uOttawa

SOFTWARE SECURITY RESEARCH GROUP

In Collaboration With IBM

Iterative Publication of Phishing Sites

Qian Cui, Guy-Vincent Jourdan, Gregor v. Bochmann , Iosif Viorel Onut

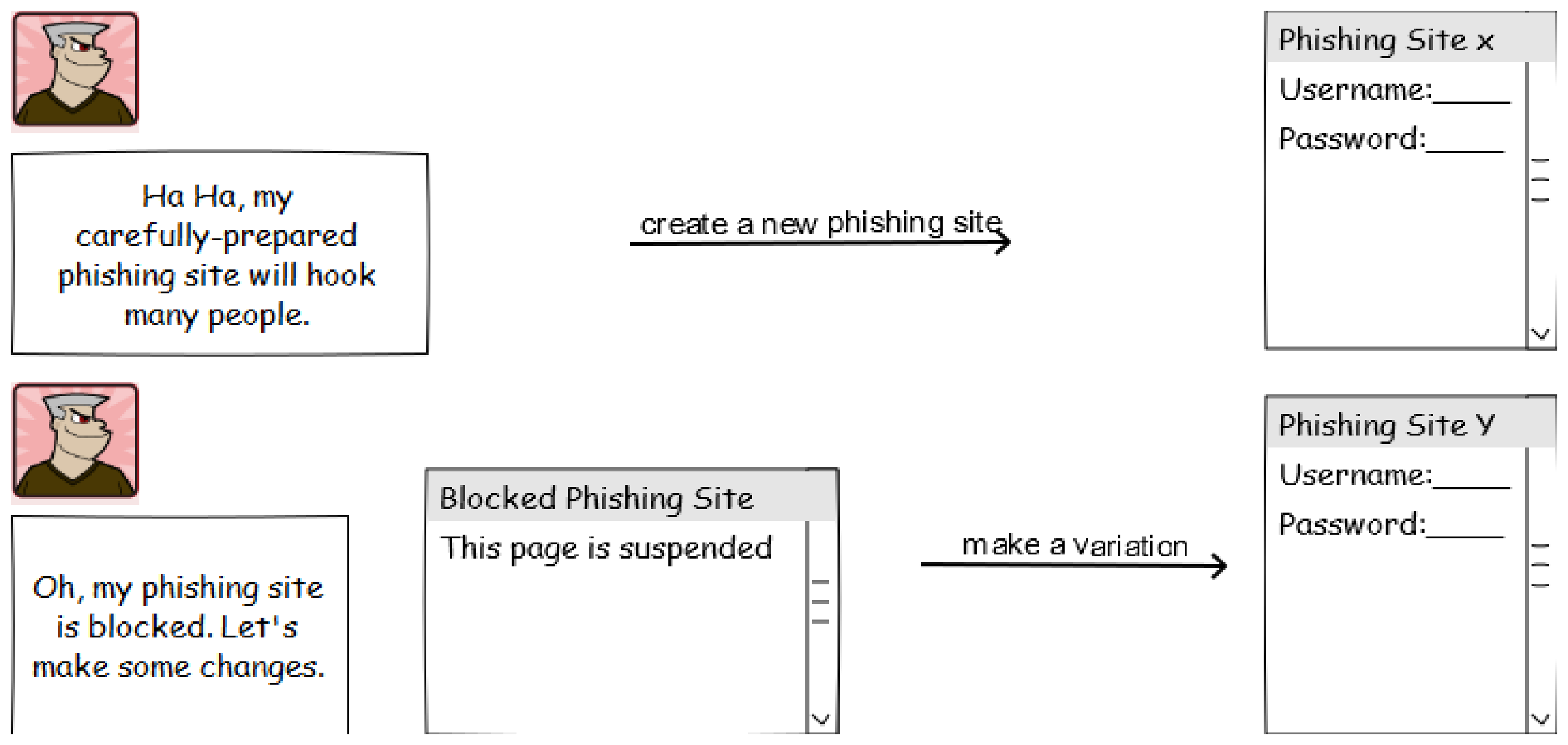
School of Information Technology and Engineering - University of Ottawa

Phishing Attacks

Phishing attacks are one of the important threats to individuals and corporations in today's Internet.

Phishing sites have the following features:
Very Short Lifespan: Slightly more than 10 hours on average
Rapid Update Cycle: Phishers constantly “refresh” their attacks by publishing new phishing sites.

