# Phase Transition in the Recoverability of Network History

Jean-Gabriel Young[1,2,3,*], Guillaume St-Onge,[1,2] Edward Laurence,[1,2] Charles Murphy,[1,2]
Laurent Hébert-Dufresne,[1,4,5] and Patrick Desrosiers[1,2,6]

[1]*Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec G1V 0A6, Canada*
[2]*Centre Interdisciplinaire de Modélisation Mathématique de l'Université Laval,
Québec G1V 0A6, Canada*
[3]*Center for the Study of Complex Systems, University of Michigan, Michigan 48109, USA*
[4]*Department of Computer Science, University of Vermont, Vermont 05405, USA*
[5]*Vermont Complex Systems Center, University of Vermont, Vermont 05405, USA*
[6]*Centre de Recherche CERVO, Québec G1J 2G3, Canada*

Network growth processes can be understood as generative models of the structure and history of complex networks. This point of view naturally leads to the problem of network archaeology: reconstructing all the past states of a network from its structure—a difficult permutation inference problem. In this paper, we introduce a Bayesian formulation of network archaeology, with a generalization of preferential attachment as our generative mechanism. We develop a sequential Monte Carlo algorithm to evaluate the posterior averages of this model, as well as an efficient heuristic that uncovers a history well correlated with the true one, in polynomial time. We use these methods to identify and characterize a phase transition in the quality of the reconstructed history, when they are applied to artificial networks generated by the model itself. Despite the existence of a no-recovery phase, we find that nontrivial inference is possible in a large portion of the parameter space as well as on empirical data.

## I. INTRODUCTION

Unequal distributions of resources are ubiquitous in the natural and social world [1]. While inequalities abound in many contexts, their impact is particularly dramatic in complex networks, whose structures are heavily constrained in the presence of skewed distributions of resources, such as edges or cliques [2]. For instance, the aggregation of edges around a few hubs determines the outcome of diseases spreading in a population [3], the robustness of technological systems to targeted attacks and random failures [4], or the spectral property of many networks [5]. It is therefore not surprising that much effort has been devoted to understanding how skewed distributions come about in networks. Many of the satisfactory explanations uncovered thus far have taken the form of constrained growth processes: the rich-get-richer principle [6], sampling space reduction processes [7], and latent fitness models [8].

A common characteristic shared by these processes is that they do not—nor are they expected to—give a perfect

*[*]jgyou@umich.edu

account of real complex systems and networks [9]. Their rules are simple and only capture the essence of the mechanisms at play, glossing over details [10]. But despite these simplifications, growth processes endure as useful models of real complex systems. At the level of macroscopic distributions, their predictions have often been found to fit the statistics of real networks to surprising degrees of accuracy [11]. At the level of detailed mechanisms, they have been shown to act effectively as generative models of complex networks [12,13], i.e., as stochastic processes that can explain the minutia of a network's growth [14,15]. This point of view has led, for example, to powerful statistical tests that can help determine how networks evolve and change [16,17].

The notion of growth processes as generative models is now being pushed further than ever before [17]. The burgeoning field of network archaeology [18], in particular, builds upon the idea that growth processes are generative models of the history of complex networks, able to reveal the past states of statically observed networks. This point of view is perhaps the most clearly stated in the bioinformatics literature, which seeks to reconstruct ancient protein-protein interaction (PPI) networks to, e.g., improve PPI network alignment algorithms [19,20] or understand how the PPI networks of organisms are shaped by evolution [21]. Indeed, almost all algorithmic solutions to the PPI network archaeology problem are based on explicit models of network growth (variations on the

duplication-divergence principle) and take the form of parsimonious inference frameworks [21–23]; greedy local searches informed by models [18,24–26]; or maximum likelihood inference of approximative [27], graphical [20], and Bayesian [28] models of the networks' evolution.

Less obvious is the fact that a second body of work, rooted in information theory and computer science, also makes the statement that growth processes can generate the history of real complex networks. This second strand of literature [29–36] focuses on temporal reconstruction problems on treelike networks generated by random attachment processes [6,37]. It has led to efficient root-finding algorithms (whose goal is to find the first node) [29–32] and to approximate reconstruction algorithms on trees [33–35]. Applying any of these algorithms to a real network amounts to assuming that growth processes—here, random attachment models—are likely generative models.

The goal of this paper is to investigate classical growth processes as generative models of the histories of networks, from the point of view of Bayesian statistics and hidden Markov processes. This investigation is made possible by recent advances in particle filtering methods as applied to temporal reconstruction from static observations [28,38,39]. Our contribution is threefold. First, we give a latent variable formulation of the network archaeology problem for a generalization of the classical preferential attachment (PA) model [6,40–42]. We derive all the tools necessary to infer history using the model, including a sampling algorithm for its posterior distribution adapted from Ref. [38], provably optimal estimators of the history, as well as efficient heuristics well correlated with these estimators. Second, we establish the extent to which complete history recovery is possible and, in doing so, identify a phase transition in the quality of the inferred histories (i.e., we find a phase where recovery is impossible and a phase where it is achievable in large networks). Third, we demonstrate with numerical experiments that we can extract temporal information from a statically observed network not generated by the model, including an aging model [43] and the phylogenetic tree of the Ebola virus. We conclude by listing a number of important open problems.

## II. METHODS FOR NETWORK ARCHAEOLOGY

### A. The problem

A network $G$ generated by a growth process is, by construction, associated with a history $X$, i.e., a series of events that explains how $G$ evolved from an initial state $G_0$. We consider the loosely defined goal of reconstructing $X$, using the structure of $G$ and the fact that it came from a growth process as our only sources of information. Formally, this is an estimation problem in which the history $X$ is a latent variable, determined by the structure of the network (see Fig. 1). The relationship between the network and its history is expressed using Bayes' formula as



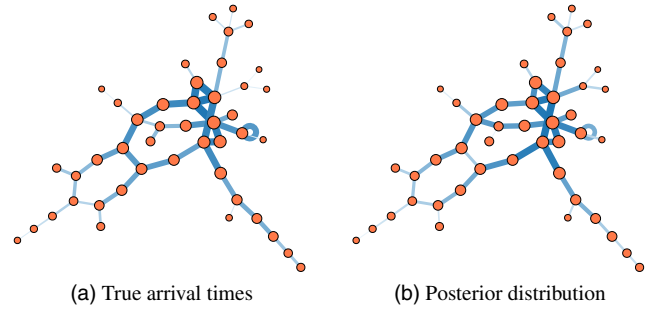(a) True arrival times    (b) Posterior distribution

FIG. 1. Reconstructing the history of a growing network. (a) Artificial network generated by our generalization of the preferential attachment model (with parameters $\gamma = -1.1$, $b = 0.9$, $T = 50$; see text). Since the network is artificial, its true history—i.e., the time of arrival of its edges in time—is known. The width and color of edges encode this history; older edges are drawn with thick, dark strokes, while younger edges are drawn using thin, light strokes. The age of the nodes is encoded in their radius. Our goal is to infer this history as precisely as possible, using the (unlabeled) network structure as our only input. (b) Accurate reconstruction obtained with $10^5$ samples of the posterior distribution over possible histories.

$$P(X|G, \theta) = \frac{P(G|X)P(X|\theta)}{P(G|\theta)}, \qquad (1)$$

where we assume, for the sake of simplicity, that the growth process parameters (denoted by the tuple $\theta$) can be estimated reliably and separately from $G$ (see the Appendix A, where we test this assumption).

A complete specification of the probabilities appearing in Eq. (1) is obtained upon choosing a growth process as our model: This choice fixes the unconditional probability $P(X|\theta)$ of the histories, the likelihood $P(G|X)$ of the network given $X$, and the evidence $P(G|\theta)$ given as the sum of $P(G|X)P(X|\theta)$ over all histories $X$.

### B. Random attachment model

For the sake of concreteness, we carry out our analysis of the network archaeology problem in the context of a specific growth process. We use a variant of the classic PA model that incorporates both a nonlinear attachment kernel [41] and densification events, i.e., attachment events between existing nodes [40,42,44,45].

#### 1. Model description

In this model, a new undirected edge is added at each time step $t = 1, 2, \ldots, T-1$, starting from an initial network $G_0$ comprising a single edge. With probability $1 - b$, the new edge connects two existing nodes, and it connects an existing node to a new node with complementary probability $b$. Whenever an existing node $i$ needs to be selected, it is chosen randomly with probability proportional to $k_i^\gamma(t)$, where $k_i(t)$ is its degree at time $t$ (prior to any modifications to the network) and where $\gamma$ is called the

exponent of the attachment kernel. Hence, the probability of choosing node $i$ when an existing node is needed is

$$u_i(\gamma) = \frac{k_i^\gamma(t)}{\sum_{j \in V_t} k_j^\gamma(t)} = \frac{k_i^\gamma(t)}{Z(t;\gamma)}, \qquad (2)$$

where $V_t$ is the set of nodes in $G$ prior to adding the new edge.

The parameter $b \in [0,1]$ controls the density, while $\gamma \in \mathbb{R}$ controls the strength of the rich-get-richer effect. We refer to these parameters collectively with $\theta = (\gamma, b)$. We recover the classic PA model by setting $(\gamma = 1, b = 1)$ [6,46]; the random attachment model with $(\gamma = 0, b = 1)$ [37]; one of the models of Aiello, Chung, and Lu with $\gamma = 1, b \in [0,1]$ [45,47]; and an undirected version of the Krapivsky-Redner-Leyvraz generalization if $\gamma$ is free to vary and $b = 1$ [41]. The model technically generates multigraphs for any $b < 1$, although numerical experiments show that the proportion of redundant edges and self-loops decreases rapidly with increasing network size, for all $b > 0$, when $\gamma < 1$. It is thus an appropriate model of sparse multigraphs, but also a reasonable approximation of large sparse networks, with few or no redundant edges and self-loops.

### 2. Posterior distribution over histories

The posterior distribution over histories $P(X|G, \gamma, b)$ is proportional to the product of two terms; see Eq. (1).

The first term, the likelihood $P(G|X)$, only weakly depends on the details of the model. Its role is only to enforce consistency: It equals 1 if $X$ is a possible history of $G$ under the model (in which case, we say that $X$ is consistent with $G$), and it equals 0 otherwise. In *any* model where growth events consist of attaching new edges or nodes to an already existing network, a consistent history $X$ is one that describes a sequence of connected subgraphs of $G$, each a subgraph of the next. Our model is one such process, which settles the issue of calculating $P(G|X)$.

The second term, the probability of a history $P(X|\gamma, b)$, is more sensitive to the choice of model. We notice that the growth model defined above is a Markov process because the probability of growth events is defined in terms of the current state of the network and nothing else. Hence, $P(X|\gamma, b)$ can be written as a product of transition probabilities. Denoting by $X_t$ the state of the history at time $t$, we write $P(X|\gamma, b) = \prod_{t=1}^{T-1} P(X_t|X_{t-1}, \gamma, b)$. Combining Eq. (2) with the fact that a new node is involved in the growth event with probability $b$, we find that the transition probability is given by

$$P(X_t|X_{t-1}, \gamma, b) = \xi \cdot b \cdot u_{v_1}(\gamma) + (1 - \xi)$$
$$\cdot (1 - b) \cdot u_{v_1}(\gamma) \cdot u_{v_2}(\gamma), \qquad (3)$$

where $v_1$ is the existing node and $v_2$ is the incoming node or the node chosen to close a loop. In this equation, we use

$\xi = 1$ to indicate that $v_2$ is new in the transition $X_{t-1} \to X_t$, and we set $\xi = 0$ otherwise.

Using Eq. (1), we can then write the posterior probability $P(X|G, \gamma, b)$ of a history $X$ as

$$P(X|G, \gamma, b) = \frac{\prod_{t=1}^{T-1} P(X_t|X_{t-1}, \gamma, b)}{P(G|\gamma, b)} \mathbb{I}[X \in \Psi(G)], \qquad (4)$$

where $\mathbb{I}[X \in \Psi(G)]$ is an indicator function equal to 1 if $X$ is in the set $\Psi(G)$ of histories consistent with $G$ (and 0 otherwise), and where the normalization is given by a sum over histories in $\Psi(G)$,

$$P(G|\gamma, b) = \sum_{X \in \Psi(G)} P(X|\gamma, b). \qquad (5)$$

### C. Inference: Goal and algorithms

In the latent variable formulation of the network archaeology problem introduced in Eqs. (1)–(4), reconstructing the past amounts to extracting temporal information from $G$ via the posterior distribution $P(X|G, \gamma, b)$. We need to set our goals carefully, however, since not all problems of this type are solvable. For example, the posterior distribution in Eq. (4) is heavily degenerate—and even uniform over the set of all histories consistent with $G$ [36] for some choices of parameters $\gamma$ and $b$ (see Appendix B for details). As a consequence, an attainable goal cannot be to find the one true history $\tilde{X}(G)$ of $G$ because this history is generally not identifiable [36].

To find a better inference task, we notice that, according to the model, every growth event marks the arrival of precisely one new edge (see Sec. II B). Consequently, we can represent histories $X$ compactly as an ordering of the edges of $G$ in discrete time $t = 0, \ldots, T - 1$. And, in turn, this representation suggests a natural inference task, which we will henceforth adopt as our inference goal: estimating the *individual* arrival times $\tau(e)$ of the edges $e \in E(G)$ of the network. We can hope to get good estimates in this case because we know of a number of network properties that correlate with the age of nodes and edges in growth models [48].

