{Y}_t$ to show that the coupling can be continued to work at time $t+1$, with coupling error bounded above by Lemma 1.

For this same pair of processes, it follows that

$$\mathbb{P}\left\{\widetilde{Y}_{[\tau,T]} \neq \breve{Y}_{[\tau,T]}\right\} \leqslant \frac{\vartheta n + \vartheta^2(2n+1)n}{\tau - 1} + \mathbb{P}\left\{\breve{Y}_T \geqslant n\right\}$$

Since $\breve{Y}_T = Z_{\ln(T/\tau)}$, the distribution of $\breve{Y}_T$ is $\mathsf{negbinom}\left(m, \left(\frac{\tau}{T}\right)^\vartheta\right)$, and the set of such distributions is tight under the limiting regime of the proposition. In other words, $\lim_{n\to\infty} \limsup_{\tau,T\to\infty} \mathbb{P}\left\{\breve{Y}_T \geqslant n\right\} = 0$ under the assumption T/τ is bounded. The statement (7) follows. □

The proof of (6), given next, is based on the following lemma.

Lemma 2. *Given a positive integer* y, $(\theta_{u,v}) \in \mathbb{R}^{r\times r}_{>0}$, ρ, *and* $(\theta^*_v : v \in [r]) \in \mathbb{R}^r_{>0}$, *let* $\theta_v = \sum_u \rho_u \theta_{u,v}$. *Suppose* $t \geqslant 1$ *and* $v \in [r]$ *such that* $\frac{\theta_{u,v} y}{mt} \leqslant 1$ *for all* u, $\frac{\theta_v y}{t} \leqslant 1$, *and* $\frac{\theta^*_v y}{t} \leqslant 1$. *Then*

$$\begin{aligned}&d_{TV}\left(\sum_u \rho_u \mathsf{binom}\left(m, \frac{\theta_{u,v} y}{mt}\right), \mathsf{Ber}\left(\frac{\theta^*_v y}{t}\right)\right) \\ &\qquad\leqslant \frac{\theta^2_{\max} y^2}{t^2} + \frac{|\theta_v - \theta^*_v| y}{t},\end{aligned}$$

where $\theta_{\max} = \max_{u,v} \theta_{u,v}$.

Proof. By the triangle inequality,

$$
\begin{aligned}
&d_{TV}\left(\sum_u \rho_u \mathsf{binom}\left(m, \frac{\theta_{u,v} y}{mt}\right), \mathsf{Ber}\left(\frac{\theta^*_v y}{t}\right)\right) \\
&\leqslant d_{TV}\left(\sum_u \rho_u \mathsf{binom}\left(m, \frac{\theta_{u,v} y}{mt}\right), \mathsf{Ber}\left(\frac{\theta_v y}{t}\right)\right) \qquad (32) \\
&\qquad + d_{TV}\left(\mathsf{Ber}\left(\frac{\theta_v y}{t}\right), \mathsf{Ber}\left(\frac{\theta^*_v y}{t}\right)\right) \qquad (33)
\end{aligned}
$$

To bound the term on line (32), we appeal to Remark 5(ii). Note that the probability masses at 0 for the two distributions inside d_{TV} are ordered as:

$$
\begin{aligned}
\sum_u \rho_u \mathsf{binom}\left(m, \frac{\theta_{u,v} y}{mt}\right)\Big|_0 &= \sum_u \rho_u \left(1 - \frac{\theta_{u,v} y}{mt}\right)^m \\
\geqslant \sum_u \rho_u \left(1 - \frac{\theta_{u,v} y}{t}\right) &= 1 - \frac{\theta_v y}{t} = \mathsf{Ber}\left(\frac{\theta_v y}{t}\right)\Big|_0
\end{aligned}
$$

So the term in (32) is the difference of the probability masses at 1:

$$
\begin{aligned}
&\frac{\theta_v y}{t} - \sum_u \rho_u \mathsf{binom}\left(m, \frac{\theta_{u,v} y}{mt}\right)\Big|_1 \\
&= \frac{\theta_v y}{t} - \sum_u \rho_u \frac{\theta_{u,v} y}{t}\left(1 - \frac{\theta_{u,v} y}{mt}\right)^{m-1} \\
&= \sum_u \rho_u \frac{\theta_{u,v} y}{t}\left(1 - \left(1 - \frac{\theta_{u,v} y}{mt}\right)^{m-1}\right) \\
&\leqslant \sum_u \rho_u \left(\frac{\theta_{u,v} y}{t}\right)^2 \leqslant \frac{\theta^2_{\max} y^2}{t^2}
\end{aligned}
$$

The term on line (33) is equal to $\frac{|\theta_v - \theta^*_v| y}{t}$. □

Proof of (6). Let n be a positive integer with $n \geqslant m$. Since the entries of β are assumed to be strictly positive, there is a finite value $\theta_{\max}$ such that $\theta_{u,v,t} \leqslant \theta_{\max}$ for all t. Given $\epsilon > 0$ let F be the event defined by $F = \{|\theta_{v,t} - \theta^*_v| > \epsilon \text{ for some } t \geqslant \tau\}$. We appeal to Lemma 2 to show that $Y_{[\tau,T]}$ and $\widetilde{Y}_{[\tau,T]}$ can be coupled (i.e. constructed on the same probability space) such that the probability of coupling failure before both processes reach state n is bounded as follows:

$$
\begin{aligned}
&\mathbb{P}\left\{Y_{[\tau,T]} \wedge n \neq \widetilde{Y}_{[\tau,T]} \wedge n\right\} \\
&\leqslant \mathbb{P}\{F\} + \sum_{t=\tau}^{T-1}\left(\frac{\theta^2_{\max} n^2}{t^2} + \frac{|\theta_{v,t} - \theta^*_v| n}{t}\right) \\
&\leqslant \mathbb{P}(F) + \frac{\theta^2_{\max} n^2}{\tau - 1} + \epsilon n \ln \frac{T}{\tau - 1}
\end{aligned}
$$

For this same pair of processes, it follows that

$$\begin{aligned}&\mathbb{P}\left\{Y_{[\tau,T]} \neq \widetilde{Y}_{[\tau,T]}\right\}\\ &\leqslant \mathbb{P}\{F\} + \frac{\vartheta n + \vartheta^2(2n+1)n}{\tau - 1} + \mathbb{P}\left\{\widetilde{Y}_T \geqslant n\right\}\end{aligned}$$

By Proposition 2 (almost sure convergence of $\eta_t \to \eta^*$) $\mathbb{P}(F) \to 0$ as $\tau \to \infty$. By (7), already proved,

$$\lim_{n\to\infty} \limsup_{\tau,T\to\infty} \left| \mathbb{P}\left\{\widetilde{Y}_T \geqslant n\right\} - \mathbb{P}\left\{\check{Y}_T \geqslant n\right\} \right| = 0,$$

so, just as for $\check{Y}_T$, the set of distributions of $\widetilde{Y}_T$ is tight under the limiting regime of the proposition. In other words, $\lim_{n\to\infty} \limsup_{\tau,T\to\infty} \mathbb{P}\left\{\widetilde{Y}_T \geqslant n\right\} = 0$ under the assumption T/τ is bounded. The statement (6) follows. □

APPENDIX C
PROOF OF PROPOSITION 5

The proof is similar to the proof of Proposition 3. Before proving the proposition we introduce some notation and present a lemma that is used to bound the coupling failure probability at a given step in the construction. A *subprobability vector* for a set $[d] = \{1, \ldots, d\}$ is a d-tuple of the form $\underline{a} = (a_i : i \in [d])$ such that $a_i \geqslant 0$ for $i \in [d]$ and $\sum_{i\in[d]} a_i \leqslant 1$. Let r and J be positive integers. Suppose ρ is a probability distribution on $[J]$. Suppose $\underline{p}, \underline{p}'$, and $\underline{q}_{u,\cdot}$ for all $u \in [r]$ are subprobability vectors for $[J]$.

