Large Scale DNS Analysis

Samuel Marchal and Thomas Engel

SnT - University of Luxembourg

UNIVERSITÉ DU LUXEMBOURG

CETREL a SIX company

SnT

SNT

Outline

1. Motivations
2. What data to analyse and how?
3. Related work
4. DNSSM
5. Conclusion

SnT

Outline

1. Motivations
2. What data to analyse and how?
3. Related work
4. DNSSM
5. Conclusion

SnT

Overview of DNS

- DNS (Domain Name System) is the service that maps a domain name to its associated IP addresses

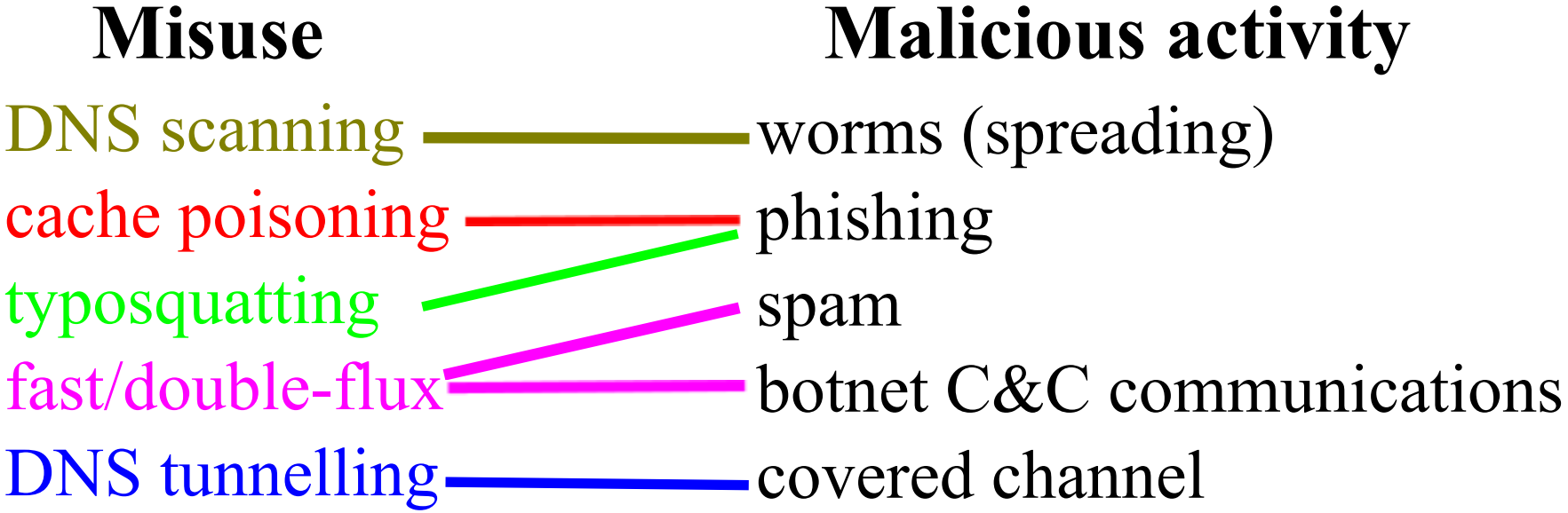
  `www.example.com` $\Longrightarrow$ `123.45.6.78`

- DNS allows to find any information about a domain :
  - A : IPv4 address
  - AAAA : IPv6 address
  - MX : Mail server
  - NS : Authoritative DNS server
  - TXT : any information

SNT

Why DNS monitoring ?

