Information Cascades in Email Forwarding Networks

CS224W Project Report
Group #41
Connor Gilbert (connorg@cs.stanford.edu) and
Aliya Deri (aderi@cs.stanford.edu)

10 December 2012

1 Introduction

Recent information diffusion research has often focused on social networking sites like Facebook and Twitter, such as the analysis of reposting or hashtags as trends go "viral." Partially due to its more private nature, email has been less popular of late. However, email's near universal adoption as a means of online communication, its prevalence in commercial and academic environments, and the well-established use of forwarding to spread information make information diffusion in email networks — and specifically the mechanics of forwarding behavior — a relevant and important area to explore. This research could provide insight about, for instance, whether introducing an email to a specific part of the network would make it more likely to spread through the rest of the network, with implications for emergency response (spreading important information), political campaigning (spreading a candidate's message), and computer security (spreading a potentially malicious file or URL).

Previous work, described below, produces strikingly different social characteristics for email forwarding networks. "Forwarding trees" — essentially cascades that model the spread of a specific message through a network — have been found to have two mutually exclusive structures in different contexts: short and wide in business mail and small-world random graphs [1], but long and narrow in chain letters [2]. We posit that this difference is due to an interaction with the properties of the network; intuitively, one can imagine that someone mass-forwarding a chain letter to their entire address book will cross many community boundaries, while a work-related message will likely be relevant only to a smaller, more defined community, so such broad community-bridging forwarding is less likely.

We seek to explain this difference with a hypothesis called "Community Saturation." We define community saturation as the portion of nodes within a node's detected community that have already been exposed to a message. We hypothesize that as community saturation increases, a node is decreasingly likely to forward the message. This fits our intuition since it is unlikely that a chain letter will have reached most of a person's address book before they forward it; if it did, they will likely remove those people from the distribution. In the context of business networks, employees likely would check the previous distribution of a message before blindly forwarding it to their team or professional community. So, community saturation provides an intuitive yet concrete model for forwarding behavior in email networks.

In this paper, we describe our analysis of various datasets, while recognizing that detailed email traffic is difficult to obtain for research because of the general expectation of privacy inherent in email. We first cover related work. Then, we describe our experimental method. Next, we characterize the forwarding behavior of individuals in the datasets and compare them to our hypothesis. Finally, we find that the data support our hypothesis — in general, users are less likely to forward messages as community saturation increases — and we elaborate on the meaning of our findings and the limitations imposed by our experimental datasets.

2 Relevant Prior Work

Email forwarding is a distinct form of information sharing from others, including more public means like Twitter. Email also reflect social structures: as noted in Wang et al. [1], managers forward email more often; and forwarding can be an important way of maintaining personal and professional relationships [3].

2.1 Wang et al. 2011 [1]

Wang et al. studied a very large email network consisting of over 2 million emails (edges) and nearly 9,000 volunteer employees (nodes). The paper analyzes the microscopic forwarding behavior of users — will a single user decide to forward an email? — and the macroscopic information diffusion structure these actions form

when added together. The authors empirically find that most message forwarding trees are wide and shallow, with size growing linearly with the depth of the tree. The paper then presents a model which closely matches the observed data without relying on the underlying social network structure, a surprising result.

The paper's results may be difficult to generalize outside of a single-company corporate setting, where the social network is somewhat homogeneous given that all users are professionally connected in some way.

2.2 Liben-Nowell and Kleinberg 2008 [2]

Liben-Nowell and Kleinberg analyze the propagation of just two Internet chain letters to analyze information diffusion over email forwarding. They find, in contrast with Wang et al., a very deep and narrow graph, with median root-to-leaf distance of approximately 300 and over 90% of nodes having forwarded the email at least once. As in Wang et al., the authors attempt to present a mathematical model to explain the empirical data they observed; Liben-Nowell and Kleinberg's is based on imitating asynchronous response times and the ability of the message to move "laterally" across the social network.

We note that chain letters are specifically designed to be "spammed", which could cause some users to be more likely to spread the message as widely as possible, while a more serious or personal message would engender more restraint. The fact that the message is a chain letter and the data collection procedure — relying on copies of the email gathered from the Internet — limit the applicability of these conclusions to general email traffic.

