Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 825–865
https://doi.org/10.1093/jrsssb/qkad102
Advance access publication 26 September 2023
Discussion Paper

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Root and community inference on the latent growth process of a network

Harry Crane and Min Xu

Department of Statistics, Rutgers University, New Brunswick, NJ, USA

Address for correspondence: Min Xu, Department of Statistics, Rutgers University, New Brunswick, NJ 08854, USA.
Email: mx76@stat.rutgers.edu

Read before The Royal Statistical Society in London at the Discussion Meeting organized by the Research Section on Wednesday, 6 December 2023, Dr Robin J Evans in the Chair.

Abstract

Many statistical models for networks overlook the fact that most real-world networks are formed through a growth process. To address this, we introduce the Preferential Attachment Plus Erdős–Rényi model, where we let a random network $\boldsymbol{G}$ be the union of a preferential attachment (PA) tree $\boldsymbol{T}$ and additional Erdős–Rényi (ER) random edges. The PA tree captures the underlying growth process of a network where vertices/edges are added sequentially, while the ER component can be regarded as noise. Given only one snapshot of the final network $\boldsymbol{G}$, we study the problem of constructing confidence sets for the root node of the unobserved growth process; the root node can be patient zero in an infection network or the source of fake news in a social network. We propose inference algorithms based on Gibbs sampling that scales to networks with millions of nodes and provide theoretical analysis showing that the size of the confidence set is small if the noise level of the ER edges is not too large. We also propose variations of the model in which multiple growth processes occur simultaneously, reflecting the growth of multiple communities; we use these models to provide a new approach to community detection.

Keywords: community detection, Gibbs sampling, network data analysis, preferential attachment model, root inference

1 Introduction

Network data is ubiquitous. To analyse networks, there are a variety of statistical models such as Erdős–Rényi, stochastic block model (SBM) (Abbe, 2017; Amini et al., 2013; Karrer & Newman, 2011; Xu et al., 2018), graphon (Diaconis & Janson, 2007; Gao et al., 2015), random dot product graphs (Athreya et al., 2017; Xie & Xu, 2019), latent space models (Hoff et al., 2002), configuration graphs (Aiello et al., 2000), and more. These models usually operate by specifying some structure, such as community structure in the case of SBM, and then adding independent random edges in a way that reflects the structure. The order in which the edges are added is of no importance to these models.

In contrast, real-world networks are often formed from growth processes where vertices and edges are added sequentially. This motivates the development of Markovian preferential attachment (PA) models for networks (Barabási, 2016; Barabási & Albert, 1999) which produce a sequence of networks $G_1, G_2, \ldots, G_n$ where G_1 starts as a single node which we call the root node and, at each iteration, we add a new node and new edges. PA models naturally produce networks with sparse edges, heavy-tailed degree distributions, and strands of chains as well as pendants (several degree 1 vertices linked to a single vertex), which are important features of real-world networks that are difficult to reproduce under a non-Markovian model, as observed by Bloem-Reddy and Orbanz (2018).

Although Markovian models are often more realistic, they have not been as widely used in network data analysis as, say SBM, because, whereas SBM is useful for recovering the community structure of a network, it is not obvious what structural information Markovian models could

Received: July 14, 2021. **Revised:** January 9, 2023. **Accepted:** January 25, 2023

extract from a network. Recently, however, seminal work from a series of applied probability papers (e.g. Bubeck, Devroye et al., 2017; Bubeck et al., 2015) demonstrate that Markovian models can indeed recover useful structure: these papers show that, surprisingly, when $\boldsymbol{G}_n$ is a random PA tree, one can infer the early history of $\boldsymbol{G}_n$, such as the root node, even as the size of the tree tends to infinity. Although these results are elegant, they are theoretical; their confidence set construction involves large constants that render the result too conservative. Moreover, most algorithms apply only to tree-shaped networks, which prohibitively limits their application since trees are rarely encountered in practice.

To overcome these problems, we propose a Markovian model for networks which we call Preferential Attachment Plus Erdős–Rényi, or PAPER for short. We say that $\boldsymbol{G}_n$ has the PAPER distribution if it is generated by adding independent random edges to a preferential attachment tree $\boldsymbol{T}$. The latent PA tree captures the growth process of the network whereas the ER random edges can be interpreted as additional noise. Given only a single snapshot of the final graph $\boldsymbol{G}_n$, we study how to infer the early history of the latent tree $\boldsymbol{T}$, focussing on the concrete problem of constructing confidence sets for the root node that can attain the nominal coverage. We give a visual illustration of the PAPER model and the inference problem in Figure 1.

Because we do not know which edges of $\boldsymbol{G}_n$ correspond to the tree and which are noise, most existing methods are not directly applicable. We therefore propose a new approach in which we first give the nodes new random labels which induce, for a given observation of the network $\boldsymbol{G}_n$, a posterior distribution of both the latent tree and the latent arrival ordering of the nodes. Then, we sample from the posterior distribution to construct a credible set for the inferential target, e.g. the root node. Bayesian inference statements usually do not have frequentist validity but we prove in our setting that that the level $1 - \epsilon$ credible set for the root node has frequentist coverage at exactly the same level.

In order to efficiently sample from the posterior distribution of the latent ordering and the latent tree, we present a scalable Gibbs sampler that alternatingly samples the ordering and the tree. The algorithm to generate the latent ordering is based on our previous work (Crane & Xu, 2021) which studies inference in the tree setting. The algorithm to generate the latent tree operates by updating the parent of each of the nodes iteratively. The overall runtime complexity of one iteration of the outer loop is generally $O(m + n \log n)$ (where m is the number of edges) and the algorithm can scale to networks of up to a million nodes.

Since a trivial confidence set for the root node is the set of all the nodes, it is important to be able to bound the size of a confidence set. In particular, the presence of noisy Erdős–Rényi edges in the PAPER model motivates an interesting question: how does the size of the confidence set increase with the noise level? In this paper, we give an initial answer to this question under two specific settings of the preferential attachment mechanism: linear preferential attachment (LPA) and

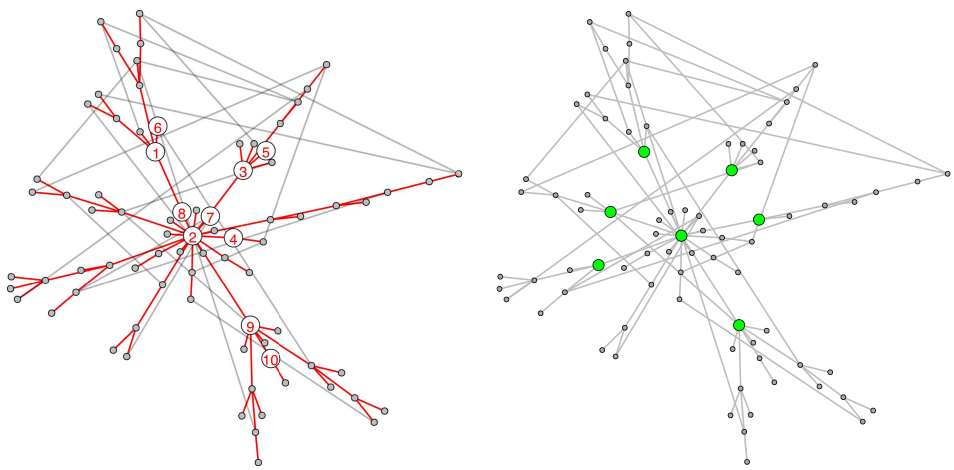


Figure 1. Left: Illustration of PAPER model; nodes have latent time ordering (only first 10 orderings shown); the dark red edges form the latent tree while light grey edges are Erdős–Rényi. Right: 80% confidence set for the root node (node number 1) constructed from the unlabelled graph.

uniform attachment (UA). For LPA, we prove that the size of our proposed confidence set does not increase with the number of nodes n so long as the noisy edge probability is less than $n^{-1/2}$ and for UA, we prove that the size is bounded by n^{γ} for some $\gamma < 1$ so long as the noisy edge probability is less than $\log(n)/n$. Our analysis shows that the phenomenon discovered by Bubeck, Devroye et al. (2017), that there exists confidence sets for the root node of $O(1)$ size, is robust to the presence of noise.

Many real-world networks often have community structures. In such cases, it would be unrealistic to assume that the network originates from a single root node. We therefore propose variations of the PAPER model in which K growth processes occur simultaneously from K root nodes. Each of K root nodes can be interpreted as being locally central with respect to a community subgraph. In the multiple roots model, there is no longer a latent tree but rather a latent forest (union of disjoint trees), where the components of the forest can naturally be interpreted as the different communities of the network. We provide model formulation that allows K to be either be fixed or random. To analyse networks with multiple roots, we use essentially the same inferential approach and Gibbs sampling algorithm that that we develop for the single root setting, with minimal modifications.

By looking at the posterior probability that a node is in a particular tree–community, we can estimate the community membership of each of the nodes. Compared with say the stochastic block model, the PAPER model approach to community recovery has the advantage that the inference quality improves with sparsity, that we can handle heavy-tailed degree distribution without a high-dimensional degree correction parameter vector, and that the posterior root probabilities also identify the important nodes in the community. Empirically, we show that our approach has competitive performance on two benchmark datasets and we find that our community membership estimate is more accurate for nodes with high posterior root probability than for the more peripheral nodes. We also use the PAPER model to conduct an extensive analysis of a statistician co-authorship network curated by Ji and Jin (2016) where we recover a large number of communities that accurately reflect actual research communities in statistics.

We have implemented our inference algorithm in a Python package called `paper-network`, which can be installed via command `pip install paper-network`. The code, example scripts, and documentation are all publicly available at https://github.com/nineisprime/PAPER.

1.1 Outline for the paper

In Section 2, we define the PAPER model in both the single root and multiple roots setting. We also formalise the problem of root inference and review related work. In Section 3, we describe our approach to the root inference problem, which is to randomise the node labels and analyse the resulting posterior distribution. We also show that the Bayesian inferential statements have frequentist validity. In Section 4, we give a sampling algorithm for computing the posterior probabilities. In Section 5, we provide theoretical bounds on the size of our proposed confidence sets and in Section 6, we provide empirical study on both simulated and large scale real-world networks.

We use the following notation throughout the paper:

- We take all graphs to be undirected. Given two labelled graphs $\boldsymbol{g}$ and $\boldsymbol{g}'$ defined on the same set of nodes, we write $\boldsymbol{g} + \boldsymbol{g}'$ as the resulting graph if we take the union of the edges in $\boldsymbol{g}$ and $\boldsymbol{g}'$ and collapse any multi-edges. We also write $\boldsymbol{g} \subset \boldsymbol{g}'$ if $\boldsymbol{g}$ is a subgraph of $\boldsymbol{g}'$.
- For a labelled graph $\boldsymbol{g}$, we write $D_{\boldsymbol{g}}(u)$ as the degree of node u in graph $\boldsymbol{g}$ and $N_{\boldsymbol{g}}(u)$ as the set of neighbours of u (all nodes directly connected to u) with respect to $\boldsymbol{g}$; we write $V(\boldsymbol{g})$ and $E(\boldsymbol{g})$ as the set of vertices and edges of $\boldsymbol{g}$, respectively.
- For an integer n, we write $[n] := \{1, 2, \ldots, n\}$. For a countable set A, we write $|A|$ as the cardinality of A. For two sets A, B of the same cardinality, we write $\mathrm{Bi}(A, B)$ as the set of bijections between them. For a vector π, we let $\pi_{1:K}$ be the sub-vector $(\pi_1, \pi_2, \ldots, \pi_K)$.
- Given a finite set V' of the same cardinality of $V(\boldsymbol{g})$ and given a bijection $\rho \in \mathrm{Bi}(V(\boldsymbol{g}), V')$, we write $\rho\boldsymbol{g}$ to denote a relabelled graph where a pair $(u', v') \in V' \times V'$ is an edge in $\rho\boldsymbol{g}$ if and only if $(u, v) \in V(\boldsymbol{g}) \times V(\boldsymbol{g})$ is an edge in $\boldsymbol{g}$.
- Throughout the paper, we use capital font (e.g. $\boldsymbol{G}$) to denote random objects and lower case font to denote fixed objects. Graphs are represented via bold font.

2 Model and problem

We first describe the model and inference problem in the single root setting and then extend the definition to the setting of having fixed K roots and having random K roots.

2.1 PAPER model

Definition 1 The affine preferential attachment tree model, which we denote by APA(α, β) for parameters $\alpha, \beta \in \mathbb{R}$, generates an increasing sequence $\boldsymbol{T}_1 \subset \boldsymbol{T}_2 \subset \cdots \subset \boldsymbol{T}_n$ of random trees where $\boldsymbol{T}_t$ is a tree with t nodes and where nodes are labelled by their arrival time so that $V(\boldsymbol{T}_t) = [t]$. The first tree $\boldsymbol{T}_1 = \{1\}$ is a singleton node, which we refer to as the *root node*, and for $t > 2$, we define the transition kernel $\mathbb{P}(\boldsymbol{T}_t \mid \boldsymbol{T}_{t-1})$ in the following way: given $\boldsymbol{T}_{t-1}$, we add a node labelled t and a random edge (t, w_t) to obtain $\boldsymbol{T}_t$, where the existing node $w_t \in [t-1]$ is chosen with probability

$$\frac{\beta D_{T_{t-1}}(w_t) + \alpha}{\beta 2(t-2) + \alpha(t-1)}. \tag{1}$$

To ensure that equation (1) is always non-negative, we require either $\alpha, \beta \geq 0$ or, if $\beta < 0$, then $\alpha = -c\beta$ for some integer $c > 0$. We may verify that (1) describes a valid probability distribution by noting that $\boldsymbol{T}_{t-1}$ always has $t-2$ edges and $t-1$ nodes. Before continuing onto the PAPER model, we consider some specific examples of APA trees:

1. Setting $\alpha = 1, \beta = 0$ means that we select w_t uniformly at random from $V(\boldsymbol{T}_{t-1})$. This yields the UA random tree. The resulting degree distribution has exponential tail and the maximum degree is of order $\log n$ (Addario-Berry & Eslava, 2018; Na & Rapoport, 1970).
2. Setting $\alpha = 0, \beta = 1$ means that we select w_t with probability proportional to the degree $D_{T_{k-1}}(w_t)$. This yields the LPA random tree. Linear preferential attachment has heavy-tailed degree distribution and a maximum degree is of order $\sqrt{n}$ (Bollobás et al., 2001; Peköz et al., 2014).
3. We may also set β as -1 and α as some positive integer so that the maximum degree of any node is α. This may be interpreted as a UA tree growing on top of a background infinite α-regular tree (Khim & Loh, 2017).

We may generalise Definition 1 by defining a nonparametric function $\phi : \mathbb{N} \to [0, \infty)$ and choose w_t with probability proportional to $\phi(D_{T_{t-1}}(w_t))$. In this paper however, we focus only on the case where ϕ is an affine function.

Definition 2 To model a general network, we define the PAPER(α, β, θ) (PAPER) model parametrised by $\alpha, \beta \in \mathbb{R}$ and $\theta \in [0, 1]$. We say that a random graph $\boldsymbol{G}_n$ distributed according to the PAPER(α, β, θ) model if

$$\boldsymbol{G}_n = \boldsymbol{T}_n + \boldsymbol{R}_n,$$

where $\boldsymbol{T}_n \sim$ APA(α, β) and $\boldsymbol{R}_n \sim$ Erdős–Rényi(θ) are independent random graphs defined on the same set of vertices $[n]$.

Since we collapse any multi-edges that occur when we add $\boldsymbol{R}_n$ to $\boldsymbol{T}_n$, we may view $\boldsymbol{R}_n$ equivalently as an ER random graph defined on potential edges excluding those already in the tree $\boldsymbol{T}_n$. The PAPER model can produce networks with either light-tailed or heavy-tailed degree distribution depending on the choice of the parameters α and β. It produces features that are commonly seen in real-world networks but absent from non-sequential models like SBM, such as pendants (a node with several degree-1 node attached to it) and chains of nodes; see Figure 2. It also assigns a non-zero probability to any connected graph, in contrast to the general preferential attachment graph model where a fixed $m > 1$ edges are added at every iteration

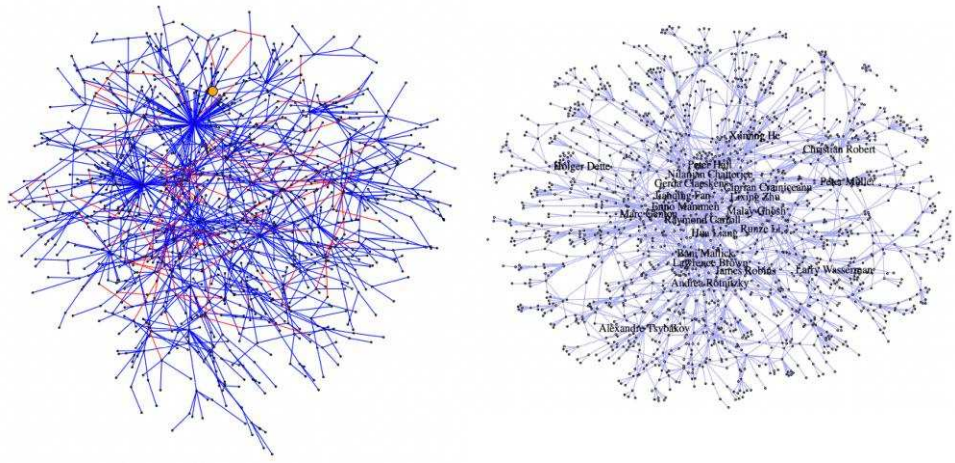


Figure 2. Left: PAPER graph with $\alpha = 1, \beta = 1$; **Right:** co-authorship graph from Ji and Jin (2016) (reprinted with permission from the Insitute of Mathematical Statistics).

(Barabási & Albert, 1999). In computer science terminology, $\boldsymbol{G}_n$ is a *planted tree model* where the signal $\boldsymbol{T}_n$ is planted in an ER random graph $\boldsymbol{R}_n$ in the same sense that SBM is often referred to as the planted partition model.

An alternative way to define the PAPER model is to specify the total number of edges m in the final graph and generate $\boldsymbol{R}_n$ as a uniformly random graph with $m - (n-1)$ edges (since a tree with n nodes always has $n-1$ edges). This is equivalent to the PAPER(α, β, θ) model where we *condition* on the event that the final graph $\boldsymbol{G}_n$ has m edges. To simplify exposition, we use PAPER to refer to this conditional model as well.

Remark 1 We may view the PAPER(α, β, θ) model as a Markovian process over a sequence of networks $\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_n$. We define the transition kernel $\mathbb{P}(\boldsymbol{G}_t \mid \boldsymbol{G}_{t-1})$ for $t \geq 3$ by first adding a new node labelled t, then adding a new tree edge (t, w_t) where w_t is chosen with probability (1), and then, for each existing node $j \in [t-1]$ not equal to w_t, we independently add a noise edge (t, j) with probability θ.

Interestingly, when $\alpha = 1$ and $\beta = 0$, we see that the PAPER model is the conditional distribution of an Erdős–Rényi graph $\boldsymbol{G}$ conditional on the event that, for some fixed ordering ρ of the nodes, the sequence of induced subgraphs $\boldsymbol{G} \cap \{\rho_1, \ldots, \rho_t\}$ for $t \in [n]$ are all connected. In Section 2.3, we extend the PAPER model so that the noise edge probability is allowed to depend on the time t and the state of the tree at time t.

Remark 2 Under APA(α, β) model, the probability of generating a given tree has a closed form expression: $\mathbb{P}(\boldsymbol{T}_n = \boldsymbol{t}_n) = \frac{\prod_{v \in [n]} \prod_{j=1}^{D_{t_n}(v)-1} (\beta j + \alpha)}{\prod_{t=3}^{n} 2(t-2)\beta + (t-1)\alpha}$. The important consequence is that the likelihood depends on the tree $\boldsymbol{t}_n$ only through its degree distribution $D_{t_n}(\cdot)$. Hence, any two trees with the same degree distribution has the same likelihood; Crane and Xu (2021) refer to this property as *shape-exchangeability*. We give the likelihood expression for the multiple roots models and the PAPER model in Section S1.1 of the online supplementary material.

Remark 3 It is known that the degree distribution of an APA(α, β) tree has an asymptotic limit. For example, if $\beta = 1$ and $\alpha > 0$, then we have by Van Der Hofstad (2016, Theorem 8.2) that $\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{D_{T_n}(t) = k\} \to \frac{2+\alpha}{3+2\alpha} \prod_{j=1}^{k-1} \frac{j+\alpha}{j+3+2\alpha}$ as $n \to \infty$ uniformly over all k. The limiting distribution is approximately a power law where the

number of nodes with degree k is proportional to $k^{-(3+\alpha)}$ (see Van Der Hofstad, 2016, Section 8.4). Since the ER graph $\boldsymbol{R}_n$ only adds an expected additional degree of at most $n\theta$ to every node, we see that, when θ is small, the PAPER graph can have heavy-tailed degree distribution without any additional degree correction parameters.

2.1.1 *Single root inference problem*

Let $\boldsymbol{G}_n \sim \text{PAPER}(\alpha, \beta, \theta)$ be a random graph. As the nodes of $\boldsymbol{G}_n$ are labelled by their arrival time, our observation is the unlabelled shape $\text{sh}(\boldsymbol{G}_n)$, that is, the network $\boldsymbol{G}_n$ with the labels removed. Our goal is to construct a subset of nodes that is guaranteed to contain the true root node (node with arrival time 1) with probability at least $1-\epsilon$. Since we need to refer to specific nodes of $\text{sh}(\boldsymbol{G}_n)$, we give the nodes of $\text{sh}(\boldsymbol{G}_n)$ names from an arbitrary alphabet $\mathcal{U}_n$ of n elements to form a labelled graph $\boldsymbol{G}_n^*$ such that $V(\boldsymbol{G}_n^*) = \mathcal{U}_n$. We take $\boldsymbol{G}_n^*$ as our observation from this point on.

We note that there exists an unobserved label bijection $\rho \in \text{Bi}([n], \mathcal{U}_n)$ such that $\rho \boldsymbol{G}_n = \boldsymbol{G}_n^*$. This unobserved ρ captures precisely the arrival time of the nodes in that for any time $t \in [n]$, the node with label ρ_t in $\boldsymbol{G}_n^*$ is exactly node with arrival time t in $\boldsymbol{G}_n$. In particular, node ρ_1 of the observed graph $\boldsymbol{G}_n^*$ is the true root node. To illustrate the setting clearly, we provide a concrete example in Figure 3.

Definition 3 For $\epsilon \in (0, 1)$, we say that a set $C_\epsilon(\boldsymbol{G}_n^*) \subset \mathcal{U}_n$ is a level $1-\epsilon$ confidence set for the root node if

$$\mathbb{P}\big(\rho_1 \in C_\epsilon(\boldsymbol{G}_n^*)\big) \geq 1 - \epsilon. \tag{2}$$

One may construct a trivial confidence set for the root nodes by taking the set of all the nodes. We aim therefore to make the confidence set $C_\epsilon(\cdot)$ as small as possible. Although we focus on the problem of root inference, the approach that we develop is applicable to more general problems such as inferring the first two or three nodes or inferring the arrival time of a particular node.

Remark 4 It is important to note that $\boldsymbol{G}_n^*$ may have multiple nodes that are indistinguishable once the node labels are removed, which may lead to the paradoxical scenario that which node of $\boldsymbol{G}_n^*$ correspond to the true root node depends on the choice of the label bijection ρ. Luckily, this is a technical issue that does not pose a problem so long as we restrict ourselves to confidence sets $C_\epsilon(\cdot)$ that are labelling equivariant in that they do not depend on the specific node labelling. Labelling equivariance is a very weak condition that only rules out confidence sets that can access side information about the nodes somehow.

Formally, we note that there may exist $\rho, \rho' \in \text{Bi}([n], \mathcal{U}_n)$ where $\rho_1 \neq \rho'_1$ but both satisfy $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n = \rho' \boldsymbol{G}_n$; in other words, root node can only be well-defined up to an automorphism. We illustrate a concrete example in Figure 4. We define $C_\epsilon(\cdot)$ to be *labelling equivariant* if, for all $\tau \in \text{Bi}(\mathcal{U}_n, \mathcal{U}_n)$,

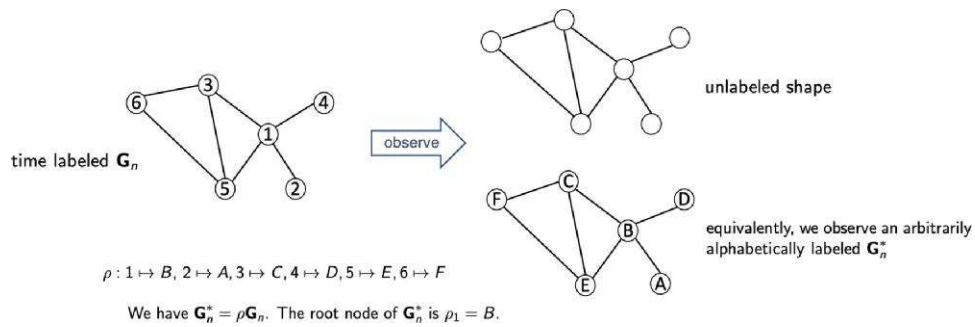


Figure 3. Our observation is the unlabelled shape or alphabetically labelled $\boldsymbol{G}_n^*$ instead of time labelled $\boldsymbol{G}_n$. There exists an unobserved ordering $\rho \in \text{Bi}([n], \mathcal{U}_n)$ such that $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n$.

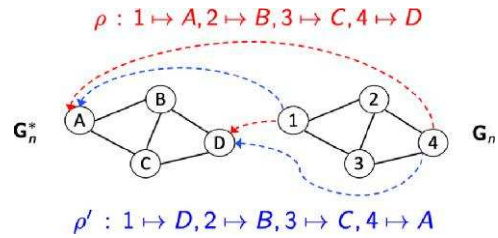


Figure 4. Both ρ and ρ' are distinct bijections in $\mathrm{Bi}([n], \mathcal{U}_n)$ but they both satisfy $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n = \rho' \boldsymbol{G}_n$. The root node is D according to ρ but A according to ρ'. Note that nodes A and D are indistinguishable if the labels are removed.

we have $\tau C_\epsilon(\boldsymbol{G}_n^*) = C_\epsilon(\tau \boldsymbol{G}_n^*)$; if the confidence set algorithm contains randomisation (to break ties for example), then we say it is labelling equivariant if $\tau C_\epsilon(\boldsymbol{G}_n^*) \overset{d}{=} C_\epsilon(\tau \boldsymbol{G}_n^*)$ for all $\tau \in \mathrm{Bi}(\mathcal{U}_n, \mathcal{U}_n)$. If a confidence set $C_\epsilon(\cdot)$ is labelling equivariant, then for any $\rho, \rho' \in \mathrm{Bi}([n], \mathcal{U}_n)$ such that $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n = \rho' \boldsymbol{G}_n$, we have that $(\rho' \circ \rho^{-1}) \boldsymbol{G}_n^* = \boldsymbol{G}_n^*$ and hence,

$$\rho_1 \in C_\epsilon(\boldsymbol{G}_n^*) \Leftrightarrow (\rho' \circ \rho^{-1})\rho_1 \in (\rho' \circ \rho^{-1}) C_\epsilon(\boldsymbol{G}_n^*) \Leftrightarrow \rho_1' \in C_\epsilon((\rho' \circ \rho^{-1})\boldsymbol{G}_n^*) \Leftrightarrow \rho_1' \in C_\epsilon(\boldsymbol{G}_n^*).$$

Therefore, the coverage probability (2) does not depend on the choice of ρ.

2.2 Multiple roots models

Many real-world networks have multiple communities that grow simultaneously form multiple sources. The APA model allows for only one root node in the graph but we can augment the model to describe networks that grow from multiple roots. When there are K roots, we start the growth process with an initial network of K singleton nodes and attach each new node to an existing node w_t with probability proportional to $\beta \cdot (\text{degree of } w_t) + \alpha$ as before.

However, one complication is that when $\alpha = 0$, the probability of attaching to a singleton node is 0. Thus, for convenience, we give each root node an unobserved imaginary self-loop edge for the purpose of computing the attachment probabilities.

Definition 4 We first define the APA(α, β, K) model for a random forest of K disjoint component trees: let $K \in \mathbb{N}$ and for $t \in S := \{1, 2, \ldots, K\}$ (the set S is the set of root nodes), let $\boldsymbol{F}_t$ be the set of singleton nodes $1, 2, \ldots, t$. For $t > K$, we define the transition kernel $\mathbb{P}(\boldsymbol{F}_t \mid \boldsymbol{F}_{t-1})$ in the following way: given $\boldsymbol{F}_{t-1}$, we add a new node t and a new random edge (t, w_t) where the existing node $w_t \in [t-1]$ is chosen with probability

$$\frac{\beta D_{F_{t-1}}(w_t) + 2\beta \mathbb{1}\{w_t \in S\} + \alpha}{(2\beta + \alpha)(t-1)}. \tag{3}$$

We then say that a random graph $\boldsymbol{G}_n \sim \mathrm{PAPER}(\alpha, \beta, K, \theta)$ if $\boldsymbol{G}_n = \boldsymbol{F}_n + \boldsymbol{R}_n$ where $\boldsymbol{F}_n \sim \mathrm{APA}(\alpha, \beta, K)$ and $\boldsymbol{R}_n \sim \mathrm{ER}_\theta$ is an Erdős–Rényi random graph independent of $\boldsymbol{F}_n$ defined on the same set of nodes $[n]$. We refer to this setting as the *fixed* K *setting*. In contrast, we refer to the PAPER(α, β, θ) model in Section 2.1 as the *single root setting*.

We can verify the normalisation term (3) by noting that each root node starts with one imaginary self-loop and that we add one node and one edge at every iteration. The theory of Polya's urn immediately implies that the number of nodes in each of the K component trees, divided by n, has the asymptotic distribution of Dirichlet($\frac{1}{K}, \ldots, \frac{1}{K}$).

To deal with networks in which the number of roots K is unknown, we propose a variation of the PAPER model with random K number of roots. We can express the model as a sequential

growth process where every newly arrived node has some probability of becoming a new root. Similar to the fixed K setting, we give each new root node an imaginary self-loop edge for the purpose of determining the attachment probabilities.

Definition 5 We first define the APA(α, β, α_0) model for a random forest graph: let $\boldsymbol{F}_1$ be a singleton node and let $S = \{1\}$. For $k > 1$, we define the transition kernel $\mathbb{P}(\boldsymbol{F}_t \mid \boldsymbol{F}_{t-1})$ in the following way: given $\boldsymbol{F}_{t-1}$, we add a new node t. With probability

$$\frac{\alpha_0}{(2\beta+\alpha)(t-1)+\alpha_0},$$

we let t be a new root node to form $\boldsymbol{F}_t$ and add t to set S. Or, we add a new edge (t, w_t) to $\boldsymbol{F}_{t-1}$ to obtain $\boldsymbol{F}_t$ where the existing node $w_t \in [t-1]$ is chosen with probability

$$\frac{\beta D_{\boldsymbol{F}_{t-1}}(w_t) + \alpha + 2\beta \mathbb{1}\{w_t \in S\}}{(2\beta+\alpha)(t-1)+\alpha_0}.$$

Note that the resulting set of root nodes $S \subset [n]$ of $\boldsymbol{F}_n$ is a random set.

We then say that a random graph $\boldsymbol{G}_n$ has the PAPER(α, β, α_0, θ) distribution if $\boldsymbol{G}_n = \boldsymbol{F}_n + \boldsymbol{R}_n$ where $\boldsymbol{F}_n \sim$ APA(α, β, α_0) and $\boldsymbol{R}_n \sim$ ER(θ) is an Erdős–Rényi random graph independent of $\boldsymbol{F}_n$ defined on the same set of nodes $[n]$. We refer to this setting as the *random K setting*.

In the random K setting, each node has some probability of becoming a new root node and creating a new component tree in the same way as the Dirichlet process mixture model, which is often called the Chinese restaurant process. Therefore, the expected number of component trees is $(1 + o(1))\frac{\alpha_0}{(2\beta+\alpha)} \log n$ (Crane, 2016, Section 2.2).

2.2.1 *Multiple roots inference problem*

We observe $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n$ for an unknown label bijection $\rho \in \mathrm{Bi}([n], \mathcal{U}_n)$. In both the APA($\alpha$, β, K) and the APA(α, β, α_0) models, the root nodes is a set S which is fixed to be $[K]$ in the first model and random in the second model. Intuitively, we interpret S as a set of *local* roots, where each root is central with respect to a specific community or sub-network represented by a component tree in the forest $\boldsymbol{F}_n$ in Definition 4 or 5. The root inference problem is then, for a given $\epsilon \in (0, 1)$, to construct a confidence set $C_\epsilon(\boldsymbol{G}_n^*)$ such that

$$\mathbb{P}\big(\rho S \subseteq C_\epsilon(\boldsymbol{G}_n^*)\big) \geq 1 - \epsilon.$$

We illustrate this notion of local roots in a synthetic example in Figure 5.

Remark 5 (Interpretation of community under the PAPER model). The disjoint component trees of $\boldsymbol{F}_n$ induce a community structure on the graph $\boldsymbol{G}_n$. This way of modelling community by adding Erdős–Rényi noise to disjoint subgraphs follows the same spirit as SBM: an SBM with K communities, p as the within-community edge probability, and $q < p$ as the between-community edge probability can be similarly defined as first generating K disjoint ER($\frac{p-q}{1-q}$) graphs on each of the communities and then taking the union of that with ER(q) noisy edges on all the nodes, collapsing multi-edges.

The PAPER notion of community is however different from that described by SBM. The PAPER notion of community is based on Markovian growth process and intuitively characterised by the imbalance of spanning trees on a network, that is, we believe a network to contain multiple communities if the spanning trees of the network tend to be highly imbalanced (see

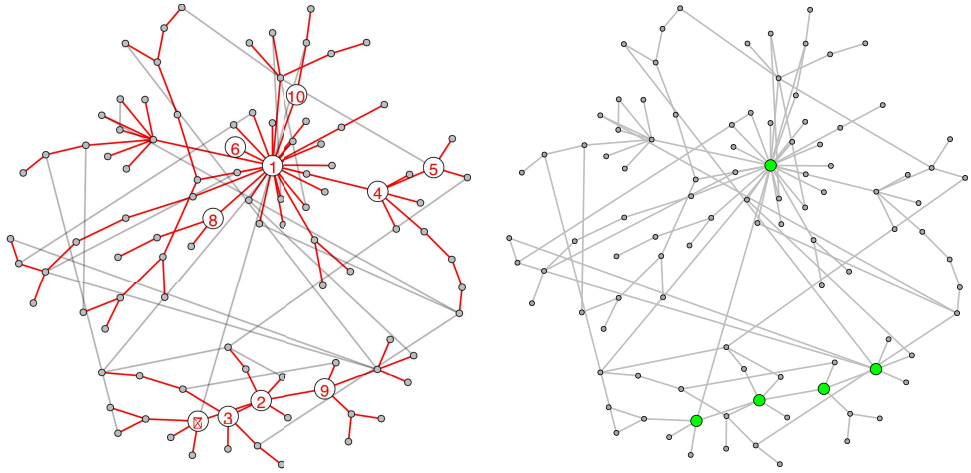


Figure 5. Left: Illustration of PAPER model with $K = 2$ underlying trees; nodes have latent time ordering (only first 10 orderings shown); the dark red edges form the latent tree, while light grey edges are Erdö s–Rényi. Right: 80% confidence set for the set of root nodes (node number 1 for tree 1 and node number 2 for tree 2) constructed from the unlabelled graph.

