

On the Penetration of Business Networks by P2P File Sharing (DRAFT)

Kevin Lee
School of Computer Science,
University of Manchester,
Manchester, UK.
+44 (0) 161 2756132
klee@cs.man.ac.uk

Danny Hughes
Computing, InfoLab21,
Lancaster University,
Lancaster, UK.
+44 (0) 1524 510351
danny@comp.lancs.ac.uk

James Walkerdine
Computing, InfoLab21,
Lancaster University,
Lancaster, UK.
+44 (0) 1524 510352
walkerdi@comp.lancs.ac.uk

Abstract

P2P file-sharing poses a number of problems for system administrators including unpredictable network usage, increased vulnerability to security threats and the danger of legal action. Because of these problems, the majority of businesses attempt to restrict the use of P2P software, but how successful have these measures been? This paper analyzes the degree to which business networks have been penetrated by P2P applications and shows that despite significant security risks, business networks participate significantly in P2P file-sharing systems. Furthermore, we find that participation is greatest amongst small businesses and that viral media has a significant effect on the participation of large businesses.

1. Introduction

Today, peer-to-peer (P2P) file sharing systems have millions of users and generate the majority of Internet traffic [1]. Unfortunately, the use of P2P file sharing systems causes a number of serious problems for network administrators:

- **Unpredictable Network Usage:** Files available on P2P networks tend to be significantly larger than those available on the web and file popularity is highly concentrated and dynamic [2]. The confluence of these factors is that P2P applications generate large and unpredictable network loads that have the potential to swamp network links and interfere with critical Internet services.
- **Security Threats:** Viruses and Trojan-horse programs are widespread on P2P file sharing systems. Furthermore, P2P users have been observed accidentally sharing private files, including sensitive corporate information [3].

- **Danger of Legal Action:** P2P file sharing systems are used to support a variety of illegal activities. The majority of files available on these systems are copyrighted and in many countries access providers may be held responsible for the copyright infringement of their users. For example, following the 2005 ruling of the United States Supreme Court [4], access providers may be held responsible for copyright infringement offenses committed by their users.

Because of the significant problems caused by P2P file sharing systems, most businesses restrict their use. However, identifying and blocking P2P traffic is difficult due to the large number of P2P protocols and their rapidly evolving nature [5]. This is likely to become ever more problematic as recent research has shown that users are migrating towards P2P file sharing systems that are more anonymous and decentralized [6].

This paper analyzes the penetration of business networks by P2P file-sharing applications using a two week trace of the Gnutella network [7]. Gnutella is a popular decentralized file-sharing protocol with a large and well-studied user base. While the participation of business networks in Gnutella cannot be considered to be truly representative of business participation in other P2P systems, our results may be viewed as generally indicative of levels of participation. Furthermore, as Gnutella is a long-established and well understood protocol, it is likely that system administrators are more successful at blocking traffic from Gnutella than from newer, more anonymous protocols [5], on which participation may be significantly higher than we observed. Our experiments show that, despite the severe security risks associated with P2P file-sharing, business networks generate a significant volume of P2P traffic.

Our experiments also indicate a strong link between the number of employees in a business and the likelihood of participation in P2P file-sharing networks. We find firstly

that small businesses are more likely to participate in P2P file-sharing networks and secondly, that the participation of larger businesses is highly dependent upon the effects of ‘viral media’.

The remainder of this paper is structured as follows: Section 2 describes the Gnutella network. Section 3 describes our experimental methodology. Section 4 presents the results of our experiments. Finally, section 5 summarizes our findings and discusses directions for future work.

2. The Gnutella Protocol

Gnutella is an open protocol which supports peer-to-peer file-sharing. The protocol builds a hierarchical decentralized overlay network [8] in which each host is required to forward both resource discovery and network maintenance messages. The protocol uses just five message types as follows:

- **PING** is used in peer discovery. A peer receiving a Ping responds with a Pong message.
- **PONG** is a response to a Ping. It contains responding peer’s address and the amount of data it serves.
- **QUERY** is a ‘search’ message. If a peer receiving a Query has matching data, it generates a QueryHit.
- **QUERYHIT** is a response to a Query. It contains information required to acquire the requested data.
- **PUSH** is a mechanism to support downloads from fire walled peers.

