Security Metrics in the field of Malware Domains

Niels Gijsen, Noah Goldsmid, Samiksha and Jeroen Weener

September 2018

1 Assignment description

Security metrics can gear up a system to defend against malicious actors, while weak security metrics can damage the system severely! They are the defensive pillars of any electronic machinery that is prone to malware attacks.The strength of security measures against any malwares defines whether the right amount of money has been spent on it or not, whether the time devoted to install those measures is worth or not or be it any extra effort added for it. Any program without security measures is vulnerable to attacks (intense damage to property/person!), thus security measures give the programs a right direction, by right direction we mean protection against virus, malwares, etc. Distinct metrics need to be applied on different modules of a program and for that prioritizing each metric is essential- therefore [6]"Security is a process". During Block 2, we learned the importance of measuring cybersecurity and the challenges to create meaningful metric. However, metrics are necessary to show how security activity contributes directly to security goals; measure how changes in a process contribute to security goals; detect significant anomalies in processes and inform decisions to fix or improve processes.

2 Introduction

In this report the reader will be presented with security metrics in the field of malware domains. First, we take a look at the context of malware domains. Then, a look will be taken at the different security metrics. We start by showing what metrics would be ideal and explain why they are not usable in practice. We will then give an overview of metrics that actually are used in practice. Lastly, we provide the reader with a few proposed metrics. The usability and effectiveness of these metrics will be discussed in the Conclusion.

3 Context

Malware or malicious software is software used by cyber criminals designed to infect computers. Malware can be used for different things e.g. to set them up for a bot net or to ransom the owner's data. A common way for cyber criminals to spread malware is by directing people to a domain on the web where they uploaded the malicious software. This report looks into a data set of malware domains from `http://www.malwaredomainlist.com/`. The data set consists of domain names that are known to host malware. Most of the domains in the list are involved in phishing, bot nets or other malicious activities. Accurate measurements on these data sets can help improve automated detection of malicious domains. This is useful for end-user protection.

4 Actors

- TLD: Top Level Domain refers to thhe last segment of the domain name
- Hosting provider provisions pool of remote and internet based serviced to the individuals
- victims of phishing/DDoS

5 What security issue does the data speak to?

The dataset provided signifies the cyber issue of malware attacks. The malwares are generated either to damage property/personal data or to extract revenue. And the victims targeted are the actors in this attack. The cyber criminals (bad actors) spread the created virus, trojan, worm, etc. It could be done by social engineering, injecting user's system without his/ her knowledge or a combination of both. The failure to uncover the attackers in a due time leads to victim's data breaching which causes a heavy loss.

The reason of these malware attacks could be lack of awareness of using unknown URLs. Usually the attackers don't use HTTPS or SSL to create the malicious websites. the other security leap could be responding to the fake ads where they ask for the credential information and later misuses it. It is certainly tough to identify whether a website is fake or real but blocking such ads could definitely help.

Malware has become a consistent threat that it can bypass advanced detection, prevention and antivirus techniques. This brings a need of robust and stable malware security measures.

6 What would be the ideal metrics for security decision makers?

These fraud mechanics require strongly built up security measures where raising awareness must be essential. The ideal security metrics should not only be able to guard the system from unwanted attacks but also be able redirect to the information of the resources which generated such attacks. Even a single 0/1 bit change, SQL injection or a minor PHP script change can bring the system into an indefinite loop.

To create a security metric, it is necessary to investigate your system as if it is already going through an attack. Presumptuous metrics can always be supportive, but an ideal metric should be the one that can be brought into practice. It must limit to the most useful and important data. What data can or cannot be used later brings in-feasibility of data storage. Thus the data should be in a standardized form such as it could be compressed without any loss of data.

The ideal security metrics that can be practiced from a user's end can be as follow:

- Install an anti-virus software
- Update the installed anti-virus software
- Keep a check on the browsing URLs
- Install a firewall
- Must stay aware about the phishing emails and fake ads.