2.3 Other work

The social dynamics of email forwarding are explored by Smith, Ubois, and Gross [3], who present a variety of issues that a user might consider when forwarding messages. This leads one to suspect that a user's forwarding behavior *does* in some way depend on their position in the network, in contrast with Wang et al.'s assertions.

Huberman and Adamic [11] present interesting findings based on email traffic at HP Labs. They highlight that "the tendency of individuals to associate according to common interests influences the way that information spreads ... it spreads quickly among individuals to whom it is relevant, but unlike a virus, is unable to infect a population indiscriminately." This is an especially relevant result for our analysis because it presents evidence that community structure may impact the spread of information through email forwarding networks.

Lerman and Ghosh conduct related research on a different form of online forwarding: the process of news stories spreading through Digg and Twitter [12]. Their paper's findings note that users' social network structure strongly influences sharing behavior, describing a process in which "users watch their friends' activities — what they tweet or vote for — and by their own tweeting and voting actions they make this information visible to their own fans or followers." While social network sites have more transparent community structures than private email and "retweeting" is not as direct an action as sending a message to a specific person, these findings indicate that people do consider the perception of their actions as they spread information.

3 Experimental Method and Algorithms

3.1 Community Detection Algorithms

Our analysis relies heavily on community detection — after all, what is community saturation without the community? Although most community detection study focuses on unweighted, undirected graphs, we use both weighted and unweighted community detection measures to determine whether our datasets exhibit different behavior when the volume of communications between nodes is considered. The potential for different community detection behavior when weighting is considered on various types of networks is noted by Newman [5]. As Newman writes, disregarding edge weight is a process that inherently loses data; we include this data and determine whether the connection weights are critical to understanding the cascades in email forwarding networks.

We began analysis using the Girvan-Newman algorithm [4]. We extended the SNAP implementation to take into account network weighting as specified in Newman's 2004 paper [5]. Briefly, this approach involves running the betweenness calculations on the network as if it were unweighted, then dividing the betweenness score for each edge by its weight before choosing the one to delete. The runtime for Girvan-Newman is $O(n^3)$, so Girvan-Newman is not a feasible candidate for large graphs like our Syria network, described below.

We use the Clauset-Newman-Moore (CNM) community detection algorithm (also built into SNAP) due to its markedly better asymptotic runtime (roughly $O(n \log^2 n)$ for large, sparse networks). [6]. We use

CNM community detection on all of our experimental datasets.

3.2 Data

3.2.1 Enron

Our first dataset is the Enron email network released by the Federal Energy Regulatory Commission, and made available by W. Cohen [7] and processed into a MySQL database by J. Shetty & J. Adibi [8]. Summarized representations of this network are available (e.g. in SNAP [10]) but do not provide the level of detail necessary to determine whether messages are forwarded, and only provide unweighted edges.

The database generated by Shetty and Adibi [8] contains identifiers for the 151 employees for whom data was released by FERC. The database contains messages from 1999-2002 and contains 252,759 messages. Most of the messages came from or were sent to people not in Shetty and Adibi's `employeelist` table, indicating that the contents of their mailboxes were not released by the government. In our analysis, we used only emails for which the sender and all receivers are among the 151 people, since we would not be able to track forwards outside of this group. This reduces the number of messages to 21,254.

3.2.2 Wikileaks Data — Stratfor "GI Files" and Syria

Because there are relatively few publicly available email datasets that include all of the content we require — subject, send time, body, and distribution for each message — we also analyzed email traffic compromised by the Wikileaks organization and exfiltrated from "global intelligence" company Stratfor [13] and multiple Syrian government and commercial organizations [15]. The released message traffic is a subset of the data compromised by Wikileaks.

The character of the Stratfor data is internal corporate email, while the Syria data often crosses corporate and governmental boundaries; each dataset contains some limited personal correspondence conducted on corporate networks.

