Peer-to-peer workload characterization: techniques and open issues

Mauro Andreolini
University of Rome “Tor Vergata”

Michele Colajanni
University of Modena and Reggio Emilia

Riccardo Lancellotti
University of Modena and Reggio Emilia

Overview of File sharing networks

- File sharing is the killer application of P2P
- Peer-to-peer systems
  - Node are peers (***Servents***)
  - Use of overlay network
- Two functions:
  - Network management and query function
  - Download
- → Two protocols

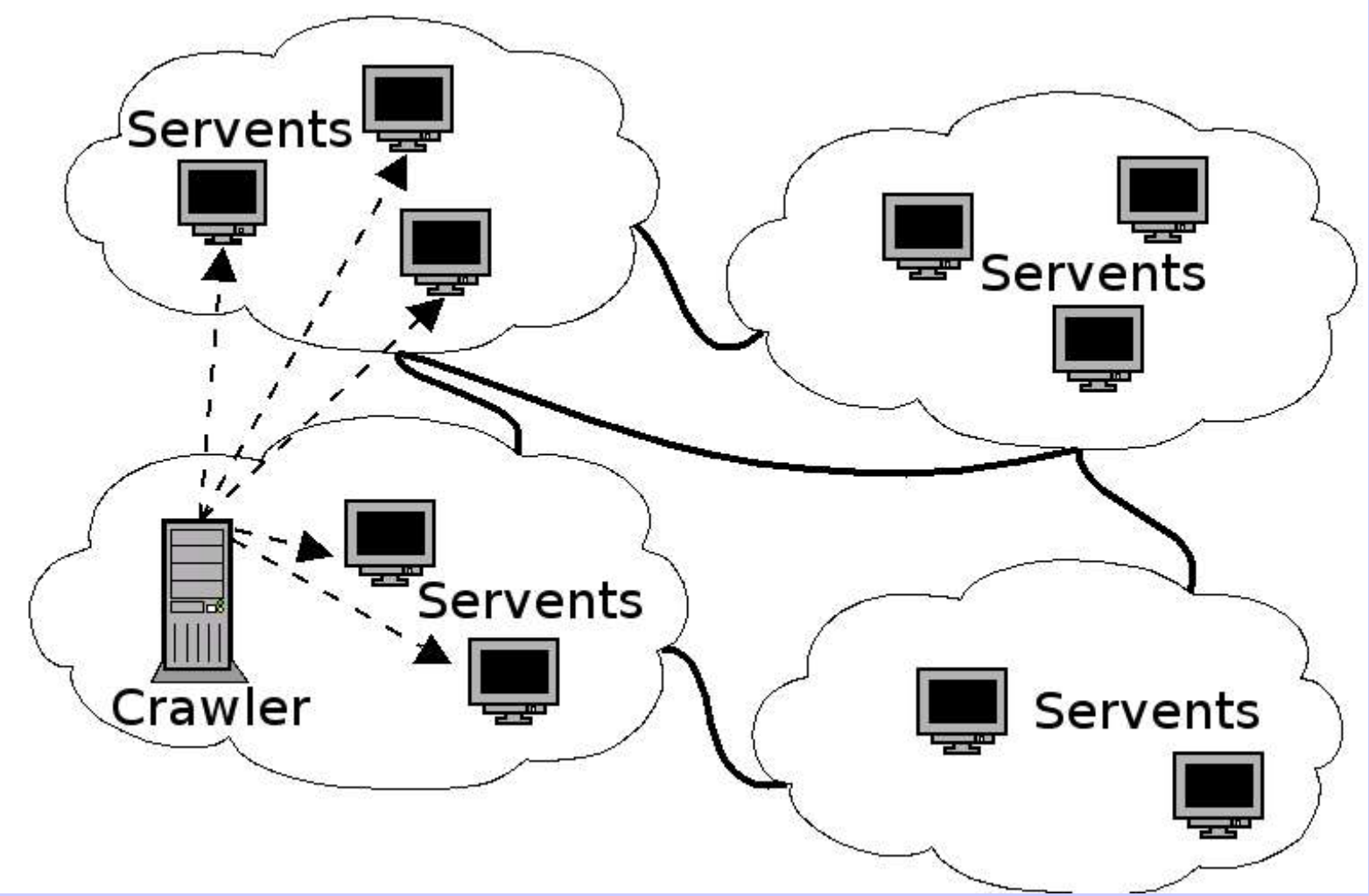
Overview of File sharing networks

- Multiple networks
- FastTrack/Kazaa
  - Closed management protocol (difficult rev. eng. [Ross])
  - HTTP-based download
- Gnutella
  - Open management protocol, O.S. servent
  - HTTP-based download