7 Data set

The data set is available at `http://www.malwaredomainlist.com/mdlcsv.php` and contains 2289 entries. It does not list the column names. These are (in order):

1. Date (UTC)
2. Domain
3. IP
4. Reverse Lookup
5. Description
6. -
7. ASN (autonomous system number)
8. -
9. Country Code

8 Metrics

8.1 Ideal Metrics

Each metric has its own biases, therefore an ideal metric would be a aggregation of multiple robust metric and prevention strategies. A comprehensive security plan is the most powerful way to gauge security.

In the book Security Metrics, Andrew Jaquith highlights the following characteristics of a good metric, stating that it needs to be:

“Consistently measures, without subjective criteria Cheap to gather, preferably in an automated way Expressed as a cardinal number or percentage, not with qualitative labels like ‘high,’ ‘medium,’ and ‘low’ Expressed using at least one unit of measure, such as ‘defects,’ ‘hours,’ or ‘dollars’ Contextually specific—relevant enough to decision-makers so that they can take action”

What percentage of domains on the internet is infected with malware? What percentage of domains in a certain country is infected with malware? What are the annual losses to society of malicious domains expressed in dollars. Incident response times. For how long do malicious domains remain online? E.g. how long to detect, how long to respond.

8.2 Metrics used in Practice

A substantial amount of literature is available on the detection of malware domains. [4], [2] and [10] propose means to determine whether a domain is malicious or benign based of the URL structure. Others have taken different approaches. [9] comes up with a system that uses classic malware detection systems on PC’s. Upon detection of malicious code the system looks at the download log of the PC to determine where the malware came from. [1] designed Kopis, a system that aims to detect malware domains by analyzing global DNS query resolution patters from the upper DNS hierarchy. [3] and [5] come up with ways to measure bot net network traffic and identify illegitimate DNS servers respectively.

[7] introduces a handful of metrics for comparing TLDs against their market. First, they present occurrence security metrics. They start with looking at the amount of unique blacklisted domains. This would be the most intuitive metric, but they argue that one domain could be used for multiple malicious activities. Therefore they come up with a second, complementary metric, the number of unique FQDNs (Fully Qualified Domain Name). Since this does not quantify the amount of ’badness’, they come up with a third metric: unique blacklisted URLs. All these metrics are normalized by the size of the TLD. They also present an uptime security metrics that indicates how long it takes for TLDs to respond once a phishing domain is abused. While the mean uptime of a domain is an intuitive metric, they argue that the median uptime is more reflective since the mean uptime does a poor job when confronted with long-lived malicious domains.

[8] discusses security reputation metrics for hosting providers. While usually

the unit of abuse is defined as the amount of distinct IP addresses, they consider the use of 2nd-level domain-IP pairs (2LD, IP). Just as [7], they argue that counting distinct IP addresses would underestimate abuse since cyber criminals may use the same IP address for multiple malicious intents. They mention that in certain cases it is better to consider the pairs (FQDN, IP) or (URL, IP). Second, they talk about what data feeds to use to measure the abuse happening at hosting providers. They argue that a combination of different data feeds should be used and that special care should be taken to make sure that the data sets provide enough coverage and purity. Lastly, they state that when normalizing measurements by the size of hosting providers it is important that the size estimator is accurate. While advertised IP space is an attractive estimator, not all IP addresses are in use or used for hosting. Two additional size estimators are proposed: the number of hosted 2LDs and the number of IP addresses to host content.

8.3 Metrics that can be designed from the data set

- **Number of malicious domains per domain extension (.com .org .ru etc.) divided by the size of the domain extension.**

  By looking at the weighed average of the number of malicious domains per domain extension, it can give an indication of which domain extensions might be most susceptible to attacks.

- **Number of malicious domains per geographical location divided by the size of the network in the geographical location.**

  Looking at the geographical location of a domain can tell us more about the locations of hosting providers preferred by attackers. Dividing by the size of the network is necessary in order to fairly compare locations.

