



Monitoring in Large Scale Networks

OUTLINE

P2P Networks

Motivations

Contributions in P2P networks

Synthesis

Future work

P2P Networks

Concepts

- ▶ Distributed Systems
- ▶ Direct exchange: peers are clients and servers

Characteristics

- ▶ Scalability (resources aggregation)
- ▶ Fault tolerance (replication)
- ▶ Lower infrastructure costs (no center)

Usage

- ▶ Initially: file sharing Napster (1999), Gnutella (2000)
- ▶ Now: many services (voice over IP, streaming,..)

Why Monitor Content in P2P Networks?

Motivation

- ▶ Lack of central control
- ▶ Autonomous peer behavior

⇒ P2P networks are vectors for and victims of malicious activities

Security issues

- ▶ Malicious content
 - ▶ illegal content diffusion: malware, paedophilia content
 - ▶ fake files and source insertion: pollution
- ▶ Privacy issues
 - ▶ attackers monitoring shared content (passive attacks)
- ▶ Denial of service issues
 - ▶ eclipse attack removing information

Our Contributions

A dual approach

- ▶ Monitoring at large scale
 - ▶ Target P2P network: KAD Network
 - ▶ Target illegal content: Paedophilia Content
- ▶ Countermeasure to protect P2P networks
 - ▶ Protection against P2P Eavesdropping (Sybil attack)
 - ▶ Protection against Pollution

KAD Monitoring

Challenge

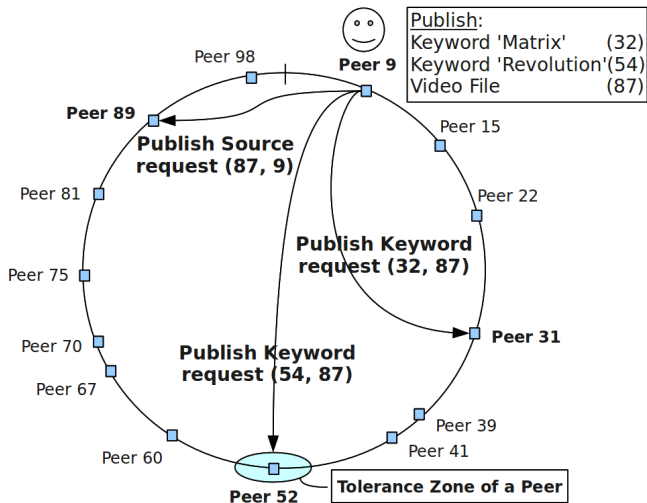
Design algorithms and a support framework to monitor the KAD network at Internet scale and to supervise content

KAD Network

- ▶ Used for file sharing
- ▶ Implemented by open source clients (eMule and aMule)
- ▶ Widely deployed (+3 million simultaneous users)
- ▶ A fully distributed P2P network
 - ▶ No server knows "Who is sharing What"
 - ▶ Each participant is responsible of a part of the overall indexation of content

KAD Network

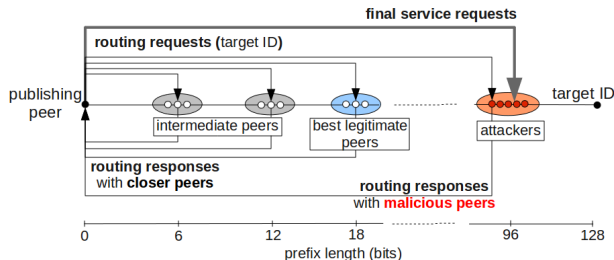
- Double indexation : keywords \rightarrow files \rightarrow sources



A KAD Monitoring Architecture

Achievements [AIMS'09, ICC'10]

- ▶ A novel monitoring algorithm scheme bypassing existing protections :
 - ▶ place peers (to 20) close to the target ID
 - ▶ attract all searches and publications for this ID
- ▶ An operational and distributed monitoring architecture



Application to Paedophilia Content

Goal

- ▶ Understand and characterize paedophilia activities
- ▶ Large scale application

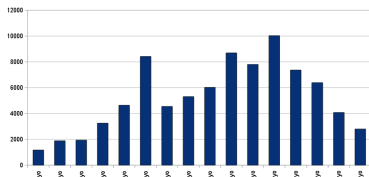
Context

Project ANR MAPE in collaboration with LIP6

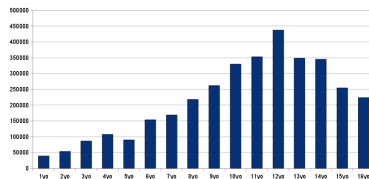
Experiment

- ▶ Passive monitoring of 75 keywords: age (1yo-16yo), pedo (pthc, hussyfan, qqaazz), pedo-linked (boy, girl, dad) or normal
- ▶ One two-weeks experiment: 28GB collected in High Security Lab (158M files, 36M publications, 12.8M peers, etc)
- ▶ Five probes per keyword, 360 deployed probes, 35 machines PlanetLab

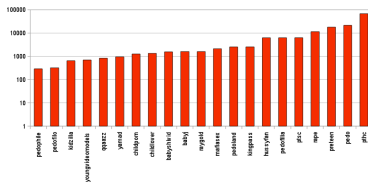
Paedophilia Content: Some Results



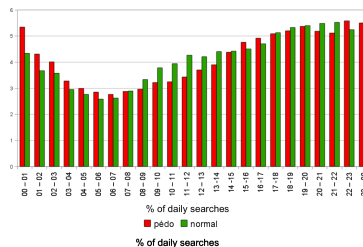
Searches Number and Ages



Publications Number and Ages



Searches Number and Keywords



% of daily searches

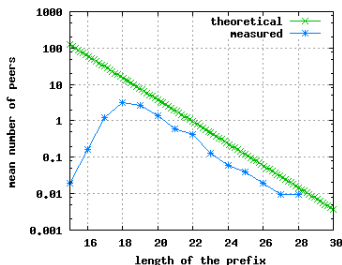
pedo normal

% of daily searches

Self-Protecting Mechanisms Against P2P Eavesdropping

Achievements [HotP2P'10,
Springer's PPNA Journal
(under minor revision)]