Both Wikileaks datasets are more recent than Enron. The Stratfor email dataset is from 2004-2007 and contains 2,934 emails between 564 addresses, and 443 attachments. The Syria dataset is from 2006-2012 and contains 2,434,899 emails from 680 domains, 678,752 sender email addresses, and 1,082,447 recipient email addresses; the released subset we analyzed contains 16,419 email addresses and 86,591 emails.

3.3 Experimental Method

Our method consists of four basic steps, outlined below, plus a step to compare with a random network structure based on the same nodes.

3.3.1 Find Forwarded Content

We track three different types of information spread through forwarding: message contents, URLs, and attachments.

In each case, we generate output with an identifier for the tree (a subject, a URL, or a SHA-1 file hash) and a time-ordered list of times the information was forwarded. We say that a node has decided to forward if we observe a forwarded message from the user, and that they have not forwarded if we do not possess a further message containing the same information for which they are the sender.

Message contents: We track message forwards (what email users would achieve by simply clicking "Forward" in their email client) to follow messages. In the absence of unambiguous features like the `References` or `In-Reply-To` headers now common on forwarded email, we must rely on heuristics to find forwarded mail, such as the presence of "Fwd", "Fw", or other similar signal phrases in the subject. We note that even Wang et al.'s highly empirical study states that it relies on similar heuristics [1]. For each message in the network, we search for potential forwards of the message by looking for messages that include the entire subject, plus an optional mailing list prefix (e.g. [Eurasia] for a Stratfor list), plus a forward signal.

To avoid spurious forwards, we eliminate any message whose subject is solely a signal phrase ("Re:" or "Fw(d):") or is wholly blank. We ensure that potential forwards are in increasing order of time. We then "hook" together forwards identified in the previous step to find any multi-level forwards (a single message is forwarded, then forwarded again).

URLs: Inspired by Huberman and Adamic [11], we find specific URLs and track their spread through the network. This could provide insight on how a network attacker might convince a large number of employees of an organization to visit a deceptive malicious website, for instance, and is easier to detect than entire messages, which may be modified by mail clients or users along the way.

To track a URL's progress through the network, we search all messages for potential URLs, applying a regular expression to find URLs that begin with http:// or https://. We note that this is rather conservative, since

many email users will not type http:// into a message unless they are copying a link into the message, for instance from another message or a browser. However, this conservative behavior helps us avoid false-positive URL detections and has the side effect of focusing more on information spread, since it is more likely that a user is taking this information from elsewhere and spreading it through the email network. As in the above section, we ensure that potential spreading is in chronological order and eliminate multiple messages that pass the same information from the same sender to the same list of recipients. We eliminate URLs that are only sent once, reasoning that one appearance of a piece of information does not constitute a spread.

Attachments: We also track the spread of attachments, which has obvious implications for security and could also provide insight on how whether people are more or less likely to forward emails that contain attachments compared with simple text. One might, for instance, be more cautious about forwarding attachments if mailbox sizes are limited.

We follow a similar process to track attachments as they are forwarded. The Enron dataset does not include attachments, but both Wikileaks datasets do. The Wikileaks datasets provide the SHA-1 hashes of attachment files associated with each message. So, we search for these SHA-1 hashes in the metadata summary provided by Wikileaks and construct the flow of attachments through the network in chronological order. We apply the tree-culling procedures to eliminate files sent only once or sent repeatedly but to the exact same recipients by the same sender.

3.3.2 Construct Email Network

We construct a social graph based on message traffic. An email address is taken to represent a single user and we record an edge for each message exchanged between a pair of users. The weights will be used in portions of the next step.

3.3.3 Run Community Detection

We run both CNM and weighted GN on the networks, identifying and recording communities.

3.3.4 Follow Forwarded Messages through Network

For each of the forwarded pieces of information (message, identified by subject; URL; or attachment, identified by SHA-1), we consider each information-spreading message in chronological order.

For each message, we read the output of the forward-finding steps above and identify whether each recipient user forwards the message further. This is determined by considering the future (chronologically later) forwards for this piece of information and finding whether the given recipient is the sender of any later information-spreading message. We store this information along with the message so we can use it in the following steps.