One possible estimator $\hat{\tau}(e)$ of the arrival time of edge $e$ is the posterior average:

$$\hat{\tau}(e) = \langle \tau(e) \rangle = \sum_X \tau_X(e) P(X|G, \gamma, b), \qquad (6)$$

where $\tau_X(e)$ denotes the arrival time of $e$ in history $X$. It is straightforward to show that $\langle \tau(e) \rangle$ minimizes the expected mean-squared error (MSE) on $\tau(e)$, and we therefore refer to it as the MMSE estimator of the arrival time. It turns out that this estimator also maximizes the correlation of the full set of estimates $\{\hat{\tau}(e)\}$ and the true arrival times, when $G$ is generated by the model (see Appendix C for a proof)—it is therefore *optimal* in some sense.

Since we are working with a distribution over histories, we can do much more than simply estimate $\langle \tau(e) \rangle$. One posterior estimate is particularly informative: the variance on $\tau(e)$, calculated as

$$\sigma^2(e) = \langle \tau(e)^2 \rangle - \langle \tau(e) \rangle^2$$
$$= \sum_X \tau_X^2(e) P(X|G, \gamma, b) - \left( \sum_X \tau_X(e) P(X|G, \gamma, b) \right)^2. \tag{7}$$

It can tell us how much we should trust our estimate of $\tau(e)$. A small variance means that the (unimodal) marginal distribution $P(\tau(e) = t|G, \gamma, b)$ is peaked on a few values of time $t$, i.e., that we should be confident of our estimate of $\tau(e)$. But conversely, if it is large—say, in the extreme case of a uniform marginal distribution over $t = 0, \ldots, T - 1$—then we do not know much about $\tau(e)$ and our estimate should not be trusted. In the applications of Sec. III, we quantify our uncertainty by calculating this variance alongside our estimate. To summarize this information in a single number, we compute the normalized variance per edge,

$$U(G) = \frac{12}{T(T^2 - 1)} \sum_{e \in E(G)} \sigma^2(e), \tag{8}$$

where the leading factor bounds $U(G)$ in the range [0, 1], with 1 corresponding to the maximal overall variance, i.e., maximal uncertainty.

We are not ready to move on to applications, however, as the computation of a complete set of MMSE estimators and the associated uncertainty score $U(G)$ is unfortunately intractable. Explicit summation is impossible because there are far too many histories consistent with networks of even moderate size (the upper bound $|\Psi(G)| = T!$ holds, sometimes tightly). In general, we cannot exploit some special symmetries of $G$ to evaluate the sum since the network is an input of the problem and therefore arbitrary. Hence, we have to resort to approximations, which we now introduce.

### 1. Sequential importance sampling

Following the standard practice in Bayesian statistics, we use a Monte Carlo approximation to evaluate the MMSE estimators as

$$\hat{\tau}(e) \approx \frac{1}{n} \sum_{i=1}^{n} \tau_{x_i}(e), \tag{9}$$

where $x_i$ is a random history drawn from the posterior distribution $P(X|G, \gamma, b)$. In theory, the error on the average decreases rapidly as $O(1/\sqrt{n})$, such that we can calculate $\hat{\tau}(e)$ to a good approximation quite easily.

Things are not so simple in practice, however, because it is hard to sample from the posterior distribution

$P(X|G, \gamma, b)$ directly. The consistency constraint $X \in \Psi(G)$ depends on the minutia of the structure of $G$ and makes generating samples from $P(X|G, \gamma, b)$ a difficult endeavor. For this reason, and following earlier work on network history sampling [49,50], we prefer a simple transformation of Eq. (9) that allows for more straightforward sampling.

The transformation relies on the introduction of a second distribution $Q(X|G)$ over the consistent histories $\Psi(G)$. The idea is to reexpress the MMSE estimators as

$$\hat{\tau}(e) = \sum_X \tau_X(e) P(X|G, \gamma, b)$$
$$= \sum_{X \in \Psi(G)} \frac{\tau_X(e) P(X|\gamma, b)}{P(G|\gamma, b)} \frac{Q(X|G)}{Q(X|G)}$$
$$= \frac{\langle \tau_X(e) \omega(X|G, \gamma, b) \rangle_Q}{P(G|\gamma, b)}, \tag{10}$$

where the average is now computed over $Q$ and where $\omega(X|G, \gamma, b) = P(X|\gamma, b)/Q(X|G)$ is called the (unnormalized) weight of history $X$. Equation (10) is a useful reformulation of the average in that the distribution $Q(X|G)$ is now arbitrary. Hence, in particular, we are free to choose a distribution that is easy to sample, which allows us to evaluate Eq. (10) as

$$\hat{\tau}(e) \approx \frac{1}{nP(G|\gamma, b)} \sum_{i=1}^{n} \tau_{x_i}(e) \omega(x_i|G, \gamma, b), \tag{11}$$

where the set of $n$ histories $\{x_i\}$ is now drawn from the distribution $Q$, known in this context as the proposal distribution [51]. We can safely ignore the intractable normalization $P(G|\gamma, b)$ introduced by the transformation because the estimators $\langle \tau(e) \rangle$ must satisfy the sum

$$\sum_{e \in E(G)} \langle \tau(e) \rangle = \sum_{e \in E(G)} \sum_{X \in \Psi(G)} \tau_X(e) P(X|G, \gamma, b)$$
$$= \sum_{X \in \Psi(G)} P(X|G, \gamma, b) \sum_{e \in E(G)} \tau_X(e)$$
$$= \sum_{X \in \Psi(G)} P(X|G, \gamma, b) \sum_{t=0}^{T-1} t = \binom{T}{2}, \tag{12}$$

where the last equality follows from the normalization of $P(X|G, \gamma, b)$. As a result, we may compute $\hat{\tau}(e)$ up to a multiplicative constant and use Eq. (12) to set the global scale.

It is advantageous to choose a Markov process as the proposal distribution $Q$, i.e., one that factorizes as $Q(X|G) = \prod_{t=1}^{T-1} Q(X_t|X_{t-1}, G)$, as we then obtain what is known as a sequential importance sampling (SIS) method [52]. The method is said to be "sequential" because all computations can now be done on the fly. We generate $X_t$

from $X_{t-1}$ with $Q(X_t|X_{t-1},G)$ and update the sample weight as

$$\omega(X_{0:t}|G,\gamma,b) = \omega(X_{0:t-1}|G,\gamma,b)\frac{P(X_t|X_{t-1},\gamma,b)}{Q(X_t|X_{t-1},G)}, \quad (13)$$

where $X_{0:t} = \{X_0, X_1, ..., X_t\}$ refers to a history, with all states included up until time $t$. Neither the transition $X_{t-1} \to X_t$ nor the weight update equation makes use of old states. Hence, after the weights have been updated, we can simply discard $X_{t-1}$ and save memory—an important benefit when the network $G$ is large.

With the above considerations in mind, we propose to use a variation on snowball sampling [53,54] as the proposal distribution $Q(X|G)$. Informally, a snowball sample mimics the growth process itself, by enumerating the edges of $G$, radiating outward from a random starting point $e_0$ (the seed). We determine the next edge $e_1$ of the history by drawing uniformly at random from the set of edges that share at least one node with $e_0$ (excluding $e_0$ itself). The next edge after that is picked from the set of edges that share at least one node with $e_0$ *or* $e_1$ (excluding $e_0$ and $e_1$) and so on until the graph is exhausted.

We can give a more formal definition of $Q(X|G)$ by defining the *boundary* $\Omega(X_t)$ of $X_t$ as the set of all edges that share at least one node with edges already appearing in $X_t$ but that are themselves not in $X_t$. With this notation, a snowball sample is obtained by repeatedly drawing an edge from $\Omega$ uniformly at random and updating $\Omega$ accordingly. It is then easy to see that the transition probability associated with this proposal distribution is

$$Q_{\text{sb}}(X_t|X_{t-1},G) = [|\Omega(X_t)|]^{-1}, \quad (14)$$

with the convention that $Q(X_0|X_{-1},G)=1/|\Omega(X_0)|=1/|E|$.

Our choice of proposal distribution is motivated by the fact that (i) it is a Markov process such that we can use SIS, (ii) it only generates histories that are consistent with $G$, and (iii) transitions $X_t \to X_{t+1}$ can be computed efficiently. This choice leads to an overall sampling algorithm that is itself efficient. The worst-time complexity of generating one sample is $O(|E| \times k_{\max})$, where $k_{\max}$ is the maximal degree of $G$ (due to the boundary updates). When this maximal degree is a slowly varying function of $|E|$, as is the case for a large portion of the parameter space [41], snowball sampling generates samples in near linear time in the number of edges—as fast as possible.

### 2. Sequential Monte Carlo algorithm

The SIS method described above is not perfect, however. The error on the true average, for example, no longer decreases as $O(1/\sqrt{n})$ because we are not sampling from the true posterior distribution anymore. To get an intuition as to why, notice that probability mass is conserved, such that any proposal distribution $Q(X|G)$ that is not equal to

the posterior itself must, by necessity, place a low probability on some histories that have a high posterior probability. As soon as we generate one of these rare samples, the sum in Eq. (11) becomes dominated by a single term because of its large weight $P(X|G,\gamma,b)/Q(X|G)$, which essentially washes out the contribution of other terms.

The impact of these high weight samples on our estimates can be quantified with the effective sample size [55]

$$\text{ESS}(\{x_i\}|G,\gamma,b) := \frac{[\sum_{i=1}^{n} \omega(x_i|G,\gamma,b)]^2}{\sum_{i=1}^{n} \omega(x_i|G,\gamma,b)^2}. \quad (15)$$

An ESS close to $n$ tells us that all samples contribute roughly equally, while an ESS close to 0 tells us that we only have a few useful samples at hand. It turns out that for problems with the structure of network archaeology, the ESS of a population of samples generated with a SIS algorithm will go to 0 unless $Q$ is extremely close to $P$ [52]. Hence, the SIS algorithm introduced above tends to evaluate its estimators with very few effective samples. A natural extension of SIS, called the adaptive sequential Monte Carlo (SMC) algorithm, is designed to address this problem [52].

In a SMC algorithm, one still generates samples using the imperfect proposal distribution $Q$, but this is now done *in parallel*. In other words, we first pick $e_0$ for a set of $n$ histories, then $e_1$ for all these histories, and so on, all the while updating the weights with Eq. (13). We could do the same with the SIS algorithm, so the two algorithms are not truly different in this regard. The defining difference between SIS and SMC comes from the way we handle these samples. Denote the set of parallel histories evolved up until step $t$ as $H(t) = \{X_{0:t}^{(i)}\}_{i=1,...,n}$ and the associated set of weights by $W(t) = \{\omega_i(X_{0:t}^{(i)})\}_{i=1,...,n}$. As we evolve the parallel histories, we monitor the ESS as a function of $t$, using Eq. (15) and the partial weights $W(t)$. Whenever the ESS becomes too small and crosses a threshold ESS*, we perform an additional *resampling* step. The goal of this additional step is to eliminate histories that look like they will not contribute much to our estimators. It is implemented by creating a new set $H'(t)$ of uniformly weighted histories from $H(t)$, obtained by randomly duplicating histories with probability proportional to their current weight. Following standard practice [52], we choose ESS* $= n/2$ as the threshold that triggers a resampling step, and we implement resampling by drawing $n$ indexes $\{a_i\}_{i=1,...,n}$ from the multinomial distribution of probabilities $\tilde{W} = \{\omega_i/\sum_j \omega_j\}_{i=1,...,n}$ and setting $H' = \{X_{0:t}^{(a_i)}\}$. One can show that this resampling step does not bias the estimators in the limit of large $n$ [52], which gives us a method to obtain unbiased estimators of $\tau(e)$ calculated with a high ESS.

It is important to realize that while the resampling step increases the ESS by design, it does so at a cost. A history

that evolved from an unlikely starting point or in an unlikely direction has a significant probability of getting overwritten. Erasing this history can be a "mistake" in that it could eventually evolve to a high weight state, which we never get to see. We can never make such a mistake when the number of samples $n$ is extremely large, as a few low weight histories will survive repeated resampling. But for finite $n$, this effect—known as path degeneracy [52]—can actually lead to poor inference results in practice. We somewhat mitigate path degeneracy by resampling only when the ESS becomes small, as opposed to at every step as is done recently in Ref. [38]. Even then, numerical experiments suggest that the downsides of path degeneracy outweigh the benefit of an increased ESS on loopy networks and very heterogeneous trees (see Supplemental Material [56]). As a result, we use resampling only when $b = 1$ and $\gamma < 1$.

This resampling technique completes the set of methods needed to draw samples from the distribution over histories. The derivation is complex, but the end result is straightforward. To summarize, we (i) initialize $n \gg 1$ histories by setting their weights to 1 and by drawing, for each, a seed uniformly at random from the edge set $E(G)$. Then, we (ii) evolve the $n$ histories, in parallel, using the snowball proposal distribution of Eq. (14) to generate random moves and Eq. (13) to update the weights. When $b = 1$ and $\gamma < 1$, we (iii.a) keep track of the ESS with Eq. (15) and trigger a resampling step whenever it drops below $n/2$. In all other cases, we (iii.b) never resample, so we do not need to calculate the ESS. Once we have our final set of histories, we (iv) approximate $\langle \tau(e) \rangle$ for all edges using Eqs. (11) and (12) and our set of samples, and we calculate the uncertainty appearing in Eq. (8) in the same way. Our reference implementation of this method is freely available online [57].

