f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

# Free Riding on Gnutella

by Eytan Adar and
Bernardo A. Huberman

*An extensive analysis of user traffic on Gnutella shows a significant amount of free riding in the system. By sampling messages on the Gnutella network over a 24-hour period, we established that almost 70% of Gnutella users share no files, and nearly 50% of all responses are returned by the top 1% of sharing hosts. Furthermore, we found out that free riding is distributed evenly between domains, so that no one group contributes significantly more than others, and that peers that volunteer to share files are not necessarily those who have desirable ones. We argue that free riding leads to degradation of the system performance and adds vulnerability to the system. If this trend continues copyright issues might become moot compared to the possible collapse of such systems.*

## Contents



## Introduction

The sudden appearance of new forms of network applications such as Gnutella [Gn00a] and FreeNet [Fr00], holds promise for the emergence of fully distributed information sharing systems. These systems, inspired by Napster [Na00], will allow users worldwide access and provision of information while enjoying a level of privacy not possible in the present client-server architecture of the Web.

While a lot of attention has been focused on the issue of free access to music and the violation of copyright laws through these systems, there remains an additional problem of securing enough cooperation in such large and anonymous systems so they become truly useful. Since users are not monitored as to who makes their files available to the rest of the network (produce) or downloads remote files (consume), nor are statistics maintained, the possibility exist that as the user community in such networks gets large, users will stop producing and only consume. This free riding behavior is the result of a social dilemma that all users of such systems confront, even though they may not be aware of its existence.

In a general social dilemma, a group of people attempts to utilize a common good in the absence of central authority. In the case of a system like Gnutella, one common good is the provision of a very large library of files, music and other documents to the user community. Another might be the shared bandwidth in the system. The dilemma for each individual is then to either contribute to the common good, or to shirk and free ride on the work of others.

Since files on Gnutella are treated like a public good and the users are not charged in proportion to their use, it appears rational for people to download music files without contributing by making their own files accessible to other users. Because every individual can reason this way and free ride on the efforts of others, the whole system's performance can degrade considerably, which makes everyone worse off - the tragedy of the *digital* commons [Ha68].

The second problem caused by free riding is to create vulnerabilities for a system in which there is risk to individuals. If only a few individuals contribute to the public good, these few peers effectively act as centralized servers. Users in such an environment thus become vulnerable to lawsuits, denial of service attacks, and potential loss of privacy. This is relevant in light of the fact that systems such as Gnutella, Napster, and FreeNet are depicted as a means for individuals to rally around certain community goals and to "hide" among others with the same goals. These may include providing a forum for free speech, changing copyright laws, and providing privacy to individuals.

Given these concerns we decided to conduct a set of experiments to determine the amount of free riding present in the Gnutella system. As we show below, a large proportion of the user population, upwards of 70%, enjoy the benefits of the system without contributing to its content.

In what follows we describe the basic architecture of Gnutella and the experiments that we performed. We then provide an analysis of the data and show ways in which such rampant free riding can impact distributed systems. Finally we propose some mechanisms that can counter free riding.

# Gnutella

People who wish to use the Gnutella network will download [Gn00a] or develop [Gn00b] an application that adheres to the Gnutella protocol. This application acts as either a *client* (a consumer of information) or a *server* (a supplier of information), as well as a high-level *network*, connecting and routing information between clients and servers. Each instance of an application is called a *peer*. We will use peer interchangeably with *host* in the following discussion.

Gnutella boasts a number of features that make it attractive to certain users. For example, Gnutella provides for anonymity by masking the identity of the peer that generated a query. Additionally, Gnutella provides the mechanism by which ad-hoc networks can be formed without central control.

Since there are no central servers in the Gnutella network, in order to join the system a user initially connects to one of several known hosts that are almost always available (although these generally do not provide shared files). These hosts then forward the IP and port address information to other Gnutella peers.

