Seeing only the successes: The power of selection bias in explaining the structure of observed Internet diffusions

Benjamin Golub Matthew O. Jackson*

January 18, 2010

Abstract

Recently, large data sets 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 median 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 the data. This demonstrates that accounting for selection biases of which data we observe can radically change the estimation of classical diffusion processes.

1 Introduction

As social network data have become increasingly available in electronic form, researchers are developing more detailed and accurate pictures of the patterns of social interactions. This is 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 either 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 data sets become available because of the electronic storage of various forms of interaction via the Internet; and these then become useful testbeds for theories of social networks. In this article, we focus on the explanatory power of one inherent selection bias that comes along with many such new data sets. Specifically, we examine a selection bias that arises from looking at unusually *large* data sets formed by diffusion processes — with a particular application to Internet chain letters.

In a recent article [6], Liben-Nowell and Kleinberg provide an important and interesting examination of two chain letters that had wide circulation on the Internet: one as a petition in support of public radio and television that began circulating in 1995 and another as a petition against the eventual war in Iraq that circulated in 2002 and 2003. By obtaining many copies of the emails 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. This is a branching process where each node has a random number of children, drawn independently according to the same distribution. It can also serve as a model of an epidemic, where the number of children is the

[*] Golub: Graduate School of Business, Stanford University, Stanford, CA, 94305-5015, bgolub@stanford.edu. Jackson: Department of Economics, Stanford University, Stanford, CA, 94305-6072, and the Santa Fe Institute, Santa Fe, NM 87501. We thank Jon Kleinberg for 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. Correspondence should be addressed to Matthew O. Jackson (jacksonm@stanford.edu).

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 a petition directly below that sender's signature. It is well-known [10] that the key quantity for 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 subcriticial 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 the population will survive with positive probability, in which event it will almost surely grow to arbitrarily large sizes; we ignore the 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 data sets that Liben-Nowell and Kleinberg observe have more than ten thousand nodes each, whereas it is quite rare for a subcritical branching process with reasonable parameters to have more than a dozen nodes. As a result of this, 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, as reported in [6], then the trees that emerge have huge breadth, branch very frequently, and do not have the long chains that are observed in the data. As Liben-Nowell and Kleinberg put it: "The central problem is that this style of random epidemic process seems unable to produce trees whose observable portions are very large, yet with a number of children per node so highly concentrated around 1." To address this, Liben-Nowell and Kleinberg develop a richer underlying model of chain letter distribution with two additional 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 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. The only thing that we add to the basic model is a selection effect — namely, that most diffusions go extinct without ever growing large enough to be observed, and so, like Liben-Nowell and Kleinberg, we compare the data only with those realizations of the proposed process that grow to be as large as the observed ones.

In our analysis, we first fit the relevant parameters of a Galton-Watson process from the data using maximum likelihood estimation. 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. Simulated outcomes from this conditional distribution match the real observations closely. The fit suggests that the process underlying the observed data is a barely subcritical one. It also highlights the fact that while the trees have a simple Galton-Watson global structure after accounting for selection bias, their local behavior presents some interesting questions, which we mention in the Discussion.

While the specifics of the model and analysis which 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 which generated them, but instead exceptional realizations. Although implications of selection bias on data are well-understood in some settings, they have not traditionally played a significant role at the data-set level in social network analysis, and this is a particularly stark example of how this perspective can 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.

2 Results

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 root r where the probability of a node having k children is $p(k)$ for $k = 0, 1, 2, \ldots$, and the distribution is identical across nodes. This distribution is the fundamental parameter in the model. It is a simple matter to fit it 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 which 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

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

Table 1: The distribution of descendants per node estimated from the data.

k	$p(k)$
0	0.0250
1	0.9525
2	0.0213
3	0.0012
≥ 4	0

Maximizing the log-likelihood with respect to p implies that if $f(k, x) = 0$ then $p(k) = 0$, and otherwise setting the derivative with respect to $p(k)$ to be equal to 0 implies that for every k,

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

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

$$p(k) = f(k, x) / \sum_k f(k, x).$$

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 these values of $p(k)$ yield a global maximum of the likelihood function.

It is worth noting that we did not explicitly model the observation process — 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.