### 3. Structural estimators

We put the SIS and SMC sampling algorithms to the test in Sec. III, but before we do, we introduce two last—much simpler—estimation algorithms as baselines. These algorithms only rely on the structure of $G$ to estimate the arrival times $\tau(e)$, instead of an explicit knowledge of the posterior distribution $P(X|G, \gamma, b)$. They follow the same overall pattern: We rank the edges in descending order, based on some network property $\mathcal{P}$ that is known to positively correlate with the age of edges, and we output these ranks as our estimated arrival times. Whenever the edges of a subset $S \subseteq E$ are indistinguishable according to property $\mathcal{P}$, we cannot order them reliably with $\mathcal{P}$, so we instead give them the same rank $\lambda(S)$. To set $\lambda(S)$, we require that the average time of arrival $\langle \tau \rangle$ be preserved; this constraint forces $\lambda(S) = t + (m + 1)/2$, where $m = |S|$.

Our first structural estimator is based on the observation that the nodes that arrive earlier in a growth process have, on average, a larger degree [6,48]. This result is a consequence of the fact that older nodes have many more

opportunities to acquire new neighbors as the growth process unfolds than nodes that arrived at the very end. Because we want to order edges and not nodes, we use the degree of nodes to induce a ranking of edges as follows. We define $(k_e^{\text{low}}, k_e^{\text{high}})$ as the degree of the nodes connected by edge $e$, with $k_e^{\text{low}} \leq k_e^{\text{high}}$. We rank edges in descending order of $k_e^{\text{high}}$ and break ties with $k_e^{\text{low}}$. The idea behind this strategy is that an edge connected to at least one high-degree node is likely to be older than an edge connected to two nodes of a lower degree.

We also know that the model generates networks that nucleate from a core, such that their central nodes tend to be older. Hence, our second structural estimator makes use of a centrality measure to order edges. We use a recursive peeling method known as the onion decomposition (OD) [58], where we create a sequence of nested subnetworks by repeatedly removing the nodes of the lowest degree. At step $t$ of this peeling process, all nodes with degree $k_{\min}$, the current lowest degree, are removed simultaneously and assigned a layer number. We turn these numbers into the time of arrival of nodes by assuming that nodes in the outermost layers appeared last. A simple modification allows the algorithm to order edges: An edge is assigned to a class as soon as one of its nodes is peeled away. All edges removed in the same pass are declared as tied. We note that OD is closely related to the peeling method introduced in Ref. [35] to tackle the archaeology problem in the special case ($\gamma = 1$, $b = 1$), although the method of Ref. [35] removes the lowest-degree nodes without batching, which leads to a slightly different ordering.

## III. RESULTS

### A. Inference on artificial networks

To calibrate the methods and understand the conditions under which they perform well, we first apply our algorithms to networks drawn from the generative model itself. In this situation, we know the ground truth $\tilde{X}$ and therefore the true arrival times $\tau_{\tilde{X}}(e)$ of all the edges. As a result, we can compute the quality of our estimates $\{\hat{\tau}(e)\}$ with the Pearson product-moment correlation as

$$\rho = \frac{\sum_{e \in E(G)} (\hat{\tau}(e) - \langle \tau \rangle)(\tau_{\tilde{X}}(e) - \langle \tau \rangle)}{\sqrt{\sum_{e \in E(G)} (\hat{\tau}(e) - \langle \tau \rangle)^2} \sqrt{\sum_{e \in E(G)} (\tau_{\tilde{X}}(e) - \langle \tau \rangle)^2}},$$

(16)

where $\langle \tau \rangle = (T - 1)/2$ is the average arrival time, which is fixed by the choice of timescale. This correlation takes values in $[-1, 1]$, where $|\rho| = 1$ indicates a perfect recovery up to a time reversal and where $|\rho| = 0$ indicates that no information is extracted from the graph at all. It is not affected by an arbitrary linear transformation of the timescales, it penalizes spurious ordering of tied events, and it is robust to small perturbations of the estimators [59].
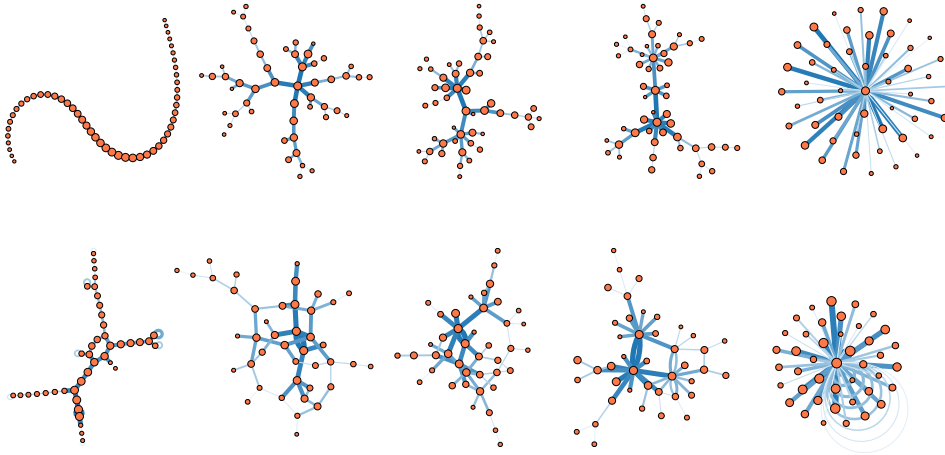
FIG. 2. Network zoo. Examples of networks generated by the process with $b = 1$ (top row) and $b = 0.75$ (bottom row), and $\gamma \in \{-10, -1, 0, 1, 10\}$ (from left to right). The width and color of edges encode this history; older edges are drawn with thick, dark strokes, while younger edges are drawn using thin, light strokes. The age of the nodes is encoded in their radius.

### 1. Inference on artificial trees

We consider the regime $b = 1$, i.e., the regime where the model generates trees, for our first set of experiments. The generative model is well understood in this case, which will help us interpret the inference results more readily. In Ref. [41], it is shown that different values of $\gamma \in \mathbb{R}$ correspond to different phases that are characterized by different degree distributions (see also Fig. 2, upper row). In the limit $\gamma \to -\infty$, the model generates long paths, where every node has degree 2 except for the two end nodes, of degree 1. For all negative values of $\gamma$, the model favors homogeneous degrees. When $\gamma = 0$, the degree distribution is geometric, of mean 2 (since we recover the uniform attachment model [37]). In the interval $0 < \gamma < 1$, the degree distribution takes the form of a stretched exponential, with an asymptotic behavior fixed by $\gamma$. At precisely $\gamma = 1$, the attachment kernel becomes linear, and the networks are scale-free: The degree distribution follows a power law of exponent $-3$ [6]. In the interval $1 < \gamma < 2$, the networks *condensate* in a rapid succession of connectivity transitions at $\gamma_m = (m + 1)/m$ for $m \in \mathbb{N}^*$. When $\gamma > \gamma_m$, the number of nodes of degree greater than $m$ becomes finite. As a result, an extensive fraction of the edges aggregates around a single node—the condensate—and this fraction grows with increasing $\gamma$ [60]. The condensation is complete at $\gamma = 2$, where the model enters a winner-takes-all scenario characterized by a central node that monopolizes nearly all the edges.

The average of the correlation attained by our various inference algorithms is shown as a function of the attachment kernel $\gamma$ in Fig. 3(a), on small networks ($T = 50$). As previously stated in Sec. II, we assume that the parameters are known when we compute the MMSE estimators, and we therefore use the true values of $(\gamma, b)$ in our calculations.

We distinguish two broad regimes based on the inference results: the regime $\gamma < 0$, characterized by a homogeneous
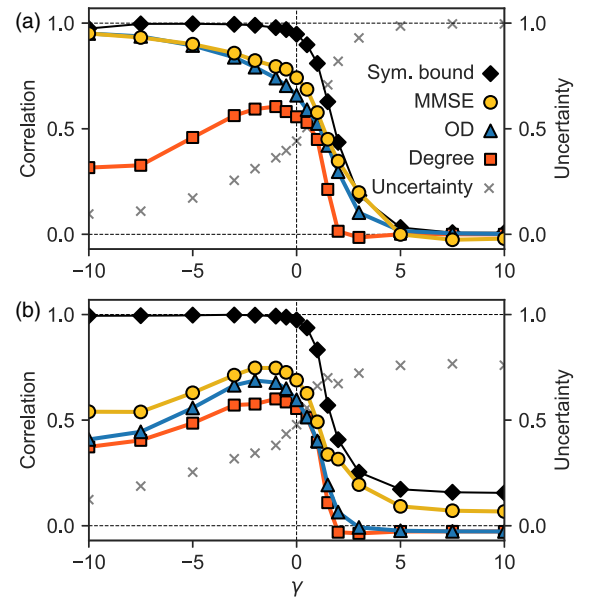


FIG. 3. Effect of the rich-get-richer phenomenon on recovery. We analyze artificial networks of $T = 50$ edges, generated with our generalization of the preferential attachment and the parameters $\gamma \in [-10, 10]$, (a) $b = 1$ (trees), and (b) $b = 0.75$ (loopy networks). We plot the average correlation attained by the Monte Carlo approximation of the minimum mean-squared error (MMSE) estimators and two efficient methods based on network properties (a degree-based method and the onion decomposition [58]). A loose upper bound that accounts for network symmetry is also shown [see Eq. (17)] as well as a measure of our uncertainty of the MMSE estimators (right axis) [see Eq. (8)]. Each point is obtained by averaging the results over $m$ different network instances, where ($b = 1$) $m = 40$ and ($b = 0.75$) $m = 250$. We use the true parameters ($\gamma$, $b$) and $n$ Monte Carlo samples to approximate the MMSE estimators, where $n = 10^5$ ($b = 1$) and $n = 5 \times 10^6$ ($b = 0.75$).

distribution of degrees, and the heterogeneous regime $\gamma > 0$. The three methods behave similarly in the first regime: They first yield a relatively large correlation at $\gamma = 0$, but their quality quickly plummets with growing $\gamma$, ultimately converging to a null average correlation for sufficiently large values of $\gamma$. The MMSE estimators remain slightly superior to the OD estimators throughout, and they both outperform the degree estimators by a significant margin. The degree estimators perform even worse in the homogeneous regime. While the quality of the OD and MMSE estimators increases with decreasing $\gamma$, the correlation achieved by the degree estimators goes in the opposite direction and shrinks with $\gamma$. From these results, we can already draw the conclusion that the MMSE and OD estimators should be preferred on trees—MMSE for their accuracy and OD for their speed. [It can be shown that the OD algorithm returns its final point estimates in $O(|E| \times \log|V|)$ steps [58], whereas a single Monte Carlo sample takes roughly as many steps to generate.]

It is natural to ask whether these results are good. To answer this question, we also show in Fig. 3(a) two measures that can help us assess the results.

The first, shown as gray crosses in the figure (right axis), is the uncertainty $U(G)$ defined in Eq. (8). It correctly increases as our ability to infer correlated histories decreases, reaching its maximal value of 1 in the condensed phase $\gamma > 2$ where inference does not seem possible.

The second, drawn in black, shows an upper bound on the average correlation as a function of $\gamma$. This bound takes the symmetries of the generated graphs into account and shows that they can seriously hamper our ability to conduct network archaeology.

Before we describe how this bound is derived, it is helpful to understand where symmetries come from and why they matter. Notice that some edges fulfill the same structural role in a network, say, the three edges of a triangle or the first two edges of the small network $E = \{(a, b), (a, c), (b, c), (a, d)\}$. If we were to draw these networks twice with different layouts and the labels removed, we would not be able to tell which of these edges is which. In other words, some edges can only be identified because we have labels on the nodes.

Since labels represent an arbitrary choice [61], estimators cannot rely on them to make predictions. As a consequence, structurally equivalent edges are impossible to order reliably. Hence, the more equivalent edges there are, the harder the archaeology problem becomes.

Now, to actually compute the symmetry bound, we first find all the structurally equivalent edges in the generated graph $G$, using a method discussed in Appendix D. We then construct an estimator $\tau^*(e)$ of the arrival time of $e$ by averaging the true time of arrival $\tau_{\tilde{X}}(e)$ of all edges in its equivalence class, as

$$\tau^*(e) = \frac{1}{|C(e)|} \sum_{e' \in C(e)} \tau_{\tilde{X}}(e'), \tag{17}$$

where $C(e)$ is the set of edges indistinguishable from $e$. We finally compute the correlation between $\{\tau^*(e)\}$ and the ground truth, using Eq. (16). The resulting bound corresponds to the correlation we would have obtained had we known the true arrival time of edges, without the labeling of $G$.

