How Much Information? 2003

Summary
Exec Summary

Stored Information
Paper | Film | Magnetic | Optical

Information Flows
Broadcast | Telephony | Internet

Wrap-up
Thanks | Printable (PDF)

Internet

8. INTERNET

Although the Internet is the newest medium for information flows, it is the fastest growing new medium of all time, and becoming the information medium of first resort for its users. Note that the Web consists of the surface web (fixed web pages) and what Bright Planet calls the deep web (the database driven websites that create web pages on demand).

Table 8.1: The size of the Internet in terabytes.

Medium	2002 Terabytes
Surface Web	167
Deep Web	91,850
Email (originals)	440,606
Instant messaging	274
TOTAL	**532,897**

Source: How much information 2003

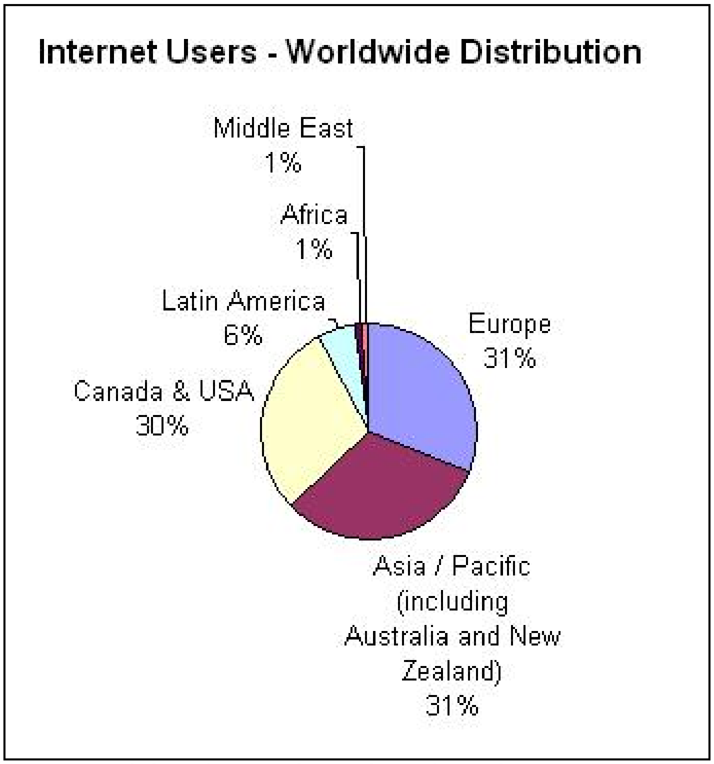
I. General Internet Usage Statistics

The CyberAtlas Stats Toolbox provides pointers to a tremendous assortment of Internet-related usage data. Here are a few topics of general interest.

A. What is the international distribution of Internet users?

According to Nielsen/NetRatings, there is a worldwide Internet population of 580 million users, as of 2002. The International Telecommunications Union provides a 15 percent higher estimate of 665 million users.

Of these, roughly 30 percent are in North America, 31 percent are in Europe, and 38 percent are in other parts of the world, as shown in the following chart.



Source: Nielsen/NetRatings via CyberAtlas

B. How many hours do individuals spend online?

The average global Internet user spends 11 hours and 24 minutes online per month, according to Nielsen/NetRatings. The average user in the United States spends more than twice that amount of time online: on average, 25 hours and 25 minutes at home and 74 hours and 26 minutes at work.

C. What activities do people do while online?

The Pew Internet and American Life Project reports that on an average day, about 72 million United States users go online. Here are the kinds of activities they do online (top 10 shown here; see the Pew site for the full list).

Table 8.2: Daily Online Activities

Activity	Percent of those with Internet Access	Most recent survey date
Send email	52	March - May 2003
Get news	32	March - May 2003
Use a search engine to find information	29	January 2002
Surf the web for fun	23	March - May 2003

Look for info on a hobby	21	March - May 2003
Do an Internet search to answer a specific question	19	September 2002
Do any type of research for your job	19	November 2002
Research a product or service before buying it	19	December 2002
Check the weather	17	March - May 2002
Send an instant message	14	March - May 2003

Source: Pew Internet and American Life Project

D. How many web searches are conducted per day?

According to SearchEngineWatch.com, as of January 2003, there were 319 million searches per day at the major search engines. This figure is calculated using the Nielsen//NetRatings "search hours;" the total time spent by all visitors searching at each engine.

Table 8.3: Search statistics

Search engine	Search hours per month (in millions)	Search minutes per day (in millions)	Searches per day (in millions)
Google	18.7	37	112
AOL Search	15.5	31	93
Yahoo	7.1	14	42
MSN Search	5.4	11	32
Ask Jeeves	2.3	5	14
InfoSpace	1.1	2	7
AltaVista	0.8	2	5
Overture	0.8	2	5
Netscape	0.7	1	4
Earthlink	0.4	1	3
Looksmart	0.2	0	1

Lycos	0.2	0	1
TOTALS	**53.2**	**106**	**319**

Source: SearchEngineWatch.com, Feb. 25, 2003

II. World Wide Web

A. Published Estimates on Size and Character

Most size studies of the World Wide Web have focused on the number of hosts connected to the network. Preliminary estimates of the amount of data on the web have been made but have not been kept up to date or do not have defined measurement methods. In our last survey we quoted a page size estimate of 18.7 KB from a 1999 article in Nature Magazine that was generated by statistically sampling web servers.

Table 8.4 shows a range of approaches to sizing the Web:

Table 8.4: Methods for Sizing the World Wide Web

Company	Domains	Method	Frequency
www.whois.net	31,987,198	Domain name registration, this number changes continuously	Continuously
www.netcraft.com	42,800,000	Web servers responding to HTTP request. Each domain name is counted as a server.	Monthly
http://wcp.oclc.org/	9,040,000	IP addresses responding to HTTP request (each IP can maintain many virtual domain names)	Yearly