In addition, Gnutella is structured into three phases: *connecting to the network*; *discovering resources*; and *transferring resources*.

2.1 Connecting to the Gnutella Network

Acquisition of an initial host address, used to bootstrap network entry occurs outside of the Gnutella protocol; typically via a ‘GWebCache’ [7]. A newly-arriving peer connects to a peer discovered in this way by initiating TCP connections to that host. Incoming peers may connect to the network as ‘*leaf nodes*’ which do not accept incoming connections, or ‘*ultra peers*’, which accept connections and are therefore responsible for routing a greater proportion of network traffic. Leaf nodes upload a list of the files they are sharing at connection time, allowing ultra-peers to proxy for leaf nodes as described in section 2.2.

Following initial connection, further peers are discovered by broadcasting a PING message across the network. All peers that receive a PING message should respond by sending a PONG message, which is forwarded back along the path of the incoming PING to the originating

peer. PONG messages contain the network address and port on which the sending peer is listening for incoming Gnutella connections and information regarding the amount of data and the number files this peer is making available to the network.

2.2 Discovering Resources

In order to support file discovery, ultra-peers listen for incoming QUERY messages, and participate in their broadcast across the network by flooding them to each of their ultra-peer neighbors. Where a QUERY matches a file available on an ultra-peers leaf node, the QUERY message will be forwarded to that node. In this way, the bandwidth of leaf-nodes is conserved.

If any peer is able to satisfy a QUERY, it should respond by sending a QUERYHIT message back along the path of the incoming QUERY. QUERYHIT messages contain the network address and port on which the responding peer is listening for HTTP file-transfer connections. QUERYHIT messages also include the speed of the peer’s Internet connection, and a set of ‘hits’ (matching file-names) which satisfy this QUERY.

2.3 Transferring Resources

File transfer itself occurs outside of the Gnutella protocol. When a searching peer receives a QUERYHIT message, it can attempt to initiate a direct download from the target peer (whose port and IP address were specified in the QUERYHIT message) via HTTP. However, if the target peer is behind a firewall, the requesting peer can instead send a PUSH message to the target, containing details of the file requested and the network address and port to which the file should be *pushed*. On receiving a PUSH, the target peer establishes the HTTP connection and sends the file to the requesting peer.

3. Experimental Methodology

In order to explore the penetration of business networks by P2P file sharing applications, we analyzed a two-week trace of traffic on the Gnutella network [7] gathered between the 30th of March and the 12th of April.

Gnutella is a popular and open P2P file sharing protocol, with a large and well-studied user base. From this trace data, we isolated PONG *connection* messages and QUERYHIT *search-response* messages. We then resolved the IP addresses in these messages and isolated those messages which originated from businesses. Based upon this data, we were able to analyze the scale and characteristics of business participation in Gnutella. Section 3.1 describes our tracing methodology and section 3.2 describes our data archiving and querying approach.

3.1 Tracing Methodology

We used a *passive application-level tracing methodology* to gather our trace data as described in [1]. Passive application-level traces are performed by monitoring P2P messages passed at the application level. As all Gnutella peers participate in routing network maintenance and resource discovery messages, passive application-level monitoring can be performed simply and transparently by modifying a Gnutella peer to log the resource discovery and network maintenance messages that it is required to route.

Our experimental work was based on intercepting and analyzing PONG and QUERYHIT messages on the Gnutella network. Both of these message types contain the IP address of originating peers and based upon this, it is possible to isolate those messages originating from businesses. In order to support our experiments we modified the Jtella [9] base classes to implement a specialized monitoring peer.

Each PONG message originating from a business network shows that a peer within that network has connected to the Gnutella network as an ultra-peer and is allowing incoming connections. Furthermore, PONG messages contain the number of files that a peer is sharing, thus by comparing PONG messages from business networks to general PONG messages, it is possible to assess the relative level of user participation through business networks. Each QUERYHIT message originating from a business network essentially shows us that a Gnutella peer within that network is advertising a file in response to a search. This can be used to assess extent to which business networks are used to distribute files on Gnutella.

Using our specialized monitoring client, we monitored Gnutella traffic over a two week period between March 30th and April 12th. We maximized the size and typicality of our sample base by connecting to the network as an ultra-peer [7], maintaining a large number of incoming and outgoing connections, and by periodically re-connecting to different areas of the network. All messages during this period were parsed and logged to an SQL database as described in section 3.2.