There is no reason to believe that we can recover such a precise temporal reconstruction from $G$ alone, which means that the bound is probably loose. That said, as shown in Fig. 3(a), it does a good job of explaining the maximal correlation attained in the extreme regimes—our estimators perform as well as possible when $|\gamma| \gg 0$. In the large positive-value regime, the networks condensate, and symmetries upper bound the correlation at 0. In the large negative-value regime, the networks are effectively grown as random paths, where all nodes are of degree 2 except the two end nodes, which are of degree 1 (see Fig. 2). All edges are thus ordered up to a mirror symmetry around the middle, such that the equivalence classes are of size 2. Standard concentration inequalities then tell us that the time of arrival of any edge can be identified with a variance that vanishes in the large-$T$ limit. Near-perfect recovery is therefore trivial: Peeling the path symmetrically from both sides yields a close approximation of the arrival time of every edge.

### 2. Inference on artificial loopy networks

Figure 3(b) shows the outcome of the same experiments, in a case where loops are allowed ($b = 0.75$). The phenomenology of the generative model is not the same as in the case of trees—we do not know that there are sharp structural connectivity transitions at many values of $\gamma$, for example. That said, the same general principle still holds (see Fig. 2, bottom row): Large positive values of $\gamma$ still mean that the network condensates on a few nodes, and negative values of $\gamma$ lead to a homogeneous distribution of degrees.

Comparing the results of Fig. 3(b) with the case of trees shown in Fig. 3(a), we find a number of noteworthy differences. The most noticeable differences perhaps concern the homogeneous regime $\gamma \ll 0$: We find that near-perfect recovery is no longer possible, that the symmetry bound is much looser, and that the average uncertainty is not predictive of the reconstruction accuracy anymore. Other important differences include an increased gap between the quality of the MMSE estimators and the structural methods (OD, degree) for all $\gamma$ and the fact that nontrivial inference remains possible deep into the condensation phase $\gamma \gg 0$ with MMSE estimators.

Starting with this last difference, let us analyze the condensation phase of the generative model with loops ($b < 1$), which bears a strong resemblance to the analog phase in the case of trees ($b = 1$). A typical network realization in this regime is comprised of the following: many self-loops centered on the condensate, a number of

parallel edges connecting high-degree nodes, and starlike node arrangements around high-degree nodes. In the regime $\gamma \gg 0$, in particular, the typical network becomes a star with $(1 - b)|E|$ self-loops on average (see bottom-right diagram of Fig. 2). These self-loops help the degree of the condensate grow faster (incrementing its degree by 2 every time instead of 1), which leads to a more pronounced condensation for fixed $T$, $b < 1$, and $\gamma > 1$. One might be tempted to conclude that as a result, inference becomes harder as we decrease $b$, but this would be discounting the fact that self-loops carry some temporal information. Indeed, the MMSE estimators achieve a slightly positive correlation by exploiting the difference between the self-loops and the spokes of the star in the condensation regime. Because the growth process starts from a single edge and a geometric distribution of mean $1/(1 - b)$ determines the time step at which the first self-loop is created, it is possible to obtain a positively correlated time of arrival by guessing that self-loops are slightly younger than the spokes, on average. Neither the degree nor the OD estimators can detect this difference, and they therefore declare all edges as tied.

A different phenomenon explains the disappearance of near-perfect recovery in the regime $\gamma \ll 0$ when we set $b < 1$. The large gap between the symmetry bound and the best inference results shows that the symmetries are not at fault: One can clearly distinguish every edge, yet inference is still difficult. We argue that the poor performance of the estimators is instead imputable to the appearance of random long-range connections not found in other regimes. One such edge appears when two low-degree nodes, typically located in the outermost layers of the network, are chosen as the end points of a new connection—an event that is only possible when $\gamma \ll 0$ and $b < 1$. These connection in turn (i) increase the number of histories consistent with $G$ and (ii) introduce uncertainties in the ordering of large subsets of edges, for example, when a long-range connection closes a long path. The inference problem becomes harder as a result.

The correlation attained by the MMSE in the regime $\gamma \ll 0$, $b < 1$ is probably not too far from its optimum—despite what the symmetry bound says. The uncertainty estimates $U(G)$ are small, which tells us that the MMSE estimators are as precise as one could have hoped. Hence, a bound of a completely different nature—one that accounts for cycles—would be needed to explain the diminishing correlation as $\gamma$ goes to negative infinity. We leave the issue of finding this bound to future work.

### 3. Phase transitions in heterogeneous networks

In the previous sections, we have shown that when we apply our methods to networks $G$ drawn from the generative model itself, the history encoded in the network's structure can be recovered to varying degrees of accuracy, depending on the exponent value of $\gamma$. Focusing on the

heterogeneous regime $\gamma > 0$, we have seen in Fig. 3 that robust inference only seems possible when $\gamma$ is positive but close to 0. These results are suggestive of a phase transition in our reconstruction capabilities—although we cannot yet reach any conclusions because our analysis has so far been limited to small networks of $T = 50$ edges, fraught with possible finite-size effects. Therefore, to get a better numerical portrait of the dependence of the attained correlation on $\gamma$, we now turn to large networks. Our goal is to uncover a single critical threshold $\gamma_c$ marking the onset of this transition. We define the threshold precisely as follows: On one side of $\gamma_c$, there exists an algorithm that returns estimators $\{\hat{\tau}(e)\}$ attaining a nonvanishing average correlation with the ground truth in the limit of large network sizes, while there is no such method on the other side of the divide, in the no-recovery phase.

To find the location of $\gamma_c$, we run a finite-size scaling analysis of the correlation attained by different methods, on increasingly larger networks generated at different values of the structural transitions $\gamma_m = (m + 1)/m$ [41]. First, we apply the OD method to these networks. Our goal is to find a value of $\gamma_m$ for which the correlation attained by OD is independent of $T$. If we find one such $\gamma_m$, then we have evidence for a *lower* bound on $\gamma_c$ since we then know of at least one method (OD) that returns correlated estimates in the large-network limit. Second, we compute the scaling of the symmetry bound for the same networks. If we can find another value of $\gamma'_m$ for which the bound goes to zero as $T$ increases, then we also have an *upper* bound on $\gamma_c$ since no method can outperform the symmetry bound by definition. Hence these two methods combined can help us bracket $\gamma_c$.

The outcomes of these experiments are shown in Fig. 4. We find that for most values of $\gamma > 1$, the average correlation attained by the OD decreases as $T^{-\delta(\gamma)}$, with $\delta(\gamma) > 0$. If $\gamma$ is close enough to 1, however, the average correlation becomes independent of $T$ (verified for network sizes up to $T \leq 3 \times 10^5$). In the case of trees ($b = 1$), we find that the correlation is independent of $T$ when $\gamma = 11/10$, but that is not the case for $\gamma = 5/4$, which gives us a lower bound of $\gamma_c > 11/10$. In the case of loopy graphs with $b = 0.75$, we find a smaller lower bound of $\gamma_c > 1$, with the correlation decreasing slowly for kernel exponents as small as $\gamma = 11/10$.

Interestingly, the upper bound on $\gamma_c$ provided by the second scaling analysis [Fig. 4(c)] shows that the OD results are nearly optimal: When the symmetry bound decreases, the correlation attained by OD also decreases. Conversely, when the symmetry bound stays steady, the correlation attained by OD becomes independent of $T$. Hence, while we have not computed the scaling analysis for all values of $\gamma$, our results suggest that $\gamma_c$ is bounded away from—but close to—$\gamma = 1$.

This raises the question of where exactly the critical threshold $\gamma_c$ lies. As we have mentioned previously, the generative model has infinitely many connectivity
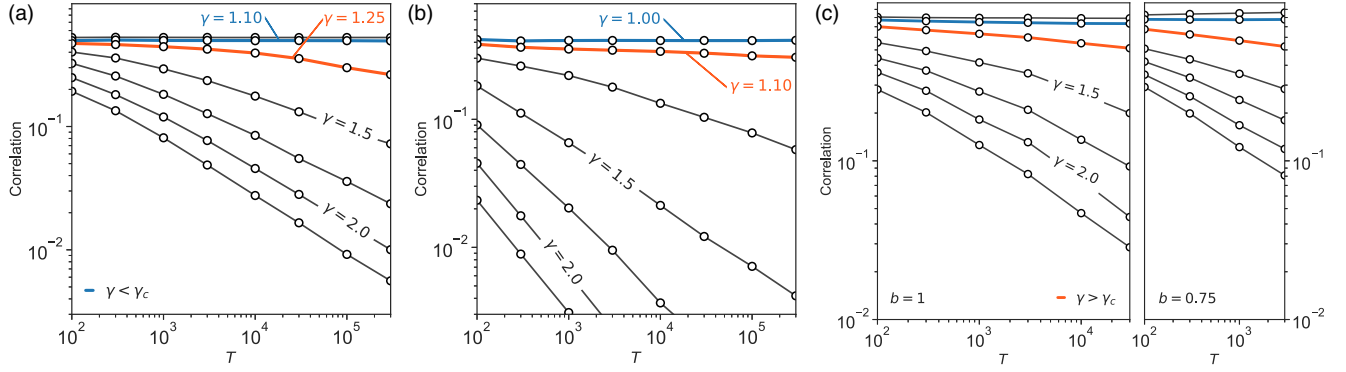
FIG. 4. Finite-size scaling analysis for the results of Fig. 3. The average correlation attained by OD is shown against the size $T$ of networks generated using (a) $b = 1$ and (b) $b = 0.75$, for $\gamma \in \{1.00, 1.10, 1.25, 1.50, 1.75, 2.00, 2.25\}$ (from top to bottom). Two scaling curves are highlighted and annotated in each figure: the curve associated with the largest value of $\gamma$ that does not decline with $T$ (blue), and the first curve that varies with $T$ after that (orange). (c) Scaling of the symmetry bound for the same values of $\gamma$ in the case of trees (left) and dense graphs (right). The scaling is not computed for large values of $T$ due to computational costs. Note that there are many known structural transitions between the highlighted curves [41], which are not shown for the sake of clarity.

transitions in the case of trees, at $\gamma_m = (m + 1)/m$ for $m = 1, 2, \ldots$ [41]. Our numerical results suggest that there is a *single* important value of $\gamma$ that matters in the infinite size limit; as such, our parsimonious hypothesis is that one of the critical values $\gamma_m$ aligns with $\gamma_c$ when $b = 1$. Combining this simple observation with our numerical bounds leaves $m = 5, \ldots, 10$ as options when $b = 1$ and all values of $\gamma \in (1, 1.10)$ when $b = 0.75$.

The appearance of a condensate from $\gamma > 1$ onwards gives a nice qualitative explanation as to why there should be a phase transition in the recovery quality. When an edge attaches to the condensate, the temporal information it carries becomes inaccessible. Furthermore, because these edges are added throughout the growth process, any estimation technique that tries to find a total ordering will conflate old and new edges in a single class. The diminishing correlation of the estimators in the regime $\gamma > 1$ is hence, at least in part, attributable to the presence of this condensate.

To quantify the impact of equivalent edges on the structure, we run a second scaling analysis and verify how the average information content (IC) of the generated networks scales with network size. The IC is, in a nutshell, an information theoretic quantity that measures the prevalence of equivalent edges [62]—the same edges we have used to define the symmetry bound. It gives us a single number that summarizes the abundance but also the heterogeneity in size of the sets of equivalent edges. It is defined as

$$S(G) = -\sum_{i=1}^{q} \frac{|C_i|}{|E|} \log \frac{|C_i|}{|E|} = \log |E| - \frac{1}{|E|} \sum_{i=1}^{q} |C_i| \log |C_i|,$$
(18)

where $C_1, \ldots, C_q$ are the $q < |E|$ sets of equivalent edges [63] (see Appendix D for how to find these edges). If all the sets of equivalent edges are finite, then the information content of $G$ is of order $\log |E|$. Conversely, if one extensive

set accounts for the totality of edges—e.g., when $G$ is a star graph—then $S(G)$ is zero. In general, if there are $\ell$ extensive sets accounting for a nonvanishing fraction $\alpha = \alpha_1 + \cdots + \alpha_\ell$ of all edges, then $S(G) \approx (1 - \alpha) \log |E| - \sum_i \alpha_i \log \alpha_i$. Hence, the scaling of $S(G)$ with $|E|$ tells us how fast new sets of distinguishable edges are created as the network grows. What we want to verify is whether good performance correlates with the presence of many distinguishable sets of edges, i.e., a large IC.

Our results are shown in Fig. 5(a). The scaling behavior of the IC confirms that there is an extensive number of equivalence classes when $\gamma = 1$. Figure 5(b) shows the difference between the true information content and the information content obtained by assuming that the equivalence classes of edges are determined *only* by the degree of the nodes at the end of the edges (a coarsening of the true equivalence classes). Because the difference is close to zero for high values of $\gamma$, this second figure tells us that most of the equivalence classes are degree classes in this regime. The figure also tells us that many new equivalence classes are created as $\gamma$ approaches 1, precisely in the regime where OD does well. Coupled with Fig. 4, Fig. 5 shows that an abundance of equivalent edges—specifically those of the condensate—drives the recoverability transition.

