Supporting Information

Golub and Jackson 10.1073/pnas.1000814107

SI Text

In this appendix, we show how a simple process of independent decisions about whether to send a chain letter, along with two forms of bias introduced by observation, can give rise to the local and global behavior of the chain letter trees reconstructed by Liben-Nowell and Kleinberg. The two forms of observation bias are driven by the following phenomena: (*i*) Only some copies of a chain letter are posted where they can be collected, and (*ii*) a chain letter is considered to be "observed" only when the part of it recovered through these observations is reasonably large.

The Process. Our process for generating trees can be described in several steps.

First, we generate a "true" underlying tree (which may or may not end up being observed, and then only partially) via a Galton–Watson branching process with a binomial offspring distribution. In particular, we begin with a root node and follow the following procedure iteratively for each node. For some integer $d > 0$ and scalar $q \in (0,1)$, the probability that a node has k children is

$$\binom{d}{k} q^k (1-q)^k.$$

The scalar q represents the chance that a given node that has been sent the letter decides to sign the letter and send it on. A node that signs and forwards the letter then sends it to d other nodes who have not received it before.* Each of those nodes then independently decides whether to sign and forward it. Thus, k is the (random) number of recipients of a given sender's letters who will continue the chain. In our simulations, the first step is to generate a random true tree $T = (V,E)$ by following this procedure.

The observed tree is not the same as the true one. To get the observed tree, called T', we randomly sample nodes from the true tree (corresponding to the type of Web search for chain letter instances performed by Liben-Nowell and Kleinberg) at a sampling rate s. That is, each node of the true tree is included in the sampled set S with probability s, independently of other nodes.

Given that sample, we reconstruct as much of the true tree as we can to get the "reconstructed" tree $T' = (V',E') \subset T = (V,E)$, which we also call an observed tree. Formally, given a sample S of nodes from the true tree, we go through every node $v \in S$ and include in the vertex set V' of the observed tree all the nodes $w \in V$ such that w is an ancestor of v in the true tree T, as well as v itself. Then T' is the graph induced on V' by T. Intuitively, this procedure corresponds to discovering the ancestors of each observed node $v \in S$ by looking at the petition in the corresponding instance of the chain letter and then reconstructing the tree as best we can from that information.

Last, we condition on the reconstructed tree being of the right size, as in the main analysis in the paper. That is, we throw away T' unless $2{,}442 \leq |V'| \leq 3{,}250$. The reconstructed trees in this size range form the output from the simulation. We can then examine whether particular values of d and q generate observed trees consistent with the actual observed chain letters.

Parameters. For a given true tree having $n \geq 70$ nodes, we choose $s = 70/n$. This is because, in the reconstructed tree that we focus on [the 2,442-node National Public Radio (NPR) petition component], there were 70 letters that were directly observed, as opposed to being inferred. Thus, we set the sampling rate to generate about this number of sampled nodes.

We chose $d = 30$ and $q = \frac{2{,}441/2{,}442}{30}$. The choice of d, the number of contacts to whom every activated node sends the petition, is essentially arbitrary but seems to us a reasonable estimate of the number of people to whom a typical sender would direct the letter and ends up working well. The choice of q is determined by a rough method of moments calculation. We know that in any tree of n nodes, the expected number of children per node† is $(n-1)/n$. On the other hand, for this binomial process the expected number of children per node in the true tree (without any observation bias) is qd. Thus, we chose q to satisfy

$$dq = (n-1)/n \quad \text{for } n = 2{,}442$$

which is the size of the observed tree that we focus on from the Liben-Nowell and Kleinberg data. Our selection of these parameters is somewhat ad hoc: We did some experiments to explore the parameter space and found, after a fairly short search, that these worked well. However, this selection of parameters could be motivated and performed more carefully—for example, by using maximum-likelihood methods or a more precise method of moments approach. Our purpose here is simply to demonstrate that this type of process can match the data closely. Nevertheless, it seems quite important that dq, the expected number of children per node, be just a little below 1. If it is much below it, then it is extremely rare to get trees of a large size, and if it is bigger than 1, then the trees have the wrong shapes; additionally, their tendency to grow infinite in this case makes simulating them a challenge.

Results. We generated 10,000 reconstructed trees by using the procedure outlined above with these parameters and studied their properties as in the main paper. Histograms of median depth and width are shown in Fig. S1. A two-dimensional density plot is depicted in Fig. S2. These figures show that these global statistics of the simulated trees match the corresponding statistics of the real trees fairly closely. A region of a typical observed tree generated by the simulation is shown in Fig. S3.

In terms of local behavior, the simulated reconstructed trees are also quite similar to those that were reconstructed in the empirical exercise performed by Liben-Nowell and Kleinberg. To explore this similarity more precisely, we computed the empirical offspring distribution of each simulated tree and compared it against the empirical offspring distribution of the NPR tree with 2,442 nodes. The notion of distance we used was the total variation norm, under which the distance between two probability distributions π and σ over a countable set X is

$$\|\pi - \sigma\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in X} |\pi(x) - \sigma(x)|.$$

The set X here is the set of possible numbers of children $X = \{0,1,2,\ldots\}$.