Once attached to the network, peers interact with each other by means of messages. Peers will create and initiate a broadcast of messages as well as re- *broadcasting* others (receiving and transmitting to neighbors). The messages allowed in the network are:

- *Ping Messages* - Essentially, an "are you there?" message directed at a host.
- *Pong Messages* - A reply to a ping ("yes, I'm here"). The pong message contains information about the peer such as their IP address and port as well as the number of files shared and the total size of those files. Peers forward this kind of message to their neighbors so that it is possible to later find other peers. This is needed in case there is a disconnect in the network.
- *Query Messages* - These are messages stating, "I am looking for x" and can get forwarded throughout the entire network (at least theoretically). Query messages are uniquely identified, but their source is unknown.
- *Query Response Messages* - These are replies to query messages, and they include the information necessary to download the file (IP, port, and other location information). Responses also contain a unique client ID associated with the replying peer. These messages are propagated backwards along the path that the query message originally took. Since these messages are not broadcast it becomes impossible to trace all query responses in the system.
- *Get/Push Messages* - Get messages are simply a request for a file returned by a query. The requesting peer connects to the serving peer directly and requests the file. Certain hosts, usually located behind a firewall, are unable to directly respond to requests for files. For this reason the Gnutella protocol includes push messages. Push messages request the serving client to initiate the connection to the requesting peer and upload the file. However, if both peers are located behind a firewall a connection between the two will be impossible.

Several features of Gnutella's protocol prevent messages from being re- broadcast indefinitely through the network. One such feature includes a short memory of messages that have been routed through a peer (thus preventing re- broadcasting). Additionally, messages are flagged with a time-to-live (TTL) field. At each hop (re-broadcast) the TTL is decremented. As soon as a peer sees a message with a TTL of zero, the message is dropped (i.e. it is not re- broadcast).

## Free riding in Gnutella

In our analysis we consider two types of free riding. In the first type, peers that free ride on Gnutella are those that only download files for themselves without ever providing files for download by others. The second definition of free riding considers not only the amount of downloadable content a producer has, but how much of that content is actually desirable content. This is essentially a quantity versus quality argument that also poses a social dilemma when there is a cost to the provider to make desirable files available to others. In the "old days" of the modem-based bulletin board services (BBS), users were required to

upload files to the bulletin board before they were able to download. In response to this requirement users would upload their own bad artwork or randomly generated text files and would be able to download high quality content generated by others. In the experiments described below we address both kinds of free riding.

# Experiments

In the following section we describe the experiments used to test the following three hypotheses:

- Hypothesis 1: A significant portion of Gnutella peers are free riders.
- Hypothesis 2: Free riders are distributed evenly across different domains (and by speed of their network connections).
- Hypothesis 3: Peers that provide files for download are not necessarily those from which files are downloaded.

## Measuring downloads

One of the features that attract users to Gnutella is the difficulty in associating queries to any particular peer/user. Given a query message it is virtually impossible (unless some large percentage of peers collude) to find the peer that originated the query. The unfortunate side effect of this property is to make it impossible to experimentally measure the number of queries and files downloaded by each client. This forces us to make assumptions about downloads in order to measure them.

One possible assumption is that users that share a high number of files had to have downloaded them, so those that share more also download more. In this case, there is no free riding. The other possible assumption is that users who have no files are those that will try to access them. Therefore the fewer files a user has the more likely he is to download them, resulting in rampant free riding.

Since we unfortunately have no way of knowing which of these two extremes is closest to reality, we assume that the truth is somewhere in between.

## Experimental Setup

In order to perform monitoring experiments on the Gnutella network it was necessary to modify a Gnutella client to log messages flowing through the system. We elected to use the Java based Furi client [Fu00] which was a full featured implementation, with numerous hooks for logging.

The Furi client was then executed for a 24-hour period over a weekend in August of 2000 (Saturday 1pm to Sunday 1pm) [1]. During this time period we collected both pong and query response messages from normal Gnutella users. A shorter trace during a weekday shows results consistent with the weekend findings. In the 24-hour period we observed 35,352 hosts issuing ping messages, which shared a total of 3,304,046 files.