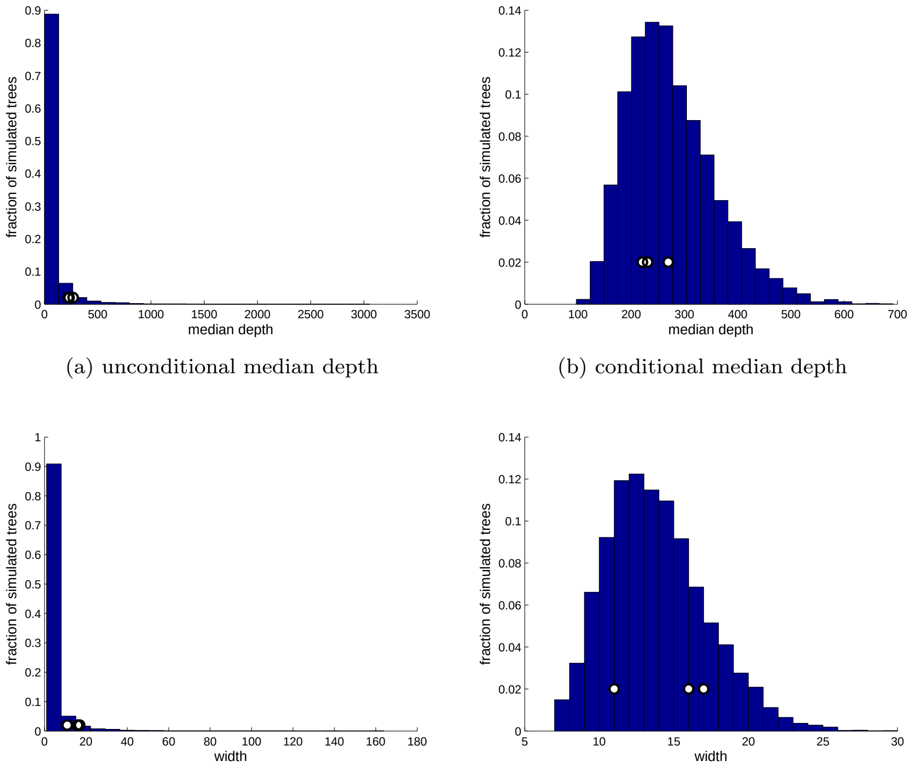
We also did not explicitly include a model of the network over which the diffusion is happening. The reason is that the only thing that matters for the Galton-Watson process is the number of observable children that each node has. While the network structure will certainly influence this random variable, the mechanisms of that can be complicated and we do not need to analyze them for our purposes.

We applied this fitting procedure to an example from the Supporting Information Appendix of [6]. Specifically, we used the NPR petition, whose observed portion had three components and used the function f for the component with 2442 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 p that was estimated is reported in Table 1. Its expectation is 0.9988, corresponding to a process that is just slightly subcritical. After estimating the distribution, we simulated the branching process with this distribution and analyzed only realizations whose number of nodes were between those of the largest and smallest observed components in the NPR data — between 2,442 and 3,250 nodes. The most relevant histograms from the analysis are shown in Fig. 1. The statistics we focus on are the median depth (distance of nodes from the origin) as well as the tree width (maximum number of nodes at the same depth).

The key fact in these histograms is that while the median node 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 .95



(a) unconditional median depth

(b) conditional median depth

(c) unconditional width

(d) conditional width

Figure 1: Histograms of unconditional and conditional statistics from the simulated trees. Figures (a) and (b) show median node depths and figures (c) and (d) show tree widths. The figures on the left, (a) and (c), correspond to the unconditional distribution, while 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.

