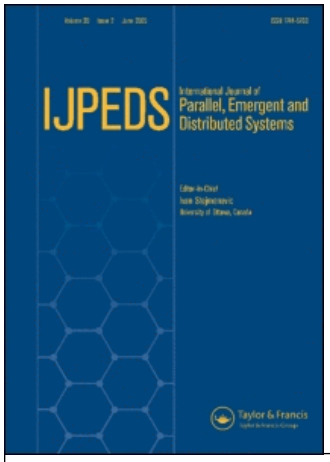

International Journal of Parallel, Emergent and Distributed Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713729127

On the session lifetime distribution of Gnutella

Kevin Ho [a]; Jie Wu [b]; John Sum [c]
[a] †Department of Computer Science and Communication Engineering, Providence University, Taichung, Taiwan, People's Republic of China
[b] Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, Florida, USA
[c] Department of Information Management, Chung Shan Medical University, Taichung, Taiwan

Online Publication Date: 01 February 2008

To cite this Article: Ho, Kevin, Wu, Jie and Sum, John (2008) 'On the session lifetime distribution of Gnutella', International Journal of Parallel, Emergent and Distributed Systems, 23:1, 1 — 15

To link to this article: DOI: 10.1080/17445760701324838
URL: http://dx.doi.org/10.1080/17445760701324838

The International Journal of Parallel, Emergent and Distributed Systems,
Vol. 23, No. 1, February 2008, 1–15

Taylor & Francis
Taylor & Francis Group

On the session lifetime distribution of Gnutella

KEVIN HO†, JIE WU‡ and JOHN SUM¶*

†Department of Computer Science and Communication Engineering, Providence University, Sha-Lu, Taichung, Taiwan, People's Republic of China
‡Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, Florida 33431, USA
¶Department of Information Management, Chung Shan Medical University, Taichung 402, Taiwan

(Received 9 November 2005; revised 4 October 2006; in final form 10 January 2007)

Over the past few years, capturing the characteristics of the Gnutella overlay network has been one of the major research activities in distributed computing. While most of the works are focusing on the node degree distribution, upstream/downstream traffics and file distribution, only a few have been done on modeling the session lifetime distribution. A good model of the session lifetime distribution is basically a key to understand the formation and evolution of the topological structure of Gnutella. In this paper, the main objective is to present a measurement result we did in 2003 on the session lifetime distribution and make comparison with the results obtained by other researchers during 2001–2005. In accordance with our measurement, it is found that the lifetime distribution fits to a power law distribution with exponential cut-off. Specifically, for all $t \geq 15$ min, $P_l(t) \propto (t/780)^{-1.12}\exp(-t/780)$. By inspecting the slopes of the log–log curve (i.e. $\partial \log P_l(t)/\partial \log t$) at 20 min (i.e. short lifetime) and 600 min (i.e. long lifetime), it is also found that the shape of our model matches to those curves obtained by other researchers based on the measurements in 2001, 2002 and 2004. Therefore, it is observed that the session lifetime distribution of Gnutella might be an invariant characteristic independent of the technological advancement or protocol change during the period 2002–2005.

Keywords: Gnutella; Peer-to-peer; Session lifetime distribution; Protocol

1. Introduction

Understanding the underlying structure of a network is not a new problem. Since the discovery that a number of natural networks have the small-world phenomena [16], i.e. the node degree distribution follows power law, a number of measurement studies on various different kinds of networks have been conducted. The Faloutsos brothers have found that our Internet is a small world network [5]. Albert *et al.* [1] found that the World Wide Web is a small world network. Since small world networks are claimed to be natural networks, a number of measurement studies regarding the network topology have thus been done in the past few years.

*Corresponding author. Email: pfsum@yahoo.com.hk

The International Journal of Parallel, Emergent and Distributed Systems
ISSN 1744-5760 print/ISSN 1744-5779 online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/17445760701324838

Gnutella is a peer-to-peer overlay network for file sharing. In terms of number of users, Gnutella is the third largest peer-to-peer network after eDonkey and FastTrack†. In spite of an interest in the power law topology, Jovanovic did a crawl on a subset of Gnutella peers and counted their node degrees. In Ref. [6], he reported that the node degree distribution of Gnutella is basically similar to a power law distribution. He then attempted to re-construct the network topology of Gnutella based on the neighbor relationships amongst the peers' IPs being crawled. As Gnutella is a dynamic network, it is hardly to validate how matches the Gnutella at any instance. Ripeanu *et al.* [8], by comparing their measurement statistics obtained in November 2000 and May 2001, found that the node degree distribution of Gnutella was changing over the period of measures. It changed from a power law distribution to a kind of mixture between power law and Poisson. Since Gnutella protocol v6.0 has been released, Sum *et al.* [11] did a crawl in 2003 and found that the node degree distribution of the ultrapeer–ultrapeer subnet follows a two power-laws distribution.

Owing to its popularity and the consumption large amounts of network resources, various measurement studies have been carried out in the last few years aiming at improving the file sharing capabilities as well as better monitoring the network traffic generated by the peers [3,9,10].

Saroiu *et al.* [9] and Sen and Wang [10] independently did measurements in 2001 capturing various characteristics of Gnutella. Using a comprehensive understanding on the extra traffic being generated by the Napster and Gnutella, their numbers of file shared, upload/download statistics and how often a peer connect and disconnected from the system, Sarious *et al.* did an active measure by periodically (in a 7 min interval) crawling each system to collect snapshots. They found that there is significant heterogeneity and lack of cooperation across peers participating in these systems. Ignoring the sessions with uptime less than 7 min‡, they have found that most sessions are of short uptime and the median session duration is approximately 60 min. Also as many as 25% of peers in which has no file shared (the so-called free riders). They simply connect to the Gnutella download files and then disconnect.

