Don’t Let One Rotten Apple Spoil the Whole Barrel: Towards Automated Detection of Shadowed Domains

> Daiping Liu, Zhou Li, Kun Du, Haining Wang, Baojun Liu, Hai-Xin Duan. CCS 2017.

Lyu Jiuyang, Oct 30th, 2021.

Task

domain shadowing: 破坏合法域名并在其下产生恶意子域。

Target: 检测合法域下的多个恶意子域的存在。

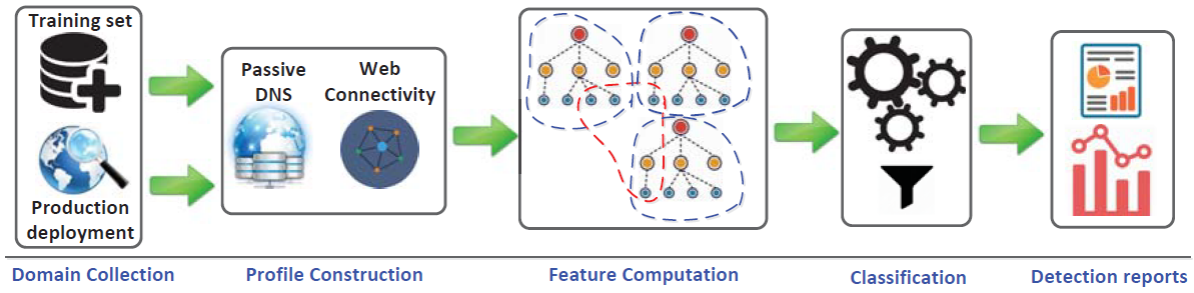

Figure 3: Workflow of Woodpecker.

Dataset：(manual, undisclosed)

Format: $D = s_i | s_i := < name_i, rrtype, rdata, t_f, t_l, count >$

Description: FQND; type fields and data fields returned by DNS servers; the time when an individual $rdata$ is first and last seen; the number of DNS queries that receive the $rdata$ in response.

Example: `{"rrname": "eu.account.amazon.com", "rrtype": "A", "rdata": "52.94.216.25;", "count": 31188, "time_first": 1477960509, "time_last": 1494290720}`

Components:

Dataset	Category
$D_{shadowed}$	Shadowed
$D_{unknown}$	Unlabeled siblings of $D_{shadowed}$
D_{pop}	Legitimate popular
D_{nonpop}	Legitimate unpopular
D_{vt}	Daily feeds from VirusTotal

Features

- Deviation from legitimate domains under the same apex.
- Correlation among shadowed domains under a different apex.

Category	Feature ID	Feature Name	Dimension	Novel
Subdomain Usage	F1	Days between 1st non-www and apex domain	D	√
	F2	Ratio of popular subdomains under the same apex domain	D	√
	F3	Ratio of popular subdomains co-hosted on the same IP	C	√
	F4	Web connectivity of a subdomains	D	√
	F5	Web connectivity of subdomains under the same apex domains	D	√
	F6	Web connectivity of subdomains co-hosted on the same IP	C	√
Subdomain Hosting	F7	Deviation of a subdomain's hosting IPs	D	√
	F8	Average IP deviation of subdomains co-hosted on the same IP	C	√
	F9	Correlation ratio in terms of co-hosting subdomain number	C	[14]
	F10	Correlation ratio in terms of co-hosting apex number	C	[14]
Subdomain Activity	F11	Distribution of first seen date	C	√
	F12	Distribution of resolution counts among subdomains on the same IP	C	√
	F13	Reciprocal median of resolution counts among subdomains on the same IP	C	√
	F14	Distribution of active days among subdomains on the same IP	C	√
	F15	Reciprocal median of active days among subdomains on the same IP	C	√
Subdomain Name	F16	Diversity of domain levels	C	√
	F17	Subdomain name length	C	[11, 33]

Table 3: Features used in our approach to detect shadowed domains. Feature dimensions D and C denote Deviation and Correlation, respectively. Although some features use the same data source as previous work, e.g., resolution counts as in [4, 51], we model them in different ways.

Days between first non-www and apex domain.

s is the first non-www subdomain under its apex.

$$F1 = \frac{1}{log(Date(s) - Date(apex(s)) + 1)}$$

Ratio of popular subdomains

Shadowed domains names tend to avoid being overlapped with popular subdomain names.

$$F2 = \frac{|\{POP\,(d_i)\}|}{|\{d_i \mid 2LD\,(d_i) == 2LD(s)\}|}$$

$$F3 = \min_{j=1\ldots n} \left\{ \frac{\left|\left\{POP\left(\hat{d}_i\right)\right\}\right|}{|\{d_i \mid IP\,(d_i) == IP_j(s)\}|} \right\}$$

2LD：二级域名。

Web connectivity

通过公共数据集检查是否被索引。

$$F5 = \frac{\sum WEB\left(d_i\right)}{|\{d_i \mid 2LD\left(d_i\right) == 2LD(s)\}|}$$
$$F6 = \min_{j=1\ldots n}\left\{\frac{\sum WEB\left(d_i\right)}{\{d_i \mid IP\left(d_i\right) == IP_j(s)\}|}\right\}$$

Deviation of hosting IP

$$F7 : \mathrm{Dev}(A,S) = \max_{j=1..m}\left\{\min_{i=1..n}\left\{\psi\left(A_i,S_j\right) \mid A\left(f_i\right) < S\left(f_j\right)\right\}\right\}$$

$$\psi\left(A_i,S_j\right) = \sum_{C\in\{IP,ASN,CC\}} w_k\left(C\left[A_i\right] \neq C\left[S_j\right]\right)$$

$F8$: average deviation of all subdomains hosted on the same IP.

Correlation ratio

Compute how many subdomains are co-hosted with s.

$$F9 = \min_{j=1..n}\left\{\frac{1}{\log(|\{d_i \mid IP_j\left(d_i\right) == IP_j(s)\}|+1)}\right\}.$$

Count the distinct apex whose subdomains are hosted together with s.

$$F10 = \min_{j=1..n}\left\{\frac{1}{\log(|\{2LD\left(d_i\right) \mid IP_j\left(d_i\right) == IP_j(s)\}|+1)}\right\}$$

Distribution of first seen date

$F11$: the Jeffrey divergence of the first seen date (in the format of MMDD-YYYY) among all subdomains hosted together with s.

Resolution count

the Jeffrey divergence ($F12$) and the reciprocal of median ($F13$) of resolution count.

Active days

the Jeffrey divergence ($F14$) and the reciprocal of median ($F15$) of active days.

Diversity of domain name levels

the Jeffrey divergence ($F16$) for all of the subdomains hosted together.

Subdomain name length

$$F17 = \frac{\sum_{i=1}^{m} Jeffrey\left(N_i\right)}{m}$$

They assess the importance of our features through a standard metric in the RandomForest model, namely mean decrease impurity (MDI), which is defined as

$$\mathrm{MDI}(X_m) = \frac{1}{N_T} \sum_{T} \sum_{t \in T:(s_t)=X_m} p(t) \Delta f\left(s_t, t\right)$$

Train

- classifiers, 10-fold cross-validation
- metrics: ROC, FPR, TPR
- RandomForest performs best. RandomForest and Neutral Network consistently outperform Logistic Regression and linear SVM.

个人感觉

- 论文重心放在了feature的构建上而非模型结构上，部分构建feature的角度、5.1阴影域名的其他应用有参考意义。