- Let $\mathsf{sel}(\underline{p})$ represent the *selector* distribution on $\mathbb{Z}_+^J$ with probability mass p_j on the vector e_j, and probability mass $1 - \sum_j p_j$ on the zero vector.
- Let $\mathsf{sel}^{*m}(\underline{p})$ denote the distribution of the sum of m independent random vectors, each with the distribution $\mathsf{sel}(\underline{p})$. In other words, $\mathsf{sel}^{*m}(\underline{p})$ is the m-fold convolution of $\mathsf{sel}(\underline{p})$.
- Let $\sum_u \rho_u \mathsf{sel}^{*m}(\underline{q}_{u,\cdot})$ denote the distribution that is a mixture of the distributions $\mathsf{sel}^{*m}(\underline{q}_{u,\cdot})$ as u varies with selection probability distribution ρ.
- Let $\otimes_{j=1}^J \mathsf{Ber}(p_j)$ denote the distribution of a random J vector with independent coordinates, with coordinate j having distribution $\mathsf{Ber}(p_j)$.

Lemma 3. *Suppose ρ is a probability distribution on $[J]$. Suppose $\underline{p}, \underline{p}'$, and $\underline{q}_{u,\cdot}$ for all $u \in [r]$ are subprobability vectors for $[J]$.*

$$d_{TV}\left(\otimes_{j=1}^{J}\mathsf{Ber}(p_j), \mathsf{sel}(\underline{p})\right) \leqslant \left(\sum_{j\in[J]} p_j\right)^2 \tag{34}$$

$$d_{TV}(\mathsf{sel}(\underline{p}), \mathsf{sel}(\underline{p}')) \leqslant \sum_{j\in[J]} |p_j - p'_j| \tag{35}$$

$$d_{TV}\left(\mathsf{sel}\left(\sum_u \rho_u \underline{q}_{u,\cdot}\right), \sum_{u\in[r]} \rho_u \mathsf{sel}^{*m}\left(\frac{1}{m}\underline{q}_{u,\cdot}\right)\right)$$
$$\leqslant \sum_{u\in[r]} \rho_u \left(\sum_{j\in[J]} q_{u,j}\right)^2 \tag{36}$$

Proof. Inequality (35) follows easily from the definitions. The proofs of the other two inequalities rely on Remark 5(ii). Note that the distribution $\mathsf{sel}(\underline{p})$ is supported on $J+1$ points in $\mathbb{Z}^J$, namely, $\underline{0}, e_1, \ldots, e_J$. Also,

$$\otimes_{j=1}^{J}\mathsf{Ber}(p_j)\Big|_{\underline{0}} = \prod_{j\in[J]}(1-p_j) \geqslant 1 - \sum_{j\in[J]} p_j = \mathsf{sel}(p)\Big|_{\underline{0}}.$$

Thus, by Remark 5(ii),

$$d_{TV}\left(\otimes_{j=1}^{J}\mathsf{Ber}(p_j), \mathsf{sel}(\underline{p})\right) = \sum_{j\in[J]} p_j \left[1 - \prod_{j'\in[J], j'\neq j}(1-p_{j'})\right]$$
$$\leqslant \sum_{j\in[J]} p_j \sum_{j'\in[J], j'\neq j} p_{j'} \leqslant \left(\sum_{j\in[J]} p_j\right)^2,$$

which establishes (34). The proof of (36), given next, is similar. The probability masses the two distributions on the lefthand side of (36) place at zero is ordered as follows:

$$\sum_{u\in[r]} \rho_u \left(1 - \frac{1}{m}\sum_{j\in[J]} q_{u,j}\right)^m \geqslant 1 - m \sum_{u\in[r]} \rho_u \sum_{j\in[J]} q_{u,j}.$$

Therefore, by Remark 5(ii),

$$d_{TV}\left(\sum_{u\in[r]} \rho_u \mathsf{sel}^{*m}\left(\frac{1}{m}\underline{q}_{u,\cdot}\right), \mathsf{sel}\left(\sum_u \rho_u \underline{q}_{u,\cdot}\right)\right)$$
$$= \sum_{j\in[J]}\sum_{u\in[r]} \rho_u q_{u,j}\left[1 - \left(1 - \frac{1}{m}\sum_{j'\in[J]:j\neq j} q_{u,j'}\right)^{m-1}\right]$$

$$\leqslant \sum_{j\in[J]}\sum_{u\in[r]} \rho_u q_{u,j}\left[\sum_{j'\in[J]} q_{u,j'}\right]$$
$$\leqslant \sum_{u\in[r]} \rho_u \left(\sum_{j\in[J]} q_{u,j}\right)^2,$$

which establishes (36). □

Lemma 4. *Suppose the conditions of Lemma 3 hold, and, in addition,* $\sum_u \rho_u q_{u,j} = p'_j$ *for all* $j \in [J]$. *Then*

$$d_{TV}\left(\otimes_{j=1}^{J}\mathsf{Ber}(p_j), \sum_{u\in[r]} \rho_u \mathsf{sel}^{*m}\left(\frac{1}{m}\underline{q}_{u,\cdot}\right)\right) \tag{37}$$
$$\leqslant \left(\sum_{j\in[J]} p_j\right)^2 + \sum_{j\in[J]} |p_j - p'_j| + \sum_{u\in[r]} \rho_u \left(\sum_{j\in[J]} q_{u,j}\right)^2.$$

Proof. The lefthand side of (37) is less than or equal to the sum of the lefthand sides of (34)-(36) by the triangle inequality for d_{TV}. The righthand side of (37) is the sum of the righthand sides of (34)-(36). So the lemma follows from Lemma 3. □

Proof of Proposition 5. By the tightness of $\mathcal{L}(Y_T^j|\ell_\tau = v_j)$ for each j in the limit regime of the proposition, implied by Proposition 3 and the known distribution of $\breve{Y}_T$, it suffices to prove the proposition with $Y_{[1,T]}^{[J]}$ and $\widetilde{Y}_{[1,T]}^{[J]}$ each replaced by versions of the same processes that are stopped when the sum of the vertex degrees (i,e. the coordinates) of the process first becomes greater than or equal to a fixed, positive integer n. So let n be a fixed, positive integer. Let the process $Y_{[1,T]}^{[J]}$ be given, defined on some probability space. By enlarging the probability space, we can construct $\widetilde{Y}_{[1,T]}^{[J]}$ on the same space, and the total variation distance is upper bounded by the probability the processes are different from each other at some time before the sum of coordinates is greater than n or before time $T+1$. The construction is done sequentially in time. For each time t in the range $1 \leqslant t \leqslant T-1$, once the random variable $\widetilde{Y}_t$ is constructed, we appeal to Lemma 3 with $q_{u,j}$ in the lemma given by $\frac{\theta_{u,v_j,t}\widetilde{Y}_t^j}{t}$. Since the entries of β are assumed to be strictly positive, there is a finite value $\theta_{\max}$ such that $\theta_{u,v,t} \leqslant \theta_{\max}$ for all t. Given $\epsilon > 0$ let F be the event defined by $F = \{|\theta_{v,t} - \theta_v^*| > \epsilon \text{ for some } t \geqslant \tau\}$.