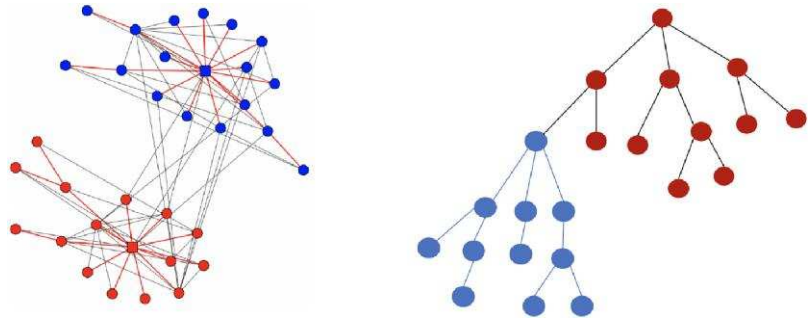


Figure 6. The karate club network (left) has two true communities. Most spanning trees of the whole karate club network would be imbalanced (such as the tree on the right), showing that the karate club network is very unlikely to have been formed from a single homogeneous growth process and hence very likely to contain multiple communities.

Figure 6), which would suggest that the network is very unlikely to have been formed from a single homogeneous growth process.

The PAPER model also produces more within-community edges than between-community edges because each community has a spanning tree. However, since a tree on n nodes only has $n - 1$ edges, the difference in the within-community edge density and the between-community edge density is diminishingly small when the noise level θ is of an order larger than $\omega(\frac{1}{n})$. In this case, the peripheral leaf nodes of a community-tree become impossible to cluster but it is still possible to recover the root node of each of the community-trees, as our experimental results show. One disadvantage of the PAPER notion of community is that it is not able to capture non-assortative clusters where nodes in the same clusters are unlikely to form edges.

The PAPER notion of community is appropriate in many application. For example, for a co-authorship network where there exists an underlying growth process, our empirical analysis in Section 6.5 shows that the PAPER model captures clusters that accurately reflect salient research communities. We can also combine both notions by a PAPER–SBM mixture model, where we generate a

preferential attachment forest $\boldsymbol{F}_n$ via the mechanism described in Definition 4 or 5, then, for every pair of nodes u and v, we add a noisy edge (u, v) with probability θ_1 if u and v belong to the same tree in $\boldsymbol{F}_n$ and with a different probability θ_2 if u and v belong to different trees. The inference method and algorithm that we develop in this manuscript can extend to such a PAPER–SBM mixture model, but the computational run-time would be substantially slower. We relegate a detailed study of a PAPER–SBM mixture model to a future work.

2.3 Sequential noise models

As suggested in Remark 1, PAPER model is a special case of a general Markovian process over a sequence of networks $\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_n$ based on a latent sequence of trees $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n$. In the general framework, we specify the transition kernel $\mathbb{P}(\boldsymbol{G}_t \mid \boldsymbol{G}_{t-1})$ by specifying two stages:

1. (tree stage) $\mathbb{P}(\boldsymbol{T}_t \mid \boldsymbol{T}_{t-1}, \boldsymbol{G}_{t-1})$ which adds one node t and one tree edge and
2. (noise stage) $\mathbb{P}(\boldsymbol{G}_t \mid \boldsymbol{T}_t, \boldsymbol{G}_{t-1})$ which adds more random edges to obtain $\boldsymbol{G}_t$.

We can of course define $\mathbb{P}(\boldsymbol{G}_t \mid \boldsymbol{G}_{t-1})$ without having an underlying tree but the key insight of our approach is that augmenting the model with the latent tree $\boldsymbol{T}_n$ greatly facilitates the design of tractable models and inference algorithms because calculations on trees are easy and fast. In addition, the latent tree has a real-world interpretation as the recruitment history—a tree edge between nodes (u, v) implies that node u recruited node v into the network.

In the noise stage, if we independently adds noise edges between the new node t and the existing nodes with the same probability θ, then we get back the single root PAPER model. More generally, we can let the noise edge probability depend on the time t and the state of the graph at time t. We define the following extension which we refer to as the seq-PAPER model with parameters $(\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta})$:

Definition 6 We start with a singleton root node $\boldsymbol{T}_1 = \boldsymbol{G}_1 = \{1\}$. At time $t = 2$, we add node 2 and attach it to node 1. At time $t \geq 3$:

1. (tree stage) We add new node t; we select node an existing node $w_t \in [t-1]$ with probability $\frac{\beta D_{T_{t-1}}(w_t)+\alpha}{2(t-2)\beta+(t-1)\alpha}$ and add edge (t, w_t) to $\boldsymbol{T}_{t-1}$ to form $\boldsymbol{T}_t$;
2. (noise stage) for each existing node $j \in [t-1]$, we add edge (t, j) independently with probability

$$q_j := \theta \frac{\tilde{\beta} D_{T_{t-1}}(j) + \tilde{\alpha}}{2(t-2)\tilde{\beta} + (t-1)\tilde{\alpha}} \wedge 1. \tag{4}$$

It is possible that we add the tree edge (j, w_t) in the noise stage in which case we collapse the multi-edge.

In general, we may take $\tilde{\beta} = \beta$ and $\tilde{\alpha} = \alpha$ but we allow them to be distinct in the model definition for greater flexibility. We discuss parameter estimation in Section S3.5.4 of the online supplementary material.

When t is large, the independent Bernoulli generative process approximates a Poisson growth model (see, e.g. Sheridan et al., 2008) where we first generate $M \sim \text{Poisson}(\theta)$, and then repeat M times the procedure where we draw an existing node $j \in [t-1]$ with probability q_j (also with replacement) and then add the edge (t, j) to the random network, collapsing multi-edges if any are formed. We thus add an average of approximately θ noise edges at each time step. In contrast, under the PAPER model where the noise edge probability is θ, we add on average $(t-2) \cdot \theta$ noise edges at time t.

The approximation error between the Bernoulli mechanism and the Poisson mechanism, in each iteration t, converges to 0 in total variation distance as t increases; see rigorous statement and proof in Proposition S4 of Section S1.2 in the online supplementary material. However, it is important to note that the two mechanisms could still produce final random graphs whose overall distributions have total variation distance bounded away from 0. For example, UA or LPA trees

are known to be sensitive to initialisation so that different initial seeds could lead to very different distributions over the final observed graph, see, e.g. Bubeck et al. (2015) and Curien et al. (2015). In this work, we prefer the Bernoulli generative process in order to simplify the inference algorithm. Even with the Bernoulli approximation however, inference under the sequential setting is much more computationally intensive than the vanilla PAPER model.

A more realistic extension of the seq-PAPER model is to replace the tree degree $D_{T_{t-1}}(j)$ with the graph degree $D_{G_{t-1}}(j)$ in the noise probability 4. This small change unfortunately leads to additional significant slowdown in the resulting inference algorithm; see Remark 9 for more detail. We note that an even more sophisticated model of sequential noise is one where the additional noise edges are generated by a random-walk mechanism (Bloem-Reddy & Orbanz, 2018); Bloem-Reddy and Orbanz (2018) propose a sequential Monte Carlo inference method which may not scale well to large networks.

We have so far considered additive noise where new edges are added to the network. We can also model deletion noise where each tree edge is removed from the observed network independently with some probability $\eta > 0$. Having deletion noise under the vanilla PAPER model can adversely increase the size of the confidence set for the root node. However, the seq-PAPER model is much more resilient to deletion noise, especially when $\tilde{\beta} = \beta$ and $\tilde{\alpha} = \alpha$ since the noise edges also contain sequential information. To be precisely, we define the seq-PAPER$^*(\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta}, \eta)$ as the model where we first generate $\boldsymbol{G}_n$ according to the seq-PAPER$(\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta})$ model with latent spanning tree $\boldsymbol{T}_n$; we then remove each edge of $\boldsymbol{T}_n$ from the final graph $\boldsymbol{G}_n$ independently with probability η.

2.4 Related work

Many researchers in statistics (Kolaczyk, 2009), computer science (Bollobás et al., 2001), engineering, and physics (Callaway et al., 2000) have been interested in the probabilistic properties of various random growth processes of networks, including the preferential attachment model (Barabási & Albert, 1999). Recently, however, the specific problem of root inference on trees has received increased attention.

These efforts began with the ground-breaking work of Bubeck, Devroye et al. (2017), Bubeck et al. (2015), and Bubeck, Eldan et al. (2017), which shows that, given an observation of an LPA or UA tree of size n, for any $\epsilon \in (0, 1]$, one can construct asymptotically valid confidence sets for the root node with size $K_{LPA}(\epsilon)$ and $K_{UA}(\epsilon)$ for LPA or UA trees respectively. Importantly and surprisingly, $K_{LPA}(\epsilon)$ and $K_{UA}(\epsilon)$ do not depend on n so that the confidence set have size that is $O(1)$. To construct the confidence sets, Bubeck, Devroye et al. (2017) compute a centrality value for every node, which can for instance be based on inverse of the size of the maximum subtree of a node (a concepted sometimes called Jordan centrality on trees, different from the notion of a Jordan centre, which is the node with the minimum farthest distance to the other nodes); they then sort the nodes by centrality and take the top $K(\epsilon)$ nodes where the size $K(\epsilon)$ is determined by probabilistic bounds.

Khim and Loh (2017) further extend these results to the setting of UA over an infinite regular tree. Banerjee and Bhamidi (2020) improve the analysis of Jordan centrality on trees and derives tight upper and lower bounds on the confidence set size. Devroye and Reddad (2018) and Lugosi and Pereira (2019) study the more general problem of seed-tree inference instead of root node inference. The aforementioned results apply only to tree shaped networks but very recently, Banerjee and Huang (2021) study confidence sets constructed from the degrees of the nodes which applies to preferential attachment models in which a fixed m edges are added at every iteration. After the completion of this paper, Briend et al. (2022) propose confidence sets for the root node on a class of UA-based general Markovian graphs by detecting anchors of double-cycle subgraphs within the network; they show the confidence set sizes to be $O(1)$ and give explicit bounds in terms of confidence level ϵ.

A line of work in the physics literature also explores the problem of full or partial recovery of a tree network history (Cantwell et al., 2019; Sreedharan et al., 2019; Young et al., 2019). In computer science and engineering, researchers have studied the related problem of estimating the source of an infection spreading over a background network Shah and Zaman (2011), Fioriti et al. (2014), and Shelke and Attar (2019), with approaches that range from using Jordan centres, eigenvector centrality, and belief propagation (see survey in Jiang et al., 2016).

3 Methodology

Our approach to root inference and related problems is to randomise the node labels, which induces a posterior distribution over the latent ordering.

3.1 Label randomisation

Suppose $\boldsymbol{G}_n$ is a time labelled graph distributed according to a PAPER model and $\boldsymbol{G}_n^*$ is the alphabetically labelled observation where $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n$ for some label bijection $\rho \in \mathrm{Bi}([n], \mathcal{U}_n)$. We may independently generate a random bijection $\Lambda \in \mathrm{Bi}(\mathcal{U}_n, \mathcal{U}_n)$ and apply it to $\boldsymbol{G}_n^*$ to obtain a randomly labelled graph

$$\tilde{\boldsymbol{G}}_n := \Lambda \boldsymbol{G}_n^* = \underbrace{(\Lambda \circ \rho)}_{\Pi} \boldsymbol{G}_n.$$

By defining $\Pi = \Lambda \circ \rho$, we see that $\tilde{\boldsymbol{G}}_n = \Pi \boldsymbol{G}_n$ where Π is a random bijection drawn uniformly in $\mathrm{Bi}([n], \mathcal{U}_n)$ independently of $\boldsymbol{G}_n$ (see Figure 7). We define the randomly labelled latent forest $\tilde{\boldsymbol{F}}_n = \Pi \boldsymbol{F}_n$. We may view label randomisation as an augmentation of the probability space. An outcome of a PAPER model is a time labelled graph $\boldsymbol{g}_n$ whereas an outcome after label randomisation is a pair $(\tilde{\boldsymbol{g}}_n, \pi)$ where $\tilde{\boldsymbol{g}}_n$ is an alphabetically labelled graph and π is an ordering of the nodes. See Table 1 for a summary of the notation. We now make two simple but important observations regarding label randomisation.

Our first key observation is that, with respect to $\tilde{\boldsymbol{G}}_n$, the random labelling Π describes the arrival time of the nodes in the sense that if $\Pi_t = u$, then the node with alphabetical label u in $\tilde{\boldsymbol{G}}_n$ has the true arrival time t. Therefore, in the single root setting, we may infer the root node if we can infer Π_1; in the multiple roots setting, we may infer the set of root nodes if we can infer ΠS.

Our second key observation is that label randomisation allows us to define the posterior distribution

$$\mathbb{P}(\Pi = \pi \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) = \frac{\mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \,|\, \Pi = \pi)}{\sum_{\pi' \in \mathrm{Bi}([n], \mathcal{U}_n)} \mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \,|\, \Pi = \pi')} \tag{5}$$

which follows because $\mathbb{P}(\Pi = \pi) = \frac{1}{n!}$. This posterior distribution is supported on the subset of bijection π such that $\pi^{-1} \tilde{\boldsymbol{g}}_n$ has non-zero probability under the PAPER model. In the case of the single root PAPER or seq-PAPER model, the support of equation (5) has a simple characterisation: for every time point $t \in [n]$, define $\pi_{1:t} \cap \tilde{\boldsymbol{g}}_n$ as the subgraph of $\tilde{\boldsymbol{g}}_n$ restricted to nodes in $\pi_{1:t}$. Then, $\mathbb{P}(\Pi = \pi \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) > 0$ if and only if $\pi_{1:t} \cap \tilde{\boldsymbol{g}}_n$ is connected for all $t \in [n]$.

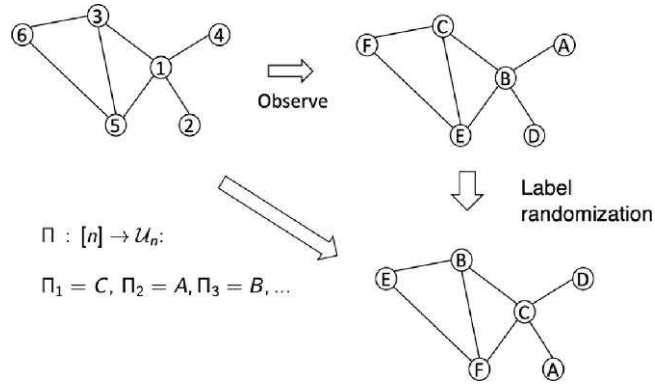


Figure 7. Label randomisation induces a random latent arrival ordering Π.

Table 1. Quick reference of important notation and definitions

G_n	Time labelled graph (unobserved)	F_n	Latent time labelled forest
G_n^*	Observed alphabetically labelled graph	F_n^*	Latent alphabetically labelled forest
$\tilde{G}_n$	Randomly alphabetically labelled graph	$\tilde{F}_n$	Latent randomly alphabetically labelled forest
ρ	Fixed unobserved ordering; $G_n^* = \rho G_n$	Π	Latent random ordering; $\tilde{G}_n = \Pi G_n$
S	Time labelled root nodes of G_n	$\tilde{S}$	Latent alphabetically labelled root nodes; $\tilde{S} = \Pi S$

From a Bayesian perspective, label randomisation adds a uniform prior distribution on the arrival ordering of the nodes in the observed alphabetically labelled graph G_n^*; this is sometimes used in Bayesian parameter inference on network models (Bloem-Reddy et al., 2018; Sheridan et al., 2012). This prior however is not subjective. Indeed, we will see in Theorem 7 that Bayesian inference statements in our setting directly have frequentist validity as well and, from online supplementary Section S2.1, that the posterior root probability of a node is equal to the likelihood of that node being the root node up to normalisation.

We describe how to compute equation (5) tractably in Section 4. For computation, we will also be interested in the posterior probability over both the ordering Π as well as the latent forest $\tilde{F}_n$:

$$\mathbb{P}(\Pi = \pi, \tilde{F} = \tilde{f}_n \,|\tilde{G}_n = \tilde{g}_n). \tag{6}$$

In the single root setting, $\tilde{f}_n$ is actually a tree, which we may write as $\tilde{t}_n$. It is then clear that equation (6) is non-zero only if $\tilde{t}_n$ is a *spanning tree* of $\tilde{g}_n$, i.e. $\tilde{t}_n$ is a connected subtree of $\tilde{g}_n$ that contains all the vertices.

3.2 Confidence set for the single root

To make the idea clear, we first consider the single root model. Since the root node is the node labelled Π_1 after label randomisation, a natural approach is to first construct a level $1-\epsilon$ Bayesian *credible set* for the node Π_1 by using its posterior distribution, which we call the posterior root distribution.

More concretely, let $\tilde{g}_n$ be an alphabetically labelled graph. For each node $u \in \mathcal{U}_n$ of $\tilde{g}_n$, we define the posterior root probability as $\mathbb{P}(\Pi_1 = u \,|\, \tilde{G}_n = \tilde{g}_n)$. We sort the nodes $u_1, \ldots, u_n$ so that

$$\mathbb{P}(\Pi_1 = u_1 \,|\tilde{G}_n = \tilde{g}_n) \geq \mathbb{P}(\Pi_1 = u_2 \,|\tilde{G}_n = \tilde{g}_n) \cdots \geq \mathbb{P}(\Pi_1 = u_n \,|\tilde{G}_n = \tilde{g}_n),$$

and define

$$L_\epsilon(\tilde{g}_n) = \min\left\{k \in [n] : \sum_{i=1}^{k} \mathbb{P}(\Pi_1 = u_i \,|\tilde{G}_n = \tilde{g}_n) \geq 1-\epsilon\right\}. \tag{7}$$

We then define the ϵ-credible set as

$$B_\epsilon(\tilde{g}_n) = \{u_1, u_2, \ldots, u_{L_\epsilon(\tilde{g}_n)}\}, \qquad \text{(breaking ties at random)}. \tag{8}$$

By definition, $B_\epsilon(\tilde{g})$ is the smallest set of nodes with Bayesian coverage at level $1-\epsilon$ in that $\mathbb{P}(\Pi_1 \in B_\epsilon(\tilde{g}_n) \,|\, \tilde{G}_n = \tilde{g}_n) \geq 1-\epsilon$. In general, credible sets do not have valid frequentist confidence coverage. However, our next theorem shows that in our setting, the credible set B_ϵ is in fact an honest confidence set in that $\mathbb{P}\{\text{root node} \in B_\epsilon(G_n^*)\} \geq 1-\epsilon$.

Theorem 7 Let $G_n \sim$ PAPER(α, β, θ) or seq-PAPER(α, β, θ, $\tilde{\alpha}$, $\tilde{\beta}$) and let G_n^* be the alphabetically labelled observation. Let $\rho \in \mathrm{Bi}([n], \mathcal{U}_n)$ be any label bijection such

that $\rho \mathbf{G}_n = \mathbf{G}_n^*$. We have that, for any $\epsilon \in (0, 1)$,

$$\mathbb{P}\{\rho_1 \in B_\epsilon(\mathbf{G}_n^*)\} \geq 1 - \epsilon.$$

The proof is very similar to that of Crane & Xu (2021, Theorem 1). Since the proof is short, we provide it here for readers' convenience.

Proof. We first claim that $B_\epsilon(\cdot)$ is labelling equivariant (cf. Remark 4) in the sense that for any $\tau \in \mathrm{Bi}(\mathcal{U}_n, \mathcal{U}_n)$ and any alphabetically labelled graph $\tilde{\boldsymbol{g}}_n$, we have that $\tau B_\epsilon(\tilde{\boldsymbol{g}}_n) \overset{d}{=} B_\epsilon(\tau \tilde{\boldsymbol{g}}_n)$ (note that $B_\epsilon(\cdot)$ uses randomisation to break ties). Indeed, since $(\Pi, \tilde{\boldsymbol{G}}_n) \overset{d}{=} (\tau^{-1} \circ \Pi, \tau^{-1} \tilde{\boldsymbol{G}}_n)$, we have that, for any $u \in \mathcal{U}_n$,

$$\mathbb{P}(\Pi_1 = u \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) = \mathbb{P}(\Pi_1 = \tau(u) \,|\, \tilde{\boldsymbol{G}}_n = \tau \tilde{\boldsymbol{g}}_n).$$

Therefore, for any $u, v \in \mathcal{U}_n$, we have that $\mathbf{P}(\Pi_1 = u \,|\, \tilde{\mathbf{G}}_n = \tilde{\mathbf{g}}_n) \geq \mathbf{P}(\Pi_1 = v \,|\, \tilde{\mathbf{G}}_n = \tilde{\mathbf{g}}_n)$ if and only if $\mathbf{P}(\Pi_1 = \tau(u) \,|\, \tilde{\mathbf{G}}_n = \tau \tilde{\mathbf{g}}_n) \geq \mathbf{P}(\Pi_1 = \tau(v) \,|\, \tilde{\mathbf{G}}_n = \tau \tilde{\mathbf{g}}_n)$. Since $B_\epsilon(\mathbf{G}_n^*)$ is constructed by taking the top elements of $\mathcal{U}_n$ that maximise the cumulative posterior root probability, the claim follows.

Now, let $\rho \in \mathrm{Bi}([n], \mathcal{U}_n)$ be such that $\rho \mathbf{G}_n = \mathbf{G}_n^*$ and let Λ be a random bijection drawn uniformly in $\mathrm{Bi}(\mathcal{U}_n, \mathcal{U}_n)$ and let $\Pi = \Lambda \circ \rho$. Then,

$$\begin{aligned}
\mathbb{P}(\rho_1 \in B_\epsilon(\mathbf{G}_n^*)) &= \mathbb{P}(\rho_1 \in B_\epsilon(\rho \mathbf{G}_n)) \\
&= \mathbb{P}\{(\Lambda \circ \rho)_1 \in B_\epsilon((\Lambda \circ \rho)\mathbf{G}_n) \,|\, \Lambda = \mathrm{Id}\} \\
&= \mathbb{P}\{(\Lambda \circ \rho)_1 \in B_\epsilon((\Lambda \circ \rho)\mathbf{G}_n)\} \\
&= \mathbb{P}(\Pi_1 \in B_\epsilon(\tilde{\mathbf{G}}_n)) \geq 1 - \epsilon,
\end{aligned}$$

where the penultimate equality follows from the labelling equivariance of B_ϵ and where the last inequality follows because $\mathbf{P}(\Pi_1 \in B_\epsilon(\tilde{\mathbf{G}}_n) \,|\, \tilde{\mathbf{G}}_n = \tilde{\mathbf{g}}_n) \geq 1 - \epsilon$ for any labelled tree $\tilde{\mathbf{g}}_n$ (with labels in $\mathcal{U}_n$) by the definition of B_ϵ. □

Remark 6 We show in Theorem S5 of the online supplementary material that the posterior root probability $\mathbb{P}(\Pi_1 = u \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)$ is equal to the likelihood of node u being the root node on observing the unlabelled shape of $\tilde{\boldsymbol{g}}_n$. Therefore, the set $B_\epsilon(\tilde{\boldsymbol{g}}_n)$ is in fact the maximum likelihood confidence set. Because the likelihood in this setting is complicated to even write down, we leave all the details to Section S2.1 of the online supplementary material.

Remark 7 One may see from the proof that Theorem 7 applies more broadly then just PAPER models. It in fact applies to any random graph $\boldsymbol{G}_n$ whose nodes are labelled by $\{1, 2, \ldots, n\}$. For the PAPER model, the integer labels encode arrival time and thus contain information about the graph. In a model where the integer labels are uninformative of the graph connectivity structure, Theorem 7 is still valid although the posterior probability $\mathbb{P}(\Pi_1 = \cdot \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)$ would be uniform. A reviewer of this paper also pointed out that Theorem 7 is related to the classical literature on invariant/equivariant estimation where credible sets constructed from uniform (Haar) priors may also be valid confidence sets; see, e.g. Schervish (1995, Theorem 6.78).

3.3 Confidence set for multiple roots

First consider the fixed K setting where $\boldsymbol{G}_n \sim \mathrm{PAPER}(\alpha, \beta, \theta, K)$; let Π be a uniformly random ordering in $\mathrm{Bi}([n], \mathcal{U}_n)$ and let $\tilde{\boldsymbol{G}}_n = \Pi \boldsymbol{G}_n$. The latent set of root nodes of $\tilde{\boldsymbol{G}}_n$ in this case is

$\tilde{S} := \Pi S = \{\Pi_1, \ldots, \Pi_K\}$. We then define the posterior root probability for any node $u \in \mathcal{U}_n$ as

$$\mathbb{P}(u \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n),$$

that is, the probability that node u is an element of the latent root set $\tilde{S}$.

To form the credible set $B_\epsilon(\tilde{\boldsymbol{g}}_n) \subseteq \mathcal{U}_n$, we sort the nodes by the posterior root probabilities

$$\mathbb{P}(u_1 \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \geq \mathbb{P}(u_2 \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \geq \cdots \geq \mathbb{P}(u_n \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n). \tag{9}$$

We may then take $B_\epsilon(\tilde{\boldsymbol{g}}_n)$ to be the smallest set of nodes such that $P(\tilde{S} B_\epsilon(\tilde{\boldsymbol{g}}_n) \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \leq \epsilon$. More precisely, define the integer

$$L_\epsilon(\tilde{\boldsymbol{g}}_n) = \min\left\{ k \in [n] : \sum_{i=k+1}^{n} \mathbb{P}(u_i \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \leq \epsilon \right\} \tag{10}$$

and then define the credible set as

$$B_\epsilon(\tilde{\boldsymbol{g}}_n) = \left\{u_1, u_2, \ldots, u_{L_\epsilon(\tilde{\boldsymbol{g}}_n)}\right\} \quad \text{(breaking ties at random)}. \tag{11}$$

In the PAPER$(\alpha, \beta, \alpha_0, \theta)$ model where the number of roots K is random, the set of root nodes is $\tilde{S} = \Pi S$ which comprises, according to the ordering Π, of the node that is first to arrive in each of the component trees of $\tilde{\boldsymbol{F}}_n$. We may then sort the nodes as in equation (9), compute $L_\epsilon(\tilde{\boldsymbol{g}}_n)$ as in equation (10), and $B_\epsilon(\tilde{\boldsymbol{g}}_n)$ as in equation (11).

Similar to Theorem 7, we may show that $B_\epsilon(\cdot)$ in fact also has frequentist coverage at the same level $1 - \epsilon$.

Theorem 8 Let $\boldsymbol{G}_n \sim$ PAPER$(\alpha, \beta, K, \theta)$ or PAPER$(\alpha, \beta, \alpha_0, \theta)$ and let $\boldsymbol{G}_n^*$ be the alphabetically labelled observation. Let $\rho \in \mathrm{Bi}([n], \mathcal{U}_n)$ be any label bijection such that $\rho \boldsymbol{G}_n = \boldsymbol{G}_n^*$ and let $S \subset [n]$ be the time labels of the root nodes (see Definitions 4 and 5). We have that, for any $\epsilon \in (0, 1)$,

$$\mathbb{P}\left\{\rho S \subseteq B_\epsilon(\boldsymbol{G}_n^*)\right\} \geq 1 - \epsilon.$$

Proof. The proof is very similar to that of Theorem 7. First, since the random set $\tilde{S}$ is a function of the random ordering Π in the fixed K setting and a function of both the random ordering Π and the random forest $\tilde{\boldsymbol{F}}_n$, we write $\tilde{S}(\Pi)$ or $\tilde{S}(\Pi, \tilde{\boldsymbol{F}}_n)$ to be precise.

We then observe that $\tilde{S}(\Pi)$ in the fixed K setting or $\tilde{S}(\Pi, \tilde{\boldsymbol{F}}_n)$ in the random K setting, are labelling equivariant in that for any $\tau \in \mathrm{Bi}(\mathcal{U}_n, \mathcal{U}_n)$, we have that $\tilde{S}(\tau^{-1}\Pi) = \tau^{-1}\tilde{S}(\Pi)$ or, in the random K setting, $\tilde{S}(\tau^{-1}\Pi, \tau^{-1}\tilde{\boldsymbol{F}}_n) = \tau^{-1}\tilde{S}(\Pi, \tilde{\boldsymbol{F}}_n)$. Therefore, since $(\Pi, \tilde{\boldsymbol{G}}_n) \stackrel{d}{=} (\tau^{-1}\Pi, \tau^{-1}\tilde{\boldsymbol{G}}_n)$ for any $\tau \in \mathrm{Bi}(\mathcal{U}_n, \mathcal{U}_n)$, we have $\tilde{S}(\Pi, \tilde{\boldsymbol{F}}_n) \stackrel{d}{=} \tau^{-1}\tilde{S}(\Pi, \tilde{\boldsymbol{F}}_n)$ and thus, for any $u \in \mathcal{U}_n$,

$$\mathbb{P}(u \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) = \mathbb{P}(\tau(u) \in \tilde{S} \,|\, \tilde{\boldsymbol{G}}_n = \tau\tilde{\boldsymbol{g}}_n).$$

The rest the proof proceeds in an identical manner to that of Theorem 7. □

When there are multiple roots, an alternative way of inferring the root set is to construct the confidence set $B_\epsilon(\cdot)$ as a set of subsets of the nodes and then require that $\tilde{S} \in B_\epsilon$ with probability at least $1 - \epsilon$. We can take the same approach to construct such confidence set over sets but it becomes much more computationally intensive to compute them in practice.

3.4 Combinatorial interpretation

Before we describe the Gibbs sampling algorithm for computing the posterior root probabilities $\mathbb{P}(\Pi_1 = u \mid \tilde{G}_n = \tilde{g}_n)$, we provide an intuitive combinatorial interpretation of the posterior root probability in the single root PAPER model (Definition 2). The definitions and calculations here are also important for deriving the algorithm in Section 4.

3.4.1 The noiseless case

We first consider the simpler setting in which we can observe the tree $\tilde{T}_n$ (with a single root) distributed according to the APA model. In this case, we have

$$\mathbb{P}(\Pi_1 = \cdot \mid \tilde{T}_n = \tilde{t}_n) = \sum_{\pi:\pi_1=u} \mathbb{P}(\Pi = \pi \mid \tilde{T}_n = \tilde{t}_n).$$

Recall that $\tilde{T}_n = \Pi T_n$ where T_n is a random time labelled tree with APA(α, β) distribution and Π is an independent uniformly random ordering in Bi($[n], \mathcal{U}_n$). The distribution $\mathbb{P}(\Pi = \pi \mid \tilde{T}_n = \tilde{t}_n)$ is supported on a subset of the the bijections Bi($[n], \mathcal{U}_n$) because $\pi^{-1}\tilde{T}_n$ must be a valid time labelled tree (also called *recursive tree* in discrete mathematics). To be precise, we define the histories of $\tilde{t}_n$ as

$$\begin{aligned} \mathrm{hist}(\tilde{t}_n) &:= \left\{\pi \in \mathrm{Bi}([n], \mathcal{U}_n) : \mathbb{P}(T_n = \pi^{-1}\tilde{t}_n) > 0\right\}, \text{ and} \\ h(\tilde{t}_n) &:= |\mathrm{hist}(\tilde{t}_n)| \end{aligned}$$

as the number of distinct histories. Since the APA tree distribution assigns a non-zero probability to any valid time labelled trees, we see that $\mathrm{hist}(\tilde{t}_n)$ contains the elements π of Bi($[n], \mathcal{U}_n$) such that for all $t \in [n]$, the subtree restricted only to nodes in $\pi_{1:t}$, i.e. $\tilde{t}_n \cap \pi_{1:t}$, is connected. Thus, $\mathrm{hist}(\tilde{t}_n)$ is the set of bijections π which represent a valid arrival ordering for the nodes of the given tree $\tilde{t}_n$. Similarly, we define, for any node $u \in \mathcal{U}_n$,

$$\begin{aligned} \mathrm{hist}(u, \tilde{t}_n) &:= \left\{\pi \in \mathrm{hist}(\tilde{t}_n) : \pi_1 = u\right\} \\ h(u, \tilde{t}_n) &:= |\mathrm{hist}(u, \tilde{t}_n)|, \end{aligned}$$

as histories of $\tilde{t}_n$ that start at node u. We illustrate an example of the set of histories for a simple tree in Figure 8.

By definition, $\mathbb{P}(\Pi = \cdot \mid \tilde{T}_n = \tilde{t}_n)$ is supported on $\mathrm{hist}(\tilde{t}_n)$. For most values of α and β, the posterior distribution is in fact uniform over $\mathrm{hist}(\tilde{t}_n)$:

Proposition 9 (Crane & Xu, 2021, Theorem 4 and Proposition 3). Let α, β be two real numbers such that either (1) $\beta \geq 0$ and $\alpha \geq -\beta$ or (2) $\beta < 0$ and $\alpha = -D\beta$ for some integer $D \geq 2$. Suppose $T_n \sim \mathrm{APA}(\alpha, \beta)$. Let Π be a uniformly random ordering taking value in Bi($[n], \mathcal{U}_n$) and let $\tilde{T}_n = \Pi T_n$. Then,

$$\mathbb{P}(\Pi = \pi \mid \tilde{T}_n = \tilde{t}_n) = \frac{1}{h(\tilde{t}_n)} 1\{\pi \in \mathrm{hist}(\tilde{t}_n)\}. \tag{12}$$

The full proof of Proposition 9 is in Crane and Xu (2021) but we give a short justification here: the posterior is uniform because $\mathbb{P}(\Pi = \pi \mid \tilde{T}_n = \tilde{t}_n) = \frac{\mathbb{P}(\tilde{T}_n = \tilde{t}_n \mid \Pi = \pi)\frac{1}{n!}}{\mathbb{P}(\tilde{T}_n = \tilde{t}_n)} = \frac{\mathbb{P}(T_n = \pi^{-1}\tilde{t}_n)\frac{1}{n!}}{\mathbb{P}(\tilde{T}_n = \tilde{t}_n)}$. Moreover, the

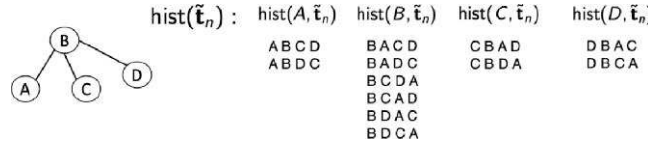


Figure 8. All histories of a tree with 4 nodes.

