Network archaeology: phase transition in the recoverability of network history Supplementary Information

Jean-Gabriel Young[a,b,c,1], Guillaume St-Onge[a,b], Edward Laurence[a,b], Charles Murphy[a,b], Laurent Hébert-Dufresne[a,d,e], and Patrick Desrosiers[a,b,f]

[a]Département de physique, de génie physique, et d'optique, Université Laval, Québec (QC), Canada
[b]Centre interdisciplinaire de modélisation mathématique de l'Université Laval, Québec (QC), Canada
[c]Center for the Study of Complex Systems, University of Michigan, MI, USA
[d]Department of Computer Science, University of Vermont, Burlington, VT, USA
[e]Vermont Complex Systems Center, University of Vermont, Burlington, VT, USA
[f]Centre de recherche CERVO, Québec (QC), Canada
[1]Corresponding author: jgyou@umich.edu

April 16, 2019

Contents

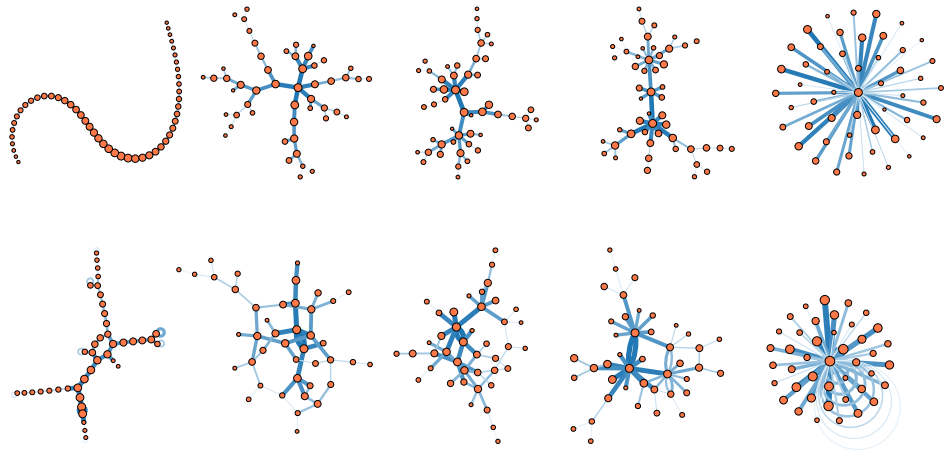

Figure 1: **Network zoo**. Examples of networks generated by the process with $b = 1$ (top row) and $b = 0.75$ (bottom row), and $\gamma \in \{-10, -1, 0, 1, 10\}$ (from left to right). The width and color of edges encode this history; older edges are drawn with thick, dark strokes, while younger edges are drawn using thin, light strokes. The age of nodes is encoded in their radius.

1 Characterization of the generative model

The generative model of the main text is a generalization of the classical preferential attachment model [1]. The salient features of the generalization are a non-linear attachment kernel k^γ [2] and the possibility for new links to connect pairs of existing nodes [3–7]. See the "Materials and Methods" section of the main text for a precise definition and Fig. 1 for an example of networks generated by the model with various parameters.

1.1 Degree distribution

Denote by $N_k(t)$ the number of nodes of degree k at time t and define the normalization $Z_\gamma(t) = \sum_k N_k(t)k^\gamma$. The evolution of $\{N_k(t)\}$ is approximately governed by the set of differential equations

$$\frac{dN_1(t)}{dt} = b - \frac{N_1(t)}{Z_\gamma(t)}[1 + (1-b)] \,, \tag{1a}$$

$$\frac{dN_k(t)}{dt} = [1 + (1-b)]\frac{\left[N_{k-1}(t)(k-1)^\gamma - N_k(t)k^\gamma\right]}{Z_\gamma(t)} \,, \tag{1b}$$

$$\frac{dZ_\gamma(t)}{dt} = \sum_k k^\gamma \frac{dN_k(t)}{dt} \,. \tag{1c}$$

The first term of Eq. (1a) accounts for the influx of new nodes of degree 1—attributable to node creation events—while the second term accounts for the loss of degree 1 nodes—through their acquisition of new edges. The positive term of Eq. (1b) tracks this growth in the next compartment, while the negative term represents the loss of nodes—again through their acquisition of edges. The same phenomenology applies for all k, and the system is therefore complete.

The specific form of the rate of change of the population in compartment k comes from the definition of the growth mechanism. At least one existing node is selected for growth at each time-step, and a second existing node can be selected, with probability $1 - b$, yielding a base rate of $[1 + (1 - b)]$. To obtain the complete expression of the rate for compartment k, $[1 + (1 - b)]$ is multiplied by the fraction of events that affect nodes in this compartment. In a preferential model with kernel k^γ, the probability that a growth event will affect a node of degree k is equal to $N_k k^\gamma / Z_\gamma(t)$, by definition of the model. This yields the final form of Eq. (1b).

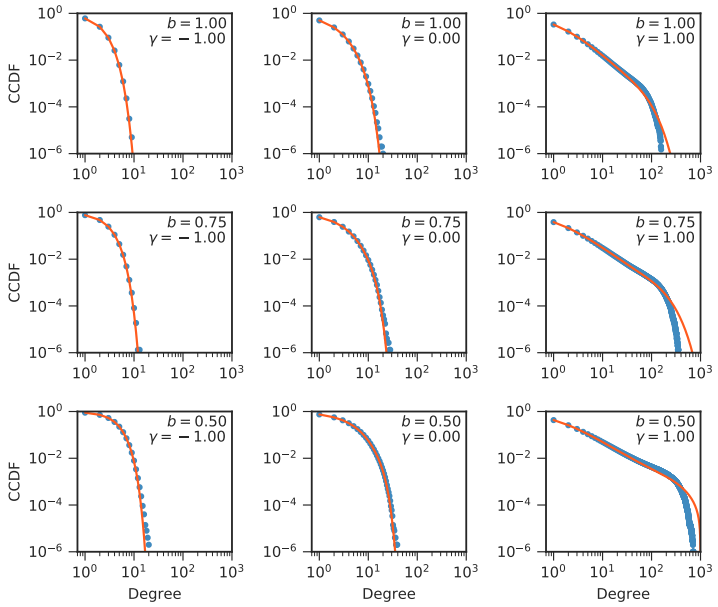

Figure 2: **Degree distributions of the (γ, b) generalization of PA.** Average empirical complementary cumulative distribution functions (blue symbols) versus the solution of Eqs. (1) (orange solid lines), for $T = 1000$, $b \in \{1, 0.75, 0.5\}$ (from top to bottom), and $\gamma \in \{-1, 0, +1\}$ (from left to right). Each empirical CCDF is computed on a concatenated degree sequences, obtained from 1 000 instances of the generative model.

We validate the mean-field equations in Fig. 2. We find that the predictions are accurate in most regimes, with the exception of $\gamma = 1$, $b < 1$, where we find a significant deviation in the tail of the distribution. This discrepancy can be traced back to a strong "peloton dynamics" [8–10], i.e., a phenomenon whereby a few individuals quickly accumulate a large fraction of the available resources. This effect is notably hard to capture with compartmental mean-field descriptions, and its impact is most felt in the regime $\gamma = 1$, $b < 1$.

1.2 Nearly simple graphs in the large network limit

It is clear that the model generates multigraphs with probability bounded away from 0 for all $b < 1$. This is simply due to the fact that new connections between existing nodes are made without regard to the identity of the nodes—doing otherwise by, e.g., rejecting proposed multi-edges and self-loops, would lead to identifiability issues for the parameter and the time-scale [11]. The net effect is that during the first several time-steps, a few densification events will invariably connect nodes to themselves and add redundant connections. But for most choices of exponent γ, this tendency eventually dies out, and the fraction of edges that are redundant edges or self-loops goes to zero with $T \to \infty$ (see Fig. 3). Thus, while they are technically multigraphs, the instances of the model are in fact *nearly* simple in the large network limit. Our generalization of PA therefore appears a reasonable model of multigraphs, but also a good approximation of large, sparse networks, with few or no redundant edges and self-loops. This being said, the fraction of redundant edges and self-loops diminishes extremely slowly when b is small; one should not use this generalization of PA as a model if the ratio $|V|/|E| \propto \langle k \rangle^{-1} \ll 1$ and the network is small.

Our numerical results show that the transition to the nearly simple regime occurs somewhere between $\gamma = 1$ and $\gamma = 2$, at a value $\gamma_c(b)$ that potentially depends on b. We do not derive rigorous bounds here, but note that the analysis of Ref. [2] implies the bound $\gamma_c(b) \leq 2$ independent from b. Indeed, we know that the networks *condensate* at $\gamma = 2$ when $b = 1$ (i.e., only a finite fraction of the nodes have degree greater

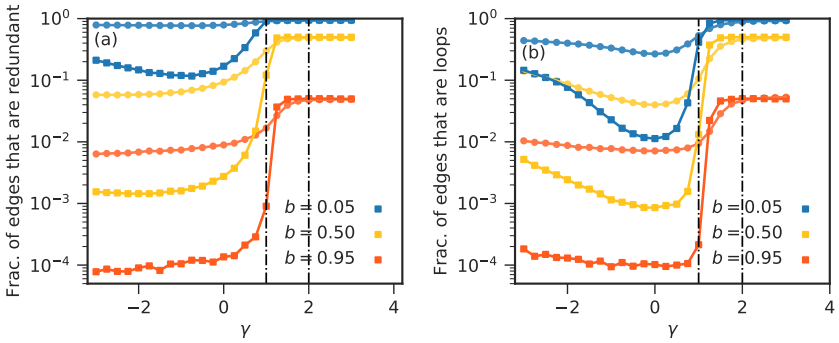

Figure 3: **Nearly simple graph transition.** Average fraction of the total number of edges that are **(a)** multi-edges, and **(b)** self-loops, as a function of γ, for three different values of b (denoted by colors), and two different sizes (denoted by marker types). Circles show the fractions for networks of $T = 10^2$ edges, while squares denote networks of $T = 10^4$ edges. In the special case of $b = 1$, it is known that networks start to "condensate" at $\gamma > 1$ (i.e., a finite fraction of the nodes have a macroscopic portion of the edges), and that they fully collapse at $\gamma = 2$ [2]. These two critical values of γ are shown with vertical lines. The curves are obtained by averaging over 1000 instances at each point for $T = 10^2$, and 100 instances at each point for $T = 10^4$.

than 1, and we are in a "winner-takes-all" scenario). It is clear that a condensate state also exists at $\gamma = 2$ when $b < 1$, since changing b only gives more opportunities to the leader to gain new connections (with itself, forming self-loops). This condensate will contain an extensive number of loops, which means that the graphs are never nearly simple at $\gamma = 2$, independent from b.

1.3 Endogenous correlations

The simple so-called "structural estimators" introduced in the main text relies on the natural correlations that arise between an edge's arrival time and its structural role in the network. We illustrate these correlations numerically, in Figs. 4–5. The results shown in Figs. 4 are in line with the well-known fact that there are endogenous correlations between the age and the degree of nodes in attachment models (see, e.g., Refs. [1, 2, 12, 13]). Figure 5 shows that the onion layer of an edge is even more strongly correlated with its age than the degree of its nodes. This partially explains why the layer-based estimation method performs better than the degree-based method.

We also note that Fig. 4–5 provide a visual explanation of the behavior of the estimators in the extreme regimes of $\gamma \to -\infty$ and $\gamma \to \infty$. In the regime $\gamma \ll 0$, the degree-based estimators perform poorly—but still extract *some* information—because the correlations start to vanish; notice how there are fewer classes the more negative γ becomes. This does not happen with OD; in fact the number of classes appears to grow with increasingly negative values of γ. Thus, the OD-based method can actually better differentiate edges in the regime $\gamma \ll 0$, using the endogenous correlations alone. In contrast, both estimators fail in the regime $\gamma \gg 0$ (see main text). Figures 4–5 show that this is due to the fact that almost all correlations vanish at the condensation threshold $\gamma = 2$.

