Community Structure and Information Flow in Usenet: Improving Analysis with a Thread Ownership Model

Mary McGlohon*
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
mmcgloho@cs.cmu.edu

Matthew Hurst
Microsoft Live Labs
1 Microsoft Way
Redmond, WA 98005
mhurst@microsoft.com

Abstract

Studying Usenet provides unique insight into online communities, since there is often a pre-defined space for community interaction. Here, we examine a large set of posts in nearly 200 politically-oriented newsgroups over a period of 4 years, using a unique and novel approach to the analysis. Our study is multi-scale: not only do we examine the newsgroups individually and compare measurements of different groups, but we also examine the relationships between the groups. Since users often post to multiple groups, shared participating authors or cross posts may be a way to assess how closely related, content-wise, two groups may be. However, this also causes some confusion of relevance. To combat this effect, we develop an "ownership" measure of an article to the set of cross-posted groups, based on the posting activity of the author. We show that this ownership measure can greatly improve assessment of information diffusion within and between groups, and suggest using it to improve analysis in other problems.

Introduction

Social networks, both on- and off-line, are rich structures of communities and communities-within-communities. An individual may be a member of multiple social circles. While this property enhances the flow of communication across networks, it makes community identification difficult in most on-line social network data. Unlike many Web 2.0 communities, Usenet has a pre-defined structure for topics of discussion, which allows us to identify which individuals are most responsible for bridging communities and aiding in information diffusion not only within, but also *between* communities. In this work we examine the structure of communities and diffusion patterns of nearly 200 politically-oriented newsgroups, both the interactions inside newsgroups and, in particular, at the borders of them, where membership, interests, and topics of discussion overlap.

Studying the pre-defined Usenet groups allows one to bypass the obstacle of community detection. This advantage, however, presents a host of interesting challenges, as the borders do tend to blur. *Cross-posting*, where a single article is posted into several groups simultaneously, is frequent in Usenet. While studying cross-posts can aid us in finding gateways for information transfer, an improper cross-post leads to confusion of relevance. Since users can simultaneously (and nearly without cost) "spam" multiple groups, and often times respondents to an article will "reply-to-all", an entire conversation can appear to happen in a group when none of its regular readers are taking part. To combat this, we propose a framework for assessing *ownership* of an article, or post.

Our work is one of the first principled approaches towards analyzing diffusion patterns in Usenet. Our contributions are the following: We perform a study of a large set of Usenet newsgroups over an extended period of time, comparing the structure of the induced social networks. We find that induced networks of groups obey a form of the densification power law, with slope of 1.2. However, despite this structural similarity, we find that reciprocity and degree distribution varies in the different groups. Understanding these structures helps us properly assess similarities in newsgroups based on membership and cross-posting activity. We then present a framework for assessing which of many cross-posted newsgroups is responsible for most of the activity in a thread, and which ones are responsible for influencing other groups. Using this framework, we show how cross-posting later in a conversation induces higher activity, which illustrates the flow of information between communities, and observe some precise diffusion patterns between Usenet groups.

Related Work

The rise of Web 2.0 has provided extensive data for studying how information and ideas travel through a social network. Researchers have largely focused on blogs in this area (Adamic and Glance 2005; Adar and Adamic 2005; Leskovec et al. 2007), but have also applied the same ideas to recommendation systems (Leskovec and Kleinberg 2006), and email (Kossinets, Kleinberg, and Watts 2008; Nowell and Kleinberg 2008).

Microsoft's Netscan Project has conducted a very thorough study of Usenet discussion patterns, depicting hierarchy of newsgroups and their changes between 2000 and 2004(Turner et al. 2005); studying the *social roles* of Usenet authors (Fisher, Smith, and Welser 2006); and creating a vi-

*Some work completed while on internship at Microsoft Live Labs.

sualization tool for different author roles identified (Viegas and Smith 2004). Other studies have focused specifically on discussion forums similar to Usenet. Blog comments can serve as forums for a specific topic, and can be used to assess controversy of blog posts (Mishne and Glance 2006). A similar study was applied to the Slashdot.org community, suggesting using a controversy measure based on the patterns in the threaded network (Gómez, Kaltenbrunner, and López 2008).

Preliminaries

One of the first online forums, Usenet originated in 1979, preceding Web 2.0 by decades. While overall its activity is declining, Usenet is still in use and there are many very active communities (Turner et al. 2005), making it an excellent resource for social network analysis. We next describe the data set and methods we used to extract threads for analysis.

