

Server

DHT

wt

A Peer Activity Study in eDonkey & Kad

Thomas Locher
David Mysicka
Stefan Schmid
Roger Wattenhofer

[Home](#)
[Call for Papers](#)
[Invited Speakers](#)
[Committee](#)
[Important Dates](#)

[Program](#)

[Submission](#)

[To FCT Webpage](#)

[Conference Venue](#)
[Info for Visitors](#)



International Workshop on DYnamic Networks: Algorithms and Security

September 5, 2009, Wroclaw, Poland

Place: Wroclaw University of Technology, building D1, room 215

DYNAS is about dynamic networks such as Peer-to-Peer, Sensor and Ad-Hoc Networks. We will gladly welcome descriptions of original, well defined (possibly not yet completely solved) problems in the following areas.

Algorithms for construction and maintenance of Peer-to-Peer, Sensor and Ad-Hoc Networks.

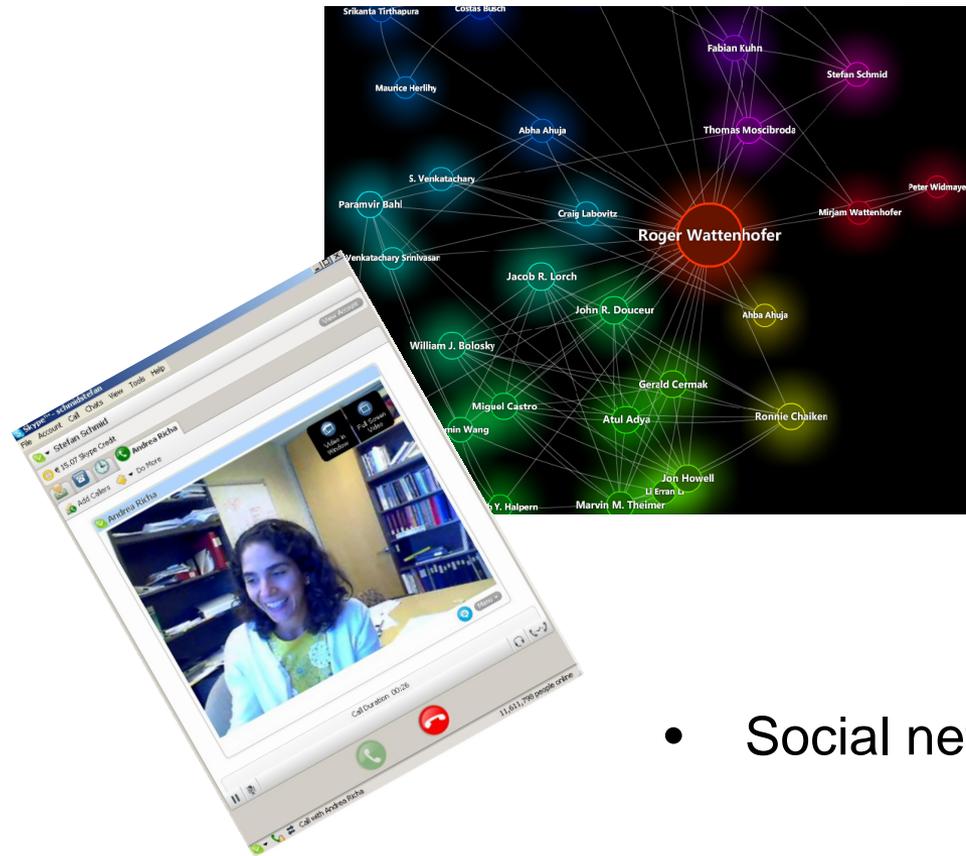
Algorithms working on top of networks.

Security and anonymity issues in systems of computationally weak devices (like RFID-tags).

The workshop will be colocated with FCT 2009. We plan to have sessions interleaving with longer time for discussions (possibly in small groups) and we will possibly have invited talks or tutorials. We especially encourage young active computer scientists to participate.

Dynamic Systems

- Many dynamic systems exist!

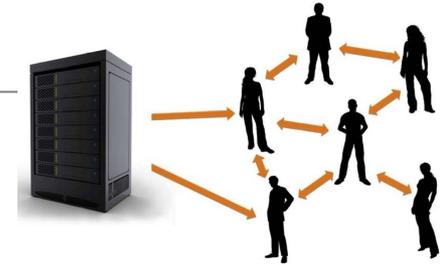


- Social networks, peer-to-peer systems...



Peer-to-Peer Systems

- In this talk: **peer-to-peer systems**



- How to design&organize an open distributed system?
- Centralized (e.g., Napster)
- „Random“ (e.g., Gnutella),
- DHT-like (e.g., Kad)

**What is better?!
It depends...**

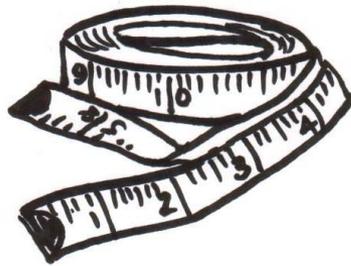
This Paper: Server vs DHT

- We performed measurements on two popular systems:
 - The server-based **eDonkey** system
 - The **Kad** network (essentially a DHT)

Both are accessed by **eMule client**:

