Tracing information flow on a global scale using Internet chain-letter data

David Liben-Nowell*† and Jon Kleinberg†‡

*Department of Computer Science, Carleton College, Northfield, MN 55057; and ‡Department of Computer Science, Cornell University, Ithaca, NY 14853

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved January 25, 2008 (received for review September 6, 2007)

Although information, news, and opinions continuously circulate in the worldwide social network, the actual mechanics of how any single piece of information spreads on a global scale have been a mystery. Here, we trace such information-spreading processes at a person-by-person level using methods to reconstruct the propagation of massively circulated Internet chain letters. We find that rather than fanning out widely, reaching many people in very few steps according to "small-world" principles, the progress of these chain letters proceeds in a narrow but very deep tree-like pattern, continuing for several hundred steps. This suggests a new and more complex picture for the spread of information through a social network. We describe a probabilistic model based on network clustering and asynchronous response times that produces trees with this characteristic structure on social-network data.

social networks | algorithms | epidemics | diffusion in networks

The dissemination of information is a ubiquitous process in human social networks. It plays a fundamental role in settings that include the spread of technological innovations (1, 2), word-of-mouth effects in marketing (3–5), the spread of news and opinion (6–8), collective problem-solving (9, 10), and sampling methods for hidden populations (11, 12). The basic models for studying such phenomena posit that information will diffuse from person to person in the style of an epidemic (13–16), expanding widely in a short number of steps according to "small-world" principles (17, 18). However, despite recent studies in online domains (5–8), it has been difficult to obtain detailed traces of the dissemination of a single piece of news or information on a global scale to assess the predictions of these models. As such, it has remained an open question whether the spreading of information truly proceeds with a rapid, epidemic-style fan-out or whether it follows a potentially more complex structure. The difference between these possibilities has consequences not only for the models that are used to capture their essential properties but also potentially for the "life cycle" of a piece of information as it spreads through the global social network.

Here, we trace these types of large-scale information-spreading processes at a person-by-person level using methods to reconstruct the propagation of massively circulated Internet chain letters, and from these observations we propose a new set of principles for how such processes work. We focus in particular on two such chain letters, which exhibit tree-like patterns of dissemination that are quite similar to each other but are initially in conflict with the intuitive picture of how information spreads in these settings. Rather than expanding to many individuals in a few steps, the trees are very narrow and continue reaching people several hundred levels deep. We describe a mathematical model that produces trees with this characteristic structure, grounded fundamentally in the observations that social networks are highly clustered and that information can take widely varying amounts of time to traverse different edges in the network. The simple structure of the model, and the fact that it is based on earlier empirical studies of human response times (19–21), thus suggests a possible basis for this narrow and deeply reaching style of information transmission in the local dynamics of communication within highly clustered social networks.

Reconstructing the Spread of Internet Chain Letters

To reconstruct instances in which specific pieces of information spread through large, globally distributed populations, we analyzed the dissemination of petitions that circulated widely in chain-letter form on the Internet over the past several years. The petitions instruct each recipient to append his or her name to a copy of the letter and then forward it to friends. Each copy will thus contain a list of people, representing a particular sequence of forwardings of the message; and hence different copies will contain different but overlapping lists of people, reflecting the paths they followed to their respective current recipients. This forwarding process is a readily recognizable mechanism by which jokes and news clippings can also achieve wide circulation through the global e-mail network; the explicit lists of names in the petition format, however, make it much easier to trace the propagation of the messages. The main chain letter that we analyze is based on a widely circulated petition from 2002–2003 claiming to organize opposition to the impending war in Iraq. We obtained copies via Internet searches of mailing-list archives in which they were publicly posted; these searches resulted in 637 copies with distinct chains of recipients, representing nearly 20,000 distinct signatories in aggregate. [See supporting information (SI) *Appendix* for the specifics of the data-collection process.]

We performed a similar analysis for a second chain letter, a petition that began circulating in 1995, purporting to organize political support for continued United States governmental funding of National Public Radio (NPR) and the Public Broadcasting System (PBS). Through similar means to those used for the Iraq petition, we acquired 316 distinct copies of the NPR petition, comprising a total of 13,052 people. The dissemination of the two chain letters exhibited qualitatively very similar structures, and for purposes of the discussion here, we focus on the analysis of the chain letter associated with the Iraq petition. Although both petitions in fact had their origins in hoaxes and naive misunderstandings, as a large fraction of the most widespread Internet chain letters do (22, 23), this fact is immaterial to our purposes, especially because almost all signatories to each appeared to believe them to be authentic; hence, we are studying genuine instances of the dissemination of individual pieces of information along links in the global social network.