Data description

We collected data from nearly 200 newsgroups with posts between 2004 and 2008, using a subscription service. In the interests of capturing a representative subset of data relating to political discussions, we selected all newsgroups available with the substring "polit" in the name[1]. We chose to focus on political newsgroups because politics is a topic that permeates most cultures, and can be used to compare cross-cultural groups. Indeed, there were many different regions of the world represented, including some groups for specific U.S. states. Around 70 were `alt.politics.*` subgroups, on topics such as political parties or regions, with another 20 topical groups under `talk.politics.*`. Others were devoted to regional discussion, either for local areas or topics. 22 were local United States (`va.politics`, `seattle.politics`, etc.), 6 were local Canadian groups, (`edm.politics`, `bc.politics`, etc). 3 from `de`, 4 from `dk`, 3 from `es`, 7 from `it`, 4 from `tw`, and 9 from `uk`. In addition there were several other international domains with one or two groups represented. Of these newsgroups, there were 19.6 million unique articles, and 6.2 million of these were cross-posted to multiple groups in the data set.

Thread and author network construction

One method of looking at patterns of information diffusion is extracting *threads*, conversation trees of replies. The algorithm for thread induction is simple. Each post is labeled with a *message-id* and *references*. References may be several posts– for our purposes we take the last one on the list, as it is the most recent and therefore the direct reply. Other references already occur further up in the tree. This forms several *cascades* (as they are referred to in related work), or conversation trees. Each message has at most one parent, and of the entire network of posts each connected component represents a thread, which may stretch across several groups (thanks to cross-posts).

From the post-reply trees one can induce a social network of *authors*. Every message has an e-mail address to identify the message author. The resultant social network is weighted for multiple links between two authors. This is similar in spirit to inducing a network of blogs based on citations of posts (as in (Leskovec et al. 2007)). As a point of reference, there are around 0.5 million authors total, and 4.7 million unique edges between them.

Structural Analysis of Communities

To provide context for diffusion experiments, we will first provide a structural analysis of the different communities represented in our data set. We will first examine newsgroups on an individual level– since this study is primarily for understanding the nature of the data, we have condensed our results for space. We also examine the groups on a larger scale by creating clusters of groups based on similarity of authors or posts.

Comparing structure in newsgroups

First we examine structural properties within each newsgroup. Here, we make an induced social network $G = (N, E)$ of authors based on replies to posts. In each group, if author a_n replies to a post by author a_m, there is a directed edge e_{mn} from a_n to a_m.

Size The size that a group reaches is one key feature examined. Interestingly, groups seem to mimic the *densification power law* discovered by Leskovec et. al– as a graph size grows in nodes, the number of edges increases superlinearly (Leskovec, Kleinberg, and Faloutsos 2005). However, while the densification law is traditionally applied to several snapshots of the same graph at different points in time, here we observe several different groups at the same point in time. The plot of edges vs. nodes is shown in Fig. 1. The weighted graph, where the weight on an edge is the total number of replies, also follows densification with exponent of 1.3 (plot omitted for space). There are some notable, interesting anomalies– the points far below the fitting line (with abnormally low reply rates) are `tw` domains. The ones above the fitting line (high reply rates) tend to be in European domains.

Degree and reciprocity We have shown that groups tend to maintain a certain edge to node property, but how are these edges distributed? The degree distribution indicates how skewed interactions are– a steeper slope on a power-law fit implies a higher proportion of activity by the "core" authors. Fig 2 shows the in-degree vs. out-degree power law exponents for groups that did fit such a distribution, based on log-binning of histogram data. Among groups that had a fit value of $R^2 > .95$, the power-law exponent ranged from -0.95 (`no.samfunn.politikk.diverse`) to -1.5 (`alt.politics.conservative` for in-degree. Out-degree power law exponent ranged from -0.86 (`no.samfunn.politikk.diverse`) to -1.8

[1]While a number of other sampling methods were considered, we chose this one for simplicity; due to the structured nature of Usenet, this was a reasonable method. The complete list of newsgroups used may be found at `www.cs.cmu.edu/~mmcgloho/data/usenet.html`.

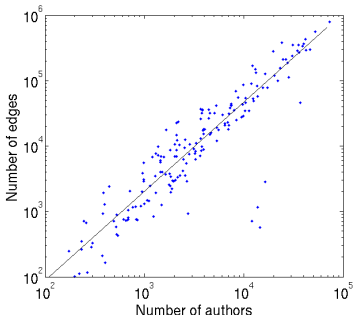

Figure 1: Number of author-to-author edges (interaction pairs) in groups vs. number of nodes (authors) in groups, based on replies. The power-law exponent is 1.2.

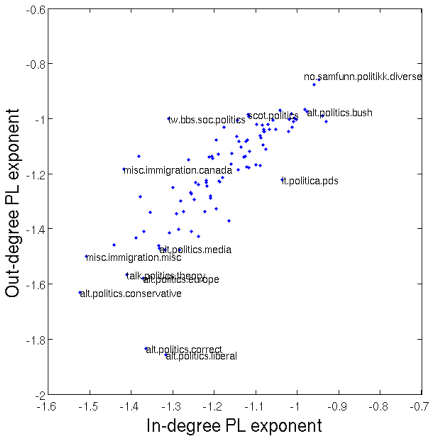

Figure 2: In-degree power law exponent vs. out-degree power law exponent, for groups with an R^2 fit of greater than 0.95. Some outliers are labeled. There is a general correlation of in-and out-degree, but there is a great deal of range in the steepness of slopes in the degree distribution.

