Using selection bias to explain the observed structure of Internet diffusions

Benjamin Golub[a,1] **and Matthew O. Jackson**[b,c]

[a]Graduate School of Business, Stanford University, Stanford, CA 94305; [b]Department of Economics, Stanford University, Stanford, CA 94305; and [c]External faculty of the Santa Fe Institute, Santa Fe, NM 87501

Edited by Ronald L. Graham, University of California, San Diego, La Jolla, CA, and approved May 11, 2010 (received for review January 29, 2010)

Recently, large datasets stored on the Internet have enabled the analysis of processes, such as large-scale diffusions of information, at new levels of detail. In a recent study, Liben-Nowell and Kleinberg [(2008) *Proc Natl Acad Sci USA* 105:4633–4638] observed that the flow of information on the Internet exhibits surprising patterns whereby a chain letter reaches its typical recipient through long paths of hundreds of intermediaries. We show that a basic Galton–Watson epidemic model combined with the selection bias of observing only large diffusions suffices to explain these patterns. Thus, selection biases of which data we observe can radically change the estimation of classical diffusion processes.

diameter | chain letters | Galton–Watson process | maximum likelihood estimation | social networks

As social network data become increasingly available in electronic form, researchers are developing more detailed and accurate pictures of the patterns of social interactions. These empirical investigations are of primary importance given the multitude of ways in which social networks affect our lives (1). However, such data come with their own idiosyncrasies. Most notably, in the past most data on social networks were obtained via questionnaires (2), interviews (3), experiments (4), or observations directly made by researchers (5), and so it was the researcher who chose the data. More recently, the availability of electronic data has made it more common for the data to choose the researcher. That is, often large and interesting datasets become available because of the electronic storage of various forms of interaction occurring via the Internet, and these then become useful test beds for theories of social networks. In this paper, we focus on the explanatory power of one inherent selection bias that comes along with many such datasets. Specifically, we examine a selection bias that arises from looking at unusually *large* instances of diffusion processes—with a particular application to Internet chain letters.

In a recent paper (6), Liben-Nowell and Kleinberg provided an important and interesting examination of two chain letters that had wide circulation on the Internet: a petition in support of public radio and television that began circulating in 1995 and a petition against the eventual war in Iraq that circulated in 2002 and 2003. By obtaining many copies of the e-mails and tracing through the ordered lists of names added to each petition, Liben-Nowell and Kleinberg were able to reconstruct large portions of the trees of dissemination of these chain letters. The remarkable aspect of Liben-Nowell and Kleinberg's findings is that these trees do not exhibit the short distances between nodes that are characteristic of many social networks (7, 8). Instead, these trees have very small widths (i.e., many nodes have a single offspring), and the median node receives the letter after it has been through hundreds of intermediaries.

To understand why the paths of chain letter dissemination that Liben-Nowell and Kleinberg reconstruct are puzzling, let us discuss what seems to be the most natural and simple model of how such a process would operate. That model is the classical one of Galton and Watson (9), which was developed in the 1870s to study the longevity of family names in a patrilineal system. Galton and Watson proposed a branching process where each node has a random number of children, drawn independently according to the same distribution. The process can also serve as a model of an epidemic, where the number of children is the number of others a given node infects. In this application, the number of children of a given sender is the number of other people who sign the petition immediately below that sender's signature in various versions of the letter that branch off as various recipients sign. It is well known (10) that the key quantity in characterizing the asymptotic properties of this process is the expected number of children per node. If this quantity is below the threshold of unity, then the process is called subcritical and with probability one it will end in extinction after a finite number of steps. If the expected number of children is more than one, then the process is called supercritical and will continue forever with positive probability. (We ignore the nongeneric borderline case in which each node has exactly one child in expectation.)

The puzzle is that neither regime seems to explain the observed data. The two datasets that Liben-Nowell and Kleinberg study have more than 10,000 nodes each, whereas it is quite rare for a subcritical branching process with reasonable parameters to have more than a dozen nodes. Thus, the typical subcritical tree is a poor match to the data on many dimensions. On the other hand, if one tries to fit the data with a supercritical process, then the trees that emerge have huge breadths, branch very frequently, and do not have the long chains that are observed in the data. In view of this, Liben-Nowell and Kleinberg developed a richer network-based model of chain letter distribution with two important features: asynchronous response times and group replies. The realizations of their process that are as large as the observed diffusions have the correct shapes.

