Semantic based DNS Forensics

Samuel Marchal, Jérôme François, Radu State and Thomas Engel

UNIVERSITÉ DU LUXEMBOURG

SnT

NETLAB

SMT

Outline

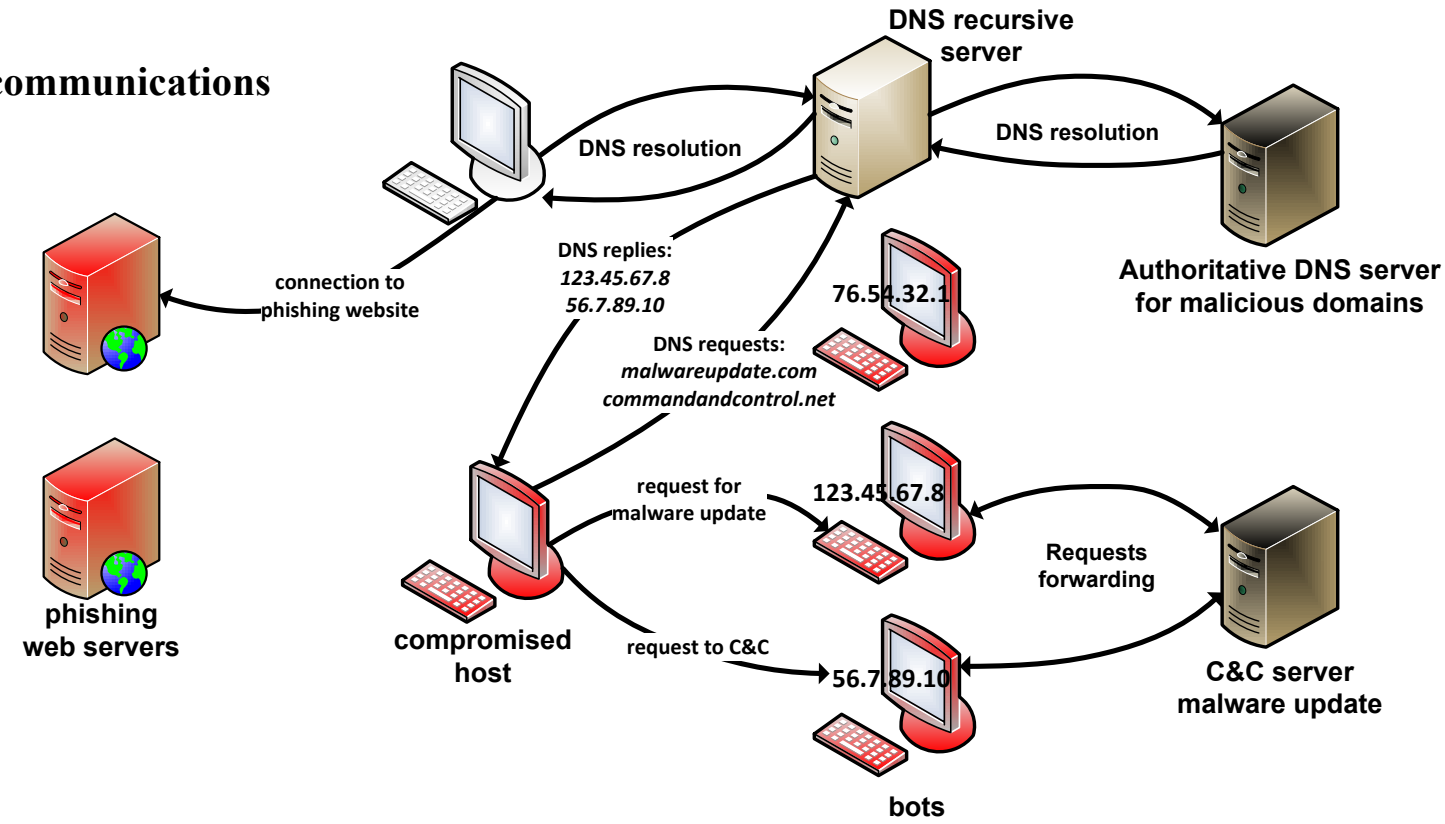
SMT

Outline

DNS misuse

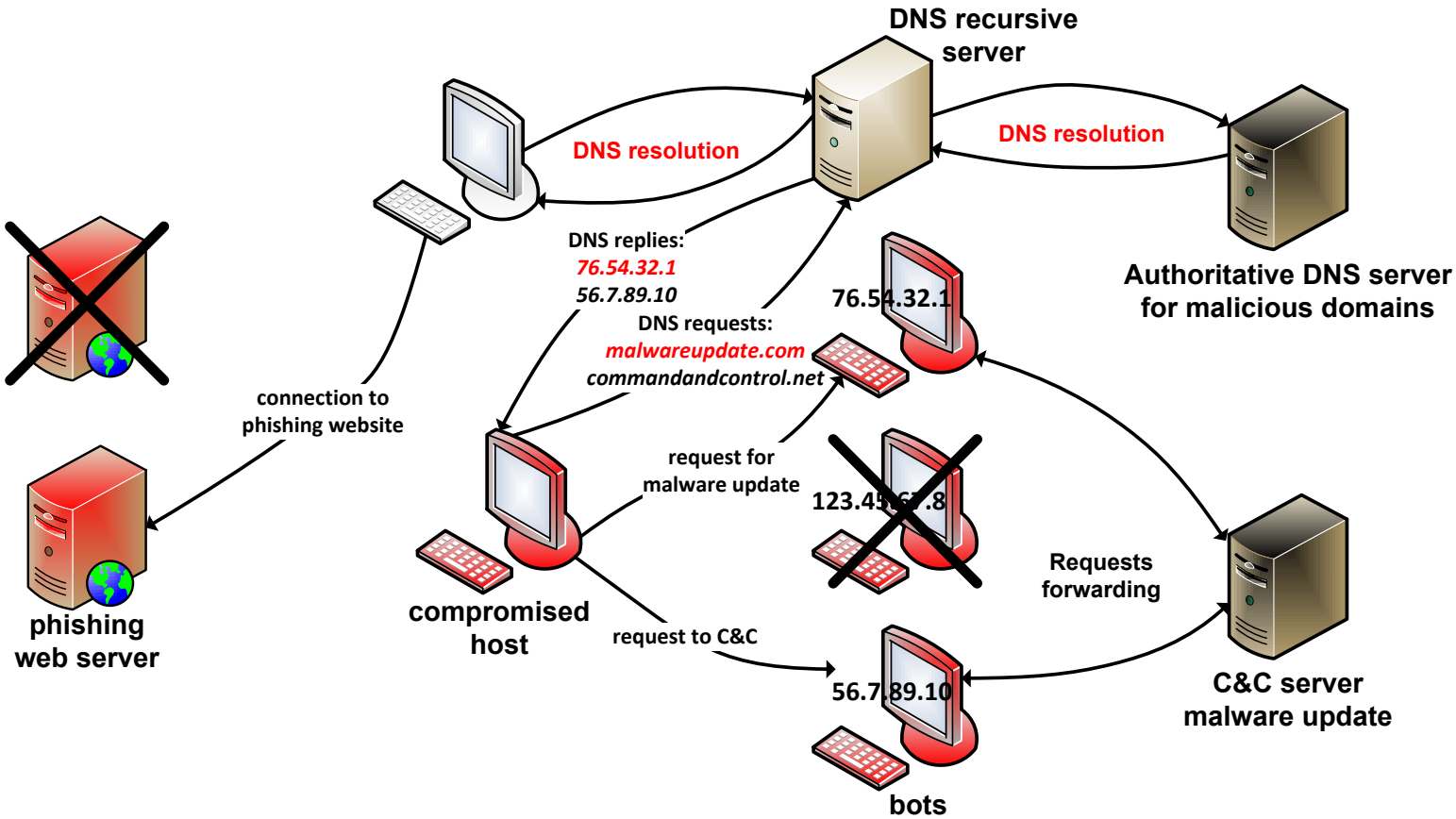
DNS: Domain Name System is the support of many malicious activities

- **malware updates**
- **botnet C&C**
- **phishing**
- **backdoor communications**
- **etc.**



DNS misuse

DNS: Domain Name System is the support of many malicious activities



SNT

DNS for forensic

Why proceed DNS analysis for forensic purposes ?

- find proof of infection (malicious domains requests)
- reduced amount of data to analyse: DNS is a meager subset of network traffic
- DNS analysis keeps users' anonymity

$\Longrightarrow$ *useful as a first step before in-depth analysis*

SNT

DNS for forensic

Why proceed DNS analysis for forensic purposes ?

- find proof of infection (malicious domains requests)
- reduced amount of data to analyse: DNS is a meager subset of network traffic
- DNS analysis keeps users' anonymity

$\Longrightarrow$ *useful as a first step before in-depth analysis*

Issue: How do we know if a domain is malicious ?

