Geographical Statistics and Characteristics of P2P Query Strings

Adam Shaked Gish
P2P Networking Research,
Skyrider, Inc., Israel
Email: adams@skyrider.com

Yuval Shavitt Tomer Tankel
School of Electrical Engineering,
Tel-Aviv University, Israel
Email: {shavitt,tankel}@eng.tau.ac.il

***Abstract*— P2P file-sharing applications are quickly being adopted by a wider and more mainstream audience. There is much to be learned from keyword searches users perform in order to retrieve content from these networks.**

This paper presents a large-scale measurement study of search terms in the modern Gnutella network. We developed a highly parallelized architecture capable of capturing an unprecedented amount of geographical-identified queries. We applied this architecture to generate daily logs of search queries. We collected over 25 such daily Gnutella logs, with more than 15 millions unique queries in each, over a three month period. We analyzed both static snapshots of the Gnutella networks, as well as the dynamics of the network over time. In particular, we look at the geographic location and the trends of searches to better understand the dynamics.

I. Introduction

P2P file sharing applications have been becoming increasingly popular as means for multimedia file sharing for several years now. Although P2P vendors and users have been facing legal pressure from copyright owners, it seems that the P2P community at large remains strong and healthy, with an ever growing numbers of avid users. Copyright owners seem to be accepting that P2P is here to stay, and are expected to start focusing on building business models that will allow them to generate revenue from P2P activity, rather than attempting to shut P2P down. For this to happen, the next generation of P2P protocols and applications will be developed, such that they will improve the relevancy of the end user's P2P experience. Users will be able to more easily find relevant content that matches their tastes, and advertisers will be able to expose them to relevant targeted advertising content. In order to improve advertising relevance, it is critical to analyze the behavior and preferences of P2P users, taking into account the user's geographic location and temporal behavior.

Most measurement studies on P2P content have focused either on analyzing ISP data traffic [1], [2] or the content in user's shared folders [3], [4], [5], [6]. The popularity of files [3], [5], [6] follows a Zipf distribution, that is power-law. A log-quadratic distribution, or a second-order Zipf distribution, was found by [3]. However, content in shared folder accumulates over time, and is actually an integration of shifting interests over an extended period of time. A much more valuable approach would be to analyze user queries as they propagate the network. This can provide strong insight into the current interests of P2P users. Previous measurement studies on query strings [7], [8] were limited in scale, both in the number of queries that were sampled and the in the short time of the data collection period. Moreover, [8] used a modified Mutella client to perform their research. This is not an optimal choice, as according to [9] and our own measurements the vast majority of the network is comprised of Limewire clients ($80-85\%$) and Bearshare clients ($6-10\%$). Leaf nodes of both these cliets have a strong preference to connect to their own kind. Thus, it is possible that a small minority of unpopular clients is severely overrepresented in their data.

In this paper we empirically characterize geographically-identified Gnutella queries captured by Skyrider systems[1]over a period of three and a half months. We start by explaining how it is at all possible to capture significant numbers of geographically-identified queries. Changing focus to analysis we begin by observing the weekly pattern in the quantities of queries intercepted concentrating on the coutries, which generated the majority of the queries. We continue by examining the breakdown of queries by countries, and then examine the diurnal pattern of user behavior in a few of these countries. Next we analyze the temporal behavior of queries over the logging period and classify two types of queries, ”constant” and ”volatile”. This is followed by a comparison of the top query strings in different countries, which allows us in a sense to develop a measure for determining the ”cultural similarity” of one country to another. Using these findings we demonstrate that the query's popularity rank, as shown in [7], [8], and its frequency exhibits a power-law relationship.