- **Distribution of the percentage of malicious activities over time. (malware/phishing/spam etc.)**

  The percentage of malicious activity type over time tells us about the popularity of a certain attack. Knowing the popularity can help when creating risk assessments for actors.

- **Distribution of malicious type per country**

  Mapping the type of malicious activity to the geographical location of the domain can give security researcher insight in where certain treats are coming from

- **Distribution of malicious type per domain extension**

  Linking malicious types to top level domains can give us an insight in to where attackers are registering their domains.

- **Distribution of file extensions per malicious type (.html .exe etc.)**

By connecting the file extention to malicious types, it will tell us which file types are most used in each type of attack. This can possibly mitigate risks, for example expecting a .exe to be a Trojan.

- **Number of reports over time divided by the growth of the internet over time**

  By comparing the reports over time taking the growth of the internet into account, could give us an indication if the number of malicious domains is growing or declining.

- **Domain extensions over time**

  This tells us if certain top level domains have become more or less popular over time.

- **Number of domains that are still online**

  By listening for a response of the IP addresses in the list, we can determine which domains are still up and running.

9 Conclusion

We can conclude that the problem of malicious domains raises multiple security issues for different actors. Not only at the level of the internet user, also at higher levels such as organizations or hosting providers. Metrics on malicious domains are therefore important to give valueable insights in to the security level. Ideal metrics would tell us something about the whole internet space, however due to limited resources ideal metrics are most often not seen in the real world. When looking to literature available on security metrics for malware domains we noticed that most scholars focus on metrics that can differentiate benign domains form malicous ones.

Based on the dataset provide we also created some metrics of our own. Using this dataset has multiple limitation, however we found some interesting findings, such as the "nf" top level domain to have a significantly higher percentage of malicious domains. Mapping incidents over time has showed us that the malicous domains are diversifying, while exploits used to rule the list, new threads such as ransomware started emerging in 2013.

10 Limitations

Most limitations to this research is associated to the data used. The dataset provided is not perfects, as is most often the case with data. First of all size of the dataset is limiting the the research, 2000 entries distributed over 9 year, is a small sample of all the malicious domains in the whole internet space. Furthermore the data is heavily biased while all the entries are manually submitted by contributing members of an internet forum. Also the descriptions of incidents is often not coherent to others, giving different names to same types of attacks.

References

[1] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. Detecting malware domains at the upper dns hierarchy. In *USENIX security symposium*, volume 11, pages 1–16, 2011.

[2] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. Exposure: Finding malicious domains using passive dns analysis. In *Ndss*, 2011.

[3] M. Grill and M. Rehák. Malware detection using http user-agent discrepancy identification. In *Information Forensics and Security (WIFS), 2014 IEEE International Workshop on*, pages 221–226. IEEE, 2014.

[4] M. A. Hart, J. S. Wilhelm, and S. Sundaram. Techniques for identifying potential malware domain names, Jan. 14 2014. US Patent 8,631,498.

[5] K. D. Himberger and B. M. Parees. System and method for identification and blocking of malicious use of servers, May 29 2012. US Patent 8,191,137.

[6] A. Jaquith. *Security metrics: replacing fear, uncertainty, and doubt*. Pearson Education, 2007.