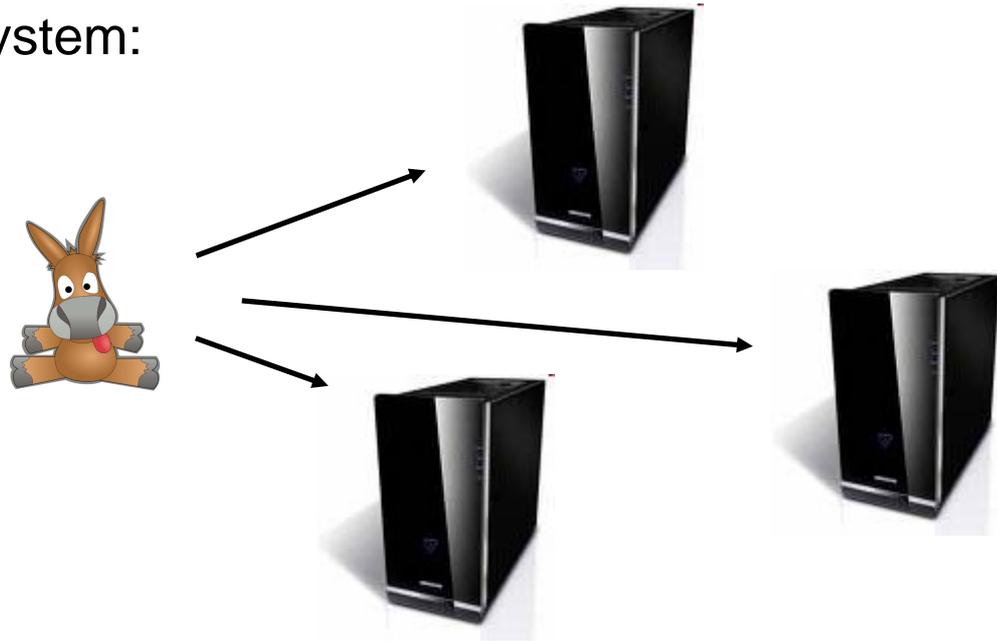


- How to measure?



eDonkey (Simplified!)

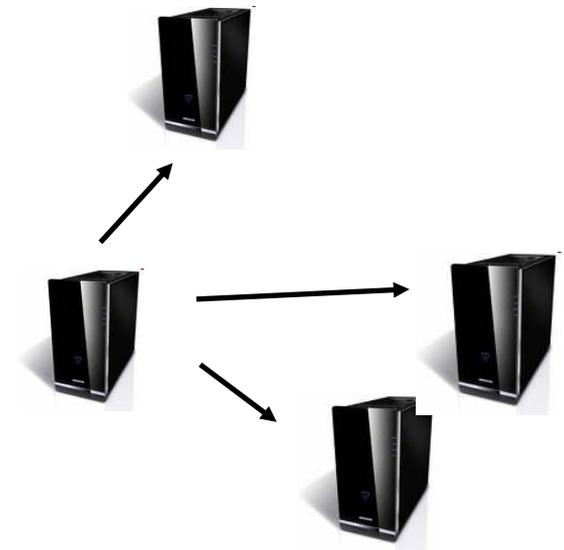
The eDonkey system:



We **reverse engineered** the *lugdunum* software (not open source to prevent fake servers),
set up our **own servers**,
and **published** them in the system.

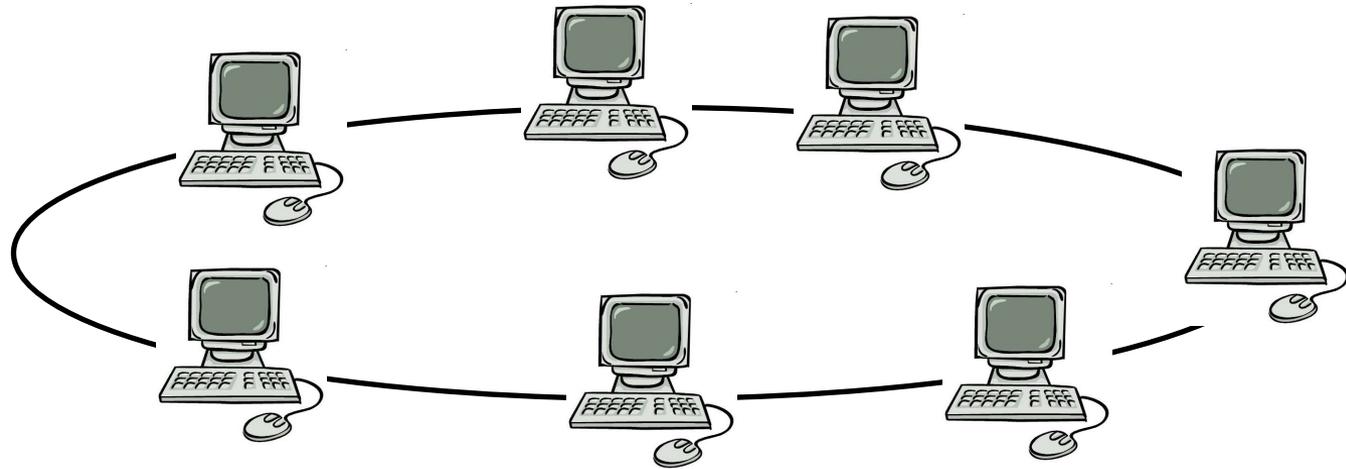
eDonkey (Simplified!)

