A Gibbs sampling algorithm for root inference on a growing network

Nilava Metya

1 Introduction

Real-world networks, such as those modeling disease transmission, fake news propagation, or computer virus spread, can be effectively represented using labeled growing trees. This study focuses on a time-labeled tree, initiated with an infected individual at $t = 1$ (the root). At each subsequent time step $t = 2, 3, 4, \cdots$, individuals labeled t are infected from w_t and are added to the growing tree with an edge between t, w_t.

The study involves looking at a sequence of trees $\boldsymbol{T}_1 \subset \boldsymbol{T}_2 \subseteq \cdots$ where $\boldsymbol{T}_t$ has t labeled nodes (indicating their time of arrival). The structure of the new tree depends on a probability distribution linked to the pre-existing tree. The key condition is that the subgraph of T_i induced by vertices $\{1, \cdots, i-1\}$ remains connected for each $i \geq 2$. The primary interest lies in understanding the initial infected individuals, especially the root node.

This report delves into the model for a tree growing from a single source, referencing [CX23; CX21]. A presented Gibbs algorithm computes conditional probabilities to identify a set of nodes containing the root with confidence. The algorithm is applicable to unlabeled graphs, which may not be trees due to data collection errors introducing unwanted edges. The input graph is unlabeled since the order of infection is unknown during observation. An online presentation is accessible at [Met23].

2 The Model

The modeling process begins with the growth of a tree, followed by the addition of error edges based on a probability distribution. The tree part adheres to the affine preferential attachment (APA) model, and the error follows the Erdös-Rényi distribution.

Definition 1 (Affine Preferential Attachment)**.** The APA(α, β) tree model generates an increasing sequence $\boldsymbol{T}_1 \subset \boldsymbol{T}_2 \subseteq \cdots$ of random trees where $\boldsymbol{T}_i$ is a labeled tree with $V(\boldsymbol{T}_i) = [i]$. Starting with $\boldsymbol{T}_1 = (\{1\}, \varnothing)$, subsequent trees are built by introducing a node t and an edge $e = \{t, w_t\}$ at each step t, where $w_t \in V_{t-1}$ is chosen with probability $\frac{\beta \cdot \deg_{\boldsymbol{T}_{t-1}}(w_t) + \alpha}{2\beta(t-2) + \alpha(t-1)}$.

Definition 2 (Erdös-Rényi)**.** The Erdös-Rényi model on n vertices with parameter $\theta \in [0, 1]$, denoted by ER(n, θ), involves constructing undirected simple edges randomly among n labeled nodes, where the probability of an edge's existence is θ. Alternatively, the probability of generating a graph with n nodes and m edges is $\theta^m (1-\theta)^{\binom{n}{2} - m}$.

The model used for networks in disease transmission or the spread of fake news is the union of graphs from the above two models.

Definition 3 (PAPER)**.** The PAPER(α, β, θ) (Preferential Attachment Plus Erdös-Rényi) generates a sequence of graphs $\{\boldsymbol{G}_n\}_{n \in \mathbb{N}}$ such that $\boldsymbol{G}_n = \boldsymbol{T}_n \cup \boldsymbol{R}_n$, where $\{\boldsymbol{T}_1 \subset \boldsymbol{T}_2 \subset$

$\cdots\}$ is generated according to APA(α, β) and $\{\boldsymbol{R}_n\}_{n\in\mathbb{N}}$ is generated with $\boldsymbol{R}_n \sim \mathrm{ER}(n, \theta)$ independently, with $V(\boldsymbol{T}_n) = V(\boldsymbol{R}_n) = [n]$ for each $n \in \mathbb{N}$. Here, this is the union of two graphs on the same set of vertices, taking the union of edges and ignoring multi-edges.

3 The Root Inference Problem

Let $\boldsymbol{G}_n \sim \mathrm{PAPER}(\alpha, \beta, \theta)$ be a random graph. Only the unlabeled shape is observed. This graph is labeled using alphabets $\mathcal{A}_n$ to obtain $\boldsymbol{G}_n^*$ via the bijection $\rho : [n] \xrightarrow{\sim} \mathcal{U}_n$, inducing a graph isomorphism $\rho_* : \boldsymbol{G}_n \xrightarrow{\sim} \boldsymbol{G}_n^*$. The goal is to infer $\rho_1 \coloneqq \rho(1)$.

Without information about ρ, finding ρ_1 is impossible. Therefore, a confidence set is sought for a given confidence level.

Definition 4 (Confidence Set)**.** Let $\varepsilon \in (0, 1)$. $C_\varepsilon(\boldsymbol{G}_n^*) \subseteq \mathcal{U}_n$ is a $1-\varepsilon$ level confidence set for the root if $\mathbb{P}\left(\rho_1 \in C_\varepsilon(\boldsymbol{G}_n^*)\right) \geq 1-\varepsilon$.

A confidence set exists because $C_\varepsilon(\boldsymbol{G}_n^*) = \mathcal{U}_n$ works $\forall\varepsilon \in (0, 1)$, allowing for the smallest one to be exist by the well-ordering principle. Smaller confidence sets provide more information about the root. Consequently, our objective is to identify the smallest confidence set that meets the aforementioned confidence criteria.