Now that the outcome of each forward to each recipient has been determined, we "follow" its spread in chronological order. We defined a set of exposed nodes, that is, the set of nodes that have received a message that contains the given information (message, URL, or attachment). For each message, we first find the community to which the sender belongs by iterating through the communities found using community detection; once the community is found, we determine the community saturation by counting the number of nodes in the exposed set that are in the community and dividing by the total size of the community. We record that, at this community saturation level, the user chose to forward the message.

We then add all of the users who were exposed by this message to the exposed set. Then, for each destination user who does not further forward the information, we repeat our analysis above: we find the community to which the user belongs, find the saturation of that community, and record that the user did not forward the message.

We continue this process for each information-spreading message, and for each piece of information that we have detected being spread through the network.

3.3.5 Compare with Randomized-Edged Network

In addition, to determine the significance of our results, we complete the following procedure. We first construct a graph with the same nodes as the network under study. Then, we insert the same number of edges between the nodes, but assign the endpoints of these edges randomly. We repeat the same analysis above and compare results, using the same forwarding paths; what has changed is the network structure. For simplicity (to avoid recalculating all of the edge weights), we use only the unweighted CNM community detection for this step.

4 Findings

4.1 Community Saturation v. Probability of Forward

We found that the probability of a user forwarding a message fell as community saturation increased. This finding is observed in each network on each type of forwarded information.

These results support our hypothesis that community saturation affects whether a user will forward a message or not. As predicted, as more members in a user's community see the forwarded email, the user becomes less likely to forward that same message. There is some "noise" at small saturation values but the trend is clear.

These findings are plotted in Figures 1-3. Figures 1-6 are binned by saturation in bins of with .05.

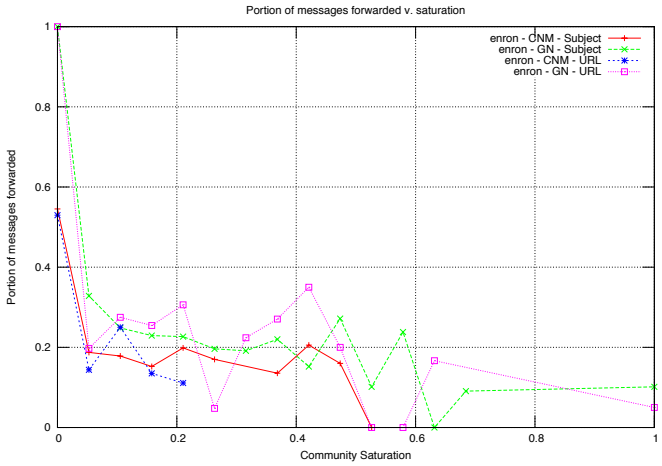

Figure 1: **Enron**

4.2 Behavior with Messages, URLs, and Attachments

We find approximately the same behavior for users spreading messages, URLs, and attachments through the network. This is illustrated in the Figures 1-3.

4.3 Effect of Network Community Structure

Our random edge-shuffling experiment shows that the randomly-shuffled graph produces much less meaningful results. In one case — Enron subject-based message forwarding — the trend is reversed, with a convex graph in contrast with all other results; note that the saturation values take a very small range. In other cases, the same