- Peers iterate over **list of servers**, sending keyword requests
 - until **300 answers** have been received
- Our fake server **announces** itself to many servers
 - they will send this info to the peers
- We answer **status requests** from servers, but do not allow peers to **log in** (reply we are full)
- We pretend to have **many users and files**



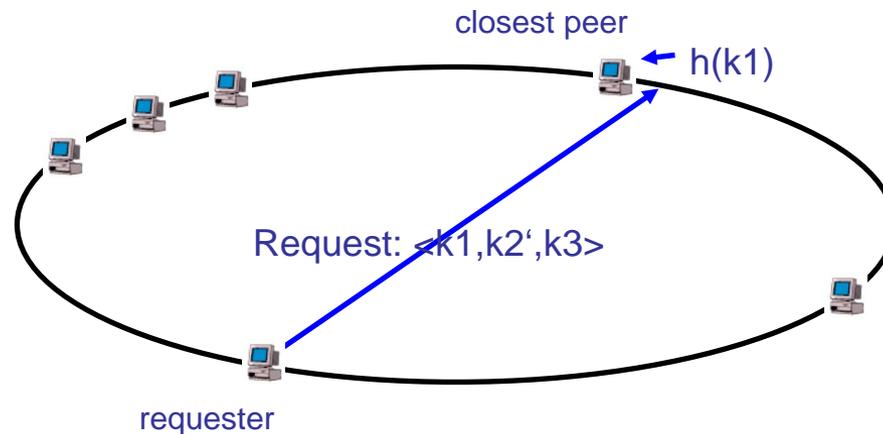
Kad (Simplified!)

The Kad system:



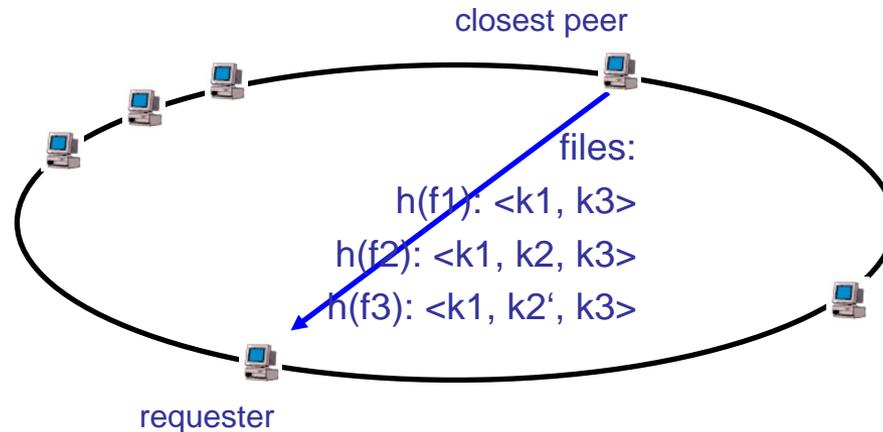
We **generated an overlay ID** at an interesting position
(„weakness“ of Kad)

Background: Kad Keyword Request



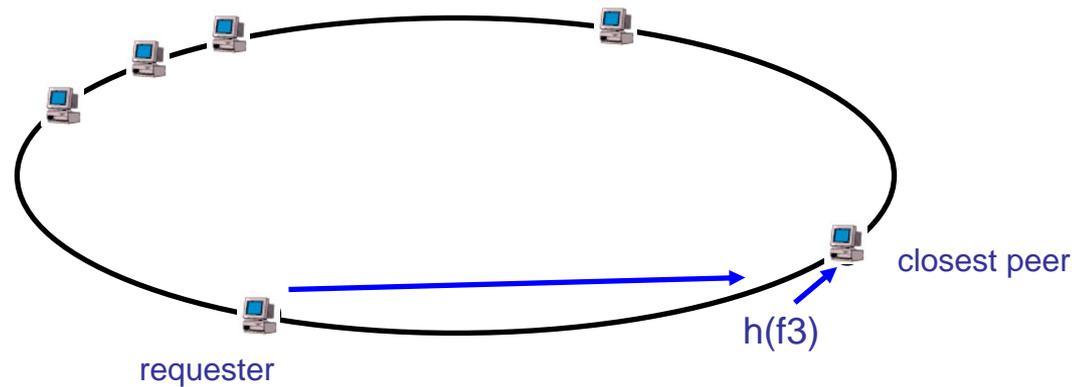
Lookup only with **first keyword** in list. Key is hash function on this keyword, will be routed to peer with Kad ID closest to this hash value.

Background: Kad Keyword Request



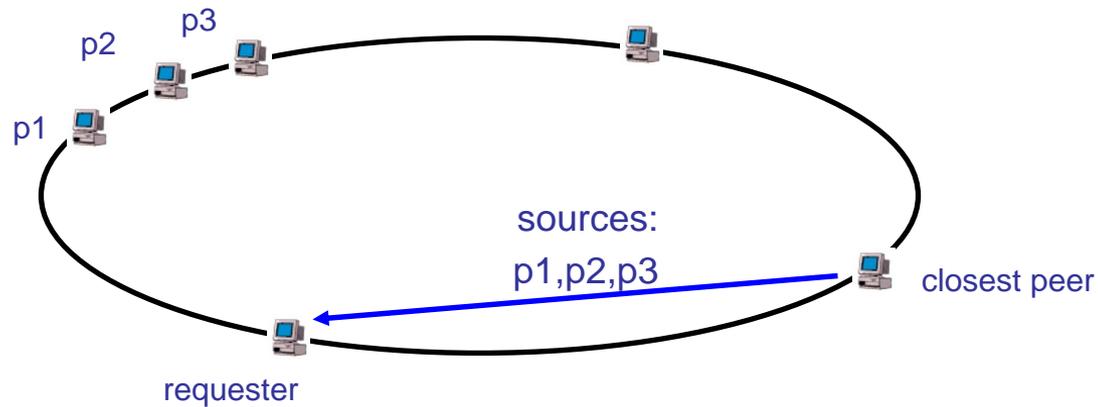
Peer **responsible** for this keyword returns different sources together with keywords.
(remark: only those files with entries that include remaining keywords of request are returned, see later)