Fix t with $\tau_1 \leqslant t \leqslant T$. Let $A_t = \{j : \tau_j \leqslant t\}$, so that A_t is the set of vertices in $[J]$ that are active at time t. For $j \notin A_t$ the values of Y_{t+1}^j and $\widetilde{Y}_{t+1}^j$ are deterministic and they are equal.

If $t+1=\tau_j$ for some j we call t an *exceptional* time. Exceptional times must be handled differently than other times because for such a time, conditioning on $(\ell_{\tau_j}=v_j)$, or, equivalently, on $(\ell_{t+1}=v_j)$, effects the distribution of $(Y_{t+1}^{j'}-Y_t^{j'}: j'\in A_t)$, and Lemma 4 doesn't apply. The effect of such exceptional times on coupling error can be bounded as follows. First, there are less than or equal to J exceptional times. Secondly, for such an exceptional time t,

$$\mathbb{P}\left\{Y_{t+1}^{j'}-Y_t^{j'}\neq 0 \text{ for some } j'\in A_t|\ell_{\tau^j}=v_j\right\}\leqslant\frac{n\theta_{\max}}{t}$$

and also

$$\mathbb{P}\left\{\widetilde{Y}_{t+1}^{j'}-\widetilde{Y}_t^{j'}\neq 0 \text{ for some } j'\in A_t\right\}\leqslant\frac{n\theta_{\max}}{t}$$

so that if $Y^{[J]}$ and $\widetilde{Y}^{[J]}$ are coupled up to time t, the coupling can be extended to to time $t+1$ with additional probability of coupling error at most $\frac{n\theta_{\max}}{t}$. The overall increase in the probability of coupling failure due to the exceptional times is less than or equal to $\frac{\theta_{\max}nJ}{\tau_0}\to 0$.

Next, suppose t is not an exceptional time. Let $y\in\mathbb{Z}_+^J$ such that $\sum_j y^j\leqslant n$ and $y^j=0$ for $j\notin A_t$. Lemma 4 with $p_j=\frac{\vartheta_j y^j}{t}$, $p'_j=\frac{\theta_{v_j,t}y^j}{t}$, and $q_{u,j}=\frac{\theta_{u,v_j,t}y^j}{t}$ for $j\in A_t$ implies that the error for attempting to couple $\widetilde{Y}_{t+1}^{[J]}$ to $Y_{t+1}^{[J]}$ given $\widetilde{Y}_t^{[J]}=Y_t^{[J]}=y^{[J]}$ is less than or equal to

$$\frac{\left(\sum_{j\in[J]}\vartheta_j y^j\right)^2}{t^2}+\frac{\sum_{j\in[J]}|\vartheta_j-\theta_{v_j,t}|y^j}{t}$$
$$+\sum_{u\in[r]}\rho_u\frac{\left(\sum_{j\in[J]}\theta_{u,v_j,t}y^j\right)^2}{t^2}$$

Hence, the probability of coupling failure, before the sum of degrees is n and before time $T+1$, is less than or equal to

$$\mathbb{P}(F)+\frac{\theta_{\max}nJ}{\tau_0}+\sum_{t=\tau_0}^{T}\left(\frac{J^2\theta_{\max}^2n^2}{t^2}+\frac{n\epsilon}{t}+\frac{\theta_{\max}^2n^2}{t^2}\right),$$

which can be made arbitrarily small as in the proof of Proposition 3. □

APPENDIX D

APPENDIX: ALTERNATIVE PROOF OF PROPOSITION 2

This section gives an alternative proof of Proposition 2, but only for convergence in probability, based on Corollary 1. The same method can be used to prove Proposition 8(b), concerning the convergence in probability of the fraction of label errors made by two recovery algorithms. We use the notation given just before the statement of Proposition 2.

Since the labels of the vertices are independent with distribution ρ, by the law of large numbers,

$$\lim_{T\to\infty} \frac{H^v(T)}{T} = \rho_v \qquad \text{(a.s. and in probability)} .$$

Thus, it suffices to show that for fixed $n \geqslant m$,

$$\lim_{T\to\infty} \frac{N_n^v(T)}{T} = \rho_v p_n(\theta_v^*, m) \qquad \text{(in probability).}$$

By the Chebychev inequality, for that it suffices to show the following two conditions:

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[N_n^v(T)\right]}{T} = \rho_v p_n(\theta_u^*, m) \tag{38}$$

$$\lim_{T\to\infty} \mathsf{var}\left(\frac{N_n^v(T)}{T}\right) = 0. \tag{39}$$

Write $N_n^v(T) = \sum_{\tau=1}^T \chi_\tau$, where $\chi_\tau = 1$ if $\ell_\tau = v$ and the degree of vertex τ at time T is n, and $\chi_\tau = 0$ otherwise. Then $\left|\mathbb{E}\left[N_n^v(T)\right] - \sum_{\tau=t_o+1}^T \mathbb{E}\left[\chi_\tau\right]\right| \leqslant t_o$. By Corollary 1 with $J = 1$, $t = T$, and $v \in [r]$,

$$\lim_{\tau_0\to\infty} \sup_{\tau, T:\tau>\tau_0 \text{ and } T>\tau_0} \left|\mathbb{E}\left[\chi_\tau\right] - \rho_v \pi_n\left(\ln(T/\tau), \theta_v^*, m\right)\right| = 0. \tag{40}$$

Therefore, by the bounded convergence theorem, (38) holds with

$$\begin{aligned} p_n(\theta, m) &= \frac{1}{T}\int_0^T \pi_n(\ln(T/t), \theta, m) dt \\ &\overset{(a)}{=} \binom{n-1}{m-1} \int_0^1 u^{m\theta}(1-u^\theta)^{n-m} du \\ &\overset{(b)}{=} \frac{1}{\theta}\binom{n-1}{m-1} \int_0^1 v^{m-1+\frac{1}{\theta}}(1-v)^{n-m} dv \\ &\overset{(c)}{=} \frac{1}{\theta}\binom{n-1}{m-1} B\left(m+\frac{1}{\theta}, n-m+1\right) \\ &\overset{(d)}{=} \frac{\Gamma\left(\frac{1}{\theta}+m\right)\Gamma(n)}{\theta\Gamma(m)\Gamma\left(n+\frac{1}{\theta}+1\right)}, \end{aligned} \tag{41}$$

where (a) follows by the definition of the negative binomial distribution and change of variable $u = t/T$, (b) follows by the change of variable $v = u^\theta$, and (c) and (d) follow from standard formulas for the beta function, B.