People may in general receive a copy of the chain letter multiple times, but if each appends his or her name to just one copy, then the full propagation of the letter can be represented as a tree

Author contributions: D.L.-N. and J.K. 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.

Freely available online through the PNAS open access option.

†To whom correspondence may be addressed. E-mail: dlibenno@carleton.edu or kleinber@cs.cornell.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708471105/DC1.

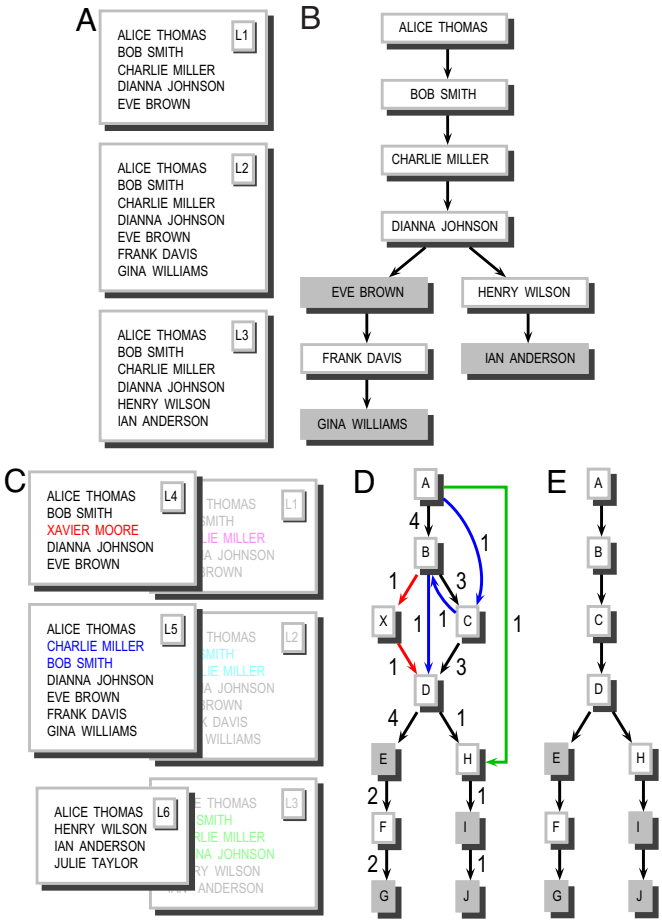

Fig. 1. Schematic view of the data-processing method. (*A*) Copies of a petition were acquired from the Web, consisting of distinct lists of signatories. (*B*) A network is formed by connecting signatory *x* to signatory *y* if in at least one petition copy *x* immediately precedes *y*. The node for the final signatory on each list is shaded gray to indicate that he or she publicly posted a copy. (*C*) The full set of petition copies may not result in a tree because of sequence rearrangements including point mutation, transposition, and block insertion/deletion in some copies of the lists. To handle minor variations in signatories' names in different petition copies, the names of the signatories were replaced by unique identifiers; we deem two nonidentical signatories' names equivalent if they are preceded by equivalent names and their names are within a small edit-distance threshold. (*D*) The network that results can deviate from a tree structure. The weight next to each edge indicates the number of petitions that exhibit that edge). (*E*) A tree is formed from this network by (*i*) running a maximum-weight spanning arborescence algorithm, which excises connections inconsistent with a tree in the least consequential way possible, using the above weights; and (*ii*) pruning any nodes that are not on a path from the root to a shaded gray node.

structure: recipients are nodes, the originator is the root, and node *w* is a child of node *v* if *w* appends its name directly below *v*'s. Moreover, if this is the case, then each copy of the letter represents a path through the propagation tree, and the observable portion of the tree can be reconstructed simply by superimposing these paths (Fig. 1 *A* and *B*). Inspection of the chain letters indicates that recipients in the observable portion did appear to almost uniformly forward the letter just once, and hence reconstruction of a tree provides a reasonable approximation to the actual propagation process. However, the superposition of the lists on the 637 letters deviates from a tree structure because of extensive noise in the data: Some recipients reordered the list of names on their copy of the letter in ways closely analogous to the kinds of chromosomal rearrangements one finds due to sequence mutation events in biological settings (24, 25) (Fig. 1*C*). We observed examples of point mutations (in some petition copies, names were replaced by the names of political figures), insertion/deletion events (there were a number of small blocks of 1–5 names that were present in the middle of the list in some petition copies and absent in other copies), duplication events (blocks of 2–20 names that were duplicated in some petition copies, sometimes immediately adjacent within the list and sometimes hundreds of names later), block rearrangements (in one petition, two pairs of blocks of 2–3 names were swapped relative to their position in all other copies that contained the same names), and one hybridization event (the names at the ends of two copies of the petition were intermingled after their common prefix in a third copy).