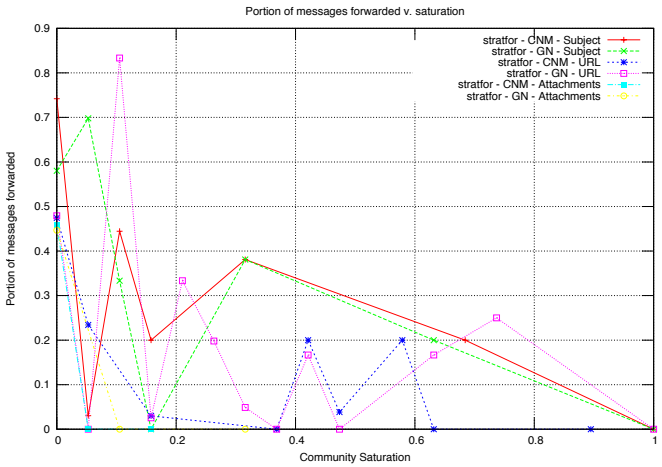

Figure 2: **Stratfor**

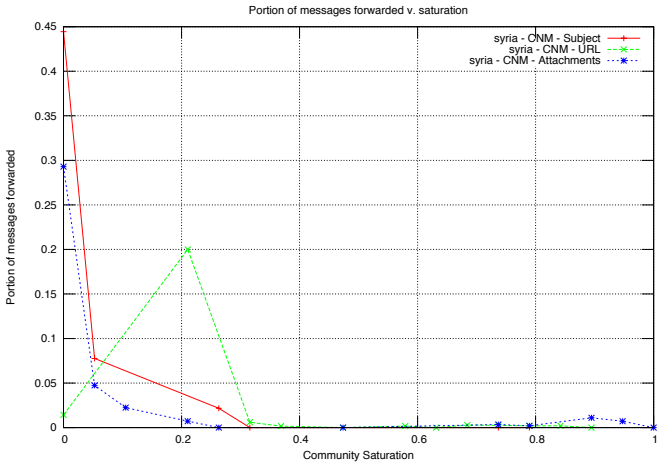

Figure 3: **Syria**

general shape of the graph is observed but the proportion of forwards begins lower and declines more rapidly. Since the community structure of the random-edge network tends to have more evenly-sized communities, a much smaller number of messages leave the community. These results are illustrated in Figures 4-6. We also note that CNM and GN produce similar large-scale trends, but different behavior at smaller scales.

4.4 Limitations and Areas for Further Study

The proportion of messages forwarded outside of the community of the sender limits our ability to assert a causal relationship between community saturation and forwarding. The percent of informatin-spreading [sender, receiver] pairs that leave the sender's community are listed in the table.