One of the difficulties in measuring Network Address Translation (NAT) [Nat00] based peers is that it is possible that multiple machines will report the same address. In our study we witnessed 2,017 peers (or about 5% of the total) reporting a NAT address in ping messages. In analyzing query response which also utilize a unique client identifier (in addition to an IP address) we saw 937 out of 5,699 hosts (16% of the total) using NAT addresses.

While the possible range of 5% to 16% seems high, we find that the characteristics (in terms of files shared) of NAT based hosts is in line with non- NAT hosts and thus it is safe to remove them from the sample[2]. This leaves with a final count of 33,335 hosts sharing 3,100,464 files.

Although we could not capture all query response messages it was nonetheless possible to sample a wide selection by shifting locations (i.e., by reattaching to different hosts) within the Gnutella network. Over the 24-hour period, we were thus able to capture 87,668 query response messages.

## Results

Figure 1 illustrates the number of files shared by each of the 33,335 peers we counted in our measurement. The sites are rank ordered (i.e. sorted by the number of files they offer) from left to right. These results indicate that 22,084, or approximately 66%, of the peers share no files, and that 24,347 or 73% share ten or less files.

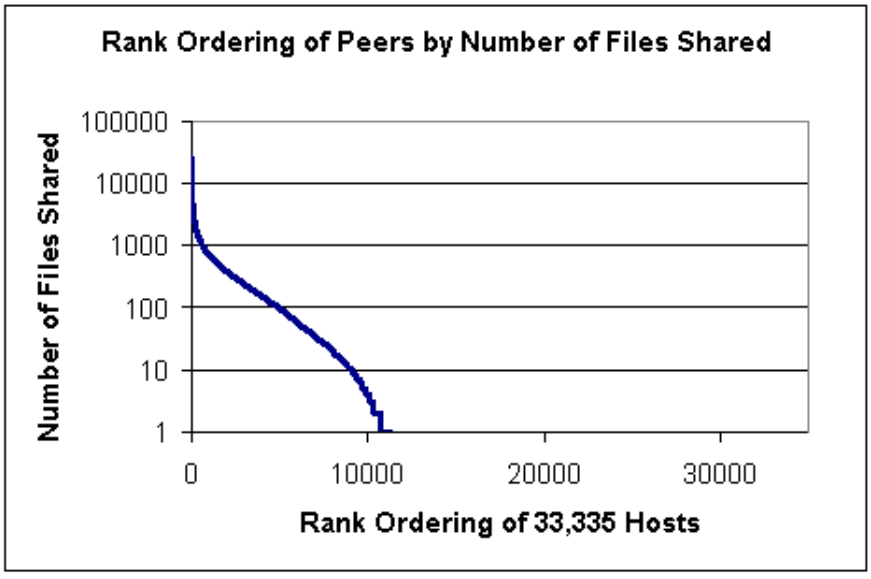


*Figure 1*

Although NAT allows firewalled hosts to share files, if both the sharing host and downloading host have NAT addresses the transaction cannot be completed. Thus, as the number of NAT based hosts on the network increases the number of completed transactions decreases. With 5% of hosts using NAT, this is a trivial .25%. However, as we approach

16% this turns into over 2% of transactions. While this in not "intentional" free riding, it is nonetheless important to consider. These probabilities push the zero share statistics up to 69%.