To reconstruct an approximation to the real propagation process from the data, we thus need to infer a tree in the presence of these sources of noise, and we perform this inference as follows (Fig. 1 *D* and *E*). We begin by representing the observed dissemination of the letter using a structure more complex than a tree, namely a directed graph G on the set of recipients in which there is an edge $e = (v, w)$ whenever w appears directly after v on at least one list. In the case of the Iraq chain letter, this graph G has 19,302 distinct names and 19,784 edges, where we applied a heuristic based on sequence alignment (24, 25) to declare two names with a common list predecessor and very small typographical variations to be equivalent. (One pseudonymized example from the data is the appearance in various copies of the signatories "John Smith Santa Monica Calif," "John Smith Santa Monica USA," and "John Smith Santa Monica Calif USA" with identical predecessors and successors.) Manual inspection verified that this heuristic did not in general conflate two distinct people who had the same predecessor.

Because any tree on this node set would have 19,301 edges (one fewer than the number of nodes), we need to delete a proportionally small number of edges (483) from the graph G to produce a tree. We do this deletion in a way that removes links inconsistent with a tree in the least consequential way possible. Specifically, for each edge $e = (v, w)$, we define the evidence for e to be the number of distinct copies of the petition that exhibit edge e. Using the evidence for each edge as its weight, we compute a directed spanning tree of G (also known as a branching or arborescence) of the maximum possible weight; this computation can be done efficiently at the scale of our data using an algorithm due to Edmonds (26). (We use an implementation from the LEMON project, http://lemon.cs.elte.hu.) Thus, we produce a spanning tree in which the total evidence for all edges, under our definition, is as large as possible. Finally, after the construction of the spanning tree, certain nodes no longer lie on a path from the root node to an individual who posted their copy of the letter. We delete such nodes, producing the final tree we use for our analysis; this tree contains 18,119 total nodes with 557 leaves, all of whom posted their copy of the letter, and 63 internal nodes that also posted.

The Structure of the Dissemination Tree

Inspection of the few messages that contained intact addressee lists indicates that recipients generally forwarded copies of the letter to a large number of other individuals. This observation is consistent with a form of information-spreading in which each person, upon receiving the information, proceeds to inform a large number of his or her neighbors in the social network. Epidemic-style models based directly on this observation suggest that the propagation tree, if it does not die out quickly, should have nodes with many children and very short paths from the root.

The tree reconstructed from the data, however, reveals a structure that is very different from the picture suggested by simple epidemic models: the median distance to the root over all nodes is nearly 300, and >90% of the nodes have exactly one child. Fig. 2 depicts the full tree, with a zoomed-in view of the tree in Fig. 3 to illustrate the characteristic structure. *SI Appendix* contains a high-resolution image of the full tree. The full superposition of the lists

Fig. 2. Tree derived from a large-scale chain-letter petition protesting the start of the war in Iraq, produced as described in Fig. 1. This tree has 18,119 nodes, of which 17,079 (94.26%) have exactly 1 child. The median node depth is 288 and the width of the tree is 82.

without correction for noise, although it deviates from a precise tree topology, exhibits a qualitatively very similar structure, indicating that these properties are intrinsic to the spreading of the chain letter, and not an artifact of the reconstruction process. Moreover, qualitatively similar structures are exhibited in the propagation of the other large-scale chain letter for which we have data. (See *SI Appendix* for more detail and an image of the tree associated with this other chain letter.)

Understanding why the reconstructed trees have this unusual structure thus poses a challenge; in contrast, for example, to the results of both older and more recent large-scale small-world experiments (17, 27), in which no chains ran for more than a few steps, we have a case in which most chains are hundreds of steps long and most recipients produce exactly one child in the observable tree. How could such a structure come about? With more detailed information about the e-mail messages themselves—for example, with complete message headers showing addressee lists and time-stamps—we could begin inferring not just the tree structure but also the sequences of actions taken by individuals to forward the letter. However, we have very few messages with such headers and almost no pairs that are close together in the tree (as would be needed to start inferring a sequence of actions by directly communicating individuals); primarily, we have just the sequences of names from the different letters.