A. Measurement Goals and Methodology

Our goal was to capture large quantities of geographically-identified human generated queries. while it is possible to capture a large quantity of queries by deploying several hundred ultra peer nodes, it will not be possible to tell the origin of most of these captured queries. The basic problem in identifying the origin of captured queries is that queries do not in general carry information regarding their origin. What they do usually carry is an ”Out of Band” return IP address. This address allows clients that have content matching a query to respond to a location close to the origin of the query, without having to backtrack the path taken by the query message. However, as most queries come from firewalled clients, in most cases the out of band address will belong to the ultra peer connected to the query origin, acting as a proxy on behalf of the query originator.

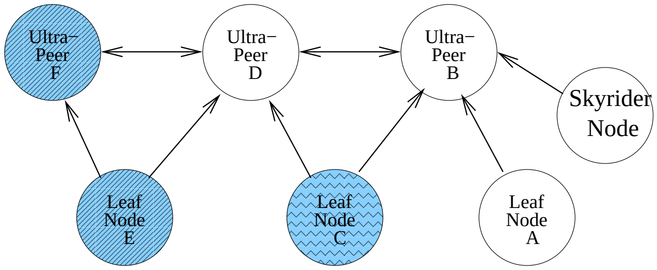


Fig. 1. Geo-Aware Query Measurement in a Two-Tier Overlay

Fig. 1 depicts a small network segment containing a Skyrider node, along with other ultrapeers and leafs. Let's assume that leafs and ultrapeers send queries that are eventually intercepted by the Skyrider node, and analyze the availability of geographic information. Ultrapeer B is directly connected to the Skyrider node. Thus any query that traversed only a single hop must have come from it, and we thus know its IP address. Leaf A, leaf C (firewalled), and ultrapeer D are at a distance of two hops away. We cannot easily distinguish between queries coming from A, C, and D. Furthermore queries originating at C will contain B's out of band return address as C is firewalled or otherwise unable to accept incoming connections. However as we are directly connected to ultrapeer B, we can simply compare the query's out of band address with B's address. If they are not identical, the query must have come from A or D, and the address is guarantied to be the origin's address. If the query contain's B's address but passed two hops, it must be acting as a proxy for C. In this case C's address is not available, and the query is not recorded. Ultrapeer F and leaf E are at a distance of 3 hops away. When we intercept their queries we cannot know whether the out of band IP address belong to them, or perhaps to ultrapeer D acting as a proxy for E. Thus any query that traversed 3 hops or more is discarded. As a result, a Skyrider node records traffic originating from its immediate neighborhood only (having a hop count ≤ 2), thus requiring a massive deployment of such nodes. It is important to state that the described setting eliminates most of the bias against popular queries that travel only short distances before been satisfied, as we discard queries that travel more than 2 hops. However, this

[1] Skyrider is a startup company dedicated to providing enhanced services to users of peer-to-peer (P2P) networks and to make these networks more useful to the consumer and business communities.

setting does introduces a bias against queries from firewalled clients, as we record only queries that can receive incoming connections.

In order to significantly reduce the amount of queries recorded that were not generated by humans we captured only queries originating from Limewire clients. The Limewire client does not allow users to perform any kind of automatic or robotic queries. It does not allow queries with the SHA1 extension, nor does it allow the automatic resending of queries. When it does send duplicate queries, it uses a constant Message ID which enables us to efficiently remove duplications. Thus, we avoid many of the difficulties experienced by [8]. As mentioned above, Limewire is by far the most popular Gnutella client, thus we do not loose much by excluding all other clients. Capturing only Limewire queries is an easy task as Limewire ”signs” the message ID associated with each message it sends. This signature can be easily verified by the intercepting node, and thus can be used to eliminates queries from all other clients who do not employ the Limewire signature scheme.

The large scale deployment of our system, combined with the features explained above, provides us with unprecedented amounts of records of human generated geographically-identified queries.

B. Data Processing

Our system simulates a vast number of nodes running on multiple computers. Logs of geographically-identified queries, which are intercepted by the system, are generated on each computer and IP information is resolved into geographical information using the MaxMind database (a country code for the purposes of this article). These log files are transferred daily to a central repository. Due to the fact that the same query may be intercepted by more than one Skyrider node we run a de-duplication script that removes duplicate instances of the same query by using the Gnutella message ID, as this ID is constant per query instance when the out of band return address belongs to the query originator. The final result of this preprocessing stage is a single daily log file, containing a single record for each geographically identified query captured. These files are used as input for further analysis.