It remains to verify (39). First note that

$$\mathsf{var}(N_n^v(T)) = \sum_{\tau_1=1}^T \sum_{\tau_2=1}^T \mathrm{Cov}(\chi_{\tau_1}, \chi_{\tau_2}). \tag{42}$$

Note that

$$\mathbb{E}\left[\chi_{\tau_1}\chi_{\tau_2}\right] = \rho_v^2 \mathbb{P}\left\{ Y_T^1 = n, Y_T^2 = n \middle| \ell_{\tau_1} = \ell_{\tau_2} = v \right\},$$

and by Corollary 1 with $J = 2$, $t = T$, and $v_1 = v_2 = v$,

$$\lim_{\tau_0 \to \infty} \sup_{\tau_1, \tau_2, T : \tau_0 \leqslant \tau_1 < \tau_2 \text{ and } T \geqslant \tau_0} \left| \mathbb{E}\left[\chi_{\tau_1}\chi_{\tau_2}\right] - \rho_v^2 \pi_n\left(\ln \frac{T}{\tau_1}, \theta_v^*, m\right) \pi_n\left(\ln \frac{T}{\tau_2}, \theta_v^*, m\right) \right| \to 0.$$

So, in view of (40) and the fact $\mathrm{Cov}(\chi_{\tau_1}, \chi_{\tau_2}) = \mathbb{E}\left[\chi_{\tau_1}\chi_{\tau_2}\right] - \mathbb{E}\left[\chi_{\tau_1}\right]\mathbb{E}\left[\chi_{\tau_2}\right]$,

$$\lim_{\tau_0 \to \infty} \sup_{\tau_1, \tau_2, T : \tau_0 \leqslant \tau_1 < \tau_2 \text{ and } T \geqslant \tau_0} |\mathrm{Cov}(\chi_{\tau_1}, \chi_{\tau_2})| = 0.$$

Using this to bound the terms on the righthand side of (42) with $\tau_1, \tau_2 \in [\tau_0, T]$ and $\tau_1 \neq \tau_2$, and bounding the other terms by one, yields:

$$\mathsf{var}(N_n^v(T)) \leqslant 2T\tau_0 + T + T^2 \left(\sup_{\tau_1, \tau_2, T : \tau_0 \leqslant \tau_1 < \tau_2 \leqslant T} |\mathrm{Cov}(\chi_{\tau_1}, \chi_{\tau_2})| \right) = o(T^2).$$

if $T, \tau_0 \to \infty$ with $\tau_0/T \to 0$. This implies (39), completing the alternative proof of the Proposition 2 (for convergence in probability).

Remark 6. *In essence, the calculation in* (41) *demonstrates that the limiting empirical distribution of degree for vertices of a given label v at a large time T, is the marginal distribution for the following joint distribution: the vertex time of arrival is uniform over $[0, T]$ and, given the arrival is at time τ, the conditional distribution of degree is* $\mathsf{negbinom}\left(m, \left(\frac{\tau}{T}\right)^{\theta_v^*}\right)$.

APPENDIX E
CONSISTENT ESTIMATION OF THE GROWTH RATE PARAMETER FOR A GIVEN VERTEX

Proposition 6 is proved in this section and evidence for Conjecture 1 is given. First a different method for estimating the rate parameter of Y is established. Consider the Barabási-Albert model with communities. Fix $\tau_o \geqslant 1$ and τ with $\tau \geqslant \max\{\tau_o, t_o\}$ (recall that t_o is the number of vertices in the initial graph). Let Y_t denote the degree of τ_o in G_t for all $t \geqslant \tau$. To avoid triviality associated with an isolated vertex in G_{t_o}, suppose $Y_\tau \geqslant 1$. We also suppose $Y_\tau \leqslant m\tau$, so by induction on t, $\frac{Y_t}{t} \leqslant m$ for all $t \geqslant \tau$. Let $\vartheta = \theta_v^*$ where v is the label of τ_o.

Proposition 10. *(Consistent estimation of rate parameter) The estimator $\widehat{\vartheta}_T$ defined by*

$$\widehat{\vartheta}_T = \frac{Y_T - Y_\tau}{\sum_{t=\tau}^{T-1} \frac{Y_t}{t}} \tag{43}$$

is consistent. In other words, $\lim_{T\to\infty} \widehat{\vartheta}_T = \vartheta$ *a.s.*

To prove the proposition we first examine a sequential version of $\widehat{\vartheta}_T$. Given a positive constant M with $M > m$, let T_M denote the stopping time defined by

$$T_M = \min \left\{ T \geqslant \tau : \sum_{t=\tau}^{T_M} \frac{Y_t}{t} \geqslant M \right\}$$

Let $\widehat{\widehat{\vartheta}}_M$ be $\widehat{\vartheta}_T$ for $T = T_M$, or, in other words,

$$\widehat{\widehat{\vartheta}}_M = \frac{Y_{T_M} - Y_\tau}{\sum_{t=\tau}^{T_M-1} \frac{Y_t}{t}}.$$

Lemma 5. *Under the idealized assumption $\eta_t \equiv \eta^*$, for any $\epsilon > 0$,*

$$\mathbb{P}\left\{ \left| \widehat{\widehat{\vartheta}}_M - \vartheta \right| \geqslant \epsilon \right\} \leqslant \frac{m\vartheta M}{\epsilon^2 (M-m)^2.}$$

Proof. Notice that the denominator of $\widehat{\widehat{\vartheta}}_M$ is in the interval $[M-m, M]$ with probability one. Also,

$$Y_T - Y_\tau - \vartheta \left(\sum_{t=\tau}^{T-1} \frac{Y_t}{t} \right) = \sum_{t=\tau}^{T-1} \left(Y_{t+1} - Y_t - \frac{\vartheta Y_t}{t} \right),$$

so that $\left(Y_T - Y_\tau - \vartheta \sum_{t=\tau}^{T-1} \frac{Y_t}{t} : T \geqslant \tau \right)$ is a martingale. Since T_M is a bounded optional sampling time, the martingale optional sampling theorem can be applied to yield

$$\mathbb{E}\left[Y_{T_M} - Y_\tau \right] = \mathbb{E}\left[\vartheta \sum_{t=\tau}^{T_M-1} \frac{Y_t}{t} \right] \in [\vartheta(M-m), \vartheta M].$$

Next we bound the second moments. It is easy to show that a random variable U with values in $[0, m]$ and mean μ satisfies $\mathsf{var}(U) \leqslant m^2 \frac{\mu}{m}\left(1 - \frac{\mu}{m}\right) \leqslant m\mu$. For any $t \geqslant \tau$, $Y_{t+1} - Y_t$

takes values in $[0, m]$ and, given the past $\mathcal{F}_t$, it has conditional mean $\frac{\vartheta Y_t}{t}$. It follows that $\mathbb{E}\left[\left(Y_{t+1} - Y_t - \frac{\vartheta Y_t}{t}\right)^2 \middle| \mathcal{F}_t\right] \leqslant \frac{m\vartheta Y_t}{t}$. Therefore, again using the optional sampling theorem,

$$\begin{aligned}
&\mathbb{E}\left[\left(Y_{T_M} - Y_\tau - \vartheta\left(\sum_{t=\tau}^{T_M-1} \frac{Y_t}{t}\right)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{t=\tau}^{T_M-1} \mathbb{E}\left[\left(Y_{t+1} - Y_t - \frac{\vartheta Y_t}{t}\right)^2 \middle| \mathcal{F}_t\right]\right] \\
&\leqslant m\vartheta\mathbb{E}\left[\sum_{t=\tau}^{T_M-1} \frac{Y_t}{t}\right] \leqslant m\vartheta M
\end{aligned}$$

Thus, for any $\epsilon > 0$, the Chebychev inequality yields

$$\begin{aligned}
&\mathbb{P}\left\{\left|Y_{T_M} - Y_\tau - \vartheta\left(\sum_{t=\tau}^{T_M-1} \frac{Y_t}{t}\right)\right| \geqslant \epsilon(M-m)\right\} \\
&\leqslant \frac{m\vartheta M}{\epsilon^2(M-m)^2},
\end{aligned}$$

which implies the conclusion of the proposition. □

Proof of Proposition 10. Since $Y_t \geqslant 1$ for all $t \geqslant \tau$, $\sum_{t=\tau}^{T-1} \frac{Y_t}{t} \to \infty$ a.s. as $T \to \infty$. Therefore, for τ_o fixed (the vertex for which we want to estimate the rate parameter), whether $\widehat{\vartheta}$ is consistent does not depend on the choice of τ. For any given $\epsilon > 0$, by taking τ very large, we can thus ensure $|\eta_t - \eta^*| \leqslant \epsilon$ for all $t \geqslant \tau$ with probability at least $1 - \epsilon$. Therefore, it suffices to prove the proposition under the added assumption $\eta_t \equiv \eta^*$ for all $t \geqslant \tau$. It follows that it suffices to prove that $\widehat{\widehat{\vartheta}}_M$ is a consistent family of estimators of ϑ.