Workload characterization

- **Data of interest**
  - Resource working set
  - User behavior
  - Network structure
- **Collection techniques**
  - Active probing (crawling)
  - Passive probing (traffic interception and analysis)

Crawling



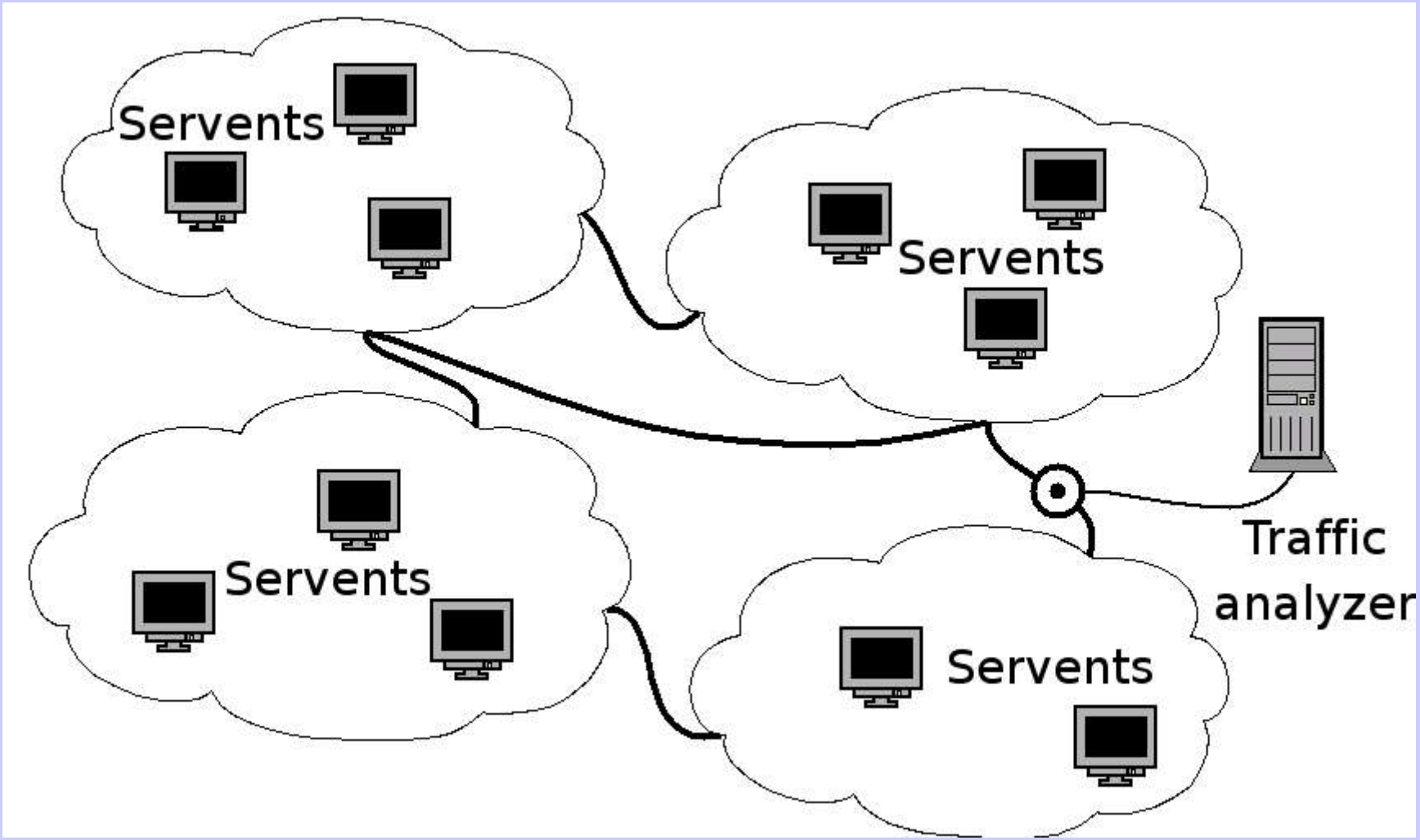
Crawling

- **Issues**
  - Queries can be rejected (e.g., “*” queries)
  - Queries can be deleted after a short amount of time (~15 min). Queries need rejuvination

Crawling

- **Pros**
  - Easy to deploy (available O.S. sw)
  - Can run from the network edge
  - Takes a snapshot of the network
  - Allows to collect interesting metadata (e.g. hash)
- **Cons**
  - Difficult to analyze dynamic aspects of the network
  - Needs open protocols
  - Difficult to detect poisoning