We show that, despite their surprising appearances, the observed trees have a global structure that corresponds to a basic and classical process. In particular, the simple Galton–Watson epidemic model suffices to generate trees reaching many nodes yet having long chains as in the data. To show this, we first fit the parameters of a Galton–Watson process by using maximum-likelihood estimation on the basis of one of the trees inferred by Liben-Nowell and Kleinberg. Then we simulate the process and examine only the rare outcomes in which a chain letter with these parameters spreads as widely as those that were observed. Most realizations are very small and have virtually no chance of being observed; we are interested in the properties of those rare ones that are big enough to match the public radio and war petitions described above. Simulated outcomes from this conditional distribution match the real observations closely on global dimensions such as tree depth, width, and the distribution of children per node. The seeming obstacle discussed above—that neither of

Author contributions: B.G. and M.O.J. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: bgolub@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1000814107/-/DCSupplemental.

the regimes of the Galton–Watson process seems to fit when we look at the unconditional distribution—is overcome by conditioning our subcritical process on the rare event of growing large, as Liben-Nowell and Kleinberg also do in their model. The main difference between their approach and ours is that we do not explicitly model the network or detailed mechanics of the distribution process. We focus only on the random variable describing how many children each node has and on the selection bias. Those two ingredients alone suffice to produce a conditional distribution concentrated on trees with the right shapes.

Whereas the specifics of the model and analysis that follow are particular to the Galton–Watson process, the broader point is worth emphasizing. Large-scale network phenomena that we observe may not be typical instances of the processes that generated them but instead exceptional realizations. Although implications of selection bias are well understood in some settings, they have not traditionally played a significant role at the dataset level in social network analysis. Our study of the chain letters provides a particularly stark example of how this perspective can, in a simple model, explain a great deal about the observations. This points to the need for a richer theoretical understanding of how selection modifies the structure of important classical processes.

Model

We begin by stating our formal model of chain letter propagation and discussing the fitting of its key parameters. Let X be a Galton–Watson random tree generated by the branching process starting at one root where the probability of any node having k children is $p(k)$ and the distribution is identical across nodes. This distribution is the fundamental parameter in the model. It is a simple matter to fit it by using maximum-likelihood estimation given an observed tree. The key fact is that, because the number of children is conditionally independent and identically distributed across nodes (given the past), this fitting goes through even when there is a size selection bias in the trees that are observed.

Let $L(p;x)$ be the probability of observing a specific tree x under the model—that is, the probability that $X = x$. For any rooted tree x, let $f(k;x)$ refer to the total number of nodes in x with k children. It follows directly that $L(p;x) = \prod_k p(k)^{f(k;x)}$, and so the log-likelihood function is

$$\ell(p;x) = f(0;x)\log\left(1 - \sum_{k>0} p(k)\right) + \sum_{k>0} f(k;x)\log p(k).$$

Maximizing the log-likelihood with respect to p and denoting the maximizer $\hat{p}$, we see that if $f(k;x) = 0$, then $\hat{p}(k) = 0$; otherwise, setting the derivative with respect to $p(k)$ to be equal to 0 implies that, for every k,

$$\frac{f(k;x)}{f(0;x)} = \frac{\hat{p}(k)}{\hat{p}(0)}$$

[noting that $f(0;x) > 0$ for any finite tree]. Therefore, for every $k \geq 1$,

$$\hat{p}(k) = f(k;x) / \sum_k f(k;x).$$

In other words, the estimated probability of having k children is just the fraction of nodes with k children in the data. It is straightforward to verify that $\hat{p}(k)$ yields a global maximum of the likelihood function. The same procedure can be carried out on many trees at once.

