Exploring the Community Structure of Newsgroups

[Extended Abstract]

Christian Borgs Jennifer Chayes * Mohammad Mahdian † Amin Saberi ‡

ABSTRACT

We propose to use the community structure of Usenet for organizing and retrieving the information stored in newsgroups. In particular, we study the network formed by cross-posts, messages that are posted to two or more newsgroups simultaneously. We present what is, to our knowledge, by far the most detailed data that has been collected on Usenet cross-postings. We analyze this network to show that it is a small-world network with significant clustering. We also present a spectral algorithm which clusters newsgroups based on the cross-post matrix. The result of our clustering provides a topical classification of newsgroups. Our clustering gives many examples of significant relationships that would be missed by semantic clustering methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; G.2.2 [**Discrete Mathematics**]: Graph Theory

General Terms

Algorithms, Theory

Keywords

Spectral Method, Usenet, Clustering

1. INTRODUCTION

There has recently been a tremendous interest in the structure of self-organized networks, including the internet [5], the world wide web [10, 8], and various social networks [11]. These networks are very different from each other, but they all share the property that their structures are not engineered, but rather are the result of dynamic non-Markovian processes of individual decisions. The networks also share striking observed properties: a broad ("power-law" or "scale-free") distribution of connections, short paths between two given points ("small world phenomenon"), and the presence of many small dense subnetworks ("communities" or "clusters"). An understanding of this structure has enabled us to model and search these networks effectively, the greatest success having been in searches of the of the world wide web [10, 12], which has by now become our primary repository of information and misinformation.

In this paper, we consider another large network, intermediate between the internet and a social network: Usenet, the network of topic-oriented newsgroups on the internet, comprising tens of thousands of newsgroups and hundreds of millions of postings by millions of authors throughout the world.

Here we propose to explore and search the community structure of Usenet using what we call the cross-post graph, which is a graph containing information on instances when messages are posted to two or more newsgroups simultaneously. Past attempts to explore the structure of Usenet focused on semantic properties – principally the names of the newsgroups, but also sometimes the words in the subject headings of the messages. In this sense our work is analogous to the use of the hyperlink structure of the web, rather than the actual content of web pages, to explore and search the web [10, 12] – an approach that has been spectacularly successful. In both cases, the information defining the structure reflects individual decisions on relationships, rather than individual decisions on wording. In addition to being less dependent on the vagaries of language, such an approach scales much better than semantic approaches.

The basic workings of Usenet are as follows. Each of the over fifty thousand newsgroups has a unique name, with the names groups into trees. Some of the more common roots of these trees include `alt.`, `biz.`, and `rec.`, at least the first of which is probably familiar to many readers. Within a newsgroup, the messages are organized in threads. Each message is written by a single author; individuals may author more than one message along a thread. Each thread originates in a single message with a subject heading usually reflecting the content of the message; later messages in the thread, of which there can be thousands, are posted as responses to the original message. Any message along a thread can

*Microsoft Research, One Microsoft Way, Redmond, WA 98122. Email: {borgs, jchayes}@microsoft.com

†Laboratory for Computer Science, MIT, Cambridge, MA Email: mahdian@mit.edu

‡College of Computing, Georgia Institute of Technology, Atlanta, GA. Email: saberi@cc.gatech.edu. This work was done while the last two authors were visiting Microsoft Research.

KDD'04, August 22–25, 2004, Seattle, Washington, USA.

be cross-posted, by its author, to any number of additional newsgroups. It is this cross-posting on which our analysis will focus. The decision to cross-post the message to additional newsgroups is a reflection of the author's judgement that the message will, or at least, should be of interest to the readership of the additional newsgroups. Cross-posts are thus in some sense similar to hyperlinks on a webpage, which reflect a webpage author's judgement that additional webpages may be of interest to the readership of the original webpage.