Source: How much information 2003

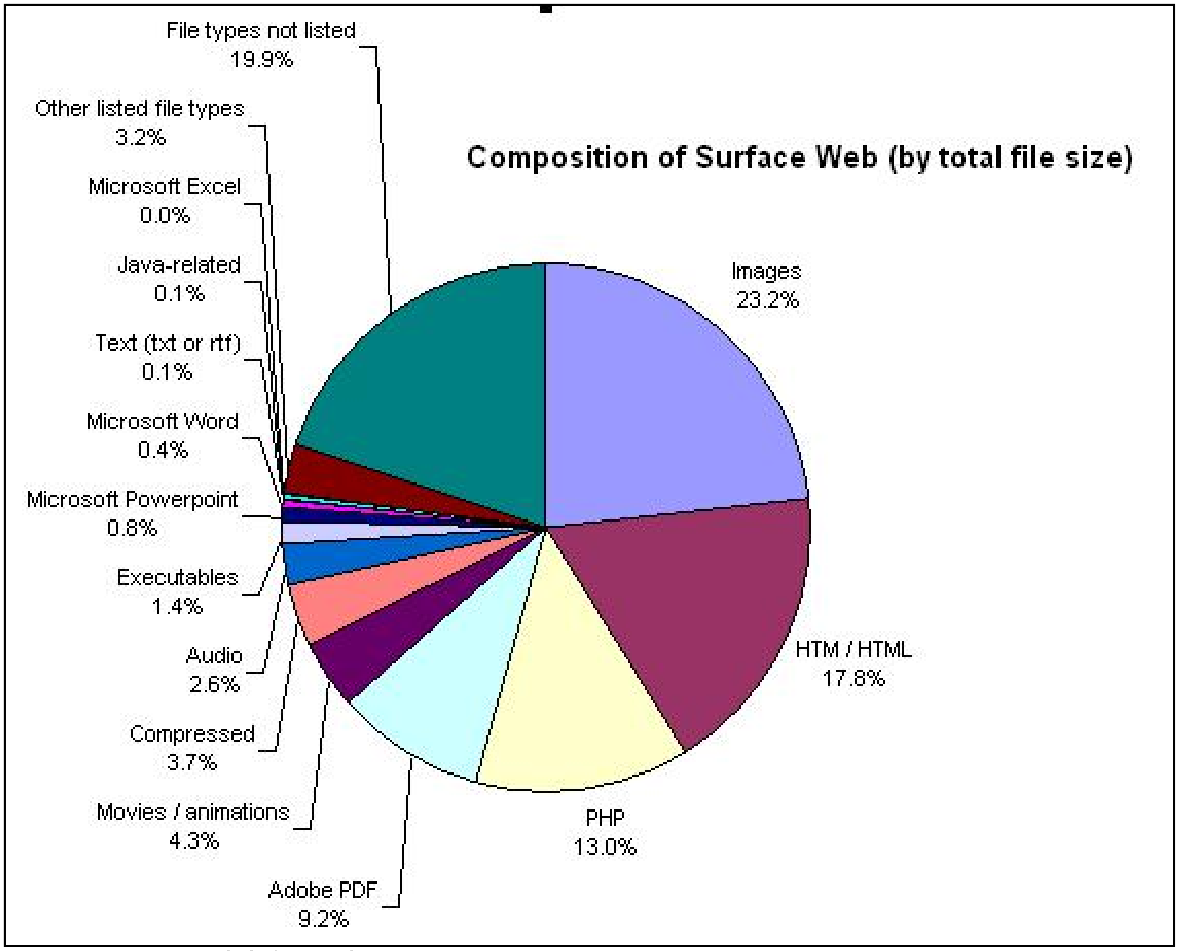
B. Web Sampling Study

We downloaded and analyzed the contents of 9,800 websites in order to estimate the size of an average webpage and the contents of an average website. These sites were chosen randomly from a list of 61 million URLS compiled by the Internet Archive. After confirming that a URL was listed in a DNS registry, we downloaded each registered site in its entirety using "wget," a website mirroring tool. We mirrored each site recursively, following only links that were relative to the original domain name. Once an entire site was completely mirrored we used pattern matching on the file type extensions, and recorded the size of the files. For each site we generated a total size for the site, the total number of files, and file and size totals for a variety of common file types.

Note that we were only able to sample the “surface web”—the static, publicly available web pages which represent a relatively small portion of the entire Web. We were not able to download or measure the dynamic, database-driven websites which comprise the “deep web.” As quantified in a landmark study by BrightPlanet in 2000, the “deep Web” is perhaps 400 to 550 times larger than the information on the “surface.”

The sum of all the web site file sizes in our sample equals **33.1 GB**. As this sample of 9,806 sites represents 0.02 percent of the 42.8 million web servers (according to the NetCraft Survey, as of August 2003), we may estimate the total size of the surface web as **167 TB** (95% confidence, +/- 1). Therefore the "deep Web" may be between **66,800** and **91,850 TB**.

We counted the number of files of various types (e.g. HTML, images, audio) and found the file types in our sample to be distributed as follows:



Source: How much information 2003

We also looked at the functionality and content of the web pages in our sample by matching keyword indicators to text found on the index page of each site. Highlights of our results include:

- **Search**. In 29.8% of the sites sampled, we found the keyword “search” on the index page. Presence of a search function suggests a large or complex site.
- **Form**. The keyword "form" appears in 31.9% of the sampled index pages.
- **Javascript**. The keyword “javascript” appeared in 19.1% of the sites sampled, indicating a degree of website sophistication and interaction.
- **Protection**. In 7.7% of the sites sampled, we found the keywords “password” or “login,” indicating some degree of protection for the content.
- **E-commerce**. On the index page of each site, we checked for words which typically indicate an e-commerce site: cart, shopping, and checkout. In 5.4% of our sampled sites, we found one or more of these words.
- **Porn**. 2,743 sites (or 28%) appeared to contain pornographic content. To generate this statistic, we matched a list of 94 pornographic stopwords to terms in the associated URL and the index page.

For more details on our methodology and findings, see the full writeup (to be submitted for publication

11/2003).

III. Blogs (web logs)

A. What is a blog?

Short for “web log,” a blog is a Web page that serves as a publicly-accessible personal journal for an individual. Typically updated daily, blogs often reflect the personality of the author. A more complete definition is offered by Jill Walker. In her entry for Routledge’s Encyclopedia of Narrative Theory, Walker writes that a blog is:

> “a frequently updated website consisting of dated entries arranged in reverse chronological order so that the reader sees the most recent post first. The style is typically personal and informal. Freely available tools on the World Wide Web make it easy for anybody to publish their own weblog, so there is a lot of variety in the quality, content and ambition of weblogs, and a weblog may have anywhere from a handful to tens of thousands of daily readers. Weblogs first appeared in the mid-nineties and became more widely popular as simple and free publishing tools such as Blogger.com became available towards the turn of the century.
>
> Examples of the genre exist on a continuum from online diaries that relate the writer’s daily activities and experiences to less confessional weblogs that comment and link to other material, discuss a particular theme or function as soapboxes. In addition to the dominant textual form of weblogs there are experiments with adding sound, images and videos to the genre, resulting in photoblogs, videoblogs and audioblogs.
>
> Each entry in a weblog tends to link to further information. Weblog authors also link to other weblogs that have dealt with similar topics, allowing readers to follow conversations between weblogs by following links between entries on related topics. Readers may start at any point of a weblog, seeing the most recent entry first, or arriving at an older post via a search engine or a link from another site. Once reading a weblog, readers can read in several orders: chronologically, thematically or searching by keywords. Weblogs also generally include a blogroll, which is a list of links to other weblogs the author recommend, and many weblogs allow readers to enter their own comments to individual posts.”

Klogs or knowledge-logs are a subset of web logs. According to klogger Spike Hall, a k-log (knowledge log) is "a weblog but also demonstrates or documents a knowledge claim and/or it documents and illustrates the dynamic individual process of a quest for knowledge.”

B. How many blogs are there?

Blogcount.com addresses the “blogosphere,” asking “What is its size, shape, color, true nature?” Blogcount collects and organizes the best reports and analyses on this subject. As of June 23, 2003, Phil Wolff estimates that there are 2.4 to 2.9 million active web logs. He bases his estimate upon statistics from centrally hosted weblogs:

Table 8.5: Hosted blog statistics

Web log host	Registered	Active	As of
LiveJournal	1,121,464	526,535	23 June 2003
Blogger	1,500,000	705,000	9 June 2003

Diaryland	850,000	400,000	March 2003
	Total:	1,631,535	

Source: Blogcount.com

He notes that this figure does not include smaller hosts such as Radio and Moveable Type or blogs hosted behind firewalls on private intranets.