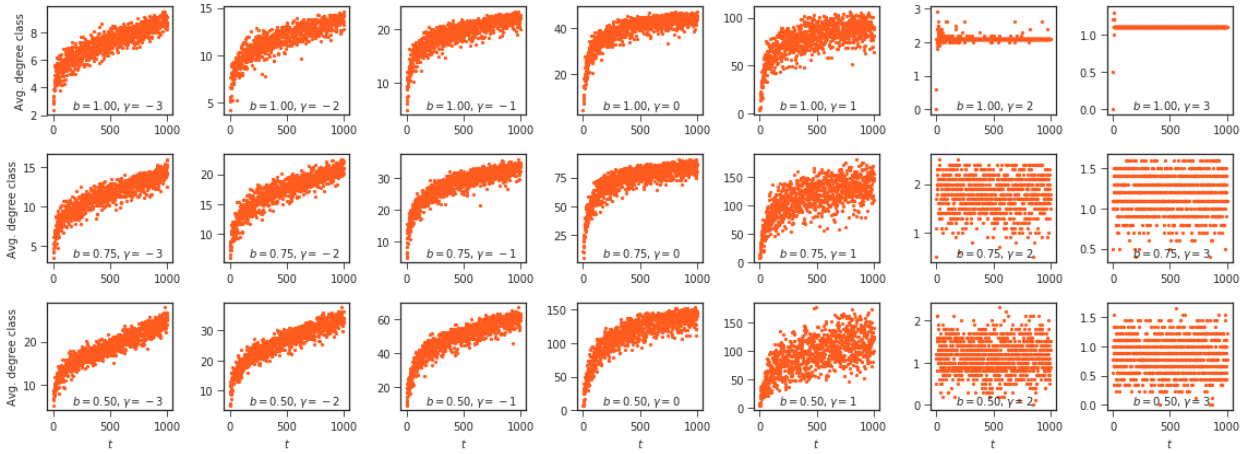

Figure 4: **Correlation between the age of edges and their degree classes.** True time of arrival of an edge t versus its expected degree class, as determined by the degree sorting algorithm presented in the Materials and Methods section of the main text, for $\gamma =\in \{-3, -2, -1, 0, 1, 2, 3\}$ (from left to right) and $b \in \{0.50, 0.75, 1.00\}$ (from bottom to top). Ten realizations of the growth and inference process are used to compute the expected classes.

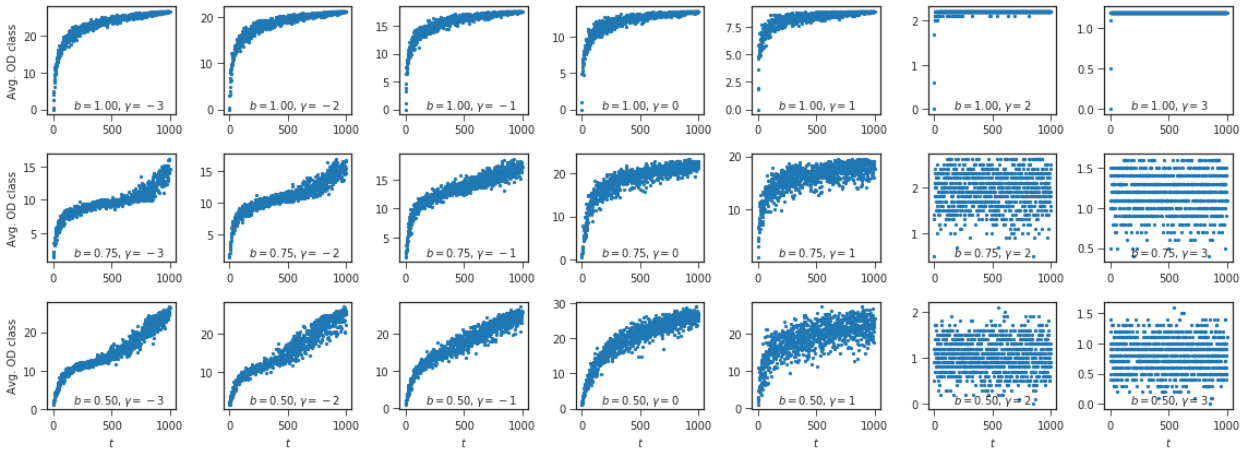

Figure 5: **Correlation between the age of edges and their onion decomposition classes.** True time of arrival of an edge t versus its expected onion decomposition (OD) class, as determined by the OD sorting algorithm presented in the Materials and Methods section of the main text, for $\gamma =\in \{-3, -2, -1, 0, 1, 2, 3\}$ (from left to right) and $b \in \{0.50, 0.75, 1.00\}$ (from bottom to top). Ten realizations of the growth and inference process are used to compute the expected classes.

2 On uniform posterior distributions

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 that this epitomizes why posterior maximization (MAP) is not a suitable method to extract information from G. We will now demonstrate that the Krapivsky-Redner-Leyvraz generalization of preferential attachment [2] is an extreme example of this problem. Namely, we will explicitly show that when $\gamma \in \{0,1\}$, the posterior distribution of this model is uniform over all histories in $\Psi(G)$ [1]. We will then further argue that there exist large equivalence classes for general $\gamma \in \mathbb{R}$ and $b \in [0,1]$.

2.1 Strict uniformity on trees

Recall from the Materials and Methods that in the model where $b = 1$, the logarithm of the prior distribution is given by

$$\log P(X_{0:T-1}|\gamma) = \sum_{t=1}^{T-1} \log u_{a_t}(\gamma, V_t) \;, \qquad u_i(\gamma, V_t) = \frac{k_i^\gamma(t)}{\sum_{i\in V_t} k_i^\gamma(t)} \;, \tag{2}$$

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 $Z(t;\gamma) = \sum_{i\in V_t} k_i^\gamma(t)$ and rewrite

$$\log P(X_{0:T-1}|\gamma) = \sum_{t=1}^{T-1} \left[\log k_{a_t}^\gamma(t) - \log Z(t;\gamma)\right] \;. \tag{3}$$

Now, in the special cases of uniform attachment and of 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:

$$Z(t;\gamma=0) = \sum_{i\in V_t} k_i^0 = |V_t| = t+1 \;, \tag{4a}$$

$$Z(t;\gamma=1) = \sum_{i\in V_t} k_i = 2t \;. \tag{4b}$$

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. (3), we are left with

$$\log P(X_{0:T-1}|\gamma) \propto \begin{cases} \text{Constant} & \gamma = 0, \\ \sum_{t=1}^{T-1} \log k_{a_t}(t) & \gamma = 1. \end{cases} \tag{5}$$

This last equation directly shows that the prior 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 this, 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 two, etc. Hence, every history consistent with G is associated with some permutation of a same sum. Obviously, a permutation does not change the value of a sum; therefore, the posterior distribution is uniform over all histories consistent with G.

[1] Lemma 5.3 of Ref. [14] states this fact without proof.

2.2 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 γ. 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 γ is arbitrary. But, as we now show, this does not preclude the existence of 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 $Z(1;\gamma) \to Z(2;\gamma) \to \ldots Z(T-1;\gamma)$, 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 at any point of the histories, and that they interact combinatorially: If there are m such equivalent sets of edges, of sizes $k_1, ..., k_m$, then each different point of the posterior is degenerated $k_1! \times ... \times k_m!$ times. This argument trivially extends to any $b \in [0, 1]$.

2.3 Connection with recursive random trees

The uniformity of the posterior distribution in the case ($\gamma \in \{0, 1\}, b = 1$) can also be explained in terms of random recursive trees and growth processes [15, pp. 14–16]. Specifically, it is well-known that generating a graph of T edges with ($\gamma = 0, b = 1$) is equivalent to drawing a graph uniformly from the set of all non-plane recursive trees (of T edges), while generating a graph of T edges with ($\gamma = 1, b = 1$) is equivalent to drawing a graph uniformly from all plane recursive trees (of T edges). In other words, $P(X_{0:T-1}|\gamma = 0)$ and $P(X_{0:T-1}|\gamma = 1)$ are uniform over the larger set of *all* trees. By conditioning on G via the likelihood $P(G|X_{0:T-1}, \gamma)$, we are merely renormalizing the distribution to the set $\Psi(G)$ of histories consistent with G. The uniformity of the posterior distribution follows from the fact that renormalizing preserves uniformity.

3 Properties of the average posterior time of arrival

In this section, we discuss why $\sum_{X\in\Psi(G)}\tau_X(e)P(X|G,\gamma)$ is a good estimator of the time of arrival. In doing so, we suppress the time subscripts $0:T-1$ for the sake of notational clarity.

3.1 Minimum mean-square error

In the main text, we mention without proof that the posterior average of $\tau_X(e)$ minimizes the mean-square error on $\tau_{\tilde{X}}(e)$. The proof is standard and goes as follows.

We first assume that the true arrival time of edge e is determined by its posterior distribution $p_e(t|G,\gamma,b)$. Without access to the ground truth, this is the best guess we can make since the posterior distribution extracts all available information from the graph G. The mean-square error (MSE) associated with some estimator $\hat{\tau}(e)$ of the true arrival time $\tau_{\tilde{X}}(e)$ of edge e is then

$$\begin{aligned}\text{MSE} &= \int dt(\hat{\tau}(e)-t)^2 p_e(t|G,\gamma,b)\\ &= \sum_{X\in\Psi(G)} P(X|G,\gamma,b)\int dt\big(\hat{\tau}(e)-t\big)^2\delta(t-\tau_X(e))\\ &= \sum_{X\in\Psi(G)} P(X|G,\gamma,b)\big(\hat{\tau}(e)-\tau_X(e)\big)^2\,,\end{aligned}$$

where we have defined the marginal $p_e(t|G,\gamma,b)=\sum_{X\in\Psi(G)}P(X|G,\gamma,b)\mathbb{I}[\tau_X(e)=t]$ for edge e, and where we have written the indicator function as Dirac's delta.

Even if the timescale of the process is by definition discrete, we resort to continuous estimators of $\tau_{\tilde{X}}(e)$. This is motivated not only by the simplifications that this decision brings to the mathematical treatment of the problem, but also by the fact that the generated graphs almost never encode the temporal ordering perfectly [16]. In short, total ordering won't do [14], and continuous estimators are needed to encode our imperfect guesses.

With this in mind, the estimator $\hat{\tau}^{\text{MMSE}}(e)$ that minimizes the mean-square error must solve

$$\left[\frac{\partial}{\partial\hat{\tau}(e)}\sum_{X\in\Psi(G)}P(X|G,\gamma,b)\big(\hat{\tau}(e)-\tau_X(e)\big)^2\right]_{\hat{\tau}(e)=\hat{\tau}^{\text{MMSE}}(e)}=0\,, \tag{6}$$

or more explicitly

$$\sum_{X\in\Psi(G)}P(X|G,\gamma,b)\big(\hat{\tau}^{\text{MMSE}}(e)-\tau_X(e)\big)=0\,. \tag{7}$$

Using the normalization of the posterior distribution, one easily obtains

$$\hat{\tau}^{\text{MMSE}}(e)=\langle\tau_X(e)\rangle\,, \tag{8}$$

where the average is taken over the posterior distribution for X.