The data also shows that the top 1 percent (333 hosts) represent approximately 37 percent of the total files shared. This quickly escalates to the top 20 percent (6,667 hosts) sharing 98% of the files. Table 1 shows the values of the in-between data points.

| The top | Share | As percent of the whole |
|---|---|---|
| 333 hosts (1%) | 1,142,645 | 37% |
| 1,667 hosts (5%) | 2,182,087 | 70% |
| 3,334 hosts (10%) | 2,692,082 | 87% |
| 5,000 hosts (15%) | 2,928,905 | 94% |
| 6,667 hosts (20%) | 3,037,232 | 98% |
| 8,333 hosts (25%) | 3,082,572 | 99% |

*Table 1*

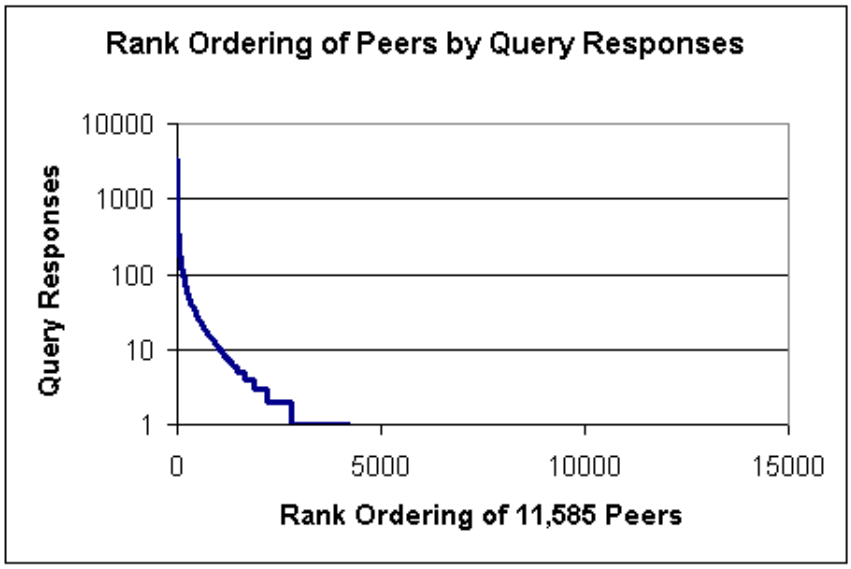


*Figure 2*

As per our second definition of free riding we determined which hosts provide files and which hosts provide files that are actually downloaded. We attempted to capture this by analyzing the query response traffic. The difficulty with analyzing this data is that it is unclear for how long each peer was actually connected to the network. However, we can assume again that due to the large sample, network connectivity averages out to some degree. As we show below, bandwidth appears not to have a significant effect on free riding. Using the lower bound estimate of NAT based hosts of 5% we find that after eliminating hosts that provide no downloadable files we were left with a set of 11,585 hosts.

Again, we measured a considerable amount of free riding on the Gnutella network. Out of the sample set, 7,349 peers, or approximately 63%, never provided a query response. These were hosts that in theory had files to share but never responded to queries (most likely because they didn't provide "desirable" files).

Figure 2 illustrates the data by depicting the rank ordering of these sites versus the number of query responses each host provided. We again see a rapid decline in the responses as a function of the rank, indicating that very few sites do the bulk of the work. Of the 11,585 sharing hosts the top 1 percent of sites provides nearly 47% of all answers, and the top 25 percent provide 98%.

# Who Shares Files?

In our second experiment we verified the hypothesis that files and query responses (and therefore free riders) are shared equally across different domains. The implication is that hosts based in domain *a* do not contribute more than hosts in domain *b* in terms of the ratio of peers on the network to files and responses offered. This does not imply that certain domains contribute more or less *total* hosts to the network, but simply that free riders are distributed equally. Additionally, domains can function as a proxy for bandwidth (for example aol.com hosts tend to operate on modems, and rr.com on cable modem connections). Therefore, if our hypothesis holds, the speed of a peer's internet connection will not influence the likelihood to free ride.

In order to do this analysis we filtered our initial test set to 26,014 peers. These were hosts with IP addresses that were readily converted to host names. We then counted the number of hosts in each *domain* (mit.edu, home.com, etc.) as well as the number of hosts in each *top-level domain*, or TLD (.edu, .com, .net, etc.).

