WebAA: Website Association Analysis via Multi-Resource Similarity Computation

Taiyao Zhang, Dongzheng Jia, Xingyu Fu, Zhihao Zhang, Qingyun Liu

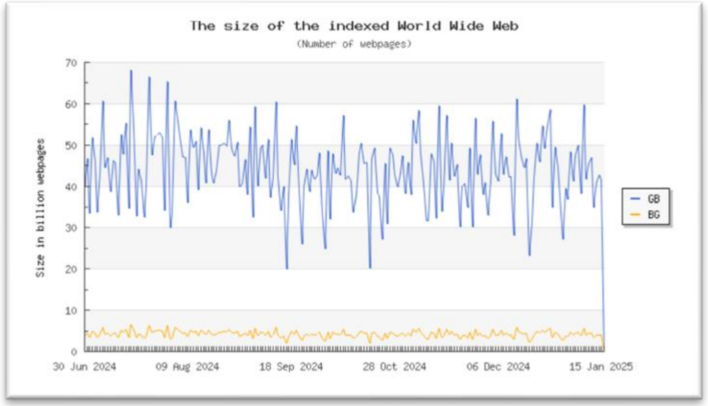
Institute of Information Engineering, Chinese Academy of Sciences

School of Cyber Security, University of Chinese Academy of Sciences

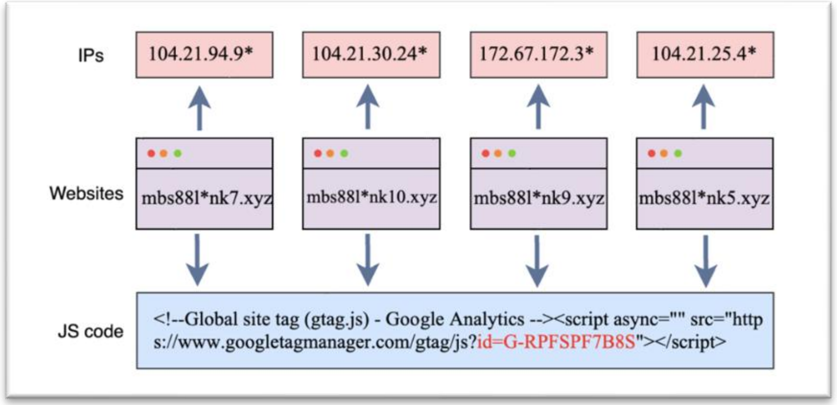
National Computer Network Emergency Response Technical Team/Coordination Center of China

Research background and significance

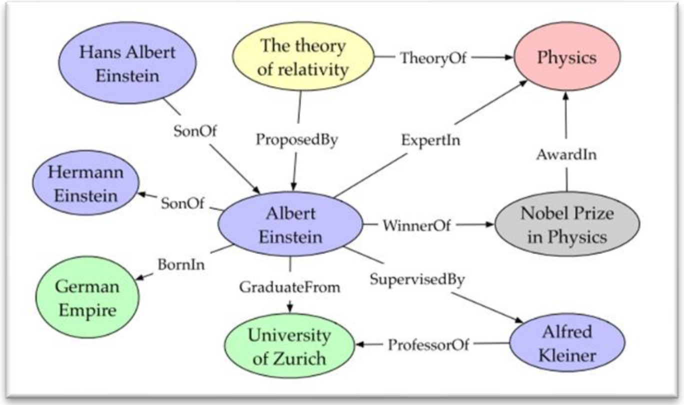
- The scale of the Internet remains vast throughout the year, presenting challenges for network management.
- Malicious organizations often create multiple associated websites to conceal their identities or expand their scope of influence.
- Different websites within the same organization may contain complementary resources. By associating these websites at the organizational level, a more comprehensive and efficient knowledge network can be established.