### B. A different task: Root finding

In Sec. II C, we briefly mentioned that inferring the complete history of a network is only one of many possible problems that fits within the Bayesian formulation of network archaeology. Any other temporal inference task that makes use of the posterior distribution $P(X|G, \gamma, b)$ can be solved with the same set of tools—like evaluating the uncertainty $U(G)$. As a further example of the versatility of this framework, we now briefly turn to another problem: finding the first edge (root) of $G$ [32].
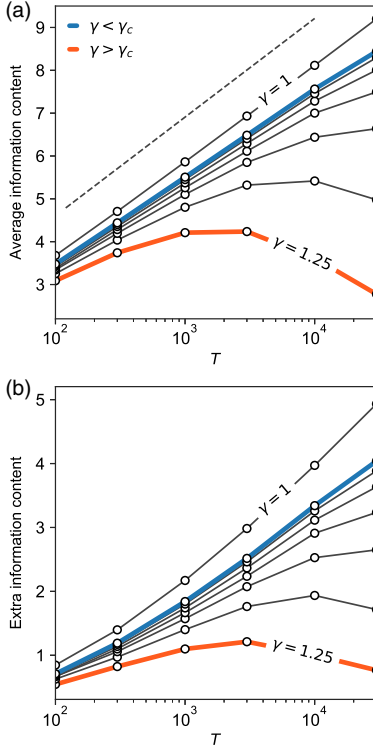
(a)

(b)

FIG. 5. Finite-size scaling of the average information content. (a) Average information content $\langle S(G) \rangle$ of tree networks ($b = 1$), as a function of network size $T$, for exponents $\gamma \in \{\frac{5}{4}, \ldots, \frac{11}{10}, 1\}$ of the nonlinear kernel. Color matches what is highlighted in Fig. 4(a). We know that these are upper and lower bounds on $\gamma_c$ due to the scaling analysis. The dotted line is the upper bound $\log(T)$ on the information content. (b) Information content not accounted for by the degree classes, i.e., $\langle S(G) - S_{\deg}(G) \rangle$, where $S_{\deg}(G)$ is the Shannon entropy of the partition obtained by classifying edges according to their nodes' degree (see text).

The most comprehensive analysis of a root-finding algorithm is put forward by the authors of Ref. [29], where the goal is to find the first *node* of a growing tree in the case $\gamma = 0$ and $\gamma = 1$. Their strategy is to compute the number $\varphi(v) = |\{X | \tilde{\tau}_X(v) = 0\}|$ of histories rooted on $v$ and to return the $K$ nodes with the largest $\varphi(v)$. They show that this algorithm can be employed to construct sets of constant size that contain the root with a fixed error rate $\varepsilon < 1$ as $T$ goes to infinity and that the case $\gamma = 0$ is easier than the case $\gamma = 1$ (smaller sets are needed to attain the same error rate $\varepsilon$). In line with this reference and following Refs. [29–31], we can give a solution to the problem in terms of sets—with the crucial difference that we look for the first edge instead of the first node.

We use the marginal distribution $P(X | G, \theta)$ over time steps $P(\tau(e) = t | G, \gamma, b)$ to construct the set $\mathcal{R}$ of likely roots. More precisely, we use the SMC sampler to approximate the probability that an edge $e$ is the first,

$$P(\tau(e) = 0 | G, \gamma, b) = \sum_X \mathbb{I}[\tau_X(e) = 0] P(X | G, \gamma, b), \quad (19)$$
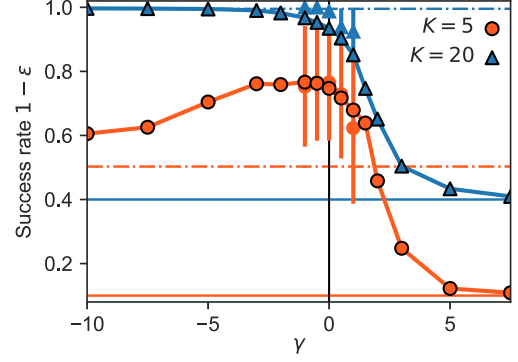


FIG. 6. Root finding on artificial trees. We show the success rate of the root-finding algorithms, with sets $\mathcal{R}$ of sizes $K = 5$ and $K = 20$ on artificial networks of $T = 50$ edges. The results of OD are shown with solid lines and symbols, while the sampling results are shown with symbols and error bars of 1 standard deviation. The horizontal solid lines show the accuracy of randomly constructed sets $\mathcal{R}$ (no information retrieval), while the horizontal dotted lines show the expected success rate in the limit $\gamma \to -\infty$, where a simple peeling technique is optimal. The sampled root sets are computed for $\gamma \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$, using $n = 10^5$ samples.

where $\mathbb{I}[S]$ is an indicator function, which is equal to 1 if the statement $S$ is true and equal to 0 otherwise. We then define $\mathcal{R}$ as the set formed by the $K$ edges that have the largest posterior probability $P(\tau(e) = 0 | G, \gamma, b)$, which gives us our prediction. For comparison, we also infer the root with the much-faster onion decomposition by constructing $\mathcal{R}$ with the $K$ most central edges (with ties broken at random).

The accuracy of the resulting algorithms is shown as a function of $\gamma$ in Fig. 6. We distinguish, again, two main regimes: Accurate recovery is possible in the strongly homogeneous regime $\gamma \ll 0$, but the success rate diminishes with growing $\gamma$, reaching a noninformative limit in the regime $\gamma \gg 0$. This result shows that, much like full temporal reconstruction, root finding is also negatively affected by the presence of equivalent edges.

It is worth noting that our results (Fig. 6) put the work of Refs. [31,64] in the broader context of Bayesian inference. Their counting strategy, based on calculating the number $\varphi(v)$ of histories rooted on $v$, for example, can be formally related to our posterior inference formulation via

$$\varphi(v) \propto \sum_{X \in \Psi(X)} \mathbb{I}[\tilde{\tau}_X(v) = 0] P(X | G, \gamma) \equiv P(v \text{ is first} | G, \gamma, b)$$

because the posterior distribution $P(X | G, \gamma, b)$ is uniform over all histories (see Appendix B for a proof) in the case studied in those works, defined by $b = 1$ and $\gamma = 0, 1$. Thus, the estimators proposed in Refs. [29,30,32] can be seen as outputting the $K$ nodes that have the largest marginal distribution at $t = 0$—assuming a uniform distribution over histories. In contrast, our method remains correct for arbitrary values of $\gamma$ (a generalization suggested

in Ref. [29]) but also for any choice of $b$. It can be easily extended to seeds that are not edges but instead small subgraphs [31]—we simply have to change the initial distribution over states used in the SMC sampler.

## C. Application: Effective modeling and the phylogenetic tree of the Ebola virus

The end goal of network archaeology is to uncover temporal information from statically observed networks not explicitly generated by growth processes. It was recently shown that it can sometimes be difficult to tell growth processes apart, even when perfect temporal data are available [17]. Turning this observation on its head, there are situations where the details of the growth process do not matter much—many mechanisms can explain the same data equally well. One consequence of this finding is that our generalization of PA—and many more models—could find applications as effective models of the growth of real networks, without getting all the details right. In other words, we can hope to make reasonable temporal inference even if we use an otherwise simplistic growth process as our model.

To put this hypothesis to the test, we first use our methods to reconstruct the past of an artificial model that is *not* generated by our generalization of preferential attachment. To this end, we consider a growth model that combines preferential attachment and added memory effects [43]. At each time step, a new node is added to the network, and it chooses its neighbor with probability proportional to

$$u_i(\alpha) = k_i(t)(t - \tau(v_i) + 1)^{-\alpha}, \qquad (20)$$

where $k_i(t)$ is, again, the degree of node $v_i$ at time $t$, $\tau(v_i)$ is its time of arrival, and $\alpha \in \mathbb{R}$ is a parameter of the model. Much like our generalization of PA, this model has a rich phenomenology (see the top row of Fig. 7). The classical preferential attachment model [6] is recovered by setting $\alpha = 0$. Older nodes are chosen preferentially in the regime $\alpha < 0$, whereas the newer ones are preferred in the regime $\alpha > 0$. This effect is the strongest in the regime $\alpha \geq 1$, where the network is no longer scale-free and tends to organize in long chains as $\alpha$ goes to infinity [43].

Our inference results are shown in Fig. 7, where we apply our methods to a few typical networks generated with this model. Since the parameters $(\gamma, b)$ are not known in this case, we estimate them with a method detailed in Appendix A—this method allows us to compute the probability $P(X|G, \gamma, b)$ when we approximate the MMSE estimators. For the network generated with $\alpha = -2$, we find that the best fit is given by the exponent $\hat{\gamma} = 1.76$. In other words, in this case, aging helps nodes accumulate neighbors faster than PA would, which leads to networks reminiscent of those generated by generalized PA in the superlinear regime of Ref. [41]. As a result, even though our model is, strictly
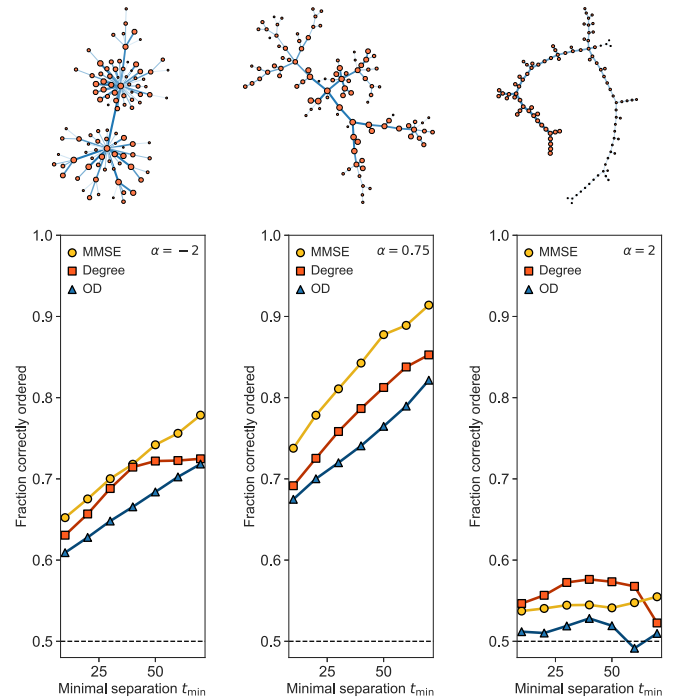


FIG. 7. Application to synthetic networks grown with a model that includes memory effects. (Top row) Single network instances with various values of the memory parameters $\alpha$ [43]. Edge thickness and node size encode age according to the ground truth (thicker and bigger equals older). Older nodes have a large probability of attracting new neighbors for the network shown on the left (memory parameter $\alpha = -2$), while newer nodes are advantaged in the other two cases ($\alpha = 0.75$ and 2). The negative effect of memory on attachment probability is strongest for the network shown on the right, with $\alpha = 2$. We find that these networks are best modeled with $\hat{b} = 1$ and, from left to right, $\hat{\gamma} = 1.76, 0.00, -1.61$ for the growth model (see Appendix A for details). The KS statistics $D^*$ on our estimates of $\gamma$ are $D^* = 0.044, 0.055, 0.052$, and the significance levels are $P(D > D^*) = 0.47, 0.45, 0.39$, signaling a good fit (see Appendix A for details). The uncertainty scores are $U(G) = 0.24, 0.39$, and 0.22 [see Eq. (8)]. (Bottom row) Fraction of edge pairs correctly ordered by the estimators, when separated by at least $t_{min}$ time steps according to the ground truth, for the single instances of the model shown in the top row. Ties are broken at random for the degree and OD estimators. MMSE estimators are calculated from $n = 200\,000$ Monte Carlo samples, using the estimated parameters.

speaking, the wrong one, it can still give useful inference since the mechanisms behave similarly (with the difference that there is now more than one condensate node). For the network generated with $\alpha = 0.75$, we find that the best fit is given by the exponent $\hat{\gamma} \approx 0$. Older nodes now fade out of memory so fast that the beneficial effect of PA barely allows them to attract new neighbor. This case leads to an attachment that is now more or less uniform, which again allows us to calculate useful estimates with an effective model.

The case of $\alpha = 2$ is interesting—it shows that we should be careful when we interpret the results of temporal

inference, in the situation where we know nothing of the growth mechanism that actually produced the network. When $\alpha$ is large, new nodes are moved out of memory immediately. As a result, the model generates long chains similar to those generated by our model when we set $b = 1$ and $\gamma \ll 0$ (see Fig. 7, top right), best fitted by $\hat{\gamma} - 1.61$. A crucial difference, however, is that the oldest nodes are now concentrated on one side of the chain instead of in its center. Our model is a poor description of this process since it assumes that central nodes are the oldest, which leads us to estimates that are close to the random baseline (see Fig. 7).

With this cautionary tale in mind, we can apply network archaeology to real data—provided that we carefully choose a good growth model based on the data. An area where network archaeology could be useful is the study of how biological systems evolved. Biological analyses often call for an estimation of the sequence in which the different constituents emerged, based on the currently observed biological diversity. This information is typically encoded in phylogenetic trees that track how species evolved and diversified from common unobserved ancestors. Much work has gone into inferring the *structure* of these trees from current observations, but this structure does not necessarily tell us about the *ordering* of speciation events. Hence, as a proof of concept, we apply our method to the inferred phylogenetic tree of the Ebola virus for the 2013–2016 West African Ebola epidemic [65,66]. The relatively short duration of the epidemics and the lack of selective pressure on the virus means that Ebola underwent neutral evolution. As a result, new lineage of strains can sprout from new and old strains alike, such that a simple attachment process is a good model of growth in this case. The extensive coverage of the surveillance and sequencing effort for this epidemic [67] means that on top of the structure, we have access to temporal metadata that are a close approximation for strain emergence. In this case, we can treat the phylogenetic tree as ground truth; our goal is to find an ordering of the emergence of all strains consistent with the metadata.