The web certainly contains a tremendous amount of information, much of which is useful. However, without an understanding of the hyperlink structure of the web, and the development of search engines reflecting that hyperlink structure, the vast majority of this information would be inaccessible. Similarly, Usenet contains a great deal of information, again with some, but not all of it being useful. It is our hope that the development of methods to explore the structure of Usenet, and to search Usenet according to this structure, will enable us to access the useful information. We expect that this may also lead to a substantial increase in the size, and hopefully the seriousness, of Usenet. Indeed, once the web became efficiently searchable, many more individuals, businesses and institutions were encouraged to devote the necessary resources to write webpages. Given that it is much easier to post information on Usenet – information is posted in the form of simple messages – the effect of efficient searching algorithms should be felt all the more quickly.

We organize the information on cross-postings into a cross-post matrix or multigraph. Let $N = N(t)$ be the number of newsgroups on Usenet at time t. The cross-post matrix $A = A(t, \delta t)$ is a symmetric $N \times N$ matrix, with each row representing a different newsgroup, in some arbitrary but fixed order. The non-negative integer components A_{ij} of A represent the total number of cross-posts between all messages on newsgroup i and newsgroups j over the time interval δt before time t. We can similarly represent this information as a multigraph (i.e., a graph in which there may be multiple edges between vertices). In this representation, the vertices of the multigraph represent distinct newsgroups and the edge E_{ij} represent the total number of cross-postings between newsgroup i and newsgroup j.

Our contributions in this work are of several types. First, we will present what is, to our knowledge, by far the most detailed data that has been collected on Usenet cross-postings. Second, we analyze the cross-post matrix to show that Usenet is indeed a scale-invariant small-world network with significant clustering. We give specific measurements of parameters characterizing this structure. Third, we present a spectral algorithm which clusters newsgroups based on the cross-post matrix or graph. This clustering should provide a wealth of information to sociologists and others studying the social structure of Usenet. In particular, our clustering gives many examples of significant relationships that would be missed by semantic clustering methods. Finally, we propose a search engine to find newsgroups of relevance in specific contexts.

2. USENET

Usenet is a world-wide distributed discussion system. It consists of a set of over fifty thousand *newsgroups* covering a variety of topics. Each newsgroup has a hierarchical name like `alt.music.rock-n-roll` or `microsoft.public.word`. The names are grouped into trees with different roots such as `alt.`, `biz.`, and `rec.`. *Articles* or *messages* are *posted* to these newsgroups by users. These messages are distributed to other interconnected computer systems via a wide variety of networks.

Within a newsgroup, the messages are organized in *threads*. Each message is written by a single author; individuals may author more than one message along a thread. Each thread originates in a single message with a subject heading usually reflecting the content of the message; later messages in the thread, of which there can be thousands, are posted as responses to the original message. Any message along a thread can be cross-posted, by its author, to any number of additional newsgroups.

Over the time, Usenet has become a huge repository of information. However, its rapid growth and chaotic structure makes it a challenging task to organize this information and make it more accessible. Past attempts to explore the structure of Usenet has focused on semantic properties e.g. the names of the newsgroups, the words in the subject headings of the messages, etc. In this work, we are using the cross-post structure of Usenet for organizing and retrieving information stored in newsgroups. In that sense, our work is analogous to the use of the hyperlink structure of the web, rather than the semantic content of web pages, to explore and search the web [10, 12].

In particular, we will provide a topical classification of the newsgroups that can be used to help users to find the right newsgroup in which to post a message or to find the right discussion. The semantically-based name hierarchy is not suitable for this purpose for the following reasons:

1- In many situations, the name of a newsgroups is not descriptive of its content. It may be because the name is not chosen carefully in the beginning, or because the topic of discussion in that newsgroups has changed over time.

2- Two similar newsgroups may have different root names such as `alt.macromedia.flash` and `macromedia.flash.sitedesign`. While this difficulty could easily be overcome by algorithms which search for overlap of names, the name-based hierarchical trees used in current usenet archives, would put such newsgroups in different classes. More problematically, the names of two close newsgroups might not have any word in common for example `alt.disney.disneyworld` and `rec.parks.theme`.

3. CROSS-POST GRAPH

The structure of various social and technological networks such as the Internet or World Wide Web has been the subject of much recent research [17, 18, 3]. Despite numerous differences between the nature and the origin of these networks, many common characteristics has been observed. These common properties include the power-law distribution of the degree sequences [3], the small-world effect [18], and large clustering coefficients [18].