As a result, we frame the problem of modeling the tree as follows: Is there a class of simple and plausible generative processes that, when run on real social networks, produce synthetic trees of the characteristic structure we observe—deep, narrow, and with most nodes having one child? A negative answer would suggest that what we are seeing is the result of unobserved idiosyncrasies in the collective behavior that produced the lists. If the answer is positive, however, it argues that this type of structure is in fact achievable by natural mechanisms, suggesting that deep patterns of transmission are in fact a robust form of information-spreading and potentially focusing the search for more detailed theories about why we observe it in real life.

Modeling the Structure of the Dissemination Tree

To evaluate message-forwarding models that may capture the structural properties of the observed tree, we simulate a sequence of probabilistic models on a social network with 4.4 million individuals gathered from the online community LiveJournal (LJ). Previous research has shown this network to have characteristics consistent with other large-scale social networks, and the organization of the online links is analogous to the e-mail contact lists that were used to propagate the chain letter (28, 29). To avoid reliance on a single model network, however, we also perform the simulations on two other sources of real social-interaction data, from different domains: the communication network of Wikipedia editors, and the co-authorship network built from a large bibliographic database (DBLP) of computer science authors. The results on these networks are qualitatively closely consistent with the results we report below for LJ, though scaled down because the Wikipedia and DBLP networks are each only approximately one-tenth of the size of LJ. The fact that we obtain broadly similar results from simulations on diverse datasets suggests that we are observing properties of the probabilistic process itself and not of the specific networks on which it is running.

Our models will start at a randomly chosen initiating node and construct trees spreading outward from this node, with portions of the tree made visible by some nodes posting their copy of the message. We will then assess how closely the structure of the observable portion of the constructed trees resembles the propagation tree of the real chain letter, using three metrics: the median node depth, the width, and the fraction of nodes with exactly one child. Here, the depth of a node is defined as its distance from the root, and the width of a tree is defined as the maximum size of a set of nodes that all possess the same depth. In all cases, the metrics will be averaged over multiple independent simulation runs on the LJ network: Each simulation run is continued until the observable tree first reaches the size of the real chain-letter propagation tree,

