”Dude, where’s my Peer?”

Anirban Banerjee, Abhishek Mitra and Michalis Faloutsos
Department of Computer Science and Engineering
University of California, Riverside.
Email: anirban, amitra, michalis@cs.ucr.edu

Abstract*— Peer to Peer (P2P) flows constitute a large portion of Internet traffic meandering through different ISP domains. Hence, it is of prime concern for ISPs to try and gauge the number of active P2P users and where they are located, both inside and outside their respective domains. An analysis of where P2P users are located in the Internet provides an insight into understanding which ISPs harbor a majority of P2P peers, which ones afford most transit to P2P flows, and possibly which ISPs should focus on anti-P2P policies the most. We observe an extremely skewed distribution, approximately 92 to 98% of P2P flows ending in tier 1 and tier 4 ISPs, and just 2 to 8% ending in tier 2 and 3 ISPs. We quantify the role of ISPs in allowing P2P flows to traverse through their domains and observe a similar skewed distribution wherein tier 1 and tier 4 ISPs contribute 92 to 95% of all hops on most P2P flows. Moreover, we compare P2P traffic with http and Internet radio traces to uncover potential parameters to differentiate between these types of flows. We present detailed results based on active measurements taken over a 30 day period, spanning over half a million P2P flows, collected from measurements employing a number of popular P2P clients hosted inside two popular ISPs.

I. Introduction

P2P networks have emerged as one of the most prevalent entities on the Internet. These networks allow for large groups of users, employing small, easily available and royalty free, clients to share a vast plethora of resources. Such resources can range from legal content such as Linux distributions to exchange of copyrighted material in the form of songs, movies and software. Such file sharing networks generate a significant amount of traffic when users attempt to share resources among themselves [7]. This is a source for concern to ISPs, since P2P algorithms have been shown to be ISP unfriendly [7] generating large amounts of traffic crossing over inter-AS boundaries, increasing AS-AS traffic and hence resulting in higher operational costs for the service providers.

P2P networks such as Gnutella, Fastrack, Bittorrent (BT), eDonkey, [1], [3], [5], [8] are rampant throughout the Internet today. They are accessed using their vanilla mainline clients and also with a humongous list of their variants. Resources shared among users of such networks are not trivial, either in content, the veracity of which can be gauged by significant legal action against a subset of users of some P2P networks [10]; or in the amount of data that is being transferred to and from clients [4], [6], [9] quietly chugging away. The primary motivation for these networks being: to allow users to share resources effectively and possibly fairly. Naturally, they do not have any consideration for utilizing resources, owned by the ISPs benevolently. It is thereby of utmost importance for ISPs to try and understand the extent of such P2P networks throughout their domains, mainly which ISPs harbor large clusters of users and what methods may be employed to detect such traffic flowing under the hood.

Our research asks the following questions:

1) What kind of network-wide spatial behavior do P2P users display, and which ISPs host large numbers of P2P peers?
2) Which ISPs allow most P2P traffic to pass through their domains?
3) Is the spatial behavior for P2P traffic different from other kinds of traffic, such as http, Internet radio?

We present our research based on profiling P2P flows weaving their way through the ASs in the Internet to understand which ISPs shelter large numbers of P2P users within their domain. This is imperative to understand which ISPs should possibly implement anti-P2P policies more vehemently than others. Additionally, with P2P based content distribution networks becoming a reality [7], this study is even more pertinent to understand which ISPs could cache content for swift delivery to P2P users through these overlay networks. Furthermore, we compare P2P traffic flows with more traditional traffic such as http and Internet radio, based on profiling results, to see if different applications display different network-wide spatial behavior. We slice up the AS structure according to a simple degree based classification, pivoting on CAIDAs AS-degree ranking [15], wherein we label the top 8 ISPs as tier 1, the next 24 as tier 2, the following 48 as tier 3, and the rest as tier 4 since most ISPs at this level have very few number of connections in comparison to the other ISPs in higher tiers. Each separation point in this simple classification represents a relatively sharp change in AS-degree in the CAIDA dataset, and intuitively differentiates the ISPs among each other. Our contribution can be summed up as follows:

1) We profile over half a million P2P flows, spread over a 30 day period, employing Yahoo DSL and Charter Communications as our primary ISPs for trace collection.
2) We quantify the network-wide spatial behavior of P2P users located in various ISPs, to find that tier 1 [16], [17], [18] and tier 4 ISPs host about 92-98% of all P2P IPs identified from our traces while tier 2 and 3 ISPs seem to host hardly any peers.
3) We identify which ISPs allow large numbers of P2P flows to traverse through their domains, to find again tier 1 and tier 4 ISPs contributing 92-95% of the number of

hops on most P2P flows.