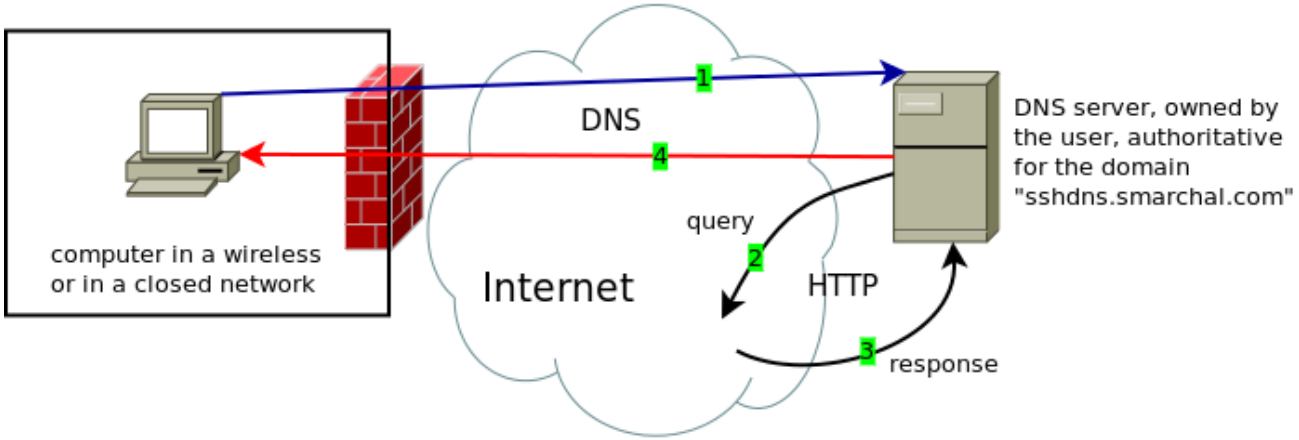
- DNS: critical and essential Internet service
- Used by attackers to enhance malicious activities



DNS Tunnel

DNS A Query :
fvrwk4tufv3damkan5ygk3ttonuc4y3pnuwgky3eonqs243imezc23tjon2h.amzygqwwgzlsoqwxmmb