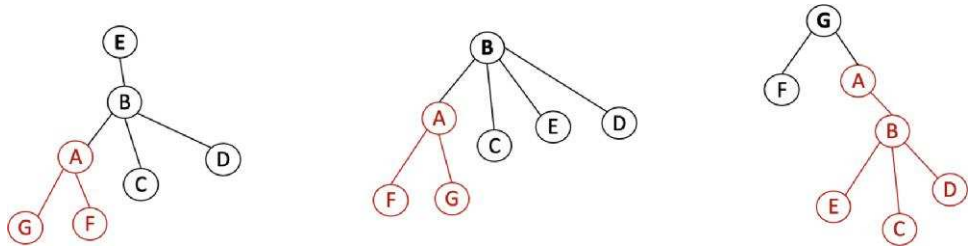


Figure 9. Same tree $\tilde{\boldsymbol{t}}_n$ in three rooted orientations. Left: $\tilde{\boldsymbol{t}}_n^{(E)}$ rooted at E; the subtree of A (denoted as $\tilde{\boldsymbol{t}}_A^{(E)}$) contains nodes A, F, G; node A is the parent of F, G. Centre: $\tilde{\boldsymbol{t}}_n^{(B)}$ rooted at B; the subtree of A (denoted as $\tilde{\boldsymbol{t}}_A^{(B)}$) contains nodes A, F, G; node A is the parent of F, G. Right: $\tilde{\boldsymbol{t}}_n^{(G)}$ rooted at G; the subtree of A (denoted as $\tilde{\boldsymbol{t}}_A^{(G)}$) contains nodes A, B, E, C, D; node A is the parent of B.

probability $\mathbb{P}(\boldsymbol{T}_n = \pi^{-1}\tilde{\boldsymbol{t}}_n)$ is actually the same for any $\pi \in \mathrm{hist}(\tilde{\boldsymbol{t}}_n)$ by online supplementary Proposition S1.

By Proposition 9, we have that

$$\mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n) = \frac{h(u, \tilde{\boldsymbol{t}}_n)}{h(\tilde{\boldsymbol{t}}_n)}.$$

Therefore, we need only count the histories $h(u, \tilde{\boldsymbol{t}}_n)$ for every node $u \in \mathcal{U}_n$. We give a well-known characterisation of $h(u, \tilde{\boldsymbol{t}}_n)$ that leads to a linear time algorithm for counting the size of the histories: define, for any node u, $v \in \mathcal{U}_n$, the tree $\tilde{\boldsymbol{t}}_v^{(u)}$ as the subtree of node v where we view the whole tree as being rooted (hanging from) node u; $\tilde{\boldsymbol{t}}_u^{(u)}$ is thus the entire tree rooted at u. See Figure 9 for an example. We then have that, by Knuth (1997) or Shah and Zaman (2011),

$$h(u, \tilde{\boldsymbol{t}}_n) = n! \prod_{v \in \mathcal{U}_n} \frac{1}{|\tilde{\boldsymbol{t}}_v^{(u)}|}. \tag{13}$$

Therefore, we can compute $h(u, \tilde{\boldsymbol{t}}_n)$ by viewing $\tilde{\boldsymbol{t}}_n$ as being rooted at u and taking the product of the inverse of the sizes of all the subtrees. By using the fact that $h(u, \tilde{\boldsymbol{t}}_n)$ can be directly computed from $h(u', \tilde{\boldsymbol{t}}_n)$ for any neighbour u' of u, Shah and Zaman (2011) derive an $O(n)$ algorithm for computing the size of the histories over all roots $\{h(u, \tilde{\boldsymbol{t}}_n)\}_{u \in \mathcal{U}_n}$, which we give in Section S2 of the online supplementary material for readers' convenience.

3.4.2 The general case

Now suppose we have the label randomised graph $\tilde{\boldsymbol{G}}_n$ from the PAPER model. We then have that

$$\begin{aligned}
\mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) &= \sum_{\tilde{\boldsymbol{t}}_n \subseteq \tilde{\boldsymbol{g}}_n} \sum_{\pi \in \mathrm{hist}(u, \tilde{\boldsymbol{t}}_n)} \mathbb{P}(\Pi = \pi, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n \mid \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \\
&\propto \sum_{\tilde{\boldsymbol{t}}_n \subseteq \tilde{\boldsymbol{g}}_n} \sum_{\pi \in \mathrm{hist}(u, \tilde{\boldsymbol{t}}_n)} \mathbb{P}(\Pi = \pi, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n) \underbrace{\mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \mid \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n, \Pi = \pi)}_{\binom{n(n-1)/2-(n-1)}{m-(n-1)}^{-1}}. \\
&\propto \sum_{\tilde{\boldsymbol{t}}_n \subseteq \tilde{\boldsymbol{g}}_n} \sum_{\pi \in \mathrm{hist}(u, \tilde{\boldsymbol{t}}_n)} \mathbb{P}(\tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n \mid \Pi = \pi) = \sum_{\tilde{\boldsymbol{t}} \subseteq \tilde{\boldsymbol{g}}_n} \sum_{\pi \in \mathrm{hist}(u, \tilde{\boldsymbol{t}})} \mathbb{P}(\boldsymbol{T}_n = \pi^{-1}\tilde{\boldsymbol{t}}_n),
\end{aligned} \tag{14}$$

where, in the outer summation, we require $\tilde{\boldsymbol{t}}_n$ to be a subtree of $\tilde{\boldsymbol{g}}_n$ with n nodes, that is, we require $\tilde{\boldsymbol{t}}_n$ to be a spanning tree of $\tilde{\boldsymbol{g}}_n$ (see equation (16)). If $\boldsymbol{T}_n$ has the uniform attachment distribution ($\alpha = 1, \beta = 0$), then we have that $\mathbb{P}(\boldsymbol{T}_n = \pi^{-1}\tilde{\boldsymbol{t}}_n) = \frac{1}{(n-1)!}$ by online supplementary Proposition S1

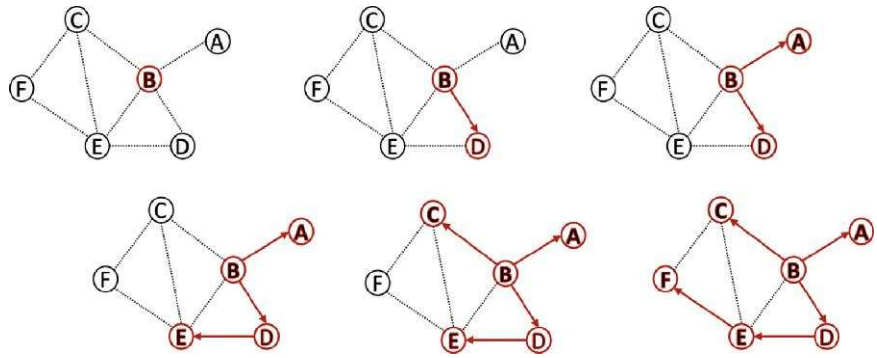


Figure 10. One possible growth realisation starting from node B.

and hence,

$$\mathbb{P}(\Pi_1 = u \,|\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \propto \sum_{\tilde{\boldsymbol{t}}_n \subseteq \tilde{\boldsymbol{g}}_n} h(u, \tilde{\boldsymbol{t}}_n).$$

Thus, the posterior root probability of u is simply proportional to the number of all possible realisations of growth process that start from node u and end up with graph $\tilde{\boldsymbol{g}}_n$; see Figure 10. When $\boldsymbol{T}_n$ has the LPA distribution ($\alpha = 0, \beta = 1$), then $\mathbb{P}(\boldsymbol{T}_n = \pi^{-1}\tilde{\boldsymbol{t}}_n)$ depends on the degree sequence of the tree $\tilde{\boldsymbol{t}}_n$ so that the posterior root probability is proportional to a weighted count of all possible growth realisations.

4 Algorithm

The inference approach that we described in Sections 3.2 and 3.3 requires computing posterior probabilities such as the posterior root probability $P(\Pi_1 = u \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)$ for a fixed alphabetically labelled graph $\tilde{\boldsymbol{g}}_n$. In this section, we derive a Gibbs sampling algorithm to generate an ordering $\pi \in \mathrm{Bi}([n], \mathcal{U}_n)$ and a forest $\tilde{\boldsymbol{f}}_n$ according to the posterior probability

$$\mathbb{P}(\Pi = \pi, \tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n \,|\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n). \tag{15}$$

As discussed towards the end of Section 3.1, in the single root setting, the posterior probability (15) over Π, $\tilde{\boldsymbol{F}}_n$ is non-zero only if $\tilde{\boldsymbol{f}}_n$ is a spanning tree of the graph $\tilde{\boldsymbol{g}}_n$. We formally define the set of spanning trees of a connected graph $\tilde{\boldsymbol{g}}_n$ as

$$\mathcal{T}(\tilde{\boldsymbol{g}}_n) := \left\{\tilde{\boldsymbol{f}}_n : \tilde{\boldsymbol{f}}_n \text{ is connected subtree of } \tilde{\boldsymbol{g}}_n \text{ and } V(\tilde{\boldsymbol{f}}_n) = V(\tilde{\boldsymbol{g}}_n)\right\}. \tag{16}$$

We note that $\mathcal{T}(\tilde{\boldsymbol{g}}_n)$ is non-empty if and only if $\tilde{\boldsymbol{g}}_n$ is connected. For the multiple roots setting, we define the spanning forest of $\tilde{\boldsymbol{g}}_n$ with K components as

$$\mathcal{F}_K(\tilde{\boldsymbol{g}}_n) := \left\{\tilde{\boldsymbol{f}}_n : \tilde{\boldsymbol{f}}_n \text{ is sub-forest of } \tilde{\boldsymbol{g}}_n \text{ with } K \text{ disjoint component trees and } V(\tilde{\boldsymbol{f}}_n) = V(\tilde{\boldsymbol{g}}_n)\right\}$$

so that $\mathcal{F}_1(\tilde{\boldsymbol{g}}_n) = \mathcal{T}(\tilde{\boldsymbol{g}}_n)$. Then, for the fixed K roots model, the posterior probability (15) is non-zero only if $\tilde{\boldsymbol{f}}_n \in \mathcal{F}_K(\tilde{\boldsymbol{g}}_n)$ and for the random K roots model, probability (15) is non-zero only if $\tilde{\boldsymbol{f}}_n \in \mathcal{F}(\tilde{\boldsymbol{g}}_n) := \cup_{K=1}^{n} \mathcal{F}_K(\tilde{\boldsymbol{g}}_n)$.

The value of the posterior probability (15) depends on the parameters of the model, e.g. α, β, θ in the single root setting. We provide an estimation procedure for these parameters in online supplementary Section S3.1 but for now, to keep the presentation simple, we assume that all parameters are known.

Our Gibbs sampler alternates between two stages:

(a) We fix the forest $\tilde{f}_n$ and generate an ordering π with probability $\mathbb{P}(\Pi = \pi \mid \tilde{G}_n = \tilde{g}_n, \tilde{F}_n = \tilde{f}_n)$.
(b) We fix the ordering π and generate a new forest $\tilde{f}_n$ by iteratively sampling a new parent for each of the nodes.

We give the details for stage A in the next section and for stage B in Section 4.2.

Remark 8 In online supplementary Section S3.3, we give an alternative collapsed Gibbs sampling algorithm in which we collapse stage (A) so that we only sample the roots instead of the whole history π. The collapsed Gibbs sampler requires fewer iterations to converge but each iteration is more computationally intensive. Practically, the sampling algorithm that we present in Sections 4.1 and 4.2 appears to be faster except for the random K roots model on some data sets.

4.1 Sampling the ordering

In this section, we provide an algorithm for the first stage of the Gibbs sampler where we sample an ordering. We fix a spanning forest $\tilde{f}_n$ of the observed graph $\tilde{g}_n$, let K be the number of component trees of $\tilde{f}_n$, and let $m = |E(\tilde{g}_n)|$ be the number of edges of g_n. We have that

$$\mathbb{P}(\Pi = \pi \mid \tilde{G}_n = \tilde{g}_n, \tilde{F}_n = \tilde{f}_n) \propto \mathbb{P}(\Pi = \pi \mid \tilde{F}_n = \tilde{f}_n)\mathbb{P}(\tilde{G}_n = \tilde{g}_n \mid \tilde{F}_n = \tilde{f}_n, \Pi = \pi). \tag{17}$$

Under the non-sequential noise PAPER models, since the non-forest edges of $\tilde{G}_n$ are independent Erdős–Rényi random edges, we have $\mathbb{P}(\tilde{G}_n = \tilde{g}_n \mid \tilde{F}_n = \tilde{f}_n, \Pi = \pi) = \binom{\binom{n}{2}-(n-K)}{m-(n-K)}^{-1}$ and may thus ignore the non-forest edges and consider only on the posterior probability $\mathbb{P}(\Pi = \pi \mid \tilde{F}_n = \tilde{f}_n)$ when sampling π. In the sequential noise seq-PAPER model, the $\mathbb{P}(\tilde{G}_n = \tilde{g}_n \mid \tilde{F}_n = \tilde{f}_n, \Pi = \pi)$ term must be taken into account but can be computed efficiently. We give the detailed algorithms for each of the settings.

4.1.1 Single root setting

In the single root setting, $\tilde{f}_n$ is connected and hence a tree; we thus change to the notation $\tilde{t}_n := \tilde{f}_n$ to be consistent with the notation used in Definition 1.

Hence, by our discussion in Section 3.4, sampling π according to $\mathbb{P}(\Pi = \cdot \mid \tilde{T}_n = \tilde{t}_n)$ is equivalent to sampling π uniformly from hist($\tilde{t}_n$). Crane and Xu (2021) and also Cantwell et al. (2021) derive a procedure to sample uniformly from hist($\tilde{t}_n$) and we provide a concise description of the procedure here for the readers' convenience.

To generate π uniformly from hist($\tilde{t}_n$), we generate the first node π_1 by taking the set of all nodes and drawing a node u with probability

$$\mathbb{P}(\Pi_1 = u \mid \tilde{T}_n = \tilde{t}_n) = \frac{h(u, \tilde{t}_n)}{h(\tilde{t}_n)}. \tag{18}$$

The entire collection $\{h(u, \tilde{t}_n)\}_{u \in \mathcal{U}_n}$ can be computed in $O(n)$ time (c.f. Section 3.4 and online supplementary Section S2) and thus we require at most $O(n)$ time to generate the first node π_1.

To generate the subsequent ordering $\pi_{2:n}$, we view the tree $\tilde{t}_n$ as being rooted at π_1 and use the notation $\tilde{t}_n^{(\pi_1)}$ make the root explicit. For each node $v \in \mathcal{U}_n$, we define $\tilde{t}_v^{(\pi_1)}$ as the subtree of the node v, viewing the whole tree as being rooted at node π_1. We give an example of these definitions in Figure 9.

Then, by Crane and Xu (2021, Proposition 9), for every $t \in [n-1]$,

$$\mathbb{P}(\Pi_{t+1} = v \mid \tilde{T}_n = \tilde{t}_n, \Pi_{1:t} = \pi_{1:t}) = \begin{cases} \frac{|\tilde{t}_v^{(\pi_1)}|}{n-t+1} & \text{if } v \text{ is a neighbour of } \pi_{1:t} \text{ in } \tilde{t}_n \\ 0 & \text{else} \end{cases} \tag{19}$$

Algorithm 1 Generating $\pi \in \text{hist}(\tilde{\boldsymbol{f}}_n)$ according to $\mathbb{P}(\Pi = \pi \mid \tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n)$ in ER noise settings.

Input: Labelled forest $\tilde{f}_n$ with K trees, denoted $\tilde{t}^1, \ldots, \tilde{t}^K$.

Output: $\pi \in \text{hist}(\tilde{f}_n)$.

1: **for** $k = 1, 2, \ldots, K$ **do:**

2: Choose node $u^k \in V(\tilde{t}^{(k)})$ with probability (18) with PAPER(α, β, θ) model and with probability (20) under PAPER(α, β, K, θ) or PAPER($\alpha, \beta, \alpha_0, \theta$).

3: **end for**

4: Let $\tilde{s} = \{u^1, u^2, \ldots, u^K\}$ be the set of roots, and

- under PAPER(α, β, θ), let $\pi_1 = u^1$ and let $t_0 = 2$,
- under PAPER(α, β, K, θ), let $\pi_{1:K} = \tilde{s}$ in a random ordering and let $t_0 = K + 1$.
- under PAPER($\alpha, \beta, \alpha_0, \theta$), choose $u^k \in \tilde{s}$ with probability $|\tilde{t}^k|/n$, let $\pi_1 = u^k$, let $t_0 = 2$.

5: Generate $\pi_{t_0:n}$ as a uniformly random permutation of $\mathcal{U}_n \backslash \pi_{1:(t_0-1)}$.

6: **for** $t = t_0, t_0 + 1, \ldots, n$ **do:**

7: Let $v_1 = \pi_t, v_2 = \text{pa}(v_1), \ldots, v_k = \text{pa}(v_{k-1})$ where k is the largest integer such that $v_1, v_2, \ldots, v_k \notin \pi_{1:(t-1)}$.
▷ $\text{pa}(v)$ denotes the parent of v with respect to $\tilde{f}_n$ rooted at $\tilde{s}$.

8: Set $\pi_t = v_k$, $t_k = \pi^{-1}(v_k)$, and $\pi_{t_k} = v_1$.

9: **end for**

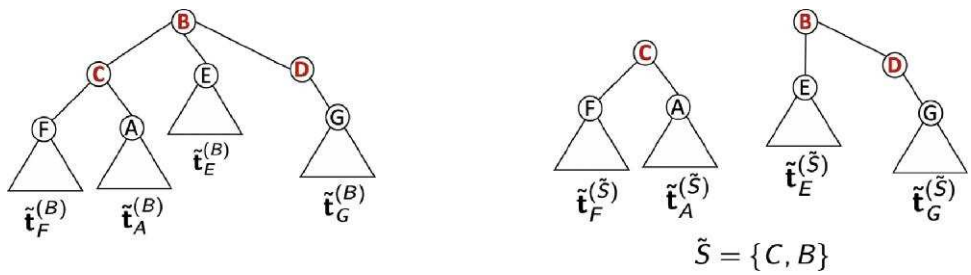


Figure 11. Example of sampling an ordering. In both cases, suppose $\pi_{1:3} = \{B, C, D\}$, then draw π_4 from the neighbours $\{F, A, E, G\}$ with probability proportional to the size of their subtrees.

One may verify this by showing that the probability of generating a particular ordering is $\frac{1}{n!}\prod_{v \in \mathcal{U}_n} |\tilde{t}_n^{(u)}| = \frac{1}{h(u, \tilde{t}_n)}$ by equation (13).

Thus, we may generate π_2 by considering all neighbours of π_1 in $\tilde{t}_n$ and drawing a node v with probability proportional to the size of its subtree $|\tilde{t}_v^{(u_1)}|$ and similar for π_3, π_4, etc. The entire sampling process can be efficiently done by generating a permutation uniformly at random and modifying it in place so that it obeys the hist($\tilde{f}_n$) constraint. We summarise this in Algorithm 1 with $K = 1$ and also give a visual illustration in Figure 11. The runtime of the sampling algorithm is upper bounded by $O(n\text{diam}(\tilde{t}_n))$ (Crane & Xu, 2021, Proposition 10). Trees generated by the APA(α, β) model have diameter $O_p(\log n)$ (see, e.g. Drmota, 2009, Theorem 6.32, and Bhamidi, 2007, Theorem 18) and the overall runtime is therefore $O(n \log n)$. The computational complexity is the same under the fixed K setting and the random K setting.

4.1.2 Fixed K *roots setting*

For the PAPER(α, β, K, θ) model, we may generate from $\mathbb{P}(\Pi = \cdot \mid \tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n)$ in a similar way. In this case, $\tilde{f}_n$ is a forest that contains K disjoint component trees, which we denote by $\tilde{t}^1, \ldots, \tilde{t}^K$. We first generate a root for each component tree. For each $k \in [K]$, we draw $u^k \in V(\tilde{t}^k)$ with

probability

$$\frac{h(u^k, \tilde{\boldsymbol{t}}^k)(\beta D_{\tilde{\boldsymbol{t}}^k}(u^k) + \beta + \alpha)(\beta D_{\tilde{\boldsymbol{t}}^k}(u^k) + \alpha)}{\sum_{v \in V(\tilde{\boldsymbol{t}}^k)} h(v, \tilde{\boldsymbol{t}}^k)(\beta D_{\tilde{\boldsymbol{t}}^k}(v) + \beta + \alpha)(\beta D_{\tilde{\boldsymbol{t}}^k}(v) + \alpha)}. \tag{20}$$

We note that equation (20) is different from the corresponding probability in the single tree setting (18) because we give each root node an imaginary self-loop edge. We leave the detailed derivation of equation (20) to Section S3.2 of the online supplementary material.

We let $\tilde{s} = \{u^1, \ldots, u^k\}$ denote the set of roots that we have generated. By the definition of the PAPER(α, β, K, θ) model (Definition 4), the root nodes $\tilde{s}$ occupy the first K positions of the ordering π and we thus let $\pi_{1:K}$ be the elements of $\tilde{s}$ placed in a random ordering.

Next, we view each component tree $\tilde{\boldsymbol{t}}^k$ as being rooted at u_k and, for every node $v \in V(\tilde{\boldsymbol{f}}_n)$, we denote the subtree of node v by $\tilde{\boldsymbol{t}}_v^{(\tilde{s})}$. We then generate $\pi_{(K+1):n}$ according to probability (19) where we use the size of the subtree $|\tilde{\boldsymbol{t}}_v^{(\tilde{s})}|$. This is equivalent to generating a full history (excluding the root node) for every tree and then interleaving them at random. We again summarise the whole procedure in Algorithm 1.

4.1.3 Random K *roots setting*

Now consider the random K roots setting with the PAPER($\alpha, \beta, \alpha_0, \theta$) model and suppose $\tilde{\boldsymbol{f}}_n$ comprises of K disjoint trees $\tilde{\boldsymbol{t}}^1, \ldots, \tilde{\boldsymbol{t}}^K$. We again generate the set of roots $\tilde{s} = \{u^1, \ldots, u^K\}$ by drawing u^k from $\tilde{\boldsymbol{t}}^k$ with probability (20). In contrast with the fixed K roots setting, the root nodes $u^1, \ldots, u^K$ need not occupy the first K positions of the ordering π.

To generate the ordering π, we first choose $u^k \in \tilde{s}$ with probability $|\tilde{\boldsymbol{t}}^k|$ and set $\pi_1 = u^k$. We then draw $\pi_{2:n}$ iteratively using the conditional distribution

$$\mathbb{P}(\Pi_{t+1} = v \,|\, \tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}, \Pi_{1:t} = \pi_{1:t}) = \begin{cases} \frac{|\tilde{\boldsymbol{t}}_v^{(\tilde{s})}|}{n-t+1} & \text{if } v \text{ is a neighbour of } \pi_{1:t} \text{ in } \tilde{\boldsymbol{f}}_n \text{ or if } v \in \tilde{s} \\ 0 & \text{else} \end{cases} \tag{21}$$

We note that for a root node $u^k \in \tilde{s}$, the subtree $\tilde{\boldsymbol{t}}_{u^k}^{(\tilde{s})}$ is precisely the whole tree $\tilde{\boldsymbol{t}}^k$. We summarise this procedure in Algorithm 1.

4.1.4 Sequential noise setting

Under the seq-PAPER model described in Section 2.3, we no longer have a direct sampling algorithm to draw from $\mathbb{P}(\Pi = \cdot \,|\, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n)$ because we have to take into account the $\mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \,|\, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n, \Pi = \pi)$ term in equation (17). For seq-PAPER models, we propose instead a Metropolis–Hastings algorithm to update π by sampling new transpositions.

Let π be the current sample of arrival ordering. To generate a new proposal π^*, we randomly choose a pair $j, k \in \{2, \ldots, n\}$ and construct π^* by swapping the j-th and the k-th entries of π, that is, $\pi_j^* = \pi_k$ and $\pi_k^* = \pi_j$ and all other entries are equal. If $\pi^* \notin \text{hist}(\tilde{\boldsymbol{t}}_n)$, then we reject the proposal; otherwise, we accept it with probability

$$1 \wedge \frac{\mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \,|\, \Pi = \pi^*, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n)}{\mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \,|\, \Pi = \pi, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n)}, \tag{22}$$

which follows because $\mathbb{P}(\Pi = \pi \,|\, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n) = \mathbb{P}(\Pi = \pi^* \,|\, \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n)$. The ratio in equation (22) has a complicated expression but can be computed in time proportional to only the degrees, with respect to $\tilde{\boldsymbol{g}}_n$, of π_j, π_k, and the parent nodes $\text{pa}(\pi_j)$, $\text{pa}(\pi_k)$, where the notion of parent node is defined in equation (23). We give a detailed description of how to efficiently compute (22) and determine whether $\pi^* \in \text{hist}(\tilde{\boldsymbol{t}}_n)$ in Section S3.5 of the online supplementary material; in particular, see online supplementary Section S3.5.2 which uses results from online supplementary Section S3.5.1. Even with our efficient implementation however, updating π by sampling transpositions is

considerably slower than sampling π directly via equation (19).

The transposition sampler does not change the root node since j, k are not allowed to take on the value 1. To sample a new root node, we fix $k_0 \in \mathbb{N}$ and generate a new proposal π^* by shuffling the first k_0 entries of π. We then accept π^* if it is a valid history and with probability (22). Finally, we note that under the seq-PAPER* model with tree edge removal, our method for sampling π is exactly the same. Since we condition on $\tilde{T}_n$, it makes no difference whether we have deletion noise or not.

Remark 9 Sheridan et al. (2012) and Bloem-Reddy et al. (2018) use the idea of swapping *adjacent* elements of an ordering π for a Poisson growth attachment models and a sequential edge-growth model referred to as Beta Neutral-to-the-Left, respectively. In contrast, under the seq-PAPER model, we can compute non-adjacent swap proposal probabilities efficiently and hence, we can explore the permutation space of π faster. This is because the seq-PAPER is a simpler model and also because we restrict ourselves to a spanning tree, which simplifies many parts of the calculations. We note that sampling π through non-adjacent pair swaps can also be used for the model $\boldsymbol{G}_n = \boldsymbol{T}_n + \boldsymbol{R}_n$ where $\boldsymbol{T}_n$ is not shape-exchangeable, for instance when the attachment probability is $\phi(D_{T_{t-1}}(w_t))$ for some non-affine function $\phi(\cdot)$ instead of the affine expression given in equation (1). Finally, We emphasise that inference for the vanilla PAPER model is significantly faster than any form of swapping-based Metropolis samplers since it directly samples the entire ordering.

4.2 Sampling the forest

In this section, we describe stage B of the Gibbs sampling algorithm. For a fixed ordering π and a spanning forest $\tilde{f}_n$, we may obtain a set of roots $\tilde{s}$ for each of the component trees of $\tilde{f}_n$ by taking the earliest node (according to π) of each tree. Viewing $\tilde{f}_n$ as being rooted at $\tilde{s}$ induces parent–child relationships between all the nodes.

To define the parent–child relationship formally, let $\tilde{f}_n$ be a forest with disjoint component trees $\tilde{t}^1, \ldots, \tilde{t}^K$ and let $\tilde{s} = \{u^1, u^2, \ldots, u^K\}$ be a set of root nodes such that $u^k \in V(\tilde{t}^k)$. Let u be any node not in $\tilde{s}$ and suppose $u \in V(\tilde{t}^k)$. There exists a unique node $v \in V(\tilde{t}^k)$ such that v is a neighbour of u in $\tilde{f}_n$ and that the unique path from u to the root u^k contains v. We say v the *parent node* of u and write

$$pa(u) \equiv pa_{\tilde{f}_n^{(\tilde{s})}}(u) = \text{parent of } u \text{ with respect to } \tilde{f}^{(\tilde{s})}. \tag{23}$$

For a root node $u \in \tilde{s}$, we let $pa(u) := \emptyset$ for convenience. Since every edge in $\tilde{f}_n$ is between a node and its parent, the set of parents $\{pa(u)\}_{u \in \mathcal{U}_n}$ specifies the $n - K$ edges in $\tilde{f}_n$ and hence uniquely specifies the forest $\tilde{f}_n$ and the root nodes $\tilde{s}$.

Our Gibbs sampler updates the forest $\tilde{f}_n$ by iteratively updating the parent of each of the nodes, which adds and removes a single edge from $\tilde{f}_n$ (it is possible to add and remove the same edge so that the forest does not change) or, in the random K setting, we may remove a single edge and add a new root node or remove a root node and add a single edge.

To be precise, the latent tree $\tilde{F}_n$ and root set $\tilde{S}$ induces a latent parent of each node which we denote $pa_{\tilde{F}_n^{(\tilde{S})}}(\cdot)$. For every node u, we generate a new parent u' according to the conditional distribution

$$Q_u(u') := \mathbb{P}\left(pa_{\tilde{F}_n^{(\tilde{S})}}(u) = u' \,\middle|\, \Pi = \pi, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n, \left\{ pa_{\tilde{F}_n^{(\tilde{S})}}(v) = pa_{\tilde{f}_n^{(\tilde{s})}}(v) \right\}_{v \neq u} \right), \tag{24}$$

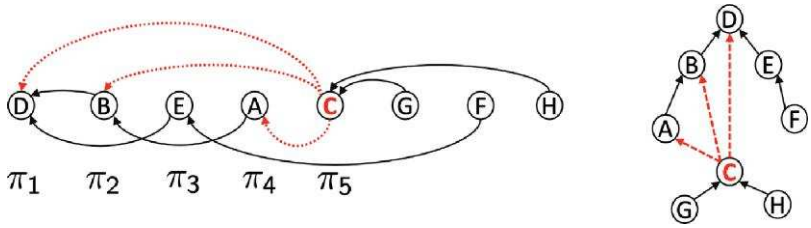


Figure 12. Sampling a parent for π_5 (node C).

and then replace the old edge $(u, pa(u))$ with (u, u'). Since we condition on the arrival ordering Π, probability (24) is non-zero only when u' arrives prior to u, i.e. $\pi^{-1}u' < \pi^{-1}u$, and $(u, u') \in E(\tilde{\boldsymbol{g}}_n)$. In other words, if $\pi^{-1}u = t$, then $Q_u(\cdot)$ is supported on the set of nodes $\pi_{1:(t-1)} \cap N_{\tilde{\boldsymbol{g}}_n}(u)$. In the random K setting, u' is allowed to be empty in which case $Q_u(\cdot)$ is supported on $\{\emptyset\} \cup (\pi_{1:(t-1)} \cap N_{\tilde{\boldsymbol{g}}_n}(u))$ where $N_{\tilde{\boldsymbol{g}}}(u)$ is the set of neighbours of u on the graph $\tilde{\boldsymbol{g}}_n$. Our sampling procedure then generate the parents for $\pi_1, \pi_2, \pi_3, \ldots$ sequentially. In Figure 12, we illustrate how we may generate a new parent for π_5 (node C) by choosing one of the edges that connects π_5 with one of the earlier nodes $\pi_{1:4}$.

At iteration t, to compute $Q_{\pi_t}(\cdot)$ with respect to π_t, for each node v in the support of $Q_{\pi_t}(\cdot)$, we let $\tilde{\boldsymbol{f}}_n^{(v,\pi_t)}$ denote the forest formed by removing the old edge $(\mathrm{pa}(\pi_t), \pi_t)$ and adding the new edge (v, π_t). We note that v is allowed to be the old parent so that we may have $\tilde{\boldsymbol{f}}_n = \tilde{\boldsymbol{f}}_n^{(v,\pi_t)}$. Then, for any w_t in the support of $Q_{\pi_t}(\cdot)$, we have

$$Q_{\pi_t}(w_t) = \frac{\mathbb{P}(\tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n^{(w_t,\pi_t)} \mid \Pi = \pi, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)}{\sum_v \mathbb{P}(\tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n^{(v,\pi_t)} \mid \Pi = \pi, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)}. \tag{25}$$

In the PAPER models with Erdős–Rényi edges, We can compute the conditional distribution $\mathbb{P}(\tilde{\boldsymbol{F}}_n = \cdot \mid \Pi = \pi, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)$ by using the fact that once when we condition on $\tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n$, the remaining edges of $\tilde{\boldsymbol{G}}_n$ are uniformly random and the fact that Π and $\boldsymbol{F}_n$ are independent. Thus,

$$\begin{aligned}
&\mathbb{P}(\tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n \mid \Pi = \pi, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \\
&\quad \propto \mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \mid \tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n, \Pi = \pi) \mathbb{P}(\tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n \mid \Pi = \pi) \\
&\quad = \binom{\binom{n}{2} - (n - K(\tilde{\boldsymbol{f}}_n))}{m - (n - K(\tilde{\boldsymbol{f}}_n))}^{-1} \mathbb{P}(\boldsymbol{F}_n = \pi^{-1}\tilde{\boldsymbol{f}}_n) \mathbb{1}\{\tilde{\boldsymbol{f}}_n \in \mathcal{F}(\tilde{\boldsymbol{g}}_n)\} \\
&\quad \propto \left\{ \prod_{k=1}^{K(\tilde{\boldsymbol{f}}_n)} \frac{n(n-1)/2 - n + k)}{m - n + k} \right\} \mathbb{P}(\boldsymbol{F}_n = \pi^{-1}\tilde{\boldsymbol{f}}_n) \mathbb{1}\{\tilde{\boldsymbol{f}}_n \in \mathcal{F}(\tilde{\boldsymbol{g}}_n)\}.
\end{aligned} \tag{26}$$

We now discuss the sampling procedure in detail in all the settings.

