Content Pollution Quantification in Large P2P networks - a Measurement Study on KAD -

Guillaume Montassier Thibault Cholez Guillaume Doyen
Rida Khatoun Isabelle Chrisment Olivier Festor

University of Technology of Troyes, ERA team
LORIA-INRIA Nancy, MADYNES team

August 31th 2011

Nancy-Université Université Henri Poincaré

utt université de technologie Troyes

INRIA

Context

P2P networks: target & support of malicious activities

- Weakness: no central authority & malicious peer's behaviors
- Pollution decreases P2P QoE
- Pollution can be used to spread malicious content (paedophile content, malware, etc.)

Contributions

- Identification of a new form of pollution
- Design of detection mechanisms
- Quantification through large scale data measurements

Example of a clean file

General
Full Name : The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.ENG.-[...]
Hash : C0F8BFA37E0DD0A4585CD3B90B9F4D26
Filesize : 185108504 bytes (176.53 MB) Partfilestatus : Waiting

Transfer
Found Sources : 106 Transferring Sources : 0

File Names

File Name	Sources
The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.ENG.-.sub.FR...	22
The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.VOSTFR.HDTV...	13
The.Big.Bang.Theory.S04E09.VOSTFR.HDTV.XviD-TheOdusseus.avi	4
The Big Bang Theory 4x09 The Boyfriendplexity Vostfr Hdtv Xvid-Th...	3
The.Big.Bang.Theory.S04E09.VOSTFR.HDTV.XviD.avi	2
409 The.Big.Bang.Theory.4x09.The.Boyfriend.Complexity.VOSTFR.H...	1
The Big Bang Theory 4x09 The Boyfriendplexity Eng - Sub Fr Hdtv Xvi...	1

Figure: Example of consistent filenames retrieved from the responding sources for a clean file

Example of a polluted file

General
Full Name : Indiana Jones Et Les Aventuriers De L'Arche Perdue-Fr-D[...]
Hash : 7B9F403468CD821C38885E7777153C1C
Filesize : 925913600 bytes (883,02 MB) Partfilestatus : Waiting

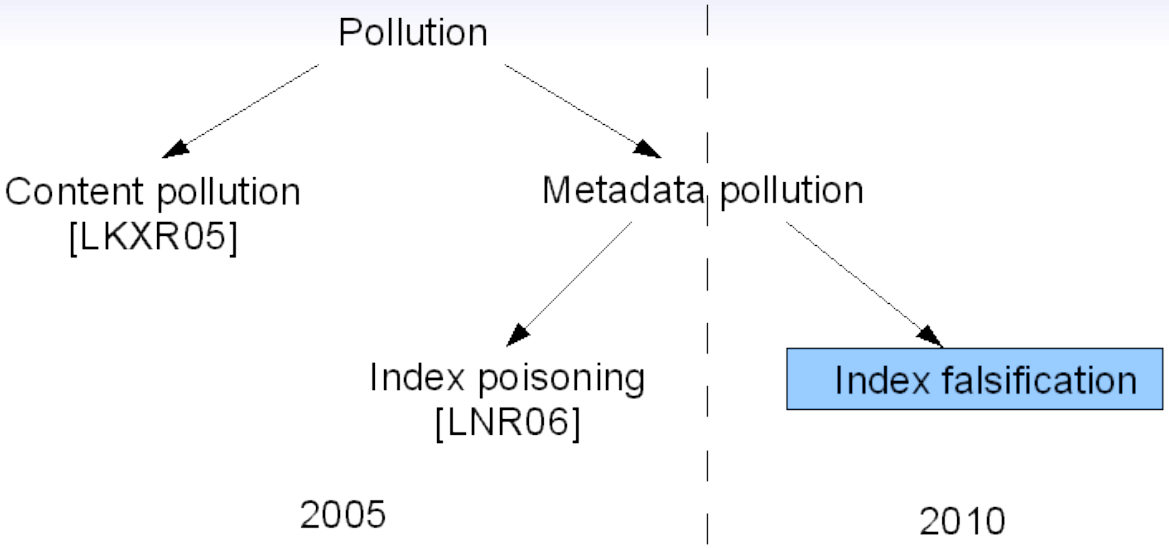
Transfer
Found Sources : 229 Transferring Sources : 0

File Names

File Name	Sources
Audio Libro. El Quijote Parte I (Completo-Audiolibro (Spanish)(Vo...	1
avatar(1).avi	1
Batman Begins Espanyol Dvdrip Spanishshare By Cdgroup.mpg	1
cars(2).avi	1
Casino No Limit (啄木鸟 赌场风云-13女星共演的梦幻巨作)[CD2].avi	1
Cine Belico - Los Panzers De La Muerte.avi	1
Culturismo - Arnold Schwarzenegger Total Rebuild.mpg	1

Figure: Part of conflicting filenames retrieved from the responding sources for a polluted file

Pollution



The index falsification pollution

- Specific to KAD's double indexing scheme
- Consists in indexing a file through many unrelated filenames
- Wrong filenames are made popular
- Very harmful: can link regular names to malicious contents

Similarity metric for pollution detection

Tversky similarity coefficient

- X and Y two sets of keywords
- X: keywords composing the desired filename
- Y: keywords composing a filename retrieved from a source

$$S(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + \alpha * |X - Y| + \beta * |Y - X|} \in [0,1] \quad (1)$$