The huge scale of the Internet



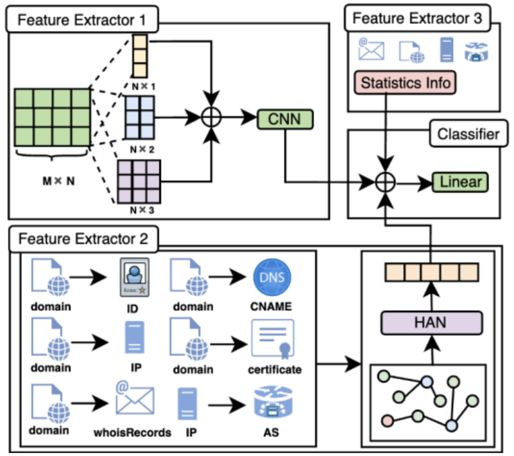
A organization operates multiple websites



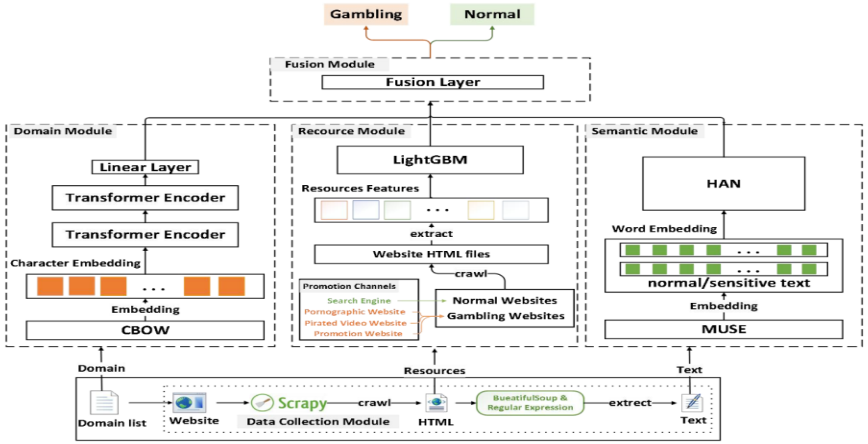
Knowledge complementarity

Related Work

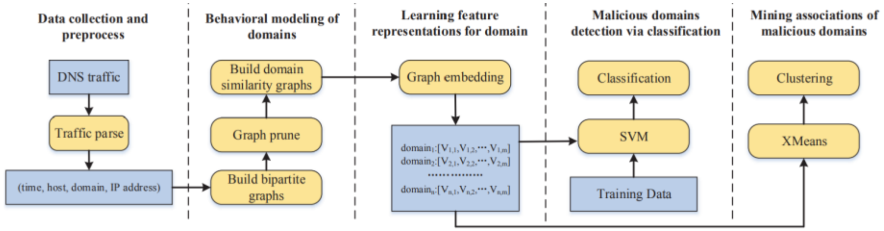
- **IDTracker:** Through third-party service IDs to associate malicious websites at the organizational level.
- **DRSDetector:** Uses CBOW, LightGBM and HAN to analyze various resources of the website and determine whether it is malicious.
- **Lei et al:** Capture and characterize the domain name resolution process using a bipartite graph and GNN to detect malicious websites.



IDTracker [1]



DRSDetector [2]



Lei et al [3]

[1]. Idtracker: Discovering illicit website communities via third-party service ids, Chenxu Wang et al. In DSN 2023
[2]. Drsdetector: Detecting gambling websites by multi-level feature fusion, Yuxin Zhang et al. In ISCC 2023
[3]. Detecting malicious domains with behavioral modeling and graph embedding, Kai Lei et al. In ICDCS 2019