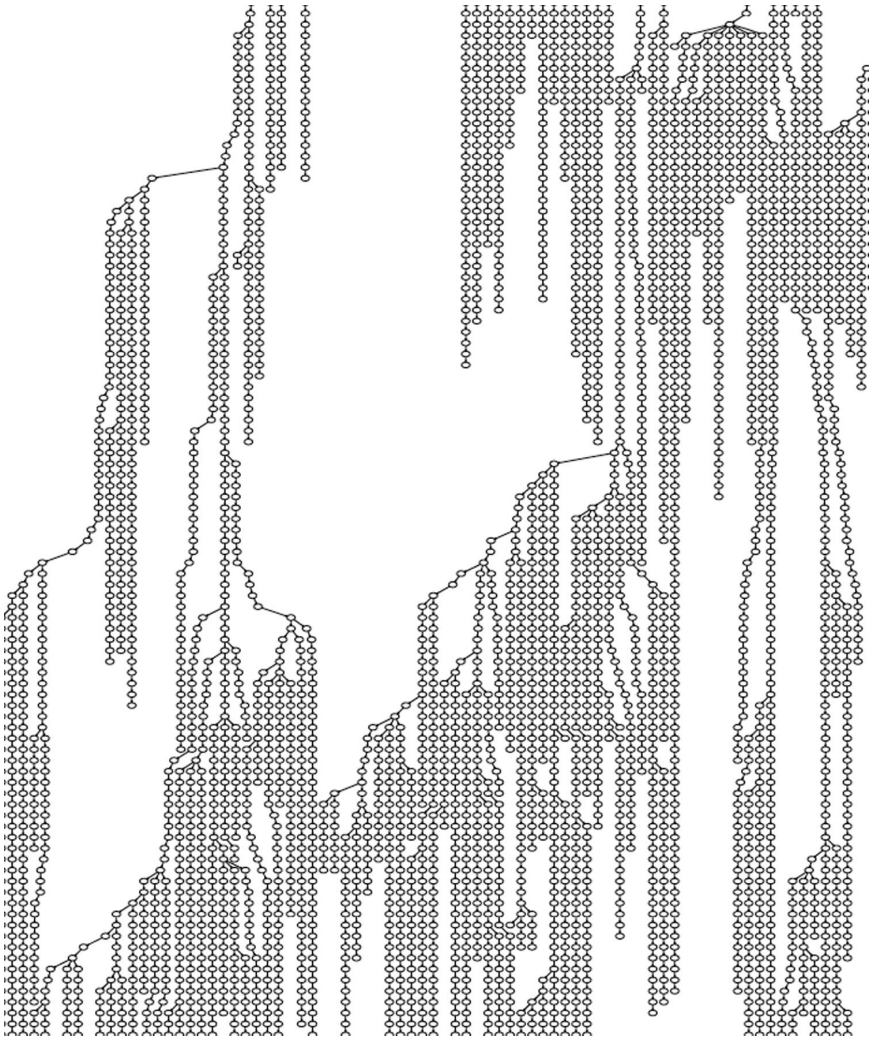

Fig. 3. Close-up of a portion of the tree in Fig. 2.

and those runs in which the tree fails to reach this size are omitted. Omitting simulated trees that fail to grow large enough is consistent with our goal of studying properties of information diffusion conditional on reaching a large population; in real life, most circulated e-mail messages never spread widely, but we are interested in the structure of those that do.

Our models all incorporate the following two principles: Many recipients may choose not to forward the letter at all, and only a few recipients will choose to post the letter publicly. Thus, we introduce a discard-rate parameter δ, specifying the probability that a given recipient discards the message and takes no further action on it, and a post-rate parameter π, specifying the probability that each recipient publicly posts his or her copy of the letter. In keeping with findings from earlier experiments based on e-mail forwarding (27), we set the discard-rate to the default value 0.65, although we find that reasonable variations do not qualitatively change the findings. The post-rate is a parameter that we will more explicitly vary. Public posting is the only means by which portions of the tree become observable: When a recipient posts the letter, his or her full path from the root becomes visible, and hence in general a node on the tree is observable at the end of the process if and only if one of its descendants posted a copy of the letter. We will be studying the structure of the observable portions of the trees produced by our models, as we do with real chain letters.

We first consider a model based on a direct application of these probabilistic ingredients. We choose a random root node and construct a tree in unit time steps. In each step, each new recipient of the letter discards it independently with probability δ and otherwise forwards it to all neighbors (posting with probability π). Any neighbor w that has not received the letter previously becomes a new recipient in the next time step; if w receives the letter from multiple senders, it chooses one of these senders arbitrarily as its parent in the tree. Finally, once the process terminates, we look at the observable portion of the tree, consisting of the union of all paths from the root to the nodes that posted their copy of the letter.

Although such a model is very natural, it produces trees that compare poorly to the real chain-letter data. Simulating this model on the LJ network, the observable portion of the tree has a median depth 5.0, width 9,625, and single-child fraction 19.04% (averaged over 10 independent runs) with $\pi = 0.10$, and very similar properties for other small values of π. This wide divergence from the real data cannot be remedied simply by having recipients send to a smaller set of neighbors; if each recipient who forwards does so to a random subset of 4 or 5 of his or her neighbors, then the width remains in the thousands, the median depth remains <50, and the single-child fraction remains $<70\%$. 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.