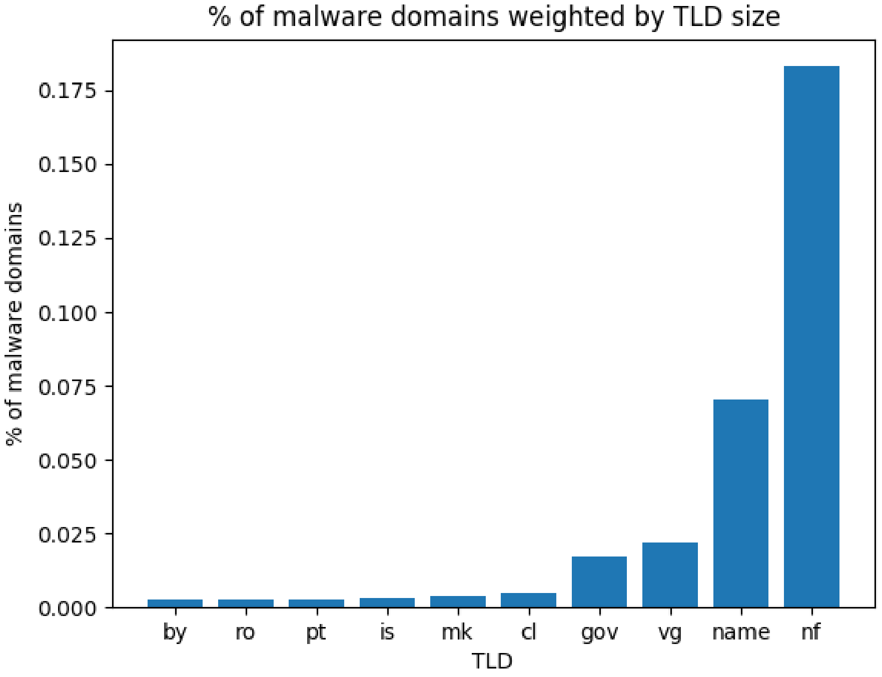
[7] M. Korczynski, S. Tajalizadehkhoob, A. Noroozian, M. Wullink, C. Hesselman, and M. v. Eeten. Reputation metrics design to improve intermediary incentives for security of tlds. In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 579–594, April 2017.

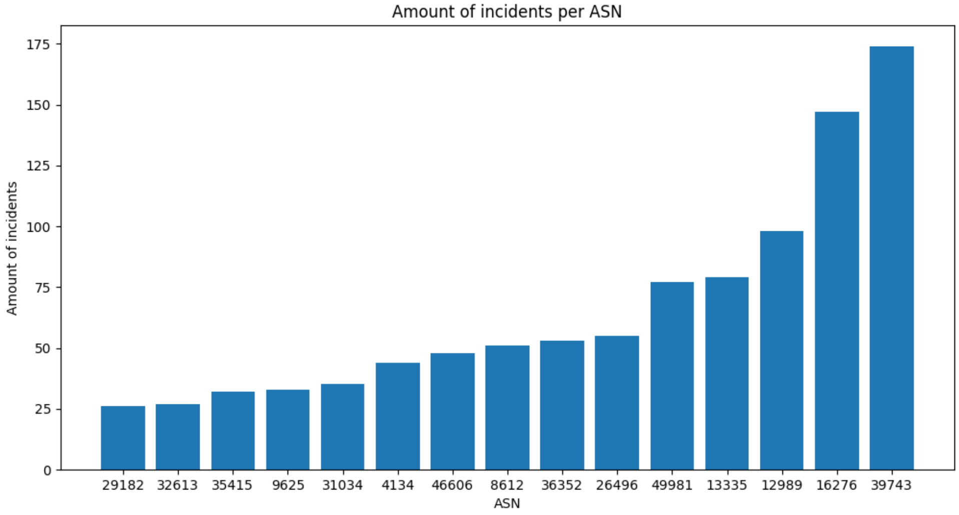
[8] A. Noroozian, M. Korczynski, S. Tajalizadehkhoob, and M. van Eeten. Developing security reputation metrics for hosting providers. In *8th Workshop on Cyber Security Experimentation and Test (CSET 15)*, 2015.

