Using selection bias to explain the observed structure of Internet diffusions

Benjamin Golub* and Matthew O. Jackson†

*Graduate School of Business, Stanford University, Stanford, CA 94305; †Department of Economics, Stanford University, Stanford, CA 94305; and ‡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.

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.

* 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.

www.pnas.org/cgi/doi/10.1073/pnas.1000814107

PNAS | June 15, 2010 | vol. 107 | no. 24 | 10833–10836
the regimes of the Galton–Watson process seems to fit when we
look at the unconditional distribution—is overcome by condition-
ing 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 distri-
bution 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 ob-
serve 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 distribu-
ted 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

\[
\ell(p;x) = f(0;x) \log \left( 1 - \sum_{k=0}^\infty p(k) \right) + \sum_{k=1}^\infty f(k;x) \log p(k).
\]

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 straight-
forward to verify that \( \hat{p}(k) \) yields a global maximum of the like-
lihood function. The same procedure can be carried out on many
trees at once.

It is worth noting that we did not explicitly model the observa-
tion 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 defin-
ing 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 off-
spring random variable, the mechanisms of that can be complicated.
What kinds of network processes underlie \( p \) is an interesting ques-
tion 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 compo-
ents and used the function \( f \) for the component with 2,442
observed nodes. Of course, the real NPR dissemination tree pre-
sumably had only one component and the pieces in the recon-
struction arose from an inability to reconstruct the tree all the
way to its origin. We simply take some of the subtrees as a single
instance of the diffusion process.

The distribution \( p \) that was estimated is reported in Table 1. Its
expectation is 2.441/2.442 = 0.9996. Indeed, it is immediate to
verify on the basis of the formulas above that the estimated dis-
tribution \( \hat{p} \) will have expectation \( n - 1 \)/\( n \) in any tree of \( n \) nodes.
The implications of this rather simple fact deserve some com-
ment. 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-homoge-
neous 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>
<thead>
<tr>
<th>( k )</th>
<th>( \hat{p}(k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0246</td>
</tr>
<tr>
<td>1</td>
<td>0.9525</td>
</tr>
<tr>
<td>2</td>
<td>0.0217</td>
</tr>
<tr>
<td>3</td>
<td>0.0012</td>
</tr>
<tr>
<td>≥4</td>
<td>0</td>
</tr>
</tbody>
</table>
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.

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 \(q\), wherein 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, 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.