(`alt.politics.liberal`). While correlated, there was a wide range of exponents, and some did not even appear to be heavy-tailed, which was surprising.

Reciprocity between groups represents whether most users reply to each other. The formula for reciprocity may be found in (Bollobas 1998), but it is essentially a ratio of the number of pairs of nodes that have a mutual edge to the number of pairs of nodes that have a non-mutual edge (one that goes only one direction). A group with no reciprocated edges would have reciprocity 0, and a group where all edges are reciprocated would have a reciprocity of 1. The most reciprocated group (`hun.politika`) had a reciprocity of up to 0.58, and the least reciprocated group `tw.bbs.soc.politics`, had a reciprocity of 0.057. Interestingly, with the exception of `hsv.politics` (Huntsville, Alabama), all of the top 20 high-reciprocity groups were European, and most of these highly-reciprocity groups did not fit a power-law degree distribution at all. The low-reciprocity groups generally had low traffic (fewer than 100 authors in any given year, with the exception of `tw.bbs.soc. politics`). All of Taiwan-based groups in our data had very low reciprocity.

Similarity Measures Between Newsgroups

We have now compared the individual groups and showed some of their differences. But how can we draw similarities between the groups? Cross-posting may help provide us with information on how related different groups are, by making the assumption that if authors regularly post the same articles into multiple groups, then the groups share those related articles and are likely of similar motivation. Likewise, groups with shared authors may be related.

We first measured how often cross posts occurred. For this, we use the Jaccard coefficient: the ratio of intersecting articles to the union of articles in both groups.

$$Sim(g_1, g_2) = \frac{|Articles(g_1) \bigcap Articles(g_2)|}{Articles(g_1) \bigcup Articles(g_2)}$$

Fig. 3 is a visualization of the resulting network [2], where an edge represents similarity greater than 0.10, and a thick edge similarity greater than 0.20. There are some interesting groups forming: the large cluster on the right includes most of the Canada local groups joined with thick edges. Notably, the group `qc.politique` was missing– we found that it actually had a higher similarity with `fr.politique` than with any of the other Canadian groups, likely due to language. Also joined to the Canada cluster (green) are other general politics groups for English speaking countries, such as the U.K., Australia, and New Zealand. In the center there is a cluster largely devoted to the U.S., with most of the regional and statewide groups on the bottom (blue). There is a surprising rate of cross-posts in this area; however, some of the less-well-connected regional groups tend to be connected in an intuitive manner– for instance, `sdnet.politics` (San Diego, Cali.) and `ba.politics` (Bay-Area, Cali.) are connected, and `houston.politics`, `dfw.politics`, and `austin.politics`, three groups for major cities in the state of Texas, along with `tx.politics`, form a clique. Above the local-U.S. cluster (in red) is a cluster of most of the `alt.politics.*` hierarchy– cross-posting is very high among these groups. To the left is a fourth cluster (yellow), mainly centered around topical groups such as guns, drugs, or specific political philosophies, with fairly intuitive connectedness. Otherwise, groups joined by language or physical borders tend to cluster together. Groups focused on Sweden, Taiwan, Norway, Hong Kong/China, and Netherlands/Belgium are related. About half of the groups are not shown, as they had no edges above the threshold.

We also measured similarity based on Jaccard coefficient of the author participation in each newsgroup, where similarity is the ratio of the size of the intersection of authors in

[2] All network visualizations in this work, including illustrations of threads later, use Eytan Adar's GUESS Graph Exploration tool, (Adar 2006).