It is worth noting that we did not explicitly model the observation process here—as Liben-Nowell and Kleinberg did—in which some but not all nodes post the chain letter on the Internet where it can then be found and its propagation traced back to the root. It turns out that this aspect of the process can be omitted without loss of generality. Formally, our approach corresponds to defining a node as an *observable* node—that is, a node that forwarded the letter *and* one of whose descendants posted a later version. As long as the number of children and the decision of whether to post are independent of each other and identically distributed across nodes, then the random variables describing how many *observable* children each node has also satisfy the assumptions of the simple Galton–Watson process, albeit with a different distribution.

We also did not explicitly include a model of the network over which the diffusion is happening. The reason is that the only thing that matters for the Galton–Watson process is the number of observable children that each node has, and our reduced-form approach focuses on this process rather than on the network. Although the network structure will certainly influence the offspring random variable, the mechanisms of that can be complicated. What kinds of network processes underlie p is an interesting question whose analysis is the focus of Liben-Nowell and Kleinberg's work mentioned earlier. In the *Discussion* and *SI Text*, we propose one micromodel that can generate our Galton–Watson process with an offspring distribution matching the data.

Results

We applied this fitting procedure to an example from the *SI Appendix* of ref. 6. Specifically, we used the National Public Radio (NPR) petition, whose observed portion had three components and used the function f for the component with 2,442 observed nodes. Of course, the real NPR dissemination tree presumably had only one component and the pieces in the reconstruction arose from an inability to reconstruct the tree all the way to its origin. We simply take one of the subtrees as a single instance of the diffusion process.

The distribution $\hat{p}$ that was estimated is reported in Table 1. Its expectation is $2{,}441/2{,}442 \approx 0.9996$. Indeed, it is immediate to verify on the basis of the formulas above that the estimated distribution $\hat{p}$ will have expectation $(n-1)/n$ in any tree of n nodes. The implications of this rather simple fact deserve some comment. It entails that maximum-likelihood estimation of the kind we perform above, on the basis of a finite tree, will always infer the process to be subcritical. The extent of subcriticality (the gap between the expected number of children and 1) will depend on the size of the observed tree. On the other hand, a confidence interval for the expected number of children per node *would* include values exceeding unity. In any case, this feature of the estimator—always estimating $(n-1)/n$ as the expected number of children—is an artifact of using a very simple time-homogeneous branching process. Including realistic features that limit the spread of a chain letter once it reaches a large size would, we conjecture, lead to a different maximum-likelihood estimator of this quantity, which could exceed unity even for finite trees.

After estimating the distribution, we simulated the branching process with this distribution and analyzed only realizations whose sizes were between those of the largest and smallest observed components in the NPR data—between 2,442 and 3,250 nodes. We generated 10,000 of these realizations. The most relevant histograms from the analysis are shown in Fig. 1. The statistics we compute for each tree are the median node depth (distance from the origin) as well as the width (maximum number of nodes at the same depth).

Table 1. The distribution of the number of children per node, estimated from the data

k	$\hat{p}(k)$
0	0.0246
1	0.9525
2	0.0217
3	0.0012
≥4	0

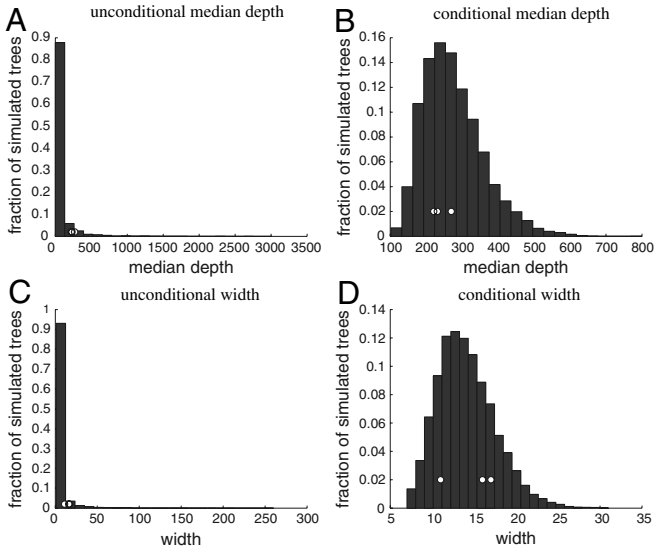

Fig. 1. Histograms of unconditional and conditional statistics from the simulated trees. (*A*) and (*B*) Median node depths. (*C*) and (*D*) Tree widths. The figures on the left, (*A* and *C*), correspond to the unconditional distribution, whereas the right ones, (*B* and *D*), come from the distribution conditional on the bounds on the number of nodes. The white circles correspond to the three observed components of the NPR chain letter data.