ribxxazloonzwqltdn5wsyzldmrzwcIltnbqtello.nfzxi4bvgiys2y3foj2c25rqgfag64dfnzzxg2bomnxw2
ldtonuc2zdt.55153-0.id-3907.up.sshdns.smarchal.com: type A, class IN

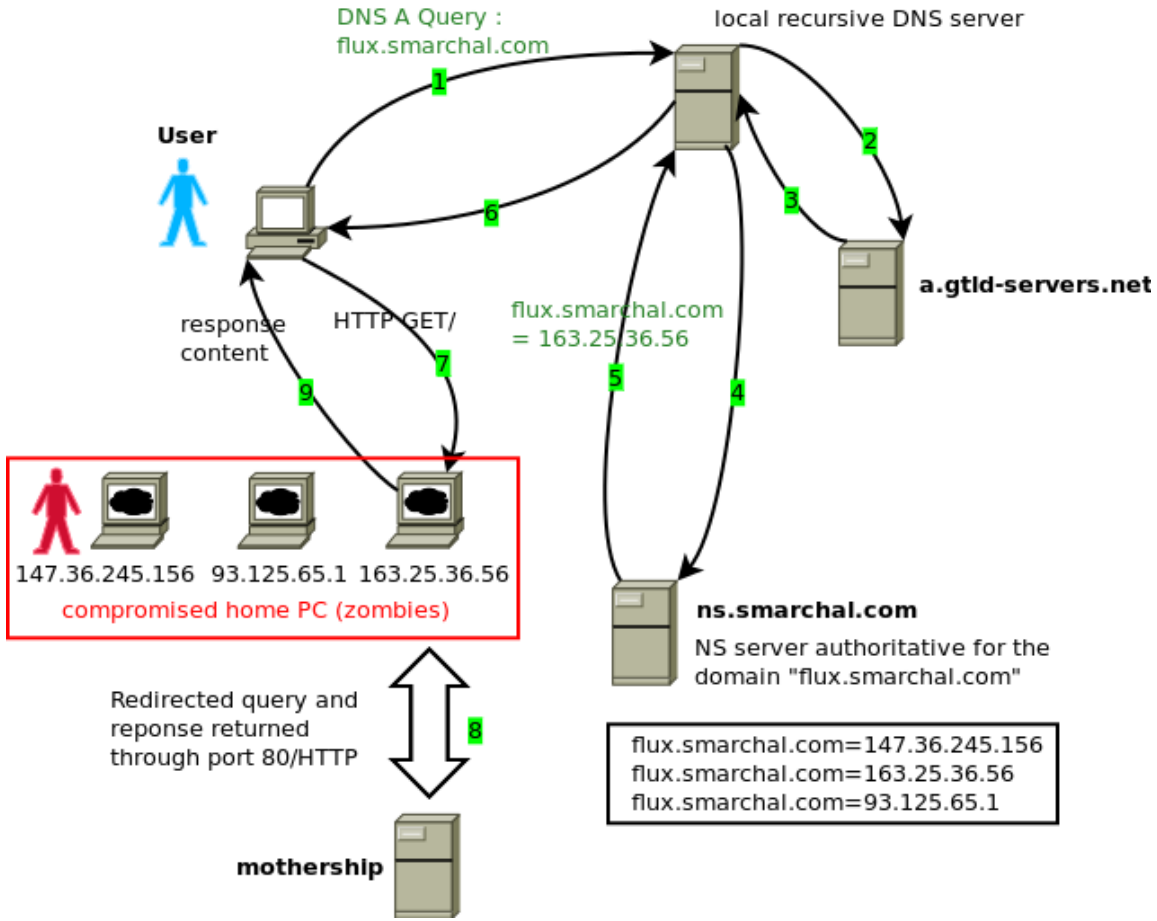
DNS TXT Query :
472-55153.id-3907.down.sshdns.smarchal.com: type TXT, class IN