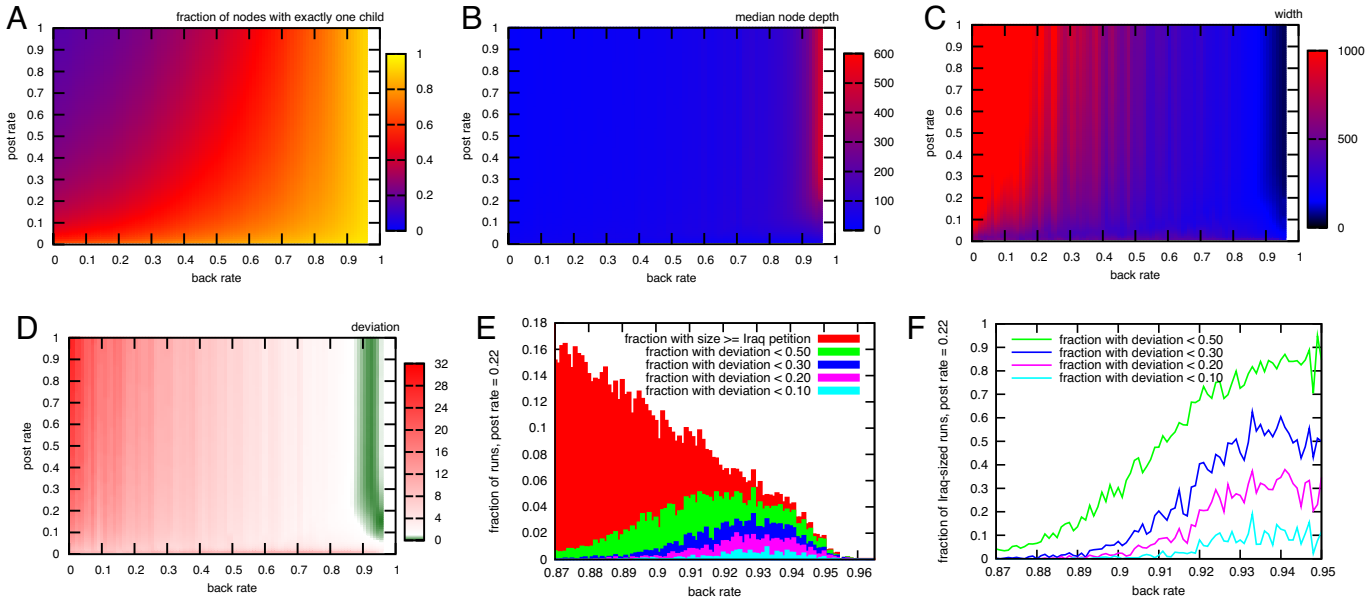

Fig. 4. Measurements of the quality of simulated trees. For each back-rate $\beta = 0.00, 0.01, \ldots, 1.00$ and post-rate $\pi = 0.00, 0.01, \ldots, 1.00$, a set $S_{\beta,\pi}$ of trees was generated using our model. Any generated tree that failed to reach 18,119 recipients—the number of observable recipients in the real chain letter—was discarded; the remaining trees were trimmed to include only the first 18,119 people whose names appeared in posted copies of the petition. (*A–C*) Median value on $S_{\beta,\pi}$ measured using each of our three metrics: single-child fraction, median depth, and width. (*D*) Median deviation—the maximum over the three metrics of the ratio $|x - y|/\min(x, y)$, where x is the value of the metric on the simulated tree and y is the value of the metric on the real chain letter—of trees in $S_{\beta,\pi}$. (Deviations between 0 and 1 are shown in green.) Once the back-rate reaches $\approx$0.9, we obtain trees that approximately match the real data in all three metrics, exhibiting high depth, low width, and a high fraction of nodes with exactly one child. At least a dozen trees were generated for each β and π, and at least 2,000 trees were generated in the region of parameter space where the match is closest, with $\beta = 0.870, 0.871, \ldots, 0.959$. (*E*) Fraction of simulated trees in this region that reach the size of the real chain letter and the fraction of simulated trees that achieved a deviation of less than 0.10, 0.20, 0.30, and 0.50 for post-rate $\pi = 0.22$. Of runs that reach the size of the real tree, *F* shows the fraction of simulated trees that achieve these small deviations.

Models Based on Asynchronous Response Times

To produce trees that approximately match the chain-letter data, we introduce two further extensions to the mechanism. The first of these extensions is based on asynchronous response times. Rather than assuming that the letter spreads in fixed unit time steps, we model each recipient as waiting a length of time τ before acting on the message, where τ is distributed according to the density function $f(x) = x^{-\alpha}$ for an exponent α. This accords with the findings of recent studies of human response times to a spectrum of communication types including e-mail (19–21), which find such distributions with exponents α ranging between 1 (with cut-off) and 3/2. We find that our results remain qualitatively consistent across this range; for the results described here, we use $\alpha = 3/2$ as a default.

The specifics of the model with asynchronous response times are as follows. Time proceeds continuously, rather than in discrete steps, and when a given node w in the network first receives a copy of the letter, at time t, it first decides whether to participate in the process at all, choosing to do so with probability $1 - \delta$. If w chooses to participate, it then chooses a random waiting time τ distributed as above. Between times t and $t + \tau$, node w may receive multiple copies of the letter (including the initial one it received at time t). At time $t + \tau$, node w selects the copy of the letter it has received with the longest list of names (breaking ties arbitrarily), forwards it to all its neighbors, and publicly posts this copy with probability π. As before, when the process terminates, we consider the observable portion of the tree.

This asynchronous pattern of response has a "serializing" effect in networks with large clustering coefficient (18), as the LJ network has: If the neighbors of a forwarding node are mutually connected, then they will forward the letter to each other as they act on it in order, producing a single long list with all of their names rather than many distinct shorter lists, each containing one of their names. In the observable tree, this change will tend to produce deeper "runs" of nodes in which each node has exactly one child, precisely the structure that we observe. This way in which real-valued response times produce paths with a greater number of hops is analogous to phenomena in the analysis of shortest paths in graphs with random edge lengths (30), although the two types of models have different structures, arising from different generative mechanisms.