So it remains to prove consistency of the family of estimators $\widehat{\widehat{\vartheta}}_M$ as $M \to \infty$. For that purpose, it suffices to show that for arbitrarily small $\epsilon > 0$, along the sequence of M values $M_k = (1+\epsilon)^k$, the estimation error is greater than or equal to ϵ for only finitely many values of k, with probability one. That follows from Lemma 5, because the error probability in Lemma 5 is $O(1/M)$ and $\sum_{k=1}^{\infty} 1/M_k < \infty$, so the Borel Cantelli lemma implies the desired conclusion. □

Proposition 6 will follows from Proposition 10 and the following lemmas, which are essentially Grönwall type inequalities.

Lemma 6. *Suppose $(f(s) : s \in \mathbb{R}_+)$ is a positive nondecreasing function such that for some $\vartheta > 0$,*

$$\lim_{S\to\infty} \frac{f(S) - f(0)}{\int_0^S f(u)du} = \vartheta.$$

Then

$$\lim_{S\to\infty} \frac{\ln f(S)}{S} = \vartheta.$$

Proof. Given any $\epsilon > 0$, there exits S_ϵ such that

$$f(S) \geqslant f(0) + (\theta - \epsilon) \int_0^S f(u)du \quad \text{for } S \geqslant S_\epsilon.$$

Since $f(u) \geqslant f(0)$ for all u,

$$f(S) \geqslant C + (\theta - \epsilon) \int_{S_\epsilon}^S f(u)du \quad \text{for } S \geqslant S_\epsilon$$

where $C = f(0)(1 + (\theta - \epsilon)S_\epsilon)$. Thus, for any $s \geqslant 0$, setting $S = s + S_\epsilon$, yields

$$f(s + S_\epsilon) \geqslant C + (\theta - \epsilon) \int_0^s f(s + S_\epsilon)du \quad \text{for } s \geqslant 0.$$

By induction on k it follows that $f(s+S_\epsilon) \geqslant C \sum_{j=0}^k \frac{((\vartheta-\epsilon)s)^j}{j!}$, so that $f(s+S_\epsilon) \geqslant C\mathrm{e}^{s(\vartheta-\epsilon)}$ for all $s \geqslant 0$. Therefore, $\liminf_{S\to\infty} \frac{\ln f(S)}{S} \geqslant \vartheta$. It can be proved similarly that $\limsup_{S\to\infty} \frac{\ln f(S)}{S} \leqslant \vartheta$, establishing the lemma. □

Lemma 7. *Let* $(y_t : t \in \{\tau, \tau + 1, \ldots\})$ *be a sequence of positive numbers such that* $y_{t+1} - y_t \in [0, m]$ *for all* $t \geqslant \tau$*, and such that*

$$\lim_{T\to\infty} \frac{y_T - y_\tau}{\sum_{t=\tau}^{T-1} \frac{y_t}{t}} = \vartheta$$

Then

$$\lim_{T\to\infty} \frac{\ln y_T}{\ln(T/\tau)} = \vartheta$$

Proof. We shall apply the previous lemma by switching to a continuous parameter and then applying a change of time. Note that $0 \leqslant \frac{1}{t} - \int_t^{t+1} \frac{1}{s}ds = \frac{1}{t} - \ln(1 + \frac{1}{t}) \leqslant \frac{1}{2t^2}$. Hence

$$0 \leqslant \sum_{t=\tau}^{T-1} \frac{y_t}{t} - \int_\tau^T \frac{y_{\lfloor t\rfloor}}{t}dt \leqslant \frac{1}{2} \sum_{t=\tau}^{T-1} \frac{y_t}{t^2} = o\left(\sum_{t=\tau}^{T-1} \frac{y_t}{t}\right).$$

The hypotheses thus imply

$$\lim_{T\to\infty} \frac{y_T - y_\tau}{\int_\tau^T \frac{y_{\lfloor t\rfloor}}{t}dt} = \vartheta.$$

Letting $f(s) = y_{\lfloor \tau \mathrm{e}^s \rfloor}$, the change of variable $u = \ln(t/\tau)$ yields

$$\frac{y_T - y_\tau}{\int_\tau^T \frac{y_{\lfloor t\rfloor}}{t}dt} = \frac{f(\ln(T/\tau)) - f(0)}{\int_\tau^T \frac{f(\ln(t/\tau))}{t}dt} = \frac{f(\ln(T/\tau)) - f(0)}{\int_0^{\ln(T/\tau)} f(u)du},$$

so the hypotheses of Lemma 6 hold. Lemma 6 yields

$$\lim_{S\to\infty} \frac{\ln y_{\lfloor \tau \mathrm{e}^S \rfloor}}{S} = \vartheta,$$

which by the change of variable $S = \ln(T/\tau)$, is equivalent to the conclusion of the lemma. □

Proof of Proposition 10. Proposition 10 follows directly from Proposition 10 and Lemma 7. □

Evidence for Conjecture 1 The Kesten-Stigum theorem [22] in the case of single-type branching processes implies that $\lim_{s\to\infty} Z_s \mathrm{e}^{-\vartheta s} = W$ a.s. for some random variable W such that $\mathbb{P}\{W > 0\} = 1$ and $\mathbb{E}[W] = Z_0 = m$. (This follows from the fact that Z restricted to multiples of any small positive constant $h > 0$ is a discrete-time single-type Galton Watson branching process with number of offspring per individual per time period, represented by a random variable L_h, such that L_h has the $\mathsf{negbinom}(m, \mathrm{e}^{\vartheta h})$ distribution. Note that $\mathbb{P}\{L_h \geqslant 1\} = 1$ and $\mathbb{E}[L_h \ln L_h] < \infty$.) Since $Z_t \mathrm{e}^{-\vartheta s}$ also converges in distribution to the Gamma distribution with parameters m and ϑ, it follows that W has such distribution. It follows that (11) holds if the process Y is replaced by the process $\breve{Y}$.

APPENDIX F
PROOF OF PROPOSITION 7

The process Z with parameters λ, m represents the total population of a branching process starting with m root individuals at time 0, such that each individual in the population spawns new individuals at rate λ. And A_s represents the sum of the lifetimes, truncated at time s, of all the individuals in the population. The joint distribution of (Z, A) with parameters λ, m is the same as the distribution of the sum of m independent versions of (Z, A) with parameters $\lambda, 1$, Hence, it suffices to prove the lemma for $m = 1$.