4) We profile P2P flows and compare it with other prevalent Internet traffic as http and Internet radio, to ascertain if different applications display different spatial characteristics. We succeed in mining such metrics, such as the IR metric, which may be employed as a first step in conjunction with standard flow identification techniques to home in on suspected P2P traffic.

II. Related Work

P2P networks and their behavior have been the focus of active research efforts over the recent past. Efforts have been made to try and fathom the models being used by popular networks, [5], such as Gnutella, Edonkey [8] and BT [1], [3]. Studies carried out on such P2P networks, as highlighted in [4],[5],[6], [7] provide an in-depth perspective on how to discriminate traffic emanating due to P2P networks versus other Internet traffic. These methodologies range from payload identification, which involves filtering traces for particular hex strings, known beforehand, in the payloads of the packets captured. Other mechanisms employ parameters such as TCP flow holding time, average downloaded data size and others, to home in on possible P2P flows [7]. Research work regarding AS-AS interactions and P2P traffic have concentrated on interactions between a pair of ASs, while we attempt to develop a birds eye view mapping of where P2P users are located in the AS hierarchy. Furthermore, we compare P2P traffic with http traffic and Internet radio traces and highlight the differences between them. We employ custom designed tools interfaced with ethereal [2] in order to extract the AS information for each P2P flow.

III. Where are my Peers?

P2P peers are distributed throughout the AS hierarchy. We concentrate on ascertaining which ASs host the most end points for P2P flows. For our experiments we chose two popular ISPs, Charter Communications Inc. and Yahoo DSL, from which to initiate connections to various P2P networks. Both these tier 4 providers were chosen for the simple reason that, if we were to choose a tier 1 ISP from which to collect traces we would probably miss out on the spatial behavior displayed by P2P flows as they rise up from lower tiers to tier 1. We would only be able to observe end point distribution but not P2P flow behavior exhibited as the connections traverse towards tier 1 through tiers 2 and 3. We employed a number of clients feeding off Gnutella, FastTrack, and Edonkey networks such as Bearshare, eMule, Limewire, Phex, Gnucleus, Xolox, Kazaa lite, iMesh, and mlDonkey. Traces were collected on 3Mbps links for a period of 30 days and more than half a million P2P flows were analyzed in the process. For trace collection we employed Ethereal as our primary data logging tool, feeding off traces from 22 clients . Custom scripts were used to filter and mine logged data to extract relevant statistics. Lists of popular music files, and videos compiled from well known listings on the Internet [19], [20], were used to inject queries into the P2P network.

Traces were logged for observation intervals (OIs) of 1, 2, 3, 5, 10, 15 and 30 minutes each. No two OI's for the same or different duration overlap. This was done primarily to determine the temporal robustness of any metrics we develop for comparing P2P versus non-P2P traffic, e.g. to observe if the statistical behavior displayed during a 1 minute OI is the same as displayed within a 5 minute OI. This is critical for developing a robust metric which can be employed for successful identification of P2P flows over a range of observation periods.

We use the latest AS rank data from CAIDA [15], to obtain a complete map of ISPs in the various tiers and employ BGP dumps from [11], [12], [13], [14] for IP to AS lookup. Here we define the end point of a flow to be the final destination IP for that flow and the sink to be the AS at which the flow terminates. We observe a significant percentage of P2P flow end points concentrated in tier 1, and tier 4 ISPs as illustrated in Fig. 1 and 2. Table 1 lists out the end point distribution in the various tiers. We infer, for an observation period ranging from 1 minute to a 5 minute OI, the percentage of tier 1 end points varies from 6.1% to 17.7% of the total number of end points logged for that duration, for P2P flows. Tier 4 ASs consistently contribute a majority of end points, ranging from 79.03% to 87.39%, over the complete range of measurements. The fluctuations in values observed can be related to the fact that with each incrementally increasing OI more P2P peers are contacted in comparison to smaller OI durations, this leads to differences in how many P2P flows end in the various tiers. Surprisingly, ASs in tiers 2 and 3 contribute end points meagerly. For other OI's with durations larger than 5 minutes we observe a similar trend. This skewed behavior is extremely intriguing and poses the following question. Since a large chunk of customers for tier 1 ISPs are large commercial organizations, do these results suggest that large corporate entities may unknowingly be harboring P2P clients on their machines?