4.2.1 Single root setting

In the single root setting, we again use the notation $\tilde{\boldsymbol{t}}_n = \tilde{\boldsymbol{f}}_n$ to be consistent with Definition 1. The first term of equation (26) is a constant since $K(\tilde{\boldsymbol{t}}_n) = 1$ and may thus be ignored. Using the likelihood of APA trees (see Remark 2 as well as Proposition S1 from the online supplementary material) and using the fact that $\mathbb{P}(\boldsymbol{T}_n = \pi^{-1}\tilde{\boldsymbol{t}}_n) > 0$ when $\pi \in \mathrm{hist}(\tilde{\boldsymbol{t}}_n)$, we have that, for any

Algorithm 2 Generating spanning forest $\tilde{\boldsymbol{f}}_n$ of $\tilde{\boldsymbol{g}}_n$ under either PAPER(α, β, θ) or PAPER(α, β, K, θ)

Input: Graph $\tilde{g}_n$, ordering $\pi \in \mathrm{Bi}([n], \mathcal{U}_n)$, and a spanning forest $\tilde{f}_n$ with K component trees.

Effect: Modifies $\tilde{f}_n$ in place.

1: **for** $t = K+1, \ldots, n$ **do:**

2: Remove old edge $(\pi_t, pa(\pi_t))$ from $\tilde{f}_n$ to obtain $\tilde{f}_n^{(\cdot,\pi_t)}$.

3: Choose a node $w_t \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)$ with probability proportional to

$$\begin{cases} \beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(w_t) + \alpha & \text{under PAPER } (\alpha, \beta, \theta) \\ \beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(w) + 2\beta \mathbb{1}\{w \in \pi_{1:K}\} + \alpha & \text{under PAPER } (\alpha, \beta, K, \theta) \end{cases}$$

4: Add new edge (π_t, w_t) to $\tilde{f}_n$.

5: **end for**

$w_t \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)$,

$$Q_{\pi_t}(w_t) = \frac{\beta D_{\tilde{t}_n^{(\cdot,\pi_t)}}(w_t) + \alpha}{\sum_{v \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)} \beta D_{\tilde{t}_n^{(\cdot,\pi_t)}}(v) + \alpha},$$

where $\tilde{t}_n^{(\cdot,\pi_t)}$ is the disconnected graph obtained by removing the old edge $(\mathrm{pa}(\pi_t), \pi_t)$ from $\tilde{t}_n$. We summarise the resulting procedure in Algorithm 2. Since we visit every node once and, for a single node u, it takes time $O(D_{\tilde{g}_n}(u))$ to generate a new parent, the overall runtime of the second stage of the algorithm is $O(m)$. The computational complexity is the same under the fixed K setting and the random K setting.

4.2.2 *Fixed* $K > 1$ *setting*

Since the number of trees K is fixed, the first term of equation (26) is again a constant. Using likelihood of APA trees again (see Proposition S2 from the online supplementary material), we have that for any $w_t \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)$,

$$Q_{\pi_t}(w_t) = \frac{\beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(w_t) + 2\beta \mathbb{1}\{w_t \in \pi_{1:K}\} + \alpha}{\sum_{v \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)} \beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(v) + 2\beta \mathbb{1}\{v \in \pi_{1:K}\} + \alpha},$$

where, as with the single root setting, $\tilde{f}_n^{(\cdot,\pi_t)}$ is the forest obtained by removing the old edge $(\mathrm{pa}(\pi_t), \pi_t)$ from $\tilde{f}_n$. The only difference from the single root setting is that we have a higher probability to attach to a root node because of the imaginary self-loop edge. We summarise the procedure in Algorithm 2.

4.2.3 *Random* K *roots setting*

Under the PAPER(α, β, α_0, θ) model, a node may become a new root in the sampling process and thus we must take into account the first term of equation (26). Moreover, in this setting, $Q_{\pi_t}(\cdot)$ for node π_t is supported on $\{\emptyset\} \cup (\pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t))$ since we may turn the node π_t into a new root node, in which case we set its parent to $\emptyset$ by convention. Define $\tilde{\alpha}_0 := \alpha_0 \frac{m-n+K+\mathbb{1}\{\pi_t \notin \tilde{s}\}}{n(n-1)/2-n+K+\mathbb{1}\{\pi_t \notin \tilde{s}\}}$; we then have that, by online supplementary Proposition S3, for any $w_t \in \{\emptyset\} \cup (\pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t))$,

$$Q_{\pi_t}(w_t) = \frac{\tilde{\alpha}_0}{\tilde{\alpha}_0 + \sum_{v \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)} \beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(v) + 2\beta \mathbb{1}\{v \in \tilde{s}\} + \alpha} \quad \text{if } w_t = \emptyset$$

$$\text{and } Q_{\pi_t}(w_t) = \frac{\beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(w_t) + 2\beta \mathbb{1}\{w_t \in S\} + \alpha}{\tilde{\alpha}_0 + \sum_{v \in \pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t)} \beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(v) + 2\beta \mathbb{1}\{v \in \tilde{s}\} + \alpha} \quad \text{if } w_t \neq \emptyset,$$

Algorithm 3 Generating spanning forest $\tilde{\boldsymbol{f}}_n$ of $\tilde{\boldsymbol{g}}_n$ under PAPER(α, β, α_0, θ)

Input: Graph $\tilde{g}_n$, ordering $\pi \in \mathrm{Bi}([n], \mathcal{U}_n)$, and a spanning forest $\tilde{f}_n$.

Effect: Modifies $\tilde{f}_n$ in place.

1: Let $\tilde{s}$ be the set of root nodes.

2: **for** $t = 2, 3, \ldots, n$ **do:**

3: If $\pi_t \notin \tilde{s}$, remove edge $(\pi_t, pa(\pi_t))$ from $\tilde{f}_n$ to get $\tilde{f}_n^{(\cdot,\pi_t)}$. Else, let $\tilde{s} = \tilde{s} \backslash \{w_t\}$ and let $\tilde{f}_n^{(\cdot,\pi_t)} = \tilde{f}_n$.

4: Choose a node $w_t \in \{\emptyset\} \cup (\pi_{1:(t-1)} \cap N_{\tilde{g}_n}(\pi_t))$ with probability proportional to

$$\begin{cases} \alpha_0 & \text{for } w_t = \emptyset \\ \beta D_{\tilde{f}_n^{(\cdot,\pi_t)}}(w_t) + 2\beta \mathbb{1}\{w_t \in s\} + \alpha & \text{for } w_t \neq \emptyset \end{cases}$$

5: If $w_t \neq \emptyset$, let $\tilde{f}_n = \tilde{f}_n^{(\cdot,\pi_t)} \cup (\pi_t, w_t)$. Otherwise, let $\tilde{s} = \tilde{s} \cup \{\pi_t\}$ and $\tilde{f}_n = \tilde{f}_n^{(\cdot,\pi_t)}$.

6: **end for**

where, if π_t is not a root node, $\tilde{f}_n^{(\cdot,\pi_t)}$ is the forest obtained by removing the old edge $(\pi_t, \mathbf{pa}(\pi_t))$ and if π_t is a root node, then $\tilde{f}_n^{(\cdot,\pi_t)} = \tilde{f}_n$. We summarise the resulting procedure in Algorithm 3.

4.2.4 Sequential noise setting

Under the seq-PAPER setting, we use the same sampling procedure but the sampling probabilities become more complicated. From equation (25), we see that, for $w \in N_{\tilde{g}_n} \cap \pi_{1:(t-1)}$,

$$\begin{aligned} Q_{\pi_t}(w) &\propto \mathbb{P}(\tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n^{(w,\pi_t)} \mid \Pi = \pi, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) \\ &\propto \underbrace{\mathbb{P}(\tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n \mid \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n^{(w,\pi_t)}, \Pi = \pi)}_{\text{noise term}} \mathbb{P}(\tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n^{(w,\pi_t)} \mid \Pi = \pi). \end{aligned}$$

Under the seq-PAPER model, the noise term also depends on w since choosing a new parent for π_t would change the tree degrees of some of the nodes. Naively computing $Q_{\pi_t}(w)$ takes time $O(n)$, but in Section S3.5.3 of the online supplementary material (using results from online supplementary Section S3.5.1), we give a detailed algorithm to compute $Q_{\pi_t}(w)$ in time $O(D_{\tilde{g}_n}(w))$ so that overall, we can sample a new parent for π_t in time proportional to the number of neighbours of neighbours of π_t.

When we have deletion noise, as the case of the seq-PAPER* model, the latent tree $\tilde{\boldsymbol{T}}_n$ need not be a subgraph of $\tilde{\boldsymbol{G}}_n$ and hence, when sampling a new parent for π_t, we must consider all of $\pi_{1:(t-1)}$ and not just graph neighbours of π_t. Thus, we draw $w \in \pi_{1:(t-1)}$ with probability $Q_{\pi_t}(w)$ and set $\mathrm{pa}(\pi_t) = w$. We give the detailed algorithm for computing $Q_{\pi_t}(w)$ in Section S3.5.3 of the online supplementary material.

4.3 Other aspects of the algorithm

4.3.1 Parameter estimation

To estimate α and β, we derive an EM algorithm in Section S3.1 of the online supplementary material. The noise level θ is easy to estimate via $\hat{\theta} = \frac{m-(n-1)}{n(n-1)/2-(n-1)}$ in the single root setting. The inference algorithm in fact does not require knowledge of θ since it conditions on the number of edges m of the observed graph. We discuss some ways to select the number of trees K in the fixed K root setting and ways to estimate α_0 in the random K roots setting in Section S3.4 of the online supplementary material.

4.3.2 Inference from posterior samples

The Gibbs sampler described in Sections 4.1 and 4.2 generates a Monte Carlo sequence $\{(\pi^{(j)}, \tilde{f}_n^{(j)})\}_{j=1}^J$ where J is the number of Monte Carlo samples. A straightforward way to

approximate the posterior root probability is to use the empirical distribution based on all the $\pi^{(j)}$'s. However, we can construct a much more accurate approximation by taking advantage of the fact that the posterior root probability is easy to compute on a tree.

Consider the single root setting for simplicity where the posterior root probability is $\mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n)$ for any node u. In this case, we may compute distributions $Q^{(1)}, Q^{(2)}, \ldots, Q^{(J)}$ over the nodes by

$$Q^{(j)} = \mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n^{(j)}, \tilde{\boldsymbol{G}}_n = \tilde{\boldsymbol{g}}_n) = \mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{T}}_n = \tilde{\boldsymbol{t}}_n^{(j)}) = \frac{h(u, \tilde{\boldsymbol{t}}_n^{(j)})}{h(\tilde{\boldsymbol{t}}_n^{(j)})}.$$

Then, we output $\frac{1}{J}\sum_{j=1}^{J} Q^{(j)}$ as our approximation of the posterior root distribution. In the multiple roots setting, we use the same procedure except that we compute $u \mapsto \mathbb{P}(u \in \tilde{S} \mid \tilde{\boldsymbol{F}}_n = \tilde{\boldsymbol{f}}_n^{(j)})$ and then average across $j \in \{1, 2, \ldots, J\}$.

In the multiple roots setting, each Monte Carlo sample of the forest $\tilde{\boldsymbol{f}}_n^{(j)}$ contain either K disjoint trees in the fixed K setting or a random number of disjoint trees in the random K setting. These disjoint trees provide a posterior sample of the communities on the network and using them, we may estimate the community structure of the network. We provide details on one way of using posterior samples for community recovery in Sections 6.3 and 6.4.

The Gibbs sampling algorithm scales to large networks. We are able to run it on networks of up to a million nodes (c.f. Section 6.2.2) on a single 2020 MacBook Pro laptop. To give a rough sense of the runtime, it takes about 1 second to perform one outer loop of the Gibbs sampler on a graph of 10,000 nodes and 20,000 edges. In Section S3.4 of the online supplementary material, we provide more details on practical usage of the Gibbs sampler such as convergence criterion.

4.3.3 Initialisation

In the single root setting, to initialise the Gibbs sampling algorithm, we recommend generating the initial tree $\tilde{\boldsymbol{t}}_n$ uniformly at random from the set of spanning trees $\mathcal{T}(\tilde{\boldsymbol{g}}_n)$ of the observed graph, which can be efficiently done via elegant random-walk-based algorithms such as the Aldous–Broder algorithm (Aldous, 1990; Broder, 1989) or Wilson's algorithm (Wilson, 1996). We then initialise π by drawing an ordering uniformly from the history of the initial tree. This initialisation distribution is guaranteed to be overdispersed and works very well in practice. The same initialisation works for the random K setting. For the fixed K setting, we can form the initial forest by constructing uniformly random spanning tree $\tilde{\boldsymbol{t}}_n$ and uniformly random ordering π as usual, taking the first K nodes of the π as the root nodes, and removing all tree edges between them to obtain an initial $\tilde{\boldsymbol{f}}_n$. We use Wilson's algorithm in our implementation.

5 Theoretical analysis

We provide theoretical support for our approach by deriving bounds on the size of our proposed confidence sets when the observed graph has the PAPER distribution. In particular, we aim to quantify how the quality of inference deterioriates with the noise level θ, that is, how the size of the confidence set increases with θ. For simplicity, for consider only the single root setting and we do not take into account approximation errors introduced by the Gibbs sampler, that is, we analyse the confidence set constructed from the exact posterior root probabilities.

We begin with a type of optimality statement which shows that the size of the confidence set $B_\epsilon(\cdot)$, as defined in equation (8), is of no larger order than any other asymptotically valid confidence set. Intuitively, this is because $B_\epsilon(\cdot)$ can be interpreted as a 'Bayes estimator' for the root node.

Lemma 10 Let ϵ be in $(0, 1)$, let $\boldsymbol{G}_n \sim \text{PAPER}(\alpha, \beta, \theta)$, and let $\boldsymbol{G}_n^* = \rho \boldsymbol{G}_n$ be the observed alphabetically labelled graph for some $\rho \in \text{Bi}([n], \mathcal{U}_n)$. Let $B_\epsilon(\boldsymbol{G}_n^*)$ be defined as in equations (7) and (8). Fix any $\delta \in (0, 1)$ and let $C_{\delta\epsilon}(\boldsymbol{G}_n^*)$ be any confidence set for the root node that is labelling equivariant and has asymptotic coverage

level $1-\delta\epsilon$, that is, $\limsup_{n\to\infty}\mathbb{P}(\rho_1 \notin C_{\delta\epsilon}(\mathbf{G}_n^*)) \le \delta\epsilon$. Then, we have that

$$\limsup_{n\to\infty}\mathbb{P}\big(|B_\epsilon(\mathbf{G}_n^*)| \ge |C_{\delta\epsilon}(\mathbf{G}_n^*)|\big) \le \delta.$$

We provide the proof of Lemma 10 in Section S4 of the online supplementary material.

Ideally, we would compare the size of $B_\epsilon(\cdot)$ with $C_\epsilon(\cdot)$ at the same level. It is however much easier to compare with the more conservative $C_{\delta\epsilon}(\cdot)$. In many cases, the size of a confidence set $|C_\epsilon(\cdot)|$ has bounds of the form $f(n)g(\epsilon^{-1})$ for some functions f and g (see, e.g. Banerjee & Bhamidi, 2020) so that comparing with $C_{\delta\epsilon}(\cdot)$ adds only a multiplicative constant to the bound.

Lemma 10 is useful because it is difficult to directly bound the confidence set $B_\epsilon(\cdot)$ as a function of n and the parameters; Lemma 10 shows that we can indirectly upper bound it by analysing a simpler asymptotically valid confidence set. Our strategy then is to construct confidence sets based on the degree of the nodes whose size is much easier to bound through well-understood probabilistic properties of preferential attachment trees. This leads to our next result which provides explicit bounds on the size of the confidence set $B_\epsilon(\cdot)$ when the underlying tree is LPA.

Theorem 11 Let $\mathbf{G}_n \sim \text{PAPER}(\alpha, \beta, \theta)$ for $\beta = 1$, $\alpha = 0$, and $\theta \in [0, 1]$. For $t \in [n]$, let $D_{G_n}(t)$ be the degree of node with arrival time t and for $k \in [n]$, let $k\text{-max}(D_{G_n})$ be the k-th largest degree of $\mathbf{G}_n$. Let $\delta > 0$ be arbitrary and suppose $\theta \le n^{-\frac{1}{2}-\delta}$. Then, for any $\epsilon > 0$, there exists $L_\epsilon \in \mathbb{N}$ (dependent on δ but not on n) such that

$$\limsup_{n\to\infty}\mathbb{P}\big\{D_{G_n}(1) \le L_\epsilon - \max(D_{G_n})\big\} \le \epsilon. \tag{27}$$

As a direct consequence, if $\theta = O(n^{-\frac{1}{2}-\delta})$ for any $\delta > 0$, then, for any $\epsilon \in (0, 1)$,

$$|B_\epsilon(\mathbf{G}_n^*)| = O_p(1).$$

We relegate the proof of Theorem 11 in Section S4.1 of the online supplementary material and provide a short sketch here: we use results from Peköz et al. (2014) which show that the degree sequence of an LPA tree, when normalised by $\frac{1}{\sqrt{n}}$, converges to a limiting distribution in the ℓ_q sequential metric sense, which shows that equation (27) holds for the tree degree $D_{T_n}(\cdot)$, that is, the degree of the root node is one of the highest among all the nodes. Since $D_{G_n} = D_{T_n} + D_{R_n}$, we show that if the noise level θ is less than $n^{-1/2-\delta}$ for some $\delta > 0$, then the degree of the noisy edges D_{R_n} has a second-order effect and equation (27) remains valid.

We know from existing results (such as Bubeck, Devroye et al., 2017, Theorem 6; see also Crane & Xu, 2021, Corollary 7) that $|B_\epsilon(\boldsymbol{T}_n^*)|$ is $O_p(1)$ in the $\theta = 0$ case where we observe the LPA tree $\boldsymbol{T}_n^*$. Theorem 11 shows that this phenomenon is quite robust to noise. Indeed, when $\theta = n^{-1/2-\delta}$, the observed graph would have approximately $n^{3/2-\delta}$ noisy edges and only $n-1$ tree edges.

The situation is different when the underlying latent tree has the UA distribution, where $\alpha = 1$ and $\beta = 0$. In this case, we have the following result:

Theorem 12 Let $\mathbf{G}_n \sim \text{PAPER}(\alpha, \beta, \theta)$ for $\alpha = 1$, $\beta = 0$, and $\theta \in [0, 1]$. For $t \in [n]$, let $D_{G_n}(t)$ be the degree of node with arrival time t and for $k \in [n]$, let $k\text{-}\max(D_{G_n})$ be the k-th largest degree of $\mathbf{G}_n$. Suppose $\theta = o(\frac{\log n}{n})$ and let $\epsilon \in (0, 1)$ be arbitrary. For any $\eta \in (0, 1)$, define $L_{\eta,n,\epsilon} := n^\eta + \epsilon^{-1}n^{1-(2-\eta)h(\frac{\eta}{2-\eta})}$ where $h(x) = (1+x)\log(1+x) - x$ for $x \ge 0$. Then, we have that

$$\limsup_{n\to\infty}\mathbb{P}\big\{D_{G_n}(1) \le L_{\eta,n,\epsilon}\text{-}\max(D_{G_n})\big\} \le \epsilon. \tag{28}$$

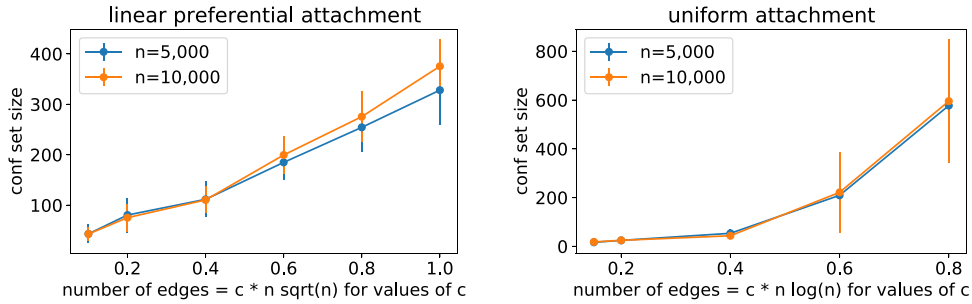


Figure 13. Size of the confidence set vs. the number of edges.

As a direct consequence, if $\theta = o(\frac{\log n}{n})$, then, for some $\gamma \leq 0.8$, we have that

$$n^{-\gamma}\epsilon^{-1}|B_\epsilon(\boldsymbol{G}_n^*)| = O_p(1) \quad \text{for any } \epsilon \in (0, 1).$$

We relegate the proof of Theorem 12 to Section S4.2 of the online supplementary material. The proof technique is similar to that of Theorem 11 except that we use concentration inequalities to derive equation (28).

Comparing Theorem 12 with Theorem 11, we see two important differences. First, even if the noise level is small, we can no longer guarantee that $|B_\epsilon(\boldsymbol{G}_n^*)|$ is bounded even as n increases. Instead, we have the much weaker bound that $|B_\epsilon(\boldsymbol{G}_n^*)|$ is less than $O(n^\gamma)$ for some $\gamma < 0.8$. We believe this bound is not tight; we observe from simulations in Section 6.1 (see Figure 13) that the size of the confidence set $B_\epsilon(\cdot)$ is indeed $O_p(1)$ even when the noise level is of order $\frac{\log n}{n}$. The bound is sub-optimal because the degree of the nodes is not informative of their latent ordering when the latent tree has the UA distribution; hence, $B_\epsilon(\cdot)$ could be much smaller than confidence sets constructed solely from degree information. Intuitively, this is because largest degree nodes do not persist in UA as opposed to linear preferential attachment (Dereich & Mörters, 2009; Galashin, 2013).

The second difference is that the noise tolerance is much smaller. We require θ to be smaller than $\frac{\log n}{n}$ rather than $n^{-1/2}$. We conjecture that these rates are tight in the following sense:

Conjecture 13 Let $\boldsymbol{G}_n \sim \text{PAPER}(\alpha, \beta, \theta)$ for $\alpha = 1$, $\beta = 0$, and $\theta \in [0, 1]$.

1. Suppose $\alpha = 0$ and $\beta = 1$ (LPA). If $\theta = o(n^{-1/2})$, then $|B_\epsilon(\boldsymbol{G}_n^*)| = O_p(1)$ and if $\theta = \omega(n^{-1/2})$, then every asymptotically valid confidence set has size that diverges with n.
2. Suppose $\alpha = 1$ and $\beta = 0$ (UA). If $\theta = o(\frac{\log n}{n})$, then $|B_\epsilon(\boldsymbol{G}_n^*)| = O_p(1)$ and if $\theta = \omega(\frac{\log n}{n})$, then every asymptotically valid confidence set has size that diverges with n.

We provide empirical support for this conjecture in Section 6.1, particularly Figure 13. In those experiments, we see that, when the latent tree has the LPA distribution and when $\theta = cn^{-1/2}$ where $c > 0$ is small, the size of B_ϵ does not increase with n; however, when c (and hence θ) is large, B_ϵ is larger when the size of the graph n is larger. The same phenomenon holds when the latent tree has the UA distribution when $\theta = c\frac{\log n}{n}$.

6 Empirical studies

We have implemented the inference approach in Section 3 and the sampling algorithm in Section 4 in a Python package named `paper-network`, which can be installed via command line `pip install paper-network` on the terminal and then imported in Python via `import PAPER`. The source code of the package, along with examples and documentation, are available at the website https://github.com/nineisprime/PAPER. All the code used in this Section are also available

there under the directory `paperexp`. We also give detailed sampler diagnostics information in Section S5.4 of the online supplementary material.

6.1 Simulation

6.1.1 Frequentist coverage in the single root setting

In our first simulation study, we empirically verify Theorem 7 by showing that a level $1-\epsilon$ credible set for the root node constructed from the posterior root probabilities has frequentist coverage at exactly the same level $1-\epsilon$. We consider three different settings of parameters: $\alpha=0, \beta=1$ (LPA), $\alpha=1, \beta=0$ (UA), and $\alpha=8, \beta=1$. We generate $\boldsymbol{G}_n^*$ according to the PAPER(α, β, θ) model with $n=3{,}000$ nodes and $m=7{,}500$ edges. We then estimate α and β using the method given in online supplementary Section S3.1, compute the level $\epsilon \in \{0.2, 0.05, 0.01\}$ credible sets, and record whether they cover the true root node. We repeat the experiment over 300 independent trials and report the results in Table 2. We observe that the credible sets attain the nominal coverage and that the size of the credile sets are small compared to the number of nodes n.

6.1.2 Size of the confidence set

In our second simulation study, we study the effect of the sample size n and the magnitude of the noisy edge probability θ on the size of the confidence set. We let $\boldsymbol{G}_n^*$ be the observed graph with n nodes and m edges according to the PAPER(α, β, θ) model where we consider $(\alpha, \beta) = (0, 1)$ (LPA) or $(1, 0)$ (UA). Since a tree with n nodes always contains $n-1$ edges, $\frac{n^2}{2}\theta + n$ is approximately equal to the number of edges m in the observed graph $\boldsymbol{G}_n^*$.

We empirically show that the confidence set size does not depend on n so long as θ is much smaller than $n^{-1/2}$ for LPA and much smaller than $\frac{\log n}{n}$ for UA. To that end, we set $m = cn\sqrt{n}$ for $c \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1\}$ for LPA and $m = cn\log n$ for $c \in \{0.15, 0.2, 0.4, 0.6, 0.8\}$ for UA. We then plot the average size of the confidence set with respect to c for $n \in \{5{,}000, 10{,}000\}$. We plot the curve for $n=5{,}000$ and for $n=10{,}000$ on the same figure and observe that, when c is small, the two curves overlap completely but when c is large, the $n=10{,}000$ curve lies above the $n=5{,}000$ curve. This provides empirical support to Theorems 11 and 12. In fact, this experiment shows that the bound of n^{γ} on the size of the confidence set in Theorem 12 is loose; the actual size does not increase with n. The fact that the confidence set size seems to diverge with n when c is larger supports Conjecture 13 and suggests that the problem of root inference exhibits a phase transition when $\theta \approx \frac{1}{\sqrt{n}}$ under the LPA model and $\theta \approx \frac{\log n}{n}$ under the UA model.

Table 2. Empirical coverage of our confidence set for the root node

(α, β)	(0, 1)	(1, 0)	(8, 1)	(0, 1)	(1, 0)	(8, 1)	(0, 1)	(1, 0)
Theoretical coverage	0.8	0.8	0.8	0.95	0.95	0.95	0.99	0.99
Empirical coverage	0.8	0.823	0.82	0.937	0.943	0.94	0.983	0.993
Ave. conf. set size	7	12	9	42	42	31	183	115

Note. We report the average over 300 trials. Graph has $n=3{,}000$ nodes and $m=7{,}500$ edges in all cases.

Table 3. Empirical coverage of our confidence set for the seq-PAPER($\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta}$) model without deletion noise, with $\theta=1.5$ and $\tilde{\alpha}=\alpha$ and $\tilde{\beta}=\beta$

(α, β) (with $\tilde{\alpha}=\alpha, \tilde{\beta}=\beta$)	(0, 1)	(1, 0)	(0, 1)	(1, 0)	(0, 1)	(1, 0)
Theoretical coverage	0.8	0.8	0.95	0.95	0.99	0.99
Empirical coverage	0.795	0.895	0.935	0.965	0.970	0.995
Ave. conf. set size	7	7	25	16	56	28

Note. We report the average over 200 trials. Graph has $n=600$ nodes and around $m \approx 1{,}500$ edges in all cases.

Table 4. Empirical coverage of our confidence set for the seq-PAPER*$(\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta}, \eta)$ model with deletion noise, with $\alpha = 0$, $\beta = 1$, $\tilde{\alpha} = 8$, $\tilde{\beta} = 1$, $\theta = 1.5$ in all cases

η (tree edge deletion probability)	0	0	0.04	0.04	0.08	0.08
Theoretical coverage	0.8	0.95	0.8	0.95	0.8	0.95
Empirical coverage	0.825	0.96	0.84	0.95	0.85	0.98
Ave. conf. set size	5.9	14.1	6.3	15.0	6.7	15.9

Note. We report the average over 200 trials. Graph has $n = 300$ nodes and around $m \approx 750$ edges in all cases.

Table 5. Empirical coverage of our confidence set for the set of $K = 2$ root nodes

(α, β)	(0, 1)	(1, 0)	(0, 1)	(1, 0)	(0, 1)	(1, 0)
Theoretical coverage	0.8	0.8	0.95	0.95	0.99	0.99
Empirical coverage	0.826	0.826	0.933	0.964	0.974	0.985
Ave. conf. set size	5	57	12	155	31	295

Note. We report the average over 200 trials. Graph has $n = 700$ nodes and $m = 1{,}000$ edges in all cases.

6.1.3 *Frequentist coverage under sequential noise models*

In our third simulation study, we verify Theorem 7 for the seq-PAPER model with sequential noise described in Section 2.3. We generate $\mathbf{G}_n^*$ according to both the seq-PAPER$(\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta})$ model and the seq-PAPER*$(\alpha, \beta, \theta, \tilde{\alpha}, \tilde{\beta}, \eta)$ model with deletion noise. We then construct the credible sets for the root node from posterior root probabilities computed via the algorithm given in Section 4. We repeat the experiment over 200 independent trials and report the results in Tables 3 and 4. We observe that the credible sets attain the nominal coverage. We also note that Table 4 shows that the seq-PAPER* model can tolerate tree deletion probability up to $\eta = 0.08$ without significant increase in the confidence set sizes.

6.1.4 *Frequentist coverage for multiple roots*

Our next simulation study is similar to the first except that we generate graphs from the PAPER$(\alpha, \beta, K, \theta)$ model with $K = 2$. We construct our credible sets as described in Section 3.3 and verify Theorem 8 by showing that the credible set at level $1 - \epsilon$ also has frequentist coverage at exactly the same level. We consider two different settings of parameters: $\alpha = 0$, $\beta = 1$ (LPA) and $\alpha = 1$, $\beta = 0$ (UA). We generate $\mathbf{G}_n^*$ according to the PAPER$(\alpha, \beta, K, \theta)$ model with $n = 700$ nodes, $m = 1{,}000$ edges, and $K = 2$. We then estimate α and β using the method given in online supplementary Section S3.1, compute the level $\epsilon \in \{0.2, 0.05, 0.01\}$ credible sets, and record whether they contain the true set of root nodes. We repeat the experiment over 200 independent trials and report the results in Table 5. We observe that the credible sets attain the nominal coverage. In the LPA setting, the size of the credible sets are small but in the UA setting, the sizes of the credible sets become much larger. We relegate an in-depth analysis of this phenomenon to future work.

6.1.5 *Posterior on* K *in the random* K *roots setting*

In our last simulation experiment, we generate PAPER graphs with $K = 2$ roots but perform posterior inference using the PAPER$(\alpha, \beta, \alpha_0, \theta)$ model and study resulting posterior distribution over the number of roots K. We consider two different settings of parameters: $\alpha = 0$, $\beta = 1$ (LPA) and $\alpha = 1$, $\beta = 0$ (UA). We generate $\mathbf{G}_n^*$ according to the PAPER$(\alpha, \beta, K, \theta)$ model with $n = 700$ nodes, $m = 1{,}000$ edges, and $K = 2$. We report the posterior distribution over K, averaged over 20 independent trials, in Figure 14. We observe that, in both cases, the mode of the posterior distribution over K is 2, which is the true number of roots. However, the distributions exhibits high

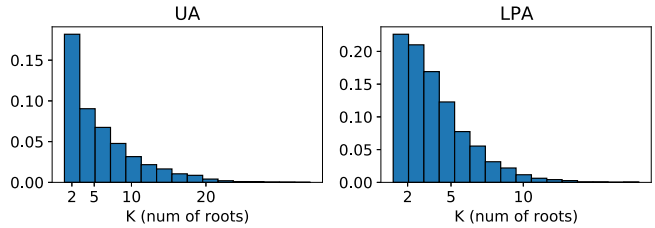


Figure 14. Posterior distribution over K averaged across 20 independent trials. Left: Networks have two latent UA trees. Right: Networks have two latent LPA trees.

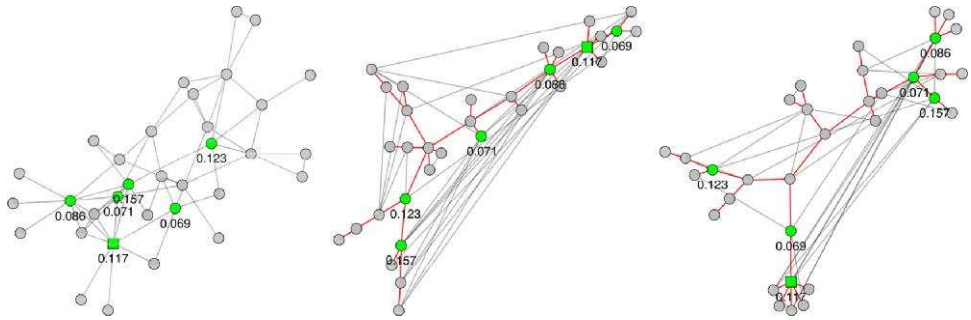


Figure 15. Left: Contact network among 32 students in a flu outbreak. Centre and right: Two examples of the latent tree generated by the Gibbs sampler.

variance, which could be due to the fact that the two true latent trees may have significantly different sizes.

6.2 Single root analysis on real data

We now apply the single root PAPER model on real-world networks. In a few cases (Section 6.2.1), we can ascertain from domain knowledge that the network originated from a single root node but more often, we use the single root model to identify important nodes and subgraphs (Section 6.2.2).

6.2.1 Flu transmission network

We analyse a person-to-person contact network among 32 students in a London classroom during a flu outbreak (Hens et al., 2012). We extract the data from Figure 3 in Hens et al. (2012) and illustrate the network in the left sub-figure of Figure 15. Public health investigation revealed that the outbreak originated from a single student, which is the true patient zero and shown as the orange node in Figure 15. We apply the PAPER model with a single root to this network. We estimate that $\beta = 1$ and $\alpha = 53.06$ using the method described in online supplementary Section S3.1 and compute the 60%, 80%, 95%, and 99% confidence sets. All the confidence sets contain the true patient zero and their sizes are as follows:

$$60\%: 6 \text{ nodes} \quad 80\%: 10 \text{ nodes} \quad 95\%: 19 \text{ nodes} \quad 99\%: 27 \text{ nodes}.$$

We provide the approximate posterior root probabilities of the top 7 nodes in Figure 15. The true patient zero has a posterior root probability of 0.11 is the node with the 3rd highest posterior root probability. In the centre and right sub-figure of Figure 15, we also show two of the latent trees $\tilde{T}_n$ that were generated by the Gibbs sampler.