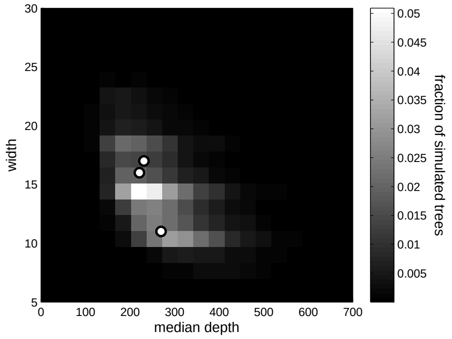
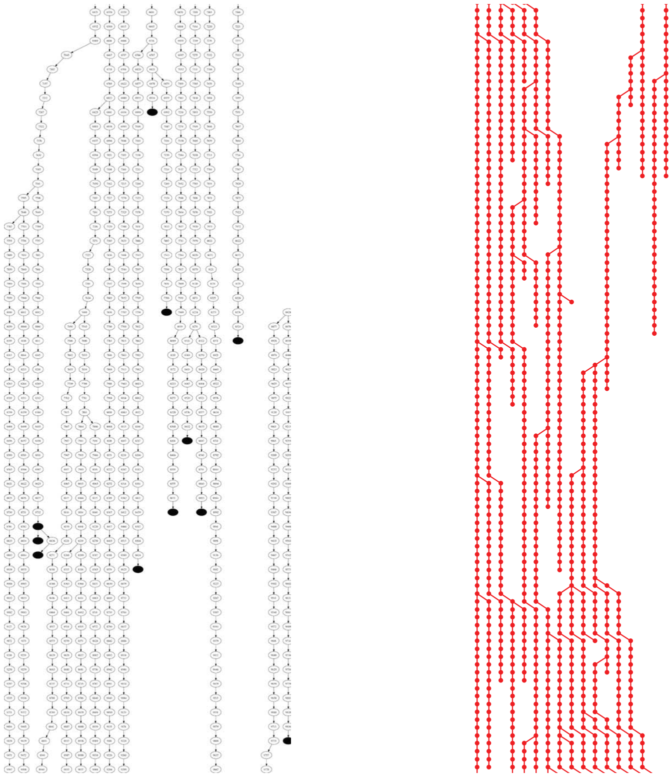


Figure 2: The joint distribution of median node depth and width conditional on the size bounds. The white circles correspond to the three observed components of the NPR chain letter data.



(a) actual

(b) simulated

Figure 3: Panel (a) shows a piece of the real NPR petition propagation tree as reconstructed by Liben-Nowell and Kleinberg [6], and panel (b) shows a piece of typical tree from the conditional distribution in our simulations.

(mean .9527 and standard deviation 0.0040), exactly matching the data. This is unsurprising, since $p(1)$, the probability of having one child, was a parameter of the model that was estimated from the observations.

The conditional distribution is a small slice of the whole distribution, but not too small. Out of 10^6 simulated trees, $51,047$ (5.10 percent) exceeded the lower size bound of 2,442 and $9,483$ (0.948 percent) fell within both bounds. This corresponds to 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 [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, since there are also forces that constrain the sizes of the (reconstructed) trees from above and we wish to apply the appropriate conditionals. These 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 email traffic.

The idea of selection sheds light on why this observed chain letter is so curiously close to the critical threshold between the regimes of extinction and infinite replication. A chain letter far below this 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 email 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 quickly detect and put an end to such traffic. Thus, while 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.

3 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 that of 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. Indeed, the distribution of children per node that fits best has interesting features; it is not readily explained by underlying behavior in which every node broadcasts the message to all its contacts, who then do the same with some probability, independently of each other. That process would result in more branching than in the trees that are observed. Thus, more complex behaviors are required to generate the Galton-Watson process that we have estimated.

Our contribution is to point out that while the local features of this process may be complicated *substantively*, its resulting global patterns can be explained quite simply in analyzing it *statistically.* This also focuses the explanatory burden of a more detailed analysis of the process on describing how the 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 local phenomena 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.

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 data set can influence the observed structure of a process, especially when it is a complex, probabilistic, and dynamic one such as diffusion in a large network. 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 data set was observed.

References

[1] Granovetter, M. (2005) *The Journal of Economic Perspectives* **19**, 33–50.

[2] Coleman, J, Katz, E, & Menzel, H. (1957) *Sociometry* **20**, 253–270.

[3] Galaskiewicz, J. (1985) *American Sociological Review* **50**, 639–658.

[4] Friedkin, N. E & Cook, K. S. (1990) *Sociological Methods & Research* **19**, 122–143.

[5] Freeman, L. C, Freeman, S. C, & Michaelson, A. (1988) *Journal of Social and Biological Structures* **11**, 415–425.

[6] Liben-Nowell, D & Kleinberg, J. (2008) *Proceedings of the National Academy of Sciences* **105**, 4633–4638.

[7] Watts, D. J & Strogatz, S. H. (1998) *Nature* **393**, 440–442.

[8] Newman, M. E. J. (2003) *SIAM Review* **45**, 167–256.

[9] Watson, H. W & Galton, F. (1875) *Journal of the Anthropological Institute of Great Britain* **4**, 138–144.

[10] Durrett, R. (2005) *Probability: Theory and Examples, Third Edition.* (Thomson, Belmont, CA).