We believe that the reason for such skewed statistics are as follows: Most P2P users obtain Internet connectivity via smaller tier 4 ISPs, and it is natural to observe a large concentration of end points in tier 4 ISPs. Some tier 1 ISPs host large numbers of modem based dial-up customers, and sell bandwidth to corporate entities at the same time. We believe that a large part of the contribution from tier 1 ISPs is due to residential customers hooking on via their dial-up connections and joining up with P2P communities. Additionally, we organize end points in bins based on an intuitive sliding scale detailed in Table 2 and observe the same skewed behavior as displayed by the tier-wise classification. Again, the largest and the smallest of ISPs seem to contribute most significantly to the number of P2P flow end points, displayed in Fig. 3. We however do not have a good explanation for why tier 2 and 3 ISPs do not contribute a larger share of P2P end points unlike tier 1 and tier 4 ISPs.

As will be discussed in section IV, this metric, for P2P flows is quite different from non-P2P flows such as Internet radio and http, and may be employed as a low-computation first line

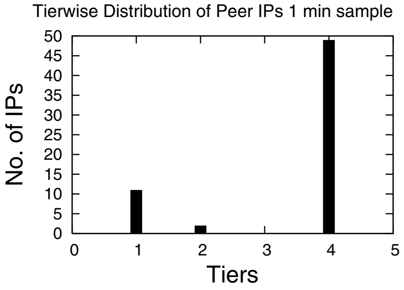

Fig. 1. P2P end-point tier-wise distribution for a 1 min trace.

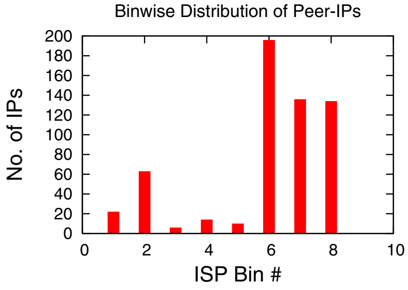

Fig. 3. Binwise P2P end-point distribution for a 1 min trace.

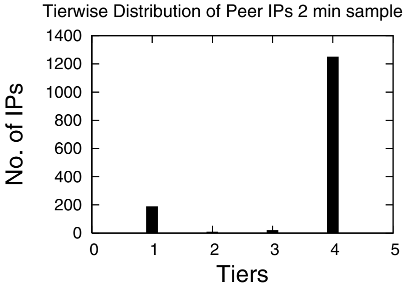

Fig. 2. P2P end-point tier-wise distribution for a 2 min trace.

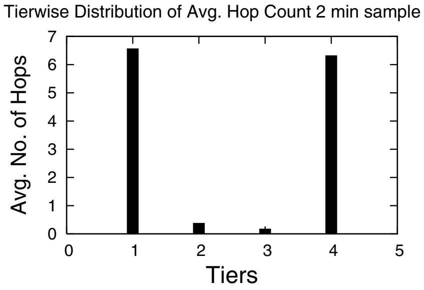

Fig. 4. Tier-wise distribution of average number of hops of P2P flows, 2 min duration.

of inspection for identifying P2P flows from the huge amount of network traffic generated by a node.

OI	$Tier1$	$Tier2$	$Tier3$	$Tier4$
1 min	17.7	3.20	0.07	79.03
2 min	12.8	0.67	1.54	84.99
3 min	6.10	6.20	0.55	87.15
5 min	7.73	3.15	1.73	87.39

Table 1: *Tier-wise (percentage) distribution of P2P end-points.*

$Bin-No.$	$ISP-rank$
1	1-3
2	4-10
3	11-50
4	51-100
5	101-200
6	201-500
7	501-1000
8	1001+

Table 2: *Bin-wise distribution of ISPs according to No. of connections. Data sourced from CAIDA [15].*