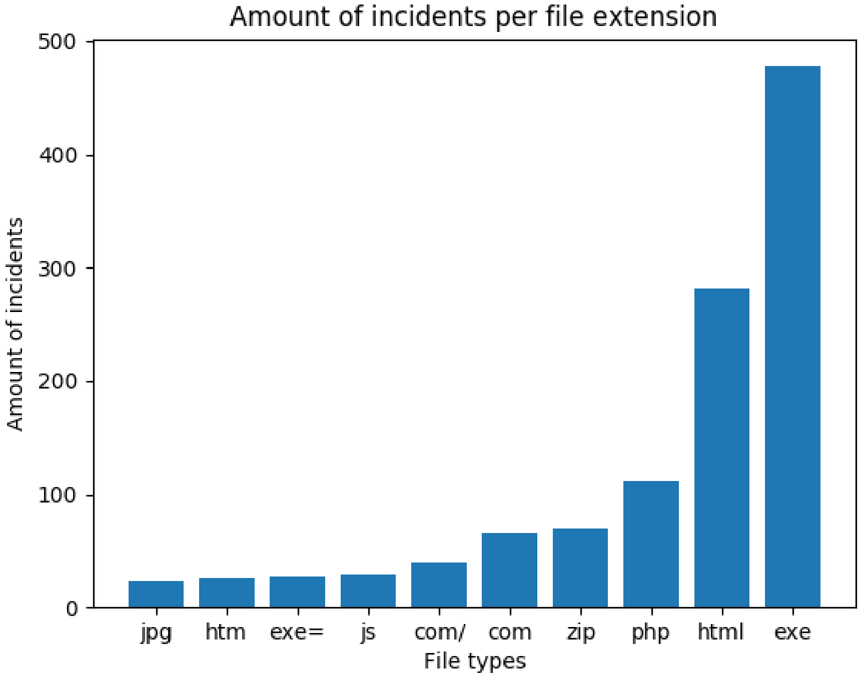
[9] P. Piccard. Systems and methods for identifying malware distribution, June 4 2009. US Patent App. 11/171,924.

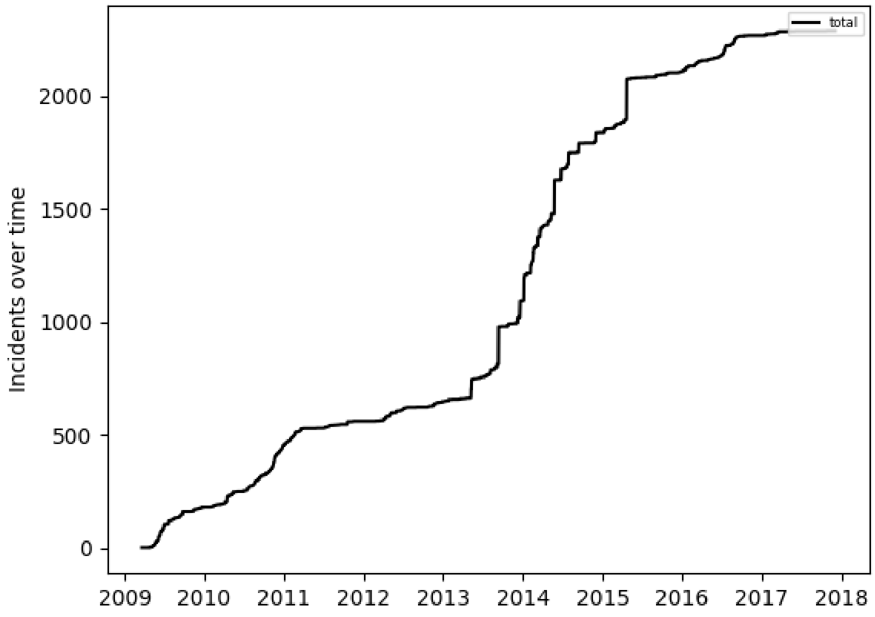
[10] S. Yadav, A. K. K. Reddy, A. Reddy, and S. Ranjan. Detecting algorithmically generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 48–61. ACM, 2010.

