The Index Poisoning Attack in P2P File Sharing Systems

Shumanski, Andrei
Trigonakis, Vasileios

Agenda

- P2P File Sharing
- Systems under Evaluation
- Types of Attacks
- Data Gathering Methodology
- Measurements & Results
- Conclusions

P2P File Sharing

- One of the most important applications in the Internet

↓

- Huge cost for the “copyright industry”

↓

- **Sharing systems under attack**

Terminology

- **Title** is a specific song or video
- A given title can have many different **versions**
- Each version has one **identifier** (hash of the version)
- Multiple **copies** of identical versions in the system
- **Advertisements** about the copies
- **Keyword search** is used

Agenda

- P2P File Sharing
- Systems under Evaluation
- Types of Attacks
- Data Gathering Methodology
- Measurements & Results
- Conclusions

Systems under Evaluation

- **Overnet**:
  - used in eDonkey2000
  - DHT-based file sharing system

- **FastTrack**:
  - two-tier unstructured file sharing system
  - index distributed over a small fraction of the nodes

Agenda

- P2P File Sharing
- Systems under Evaluation
- Types of Attacks
- Data Gathering Methodology
- Measurements & Results
- Conclusions

Types of Attacks

- **Pollution attack**: corrupting the targeted content, rendering it unusable, and then making this polluted content available for sharing in large volumes.
  - Resource intensive attack

- **Index poisoning attack**: inserting massive numbers of bogus records into the index. (i.e. randomly chosen file identifiers)
  - Structured & unstructured systems
  - Non resource intensive attack

The Index Poisoning Attack

- Typically, **no authentication** for the files' advertisements
- Attack by falsely advertising copies of the targeted titles

- **Possible types**:
  - non-existing, random ids (mostly used)
  - non-existing IPs
  - unavailable service port numbers

Index poisoning attack in FastTrack

- **Decentralized** & **unstructured** (two-tier)
- Two classes of nodes:
  - Ordinary Nodes (ONs)
  - Super-Nodes (SNs)
- SN overlay – long-lived TCP connections
- Index kept by the SNs

- **Attack by**:
  - inserting bogus records into the indexes of the SNs

Index poisoning attack in Overnet

- Based on **Kademlia**
- All nodes equal
- UDP messages
- Version ids & keyword hashes stored

- **Attack by**:
  1. defining the target keywords and hash them
  2. random id, not derived by some existing file
  3. periodically refresh this information

Agenda

- P2P File Sharing
- Systems under Evaluation
- Types of Attacks
- Data Gathering Methodology
- Measurements & Results
- Conclusions

Data Gathering Methodology

- Downloading of files too expensive
- **Solution**:
  1. **Harvesting**: collect the version ids and publisher node data & create a list of the advertised versions and a list of the distinct copies of each version. Done by:
     - **FastTrack**: a crawler
     - **Overnet**: inserting a node in the DHT with the target keywords hash as id
  2. Classify the versions (**clean, polluted, poisoned**)
  3. Determine the pollution and poison levels for the versions and copies

Classifying the Versions

- **Observation**: "*Among the users that have at least one version of the title, the large majority of users advertise at most a few versions* (**Light users**) *and a relatively small number of users advertise a large number of versions* (**Heavy users**)*.*"
- **Heuristic**:
  - V → set of all the advertised versions
  - V_H → by heavy users
  - V_L → by light users
  - $V_X = V_H \cap V_L$ → **polluted versions**
  - $V_H^* = V_H - V_X$ → **poisoned versions**
  - $V_L^* = V_L - V_X$ → **clean versions**

Poisoning & Pollution Levels

- **poisoning**:
  $|V_H^*| \ / \ |V|$
- **pollution**:
  $|V_X| \ / \ |V|$
- **clean**:
  $|V_L^*| \ / \ |V|$