At this juncture we ask, *given these statistics would it be prudent to assume that ISPs in tiers 1 and 4 should be the ones to implement anti-P2P policies more vehemently than tier 2 and 3 ISPs?* To answer this question, it is imperative to examine which ISPs allow a large number of P2P flows to pass through their domains. This affords us a more informed view regarding which ISPs should perhaps implement anti-P2P policies more industriously than others. Fig. 4 and 5 provide an idea of how much transit is provided by ISPs in the various tiers to P2P flows. We say that an ISP provides transit to P2P flows if it allows such flows to pass through its domain. The average number of router hops, distributed tier-wise, for all P2P flows captured during the various time durations provides an insight into which tiers provide more transit than others. Again we observe a skewed distribution, tiers 1 and 4 contain most of the hops in the P2P flows. Apparently, *P2P flows seem to traverse through tiers 2 and 3 rapidly while seemingly staying for a longer number of hops in tiers 1 and 4.*

OI	$Tier1$	$Tier2$	$Tier3$	$Tier4$
1 min	48.12	5.0	2.5	44.38
2 min	48.70	2.86	1.24	47.2
3 min	45.83	3.6	1.92	48.65
5 min	45.7	5.0	1.91	47.39

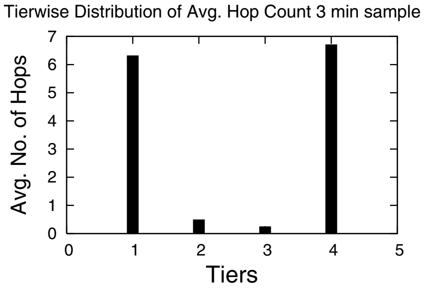

Fig. 5. Tier-wise distribution of average number of hops of P2P flows, 3 min duration.

Table 3: *Tier-wise (percentage) distribution of average number of hops of P2P flows.* This possibly implies that those *ISPs which act as large sinks for P2P flows also provide maximum transit for P2P connections*. Table 3 depicts in detail the contribution of each tier in providing transit to P2P flows.

One interesting statistic we observe is that, approximately 98% of all P2P flows traverse tier 1 ISPs and only a very small number of flows do not pass at all through tier 1 ISPs. This alludes towards the hypothesis that tier 1 and tier 4 ISPs not only act as sinks for P2P traffic but also carry most of these flows. This observation suggests that ISPs in tier 1 should implement P2P detection policies hand in hand with tier 4 ISPs. In the following section we compare P2P flows with other kinds of common Internet traffic.

IV. P2P Traffic: A Comparison

In order to further develop an insight into how P2P traffic weaves its way through the AS structure of the Internet we compare it with other forms of prevalent Internet traffic such as http and Internet radio. In this section we present our findings which conclusively prove that that P2P traffic displays a different spatial behavior from these other forms of traffic in the Internet and quantify the characteristics which enable us to differentiate P2P flows from the rest.

Http and Internet radio traces were captured using Ethereal, running on the same machines with connections through the same ISPs which were used to gather traces for P2P flows. The top 500 websites, compiled from resources on the web, were accessed using automated scripts. Winamp Shoutcast, Yahoo Radio and Real Radio were primary resources for compiling Internet Radio traces. We present Fig. 6 and 7, detailing out the tier-wise distribution of flow end points of http and Internet radio flows. We observe that this statistic for P2P flows is different from http and Internet radio flows. We define the End-Point-Ratio (EPR) as being the ratio of end points in two tiers, e.g. EPR $_{t1,t4}$ represents the ratio of end points in tier 1 Vs those in tier 4. This provides us with a simple metric with which to compare these traffic flows. EPR $_{t1,t4}$ for Http flows was found to be approximately 0.533, while for Internet Radio applications it was about 0.466. For P2P flows EPR $_{t1,t4}$ varies from approximately 0.0699 to 0.223, significantly different from other kinds of traffic. We also compare how much transit is provided to http and Internet radio flows by ISPs in various tiers of the Internet and compare with statistics obtained for P2P flows. We present Fig. 8 and 9 which depict the tier-wise average hop count at each tier for http and Internet radio flows. We observe that for P2P flows tiers 2 and 3 provide transit, ranging from 1.2 to 5% of the total number of hops per flow. While for http and Internet radio tiers 2 and 3 contribute a miniscule 0.3 to 1.1%. For http and Internet radio connections only tiers 1 and 4 provide significant transit contributing about 98.9% of the total number of hops for each flow, and for Internet radio about 99.7%. While, for P2P flows, tiers 1 and 4 contribute about 92-95% of all hops per flow. This behavior can be explained by the fact that most popular http sites accessed are either cached by local content providers with servers in local tier 4 ISP domains or exist on large high speed content distribution networks as those hosted by the likes of Akamai, a large portion of which possibly resides in tier 1 ISPs. The same could hold true for Internet radio flows.

