Digging Into KAD Users' Shared Folders

Marcin Pietrzyk
Orange Labs, France
marcin.pietrzyk@orange-ftgroup.com

Guillaume Urvoy-Keller
Institut Eurecom, France
urvoy@eurecom.fr

Jean-Laurent Costeux
Orange Labs, France
jeanlaurent.costeux@orange-ftgroup.com

ABSTRACT

Characterizing peer-to-peer overlays is crucial for understanding their impact on service provider networks and assessing their performance. Most popular file exchange applications use distributed hash tables (DHTs) as a framework for managing information. Their fully decentralized nature makes monitoring and users tracking challenging. In this work, we analyze KAD, a widely deployed DHT system. Thanks to the unique possibility to monitor a large population of about 20,000 ADSL clients at the edge of the network, we are able to characterize the content downloaded and shared by local users. We devised a passive content monitoring toolkit to reliably track users between sessions despite dynamic IP allocation. We applied our tool over one month of data. Our main findings are: (i) Over half a TB of fresh data is downloaded every day by the users we monitor, (ii) A significant fraction of peers (20%) regulary change their ID in the KAD overlay, either on a session basis or on a sub-session basis, which can be detrimental to the proper functioning of the DHT, (iii) Those users, that we term Chameleon users, are connected longer than regular users, and they (claim to) have less data in their shared folder than regular peers and (iv) As a consequence, even a non biased observation of the users shared folder can only provide a lower bound of the content downloaded and shared by a population of ADSL users.

Categories and Subject Descriptors: H.3 INFORMATION STORAGE AND RETRIEVAL Systems and Software: Distributed systems

General Terms: Measurements, Performance.

1. BACKGROUND ON KAD

KAD is a Kademlia-based [4] peer-to-peer DHT routing protocol implemented by several peer-to-peer applications such as eMule [2] and aMule [1]. KAD is used as an overlay network for searching the content, where all the objects - nodes, files and keywords are identified by IDs in a common space. Each KAD client listens on one TCP port for data transfers and UDP port for KAD signaling. Recently, the transfer phase of the protocol is obfuscated, which makes detection difficult. However, we leverage the fact that signaling packets can still be identified by a fixed string placed in the beginning of each message. Each node in the network is responsible for advertising its shared content to the other nodes by publishing references. A reference contains file metadata and a pointer to the node that physically owns the file, so that, any node can reach the content.

SIGCOMM'08, August 17–22, 2008, Seattle, Washington, USA.
ACM 978-1-60558-175-0/08/08.

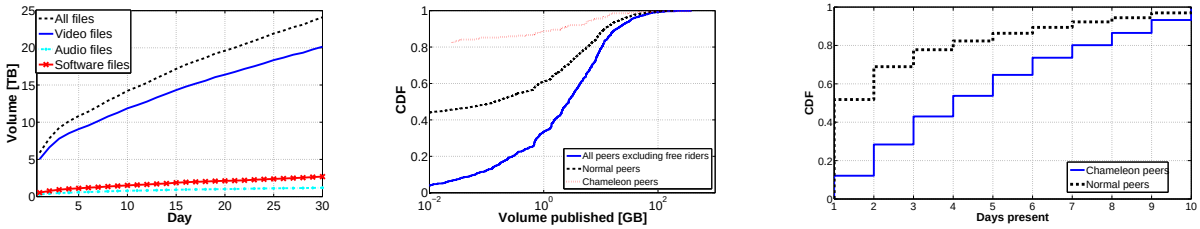
2. MEASUREMENT STRATEGY

Several studies over KAD have been performed. Authors in [5] have been crawling KAD for six months and obtained information about the total number of peers online and their geographical distribution. The same authors in [6] attempt to monitor content exchanged in the network by inserting fake identities - Sybils [7] to the overlay. However, to characterize the content exchanged by a particular user they would need to monitor the entire overlay, which is challenging from a bandwidth point of view. Furthermore, there is no functionality in the protocol, which could be used by any crawler to obtain the list of files shared by a peer, like it was done in case of Gnutella in [8]. Moreover, previous studies were also constrained by the lack of possibility of reliable end-user tracking as they rely on the IP addresses of peers, which are dynamical for most ADSL users.

We present an alternate approach. We have been capturing all KAD signaling packets continuously for one month, on an ADSL access link of a major ISP in France, connecting more than 20,000 users. To process captured messages, we designed and implemented a passive content parser based on protocol implementations of eMule and aMule. Applying the parser on our dumps, we were able to constantly monitor the content in local KAD users' shared folders. This includes file hashes, filenames, file sizes, file types and other metadata. In addition, in order to keep track of the clients we capture Radius [3] traffic. Tickets are persistent between sessions and uniquely identify each ADSL client. Once obtained the mapping between dynamic IP and constant Radius ticket, we track peers regardless their changing IP address and KAD overlay identity. Both, the Radius and data trace were fully anonymized prior to the analysis, in order to avoid privacy issues.