DNS TXT Response :
472-55153.id-3907.down.sshdns.smarchal.com: type TXT, class IN
Text: 4PR7rSAb6lACp7ioEm4+VNWnPRZECBFxtM1ZqpT7C7npsmNDbMrPVzMRtA3V4TsVfyAWOBKhOT0R
k7efFbCxWEYk/xbwGYTKzaEDdACzi4daWutOoDAGxHbA7FTnQ5sB3Z5oAnaisAn+INxWFBw=
Text: 4NbAGnqBcd53QBVVscpU3lloCkS5GNz/xqgdMVLdxGBXSDn4oyInz4lK8XZtyrMjZxh7N4VZTFXA
KXs61eHQvtbHtgEz11sWMPc=

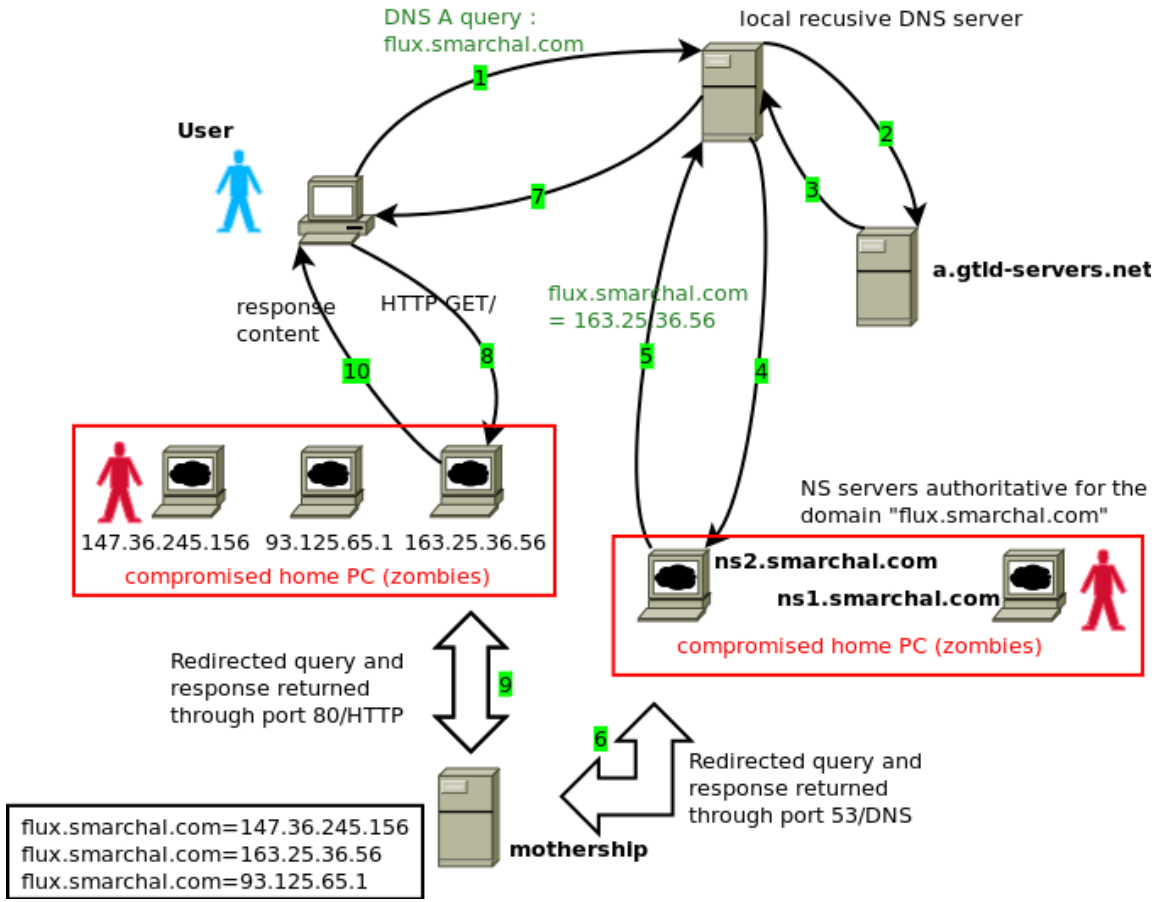
- huge number of DNS A queries for the same domain
- abnormal number of DNS TXT queries and responses over time

