

Where in the world is Carmen BitDiego? And who is she, anyways...

Alexandru Iosup

A.Iosup@ewi.tudelft.nl

TU Delft, Dept. of Electrical Engineering, Mathematics and Computer Science,
Mekelweg 4, 2628 CD, Delft, The Netherlands

February 2, 2005

Abstract

Peer-to-Peer (or Client-to-Client) systems such as Gnutella and BitTorrent have recently attracted the interest of the Internet audience because of their ability to share content at remarkably high speeds while lowering the burden on the initial data owner. Besides optimizing the transfers, researchers have been focusing on issues like robustness, scalability, and high tolerance to faults. All these aspects are inextricably linked with the location and social behavior of the users. In this study we analyze a large BitTorrent data set covering over 40 million direct observations made for over 100 different media streams. Our analysis offers three unique views of the location and social behavior of the BitTorrent community: geographical, geopolitical and organizational. We show that these views yield non-trivial results and that their use could greatly benefit the world of P2P systems.

Keywords: P2P analysis, geographical view, geopolitical view, organizational view, user-network view.

1 Introduction

Peer-to-Peer file sharing systems are becoming more and more present in the day-to-day life of content consumers. Recent studies show that P2P file sharing traffic occupies as much as one third and over of the global data traffic [3, 4, 6], which makes it the most important data traffic in the world. By comparison, Web surfers produce 10 times less traffic than P2P users [6]. The main challenge with such systems is to make the dynamic, ad-hoc networks of users to efficiently share data, a task which relies heavily on

the peer characteristics of such networks [8].

With BitTorrent becoming the dominant P2P file-sharing system (in terms of bandwidth consumption), a complete study of its geographical, geopolitical, and organizational components is becoming critical to understanding the P2P world at large. Indeed, establishing patterns can result in understanding, and most likely optimizing such systems. For instance, a P2P system could exploit the possibility that users from a given country would develop communities based on common taste, or that users from the same country would exhibit a similar time pattern for data access. On the same level of adaptation, some ISPs could realize that caching P2P data transfers would result in huge savings, due to optimized use of the network capacity. Also, with shared media being in many cases duplicated in the eyes of the consumers, it is imperative to study the behavior of users in such cases (see the discussion in Section 2.2.3).

However, obtaining accurate information about a P2P network (*tracking*) is much like the anecdotic search for Carmen SanDiego, the world's most famous spy. Carmen used to hide herself from the world's police and fans alike; from time to time, she would leave a sign on a hotel or at some other public place, just for fun. The sign would be a symbol of the next possible location of another sign, which would reveal another sign, and so on. Only a mind accustomed to the local traditions from the places situated all across the world would know how to link the clues to future locations. Even if Carmen was never caught, finding all the clues and establishing the relationship between them would have eased considerably that task; but this was also never fully accomplished. We argue that this is also the case for BitTorrent. To complete the parallel between Carmen SanDiego and BitTorrent, it is sufficient to call the set of clues

set of peers and the search for all possible locations *tracking* the clues – we are now looking for Carmen BitDiego, the (currently) world’s most famous P2P network.

In this study we analyze BitTorrent data covering several months, over 40,000,000 contacts, over 200,000 unique users, for a number of 120 files and 9 specific media type categories. Our analysis offers three unique views of the location and social behavior of the BitTorrent community: geographical, geopolitical and organizational. We think that such a study would bring a new and non-trivial light upon the patterns that are forming up in the world of BitTorrent and, possibly, other P2P networks.

The main contributions of our work are:

- The first unitary set of geographical, geopolitical, and organizational views of the BitTorrent network for a large set of files coming from very different media categories;
- The first view of BitTorrent’s *alias media* phenomenon (see Section 2.2.3) for a very popular file;
- The proof that BitTorrent shared media presents trivial and non-trivial locality features.

The paper is organized as follows. In Section 2 we give some preliminary information about this work. In Section 3 we present the methods employed for analyzing the BitTorrent network, and the expected results. In Section 4 we discuss our results. Section 5 compares our work with a number of related studies. Finally, section 6 presents our conclusions and a possible road map for future work.

2 Preliminaries

This section introduces the subject of our work, the BitTorrent P2P network, and the vehicle of our work, the data set.