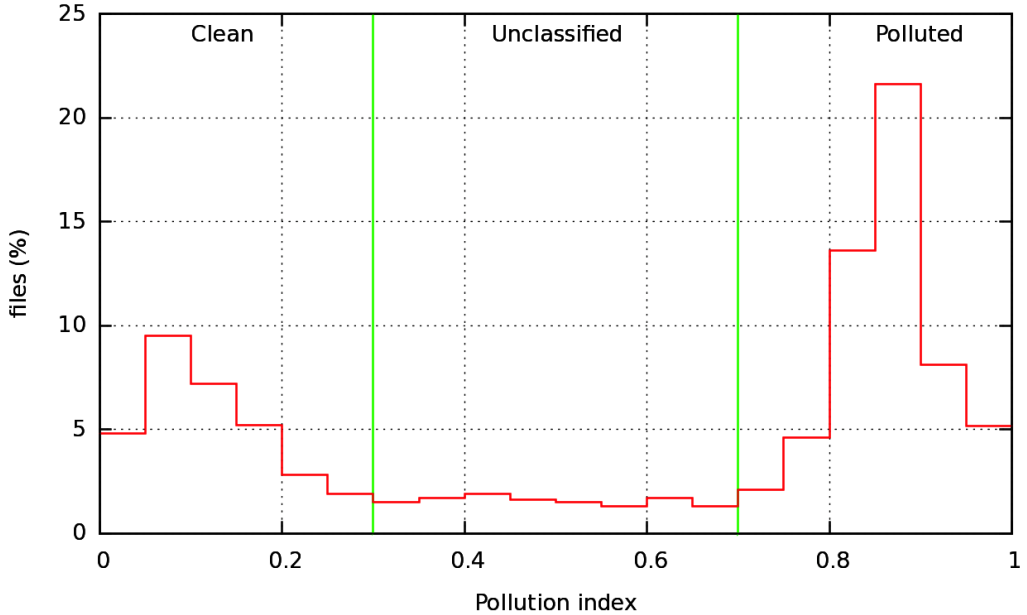
Pollution coefficient

- Pollution coefficient P computed for each file X
- From all filenames Y_i retrieved from the sources

$$P(X) = 1 - \frac{\sum_{i=1}^{n} S(X, Y_i)}{n} \quad (2)$$

Metric's parameters and validation

- Experts evaluate the different filenames and tag files as polluted, clean or unclassified
- Parameters to best match experts' votes:
  $P(X) < 0.3 =$ clean; $P(X) > 0.7 =$ polluted
- Error rates: false positives: 3.5%, false negatives: 0.8%
- False positives: because of localized names



Data collection

Large data collection

- Based on a top 100 of the most downloaded contents in 2010
- From a major BitTorrent indexation website[a] (> one hundred million searchs a year)
- For each content: investigation of the 20 most popular files
- Collection of all existing filenames for 2000 popular files

[a] `http://www.kickasstorrents.com/`

Filenames retrieval

- External to the KAD DHT
- Need one TCP connection by source
- 300s to find almost all sources

Quantification of KAD's pollution

Type	Quantification (%)
No responding source (index poisoning)	20.5
Polluted (index falsification)	41.1
Clean	28.6
Unclassified	9.6

Table: Global pollution quantification

Contents	Quantification (%)
Child pornography	8.8
Pornography	55.7
Other	35.3

Table: Content terms found in index falsification

Conclusion

Our detection metric

- Is efficient: low error rate and computional cost
- Can be applied online

Quantification of pollution

- KAD popular contents are highly polluted ($> 41\%$) by index falsification
- Pollution can spread harmfull contents

Future work

- Time evolution of pollution
- Countermeasures against index falsification

Cristiano Costa and Jussara Almeida.
Reputation systems for fighting pollution in peer-to-peer file sharing systems.
In *P2P '07: Proceedings of the Seventh IEEE International Conference on Peer-to-Peer Computing*, pages 53–60, Washington, DC, USA, 2007. IEEE Computer Society.

Thibault Cholez, Isabelle Chrisment, and Olivier Festor.
Efficient DHT attack mitigation through peers' ID distribution.
In *Seventh International Workshop on Hot Topics in Peer-to-Peer Systems - HotP2P 2010*, Atlanta USA, 04 2010. IEEE International Parallel & Distributed Processing Symposium.

Hun Jeong Kang, Eric Chan-Tin, Nicholas Hopper, and Yongdae Kim.
Why kad lookup fails.
In *Peer-to-Peer Computing 09*, pages 121–130, Atlanta, USA, 09 2009. IEEE.

Jian Liang, Rakesh Kumar, Yonjian Xi, and Keith W Ross.
Pollution in p2p file sharing systems.
In *INFOCOM*, pages 1174–1185. IEEE, 2005.

Thomas Locher, David Mysicka, Stefan Schmid, and Roger Wattenhofer.
Poisoning the Kad Network.
In *11th International Conference on Distributed Computing and Networking*, Kolkata, India, 01 2010.

J. Liang, N. Naoumov, and K. W. Ross.
The Index Poisoning Attack in P2P File Sharing Systems.
In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications*, pages 1–12. IEEE, 2006.

Moritz Steiner, Taoufik En-Najjary, and Ernst W. Biersack.
Exploiting kad: possible uses and misuses.
SIGCOMM Comput. Commun. Rev., 37(5):65–70, 2007.

Moritz Steiner, Taoufik En-Najjary, and Ernst W Biersack.
A global view of kad.
In *IMC 2007, ACM SIGCOMM Internet Measurement Conference, October 23-26, 2007, San Diego, USA*, 10 2007.

Kyuyong Shin, Douglas S. Reeves, Injong Rhee, and Yoonki Song.
WINNOWING : Protecting P2P Systems Against Pollution By Cooperative Index Filtering.
Tech report TR-2009-2, Department of Computer Science, North Carolina State University, 2009.

Peng Wang, James Tyra, Eric Chan-Tin, Tyson Malchow, Denis Foo Kune, Nicholas Hopper, and Yongdae Kim.
Attacking the kad network.
In *SecureComm '08*, pages 1–10, New York, NY, USA, 2008. ACM.