In Fig. 8, we show that all the inference methods recover some level of temporal information; statistical inference, however, performs much better than the others, regardless of the measure of quality used. The naive estimators yield correlations of $\rho_{\text{degree}} = 0.152$ and $\rho_{\text{OD}} = 0.150$ with the known metadata, while we find $\rho_{\text{MMSE}} = 0.456$ with the sampling method (using 25000 samples). Furthermore, while all the methods resolve pairs of mutations separated by any number of time steps better than chance, the MMSE estimators outperform the other techniques. This performance gap is in part due to the presence of equivalent edges; the OD identifies only 26 sets of distinguishable edges, while the degree estimators identify 47. In contrast, the true MMSE estimators can order all pairs of edges that are not equivalent (there are 1588 sets of distinguishable edges), a property that is retained by the subsampled estimators.
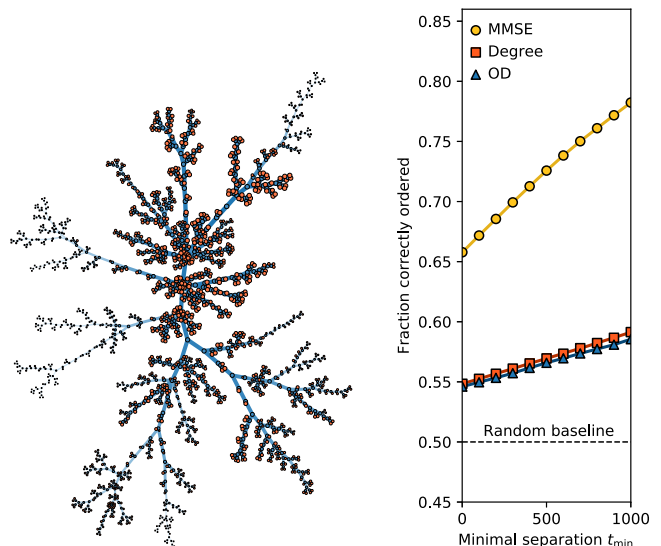


FIG. 8. Application to the phylogenetic tree of strains of the Ebola virus. (Left diagram) Leaves ($n = 1238$, in orange) represent strains of Ebola sequenced during the 2013–2016 West African outbreak [65], while the remainder of the nodes represent inferred common ancestors ($n = 959$, in blue), and edges are most likely mutations [66]. Edge thicknesses indicate age according to the ground truth. We find that this network is best modeled with $\hat{b} = 1$ and $\hat{\gamma} = -0.71$, associated with a KS statistic of $D^* = 0.17$ and a significance level of $P(D > D^*) = 0.53$ under the random model, signaling a good fit (see Appendix A for details). The uncertainty score of the MMSE estimators is equal to $U(G) = 0.001$ [see Eq. (8)]. (Right diagram) Fraction of edge pairs correctly ordered by the estimators, when separated by at least $t_{\text{min}}$ time steps according to the metadata. Ties are broken at random for the degree and OD estimators. MMSE estimators are calculated with the $n = 25\,000$ Monte Carlo samples.

Similar analyses are not possible for most epidemics due to a lack of data. For epidemics that are more sparsely monitored than the 2013–2016 Ebola outbreak, reconstruction typically relies on models that are hard to parametrize (transmission rates, mutation rates, demographics, etc.) [68]. The quality of our results suggests that reconstructing the history of phylogenetic trees should be one of the interesting avenues for future network archaeology, especially since the method does not rely on parameters that require additional data sources.

## IV. CONCLUSION

In this paper, we have addressed the network archaeology problem from the point of view of Bayesian inference, with generalized preferential attachment as our generative model. In doing so, we have shown that the equivalent edges that appear in random growing trees make inference difficult, to the point where inference becomes impossible in some regimes. This difficulty does not mean that the problem is generally impossible, however, since we

have also shown that reconstruction is possible in both artificial and real systems.

The opportunities brought about by network archaeology are tantalizing. In bioinformatics alone—the only field where it has found widespread use thus far [28,69]—network archaeology with the divergence-duplication models has already yielded insights into the past states of real PPI networks [18,21] and improved on network alignment [19,20]. Generalization to models that are relevant beyond bioinformatics will allow us to answer new questions about the past of statically observed systems and improve network analysis techniques [70].

Our paper shows an example of how to carry this analysis almost automatically. One simply has to define a Markov growth process; our adaptive sequential Monte Carlo algorithm—an extension of the algorithms proposed in Refs. [38,49,50]—can then generate weighted histories that can be aggregated as MMSE estimators to yield optimal estimators of the history of the network.

Our analysis is, of course, far from complete, and it leaves a number of important theoretical and computational problems open. First, while we have provided compelling evidence for the existence of a scalable inference phase and a no-recovery phase, we have yet to pinpoint the location $\gamma_c$ of the transition that separates them. Our numerical analysis suggests that it lies at some rational value $\gamma_c = (m+1)/m$ when $b = 1$, but finding the exact location will require further analytical work, perhaps relying on a detailed analysis of the expected information content of the generated graphs, for example, using the counting techniques introduced in Ref. [71]. Second, we have used correlation as our notion of inference quality, but a recent analysis of the special cases $\gamma = 0$, 1 [35] instead emphasizes trade-offs between precision—how many elements are correctly ordered—and density—how many elements we *can* order. It would be useful to study the full range of parameters $\gamma$ under this alternate definition of quality. Third, the phenomenology of the observed phase transition is strikingly similar to that observed in many disorder models [72,73]. While growth models are formally out-of-equilibrium processes—and thus cannot be obviously mapped onto disorder models—it will be important to establish how the network archaeology phase transition fits within the broader family of phase transitions in Bayesian inference problems. The tools introduced in Ref. [36], further work on information content, and the concept of exchangeability [74,75] all might offer insights into this issue. Fourth, we have shown that nonlinear preferential attachment can sometimes act as a useful effective model of a network's growth, even when the network is definitely not generated by this model. A systematic study of how generalizable these conclusions are could also be an interesting direction for future work. Finally, we have used the OD algorithm [58] in large networks because sequential Monte Carlo is, ultimately, not very scalable. We have shown that this

substitution does not work in all cases because it is based on a correlation between the true arrival times and the order of peeling of a network specific to the class of models studied. As a result, the next step for general network archaeology should be to derive efficient approximation methods that work with general models to allow for flexible network archaeology. These methods will have to handle models specified as chains $P(X|\theta)$ with some arbitrary notion of consistency $P(G|X)$. The relaxation technique of Ref. [76] for permutation inference comes to mind, but one could also consider the message-passing algorithm [77] and its dynamical variant [78].

## APPENDIX A: MODEL PARAMETERS

### 1. Estimation

Throughout the text, we assumed that the parameters $(\gamma, b)$ are known or can be estimated, which has allowed us to use the simpler conditional posterior $P(X|G, \gamma, b)$ (as opposed to a joint distribution over histories and parameters). We now give a method to calculate these parameters in the cases where they are not known (such as with a real graph; see Sec. III C).

Estimating $b$ is easy because the observed graph can be seen as the outcome of $|E| - 1$ independent and identically distributed Bernoulli trials of success probability $b$. Every excess edge beyond the minimum needed to ensure connectivity is seen as a failure, and the total number of edges gives the number of trials. After adjusting for the initial conditions, we obtain the unbiased estimator

$$\hat{b}(G) = \frac{|V(G)| - 2}{|E(G)| - 1}. \tag{A1}$$

The exponent $\gamma$ is, in theory, more difficult to estimate because the degree distribution is only a sufficient statistics for $\gamma$ once it is conditioned on the arrival times [79,80]. A principled estimation technique should therefore rely on a known history [13,81] or a joint sampling mechanism for $X$ and the parameters; see Ref. [38] for a general particle MCMC method. But as we show in Fig. 9, a simple non-Bayesian heuristic—the Kolmogorov-Smirnov (KS)
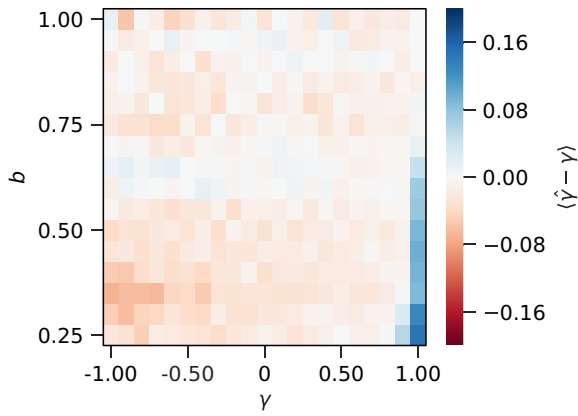
FIG. 9. Consistency of the estimator of $\gamma$ in small networks. We show the average difference between the true value and estimated value of $\gamma$, for networks of $T = 1000$ edges generated with parameters $\gamma \in [-1, 1]$ and $b \in [0.25, 1.00]$. This difference is averaged over 20 different network realizations at each point of the parameter space. We use $n = 1000$ samples each time we evaluate the average KS statistic.

minimization of Ref. [82]—yields accurate enough estimates from a single snapshot.

The KS statistic of a pair of distributions $(P, Q)$ is given by the supremum of the difference of their cumulative distribution function (CDF), i.e.,

$$D(P, Q) = \sup_k |f_P(k) - f_Q(k)|, \qquad (A2)$$

where $f_P(k)$ is the CDF of $P$ at point $k$. Given an empirical degree distribution $P(G)$ derived from an observed network $G$, we estimate $\gamma$ by minimizing the KS statistic averaged over a set of $n$ random degree distributions $\{Q^{(i)}(\gamma)\}_{i=1,...,n}$ generated by the model of parameters $(\gamma, \hat{b})$. The minimum $D^*(G)$ can be found efficiently using Brent's method [83] since the average KS statistic is convex. We use $n \gg 1$ random network instances to compute the average $D$ at each probed $\gamma$, which can be costly when $T$ is large. Therefore, in practice, we first compute the expected degree distribution of the model using mean-field equations and then draw $n$ finite samples from the resulting distribution. This approach is equivalent to—but much faster than—direct simulations. Note that the above framework also provides a natural (non-Bayesian) notion of goodness of fit for $\gamma$ [82]. It can be assessed by generating random degree distributions with the estimated parameters $(\hat{\gamma}, \hat{b})$, to which we apply the complete testing procedure. This method provides a null distribution for $D$, which tells us whether $D^*(G)$ is an extreme value of the average KS statistic or not. Following the standard, we assume that if $P[D > D^*(G)] > 0.1$, then the fit is good [82].

### 2. Sensitivity

To show that conditioning on point estimates of the parameter does not alter temporal inference significantly,

we run a sensitivity analysis in which we generate synthetic networks with some parameters $(\gamma, b)$, perturb the parameters $(\gamma, b) \to (\gamma', b')$, and then run the SMC algorithm with the perturbed parameters. We use Fig. 9 to select meaningful perturbation sizes: We investigate differences $\Delta_\gamma = \gamma' - \gamma$ that are much larger than the maximal error made in estimating $\hat{\gamma}$ (the error is overwhelmingly bounded to $\hat{\gamma} - \gamma \in [-0.05, 0.05]$ in Fig. 9). We also investigate large perturbations $\Delta_b$ to the probability $b$, but we note that large errors are unlikely since the standard deviation on the success rate of Bernoulli trials varies as $O(1/\sqrt{n})$.

The analysis, shown in Fig. 10, confirms that we can safely conduct temporal inference by conditioning on point estimates of the parameters.
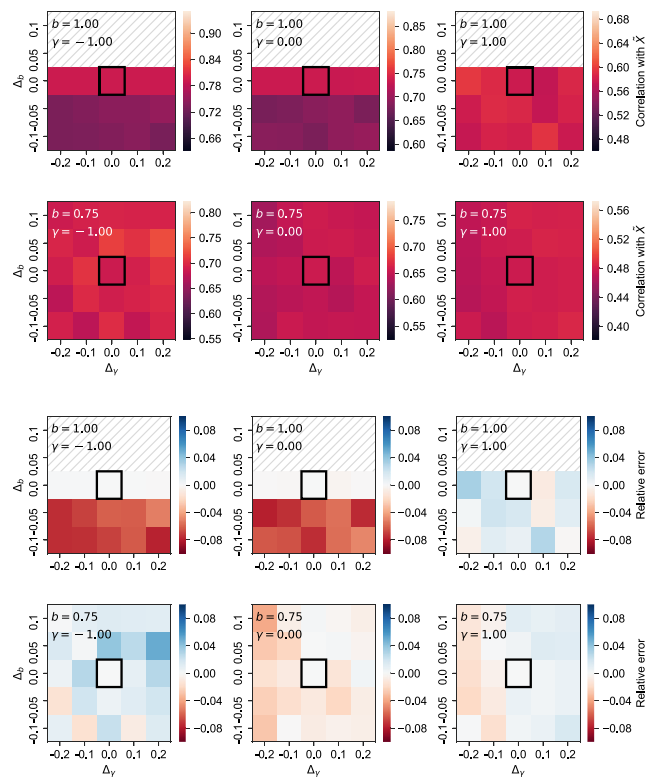


FIG. 10. Sensitivity analysis of the parameter estimators. Average correlation with the ground truth (top six panels) and relative error on the true correlation (bottom six panels) of the histories estimated by running the inference with misspecified parameters. These results are produced on trees ($b = 1$, top rows) and on loopy graph ($b = 0.75$, bottom rows), at various levels of heterogeneity ($\gamma \in \{-1, 0 + 1\}$ from left to right); see also inset text. The bold square shows the reference value (no perturbation), while the rest of the heat map shows results for absolute perturbations of magnitude $\Delta_\gamma, \Delta_b \in \{-0.2, -0.1, 0.1, 0.2\}$. Estimates above $\hat{b} = 1$ are impossible and therefore not computed (shown as a hatched region). We use the sampling level most appropriate for each case and $n = 100\,000$ samples, and we average the results over 35 network instances.