2.1 A brief introduction to BitTorrent

BitTorrent is a second generation peer-to-peer network. As such, its main feature is the high data

transfer speed. Data exists in the network in the form of torrents (files or archives with multiple files). BitTorrent employs a *tit-for-tat* mechanism for data sharing – each user has to contribute to the network in order to obtain something from it; this leads in turn to data spreading at much higher speeds. To facilitate the exchange process, files are split in smaller parts, called *chunks*; the whole *tit-for-tat* mechanism is used to exchange *chunks*. Another important BitTorrent design issue is the lack of a contents search system at the peer level.

BitTorrent’s world is composed of peers, trackers, and web sites. As in any other P2P network, *peers* have theoretically equal rights to download files. However, to ensure high speed transfers, users with higher bandwidths are favored by the data transfer algorithm. The *trackers* are semi-centralized components (their activity is centralized but their number makes the whole level work as a decentralized network) used to keep track of the network transfers, as such, each file (content) in the network is monitored (*tracked*) by a tracker. By using the services of a tracker, BitTorrent peers can find out who are the other peers that can chunks of the desired file, so trackers also act as *redirector services* for the peers network. *Web sites* are used to locate data. For this, the web sites provide pages with BitTorrent digital content description (movie and music titles, authors, number of user that have the complete file or parts of it, and so on) that a network user may need to retrieve the desired digital data. The most important aspect regarding web sites is that they help providing *moderated* contents – that is, the web site supervisors verify the content that is inserted on the web and do not allow polluted data to be shared with unsuspecting users. Because of this, the BitTorrent network is virtually pollution-free.

To summarize, the life of a BitTorrent user would look very much like the following: the user would first go to a specialized BitTorrent web site, find the files that she desires (movies, music or any other digital content), download the description of the file tracker, in the form of a `.torrent` file, and then open the tracker description with her BitTorrent client (peer software). From this, the user would just wait online until the complete file is downloaded from the net-

| Group name | Group size (# of Files) | # of Records (in M) |
|------------|----------------------------|------------------------|
| All | 120 | 42.81 |
| Big | 12 | 34.83 |
| Small | 108 | 7.98 |

Table 1: The major groups of files.

work; the client software automatically handles the BitTorrent protocol and *tit-for-tat* mechanism.

2.2 The data set

This section describes the data set used in this study. A complete description on the data acquisition methods can be found in [7]. It is important for this study that the data set covers over 95% of the available peers [7].

2.2.1 Data set structure

The data set tracks the peer behavior during the download of 120 files shared using the BitTorrent network. File *tracking* data is stored in *trace files*, which consist of time-stamped observations. Each observation consists of an (*IP, port, number of chunks*) record.

The data set was acquired between December 2003 and March 2004. A number of 12 big *trace files* have been acquired in the period December 2003-January 2004. Another 108 small *trace files* have been acquired during March 2004. The terms *Big* and *Small* refer to the number of time-stamped observations, with *Big* files containing over 500,000 time-stamped observations, and *Small* files containing under 500,000 time-stamped observations.

2.2.2 Data set distribution

The 120 *trace files* in the data set cover a wide set of file types, from games to movies and from music to computer applications. There are 3 general groups: all the files available in the data set—group *All*—, all the files in the big files data subset—group *Big*—, and all the files in the small files data subset—group *Small*.

| Category name | Category size (# of Files) | # of Records (in M) |
|----------------------|-------------------------------|------------------------|
| Movie, popular | 10 | 19.73 |
| In English | 7 | 9.54 |
| In German | 2 | 9.67 |
| In French | 1 | 0.51 |
| Movie, themed | 1 | 2.12 |
| Movie, animé | 2 | 0.01 |
| Game, popular | 1 | 13.14 |
| Game, themed | 1 | 0.17 |
| Music, popular | 1 | 0.15 |
| Music, album | 1 | 0.05 |
| Application, popular | 1 | 0.02 |
| App, educational | 1 | 0.03 |

Table 2: The nine categories of files. The group *Movie, popular* is sub-divided into the English, German, and French versions.

We have also split the data set into 9 specific categories (e.g. popular movies or themed games). The files belonging to the 9 specific categories are also part of the larger groups. Table 2 shows various statistics about the categories.

2.2.3 Aliased Media

From each of group of files we have selected the most representative file as the file with the highest number of unique downloaders. As an observation, the same media contents (e.g. movie "X") can be available in the same P2P network in several forms, in very similarly, but not identically, named files (e.g. movie "X" may appear as "X.by.YYY", "X.created.on.03.2004" and so on). We call the whole set of files regarding the same media contents as *alias media*. In the case one of our selected files referred to such a media content, we have selected all the traces corresponding to its *alias media*. Table 1 shows the 9 major categories and the cardinality of the *alias media* for that category (column *Group size*).