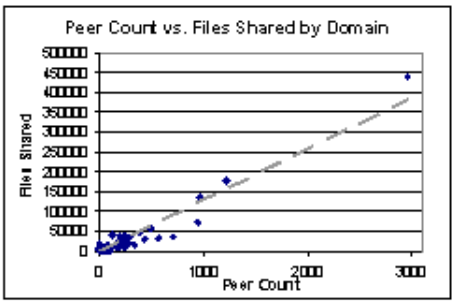


Figure 3a

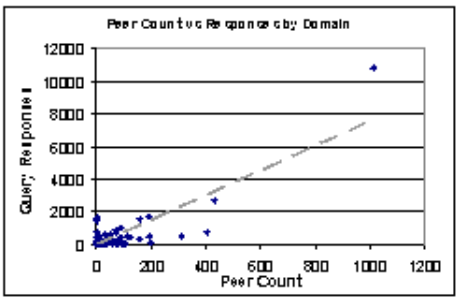


Figure 3b

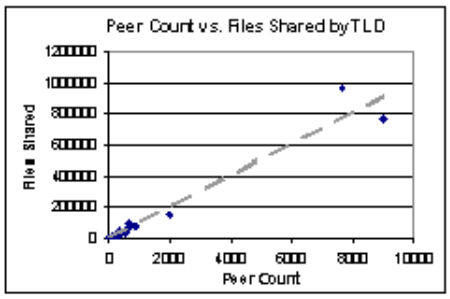


*Figure 4a*

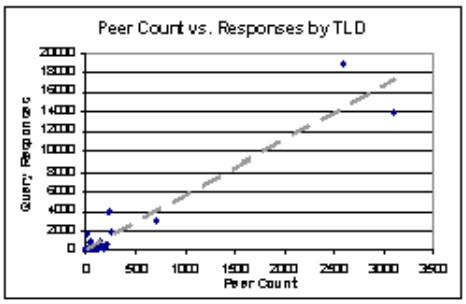


*Figure 4b*

In our set of hostnames there were 2,538 unique domains. The range of peers in each ranged from 1 to a maximum of 2,951. Figure 3a above illustrates this data. Each of the points in the figure represents a domain in terms of the number of peers (the x-axis) and the total number of files shared (the y-axis). The dashed line is the trend line for this data. A regression of the two dimensions yields an r-squared value of 0.927, indicating that peer count is linearly related to the number of files shared independent of the domain.

Figure 3b depicts the relationship between query responses and peer count. Again, a regression on this sample of 1,276 domains reveals a fairly linear relationship between the two dimensions (with an r-squared of 0.922). We consider this evidence of an even distribution of free riders [3].

Figures 4a and 4b display the equivalent data sets for TLDs (edu, net, org, etc.). Figure 4a represents the 77 top-level domains in terms of peer count to the number of files shared. Figure 4b represents 61 top-level domains in terms of peer count to query responses. Again, there appears to be a linear relationship in both figures with the regression fitting with an r-squared of 0.953 and 0.958 for figures 4a and 4b respectively.

## Quality vs. Quantity

In the final experiment we tested our hypothesis that the number of queries answered is not necessarily proportional to the number of files offered. This provides a test of the "quality" vs. quantity argument. The intuition is that the kinds of queries that are issued by the bulk of Gnutella users are very concentrated on particular topics. The files that are returned for these queries are therefore more desirable, which defines their quality. Therefore, only a small number of peers will actually share anything that is considered to be high "quality."