3.2 Maximal correlation

A similar line of reasoning can be used to show that the MMSE estimators maximize the correlation, on average. 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

$$\begin{aligned}\langle\rho(X,Y)\rangle &= \sum_{X\in\Psi(G)} P(X|G,\gamma,b)\rho(X,Y) \\ &= \frac{1}{\sigma_X\sigma_Y}\sum_{X\in\Psi(G)} P(X|G,\gamma,b)\sum_{e\in E(G)}\big(\tau_X(e)-\langle\tau\rangle\big)\big(\tau_Y(e)-\langle\tau\rangle\big) \\ &= \frac{1}{\sigma_X\sigma_Y}\sum_{e\in E(G)}\Big[\langle\tau_X(e)\rangle-\langle\tau\rangle\Big]\Big[\tau_Y(e)-\langle\tau\rangle\Big]\,, \end{aligned} \tag{9}$$

where we have defined $\sigma_X^2 := \sum_{e\in E(G)}(\tau_X(e)-\langle\tau\rangle)^2$, and σ_Y similarly. These standard deviations can be taken out of the sum because σ_Y is independent of X and the value of σ_X is constant for all X. The uniformity of σ_X comes from the fact that one edge must occupy each 'time slot' $t = 0,\dots,T-1$ by definition, which implies $\sigma_X^2 = \sum_{t=0}^{T-1}(t-\langle\tau\rangle)^2$. Again, somewhat stretching the notation, we define as Z the history constructed with the MMSE estimators, i.e, the history such that $\tau_Z(e) = \langle\tau_X(e)\rangle$. This allows us to express the expected correlation compactly as

$$\langle\rho(X,Y)\rangle = \frac{\sigma_Z}{\sigma_X}\rho(Z,Y)\,, \tag{10}$$

where $\sigma_Z \neq \sigma_X$ in general. In other words, we find 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. This implies that the expected overall correlation is maximized if we choose Y to be the MMSE estimators of the arrival times.

Equation (10) has additional interesting consequences. First, it gives a compact expression of the expected correlation achieved by the MMSE estimators, since using the MMSE estimators amounts to setting $Z = Y$, yielding

$$\langle\rho(X,Z)\rangle = \frac{\sigma_Z}{\sigma_X}\,. \tag{11}$$

Second, Eq. (10) confirms the intuition that a good correlation can only be achieved by reliably ordering all the edges. Indeed, for the MMSE estimators the ratio σ_Z/σ_X gives an expected correlation, i.e., an average of numbers in $[-1,1]$. It follows that this ratio can never be greater than 1. This implies $\sigma_Z \leq \sigma_X$, where σ_X^2 is a known variance (see above). Now, turning this statement around: The ratio σ_Z/σ_X will be maximized if the (MMSE) estimator achieves a variance σ_Z^2 equal to the variance σ_X^2, obtained by ordering all edges. Grouping edges in equivalence classes reduces the variance σ_Z (since grouping two MMSE estimators imply averaging their ranking); this implies a degradation of the correlation. Thus, the maximum correlation $\langle\rho(X,Z)\rangle = 1$ can only be achieved if the MMSE estimators give a total ordering of the edges.

4 Properties of the correlation

In this section, we discuss why the Pearson product-moment correlation coefficient is a good measure of the quality of a collection of estimators $\{\tau(e)\}$ in the context of network archaeology. In doing so, we suppress the time subscripts $0:T-1$ for the sake of notational clarity.

4.1 Impossibility of improving on equivalence classes

Suppose that some estimator of $\tau_{\tilde{X}}(e)$ cannot differentiate the edges in a subset $S \subset E$, and that according to the estimation procedure, these edges would occupy time slots $t+1$ through $t+m$ (where $m = |S|$). We now show that trying to arbitrarily break the tie will not improve correlation on average.

We define the history Y as the one where $\tau_Y(e) = t + (m+1)/2$ if $e \in S$. We also consider the $m!$ different ways of randomly breaking the ties for S (i.e., the permutations of S). We denote these histories with $\{Z_{\pi_i}(X)\}$. They satisfy the property that $\tau_{Z_{\pi_i}}(e) = \tau_Y(e)$ if $e \notin S$, and that the edges of S are assigned some permutations of the times of arrival $t+1, \ldots, t+m$, indexed by i.

If we break the ties randomly, we will achieve, on average, a correlation of

$$\langle \rho(X, Z_\pi) \rangle_\pi := \frac{1}{m!} \sum_{i \in \pi(S)} \rho(X, Z_{\pi_i}) \tag{12}$$

with the ground truth X (here noted without tilde to simplify the notation). Then, to prove that randomly breaking ties cannot improve the correlation, we must show that

$$\langle \rho(X, Z_\pi) \rangle_\pi - \rho(X, Y) := \Delta \leq 0 \,. \tag{13}$$

By symmetry of the permutation group, we can compute the average as

$$\langle \rho(X, Z_\pi) \rangle_\pi = \frac{1}{2}[\rho(X, Z_{\pi_+}) + \rho(X, Z_{\pi_-})] \,, \tag{14}$$

where Z_{π_-} is the history constructed with some permutation of S and where Z_{π_+} is the inverse permutation. Noting that $\sigma_{Z_{\pi_-}} = \sigma_{Z_{\pi_+}}$, we write $\sigma_X \sigma_{Z_{\pi_+}} \Delta$ as

$$\begin{aligned} \frac{1}{2}\left[\sum_{e \in E(G)} \big(\tau_{Z_{\pi_-}}(e) - \langle\tau\rangle\big)\big(\tau_X(e) - \langle\tau\rangle\big) + \sum_{e \in E(G)} \big(\tau_{Z_{\pi_+}}(e) - \langle\tau\rangle\big)\big(\tau_X(e) - \langle\tau\rangle\big) \right] \\ - \frac{\sigma_{Z_{\pi_+}}}{\sigma_Y} \left[\sum_{e \in E(G)} \big(\tau_Y(e) - \langle\tau\rangle\big)\big(\tau_X(e) - \langle\tau\rangle\big) \right] \,. \end{aligned} \tag{15}$$

The standard deviations σ_X and $\sigma_{Z_{\pi_+}}$ are non-negative coefficients by definition. It therefore suffices to show that the above expression is smaller than or equal to 0.

To do so, we first note that $\sigma_{Z_{\pi_+}} > \sigma_Y$[2]. This implies that

$$\begin{aligned} \sigma_X \sigma_{Z_{\pi_+}} \Delta < \sum_{e \in S} \Big[\frac{1}{2}\big(\tau_{Z_{\pi_-}}(e) - \langle\tau\rangle\big)\big(\tau_X(e) - \langle\tau\rangle\big) \\ + \frac{1}{2}\big(\tau_{Z_{\pi_+}}(e) - \langle\tau\rangle\big)\big(\tau_X(e) - \langle\tau\rangle\big) - \big(\tau_Y(e) - \langle\tau\rangle\big)\big(\tau_X(e) - \langle\tau\rangle\big) \Big] \end{aligned} \tag{16}$$

[2]This can also be shown directly by comparing the *variances*; one finds that $\sigma^2_{Z_{\pi_+}} - \sigma^2_Y = \sum_{e \in S}[\tau^2_{Z_{\pi_+}}(e) - \tau^2_Y(e)] > 0$, since $\tau_Y(e) = t + (m-1)/2$ for all $e \in S$ and the $\tau_{Z_{\pi_+}}$ goes from $t+1$ to $t+m$.

where we have used $\sigma_{Z_{\pi_+}}/\sigma_Y > 1$ and the fact that the permuted histories are identical to Y for all $e \notin S$. To obtain an explicit numerical expression for the right-hand side, we index the edges of S by their true time of arrival (i.e., their arrival time in X). That is to say, we choose

$$\tau_X(e_1) < \tau_X(e_2) < \ldots < \tau_X(e_m)\ . \tag{17}$$

With an ordering defined, we choose the following permutations to construct Z_{π_-} and Z_{π_+}:

$$\begin{array}{lcccc} Z_{\pi_+}: & t+1 & t+2 & \ldots & t+m \\ & \downarrow & \downarrow & \ldots & \downarrow \\ & e_1 & e_2 & \ldots & e_m \end{array}$$

$$\begin{array}{lcccc} Z_{\pi_-}: & t+1 & t+2 & \ldots & t+m \\ & \downarrow & \downarrow & \ldots & \downarrow \\ & e_m & e_{m-1} & \ldots & e_1 \end{array}$$

In other words, in Z_{π_+}, edge e_1 is assigned time $t+1$, edge e_2 is assigned time $t+2$, etc. Then, recalling that $\tau_Y(e) = t + (m+1)/2$ for all $e \in S$, we obtain for the right-hand side of Eq. (16):

$$\begin{aligned} \sum_{i=1}^{m} \Big[& \frac{1}{2}\big(t+m-i+1-\langle\tau\rangle\big)\big(\tau_X(e_i)-\langle\tau\rangle\big) \\ & + \frac{1}{2}\big(t+i-\langle\tau\rangle\big)\big(\tau_X(e_i)-\langle\tau\rangle\big) - \big(t+(m+1)/2-\langle\tau\rangle\big)\big(\tau_X(e_i)-\langle\tau\rangle\big)\Big] \\ = \sum_{i=1}^{m} & \Big[\frac{1}{2}\big(t+m-i+1-\langle\tau\rangle\big) + \frac{1}{2}\big(t+i-\langle\tau\rangle\big) - \big(t+(m+1)/2-\langle\tau\rangle\big)\Big]\big(\tau_X(e_i)-\langle\tau\rangle\big) = 0 \end{aligned}$$

This completes the proof that randomly breaking ties will on average yield a worst outcome than merely guessing $\tau_Y(e) = t + (m+1)/2$ for the edges $e \in S$.

4.2 On the placement of equivalence classes

Consider again a set of tied edges S with prescribed arrival times $t+1, ..., t+m$ (if they were not tied). We have shown in the above section that assigning $\tau_Y(e) = t + (m+1)/2$ for all $e \in S$ is a better choice, on average, than breaking the ties at random. We now show this choice is in fact the best we can do.

Our overall proof strategy is to assign the time $\lambda + \varepsilon$ to the edges in S, where $\lambda = t + (m+1)/2$ and where ε is a small perturbation away from this guess. We then compute the average correlation with the ground-truth as a function of ε. Finally, we show that setting $\varepsilon \neq 0$ might actually decrease the correlation with the ground-truth, such that $\varepsilon = 0$ is the better choice.

Setting $\lambda + \varepsilon$ for the edges in S, and assuming that all edges *not in* S are assigned a unique time, or that the other equivalence classes have been collapsed on *their* averages $\tilde{\lambda}$, the overall average time of arrival of the edges in history Y is

$$\langle\tau(\varepsilon)\rangle = \frac{1}{T}\left[\sum_{i=0}^{t} i + m(\lambda+\varepsilon) + \sum_{i=t+m+1}^{T-1} i\right] = \frac{m}{T}(\lambda+\varepsilon) + C =: \langle\tilde{\tau}\rangle + \frac{m}{T}\varepsilon \tag{18}$$

where C is a constant and where we define $\langle\tilde{\tau}\rangle$ ast the average time, were ε set to 0. Now, the variance of the arrival times in $Y(\varepsilon)$ is

$$\sigma_Y^2(\varepsilon) = \sum_{e\in E}\left(\tau_Y(e) - \langle\tilde{\tau}\rangle - \frac{m\varepsilon}{T}\right)^2 = \sigma_Y^2(0) + 2m\varepsilon(\lambda - \langle\tilde{\tau}\rangle) + m\epsilon^2\left(1+\frac{m}{T}\right)\ , \tag{19}$$

such that the standard deviation is, to the first order in ε,

$$\sigma_Y(\varepsilon) = \sigma_Y(0)\left[1 + \frac{m\varepsilon(\lambda - \langle\tilde{\tau}\rangle)}{\sigma_Y^2(0)} + O(\varepsilon^2)\right] . \tag{20}$$

This variance can be substituted in Eq. (10) to compute the expected correlation of the history $Y(\varepsilon)$ with the ground truth X, as a function of ε.