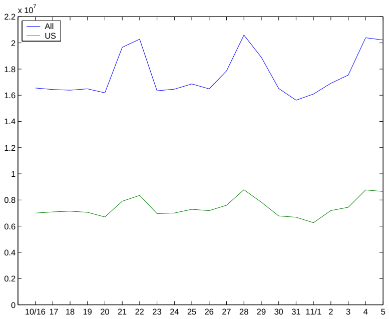


Fig. 2. Total Unique Queries Collected

IPs	**US**	**CA**	**GB**	**JP**	**NL**	**BR**
10/17	1175	269	248	92	77	89
10/24	1147	259	250	89	78	84
10/31	1080	240	201	89	74	86
Queries						
10/17	7101	1461	1852	1330	461	391
10/24	7020	1444	1950	1246	487	375
10/31	6691	1319	1521	1281	356	386
Percent[2]	42.8	8.3	9.3	7.9	2.8	2.5
$avg \frac{\text{Queries}}{\text{IPs}}$	6.12	5.50	7.61	14.25	6.05	4.45

TABLE I

NUMBER OF UNIQUE IP ADDRESSES (THOUSANDS) AND NUMBER OF UNIQUE QUERIES (THOUSANDS) PER COUNTRY PER DATE

II. DATA-SET STATISTICS

We have captured query traffic for 38 days, beginning on 7/15/2006 and ending on 11/5/2006. The total number of unique geographically identified queries intercepted is 665 million queries.

Fig. 2 shows the daily count of queries captured over a three week period beginning October 16th, globally, and in the US. It is interesting to note the increase in the number of intercepted queries on Saturdays and Sundays. This is probably due to the increase in the number of connected users over the weekend, and a probable increase in their session lengths as users are away from work or school. The slowest days during this period was Oct. 31st when many nations celebrate Halloween. In the US the slowest date is Nov. 1st, since we use midnight UTC for day border, thus much of Halloween activities in the US occur on Nov. 1st UTC.

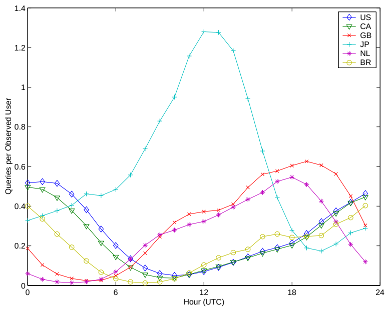


Fig. 3. Queries per hour per user

Table I [2] provides details on the number of unique IP addresses and the number of unique queries observed on three different dates from six different countries. Note the dominance of US queries, 42.78%, and of queries from mainly English speaking countries (US, CA, GB, AU) ,63.24%. The top 13 countries account for 86.89% of all queries.

Fig. 3 examines the diurnal patterns of query activity in the six countries with most queries. The X axis is the time of day (UTC) and Y axis denotes the number of queries observed per hour, divided by the total number of unique IPs observed during the entire day. This is calculated on 3 different days (Oct. 17, 24, and 31) and an average is used. Normalizing this way allows us to factor away the population size, and more easily compare the behavior of the average user in different countries. Similarities between the different countries are immediately evident. In all countries plotted, peak activity occurs in the local evening hours. At the very late night hours query activity drops to about $1\% - 13\%$ of peak activity. Excluding Japan, the remaining 5 countries are similar in the activity level of their users. At peak hours we record an average of 0.5-0.6 queries per non Japanese user per hour and $0.01 - 0.05$ queries in the trough hours. For Japanese users we record a peak of above 1.2 queries per user per hour and 0.17 in the trough hours. At this stage we can only speculate on the reasons for the rather higher activity levels of Japanese users throughout the day.