SNT

State of the art

Identification of malicious domains:

- User reports + manual checking
- DNS packet fields analysis + classification via machine learning algorithm:
  - domain records removed: data is no longer available
    $\Longrightarrow$ problematic for forensic analysis
- Domain name based analysis:
  - number of domain levels
  - relative position of labels
  - domain length
  - etc.

SMT

Outline

SNT

Analyse domain semantic

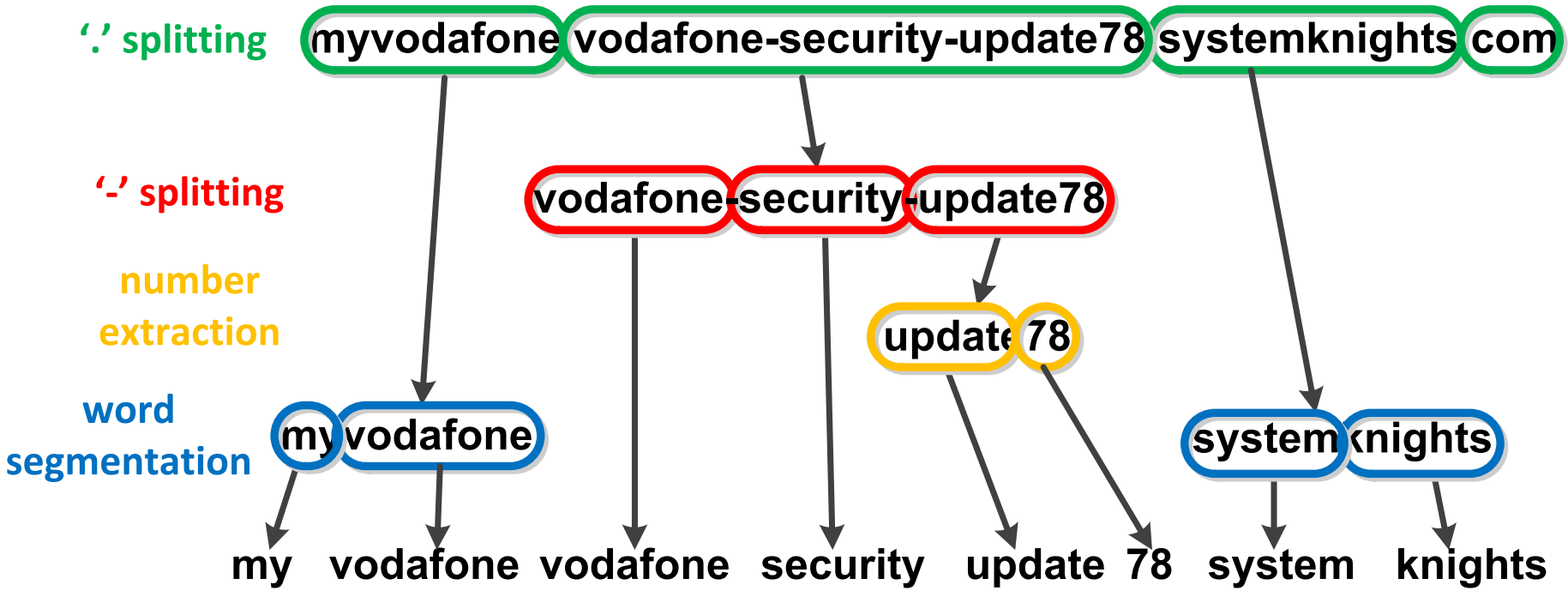
- Domain names are meant to be meaningful
- Observations: malicious domains often use words from the same semantic fields:
  - www.visa-sweden.mastercard.forever4c.com
  - myvodafone.vodafone-security-update78.systemknight.com
  - paypal.com-us.webscr.cmd-homeelocale.gumuspena.com
- Issue: single domains are not significant enough
- $\Longrightarrow$ Group domains according to common features (IP address, etc.)
- Knowing group of malicious and legitimate domains

$\Longrightarrow$ **deduce if an unknown group is malicious or not**

Features extraction

Splitting of domain name:

`myvodafone.vodafone-security-update78.systemknights.com`



- $distword = \{(my, 0.125), (vodafone, 0.25), (security, 0.125), ...\}$

SMT

Semantic relatedness evaluation

How to evaluate semantic similarity between two sets of domain names ?