If each blog is 50 KB, then the total size of the active blogosphere is **81 GB**.

C. Who is blogging?

According to Jupiter Research, about 2 percent of Internet users have created a blog. The majority of bloggers use dial-up access to get online, and more than half have a household income below $60,000 per year. Jupiter also found that blogging is split evenly between the genders and that 70 percent of the bloggers have used the Internet for more than 5 years. (Source: Blogging by the Numbers)

More than 50 percent (350,000) of the 655,000 web logs crawled in National Institute for Technology and Liberal Education (NTILE) web log census are written in English. The rest of the top 10 languages for blogs are (in order): Portuguese, Polish, Farsi, French, Spanish, German, Italian, Dutch and Icelandic.

D. Who is reading web logs?

Jupiter Research estimates that only 4 percent of the online community read blogs. “Blogs seem to be read mostly by men (60 percent vs. 40 percent women), in homes where the total income is more than $60,000 per year (61 percent). Dial-up remains the connection of choice (54 percent compared to 46 percent broadband), and the majority (73 percent) of blog readers have been online for more than 5 years.” (Source: Blogging by the Numbers)

E. More interesting blog facts

- 10,000 domains listed in the whois registry have “blog” in their names
- On average, blogs are updated every 3 days, according to a brief informal study reported at BlogCount.
- About four percent of online Americans report that they have gone to blogs for information and opinions related to the war in Iraq.
- Nielsen/NetRatings says that in May 2003, LiveJournal was the 650th most popular site on the Internet by unique audience. The 184,000 people visiting about every ten days (3.13 visits per person during the period) were so active that LiveJournal was number 213 by pages viewed. When people came, they spent 22 minutes at the site, hitting the back button about 26 percent of the time.
- During the summer of 2003, America Online introduced AOL Journals, allowing members to update blogs by instant message. As of December 2002, AOL has 35.2 million paid members, using 8 languages. The journals will be integrated with AOL community forums, instant messaging, photo albums, and home page construction kits.

IV. Email and Spam

Email is currently one of the most widespread methods of communication. Email users comprise:

- 35% of the total U.S. population (eMarketer)

- 50% of U.S. consumers (Forrester Research)
- 94% of U.S. Internet users (eMarketer)
- 98% of employed Americans with Internet access– 57 million adults (Pew Internet)

Forrester also reports that email use accounts for over 35 percent of all time spent on the Internet, while a PricewaterhouseCoopers survey found people spending 84 percent of their time on the Internet for email.

The Pew Internet and American Life Project recently published their research on email at work. They found the following (through participant self-reports on activity):

- 60% of work emailers receive 10 or fewer messages on an average day; 23% receive more than 20 and only 6% more than 50.
- 78% of work emailers send 10 or fewer messages on an average day; 11% send more than 20.
- 73% of work emailers spend an hour or less per day on their email. That includes 23% of all work emailers who spend fewer than 15 minutes per day handling email.
- 46% of work emailers say their work email volume has stayed the same over the past year.
- 48% say their email volume has increased over the past year.

Since 1995, email volume has increased, though sources differ as to the degree of the increase. A study by Rogen International and Goldhaber Research Associates found that in 1995, employees sent an average of 3 emails a day and received five. As of 2002, employees were sending 20 a day and receiving 30. A 1999 study reported in Newsweek estimated that “a white-collar worker receives about 40 email messages in his office every day.”

Daily email traffic is expected to almost double by 2006, from 31 billion today to 60 billion, according to a new study by International Data Corporation (IDC). If each email is 59 KB (Source: Forrester Research), the daily flow of emails worldwide is currently **1,829 TB**. Over the course of a year, the total would be **3.35 petabytes**.

Table 8.6: Worldwide email messaging

Year	Emails per day	Emails per year
1999	5 billion	1.4 trillion
2000	10 billion	
2001		
2002	14.9 billion or 31 billion	4 trillion
2003		

2004		
2005		
2006	60 billion	

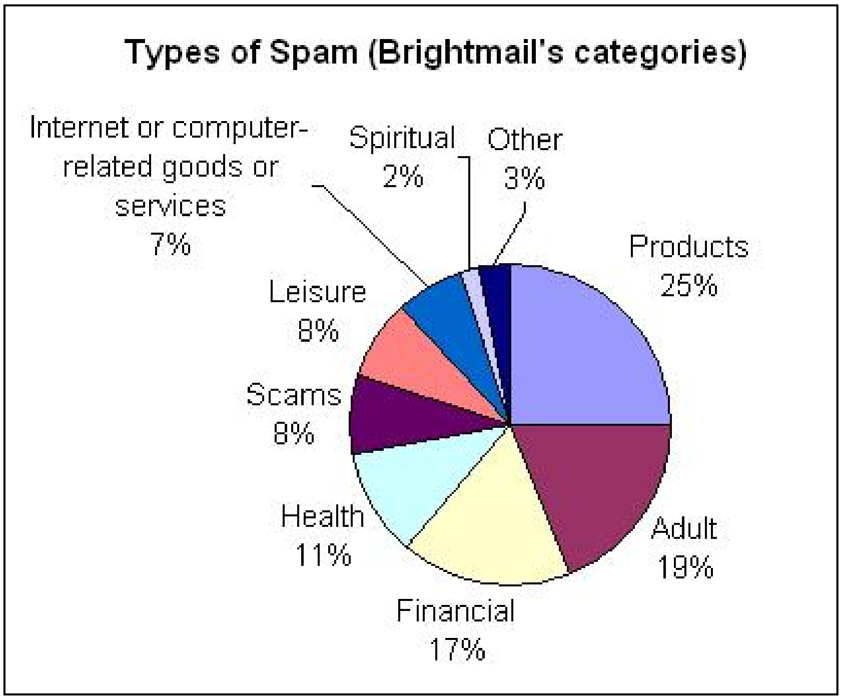
Source: IDC via [Channel One Market Overview](#)>

IDC also estimates that only half of the email traffic will be personal messages. Unsolicited email (also known as spam), commercial notifications and news alerts account for one-third of today's email load and will comprise nearly half of the traffic four years from now, the report said. Therefore we estimate the upper bound of original content in emails as **440,606 terabytes** (uncompressed), lower bound as **333,792 terabytes**.

Mailing lists can be viewed as a subcategory in email. It is hard to determine the number of mailing lists in existence, but we can approximate it based on some available statistics. One of the most frequently used mailing list managers - LISTSERV - is used to send 25 million messages per day in approximately 300,000 mailing lists. A sample of mailing lists has shown that 30 percent of them are managed using LISTSERV. Using this information, we would estimate the total number of mailing list messages at 30 billion per year with aggregate volume of 563 terabytes.