Limitaitons

- **Resource limitations**

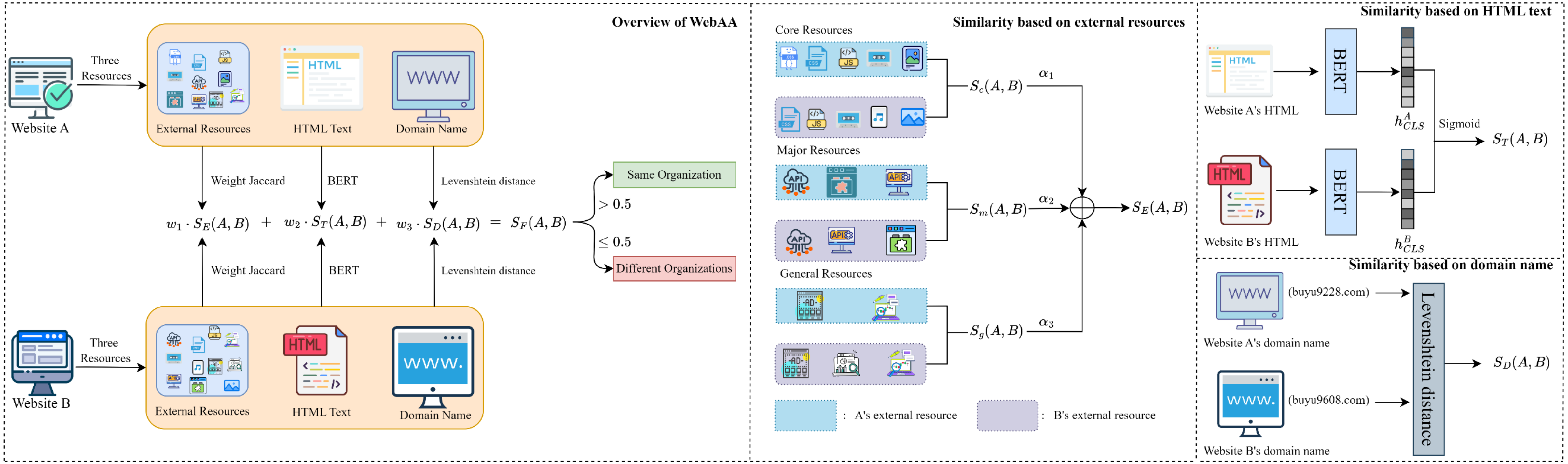
Existing methods perform well in identifying malicious websites, but some of the resources they rely on (such as IP addresses, WHOIS information, etc.) are difficult to obtain.

- **Only website-level detection**

Existing methods are focused solely on the website level, which is insufficient for effective website governance.

WebAA

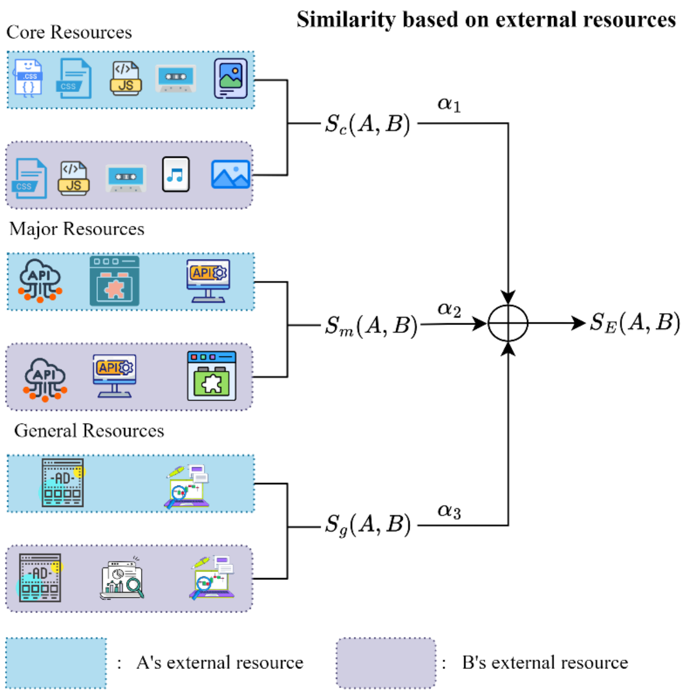
- **WebAA: Web**site **A**ssociation **A**nalysis via Multi-Resource Similarity Computation



The overview of WebAA

WebAA

Similarity based on External Resources



Resource Classification:

core resources: e.g. images, audio and video, CSS files

major resources: e.g. third-party plug-ins

general resources: e.g. advertising and analytics tools

Weighted Jaccard Similarity:

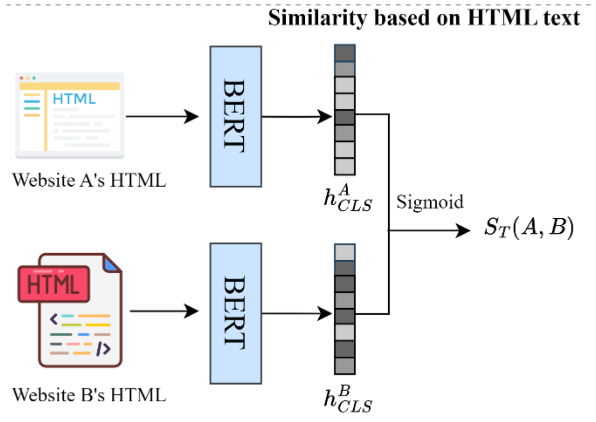
$$S_c\ (A,B)=\frac{\sum_{x\in D_A^{\text{core}}\cap D_B^{\text{core}}}\min\left(w_x^A,w_x^B\right)}{\sum_{x\in D_A^{\text{core}}\cup D_B^{\text{core}}}\max\left(w_x^A,w_x^B\right)}$$

External Resources Similarity:

$$S_E(A,B)=\alpha_1 S_c(A,B)+\alpha_2 S_m(A,B)+\alpha_3 S_g(A,B)$$

WebAA

Similarity based on HTML Texts



HTML Embedding based on BERT:

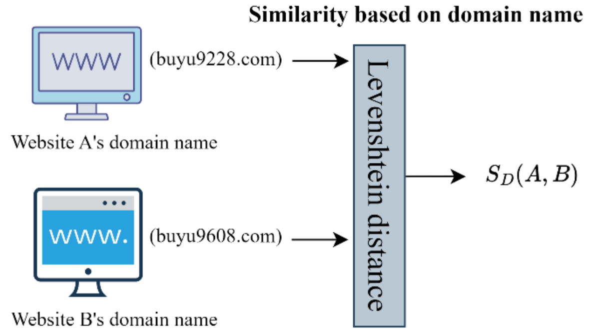
$$H^l = Transformer(H^{l-1})$$

$$H = \mathrm{BERT}(T) \quad \Rightarrow \quad h_{\mathrm{CLS}} = H[0]$$

HTML Similarity:

$$S_T(A,B) = \sigma(W \cdot [h^A_{CLS} \oplus h^B_{CLS}] + b)$$

Similarity based on Domain Names



Levenshtein Distance:

$$\mathrm{lev}_{d_A,d_B}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j)=0 \\ \min \begin{cases} \mathrm{lev}_{d_A,d_B}(i-1,j)+1 \\ \mathrm{lev}_{d_A,d_B}(i,j-1)+1 \\ \mathrm{lev}_{d_A,d_B}(i-1,j-1)+1_{(d_{Ai}\neq d_{Bj})} \end{cases} & \text{otherwise.} \end{cases}$$

Domain Name Similarity:

$$S_D(A,B) = \frac{1}{lev_{d_A,d_B}(|d_A|,|d_B|)+1}$$

Experiments

Datasets

Dataset	Web Num	Org Num	Largest Org	Smallest Org
D_{Legal}	4,746	3,127	281	1
$D_{Illegal}$	8,998	1,606	546	1

D_{Legal} : Legitimate website dataset, constructed based on website registration number [1]

$D_{Illegal}$: The illegal website dataset released by Wang et al. [2]

Metrics

- ACC, Recall, F1, and Time

Questions

– **(RQ1)** In the real dataset, are there correlations between websites under the same organization in terms of external resources **(RQ1-1)**, HTML text **(RQ1-2)**, and domain names **(RQ1-3)**?

– **(RQ2)** How does WebAA model perform on real datasets?

– **(RQ3)**How do different modules of the model affect the model performance?