Here we study the cross-post graph, which is a graph containing information on instances when messages are posted to two or more newsgroups simultaneously. We will use these cross-posts as evidence of a close relationship between the content of the newsgroups to which they are posted.

We define the cross-post graph as a weighted graph with vertices representing the newsgroups and weights of edges representing the number of cross-posts between the corresponding endpoints. This graph exhibits many interesting

properties similar to those observed for other social and technological networks [3].

Figure 1 shows the distribution of the weighted degrees of the vertices of the cross-post graph (i.e., the number of cross-posts between a newsgroup and all other newsgroups) in linear and log-log scale. This degree sequence appears close to a *power-law* distribution. That is, the probability that a newsgroup has x cross-posts with other newsgroups is proportional to $x^{-\alpha}$; here $\alpha \approx 1.3$. A similar observation about the distribution of the number of authors that have posted to a newsgroup (see Figure 2) shows that it is close to a power-law distribution with $\alpha \approx 1.2$.

We also observe the small world effect in the cross-post graph. The graph consists of a giant connected component, containing more than 98% of the vertices, and a few hundred components with average size less than 5. The maximum and average distance between any two vertices in the giant component is 13 and 3.8, respectively. This can be compared to the average distance of 19 in the World-Wide Web graph [1].

Another interesting property of this graph is its high transitivity, also known as high clustering coefficient. The clustering coefficient of a graph is the probability that two random neighbors of a randomly chosen vertex are neighbors themselves. The clustering coefficient of the cross-post graph is 0.4492 although the density of edges are as low as 0.0016.

4. SPECTRAL CLUSTERING ALGORITHM

Spectral graph partitioning is a powerful tool based on techniques introduced by Fiedler [6, 7] in 1970's and popularized in 1990 by Pothen et al. [14]. It is used in many applications in computer science like assigning a set of tasks among processors so as to balance the load and minimize the communications [15], data mining in large data sets [2], and web page classification [8, 9].

We will denote the cross-post graph as $G = (V, E)$ where V is the set of vertices corresponding to newsgroups and E is the set of edges corresponding to cross-posts. Note that G is a multigraph i.e. there may be several edges between two vertices of G.

The goal of clustering is to partition the network into components such that each component is well-connected within itself, but the cut defined between two components is relatively sparse. For example, if we want to partition V into S and $\bar{S}$, the following ratio is a commonly used measure of the quality of the cut between S and $\bar{S}$:

$$\frac{\text{cut}(S, \bar{S})}{\min(W(S), W(\bar{S}))} \tag{1}$$

Here $\text{cut}(S, \bar{S})$ is the total number of edges between S and $\bar{S}$. $W(S)$ and $W(\bar{S})$ are the number of edges incident to vertices in S and $\bar{S}$, respectively. In a general graph, it is NP-hard to find the cut that minimizes the above ratio. Therefore, we will use a heuristic algorithm for finding a cut with a ratio close to the minimum. Our heuristic algorithm is based on spectral techniques which are the heart of many algorithms for finding sparse cuts in a graph [16].

Spectral analysis reduces to the analysis of eigenvectors of a normalized version of the adjacency matrix of the graph. Consider the matrix A with a_{ij} equal to the number of cross-posts between newsgroups i and j. Here we look at the Laplacian of the matrix A which is defined as $L = D - A$, where D is a diagonal matrix with $d_{ii} = \sum_j a_{ij}$. The Fiedler vector v of A is the eigenvector corresponding to the second smallest eigenvalue of L. Here we use a variant of the Fiedler vector, introduced by Chung [4], which is the solution to the generalized eigenvector equation $(D - A)v = \lambda Dv$. Equivalently, v is the second largest eigenvector of $D^{-1/2}AD^{-1/2}$ multiplied by $D^{-1/2}$.