Remark 5. Note that if $\rho, \rho' : [n] \xrightarrow{\sim} \mathcal{U}_n$ induce isomorphisms $\rho_* = \rho'_* : \boldsymbol{G}_n \xrightarrow{\sim} \boldsymbol{G}_n^*$, then $\rho_i \in C_\varepsilon(\boldsymbol{G}_n^*) \iff \rho'_i \in C_\varepsilon(\boldsymbol{G}_n^*)$ for any node i. This means that one would want the C_ε to be *labeling equivariant*: $\tau C_\varepsilon(\boldsymbol{G}_n^*) = C_\varepsilon(\tau\boldsymbol{G}_n^*)\forall\tau : \mathcal{U}_n \xrightarrow{\sim} \mathcal{U}_n$. In the presented algorithm, randomization is used to break ties, so $\tau C_\varepsilon(\boldsymbol{G}_n^*) \overset{d}{=} C_\varepsilon(\tau\boldsymbol{G}_n^*)\forall\tau : \mathcal{U}_n \xrightarrow{\sim} \mathcal{U}_n$.

4 Methodology

This section begins with the motivation that *no label is better than any other* since only an unlabeled graph is observed. Let $\boldsymbol{G}_n^* = \rho_*\boldsymbol{G}_n$ be a labeled observation, and let $\Lambda : \mathcal{U}_n \xrightarrow{\sim} \mathcal{U}_n$ be a random relabel. Denote $\Pi = \Lambda\rho : [n] \xrightarrow{\sim} \mathcal{U}_n$ and $\Pi_* \coloneqq \Lambda_*\rho_*$. Consider $\tilde{\boldsymbol{G}}_n \coloneqq \Pi_*\boldsymbol{G}_n$. Since Λ is uniformly random, Π is also uniformly random. Recall that $\boldsymbol{G}_n$ has the structure of $\boldsymbol{T}_n \cup \boldsymbol{R}_n$. Define $\tilde{\boldsymbol{T}}_n \coloneqq \Pi_*\boldsymbol{T}_n$ as the randomly labeled latent tree. Based on such a randomly labeled $\tilde{G}$, a method is described to find a C_ε for the root Π_1. Suppose the probabilities $p_u \coloneqq \mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g})$ are known. Sort the nodes $u_1, \cdots, u_n$ such that $p_{u_1} \geq \cdots \geq p_{u_n}$. Propose the credible set $B_\varepsilon(\boldsymbol{g}) \coloneqq \{u_1, \cdots, u_k\}$ (breaking ties at random), where k is the minimum such that $\sum_{i=1}^k p_{u_i} \geq 1-\varepsilon$. This Bayesian credible set is, in fact, an honest confidence set, as stated in the following theorem.

Theorem 6. Let $\boldsymbol{G}_n \sim \mathrm{PAPER}(\alpha, \beta, \theta)$, and let $\rho : [n] \to \mathcal{U}_n$ be a relabeling such that $\boldsymbol{G}_n^* = \rho_*\boldsymbol{G}_n$. Then for any $\varepsilon \in (0, 1)$, $\mathbb{P}\left(\rho_1 \in B_\varepsilon(\boldsymbol{G}_n^*)\right) \geq 1-\varepsilon$.

Note that in the above, it was assumed p_u is known. This is crucial to compute. It is observed that $\mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g}) = \sum_\pi \mathbf{1}_{\{\pi_1=u\}} \cdot \mathbb{P}(\Pi = \pi \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g})$. However, the RHS term is supported on only a subset of $\{\pi \mid \pi_1 = u\}$ because not all π are valid labels for growing trees. For example, 1—3—2 is invalid. This computation involves summing over all spanning trees $\tilde{\boldsymbol{T}}_n$. Thus $p_u = \sum_{\boldsymbol{t}\subset\boldsymbol{g}} \sum_{\substack{\pi\in\mathrm{hist}(\boldsymbol{t}) \\ \pi_1=u}} \mathbb{P}(\Pi = \pi, \tilde{\boldsymbol{T}}_n = \boldsymbol{t} \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g})$, where $\mathrm{hist}(\boldsymbol{t}) = \{\pi : [n] \xrightarrow{\sim} \mathcal{U}_n \mid \boldsymbol{t} \cap \pi([i]) \text{ is connected } \forall i \in [n]\}$ is the set of bijections π

representing a valid arrival ordering for the nodes of $\boldsymbol{t}$, and the outer sum is taken over all spanning trees $\boldsymbol{t}$ of $\boldsymbol{g}$. This already reduces the space over which the sum is taken. A Gibbs sampling algorithm will compute $\mathbb{P}(\Pi = \pi, \tilde{\boldsymbol{T}}_n = \boldsymbol{t} \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g})$.