The key fact in these histograms is that, whereas the median depths and widths of the trees corresponding to the observed petitions are in the tails of the unconditional distribution of trees, they are centered within the distribution that conditions on containing the appropriate number of nodes. Fig. 2 takes this point further, showing that, after conditioning on having the appropriate number of nodes, the observed trees are near the thickest part of the joint distribution of median depth and width. A region of a typical simulated tree is shown in Fig. 3.

The third statistic that Liben-Nowell and Kleinberg report for the NPR data is the fraction of nodes with one child. In our simulations, conditional on having the appropriate number of nodes, this statistic was tightly clustered around 0.95 (mean 0.9527 and standard deviation 0.0040), exactly matching the data. This correspondence is unsurprising, because $\hat{p}(1)$, the estimated probability of having one child, was a parameter of the model that was fitted on the basis of the data.

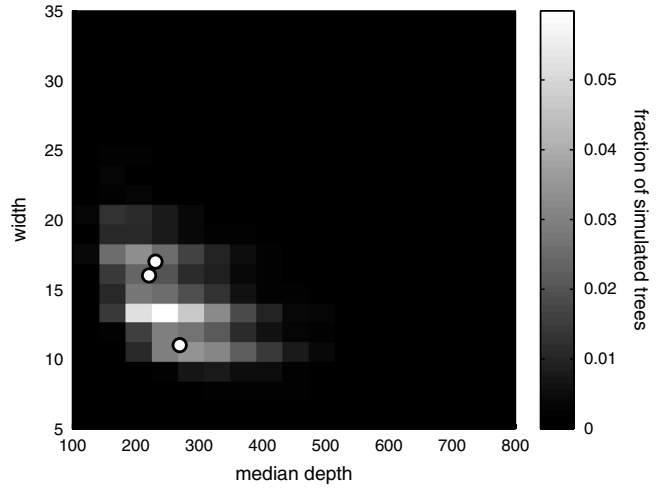

Fig. 2. The joint distribution of median node depth and width conditional on the size bounds. The white circles correspond to the three observed components of the NPR chain letter data.