Now, the idea of the heuristic algorithm is to choose a splitting value s and divide the vertices into two sets based on whether or not the value assigned to them by v is greater than s. Different heuristic algorithms are based on different choices of s; some of the popular ones are [16]:

- Bisection cut: Take s to be the median of the values assigned to vertices by v.
- Sign cut: Take $s = 0$.
- Gap cut: Take s to be a value in the largest gap in the sorted list of Fiedler vector values.
- Best cut: Take s to be the value which gives the best cut according to the cut objective function in equation (1).

We will use a variation of the last approach (best cut) in our algorithms. In order to partition the graph into more than two clusters, we can recursively use the same method until the size of each component is sufficiently small. This will give a hierarchical clustering which provides us with a classification of newsgroups at the desired level of granularity.

5. OUR RESULTS

By applying the algorithm described in the previous section, we obtained a hierarchical clustering of the newsgroups and hence a tree that allows us to study the Usenet at various levels of granularity. Although the effectiveness of a particular clustering algorithm is difficult to quantify and usually application dependent, it is clear from the output of our algorithm that it has successfully recognized many classes of newsgroups with close topics. For the convenience of the reader, we have put an output of our algorithm with 1056 clusters on the web at:
http://research.microsoft.com/~jchayes/Papers/usenet.html

One quantifiable measure of effectiveness is the percentage of cross-posts within clusters. In our clustering this percentage is 83.13%, while for a random clustering of the graph with the same distribution of cluster sizes this percentage is less than 1.53%. This comparison indicates that the cross-post graph is indeed strongly clusterable, and our algorithm has succeeded in finding a good clustering.

An examination of our results gives many examples of significant relationships that would be missed by name-based methods. For example `alt.disney.disneyworld` and `rec.parks.theme`.

Finally, it is worth noting that the clustering derived from the cross-post graph represents strong interaction among newsgroups in each cluster but does not necessarily indicate that the newsgroups are about the same topic. As an interesting example, the newsgroups `alt.microsoft.sucks` and `alt.linux.sucks` are grouped together in the same cluster. Also, usually newsgroups that share the same language, other than English, are grouped together.

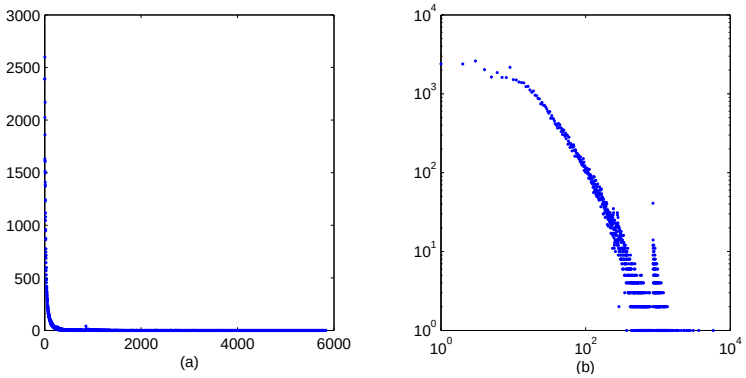

Figure 1: Distribution of the degree sequence of the cross-post graph in (a) linear (b) log-log scale