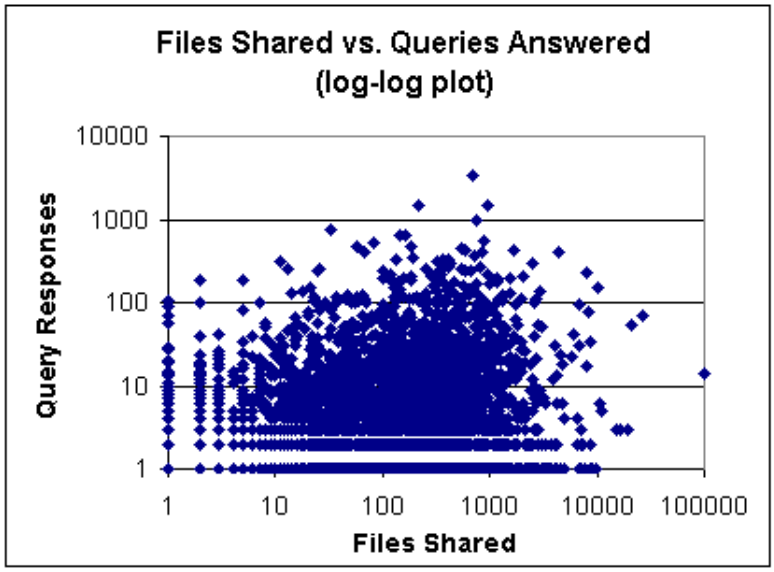


*Figure 5*

We found the degree to which queries are concentrated through a separate set of experiments in which we recorded a set of 202,509 Gnutella queries. The top 1 percent of those queries accounted for 37% of the total queries on the Gnutella network. The top 25 percent account for over 75% of the total queries. In reality these values are even higher due to the equivalence of queries ("britney spears" vs. "spears britney").

The predicted behavior is present to some extent. For example, the top responding host only hosted 695 files, but responded to 3,436 queries. The next most responsive peer hosted 956 files and responded to 1,474 queries.

Figure 5 illustrates the relationship between files hosts (the x-axis) and query responses (the y-axis) for 10,510 peers. As is apparent from the plot there is very little evidence of a relationship between quantity and quality in the Gnutella network. A regression analysis yields a very low r-squared value of 0.00105 for this data.

# Discussion

Studies of social dilemmas [Gl94] [Hu96] [Hu97] have shown that is hard to generate spontaneous cooperation in large anonymous groups. As we have shown in this paper, Gnutella is no exception to this finding, and an experimental study of its user patterns shows indeed that free riding is the norm rather than the exception.

If distributed systems such as Gnutella rely on voluntary cooperation, rampant free riding may eventually render them useless, as few individuals will contribute anything that is new and high quality. Thus, the current debate over copyright might become a non-issue when

compared to the possible collapse of such systems. This collapse can happen because of two factors, the tragedy of the digital commons, and increased system vulnerability, which we now discuss.

## The Tragedy of the Digital Commons

An ideal analysis of free riding would allow us to calculate the contribution provided by individuals in exchange for consumption (either in proportion or some fixed cost). There are two ways in which individuals on Gnutella can contribute. The first is simply by uploading files. The second is the active participation in the protocol of the network, thus providing the "glue" that holds the network together. It may be then that all peers on the network contribute even if they provide no downloadable files. However, there is a point at which peers that act only as glue provide diminishing returns to the system leading to at least two ways in which the quality of the service degrades.

First, peers that provide files are set to only handle some limited number of connections for file download. This limit can essentially be considered a bandwidth limitation of the hosts. Now imagine that there are only a few hosts that provide responses to most file requests (as was illustrated in the results section). As the connections to these peers is limited they will rapidly become saturated and remain so, thus preventing the bulk of the population from retrieving content from them.

A second way in which quality of service degrades is through the impact of additional hosts on the *search horizon*. The search horizon is the farthest set of hosts reachable by a search request. For example, with a time-to-live of five, search messages will reach at most peers that are five hops away. Any host that is six hops away is unreachable and therefore outside the horizon. As the number of peers in Gnutella increases more and more hosts are pushed outside the search horizon and files held by those hosts become beyond reach.

## Vulnerability

One argument that has appeared in the popular press regarding systems such as Gnutella [Or00] is that there is a diminished risk of the system being shut down by either lawsuit or attack. It will be impossible, users argue, for the Recording Industry Association of America (RIAA) to sue all of them. This belief, which was spread by the press, allowed users to believe that they were safe among others. Unfortunately, in light of the evidence provided above, Gnutella provides a false sense of security.