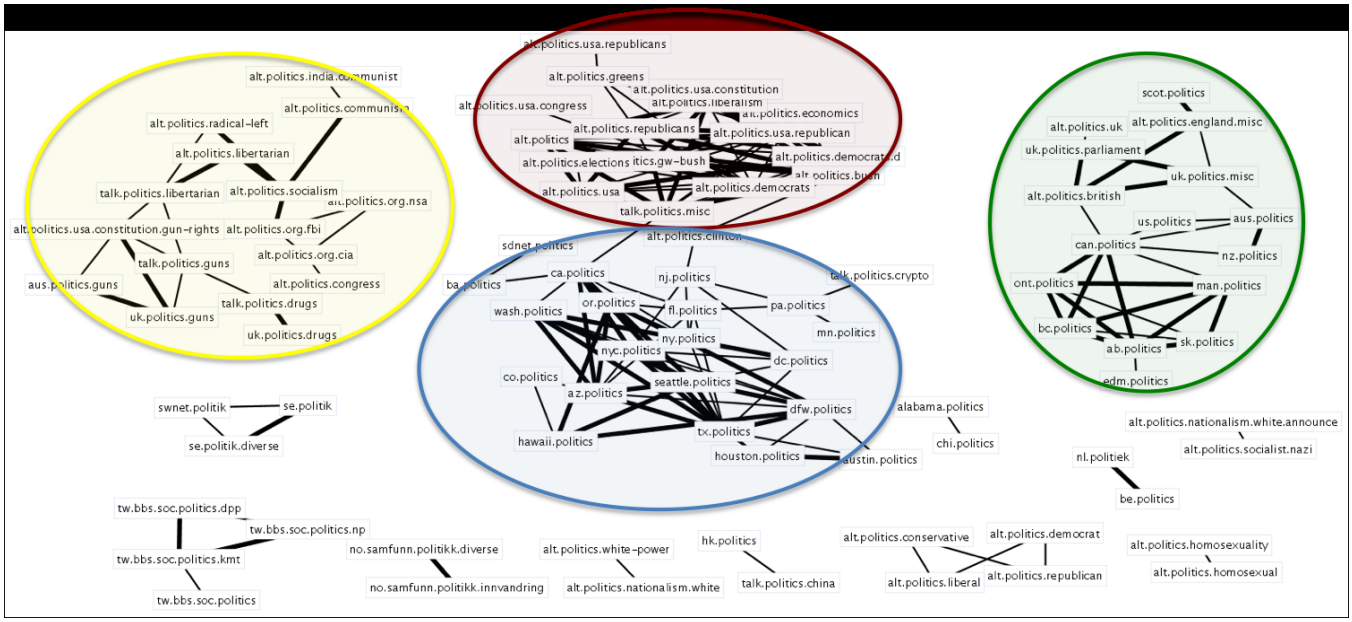

Figure 3: Newsgroups clustered by cross-posting based on Jaccard coefficient. A thin edge indicates a similarity of over 0.1, and a thick edge of over 0.2. In the center there are distinct clusters for local U.S. politics groups and the main `alt.politics` groups. On the left are topical groups for issues and some political philosophies, and on the right are clusters for local Canadian groups and for other English-speaking countries. Otherwise, groups sharing language or physical borders tend to group together.

each group to the union of authors (in the same manner we assessed cross-posts). Here we thresholded edges at an coefficient of 0.2, thick edges at 0.3, which resulted in about half of the groups being connected to at least one other group. The visualization is omitted for space; however, we found that the structure formed similar clusters to those in Fig. 3.

Next we will study patterns of diffusion, exploring whether similarity leads to more information flow.

Diffusion Inside and Between Communities

In the previous section we completed a multi-scale analysis of the Usenet sample, both contrasting differences between the groups and clustering them based on similarity measures. We next analyze threads themselves, particularly focusing on how threads move between groups. Often times even when a thread is initially posted to one or a few groups, it may be later cross-posted to others. The thread may be picked up by the new groups, but even if members in the old groups are no longer interested in the discussion (or never were), people in other groups may still cross-post to that group (the "reply-to-all" effect). Therefore, as we describe the interactions we try to consider when we can truly consider a discussion as occurring in a given group. To that end, we propose a measure of *ownership* for authors and for articles, and show how it aids in studying diffusion patterns.

Post ownership

Since nearly half of all posts are cross-posted, it is difficult to assign ownership from articles alone. However, based on the *authors'* posting patterns, we can often discern where their loyalties lie, so to speak. If an author usually posts into g_1 and only occasionally cross-posts into both g_1 and g_2, then it is a safe assumption that posts written by that author "belong" to g_1. To aid in formalization, we define the following expressions:

Author-group activity, $act(a, g)$ is defined as the percent of author a's posts that are posted into group g. These may be cross-posted, so $\sum_g act(a, g) \geq 1$.

While this may give a realistic distribution of where an author is cross-posting, we feel that in order to capture whether an author truly considers himself a member of a group, we need to determine where that author is writing unique posts, because many cross-posts are unintentional "reply-to-alls". Therefore, we define *Author-group devotion*, $dev(a, g)$, as the percent of author a's posts that are *only* posted into group g, and not cross-posted into any other groups. Therefore, $0 \leq \sum_g dev(a, g) \leq 1$. From there, we can define a group g_i's degree of ownership of a post, based on how devoted the post's author is to the groups it is posted into.

$$own(g_i, p) = \frac{dev(a(p), g_i)}{\sum_{g_j | p \in g_j} dev(a(p), g_j)}$$

A simple extension gives us the ownership of a *set* $\mathcal{P}$ of posts, taking the mean of the ownership of each post. One can apply this ownership score to the set of all posts that have occurred, whether uniquely or as a cross-post, into a group[3]. In this manner we have aggregated ownership for posts and devotedness scores for authors, and

[3]For some posts $dev(a(p), g)$ is 0 for all groups in question– this is a relatively rare occurrence, particularly on the thread level.