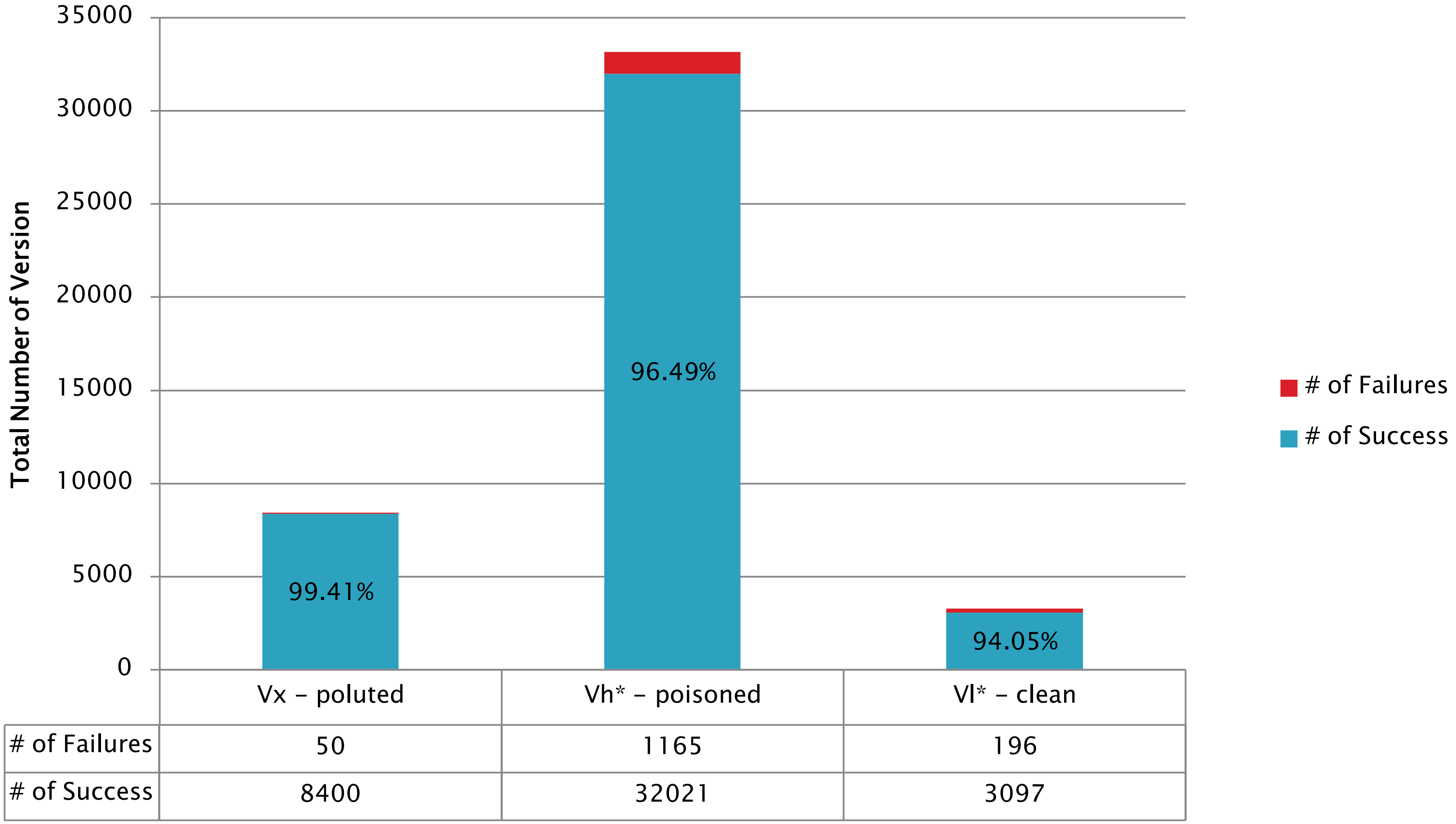
Version Levels

- **poisoning**: $\frac{\sum_{u \in V_H^*} |C_u|}{\sum_{u \in V} |C_u|}$
- **pollution**: $\frac{\sum_{u \in V_X} |C_u|}{\sum_{u \in V} |C_u|}$
- **clean**: $\frac{\sum_{u \in V_L^*} |C_u|}{\sum_{u \in V} |C_u|}$

C_u is the set of copies for version u

Copy Levels

Evaluation of the Heuristic



	Vx – poluted	Vh* – poisoned	Vl* – clean
# of Failures	50	1165	196
# of Success	8400	32021	3097