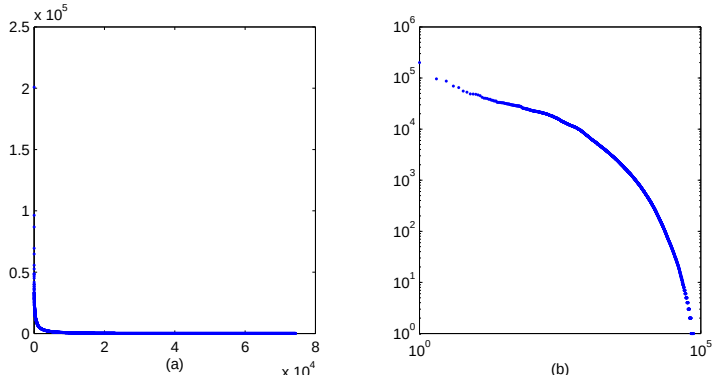

Figure 2: Distribution of the number of authors in (a) linear (b) log-log scale

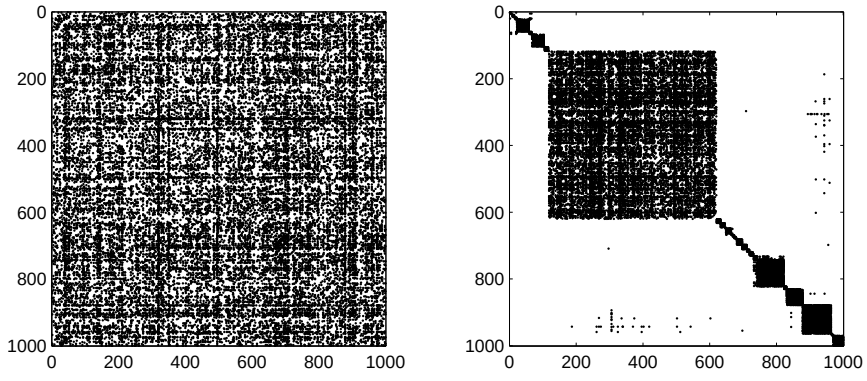

Figure 3: (a) The cross-post matrix restricted on the first one thousand newsgroups: a black point in position i and j indicates at least one cross-post between newsgroups i and j (b): The same matrix after reordering the newsgroup based on the result of our clustering. Note that the upper-right and lower-left of our matrix is now almost empty.

6. CONCLUSION AND OPEN QUESTIONS

In this paper, we proposed to use the community structure of newsgroups for the purposes of information retrieval. Similar methods in using the hyperlink structure of the web have been spectacularly successful.

In particular, we studied the network formed by cross-posts, messages that are posted to two or more newsgroups simultaneously. We analyzed this network to show that it is a small-world network with significant clustering. We also used a spectral algorithm which clusters newsgroups based on the cross-post matrix. The result of our clustering provides a topical classification of newsgroups. An instance of our clustering is available at *http://research.microsoft.com/~jchayes/Papers/usenet.html*.

The result of our algorithm can be used to help users find the right newsgroup to post their messages or find the right discussion. It can also be a source of many interesting sociological observations.

Our method can also be used for clustering authors, threads, or messages in a newsgroup or a cluster of newsgroups. Clustering authors can potentially lead to characterizing the expertise of active authors in each newsgroup. Clustering messages might be also helpful in distinguishing valuable answers from irrelevant discussions. In clustering messages, we can also use word frequencies in each message [2, 13].

7. ACKNOWLEDGEMENTS

We would like to thank the Collaborative and Multimedia Systems Group in Microsoft Research, especially Marc Smith, for posing some of the questions addressed here and providing us with the data, and for many helpful discussions.

8. REFERENCES

[1] R. Albert, H. Jeong, and A. Barabasi. Diameter of the world wide web. *Nature*, pages 130–131, 1999.

[2] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *ACM Symposium on Theory of Computing*, pages 619–626, 2001.

[3] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, pages 509–512, 1999.

[4] F.R.K. Chung. Spectral graph theory. In *Amer. Math. Society*, 1997.

[5] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[6] M. Fiedler. Eigenvectors of acyclic matices. *Czechoslovak Mathematical Journal*, 25(100):607–618, 1975.

[7] M. Fiedler. A property of eigenvectors of non-negative symmetric matrices and its application to graph theory. *Czech. Mathematical Journal*, 25(100):619–633, 1975.

[8] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring Web Communities from Link Topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pages 225–234, Pittsburgh, Pennsylvania, June 1998.

[9] R. Kannan and V. Vinay. The manjara meta-search engine.

[10] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[11] M. Newman, D. Watts, and S. Strogatz. Random graph models of social networks.

[12] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[13] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. pages 159–168, 1998.

[14] A. Pothen, H. D. Simon, and K. P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11:430–452, 1990.

[15] Horst D. Simon. Partitioning of unstructured problems for parallel processing. *Computing Systems in Engineering*, 2:135–148, 1991.

[16] Daniel A. Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *IEEE Symposium on Foundations of Computer Science*, pages 96–105, 1996.

[17] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[18] D. Watts and S. Strogatz. Collective dynamics of smallworld networks. *Nature*, 393, 1998.