5 The Gibbs Sampling Algorithm for Computing

The Gibbs sampling procedure is generally used to find the joint distribution when conditionals are known or easy to compute. The algorithm alternates between two steps:

1. Fix $\boldsymbol{t}$ and generate π from $\mathbb{P}(\Pi = \pi \mid \tilde{\boldsymbol{T}}_n = \boldsymbol{t}, \tilde{\boldsymbol{G}}_n = \boldsymbol{g})$, and
2. Fix π and generate $\boldsymbol{t}$ from $\mathbb{P}(\tilde{\boldsymbol{T}}_n = \boldsymbol{t} \mid \Pi = \pi, \tilde{\boldsymbol{G}}_n = \boldsymbol{g})$.

For the first step, $\mathbb{P}(\Pi = \pi \mid \tilde{\boldsymbol{T}}_n = \boldsymbol{t}, \tilde{\boldsymbol{G}}_n = \boldsymbol{g}) = \frac{\mathbb{P}(\Pi=\pi \mid \tilde{\boldsymbol{T}}_n=\boldsymbol{t}) \cdot \mathbb{P}(\tilde{\boldsymbol{G}}_n=\boldsymbol{g} \mid \Pi=\pi, \tilde{\boldsymbol{T}}_n=\boldsymbol{t})}{\mathbb{P}(\tilde{\boldsymbol{T}}_n=\boldsymbol{t} \mid \tilde{\boldsymbol{G}}_n=\boldsymbol{g})}$. One only needs to sample from $\mathbb{P}(\Pi = \pi \mid \tilde{\boldsymbol{T}}_n = \boldsymbol{t})$ because $\mathbb{P}(\tilde{\boldsymbol{G}}_n = \boldsymbol{g} \mid \Pi = \pi, \tilde{\boldsymbol{T}}_n = \boldsymbol{t}) = \binom{\binom{n}{2}-n+1}{m-n+1}^{-1}$ and the denominator does not involve Π. It's worth mentioning that π from $\mathbb{P}(\Pi = \pi \mid \tilde{\boldsymbol{T}}_n = \boldsymbol{t})$ is the same as sampling uniformly π from $\mathrm{hist}(\boldsymbol{t})$. The second step is carried out by iteratively sampling a new parent for each of the nodes. We leave it to the reader to look up details of the last two statements from [CX23, Sections 4.1, 4.2].

6 Appendix

1. Here is a proof of Theorem 6. Note that $p_u \coloneqq \mathbb{P}(\Pi_1 = u \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g}) = \mathbb{P}(\Pi_1 = \tau(u) \mid \tilde{\boldsymbol{G}}_n = \tau\boldsymbol{g})$. This implies $p_u \geq p_v \iff p_{\tau(u)} \geq p_{\tau(v)}$. By construction of B_ε in Section 4, it follows that $B_\varepsilon(\tau\boldsymbol{g}) \overset{d}{=} \tau B_\varepsilon(\boldsymbol{g}) \forall \tau$. Now let $\rho : [n] \xrightarrow{\sim} \mathcal{U}_n$ be such that $\boldsymbol{G}_n^* = \rho_* \boldsymbol{G}_n$ and $\Lambda : \mathcal{U}_n \xrightarrow{\sim} \mathcal{U}_n$ be random. Let $\Pi = \Lambda\rho$ and $\tilde{\boldsymbol{G}}_n = \Gamma_* \boldsymbol{G}$. Then $\mathbb{P}[\rho_1 \in B_\varepsilon(\boldsymbol{G}_n^*)] = \mathbb{P}[(\Lambda\rho)_1 = \Lambda(\rho_1) \in \Lambda B_\varepsilon(\boldsymbol{G}_n^*)] = \mathbb{P}[\Pi_1 \in B_\varepsilon(\tilde{\boldsymbol{G}}_n)] \geq 1 - \varepsilon$ where the last inequality is true because $\boldsymbol{P}[\Pi_1 \in B_\varepsilon(\tilde{\boldsymbol{G}}_n) \mid \tilde{\boldsymbol{G}}_n = \boldsymbol{g}] \geq 1 - \varepsilon$ for any $\mathcal{U}_n$-labelled graph $\boldsymbol{g}$.
2. Some notation:
   - All graphs are undirected. $V(\cdot), E(\cdot)$ represent the vertices and edges respectively.
   - $f : A \xrightarrow{\sim} B$ means a bijection in case of sets and an isomorphism in case of graphs.
   - $[n] = \{1, \cdots, n\}$ and $\mathbb{N} = \{1, 2, 3, \cdots\}$.

References

[CX21] Harry Crane and Min Xu. "Inference on the History of a Randomly Growing Tree". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.4 (July 2021), pp. 639–668. ISSN: 1369-7412. DOI: 10.1111/rssb.12428. eprint: https://academic.oup.com/jrsssb/article-pdf/83/4/639/49320709/jrsssb_83_4_639.pdf. URL: https://doi.org/10.1111/rssb.12428.

[CX23] Harry Crane and Min Xu. *Root and community inference on the latent growth process of a network*. 2023. arXiv: 2107.00153 [stat.ME].

[Met23] Nilava Metya. *Root inference on a growing network by Crane and Xu*. 2023. URL: https://youtu.be/l2dJ6WUBLzs.