

A methodology for workload characterization of file-sharing peer-to-peer networks

Diêgo Nogueira, Leonardo Rocha, Juliano Santos,
Paulo Araújo, Virgílio Almeida, Wagner Meira Jr.

Department of Computer Science
Federal University of Minas Gerais - Brazil



What is peer-to-peer (P2P)?

Class of distributed applications on the Internet

- peers act as both servers and clients (*servent*)
- servents share computational resources
- particular features:
 - dual role of servents
 - totally distributed processing nature
 - dynamic nature

Why characterize P2P networks?

- growth of P2P networks (especially file-sharing)
- lack of characterization **methodologies**
- provide important information for further research

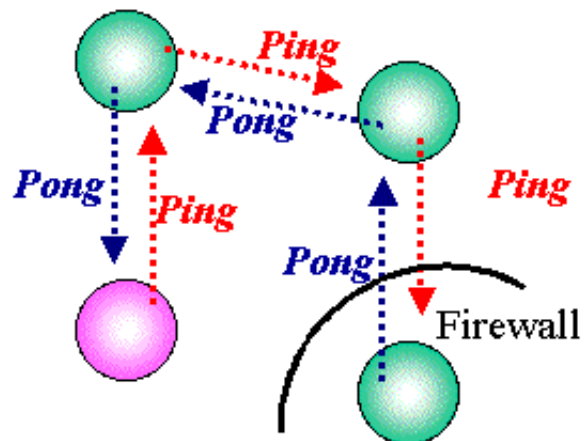
Gnutella: case study

- file-sharing open P2P network
- why Gnutella?
 - simple P2P network
 - intense traffic, with users all over the world
- previous work:
 - did not focus on standard statistical distributions
 - not in the context of general characterization methodology

Gnutella: case study

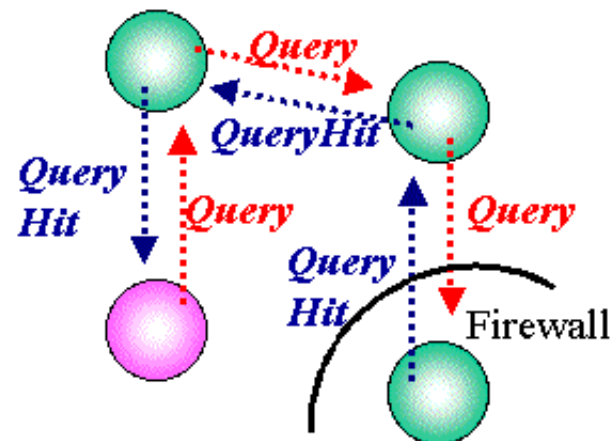
1. Ping/Pong

Who is on the network?



2. Query/QueryHit

Do you have the content?



Workload characterization methodology

- derives from the classic **client-server** characterization
- divided into:
 - Qualitative characterization
 - * conceptual definition of attributes
 - Quantitative characterization
 - * **client-side** criteria
 - * **server-side** criteria

Qualitative characterization - (1/2)

Conceptual definition of the attributes:

P2P architecture:

- type of resources shared
- communication protocol (connection + service interface)

P2P application:

- set of messages that implement the protocol

Qualitative characterization - (2/2)

P2P network:

- application implemented by the servants
- set of servants
 - the resources shared by the servant
 - the servant's neighborhood

Gnutella qualitative characterization

Gnutella architecture

- Shared resource: storage device (hard disk)
- Communication protocol:
 - connection interface: {PING, PONG}
 - service interface: {QUERY, QUERYHIT, DOWNLOAD, PUSH}

Gnutella application

Gnutella network (gNet)

Quantitative characterization

Workload characterization of a live P2P network

- collection of traffic and peer behavior data, analysis of the data

Client-side criteria

- demand for resources
- interaction pattern
- servers' connectivity

Server-side criteria

- resource availability
- service capacity

Gnutella quantitative characterization

- Data collector
 - developed over **Gnut**
 - * collected data addressed to and through peer
 - * periodically sent random QUERYS
- Experiments
 - 2 Linux workstations connected to Brazilian research network
 - connection to reference servers on Gnutella
 - results presented from 24 hours (10/02/2001)