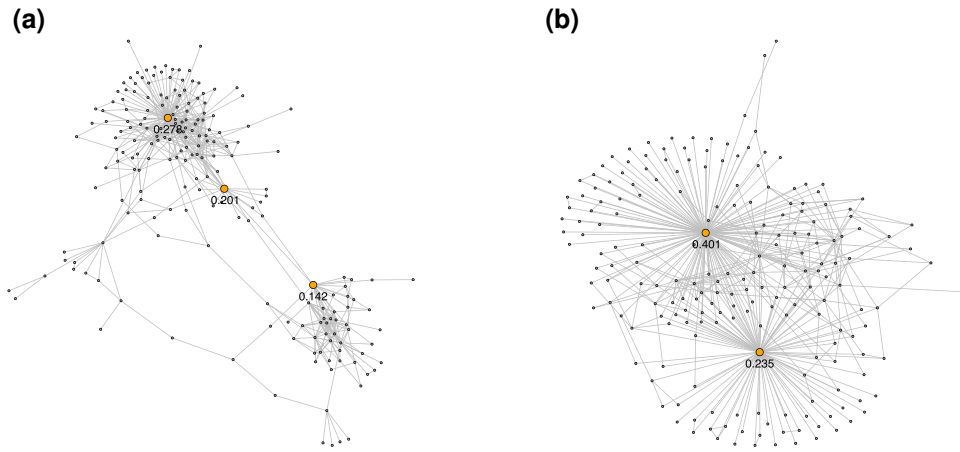


Figure 16. Subgraph of the 200 nodes with highest posterior root probabilities. (a) MathSciNet subgraph and (b) Notre Dame subgraph.

6.2.2 Visualising central subgraphs

Large-scale real graphs are difficult to visualise but one can often learn salient structural properties of a graph by visualising a smaller subgraph that contains the most important nodes. In this section, we apply the single root PAPER model on four large networks and, for each graph, display the subgraph that comprises the 200 nodes with the highest posterior root probability. We see that the result reveals striking differences between the different graphs. Unfortunately, we do not have the node labels on any of these four graphs and can only make qualitative interpretations of the results.

MathSciNet collaboration network. We first consider a collaboration network of research publications from MathSciNet, which is publicly available in the Network Repository (Rossi & Ahmed, 2015) at the link http://networkrepository.com/ca-MathSciNet.php. This network has $n = 332,689$ nodes and $m = 820,644$ edges, with a maximum degree of 496. Using the method described in online supplementary Section S3.1, we estimate $\beta = 1$ and $\alpha = 0$. The sizes of confidence sets are:

$$60\%: 3 \text{ nodes} \quad 80\%: 6 \text{ nodes} \quad 95\%: 21 \text{ nodes} \quad 99\%: 112 \text{ nodes}.$$

We display the subgraph containing the 200 nodes with the highest posterior root probability in Figure 16a. We observe that the subgraph reveals a cluster structure that may represent the different academic disciplines.

University of Notre Dame website network. We study a network of hyperlinks between webpages of University of Notre Dame (Albert et al., 1999), which is publicly available at the website https://snap.stanford.edu/data/web-NotreDame.html. This network has $n = 325,729$ nodes and $m = 1,090,108$ edges, with a maximum degree of 10,721. Using the method described in online supplementary Section S3.1, we estimate $\beta = 1$ and $\alpha = 0$. The sizes of confidence sets are:

$$60\%: 2 \text{ nodes} \quad 80\%: 21 \text{ nodes} \quad 95\%: 524 \text{ nodes} \quad 99\%: 3498 \text{ nodes}.$$

We observe that the central subgraph (shown in Figure 16b) reveals two hub nodes with many sparsely connected 'spokes'.

6.3 Community recovery with the fixed K model

In this section, we show that we can use the PAPER model with multiple roots for community recovery on real-world networks. To estimate the community membership from the posterior samples, we use a greedy matching procedure. To be precise, our Gibbs sampler outputs a sequence of forests $\tilde{f}_n^{(1)}, \ldots, \tilde{f}_n^{(J)}$ where J is the number of Monte Carlo samples. Each forest $\tilde{f}_n^{(j)}$ contains K component trees which we denote $\tilde{t}^{(1,j)}, \tilde{t}^{(2,j)}, \ldots, \tilde{t}^{(K,j)}$. We write $Q_k^{(j)}(\cdot) := \mathbb{P}(\Pi_1 = \cdot \mid \tilde{T} = \tilde{t}^{(k,j)})$ as the posterior root distribution of the k-th tree of the j-th Monte Carlo sample. Since the tree labels may switch from sample to sample, we use the following matching procedure: we maintain K distributions $Q_1(\cdot), Q_2(\cdot), \ldots, Q_K(\cdot)$ and initially set $Q_k = Q_k^{(1)}$ for all $k \in [K]$. Then, for $j = 2, 3, \ldots, J$, we use the Hungarian algorithm to compute a one-to-one matching $\sigma : [K] \to [K]$ that minimises the overall total variation distance

$$\sum_{k=1}^{K} \mathrm{TV}(Q_k^{(j)}, Q_{\sigma(k)}).$$

Once we compute the matching, we then update $Q_{\sigma(k)} \leftarrow \frac{j-1}{j} Q_{\sigma(k)} + \frac{1}{j} Q_k^{(j)}$.

In this way, we interpret $Q_1, \ldots, Q_K$ as the average posterior root distributions for the K trees across all the Monte Carlo samples and using the matching, we may also compute the posterior probability $\mathbb{P}(u \text{ in tree } k \mid \tilde{G}_n = \tilde{g}_n)$, which allows us to perform community detection – we put node u in cluster k if $\mathbb{P}(u \text{ in tree } k \mid \tilde{G}_n = \tilde{g}_n) \geq \mathbb{P}(u \text{ in tree } k' \mid \tilde{G}_n = \tilde{g}_n)$ for all $k' \neq k$. We use the greedy matching procedure for computational efficiency—slower but more principles approaches are studied by, e.g. Wade and Ghahramani (2018).

6.3.1 Karate club network

We apply the PAPER model to Zachary's karate club network Zachary (1977), which is publicly available at http://www-personal.umich.edu/mejn/netdata/. The karate club network has $n = 34$ nodes and $m = 76$ edges, where two individuals share an edge if they socialise with each other. The network has two ground truth communities, one led by the instructor and one led by the administrator (shown as rectangular nodes in Figure 17. These two communities later split into two separate clubs. In this case, we apply the PAPER model with $K = 2$ roots. For every node u, we consider the community membership probability $\mathbb{P}(u \text{ in tree } 1 \mid \tilde{G}_n)$ and assign u to community 1 if and only if this value is greater than 0.5. We show the result in in Figure 17, where each node has a colour that reflects its community membership probability.

We correctly cluster all but one node, which matches the performance of degree-corrected SBM Karrer and Newman (2011) and Amini et al. (2013) (DCSBM)—the current state-of-the-art model for community detection. The node that we misclassify has a posterior probability $\mathbb{P}(u \text{ in tree } 1 \mid \tilde{G}_n) = 0.47$, indicating that the model is indeed unsure of whether it belong in community 1 or 2. We note that the PAPER model requires only 3 parameters whereas the DCSBM for

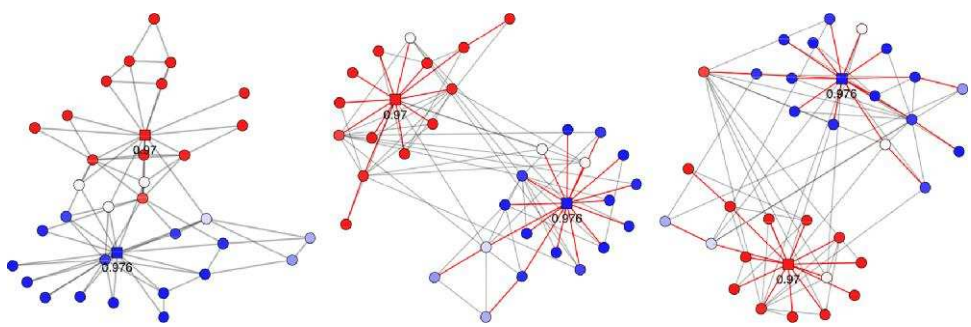


Figure 17. Left: Karate club network where node colour reflects community membership probability. Centre and right: Two examples of the latent forest generated by the Gibbs sampler.

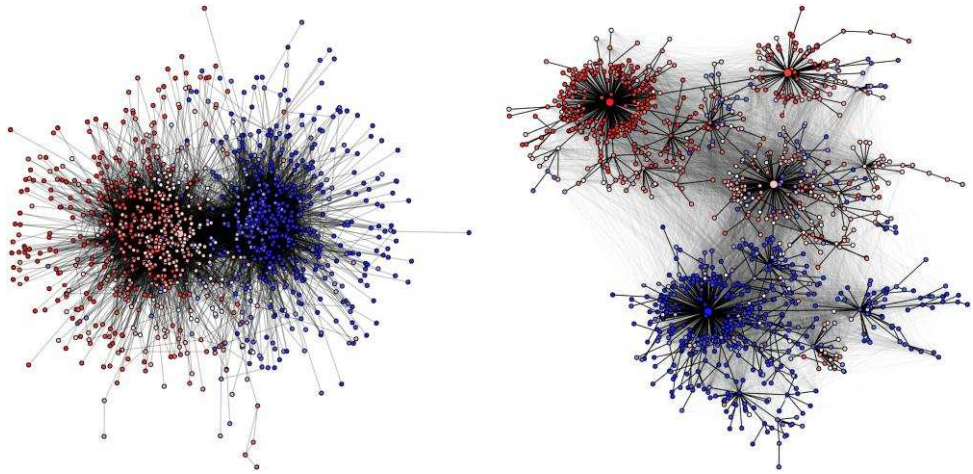


Figure 18. Left: Political blog network where node colour reflects community membership probability. Right: One example of a forest generated by the Gibbs sampler. The 5 nodes with the larger marker comprise the 95% confidence set for the roots.

this network requires 38 parameters because each node has a degree correction parameter. SBM without degree correction performs badly Karrer and Newman (2011).

6.3.2 Political blogs network

Next, we analyse a political blogs network (Adamic & Glance, 2005) that is frequently used as a benchmark for network clustering algorithms; the full network is publicly available at the website http://www-personal.umich.edu/mejn/netdata/. This network contains $m = 16{,}714$ edges between $n = 1{,}222$ blogs, where two blogs are connected if one contains a link to the other. For simplicity, we treat the network as undirected.

The network again has two ground truth communities, one that comprise of left-leaning blogs and one that comprises of right-leaning blogs. We again apply the PAPER model with $K = 2$ roots and for every node u, we compute the community membership probability $\mathbb{P}(u \text{ in tree } 1 \mid \tilde{\boldsymbol{G}}_n)$ and assign u to community 1 if and only if this value is greater than 0.5. We show the result in in Figure 18, where each node has a colour that reflects its community membership probability.

Our overall misclustering error rate is **9.1%**, which is high compared to current state-of-the-art approaches; for example, the SCORE method (Jin, 2015) attains an error rate of about 5%. However, we compute the misclustering error rate with respect to only the top 400 nodes with the highest posterior root probabilities, which can be interpreted as the most important nodes in the graph, our misclustering error rate drops to **3.5%**. This confirms our intuition that the PAPER model, when used for clustering, is more reliable for central nodes than for peripheral nodes.

6.4 Community discovery with the random K model

For networks with an unknown number of small and possibly overlapping communities, the random K model PAPER(α, β, α_0, θ) can be useful for discovering complex community structures. To extract community information from the posterior samples, we again use a greedy matching procedure. To be precise, in the random K setting, our proposed Gibbs sampler outputs a sequence of forests $\tilde{f}_n^{(1)}, \ldots, \tilde{f}_n^{(J)}$ where J is the number of Monte Carlo samples. We write each forest $\tilde{f}_n^{(j)}$, for $j \in [J]$, as a collection of trees $\{\tilde{\boldsymbol{t}}^{(1,j)}, \ldots, \tilde{\boldsymbol{t}}^{(K_j,j)}\}$ where K_j is the number of trees in $\tilde{f}_n^{(j)}$. For $j \in [J]$ and $k \in [K_j]$, we write $Q_k^{(j)}(\cdot) = \mathbb{P}(\Pi_1 = \cdot \mid \tilde{\boldsymbol{T}} = \tilde{\boldsymbol{t}}^{(k,j)})$ as the posterior root distribution of the k-th tree in the j-th Monte Carlo sample. To summarise the output in an interpretable way, we do the following:

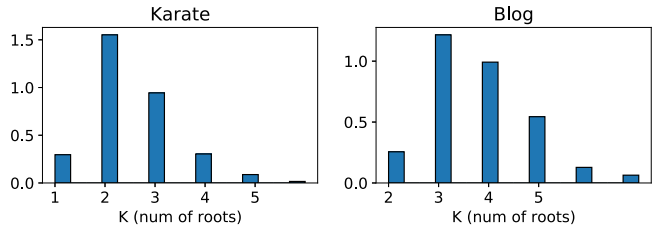


Figure 19. Posterior over K using the random K roots model on the karate club network (left) and the political blog network (right).

1. We initialise $K_{\text{all}} = \max_{j\in[J]} K_j$ and $Q_k = Q_k^{(1)}$ for $k = 1, 2, \ldots, K_1$. For $k = K_1 + 1, \ldots, K_{\text{all}}$, we initialise $Q_k(\cdot) = 0$.
2. For $j = 2, 3, \ldots, J$, we match $\{Q_1, \ldots, Q_{K_{\text{all}}}\}$ with $\{Q_1^{(j)}, \ldots, Q_{K_j}^{(j)}\}$ by computing a one-to-one matching $\sigma : [K_j] \to [K_{\text{all}}]$ that minimises
$$\sum_{k=1}^{K_j} \text{TV}(Q_k^{(j)}, Q_{\sigma(k)}).$$
For every $k \in [K_j]$, if the total variation distance between the k-th pair of the matching is too large, that is $\text{TV}(Q_k^{(j)}, Q_{\sigma(k)}) > 0.75$, then we create a new set $K_{\text{all}} \leftarrow K_{\text{all}} + 1$ and set $Q_{K_{\text{all}}+1} \leftarrow Q_k^{(j)}$; otherwise, we perform the update $Q_{\sigma(k)} \leftarrow \frac{j-1}{j} Q_{\sigma(k)} + \frac{1}{j} Q_k^{(j)}$.
3. We output $\{Q_1, \ldots, Q_{K_{\text{all}}}\}$ as the discovered clusters, represented as posterior root probability distributions.

For all of our experiments, we only include trees that contain at least 1% of the total number of nodes. For each discovered cluster Q_ℓ for $\ell \in [K_{\text{all}}]$, we also compute ρ_{Q_ℓ} as the number of Monte Carlo iteration $j \in [J]$ where we match Q_ℓ with $Q_k^{(j)}$, i.e. $\sigma(k) = \ell$, and update Q_ℓ. We then compute $\frac{\rho_{Q_\ell}}{J}$ as the posterior frequency of cluster Q_ℓ.

In order to check that the random K model is reasonable, we first apply it to the karate club and the political blog networks, which we know contain two underlying clusters, and analyse the resulting posterior distribution over the number of cluster-trees K. We provide the results for the karate club network in the left part of Figure 19, in which we see that the posterior mode is at $K = 2$. For the political blog network, the Gibbs sampler tends to produce a few large clusters and many tiny clusters of fewer than 10 nodes. Therefore, to compute the posterior over K, we count only clusters that have at least 12 nodes (1% of the total number of nodes) and give the results in the right part of Figure 19. The posterior mode in this case is $K = 3$, which is reasonably close to the ground truth.

We also analyse an air route network (Guimera et al., 2005) of $n = 3{,}618$ airports and $m = 14{,}142$ edges where two airports share an edge if there is a regularly scheduled flight between them. We remove the direction of the edges and treat the network as undirected. The dataset is publicly available at http://seeslab.info/downloads/air-transportation-networks/. Using the random K model, we discover a large central cluster containing major airports around the world and various small clusters that correspond to more remote regions such as airports on Pacific and Polynesian islands, airports in Alaska, and airports in the Canadian Northwest Territories. For sake of brevity, we defer the detailed results to Section S5.2 of the online supplementary material.

6.5 Analysis of statistician co-authorship network

We now apply PAPER models to perform an extensive analysis of a statistician co-authorship network constructed by Ji and Jin (2016). In this network, each node corresponds to a statistician and two nodes u and v have an edge between them if they have co-authored 1 or more papers in either *Journal of Royal Statistical Society: Series B*, *Journal of the American Statistical*

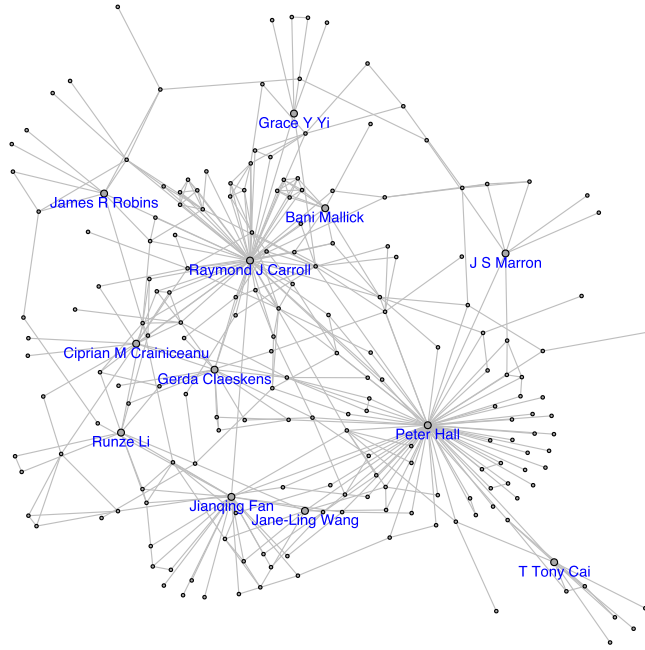


Figure 20. Subgraph of the co-authorship graph comprising the 200 nodes with the highest posterior root probabilities. We label the 12 nodes with the highest root probabilities.

Figure 21. Nine of the clusters that most frequently appear in the posterior samples. Word sizes are proportional to the posterior root probability with respect to the cluster. (a) Central super-cluster. (b) Bayesian. (c) Bayesian. (d) Theory. (e) Multivariate analysis. (f) Biostat. (g) Computation/UK. (h) Biostat. (i) Graphical models.

Association, *Annals of Statistics*, or *Biometrika* from 2002 to 2013. We consider only the largest connected component which has $n = 2{,}263$ nodes and $m = 4{,}388$ edges. Ji and Jin (2016) in their manuscript (Section 4.3) refer to this network as 'Coauthorship Network (B)'. We emphasise that since the data reflect only co-authorship in four journals in the period 2002–2013, the results that

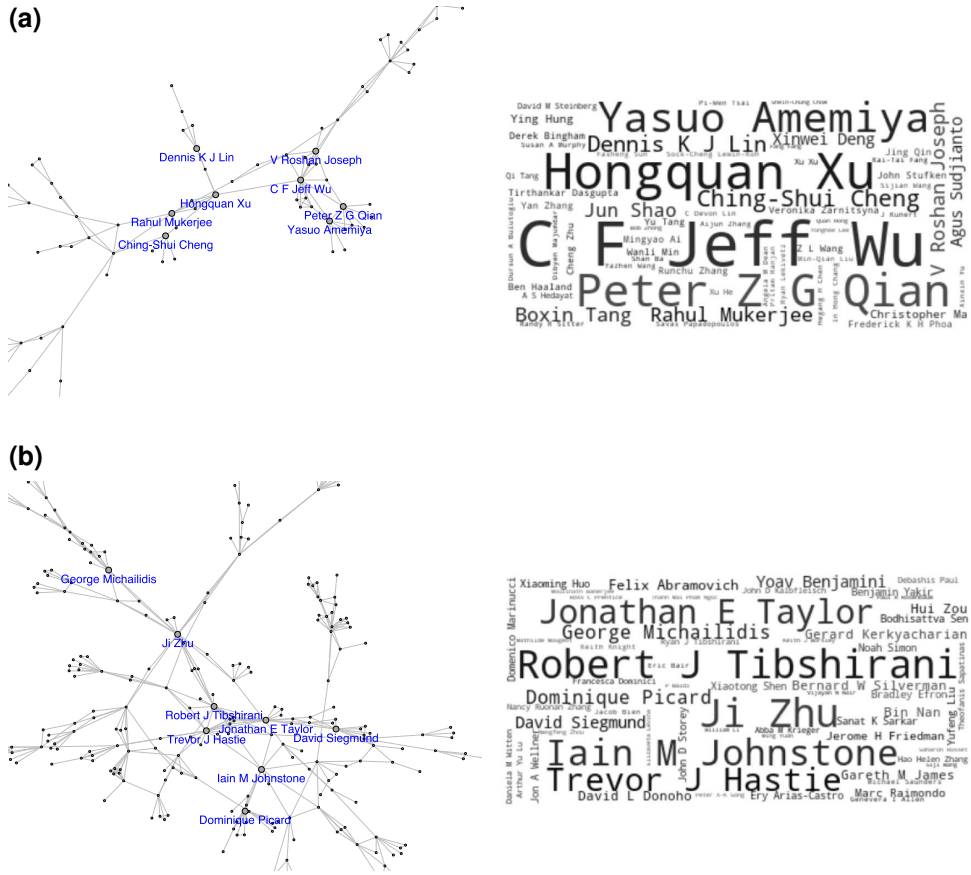


Figure 22. Two additional clusters along with the subgraphs that correspond to the clusters. In the subgraph, we label the 8 nodes with the highest posterior root probability with respect to that cluster. We observe that the subgraphs are tree-like. (a) Experimental design community. (b) High-dimensional statistics community.

we produce cannot be used to compare researchers—we use this network only to illustrate PAPER models in a setting where we can more easily assess whether the output is meaningful or not.

6.5.1 Single root analysis

We first use the single root PAPER(α, β, θ) model where we estimate $\alpha = 0, \beta = 1$ using the EM algorithm described in online supplementary Section S3.1. We find that the following 4 nodes have the highest posterior root probabilities: (1) Raymond Carroll with root probability 0.32, (2) Peter Hall with root probability 0.26, (3) Jianqing Fan with root probability 0.086, and (4) James Robins with root probability 0.048. The root probability ranking align closely with betweenness centrality ranking, in which Raymond Carroll, Peter Hall, and Jianqing Fan are also the top 3 most central nodes; see Table 2 of Ji and Jin (2016). Both the root probability ranking and the betweenness ranking differ significantly from degree ranking. We also display the subgraph of the 200 nodes with the highest posterior root probabilities in Figure 20 where we labelled the top 12 nodes with the highest root probabilities.

6.5.2 Community detection with random K *roots model*

Using our inference algorithm and the greedy matching procedure in Section 6.4, we compute clusters $\{Q_1, \ldots, Q_{K_{\text{all}}}\}$ where we find about $K_{\text{all}} \approx 40$ significant clusters. We order the clusters by their posterior frequencies and display the top 9 clusters in Figure 21, along with labels that we

curated; we display the nodes in the cluster as word clouds in which the word size is proportional to the posterior root probabilities. We display 18 additional clusters in Section S5.3 of the online supplementary material. We note that the clusters can overlap since they are constructed from a sequence of posterior samples by matching; see the first paragraph of Section 6.4.

Ji and Jin (2016) on the same network uses scree plot to conclude that there are $K = 3$ clusters, which are shown in Figures 9–11 in their paper. They refer to the three clusters as a 'high-dimensional' super-cluster, a 'biostatistics' cluster, and a 'Bayes' cluster. We find a giant super-cluster, but we also find a large number of smaller clusters which accurately reflect actual research communities in statistics. For example, we find the same 'Bayes' cluster in Ji and Jin (2016) (see Figure 21a), but we also discover other Bayesian clusters such as ones shown in Figure 21b. Similarly, we find the 'biostat' community in Ji and Jin (2016) (see Figure 21f) but we find other biostat clusters as well such as the one shown in Figure 21h and the one centred on Jason Fine and Michael Korsorok in Figure 27 in the online supplementary material. In addition, we find many other meaningful communities, such as the experimental design community or the high-dimensional statistics community shown in Figure 22, or the survey and theory community in Figure 27 in the online supplementary material. We believe that PAPER model gives highly coherent clusters for this network because the network itself is locally tree-like, as shown in two cluster subgraphs that we display in Figure 22.

7 Discussion

In this paper, we present the PAPER model for networks with underlying formation processes and formalise the problem of root inference. We extend the PAPER model to the setting of multiple roots to reflect the growth of multiple communities. There are a number of important open questions from modelling, theoretical, and algorithmic perspectives.

From a modelling perspective, an interesting direction is to suppose that the graph start not as singleton nodes but as a small subgraph. The goal then is to infer the seed-graph instead of the root node (c.f. Devroye & Reddad, 2018). Model extensions such as the PAPER–SBM mixture described in Remark 5 are also interesting; in these models, a subtle question is to what extend we have to estimate the parameters of the noise model well in order to recover the root nodes of the latent forest.

There are many open theoretical questions related to PAPER model and root inference. For instance, in Conjecture 13, we hypothesise that the size of the optimal confidence set for the root node is of a constant order if so long as the noise level is below a certain threshold. If the noise level is above the threshold, then every confidence set has size that diverges with n. The lower bound of this conjecture seems especially difficult and may require new techniques. Another interesting theoretical question is the analysis of community recovery using the PAPER model with multiple roots. Intuitively, we expect be able to correctly cluster the early nodes since they tend to have more central positions in the final graph. The late arriving nodes on the other hand would be more peripheral and difficult to cluster.

Algorithmically, we observe that the Gibbs sampler that we derived in Section 4 converges very quickly in practice (see online supplementary Section S5.4). It would be interesting to study its mixing time, especially how the mixing time depends on the noise level.

Acknowledgments

M.X. is grateful to Justin Khim for insightful discussions in the early stage of the work and to Rong Chen for helpful feedback and comments. The authors would like to thank the anonymous referees for insightful comments which helped to improve the paper.

Conflict of interest: None declared.

Funding

This work is supported by the U.S. National Science Foundation DMS grant #2113671.

Data availability

The flu transmission network is extracted from Figure 3 in Hens et al. (2012). The MathSciNet network is available at http://networkrepository.com/ca-MathSciNet.php.

The University of Notre Dame website network is available at https://snap.stanford.edu/data/web-NotreDame.html. The karate club network is available at http://www-personal.umich.edu/mejn/netdata/. The political blog network is available at http://www-personal.umich.edu/mejn/netdata/.

The airport network is available at http://seeslab.info/downloads/air-transportation-networks/. The statistician co-authorship network is available at http://zke.fas.harvard.edu/MADStat.html.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Abbe E. (2017). Community detection and stochastic block models: Recent developments. *The Journal of Machine Learning Research*, *18*(1), 6446–6531. https://doi.org/10.1561/9781680834772

Adamic L. A., & Glance N. (2005). The political blogosphere and the 2004 us election: Divided they blog. In *Proceedings of the 3rd international workshop on Link Discovery* (pp. 36–43). https://doi.org/10.1145/1134271.1134277

Addario-Berry L., & Eslava L. (2018). High degrees in random recursive trees. *Random Structures & Algorithms*, *52*(4), 560–575. https://doi.org/10.1002/rsa.v52.4

Aiello W., Chung F., & Lu L. (2000). A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of Computing* (pp. 171–180). https://doi.org/10.1515/9781400841356.259

Albert R., Jeong H., & Barabási A.-L. (1999). Diameter of the world-wide web. *Nature*, *401*(6749), 130–131. https://doi.org/10.1038/43601

Aldous D. J. (1990). The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, *3*(4), 450–465. https://doi.org/10.1137/0403039

Amini A. A., Chen A., Bickel P. J., & Levina E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Annals of Statistics*, *41*(4), 2097–2122. https://doi.org/10.1214/13-AOS1138

Athreya A., Fishkind D. E., Tang M., Priebe C. E., Park Y., Vogelstein J. T., Levin K., Lyzinski V., & Qin Y. (2017). Statistical inference on random dot product graphs: A survey. *The Journal of Machine Learning Research*, *18*(1), 8393–8484.

Banerjee S., & Bhamidi S. (2020). 'Root finding algorithms and persistence of Jordan centrality in growing random trees', arXiv, arXiv:2006.15609, preprint: not peer reviewed.

Banerjee S., & Huang X. (2021). 'Degree centrality and root finding in growing random networks', arXiv, arXiv:2105.14087, preprint: not peer reviewed.

Barabási A.-L. (2016). *Network science*. Cambridge University Press.

Barabási A.-L., & Albert R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Bhamidi S. (2007). Universal techniques to analyze preferential attachment trees: Global and local analysis. preprint: not peer reviewed. Preprint available at http://www.unc.edu/~bhamidi/preferent.pdf

Bloem-Reddy B., Foster A., Mathieu E., & Teh Y. W. (2018). Sampling and inference for beta neutral-to-the-left models of sparse networks. In *Proceedings of the thirty-fourth conference on Uncertainty in Artificial Intelligence (UAI)*, Monterey, CA (pp. 477–486).

Bloem-Reddy B., & Orbanz P. (2018). Random-walk models of network formation and sequential monte carlo methods for graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 871–898. https://doi.org/10.1111/rssb.12289

Bollobás B., Riordan O., Spencer J., & Tusnády G. (2001). The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, *18*(3), 279–290. https://doi.org/10.1002/rsa.1009

Briend S., Calvillo F., & Lugosi G. (2022). 'Archaeology of random recursive dags and cooper-frieze random networks', arXiv, arXiv:2207.14601, preprint: not peer reviewed.

Broder A. (1989). Generating random spanning trees. In *30th annual symposium on Foundations of Computer Science*, Research Triangle Park, NC. https://doi.org/10.1109/SFCS.1989.63516

Bubeck S., Devroye L., & Lugosi G. (2017). Finding Adam in random growing trees. *Random Structures & Algorithms*, *50*(2), 158–172. https://doi.org/10.1002/rsa.v50.2

Bubeck S., Eldan R., Mossel E., & Rácz M. Z. (2017). From trees to seeds: On the inference of the seed from large tree in the uniform attachment model. *Bernoulli*, *23*(4A), 2887–2916. https://doi.org/10.3150/16-BEJ831

Bubeck S., Mossel E., & Rácz M. Z. (2015). On the influence of the seed graph in the preferential attachment model. *IEEE Transactions on Network Science and Engineering*, *2*(1), 30–39. https://doi.org/10.1109/TNSE.2015.2397592

Callaway D. S., Newman M. E., Strogatz S. H., & Watts D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, *85*(25), 5468–5471. https://doi.org/10.1103/PhysRevLett.85.5468

Cantwell G. T., St-Onge G., & Young J. -G. (2019). 'Recovering the past states of growing trees', arXiv, arXiv:1910.04788, preprint: not peer reviewed.

Cantwell G. T., St-Onge G., & Young J. -G. (2021). Inference, model selection, and the combinatorics of growing trees. *Physical Review Letters*, *126*(3), 038301. https://doi.org/10.1103/PhysRevLett.126.038301

Crane H. (2016). The ubiquitous Ewens sampling formula. *Statistical Science*, *31*(1), 1–39. http://doi.org/10.1214/15-STS529

Crane H., & Xu M. (2021). Inference on the history of a randomly growing tree. *Journal of Royal Statistical Society, Series B*, *83*(4), 639–668. http://doi.org/10.1111/rssb.12428.

Curien N., Duquesne T., Kortchemski I., & Manolescu I. (2015). Scaling limits and influence of the seed graph in preferential attachment trees. *Journal de l'École polytechnique—Mathématiques*, *2*, 1–34. https://doi.org/10.5802/jep.15

Dereich S., & Mörters P. (2009). Random networks with sublinear preferential attachment: Degree evolutions. *Electronic Journal of Probability*, *14*, 1222–1267. https://doi.org/10.1214/EJP.v14-647

Devroye L., & Reddad T. (2018). 'On the discovery of the seed in uniform attachment trees', arXiv, arXiv:1810.00969, preprint: not peer reviewed.

Diaconis P., & Janson S. (2007). 'Graph limits and exchangeable random graphs', arXiv, arXiv:0712.2749, preprint: not peer reviewed.

Drmota M. (2009). *Random trees: An interplay between combinatorics and probability*. Springer Science & Business Media.

Fioriti V., Chinnici M., & Palomo J. (2014). Predicting the sources of an outbreak with a spectral technique. *Applied Mathematical Sciences*, *8*, 6775–6782. https://doi.org/10.12988/ams.2014.49693

Galashin P. (2013). 'Existence of a persistent hub in the convex preferential attachment model', arXiv, arXiv:1310.7513, preprint: not peer reviewed.

Gao C., Lu Y., & Zhou H. H. (2015). Rate-optimal graphon estimation. *Annals of Statistics*, *43*(6), 2624–2652. https://doi.org/10.1214/15-AOS1354

Guimera R., Mossa S., Turtschi A., & Amaral L. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, *102*(22), 7794–7799. https://doi.org/10.1073/pnas.0407994102

Hens N., Calatyud L., Kurkela S., Tamme T., & Wallinga J. (2012). Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. *American Journal of Epidemiology*, *176*(3), 196–203. https://doi.org/10.1093/aje/kws006

Hoff P. D., Raftery A. E., & Handcock M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Ji P., & Jin J. (2016). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, *10*(4), 1779–1812.http://doi.org/10.1214/15-AOAS896

Jiang J., Wen S., Yu S., Xiang Y., & Zhou W. (2016). Identifying propagation sources in networks: State-of-the-art and comparative studies. *IEEE Communications Surveys & Tutorials*, *19*(1), 465–481. https://doi.org/10.1109/COMST.2016.2615098

Jin J. (2015). Fast community detection by SCORE. *Annals of Statistics*, *43*(1), 57–89. https://doi.org/10.1214/14-AOS1265

Karrer B., & Newman M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, *83*(1), 016107. https://doi.org/10.1103/PhysRevE.83.016107

Khim J., & Loh P.-L. (2017). Confidence sets for the source of a diffusion in regular trees. *IEEE Transactions on Network Science and Engineering*, *4*(1), 27–40. https://doi.org/10.1109/TNSE.6488902

Knuth D. E. (1997). *The art of computer programming: Volume 1: Fundamental algorithms*. Addison-Wesley Professional.

Kolaczyk E. D. (2009). *Statistical analysis of network data: Methods and models*. Springer Series in Statistics. Springer.

Lugosi G., & Pereira A. S. (2019). Finding the seed of uniform attachment trees. *Electronic Journal of Probability*, *24*, 1–15. https://doi.org/10.1214/19-EJP268

Na H. S., & Rapoport A. (1970). Distribution of nodes of a tree by degree. *Mathematical Biosciences*, *6*, 313–329. https://doi.org/10.1016/0025-5564(70)90071-4

Peköz E. A., Röllin A., & Ross N. (2014). 'Joint degree distributions of preferential attachment random graphs', arXiv, arXiv:1402.4686, preprint: not peer reviewed.