3.2 Data Archiving and Querying

In order to ensure the flexibility of our trace data, all data that our monitoring peer routed during the test period was logged to an SQL database. This database contains a separate table for each message type (as enumerated in section 2) and each piece of information stored in these messages maps to a separate field within these tables. Each table also contains three additional fields: 'time', 'country' and 'owner'. The *time* field provides a per-message time stamp, while the *country* and *owner* fields are derived from

each messages IP address through a separate resolution process, which operates as follows:

For each message which contains an IP address, the database is firstly scanned. Where an IP address has been observed previously, the 'owner' field will be set from existing records. Where an IP address has not been observed previously, a recursive WHOIS [10] lookup is performed to determine its owner. On average, this process was capable of identifying the owner of an IP address with a 91% success rate, with 9% being irresolvable, primarily due to the use of Network Address Translation (NAT) by participating peers.

While this methodology may only be used to identify businesses which administer their own IP range, it is exactly this group of businesses where system administrators have most control over Internet access and thus where the question of P2P penetration is most critical.

4. Penetration of Business Networks by P2P

The experiments performed in this paper firstly consider the scale of participation in P2P file-sharing by users on business networks. Secondly, we explore the typical characteristics of this participation. Thirdly we illustrate the role of company size in participation. Finally, we discuss the significant effects of *viral media* on the penetration of business networks by P2P file sharing.

Section 4.1 analyses the participation of business networks in Gnutella. Section 4.2 discusses the characteristics of P2P file-sharing traffic originating from business networks. Finally, Section 4.3 discusses the effect of company size upon this participation.

4.1 P2P use through Businesses Networks

In order to explore the scale of business networks participation in P2P file sharing systems, we analyzed PONG connection messages generated throughout our trace. The *owner* field associated with each unique IP address was manually categorized as belonging to either a business or non-business network and the number of both categories of PONG message was extracted for each day. We logged a total of 712,659 PONG messages during our trace, an average of over 50,000 messages per day. Table 1 shows the number of PONG messages originating from business networks. Figure 1 shows this as a proportion of the total number of PONG messages observed.

Table 1 – PONG Messages from Business Networks

Date	From Business	Total Messages	Proportion from Business
30-03	831	53776	1.5%
31-03	549	26032	2.1%
01-04	4784	80615	5.9%
02-04	1348	24142	5.6%
03-04	1533	73320	2.1%
04-04	2119	46662	4.5%
05-04	3379	49670	6.9%
06-04	2316	60738	3.2%
07-04	1712	57879	3%
08-04	1121	49601	2.3%
09-04	1135	33359	3.4%
10-04	828	38346	2.2%
11-04	1886	45914	4.1%
12-04	3500	72605	4.8%

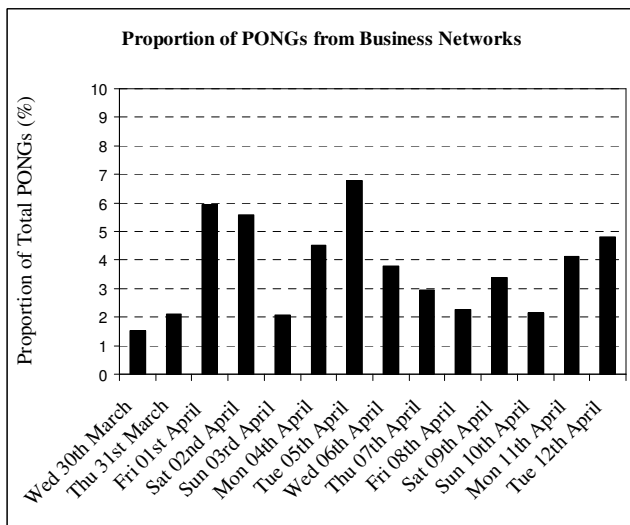


Figure 1 – Proportion of PONG messages from Business Networks

As can be seen from figure 1, business networks are responsible for a small, but significant proportion of PONG traffic on the Gnutella network. The proportion of PONG messages originating from business networks varied from a low of 1.5% on Wednesday the 30th of March to a high of 6.9% on Tuesday the 5th of April. The average proportion of PONG messages generated by businesses was 3.71%. As might be expected, PONG traffic from businesses was subject to significant day-of-week effects, with Sundays demonstrating a significantly lower than average volume of PONG messages: 2.09% on Sunday the 3rd of April and 2.15% on Sunday the 10th of April.