A. What is spam?

Spam is unsolicited bulk email. Brightmail, an anti-spam service provider, classifies spam into the following categories (percentages as of May 2003):



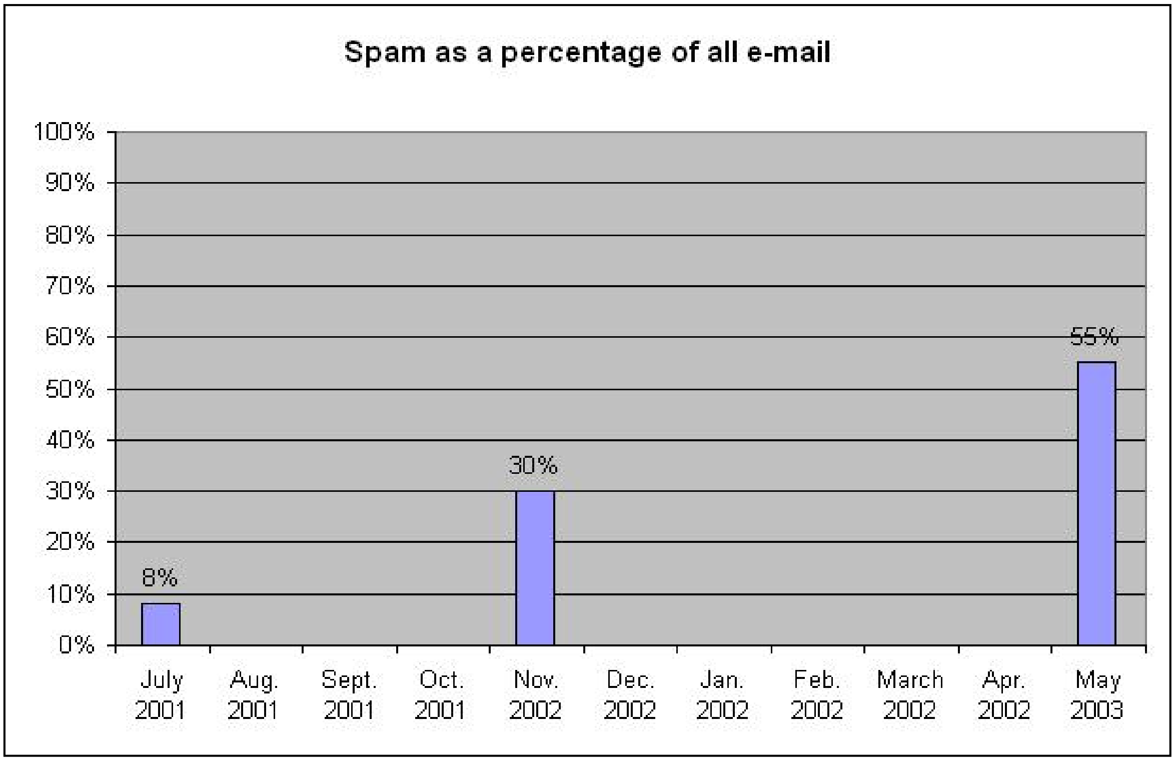
Source: BrightMail http://www.brightmail.com/spamstats.html

B. How much spam is sent each year?

Forrester estimates that by 2004, marketers will send more than 200 billion emails in the United States alone.

In December 2002, according to interception figures from Brightmail, a leading spam filtering company, unsolicited bulk email made up 40 percent of all email traveling over the Internet, up from 8 percent in 2001. MessageLabs, a U.K. spam filtering company provides slighter lower figures for 2002 (30 percent spam) but reports an increase to 55 percent spam as of May 2003. As a point of comparison, only 40 percent of United

States Postal Service mail is business marketing.



Source: MessageLabs

According to a recent CNN Techweb article, “Gartner analyst Joyce Graff says spam is a serious problem, but she disputes MessageLabs' estimate that spam will account for half of all email traffic by midyear. She says users often erroneously lump together all their annoying email into the category of spam, including both real spam, such as con games and phony businesses, and business mail that they're copied on by overzealous colleagues.”

C. How much spam is blocked or filtered by the ISPs?

The three major email service providers AOL, Microsoft and Yahoo! have more than 200 million email account holders, making them an attractive target for spammers.

The spam problem has shown a tremendous spiral, according to figures provided by AOL (see chart below). “In a single 24-hour period in March 2003, America Online says it trashed a billion spam emails using its software filters. AOL said its members used ``report spam'' buttons on their email software 5.5 million times during the same period. AOL said it blocks an average of 28 junk emails per account, per day. Graham said ``an extremely small fraction'' of the messages snagged in AOL's spam filters were legitimate ones. He declined to reveal any figures for that mail.”

Yahoo! offers comparable statistics, noting that it intercepts 1 billion spam messages a day.

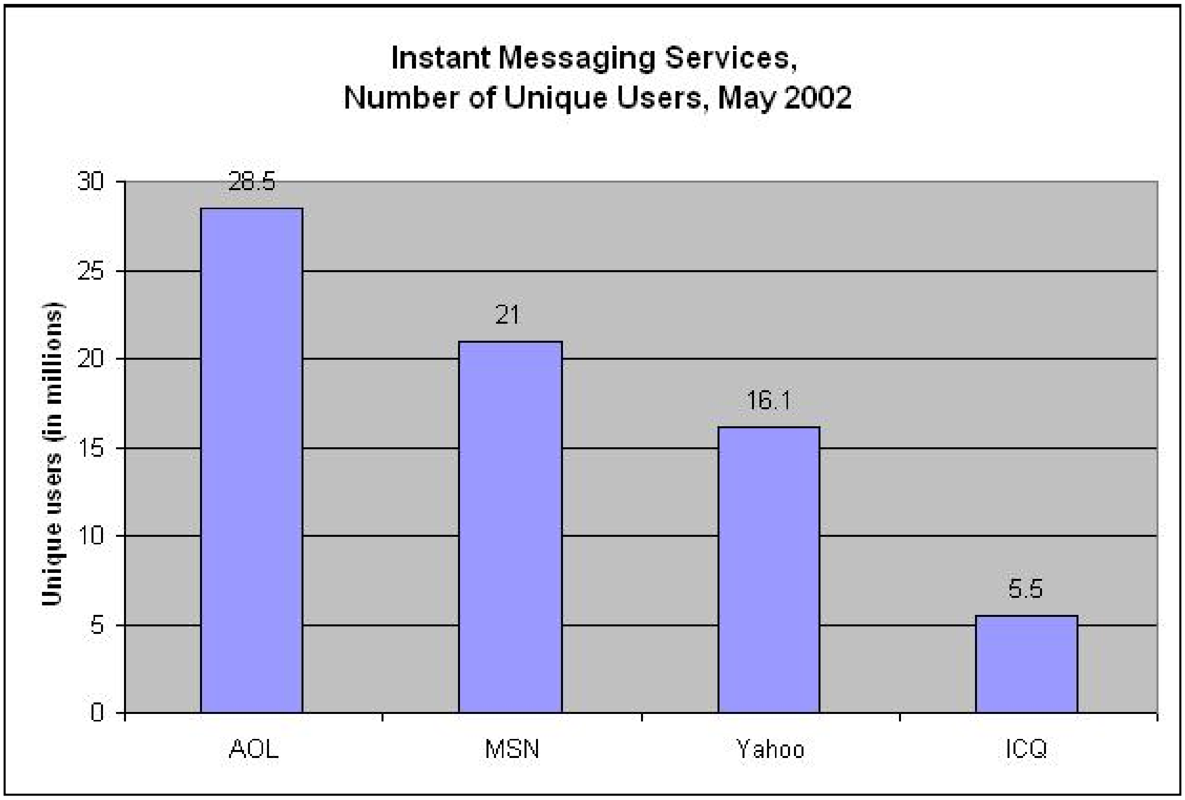
V. Instant Messaging

Instant Messaging Planet provides a concise definition of instant messaging:

> "IM is a type of communications service that enables one to create a private chat room with another individual. Typically, the instant-messaging system alerts the user to whenever somebody on their private list is online -- a capability known as "presence." They can then initiate a chat session with that particular individual. People can communicate with each other by typing with a PC, wireless device (cell phone, PDA, etc.) or other Internet appliance/device."