On the other hand, Sen and Wang [10] took a passive approach by analyzing 800 million *flow*¶ records collected at multiple routers across the ISP's network for Direct-Connect, FastTrack and Gnutella overlay networks over a period of 3 months. They have observed less than 10% of the IPs contributes around 99% of the total traffic volume. Also, many traffic characteristics such as traffic volume and connection time per unique source/destination IP pair are extremely skewed. Only a small fraction of peers are persistent over long time periods. As no actual figures or tables of results related to Gnutella have been presented, it is hard to cross validate the claim they have made with the results obtained in Ref. [9].

Chu *et al.* [3] presented a study (based on active measurement conducted over 3 months in 2002) of the popularity of files stored and transferred among peers in Napster and Gnutella indicated that the popularity of files is skewed. Furthermore, they have also found that distribution of session uptime follows a log-quadratic function. This observation has also been found in a recent measurement studied by Stutzbach and Rejaie [13] who did an active

†It is based on the August 31, 2005 statistics in Slyck.com.

‡It corresponds to the un-measurable portion.

¶In accordance with Sen and Wang, a flow is defined to be an unidirectional sequence of packets between a particular source and destination IP address pair.

probing measurement for a large number of IPs with 7 min probing cycle on October 2004 over a period of 15 h.

Although other measurement results have been reported, such as using a fast probing system [14,15] that can quickly capture millions of IPs within 7 min and on the bootstrapping time behavior [2,7], not much work has been done on the change of the session lifetime distributions over a long period of time (in terms of years). Session lifetime is of paramount importance for the understanding of the topological evolution of Gnutella, such as the node degree distribution and the network connectivity. As mentioned in Ref. [14], all these factors will as a result affect the overlay structure, the resiliency of the overlay and the simulation of such a network. Eventually, it will affect the selection of design parameters for a P2P network. In this regards, the primary objective of this paper is to present a measurement result we did in 2003 on the session lifetime of ultrapeers and give a comparative discussion on this parametric model with the models obtained by Chu *et al.* [3] and by Stutzbach and Rejaie [13]. In this paper, session lifetime and session uptime are used interchangeably. They correspond to the duration a peer is connecting to Gnutella. This is also called node availability in some other papers.

The remainder of the paper will be organized as follows. In the next section, we will give an overview on the connection mechanism of Gnutella. Then in Section 3, the methodology of measurement and results on node degree distribution and session lifetime distribution will be elucidated§. The estimated model is presented in Section 4. A discussion of the results compared with others will be presented in Section 5. The conclusion of the paper will be presented in Section 6.

2. Peer connection mechanism (bootstrapping)

In accordance with Gnutella protocol [4], while a new peer (Peer A) is trying to make connection to the existing Gnutella network, Peer A can either send a XTRY or PING message to an online Gnutella peer (Peer B) asking for the IP addresses (IPs) of its neighbors (figure 1). If XTRY is used, Peer B returns with a list of its direct neighbors. If PING is used, Peer B will broadcast the request to its direct neighbors. Its direct neighbors will then broadcast the request to their direct neighbors again and again until the message has been propagated time-to-live (TTL) hops away.

All their IPs will then be propagated backward to Peer B based on the PONG protocol. After receiving these PONG messages, Peer B will again send a PONG message to Peer A. Whether XTRY or PING–PONG protocol is used, Peer A can select a number of IPs from the list and make connections.

There are several remarks for this connection mechanism. The first remark is about the *seed IP*—where a peer can get the online peers' IPs. Once a peer, say Peer A, wants to connect to Gnutella, it needs to send a message to an online Gnutella peer. Basically, the software developer will set two default locations for a peer to find online peers' IPs. The first location is the local cache, which stores all the IPs a peer previously connected to. The second location is a well-known global server like limewire.com. It stores the IPs of all the current online peer nodes. Whenever a peer wants to connect, it can get the IPs from these locations and select one IP address as the seed for XTRY or PING.

§Throughout the paper, lifetime and session lifetime are used interchangeably.

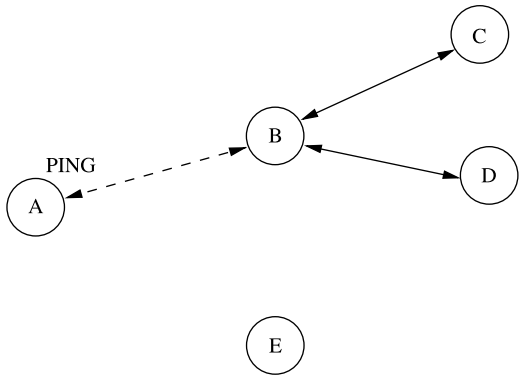

Figure 1. Peer connection.

The second is about *the number of seed nodes*. Suppose there are two available seed nodes, say *X* and *Y*, found in the local cache. In the current setting, Peer A is unlikely to ask Peer Y for further information if Peer X has been asked and replied with a list of IPs. Therefore, we can assume only one seed node will be XTRYed whenever a peer node wants to make a connection to Gnutella. This assumption has also been strengthened by a recent paper on simulated Gnutella [12]. Whenever a peer connects to two or more seed nodes, the shape of the node degree distribution in log−log scale should single modal.

The third remark is about *the connectedness of Gnutella*. If the network is not connected, with a number of isolated components, the availability of information from other components will definitely be hidden from Peer B. Obviously, Gnutella is a highly dynamic network. Peer nodes can go offline whenever they finish downloading files or encounter software or hardware failure. When a large number of peer nodes go offline, the connectedness will no longer be guaranteed. Therefore, once the network has been disconnected, asking information from only one seed node cannot help in recovering the network connectedness.

3. Measurement method and results

The measurement was carried out in August 2003. The measurements of the node degree and the node lifetime are carried out in two consecutive phases.