Noting that $\tau_Y(e) = \lambda + \varepsilon$ if $e \in S$, and that the sum of $\tau_Z(e) - \langle\tilde{\tau}\rangle$ over all edges is equal to 0, we find

$$\begin{aligned}
\langle\rho(X,Y)\rangle &\propto \frac{\sum_{e\in E}(\tau_Z(e) - \langle\tilde{\tau}\rangle)(\tau_Y(e) - \langle\tilde{\tau}\rangle - m\varepsilon/T)}{\sigma_Y(\varepsilon)\sigma_Z} \\
&= \left[\langle\tilde{\rho}\rangle + \varepsilon\frac{\sum_{e\in S}(\tau_Z(e) - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)\sigma_Z}\right]\left[1 - \frac{m\varepsilon(\lambda - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)^2} + O(\varepsilon^2)\right] \\
&= \langle\tilde{\rho}\rangle\left\{1 + \varepsilon\left[\frac{1}{\langle\tilde{\rho}\rangle}\frac{\sum_{e\in S}(\tau_Z(e) - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)\sigma_Z} - \frac{m(\lambda - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)^2}\right] + O(\varepsilon^2)\right\}
\end{aligned} \tag{21}$$

where $\langle\tilde{\rho}\rangle$ is the averaged correlation achieved when $\varepsilon = 0$, and where Z is the MMSE history.

The question is then: What values should we choose for ε? A priori, the two terms of the coefficient of ε in Eq. (21) have about the same order of magnitude, because $|S| = m$, $\langle\tilde{\rho}\rangle = O(1)$, and because $\sigma_Y(0)$ and σ_Z are variances with exactly the same number of terms. Hence, the coefficient may be positive or negative, and the only way to know is to compute the (optimal) MMSE history. But if we knew Z, then we would not need to set ε in the first place. This leaves us in a situation where we have to guess ε without calculations. Hence, $\varepsilon = 0$ is as good as any guess. We favor it because it preserves the mean arrival time.

5 Characterization of the sampling method

5.1 Proposal distributions

Sequential Monte Carlo methods make use of a proposal distribution $Q(X_{0:T-1}, \theta|G)$ to approximate $P(X_{0:T-1}|G, \gamma, b)$ (where θ is a set of parameters, potentially functions of γ and b). We here discuss several proposal distributions in depth.

5.1.1 Snowball proposal distribution

The snowball proposal distribution $Q_{\mathrm{sb}}(X_{0:t}|G)$ is introduced in the main text, see the Materials and Methods for details. In a nutshell, a snowball sample is a random recursive enumeration of a graph, rooted on a randomly selected seed [17–19]. The main advantage of this method is that it admits a simple and efficient sampling algorithm. We simply grow a boundary Ω, i.e., a set of potential edges that could be enumerated next, and store it in a variable length array that supports random access. Drawing uniformly from this boundary can be done in $O(1)$ time. The principal bottleneck is updating the boundary, which takes $O(k_{\max})$ time if we use the graph structure and tags to determine which edges should be added to Ω (this is faster than querying the boundary).

5.1.2 Truncated posterior proposal distribution

Another possible choice of proposal distribution—not discussed in the main text—is what we call the "truncated posterior distribution." It has been recently applied to a temporal graph sampling problem close to the one analyzed in the main text [20] and is frequently used in the particle filtering literature [21] due to its mathematical simplicity. The idea is to choose a proposal that simplifies the form of the weights $\omega(X_{0:t}|G, \gamma, b) = P(X_{0:t}|\gamma, b)/Q(X_{0:t}|G, \gamma, b)$, specifically

$$Q_{\mathrm{tp}}(X_{0:t}|G, \gamma, b) \propto P(X_{0:t}|\gamma, b)\mathbb{I}[X_{0:t} \in \Psi(G_t)] \,, \tag{22}$$

which transform the weight into a simple normalization over all the possible continuation of $X_{0:t-1}$ consistent with G_T. The sample weights can therefore be incremented by:

$$\sum_{e \in \Omega(X_t)} P(X_t|X_{t-1}, \gamma, b) \tag{23}$$

at every step, with $\Omega(X_t)$ the boundary at time t. Unfortunately, a naive implementation of the truncated posterior proposal distribution will return samples in $O(|E| \times |E|)$ time, since the possible transitions $X_{t-1} \to X_t$ have heterogeneous (and changing) weights that must be updated at every step.

Thankfully, an efficient $O(|E| \times k_{\max} \times \log\log k_{\max}^{2|\gamma|})$ implementation can be obtained by using a slightly complex underlying data structure. We again grow a boundary Ω. Differently from the snowball proposal distribution, we store this boundary in two separate maps, one for the *tendrils* $\mathcal{T}$ (edges in the boundary connected to only one node already in G_t) and one for the *closures* $\mathcal{C}$ (connected to two nodes in G_t). We associate a weight k_i^γ to every edge in $\mathcal{T}$, where i is the node of the tendril already part of the graph $G(X_t)$. Likewise, we associate a weight $k_i^\gamma k_j^\gamma$ to every edge in $\mathcal{C}$, where (i, j) are the two end-nodes of the closure. These weights are used as the keys of the map, and the content is a list of edges with that weight. Finally, we keep track of the sum of the weights in both trees as $\Sigma_{\mathcal{T}}$ and $\Sigma_{\mathcal{C}}$, and also update a overall degree normalization $Z(t;\gamma) = \sum_{i \in V_t} k_i^\gamma$. Upon adding a new edge to the history, we cycle through its neighboring edges, and (1) add new edges to the boundary, (2) move edges from the tendril set to the closure set if necessary, (3) update the weight of the affected edges, and (4) update all the running normalization. This can all be done in at worst $O(k_{\max})$ time using tags on edges to classify them as in the boundary, already enumerated or unseen. The bottleneck is the lookup of up to $k_{\max}$ neighbors of the newly enumerated edge.

This complex storage method pays off when it comes time to sample: We can draw an edge from the truncated posterior distribution in near constant time by using a two–steps algorithm. We first randomly

choose between the tendrils and closures, with probabilities proportional to the total weight of their elements, i.e.:

$$\Pr[\text{Next edge is tendril}] = \frac{b\Sigma_{\mathcal{T}}}{b\Sigma_{\mathcal{T}} + (1-b)\Sigma_{\mathcal{C}}/Z(t;\gamma)} , \tag{24}$$

where $Z(t;\gamma)$ accounts for the fact that a transition involving a closure (i,j) is proportional to $k_i^\gamma k_j^\gamma / Z^2$ while a transition that involves a tendril (i,k)—where k is a node that was never seen before—is proportional to k_i^γ / Z. Second, after having chosen whether the next edge is a tendril or a closure, we sample from the selected subset using probabilities proportional to the weights the edges. Since we have separated edges in weight classes, a composition-rejection algorithm can be used to sample from the relevant subset of the boundary in $O(\log\log w_{\max})$ time [22], where $w_{\max}$ is the maximal weight of the elements in that set (equal to $k_{\max}^{|\gamma|}$ for $\mathcal{T}$, and $k_{\max}^{2|\gamma|}$ for $\mathcal{C}$). It is easy to verify that the two-step process above samples from the edges in the boundary with probability proportional to $P(X_t|X_{t-1},\gamma,b)$.

Note that separating the boundary in two parts is necessary: Updates to the degrees affect the relative probability of choosing closures and tendrils differently, because the normalizations of the associated transitions differ by a factor of Z. Hence we cannot keep track of these two edges sets in a single binary decision tree without costly updates.

5.1.3 Snowball distribution with an initial bias

Recall from the main text that the probability of generating a history $X_{0:T-1}$ with the unbiased snowball proposal distribution is

$$Q_{\text{sb}}(X_{0:T-1}|G) = P(X_0|G) \times \prod_{t=1}^{T-1} \left[|\Omega(X_t)|\right]^{-1} , \tag{25}$$

where $\Omega(X_t)$ denotes the boundary at step t of history X, and where $P(X_0|G)$ is the distribution used to select the seed (the initial edge). In the classical snowball proposal distribution, we choose uniformly from $E(G_T)$ for the sake of simplicity. But the distribution $P(X_0|G)$ is in fact an optimization opportunity; it allows us to start "on the right foot," with edges that are likely to be the first. Building on the insight of Sec. 1.3, we use the parametrized distribution

$$P(X_0|G;\alpha) = \frac{(k_1(e)\times k_2(e))^\alpha}{Z(\alpha)} \qquad Z(\alpha) = \sum_{e\in E(G)} (k_1(e)\times k_2(e))^\alpha \tag{26}$$

where $\alpha \geq 0$ is a parameter and $k_1(e)$ and $k_2(e)$ the degrees of the nodes attached to e. Large values of α nudge the algorithm towards histories that begin with edges that are attached to high degree nodes.

5.1.4 Biased snowball proposal distribution: controlled evolution

Both b and γ are constants of the generative process. This implies that unbiased estimators $(\hat{b}(t),\hat{\gamma}(t))$ constructed with the first t iterations of the ground truth $\tilde{X}$ ought to settle on (b,γ), for sufficiently large t. A *biased* snowball proposal distribution can make use of this property to generate better histories, by ensuring that some running estimators of b and γ stay close to some constant baseline computed on the whole network. If the estimators fluctuate a lot throughout a history X, then X is probably not the ground truth, and it is also unlikely to be *even well correlated* with the ground truth. Ergo, it can pay to actively steer the sampler away from such histories.

In practice, running estimators of γ are hard to evaluate, let alone update efficiently. Therefore we use the estimator pair:

$$\hat{b}(t) = \frac{|V_t|-2}{|E_t|-1} \qquad \text{and} \qquad \hat{v}(t) = \frac{1}{|V_t|}\left[\sum_{i\in V_t} k_i(t)\right]^2 + \frac{1}{|V_t|^2}\sum_{i\in V_t} k_i^2(t) . \tag{27}$$

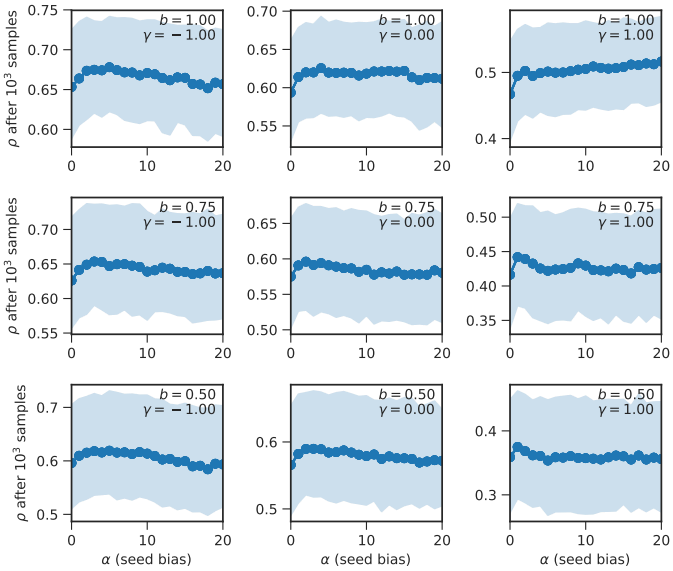

Figure 6: **Effect of bias in selecting the seed of snowball samples.** Average correlation attained with the snowball proposal distribution, when using 10^3 samples to compute MMSE estimators, as a function of the strength of the bias α used in selecting the seed. The experiments are run on artificial graphs of $T = 50$ edges generated by the model, with $b \in \{1, 0.75, 0.5\}$ (from bottom to top), and $\gamma \in \{-1, 0, +1\}$ (from left to right). The curves are averaged for 10 runs of the sampler, on 200 different instances of the model (i.e., all averages are computed with 2 000 data points). To isolate the effect of α, we do not perform any resampling (i.e., we use a SIS algorithm, and not an SMC algorithm). The shaded region contains 50 percent of the data points (from the 25^{th} percentile to the 75^{th} percentile).

where $\hat{v}(t)$, the variance of the degree sequence, is a proxy for $\hat{\gamma}(t)$. These estimators are easy to update on the fly in $O(1)$ operations; one simply needs to keep a running count of $|V_t|, |E_t|, \sum_{i \in V_t} k_i^2(t)$ and t (notice that $\sum_{i \in V_t} k_i(t) = 2t$ by definition). Now, for target values (v^*, b^*) computed on the final graph (i.e., the target baseline), and some parameters $(\beta > 0, \mu > 0)$, we give the following weight to an edge:

$$w(e) \propto \exp\Big\{ -\beta \Big[\hat{b}(t+1) - b^*\Big]^2 - \mu \left[\hat{v}(t+1) - v^*\right]^2 \Big\} , \tag{28}$$

where $\hat{b}(t+1)$ and $\hat{v}(t+1)$ are the estimators evaluated in the hypothetical history where e is selected at time t. Sampling is accomplished by, again, growing a boundary Ω and sampling from it with probability proportional to $w(e)$, i.e.:

$$Q_{\text{bsb}}(X_{0:T-1} | G; \beta, \mu) = P(X_0 | G; \alpha) \times \prod_{t=1}^{T-1} \frac{w(e_t)}{W(t)} , \tag{29}$$

where $W(t)$ is the sum of the weights on the boundaries. The unbiased snowball distribution recovered by setting $\beta = \mu = 0$, since $w(e_t)/W(t)$ is then equal to $1/|\Omega(X_t)|$.