Rossi R. A., & Ahmed N. K. (2015). *The network data repository with interactive graph analytics and visualization*. AAAI. http://networkrepository.com.

Schervish M. J. (1995). *Theory of statistics*. Springer Series in Statistics. Springer.

Shah D., & Zaman T. (2011). Rumors in a network: Who's the culprit?. *IEEE Transactions on Information Theory*, *57*(8), 5163–5181. https://doi.org/10.1109/TIT.2011.2158885

Shelke S., & Attar V. (2019). Source detection of rumor in social network—A review. *Online Social Networks and Media*, *9*, 30–42. https://doi.org/10.1016/j.osnem.2018.12.001

Sheridan P., Yagahara Y., & Shimodaira H. (2008). A preferential attachment model with poisson growth for scale-free networks. *Annals of the Institute of Statistical Mathematics*, *60*(4), 747–761. https://doi.org/10.1007/s10463-008-0181-5

Sheridan P., Yagahara Y., & Shimodaira H. (2012). Measuring preferential attachment in growing networks with missing-timelines using Markov chain Monte Carlo. *Physica A: Statistical Mechanics and its Applications*, *391*(20), 5031–5040. https://doi.org/10.1016/j.physa.2012.05.041

Sreedharan J. K., Magner A., Grama A., & Szpankowski W. (2019). Inferring temporal information from a snapshot of a dynamic network. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-38912-0

Van Der Hofstad R. (2016). *Random graphs and complex networks* (Vol. *1*). Cambridge University Press.

Wade S., & Ghahramani Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, *13*(2), 559–626. https://doi.org/10.1214/17-BA1073

Wilson D. B. (1996). Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing* (pp. 296–303). https://doi.org/10.1145/237814.237880

Xie F., & Xu Y. (2019). 'Optimal Bayesian estimation for random dot product graphs', arXiv, arXiv:1904.12070, preprint: not peer reviewed.

Xu M., Jog V., & Loh P.-L. (2018). Optimal rates for community estimation in the weighted stochastic block model. *Annals of Statistics*, *48*(1), 183–204. http://doi.org/10.1214/18-AOS1797

Young J.-G., St-Onge G., Laurence E., Murphy C., Hébert-Dufresne L., & Desrosiers P. (2019). Phase transition in the recoverability of network history. *Physical Review X*, *9*(4), 041056. https://doi.org/10.1103/PhysRevX.9.041056

Zachary W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, *33*(4), 452–473. https://doi.org/10.1086/jar.33.4.3629752

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 866–895

Journal of the Statistics Society Series B Statistical Methodology

Discussion Paper Contribution

Proposers of the vote of thanks to Crane and Xu and contribution to the Discussion of 'Root and community inference on the latent growth process of a network'

Varun Jog and Po-Ling Loh

Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK

Address for correspondence: Po-Ling Loh, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB58LE, UK. Email: pll28@cam.ac.uk

We commend the authors on an illuminating article which significantly extends the state of the art in network archaeology. Notably, whereas most other literature in the area has provided methods for root inference which are theoretically grounded only in the case of random trees, the techniques in this paper have been designed to model nontree-structured networks, as well. This has important practical implications, as many real-world networks contain one or more cycles. The main challenge faced by the probability and statistics community—for the better part of the last decade—was that it was difficult to concoct a model that was simple enough to analyse rigorously, yet complex enough to comprise a sufficiently interesting family of random graph structures. By introducing their PAPER model, Crane and Xu have proposed a model achieving an ideal balance, accompanied by a variety of attractively useful and strikingly intuitive theoretical results. The simulation results are also very impressive, in that the algorithms can be efficiently run on a single laptop even for networks with millions of nodes.

The fact that the PAPER model achieves such versatility with only three parameters is a major strength of the work. This makes parameter estimation plausible (the authors suggest several methods for parameter estimation in the appendices, based on EM and Bayesian methods). From our reading, we wondered how one might decide in practice to use a more complicated model for a given network, e.g. sequential PAPER vs. nonsequential PAPER, or fixed-K vs. random K, as the more complicated networks would of course lead to more complicated inference procedures. For instance, is anything known about degree profiles of random-K networks, and are they qualitatively different from the degree profiles of fixed-K networks? (Which would be more reasonable for specific problems such as the coauthor networks studied in this paper?) We are also curious to understand better, in practice, what the confidence sets look like for $K > 1$ (does the algorithm output K connected components?). How might the algorithm be modified if the goal is to output a set of K-tuples such that the K-tuple containing the K root nodes is contained in this set, with a certain high probability?

Section 5 of the paper presents theoretical results concerning confidence sets for the root node (or root nodes, when $K > 1$) of a random network. Along the lines of some of our previous work (Jog & Loh, 2018), we wondered how much the theory could extend to the case of constructing a confidence set for the first $M > 1$ nodes (say, for simplicity, in the case when $K = 1$). Also, can reasonable conditions be imposed under which confidence sets are proven to 'persist' when the number of nodes n tends to infinity? Finally, if the random network were initialized with a 'seed graph'

Received: January 4, 2024. **Accepted:** January 22, 2024

(Lugosi & Pereira, 2019), could confidence sets for the seed graph be constructed by an extension of the techniques in this paper, with corresponding theoretical results?

Although several variants of the PAPER model are introduced in Section 2 of the paper, the core theoretical results seem to only cover the vanilla and fixed-K cases. Thus, we wondered whether results such as the fact that credible sets have the correct frequentist coverage (Theorems 7 and 8) also holds for sequential and random-K models. Does the minimality property of confidence sets based on posterior root probabilities (Remark 6/Theorem S5) also hold for $K > 1$? If any of the answers are negative, can the authors comment on the technical difficulties involved in extending their results to these cases?

Lastly, we found the results of the paper to be rather thought-provoking in terms of their potential for opening new avenues for studying more complicated random growth models. Some examples include sequential models with a 'vertex retirement' feature, e.g. motivated by the study of coauthorship networks. Another idea is to model in deletion noise which is time-dependent, in contrast to the versions of PAPER presented in the paper, which assume that once connections are formed in the sequential model, they remain fixed forever. Other worthwhile models to study in conjunction with PAPER might be the sublinear preferential attachment model, which also exhibits persistence properties analogous to the linear preferential attachment model (Jog & Loh, 2016), and the 'superstar model', which purportedly provides a better fit to empirical data in social networks than the standard preferential attachment model (Bhamidi et al., 2015).

Conflict of interest: None declared.

References

Bhamidi S., Steele J. M., & Zaman T. (2015). Twitter event networks and the superstar model. *The Annals of Applied Probability*, *25*(5), 2462–2502. https://doi.org/10.1214/14-AAP1053

Jog V., & Loh P. (2016). Analysis of centrality in sublinear preferential attachment trees via the Crump-Mode-Jagers branching process. *IEEE Transactions on Network Science and Engineering*, *4*(1), 1–12. https://doi.org/10.1109/TNSE.2016.2622923

Jog V., & Loh P. (2018). Persistence of centrality in random growing trees. *Random Structures & Algorithms*, *52*(1), 136–157. https://doi.org/10.1002/rsa.v52.1

Lugosi G., & Pereira A. S. (2019). Finding the seed of uniform attachment trees. *Electronic Journal of Probability*, *24*(18), 1–15. https://doi.org/10.1214/19-EJP268

https://doi.org/10.1093/jrsssb/qkae043
Advance access publication 4 June 2024

Seconder of the vote of thanks to Crane and Xu and contribution to the Discussion of 'Root and community inference on the latent growth process of a network'

Patrick Rubin-Delanchy

University of Edinburgh, UK

Address for correspondence: Patrick Rubin-Delanchy, School of Mathematics, University of Edinburgh, Edinburgh EH93FD, UK. Email: patrick.rubin-delanchy@ed.ac.uk

This paper and general line of work present a fascinating array of new ideas for network inference. The central contribution is a radically new model for community structure in graphs, which has several interesting features. To pick one which is perhaps significant given the recent obvious

practical success of AI systems: It is good to see more theory challenging ideas such as the 'curse of dimensionality' or, in this instance, the 'curse of sparsity' for community detection, reminding us that these pessimistic predictions often depend on models for data that may be unrealistic.

At the meeting, I raised several points of discussion, including reproducing high triangle counts (Seshadhri et al., 2020), modelling heterophilic connectivity (Rubin-Delanchy et al., 2022), incorporating continuous latent structure (Athreya et al., 2017, 2021; Hoff et al., 2002; Rubin-Delanchy, 2020), and randomization. But my most important concern at the time was that in the applications cited—finding patient zero in a disease network, or the source of fake news in a social media network—there would almost always be timing information on the edges. It would seem highly irresponsible to ignore this in practice. In a conversation after the meeting, Prof. Xu made the compelling counterargument that existing approaches were often overreliant on time. Still, I think there could be *some* appropriate use of this information.

Since the meeting, I have remained unsure about the role of randomization in this work. The authors write: 'Our approach to root inference and related problems is to randomize the node labels, which induces a posterior distribution over the latent ordering.' I initially read this literally, thinking that we would be working exclusively on a computer-generated, uniformly random relabelling of the graph. However, pushing G^* up the equations at the bottom of page 14, it becomes clear that we are conditioning on the event (shuffled graph) $\tilde{G} = G^*$ (observed graph). In other words, for the computation of the confidence set $B(G^*)$, we are *imagining* that a computer shuffled our graph, and that it spat out G^*. In my view that is not so different and requires just as much 'double-think' as assuming the original labels of G^* were chosen uniformly at random, which I presume is what the authors were trying to avoid. In any case, at a mathematical level, randomization seems an inefficient way of stripping label information away from the problem. Below, I present a different treatment which

1. makes away with any sort of randomization, conceptual or otherwise, and associated Bayesian/frequentist explanations;
2. provides a stronger, *conditional* rather than marginal coverage guarantee;
3. shows that results such as Theorem 7 hold more generally, to any sort of data (e.g. time series, documents, complex networks) which are observed relabelled or disordered.

I would like to thank my colleague Dr Charles Cox (University of Bristol) for his help with group theory and several calculations.

Problem. $G = (\mathcal{V}, \mathcal{E})$ is an undirected random graph on the vertex set $\mathcal{V} = [n] = \{1, \ldots, n\}$, with vertex 1 described as the 'root'. G has a fully specified distribution, with probability mass function f. We will use $n = 4$, $G \sim \text{PAPER}(0, 1, 0)$ (linear preferential attachment) for illustration purposes, in which case for example $f(1\text{—}2\text{—}3\text{—}4) = \mathbb{P}(G = 1\text{—}2\text{—}3\text{—}4) = 1/2 \times 1/4 = 1/8$ (3 chooses 2 with probability 1/2, 4 chooses 3 with probability 1/4). Given a permutation π of $[n]$ (a bijection from $[n]$ to $[n]$), we write $\pi i = \pi(i)$, $\pi S = \{\pi i : i \in S\}$ and $\pi G = ([n], \{(\pi i, \pi j) : (i, j) \in \mathcal{E}\})$.

Instead of G, we observe $G^* = \rho G$, where ρ is a random permutation of $[n]$, whose conditional distribution given G is unknown. (Unlike the authors I do not find it helpful to introduce a different, alphabetical, set of vertex labels for G^*.) Given some coverage probability $1 - \epsilon$, the problem is to find a set $B(G^*)$ such that $\mathbb{P}\{\rho(1) \in B(G^*)\} \geq 1 - \epsilon$, that is, a confidence set for the root.

Solution. Given an arbitrary, fixed graph $\mathcal{G}$ on the vertex set $[n]$, define $\text{Aut}(\mathcal{G}) = \{\pi \in S_n : \pi\mathcal{G} = \mathcal{G}\}$, the automorphism group of $\mathcal{G}$, where S_n is the group of all permutations of $[n]$. The vertex set of $\mathcal{G}$ admits a unique partition into *orbits*, $o_1, \ldots, o_L$, where each $o_l = \{\pi i : \pi \in \text{Aut}(\mathcal{G})\}$ for some $i \in [n]$. It is a standard fact of group theory that $\{\pi i : \pi \in \text{Aut}(\mathcal{G})\}$ and $\{\pi j : \pi \in \text{Aut}(\mathcal{G})\}$ are either equal or disjoint. Let $\mathcal{G}_1, \ldots, \mathcal{G}_M$ denote the distinct graphs which can be obtained from $\mathcal{G}$ by relabelling, where $M = n!/|\text{Aut}(\mathcal{G})|$, that is, $\{\mathcal{G}_1, \ldots, \mathcal{G}_M\} = \{\pi\mathcal{G} : \pi \in S_n\}$.

To illustrate the constructions so far, if $\mathcal{G} = 1\text{—}2\text{—}3\text{—}4$ then $\text{Aut}(\mathcal{G})$ contains only two elements, the reverse permutation $\pi i = 5 - i$ and the identity, the orbits are $\{1, 4\}$ and

{2, 3}, and there are $4!/2 = 12$ distinct graphs $\mathcal{G}_1, \ldots, \mathcal{G}_{12}$, one of which is $\mathcal{G}$, another is 3—1—2—4, etc.

Lemma 1 Let π_2 be a permutation satisfying $\mathcal{G}_2 = \pi_2\mathcal{G}_1$ and $o^{(1)}$ an orbit of $\mathcal{G}_1$. Then $\pi_2 o^{(1)}$ is an orbit of $\mathcal{G}_2$. Let π_1 be a permutation such that $\mathcal{G}_1 = \pi_1\mathcal{G}_2$. Then $\pi_1\pi_2 o^{(1)} = o^{(1)}$.

Proof. It is clear that $\pi_2 o$ is an orbit: if we relabel a graph, we relabel its orbits. For the second part, note that $\mathcal{G}_1 = \pi_1\pi_2\mathcal{G}_1$ so that $\pi_1\pi_2$ is an element of $\mathrm{Aut}(\mathcal{G}_1)$, and so $(\pi_1\pi_2)^{-1}\mathrm{Aut}(\mathcal{G}_1) = \mathrm{Aut}(\mathcal{G}_1)$ (any element h of a group H satisfies $hH = h^{-1}H = H$). Pick some $j \in o^{(1)}$ such that $o^{(1)} = \{\pi j : \pi \in \mathrm{Aut}(\mathcal{G}_1)\} = \{\pi j : \pi \in (\pi_1\pi_2)^{-1}\mathrm{Aut}(\mathcal{G}_1)\}$. Then $\pi_1\pi_2 o^{(1)} = \{\pi_1\pi_2\pi j : \pi \in (\pi_1\pi_2)^{-1}\mathrm{Aut}(\mathcal{G}_1)\} = \{\pi j : \pi \in \mathrm{Aut}(\mathcal{G}_1)\} = o^{(1)}$. □

We can 'track' orbits across different relabelling of $\mathcal{G}$ in a way that we can't do with individual vertices. In general, it is not the case that for π_1, π_2 as above, $\pi_1\pi_2 i = i$ for every $i \in [n]$, unless $\mathcal{G}$ has no non-trivial symmetries.

Now, condition on the event G^*, $G \in \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}$ and fix an orbit o of $\mathcal{G}$. By the above, this corresponds to a well-defined orbit $o^{(1)}$ of $\mathcal{G}_1$, $o^{(2)}$ of $\mathcal{G}_2$, and so on. There is a random μ for which $G^* = \mathcal{G}_\mu$, and an associated random orbit $o^{(\mu)}$.

Theorem 1

$$p(o) := \mathbb{P}[\rho(1) \in o^{(\mu)} \mid G^*, G \in \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}] = \frac{\sum_{m\in C} f(\mathcal{G}_m)}{\sum_{m\in[M]} f(\mathcal{G}_m)},$$

$$\text{where } C = m : 1(m)C = m : 1 \in o(m)C = \{m : 1 \in o^{(m)}\}.$$

Suppose that $\mathcal{G}$ = 1—2—3—4 and $o = \{1, 4\}$. In plain English, we could describe the event $\rho(1) \in o^{(\mu)}$ as: 'one of the tail nodes of G^* is the root node of G'. Under the PAPER(0, 1, 0) model, the conditional probability above evaluates to 1/4. It's three times more likely (3/4) that the root is one of the middle nodes.

Suppose the orbits o_l of $\mathcal{G}$ are ordered by decreasing density, $p(o_l)/|o_l|$, and pick the smallest ℓ such that a) $\sum_{l\in[\ell]} p(o_l) \geq 1 - \epsilon$ and b) $p(o_\ell) > p(o_{\ell+1})$, ignoring the latter condition if the former requires $\ell = L$. Let $B(\mathcal{G}) = \cup_{l\in[\ell]} o_l$.

Let $o_l^{(m)}$ denote the corresponding orbits in $\mathcal{G}_m$. Then we can verify that $B(\mathcal{G}_m) = \cup_{l\in[\ell]} o_l^{(m)}$. Thus,

$$\begin{aligned}\mathbb{P}[\rho(1) \in B(G^*) \mid G^*, G \in \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}] &= \mathbb{P}[\rho(1) \in \cup_{l\in[\ell]} o_l^{(\mu)} \mid G^*, G \in \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}] \\ &\geq 1 - \epsilon, && \text{(conditional coverage)} \\ \Rightarrow \mathbb{P}[\rho(1) \in B(G^*)] &\geq 1 - \epsilon. && \text{(marginal coverage)}\end{aligned}$$

We could make $B(\cdot)$ smaller by randomly selecting between low-density orbits, and smaller still by randomly pruning those orbits. Personally I think this is over-obsessing about the target $1 - \epsilon$, and it would be better in practice to list the critical orbits with probability p_ℓ, reporting the proportion of nodes that *could* be removed from each.

Why hasn't this been done before?. I understand that for some probabilists this is a fairly standard way of doing things, and it is also worth noting that on simpler problems statisticians often implicitly do this too. We might, for example, observe a word sequence (or time series) $X = (X_1, \ldots, X_n)$ in disorder, $X^* = \rho X$. Here, the set of relabellings of X^* is sometimes called a 'bag', and here the 'orbits' of X^* are the sets of indices corresponding to unique values (e.g. words). Given a distribution for X, the reasoning above would give us a confidence set for the first word of the sequence. It's just that deploying a full-blown group-theoretic argument here is a bit heavy-handed.

Conflict of interest: None declared.

References

Athreya A., Fishkind D. E., Tang M., Priebe C. E., Park Y., Vogelstein J. T., Levin K., Lyzinski V., & Qin Y. (2017). Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research*, *18*(1), 8393–8484.

Athreya A., Tang M., Park Y., & Priebe C. E. (2021). On estimation and inference in latent structure random graphs. *Statistical Science*, *36*(1), 68–88. https://doi.org/10.1214/20-STS787

Hoff P. D., Raftery A. E., & Handcock M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Rubin-Delanchy P. (2020). Manifold structure in graph embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. *33*, pp. 11687–11699). Curran Associates, Inc.

Rubin-Delanchy P., Cape J., Tang M., & Priebe C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(4), 1446–1473. https://doi.org/10.1111/rssb.12509

Seshadhri C., Sharma A., Stolman A., & Goel A. (2020). The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences*, *117*(11), 5631–5637. https://doi.org/10.1073/pnas.1911030117

The vote of thanks was passed by acclamation.

https://doi.org/10.1093/jrsssb/qkae053
Advance access publication 27 June 2024

Andrej Srakar's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Andrej Srakar

Institute for Economic Research, Ljubljana, Slovenia

Address for correspondence: Andrej Srakar, Institute for Economic Research, Ljubljana, Slovenia. Email: srakara@ier.si

Paper presented by Harry Crane and Min Xu develops a rather novel area of network archaeology, which builds on findings of network science, probability theory and (in Crane and Xu's article) Bayesian statistics. My comments are directed to possible extensions for future work in this promising area of research.

Firstly, in terms of network stochastic processes (authors use a combination of preferential attachment and Erdős–Rényi models) paper does little to explain comparisons with possible alternatives. Briend et al. (2023) discuss network archaeology in a random recursive dags and Cooper–Frieze random networks contexts, and Brandenberger et al. (2022) in the more general context of Bienaymé–Galton–Watson trees. I wonder if topology of the studied network could be incorporated in more detail, for example in today often studied weak-topology context (for example in Gromov-weak or just general Skorokhod topologies). This would need different metric spaces and distance metrics where results for different types of random graphs and trees are wide and could be useful for the development of the area as well for its extensions of studying the asymptotic behaviour and estimation. Combination with graphon perspectives which authors mention in the introduction would also be interesting to explore, as well as addressing the

possibilities noted by Brandenberger et al. of extensions to k-ary, Cayley, Motzkin and planted plane trees as well as branching stochastic processes in general.

Secondly, it seems to go unexplained why the authors are using Bayesian approach. The Bayesian component in the article seems underexplored and many possibilities could be useful for future developments of the used priors as well as Gibbs sampling procedure (authors themselves mention mixing properties of the sampler). Extensions of the Bayesian part could go in the direction of intractable likelihood perspectives, computational improvements (say, using integrated nested Laplace approximation) or in the selection of priors (parametric, semiparametric, or nonparametric perspectives or even to empirical Bayes possibilities). Study of the Bayesian asymptotic properties here would be very interesting.

Finally, I miss more explicit and broad connection to the study of temporal networks and network in general. Not only root vertex and communities could be discovered, but connection to cliques, islands, and homophily as well as many possible types of networks based on characteristics of their ties, such as affiliation, weighted, multi-relational, or multi-layer networks. Possibilities for future research in this area of network science could loom large.

Conflicts of interest: none declared.

References

Brandenberger A. M., Devroye L., & Goh M. K. (2022). Root estimation in Galton– Watson trees. *Random Structures & Algorithms*, *61*(3), 520–542.

Briend S., Cavillo F., & Lugosi G. (2023). Archaeology of random recursive dags and Cooper-Frieze random networks. *Combinatorics, Probability and Computing*, *32*(6), 859–873.

The following contributions were received in writing after the meeting:

https://doi.org/10.1093/jrsssb/qkae044
Advance access publication 7 June 2024

Filippo Ascolani, Antonio Lijoi and Igor Prünster's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Filippo Ascolani[1], Antonio Lijoi[2] and Igor Prünster[2]

[1]Department of Statistical Science, Duke University, Durham, NC, USA
[2]Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy

Address for correspondence: Igor Prünster, Bocconi Institute for Data Science and Analytics, Bocconi University, via Röntgen 1, 20136 Milan, Italy. Email: igor@unibocconi.it

We congratulate the authors for their timely and insightful contribution, which introduces a novel Bayesian approach to inferring the latent structure (early history and community detection) given a current observation of a network with n nodes. The Bayesian component of the model is summarized through a uniform prior on the random relabelling associated to the observed network. The methodological innovation is remarkably complemented by strong theoretical (frequentist) guarantees concerning uncertainty quantification.

The sequential formulation of the preferential attachment component of the model exhibits many interesting connections with the Bayesian nonparametric literature. Indeed, the mechanism for sampling new edges is reminiscent of Pólya urns: the colours of the balls in the urn are the nodes labelled by the community they belong to. Moreover, in the random K roots model, the authors consider the case where K increases with n and the probability of creating a new root is the same of sampling a new value from a Dirichlet process (Ferguson, 1973) with concentration parameter $\alpha_0/(2\beta + \alpha)$; each tree corresponds to a different cluster of nodes.

Within the Dirichlet process framework, the probability of generating a new tree depends only on the number of nodes in the forest, whereas the number of existing trees does not have any impact. This generative scheme might be restrictive in many settings and it is often desirable for the probability of creating a new tree to explicitly depend on the number of existing trees. In Bayesian nonparametrics, the latter requirement corresponds to prediction rules arising from the class of Gibbs-type priors (De Blasi et al., 2015): the most popular instances are the Pitman–Yor (Pitman & Yor, 1997) and the normalized generalized gamma (Lijoi et al., 2007) processes, which both generalize the Dirichlet process and lead to an asymptotic growth of the number of trees of order n^{σ}, with $\sigma \in (0, 1)$. What would the impact of different predictive structures on the model properties be? Even more flexible behaviours can be obtained through hierarchical compositions (Camerlenghi et al., 2019, 2018), which do not lead to Gibbs-type priors. For instance, Dirichlet process hierarchies lead to iterated logarithmic behaviours, with the number of iterations equal to the number of hierarchies. This growth rate, which can be made as slow as desired but still leads to an infinite number of trees, might have noteworthy implications.

Finally, the paper addresses the case of a single network. However, in many situations one may face distinct networks with no common nodes but likely similar features: for example different academic fields may share similar co-authorship structures. Therefore being able to borrow information across different networks would often be beneficial leveraging some suitable form of probabilistic symmetry. There have been some recent proposals of partially exchangeable models for stochastic block models (see, e.g. Durante et al., 2023) that allow for prediction of the clustering of future nodes through a probabilistically coeherent sequential procedure in a spirit similar to the one proposed by the authors. However, to the best of our knowledge, nothing of the sort is available for multiple Markovian preferential attachment structures. This would be an interesting direction to explore.

Conflicts of interest: None declared.

References

Camerlenghi F., Lijoi A., Orbanz P., & Prünster I. (2019). Distribution theory for hierarchical processes. *Annals of Statistics*, *47*(1), 67–92. https://doi.org/10.1214/17-AOS1678

Camerlenghi, F., Lijoi, A., & Prünster, I. (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, *45*(4), 1062–1091. https://doi.org/10.1111/sjos.v45.4

De Blasi, P., Favaro, S., Lijoi, A. M., Mena, R.H., Prünster, I., & Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 212–229. https://doi.org/10.1109/TPAMI.2013.217

Durante D., Gaffi F., Lijoi A., & Prünster I. (2023). Partially exchangeable stochastic block models for multilayer networks. In Under revision.

Ferguson T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230. https://doi.org/10.1214/aos/1176342360

Lijoi A., Mena R. H., & Prünster I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society Series B*, *69*(4), 715–740. https://doi.org/10.1111/j.1467-9868.2007.00609.x

Pitman J., & Yor M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, *25*(2), 855–900. https://doi.org/10.1214/aop/1024404422

https://doi.org/10.1093/jrsssb/qkae045
Advance access publication 3 June 2024

Sayan Banerjee's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Sayan Banerjee

Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, USA

Address for correspondence: Sayan Banerjee, Department of Statistics and Operations Research, University of North Carolina, 353 Hanes Hall CB #3260, Chapel Hill, NC 27599, USA. Email: sayan@email.unc.edu

I congratulate the authors on their elegant and computationally tractable approach to inference problems in network archaeology. This is one of the few papers which rigorously analyses inference for noisy temporal networks. The key idea is to compute the distribution of the possible arrival orders of vertices given an unlabelled network realization (online detection) and construct the confidence set for the root by sorting vertices according to their probability of arriving first given the observed unlabelled network. A Gibbs-type algorithm is provided to compute the above distribution in $O(m + n \log n)$ time (m, n are, respectively, the number of edges and vertices). Finally, some theoretical guarantees are provided for the size of the confidence set provided the noise is not too large. I have the following comments.

(i) **Community structure for dynamic networks:** The community structure conceived in the paper by looking at a K-forest and then adding noise is rather restrictive as the individual trees in the forest are 'decoupled': they can be generated by running independent continuous-time branching processes from K ancestors till the total population size hits n. A more realistic model should comprise individual vertices exerting community-specific influences, in addition to their degrees, in the growth of the network. One such model has been recently explored in Antunes et al. (2023) under the name of *attribute network models*. In addition to obtaining local weak limits and other asymptotics, some network sampling and ranking algorithms have been explored in these papers. It would be interesting to extend the current approach to community detection problems for these networks.

(ii) **Comparison with other root finding algorithms:** The authors mention several other works where root detection algorithms are presented. They are said to be more conservative as the guarantees are asymptotic as the network size grows, and there are some non-explicit constants involved in the associated bounds. However, no direct comparison is provided for these algorithms on the PAPER model for fixed large n. It would be instructive to compare these methods computationally. To my knowledge, some of the asymptotic guarantees for these other algorithms can be made quantitative and it would be interesting to compare them to the theoretical results in this paper.

(iii) **Non-local centrality measures:** In Theorem 12 of the paper, it is shown, by 'upper bounding' the confidence set with that obtained via a degree-based criterion, that the size of the confidence set is $O(n^{\gamma})$ for some $\gamma \leq 0.8$ if the noise level $\theta = o(\log n/n)$. Moreover, a conjecture is made that the actual size of the confidence set should be $O(1)$ in this case. I believe that to prove this conjecture, one needs to analyse 'non-local' centrality measures that look beyond one-step neighbourhoods of vertices (like Jordan centrality Bubeck et al., 2017). This is because, as shown in Banerjee and Bhamidi (2021), the degree centrality lacks *persistence*, that is, the identities of the highest degree vertices keep changing infinitely often as the network grows, unlike the APA model (see also Dereich & Mörters, 2009). However, Jordan centrality exhibits persistence (Banerjee & Bhamidi, 2022). This could be interesting future work.

Conflict of interest: None declared.

References

Antunes N., Banerjee S., Bhamidi S., & Pipiras V., *Attribute network models, stochastic approximation, and network sampling and ranking algorithms*. arXiv 2304.08565. https://doi.org/10.48550/arXiv.2304.08565, 2023, preprint: not peer reviewed.

Banerjee S., & Bhamidi S. (2021). Persistence of hubs in growing random networks. *Probability Theory and Related Fields*, *180*(3–4), 891–953. https://doi.org//10.1007/s00440-021-01066-0
Banerjee S., & Bhamidi S. (2022). Root finding algorithms and persistence of Jordan centrality in growing random trees. *The Annals of Applied Probability*, *32*(3), 2180–2210. https://doi.org//10.1214/21-AAP1731
Bubeck S., Devroye L., & Lugosi G., (2017). Finding Adam in random growing trees. *Random Structures & Algorithms*, *50*(2), 158–172. https://doi.org/10.1002/rsa.v50.2
Dereich S., & Mörters P. (2009). Random networks with sublinear preferential attachment: Degree evolutions. *Electronic Journal of Probability*, *14*, 1222–1267. https://doi.org//10.1214/EJP.v14-647

https://doi.org/10.1093/jrsssb/qkae047
Advance access publication 3 June 2024

Marta Catalano, Augusto Fasano, Matteo Giordano, and Giovanni Rebaudo's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Marta Catalano[1], Augusto Fasano[2], Matteo Giordano[3] and Giovanni Rebaudo[3]

[1]Department of Economics and Finance, Luiss University, Roma, Italy
[2]Catholic University, Milano, Italy
[3]University of Torino, Torino, Italy

Address for correspondence: Marta Catalano, Department of Economics and Finance, Luiss University, Viale Romania 32, 00197, Roma, Italy. Email: mcatalano@luiss.it

We congratulate the authors for their methodological and theoretical contribution to the statistical literature on networks.

A natural extension of the proposed PAPER model is included, with K communities growing simultaneously and where new nodes are either assigned to an existing community or elected as a new root. The employed assignment rule is of Pólya-urn type, which leads to a logarithmic growth of the number of communities (Korwar & Hollander, 1973) and is known to coincide with the predictive scheme of exchangeable sequences associated with the Dirichlet process. The probability of creating a new community is then independent of the number of past ones, which is a distinctive feature of the Dirichlet process within the class of Gibbs-type priors (De Blasi et al., 2015). An interesting direction would be to allow for more flexible predictive schemes that ensure alternative asymptotics, ranging from power-law behaviours (via the Pitman-Yor process Pitman, 2006) or normalized generalized gamma completely random measures (Lijoi et al., 2007)) to slower than logarithmic growth (via the single-group hierarchical Dirichlet process (Camerlenghi et al., 2018).

An important theoretical aspect is the number of communities. While the authors provide an empirical investigation, future research could tackle the question of posterior consistency for the number of communities, along the lines of the existing results for stochastic block models (Geng et al., 2019) and in the growing literature in Bayesian nonparametric mixture models (Ascolani et al., 2023; Miller & Harrison, 2013; Nobile, 1994).

Turning to applications, the proposed model lends itself to some natural generalizations suggested by popular epidemiological models, like SIR dynamics, where at each instant the infectious

nodes can transmit the disease to their susceptible neighbours with some probability, resulting in multiple new infectious individuals at the next time. Equating new infections in SIR dynamics to added nodes in the PAPER model, a useful extension would be obtained by allowing the addition of multiple nodes at each step: for instance, if computationally feasible, a fraction of the existing nodes at that time, representing an average contact rate, or also a random number, e.g. driven by a nonhomogeneous Poisson process. The SIR analogy further suggests extensions where nodes are active (i.e. accepting newly introduced nodes as neighbours) only for a limited time, representing the period during which an infectious agent can spread the disease.

Once more, we commend the authors for an outstanding paper.

Conflict of interest: None declared.

References

Ascolani F., Lijoi A., Rebaudo G., & Zanella G. (2023). Clustering consistency with Dirichlet process mixtures. *Biometrika*, *110*(2), 551–558. https://doi.org/10.1093/biomet/asac051

Camerlenghi F., Lijoi A., & Prünster I. (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, *45*(4), 1062–1091. https://doi.org/10.1111/sjos.12334

De Blasi P., Favaro S., Lijoi A., Mena R. H., Prünster I., & Ruggiero M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 212–229. https://doi.org/10.1109/tpami.2013.217

Geng J., Bhattacharya A., & Pati D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association*, *114*(526), 893–905. https://doi.org/10.1080/01621459.2018.1458618

Korwar R. M., & Hollander M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability*, *1*(4), 705–711. https://doi.org/10.1214/aop/1176996898

Lijoi A., Mena R. H., & Prünster I. (2007). Bayesian nonparametric estimation of the probability of discovering a new species. *Biometrika*, *94*(4),769–786. https://doi.org/10.1093/biomet/asm061

Miller J. W., & Harrison M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *Advances in Neural Information Processing Systems*, *26*, 199–206. https://proceedings.neurips.cc/paper_files/paper/2013/file/f7e6c85504ce6e82442c770f7c8606f0-Paper.pdf

Nobile A. (1994). Bayesian Analysis of Finite Mixture Distributions. Ph.D. thesis, Carnegie Mellon Univ.

Pitman J. (2006). *Combinatorial stochastic processes*. Lecture Notes in Mathematics, 1875. Springer. https://doi.org/10.1007/b11601500

https://doi.org/10.1093/jrsssb/qkae051
Advance access publication 5 June 2024

Yang Feng and Jiajin Sun's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Yang Feng[1] and Jiajin Sun[2]

[1]Department of Biostatistics, School of Global Public Health, New York University, 708 Broadway, New York, NY 10003, USA
[2]Department of Statistics, Columbia University, New York, NY 10027, USA

Address for correspondence: Yang Feng, Department of Biostatistics, School of Global Public Health, New York University, 708 Broadway, New York, NY 10003, USA. Email: yang.feng@nyu.edu

We congratulate Drs. Crane and Xu for their fascinating piece of work. In the multiple roots PAPER model, one assumption is that different trees share the same affine preferential attachment (APA) growth parameters (α, β). In practice, the multiple trees in a network may have varying growth mechanisms, and it may be of interest to model each tree to have its own growth parameters. We consider such a heterogeneous affine preferential attachment (HAPA) model.

