Semantic based DNS Forensics

Samuel Marchal, Jérôme François, Radu State and Thomas Engel
Motivations

Semantic analysis

Experiments and Results

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

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

DNS requests:
- malwareupdate.com
- commandandcontrol.net
- compromised host
- etc.
DNS: Domain Name System is the support of many malicious activities
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

⇒ useful as a first step before in-depth analysis
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

⇒ useful as a first step before in-depth analysis

Issue: How do we know if a domain is malicious?
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
  \[\rightarrow\] problematic for forensic analysis
- Domain name based analysis:
  - number of domain levels
  - relative position of labels
  - domain length
  - etc.
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
- Group domains according to common features (IP address, etc.)
- Knowing group of malicious and legitimate domains

⇒ deduce if an unknown group is malicious or not
Splitting of domain name:
myvodafone.vodafone-security-update78.systemknights.com

- ‘.’ splitting
  - myvodafone
  - vodafone
  - security
  - update
  - 78
  - system
  - knights

- ‘-’ splitting
  - vodafone
  - security
  - update
  - 78

- Number extraction
  - update
  - 78

- Word segmentation
  - my
  - vodafone
  - vodafone
  - security
  - update
  - 78
  - system
  - knights

- distword = {((my, 0.125), (vodafone, 0.25), (security, 0.125), ...)}
How to evaluate semantic similarity between two sets of domain names?

⇒ 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.)

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

⇒ use this metric in new ones
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 $\text{distword}$ we have:

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

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

$$Sim'_3(A, B) = \sum_{w \in W_A} \sum_{w' \in \text{Disco}(w,n)} \text{sim}(w, w') \times \text{distword}_{w', W_B}$$

$$\implies Sim_3(A, B) = Sim'_3(A, B) + Sim'_3(B, A)$$
Outline

1 Motivations
2 Semantic analysis
3 Experiments and Results
4 Conclusion
Similarity metrics efficiency

Comparison pair-wise of domains sets \((\text{Sim}_3(A, B))\)

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

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

- 0.7 0.76 0.82 0.88 0.94 1.00
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 ?

![Graph showing the value of Sim₃ between datasets as a function of the number of domains in the dataset.](image)

- **leg**
- **mal**
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
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