=> Our working assumption here is that attackers are not creating new attack sites, but instead are constantly modifying and re-publishing existing ones.



Ha Ha, my carefully-prepared phishing site will hook many people.
create a new phishing site
Phishing Site x
Username:____
Password:____
Oh, my phishing site is blocked. Let's make some changes.
Blocked Phishing Site
This page is suspended
make a variation
Phishing Site Y
Username:____
Password:____

Detection of Phishing Variation and Duplication

Basic idea: a site which is very similar to another known phishing site is more likely itself a phishing site.

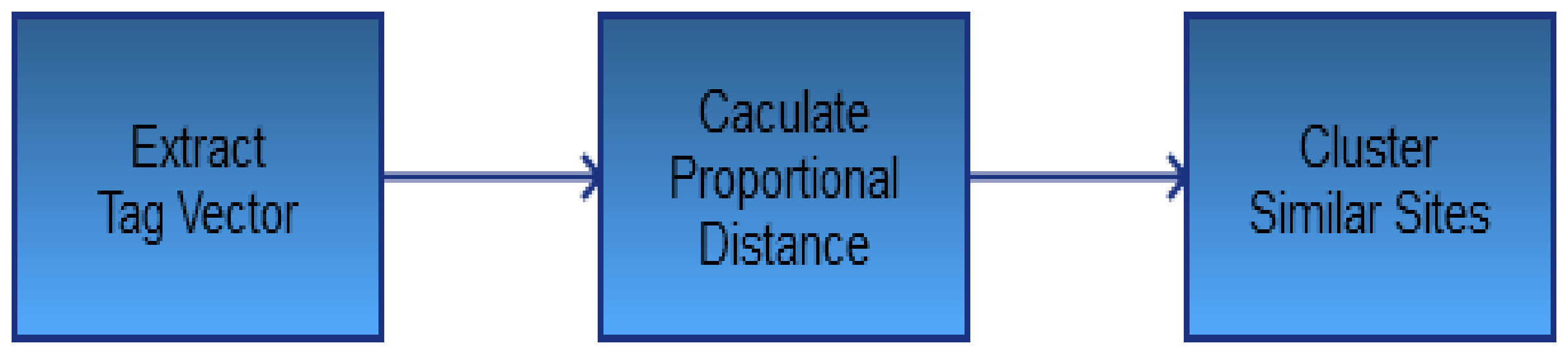
Challenges

✓ The textual content of the phishing site may be updated substantially.
✓ The structure of the phishing site may be modified, for example because some parts are being swapped.

Methodology

We have created a method for the detection of phishing duplicates which is not sensitive to the textual content of the page and will resist to slight changes to the page structure.

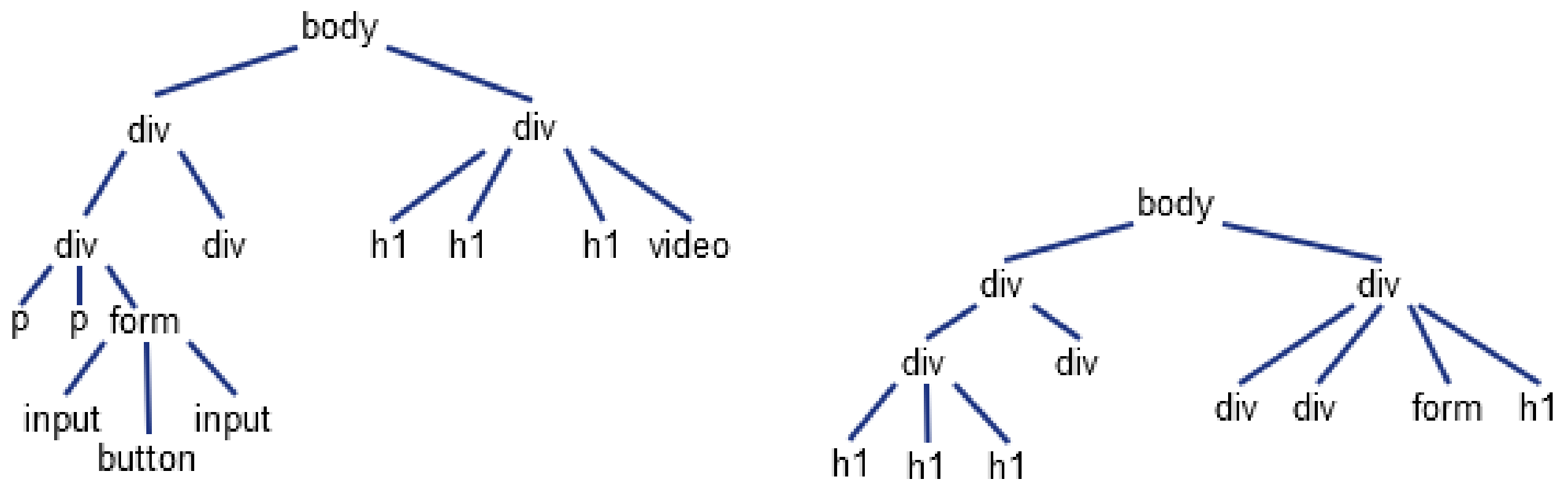
Our method has 3 steps.
✓ Extract tag vector
✓ Calculate proportional distance
✓ Cluster similar sites