We show in Section 4 that these categories of files yield non-trivial results and represent one of the major achievements of this study.

3 Methods, or how to catch 'er

3.1 Geographical analysis

Because of the locality features of the underlying network, BitTorrent needs a proper characterization of the geographical identity of its users – the continents where they are located. For this, we have mapped the tracked IPs (see Section 2.2.1 for more information on the data) to specific countries, and then to their corresponding continents. Using the observed data *number of chunks* information, we were able to also measure the users *weight* (the total consumed bandwidth) for each continent.

The novel approach is to perform this IP-continent association also for specific groups of files, instead of only for the whole data set. This approach can provide interesting insights into the coarse locality of the data, a feature that cannot be derived from a model of the network under study.

3.2 Geopolitical analysis

As an extension to the geographical description of BitTorrent users, we focused on the description of the geopolitical characteristics of the same users. For this, we mapped the tracked IPs (see Section 2.2.1 for more information on the data) to specific countries and cities. Using the observed data *number of chunks* information, we were able to also measure the users *weight* (the total consumed bandwidth) for each country or city discovered in the IP matching process.

3.2.1 Country-based analysis

Analyzing the distribution of users per country is critical for understanding the locality of shared media. The most important question to be answered by the analysis is whether there are countries which display a clear preference for specific types of media. We are interested in the trivial correlation country/preferred language, but also to more interesting ones, like the ones displayed in section 4.2. These correlations could prove interesting for optimizing traffic at the country level.

Our country analysis tools are based on MaxMind's [5] GeoIP libraries and databases. Having a local

copy of an IP-country lookup database allowed us to track our complete data set in a reasonable amount of time. Also, the novelty of the database ensured correct answers to the question *What is the distribution of users/country for all the media categories in our data set?*

3.2.2 City-based analysis

The existence of a city/shared media preference correlation could be very important for local ISPs and for optimizing the immediate (local) data sharing alike.

Our city analysis tools are based on MaxMind's GeoIP libraries and databases and on WebLog Expert's [9] databases. Again, the local availability of the IP-city matching engine was key to analyzing the huge data set at hand. However, the accuracy of the matching was below that expected, with over 30% of the IPs not being associated with a specific city. The reason is twofold: first, the database was publicly available and therefore restricted in contents, and second, the information itself (the name of the city an IP belongs to) is not always publicly available.

3.3 Organizational analysis

The organizational analysis is critical for establishing the impact of caching shared media transfers at ISP level. We are not referring to the legal aspects here - it is an established fact that P2P networks are used now also for transferring legal content (for a typical example see [2]). Using the observed data *number of chunks* information, we were able to also measure the users *weight* (the total consumed bandwidth) for each ISP identified during the analysis.

To answer the question *What are the ISPs that support the network?*, our tools rely again on MaxMind's GeoIP libraries and databases and on WebLog Expert's databases. This time the matching IP-ISP is almost perfect, with at most 1% of the observations not being matched, and with all Top20 ISPs contributing to the network, both in terms of number of users and of consumed bandwidth, being established.

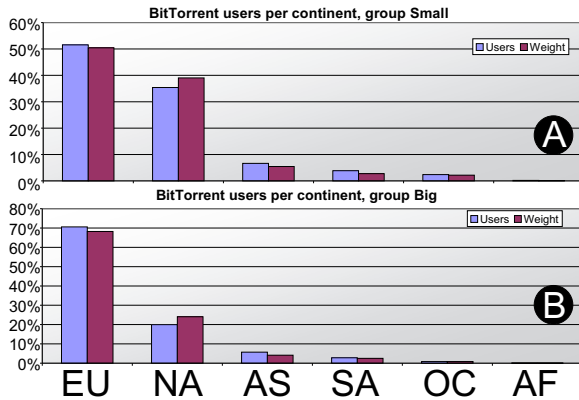


Figure 1: The distribution of users per continent: (a) for group Small; (b) for group Big.

4 Results, or how we caught 'er

4.1 Geographical analysis

Continent-based analysis The most important information in this section is that Europe is the most important contributor to the BitTorrent network (see fig.1. It is noteworthy that Europe is leading for the Big group as well as for the Small group. This means that European users dominate both the spectrum of files with lots of users, and the files with relatively few users, all from our data set.