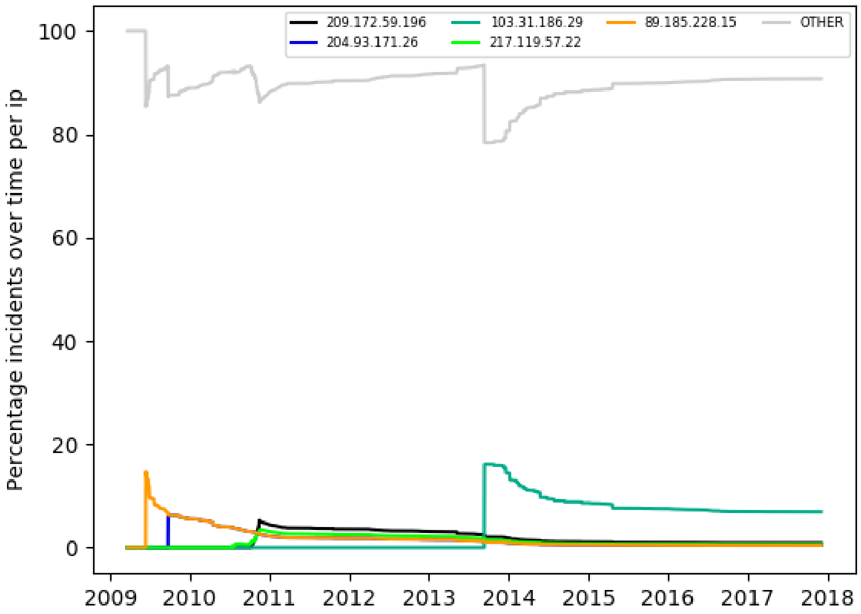
11 Appendix

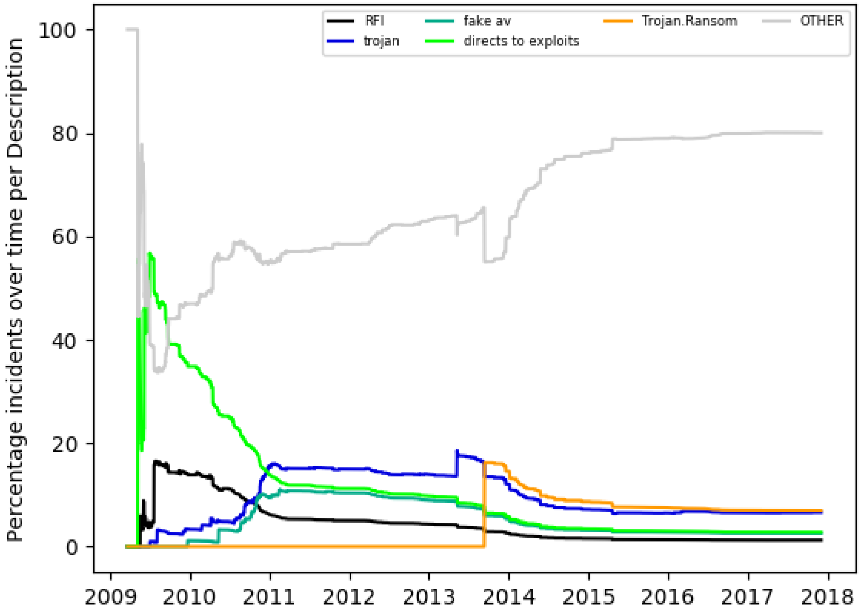

% of malware domains weighted by TLD size
% of malware domains
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
by
ro
pt
is
mk
cl
gov
vg
name
nf
TLD