Demand for resources characterization

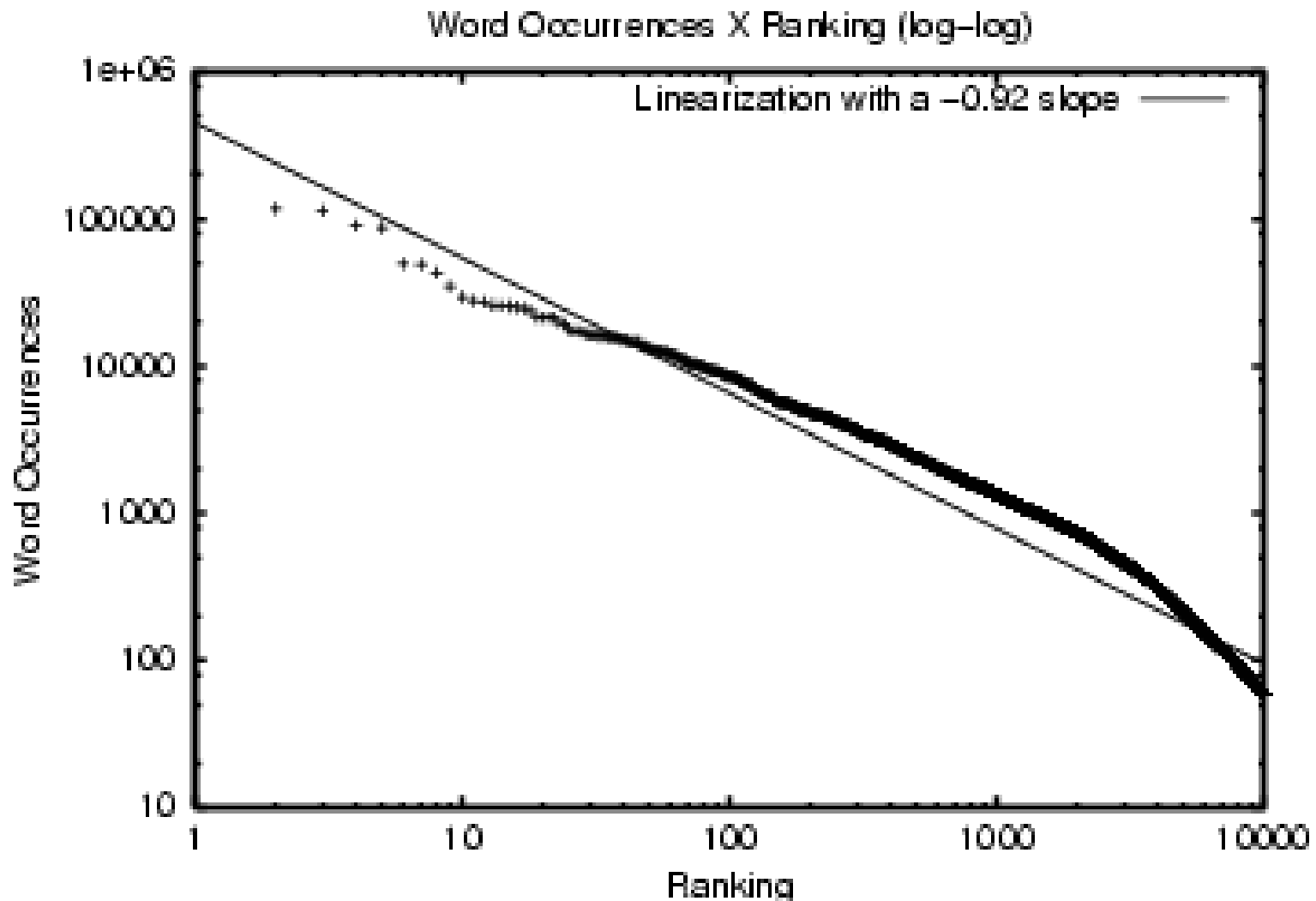
Identifies:

- the subjects of interest
- popularity of subjects among users
- temporal locality among requests

Gnutella:

- Servents' interests
 - 2,992,390 QUERYS received / 94,642 distinct words (including stop words)