$\Longrightarrow$ between two words: Wordnet, Disco:

- calculate a similarity score (semantic relatedness) between 2 words
- give the n most related words to w
- based on dictionary (Wikipedia, BNC, PubMed, etc.)

$$sim(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} I(w_1,r,w)+I(w_2,r,w)}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w)+\sum_{(r,w)\in T(w_2)} I(w_2,r,w)}$$

$\Longrightarrow$ **use this metric in new ones**

SNT

Semantic metrics

3 metrics defined to compare two sets of domains:

Assuming two domain sets A and B and the associated extracted word sets W_A and W_B with the occurrence frequencies *distword* we have:

$$Sim_1(A,B) = \sum_{w_A \in W_A} \sum_{w_B \in W_B} sim(w_A, w_B)$$

$$Sim_2(A,B) = \sum_{w_A \in W_A} \sum_{w_B \in W_B} sim(w_A, w_B) \times distword_{w_A,W_A} \times distword_{w_B,W_B}$$

$$Sim'_3(A,B) = \sum_{w \in W_A} \sum_{w' \in Disco(w,n)} sim(w, w') \times distword_{w',W_B}$$
$$\Longrightarrow Sim_3(A,B) = Sim'_3(A,B) + Sim'_3(B,A)$$

SMT

Outline

Similarity metrics efficiency

Comparison pair-wise of domains sets ($Sim_3(A, B)$)

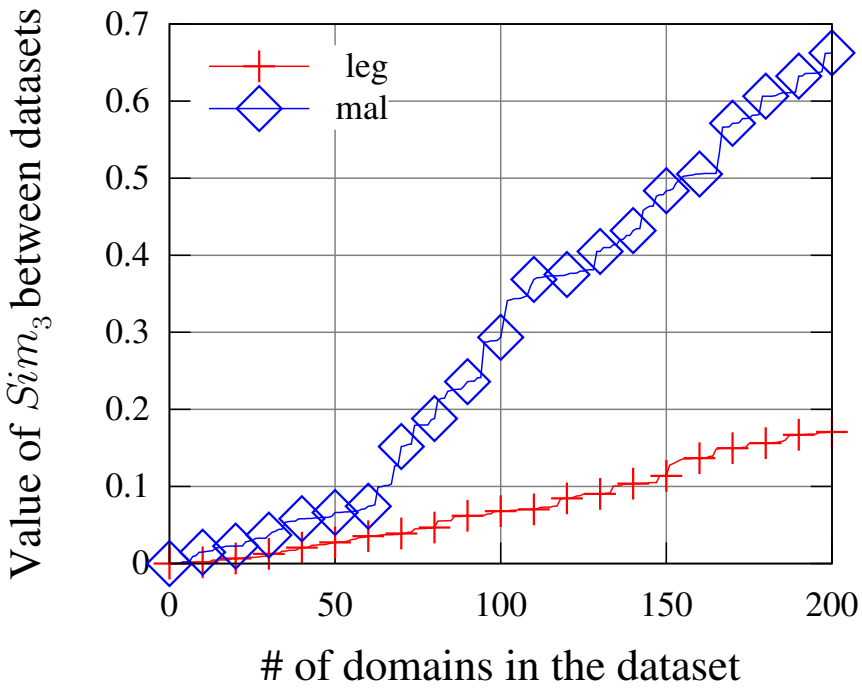
- 10 sets of around 13,000 domains each
- 5 legitimate (Alexa + passive DNS)
- 5 malicious (PhishTank, DNS-BH, MDL)

	leg-5	leg-4	leg-3	leg-2	leg-1	mal-5	mal-4	mal-3	mal-2
mal-1	0.776	0.795	0.793	0.789	0.785	0.955	0.962	0.965	0.975
mal-2	0.782	0.800	0.798	0.797	0.797	0.965	0.968	0.973	
mal-3	0.772	0.796	0.793	0.788	0.784	0.951	0.962		
mal-4	0.783	0.804	0.804	0.800	0.796	0.953			
mal-5	0.769	0.785	0.784	0.782	0.772				
leg-1	0.946	0.948	0.952	0.938					
leg-2	0.915	0.924	0.922						
leg-3	0.936	0.934							
leg-4	0.935								

Size of domains sets

Similarity metrics able to distinguish legitimate from malicious sets of domains:

- for big set (13,000 domains): ok !!
- minimum number of domains in a set to evaluate it ?



SMT

Outline

SNT

Conclusion

Technique for domains sets comparison:

- semantic similarity scoring
- apply to identification of malicious domain set
- useful for first step of forensic analysis

Results:

- able to distinguish malicious from legitimate domains...
- ... for sets of at least 10 domains

Future works:

- improve similarity metrics
- correlate with IP Flow records

Semantic based DNS Forensics

Samuel Marchal, Jérôme François, Radu State and Thomas Engel

UNIVERSITÉ DU LUXEMBOURG

SnT

NETLAB