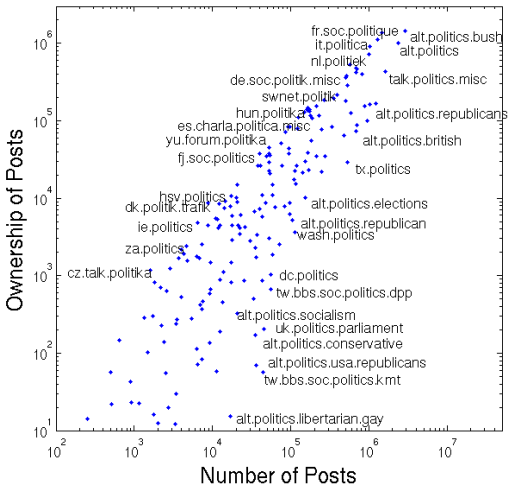

Figure 4: Post ownership vs. number of posts in groups, with some outliers marked. For example, `talk.politics.misc` had around one million posts, and an average ownership score of around 0.4, so the point occurs at $(10^6, 4 * 10^5)$.

show some comparisons of the different groups in Fig. 4, with outliers marked. We find that some groups "own" a large amount of their posts, while others have much sparser relative ownership. For instance, `fr.soc.politique` has a ratio of 0.92 while `alt.politics.bush` has an aggregate ownership score of 0.56: so under this score, `alt.politics.bush` actually has less activity. Some groups had even lower ratios of ownership– for example, `tw.bbs.soc.politics.kmt`'s was around 0.003.

We illustrate the importance of ownership using an example. In Fig. 5, we show a conversation cross-posted to several groups, and then label each node with the group that the author most "belongs" to (based on highest ownership). The original article, "Kiss the national parks goodbye", was cross-posted to several large newsgroups, including `talk.politics.misc` and `alt.politics`. The second node from the left on the second level was a reply to that post, which was cross-posted to `talk.politics.misc`, `seattle. politics`, `or.politics`, and a few other local politics groups. According to our ownership rules, the bulk of the thread was made by authors that mainly posted to `seattle.politics` (16,000 members, marked in green) and `or.politics` (10,000 members, blue). Authors posting primarily onto talk.politics.misc (a much larger group, with over 50,000 participants) are marked in red. Even though nearly all of the posts were cross-posted to `talk.politics.misc`, few of the "devoted authors" of that group participated. Considering the subject line, it is not surprising that such a subject would appeal more to members of groups in the Pacific Northwest, which has a higher concentration of national parks.

The largest thread was over 9000 posts, occurring in major `alt.politics` subgroups and `talk.politics.misc`, and focused on the 2004 election. It was cross-posted to 38 groups during its tenure– yet, 85% of ownership was concentrated in three groups.

The effects of cross-posting on threads in groups

Once we have established which groups dominate conversation for a given thread, we can develop a better understanding of how cross-posting affects how well-received a thread becomes inside a group. We can start to answer the questions: How does cross-posting affect a conversation? Does a conversation pick up when cross posted, or die off? How does a thread fare if it begins in a group, compared to when it begins elsewhere? To assess whether cross-posting helps or hurts activity in groups, we can divide conversations happening in a group g_i into the following four categories:

1. An article is initially posted to g_i and never cross-posted to other groups in our data set. (No X-post)
2. An article is initially cross-posted both to g_i and another group in the data set. (Initial X-post)
3. An article is initially posted to g_i and, later in the conversation, a reply is cross-posted to a different group. (Late X-post, original group)
4. An article is initially posted to another group, and later in the conversation debuts in g_i. (Late X-post, late group)

To compare these cases, we took the ownership of the set of posts in the thread. (In the fourth case this means taking the ownership of all posts below the point in the conversation where g_i appears). In Fig. 6, we show the distribution of thread sizes, for the different "types". All types follow a heavy-tailed distribution. However, it is clear that most of the largest threads are of the "late-cross-posting" type. Furthermore, there is not much difference in overall thread size for threads with no cross-posts and those that are only initially cross-posted to multiple groups– so simply the act of cross-posting may often be associated with spam.

We recognize that there is some correlation between natural thread size and type (by definition, threads of type 3 and 4 must be at least of size 2, for instance). We can make a better assessment by instead examining what happens not simply to the thread overall, but what happens *within each group*. If we measure the cascade size based on ownership for a given group, we can more confidently state whether the act of cross-posting induces conversation. In doing this, we find that Type 4 threads do indeed have more activity. We are only measuring the size *below* the point where it reaches the group, making it a comparable measure to types 1 and 2. The resultant PDF is shown in Fig. 6, normalized as there are relatively few Type 4 occurrences.