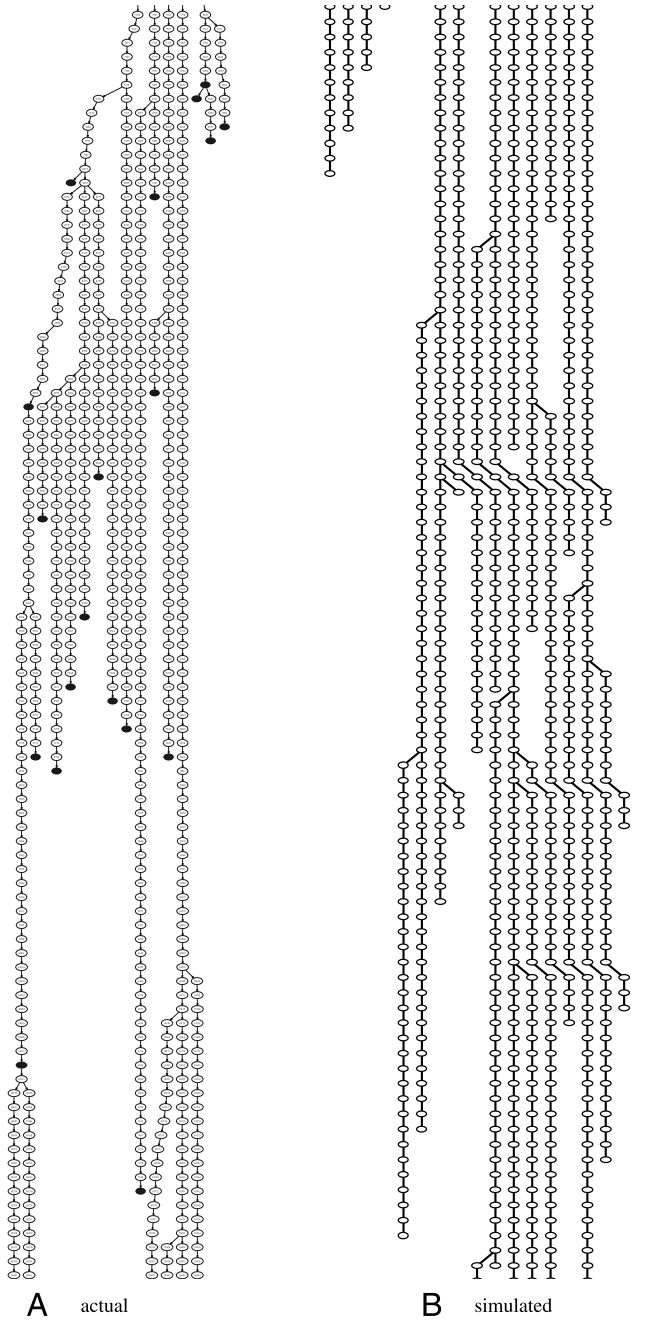

Fig. 3. A comparison of the real data and the simulations. (*A*) A piece of the real NPR petition propagation tree as reconstructed by Liben-Nowell and Kleinberg, and (*B*) a piece of typical tree from the conditional distribution in our simulations.

The conditional distribution is a small slice of the whole distribution, but not too small. Out of 10,000 simulated trees, 654 (6.54%) met or exceeded the lower size bound and 110 (1.10%) fell within both the upper and lower bounds. These statistics confirm the intuitive fact that wide dissemination is unlikely (even for this barely subcritical chain letter) but not impossible.

Why did we condition on both a lower and an upper size bound? The former, as discussed above, is explained by the fact that only sufficiently successful letters show up on researchers' radar, at least with the methods used in ref. 6. The upper bound is not important to the outcome of the analysis; when it is done without the upper bound, the histograms have fatter right tails but

the conclusions of our analysis remain unchanged. Nevertheless, we consider conditioning on an upper bound appropriate because there are also forces that constrain the sizes of the (reconstructed) trees from above and we wish to apply the appropriate conditionals. These forces may include (*i*) noise introduced into the recipient lists and the resulting limitations of the reconstruction procedure and (*ii*) network-level filtering policies that limit the spread of chain letters and other massively replicated e-mail traffic.

The idea of selection sheds light on why we might expect chain letters, in general, to be just barely subcritical when viewed as a Galton–Watson process. A chain letter far below the critical threshold has a truly negligible probability of reaching more than a few people. On the other hand, a chain letter far above it threatens the stability of e-mail servers. If it is possible to write a sufficiently persuasive chain letter to surpass the critical threshold, the continued operation of the Internet suggests that there are effective mechanisms that detect and put an end to such traffic. Thus, whereas we would be surprised by an almost-critical chain letter in the absence of selection effects, these considerations suggest that, in fact, almost-critical chain letters are essentially the only ones that we should expect to see.

Discussion

Liben-Nowell and Kleinberg's analysis of real-life network diffusion at a very large scale yields some striking patterns that seem difficult to explain within a simple model. Nevertheless, the global patterns in the data can be matched with a basic Galton–Watson process by conditioning on the process reaching a large number of nodes. Our analysis works by estimating the local behavior of the process (number of children per recipient) from the data and conditioning on the number of nodes matching the reconstructed trees in the data. These two inputs combined with the dynamics of the branching process produce depths and widths matching those observed in the data, as well as trees that look very much like the real ones. It is worth noting that no aspect of our selection explicitly constrained depth or width, and so the fact that these come out at the right values in the simulations supports the reasonableness of modeling the observed diffusion as a Galton–Watson process with a size selection bias.