While the proportion of PONG messages generated by business networks illustrates that they connect to Gnutella

in significant numbers, this does not necessarily mean they will participate by actively sharing files. To analyze the extent to which business networks distribute files we also analyzed the level of *free-riding* on business networks.

4.1.1 Free-riding on Business Networks

Anonymous peer-to-peer file-sharing networks such as Gnutella embody a social dilemma often referred to as the tragedy of the digital commons [11]. The dilemma for each individual is whether to contribute to the common good by sharing files, or to maximize their personal experience by ‘free-riding’ (i.e. downloading files while not contributing any to the network).

We measured the rate of free riding during our trace by analyzing the number of PONG messages which report no shared files. We found a significant level of free riding on both business and non-business networks. This is shown in figure 2.

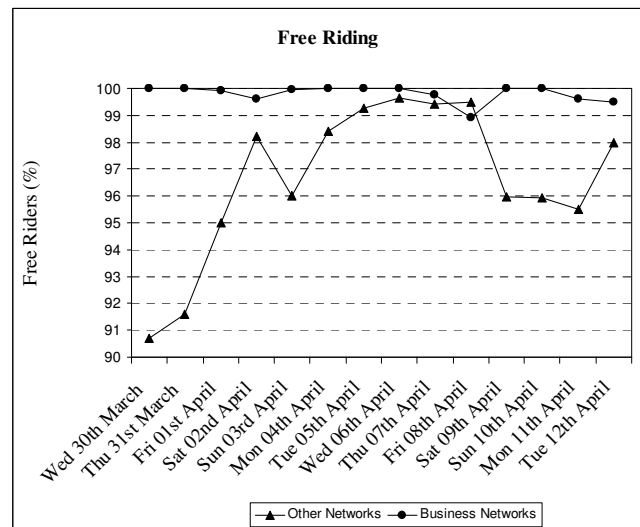


Figure 2 – Free Riding on Business Networks

Our trace revealed an average free-riding level of 96.7% and a level of 99.8% for business networks. This is in-line with existing studies, which have shown that the level of free-riding on Gnutella is significant and increasing, from 66% in 2000 [11] to 85% [12] in 2004. Our results show that since 2004, free-riding has continued to increase.

Furthermore, as Figure 2 shows, we find that PONG messages originating from business networks consistently demonstrate a higher level of free riding than non-business PONG messages. This is most likely due to prohibitions against P2P file sharing in the work-place.

Our trace data contains one anomalous result - the rate of free-riding on business networks dropped below that on

non-business networks on Thursday the 7th of April. This was due to the effect of viral media, which is discussed in more detail in section 4.3.1.

4.2 Characteristics of Business Network Traffic

In order to further explore the active participation of business networks in Gnutella, we analyzed QUERYHIT search-response messages generated throughout our trace. We logged a total of 554,693 QUERYHIT messages during our trace, an average of over 39,000 messages per day. QUERYHIT messages originating from business networks were recorded on each day of our trace. Figure 3 shows this as a proportion of the total number of QUERYHIT messages observed. Due to space constraints, the raw data is omitted.