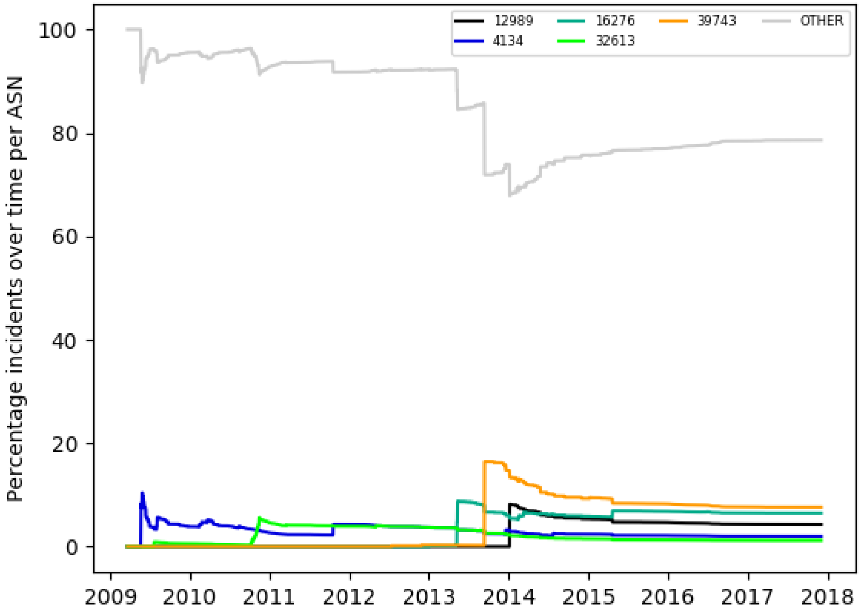

Amount of incidents per ASN
Amount of incidents
175
150
125
100
75
50
25
0
29182
32613
35415
9625
31034
4134
46606
8612
36352
26496
49981
13335
12989
16276
39743
ASN

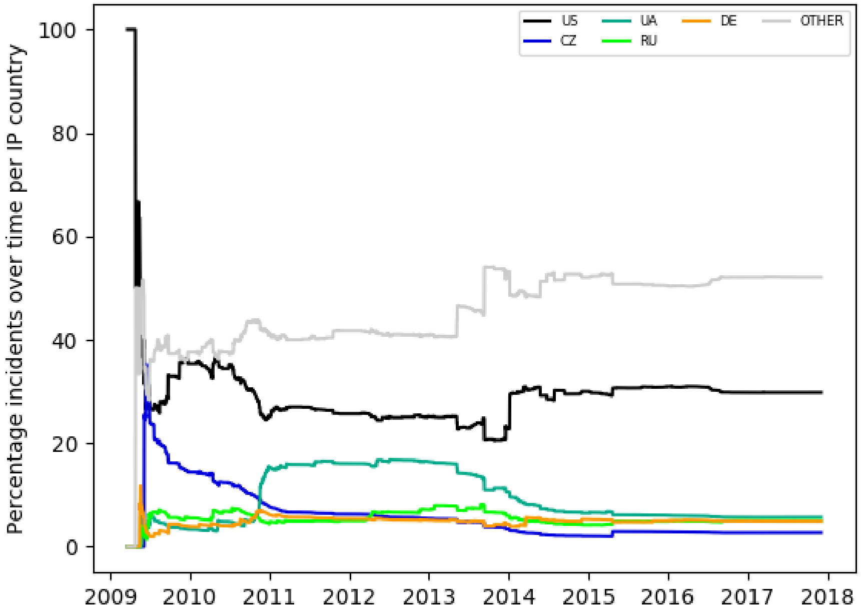

Amount of incidents per file extension
Amount of incidents
0
100
200
300
400
500
jpg
htm
exe=
js
com/
com
zip
php
html
exe
File types