3.1 Ultrapeers collection

First, a list of IPs is obtained from limewire.com. Starting from this list of IPs, we send out two messages: *Gnutella Ultrapeer: True*. Once a *Gnutella 200 OK* message has been received, we immediately send out a *PING message* to craw its neighbor IPs. Then a *Gnutella 200 OK* message is sent to these neighbor IPs. We repeat this process (crawl) until some 100,000 IPs have been collected as a master list.

3.2 Handshaking

As a Gnutella peer can define itself as either a blocked ultrapeer or non-blocked ultrapeer, node degree measurement is a bit complicated. To facilitate fast capturing, 20 threads are out simultaneously for handshaking. As the IPs collected in the last step might be outdated

in the database, our program only threads out to those IPs that have no more than 60 min of log time.

For each thread: (i) An IP address is randomly selected from the database and a bootstrapping message with *Ultrapeer: True* is sent to it. (ii) If the node sends back a message with *Gnutella 200 OK*, it confirms that this is a successful connection. Otherwise, we assume that the node is either inactive or acting as a block node. (iii) If the node is non-blocked, a message PING TTL 2 is sent to that node asking for neighbor IPs. (iv) If the node is a blocked node, a new IP is selected randomly from the remaining IPs and the same message *Ultrapeer: True* is sent to it. The received PONG replies are stored in a database.

By analysis on the data obtained, we can then identify whether a node is a leaf, a blocked ultrapeer, or a non-blocked ultrapeer. To maintain uniformity, our study only focuses on the ultrapeer–ultrapeer node degree distribution and eventually 5087 non-blocked ultrapeer nodes are captured and identified. Their node degree distribution is shown in figure 2. Observed from the curve, the distribution looks like a mixture of two power-law distributions:

$$P_n(k) \propto \begin{cases} k^{-0.8} & k \leq 5 \\ k^{-4} & k > 5 \end{cases} \tag{1}$$

where k is the node degree and $P_n(k)$ is the probability of a node with degree k.

3.3 Online session time measurement

The methodology of collecting session time information is similar to the way we explore online peers. About 7500 sample nodes are randomly selected from the database. Another database of 7500 records with $8 \times 24 \times 4$ fields in each record.

Every 15 min these IPs are tested by sending bootstrap messages to them. If a node replies with *Gnutella 200 OK* or *Gnutella 503*, it is defined as active and a "1" will be saved in the corresponding record and the corresponding time field. If there is no reply from the node a "0" will be written. Figure 3 shows a few typical cases that are captured from the measurement. Let IP_1, IP_2 and IP_3 be the IPs to be measured. In IP_1, two sessions A and B are

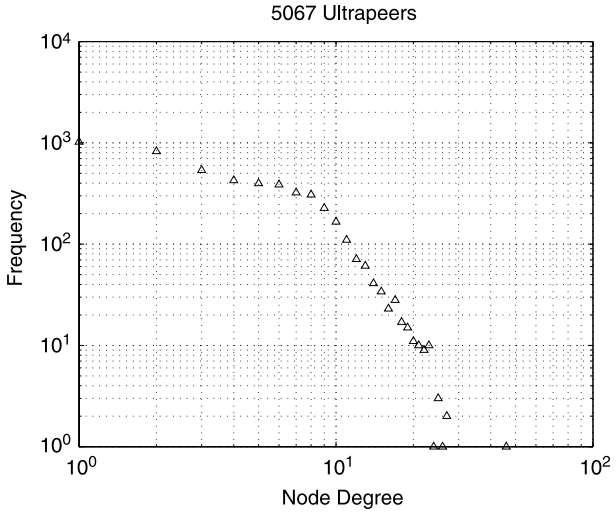

Figure 2. Node degree distribution of the ultrapeer–ultrapeer subnet of Gnutella.

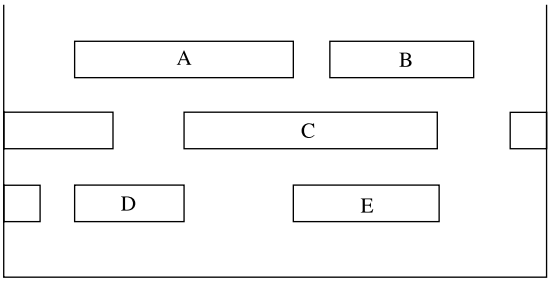

Figure 3. Typical cases that are measured.

captured. In IP_2, three sessions have been captured during the measurement duration. As the session measured in the beginning of measurement has already been alive long before the measurement started, it is excluded from our statistics accounting for the lifetime distribution. Similarly, the session that is still alive when the measurement has been ended will be excluded from our statistics. Therefore, only session C will be considered as a valid candidate. Similarly, only session D and E will be considered for lifetime distribution. After excluding all the invalid sessions, there is another problem that we need to decide. Should sessions A and B be considered as a single session or two? To give an answer to this problem, we take the following approach.

After the program has ran for 8 days (768 h), 7500 records similar to the following line were obtained.

$$1100111111011110000000000111111110000\ldots$$

The first "1" corresponds to the first measurement. Since we have no information about when it starts, we ignore the first two ticks. In the third and fourth ticks, there is no reply. We assume that during that period of time, the IP has not been assigned with an active Gnutella peer. So, the fifth tick will be the starting uptime of a peer. But there is one question left: When is the end of a session?

Like the above example, $\ldots 0011111101111000\ldots$, we assume "1 0 0" indicates the end of a session. Whenever an IP does not reply for two or more consecutive ticks (i.e. ≥ 30 min), it is assumed to be offline. Later, if the IP is found to be active again, it is treated as a new peer.

So, in this example, we assume the IPs have been occupied two times by two different Gnutella peers. Their uptimes are 11×15 and 8×15 min, respectively. Another reason why we measure at every 15 min is because we assume that active uptime of an ultrapeer should be long, in order of hours.

To confirm our assumption, the number of sessions counted for the following four cases has been shown in figure 4.