Traffic interception and analysis



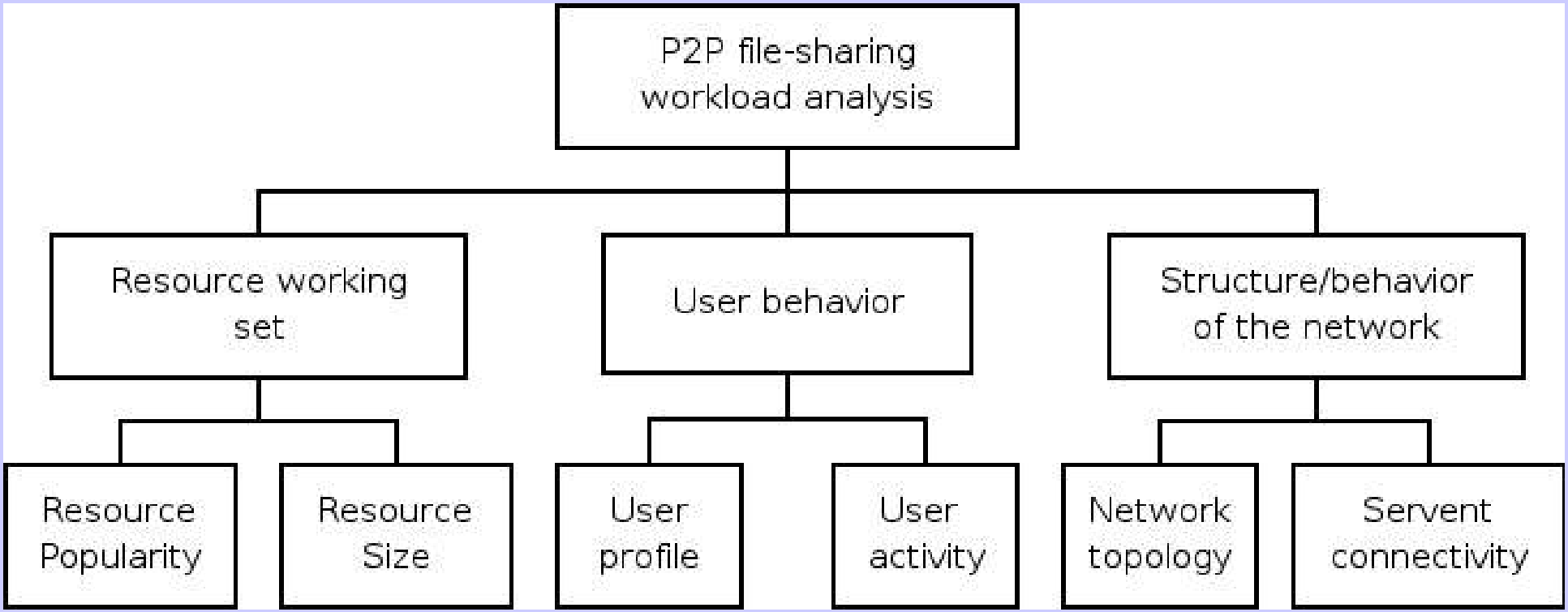
Traffic interception and analysis

- **Issues**
  - Analyze large amount of traffic
  - Capture only meaningful traffic
- **Types of meaningful traffic**
  - Download
  - Query
  - Network management

Traffic interception and analysis

- **Pros**
  - Considers actual file-sharing traffic
  - Allows the observation of dynamic characteristics of the network
- **Cons**
  - Needs representative traffic
  - Needs open protocols (mainly download traffic)