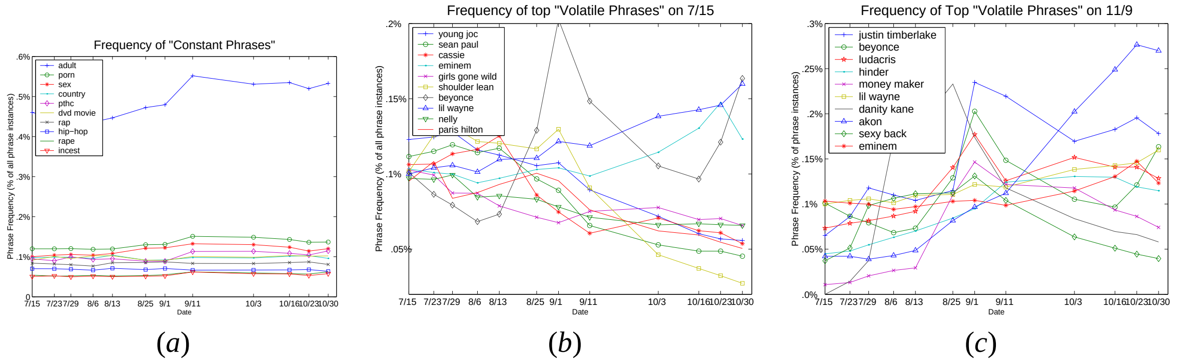
III. Temporal Behavior of the Most Popular Queries

A. *Phrases of constant and volatile popularity*

We examine the change in popularity of search terms over a 3 and a half month period beginning 7/15/2006 and ending at 10/31/2006. As we begun our analysis it became immediately evident that there are two distinct types of search phrases over the Gnutella network, we denote them as ”Constant Phrases” and ”Volatile Phrases”. Constant phrases are searches phrases aimed at finding any content of a certain type. Popular constant phrases are mainly music related terms like ”country”, ”rap” and ”hip-hop” or adult related like ”adult”, ”porn” and ”sex”. We find that over a 3 month period there are only slight changes in the frequency of these queries. On the other hand, volatile phrases are search phrases for more specific content. The top volatile search phrases are all name of performing artists. As the popularity of artists changes rapidly, the frequency of these phrases changes rapidly with time, and it is possible to track significant popularity changes over a 3 and a half month period.

Fig. 4 shows the evolution of the popularity of the top 10 phrases of each type for the 7/15/2006 and the 9/11/2006. Fig. 4(a) is sufficient to describe both dates since the top 10 constant phrases do not change between these two dates. Note how relatively unchanged the frequency of these phrases remains throughout the reported period. The order of popularity of these phrases just slightly changes. It is also interesting to observe that ”adult” stands out in being much stronger than any other search term, constant or volatile. Fig. 4(b) shows the evolution of the top 10 volatile phrases on 7/15/2006. Fig. 4(c) shows the evolution of the top 10 volatile phrases on 9/11/2006. It is interesting to note that only 3 of the 10 most popular phrases on 7/15/2006 are also popular on 9/11/2006 (”eminem”, ”beyonce”, ”lil wayne”). These three phrases enjoy a perhaps not stable, but still clear upward trend. The other 7 top phrase show a constant decline, and are out

[2] The breakdown of queries by country on 10/31/2006, in other countries is as follows: FR-3.3%, AU-2.7%, DE-2.2%, IT-1.4%, BE-1.4%, SE-1.1%, MX-1%. Other 194 countries account for 12.18% of all queries, and 0.93% of all queries remain geographically unresolved, although we do have their IP address.



(*a*) (*b*) (*c*)

Fig. 4. Top-10 Phrases Popularity Evolution. (a) constant (b) volatile topped at beginning (c) volatile topped at 0911

of the top 10 by 9/11/2006, and continue declining thereafter. Fig. 4(c) provides a view of the past and future of the top 10 phrases on that date. As might be expected, some of the phrases show a peak on the vicinity of 9/11, exhibiting their past climb into the top 10 and their future decline (see for example ”sexy back”, ”justine timberlake”, ”danity kane”, ”money maker”). ”akon” on the other hand, shows a steady climb thought out the reported period.