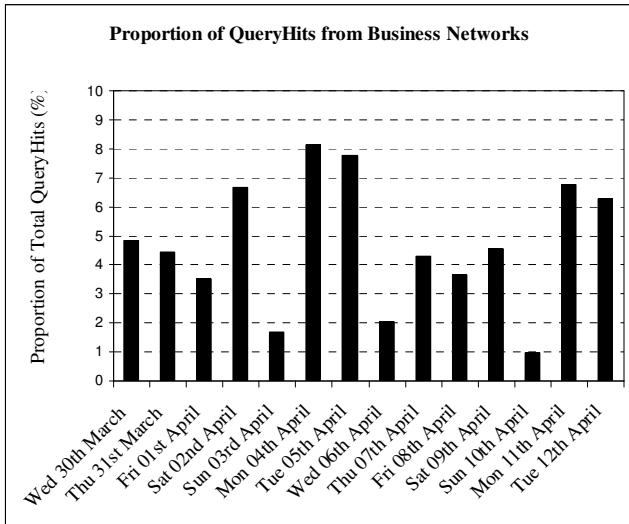


Figure 3 – Proportion of QUERYHIT messages from Businesses

As Figure 3 shows, business networks account for a significant volume of QUERYHIT traffic – an average of 4.69% per day. As was the case with PONG connection messages, QUERYHIT traffic from businesses networks is subject to significant day-of-week variations, with Sundays demonstrating a significantly lower than average volume of QUERYHIT messages: 1.68% on Sunday the 3rd of April and 0.96% on Sunday the 10th of April.

We also investigated the distribution of QUERYHITs across unique business networks. Figure 4 shows a rank ordering of the 72 unique businesses that we observed generating QUERYHIT messages, sorted by the number of QUERYHITs that they produced.

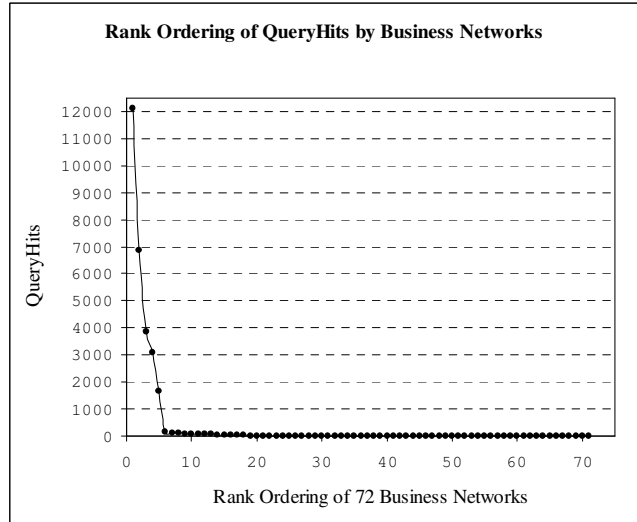


Figure 4 – Rank ordering of QueryHits from Business Networks

As can be seen from Figure 4, the extent to which business networks have been penetrated by P2P file sharing applications is highly asymmetric. The majority of businesses distribute very few files. Conversely, the majority of files are distributed by a very small number of highly active businesses.

4.3 The Effect of Company Size

We analyzed the effect of business size upon participation in the Gnutella network. Businesses were first broadly categorized according to their number of employees. The number of businesses observed from each size category was then recorded for each day in our trace. This is shown in Figure 5 below.

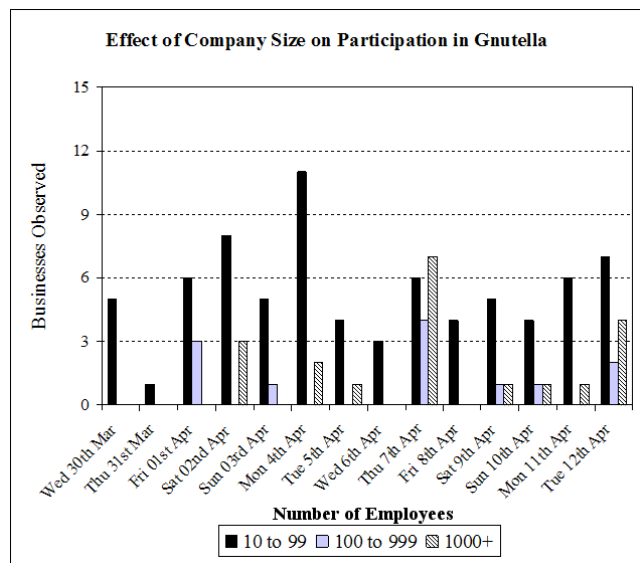


Figure 5 – The Effect of Company Size on Participation

As can be seen in Figure 5, small businesses participate most heavily in Gnutella, while the participation of large businesses is consistently and significantly lower, except in the case of Thursday the 7th of April, on which date, business networks also demonstrated an anomalously low level of free-riding (see figure 2).

To investigate these phenomena, we performed further analysis on QUERYHIT messages produced on this date and found that the sudden and dramatic increase in the participation of large companies (and likely the reduced level of free-riding) was due to the effect of ‘viral media’.