Extract Tag Vector → Caculate Proportional Distance → Cluster Similar Sites

Extracting Features of Pages: Tag Vectors

Corpus of HTML Tags: We used the complete set of HTML elements provided by the World Wide Web Consortium, and removed some of more common tags, such as body, head, html

Our **Tag Vector** is a vector of the size of the corpus, counting the number of times each tags in the corpus appearing in the page



tag vector: [1, 0 ,2 ,3, 1, 1, 2, 0, 4]

tag vector: [1, 0 ,0 ,4, 0, 0, 0, 0, 6]

tag corpus:<form> <b> <p> <h1> <button> <video> <input> <iframe>

Proportional Distance

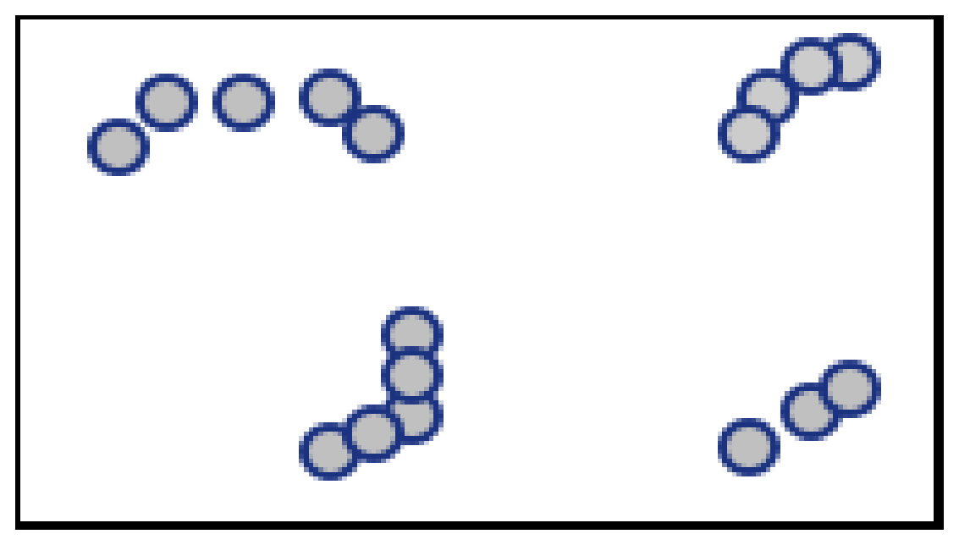
Our **Proportional Distance** is used to evaluate the similarity between two tag vectors (smaller = more similar).
Let t_1 and t_2 be two non-null tag vectors over the same corpus of size n. The proportional distance $PD(t_1,t_2)$ is defined by the following formulas:

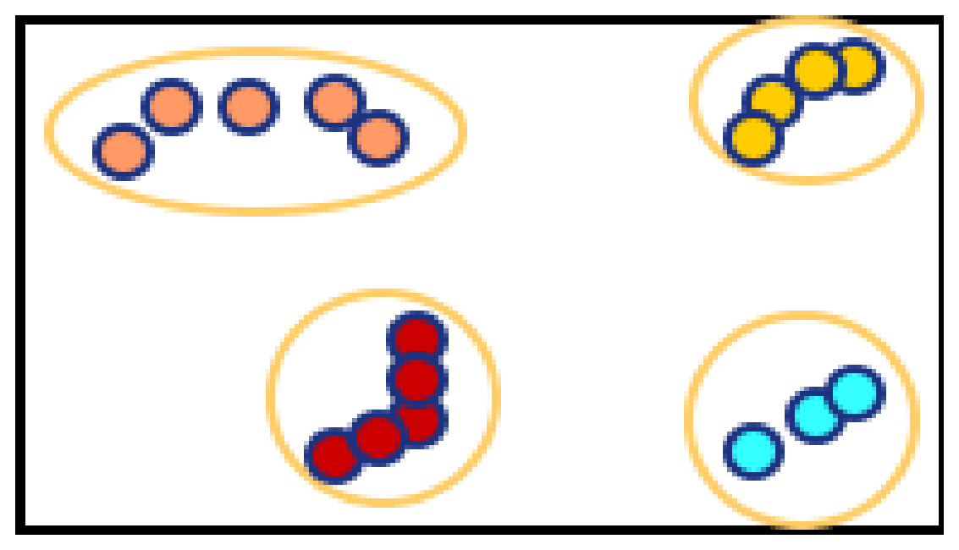
$$D(x,y) = \begin{cases} 1 & if\ x \neq y \\ 0 & if\ x = y \end{cases} \qquad L(x,y) = \begin{cases} 1 & if\ x \neq 0\ OR\ y \neq 0 \\ 0 & if\ x = 0\ AND\ y = 0 \end{cases}$$

$$PD(t_1,t_2) = \frac{\sum_{i=1}^{n} D(t_1[i], t_2[i])}{\sum_{i=1}^{n} L(t_1[i], t_2[i])}$$

Clustering Algorithm

Given a set of tag vectors and a threshold H, our clustering algorithm groups together the vectors whose proportional distance is less than H.