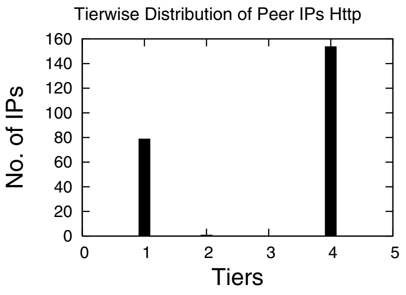

Fig. 6. Tier-wise distribution of end points of http flows.

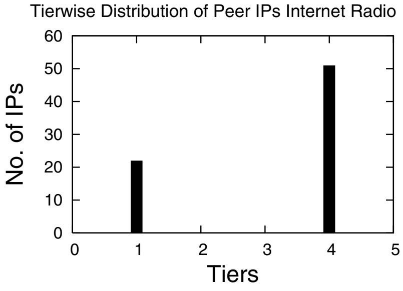

Fig. 7. Tier-wise distribution of end points of internet radio flows.

Additionally, we analyze one more interesting metric, the upslope and downslope of P2P flows versus those of http and Internet radio flows. We define the upslope of a flow as the number of hops needed by a flow to reach the highest tier, from tier 4 to tier 1. Similarly, downslope is simply the number of hops needed by a flow to reach the lowest tier from the highest. This metric, presented in Fig. 10, is especially interesting since it suggests that *P2P flows traverse a larger number of hops while weaving down the AS hierarchy, from tier 1 to lower tiers as compared to the number of hops needed to reach the topmost tiers, e.g. from tier 4 to tier 1.*

Http and Internet radio flows do not display such large imbalance in the number of hops while traversing through

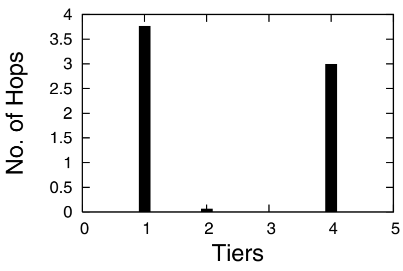

Fig. 8. Tier-wise distribution of avg. hop count of http flows.

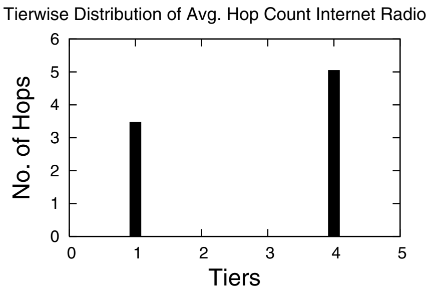

Fig. 9. Tier-wise distribution of avg. hop count of internet radio flows.

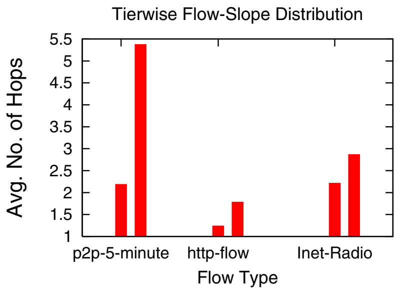

Fig. 10. Tier-wise distribution of flow slope for P2P Vs http and internet radio flows. Each pair of columns represents up-slope and down-slope for a 5-minute P2P OI, http or Inet Radio. In each pair, the first column represents up-slope and the next one depicts down-slope. Up-slope and down-slope for P2P 2 min and 3min OI's display similar behavior.

the tiers. We define the Imbalance Ratio (IR), as the ratio of number of hops traversed from tier 1 to 4, Vs the number of hops traversed from tier 4 to 1. We observe the IR for P2P flows to range from 1.8 to 2.44, while IR for http flows was observed to be 1.4 and for Internet radio was 1.27. This is a clear differentiation metric between P2P flows and other types of Internet traffic. Thus adding plausibility to the fact that P2P traffic and other prevalent forms of Internet traffic display different network wide spatial behavior. An explanation for such behavior would be, since Internet radio and popular http sites are hosted on well advertised servers, having high network visibility with entries in most network routers, once a connection reaches a tier 1 ISP it is relatively easy to find a route to the destination server. In case of P2P peers located away from the core of the net, it is but natural to hit larger number of routers in order to find a path to the other peers which definitely have much lesser network visibility than popular servers. In this section we have conclusively proved that P2P traffic displays different spatial behavioral characteristics than other forms of Internet traffic. These metrics can be employed in conjunction with other payload and non-payload based mechanisms to home in on suspect P2P flows for a closer look. In fact since our metrics do not make use of payload sniffing, they are immune to legal ramifications. Furthermore since we do not link our metrics with specific port based analysis, our mechanism can successfully target P2P clients deliberately using well known ports to mask themselves.