B. *Cumulative Top-100 Rank Drift*

We analyze change in popularity between two sets of queries captured from different populations at the same time or from same population at different times. We choose to consider just the top 100 popular queries as they reflect the interests of the compared population during the capture interval. For each of the top 100 terms we define the *ranking drift* as the difference between its rank in the compared set to its rank in the reference set. We then plot the cumulative distribution (CDF) of this difference for all top 100 terms.

Fig. 5 compare the ranking drift from the first snapshot, captured at 07/15, in the depicted 6 countries over a period of three month till 10/20. As expected the ranking drift increases for the longer compared time interval.

Fig. 6 compares the difference between the top 100 terms in several countries to the top 100 US terms. To increase confidence the CDF of ranking drift is calculated from four snapshots, captured approximately every month (at 07/15,08/13,09/11 and 10/20). First, we rank the top 100 terms of each

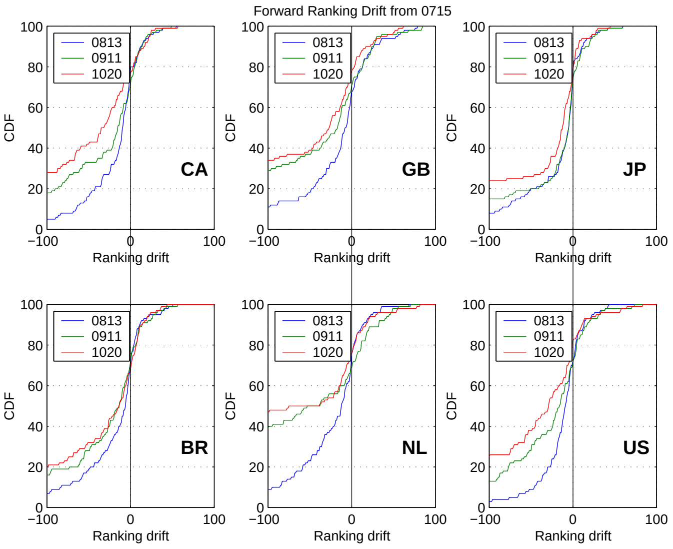


Fig. 5. Top-100 ranking drift between 07/15 and depicted dates

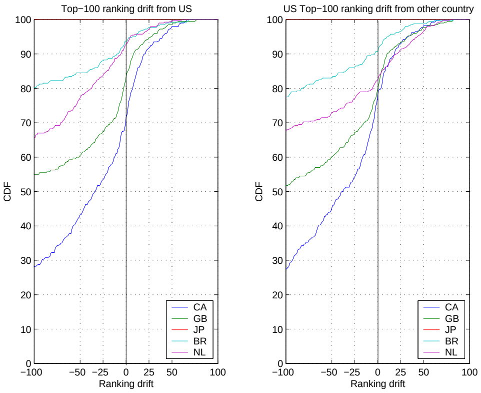


Fig. 6. Top-100 ranking drift between US and depicted countries(left) and vice versa (right) averaged on 07-10/06