Asynchronous response is a step toward trees with the correct structure, but it is not enough by itself; consequently, we introduce a second extension to the model as well. This second extension is based on the fact that recipients actually have two natural ways of reacting to the message other than discarding it: they can forward it to their neighbors in the network, as before, or they can group-reply to the set of corecipients on the e-mail message they receive; in the latter case, these corecipients each receive a copy of the letter with the recipient's name appended. Thus, we keep the details of the previous model the same as before, with one addition: for a back-rate parameter β, a nondiscarding recipient node w at time $t + \tau$ forwards the letter to its own neighbors as before with probability $1 - \beta$, and otherwise it group-replies with probability β.

Combined with asynchronous response times, group-replying further amplifies the serializing effect of having copies of the letter handled in sequence by the set of nondiscarding neighbors of a node, with each appending its name and thus producing a single long path in the tree. However, increasing the back-rate also reduces the progress of the letter to new nodes in the graph, because group-replying rather than forwarding to neighbors only provides copies of the letter to nodes that have already received it at least once. With a high back-rate, the letter is thus less likely to ever reach a large set of nodes. Thus, it becomes natural to study trade-offs in the tree structure as a function of β.

Fig. 4 *A–C* shows the median depth, width, and single-child fraction of trees produced as the back-rate β and post-rate π are each varied independently between 0 and 1 (with the discard-rate

δ fixed to 0.65, although analogous results hold for other discard-rates in the range between 0.5 and 0.75). For high back-rates around 0.95, combined with low post-rates around 0.22, we obtain trees that approximately match the propagation tree of the real chain letter in all three metrics (Fig. 4 *D–F*).

The model that produces trees approximately matching the observed diffusion patterns in our data thus involves two related ingredients: asynchronous response times and the ability of a message (via the back-rate parameter) to move "laterally" between multiple nodes receiving a message from the same source. Both of these ingredients have the effect of producing long, narrow chains of recipients, a striking and arguably unexpected property of the structure one observes in the real dissemination trees. Moreover, for the parameters at which the closest approximations to the real tree are obtained, an extremely small fraction of the simulation runs on the LJ network produce trees as large as the real chain-letter tree before dying out. In other words, the structure of the real tree corresponds to a portion of the parameter space in which large trees are rare events—as they are in real life as well.

Discussion

In essence, the progress of the Iraq-war and NPR chain letters had a type of stroboscopic effect, serving to briefly "light up" a structure—the global e-mail network—that has otherwise been largely invisible, and allowing us to observe a snapshot of this network's everyday use as a means of conveying information. The resulting analysis has exposed several themes. First, accurately reconstructing the paths followed by the information is a computational challenge in itself, given the extensive ways in which the data are mutated as they spread. Second, the spreading patterns of the real chain letters are strongly at odds with the predictions of simpler theoretical models, which posit processes that reach many more people in radically fewer steps. Finally, simple probabilistic models incorporating the speed with which individuals respond to information can produce synthetic spreading patterns that closely resemble the ones we observe in real life.

As noted earlier, the way in which the spreading pattern is made visible to us by the data—through lists of signatories—means that we lack detailed information about recipient lists and time-stamps on all but a handful of individual messages. As a result, our modeling efforts have, by necessity, focused on arguing that the unusual structures we observe are capable of arising from simple generative processes, thus suggesting that this style of information transmission can in fact have a natural basis, and attempting to expose some of its plausible qualitative ingredients. With more detailed information—for example, in an analysis that had access to many or most of the message headers—we could study the response times and overlaps in recipient lists among adjacent nodes in the tree and thus assess the alignment of these models to the detailed mechanics of message-sending, not just to global parameters (depth, width, single-child fraction) of the tree itself.

Overall, then, Internet-based snapshots of information diffusion can potentially provide us with insight into some of the global dynamics underlying social phenomena such as opinion formation and political mobilization. The fact that the observed diffusion occurs along trees that are so deep and narrow suggests that the paths traversed by information through social networks can be more complex than might have been supposed, with the large number of steps giving the diffusion a certain fragility and presenting greater opportunities for the information to be altered or lost as it spreads. The pattern of the diffusion also seems initially in conflict with the small-world nature of the social network in which it is embedded; but the models discussed here show that such patterns are capable of arising from natural processes operating in real social networks. In the end, the structure of a small world, in which most people are connected by short paths, need not be at odds with a world in which an antiwar appeal, embedded in an e-mail chain letter, can pass through several hundred intermediaries before arriving in one's inbox.