(i) Assume the session is off if no reply has been received for one time, i.e. whenever "1 0" has encountered.
(ii) Assume the session is off if no reply has been received for two times, i.e. whenever "1 0 0" has encountered.
(iii) Assume the session is off if no reply has been received for three times, i.e. whenever "1 0 0 0" has encountered.
(iv) Assume the session is off if no reply has been received for four times, i.e. whenever "1 0 0 0 0" has encountered.

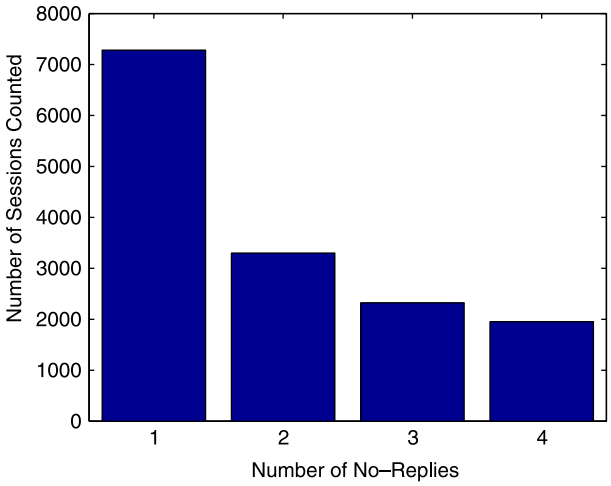

Figure 4. Number of sessions counted against the number of no-replies.

Using "1 0" as an end of a session, the short lifetime sessions will be over-counted.

The normalized lifetime distribution is shown in figures 5 and 6 shows the corresponding cumulative distribution. The distribution obtained for case 1 is a bit different. For case 1, more "short uptime" data was counted. Eventually, it makes the slope of the "log–log" plot in the "short uptime" region steeper. By observing the number of sessions being counted against the numbers of no-replies in figure 4, we assume that the number of sessions in the case "1 NR" is over-counted.

4. Estimated model

Therefore, the lifetime data corresponding to the case "2 NR" is used to model the lifetime distribution for ultrapeers. By inspecting the shape of the cumulative lifetime distribution in figure 6, a reasonable model for the distribution is power law with exponential cut off, defined as follows:

$$P_l(t;\alpha,\beta) \propto \left(\frac{t}{\beta}\right)^{-\alpha} \exp\left(\frac{-t}{\beta}\right). \tag{2}$$

To estimate the parameters of α and β, one cannot directly use least squares fit to the log–log curve of

$$\frac{P_l(kT;\alpha,\beta)}{\sum_k P_l(kT;\alpha,\beta)}$$

and the data in figure 4. Here T is the time interval between two consecutive measures. To be exact, we define normalized distribution $\bar{P}_l(kT)$ as follows:

$$\bar{p}_l(kT) = \frac{\int_0^T \int_0^T P(t_e|kT+t_s;\alpha,\beta)\mathrm{d}t_e P(t_s)\mathrm{d}t_s}{\sum_{k\geq 1}\int_0^T \int_0^T P(t_e|kT+ts)\mathrm{d}t_e P(t_s)\mathrm{d}t_s}, \tag{3}$$

Here t_s and t_e correspond to the unknown session starting offset and the session ending offset (figure 7). $P(t_e|kT+t_s;\alpha,\beta)$ is the conditional probability that the session will be alive for t_e extra time when it has already been alive for $kT+t_s$. $P(t_s)$ is probability that a new session

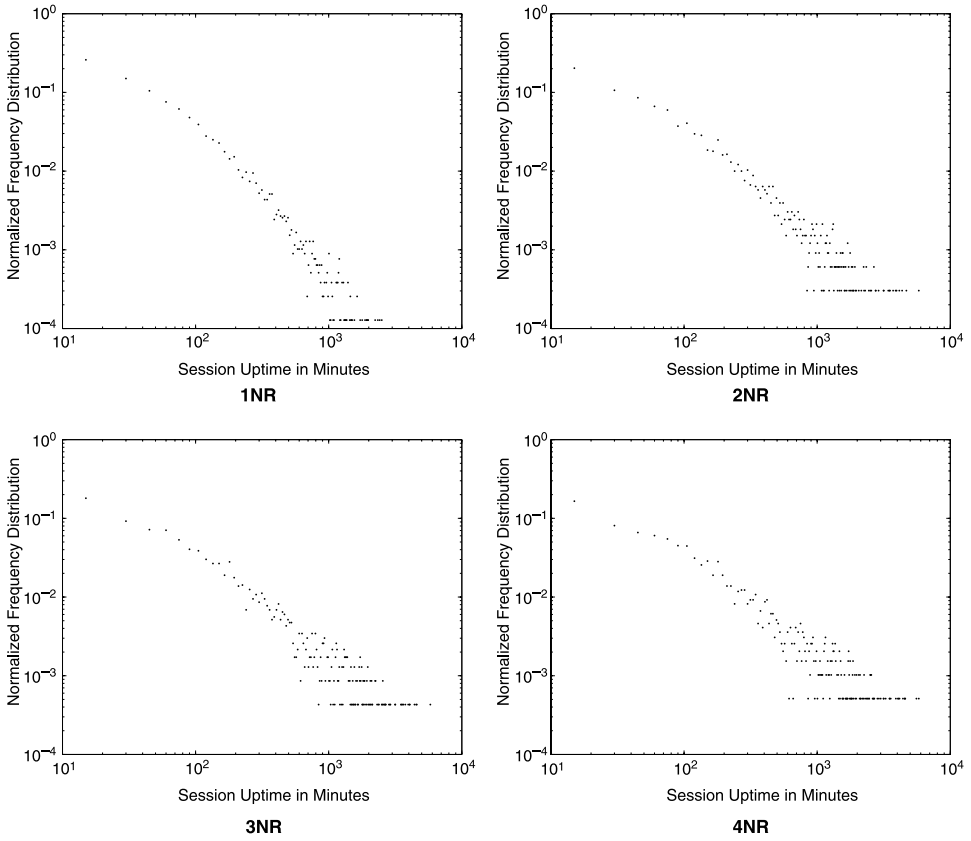

Figure 5. Lifetime distribution directly measured from Gnutella for 768 h. 1 NR: A node is assumed offline if no reply has been received for one trial. 2 NR: A node is assumed offline if no reply has been received for two consecutive trials. 3 NR: A node is assumed offline if no reply has been received for three consecutive trials. 4 NR: A node is assumed offline if no reply has been received for four consecutive trials.