The total variation distance between the offspring distribution in the simulations and the offspring distribution of the 2,442-node NPR tree was, on average, 0.0058. The standard deviation of this statistic across the simulations was 0.0038. The maximum devia-

*Here d and q are the same across all nodes. We could enrich the model by allowing different nodes to choose different numbers of nodes to forward the letter to. Effectively, given the randomness in the number who subsequently pass it on, if this was done in a binomial manner, it would simply result in a redundant parameter.

†That is, if a node is drawn from the tree uniformly at random.

tion observed was 0.038. Thus, the offspring distributions in the simulations closely matched the real one.

Discussion. This exercise shows that a simple process can give rise to the global and local patterns observed in the data. It is worth noting that the step of going from the true tree to the reconstructed tree is crucial. Fig. S4 compares the offspring distribution in a typical simulated true tree (black bars) with that of the tree reconstructed from it after sampling (white bars). The observation process introduces a heavy bias toward nodes with one child. The reason for this bias is simple. There are some rare nodes in the true tree that produce many children—as many as 8. However, such a node only ends up having many children in the reconstructed tree if many of its children are either observed directly or have descendants that are eventually observed; both of these events are unlikely. So, even though nodes that have many children are more likely to end up appearing in the reconstructed tree than a node that truly has few children, they will still tend to produce few branches from which anything is sampled and so will tend to have a small number of visible children in the reconstructed tree.

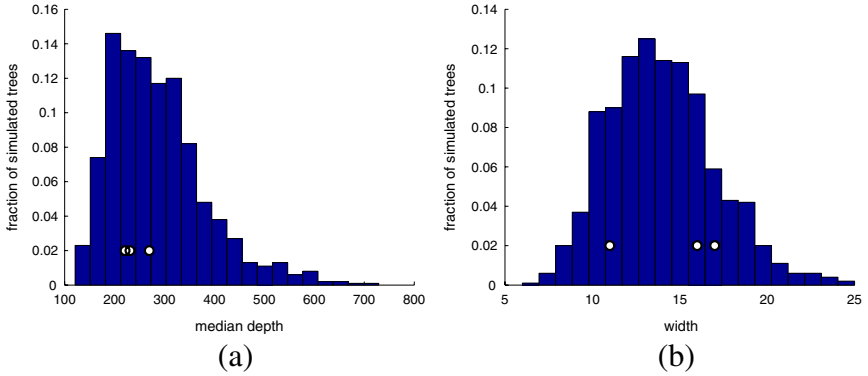

Fig. S1. Histograms of statistics from the simulated trees. The white circles correspond to the three observed components of the NPR chain letter data. (*a*) depicts median depth and (*b*) depicts width.

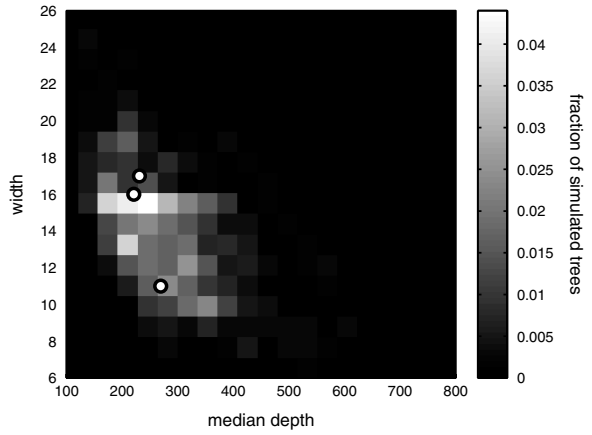

Fig. S2. The joint distribution of median node depth and width from the simulated trees. The white circles correspond to the three observed components of the NPR chain letter data.

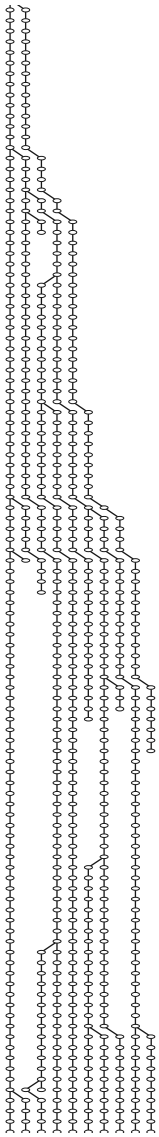

Fig. S3. Part of a typical reconstructed tree generated by the simulation procedure.

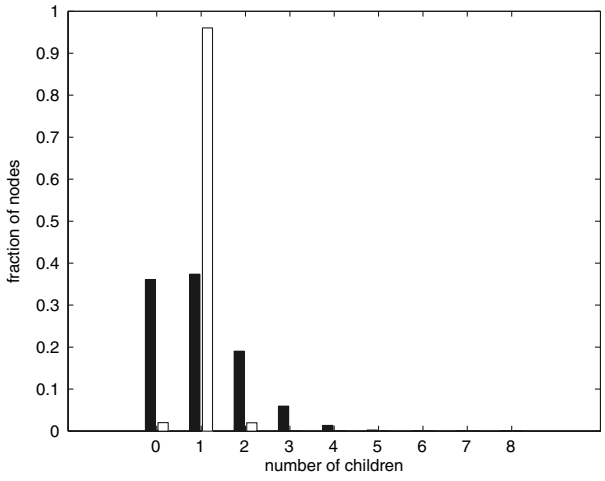

Fig. S4. A comparison of the offspring distribution of a typical true tree from one of the simulations (*Black Bars*) with that of the reconstructed tree that would be observed (*White Bars*).