In other words, mass initial cross-posting does not lead to high activity within any given group. However, if somewhere in a thread an author decides it is relevant to group g_i and cross-posts, then g_i tends to gain more activity than it would for a post that was not cross-posted at all. Perhaps this is indicative of authors "discovering" threads that are relevant to a given group, and "recommending" these threads to the group by cross-posting their replies– indeed, we find that for cases where a post is later cross-posted to a new group,

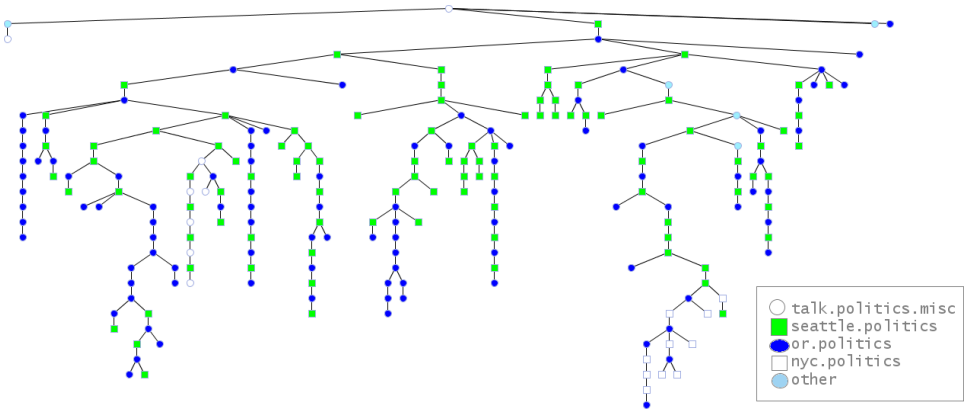

Figure 5: An example of a thread that is posted into several groups but is "owned" by a very small number. It is described in detail in the text. While the original article was cross-posted to several large newsgroups, including `talk.politics.misc` and `alt.politics`, most of the posts are from authors who primarily make their non-cross-posts into `or.politics` and `seattle.politics`.

about half the time the person who introduces the post is "devoted" to both the old group and the new group.

One example of this phenomenon occurs in a thread with subject line "The truth about British Racism & Imperialism". It begins by being cross-posted to `alt.politics.british` and `uk.politics.misc`. At one point in the conversation, one author replies saying "If you can be Scottish and British, why not Asian and Scottish?" A second author, who we have labeled as most "devoted" to `scot.politics`, then posts "Why not be Asian and Scottish? Most Asian people in Scotland consider themselves to be both." In the process of replying the author also sends the reply to `scot.politics`. At that point, there is an explosion of conversation– in fact, we find that 79 percent of the conversation occurs below this point, and largely among authors in `scot.politics`. We show a diagram of the conversation in Fig. , emphasizing the point at which the late cross-posting occurs. Taking into account this mechanism of "discovery", we next assess diffusion in terms of thread ownership.

Information flow based on post ownership

Without an idea of where posts are truly occurring, measuring how information flows across groups becomes difficult to assess. If a parent post p_p is cross-posted to g_1, g_2, g_3, and an author then replies to it by adding a child post p_c into g_4, how does one assess where the new author read the original post– that is, which group influenced her to form edge e_{pc}?

The goal is to find an *influence measure* for any two groups, based on a given edge, which we can extend to the entire set of threads. We would like a score $Infl_{e_{pc}}(g_p, g_c)$ for each possible pair of groups. Without ownership information, one might assign the influence as a distribution from all of p_p's groups and all of p_c's groups. For each pair,

$$SimpleInfl_{e_{pc}}(g_p, g_c) = \frac{1}{|(g_k|p_p \in g_k)|} * \frac{1}{|(g_l|p_c \in g_l)|}$$

Under this case, since there are three groups in the parent post, and one in the child post, $SimpleInfl_{e_{pc}}(g_1, g_4) = \frac{1}{3}$. To get an influence score between two groups over an entire group of threads, one would simply sum the influence scores for each pair of parent-child posts. However, this measure has shortcomings: it ignores the fact that some cross-posting may be meaningless to authors who post only to a certain group. Therefore, we introduce ownership. We may decide to assign influence based on how devoted the parent post's author, $a(p_p)$, and the child post's author, $a(p_c)$, are to each group. The score for any pair of groups $(g_p, g_c|p_p \in g_p, p_c \in g_c)$ is then:

$$OwnInfl_{e_{pc}}(g_p, g_c) = dev(a(, g_i) * dev(a, g_j)$$

Still, we would like to take it a step further, to answer the question, *How often do authors in g_c respond to a post they first saw in g_p?*. One would then measure not g_p's influence based on the parent distribution, but rather the child author's distribution:

$$ChildOwnInfl_{e_{pc}}(g_i, g_j) = dev(a, g_i) * dev(a, g_j)$$

These three potential measures allow us to attribute influence over the entire set of threads. Summing over each e_{pc} where an edge is a reply, and normalized based on the "influencees" we can get a total score of influence from each group to another. Under $SimpleInfl$, we find that a slim majority of the mass (57%) is along the diagonal of the adjacency matrix. By using $OwnInfl$, attributing the flow from an ownership distribution of the parent post, into an ownership distribution of the child's post, 67% of the mass is along the diagonal. Taking it a step further, by attributing influence based only on the newer author, under $ChildOwnInfl$, 85%. This would seem the most intuitive measure of influence, as one would expect most influence to occur within a group.

Based on the third measure we can claim that perhaps 15% of the time, information is traveling from one newsgroup to another. Which groups are responsible? Based on $ChildOwnInfl$, we found that the

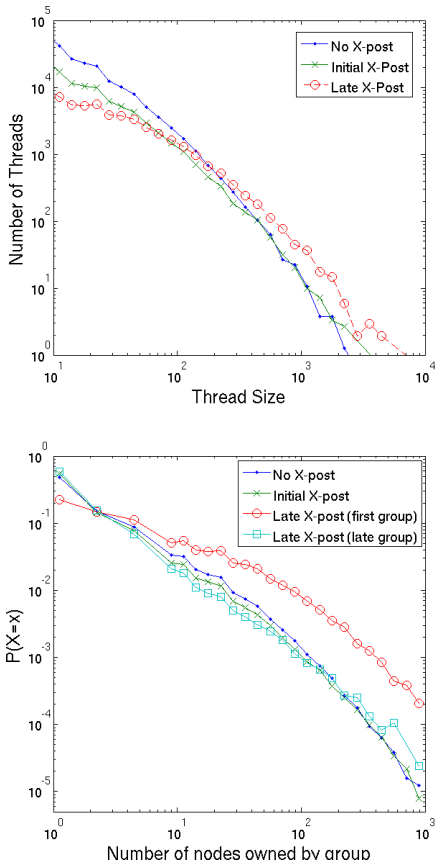

Figure 6: **Top:** Histogram of thread sizes, where each thread is either never cross-posted, cross-posted only at the root, or cross-posted later. Most of the largest threads tend to have late-occurring cross-posts. **Bottom:** PDF distribution for per-group thread ownership. Here, threads are double-counted for each group they appear in– however, posts are divided amongst the groups such that each *post* is only counted once. For the first two types, a higher proportion of the probability mass is concentrated in less activity, while late cross-posting leads to higher activity in the new groups.

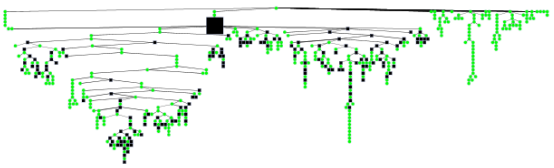

Figure 7: An example of a thread that is first posted to `alt.politics.british` and `uk.politics.misc`, but later is cross-posted into `scot.politics`. At the point which the third group is added (denoted by a large black square node), the conversation takes off, and 79 percent of all nodes occur below that point. `scot.politics`-owned posts are marked in black.

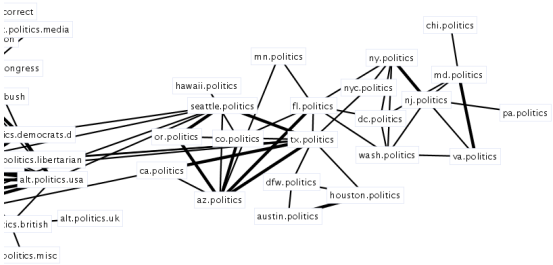

Figure 8: Similarity based on devoted authors, focusing on the local US groups. A thin edge represents a Jaccard coefficient of ≥ 0.08, and a thick edge ≥ 0.1.

most influential were often the ones with the largest mass, such as `alt.politics.bush` and `alt.politics`, but were more often simply the larger groups in a cluster, such as `can.politics` in the Canadian groups, `seattle.politics` in the local US groups, or `talk.politics.guns` for topical groups. The following edges had the highest influence scores:

Influencer	**Influencee**
`swnet.politik`	`se.politik.diverse`
`de.soc.politik.misc`	`bln.politik.rassismus`
`can.politics`	`man.politics`
`can.politics`	`ab.politics`
`can.politics`	`bc.politics`
`can.politics`	`ont.politics`
`uk.politics.misc`	`uk.politics.constitution`
`uk.politics.misc`	`uk.politics.parliament`
`talk.politics.drugs`	`uk.politics.drugs`