Background: Kad Source Request



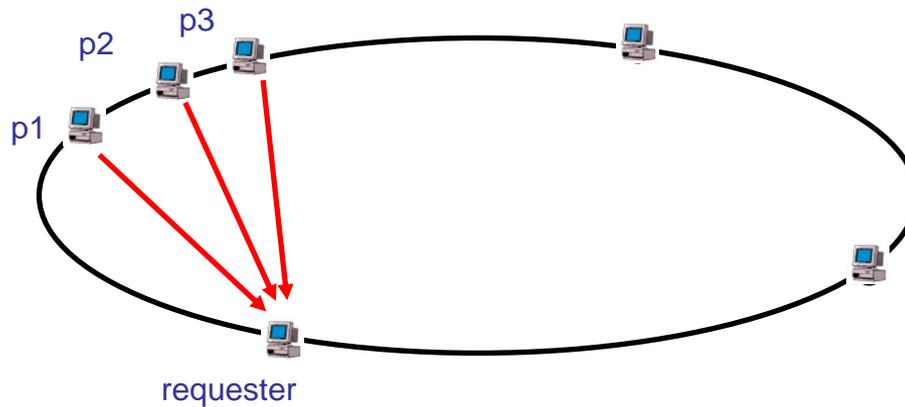
Peer can use this hash to find
peer **responsible for the file**
(possibly many with same content
/ same hash)

Background: Kad Source Request



Peer provides requester with a list of peers storing a copy of the file.