The main drawback of the biased snowball proposal is its computational complexity. The weights must be recomputed at each step, because they are functions of time-dependent quantities that are functions of the whole network, leading to a complexity of $O(|E| \times |E|)$ per sample.

5.2 Analysis of the proposal distributions

This section regroups two numerical case studies (Figs. 6–7) showcasing the advantage and drawbacks of the proposal distributions.

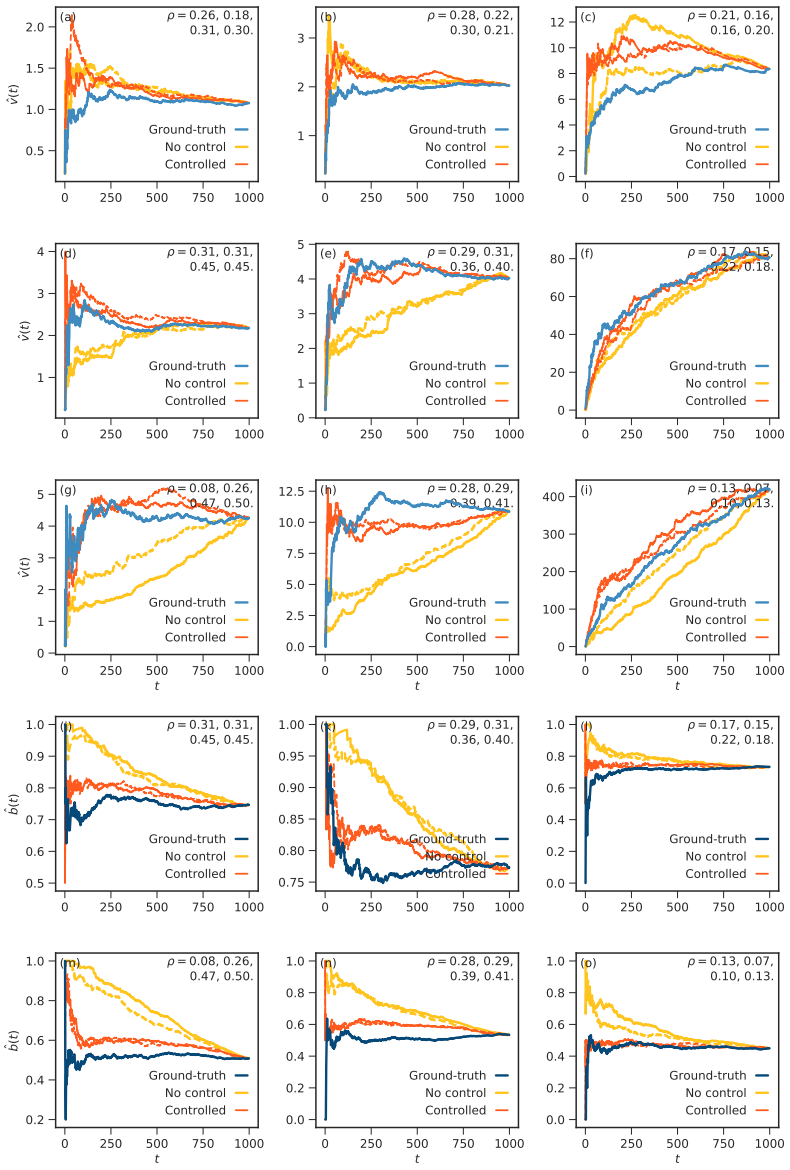

Figure 7: **Evolution of the estimators in snowball samples of large graphs.** Evolution of the ground truth (blue lines), of two uncontrolled samples (yellow lines), and of two controlled samples (orange lines), on artificial graphs of $T = 1000$ edges generated by the model. We use the kernel exponent (left column) $\gamma = -1$, (central column) $\gamma = 0$, and (right column) $\gamma = 1$, and the node creation probability equals (top row) $b = 1$, (second and fourth row) $b = 0.75$, and (third and last row) $b = 0.50$. The first 9 panels show the evolution of the estimated variance $\hat{v}(t)$ and the last 6 panels show the evolution of the estimated node creation probability $\hat{b}(t)$. The panels are matched, e.g., the dotted orange line in panels (d) and (j) illustrate the evolution of the estimators in a single sample, on the same graph. There are only six panels associated with $\hat{b}(t)$ since its evolution is trivial on trees. The number in the top-right corner of each panel indicates the correlation of the ground truth and the uncontrolled samples (two topmost numbers), and of the ground truth and the controlled samples (the two numbers below). We use $\beta = 2000, \mu = 0$ on loopy graphs ($b < 1$) and $\mu = 1$ on trees, except when $\gamma = 1$, in which case we opt for no control at all, by fear of reaching the steady state of $\hat{v}(t) = v^*$ too rapidly.

5.2.1 Effect of initial bias

In Fig. 6, we investigate the effects of the initial bias parameter α of Sec. 5.1.3 by checking the quality of the inference results on multiple artificial networks, for various of levels of initial bias. We find that non-zero values of α systematically enhance the inference. But because the increase in quality is very typically small, we opted for clarity in the main text and performed all computations with $\alpha = 0$ (i.e., we used the "simple" snowball distribution) and more samples. As a rule of thumb, however, $\alpha = 5$ yields better results across all parameters.

5.2.2 Effect of controlled evolution

In Fig. 7, we show the evolution of the estimators $(\hat{b}, \hat{v})$ for both *unbiased* and *biased* ("controlled") samples generated using the snowball proposal distribution, on artificial network instances. We superimpose the trajectories associated with the ground truth. We set an initial bias of $\alpha = 5$ for all controlled trajectories (see above discussion), and of $\alpha = 0$ in all uncontrolled trajectories. We choose different values of (β, μ) based on $(\hat{b}(T), \hat{\gamma}(T))$, see the caption [3].

The figures show that biased samples are much closer to the ground truth than a typical snowball sample. In the early stage of the enumeration, the snowball proposal distribution branches out on the background graph [23], generating an almost tree-like sample. Only in the later stages does it consolidates edges and lowers its node to edge ratio $\hat{b}(t)$. The biased snowball proposal distribution therefore appears far superior to the simple snowball sample.

Unfortunately, the complexity of the sampling process for this proposal distribution is prohibitive (several seconds are needed to generate a sample for the network shown), and prevents us from actually generating enough sample to compute posterior estimates. Hence, given a fixed time budget, we find it ultimately more practical to generate many poorly correlated (using the snowball or truncated posterior proposal distribution) than to generate much fewer high-quality samples with a biased snowball proposal distribution. An interesting direction for future research would be to obtain a variation on the biased snowball proposal distribution that admits an efficient implementation, perhaps building on the "bookkeeping techniques" introduced in Sec. 5.1.2.

5.3 Best choice of proposal distribution and resampling level

In this section, we determine which proposal distribution is best suited to network archaeology, by checking how fast the estimators $\{\hat{\tau}(e)\}$ converge to their optimal values, depending on which distribution is used. In doing so, we do not investigate the biased snowball proposal distribution because its computational cost is far too prohibitive to conduct a numerical study.

Our results are shown in Fig. 8, where we analyze the interplay between the sample size, resampling level, proposal distribution and achieved correlation. Our experiments show that, as one might expect, the SMC always outperforms the onion decomposition (OD) on average given enough samples. The number of samples needed to accomplish this is not constant depending on the combination of generating and sampling parameters. In all cases, however, the number of samples needed represent an astronomically small fraction of the number of $|\Psi(G)|$ potential histories[4].

From the results of Fig. 8, it appears that the snowball proposal distribution (triangles) always does *at least as good* as the truncated posterior proposal distribution (circles), once we choose the correct level of resampling. Furthermore, we find that resampling *does not* help in the loopy case ($b = 0.75$) or in the PA

[3] Here we evaluate γ directly—only once on the full graph—to choose values for bias parameters. This is a relatively easy minimization problem, see Materials and Methods.

[4] A loose upper bound on $|\Psi(g)|$ is $T!$, not accounting for the existence of inconsistent histories. Closer estimates can be obtained with a small modification to the sequential importance sampling method (see Ref. [24] for a related application of sampling to counting). The method relies on the identity $|\Psi(G)| = \sum_{X_{0:T-1} \in \Psi(G)} Q(X_{0:T-1}|G)/Q(X_{0:T-1}|G) = \langle 1/Q(X_{0:T-1}) \rangle_Q$. The idea is to approximate the average as $\frac{1}{n}\sum_{i=1}^{n}[Q(x_i|G)]^{-1}$ with x_i drawn from the distribution $Q(X)$, here chosen as the snowball proposal distribution for convenience. In the classical preferential attachment case where $b = 1$ and $\gamma = 1$, for example, we find $|\Psi(G)| = O(10^{54})$ when $T = 50$, a result that is ten orders of magnitude smaller than the loose bound $T! = O(10^{64})$.

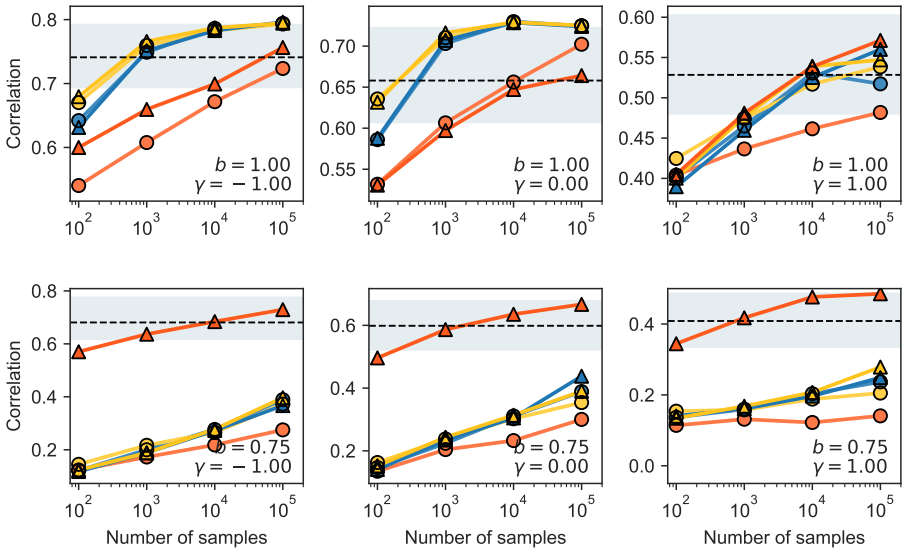

Figure 8: **Convergence of the sampling methods.** Average correlation achieved with various proposal distributions and resampling levels, on artificial networks of $T = 50$ edges generated with $b \in \{1, 0.75\}$ (from top to bottom), and $\gamma \in \{-1, 0, +1\}$ (from left to right). Results obtained with the *truncated posterior proposal distribution* are shown using circles. Results obtained with the *snowball proposal distribution* are shown with triangles. The resampling threshold used is shown with colors: systematic corresponds to the blue makers (ESS $< n$), adaptive SMC corresponds to the yellow markers (ESS $< n/2$), and SIS is shown in orange (ESS < 0). All the data points are obtained by averaging over independent realization of the model and of the sampling process. Denoting the number of samples used to evaluate the estimators as n, the number of independent trials used to compute the averages are: $n = 10^2$ (3000) $n = 10^3$ (1800) $n = 10^4$ (1200) $n = 10^5$ (120). The performance of the onion decomposition is shown for reference (average denoted by a dotted black line and 50% of the data is shown with the shaded rectangle). We omit the error bars for the sake of clarity. To avoid confounding factor, we use no root biased $\alpha = 0$ in all cases.