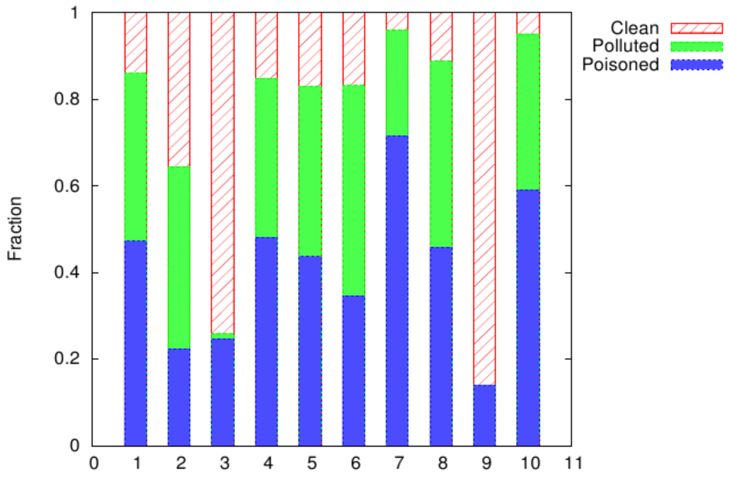
Agenda

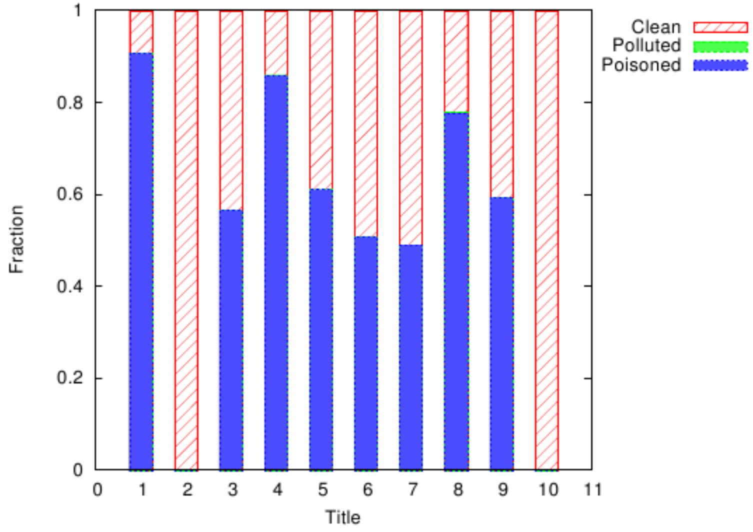
- P2P File Sharing
- Systems under Evaluation
- Types of Attacks
- Data Gathering Methodology
- Measurements & Results
- Conclusions

Measurements & Results

- **FastTrack**:
  - 38.97 copies per user
  - 8683 decoy users from 624 IPs
  - Decoyers are 7% of all users but provide 77% of all copies and 73% of all versions
- **Overnet**:
  - 11 copies per user
  - 27 decoy users from 26 IPs
  - Most of the versions and copies are provided by decoyers

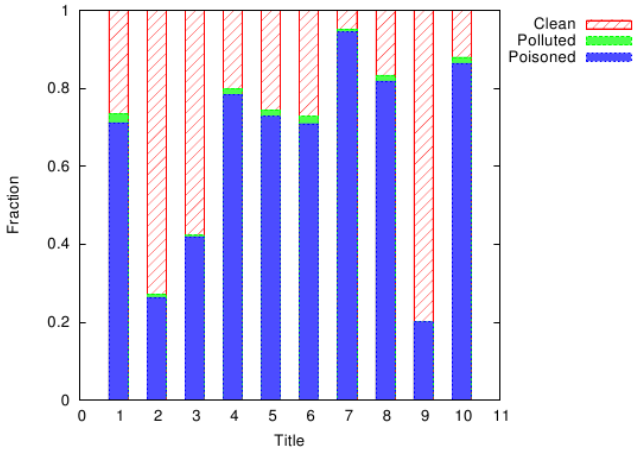
Mesurements & Results – Copies

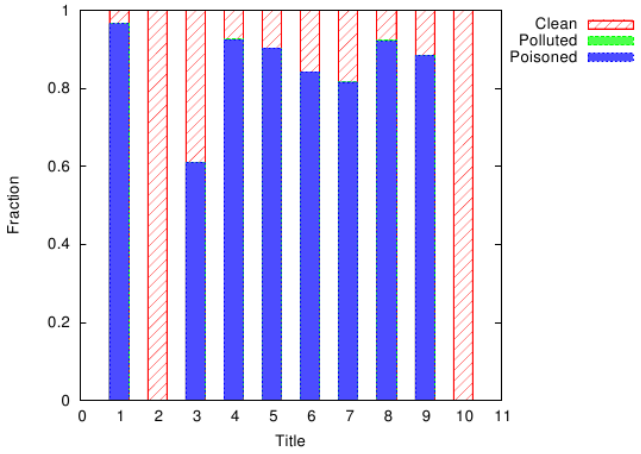




- There are different companies and techniques
- Total decoy percentage is from 50% to 95%
- Little pollution in Overnet

Mesurements & Results – Versions





- Majority of versions are poisoned
- Versions poison level is higher than copies poison level: decoyers make copies of polluted version, copies of poisoned versions do not circulate

DHT Vulnerabilities to Poisoning

- **Node insertion attack**: Overnet – can prevent users from finding clean versions

- **Poisoning**: DHT vs. Unstructured
  - Small # of titles → DHT requires less resources
  - Increasing # of titles → eventually, DHT requires more resources

- **DDoS attack** by exploiting DHT
  - pointing one node

Defending against Poisoning Attack

- **Overview of Solutions**:
  - Rating versions and advertisements – **forums**
  - Rating sources – **blacklists** of IP ranges based on reputation

Agenda

- P2P File Sharing
- Systems under Evaluation
- Types of Attacks
- Data Gathering Methodology
- Measurements & Results
- Conclusions

Conclusions

- Both structured & unstructured overlays are vulnerable
- Proposed solution can detect the polluted and poisoned versions–copies with a good approximation

References

- J. Liang, N. Naoumov, KW. Ross, *The index poisoning attack in p2p file sharing systems*, IEEE INFOCOM, 2006.

The end..

Thank you ☺