Fast Flux



- several A records over time
- RRs with low TTL

Double Flux



- several A and NS records over time
- RRs with low TTL

SnT

Aim

Identify such behaviour (DNS tunnelling, fluxing domains, DNS scanning, etc.) $\Longrightarrow$ *detection:*

- worm infected hosts
- malicious backdoor communication
- botnet participating hosts
- phishing websites hosting
- spamming domains
- etc.

$\Longrightarrow$ **Only based on DNS features**

SnT

Issues

Legitimate activities can have the same characteristics as malicious activities:

$\Longrightarrow$ Fluxing domains / Content Delivery Network (CDN):

- RRs with low TTL
- several IP addresses for the same domain name
- IP addresses scattered over several IP ranges
- algorithmically generated subdomains

$\Longrightarrow$ **Refine features selection to discriminate malicious from legitimate activities**

SNT

Outline

1 Motivations

2 What data to analyse and how?

3 Related work

4 DNSSM

5 Conclusion

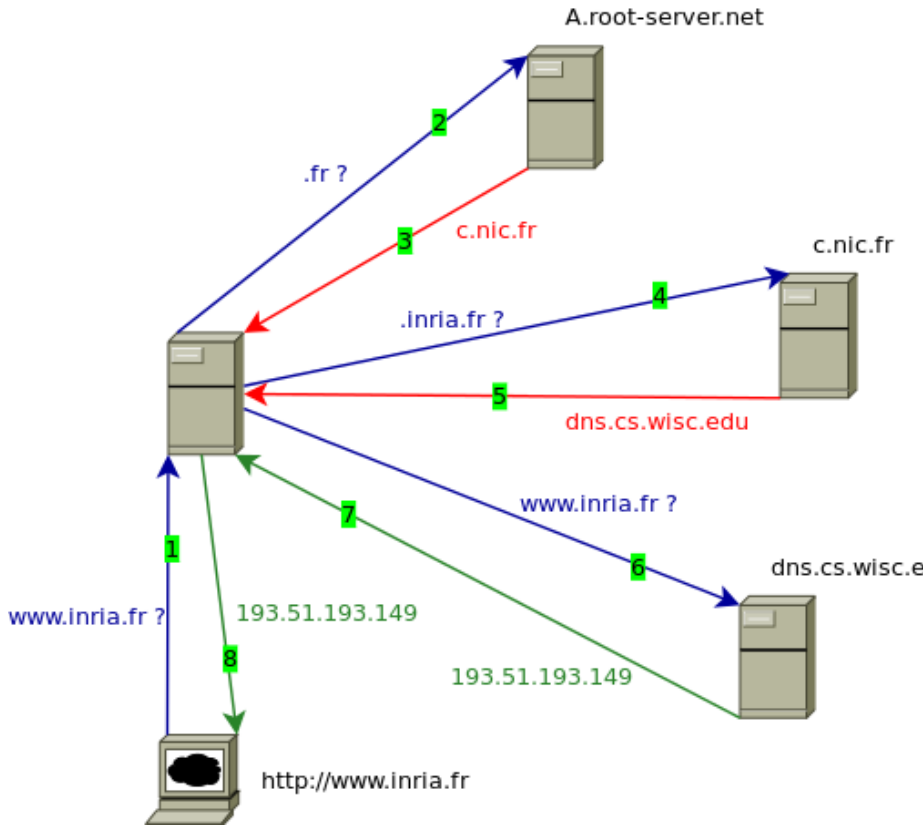
SnT

Features selection

- Extract relevant features:
  - TTL
  - Number of Resource Records (A, NS, TXT, etc.)
  - DNS querying behaviour for a single domain (number of requests)
  - Number of subdomains (both in requests and responses)
  - features from lexical analysis of domain name
  - etc.
- Information contained in :
  - domain names
  - DNS requests (name and behaviour)
  - fields in DNS responses

Data collection

- Gather both DNS requests and responses
- Where must we place probes ?
  - end user machine (no privacy, redundancy)
  - recursive DNS server (no specific user behaviour)
  - authoritative DNS server (targeted domains)
- Probe locally or worldwide



SnT

Data storage

- Large scale analysis $\Rightarrow$ recursive DNS server
- Multiple locations all over the world
- Observations (probes set up in recursive DNS servers of 2 ISPs in Luxembourg)
- 1 GB of data per day

$\Longrightarrow$ **Address the problem of data storage and data processing for scalability**

SNT

Outline

1 Motivations

2 What data to analyse and how?

3 Related work

4 DNSSM

5 Conclusion

SNT

Related work

- Apply supervised classification to DNS features ⇒ 2 classes: legitimate / malicious
- Targeted detection of malicious activity (only phishing, botnets, spam, etc.)
- Use ISC (Internet System Consortium) Passive DNS Database for learning step ⇒ only DNS responses : domain names, RR type, @IP, first and last seen
- Data storage addressed using binary tree