This approach is different from that of Liben-Nowell and Kleinberg in that it encodes all the details of the emergence and reconstruction of the observed chain letters into the key parameter of the Galton–Watson process, namely, the distribution of how many children each node has, rather than modeling signing behaviors explicitly. Those local details may be quite intricate, as suggested by Liben-Nowell and Kleinberg.

Our contribution is to point out that, whereas the local features of this process may be complicated *substantively*, its resulting global patterns can be explained quite simply in analyzing it *statistically*. This approach also focuses the explanatory burden of a more detailed analysis of the process on describing how the observed distribution of children comes to be that way, reducing a global question about a complex stochastic process to one that is essentially about local features. There are various possible micromodels that would give rise to the correct distribution of children per node, including that discussed by Liben-Nowell and Kleinberg; given one of them, a basic global model explains the data after selection is accounted for.

In the *SI Text*, we analyze a simple micromodel that generates an offspring distribution and global behavior consistent with the observations. In that model, if a node decides to forward the chain letter, it sends it to d people and each of them independently decides whether to sign and forward it with some activation probability q, whereupon the process continues. That dynamic generates what we call the "true" dissemination tree. In order to generate the "reconstructed" tree, we then sample nodes randomly from the true tree so that we get about as many as were collected in the empirical exercise of Liben-Nowell and Kleinberg. Then we reconstruct as much of the true tree as can be inferred from those samples. Finally, we condition on the reconstruction being of the appropriate size (between 2,442 and 3,250 nodes). The resulting reconstructed trees look, both locally and globally, like the ones that were reconstructed by Liben-Nowell and Kleinberg. In this exercise, we used $d = 30$ and chose the activation probability q to generate a good fit. It is worth noting that the true trees produced by this process look very different from the ones that were observed; in particular, there are some nodes with large numbers of children, in contrast to the reconstructed trees. The key aspect in obtaining reconstructed trees similar to the observed ones comes from the selection issue of observing only a subset of nodes, which, the simulations show, heavily biases the offspring distribution toward low numbers. Together with conditioning on size, this sampling process allows us to match the observations.

The features of the conditional realizations of the Galton–Watson process are perhaps unexpected. Indeed, the analysis points out how starkly selection at the level of an entire dataset can influence the observed structure of a process, especially when it is a complex, probabilistic, and dynamic one such as diffusion in a large population. Despite their power, little is known about these sorts of selection effects. To deal with such issues, a sophisticated theoretical apparatus is needed to analyze *conditional* distributions of classical processes, where the conditioning is upon the selection that determined how or why that dataset was observed.

ACKNOWLEDGMENTS. We thank Alex Frankel and Jon Kleinberg for very helpful comments on earlier drafts. We gratefully acknowledge financial support from the National Science Foundation under Grant SES-0647867 as well as the Jaedicke fellowship at the Stanford Graduate School of Business.

1. Granovetter M (2005) The impact of social structure on economic outcomes. *J Econ Perspect* 19:33–50.
2. Coleman J, Katz E, Menzel H (1957) The diffusion of an innovation among physicians. *Sociometry* 20:253–270.
3. Galaskiewicz J (1985) Professional networks and the institutionalization of a single mind set. *Am J Sociol* 50:639–658.
4. Friedkin NE, Cook KS (1990) Peer group influence. *Sociol Method Res* 19:122–143.
5. Freeman LC, Freeman SC, Michaelson A (1988) On human social intelligence. *J Soc Biol Struct* 11:415–425.
6. Liben-Nowell D, Kleinberg J (2008) Tracing information flow on a global scale using Internet chain-letter data. *Proc Natl Acad Sci USA* 105:4633–4638.
7. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. *Nature* 393:440–442.
8. Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45:167–256.
9. Watson HW, Galton F (1875) On the probability of the extinction of families. *J R Anthropol Inst GB* 4:138–144.
10. Durrett R (2005) *Probability: Theory and Examples* (Thomson, Belmont, CA), 3rd Ed.

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.