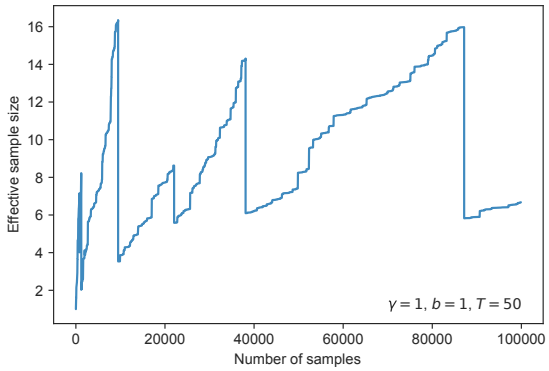

Figure 9: **Effective sample size in SIS.** Typical example of the evolution of the effective sample size (ESS, see Eq. (30)), as a function of the number of samples used. To highlight the abrupt change in ESS, the results are computed for a single instance of the PA model ($\gamma = 1, b = 1$), and a single set of samples.

case ($\gamma = 1, b = 1$). Based on these experiments, we always use the snowball proposal distribution and resample only on homogeneous trees $b < 1$.

Note that these results are somewhat unexpected, in two different ways. First, our analysis suggests that the problems afflicting SMC (path degeneracy, see Sec. 5.4.2) can outweigh its benefits in the context of network archaeology. Second, our analysis shows that while the truncated posterior proposal distribution makes use of more "immediate" knowledge of the model, it ultimately leads to *worst* results than a proposal distribution that does not even try to match the posterior $P(X_{0:T-1}|G, \gamma, b)$.

5.4 Additional results

5.4.1 Evolution of the effective sample size

As highlighted in the main text, SIS often becomes dominated by a few high weight samples, something that can be quantified with the *effective sample size* [25]

$$\mathrm{ESS}(\{x_i\}|G, \gamma, b) := \frac{\left[\sum_{i=1}^{n} \omega(x_i|G, \gamma, b)\right]^2}{\sum_{i=1}^{n} \omega(x_i|G, \gamma, b)^2} , \tag{30}$$

where $\omega(x_i|G, \gamma, b)$ is the unnormalized weight of history i. An ESS close to n indicates that all histories contribute equally.

In Fig. 9, we show a typical example of the evolution of the ESS as we sequentially add on more samples. The ESS grows slowly, and dips abruptly whenever a new high weight sample is generated. Importantly, the rate of increase of the ESS slows down after each dip, because the slowly growing set of high weight samples become harder and harder to "outcompete." This implies that obtaining a large effective sample size is impossible with the SIS; a few histories will invariably dominate the others.

The SMC method attempts to mitigate the problem by resampling the histories whenever the ESS becomes too low [26] (or even systematically [20]). In Fig. 10, we show a few typical ESS trajectories for a number of variants of the SMC, now as a function of the time-step t, since SMC calls for a fixed number of histories evolved in parallel. Our results show that while the ESS is equal to n right after the histories are resampled (by construction, not shown), the ESS actually decreases quite fast if the histories are left to evolve without resampling (yellow and orange curves). Hence constant resampling is needed to maintain some form of homogeneity. We note that the ESS peaks at a similar value regardless of the specific resampling threshold (n vs. $n/2$). This suggests that constant resampling is probably not needed, and that SMC and adaptive SMC actually achieve similar results (an observation that is also supported by the similar convergence results of Fig. 8, notice how the blue and yellow lines are nearly identical in the limit of large n). We also note that

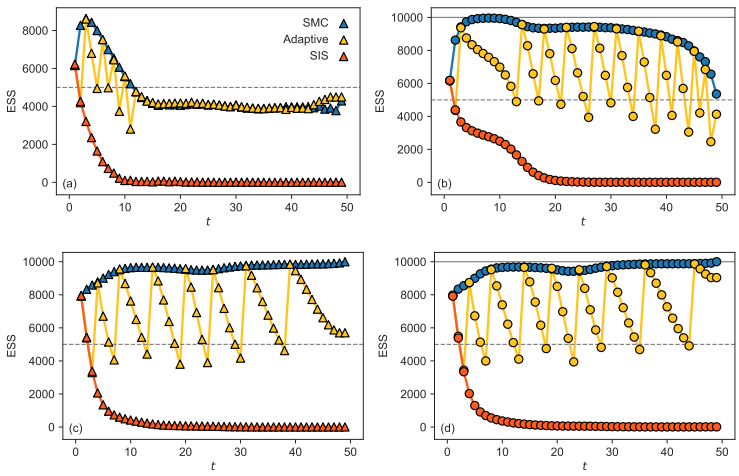

Figure 10: **Effective sample size in SMC.** Curves correspond to a single SMC sampling run with $n = 10\,000$ parallel samples, generated for a single instance of the model $b = 1$ and (a,b) $\gamma = 1$, (c,d) $\gamma = 0$. We use the snowball proposal distribution in (a,c) and the truncated posterior proposal distribution in (b,d). For SMC (blue markers), the set of histories is resampled at every step. For adaptive SMC (yellow markers), the histories are resampled whenever ESS $< n/2$ (dotted horizontal line). Finally, for SIS (orange markers), the histories are never resampled. Note that the ESS (see Eq. (30)) is calculated *before* any potential resampling step is carried out.

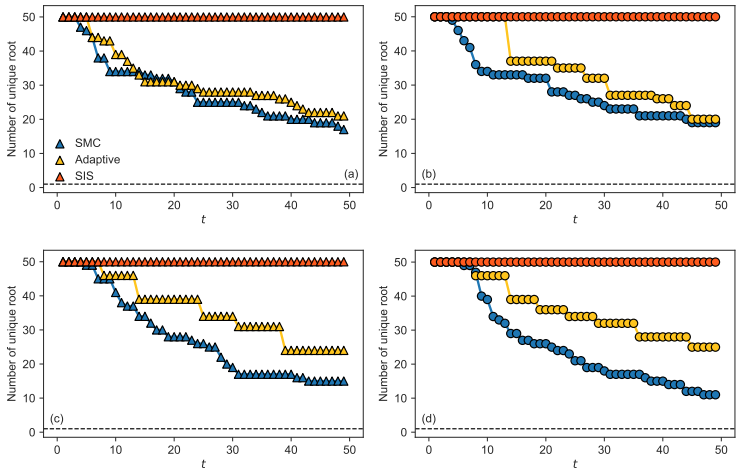

Figure 11: **Path degeneracy in SMC.** Number of distinct roots in the set of histories $H(t)$, for the simulations shown in Fig. 10. We use the same layout and color code. The minimum is indicated with a dotted line.

it is possible to achieve good results with a small ESS (SIS is on par with the other methods at $\gamma = 1, b = 1$, yet it has a small ESS).

5.4.2 Path degeneracy

The so-called path degeneracy effect [21] potentially explains why we do *not* necessarily get better results with SMC or adaptive SMC. Path degeneracy refers to the idea that resampling sometimes erases histories that could have evolved towards a high weight state, because it is "blind" to what happens down the road. Hence, SMC effectively trades variances for bias.

To investigate path degeneracy, we plot in Fig. 11 the number of distinct edges acting as the root, as a function of t. Because the number of samples n is much larger than T, every edge is represented as a root in the initial set of histories $H(t = 0)$, with high probability. However, as we let time evolve and resample, the roots associated with low weights $\omega(X_{0:t}; G, \gamma, b)$ are removed from $H(t)$ (unless the number of samples n is infinite / very large). As a result, the SMC algorithm is assigning an empirical probability of zero to a large portion of the posterior distribution. This is fine if these histories *actually* turn out not to carry much probability mass (in our experiment: on trees), but path degeneracy can also have disastrous effects (in our experiment: on loopy graphs).

In general we see that adaptive resampling alleviates some of the path degeneracy while maintaining an ESS after resampling that is similar to systematic resampling. This explains the better performance of adaptive SMC over pure SMC at low n (see Fig. 8).

5.4.3 Convergence: detailed convergence analysis in the absence of resampling

The "averaged results" of Fig. 8 should not be taken as a proof of the dominance of the sampling method over the OD on an *instance-by-instance* basis, even given infinitely many samples. Indeed, it is unclear whether the sampling method would systematically outperform the OD given a perfect knowledge of the posterior distribution $P(X|G, \gamma, b)$. The optimality proof of Sec. 3 only guarantees that the correlation is maximized on average, over the set of all histories consistent with some graph G. In other words, the proof relies on an average notion of optimality, rather than one couched, e.g., in the languages of worst-case guarantees. As a consequence, there could be instances of the model for which the OD works better than the perfect MMSE estimators (or their numerical approximations). The relevant question, then, is whether these instances are *common* or so *unlikely* that they will never be observed in practice. The latter scenario would lead to a practical dominance of the MMSE method. The results of Fig. 12 suggest that reality lies somewhere in between: There are indeed some instances where it is hard—perhaps impossible—to outperform the simple OD method, but in the majority of the cases, well approximated MMSE estimators triumph.

This could be better settled by running these experiments with even more samples and independent graphs. The sample size used here is relatively small, due to computational constraints.

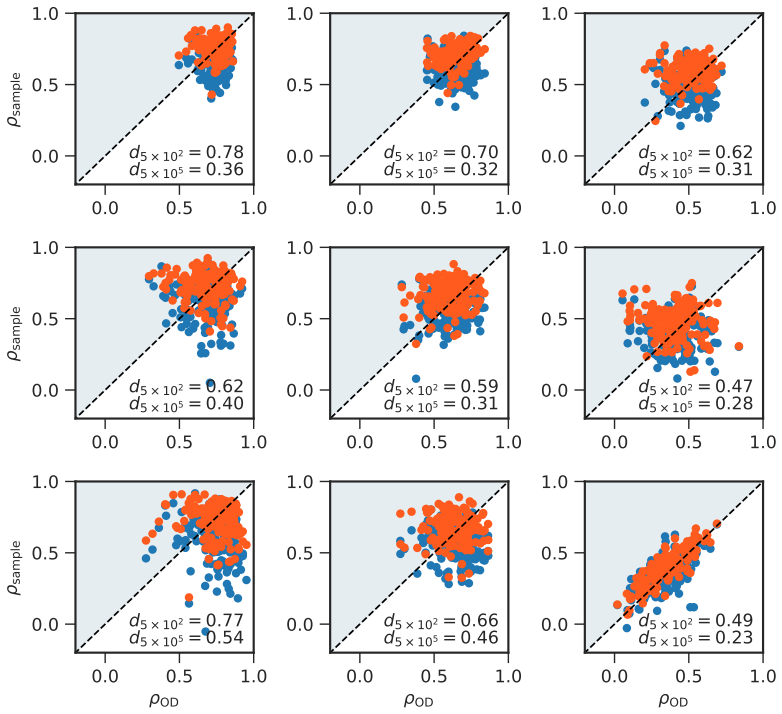

Figure 12: **Inference quality on an instance-by-instance basis.** Instance-versus-instance comparison of the results appearing in Fig. 8. Each dot corresponds to a specific network realization. We show the correlation ρ_{OD} achieved by the onion decomposition versus the correlation ρ_{sample} achieved by the sampling method (here without resampling, using the snowball proposal distribution), with 5×10^2 samples (blue) and 5×10^5 samples (orange). The fraction d_n of instances for which $\rho_{\text{OD}} > \rho_{\text{sample}}$ appears in each figure. Points in the shaded region denote instances where the sampling outperformed the onion decomposition.