starts at t_s before its first measurement. We assume that the arrival of a new peer is uniformly random. That is to say, $P(t_s) = 1/T$. For $P(t_e|kT + t_s)$, there is no simple close form that we can use. Therefore, the double integral in equation (3) is calculated numerically by the following equation:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} P_l(kT + i\delta + j\delta; \alpha, \beta),$$

where $m\delta = T$. By Law of large number, it can be proved that

$$\sum_{i=1}^{m}\sum_{j=1}^{m} P_l(kT + i\delta + j\delta; \alpha, \beta)(\delta)^2 \approx \int_0^T\int_0^T P(t_e|kT + t_s; \alpha, \beta)\mathrm{d}t_e P(t_s)\mathrm{d}t_s,$$

for small δ and for all $k \geq 1$. Thus, the corresponding cumulative distribution, denoted by $\bar{C}_l(kT; \alpha, \beta)$, can be given by

$$\bar{C}_l(kT; \alpha, \beta) = \frac{\sum_{g=1}^{k}\sum_{i=1}^{m}\sum_{j=1}^{m} P_l(gT + i\delta + j\delta; \alpha, \beta)}{\sum_{g\geq 1}\sum_{i=1}^{m}\sum_{j=1}^{m} P_l(gT + i\delta + j\delta; \alpha, \beta)}. \quad (4)$$

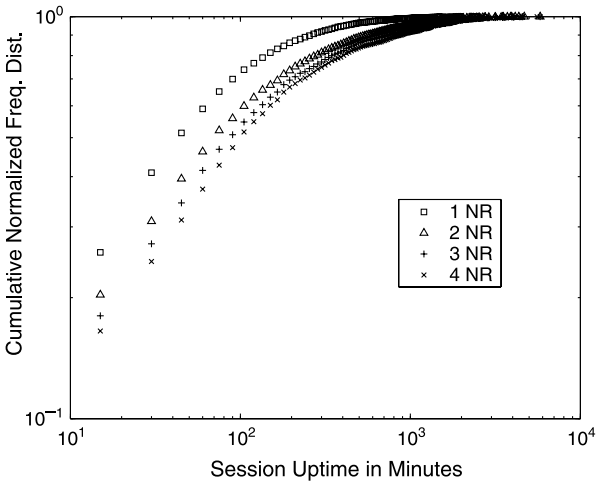

Figure 6. Cumulative lifetime distribution directly measured from Gnutella for 768 h. 1 NR: A node is assumed off-line if no reply has been received for one trial. 2 NR: A node is assumed off-line if no reply has been received for two consecutive trials. 3 NR: A node is assumed off-line if no reply has been received for three consecutive trials. 4 NR: A node is assumed off-line if no reply has been received for four consecutive trials.

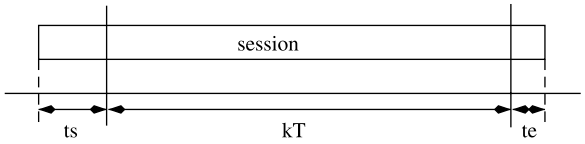

Figure 7. Definition of a session.

The best estimate of (α_0, β_0) will be defined as:

$$(\alpha_0, \beta_0) = \arg\min_{(\alpha,\beta)} \left\{ \sum_{k\geq 1} (\log \bar{C}_l(kT; \alpha, \beta) - \log C_l(kT))^2 \right\}, \tag{5}$$

where $\log C_l s$ are the values shown in figure 6.

To avoid costly computation on the gradient of $\log \bar{C}_l$, we set $m = 150$ and exhaustive search for‖

$$\alpha \in 1.00, 1.01, 1.02, \ldots, 1.50; \quad \beta \in 100, 110, 120, \ldots, 990.$$

Eventually, it is found that the best estimate for (α, β) is (1.12, 780). In other word, the session lifetime distribution is given by:

$$P_l(t; 1.12, 780) \propto \left(\frac{t}{780}\right)^{-1.12} \exp\left(-\left(\frac{\mathrm{t}}{780}\right)\right). \tag{6}$$

Figure 8 compares the curves of $\bar{P}_l(kT)$ and the normalized distribution with two NR. The comparison on their cumulative counterparts is shown in figure 9. It should be noted that the time scale in equation (6) is in minute. One can see that the value β is equal to 13 h. That means, most Gnutella peers are only alive for no more than 13 h (table 1).

‖The same results are obtained even when $m = 300$.

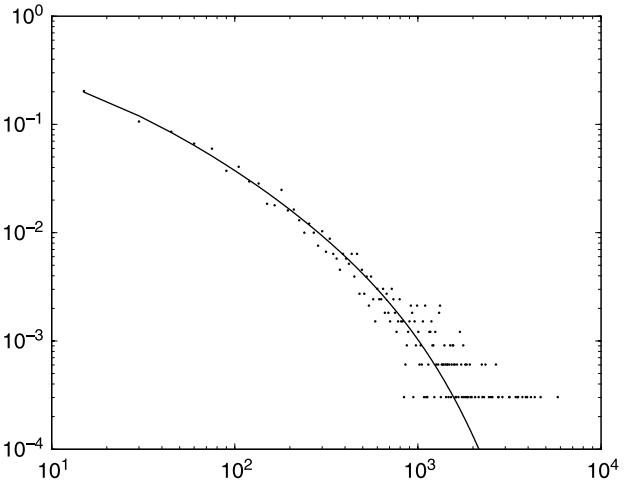

Figure 8. The normalized probability distribution of the estimated model versus the actual data measured. The solid line corresponds to the estimated model while the dots correspond the 2 NR data.