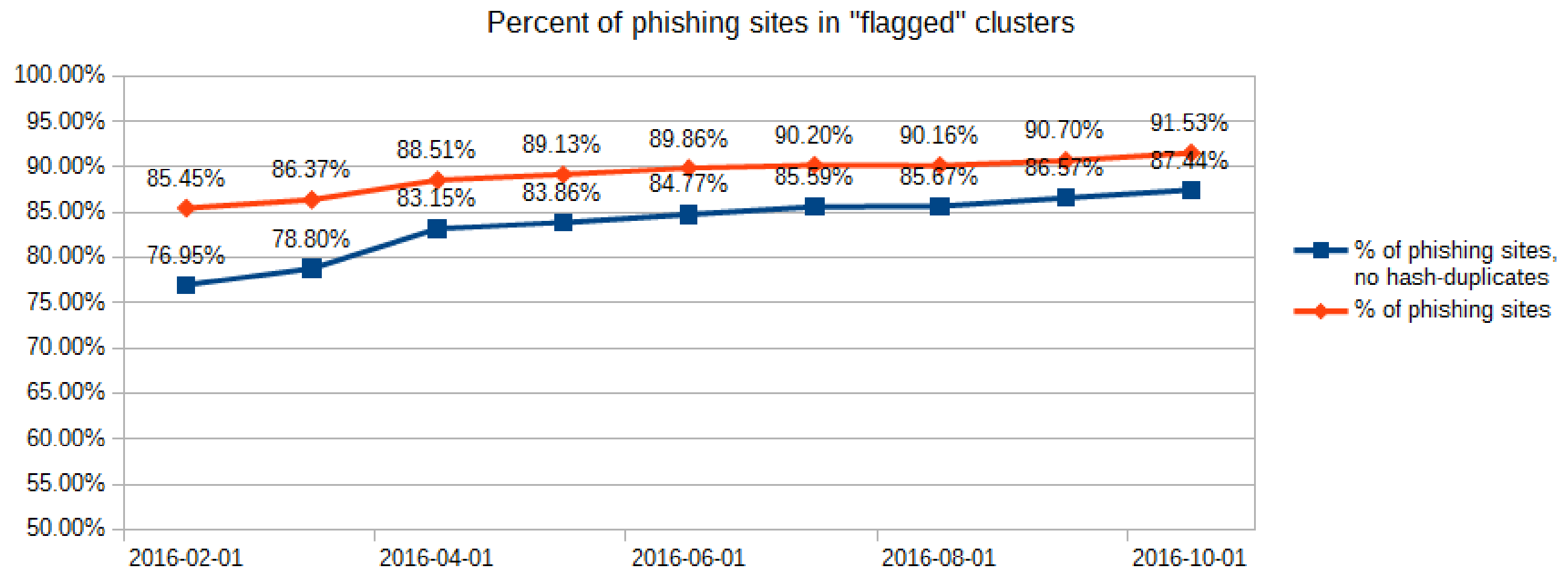

Before Clustering


After Clustering

Experiments

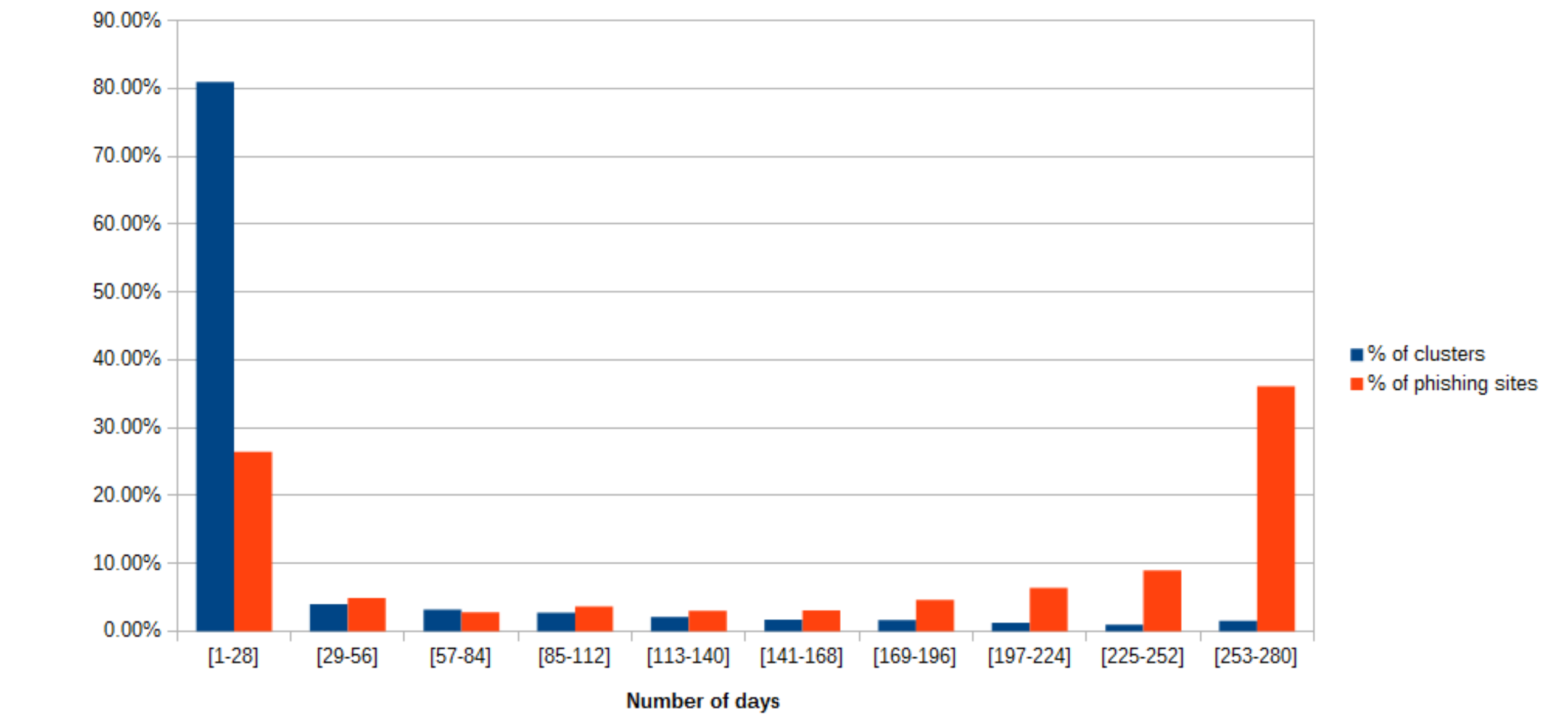
We have tested our method with 19,066 phishing sites taken from PhishTank between January 1st to October 5th , 2016.