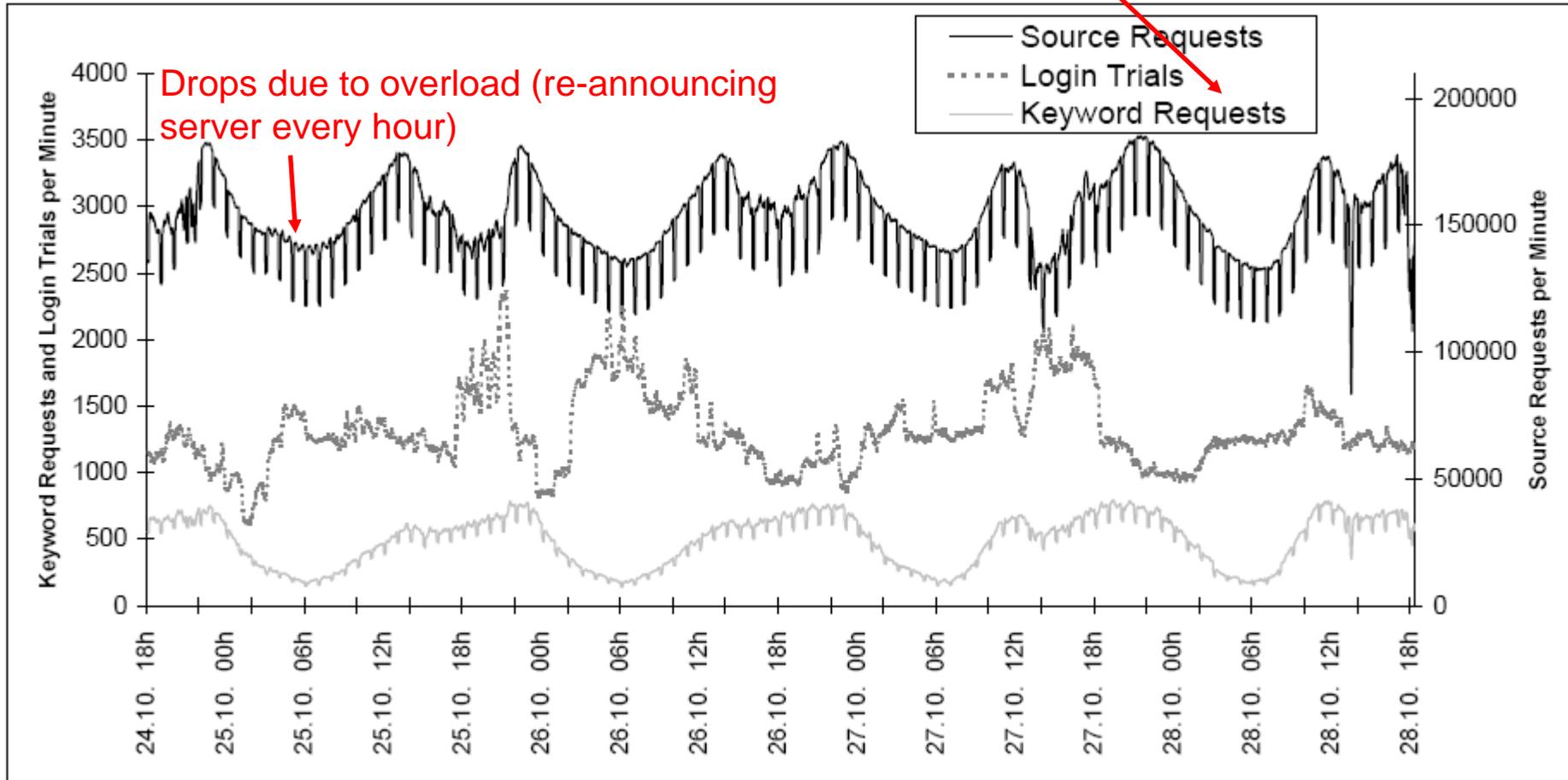
Background: Kad Download



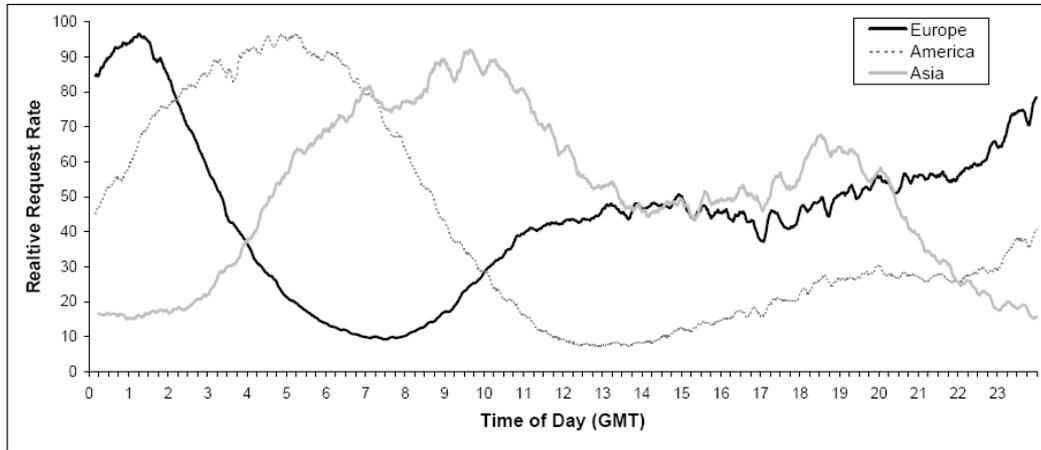
Eventually, the requester can download the data from these peers.

Activity on eDonkey

Fairly easy to make server popular...: Keyword requests entered „live“ by users (daily pattern)



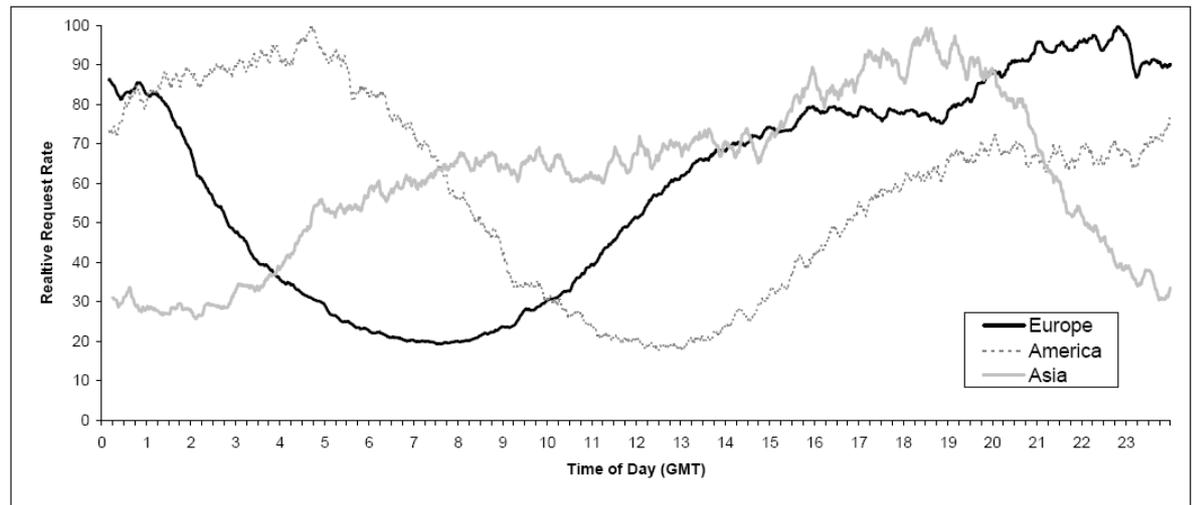
Temporal Distributions (wrt GMT)