Nearly 40 percent of the active U.S.-based Internet-using population at home logged onto one of the public instant-messaging (IM) networks at least once in May 2002, while 31 percent of U.S. business Internet users used IM in that same time frame, according to a Nielsen//NetRatings study. Gartner Group forecasts 70 percent of all enterprises will use IM in 2003, and that by 2005 IM will represent 50 percent of all business-to-client communication.

The most widely used IM services are AOL, MSN, Yahoo and ICQ, as shown in the figure below.



Source: InternetNews.com, June 17, 2002

AOL's messaging statistics illustrate the rapid growth of this communication medium. As of 1999, users were exchanging 400 million messages per day. By the end of 2000, the rate had increased to 660 million messages--double the daily traffic of the U.S. Postal Service. As of June 2003, IEEE reports, AOL messaging has exceeded 2 billion per day.

If we assume similar rates of messaging at the other IM services, we can estimate the total messaging volume: more than 5 billion messages per day. If each text-only message is 0.15 KB (average English-language message length, according to Expresso Instant Messaging WhitePaper *[???]),* then the total daily volume is **750 GB** and the annual volume is **274 terabytes**. All of this can be considered to be unique content.

VI. Peer to Peer (P2P) File-Sharing

A significant new source of storing, creating and exchanging media and data on the Internet is through P2P file sharing networks. P2P file sharing has exploded in popularity since the creation of programs such as Napster in the late 1990's. The first file sharing application, Napster, allowed users to share only music files in the popular MP3 format. Today, there are numerous applications that allow users to share any type of file on the network. In a short period of time, programs such as KaZaA have become the most popular applications on the Internet besides email and Web browsing. KaZaA is the most popular application ever downloaded on the Internet, having recently reached over 230 million downloads worldwide, with an average of 2 million more per week (Source: Download.com). In addition, users on KaZaA share almost **5,000 terabytes** of information, including over 600 million files shared by an average 3 million users active at any given time (Source: KaZaA.com).

P2P applications tell us about how information is consumed on the Internet, but unlike other parts of this study, relatively little of this information is unique. Looking at a sample of users on the P2P network (approximately 40,000 nodes, 2 million files, 14.4 Terabytes), we were able to determine what types of files were being shared and what kind of files people were sharing on their hard drives. We collected our sample over a period of 24 hours, traversing 400,000 nodes to find approximately 40,000 users were sharing files with others. Using this, we were able to describe how P2P users consume and share information. We looked specifically at the number and size of filetypes being shared per user and on the aggregate, as well as what percentage of these files were unique or similar. We did not look at the content of user files.

A. How P2P users use their hard disks

The dataset was 14.4 terabytes, averaging 7.6 megabytes, with a total file count of almost two million.

Table 8.7: KaZaA total size and count: aggregate size for the dataset

Size	Avg of Size	Min of Size	Max of Size	Std Deviation	Total Files
14.4 TB	7.6 MB	1 BYTE	1.97 GB	44168598.19	1,980,426

Source: How much information 2003

The three file types that took up the most storage space were .avi video files, mp3 audio files and .mpg video files; combined they were more than 80 percent of the sample.

Table 8.8: Summary of size by file extension for the dataset: top 25 file types sorted by aggregate file size

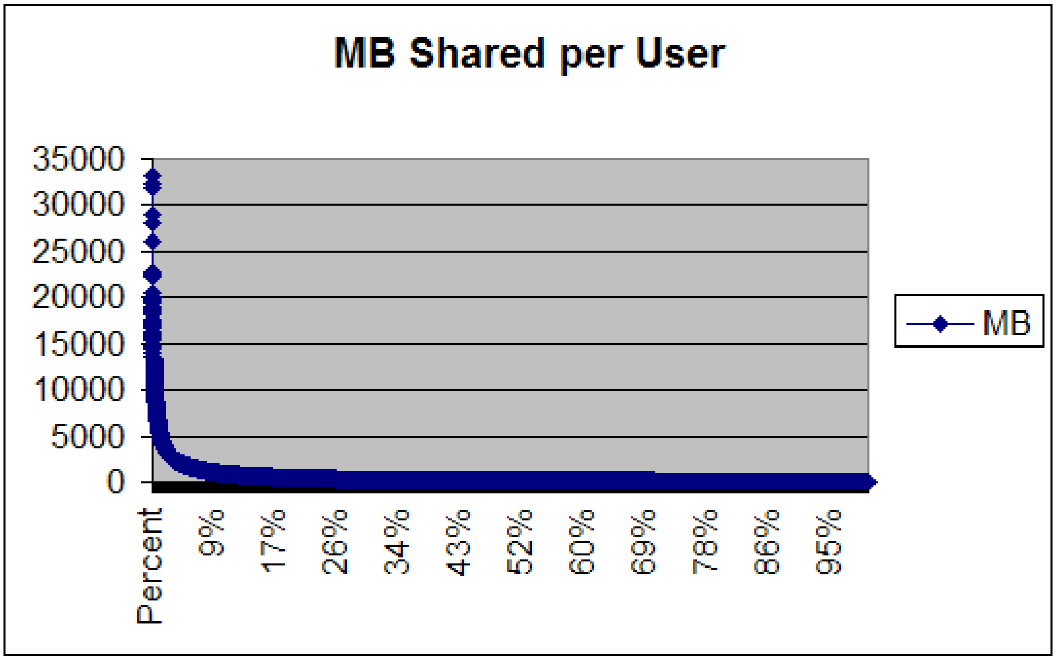
File Extension	GB	Avg of Size (bytes)	Min of Size (bytes)	Max of Size (bytes)	Number of files	Type	% of Total Size
avi	4,851.12	1.73E+08	82	2.12E+09	30,026	video	32.88%
mp3	4,564.79	4,120,597	1	5.58E+08	1,189,488	audio	30.94%
mpg	2,855.88	59,855,823	6	1.6E+09	51,231	video	19.36%
exe	550.76	7,293,783	1	8.16E+08	81,079	software	3.73%
mpeg	311.73	20,693,836	162	9.51E+08	16,175	video	2.11%
wmv	297.00	23,311,552	1316	8.75E+08	13,680	video	2.01%
asf	297.00	34,909,551	13	5.83E+08	9135	video	2.01%
none	248.37	19,767,776	1	1.67E+09	13491	unknown	1.68%
wav	243.44	15,177,476	44	1.24E+09	17,222	audio	1.65%
wma	113.79	2,732,813	111	7.64E+08	44,707	audio	0.77%
zip	63.23	21,470,521	1	1.91E+09	3,162	software	0.43%
mov	48.88	8,790,495	162	3.82E+08	5,970	video	0.33%
bin	34.29	44,311,001	10	8.45E+08	831	software	0.23%
vob	32.71	5.02E+08	8,192	1.07E+09	70	archive	0.22%

iso	27.68	3.81E+08	206,260	7.92E+08	78	archive	0.19%
cab	21.09	2,414,563	19	6.97E+08	9379	system	0.14%
rar	17.54	53,215,891	69	8.53E+08	354	archive	0.12%
pdf	13.22	5,411,649	162	1.34E+08	2,623	document	0.09%
cif	11.18	4.8E+08	38,062	8.36E+08	25	archive	0.08%
dat	8.16	3,,711,050	1	8.14E+08	2,361	system	0.06%
rmj	7.42	4,213,919	101,267	42,755,921	1,891	audio	0.05%
nrg	6.79	3.47E+08	342,172	8.16E+08	21	archive	0.05%
part	5.90	2.18E+08	21,759	6.15E+08	29	e-donkey temp files	0.04%
jpg	5.79	44,829.72	75	29,016,816	138,751	image	0.04%