## APPENDIX B: MAXIMUM POSTERIOR MAXIMIZATION DOES NOT WORK

In the main text, we mention that the posterior distribution $P(X_{0:T-1}|G, \gamma, b)$ is often uniform over large sets of histories, and this epitomizes why posterior maximization (MAP) is not a suitable method to extract information from $G$. We now demonstrate that the Krapivsky-Redner-Leyvraz generalization of preferential attachment [41] is an extreme example of this problem. Namely, we explicitly show that when $\gamma \in \{0, 1\}$, the posterior distribution of this model is uniform over all histories in $\Psi(G)$ [84]. We then further argue that there exist large equivalence classes for general $\gamma \in \mathbb{R}$ and $b \in [0, 1]$.

### 1. Strict uniformity on trees

Recall from Eqs. (2) and (3) that in the model where $b = 1$, the logarithm of the unconditional probability of a history is given by

$$\log P(X_{0:T-1}|\gamma) = \sum_{t=1}^{T-1} \log u_{a_t}(\gamma, V_t),$$

$$u_i(\gamma, V_t) = \frac{k_i^\gamma(t)}{\sum_{j \in V_t} k_j^\gamma(t)},$$

where $V_t$ is the node set of $G_t$ prior to any modification of the graph's structure and where we have denoted by $a_t$ the node selected as the fixation site at time $t$ in $X_{0:T-1}$. To demonstrate that the distribution is uniform over the set of all consistent histories, we first define the normalization $Z(t; \gamma) = \sum_{j \in V_t} k_j^\gamma(t)$ and rewrite

$$\log P(X_{0:T-1}|\gamma) = \sum_{t=1}^{T-1} [\log k_{a_t}^\gamma(t) - \log Z(t; \gamma)]. \quad \text{(B1)}$$

Now, in the special cases of uniform attachment and linear preferential attachment, corresponding to $\gamma = 0$ and $\gamma = 1$, the normalization $Z(t; \gamma)$ *always* takes a special value independent from the actual content of $V_t$, namely,

$$Z(t; \gamma = 0) = \sum_{j \in V_t} k_j^0 = |V_t| = t + 1,$$

$$Z(t; \gamma = 1) = \sum_{j \in V_t} k_j = 2t.$$

The second identity follows from the fact that exactly one edge is created at each $t$ and that the sum of all degrees is always equal to twice the number of edges. These normalizations are independent of $X_{0:T-1}$, meaning that they can be dropped as an additive constant. Using $\gamma = 0$ and $\gamma = 1$ in Eq. (B1), we are left with

$$\log P(X_{0:T-1}|\gamma) \propto \begin{cases} \text{Constant} & \gamma = 0 \\ \sum_{t=1}^{T-1} \log k_{a_t}(t) & \gamma = 1. \end{cases} \quad \text{(B2)}$$

This last equation directly shows that the distribution $P(X_{0:T-1}|\gamma)$ (and therefore the posterior distribution) is uniform over all histories when $\gamma = 0$. Less obvious is the fact that the equation also implies a uniform posterior distribution in the case $\gamma = 1$. To see why, notice that the posterior distribution is obtained by conditioning on $G$ and that this restricts the possible histories to those in which a node $i$ of degree $k_i^*$ in $G$ appears $k_i^* - 1$ times in the sum $\sum_{t=1}^{T-1} \log k_{a_t}(t)$: once as a node of degree one, once as a node of degree 2, etc. Hence, every history consistent with $G$ is associated with some permutation of the same sum. Obviously, a permutation does not change the value of the sum; therefore, the posterior distribution is uniform over all histories consistent with $G$.

### 2. Extension to all parameters

Large sets of equally likely histories also arise in the more general attachment model on trees (i.e., when $\gamma \in \mathbb{R}$ with $b = 1$). The proof that these sets of histories exist is similar in spirit to that of the special cases above. We first make use of the permutation argument again, noting that it applies to the general sum $\sum_{t=1}^{T-1} \log k_{a_t}^\gamma(t)$, regardless of the value of $\gamma$. The problem therefore reduces to the study of the evolution of the normalization constant. Different from the special cases $\gamma = 0$ and $\gamma = 1$, the normalization $Z(t; \gamma)$ does not grow at the same rate for all histories when $\gamma$ is arbitrary. But, as we now show, there are still equivalence classes with respect to the posterior distribution. For example, consider two histories identical in all respects until a last node of degree $k$ and its $k - 1$ remaining neighbors are encountered. The $(k - 1)!$ histories resulting from the enumeration of this neighborhood will have, by construction, equivalent sequences of normalization constants $Z(1; \gamma) \rightarrow Z(2; \gamma) \rightarrow \cdots Z(T - 1; \gamma)$, which imply that these histories will be associated with the same posterior probability and that they will form a small equivalence class. Broader equivalence classes can be identified by noticing that similar permutations arise not only at the end but also at any point of the histories, and that they interact combinatorially: If there are $m$ such equivalent sets of edges, of sizes $k_1, \ldots, k_m$, then each different point of the posterior is degenerated $k_1! \times \ldots \times k_m!$ times. This argument trivially extends to any $b \in [0, 1]$.

## APPENDIX C: OPTIMALITY OF MMSE ESTIMATORS

The proof that the posterior average of $\tau_X(e)$ maximizes correlation *a posteriori* goes as follows.

By a slight abuse of notation, let us refer to an estimated history constructed with some arbitrary estimators

$\{\hat{\tau}(e)\}_{e \in E(G)}$ as $Y$, such that $\tau_Y(e) = \hat{\tau}(e)$. Then, assuming again that $X$ is drawn from the posterior distribution, the expected correlation of $X$ and $Y$ is

$$\langle \rho(X, Y) \rangle = \sum_{X \in \Psi(G)} P(X|G, \gamma, b) \rho(X, Y), \qquad \text{(C1)}$$

which we rewrite as

$$\frac{1}{\tilde{\sigma}_X \tilde{\sigma}_Y} \sum_{X \in \Psi(G)} P(X|G, \gamma, b) \sum_{e \in E(G)} (\tau_X(e) - \langle \tau \rangle)(\tau_Y(e) - \langle \tau \rangle)$$
$$= \frac{1}{\tilde{\sigma}_X \tilde{\sigma}_Y} \sum_{e \in E(G)} [\langle \tau(e) \rangle - \langle \tau \rangle][\tau_Y(e) - \langle \tau \rangle],$$

where $\tilde{\sigma}_X^2 := \sum_{e \in E(G)} (\tau_X(e) - \langle \tau \rangle)^2$ and $\tilde{\sigma}_Y^2$ are sums of individual variances [the same variances we use to calculate uncertainty $U(G)$, see Eqs. (7) and (8)]. We are able to take these standard deviations out of the sum because (1) $\tilde{\sigma}_Y$ is independent of $X$ and (2) the value of $\tilde{\sigma}_X$ is constant for all $X$ since one edge must occupy each "time slot" $t = 0, \ldots, T-1$ by definition, implying $\tilde{\sigma}_X^2 = \sum_{t=0}^{T-1} (t - \langle \tau \rangle)^2$. Again, somewhat stretching the notation, we define $Z$ as the history constructed with the MMSE estimators, i.e., the history such that $\tau_Z(e) = \langle \tau(e) \rangle$. We can then express the expected correlation compactly as

$$\langle \rho(X, Y) \rangle = \frac{\tilde{\sigma}_Z}{\tilde{\sigma}_X} \rho(Z, Y), \qquad \text{(C2)}$$

where $\tilde{\sigma}_Z \neq \tilde{\sigma}_X$, in general. Equation (C2) tells us that when histories are actually drawn from the posterior distribution, the expected correlation of the arbitrary estimators $\{\tau_Y(e)\}$ is proportional to the correlation between these estimators and the MMSE estimators. The expected overall correlation is therefore maximized if we choose $Y$ to be the MMSE estimators of the arrival times.

## APPENDIX D: FINDING EQUIVALENT EDGES

We gave an intuitive definition of equivalent edges in the main text. A more formal definition can be given by using the concept of orbits [85]. An orbit is a set of nodes that map onto themselves when we take an automorphism of $G$. Hence, for example, the three nodes of a graph comprising a single triangle form a single orbit. As another example, the two end nodes ($a$ and $c$) of the graph with edges $\{(a, b), (b, c)\}$ also form an orbit, while the central node $b$ forms an orbit by itself.

Our goal is, of course, to find the equivalent edges of $G$ and not the equivalent nodes, so we need to adapt the concept slightly. One can easily define "edge orbits" in the same way, but standard graph isomorphism software tends to deal with node orbits only [85]. Instead, we take a shortcut and resort to the line graph $L(G)$ of $G$. The line graph $L(G)$ is constructed by replacing every edge of $G$ by a node and by connecting two of these nodes if the corresponding edges share an endpoint in $G$. This transformation is useful because the indistinguishable edges of $G$ are mapped to the (node) orbits of $L(G)$ by construction since the $L(G)$ preserves the graph symmetries of $G$. This method gives a straightforward algorithm to find equivalent edges: We transform $G$ into a line graph $L(G)$, run a standard orbit identification algorithm on $L(G)$, and translate the result back to $G$.

We note that the definition of the line graph is ambiguous when $G$ has self-loops and parallel edges. (Should a self-loop be connected with itself in $L(G)$? What about sets of parallel edges: Should the copies of an edge be connected with the other copies?) Since our goal is to find the distinguishable edges of $G$, we must ensure that the transformation does not erase the important symmetries of $G$. With this goal in mind, we choose to (i) *not* connect the parallel edges of $G$ among themselves and (ii) add a self-loop to the nodes of $L(G)$ that stand in for self-loops in $G$. Choice (i) allows us to distinguish the parallel edges from simple edges. As an example, suppose we have a triangle with a parallel edge $\{(a, b), (a, b), (b, c), (c, a)\}$. The resulting line graph is a clique with a missing edge between the two copies of $(a, b)$, which allows us to disambiguate the parallel and simple edges. Choice (ii) allows us to tell self-loops apart from other edges. For example, if we have a star graph $G$ with added self-loops on the central nodes, then the line graph is again a clique, but the self-loops are marked as such—again allowing us to tell them apart from the spokes.

[1] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé, *Constrained Growth of Complex Scale-Independent Systems*, Phys. Rev. E **93**, 032304 (2016).

[2] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, and L. J. Dubé, *Structural Preferential Attachment: Stochastic Process for the Growth of Scale-Free, Modular, and Self-Similar Systems*, Phys. Rev. E **85**, 026108 (2012).

[3] G. St-Onge, J.-G. Young, E. Laurence, C. Murphy, and L. J. Dubé, *Phase Transition of the Susceptible-Infected-Susceptible Dynamics on Time-Varying Configuration Model Networks*, Phys. Rev. E **97**, 022305 (2018).

[4] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Network Robustness and Fragility: Percolation on Random Graphs*, Phys. Rev. Lett. **85**, 5468 (2000).

[5] C. Castellano and R. Pastor-Satorras, *Relating Topological Determinants of Complex Networks to Their Spectral Properties: Structural and Dynamical Effects*, Phys. Rev. X **7**, 041024 (2017).

[6] A.-L. Barabási and R. Albert, *Emergence of Scaling in Random Networks*, Science **286**, 509 (1999).

[7] S. Thurner, R. Hanel, B. Liu, and B. Corominas-Murtra, *Understanding Zipf's Law of Word Frequencies through Sample-Space Collapse in Sentence Formation*, J. R. Soc. Interface **12**, 20150330 (2015).

[8] G. Bianconi and A.-L. Barabási, *Bose-Einstein Condensation in Complex Networks*, Phys. Rev. Lett. **86**, 5632 (2001).

[9] H. A. Simon, *Models of Man; Social and Rational* (Wiley, New York, 1957).

[10] M. Mitchell, *Complexity: A Guided Tour* (Oxford University, New York, 2009).

[11] L. Hébert-Dufresne, E. Laurence, A. Allard, J.-G. Young, and L. J. Dubé, *Complex Networks as an Emerging Property of Hierarchical Preferential Attachment*, Phys. Rev. E **92**, 062809 (2015).

[12] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, *Microscopic Evolution of Social Networks*, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, Las Vegas, Nevada, 2008), pp. 462–470.

[13] V. Gómez, H. J. Kappen, and A. Kaltenbrunner, *Modeling the Structure and Evolution of Discussion Cascades*, in *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (ACM, Eindhoven, 2011), pp. 181–190.