SnT

Outline

1. Motivations
2. What data to analyse and how?
3. Related work
4. DNSSM
5. Conclusion

SMT

DNSSM

Automated clustering technique for DNS on-line analysis

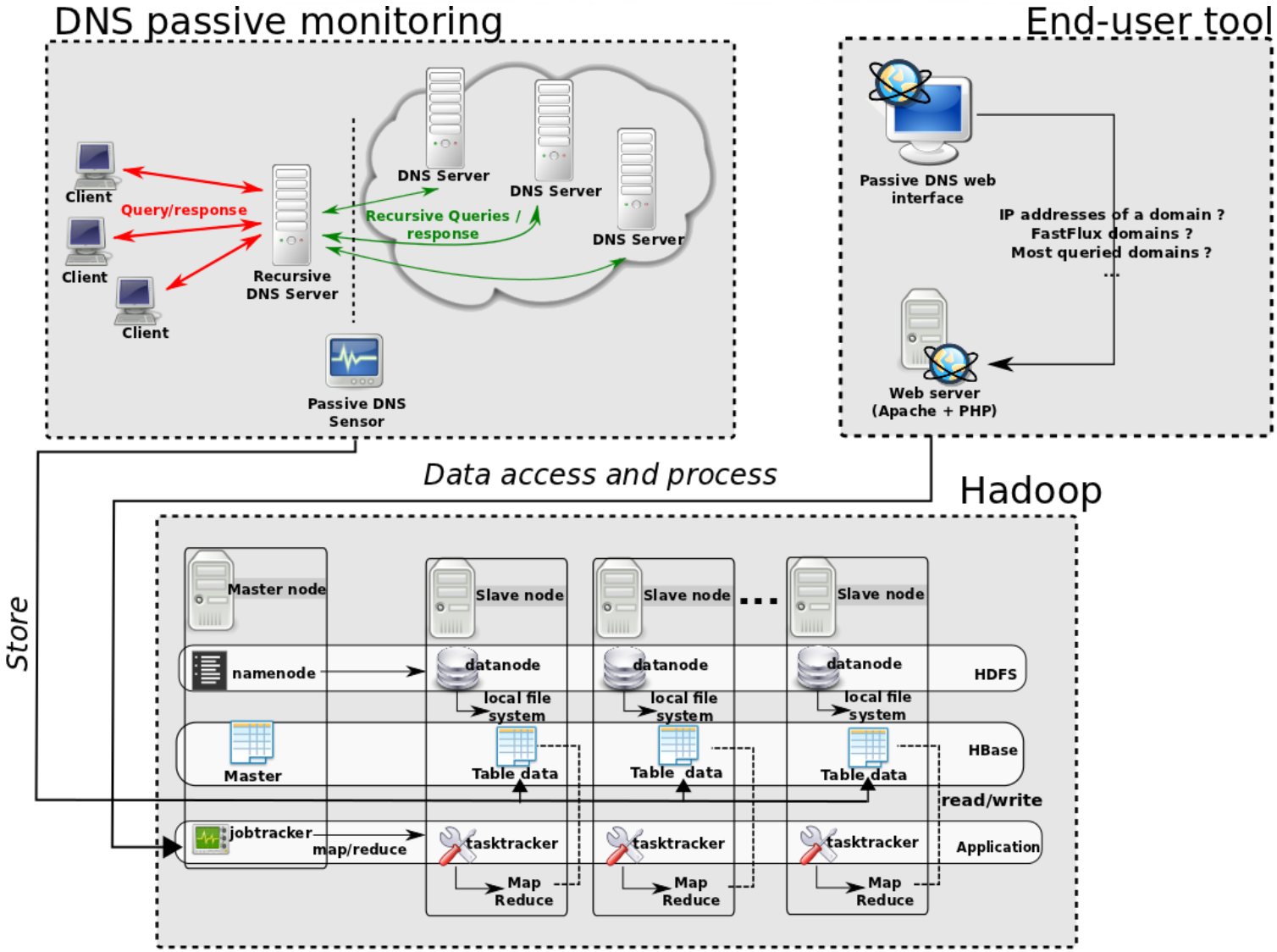
- Passively collected data at recursive DNS server level (based on Florian Weimer work)
- Extraction of 10 relevant features
- MySQL database storage
- K-means clustering : 8 clusters
- Group domains regarding their activity
- Tested on 2 datasets ($\neq$ location, $\neq$ type of network, $\neq$ users, $\neq$ quantity)

SnT

Distributed data storage

- Meet scalability requirement (10s of Gigabytes of data per month)
- Centralized MySQL database $\Longrightarrow$ distributed architecture
- Hadoop cluster implementing MapReduce design pattern:
  - distributed data storage
  - distributed computations

Architecture



SNT

Outline

1 Motivations

2 What data to analyse and how?

3 Related work

4 DNSSM

5 Conclusion

SnT

Conclusion

- Challenges for large scale DNS analysis
  - Probing, features selection, data storage, data mining
- Passive DNS monitoring solution (DNSSM)
  - Architecture for data collection, storage and mining (leveraging Hadoop)
  - Relevant features selection
  - Unsupervised clustering techniques $\Rightarrow$ domain activity
- Futur Work:
  - Apply technique to bigger datasets
  - Selection of new relevant features
  - Explore lexical and semantic composition of domain names

Large Scale DNS Analysis

Samuel Marchal and Thomas Engel

SnT - University of Luxembourg

UNIVERSITÉ DU LUXEMBOURG

CETREL a SIX company

SnT