Each continent in Figures 1 (a) and (b), has a block of 2 bars associated. The first (leftmost) bar refers to the number of users, while the second (rightmost) refers to their *weight*. The figures clearly indicate that, for a given continent, the percentage of users (from the total number of users in the world) is very close to the percentage of their weight (from the total weight in the world). The weight percentage is higher up to 5 percentage points than the number of users percentage for the continents that have a good network infrastructure, like North America, and lower below 5 percentage points for countries with a worse network infrastructure, like Asia. This leads to the (unverifiable from the data presented in this work) conclusion that users in BitTorrent have similar download patterns.

Figure 2 also demonstrates that the distribution

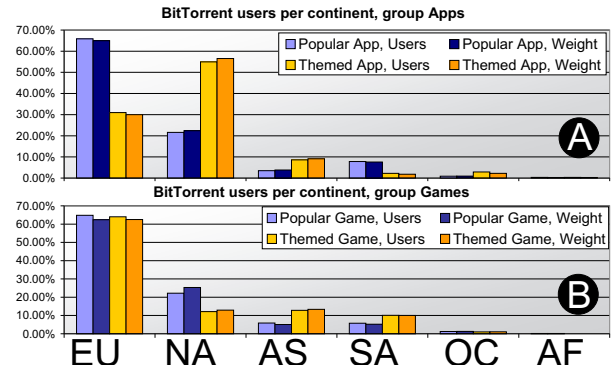


Figure 2: The distribution of users per continent: (a) for group Apps; (b) for group Games.

of users per continent varies with the data subsets. In Fig.2 (a), North America dominates the transfers of the *themed application* group, while in Fig.2 (b), Asia overcomes North America for the *themed game* group. This proves the existence of a coarse locality property for certain files or groups of files.

The same discussion as in the case of Fig.1 holds for Fig.2: the percentage of users per continent (from the total number of users in the world) is very close to the percentage of their weight (from the total weight in the world).

4.2 Geopolitical analysis

4.2.1 Country-based analysis

Unlike the situation from the distribution of users per continents, the distribution of users per country has no clear winner. Fig.3 (a), shows that, for the Small group of files, US is the clear winner, with Great Britain, Canada and Germany following from the distance. However, 3 (b), shows that, for the Big group of files, Germany takes the former position of US, with US, Great Britain and France being the distant followers. As an affective remark, The Netherlands is occupying only the 6th place, while Romania only around the 50th position.

We move now to the locality properties of the shared data. Figures 4 and 5 show that some countries present very interesting preferences. Fig.4 shows

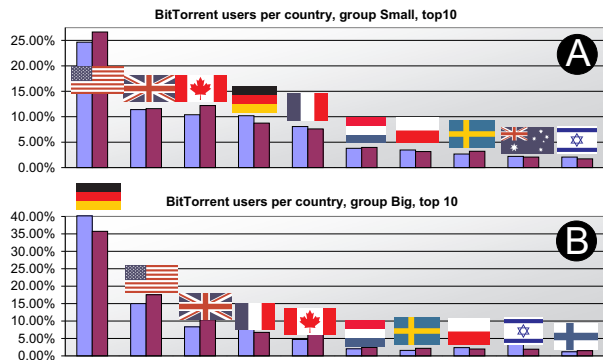


Figure 3: The distribution of users per country: (a) for group Small; (b) for group Big.

that Israel is the most important contributor to the sharing of a themed movie, while Japan is only making marks in downloading Animé Movies. Fig.5 shows that Hong Kong and Chile are surprisingly important contributors to the sharing of a themed game. We attribute (very speculatively) the presence of Hong Kong to the fact that the themed game was a soccer management simulator which has proven to model very accurately the English and other European soccer competition, which are subject to intense (and illegal) gambling in Hong Kong [1]. It is important to also note that in the case of all these countries, their interest for other similar media (movies for Israel and Japan and games for Chile and Hong Kong) was at best reduced. Figures 4 and 5 show that, in these countries, the cumulated interest for other media in the same category was at most 20% of the interest for the media which presented a distinctive locality feature.

A similar discussion as in the case of Fig.1 holds for Figures 3, 4 and 5: the percentage of users per country is very close to the percentage of their weight.