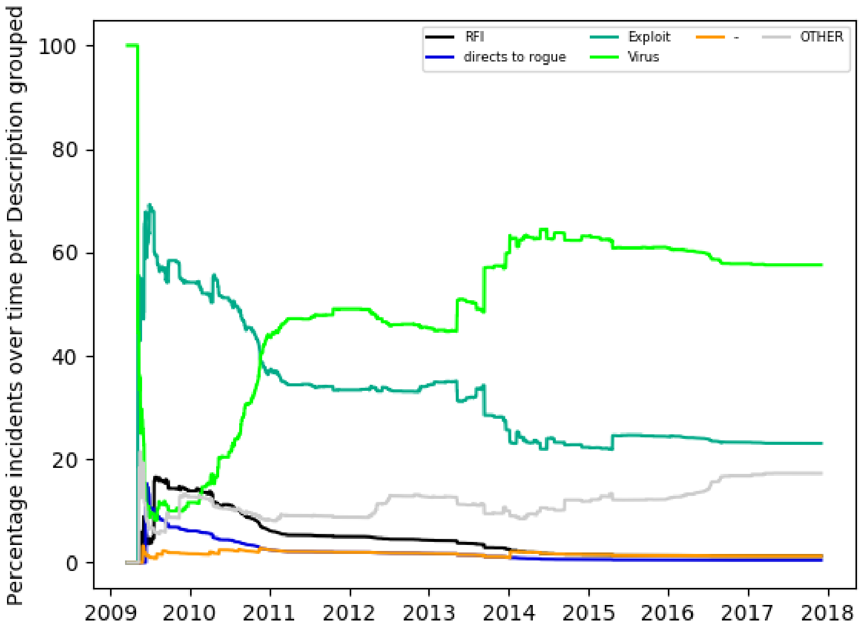

total
Incidents over time
0
500
1000
1500
2000
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

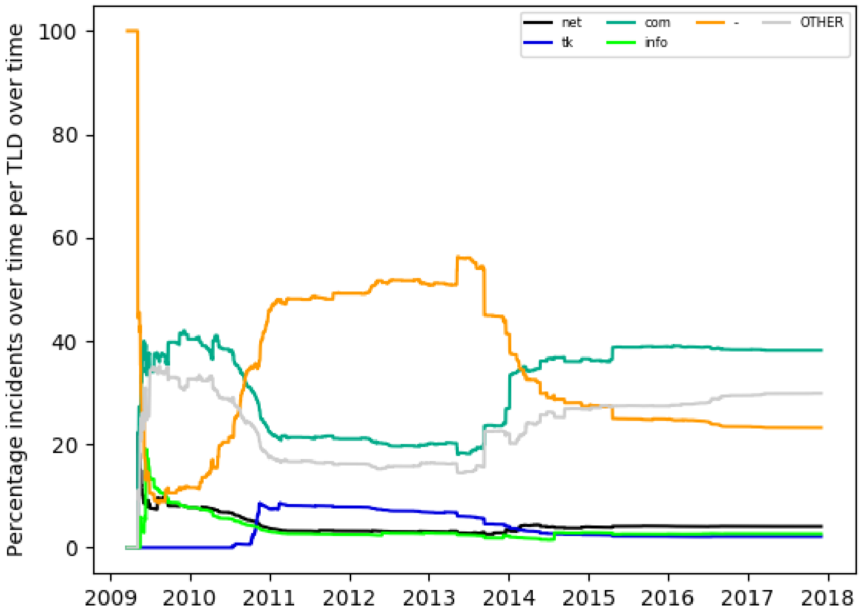

209.172.59.196
204.93.171.26
103.31.186.29
217.119.57.22
89.185.228.15
OTHER
Percentage incidents over time per ip
100
80
60
40
20
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018


RFI
trojan
fake av
directs to exploits
Trojan.Ransom
OTHER
Percentage incidents over time per Description
0
20
40
60
80
100
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018


12989
4134
16276
32613
39743
OTHER
Percentage incidents over time per ASN
0
20
40
60
80
100
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018


Percentage incidents over time per IP country
0
20
40
60
80
100
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
US
CZ
UA
RU
DE
OTHER


Percentage incidents over time per Description grouped
100
80
60
40
20
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
RFI
directs to rogue
Exploit
Virus
-
OTHER


net
tk
com
info
-
OTHER
Percentage incidents over time per TLD over time
0
20
40
60
80
100
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018