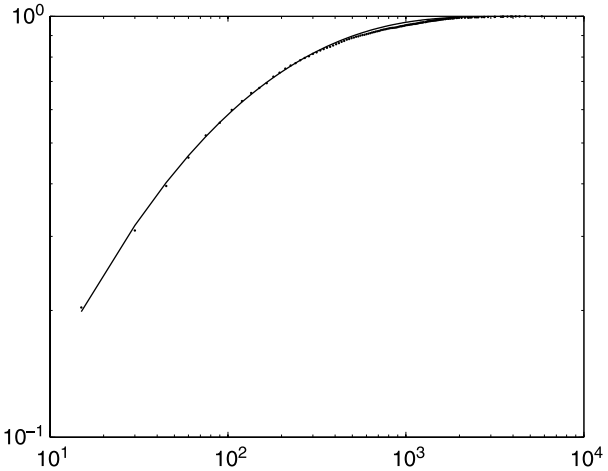

Figure 9. The cumulative distribution of the estimated model versus the actual data measured. The solid line corresponds to the estimated model while the dots correspond the 2 NR data.

5. Related works since 2001

As mentioned earlier in the paper, many measurement studies on Gnutella have been done in the last couple of years, with particular interests in node degree distribution, traffic volume, file distributions and node uptime. Only a few of them provided information about session lifetime. Table 2 summarizes some of these works since 2001.

Saroiu *et al.* [9], one of the earliest group of researchers who based their work on a create-based approach# captured the session uptimes of Gnutella nodes. In their measurement, it is found that 50% of the peers never remain online for more than 1 h and the median session

#The captured traces are divided into two halves. The reported durations are for the sessions that started in the first half, and finished in either the first or second half.

Table 1. Summary of measurement works in between 2001 and 2004, with session lifetime information.

Works	*MM/YY*	*T (mins)*	*Node size*	*Duration*
Saroiu *et al.* [9]	05/01	7	17125	2.5 days
Sen and Wang [10]	09/01	0.02	–	1 day
Chu *et al.* [3]	05/02	10	5000	3 days
This paper	08/03	15	7500	8 days
Cramer *et al.* [2]	xx/03†	10	665723‡	5 weeks
Wong [17]	11/03	10	7500¶	8 days
Stutzbach and Rejaie [13]	04/04–10/04	2, 7	1.3 M	72, 31 h
Stutzbach and Rejaie [14]	10/04–12/04	7	1.3 M	2 days

† In accordance with the year the paper has been included in CiteSeer.
‡ On Gwebcache nodes only.
¶ Since the set of probing nodes are initially crawled from Gwebcache, large proportion of the set is Gwebcache nodes.

Table 2. Summary of session lifetime models.

Works	$p(t)$	$\partial \log p(t)/\partial \log t$
Chu *et al.* [3]	$t^{-0.61}\exp(-0.07(\log t)^2)$	$-0.61 - 0.14\log t$
This paper	$t^{-1.12}\exp(-t/780)$	$-1.12 - t/780$
Wong [17]	$t^{-1.998}$	-1.998
Stutzbach and Rajaie [13]	$t^{-1.756}$	-1.756
Stutzbach and Rejaie [13]	$t^{-0.126}\exp(-0.149(\log t)^2)$	$-0.126 - 0.3\log t$

All logs are natural logarithms.

duration is approximately 60 min. As they claimed, it is similar to the time a user takes to download a small number of music files.

On the other hand, Sen and Wang [10] took a passive approach by collecting the *flow* information from Cisco's *Netflow* systems. For Fast Track, it is found that about 50% sessions are of lifetime less than 10 min. Although they have claimed that the result for Gnutella is similar, no actual figure has been reported. A drawback of their analysis is that the lifetime of each session is over-counted. As *NetFlow* collects only *Transport Layer* information, it is not possible to identify whether a *Flow* is generated after or during bootstrap. The connection duration (figure 8a in Ref. [10]) reported is essentially the subtotal of bootstrapping duration and session uptime.

In accordance with their measurement conducted in 2002, Chu *et al.* [3] is almost the first research group to suggest that session lifetime of a Gnutella peer is of log quadratic form. With reference to a measurement done in 2004, Stutzbach and Rejaie [13] re-confirm this claim. To accurate capture snapshots of Gnutella *et al.* developed a system called Cruiser which can capture a large number of peers within just a few minutes. By probing to 1.3 millions peers for 72 h with bin size 2 min**, they have identified that the session lifetime distribution of the top-level peers follows a power law distribution with exponent -1.756.

In another probing measure for 31 h with bin size 7 min, they have shown the session lifetime distribution of all peers can be characterized by either a 2-power-law distribution or a log-quadratic distribution. In Ref. [14], they have further illustrated that the shapes of the