4.2.2 City-based analysis

The city-based analysis did not result into meaningful results for our goals. Indeed, many of the users gather around several important cities, like Madrid, Paris, or Atlanta, but the fact that over 30% of the IPs could not be matched to real cities prevents us from gen-

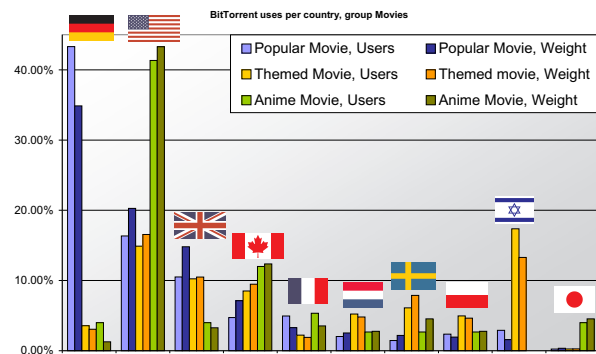


Figure 4: The distribution of users per country for group Movies.

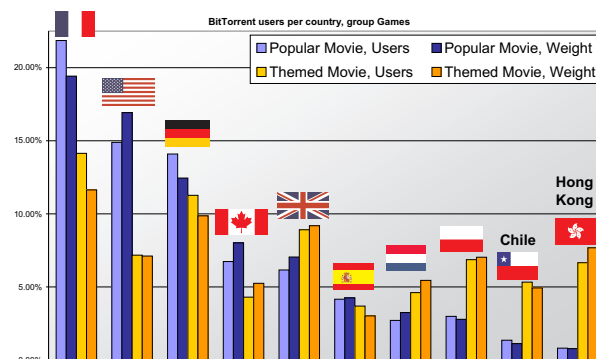


Figure 5: The distribution of users per country for group Games.

eralizing our conclusion. Another aspect has become an obstacle to creating a conclusive users/city distribution: the existence of *buffer cities*—small cities that host important network conjunction nodes, like Oldenburg and Eschborn (DE), and Herndon (US). The *buffer-cities* cover another 15% of the total IPs.

4.3 Organizational analysis

ISP-based analysis Our measurements show that the top 20 organizations are attracting over 37% of the world-wide traffic for the Small group and over 54% of the world-wide traffic for the Big group. Table 3 shows the top 10 ISPs, with regard to the weight of their users.

| Organization (Country) | Weight (%) |
|---------------------------------|------------|
| Deutsche Telekom AG (DE) | 5.51 |
| Comcast Cable (US) | 4.85 |
| America Online (US) | 2.30 |
| Road Runner (US) | 3.28 |
| British Telecommunications (GB) | 2.09 |
| Pac Bell Internet Services (US) | 2.12 |
| Proxad Free SAS (FR) | 1.90 |
| Telewest HSD Platform (GB) | 1.86 |
| Telia Network Services (SE) | 1.63 |
| Neostrada Plus (PL) | 1.23 |
| Total | 26.77 |

Table 3: The Top10 organizations by users weight, group Small.

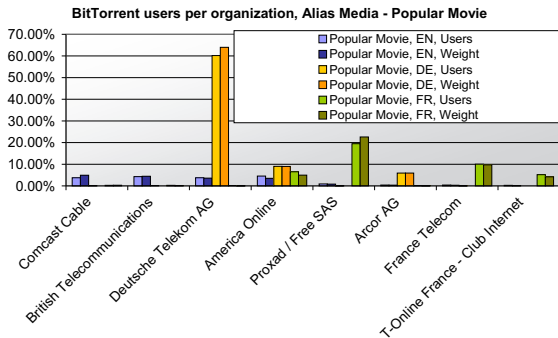


Figure 6: The distribution of users per organization for the popular movie subgroups (EN, DE, FR).

Fig.6 shows that, for some files, only a few ISPs cover the vast majority of the users. For instance, for the German version of a popular movie, over 60% of the users use the services of only one ISP, *Deutsche Telekom AG*. However, for the French version of the same movie, the top 4 ISPs, *Proxad/Free SAS*, *France Telecom*, *T-Online France - Club Internet*, and *America Online* cover only approximately 50%. It is clear that, in some cases, ISP-based optimizations are possible, while in others, despite the locality of the data in terms of country (over 60% of the users for the aforementioned French version came from France), there are too many ISPs that serve the majority of the users.

A similar discussion as in the case of fig.1 holds for

figs.6: the percentage of users per organization (from the total number of users in the world) is very close to the percentage of their weight (from the total weight in the world).