Servents' interests



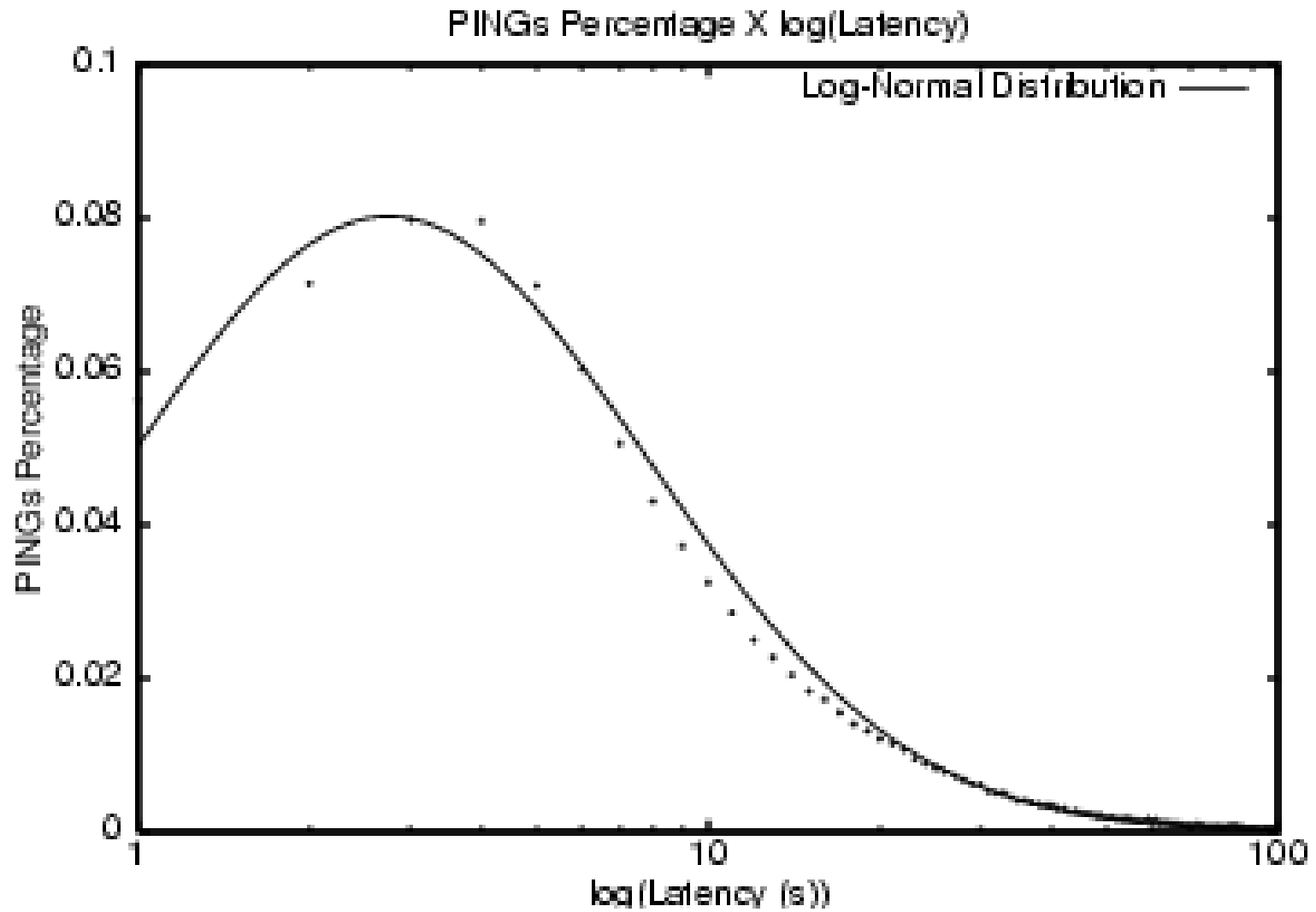
Interaction pattern characterization

- quality-of-service metric
- used to quantify the overall performance of a P2P network

Gnutella:

- Latency
 - sent TTL 1 PINGs
 - 1,823,972 registered PONGs

Latency



Servents' connectivity characterization

Quantified through:

- average number of neighbors
- network traffic associated with communication
- amount of data exchanged

Gnutella:

- Unique servents
 - number of servents varied approximately 15% across collection periods
 - 75% of peers answered