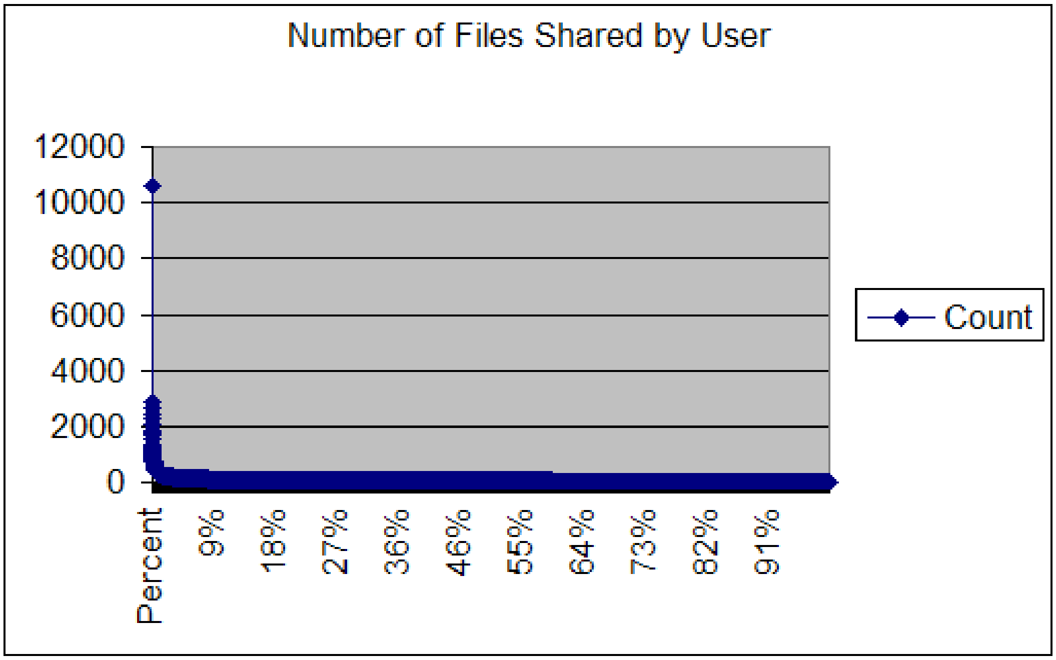
Source: How much information 2003

1. How users shared files

In addition to aggregate information, we looked at how many files each individual user had shared and how large their shares were. We discovered that the largest shared files on an individual user's hard disk was over 30 GB, while the smallest was just a couple of bytes. We discovered that 10 percent of the users accounted for about 60 percent of the total size of files shared, and 32 percent of the number of all files shared.



Source: How much information 2003



Source: How much information 2003

We found that only one user was sharing more than 10,000 files. The mean share size was between 300 and 400 files.

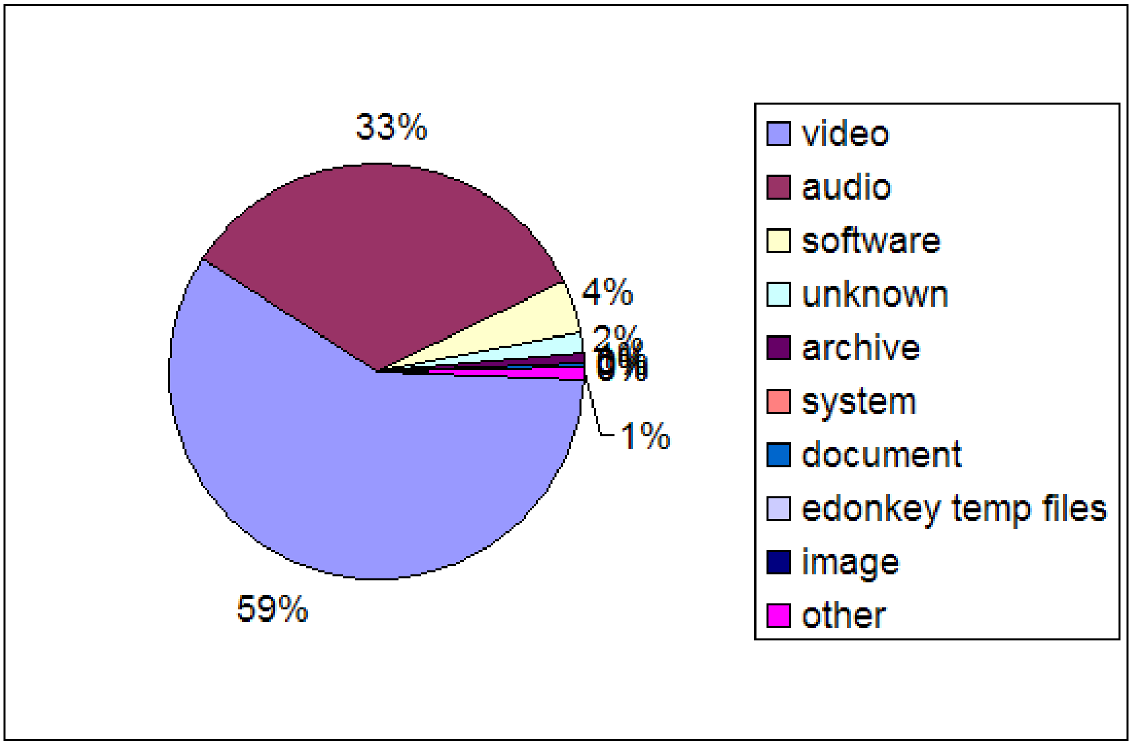
2. Summary of file size by type

We categorized the top 25 files into types and categories to get an idea of the distribution of file categories over our dataset. We found that video and audio data together accounted for about 90 percent of the total size of all files being shared.

Table 8.9: Sum of size by type

Type	Sum of Size (GB)	Number of Files	% of total size
video	8,661.60	126,217	58.70%
audio	4,929.43	1,253,308	33.41%
software	648.28	85,072	4.39%
unknown	248.37	13,491	1.68%
archive	95.91	548	0.65%
system	21.09	9,379	0.14%
document	13.22	2,623	0.09%
'.part' temp files	5.90	29	0.04%
image	5.79	138,751	0.04%
other	124.94	351,008	0.85%

Source: How much information 2003



Source: How much information 2003

3. What file types are largest?

The largest file types are .AVI video files, followed by archival .ZIP files. AVI files are video files playable on a computer. The range of these in our sample is 82 bytes to 2GB, with most being in the 100-200 MB range. Pornography seems to be a major contributor to this traffic, according to user identification of genre types.