eDonkey

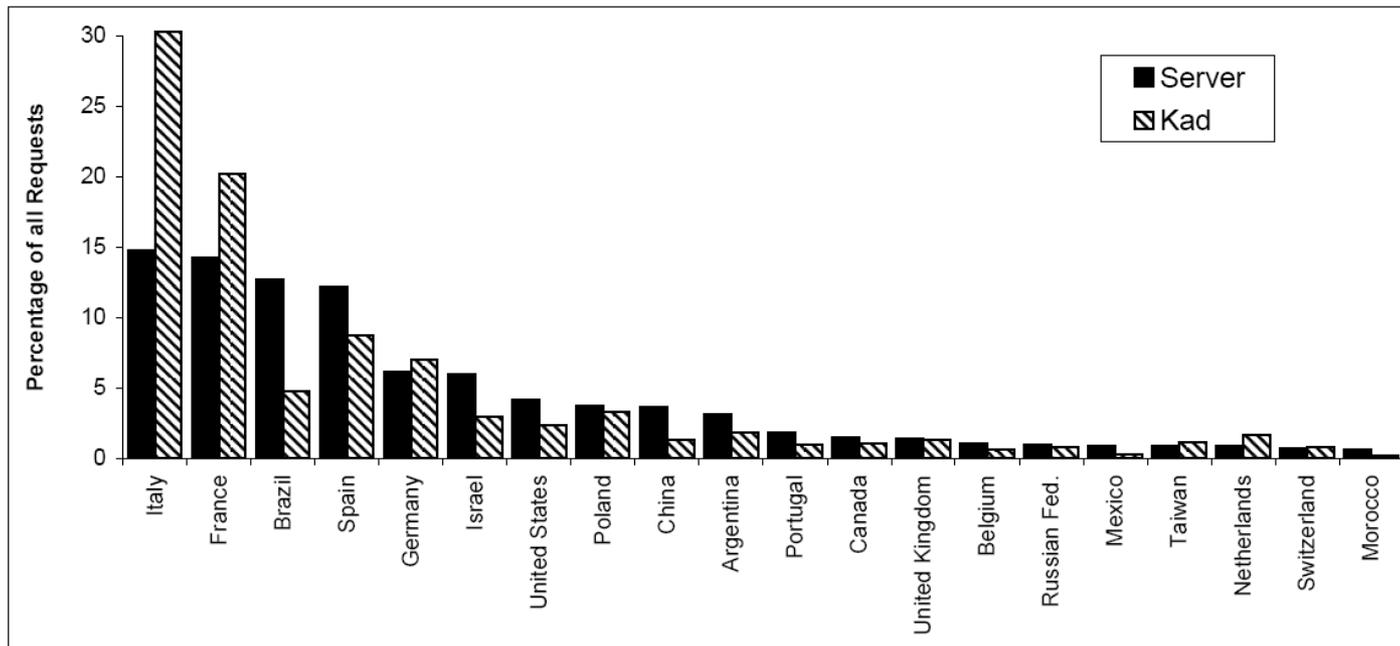
Kad

(average over 14 positions)



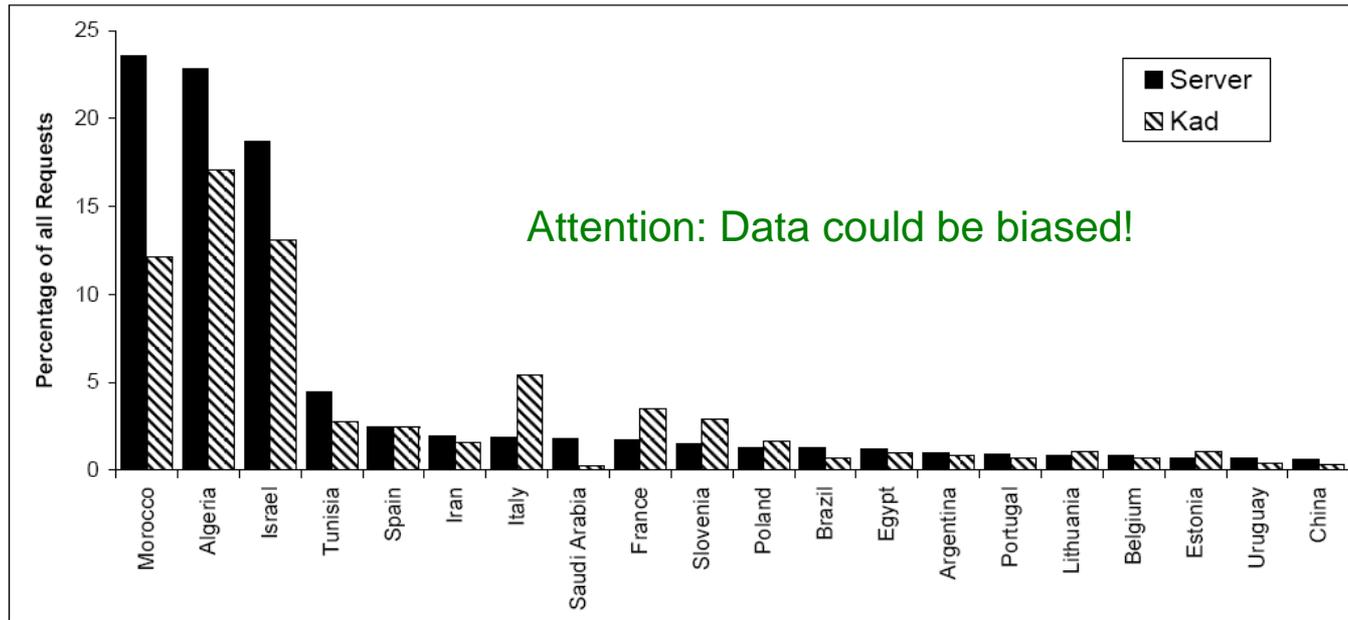
No surprise: main activity in both networks **in the evening**.

Origin of Keyword Requests (Server vs DHT)