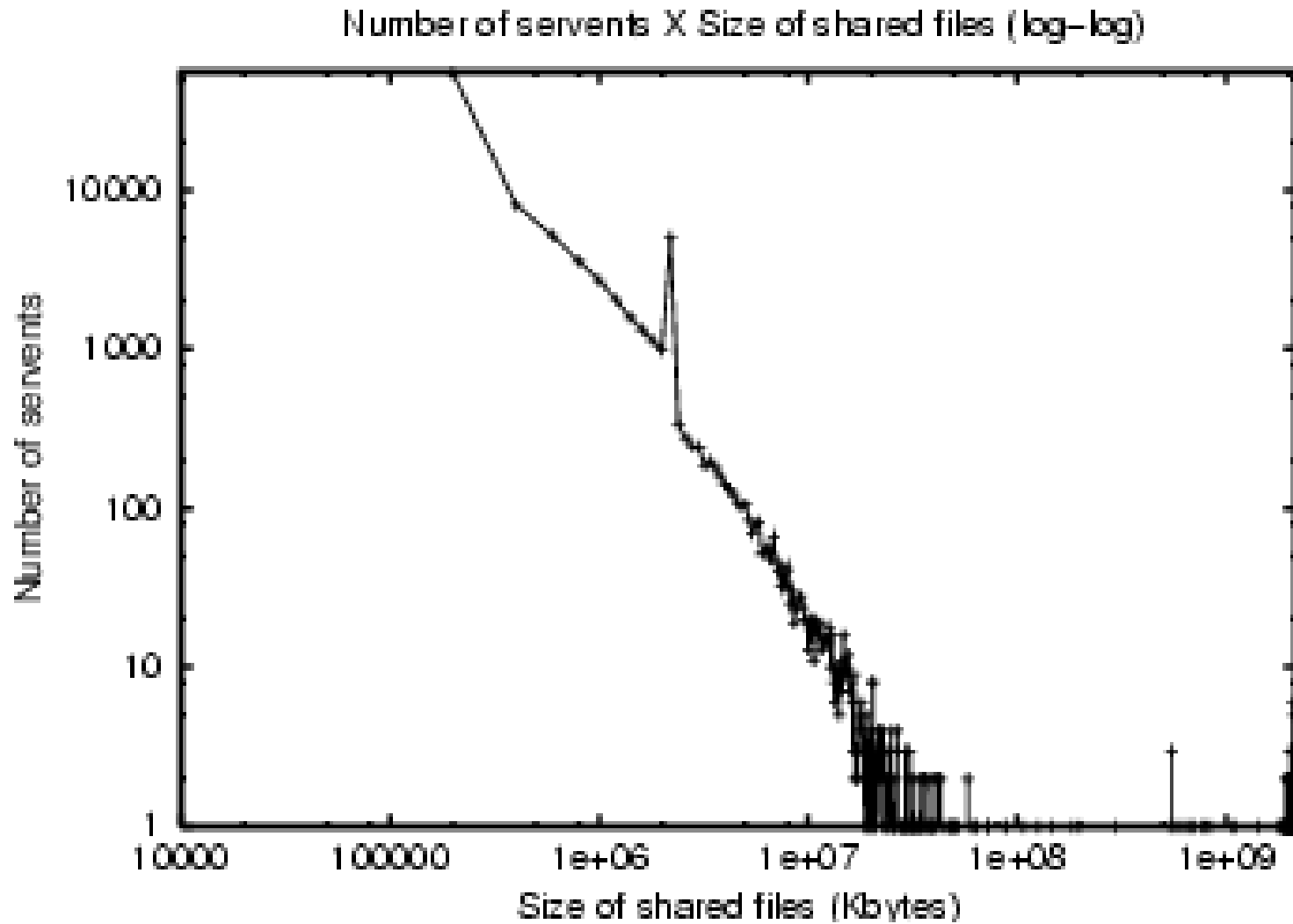
Resource availability characterization

- how dynamics of P2P affects access to information
- assess effectiveness of network in providing information
- good for comparing data distribution protocols and mechanisms

Gnutella:

- Shared kbytes
 - 120,535 addressed servents / 90,282 replied
 - information from PONG messages