**Bin size corresponds to the time interval between two consecutive probe.

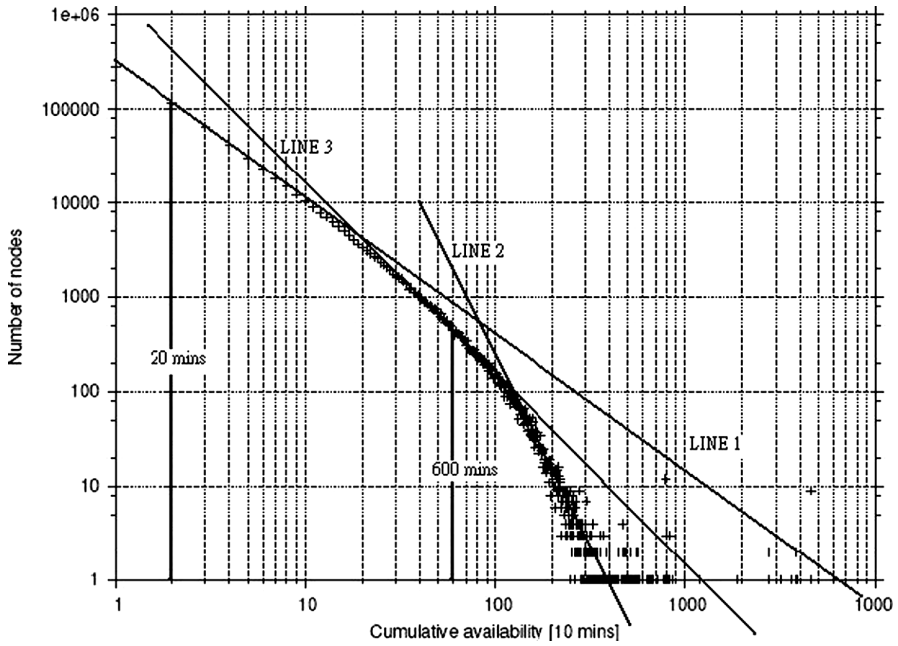

Figure 10. Gwebcache uptime distribution adapted from Cramer *et al.* [2]. Slope of LINE 1, LINE 2 and LINE 3 are approximately -1.445, -2.0 and -4 respectively.

session lifetime distribution captured at five different periods of time, from October to December of 2004, are not exactly the same. The exponents of the session lifetime in the range of (12, 1440 min) are between -1.30 and -168.

While studying the bootstrapping characteristics, Cramer *et al.* did a measure on the Gwebcache nodes [2] some time in 2003 to identify their actual availability. We reproduce the session lifetime distribution they obtained in figure 10 for reference. The straight lines (not from their original figure) are added to illustrate the slopes of the tangents at the points $t = 20$, 60 and 2000 min, respectively. Wong did a similar measure on the lifetime of nodes that were crawled from Gwebcache and its neighbors in the August of 2003 [17]. Both results show that the lifetime distribution is like a power law distribution with exponential cut-off and the exponent for lifetime in between $t = 200$, 2000 min is around 2. For the sake of clarification, the results obtained in the previous studies are summarized in tables 2 and 3.

Table 3. The exponents locally at t equals to 20 and 600 min.

Works	*MM/YY*	*T (mins)*	*20 (mins)*	*600 (mins)*
Chu *et al.* [3]	03/02–05/02	10	1.03	1.50
This paper	08/03	15	1.15	1.90
Cramer *e al.* [2]	xx/03	10	1.45[†]	2.0
Wong [17]	11/03	10	–	2.0
Stutzbach and Rejaie [13]	09/04	2	1.756	1.756
Stutzbach and Rejaie [13]	10/04	7	1.18	1.90
Stutzbach and Rejaie [14]	10/04	7	1.30, 1.34	1.30, 1.34
Stutzbach and Rejaie [14]	11/04	7	1.45	1.45
Stutzbach and Rejaie [14]	12/04	7	1.60, 1.68	1.60, 1.68

[†] The data is estimated by inspecting the curve provided by Cramer *et al.* in Ref. [2], see also figure 10 in this paper.

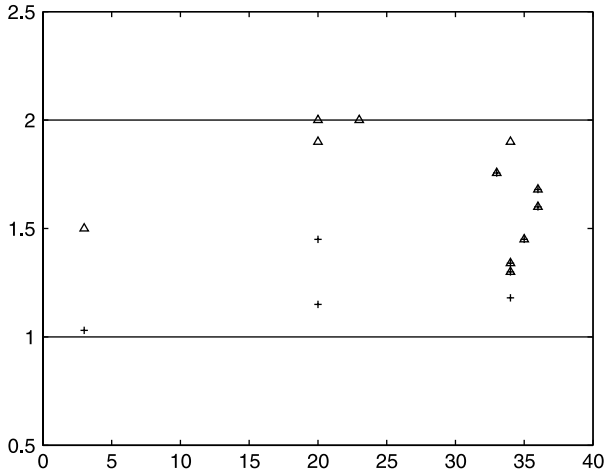

Figure 11. Exponents against the date the measurement conducted. Here January 2002 is set to be the first month. The " + " dots correspond to the values at $t = 20$ min while the triangular dots correspond to the values at $t = 600$ min.

Table 3 lists the exponents

$$\frac{\partial \log p(t)}{\partial \log t}$$

of the models at $t = 20$ min and $t = 600$ min. For Cramer *et al.*'s data, the values are estimated by inspecting (figure 10). One can observe that the model estimated in this paper (based on our measurement in August 2003) is rather consistent with the one obtained by Stutzbach and Rejaie in Ref. [13] (based on their measurement in October 2004).

Figure 11 shows the scatter plot of the exponents with respect to the date of the measurement. One can see that the exponents are within [1,2] for all $20 \text{ min} \leq t \leq 600$ min. Except the results obtained by Chu *et al.* in 2002 and by Cramer *et al.* in 2003 specifically for Gwebcache nodes, the remaining data fit well within the range. In accordance with Chu *et al.*'s results, the slope of the distribution is not that deep, while the slope in Cramer *et al.*'s distribution is deeper than others. We do not have a definite answer for this case but we believe it is due to the protocol difference and the slow Internet access rate at that time. For the latter case, we suspect that it is due to the node sampling problem.

Excluding the two special cases measured in 2002 and 2003, respectively, one could make a statement refined from Ref. [13]. Stutzbach and Rejaie have suggested that *the power law distribution of session times is independent of the application and the purpose of application. Rather it is determined by application-independent aspects of user-behavior.* Here, we add that *user-behavior in general is a time invariant characteristic.*

Before proceeding to the conclusion of the paper, a few notes should be commented about the mathematical models for the lifetime distribution.