- ▶ A new metric for sybils detection
 - ▶ ID Distribution-based
 - ▶ Kullback-Leibler Divergence test for attack detection
- ▶ A lightweight fully distributed protection scheme
- ▶ Simple to implement countermeasure



$$D_{KL}(M \mid T) = \sum_i M(i) \log \frac{M(i)}{T(i)} \quad (1)$$

Content Pollution Quantification

Context

Project GIS 3SGS ACDAP2P in collaboration with UTT (University of Technology of Troyes)

Achievements [P2P'11 (short-19%)]

- ▶ Index falsification
- ▶ A new pollution measurement metric for KAD based on Tversky index
- ▶ Collection of popular files (2000) and pollution quantification:
 - ▶ 41% of files infected by index falsification
 - ▶ More than 20% by index poisoning

$$\text{Pollution}(X) = 1 - \frac{\sum_i \text{Similarity}(X, Y_i)}{i} \quad (2)$$

$$\text{Similarity}(X, Y) = \frac{2 * |X \cap Y|}{|X| + |Y|} \quad (3)$$

- ▶ X: keywords composing the desired filename
- ▶ Y_i : keywords composing a filename retrieved from a source i.

Filename: Indiana Jones Et Les Aventuriers ... FileID: 7B9F403468CD821C38885E7777153C1C Number of responding sources: 175	
Found filenames	#
Xxx Marc Dorcel - Russian Institute Lesson 1	4
The Best Of The Doors.rar	2
[DIVX-ITA]-Disney Pixar-Wall-E-2008-Italian...	1
...	...

Synthesis of our contributions

An efficient approach to monitor content in P2P networks

- ▶ Validation at large scale: supervise paedophilia content

An original method to detect attacks by comparing real peers'ID distribution to the theoretical one

- ▶ Effective solution: simple, no additional cost and backward compatibility
- ▶ Implementations in KAD and in gtk-gnutella (Raphael Manfredi)

Detection of a new type of pollution

- ▶ Quantification of pollution: 2000 files, 2/3 polluted

Ongoing and Future Work

Monitoring other structured P2P networks (short term)

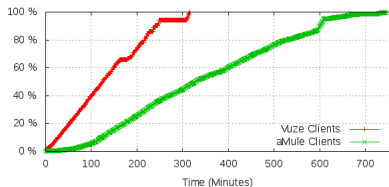
Monitoring anonymous P2P networks (medium term)

Monitoring information centric networks (long term)

Monitoring other structured P2P networks

Assessment vulnerabilities on BitTorrent's DHTs and Countermeasures

- ▶ Mainline DHT Analysis [NTMS'11]
- ▶ Vuze DHT Analysis



Multiprotocol cooperation in P2P networks

- ▶ KAD: a great DHT for indexing of content publicly available in P2P
- ▶ BitTorrent: a fast and reliable transport protocol maximizing availability

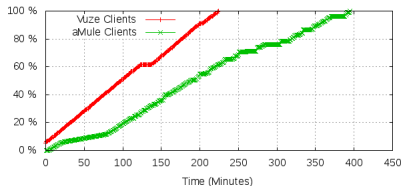
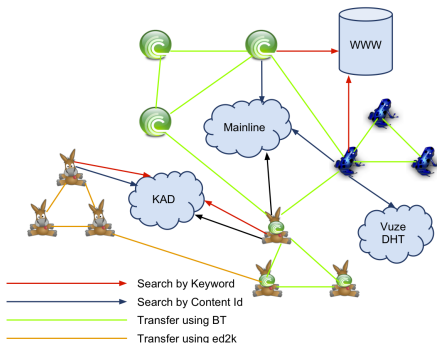


Figure: Download Times with 1 and 10 Sources

Multiprotocol cooperation in P2P networks



⇒ Provide an efficient client that indexes content using KAD and transfers it using BitTorrent and/or ed2K while being fully compatible with both networks [HotP2P'11].

⇒ Design an incentive scheme that ensures fairness
[Collaboration with University of Buenos Aires - D. Vicino/C. Righetti]

Monitoring anonymous networks

- ▶ Anonymous networks offer an excellent support for preserving users privacy but also for illegal activities

Our target: the I2P network

- ▶ formed by a group of routers - a piece of software to connect to the I2P network
- ▶ a growing network with currently 10000 routers in average
- ▶ fully distributed using a DHT-like approach.
- ▶ not designed for routing traffic outside the network (few out-proxies).
- ▶ support I2PSnark, a built-in BitTorrent client

⇒ Monitoring such a network is very challenging [TMA'12 (short paper)]: IP2's characterization, vulnerabilities detection.....

Monitoring content in Information Centric Networks

- ▶ Now network is more an information distribution system than an end-to-end communication system
- ▶ Emergence of a new paradigm : Content Centric Networks [Van Jacobson 2009]
 - ▶ Telephony \rightarrow Internet \rightarrow CCN
 - ▶ Wires \rightarrow Hosts \rightarrow Content

\Rightarrow How to monitor content in ICN: data collection, definition of metrics, topology inference...