Shared kbytes



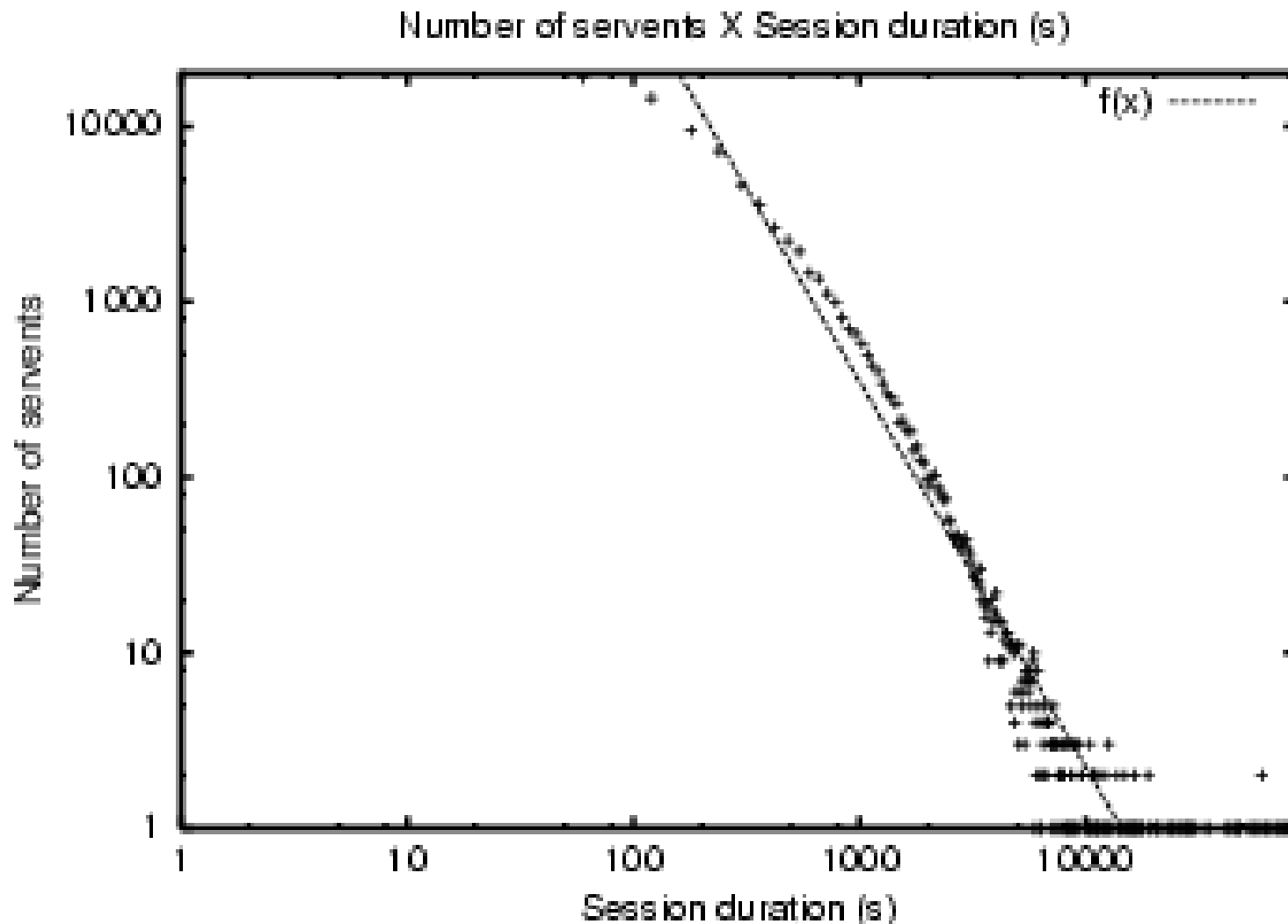
Service capacity characterization

- quantifies amount of information provided by servers
- helps to understand if idle capacity is used efficiently
- provides information to improve scalability of servers

Gnutella:

- Servers' availability
 - 98% on for at most 45 min.
 - 84% no longer than 10 min.