As we have seen in the experiments, there is a small collection of peers that provide the bulk of the shared files and answered queries. These few providers act as a rather centralized server consisting of several peers and thus the RIAA need not sue all users or even the bulk of users. They simply need to target the top-serving peers (of which there are very few that serve very many).

## Overcoming free riding

There are many ways of patching Gnutella so that it can accommodate the same privacy rules but scale more effectively.[5] It is interesting therefore to establish how different

file-sharing applications rely on technological features to induce users to share. FreeNet, for example, forces caching of downloaded files in various hosts. This allows for replication of data in the network forcing those who are on the network to provide shared files. Unfortunately, such a system is prone to replication of "bad" or illegal data and "tainting" hosts.[6] The second cost of the automatic replication as implemented in FreeNet is the unique identifiers for files that forces users to know exactly what they are looking for.

Napster, by default, downloads all files into a shared upload directory. In this way when a user downloads a file it is automatically shared. In some ways this feature addresses the FreeNet problem because users will only keep "good" files on their computers. However, users can easily circumvent this shared upload/download directory and frequently do. We have also witnessed Napster users misrepresenting the speed of their network connections (saying they are on a modem when they are on a high speed connection) in order to discourage other users from connecting to them. Both system provide their own set of solutions to the free riding but at the cost of introducing other problems to their systems.

Another possible solution to this problem is the transformation of what is effectively a public good into a private one. This can be accomplished by setting up a market-based architecture that allows peers to buy and sell computer processing resources, very much in the spirit in which Spawn was created [Wa92]. In this context we should stress that the utility to users does not necessarily have to be monetary. For instance, issues of prestige or status drive participation in open source systems like Linux [Lo00] and the same can be said of SETI@Home[Se00], where obviously to be the owner the PC that detects the first intelligent signal from outer space would constitute great utility.

Another alternative for eliminating free riding is to reduce the cost. For example the Usenet system, while allowing some degree of anonymity, provided a great advantage to individual users as their messages were distributed by an infrastructure that offloaded the bandwidth requirements for individuals. That is, the only cost to the user was the initial posting; afterwards the message was propagated by the system.

# Conclusions

In this paper we analyzed user traffic in Gnutella and concluded that there is a significant amount of free riding in the system. Specifically, we found that nearly 70% of Gnutella users share no files, and nearly 50% of all responses are returned by the top 1% of sharing hosts. Furthermore, we found that free riding is distributed evenly between domains, so that no one group contributes significantly more than others, and that peers that volunteer to share files are not necessarily those who have desirable ones.

These findings have serious implications for the future development of Gnutella and its many variants. In order for distributed systems with no central monitoring to succeed, a large amount of voluntary cooperation is required, a requirement that is very hard to fulfill in systems with large user populations that remain anonymous.

Sometimes, the logic behind the decision to cooperate or not changes when the interaction is

ongoing, since future expected utility gains will join present ones in influencing the rational individual's decision. In particular, individual expectations concerning the future evolution of the social dilemma can play a significant role in each member's decisions[Hu96]. An interesting continuation of these experiments may lead to an understanding of how free riding changes over time.

# About the Authors

Eytan Adar is a member of the Internet Ecologies Area at the Xerox Palo Alto Research Center in Palo Alto, Calif. Most recently he has been involved in Internet characterization research and the design of new systems fusing economic and computer science ideas. He holds a BS and MEng degrees from the Massachusetts Institute of Technology. E-mail: adar@parc.xerox.com

Bernardo Huberman is a Research Fellow at the Xerox Palo Alto Research Center, where he heads the Internet Ecologies Area, a group involved in studying the dynamics of distributed processes in social organizations and the Internet. His recent research has concentrated in the World Wide Web, with particular emphasis on the dynamics of its growth and use. This work helps uncover the nature of electronic markets and the law of surfing. He has been involved in the design of novel mechanisms for enforcing privacy and trust in e-commerce and negotiations.
E-mail: huberman@parc.xerox.com

# Acknowledgments

The authors would like to thank Rajan Lukose, Lada Adamic, Ed Chi, and Pam Schraedley for valuable discussions. We also thank Sara Dubowsky for her late night editorial help.