V. Conclusion

Our research clearly highlights the skewed distribution wherein a majority of P2P flows end at tier 1 and tier 4 ISPs to the tune of 92 to 98% of all P2P flows analyzed. Also, 92 to 95% of P2P flows traverse through tier 1 and tier 4 ISPs, incurring a larger number of hops in these tiers than in tiers 2 and 3. Furthermore, tier 2 and tier 3 ISPs do not seem to participate significantly in providing transit to P2P traffic neither do they act significantly as sinks for the same. Interestingly, a considerable percentage of P2P flows, nearly 98% of the complete observation set, managed to reach tier 1 ISPs and weave through their domains. These facts may encourage tier 1 and 4 ISPs to implement anti-P2P policies more vehemently than others. Moreover, we observe that P2P flows traverse a larger number of hops while weaving down the AS hierarchy, e.g. from tier 1 to tier 4 as compared to the number of hops needed to reach the topmost tiers from the lower ones. The imbalance metric referring to this observation, in conjunction with others developed throughout the paper conclusively prove that network-wide spatial behavior displayed by P2P flows is very different from other forms of prevalent Internet traffic.

References

[1] B. Cohen, *Incentives build robustness in Bittorrent*, In Workshop on Economics of Peer-to-Peer Systems, Berkeley, CA, 2003.

[2] Ethereal. http://www.ethereal.com/.

[3] M. Izal, G. Urvoy-Keller, E.W. Biersack, P.A. Felber, A. Al Hamra, and L. Garces-Erice, *Dissecting BitTorrent: Five Months in a Torrents Lifetime*, In PAM'04, Antibes Juan-les-Pins, France, 2004.

[4] T. Karagiannis, A.Broido, M. Faloutsos, and kc claffy, *Transport layer identification of P2P traffic*, In ACM Sigcomm IMC'04, Taormina, Italy, 2004.

[5] E. Markatos, *Tracing a large-scale peer to peer system: an hour in the life of gnutella*, In 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, 2002.

[6] S. Sen and J. Wang, *Analyzing Peer-to-Peer Traffic Across Large Networks*, In ACM SIGCOMM IMW, 2002.

[7] Thomas Karagiannis, Pablo Rodriguez and Dina Papagiannaki, *Should Internet Service Providers Fear Peer-Assisted Content Distribution?*, In IMC'05, Berkeley, 2005.

[8] Kurt Tutschku, *A measurement-based traffic profile of the edonkey filesharing service*, In PAM'04, Antibes Juan-les-Pins, France, 2004.

[9] Thomas Karagiannis, Dina Papagiannaki and Michalis Faloutsos, *BLINC: Multilevel Traffic Classification in the Dark*, ACM SIGCOMM, Philadelphia, PA, USA, August 2005.

[10] http://www.techspot.com/news/16394-record-labels-launch-legal-action-against-kazaa.html

[11] http://www.ripe.net

[12] http://www.lacnic.net

[13] http://www.afrinic.net

[14] http://www.arin.net

[15] http://www.caida.org

[16] Lixin Gao, *On Inferring Autonomous System Relationships in the Internet*, IEEE/ACM Transactions on Networking (TON), Volume 9 , Issue 6 (December 2001). Pages: 733 745.

[17] Ramesh Govindan and Anoop Reddy, *An Analysis of Internet Inter-Domain Topology and Route Stability*, In INFOCOM 1997

[18] L. Subramanian, S. Agarwal, J. Rexford and R. H. Katz, *Characterizing the Internet Hierarchy from Multiple Vantage Points*, In IEEE Infocom 2002.

[19] http://www.billboard.com/bbcom/charts/chart_display.jsp?fT̄he_Billboard_Hot_100

[20] http://www.mtvasia/Onair