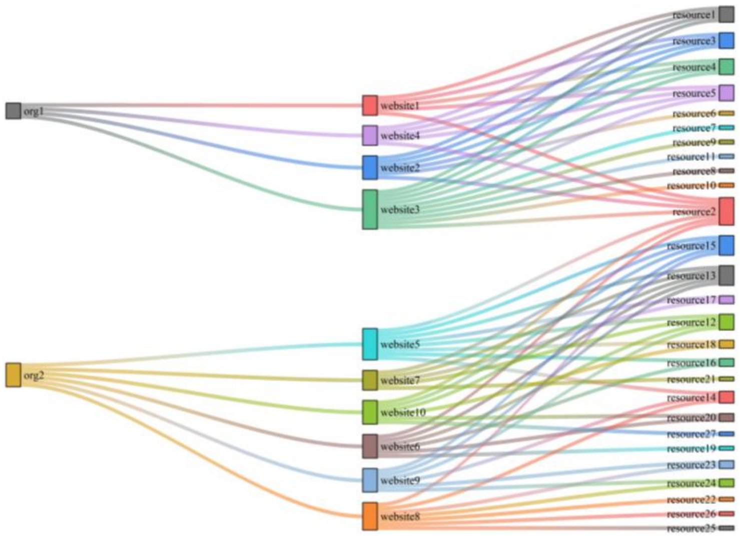
Environments

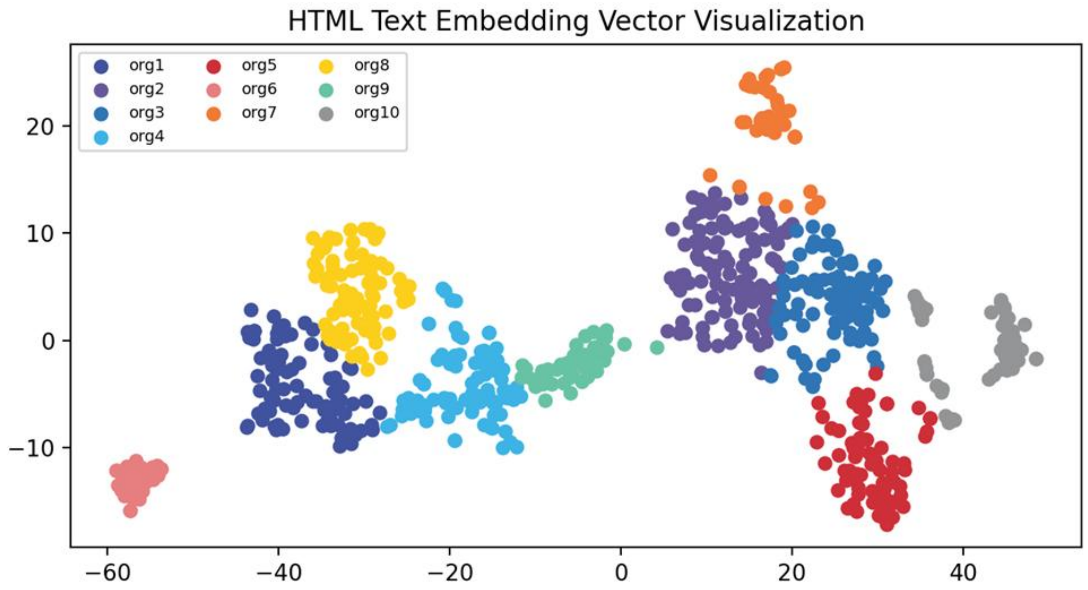
- Model build: Python 3.8.18, Pytorch 2.2.1
- Model train: Ubuntu 20.04, A100*2, CUDA 11.2

[1]. A unique identifier assigned to an organization in mainland China when it registers a website or application.
[2]. https://github.com/IDTrackerSystem/dataset

Experiments

Website Resource Analysis (answer RQ1)