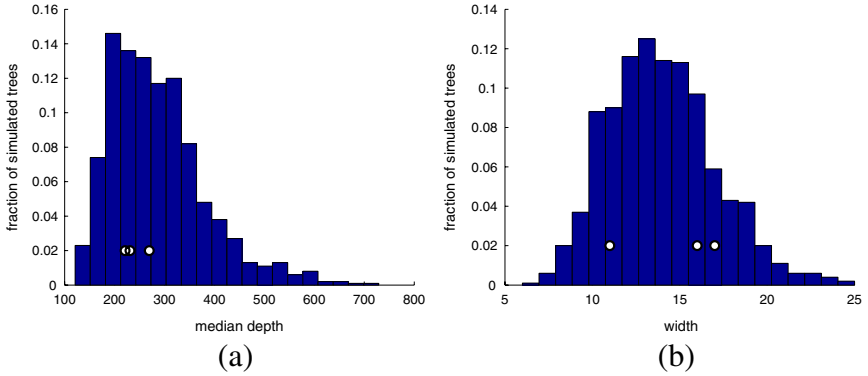

Fig. S1. Histograms of statistics from the simulated trees. The white circles correspond to the three observed components of the NPR chain letter data. (*a*) depicts median depth and (*b*) depicts width.

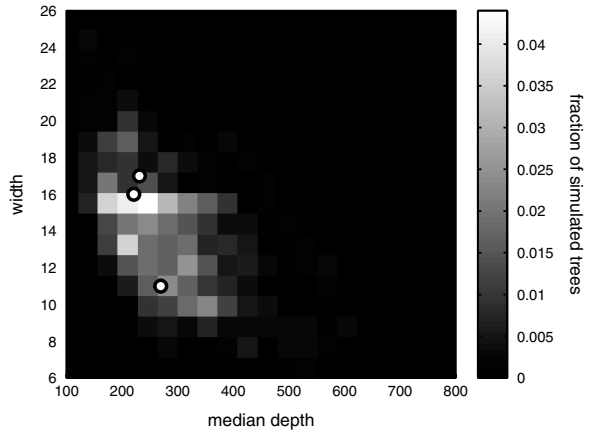

Fig. S2. The joint distribution of median node depth and width from the simulated trees. The white circles correspond to the three observed components of the NPR chain letter data.

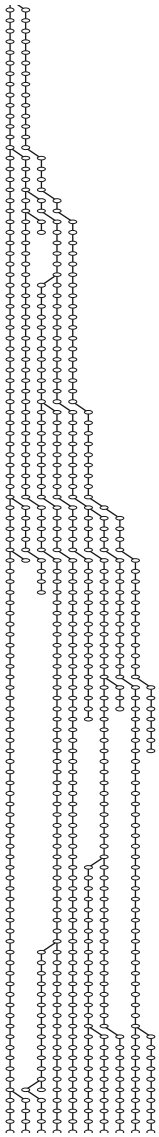

Fig. S3. Part of a typical reconstructed tree generated by the simulation procedure.

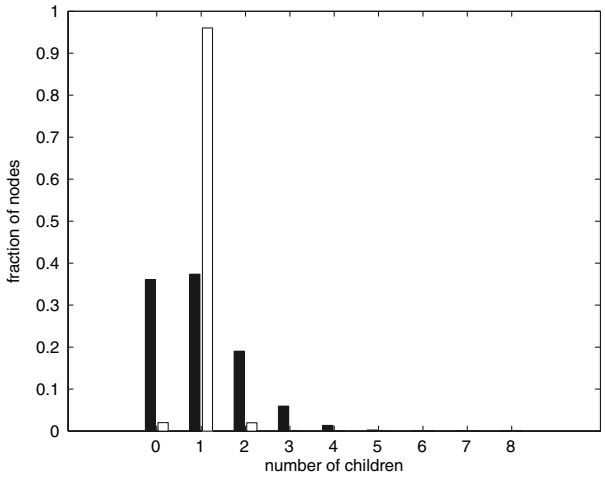

Fig. S4. A comparison of the offspring distribution of a typical true tree from one of the simulations (*Black Bars*) with that of the reconstructed tree that would be observed (*White Bars*).