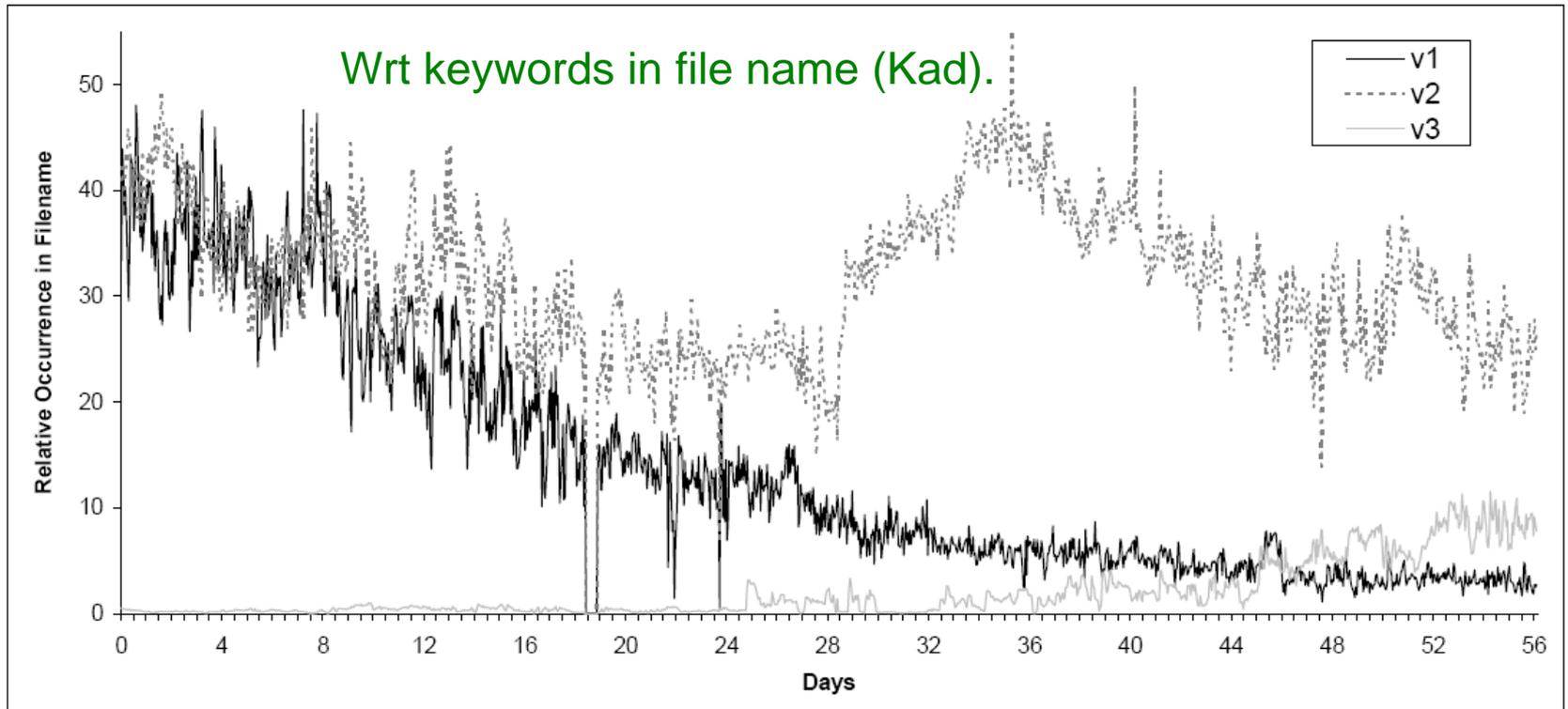
- Users can choose where to search...
- In both networks, the same countries are the most active.
- In Kad, the distribution is **more concentrated**. In particular, it is quite popular in **European countries**.

Origin of Keyword Requests (Server vs DHT)



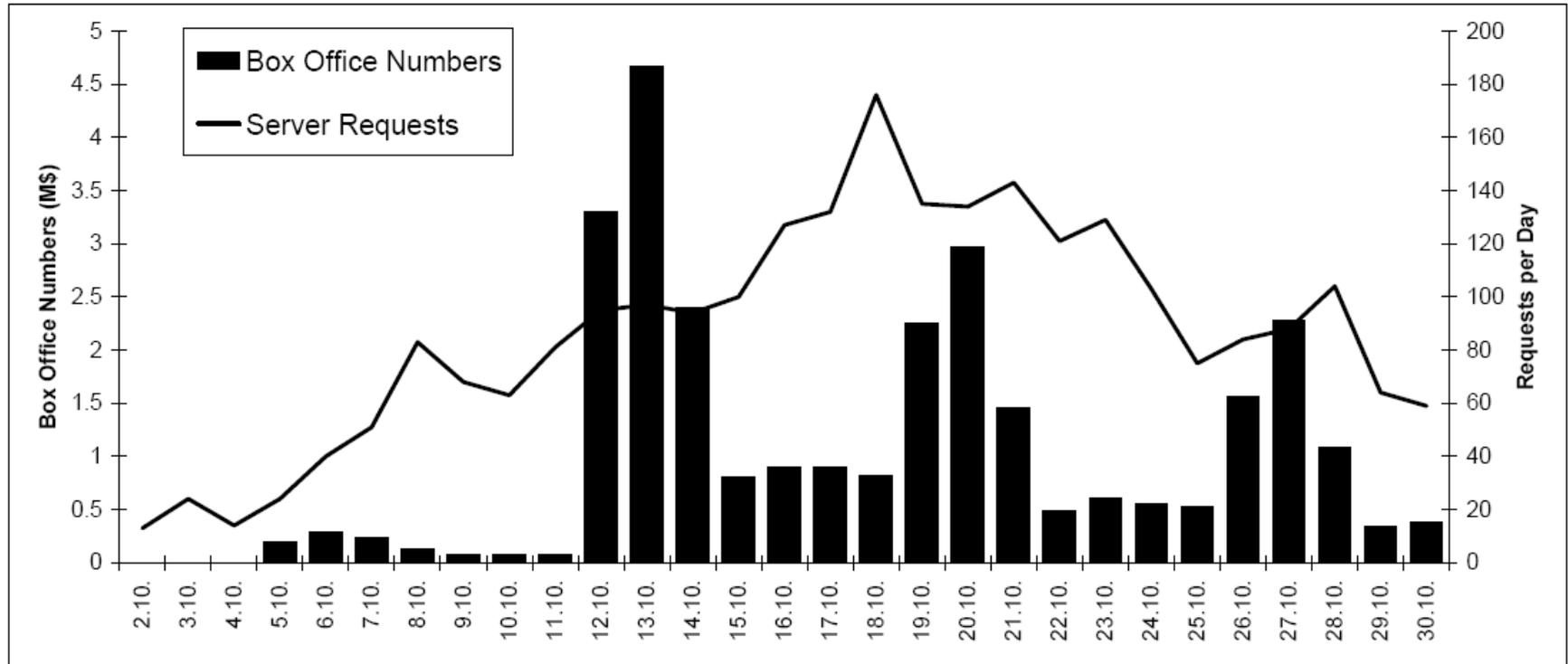
- Different countries have different **population sizes**...
- ... thus, we **normalized** the number of requests by number of Internet users in that country!
- Different picture now!
- Explanations? Because of few Internet users in **Marocco**? Because traffic is obfuscated by servers there (many requests from same IP!)?
- Popular in Europe, especially **Israel**, but not in the U.S.
- Distribution has **heavy tail**!