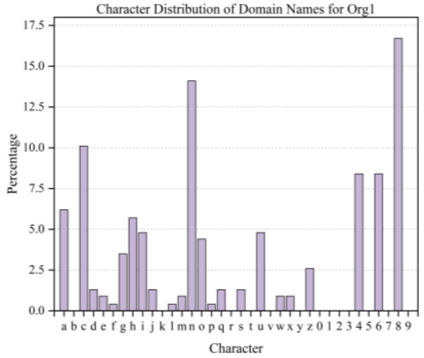
- External resource analysis
- HTML text analysis

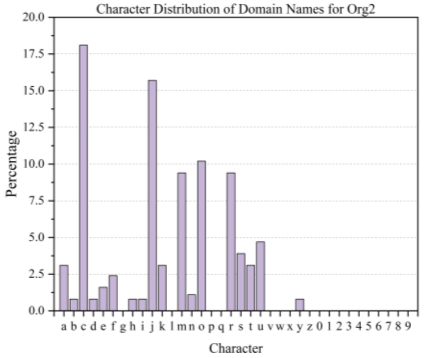
RQ 1-1: Websites within the same organization reuse a significant amount of external resources, whereas only a small number of resources are shared between websites from different organizations.

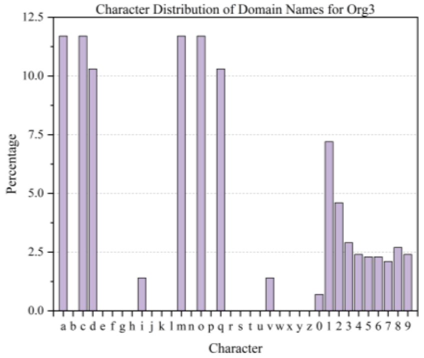
RQ 1-2: The HTML texts of websites within the same organization exhibit significant similarities, while those from different organizations are markedly distinct.

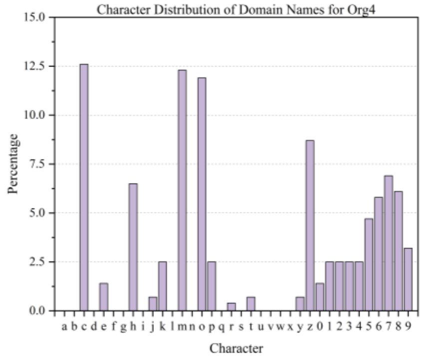
Experiments

Website Resource Analysis (answer RQ1)