So for the remainder of this proof suppose $m = 1$; there is a single root individual. Suppose there are $n(s)$ children of the root individual, produced at times $R_1, \ldots, R_{n(s)}$. Then

$$Z_s = 1 + \sum_{l=1}^{n(s)} Z^{\ell}_{s-R_l} \tag{44}$$

$$A_s = s + \sum_{l=1}^{n(s)} A^{\ell}_{s-R_l} \tag{45}$$

where $Z^{\ell}_{s-R_l}$ denotes the total subpopulation of the l^{th} child of the root, $s - R_l$ time units after the birth of the l^{th} child, and $A^{\ell}_{s-R_l}$ is the associated sum of lifetimes of that subpopulation,

truncated $s - R_l$ time units after the birth of the l^{th} child (i.e. truncated at time s). The processes (Z^l, A^l) are independent and have the same distribution as (Z, A). The variables $R_1, \ldots, R_{n(s)}$ are the points of a Poisson process of rate λ. Therefore,

$$e^{uZ_s + vA_s} = \mathrm{e}^{u+vs} \prod_{l=1}^{n(s)} \exp(uZ_{s-R_l} + vA_{s-R_l}),$$

which after taking expectations yields

$$\psi_{\lambda,1}(u, v, s) = \mathrm{e}^{u+vs} \mathbb{E}_{\lambda,1} \left[\prod_{l=1}^{n(s)} \exp(uZ^l_{s-R_l} + vA^l_{s-R_l}) \right].$$

Since $n(s)$ is a Poisson(λ) random variable, and, given $n(s)$, $R_1, \ldots, R_{n(s)}$ are distributed uniformly on $[0, s]$, the above expectation can be simplified by first conditioning on $n(s)$, and then summing over all possible values of $n(s)$ (tower property).

$$\begin{aligned}
&\psi_{\lambda,1}(u, v, s) \\
&= e^{u+vs} \sum_{k=0}^{\infty} \frac{e^{-\lambda s}(\lambda s)^k}{k!} \mathbb{E}_{\lambda,1} \left[\prod_{l=1}^{k} e^{uZ^l(s-R_l) + vA^l(s-R_l)} \right] \\
&= e^{u+vs} \sum_{k=0}^{\infty} \frac{e^{-\lambda s}(\lambda s)^k}{k!} \left(\frac{1}{s} \int_0^s \psi_\lambda(u, v, \tau) d\tau \right)^k
\end{aligned} \tag{46}$$

In the above step, the expectation of the product is the same as the product of the expectations, because the variables $(Z^l(s - R_l), A^l(s - R_l)), l = 1, \ldots, k$ are independent of each other. Moreover, the expectation of each of the k terms is identical. Denoting $F(s) \triangleq \int_0^s \psi_{\lambda,1}(u, v, \tau) d\tau$, we can write (46) as

$$\begin{aligned}
\dot{F}(s) &= e^{u+vs} e^{-\lambda s} e^{\lambda F(s)} \\
\frac{d}{ds}\left(e^{-\lambda F(s)}\right) &= -\lambda e^{(v-\lambda)s+u}; \quad F(0) = 0 \\
e^{-\lambda F(s)} &= 1 - \lambda e^u \int_0^s e^{(v-\lambda)s'} ds' \\
&= 1 + \frac{\lambda e^u}{v - \lambda} \left(1 - e^{(v-\lambda)s}\right) \\
F(s) &= -\frac{1}{\lambda} \log \left(1 + \frac{\lambda e^u}{v - \lambda} \left(1 - e^{(v-\lambda)s}\right) \right)
\end{aligned} \tag{47}$$

Finally, using $\psi_{\lambda,1}(u, v, s) = \dot{F}(s)$ yields (15) for $m = 1$, and the proof is complete.

APPENDIX G
PROOF OF PROPOSITION 9

Proof. The basic difficulty to be overcome is that the limit result $\eta_t \to \eta^*$ in Proposition 1 doesn't approximately determine the distribution of the degree evolution for vertex τ_o if $\tau_o \nrightarrow \infty$. To produce an estimator for ℓ_{τ^o} given $Y^o_{[\tau^o,T]}$, we produce a virtual degree growth process, denoted by $\breve{Y}^o_{[\tau,T]}$, which becomes arbitrarily close to $\widetilde{Y}_{[\tau,T]}$ in total variation distance as $T \to \infty$ under any of the r hypotheses about ℓ_{τ^o}, where $\tau \to \infty$ with $\tau/T \to a$ for some fixed $\delta > 0$.

Given an arbitrary $\epsilon > 0$, select $\delta \in (0,1)$ so small that $f^C_Z(\rho, \theta^*, m, \ln(1/\delta)) < \epsilon$. Suppose τ depends on T such that $\tau/T \to \delta$ as $T \to \infty$. By Proposition 8, ℓ_τ can be recovered with error probability less than ϵ from $\widetilde{Y}_{[\tau,T]}$ by using Algorithm C.

The virtual process $\breve{Y}^o_{[\tau,T]}$ has initial value $\widetilde{Y}^o_\tau = m$. Thus, although τ_o arrives before τ, the virtual process does not begin evolution until after time τ. The construction of $\breve{Y}^o$ proceeds by induction and uses a random thinning of the process Y^o, the actual degree growth process for τ^o. The thinning probability is the ratio of degrees. Specifically, for t with $\tau \leqslant t \leqslant T-1$, let

$$\mathcal{L}(\breve{Y}^o_{t+1} - \breve{Y}^o_t | \breve{Y}^o_{[\tau,t]}, Y^o_{[\tau_o,T]}) = \mathsf{binom}\left(Y^o_{t+1} - Y^o_t, \frac{\breve{Y}^o_t}{Y^o_t}\right).$$

The virtual process $\breve{Y}^o_{[\tau,T]}$ satisfies the same properties as $Y_{[\tau,T]}$ (based on the degree evolution of vertex τ) used in the proof of Proposition 5, so for $v \in [r]$,

$$d_{TV}\left((\breve{Y}^o_{[\tau,T]} | \ell_{\tau^o} = v), (\widetilde{Y}_{[\tau,T]} | \ell_\tau = v)\right) \to 0.$$

Hence, applying Algorithm C, designed for recovery of ℓ_τ, to the virtual process $\breve{Y}_{[\tau,T]}$ recovers ℓ_{τ^o} with average error probability less than ϵ for T sufficiently large. □

APPENDIX H
DERIVATION OF THE MESSAGE PASSING EQUATIONS

The initial conditions given by (23) are chosen to make the initial likelihood vector the same as produced by Algorithm C (observation of children). Equations (18) - (22) are derived in what follows in the special case $m = 1$, with the initial graph G_{t_o} consisting of a single vertex (i.e. $t_o = 1$) with a self-loop. In that case, the graph (V, E) is a tree (ignoring the self-loop incident to the first vertex) so the message passing algorithm is conceptually simpler. The equations (18) - (22) for any finite $m \geqslant 1$ are simply taken to have the same form as for $m = 1$ on the grounds

that loopy message passing is obtained by using the same equations as for message passing without loops.

Our first assumption in deriving the message passing algorithm is that the approximation λ_τ^C for the log likelihood vector based on observation of children (derived in Section III) is exact, or in other words:

$$\ln \mathbb{P}\left\{\partial\tau = \{t_1, \ldots, t_n\} | \ell_\tau = v\right\} = \lambda_\tau^C(v), \tag{48}$$

where $\Lambda_\tau^c(v)$ is given by (17). The second assumption is regarding how the distribution of $\partial\tau$ changes, given the label of another vertex. Namely,

$$\begin{aligned} &\mathbb{P}\left\{\partial\tau = \{t_1, \ldots, t_n\} | \ell_\tau = v, \ell_{\tau'} = u\right\} \\ &= \begin{cases} \mathbb{P}\left\{\partial\tau = \{t_1, \ldots, t_n\} | \ell_\tau = v\right\} \theta^*_{u,v}/\theta^*_v & \text{if } \tau' \in \partial\tau \\ \mathbb{P}\left\{\partial\tau = \{t_1, \ldots, t_n\} | \ell_\tau = v\right\} & \text{if } \tau' \notin \partial\tau \end{cases}, \end{aligned} \tag{49}$$

where the expression for the first case follows from (8).