# References

[Ch85] D. Chaum, 1985. "Security without identification: Transaction systems to make big brother obsolete," *Communications of the ACM,* volume 28, number 10, pp. 1030-1044.

[Fr00] The FreeNet home page, http://freenet.sourceforge.net/

[Fu00] The Furi home page, http://www.jps.net/williamw/furi/

[Gl94] N.S. Glance and B.A. Huberman, 1994. "Dynamics of Social Dilemmas," *Scientific American,* (March).

[Gn00a] The Gnutella home page, http://gnutella.wego.com/

[Gn00b] The Gnutella Developer home page, http://gnutelladev.wego.com/

[Ha68] G. Hardin, 1968. "The Tragedy of the Commons," *Science,* volume 162, pp. 1243-1248, also at http://dieoff.com/page95.htm

[Hu96] B.A. Huberman and N. S. Glance 1996. "Beliefs and Cooperation," In: P. Danielson (editor). *Modeling Rational and Moral Agents.* Oxford: Oxford University Press, pp. 210-235.

[Hu97] B.A. Huberman and R. Lukose, 1997. "Social Dilemmas and Internet Congestion," *Science,* volume 277, p. 535.

[Lo00] C.H. Loch, B.A. Huberman, S. and Stout, in press. "Status Competition and Performance in Work Groups", *Journal of Economic Behavior and Organizations.*

[Na00] The Napster home page, http://www.napster.com/

[Nat00] "The Network Address Translation White Paper," @Home Networks, at http://work.home.net/whitepapers/natwpaper.html

[Or00] A. Oram, 2000. "Gnutella and Freenet Represent True Technological Innovation," The O'Reilly Network, (12 May), at http://www.oreillynet.com/lpt/a/208

[Se00] SETI@Home: Search for Extraterrestrial Intelligence at Home, http://setiathome.ssl.berkeley.edu/

[St00] Stop Napster home page, http://www.stopnapster.com/

[Wa92] C.A. Waldspurger, T. Hogg, B.A. Huberman, J.O. Kephart, and S. Stornetta, 1992. "Spawn: A Distributed Computational Economy," *IEEE Transactions of Software Engineering,* volume 18, number 2 (February), pp. 103-117.

# Notes

1. A much smaller experiment during a weekday revealed that in a sample of over 300 hosts 72% of share no files, a result consistent with our extended study.

2. NAT hosts shared no files 68.7% of the time, and ten or less files 74.5% of the time. The top 1% of NAT hosts shared 37.8% of the total files, and the top 25% shared 99.4% of the total files.

3. Of tangential interest may be the top number of hosts sharing files. The top 5 domains are (from most to least) home.com, rr.com, aol.com, t-dialin.net, and mediaone.net. The top hosts in query responses are home.com, rr.com, mediaone.net, ks.us, and pacbell.net.

4. The top five domains for queries in the first-level domain in terms of files shared are: net, de, nl, edu, and ca. For queries answered they are: com, net, edu, de, and nl.

5. Hint: Mix one part mailing list, one part anonymous bulletin board (see for example [Ch85]), and one part anonymous re-mailer (add more re-mailers depending on taste for paranoia).

6. If a user requests a bad file (say a bomb or Trojan [St00]), this file is replicated between

all computers from the host uploading to the host downloading.

---

## Editorial history

Paper received 8 August 2000; revision received 19 September 2000; accepted 27 September 2000.

---

Contents Index



Free Riding on Gnutella by Eytan Adar and Bernardo A. Huberman
First Monday, volume 5, number 10 (October 2000),
URL: http://firstmonday.org/issues/issue5_10/adar/index.html