Table 8.10: Largest file types shared on P2P

File Extension	Type of File	Largest File Size (GB)	Total Size (in GB)
avi	Video	1.97	4,851
zip	Archive	1.77	63
mp2	Video/Audio	1.69	3
mpg	Video	1.48	2,855
wav	Audio	1.15	243
vob	Video	1.00	32
mpeg	Video	0.88	311
wmv	Video	0.81	297
ncd	Document	0.80	4

Source: How much information 2003

Note on Large Files (over 100 MB): We found 24,947 files of over 80 different file types that were larger than 100 MB in the sample. Although they accounted for only 1 percent of the total number of files in the dataset, they accounted for almost 50% of the size of the collection studied.

4. What file types are most common?

The most common files shared by P2P users are MP3 files, music files encoded using MP3 technology. Images (jpg, bmp) are also popular but take up much less space. The kpl files are KaZaA playlist files. Sixty percent of the files on users' hard disks were MP3 files, taking up about 30 percent of the space.

Table 8.11: Most common file types

Ext	GB	Avg of Size	Min of Size	Max of Size	Count of KaZaASharesExt	% of Total
mp3	4,564.79	4,120,597	1	5.58E+08	1,189,488	60.06%
kpl	0.36	1,942.186	20	44,789,209	196,797	9.94%
jpg	5.79	44,829.72	75	29,016,816	138,751	7.01%
exe	550.76	7,293,783	1	8.16E+08	81,079	4.09%
mpg	2,855.88	59,855,823	6	1.6E+09	51,231	2.59%
wma	113.79	2,732,813	111	7.64E+08	44,707	2.26%
avi	4,851.12	1.73E+08	82	2.12E+09	30,026	1.52%
wav	243.44	15,177,476	44	1.24E+09	17,222	0.87%
bmp	2.52	160,220.4	54	1.34E+08	16,859	0.85%
mpeg	311.73	20,693,836	162	9.51E+08	16,175	0.82%
wmv	297.00	23,311,552	1,316	8.75E+08	13,680	0.69%
none	248.37	19,767,776	1	1.67E+09	13,491	0.68%
gif	0.13	15,275.78	34	1,598,776	9,394	0.47%
cab	21.09	2,414,563	19	6.97E+08	9,379	0.47%
ini	0.02	1,797.959	1	1,015,477	9196	0.46%
asf	296.10	34,909,551	13	5.83E+08	9135	0.46%
db	3.11	398,867.7	178	85,132,446	8,369	0.42%
dll	1.78	261,220.8	18	40,111,104	7318	0.37%
mov	48.88	8,790,495	162	3.82E+08	5,970	0.30%
txt	0.27	51,415.17	1	1.61E+08	5,553	0.28%
doc	0.49	125,125.3	28	11,816,199	4,189	0.21%
lnk	0.01	3,564.337	104	3,468,472	3661	0.18%
htm	0.06	20,782.51	4	18,030,044	3,226	0.16%
zip	63.23	21,470,521	1	1.91E+09	3,162	0.16%
pdf	13.22	5,411,649	162	1.34E+08	2,623	0.13%

Source: How much information 2003

Table 8.12: Top 10 most common file types shared on P2P

File Extension	Type of File	Total # of files
mp3	Audio	1,189,488
kpl	Kazaa playlists	196,797
jpg	Image	138,751
exe	---	81,079
mpg	Video	51,231
wma	Executable	44,707
avi	MP3	30,026
wav	Video	17,222
bmp	Image	16,859
mpeg	Video	16,175
	TOTAL	**1,782,335**

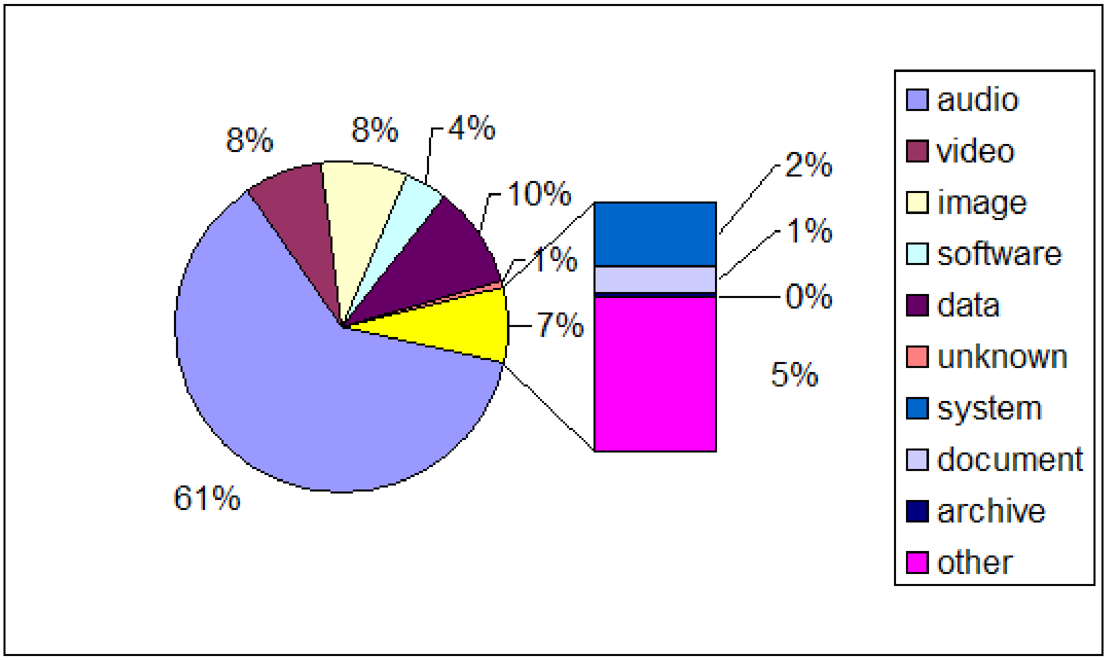
Source: How much information 2003

Audio files had the largest number of file (~62 percent) by far, followed by data files (~10 percent), images and video (~8 percent).

Table 8.13: Number of files by type

Type	Number	Percent
audio	1,220,390	61.6%
data	196,797	9.9%
image	165,004	8.3%
video	157,244	7.9%
other	89,745	4.5%
software	81,079	4.1%
system	37,923	1.9%
document	15,591	0.8%
unknown	13,491	0.7%
archive	3,162	0.2%

Source: How much information 2003



Source: How much information 2003

B. P2P consumption patterns

To look at the consumption of files we analyzed how many 'distinct' files are on the P2P network. By ‘distinct' we mean the number of files with the same content, although the same file content often is given many different names. Each file has a hash code associated with it, that associates it with other identical files. To find the top 25 most common files, as defined by their content not their names, we divided the number of unique or distinct file hash codes by the total number of files found, to come up with a number of unique or distinct files being shared by file type.