5 Related work

Our work is similar in goals to a number of recent studies on Gnutella, KaZaa and even BitTorrent. Our work complements and builds upon these previous efforts by providing and correlating information for two layers of characterization: the locative layer (geographical, geopolitical, and organizational views) and the behavioral layer (temporal, network robustness, the downloads consistency, and special peer behavior views).

A comprehensive study of BitTorrent’s high-level characteristics has been performed by the very authors of this work in [7]. This work should be regarded as the logic complement of the previous study, in that it provides detailed insights into the low-level and social characteristics of the BitTorrent network.

One of the first studies regarding the characterization of P2P file-sharing systems is the work of Saroiu et al. [8]. They characterize the one-point-to-target latency, bottleneck bandwidth, the user connection/disconnection frequency, and the number of files and correlate this data.

To our knowledge, the only complete study of the geographical aspects of a P2P network is the study of Gnutella’s global geographical, geopolitical and organizational views by Yazti et al. [10]. Our work adds the needed details for specific media types, resulting in proving the existence of a locality property of BitTorrent shared media.

In [2], the authors present the life of a file shared using BitTorrent. The file comes from the operating systems domain and is particularly large, thus being outside the targets of a common P2P file-sharing system user. Our study enhances this work with another 9 specific and 2 general categories of files, and advances the knowledge of user behavior.

6 Conclusions and future work

In this paper we have presented a threefold view of the geographical, geopolitical and organizational structure of the BitTorrent file-sharing system.

Four important lessons surface our study. The most important lesson is that BitTorrent shared media displays various levels of locality, both trivial and non-trivial. The most obvious types of locality are the communities of users speaking the same language sharing localized versions of some data. The non-trivial examples include the preference of some countries for special categories of contents (e.g. Hong Kong downloading a soccer management simulator) and the presence of single ISPs that serve over 50% of some media's users. The second lesson is that Europe is the most important contributor to BitTorrent's traffic. The third lesson is that some BitTorrent exhibits the presence of *aliased media*—the same contents presented under different names and/or languages. The final lesson is that the percentage of users per continent, country, or organization (from the total number of observed users) is very close to the percentage of their weight (from the total observed weight).

For the future, we plan to improve this study with a more recent and diverse set of data, and to add temporal and peer activity analysis. This would help us conclude on several issues that have arisen in this work, but which could not be proved because of the nature of the study. The continuation of this work would most likely also lead to a number of possible improvements of the BitTorrent network.

Acknowledgements

The author would like to thank Johan Pouwelse and Pawel Garbacki alike for their fantastic support throughout the conception and development of this study. This work could not have been completed without their support. The author would also like to thank Dick Epema for the support offered in accomplishing this task.

References

- [1] Hong Kong Govt. Gambling review: A consultation paper, Jul 2001. <http://www.info.gov.hk/archive/consult/2001/gambling-e.pdf>.
- [2] M. Izal, G. Urvoy-Keller, E.W. Biersack, P. Felber, A. Al Hamra, and L. Garcés-Erice. Dissecting BitTorrent: Five months in a torrent's lifetime. In *Passive and Active Measurements (PAM 2004)*, April 2004.
- [3] JoltID. Peercache. January 2005.
- [4] T. Karagiannis, A. Broido, N. Brownlee, k. claffy, and M. Faloutsos. Is P2P dying or just hiding? In *Global Internet and Next Generation Networks (Globecom 2004)*, Dallas, Texas, US, Dec 2004.
- [5] MaxMind, LLC. <http://www.maxmind.com/>.
- [6] Andrew Parker. The true picture of peer-to-peer file-sharing. CacheLogic Presentation, July 2004.
- [7] J.A. Pouwelse, P. Garbacki, D.H.J. Epema, and H.J. Sips. The BitTorrent p2p file-sharing system: Measurements and analysis. In *Proceedings of the 4th International Workshop on Peer-To-Peer Systems (IPTPS'05)*, Ithaca, New York, USA, February 2005.
- [8] S. Saroiu, P. Gummadi, and S. Gribble. A measurement study of peer-to-peer file sharing systems, 2002.
- [9] WebLog Expert. <http://www.weblogexpert.com/>.
- [10] Demetris Zeinalipour-Yazti and Theodoros Foflias. A quantitative analysis of the gnutella network traffic. Course Project for "Advanced Topics in Networks", with M. Faloutsos at the University of California - Riverside, Dpt. of CS, April 2002.