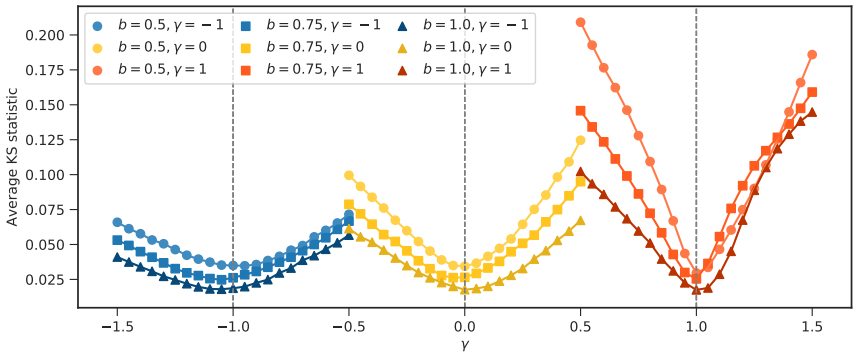

Figure 13: **Convexity of the average KS–statistic.** Each curve is associated with a single random graph, of $T = 1\ 000$ edges, with generating parameters (γ^*, b^*) indicated by the color and shape of the markers. For every instance G, we compute the average KS–statistic of the empirical degree distribution $P(G)$ and of $n = 1\ 000$ random degree distributions, generated with parameters $(\gamma, \hat{b}(G))$ in the range $\gamma \in \left[\gamma^* - \frac{1}{2}, \gamma^* + \frac{1}{2}\right]$. The figure confirms that the average KS–statistic is convex and centered near the generating parameters (indicated by vertical lines).

6 Details of the parameter estimation procedure

The statistical inference approach introduced in the main text is developed with the explicit assumption that the parameters of the generative model are either known, or correctly estimated prior to any archaeology steps. We characterize this estimation step, and show that the parameters are indeed recoverable in large instances, by way of simple calculations that do not rely on the full posterior distribution of the model.

6.1 Node creation probability

The quantity b determines whether new edges involve two existing nodes (prob. $1 - b$) or an existing node and a new node (prob. b). We therefore term it the "node creation probability."

Our estimator of b relies on the observation that a graph G can be seen as a signature of $|E(G)| - 1$ i.i.d. Bernoulli trials of parameter b. Indeed, each edge beyond the first embodies a test of whether a new node should be added, and each node beyond the two initial nodes signals a success of the trial. Therefore the classical theory of estimation applies to the node creation probability. We opt for the simple maximum a posteriori (MAP) estimator

$$\hat{b}(G) = \frac{|V(G)| - 2}{|E(G)| - 1}\,, \tag{31}$$

valid in the large graph limit or when the prior on $\hat{b}(G)$ is uniform. Note that upon generalizing the model to different seed graphs G_0 [27], the asymptotic estimator would instead read:

$$\hat{b}(G) = \frac{|V(G)| - |V(G_0)|}{|E(G)| - |E(G_0)|}\,. \tag{32}$$

6.2 Kernel exponent

The exponent γ of the attachment kernel controls the degree heterogeneity. Our goal is to learn this exponent network structure alone, without using temporal data.

The degree distribution is technically a sufficient statistic for γ only once we condition on the arrival times [20, 28], meaning that we should, in full rigor, sample from the history distribution to evaluate γ, perhaps along the line of [20]. But in practice, we know that the *final* degree distribution is well determined

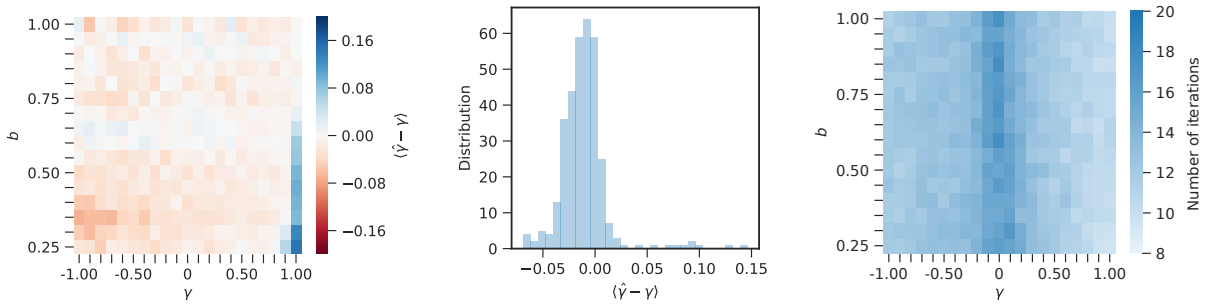

Figure 14: **Consistency of the estimator of γ in small networks. (left)** Difference between the true value and estimated value of γ, for networks of $T = 1\,000$ edges generated with parameters $\gamma \in [-1, 1]$ and $b \in [0.25, 1.00]$. The difference $(\gamma - \hat{\gamma})$ is averaged over 20 different network realizations at each point of the parameter space. **(center)** Distribution of the values appearing in the left panel. The error $(\gamma - \hat{\gamma})$ is close to zero on average, with a small systematic negative bias. However, we note that the estimator is strongly biased at $\gamma = 1$ with $b \ll 1$. This effect can be traced back to the poor fit between the model and its mean-field description, see Sec. 1.1. **(right)** Average number of evaluations needed to attain convergence in Brent's method. Exponents closer to 0 are harder to find. We use $n = 1\,000$ samples each time we evaluate the average KS–statistic.

by γ for all $\gamma < 2$, with many qualitative transitions at special values of γ, see main text and Ref. [2]. As a result, we can expect a good inference for γ if we simply match the observed empirical distribution with that of random instance of the models while tuning $\hat{\gamma}$.

This suggests the following non-Bayesian heuristic (based on a method introduced in Ref. [29]). Given some value of γ and the MAP estimator $\hat{b}$, we quantify the distance between the model and the data using the average Kolmogorov-Smirnov (KS) statistic of the empirical degree distribution $P(G)$ and of the degree distributions of random instance of the model $\{Q^{(i)}(\gamma)\}_{i=1,\dots,n}$. We obtain our estimator $\hat{\gamma}$ by minimizing this function with respect to γ. The average KS-statistic is convex with respect to γ, such that $\hat{\gamma}$ can be found efficiently via bisection or Brent's method [30] (see Fig. 13). The KS–statistic of a pair of noisy distributions (P, Q) is given by the supremum of the difference of their cumulative distribution function (CDF), i.e.,

$$D(P, Q) = \sup_k |f_P(k) - f_Q(k)| \,, \tag{33}$$

where $f_P(k)$ is the CDF of P at point k. In all our characterization tests shown in this document, we use $n = 1\,000$ samples to approximate the true average KS–statistic. When fitting real networks, we use $n = 100\,000$ samples, since we need not sweep the whole parameter space.

Because the minimization works correctly only when the KS–statistic is smooth with respect to γ, we need $n \gg 1$ samples to properly approximate it. Unfortunately, simulations become prohibitively expensive in the limit of large networks and large sample sizes. Hence, we opt for a more efficient solution: We first integrate the mean-field equations of the model (see Eqs. (1)), and then draw n finite samples from the distributions. This method is equivalent to—but much faster than—n repeated simulations. The consistency of the estimators is investigated in Fig. 14

6.3 Goodness of fit

The estimation framework provides a natural test of the goodness of fit of γ [29]. The procedure goes as follows. After we have found $\hat{\gamma}$ via the minimization of the average KS–statistic, we keep the corresponding minimum KS–statistic $D^*(G)$ in memory. We then generate n_{bs} random degree distributions $Q^{(i)}$ from the model of parameters $(\hat{\gamma}, \hat{b})$ and compute the distribution of their KS–statistic, by comparing each of them against n additional random degree distributions $\{S^{(j)}\}_{j=1,\dots,n}$. This provides a null distribution for D,

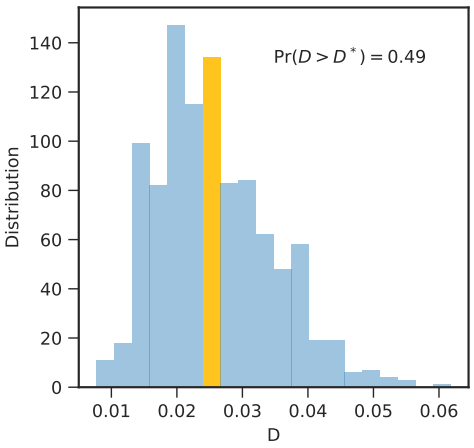

Figure 15: **Verification of the validity of the test of the goodness of fit (on artificial network).** To ensure that our test of the goodness of fit works properly, we check whether it lends evidence to the generative model when it is applied to a graph generated by the model itself. We use an artificial graph $G = (V, E)$ generated with the parameters $(T = 1\,000, \gamma = 0.7, b = 0.9)$ as our test dataset. We begin by inferring the parameters back, mimicking a typical situation where the source of G is unknown. For the particular graph used to produce the figure, the estimated parameters are $\hat{\gamma} = 0.643$, $\hat{b} = 0.893$, and they are associated with an average KS–statistic of $D^* = 0.025$. We then turn to the test of the goodness of fit [29]. We first draw $n_{bs} = 1\,000$ random degree distributions $\{Q^{(i)}\}_{i=1,\dots n_{bs}}$ from the model (of parameters $\hat{\gamma}$,$\hat{b}$), and generate $n = 1\,000$ additional degree distributions $\{S^{(j)}\}_{j=1,\dots,n}$ for each $i = 1, \dots, n_{bs}$. Then, for each i, we approximate the expected KS–statistic by averaging $D(Q^{(i)}|S^{(j)})$ over all $\{S^{(j)}\}_{j=1,\dots,n}$. This last step yields the distribution shown in the figure (with the bin where D^* falls highlighted). Here, the test of the goodness of fit shows that the KS–statistic obtained by minimization is at least as extreme as the one associated with roughly 50% of the samples, providing strong evidence for the model with parameters $(\hat{\gamma}, \hat{b})$, as expected.

which tells us whether $D^*(G)$ is an extreme value of the average KS–statistic or not. See Fig. 15 and its caption for an example.

6.4 Sensitivity analysis

For the sake of simplicity, we have treated (γ, b) as nuisance parameters and conditioned the posterior on their estimates throughout the text. In our analysis of the phase transition, our focus is history recovery for a *sequence of models* characterized by different γ, such that this conditioning is fine. "Polluting" the inference with parameter uncertainty would only muddy our analysis of the true cause for the transition (edge equivalence). That said, given a *real* system, one needs a principled way of handling parameter uncertainty. A *complete* Bayesian perspective is one way to do so, see e.g. the particle MCMC of Ref. [20].

However, as we show in Fig. 16, a simple sensitivity analysis confirms that the temporal estimates are robust to parameters misspecification. Because the parameter estimators also appear robust (see Ref. 14), we can safely conduct inference by conditioning on point estimates. The results of a much costlier full Bayesian inference process would lead to better error estimates.