Definition 1 For a random forest of K heterogeneous disjoint component trees, denote its growth parameters by $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv (\alpha_k, \beta_k)_{k=1}^K$. We define its growth process by the HAPA$(\boldsymbol{\alpha}, \boldsymbol{\beta}, K)$ model: For $k \in S = [K] \equiv \{1, \ldots, K\}$, let node k be the root of the kth component tree. For any $K \le t \le n$, $k \in [K]$, denote the kth component tree at time t by $\mathbf{T}_{k,t}$, and the time labelled forest at time t by $\mathbf{F}_t = \cup_{k=1}^K \mathbf{T}_{k,t}$. At $t = K$, $\mathbf{T}_{k,t}$ is the set of the singleton node k. For $t = K+1, \ldots, n$, given $\mathbf{F}_{t-1}$, we add a new node t and a new random edge (t, w_t) where the existing node $w_t \in \mathbf{T}_{k,t-1}$ is chosen with probability

$$\frac{\beta_k D_{\mathbf{F}_{t-1}}(w_t) + 2\beta_k \mathbb{1}_{\{w_t \in S\}} + \alpha_k}{\sum_{k=1}^K (2\beta_k + \alpha_k) X_{k,t-1}}$$

where $D_{\mathbf{F}_{t-1}}(w_t)$ is the degree of w_t in $\mathbf{F}_{t-1}$, and $X_{k,t} = |\mathbf{T}_{k,t}|$ is the number of nodes in tree $\mathbf{T}_{k,t}$.We then say that a random graph $\mathbf{G}_n \sim$ HPAPER$(\boldsymbol{\alpha}, \boldsymbol{\beta}, K, \theta)$ if $\mathbf{G}_n = \mathbf{F}_n + \mathbf{R}_n$ where $\mathbf{F}_n \sim$ HAPA$(\boldsymbol{\alpha}, \boldsymbol{\beta}, K)$ and $\mathbf{R}_n \sim$ ER(θ) is an Erdős–Rényi random graph independent of $\mathbf{F}_n$ defined on the same set of nodes $[n]$.

Definition 1 resembles the K roots model in Crane Xu (2023), with two major differences: (i) instead of homogeneous (α, β) parameters, the new model has K pairs of growth parameters, one for each tree and (ii) the likelihood of the HAPA forest

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{f}_n) = \frac{\prod_{k=1}^K \prod_{v \in \mathbf{t}_k} \prod_{j=1}^{D_{\mathbf{t}_n(v)}-1} \{\beta_k \cdot j + \alpha_k + 2\beta_k \mathbb{1}_{\{v \in S\}}\}}{\prod_{t=K+1}^n \left\{\sum_{k=1}^K (2\beta_k + \alpha_k) x_{k,t-1}\right\}} \tag{1}$$

depends on not only the degree sequence but also the tree growth history.

Algorithm 1 EM algorithm of estimation of $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\alpha_k, \beta_k)_{k=1}^K$ under the HPAPER model

Input : Graph $\tilde{\mathbf{g}}_n$; number of component trees K

Output: Parameter estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = (\hat{\alpha}_k, \hat{\beta}_k)_{k=1}^K$

1 Initialization: estimate $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ from PAPER model, sample one forest $\tilde{\mathbf{f}}^{(0)}$ from PAPER$(\hat{\alpha}, \hat{\beta}, K, \hat{\theta})$, and initialize $(\hat{\alpha}_k, \hat{\beta}_k) = (\hat{\alpha}, \hat{\beta})$, $\forall k \in [K]$.

2 Generate M Monte-Carlo samples of forest and ordering $(\tilde{\mathbf{f}}^{(m)}, \pi^{(m)})_{m=1}^M$ from the HPAPER model $\mathbb{P}(\tilde{\mathbf{f}}, \pi | \tilde{\mathbf{g}}_n, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, K)$ with Algorithm 2 described below.

3 Update the parameter estimates by

$$(\hat{\alpha}_k, \hat{\beta}_k)_{k=1}^K = \underset{(\boldsymbol{\alpha}, \boldsymbol{\beta})}{arg\,max}\, M^{-1} \sum_{m=1}^M l(\boldsymbol{\alpha}, \boldsymbol{\beta}; \tilde{\mathbf{f}}^{(m)}, \pi^{(m)}),$$

where $l(\boldsymbol{\alpha}, \boldsymbol{\beta}; \tilde{\mathbf{f}}, \pi)$ is the logarithm of the likelihood in (1).

4 Let $\tilde{\mathbf{f}}^{(0)} = \tilde{\mathbf{f}}^{(M)}$. Repeat steps 2 and 3 until the parameter estimates have converged.

Algorithm 2 Gibbs sampling of ordering and forests from the HPAPER model

Input : Graph $\tilde{\mathbf{g}}_n$; an initial forest $\tilde{\mathbf{f}}^{(0)}$ with K component trees; parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$

Output: Monte Carlo samples of forest and ordering $(\tilde{\mathbf{f}}^{(m)}, \pi^{(m)})_{m=1}^{M}$

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

Sample ordering $\pi^{(m)}$ given forest $\tilde{\mathbf{f}}^{(m-1)}$:

Sample the tree size sequence: Let $x_{k,t} = 1$ for $t = K$ and $k \in [K]$. For t from $K+1$ to n, given $\{x_{1,t-1}, \ldots, x_{K,t-1}\}$, choose a $k \in [K]$ with probability $\frac{(2\beta_k+\alpha_k)x_{k,t-1}}{\sum_{k=1}^{K}(2\beta_k+\alpha_k)x_{k,t-1}}$. Let $x_{k,t} = x_{k,t-1} + 1$, and $x_{j,t} = x_{j,t-1}$ for all $j \neq k$.

For each $k \in [K]$, sample the ordering inside the kth component tree $\tilde{\mathbf{t}}_k^{(m-1)}$ with steps 5-8 of Algorithm 1 in Crane and Xu (2023).

Combine the ordering of all component trees: Denote by $\pi_t^{(m)}$ the tth node in ordering $\pi^{(m)}$. For $t = k \leq K$, let $\pi_t^{(m)}$ be the root of $\tilde{\mathbf{t}}_k^{(m-1)}$. For $t > K$, if $x_{k,t} = x_{k,t-1} + 1$, let $\pi_t^{(m)}$ be the first node from the ordering inside $\tilde{\mathbf{t}}_k^{(m-1)}$ that has not appeared in $\pi_{1:(t-1)}^{(m)}$.

Sample forest $\tilde{\mathbf{f}}^{(m)}$ given ordering $\pi^{(m)}$:

Let $\tilde{\mathbf{f}}_n = \tilde{\mathbf{f}}^{(m-1)}$, $\pi = \pi^{(m)}$. For t from $K+1$ to n, update the parent of π_t in $\tilde{\mathbf{f}}_n$ to be $w \in \pi_{1:(t-1)} \cap N_{\tilde{\mathbf{g}}_n}(\pi_t) \cap \tilde{\mathbf{t}}_k$ with probability proportional to

$$\mathbb{P}\left(pa_{\tilde{\mathbf{f}}_n}(\pi_t) = w \middle| \pi, \tilde{\mathbf{g}}_n, \{pa_{\tilde{\mathbf{f}}_n}(v)\}_{v \neq \pi_t} \right) \propto \frac{\beta_k D_{\tilde{\mathbf{f}}_n^{(\cdot,\pi_t)}}(w) + 2\beta_k \mathbb{1}_{\{w \in \pi_{1:K}\}} + \alpha_k}{\prod_{\tau=t}^{n} \{ \sum_{k=1}^{K} (2\beta_k + \alpha_k) x_{k,\tau-1}(\tilde{\mathbf{f}}_n^{(w,\pi_t)}) \}}$$

where $N_{\tilde{\mathbf{g}}_n}(\pi_t)$, $\tilde{\mathbf{f}}_n^{(w,\pi_t)}$, $\tilde{\mathbf{f}}_n^{(\cdot,\pi_t)}$, $D_{\tilde{\mathbf{f}}_n^{(\cdot,\pi_t)}}(w)$ are defined the same as in Crane Xu (2023), and $x_{k,\tau}(\tilde{\mathbf{f}}_n^{(w,\pi_t)})$ is the number of nodes in the kth component tree of $\tilde{\mathbf{f}}_n^{(w,\pi_t)}$ at time τ. Then output $\tilde{\mathbf{f}}^{(m)} = \tilde{\mathbf{f}}_n$.

To conduct inference for the model in Definition 1, we need to address two primary tasks: estimating the $(\alpha_k, \beta_k)_{k=1}^{K}$ parameters, and sampling from the distribution of time labelled forests. Both tasks bring new challenges compared with the K roots model in Crane and Xu (2023). First, the growth parameters can be estimated in the PAPER model without knowing the class assignments of the nodes, while in the HPAPER model, one must have some knowledge or estimation of the clustering as they are necessary even for defining the growth parameters. Second, some of the smart sampling techniques for the APA model are built upon the property that the forest likelihood depends only on the degree sequences of its nodes, which is no longer the case for the HAPA likelihood (1). Following the notation in Crane and Xu (2023), we write $\tilde{\mathbf{f}}_n = \pi \mathbf{f}_n$ where $\tilde{\mathbf{f}}_n$ denotes the randomly alphabetically labelled forest and π denotes the ordering of the nodes. We describe an Expectation–Maximization (EM) algorithm framework to estimate the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and a Gibbs sampling framework to sample forest $\tilde{\mathbf{f}}_n$ and ordering π in Algorithms 1 and 2, respectively.

We plan to give a full treatment to the proposed models and algorithms in a follow-up work. In the future, it is also of interest to establish an 'HPAPER-SBM' model, in which the Erdős–Rényi parameter is allowed to be different within and between different communities, and further explore suitable sampling and estimation procedures under that model. Another interesting direction is to incorporate nodal or edge covariates to the model (Huang et al., 2023; Weng & Feng, 2022).

Conflict of interest: None declared.

References

Crane H., & Xu M. (2023). Root and community inference on the latent growth process of a network using noisy attachment models. *Journal of the Royal Statistical Society, Series B*, in press.

Huang S., Sun J., & Feng Y. (2023). PCABM: Pairwise covariates-adjusted block model for community detection. *Journal of the American Statistical Association*, 1–13. in press. https://doi.org/10.1080/01621459.2023.2244731

Weng H., & Feng Y. (2022). Community detection with nodal information: Likelihood and its variational approximation. *Stat*, *11*(1), e428. https://doi.org/10.1002/sta4.v11.1

https://doi.org/10.1093/jrsssb/qkae055
Advance access publication 12 June 2024

Yicong Jiang and Zheng Tracy Ke's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Yicong Jiang and Zheng Tracy Ke

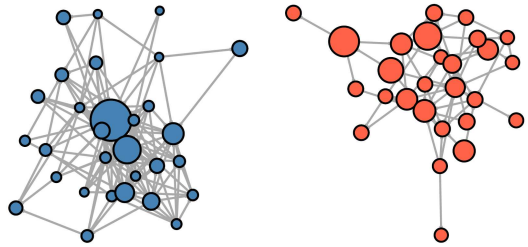
Department of Statistics, Harvard University, Cambridge, USA

Address for correspondence: Zheng Tracy Ke, Department of Statistics, Harvard University, Cambridge, USA.
Email: zke@fas.harvard.edu

We congratulate the authors on an excellent paper! Crane and Xu (2021) proposed novel methods for finding 'root nodes' from a single snapshot of a dynamic network process, with several interesting real-data examples. We now consider a new application for finding 'root papers' in a citation network. The MADStat dataset (Ji et al., 2022; Ke et al., 2023) consists of the bibtex and citation information of over 83 K papers, which we use to construct paper citation networks. Given a keyword (e.g. 'Lasso'), let V_0 be the set of papers whose titles contain this keyword, and let V be the set of papers that are either citers or citees of papers in V_0 (we only count the citations recorded in MADStat). We then build a symmetric network on V, with an edge between two papers i and j if either i cites j or j cites i; if the network is disconnected, we restrict it to its giant component. The networks for two keywords, Lasso and Bayesian, are shown in Figure 1. We apply the method in Crane and Xu (2021) to each network to obtain the posterior probability of each node being a root node. The top 6 papers with the highest posterior root probability are in Table 1. In the Lasso network, Tibshirani (1996) is ranked top 1. In the Bayesian network, Gelfand and Smith (1990) is ranked top 1. The results are meaningful and motivate a new application of the proposed method.

We also suggest some extensions of Crane and Xu (2021). First, the PAPER model is built on the Erdos–Renyi model and does not model degree heterogeneity among nodes. The Erdos–Renyi model can be generalized to accommodate degree heterogeneity [such as a DCBM model with $K = 1$; see Jin et al. (2022)]. It will be interesting to see if the PAPER model can be generalized similarly. Second, in the case of multiple roots, we may run community detection first and then apply the algorithm to each community separately. There are fast community detection algorithms [e.g. Jin et al. (2022); Jiang and Ke (2023)] equipped with data-driven choices of the number of communities (Jin et al., 2023). Combining them with the current algorithm will help reduce computational costs and avoid randomness caused by forest partition. We hope these ideas are beneficial. Congratulations to the authors again on their remarkable work!

Conflict of interest: None declared.



	Lasso	Bayesian
#Nodes	2308	18502
#Edges	7137	54794
$d_{\min}$	1	1
$d_{\max}$	1335	718
$\bar{d}$	6.18	5.92

Figure 1. The Lasso network (left graph) and the Bayesian network (right graph); only the 30 highest-degree nodes are shown. The table on the right provides the summary statistics, where $d_{\max}$, $d_{\min}$, and $\bar{d}$ are the maximum, minimum, and average degrees, respectively.

Table 1. The top 6 papers with the highest posterior root probability in the Lasso network (top) and the Bayesian network (bottom), respectively

Title	Author(s) & Year	Journal	#Citation	Root Prob.
Regression Shrinkage and Selection via the Lasso	Tibshirani (1996)	JRSSB	55448	0.50
High-dimensional Graphs and Variable Selection with the Lasso	Meinshausen and Bühlmann (2006)	AoS	4328	0.05
The Adaptive Lasso and its Oracle Properties	Zou (2006)	JASA	8245	0.03
Simultaneous Analysis of Lasso and Dantzig Selector	Bickel et al. (2009)	AoS	2800	0.01
The Bayesian Lasso	Park and Casella (2008)	JASA	3453	0.01
Sparsity and Smoothness via the Fused Lasso	Tibshirani et al. (2005)	JRSSB	3212	0.01
Sampling-based Approaches to Calculating Marginal Densities	Gelfand and Smith (1990)	JASA	9818	0.13
Bayesian Statistics in Medicine: A 25 Year Review	Ashby (2006)	SMed	295	0.11
Bayesian Computation via the Gibbs Sampler	Smith and Roberts (1993)	JRSSB	2536	0.08
And Related Markov-chain Monte–Carlo Methods				
Bayesian Experimental Design: A Review	Chaloner and Verdinelli (1995)	StSci	2354	0.06
Bayesian Computation and Stochastic-systems	Besag et al. (1995)	StSci	1548	0.05
Bayesian Measures of Model Complexity and Fit	Spiegelhalter et al. (2002)	JRSSB	14395	0.05

References

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine*, *25*(21), 3589–3631. https://doi.org/10.1002/sim.2672

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, *10*(1), 3–41. https://doi.org/10.1214/ss/1177010123

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals Statistics*, *37*(4), 1705–1732. https://doi.org/10.1214/08-AOS620

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304. https://doi.org/10.1214/ss/1177009939

Crane, H., & Xu, M. (2021). Root and community inference on the latent growth process of a network using noisy attachment models, arXiv, https://doi.org/10.48550/arXiv.2107.00153, July 1, 2021, preprint: not peer reviewed.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.1080/01621459.1990.10476213

Ji, P., Jin, J., Ke, Z. T., & Li, W. (2022). Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics*, *40*(2), 469–485. https://doi.org/10.1080/07350015.2021.1978469

Jiang, Y., & Ke, T. (2023). Semi-supervised Community Detection via Structural Similarity Metrics. In The Eleventh International Conference on Learning Representations.
Jin, J., Ke, Z. T., & Luo, S. (2022). Improvements on SCORE, especially for weak signals. *Sankhya A*, *84*, 127–162. https://doi.org/10.1007/s13171-020-00240-1
Jin, J., Ke, Z. T., Luo, S., & Wang, M. (2023). Optimal estimation of the number of network communities. *Journal of the American Statistical Association*, *118*(543), 2101–2116. https://doi.org/10.1080/01621459.2022.2035736
Ke, Z. T., Ji, P., Jin, J., & Li, W. (2023). Recent advances in text analysis. *Annual Review in Statistics and Its Applications*, *11*(1), 347–372. https://doi.org/10.1146/annurev-statistics-040522-022138
Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals Statistics*, *34*(3), 1436–1462. https://doi.org/10.1214/009053606000000281
Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337
Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(1), 3–23. https://doi.org/10.1111/j.2517-6161.1993.tb01466.x
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(1), 91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

https://doi.org/10.1093/jrsssb/qkae048
Advance access publication 3 June 2024

Tianxi Li's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Tianxi Li

School of Statistics, University of Minnesota, Minneapolis, Minnesota, USA
Address for correspondence: Tianxi Li, Ford Hall 376, 224 Church St SE, Minneapolis, MN 55455, USA. Email: tianxili@umn.edu

I want to congratulate Prof. Crane and Prof. Xu for their impressive work. The paper introduces two novel components: the PAPER model and the inference framework for the tree. Both of them bring many insights into handling complex network data and open the doors to new problems. I will discuss extension problems from both aspects.

The PAPER model assumes preferential attachment trees with additional the Erdős-Rényi (ER) edges. One nice property of the model lies in its ability to generate pendants that can often be observed in many real-world networks. The model can be interpreted as a signal (tree) + noise (ER) structure for edges. The recent core-periphery models of Elliott et al. (2020) and Miao and Li (2023) can be seen as examples of such a structure at the node level. It would be interesting to see what could be achieved by combining the two signal+noise structures. Meanwhile, the topological properties of PAPER are also worth a thorough study. For example, as both tree structures

and the ER model lack *transitivity*, we may conjecture that the PAPER also has the same limitation. How to generalize the model to incorporate such additional properties would be an important direction for future research. One potential obstacle to such generalization might be the root inference computation. The paper's inference nicely hinges on the uniform noise edges in the ER mechanism, and it is unclear if the computation can be efficiently done without it.

A related setting for root inference is the *diffusion scenario*, which was initially studied in Shah and Zaman (2011) with many follow-up studies (Dawkins et al., 2021; Kazemitabar & Amini, 2020; Khim & Loh, 2016): a fixed network (not necessarily a tree) is given and a random diffusion process is initiated within the network. This diffusion scenario differs from the tree-growing setting, but when the network in the former is a tree and the ER part is removed from PAPER (Crane & Xu, 2021), the two settings coincide. The setup of the current paper is not directly compatible with the diffusion setting. Still, the available optimality in the tree-growing scenario raises a natural question about whether we can build an optimal inference procedure in the diffusion setting. Defining a similar inference method seems plausible, but the computation is again a bottleneck. Explorations in this direction may help with broader applications in epidemiology and cyber security.

Conflict of interest: None declared.

References

Crane H., & Xu M. (2021). Inference on the history of a randomly growing tree. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(4), 639–668. https://doi.org/10.1111/rssb.12428

Dawkins Q. E., Li T., & Xu H. (2021). Diffusion source identification on networks with statistical confidence. In *International Conference on Machine Learning* (pp. 2500–2509). PMLR.

Elliott A., Chiu A., Bazzi M., Reinert G., & Cucuringu M. (2020). Core-periphery structure in directed networks. *Proceedings of the Royal Society A*, *476*(2241), Article 20190783. https://doi.org/10.1098/rspa.2019.0783

Kazemitabar S. J., & Amini A. A. (2020). Approximate identification of the optimal epidemic source in complex networks. In *Proceedings of NetSci-X 2020: Sixth International Winter School and Conference on Network Science 6* (pp. 107–125). Springer.

Khim J., & Loh P.-L. (2016). Confidence sets for the source of a diffusion in regular trees. *IEEE Transactions on Network Science and Engineering*, *4*(1), 27–40. https://doi.org/10.1109/TNSE.2016.2627502

Miao R., & Li T. (2023). Informative core identification in complex networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(1), 108–126. https://doi.org/10.1093/jrsssb/qkac009

Shah D., & Zaman T. (2011). Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, *57*(8), 5163–5181. https://doi.org/10.1109/TIT.2011.2158885

https://doi.org/10.1093/jrsssb/qkae046
Advance access publication 3 June 2024

Qing Yang and Xin Tong's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Qing Yang[1] and Xin Tong[2]

[1]International Institute of Finance, School of Management, University of Science and Technology of China, Heifei, Anhui, China
[2]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA

Address for correspondence: Xin Tong, University of Southern California, BRI 310A, Bridge Hall, 3670 Trousdale Parkway, Los Angeles, CA 90089, USA. Email: xint@marshall.usc.edu

First, we congratulate Dr. Crane and Dr. Xu for this important contribution to statistical network analysis. To address an overlooked fact that most real networks are formed through a growth process, the authors introduce a useful PAPER model and its variations and skilfully construct computationally feasible confidence sets for the root node(s). In the data analysis, the authors also use their models as a new approach to community detection.

On this note, our interest lies in the multiple roots setting, as real social networks typically encompass more than one community. Take the well-known political blogs network (Adamic & Glance, 2005) as an example. It has $K = 2$ ground truth communities—the liberal and conservative communities. At time t, the PAPER model attaches the new node t to an existing node $w_t \in \{1, \ldots, t-1\}$ with probability

$$\frac{\beta D_{\mathbf{F}_{t-1}}(w_t) + 2\beta \mathbb{I}\{w_t \in S\} + \alpha}{(2\beta + \alpha)(t-1)}, \quad S = \{1, \ldots, K\},$$

which is somehow proportional to the degree of node w_t. In reality, in addition to the node degrees which reflect the popularity, homophily (McPherson et al., 2001) is a well-documented phenomenon that warrants attention. Individuals often exhibit a preference for connecting with others who have similar characteristics. Furthermore, due to either hostility or lack of interest, individuals in the liberal group may even decline connections with specific individuals from the other group. Consequently, an important question arises regarding how to incorporate this type of information into the PAPER model. A nonrigorous intuition in the fixed K roots setting suggests attaching node t to w_t with a probability $\rho_{w_t} \frac{\beta D_{\mathbf{F}_{t-1}}(w_t) + 2\beta \mathbb{I}\{w_t \in S\} + \alpha}{(2\beta + \alpha)(t-1)}$, where $\rho_{w_t} \in (0, 1)$ is set to be higher when nodes t and w_t share similar characteristics; conversely, reducing the magnitude. Meanwhile, a selection mechanism can be employed to determine whether to retain the current edge choice or not, driven by binary options of 'Yes' or 'No' when deciding whether to follow an individual on the Internet.

Another question pertains to inferring the true number of roots K. The authors propose methodologies to obtain the posterior root distributions, then can these probabilities be utilized to construct test statistics for testing $H_0 : K = K_0$, i.e. whether the model follows PAPER(α, β, K_0, θ) or not? A related query concerns simulation studies that demonstrate promising community recovery performance, particularly with the fixed K model; therefore is it feasible to provide theoretical guarantees?

Lastly, we would like to make two suggestions regarding further dissemination. (i) to expand the readerships of this work, the authors might consider writing an R markdown file that only includes the most essential components of the model and one simple example. (ii) We feel that the proposed models and algorithms have potentials to make some real insights in empirical studies beyond what other community detection algorithms can already claim.

Conflicts of interest: None declared.

References

Adamic L. A., & Glance N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 36–43). Association for Computing Machinery.

McPherson M., Smith-Lovin L., & Cook J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/soc.2001.27.issue-1

https://doi.org/10.1093/jrsssb/qkae050
Advance access publication 4 June 2024

Fan Wang, Yi Yu and Alessandro Rinaldo's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Fan Wang[1], Yi Yu[1] and Alessandro Rinaldo[2]

[1]Department of Statistics, University of Warwick, Coventry, UK
[2]Department of Statistics and Data Science, The University of Texas at Austin, Austin, USA

Address for correspondence: Yi Yu, Department of Statistics, University of Warwick, Coventry CV47AL, UK. Email: yi.yu.2@warwick.ac.uk

We congratulate Prof. Crane and Dr Xu for introducing a very interesting and appealing statistical model for Markovian network growth. The PAPER model enjoys several qualities. In particular, (i) it strikes a rare balance between analytic elegance and tractability, and expressive power; (ii) it allows for practicable algorithms for statistical inference that scale to networks of very large size, and (iii) it is very flexible and can be readily generalized to model a variety of phenomena and features commonly observed in modern networks. As pointed out by the authors, there are many extensions and open problems related to the model that are worth considering.

In this note, we suggest possible extensions to change point analysis (CPA) for networks. CPA is a well-studied topic concerned with modelling and detecting abrupt changes in the data-generating distribution in time series data. Developing models, theories and methods for CPA in dynamic and large networks is a relatively new and exciting area of research (e.g. Wang et al., 2021; Yu et al., 2023). We believe the PAPER model provides an excellent reference framework for building powerful and realistic Markovian network models.

To provide some details, in the PAPER model, at time point $t \in [n]$, the newly added node with label t is connected to an existing node $w_t \in [t-1]$, with probability $\{\beta D_{\mathrm{T}_{t-1}}(w_t) + \alpha\}/\{\beta 2(t-2) + \alpha(t-1)\}$. The parameters (α, β) are fixed across the whole time course. It would be interesting to consider the scenario where the parameters are instead allowed to change in a piecewise manner at unknown CPs. In the simplest instantiation of the PAPER CP model, there *might* exist an unknown CP t^* such that the values of the parameters α and β change after t^*. A possible application could be, in a social network, at the booming stage of a key opinion leader, β, the parameter characterizing the 'rich gets richer' phenomenon, is positive and large. As the craze cools down, β should decrease to reflect the fading of fame. In extreme cases, β may even change the sign. To estimate the change time t^*, one could consider a likelihood-based ℓ_0-penalization (e.g. Wang et al., 2023), with the likelihood stemming from the PAPER model. A second, more sophisticated extension is to consider an APA model with multiple root nodes; at the CP(s), the number of root nodes, along, possibly, with the model parameter, change, thus accounting for the creation, elimination or even merging of tree components. Finally, the CPA tasks just outlined can be analysed in the offline settings, in which the fully grown network at a given time, say n, is observed and then analysed.

Lastly, we congratulate Prof. Crane and Dr Xu again for their excellent paper. We anticipate that the ideas and methods of the paper will provide the impetus for further research developments of the PAPER model for years to come!

Conflict of interests: None declared.

References

Wang D., Yu Y., & Rinaldo A. (2021). Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, *49*(1), 203–232. https://doi.org/10.1214/20-AOS1953

Wang D., Yu Y., & Willett R. (2023). Detecting abrupt changes in high-dimensional self-exciting Poisson processes. *Statistica Sinica*, *33*, 1653–1671. https://doi.org/10.5705/ss.202021.0221
Yu Y., Padilla O. H. M., Wang D., & Rinaldo A. (2023). Optimal network online change point localisation. *SIAM Journal on Mathematics of Data Science, to appear*.

https://doi.org/10.1093/jrsssb/qkae054
Advance access publication 7 June 2024

Jason Wyse, James Ng, Arthur White and Michael Fop's contribution to the Discussion of 'Root and community inference on the latent growth process of a network' by Crane and Xu

Jason Wyse[1], James Ng[1], Arthur White[1] and Michael Fop[2]

[1]School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[2]School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

Address for correspondence: Jason Wyse, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland. Email: wyseja@tcd.ie

We congratulate the authors on a thought provoking paper introducing innovative ideas for the statistical modelling of networks. The emphasis on identifying highly probable root nodes is both original and intriguing. In particular, using the root node credible set methods for node clustering within the PAPER(α, β, α_0, θ) random K model (Section 6.3) is an interesting perspective. We wonder how these approaches might connect with latent space model-based node clustering methods in the literature (Handcock et al., 2007) and those quantifying uncertainty in the number of clusters (D'Angelo et al., 2023; Ryan et al., 2017). Another research avenue could explore the feasibility of a model-free approach to construct confidence sets for root nodes, such as utilizing a generalized Bayesian approach with loss functions (Bissiri et al., 2016).

Regarding parameter estimation, the authors propose an approximate EM algorithm for estimation of α in a PAPER(α, β, θ) model. The algorithm detailed in S3.1 of the appendix is useful for applications, as it appears it can be run efficiently for large n. Computing the objective function relies on two approximations for a tractable approximate E-step. The first involves breaking dependence between node degrees, the second relies on the limiting distribution approximation (van der Hofstad, 2016), facilitating reasonable computation time. It may be interesting to investigate the impact of these approximations on networks of tens or hundreds of nodes, considering the potential substitution of one or both with Gibbs sampling. Although the approximate EM algorithm provides point estimates of α, we are intrigued about the authors' insights into quantifying uncertainty in an estimate $\widehat{\alpha}$. Such measures could be valuable in evaluating evidence for different attachment behaviours in networks.

The key aspect of the PAPER model generative structure is the notion of root node, one or multiple in the case of clustering. In some settings, such as networks evolving over time, this concept is natural and of high relevance; in others, we believe the notion of root node may not align with the application context. For example, in many social science applications involving data collected through questionnaires or observational studies, even including the Zachary karate club and the co-authorship network examples in the paper, relations and communities cannot be reconciled with a single or multiple originating root nodes, but rather with homophily or popularity of

certain actors in the network (D'Angelo et al., 2023; Sengupta & Chen, 2017). We are curious about the authors' thoughts on the rationale behind using the PAPER model in these contexts, in particular in relation to the goals of root node identification and their interpretation.

Conflicts of interest: None declared.

References

Bissiri P. G., Holmes C. C., & Walker S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *78*(5), 1103–1130. https://doi.org/10.1111/rssb.12158

D'Angelo S., Alfò M., & Fop M. (2023). Model-based clustering for multidimensional social networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *186*(3), 481–507. https://doi.org/10.1093/jrsssa/qnac011

Handcock M. S., Raftery A. E., & Tantrum J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301–354. https://doi.org/10.1111/j.1467-985X.2007.00471.x

Ryan C., Wyse J., & Friel N. (2017). Bayesian model selection for the latent position cluster model for social networks. *Network Science*, *5*(1), 70–91. https://doi.org/10.1017/nws.2017.6

Sengupta S., & Chen Y. (2017). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(2), 365–386. https://doi.org/10.1111/rssb.12245

van der Hofstad R. (2016). *Random graphs and complex networks, Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

The authors replied later in writing as follows:

https://doi.org/10.1093/jrsssb/qkae049
Advance access publication 3 June 2024

Authors' reply to the Discussion of 'Root and community inference on the latent growth process of a network'

Harry Crane and Min Xu

Department of Statistics, Rutgers University, New Brunswick, NJ, USA

Address for correspondence: Min Xu, Department of Statistics, Rutgers University, 110 Frelinghuysen Rd, Piscataway, NJ 08854, USA. Email: mx76@stat.rutgers.edu

We sincerely thank all the discussants for their careful thoughts and insightful contributions. We also appreciate the diversity of topics in the discussions, ranging from applied probability and algorithms to social science and Bayesian modelling—it appropriately reflects the broad thinking necessary to make advancements in network data analysis. In this article, we respond to model related issues in Section 1 and theory related issues in Section 2. In the final section, we respond to various specific points raised by each of the discussants.

1 Model

Many discussants highlighted ways in which the Preferential Attachment Plus Erdös–Rényi (PAPER) model may be unrealistic and suggested potential extensions. We acknowledge that such suggestions

could be appropriate in the right context, but emphasize that the context dictates which features are the most salient to model. Except in stylized contexts, none of which exist in networks applications to our knowledge, no model can fully account for all observed properties of the data. And one might argue that no model should even try. Instead, the model ought to be specified to adequately explain the most salient properties of the data for the purpose of a given scientific question. Beyond that, the principle of parsimony suggests that the simplest model that can address the question of interest is often the best. This is the principle we applied in suggesting the PAPER model, as a framework on which more specific and sophisticated models can be tailored to a wide range of applications.

For example, the PAPER model may be appropriate if the scientific question is to understand node centrality in a given network. The PAPER model gives a simple way to quantify uncertainty when measuring node centrality; it enables questions like 'is one node significantly more central than another?' The multiple roots model also allows one to study 'community-specific" node centrality instead of global node centrality.

In the paper, we discussed the seq-PAPER model and the deletion noise model to illustrate how one could extend the PAPER model and what would be the corresponding modifications needed for the inference algorithm. As Prof. **Li** noted, inference becomes more computationally involved once we leave the comfort of the PAPER model, but very often it is still tractable on networks with hundreds or thousands of nodes.

1.1 Community structure and homophily

Yang and Tong, **Srakar**, and **Wyse, Ng, White, and Fop** all pointed out that the PAPER notion of community does not seem to take into account the phenomenon of homophily, where nodes with similar characteristics tend to connect with each other, e.g. a person tends to become friends with others of similar cultural backgrounds.

Homophily has been the main consideration behind the design of statistical network models for the past two decades. It is a primary motivation for stochastic block models and various latent space models. Homophily is certainly important, but one of the theses of our work is that other features of network data are also important and have been overlooked. Specifically, we believe that the underlying Markovian growth process plays an equally crucial role in forming the topology of a network. Indeed, because the PAPER model is able to obtain good community detection results on real world networks, there is good reason to believe that when estimating communities in a network, one should account for not just homophily but also by the underlying growth processes of the communities.

We did not explicitly incorporate homophily in the PAPER model in order to keep the model simple. Just as it is good practice to control as many extraneous variables as possible in a scientific experiment, we believed we can illustrate our ideas most clearly by having the model focus on the growth process and omitting extraneous features. However, we agree that it would be appropriate to incorporate homophily into the PAPER model in many applications. One example is the PAPER-SBM model which we briefly discussed in the paper. We discuss a few other examples here.

