

# 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%) regularly 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 Kademia-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.

## 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