[14] A. Z. Jacobs and A. Clauset, *A Unified View of Generative Models for Networks: Models, Methods, Opportunities, and Challenges*, arXiv:1411.4070.

[15] M. Medo, *Statistical Validation of High-Dimensional Models of Growing Networks*, Phys. Rev. E **89**, 032801 (2014).

[16] T. Pham, P. Sheridan, and H. Shimodaira, *Joint Estimation of Preferential Attachment and Node Fitness in Growing Complex Networks*, Sci. Rep. **6**, 32558 (2016).

[17] J. Overgoor, A. Benson, and J. Ugander, *Choosing to Grow a Graph: Modeling Network Formation as Discrete Choice*, in *The World Wide Web Conference* (ACM, San Francisco, 2019), pp. 1409–1420.

[18] S. Navlakha and C. Kingsford, *Network Archaeology: Uncovering Ancient Networks from Present-Day Interactions*, PLoS Comput. Biol. **7**, e1001119 (2011).

[19] J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou, *Graemlin: General and Robust Alignment of Multiple Large Interaction Networks*, Genome Res. **16**, 1169 (2006).

[20] J. Dutkowski and J. Tiuryn, *Identification of Functional Modules from Conserved Ancestral Protein–Protein Interactions*, Bioinformatics **23**, i149 (2007).

[21] J. W. Pinney, G. D. Amoutzias, M. Rattray, and D. L. Robertson, *Reconstruction of Ancestral Protein Interaction Networks for the bZIP Transcription Factors*, Proc. Natl. Acad. Sci. U.S.A. **104**, 20449 (2007).

[22] R. Patro, E. Sefer, J. Malin, G. Marçais, S. Navlakha, and C. Kingsford, *Parsimonious Reconstruction of Network Evolution*, Algorithm Mol. Biol. **7**, 25 (2012).

[23] R. Patro and C. Kingsford, *Predicting Protein Interactions via Parsimonious Network History Inference*, Bioinformatics **29**, i237 (2013).

[24] S. Li, K. Choi, T. Wu, and L. Zhang, *Reconstruction of Network Evolutionary History from Extant Network Topology and Duplication History*, in *Proceedings of the 2012 International Symposium on Bioinformatics Research and Applications* (Springer, New York, 2012), pp. 165–176.

[25] S. Li, K. P. Choi, T. Wu, and L. Zhang, *Maximum Likelihood Inference of the Evolutionary History of a PPI Network from the Duplication History of Its Proteins*, IEEE T. Comput. Biol. Bioinformat. **10**, 1412 (2013).

[26] T. A. Gibson and D. S. Goldberg, *Reverse Engineering the Evolution of Protein Interaction Networks*, in *Biocomputing* (World Scientific, Singapore, 2009), pp. 190–202.

[27] X. Zhang and B. M. E. Moret, *Refining Transcriptional Regulatory Networks Using Network Evolutionary Models and Gene Histories*, Algorithm Mol. Biol. **5**, 1 (2010).

[28] A. Jasra, A. Persing, A. Beskos, K. Heine, and M. De Iorio, *Bayesian Inference for Duplication-Mutation with Complementarity Network Models*, J. Comput. Biol. **22**, 1025 (2015).

[29] S. Bubeck, L. Devroye, and G. Lugosi, *Finding Adam in Random Growing Trees*, Random Struct. Algorithms **50**, 158 (2017).

[30] M. Z. Rácz, S. Bubeck *et al.*, *Basic Models and Questions in Statistical Network Analysis*, Stat. Surv. **11**, 1 (2017).

[31] G. Lugosi and A. S. Pereira, *Finding the Seed of Uniform Attachment Trees*, Electron. J. Probab. **24** (2019).

[32] D. Shah and T. Zaman, *Rumors in a Network: Who's the Culprit?*, IEEE Trans. Inf. Theory **57**, 5163 (2011).

[33] S. M. Mahantesh, S. Iyengar, M. Vijesh, S. R. Nayak, and N. Shenoy, *Prediction of Arrival of Nodes in a Scale Free Network*, in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (IEEE Computer Society, Washington, DC, 2012), pp. 517–521.

[34] G.-M. Zhu, H. J. Yang, R. Yang, J. Ren, B. Li, and Y.-C. Lai, *Uncovering Evolutionary Ages of Nodes in Complex Networks*, Eur. Phys. J. B **85**, 106 (2012).

[35] Abram N. Magner, Jithin K. Sreedharan, Ananth Y. Grama, and Wojciech Szpankowski, *TIMES: Temporal Information Maximally Extracted from Structures*, in *WWW 2018: The 2018 Web Conference, Lyon, France* (ACM, New York, NY, 2018), pp. 389–398.

[36] A. Magner, A. Grama, J. Sreedharan, and W. Szpankowski, *Recovery of Vertex Orderings in Dynamic Graphs*, in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT)* (IEEE, Aachen, 2017), pp. 1563–1567.

[37] M. Drmota, *Random Trees: An Interplay between Combinatorics and Probability* (Springer, New York, 2009).

[38] B. Bloem-Reddy and P. Orbanz, *Random-Walk Models of Network Formation and Sequential Monte Carlo Methods for Graphs*, J. R. Stat. Soc. B **80**, 871 (2018).

[39] P. Del Moral and L. M. Murray, *Sequential Monte Carlo with Highly Informative Observations*, SIAM/ASA J. Uncertain. Quantification **3**, 969 (2015).

[40] T. Pham, P. Sheridan, and H. Shimodaira, *PAFit: A Statistical Method for Measuring Preferential Attachment in Temporal Complex Networks*, PLoS One **10**, e0137796 (2015).

[41] P. L. Krapivsky, S. Redner, and F. Leyvraz, *Connectivity of Growing Random Networks*, Phys. Rev. Lett. **85**, 4629 (2000).

[42] R. Albert and A.-L. Barabási, *Topology of Evolving Networks: Local Events and Universality*, Phys. Rev. Lett. **85**, 5234 (2000).

[43] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks with Aging of Sites*, Phys. Rev. E **62**, 1842 (2000).

[44] P. L. Krapivsky, G. J. Rodgers, and S. Redner, *Degree Distributions of Growing Networks*, Phys. Rev. Lett. **86,** 5401 (2001).

[45] W. Aiello, F. Chung, and L. Lu, *Random Evolution in Massive Graphs*, in *Handbook of Massive Data Sets* (Springer, New York, 2002), pp. 97–122.

[46] D. J. De Solla Price, *Networks of Scientific Papers*, Science **149,** 510 (1965).

[47] F. Chung, F. R. K. Chung, F. C. Graham, L. Lu, K. F. Chung *et al.*, *Complex Graphs and Networks* (American Mathematical Society, Providence, Rhode Island, 2006), p. 107.

[48] L. A. Adamic and B. A. Huberman, *Power-Law Distribution of the World Wide Web*, Science **287,** 2115 (2000).

[49] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf, *A Likelihood Approach to Analysis of Network Data*, Proc. Natl. Acad. Sci. U.S.A. **103,** 7566 (2006).

[50] A. N. Guetz and S. P. Holmes, *Adaptive Importance Sampling for Network Growth Models*, Ann. Oper. Res. **189,** 187 (2011).

[51] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, *An Introduction to MCMC for Machine Learning*, Mach. Learn. **50,** 5 (2003).

[52] A. Doucet and A. M. Johansen, *Tutorial on Particle Filtering and Smoothing: Fifteen Years Later* in *The Oxford Handbook of Nonlinear Filtering*, edited by D. Crisan and B. Rozovskii (Oxford University Press, Oxford, 2011), pp. 656–704.

[53] B. H. Erickson, *Some Problems of Inference from Chain Data*, Sociol. Methodol. **10,** 276 (1979).

[54] S. H. Lee, P.-J. Kim, and H. Jeong, *Statistical Properties of Sampled Networks*, Phys. Rev. E **73,** 016102 (2006).

[55] J. S. Liu, *Metropolized Independent Sampling with Comparisons to Rejection Sampling and Importance Sampling*, Stat. Comput. **6,** 113 (1996).

[56] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.9.041056 for a numerical analysis of the generative model and sampling method.

[57] J.-G. Young, *Network-Archaeology: C++ Algorithms for Inferring the History of Random Networks* (2019), https://github.com/jg-you/network-archaeology.

[58] L. Hébert-Dufresne, J. A. Grochow, and A. Allard, *Multiscale Structure and Topological Anomaly Detection via a New Network Statistic: The Onion Decomposition*, Sci. Rep. **6,** 31708 (2016).

[59] Even though we are comparing "rankings" of sorts, the Pearson correlation is better suited to our use-case than a rank correlation like Spearman's. The main reason for this is that the Monte Carlo approximations of the averages $\{\langle\tau(e)\rangle\}$ are invariably noisy, and the Pearson correlation is robust to these small perturbations, whereas a rank correlation heavily penalizes them. Hence, the Pearson correlation is a better choice. We further note that in the special case in which $\hat{\tau}(e)$ takes on discrete values, the Pearson correlation of Eq. (16) is actually equivalent to Spearman's rank correlation, meaning that we are, in effect, using a rank correlation for our structural estimators.

[60] R. Oliveira and J. Spencer, *Connectivity Transitions in Networks with Super-Linear Preferential Attachment*, Internet Math. **2,** 121 (2005).

[61] If the labels were not arbitrary and we used the time of arrival as labels, then network archaeology would be trivial. We place ourselves in the more realistic situations where the labels are random.

[62] N. Rashevsky, *Life, Information Theory, and Topology*, Bull. Math. Biophys. **17,** 229 (1955).

[63] The standard definition of $S(G)$ is in terms of nodes [62]; here, we have opted for edge partitions since they are the basic unit of our inference.

[64] S. Bubeck, R. Eldan, E. Mossel, M. Z. Rácz, *From Trees to Seeds: On the Inference of the Seed from Large Trees in the Uniform Attachment Model*, Bernoulli **23,** 2887 (2017).

[65] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, *Nextstrain: Real-Time Tracking of Pathogen Evolution*, Bioinformatics **34,** 4121 (2018).

[66] P. Sagulenko, V. Puller, and R. A. Neher, *Treetime: Maximum-Likelihood Phylodynamic Analysis*, Virus Evol. **4,** vex042 (2018).

[67] S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas *et al.*, *Genomic Surveillance Elucidates Ebola Virus Origin and Transmission During the 2014 Outbreak*, Science **345,** 1369 (2014).

[68] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond, *BEAST 2: A Software Platform for Bayesian Evolutionary Analysis*, PLoS Comput. Biol. **10,** e1003537 (2014).

[69] J. Wang, A. Jasra, and M. De Iorio, *Computational Methods for a Class of Network Models*, J. Comput. Biol. **21,** 141 (2014).

[70] B. Hajek and S. Sankagiri, *Recovering a Hidden Community in a Preferential Attachment Graph*, in *ISIT 2018: 2018 IEEE International Symposium on Information Theory, Vail, CO, USA* (2018), pp. 2097–2101, https://ieeexplore.ieee.org/document/8437450.

[71] A. Magner, S. Janson, G. Kollias, and W. Szpankowski, *On Symmetry of Uniform and Preferential Attachment Graphs*, Electron. J. Comb. **21,** 3 (2014).

[72] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic Analysis of the Stochastic Block Model for Modular Networks and Its Algorithmic Applications*, Phys. Rev. E **84,** 066106 (2011).

[73] F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová, *Typology of Phase Transitions in Bayesian Inference Problems*, Phys. Rev. E **99,** 042109 (2019).

[74] H. Crane and W. Dempsey, *Edge Exchangeable Models for Interaction Networks*, J. Am. Stat. Assoc. **113,** 1311 (2017).

[75] D. Cai, T. Campbell, and T. Broderick, *Edge-Exchangeable Graphs and Sparsity*, in *Adv. Neural Inf. Process. Syst.* (2016), pp. 4249–4257.

[76] S. Linderman, G. Mena, H. Cooper, L. Paninski, and J. Cunningham, *Reparameterizing the Birkhoff Polytope for Variational Permutation Inference*, in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics* (PMLR, Playa Blanca, Lanzarote, Canary Islands, 2018), Vol. 84, pp. 1618–1627.

[77] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University, New York, 2009).

[78] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, *Inferring the Origin of an Epidemic with a Dynamic Message-Passing Algorithm*, Phys. Rev. E **90,** 012801 (2014).

[79] F. Gao and A. van der Vaart, *On the Asymptotic Normality of Estimating the Affine Preferential Attachment Network Models with Random Initial Degrees*, Stoch. Proc. Appl. **127**, 3754 (2017).

[80] B. Bloem-Reddy, A. Foster, E. Mathieu, and Y. W. Teh, *Sampling and Inference for Beta Neutral-to-the-Left Models of Sparse Networks*, arXiv:1807.03113.

[81] H. Jeong, Z. Néda, and A.-L. Barabási, *Measuring Preferential Attachment in Evolving Networks*, Europhys. Lett. **61**, 567 (2003).

[82] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-Law Distributions in Empirical Data*, SIAM Rev. **51**, 661 (2009).

[83] W. H. Press, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University, Cambridge, England, 2007).

[84] Lemma 5.3 of Ref. [35] states this fact without proof.

[85] B. D. McKay and A. Piperno, *Practical Graph Isomorphism, II*, J. Symb. Comput. **60**, 94 (2014).