If we know the characteristics of the nodes that induce homophily, then the extension proposed by **Yang and Tong** is very sensible and in fact similar to a proposal by Kim and Altmann (2017), who studied the effect of accounting for homophily in the preferential attachment model. To adapt their proposal to the PAPER model, we could, when generating the tree $\boldsymbol{T}$, have a new node u connect to an existing node v with probability

$$\frac{\{\beta D(v) + \alpha\}\Lambda_{uv}}{\sum_w \{\beta D(w) + \alpha\}\Lambda_{uw}},$$

where Λ_{uv} is the affinity between node u and node v; for example, we may have $\Lambda_{uv} = 1$ if u, v are in the same community and $\Lambda_{uv} < 1$ otherwise. The sampling algorithm becomes more difficult because we cannot sample an ordering uniformly at random from the spanning tree. However, we could adapt the swapping algorithm that we proposed for inference on the seq-PAPER model.

When we do not know the characteristics of the nodes, we could try to estimate them, as **Jiang and Ke** suggested. We could also incorporate latent variables into the model in a way that combines latent space models with Markovian network models, as **Wyse, Ng, White, and Fop** suggested. For instance, we may consider a model where for each node u, we generate latent

representation $Z_u \in \mathbb{R}^d$. Then, to generate the tree $\boldsymbol{T}$, we connect a new node u to an existing node v with probability

$$\frac{Z_u^\top Z_v}{\sum_w Z_u^\top Z_w}.$$

It is unclear how one can estimate the latent representations. There may not even be enough information in the network to do so. Spectral methods that are typically used in latent space models seem unlikely to succeed here.

1.2 Triangles and transitivity

Both **Li** and **Rubin-Delanchy** pointed out that the PAPER model may produce too few triangle subgraphs, that is, three nodes that are all connected to each other. Real world networks tend to have many triangles, reflecting the fact that two people who are friends with the same person tend to be friends with each other as well. In network data analysis, this is referred to as *transitivity*. One measure of transitivity is the global clustering coefficient defined as

$$\text{global clustering coef} = \frac{\text{number of triangles}}{\text{number of connected triplets}}.$$

Another common measure is the average local clustering coefficient:

$$C_u = \frac{\text{number of triangles in the neighbourhood of } u}{\frac{1}{2}\text{Deg}(u) \cdot (\text{Deg}(u) - 1)},$$
$$\text{average local clustering coef} = \frac{1}{n}\sum_u C_u.$$

We expect a PAPER graph to produce more triangles than an Erdős–Rényi graph with the same number of nodes and edges. This is because a preferential attachment tree $\boldsymbol{T}_n$ tends to have hubs and any two 'spoke' nodes in the neighbourhood of the hub centre can form a triangle by forming an edge between each other. We perform simulation experiments which confirm that PAPER graphs produce more triangles than Erdös–Rényi (ER) graphs. The simulation results are shown in Tables 1 and 2. In these simulations, we generate a PAPER graph with n nodes and fix the number of edges to be m; that is, we generate Erdős–Rényi noise by selecting $[m - (n - 1)]$ edges from $n(n-1)/2 - (n-1)$ potential pairs at random. We then compare the global clustering coefficient and the average local clustering coefficient against an Erdős–Rényi graph with the same number of nodes and edges. We repeat the experiment 100 times to generate the results.

We observe that a PAPER graph with linear preferential attachment (LPA) tree ($\alpha = 0, \beta = 1$) has higher transitivity measure than an ER graph as expected. The transitivity measures do decrease as

Table 1. $n = 100$ nodes and $m = 300$ edges

	$\alpha = 0, \beta = 1$	$\alpha = 1, \beta = 0$	ER
Global clustering coefficient	0.065	0.058	0.059
Average local clustering coef	0.077	0.058	0.058

Table 2. $n = 1,000$ nodes and $m = 3,000$ edges

	$\alpha = 0, \beta = 1$	$\alpha = 1, \beta = 0$	ER
Global clustering coefficient	0.0068	0.0059	0.0057
Average local clustering coef	0.0087	0.0060	0.0055

the size of the graph increases, which is undesirable but also a feature of the stochastic block model. One way to increase transitivity in a Markovian network model would be to use a random-walk mechanism to generate the noise edges, as proposed in Bloem-Reddy and Orbanz (2018).

There is the question of just how important it is to match the number of triangles in a network model with that of the data. Is it worth adding extra complexity to the model? We believe one must consider the end goal of the analysis, that is, the model choice should depend on whether we are trying to estimate communities, predict links, infer root nodes, or extract other information. To help practitioners answer this question in a principled way, we believe more research is needed on goal-specific model selection methods for network data. Promising work in this direction include network resampling methods such as ones proposed by Li et al. (2020), but these require assumptions that may not hold for Markovian networks. For stochastic block model (SBM), this is related to estimating the number of communities, for which there are good methods (Jin et al., 2023). We continue our discussion on model selection more in Section 1.4.

1.3 Number of communities

Ascolani, Lijoi, and Prünster and **Catalano, Fasano, Giordano, and Rebaudo** raised the point that our treatment of multiple root nodes assumes that the number of communities grows according to the dynamics of a one-parameter Ewens process, also known as the Chinese restaurant process. Those authors went on to highlight the potential benefit of considering alternative models for the number of clusters, such as the two-parameter extension of the Ewens process (De Blasi et al., 2013). We agree that this seems a sensible suggestion that adds flexibility to the class of models presented here. As highlighted at the outset of our response, our specific choice of the one-parameter family here is intended as a starting point for introducing a framework for modelling network data that arises from a growth process, rather than as a be-all, end-all. Extensions and modifications, such as the one above regarding the distribution of the number of clusters, are often necessary when attempting to apply this model—or any model for that matter—in a practical setting.

One interesting observation is that if we apply the Pitman–Yor process prior with a particular parameter setting, we obtain a particularly simple model:

1. Generate $\boldsymbol{G}^0 \sim \mathrm{PAPER}(\alpha, \beta, \theta)$ with $K = 1$ and $\beta > 0$.
2. Remove node 1 and all edges incident on node 1 from $\boldsymbol{G}^0$ to obtain $\boldsymbol{G}$. Output $\boldsymbol{G}$ without labelling.

Once we 'retire' the original root node, each direct child of the original root node becomes a new root of a community-tree. Note that in this model, we give each root a base degree of 1 instead of a base degree of 2. This model is a specific example of 'vertex retirement' described by **Jog and Loh**.

The sampling procedure for this model is also simple. We can sample π from distribution $\mathbb{P}(\Pi = \cdot \mid \tilde{\boldsymbol{F}}, \tilde{\boldsymbol{G}})$ by sampling an ordering from the history of $\tilde{\boldsymbol{F}}$ uniformly at random, equivalent to the cases described in the paper. To sample a forest $\tilde{\boldsymbol{f}}$ from distribution $\mathbb{P}(\tilde{\boldsymbol{F}} = \cdot \mid \Pi, \tilde{\boldsymbol{G}})$, we follow Algorithm 3 in our paper except we choose a new parent $w_t \in \{\emptyset\} \cup (\pi_{1:(t-1)} \cup N_{\tilde{g}_n}(\pi_t)$ for π_t with probability proportional to

$$\begin{cases} (\beta K + \alpha)/\hat{\theta} & \text{for } w_t = \emptyset \\ \beta D_{f_n^{(\cdot,\pi_t)}}(w_t) + \alpha & \text{for } w_t \neq \emptyset, \end{cases}$$

where K is the current number of trees in $\boldsymbol{f}$. We have implemented this prior and it works as expected in simulation studies. We will study the specification of prior in the random K model more extensively in a future work.

1.4 Model misspecification and model selection

The points made by the discussants piqued our own curiosity on a question: how do we interpret frequentist inference results when we know for a fact that our model is misspecified? This question

pertains to all of statistics but it is particularly tricky for network data for two reasons. First is that networks edges have strong dependence so we cannot assume we have IID observations; the second related reason is that we typically do not have universality phenomenon such as central limit theorem. How then do we interpret confidence set for the root node? **Wyse, Ng, White, and Fop** raised a concern very similar to this.

We give several answers to this question although we acknowledge that none of them are perfectly satisfactory. First, our confidence set for the root node has a Bayesian interpretation: conditioned on the event that the observed network is the result of the PAPER generating mechanism, we can find the plausible root nodes. Second, our procedure has a combinatorial interpretation where it is computing a combinatorial centrality measure, which can be viewed as a generalization of rumour centrality (Shah & Zaman, 2011); see Section 3.4 of our paper. Third, because the posterior root probability $\mathbb{P}(\Pi_1 = u \mid \tilde{G}_n)$ of node u is proportional to the likelihood of node u being the root node, we can interpret nodes with highest posterior root probability as being the root of the Kullback–Leibler projection of the actual network generative model to the PAPER model.

Finally, the issue of model misspecification can be somewhat alleviated by allowing a Markovian network to start from a *seed graph* instead of a single root node or several root nodes, as **Jog and Loh** suggested. There are several significant works in this direction (Devroye & Reddad, 2019; Lugosi & Pereira, 2019) but a number of major challenges remain. One difficulty is computation—the number of potential seed graphs in a network increases exponentially with the size of the seed graph. Another is model selection—models with a larger seed graph contain models with small seed graph. An interesting question then is how to select or estimate the size of the seed graph. Traditional model selection criteria such as AIC do not seem directly applicable in this setting.

As we mentioned before in the discussion on transitivity, more research is needed on model selection methods for networks. **Jog and Loh** echoed this sentiment when they asked how one would choose between the different variations of the PAPER model: fixed K versus random K, sequential versus nonsequential. This question can be extended to include the change-point extension proposed by **Wang, Yu, and Rinaldo** and the heterogeneous affine preferential attachment extension proposed by **Feng and Sun**. How should we choose among these models? One idea is to conduct goodness-of-fit test and compare the p-value of each of the potential models, but this simply leads to the question of how to choose a statistically and computationally efficient test statistic. One potential approach is be to construct the test statistic based on the degree distribution or higher order subgraph counts and generate Monte Carlo p-value by simulation.

2 Theory

2.1 Root inference formulation

In this section, we respond to Prof. **Rubin-Delanchy**'s stimulating comments regarding how to best formulate the root inference problem. We thank him for bringing group theory to the fore, which has helped us clarify a lot of our own thoughts.

Root inference is a problem that is intuitively easy to understand but difficult to formalize. This is in part because the observed data is in actuality an unlabelled graph. It is not well-defined to refer to a specific node in an unlabelled graph. For example, in a chain graph with four nodes, we cannot distinguish between the two end-point nodes or the two interior nodes.

There are many different ways to formalize the root inference problem, as Prof. **Rubin-Delanchy**'s discussion demonstrates. What is remarkable (and also comforting) is that these formulations are all equivalent in that the sense that they give the same notion of conditional root probability. In the manuscript, our goal was to give a formulation that is both rigorous and also efficient in the sense that the reader can go from the problem definition to the methodology/algorithm without needing to digest new concepts. This is the main reason we introduced the random labelling device: it leads naturally to the inference algorithm. It is not just a way to strip away the node label information. In what follows, we lay out all the formulations clearly and formally establish their equivalence. We leave the readers to choose the formulation that makes the most sense to them.

Let $\boldsymbol{G}$ be a random graph with n nodes whose nodes are labelled using $[n] := \{1, 2, \ldots, n\}$. We think of $\boldsymbol{G}$ as a Markovian model whose nodes are labelled by their arrival time, although the

technical results apply to any random graph models. We observe $\boldsymbol{G}^* = \rho \boldsymbol{G}$ where $\rho \in S_n$ is an unknown permutation. In the manuscript, we supposed that the node labels of $\boldsymbol{G}^*$ take value in some alphabet to make it clear that the node labels of $\boldsymbol{G}^*$ do not correspond to the arrival time.

Random relabelling formulation:

This is the formulation we gave in the paper. Let Π be a random permutation distributed uniformly over S_n. For a node $v \in \boldsymbol{G}^*$, we define its conditional root probability as

$$p_1(v) := \mathbb{P}\{\Pi_1 = v \mid \Pi \boldsymbol{G} = \boldsymbol{G}^*\}. \tag{1}$$

We note that in the paper, we wrote $\tilde{\boldsymbol{G}} = \Pi \boldsymbol{G}$ to denote the randomly relabelled graph. As Prof. **Rubin-Delanchy** pointed out, we do not need to actually apply randomization—we simply view the observed graph $\boldsymbol{G}^*$ as being randomly labelled.

Group-theoretic formulation:

This is the formulation that Prof. Rubin-Delanchy gave. Suppose that the unobserved permutation ρ is uniformly random. Define $\mathrm{Aut}(\boldsymbol{G}^*) = \{\pi \in S_n : \pi \boldsymbol{G}^* = \boldsymbol{G}^*\}$ as the automorphism group and define $M_{\boldsymbol{G}^*} := \frac{n!}{|\mathrm{Aut}(\boldsymbol{G}^*)|}$ as the number of distinct labelled graphs of the same shape as $\boldsymbol{G}^*$; we put $\boldsymbol{G}^*$ in the subscript but we note that $M_{\boldsymbol{G}^*}$ depends only on the unlabelled shape of $\boldsymbol{G}^*$. We enumerate these distinct labelled graphs as $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{M_{\boldsymbol{G}^*}}$.

For a node $v \in \boldsymbol{G}^*$, we define its orbit $o(v, \boldsymbol{G}^*) = \{\pi v : \pi \in \mathrm{Aut}(\boldsymbol{G}^*)\}$, which is the set of nodes of $\boldsymbol{G}^*$ indistinguishable from v once the node labels are removed. We then define the conditional root probability of v as

$$p_2(v) := \mathbb{P}\{\rho(1) \in o(v, \boldsymbol{G}^*) \mid \boldsymbol{G}^*, \boldsymbol{G} \in \{\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{M_{\boldsymbol{G}^*}}\}\}. \tag{2}$$

Unlabelled shape formulation:

This is a formulation that we described in a previous work (Crane & Xu, 2021). The idea is that we do not even need to define an unobserved permutation ρ in order to formalize the root inference problem. It is well-defined to write the following conditional probability:

Indeed, the APA tree model with $\alpha = 1$ and $\beta = 0$ produces the four node chain graph on the right if it produces $1 \to 2 \to 3 \to 4$ or $2 \leftarrow 1 \to 3 \to 4$ or $3 \leftarrow 1 \to 2 \to 4$ or $4 \leftarrow 1 \to 2 \to 4$, with total probability $\frac{2}{3}$. It produces the rooted chain graph on left if it produces $2 \leftarrow 1 \to 3 \to 4$ or $3 \leftarrow 1 \to 2 \to 4$ or $4 \leftarrow 1 \to 2 \to 3$, with total probability of $\frac{1}{2}$ so that the conditional probability evaluates to $\frac{3}{4}$.

More generally, for a labelled graph $\boldsymbol{G}$, define its *shape* (unlabelled graph) as the equivalence class

$$\mathrm{sh}(\boldsymbol{G}) := \{\boldsymbol{G}' : \boldsymbol{G}' = \pi \boldsymbol{G}, \exists \pi \in S_n\}.$$

The cardinality of $\mathrm{sh}(\boldsymbol{G})$ is exactly $M_{\boldsymbol{G}}$ as defined in the group-theoretic formulation section. We note that $\mathrm{sh}(\boldsymbol{G})$ is the quotient set $S_n/\mathrm{Aut}(\boldsymbol{G})$ (it is not a quotient group since $\mathrm{Aut}(\boldsymbol{G})$ may not be a normal subgroup of S_n).

Similarly, we define the notion of a rooted shape. Let v be a node in $\boldsymbol{G}$, then define

$$\mathrm{sh}_0(\boldsymbol{G}, v) := \{(\boldsymbol{G}', v') : \boldsymbol{G}' = \pi \boldsymbol{G}, v' = \pi v, \exists \pi \in S_n\}.$$

If we define the subgroup $S_n(v) = \{\pi \in S_n : \pi_1 = v\}$ and $\mathrm{Aut}(\boldsymbol{G}, v) := \{\pi \in S_n : \pi \boldsymbol{G} = \boldsymbol{G}, \pi v = v\}$, then $\mathrm{sh}_0(\boldsymbol{G}, v)$ is the quotient group $S_n(v)/\mathrm{Aut}(\boldsymbol{G}, v)$.

For a node $v \in \boldsymbol{G}^*$, we define the conditional root probability as

$$p_3(v) := \mathbb{P}\{(\boldsymbol{G}, 1) \in \mathrm{sh}_0(\boldsymbol{G}^*, v) \mid \boldsymbol{G} \in \mathrm{sh}(\boldsymbol{G}^*)\}. \tag{3}$$

In our previous work (Crane & Xu, 2021), we define $\mathrm{Eq}(v, \boldsymbol{G}) = \{u \in V(\boldsymbol{G}) : \pi u = v,\ \pi \boldsymbol{G} = \boldsymbol{G},\ \exists \pi \in S_n\}$ as the set of nodes of $\boldsymbol{G}$ that are *equivalent* to v. This set is exactly the orbit $o(v, \boldsymbol{G})$. We note that $o(v, \boldsymbol{G})$ is the quotient set $\mathrm{Aut}(\boldsymbol{G})/\mathrm{Aut}(\boldsymbol{G}, v)$. We thus have that

$$|o(v, \boldsymbol{G})| = \frac{|\mathrm{Aut}(\boldsymbol{G})|}{|\mathrm{Aut}(\boldsymbol{G}, v)|} = n\frac{|\mathrm{sh}_0(\boldsymbol{G}, v)|}{|\mathrm{sh}(\boldsymbol{G})|}.$$

Equivalence:

The following theorem equates all three notions of conditional root probability. In the theorem below, we need to divide $p_2(v)$ and $p_3(v)$ by $|o(v, \boldsymbol{G}^*)|$ because these are probability of a entire set—the orbit of v. In contrast, $p_1(v)$ is the probability of a single node. One consequence of the theorem is that $p_1(\cdot)$ is a constant for all nodes in the same orbit.

Theorem 1 Let v be any node in $\boldsymbol{G}^*$ and let $p_1(v)$, $p_2(v)$, $p_3(v)$ be defined as in (1), (2), (3), respectively. Then, we have that

$$p_1(v) = \frac{p_2(v)}{|o(v, \boldsymbol{G}^*)|} = \frac{p_3(v)}{|o(v, \boldsymbol{G}^*)|}.$$

Proof. Given a node v of $\boldsymbol{G}^*$, define

$$\mathcal{L}(v, \boldsymbol{G}^*) = \frac{1}{|o(v, \boldsymbol{G}^*)|}\sum_{g} \mathbb{P}(\boldsymbol{G} = g)\mathbb{1}\{(g, 1) \in \mathrm{sh}_0(\boldsymbol{G}^*, v)\},$$

where the summation is taken over all graphs whose nodes are labelled with $\{1, 2, 3, \ldots, n\}$. It follows from Theorem S5 of our paper [see also Theorem 8 in Crane and Xu (2021)] that

$$p_1(v) = \frac{\mathcal{L}(v, \boldsymbol{G}^*)}{\sum_{u\in[n]} \mathcal{L}(u, \boldsymbol{G}^*)}. \tag{4}$$

Intuitively, (4) holds because $\mathrm{sh}(\boldsymbol{G}^*, v) = S_n(v)/\mathrm{Aut}(\boldsymbol{G}^*, v)$ and $o(v, \boldsymbol{G}^*) = \mathrm{Aut}(\boldsymbol{G}^*)/\mathrm{Aut}(\boldsymbol{G}^*, v)$ so that

$$\begin{aligned} p_1(v) &\propto \sum_{\pi\in S_n : \pi_1 = v} \mathbb{P}(\boldsymbol{G} = \pi^{-1}\boldsymbol{G}^*) = \sum_{(g,1)\in \mathrm{sh}_0(\boldsymbol{G}^*, v)} \sum_{\pi\in \mathrm{Aut}(\boldsymbol{G}^*, v)} \mathbb{P}(\boldsymbol{G} = \pi^{-1}g) \\ &= \sum_{(g,1)\in \mathrm{sh}_0(\boldsymbol{G}^*, v)} |\mathrm{Aut}(\boldsymbol{G}^*, v)|\mathbb{P}(\boldsymbol{G} = g) \propto \frac{1}{o(v, \boldsymbol{G}^*)} \sum_{(g,1)\in \mathrm{sh}_0(\boldsymbol{G}^*, v)} \mathbb{P}(\boldsymbol{G} = g). \end{aligned}$$

Then, we have that

$$\begin{aligned} p_3(v) &= \frac{\mathbb{P}\{(\boldsymbol{G}, 1) \in \mathrm{sh}_0(\boldsymbol{G}^*, v)\}}{\mathbb{P}\{\boldsymbol{G} \in \mathrm{sh}(\boldsymbol{G}^*)\}} \\ &= \frac{\sum_g \mathbb{P}(\boldsymbol{G} = g)\mathbb{1}\{(g, 1) \in \mathrm{sh}_0(\boldsymbol{G}^*, v)\}}{\sum_{u\in[n]}\sum_g \mathbb{P}(\boldsymbol{G} = g)\mathbb{1}\{(g, 1) \in \mathrm{sh}_0(\boldsymbol{G}^*, u)\}} = p_1(v)|o(v, \boldsymbol{G}^*)|. \end{aligned}$$

Finally, we also have that

$$p_2(v) = \frac{\sum_{(g,1)\in \mathrm{sh}_0(\boldsymbol{G}^*, v)} \mathbb{P}(\boldsymbol{G} = g)}{\sum_{g\in \mathrm{sh}(\boldsymbol{G}^*)} \mathbb{P}(\boldsymbol{G} = g)} = p_3(v).$$

The desired conclusion follows. □

2.2 Theoretical guarantees

Various discussants raise a number of questions related to bounds on the size of the confidence set which we discuss here.

Banerjee pointed out that to optimally bound the size of the confidence set when $\alpha = 1$ and $\beta = 0$, one needs a nonlocal centrality measure beyond the degree. We agree that an alternative centrality measure is needed but what that alternative should be remains an elusive question. We unsuccessfully attempted to use anchors of double cycles, which is a clever idea proposed by Briend et al. (2023) to study Cooper–Frieze networks and other related models. Cooper–Frieze network is essentially the seq-PAPER model with $\alpha = 1$, $\beta = 0$ and where the noise edge probability is $\theta_t = \theta_0 \frac{1}{t}$; it differs from the PAPER model in that early nodes tend to be more tightly connected amongst each other. Because of this difference, we could not obtain satisfactory results using the 'anchor of double-cycle' idea. We also do not see how to extend Jordan centrality (Bubeck et al., 2017) to the nontree setting. For now, the problem remains open.

For the multiple root setting, **Yang and Tong** asked what could be proved about community detection while **Jog and Loh** asked about bounds on the size of the confidence set. We conjecture that exact community recovery is impossible under the PAPER model, because it does not seem likely that we can perfectly estimate the community membership of the peripheral leaf nodes of a community-tree. We do believe, however, that, in the fixed K setting, the early nodes of each of the K communities can be consistently clustered. To be more precise, it may be possible to obtain consistency if we use a weighted misclustering measure where we weigh each node by the inverse of its arrival time. A potential approach may be to first show that the confidence set of the K root nodes are likely to comprise K disjoint subgraphs each of which correspond to a community.

Qing and Tong asked whether the posterior root distributions can be used to construct test statistics for testing the number of communities in the network. This is an interesting question. We have not studied how the posterior root distribution behaves when the specified K is either smaller or larger than the true K_0.

Jog and Loh asked about frequentist guarantee for the credible set in the sequential noise setting. This is indeed true as Theorem 7 in our paper applies to the sequential noise setting. They then ask about the what guarantees can be provided under the random K model. There are two layers to this question. First, if we assume that the Dirichlet process prior is well specified, then it follows from the conditional coverage that we would also have marginal coverage in that the credible set contains all the K root nodes with at least $1 - \epsilon$ probability. However, if we suppose that the graph is actually generated according to the PAPER model with fixed K roots but where K is unknown, then we do not expect our credible set to have frequentist guarantees.

Jog and Loh also asked about constructing confidence set for the K root nodes as a set of K-tuples. This is easy to do with our methodology. The reason we did not investigate this approach in the paper is that the resulting set of K-tuples may be too large, especially when K is large.

3 Miscellaneous points

Response to miscellaneous points in the discussion by **Banerjee**:

Banerjee suggested empirical comparison between the size of our confidence set with those constructed by probability analysis. We have conducted these comparisons in our previous paper on the tree setting (Crane & Xu, 2021) and found that the latter confidence sets are overly conservative.

Response to miscellaneous points in the discussion by **Qing and Tong**:

We are grateful for the suggestion of creating a markdown file to illustrate the model and the methodology in the simplest setting possible. We plan to implement our algorithm in R and provide such an illustrative markdown file. We are also grateful for the suggestion of using the PAPER model for tasks beyond minimizing misclustering error. We believe two promising examples include graph summarization and hierarchical clustering.

Response to miscellaneous points in the discussion by **Wang, Yu, and Rinaldo**:

We thank Wang, Yu, and Rinaldo for their interesting formulation of a change-point problem on the PAPER model. This formulation is similar to the change-point model analyzed in Banerjee et al. (2023). Banerjee et al. (2023) consider the generalized preferential attachment tree model and

study two single-change settings: one is where the change-point occurs at time γn and the other is where the change-point occurs early at time n^{γ} for some $\gamma \in (0, 1)$. They propose consistent estimators based on the empirical degree distribution. Their work, along with the question raised by Wang, Yu, and Rinaldo, show that change-point estimation in Markovian networks has unique properties which need to be better understood. The likelihood based estimation approach proposed by Wang, Yu, Rinaldo is a promising direction.

Response to miscellaneous points in the discussion by **Ascolani, Lijoi, and Prünster**:

Ascolani, Lijoi, and Prünster raised the point of considering multiple networks instead of just one, citing an example regarding co-authorship structures among different academic communities. Interestingly, we illustrated our approach on the statisticians' co-authorship network, for which there are numerous overlapping subcommunities. In that application, we showed that the joint dynamics of the communities helps in allowing us to infer structure about the network. There are a number of interesting problems we may consider if we observe multiple networks. When two network have the same set of nodes, we may consider the setting where they share parts of the latent growth history, that they co-evolved in some sense.

Response to miscellaneous points in the discussion by **Wyse, Ng, White, and Fop**:

Wyse, Ng, White, and Fop made an insightful comment regarding the EM estimation algorithm for the α, β parameters in the PAPER model. In the paper, we made two approximations in the EM algorithm. First, we approximated the conditional distribution of the tree degree given the graph $\mathbb{P}_{\alpha}\{j < D_{\tilde{T}_n}(v)|\tilde{\boldsymbol{G}}_n\}$ by $\mathbb{P}_{\alpha}\{j < D_{\tilde{T}_n}(v)|D_{\tilde{G}_n}(v)\}$, which ignores the dependence between the graph degree of all the nodes. Second, when computing $\mathbb{P}_{\alpha}\{j < D_{\tilde{T}_n}(v)|D_{\tilde{G}_n}(v)\}$, we approximated the marginal distribution of $D_{\tilde{T}_n}(v)$ by its asymptotic limit.

We do not have theoretical analysis on how significantly these approximations would affect estimation accuracy of the resulting EM procedure. We conjecture that the second approximation has only a small effect because there is uniform convergence of the finite n distribution of the degree $D_{\tilde{T}_n}(v)$ to its asymptotic limiting distribution (see (Van Der Hofstad, 2016, Theorem 8.2). The first approximation may be inaccurate if the observed graph $\boldsymbol{G}$ does not resemble a PAPER graph at all. For example, we can show that the approximation is poor if $\boldsymbol{G}$ is a cycle graph.

It would be interesting to compare the approximate EM algorithm with either Monte Carlo EM or Bayesian inference where we put a prior on the α, β parameters.

Response to miscellaneous points in the discussion by **Catalano, Fasano, Giordano, and Rebaudo**:

The susceptible-infectious-recovered (SIR)-inspired model proposed by Catalano, Fasano, Giordano, and Rebaudo is interesting. If we view the PAPER model as an infection process, then we are not making assumptions about the infection time; the arrival ordering of the nodes would reflect the order of infection. One could consider a model where in each infection 'wave', each existing (infected) node recruits (infects) some number of new nodes. Our intuition is that a network generated by such a model would provide more information about the root node—it may even enable consistent estimation of the root node. We also note that SIR is used to motivate infection process that occurs over a fixed background graph, which Li discussed.

Response to miscellaneous points in the discussion by **Feng and Sun**:

We thank Feng and Sun for their interesting discussion on the heterogeneous affine preferential attachment (HAPA) model. This proposal allows the different trees in the $K > 1$ setting to have potentially different growth dynamics, i.e. potentially different α, β parameters. To estimate the K different (α_k, β_k) parameter pairs, Feng and Sun propose a Monte Carlo EM algorithm; this could be effective even in the PAPER setting. They also suggest a way to adapt the Gibbs sampling procedure to the HAPA setting. One novel aspect is, before sampling the ordering π given a forest $\boldsymbol{f}$, to first samples a sequence of tree assignment history $\{(x_{1,t}, \ldots, x_{K,t})\}_{t=K+1}^{n}$ where $x_{k,t}$ is the number of nodes in tree k at time t. Sampling the tree assignment history conditional on the final forest $\boldsymbol{f}$ is challenging. The proposed approach may not ensure consistency of the sampled history with the forest $\boldsymbol{f}$. Sequential Monte Carlo methods may be appropriate here.

Response to miscellaneous points in the discussion by **Srakar**:

Srakar stated that it seems unexplained why we are using a Bayesian approach. The reasons are that (1) the Bayesian credible set for the root node has frequentist guarantee, (2) the Gibbs sampler

is scalable, and (3) the resulting confidence set has size of an optimal order—we discuss each of these in the paper. Srakar mentioned connections to temporal networks which we agree merit additional research.

Response to miscellaneous points in the discussion by **Jog and Loh**:

Jog and Loh suggested a number of extensions to the PAPER model. The notion of 'vertex retirement' is particularly interesting, from both practical and theoretical perspectives. Interestingly, if we 'retire' the root node in a single root PAPER model, we obtain a random K PAPER model where formation of new trees is governed by a Pitman–Yor process. See discussion in Section 1.3.

Response to miscellaneous points in the discussion by **Rubin-Delanchy**:

In addition to raising a subtle technical question which we addressed in Section 2.1, Rubin-Delanchy stated that his 'most important concern at the time was that in the applications cited [$\cdots$] there would almost always be timing information on the edges. It would seem highly irresponsible to ignore this in practice.' We comment here that our proposed model and subsequent methodological developments are framed specifically in the setting in which there is no additional information other than the contact pattern (i.e. 'shape'). In particular, we explicitly assume that time information is unavailable for the network. At no point do we suggest that one should ignore such information if it were to exist.

If exact time information is known, then the root inference problem becomes trivial of course. In most situations; however, the time information is noisy or unreliable so that it could be very helpful to also incorporate the network structure information. For example, in a disease infection network, we may only have rough guesses on the actual times of infection of all the individuals.

Response to miscellaneous points in the discussion by **Li**:

Li raised a connection of the PAPER model to a diffusion process over a fixed background graph. In that setting, there is a set of infected nodes that start as a single root and, at every iteration, infects an additional neighbouring node chosen at random.

If we assume that the root node is chosen uniformly at random, then we can also define the posterior root distribution. However, the credible set formed from the posterior root distribution will not have frequentist coverage in general. Frequentist coverage does hold when the background graph exhibits symmetry. For instance, if the background graph is an infinite regular tree or an infinite grid graph.

Response to miscellaneous points in the discussion by **Jiang and Ke**:

We thank the Jiang and Ke for showcasing the PAPER model on a citation network. In this case, because the directions of the edges are removed, we expect the posterior root distribution to assign higher probabilities to survey papers that cite many influential papers. This seems to be the case in the result that Jiang and Ke obtained.

Jiang and Ke stated that the PAPER model fails to account for degree homogeneity due to its use of the Erdős–Rényi model. We point out that the PAPER model does, in fact, model degree homogeneity as a result of its preferential attachment dynamics. Importantly, the preferential attachment component of PAPER is the dominant structural component, whereas Erdős–Rényi serves as a secondary noise distribution on top of the main structural layer. Furthermore, we point out that PAPER addresses degree homogeneity with only three parameters, whereas the suggested degree-corrected stochastic blockmodel requires a separate parameter for every node of the graph, so that the number of parameters depends on sample size.

Acknowledgments

M. Xu is supported by National Science Foundation Grants DMS-2113671 and DMS-2311299.

Conflict of interest: None declared.

References

Banerjee S., Bhamidi S., & Carmichael I. (2023). Fluctuation bounds for continuous time branching processes and evolution of growing trees with a change point. *The Annals of Applied Probability*, *33*(4), 2919–2980. https://doi.org/10.1214/22-AAP1881

Bloem-Reddy B., & Orbanz P. (2018). Random-walk models of network formation and sequential monte carlo methods for graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 871–898. https://doi.org/10.1111/rssb.12289

Briend S., Calvillo F., & Lugosi G. (2023). Archaeology of random recursive dags and cooper-frieze random networks. *Combinatorics, Probability and Computing*, *32*(6), 859–873. https://doi.org/10.1017/S0963548323000184

Bubeck S., Devroye L., & Lugosi G. (2017). Finding Adam in random growing trees. *Random Structures & Algorithms*, *50*(2), 158–172. https://doi.org/10.1002/rsa.v50.2

Crane H., & Xu M. (2021). Inference on the history of a randomly growing tree. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(4), 639–668. https://doi.org/10.1111/rssb.12428

De Blasi P., Favaro S., Lijoi A., Mena R. H., Prünster I., & Ruggiero M. (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 212–229. https://doi.org/10.1109/TPAMI.2013.217

Devroye, L., & Reddad, T. (2019). On the Discovery of Seed in Uniform Attachment Tree. *Internet Mathematics*, *1*(1). https://doi.org//10.48550/arXiv.1810.00969

Jin J., Ke Z. T., Luo S., & Wang M. (2023). Optimal estimation of the number of network communities. *Journal of the American Statistical Association*, *118*(543), 2101–2116. https://doi.org/10.1080/01621459.2022.2035736

Kim K., & Altmann J. (2017). Effect of homophily on network formation. *Communications in Nonlinear Science and Numerical Simulation*, *44*, 482–494. https://doi.org/10.1016/j.cnsns.2016.08.011

Li T., Levina E., & Zhu J. (2020). Network cross-validation by edge sampling. *Biometrika*, *107*(2), 257–276. https://doi.org/10.1093/biomet/asaa006

Lugosi G., & Pereira A. S. (2019). Finding the seed of uniform attachment trees. *Electronic Journal of Probability*, *24*, 1–15. https://doi.org/10.1214/19-EJP268

Shah D., & Zaman T. (2011). Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, *57*(8), 5163–5181. https://doi.org/10.1109/TIT.2011.2158885

Van Der Hofstad R. (2016). *Random graphs and complex networks*. (Vol. *1*). Cambridge University Press.

https://doi.org/10.1093/jrsssb/qkae052
Advance access publication 7 June 2024