Taxonomy on file-sharing workload analysis



Analysis on resource working set

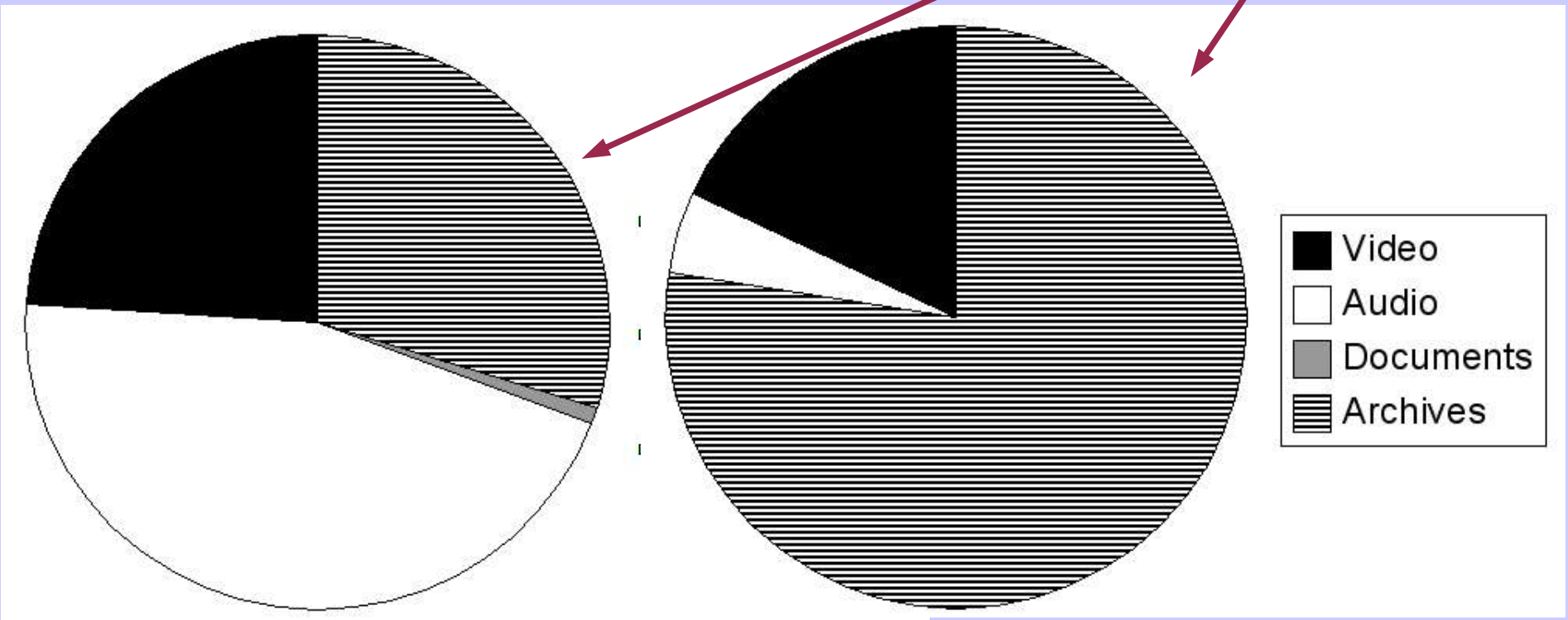
- **Studies on file popularity**
  - Resource popularity
    - [Leib] 80% of resources, 20% of downloads
    - [Andr] Zipf resource popularity
    - [Gum] Truncated Zipf popularity
  - File type popularity
    - [Leib, Andr] Audio clips most popular resource
  - Keyword popularity in shared files
    - [Makosiej] Analytical model for keyword popularity (60% files are associated with the keyword “Love”)
  - Changes of popularity rank over time
    - [Leib] 20% of files remains popular for long time

Analysis on resource working set

- **Studies on working set size**
  - Resource size in the global working set
    - [Leib] histogram of file size, 5 MB most popular size
  - Resources shared by each node
    - [Andr] analytical model of resource shared by nodes

Analysis on resource working set

- **Studies on working set size**
  - Resource size according to type
    - [Leib] correlation size/type
    - [Andr] analytical model



Analysis of user behavior

- **Definition of user profile**
  - Impact of freeloaders
    - [Tow] not always harmful
  - Download time
    - [Gum] users are patient:
      - small files: 30% > 1h, 10% ~1 day
      - large files: 50% > 1 day, 20% > 1 week
  - Aging of users
    - [Gum] After 3-4 weeks users download smaller files less frequently

Analysis of user behavior

- **User activity characterization**
  - Session length
    - [Gum] Download session
    - [Sar] Network session

    → Chunked downloads
  - Activity fraction [Gum]

	median	90-percentile
Activity fraction [Gum]	66%	100%
Download session length [Gum]	2.40 min	28.33 min
Session length [Sar]	60 min	300 min

  - Query activity
    - [Makosiej] Keywords per query, popularity of key-words in queries, types of keywords per query

Characterization of servents and of the overlay network

- **Studies on network topology**
  - Relationship between physical and overlay networks
    - [Ripe] completely different topologies
  - Topology of overlay networks
    - [Ripe, Sar] power law network
  - Impact of network topology on resilience
    - [Sar] removing 5% top nodes leads to network partition (interesting if you're interested in enforcing copyright law)

Characterization of servents and of the overlay network

- **Characterization of servent connectivity**
  - Relationship between advertised and actual bandwidth
    - [Sar] DSL-class connectivity
    - [Sar] under-advertised connectivity
  - Types of clients
    - [Sar] 15% of nodes are *Server-servent*, the remaining are *Client-servent*

Open issues

- Comparison between results obtained through crawling and traffic analysis
- Studies of local and time-related phe-nomenon impact over the network
- Improvement of packet interception analysis by means of statistical analysis (NetScope)