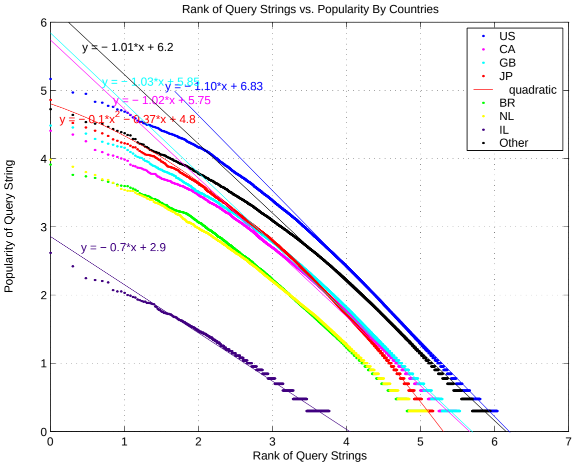


Fig. 7. Popularity Rank of Query Strings

country from each snapshot separately. The ranking drifts are then calculated for all 100 terms from each snapshot, and the CDF is calculated over all 400 ranking drifts. Fig. 6 (left), depicts the ranking drift of each country from the top 100 US terms. The ranking drift of US from the top 100 terms of each country is depicted in Fig. 6 (right). The correspondence of the countries with US rank differs a lot. For instance the percentage of US terms whose rank drift is in the range $-25 \ldots 25$ is 38% for Canada, 28% for Great-Britain, 13% for Netherlands and 9.5% for Brazil. The top 100 terms in Japan and the US has no common term!

C. *Query Popularity Distribution*

We calculate the query popularity distribution separately for each country over a relatively short time, that is one week. That way we avoid the bias of the distribution by the different popularity ranking of each country, Fig. 6 and the temporal rank drift, Fig. 5. Thus we calculated the popularity of all the queries captured continuously between the 10/29/2006 and the 04/11/2006. Fig. 7 depict the popularity of query strings as a function of their popularity rank. The data points are depicted in different colors for each country and the linear fit lines are depicted by the same color. The query strings with popularity 1 were removed to prevent tilt of the linear fit. The lower popularity y-values, $1 \ldots 1000$, follows a Zipf power-law in all countries, except Japan. Japan data fits very well to a second order Zipf distribution. We depicted also the frequency of query strings as a function of their popularity but it is omitted to save space. For US we found a fairly good fit to Zipf power-law $y = -1.89x + 6.11$, in the popularity range ≤ 200. We found a similar fit for Japan and all other countries, for the frequencies of their mid and low popularity.

IV. Concluding Remarks

We have demonstrated large-scale capturing of geographical-identified queries, which enables us to track the instant changes in P2P users interests from each geographic area. We can use the same tools and techniques to deepen our understanding of cultural difference between countries, cities and even neighborhoods in the same city. We have also started working on correlating real world events like the release of a new video clip or an appearance in a national event with changes in the popularity of artists on P2P networks in different geographic regions.

References

[1] N. Leibowitz, M. Ripeanu, and A. Wierzbicki, “Deconstructing the kazaa network,” in *WIAPP*, 2003.

[2] S. Sen and J. Wang, “Analyzing peer-to-peer traffic across large networks,” *IEEE/ACM Trans. Networking*, vol. 12, pp. 219 – 232, Apr. 2004.

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

[4] F. L. Fessant, S. Handurukande, A.-M. Kermarrec, and L. Massoulie, “Clustering in peer-to-peer file sharing workloads,” in *Intl. Workshop on Peer-to-Peer Systems*, 2004.

[5] J. Liang, R. Kumar, Y. Xi, and K. W. Ross, “Pollution in p2p file sharing systems,” in *INFOCOM*, Mar. 2005.

[6] S. Zhao, D. Stutzbach, and R. Rejaie, “Characterizing files in the modern gnutella network: A measurement study,” in *SPIE/ACM Multimedia Computing and Networking*, San Jose, Jan. 2006.

[7] K. Sripanidkulchai, “The popularity of gnutella queries and its implications on scalability,” Feb. 2001, featured on O’Reilly’s www.openp2p.com website.

[8] A. Klemm, C. Lindemann, M. Vernon, and O. P. Waldhorst, “Characterizing the query behavior in peer-to-peer file sharing systems,” in *Internet Measurement Conference*, Taormina, Italy, Oct. 2004.

[9] A. H. Rasti, D. Stutzbach, and R. Rejaie, “On the long-term evolution of the two-tier gnutella overlay,” in *IEEE Global Internet Symposium*, Barcelona, Spain, Apr. 2006.