Group similarity based on shared "devoted" authors and shared posts

This new framework of ownership brings previous measures of group similarity into a new light. We can re-assess group similarity based on "devoted" authors. By redefining group membership from "any member who posts into a group" into "any member who, at some point, single-posts into a group", and then taking the Jaccard coefficient, we paint a different picture of which groups truly share members. Naturally, the similarity scores are lower. One can also build a network using similarity of shared ownership of posts: a post is shared between two groups if $dev(a(p), g_i) > 0$ for both groups. While the general structure is similar, there are a few interesting differences. For example, the devoted-author network has a much more distinct divide in the local U.S. groups– with a couple of exceptions, the groups appear to be neatly divided between cities/states on either side of the Mississippi River (see Fig. 8).

Conclusion

We have analyzed a large set of Usenet newsgroups, comparing structures of induced social networks for each group, and considering similarity based on activity. We have shown that there is a power-law relationship between the number of nodes and the number of edges in the induced author social

network, and showed how reciprocity and the skewness of degree vary per group. At a higher level, we have showed that one can visualize similarity between newsgroups, based on membership and cross-posts. We also take the unique approach of looking at diffusion not simply between individuals, but between groups.

While cross-posting aids in analyzing similarity between groups, when it comes to assessing relevance within groups, cross-posting becomes a barrier to understanding. Therefore, we have proposed an ownership measure, which assigns posts in a thread to groups based on how "devoted" the post authors are to the various groups. Our ownership measure is an excellent tool for many applications in data analysis. By assigning ownership of posts to groups, we observed how threads evolved as cross-posts occurred. By looking at different "types" of cross-posting activity, we demonstrated that while cross-posting, when initially in a thread, does not lead to more activity, a cross-post that occurs later in the thread is correlated with higher activity. Furthermore, we were able to create an influence measure between groups, based on the ownership of parent and child threads. These experiments in cross-posting activity that examine the devoted authors and activity in groups are particularly relevant, as identifying individuals who are devoted to multiple groups serves to better understand how information is transferred across social group boundaries.

Future work using this framework abounds. We have established ways in which ownership aids diffusion analysis in Usenet. Other applications include business intelligence problems such as targeted advertising, tracking the "health" of an online group in order to predict which groups survive and which die off (as in (Backstrom et al. 2008)), or even for building a "group recommendation" system for online social networking forums, to recommend new forums for users. Other domains demand some measure of ownership and participation, the largest of which is perhaps e-mail– corporate mailing lists are essentially identical to Usenet groups, and the behavior of forwarding is similar to that of cross-posting later in a conversation. Furthermore, as the lines between social networking sites blur, thanks to applications such as Friendfeed, it is likely that the resultant networks will demand a new assessment of relevance. The framework introduced can have a number of promising applications in such domains where assessing membership and participation in groups is necessary.

References

Adamic, L., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog.

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 207–214.

Adar, E. 2006. Guess: a language and interface for graph exploration. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, 791–800. New York, NY, USA: ACM.

Backstrom, L.; Kumar, R.; Marlow, C.; Novak, J.; and Tomkins, A. 2008. Preferential behavior in online groups. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, 117–128. New York, NY, USA: ACM.

Bollobas, B. 1998. *Modern Graph Theory*. Springer.

Fisher, D.; Smith, M.; and Welser, H. T. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*. Washington, DC, USA: IEEE Computer Society.

Gómez, V.; Kaltenbrunner, A.; and López, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 645–654. New York, NY, USA: ACM.

Kossinets, G.; Kleinberg, J.; and Watts, D. 2008. The structure of information pathways in a social communication network.

Leskovec, J., and Kleinberg, J. 2006. Patterns of influence in a recommendation network. In *In Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD*, 380–389. Springer-Verlag.

Leskovec, J.; Mcglohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. *SIAM International Conference on Data Mining (SDM)*.

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 177–187. New York, NY, USA: ACM Press.

Mishne, G., and Glance, N. 2006. Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*.

Nowell, D. L., and Kleinberg, J. 2008. Tracing the flow of information on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences* 105(12):4633–4638.

Turner, T. C.; Smith, M. A.; Fisher, D.; and Welser, H. T. 2005. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication* 10(4).

Viegas, F. B., and Smith, M. 2004. Newsgroup crowds and authorlines: visualizing the activity of individuals in conversational cyberspaces. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.*, 10 pp.+.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF), Grants No. IIS-0705359 and CNS-0721736. Mary McGlohon was partially supported by a Yahoo! Key Technical Challenges Grant and a travel grant from Microsoft Live Labs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or other funding parties. The authors would like to thank Michael Gamon, Kathy Gill, Christian Konig, Alexei Maykov, Marc Smith, and anonymous reviewers for their helpful discussions.