3. CHAMELEON USERS

During our measurements we discovered that 26% of our ADSL clients were using KAD at least once (over the month of observation). Peers in the overlay are identified by the, so called, KAD ID, which is supposed to remain the same across sessions. However, we observed that over 20% of local peers change their KAD ID. We refer to those peers as *Chameleon* peers and study their behavior separately from



(a) Evolution of the volume of the files discovered during the measurement (b) Disk space published per client (c) Number of days a client was active

Figure 1:

the standard ones. Existence of the Chameleon peers as well as other unexpected phenomena can be explained by the popularity of unofficial releases of eMule clients. We found over twenty different modifications of clients offering additional functionalities. More anonymity or promise to speedup download process can explain the popularity of those modified clients. Chameleon users were first observed by Steiner et al. [5], while they were studying the global KAD overlay. They identified the existence of Chameleon users by looking at clients with fixed IP addresses. Our results are complementary to the ones in [5], as we are able to reliably monitor users even if they change IP address. We noticed two main types of behavior regarding KAD identity changes. Some clients advertise themselves to the overlay as several clients in parallel (possibly to boost performance), while others change their identity once per session or even more frequently (for just a few percent of Chameleons). The latter case can be problematic for the proper functioning of the DHT that requires a certain stability of the peers in the network. In Figure 1(c) we plot the cumulative distribution function (CDF) of number of days each peer was seen during ten days. Chameleon peers tend to be connected more often than the ordinary ones. This indicates that these clients are using peer-to-peer extensively, and they probably use modified clients to evade any monitoring activity.

4. LOOKING INTO USERS' SHARED FOLDERS

In one month, local users have published over 275 thousands distinct files of a total volume of 24 TB. Over 60% of those files are small audio files with sizes between 1-10 MB. However over 80% of bytes discovered are due to large videos with sizes around 300 MB or 700 MB. These are typical values for movie series episodes and full movies encoded using the divx format. In Figure 1(a), we depict the evolution of the volume of files discovered during one month. We identify each unique file by its content hash which is published together with the metadata. After an initial warm up period of our monitoring tool, we observe that the amount of new content introduced to the network is stable and we have on average 620 GB of data discovered every day. Fresh content is the effect of either successful downloads or manual placement of new files inside shared folder.

In Figure 1(b) we plot the CDF of disk space shared per client during a single day. There is a striking difference between normal and Chameleon users. Fraction of the peers not publishing even a single file, referred to as *free riders*, is much higher among Chameleon users: 82% for Chameleon users against 44% for standard users. Moreover, they tend to publish less than standard peers. It is most probable that modified clients take care that downloaded content does not appear in the shared folder of the user. In addition, as Chameleon users are connected more often to the overlay than standard users (see Figure 1(c)), we suspect that those users are *heavy hitters* that download much more content than standard users; and thus 620 GB of fresh content seen every day is only a lower bound on the actual volume of content downloaded using KAD.

5. CONCLUSIONS AND FUTURE WORK

We have reported our findings obtained from monitoring the signaling layer of a large population of KAD users. KAD appears to be very popular, and KAD users download a large amount of fresh content evey day. We demonstrate that reliable users tracking is more challenging than it appears due to the existence of Chameleon peers that alter the behavior of the signaling of the client. From an ISP point of view, this means that monitoring the signaling layer can only provide lower bounds of the amount of content downloaded and shared by its customers. As a future work, we intend to couple the analysis of the signaling layer with the one of the data layer. This coupling should allow us to overcome the fact that the transfer phase is obfuscated and flag TCP flows generated by KAD users.

6. REFERENCES

[1] A-Mule. http://www.amule.org/.
[2] E-Mule. http://www.emule-project.net/.
[3] RADIUS protocol. http://www.ietf.org/rfc/rfc2865.txt.
[4] P. Maymounkov and D. Mazieres. Kademlia: A Peer-to-peer informatiion system based on the XOR metric. In Proceedings of IPTPS, , Cambridge, MA, USA, 2002.
[5] M. Steiner, T. En-Najjary, E. W. Biersack A Global View of KAD In Proceedings of IMC, San Diego, USA, 2007.
[6] M. Steiner, W. Effelsberg, T. En-Najjary, E. W. Biersack Load Reduction in the KAD Peer-to-Peer System In Proceedings of DBISP2P, Vienna, Austria, 2007.
[7] J. R. Douceur, The Sybil Attack, In Proceedings of IPTPS, Cambridge, MA, USA, 2002.
[8] D. Stutzbach, S. Zhao, and R. Rejaie Characterizing Files in the Modern Gnutella Network Multimedia Systems, Volume 13, Number 1, September 2007.