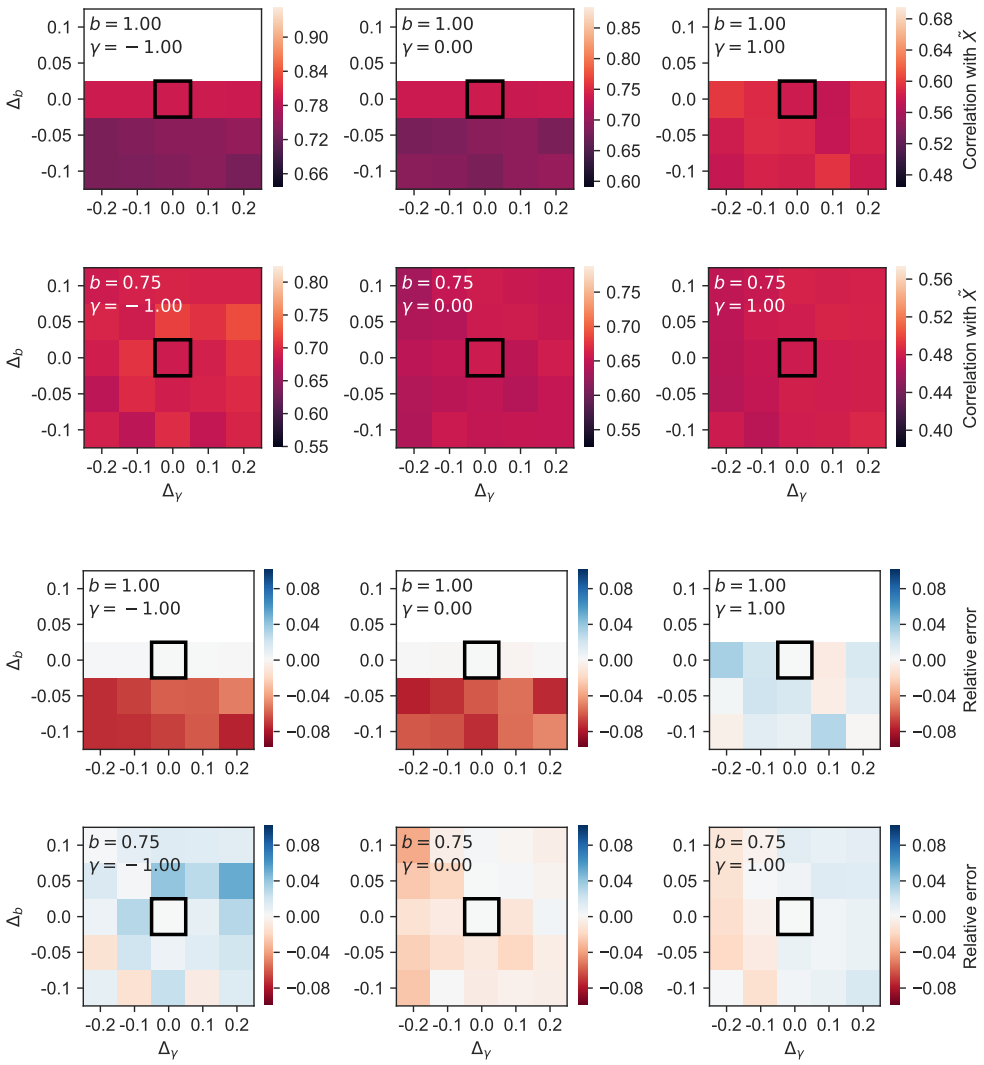

Figure 16: **Sensitivity analysis of the parameter estimators.** Average correlation with the ground truth (top 6 panels) and relative error on the true correlation (bottom 6 panels) of the histories estimated by running the inference with misspecified parameters. These results are produced on trees ($b = 1$, top rows) and on loopy graph ($b = 0.75$, bottom rows), at various level of heterogeneity ($\gamma \in \{-1, 0 + 1\}$ from left to right), also see inset text. The bold square shows the reference value (no perturbation) while the rest of the heatmap shows results for absolute perturbations of magnitude $\Delta_\gamma, \Delta_b \in \{-0.2, -0.1, 0.1, 0.2\}$. Note that the error on (γ, b) are typically much smaller than these perturbations, see Fig. 14. In particular we are not likely to make such a large mistake on b when the network is actually a tree: any $b < 1$ will produce at least an edge with probability that grows exponentially in T. Estimates above $\hat{b} = 1$ are impossible and therefore not computed. We use the snowball proposal distribution, and the sampling level most appropriate to each case (see Fig. 5.3) and $n = 100\,000$ samples. The results are averaged over 35 network instances.

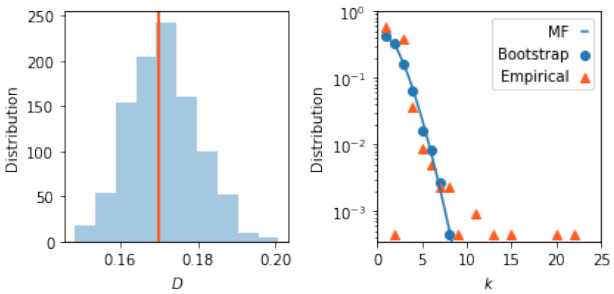

Figure 17: **Goodness of fit of the degree distribution (ebola network). (left)** Null distribution for the average KS–statistic. The value D^* associated to the best fit is indicated with an orange vertical line. This distribution is estimated with $n_{bs} = 1000$ random degree distribution, each compared against $n = 100$ additional distributions to compute the average KS-statistic. **(right)** Degree distribution the network (orange triangle) compared with the mean-field distribution of the best fit (solid line) and one bootstrap samples of this distribution (blue circles).

7 Nextstrain Ebola dataset: details

The phylogenetic tree of the Ebola virus is updated in real time and available online from https://nextstrain.org/ebola. The date at which every strain was sequenced is available as metadata, as well as the time of emergence of the common ancestors, inferred to within a confidence interval that ranges from a few days to many weeks [31]. We use the most likely date when the time of emergence is uncertain. When it was archived for the purpose of the present study, the dataset comprised 1238 unique sequenced strains, connected by 2196 mutations and 959 inferred common ancestors. We have made the archived copy used in the main text available online at github.com/jg-you/network_archaeology/data/.

The dataset already comes in the form of a time-ordered tree (we drop the exact *date* of emergence, and focus on the ordering). To estimate the parameters, we run the procedure described in the Materials and Methods as well as in Sec. 6, see Fig. 17. We find that the Ebola phylogenetic tree is best modeled by $(\hat{\gamma}, \hat{b}) = (-0.71464, 1)$, associated with an average KS–statistics of $D^* = 0.17005$. Using $n_{bs} = 1000$ random bootstrap degree distribution, each compared against $n = 100$ additional distributions, we find that the D^* is significant, with $P[D > D^*] = 0.53$ [29] (see Fig. 17). Running the inference methods on the full network, we obtain correlations of $\rho_{\text{degree}} = 0.152$, $\rho_{\text{OD}} = 0.150$, $\rho_{\text{MMSE}} = 0.456$ (with only 25 000 SMC samples).

References

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

[2] P. L. Krapivsky, S. Redner, and F. Leyvraz, "Connectivity of growing random networks," *Phys. Rev. Lett.*, vol. 85, no. 21, p. 4629, 2000.

[3] R. Albert and A.-L. Barabási, "Topology of evolving networks: local events and universality," *Phys. Rev. Lett.*, vol. 85, no. 24, p. 5234, 2000.

[4] T. Pham, P. Sheridan, and H. Shimodaira, "PAFit: a statistical method for measuring preferential attachment in temporal complex networks," *PLoS One*, vol. 10, no. 9, p. e0137796, 2015.

[5] P. L. Krapivsky, G. J. Rodgers, and S. Redner, "Degree distributions of growing networks," *Phys. Rev. Lett.*, vol. 86, no. 23, p. 5401, 2001.

[6] S. N. Dorogovtsev and J. F. F. Mendes, "Scaling behaviour of developing and decaying networks," *Europhys. Lett.*, vol. 52, no. 1, p. 33, 2000.

[7] W. Aiello, F. Chung, and L. Lu, "Random evolution in massive graphs," in *Handbook of Massive Data Sets*, pp. 97–122, Springer, 2002.

[8] 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*, vol. 85, no. 2, p. 026108, 2012.

[9] J.-G. Young, L. Hébert-Dufresne, A. Allard, and L. J. Dubé, "Growing networks of overlapping communities with internal structure," *Phys. Rev. E*, vol. 94, no. 2, p. 022317, 2016.

[10] P. S. Dodds, D. R. Dewhurst, F. F. Hazlehurst, C. M. Van Oort, L. Mitchell, A. J. Reagan, J. R. Williams, and C. M. Danforth, "Simon's fundamental rich-get-richer model entails a dominant first-mover advantage," *Phys. Rev. E*, vol. 95, no. 5, p. 052301, 2017.

[11] R. Patro, E. Sefer, J. Malin, G. Marçais, S. Navlakha, and C. Kingsford, "Parsimonious reconstruction of network evolution," *Algorithm Mol. Biol.*, vol. 7, no. 1, p. 25, 2012.

[12] P. L. Krapivsky and S. Redner, "Statistics of changes in lead node in connectivity-driven networks," *Phys. Rev. Lett.*, vol. 89, no. 25, p. 258703, 2002.

[13] L. A. Adamic and B. A. Huberman, "Power-law distribution of the World Wide Web," *Science*, vol. 287, no. 5461, 2000.

[14] A. Magner, A. Grama, J. K. Sreedharan, and W. Szpankowski, "TIMES: temporal information maximally extracted from structure," in *Proceedings of the 2018 World Wide Web Conference (WWW)*, pp. 389–398, 2018.

[15] M. Drmota, *Random trees: An Interplay Between Combinatorics and Probability*. New York: Springer, 2009.

[16] 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)*, pp. 1563–1567, IEEE, 2017.

[17] B. H. Erickson, "Some problems of inference from chain data," *Sociol. Methodol.*, vol. 10, pp. 276–302, 1979.

[18] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Phys. Rev. E*, vol. 73, no. 1, p. 016102, 2006.

[19] M. S. Handcock and K. J. Gile, "Comment: on the concept of snowball sampling," *Sociol. Methodol.*, vol. 41, no. 1, pp. 367–371, 2011.

[20] B. Bloem-Reddy and P. Orbanz, "Random-walk models of network formation and sequential monte carlo methods for graphs," *J. Royal Stat. Soc. Series B*, vol. 80, no. 5, pp. 871–898, 2018.

[21] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later,"

vol. 12, p. 3, 2009.

[22] A. Slepoy, A. P. Thompson, and S. J. Plimpton, "A constant-time kinetic monte carlo algorithm for simulation of large biochemical reaction networks," *J. Chem. Phys.*, vol. 128, no. 20, p. 05B618, 2008.

[23] D. Shah and T. Zaman, "Rumors in a network: who's the culprit?," *IEEE T. Inform. Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[24] J. Blitzstein and P. Diaconis, "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet Math.*, vol. 6, no. 4, pp. 489–522, 2011.

[25] J. S. Liu, "Metropolized independent sampling with comparisons to rejection sampling and importance sampling," *Statistics and Computing*, vol. 6, no. 2, pp. 113–119, 1996.

[26] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, no. 1-2, 2003.

[27] G. Lugosi, A. S. Pereira, *et al.*, "Finding the seed of uniform attachment trees," *Electronic Journal of Probability*, vol. 24, 2019.

[28] F. Gao and A. van der Vaart, "On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees," *Stoch. Process. Appl.*, vol. 127, no. 11, pp. 3754–3775, 2017.

[29] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[30] W. H. Press, *Numerical Recipes: The Art of Scientific Computing*. Cambridge University press, 3rd ed., 2007.

[31] P. Sagulenko, V. Puller, and R. A. Neher, "Treetime: Maximum-likelihood phylodynamic analysis," *Virus Evol.*, vol. 4, no. 1, p. vex042, 2018.