- Measurement of very short lifetime session is notoriously difficult. Hence, the modeling of distribution head is still an open problem. Without such information, all measurement results reported so far cannot truly reflect the actual distribution, whether in terms of probability density (i.e. PDF) or cumulative distribution (CDF).

- One major obstacle in using either 2-power-laws distribution, log-quadratic distribution or power law with exponential cut off is that they are not integrable, i.e.

$$\int_{t_0}^{\infty} P_l(t)\mathrm{d}t$$

is undefined when $t_0 \rightarrow 0$. This leads to the conclusion that the analysis on the growth and the size of Gnutella impossible, unless a less fitting model like gamma distribution [11] is employed.

6. Conclusion

In this paper, a measurement result obtained in 2003 regarding the session lifetime of Gnutella peers has been presented. The measurement is based on the active probing method. The results on node degree distribution and session lifetime distribution have been shown. We found that the session lifetime statistics can fit to a power law distribution with exponential cut-off, $P_l(t) \propto (t/780)^{-1.12}\exp(-t/780)$ for all session lifetime $t \geq 15$ min. The result is similar to the model estimated by Stutzbach and Rejaie in a measurement done in 2004. To study the evolutionary change in session lifetime, a survey on work done in between year 2001 and 2004 on session lifetime distribution has been elucidated. Comparing our 2003 measurement result with the others in the literature, it is also found that session lifetime seems to be a time invariant characteristic. For sessions with relatively long lifetime (around 600 min), the exponent of the power law fit is around $-1.9 - -2.0$. For sessions with relatively short lifetime (around 20 min), the exponent is around 1.2. In view of the recent findings about session lifetime, we make the following observation.

After 2002, when Gnutella v0.6 has been released and broadband access has became the common Internet access medium, the session lifetime seems to be an application independent property and it looks like an invariant property of Gnutella, which is depended solely on the user behavior. And users behavior should also be time invariant.

One difficulty in simulating the topological formation of Gnutella due to the non-integrable property of power law or power law with exponential cutoff has also been remarked. It brings out that searching for a better distribution model should be a worthwhile direction for further investigation. Finally, one should take a note on the measurement results reported by Cramer *et al.* [2] and Wong [17] on Gwebcache lifetime distribution. In their measurement, it is found that the lifetime of Gwebcache is relatively short (the exponent is -2.0) compared with Gnutella peers in general. As Gwebcache is a technique to facilitate peer discovery, its unavailability can affect the performance of the overall setup. Further investigation along this line seems to be inevitable. Moreover, whether the results apply to other networks in the same category should also be explored in the future.

Acknowledgements

The authors would like to express their gratitude to the reviewers for their valuable comments.

References

[1] Albert, H. and Jeong, A.-L., 1999, Diameter of the World Wide Web, *Nature*, **401**, 130–131.

[2] Cramer, C., Kutzner, K. and Fuhrmann, T., 2004, Bootstrapping locality-aware P2P networks, *Proceedings of the IEEE International Conference on Networks*, ICON 2004, **1**, 357–361.

[3] Chu, J., Labonte, K. and Levin, B.N., 2002, Availability and locality measurements of peer-to-peer file systems. *IT COM: Scalability and Traffic Control in IP Networks II Conferences*.

[4] Clip2 Distributed Search Services, 2001, The Gnutella protocol specification v0.4, http://dss.clip2.com.

[5] Faloutsos, M., Faloutsos, P. and Faloutsos, C., 1999, On power-law relationships of the internet topology, *Computer Communication Review*, **29**, 251.

[6] Jovanovic, M.A., Modeling large-scale peer-to-peer networks and a case study of Gnutella, master thesis department of electrical and computer engineering and computer science, University of Cincinnati, April 2001.

[7] Karbhari, P., Ammar, M., Dhamdhere, A., Raj, H., Riley, G. and Zegura, E., 2004, Bootstrapping in Gnutella: a measurement study. *Proceedings of the Passive and Active Measurements Workshop* (France: Antibes Juan-les-Pins).

[8] Ripeanu, M., *et al.*, 2002, Mapping the Gnutella network, Jan/Feb 50–57, *IEEE Internet Computing*.

[9] Saroiu, S., Gummadi, P.K. and Gribble, S.D., 2002, A measurement study of peer-to-peer file sharing systems. *Proceedings of the Multimedia Computing and Networking (MMCN)* (CA: San Jose).

[10] Sen, S. and Wang, J., 2004, Analyzing peet-to-peer traffic across large networks, *IEEE/ACM Transactions on Networking*, **12**(2), 219–232.

[11] Sum, J., *et al.*, A Study of the Connectedness of Gnutella, PDPTA'04, Las Vegas, 2004.

[12] Sum, J. and Lau, S.C., A novel connection mechanism for P2P, PDPTA'04, Las Vegas, 2004.

[13] Stutzbach, D. and Rejaie, R., 2004, Towards a better understanding of churn in peer-to-peer networks, Technical Report CIS-TR-04-06, Department of Computer Science, University of Oregon, November 2004. Available at the URL www.barsoom.org/∼agthorr/blog/Publications.shtml.

[14] Stutzbach, D. and Rejaie, R., 2006, Characterizing churn in peer-to-peer network, to appear at. *Internet Measurement Conference'06*, October 25–27, Brazil. Available at the URL www.barsoom.org/∼-agthorr/blog/Publications.shtml.

[15] Stutzbach, D. and Rejaie, R., 2005, Charaterizing the two-tier Gnutella topology. *SIGMETRIC'05*.

[16] Watts, D., 1999, *Small Worlds* (NJ: Princeton University Press).

[17] Wong, M.F., Statistical Analysis of worm spreading and resilience study in P2P, Project Report, Department of Computing Hong Kong Polytechnic University, 2004.