Servents' availability

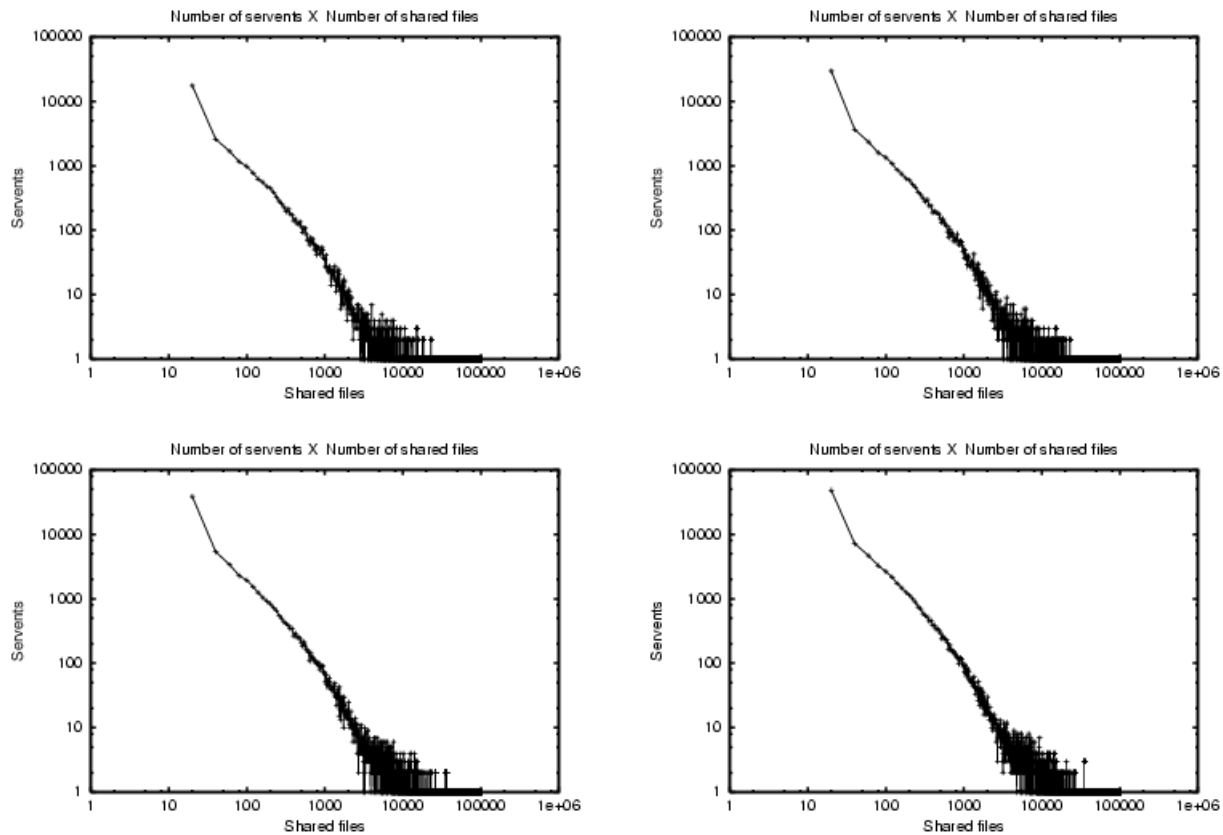


Data convergence analysis

- verify how representative the data collected and studied are
- mechanism:
 - Perform the same experiments in shorter periods
 - * 6, 12, 18 and 24 hours, for example
 - * verify the distributions from each period

Data convergence analysis

Shared files



Conclusions

Definition of a workload characterization methodology

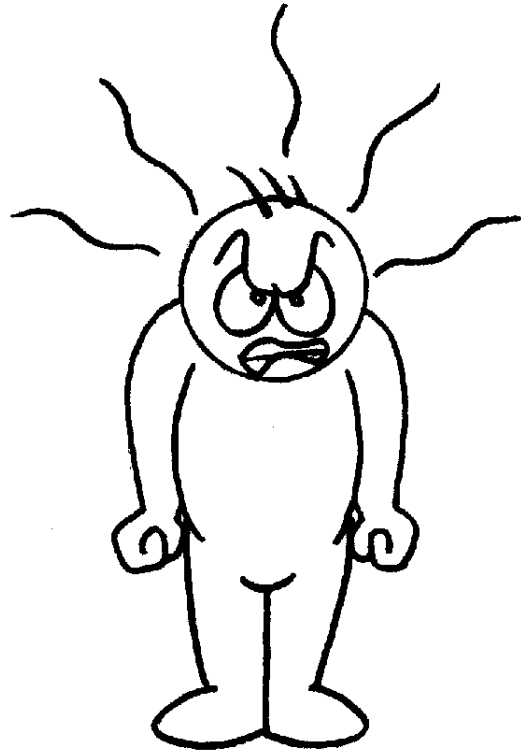
- Qualitative characterization (**conceptual definitions**)
- Quantitative characterization (**client + server side criteria**)

Successful application of methodology on Gnutella

Interesting results about Gnutella's traffic and peer behavior

- Statistical distribution analysis
- Latency distribution follows the Log-Normal distribution
- Search traffic 50 times larger than control traffic

Questions?



contact: diego@dcc.ufmg.br

Index

P2P definition

Demand for resources

Motivation

Interaction pattern

Gnutella

Servents' connectivity

Methodology introduction

Resource availability

Qualitative characterization

Service capacity

gNet qualitative characterization

Data convergence analysis

Quantitative characterization

Conclusions

gNet quantitative characterization