The third assumption is regarding the joint distribution of degree-growth processes. Observing the degree-growth process of one vertex τ changes the distribution of the degree growth process of another vertex τ' in one of two possible ways. Firstly, the children of the first vertex cannot be the children of the other (if $m = 1$). However, Proposition 5 shows this effect is insignificant. Secondly, observing the degree-growth process gives us some information about the label of each vertex. If one vertex appears as a child of the other (say $\tau' \in \partial\tau$), the probability of the given observation is affected; else it is not. In the asymptotic limit, the degree-growth processes of a finite number of vertices are indeed independent, by Proposition 5.

The following additional notation is used. Let D_τ^k denote the event of observing the subtree of (V, E) rooted at τ, and of depth k. For example, $D_\tau^1 \equiv \{\partial\tau = \{t_1, \ldots, t_n\}\}$, $D_\tau^2 \equiv \{\partial\tau = \{t_1, \ldots, t_n\}, \partial t_1 = \{t_1^1, \ldots, t_{n_1}^1\}, \ldots, \partial t_n = \{t_1^n, \ldots, t_{n_n}^n\}\}$. Further, let D_τ denote the event of observing the subtree of (V, E) rooted at τ. We call this subtree as the *descendants* of τ. The event of observing the entire graph is D_1, because the initial graph has a single vertex. Therefore:

$$\Lambda_\tau(v) = \ln \mathbb{P}\left\{E_T = E | \ell_\tau = v\right\} = \ln \mathbb{P}\left\{D_1 | \ell_\tau = v\right\} \tag{50}$$

For a vertex τ with $\tau \geqslant 2$, the event $D_1 \backslash D_\tau$ includes the information of which vertex is the parent of vertex τ. Also, for vertices τ and τ_0 with $\tau_0 < \tau$, let $\tau \to \tau_0$ denote the event there is an edge from τ to τ_0.

At this point, we make the assumption:

$$\mathbb{P}\{D_1|\ell_\tau = v\} = \mathbb{P}\{D_\tau|\ell_\tau = v\}\,\mathbb{P}\{D_1\backslash D_\tau|\ell_\tau = v\} \;\; \forall\tau \tag{51}$$

In other words, D_τ and $D_1\backslash D_\tau$ are assumed to be conditionally independent given $\ell_\tau = v$. The rationale for that also comes from ignoring the implications of the fact that the descendants of τ must be disjoint from the descendants of vertices close to τ in G_T in the direction through the parent of τ.

Let τ and τ_0 be vertices such that τ is a child of τ_0. We define the messages as follows, and then derive the message passing equations as fixed points.

$$\nu_{\tau\to\tau_0}(u) \triangleq \ln\mathbb{P}\{D_\tau|\ell_\tau = u\} \tag{52}$$

$$\mu_{\tau_0\to\tau}(v) \triangleq \ln\left(\frac{\mathbb{P}\{D_1\backslash D_\tau|\ell_\tau = 0, \ell_{\tau_0} = v\}\,\theta^*_v}{\theta^*_{0,v}}\right) \tag{53}$$

$$\widetilde{\nu}_{\tau\to\tau_0}(v) \triangleq \ln\mathbb{P}\{D_\tau|\ell_{\tau_0} = v, \tau\to\tau_0\} \tag{54}$$

$$\widetilde{\mu}_{\tau_0\to\tau}(u) \triangleq \ln\mathbb{P}\{D_1\backslash D_\tau|\ell_\tau = u\} \tag{55}$$

Remark 7. *In the definition* (53) *of $\mu_{\tau_0\to\tau}$ it is assumed that 0 represents some choice of label, but the definition for all choices of 0 are equivalent. In other words, because of* (49)*,*

$$\mu_{\tau_0\to\tau}(v) = \ln\left(\frac{\mathbb{P}\{D_1\backslash D_\tau|\ell_\tau = u, \ell_{\tau_0} = v\}\,\theta^*_v}{\theta^*_{u,v}}\right) \tag{56}$$

for any $u \in [r]$.

We show that the message passing equations (18) - (22) follow from our independence assumptions and the definitions of the messages given in (52) - (55).

Derivation of (18)*:* Start with the fact $D_\tau = D^1_\tau \cap (\cap_{t\in\partial\tau} D_t)$, and, given $\ell_\tau = v$ and D^1_τ, The events $D_t, t \in \partial\tau$ are conditionally independent. Hence,

$$\begin{aligned}
&\mathbb{P}\{D_\tau|\ell_\tau = v\} \\
&= \mathbb{P}\left\{D^1_\tau|\ell_\tau = v\right\}\prod_{t\in\partial\tau}\mathbb{P}\left\{D_t|\ell_\tau = v, D^1_\tau\right\} \\
&= \mathbb{P}\left\{D^1_\tau|\ell_\tau = v\right\}\prod_{t\in\partial\tau}\mathbb{P}\{D_t|\ell_\tau = v, t\to\tau\}.
\end{aligned}$$

So by (48) and the definition of $\widetilde{\nu}_{t\to\tau}$,

$$\ln\mathbb{P}\{D_\tau|\ell_\tau = v\} = \lambda^C_\tau(v) + \sum_{t\in\partial\tau}\widetilde{\nu}_{t\to\tau}(v). \tag{57}$$

Since $\nu_{\tau\to\tau_0}(v) = \ln\mathbb{P}\{D_\tau|\ell_\tau = v\}$ this establishes (18) for $m = 1$.

Derivation of (19)*:* Assume $\tau_0 \geqslant t_o + 1$; the proof in case $\tau_0 \leqslant t_o$ is similar. Then, also accounting for the assumption $m = 1$, (19) becomes

$$\mu_{\tau_0 \to \tau} = \lambda_{\tau_0}^C + \sum_{t \in \partial \tau_0 \backslash \{\tau\}} \widetilde{\nu}_{t \to \tau_0} + \widetilde{\mu}_{\tau_1 \to \tau_0}, \tag{58}$$

where τ_1 is the parent of τ_0. Observe that

$$\begin{aligned} D_1 \backslash D_\tau &= (D_1 \backslash D_{\tau_0}) \cap (D_{\tau_0} \backslash D_\tau) \\ &= (D_1 \backslash D_{\tau_0}) \cap D_{\tau_0}^1 \cap \left(\cap_{t \in \partial \tau_0 \backslash \{\tau\}} D_t \right) \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{P}\left\{ D_1 \backslash D_\tau | \ell_\tau = 0, \ell_{\tau_0} = v \right\} \\ &= \mathbb{P}\left\{ D_1 \backslash D_{\tau_0} | \ell_\tau = 0, \ell_{\tau_0} = v \right\} \mathbb{P}\left\{ D_{\tau_0} \backslash D_\tau | \ell_\tau = 0, \ell_{\tau_0} = v \right\} \\ &= \mathbb{P}\left\{ D_1 \backslash D_{\tau_0} | \ell_{\tau_0} = v \right\} \mathbb{P}\left\{ D_{\tau_0} \backslash D_\tau | \ell_\tau = 0, \ell_{\tau_0} = v \right\} \\ &= \mathbb{P}\left\{ D_1 \backslash D_{\tau_0} | \ell_{\tau_0} = v \right\} \mathbb{P}\left\{ D_{\tau_0}^1 | \ell_\tau = 0, \ell_{\tau_0} = v \right\} \\ &\qquad \prod_{t \in \partial \tau_0 \backslash \{\tau\}} \mathbb{P}\left\{ D_t | \ell_\tau = 0, \ell_{\tau_0} = v, t \to \tau \right\} \\ &= \mathbb{P}\left\{ D_1 \backslash D_{\tau_0} | \ell_{\tau_0} = v \right\} \mathbb{P}\left\{ D_{\tau_0}^1 | \ell_\tau = 0, \ell_{\tau_0} = v \right\} \\ &\qquad \prod_{t \in \partial \tau_0 \backslash \{\tau\}} \mathbb{P}\left\{ D_t | \ell_{\tau_0} = v, t \to \tau \right\} \end{aligned}$$