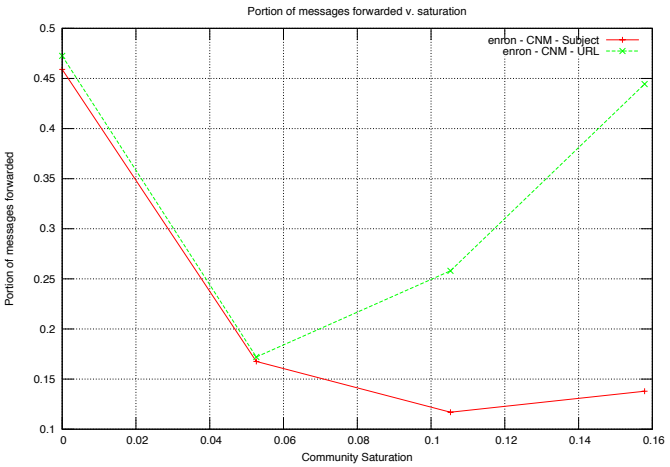

Figure 4: **Enron - random**

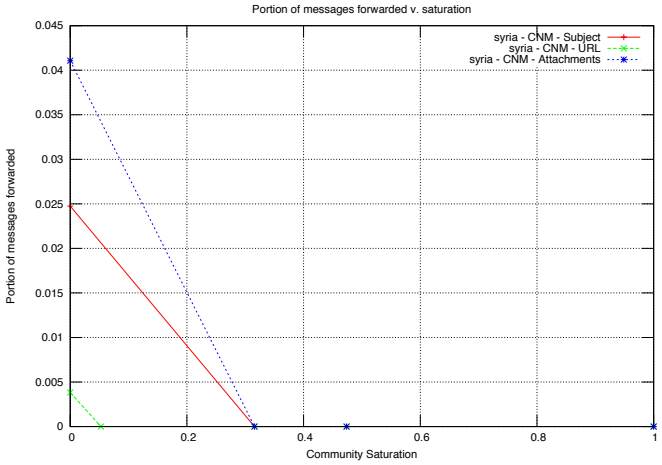

Figure 6: **Syria - random**

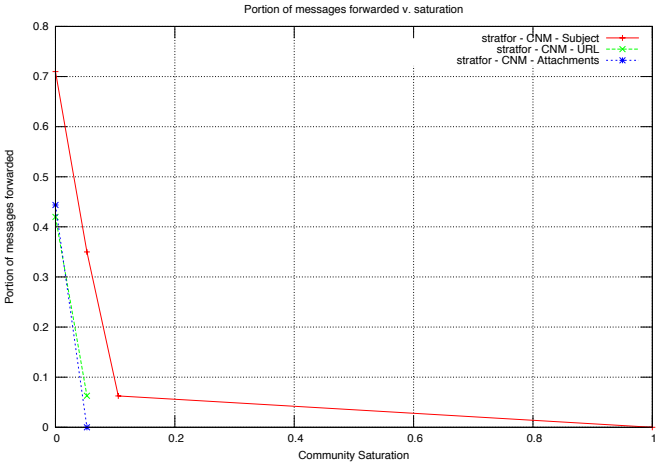

Figure 5: **Stratfor - random**

Dataset, Information, Alg.	% Out Curr. Cmty.
Enron, Subject, CNM	77.8%
Enron, URL, CNM	62.3%
Enron, Subject, GN	75.5%
Enron, URL, GN	61.1%
Stratfor, Subject, CNM	35.7%
Stratfor, URL, CNM	67.7%
Stratfor, Attachment, CNM	34.0%
Stratfor, Subject, GN	71.4%
Stratfor, URL, GN	99.1%
Stratfor, Attachment, GN	85.4%
Syria, Subject, CNM	99.2%
Syria, URL, CNM	90.9%
Syria, Attachment, CNM	98.9%

Further study may indicate a stronger match with measures of the degree to which a node bridges distinct communities or other measures.

We also note that the best empirical analyses of email traffic rely on having *full* — not selected — email traffic, ideally in a structured format. The Enron dataset is limited because only the email sent to and from 151 high-ranking individuals was released by federal regulators. The Stratfor and Syria emails are limited because they are by their nature incomplete samples of email traffic, since unauthorized users obtained them. Even so, Wikileaks writes that it has over 4 million messages from Stratfor, but the released corpus contains only 2,934 [13]. Stratfor will not confirm the authenticity or completeness of the released email, though inspection of the mail and the reaction of the company indicates that the released messages are likely to be legitimate [14].

In addition, many organizations make heavy use of email aliases. Without knowledge of the contents of these email aliases, we cannot analyze the spread of messages with complete accuracy. This is particularly apparent in the Stratfor data, and to a lesser extent in the Enron data. Without internal knowledge of the structure of the email lists, we cannot recover list membership and follow those forwarding trees with complete accuracy.

References

[1] D. Wang et al., *Information Spreading in Context.* In Proc. WWW '11, p735-744.

[2] D. Liben-Nowell and J. Kleinberg, *Tracing information flow on a global scale using Internet chain-letter*

data. PNAS March 25, 2008 vol. 105 no. 12 4633-4638.

[3] M. A. Smith, J. Ubois, and B. M. Gross, *Forward Thinking*. In Proc. CEAS 2005.

[4] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*. PNAS June 11, 2002 vol. 99 no. 12 7821-7826.

[5] M. E. J. Newman, *Analysis of weighted networks*. Phys. Rev. E 70, 056131 (2004). arXiv:cond-mat/0407503 [cond-mat.stat-mech]

[6] A. Clauset, M. E. J. Newman, C. Moore, *Finding community structure in very large networks*. Phys. Rev. E 70, 066111 (2004).

[7] `http://www.cs.cmu.edu/~enron/`

[8] J. Shetty and J. Adibi, *The Enron Email Dataset Database Schema and Brief Statistical Report*. `http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf`

[9] `http://bailando.sims.berkeley.edu/enron_email.html`

[10] `http://snap.stanford.edu/data/email-Enron.html`

[11] B. A. Huberman and L. A. Adamic, *Information Dynamics in the Networked World*. Lect. Notes Phys. 650, 371-398 (2004).

[12] K. Lerman and R. Ghosh, *Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks*. In Proc. AAAI Conference on Weblogs and Social Media 10, p90-97.

[13] `http://wikileaks.org/the-gifiles.html`

[14] `http://www.stratfor.com/hacking-news/subscriber-info`

[15] `http://wikileaks.org/syria-files/`