Search Content: Movie Quality



- Different **qualities of a movie**: no surprise, as soon as a better version is announced, users start looking for it!

Kad vs Real World



- For a specific movie, popularity in cinema and Kad exhibits a similar trend.
- with a slight delay

Open Questions / Experiments



- Is there a statistical trend **towards DHTs**?
How fast does the popularity of Kad grow?
- Given current network data, can we make **predictions** about real developments? (cf also Google trends)
- Demographic / political / sociological trends:
E.g., is there a relation between **political regimes** and usage of centralized vs decentralized computing?
- Cultural developments:
Which countries are interested in the **culture** (music, movies, ...) of each other?

What Else Can I Find in the Paper?

- More information on the measurement environment
- Discussion on representativeness of data
- Interesting related work, e.g., by **Biersack and Steiner**

Thank you for your attention!

More infos on:

<http://www.cs.uni-paderborn.de/fachgebiete/fg-ti/personen/schmiste.html>

A Peer Activity Study in eDonkey & Kad

Thomas Locher¹, David Mysicka¹, Stefan Schmid², Roger Wattenhofer¹
¹ Computer Engineering and Networks Laboratory (TIK), ETH Zurich, CH-8092 Zurich, Switzerland
(locher@tik.ee.ethz.ch, dmysicka@tik.ee.ethz.ch, wattenhofer@tik.ee.ethz.ch)
² Chair for Theory of Distributed Systems, University of Paderborn, D-33102 Paderborn, Germany
schmid@mail.upb.de

Abstract—Although several fully decentralized peer-to-peer systems have been proposed in the literature, most existing systems still employ a centralized architecture. In order to compare these two paradigms, as a case study, we conduct measurements in the eDonkey and the Kad network—two of the most popular peer-to-peer systems in use today. We re-engineered the eDonkey server system and integrated two modified servers into the eDonkey network and integrated two traffic analysis servers into the Kad network in order to monitor traffic. Additionally, we implemented a Kad client exploiting a design weakness we implemented in our servers to provide insight into the ID space. The goal of this study is to provide insight into the spatial and temporal distribution of the peers' activities and also examine the searched contents. Finally, we discuss problems related to the collection of such data sets and investigate techniques to verify the representativeness of the measured data.

I. INTRODUCTION

Today's peer-to-peer (p2p) networks come in different flavors. On the one hand, there are completely decentralized systems such as the Kad network which is based on a distributed hash table (DHT) [5], [11] where both the task of indexing the content and the content itself is distributed among the peers. Other systems still rely on centralized entities, e.g., a cluster of servers takes care of the data indexing in the eDonkey network.

In Section III, several measurement results are presented, which occur without any user intervention. Our measurements show that the temporal request distributions of the two networks are very similar, exhibiting a high activity in the evening with high loads at the eDonkey servers or at the peers being popular files in Kad. We also found that both networks are predominantly used in European countries, but there are also many active users from Israel, China, Brazil, and the U.S. Section III also investigates the content shared in the two systems. *Keywords*—peer-to-peer, measurement, traffic analysis, distributed hash table.

In order to investigate various properties of eDonkey and Kad, we collected large amounts of data from both networks. For this purpose, we reverse-engineered the eDonkey server software and published two own servers which successfully attracted a considerable amount of traffic despite the fact that our servers never returned any real content. For our Kad tests, we implemented a client that is capable of spying on the traffic at any desired position in the ID space. Section II reports on the setup of our measurement infrastructure.

In Section III, several measurement results are presented, which occur without any user intervention. Our measurements show that the temporal request distributions of the two networks are very similar, exhibiting a high activity in the evening with high loads at the eDonkey servers or at the peers being popular files in Kad. We also found that both networks are predominantly used in European countries, but there are also many active users from Israel, China, Brazil, and the U.S. Section III also investigates the content shared in the two systems. *Keywords*—peer-to-peer, measurement, traffic analysis, distributed hash table.

Representativeness

- Is our data representative?!
 - Server: Obtains all kinds of requests, but there are **other servers**
 - Kad: We only obtain the requests from the **positions** we monitor
- Server
 - eMule sends all **source requests** to bot eDonkey and Kad
 - In Kad, we obtain almost all requests for a given ID
 - Thus, we can **measure the fraction of requests** at our server!
 - There is no reason why selecting servers is biased, e.g., by geography? Distribution is same as for Kad!
 - Interestingly, it's almost **around 10%!**
- Kad
 - Monitoring as many **uniformly chosen positions** as possible
 - Attention: other peers may not be distributed uniformly, though!