Hash duplicates: remove all the spaces from DOM and replace all the default value in each INPUT field by empty strings, calculate the SHA-1 hash of resulting DOM. The phishing sites with the same hash hosted on the same IP.



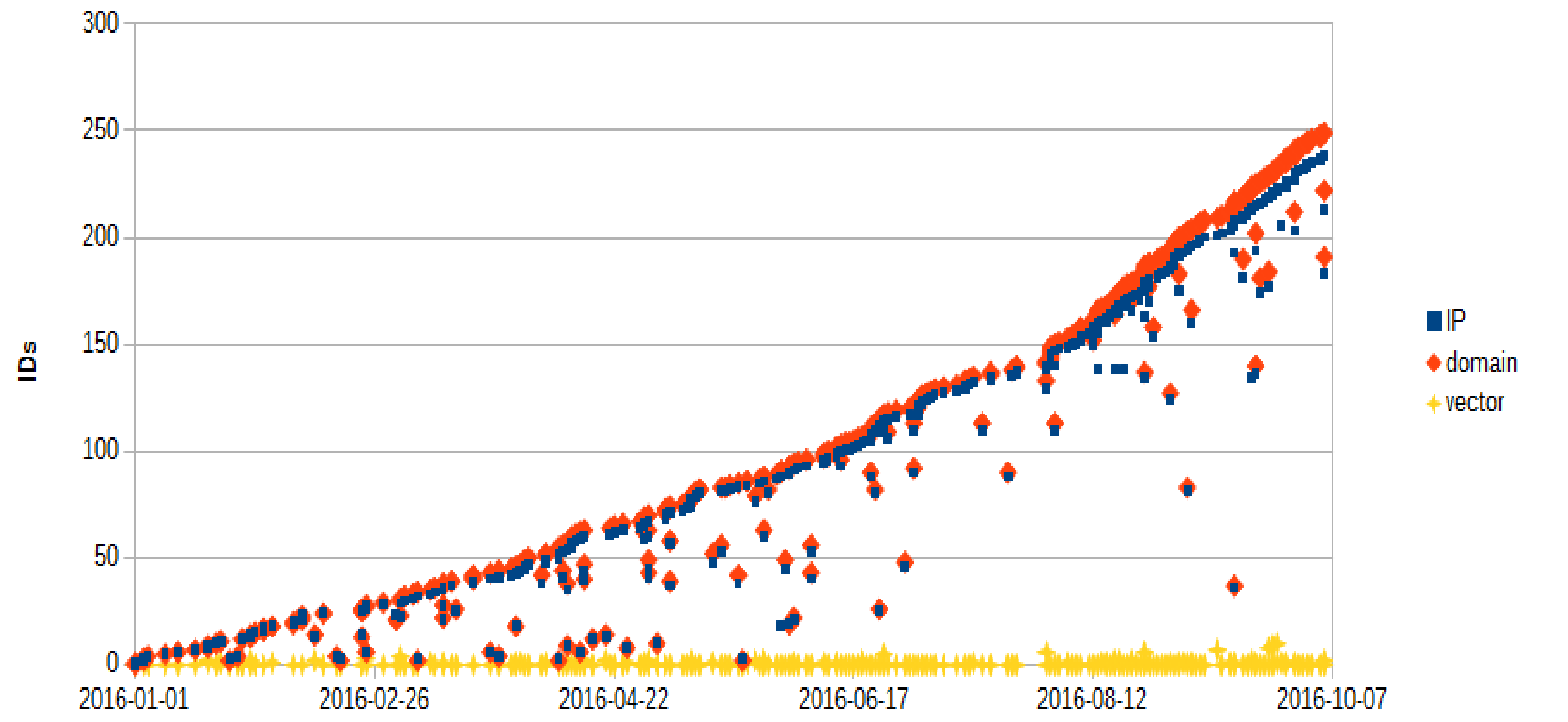
Percentage of duplicates over time

% of clusters and % of phishing sites in the clusters shows a bimodal distribution



Cluster lifespan

When we find a new record, we assign an ascending IDs to it



Number of vectors, IP addresses and domains used in clusters over time

Conclusion

Conclusion: 1) We have shown that Phishers repeatedly re-publish their attacks by minor modifications or seeking new host, which can keep the attack active for a long time 2) Our clustering method is very efficient, with about 90% replicas reported, the more phishing instances caught, the higher chance to prevent further attacks.