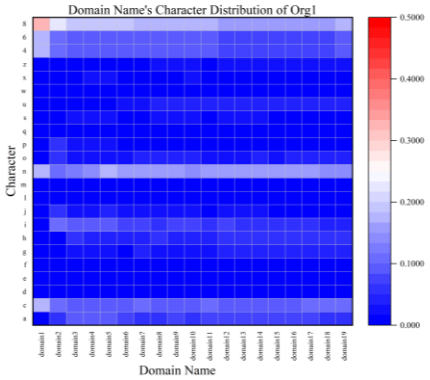


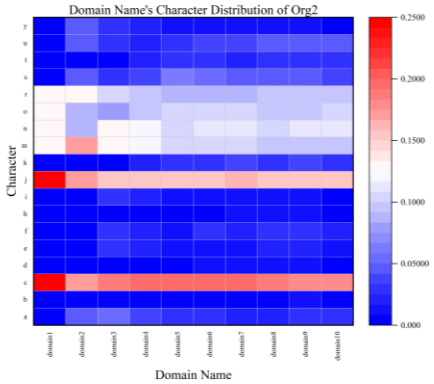


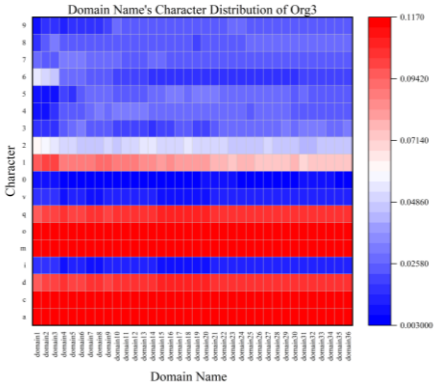


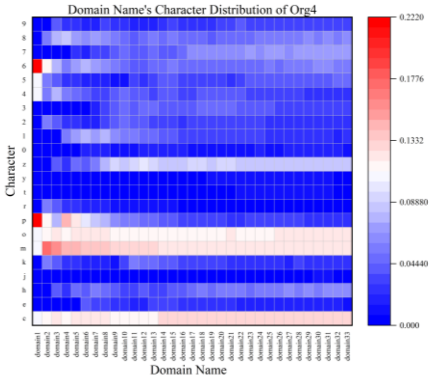


- character distribution of each organization









- character distribution of each website

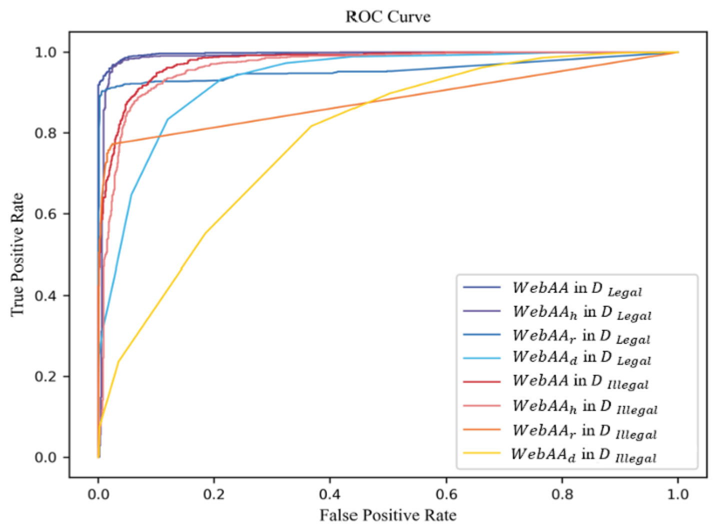
RQ 1-3: The character distribution of domain names among different organizations varies significantly, whereas the character distribution of domain names within an organization tends to be similar.

Experiments

WebAA performance (answer RQ2) and Ablation experiment (answer RQ3)

Model	D_{Legal} Acc	Recall	F1	Time(s)	$D_{Illegal}$ Acc	Recall	F1	Time(s)
WebAA	**97.15**	95.10	**97.09**	0.45	**92.95**	**92.70**	**92.93**	0.45
WebAA_r	91.75	83.70	91.03	0.03	86.90	75.80	85.26	**0.12**
WebAA_h	96.80	96.50	96.79	0.38	90.85	88.50	90.63	0.40
WebAA_d	82.45	**97.30**	84.72	**0.02**	69.80	89.80	74.83	0.14

- Experimental results on two datasets



- ROC curves

RQ 2: On both datasets, the accuracy, recall and F1 score of the WebAA model exceed 90%. The WebAA model can complete the association of thousands of website pairs within milliseconds on both datasets.

RQ 3: Among the variant models, $WebAA_h$ demonstrates the best performance. The performance of the $WebAA_d$ is relatively poor. The three variants each have their own advantages and contribute to the WebAA model to varying degrees.

Conclusion

- **New Concept**: This is the first time to propose the concept of website association, which focuses on the organizations behind websites, enabling regulators to effectively manage legitimate websites and fundamentally combat illicit ones.

- **New Technique**: We innovatively propose a website association model, WebAA, which can accurately and efficiently perform the association task by analyzing only three easily accessible resources of the target websites.

- **New Datasets**: We manually constructed two real-world datasets and are releasing them[1] to support community researchers in conducting studies related to this field.

- **New Promotion**: Extensive experiments on two real-world datasets demonstrate that our model can efficiently associate thousands of website pairs within milliseconds with an accuracy exceeding 90%.

[1]. github: https://github.com/SevenZhang123/WebAA-Datasets

Thanks for your listening

Taiyao Zhang, Dongzheng Jia, Xingyu Fu, Zhihao Zhang, Qingyun Liu

Institute of Information Engineering, Chinese Academy of Sciences

School of Cyber Security, University of Chinese Academy of Sciences

National Computer Network Emergency Response Technical Team/Coordination Center of China