Distinct Files/ Total Files = Total Distinct Files

<

Table 8.14: Total distinct files, by extension and type

Extension	Total Files	Distinct Files	Total Distinct	Type
mp3	1,139,302	752,150	66%	audio
kpl	185,614	37,722	20%	audio
jpg	132,398	65,582	50%	image
exe	77,890	18,423	24%	application
mpg	49,147	21,807	44%	video
wma	42,649	26,251	62%	audio
avi	28,785	14,514	50%	video
wav	16,662	12,643	76%	audio
bmp	16,429	7,177	44%	image
mpeg	15,365	7,183	47%	video

none	12,925	6,461	50%	unknown
wmv	12,920	6,156	48%	audio
gif	9,146	5,735	63%	image
ini	8,840	5,516	62%	system
cab	8,814	1,210	14%	system
asf	8,608	3,209	37%	video
db	7,881	7,782	99%	system
dll	7,073	4,114	58%	system
mov	5,845	2,920	50%	video
txt	5,351	3,877	72%	document
doc	3,939	3,146	80%	document
lnk	3,516	3,181	90%	system
htm	3,168	2,280	72%	document
zip	3,071	2,647	86%	document
pdf	2,490	1,917	77%	document

Source: How much information 2003

Table 8.15 shows the top twenty music and movie genres, as identified by the users, rank ordered by size of files; percent measures proportion of genre files in the dataset.

Table 8.15: Top 20 genres, by file size

Genre	Size (GB)	Count	% of Total Count
Erotica	1,516.89	42,273	2.1%
Comedy	1,002.20	22,345	1.1%
Action and Adventure	814.25	3,137	0.2%
Other	798.20	214,391	10.8%
Rock	535.52	133,749	6.8%
Rap	462.09	109,719	5.5%
Music and Musicals	450.25	13,498	0.7%
Science Fiction and Fantasy	332.16	1,427	0.1%
Series	303.61	2874	0.1%
Pop	294.00	89,399	4.5%
Drama	285.12	951	0.0%

Kids and Family	227.59	1,013	0.1%
R&B	200.75	46,714	2.4%
Horror and Suspense	192.77	932	0.0%
Hip-Hop	156.34	51,429	2.6%
Country	153.29	43,266	2.2%
Games	147.29	2,703	0.1%
Anime	143.94	1,409	0.1%
Blues	129.99	34,473	1.7%

Source: How much information 2003

VII. FTP Sites

FTP is the abbreviation for File Transfer Protocol. This protocol is used on the Internet for downloading and uploading files to a server.

To get an idea of the scale of the FTP universe we consider anonymous FTP sites. Filewatcher.org reports that as of August 2003, there are 5,700 anonymous FTP sites containing 269,177,202 files, for a total size of **97.04 TB**.

VIII. Usenet

According to Dictionary.com, Usenet (from User's Network) is a distributed bulletin board system that serves millions of readers worldwide. "Originally implemented in 1979 - 1980 by Steve Bellovin, Jim Ellis, Tom Truscott, and Steve Daniel at Duke University, and supported mainly by Unix machines, [Usenet] swiftly grew to become international in scope and, before the advent of the World-Wide Web, probably the largest decentralised information utility in existence." Many newsgroups on Usenet are devoted to sharing files, including software, music, and images.

"As of early 1993, Usenet hosted over 1,200 newsgroups and an average of 40 megabytes (the equivalent of several thousand paper pages) of new technical articles, news, discussion, chatter, and flamage every day. By November 1999, the number of groups had grown to over 37,000."

As of August 2003, approximately **280 GB** of data are posted to newsgroups on Usenet every day, resulting in a total annual flow of **102 terabytes**. According to James Robertson and Emil Sit of MIT, the size of this flow has been doubling every 10 months.

REFERENCES

World Wide Web

- Bergman, Michael K. "The Deep Web: Surfacing Hidden Value" (BrightPlanet White Paper) http://www.brightplanet.com/technology/deepweb.asp

- CyberAtlas, citing the CIA World Factbook http://cyberatlas.internet.com/big_picture/geographics/article/0,1323,5911_151151,00.html

Blogs (web logs)

- Blogcount http://dijest.com/bc/
- Blogging by the Numbers http://cyberatlas.internet.com/big_picture/applications/article/0,,1301_2238831,00.html
- Day Pop, Top Web Logs http://www.daypop.com/blogrank/
- Spike Hall, http://radio.weblogs.com/0106698/2002/10/26.html
- LISTSERV Statistics http://www.lsoft.com/news/statistics.asp
- NITLE Blog Census http://blogcensus.net./
- Pointblog http://www.pointblog.com/past/000148.htm
- USA Today article, July 8, 2003 “Welcome to the Blogosphere” http://www.usatoday.com/tech/webguide/internetlife/2003-07-08-blogs_x.htm
- Jill Walker http://huminf.uib.no/~jill/archives/blog_theorising/final_version_of_weblog_definition.html

Email and spam

- Brightmail spam statistics http://www.brightmail.com/spamstats.html
- Brightmail statistics http://news.com.com/2100-1001-955842.html
- Anti-spam technologies http://research.microsoft.com/displayArticle.aspx?id=414
- IDC Report on email http://emailuniverse.com/list-news/2002/10/01.html
- NFO WorldGroup Survey, http://one.ie/report/email/marketoverview.asp
- Pew Internet and American Life Project, “Email at work,” http://www.pewinternet.org/reports/reports.asp?Report=79&Section=ReportLevel1&Field=Level1ID&ID=346
- Forrester Research cited in http://one.ie/report/email/marketoverview.asp
- TechWeb, Dec. 12, 2002, “Spam may overtake email in 2003.” http://www.cnn.com/2002/TECH/biztech/12/12/techweb.email.swamp/index.html
- Newsweek, Aug 9, 1999, p.51, "Sending AOL a Message."
- Office of Senator Conrad Burns, Monthly Report, April 2003, http://www.asme.org/gric/Fellows/Reports/PrabhatH_04-03.html
- CNET, News.com, July1, 2003, “A call for worldwide action on spam,” http://news.com.com/2100-1028-1022552.html

Instant Messaging and Chat

- Expresso Instant Messaging White Paper http://www.virtualthere.com/expresso/ExpressoRevenueModels.pdf

- IEEE Spectrum Online, June 12, 2003, "AOL Programmer Lays Waste to Employer," http://www.spectrum.ieee.org/WEBONLY/wonews/jun03/waste.html

- Instant Messaging, Part I: Corporate Productivity Tool or Cool Toy? http://www.intranetjournal.com/articles/200305/ij_05_01_03a.html

- Instant Messaging Planet http://www.instantmessagingplanet.com/

- InternetNews.com, Jun 17, 2002, http://www.internetnews.com/stats/article.php/1366931

- Wired News, Jan. 9, 2001, "No Whiners at AOL," http://www.wired.com/news/technology/0,1282,40973,00.html

FTP

- Filewatcher.org http://filewatcher.org/ftp-list/anonymous_ftp_sites_list.html

Usenet

- ComputerWeekly.com, June 26, 2001, "Usenet - a breeding ground for viruses," http://www.computerweekly.com/Article103439.htm

- Dictionary.com, Definition of Usenet, http://dictionary.reference.com/search?q=usenet

- James Robertson and Emil Sit, "UsenetDHT: Using DHTs for storage in Usenet," MIT Laboratory for Computer Science, 1 August 2003 project-iris.net/isw-2003/papers/robertson.ps

- Statistics for Article Size of Incoming [Usenet Messages] http://newsfeed.mesh.ad.jp/flow/size.html

Release date: October 27, 2003.