Multiplying both sides of the above by $\frac{\theta_v^*}{\theta_{0,v}^*}$, using (49), and taking logarithms yields

$$\begin{aligned} \mu_{\tau_0 \to \tau}(v) = \ln \frac{\theta_v^*}{\theta_{0,v}^*} + \widetilde{\mu}_{\tau_1 \to \tau_0}(v) + \left(\Lambda_{\tau_0}^C(v) + \ln \frac{\theta_{0,v}^*}{\theta_v^*} \right) \\ + \sum_{t \in \partial \tau_0 \backslash \{\tau\}} \widetilde{\nu}_{t \to \tau}(v), \end{aligned}$$

which is equivalent to (58), so that (19) is proved for $m = 1$.

Derivation of (20)*:* Note that

$$\begin{aligned} &\mathbb{P}\left\{ D_\tau | \ell_{\tau_0} = v, \tau \to \tau_0 \right\} \\ &= \sum_{u \in [r]} \mathbb{P}\left\{ D_\tau, \ell_\tau = u | \ell_{\tau_0} = v, \tau \to \tau_0 \right\} \\ &= \sum_{u \in [r]} \mathbb{P}\left\{ \ell_\tau = u | \ell_{\tau_0} = v, \tau \to \tau_0 \right\} \mathbb{P}\left\{ D_\tau | \ell_\tau = u \right\} \\ &= \sum_{u \in [r]} \frac{\rho_u \theta_{u,v}^*}{\theta_v^*} \mathrm{e}^{\nu_{\tau \to \tau_0}(u)}, \end{aligned}$$

where for the second inequality we used

$\mathbb{P}\{D_\tau, \ell_\tau = u | \ell_{\tau_0} = v, \tau \to \tau_0\} = \mathbb{P}\{D_\tau | \ell_\tau = u\}$. Taking the logarithm of each side yields (20).

Derivation of (21)*:* The derivation is given by:

$$\begin{aligned}
&\widetilde{\mu}_{\tau_0 \to \tau}(u) = \ln \mathbb{P}\{D_1 \backslash D_\tau | \ell_\tau = u\} \\
&= \ln \sum_{v \in [r]} \mathbb{P}\{D_1 \backslash D_\tau, \ell_{\tau_0} = v | \ell_\tau = u\} \\
&= \ln \sum_{v \in [r]} \mathbb{P}\{D_1 \backslash D_\tau | \ell_\tau = u, \ell_{\tau_0} = v\} \mathbb{P}\{\ell_{\tau_0} = v | \ell_\tau = u\} \\
&= \ln \sum_{v \in [r]} \theta^*_{u,v} \frac{\mathbb{P}\{D_1 \backslash D_\tau | \ell_\tau = u, \ell_{\tau_0} = v\}}{\theta^*_{u,v}} \rho_v \\
&= \ln \sum_{v \in [r]} \theta^*_{u,v} \frac{\mathbb{P}\{D_1 \backslash D_\tau | \ell_\tau = 0, \ell_{\tau_0} = v\}}{\theta^*_{0,v}} \rho_v \\
&= g^{pc}(\mu_{\tau_0 \to \tau})(u).
\end{aligned}$$

Derivation of (22)*:* Equation (22) (for $m = 1$) follows from (50), (51), (57), and (55).

REFERENCES

[1] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[2] C. Moore, "The computer science and physics of community detection: Landscapes, phase transitions, and hardness," 2017, arXiv 1702.00467.

[3] E. Abbe, "Community detection and stochastic block models: recent developments," Mar 2017, arXiv 1703.10146.

[4] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *Proc. 3rd Intl Workshop on Link Discovery*, New York, NY, USA, 2005, pp. 36–43.

[5] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[6] J. Jordan, "Geometric preferential attachment in non-uniform metric spaces," *Electronic Journal Probability*, vol. 18, no. 8, pp. 1–15, 2013.

[7] G. Bianconi and A. Barabási, "Competition and multiscaling in evolving networks," *EuroPhysics Letters*, vol. 54, no. 4, pp. 436–442, 2001.

[8] A.-L. Barabási, *Network Science*. Cambridge University Press, 2016.

[9] A. Montanari, "Finding one community in a sparse random graph," *Journal of Statistical Physics*, vol. 161, no. 2, pp. 273–299, 2015, arXiv 1502.05680.

[10] B. Hajek, Y. Wu, and J. Xu, "Recovering a hidden community beyond the Kesten-Stigum threshold in $O(|E| \log^* |V|)$ time," *J. Applied Probability*, vol. 55, no. 2, June 2018, arXiv 1510.02786.

[11] T. Antunović, E. Mossel, and M. Racz, "Coexistence in preferential attachment networks," *Combinator. Probab. Comp.*, vol. 25, pp. 797–822, 2016. [Online]. Available: https://arxiv.org/abs/1307.2893

[12] Y. Chen, X. Li, and J. Xu, "Convexified modularity maximization for degree-corrected stochastic block models," December 2015, arXiv 1512.08425. To appear in Annals of Statistics.

[13] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, "The degree sequence of a scale-free random graph process," *Random Structures & Algorithms*, vol. 18, no. 3, pp. 279–290, 2001.

[14] S. Janson, "Limit theorems for triangular urn schemes," *Probab. Theory Related Fields*, vol. 134, pp. 417–452, 2006.

[15] E. A. Peköz, N. Ross, and A. Röllin, "Joint degree distributions of preferential attachment random graphs," 02 2014, arXiv 1402.4686. [Online]. Available: http://arxiv.org/abs/1402.4686

[16] A. D. Flaxman, A. M. Frieze, and J. Vera, "A geometric preferential attachment model of networks," *Internet Math*, vol. 3, no. 187-205, 2006.

[17] T. Luczak, A. Magner, and W. Szpankowski, "Asymmetry and structural information in preferential attachment graphs," 07 2016, arXiv 1607.04102.

[18] H. Kobayashi and J. Thomas, "Distance measures and releated criteria," in *Proc. 5th Allerton Conf. Circuit and System Theory*, Monticello, Illinois, 1967, pp. 491–500.

[19] H. Poor, *An introduction to signal detection and estimation*. Springer Science, 1994.

[20] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE transactions on communication technology*, vol. 15, no. 1, pp. 52–60, 1967.

[21] V. Borkar, "Stochastic approximation," *Cambridge Books*, 2008.

[22] H. Kesten and B. Stigum, "A limit theorem for multidimensional Galton-Watson processes," *Ann. Math. Statist.*, vol. 37, pp. 1211–1223, 1966.