ACKNOWLEDGMENTS. We thank Moses Liskov for valuable discussions in the early stages of this work. This work was supported in part by a John D. and Catherine T. MacArthur Foundation Fellowship, a Google Research Grant, a Yahoo! Research Alliance Grant, and National Science Foundation Grants CCF-0325453, IIS-0329064, CNS-0403340, BCS-0537606, and CCF-0728779.

1. Rogers E (1995) *Diffusion of Innovations* (Free Press, New York), 4th Ed.
2. Strang D, Soule S (1998) Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annu Rev Sociol* 24:265–290.
3. Domingos P, Richardson M (2001) Mining the network value of customers. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Provost F, Srikant R, Schkolnick M, Lee D (Association for Computing Machinery, New York), pp 57–66.
4. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence in a social network. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Getoor L, Senator TE, Domingos P, Faloutsos C (Association for Computing Machinery, New York), pp 137–146.
5. Leskovec J, Adamic L, Huberman B (2006) The dynamics of viral marketing. *Proceedings of the 7th ACM Conference on Electronic Commerce*, eds Feigenbaum J, Chuang JC-I, Pennock DM (Association for Computing Machinery, New York), pp 228–237.
6. Adar E, Zhang L, Adamic LA, Lukose RM (2004) Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem*. Available at www.blogpulse.com/papers/www2004adar.pdf.
7. Gruhl D, Liben-Nowell D, Guha RV, Tomkins A (2004) Information diffusion through blogspace. *Proceedings of the 13th International World Wide Web Conference*, eds Feldman SI, Uretsky M, Najork M, Wills CE (Association for Computing Machinery, New York), pp 491–501.
8. Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M (2007) Patterns of cascading behavior in large blog graphs. *Proceedings of the SIAM International Conference on Data Mining*. Available at www.siam.org/proceedings/datamining/2007/dm07_060Leskovec.pdf.
9. Kearns M, Suri S, Monfort N (2006) An experimental study of the coloring problem on human subject networks. *Science* 313:824–827.
10. Monge P, Contractor N (2003) *Theories of Communication Networks* (Oxford Univ Press, Oxford).
11. Goodman LA (1961) Snowball sampling. *Ann Math Stat* 32:148–170.
12. Heckathorn D (1997) Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl* 44:174–199.
13. Dodds P, Watts D (2004) Universal behavior in a generalized model of contagion. *Phys Rev Lett* 92:218701.
14. Jackson M, Yariv L (2005) Diffusion on social networks. *Econ Publ* 16:69–82.
15. Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45:167–256.
16. Valente T (1995) *Network Models of the Diffusion of Innovations* (Hampton, Cresskill, NJ).
17. Travers J, Milgram S (1969) An experimental study of the small world problem. *Sociometry* 32:425–443.
18. Watts DJ, Strogatz SH (1998) Collective dynamics of ''small-world'' networks. *Nature* 393:440–442.
19. Vazquez A, *et al*. (2006) Modeling bursts and heavy tails in human dynamics. *Phys Rev E* 73:036127.
20. Oliveira JG, Barabasi AL (2005) Human dynamics: Darwin and Einstein correspondence patterns. *Nature* 437:1251.
21. Leskovec J, Horvitz E (2007) *Worldwide Buzz: Planetary-Scale Views on an Instant-Messaging Network* (Microsoft, Redmond, WA), Microsoft Res Tech Rep MSR-TR-2006-186.
22. Richman S (February 9, 2003) Not in my name: Why e-mail protesting is not all that it seems. *Independent on Sunday*. Available at http://find articles.com/p/articles/mi_gn4159/is_20030209/ai_n12734548.
23. Serazio M (February 18, 2003) When armchair activism backfires. *AlterNet*. Available at www.alternet.org/story/15212.
24. Waterman MS (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes* (CRC, Boca Raton, FL).
25. Gusfield D (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology* (Cambridge Univ Press, Cambridge, UK).
26. Edmonds J (1967) Optimum branchings. *J Res Natl Bur Stand* 71B:233–240.
27. Dodds P, Muhamad R, Watts D (2003) An experimental study of search in global social networks. *Science* 301:827–829.
28. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D (Association for Computing Machinery, New York), pp 44–54.
29. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. *Proc Natl Acad Sci USA* 102:11623–11628.
30. Braunstein LA, Buldyrev SV, Cohen R, Havlin S, Stanley HE (2003) Optimal paths in disordered complex networks. *Phys Rev Lett* 91:168701.