4.3.1 The Effect of Viral Media

We performed additional analysis of QUERYHIT messages produced by large businesses on Thursday the 7th of April and found that, remarkably, all of the large businesses which appeared in our trace for the first time on that date were distributing QUERYHITs for only one, identical file – a viral video. The term *viral media* refers to media, particularly video files which suddenly become very popular and are distributed between Internet users in an epidemic fashion. The viral media that we identified was a short comedy video, (unfortunately, its title is unsuitable for reproduction here) which research revealed had been a popular viral video during our trace.

The effects of this viral video were startling: The number of large business networks distributing QUERYHIT messages jumped five-fold following its release, as can be seen in figure 5. This was similarly reflected by a notable drop in the level of free-riding, as people shared this viral video (this can be seen in figure 2). In keeping with the nature of viral media, the popularity of this file was short-lived and all of the business networks that we observed for the first time distributing this file on April 7th did not re-appear during the remainder of our trace.

5. Summary and Future Work

This paper has discussed the security implications of participating in P2P file-sharing and has analyzed the scale and characteristics of P2P penetration of business networks using a two-week trace of the Gnutella network.

Our experiments revealed that business networks are responsible for a small but significant volume of Gnutella traffic – 3.71% of PONG messages and 4.69% of QUERYHIT messages. As one might expect, we found a higher rate of free-riding for peers on business networks and that traffic from such networks was subject to significant day-of-week variations. Furthermore, we find that small businesses are far more likely to host P2P applications and that the release of viral media can

significantly impact the level of penetration of business networks.

Future work will focus upon three key areas. Firstly we will explore how the real-world characteristics of businesses affect their penetration by P2P applications. In particular we wish to investigate impact of the domain of a business and its geographical location. Secondly, we wish to further explore the impact of the release of viral media. Finally, we intend to revisit our experiments using a newly gathered trace of a broader range of P2P networks in order to validate the results reported here.

6. References

- [1] “Monitoring Challenges and Approaches for P2P File Sharing Systems”, Hughes D, Walkerdine J., Lee K., published in the proceedings of the 1st International Conference on Internet Surveillance and Protection (ICISP’06), August 2006.
- [2] “Measuring and Analyzing the Characteristics of Napster and Gnutella Hosts” Saroiu S., Gummadi K., Gribble S. D., published in Multimedia Systems 9, pp 170-184, 2003.
- [3] “P2P File Sharing”, OnGuard, available online at: <http://onguardonline.gov/p2p.html>
- [4] “Metro-Goldwyn-Mayer Studios inc. et al. v. Grockster et al.”, U.S. 9th Circuit Supreme Court Decision, June 25. http://www.eff.org/IP/P2P/MGM_v_Grokster/04-480.pdf
- [5] “Accurate, Scalable Network-level Identification of P2P Traffic Using Application Signatures” Subhabrata S., Spatscheck O., Wang D., published in the proceedings of the thirteenth international world wide web conference (WWW2004), New York, USA, 2004.
- [6] “Is P2P Dying or Just Hiding?”, Karagiannis, T., Broido, A., Brownlee, N., Faloutsos, M., In the Proceedings of Globecom 2004, Dallas, U.S., December 2004.
- [7] “The Gnutella Protocol Specification v0.6”: http://rfcgnutella.sourceforge.net/src/rfc-0_6-draft.html.
- [8] “Dependability Properties of P2P Architectures,”, Walkerdine J., Melville L. and Sommerville I., in the proceedings 2nd IEEE International Conference on Peer-to-Peer Computing (P2P ‘02), September 2002.
- [9] “The Jtella Gnutella Classes”, available online at: <http://jtella.sourceforge.net/>
- [10] “WHOIS Domain-Based Research Services”, website: <http://www.whois.net/>
- [11] “Free Riding on Gnutella”. Adar, E., Huberman, B., First Monday, October 2000. <http://www.firstmonday.dk/issues/issue5.10>.
- [12] “Free Riding on Gnutella Revisited: the Bell Tolls?”, Hughes D., Coulson G., Walkerdine J., published in IEEE Distributed Systems Online, vol. 6, no. 6, June 2005. available online at: <http://csdl2.computer.org/comp/mags/ds/2005/06/o6001.pdf>