KnowPhish:
Large Language Models Meet Multimodal Knowledge Graphs for Enhancing Reference-Based Phishing Detection

Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao,
Nay Oo, Hoon Wei Lim, Bryan Hooi

Presenter: *Yuexin Li*

33rd USENIX Security Symposium
August 14-16, 2024

Background: What is Phishing?

Phishing webpages usually

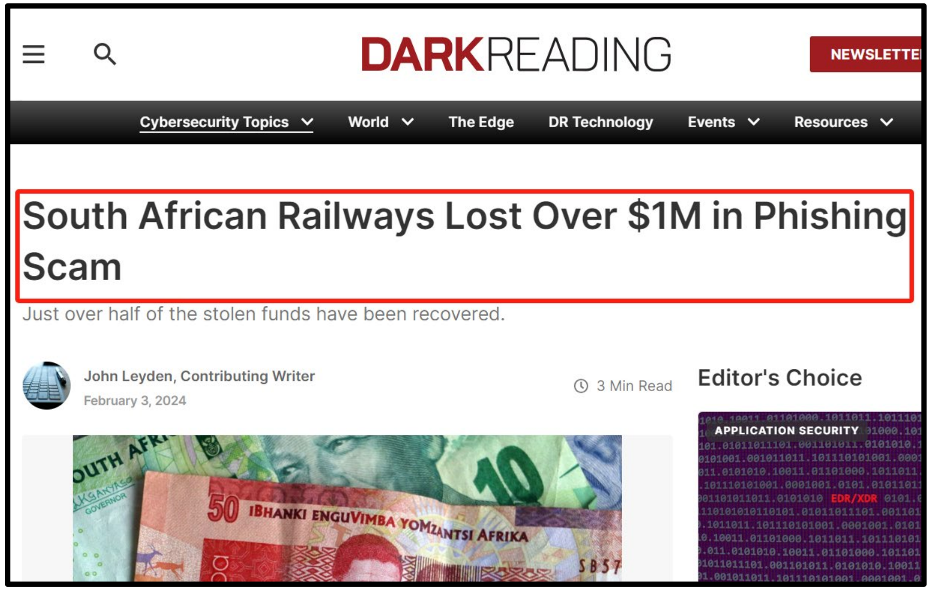
1. **Impersonate** themselves as **popular brands** (e.g. PayPal, Bank of America, DHL)
2. Use a **different domain** from the legitimate ones
3. Require users to **submit credentials**

Background: Why Phishing Detection?

Phishing attacks are ubiquitous in cyberspace with severe consequences

- Effective and efficient phishing detection systems are urgently needed

Singapore



South Africa

Norway

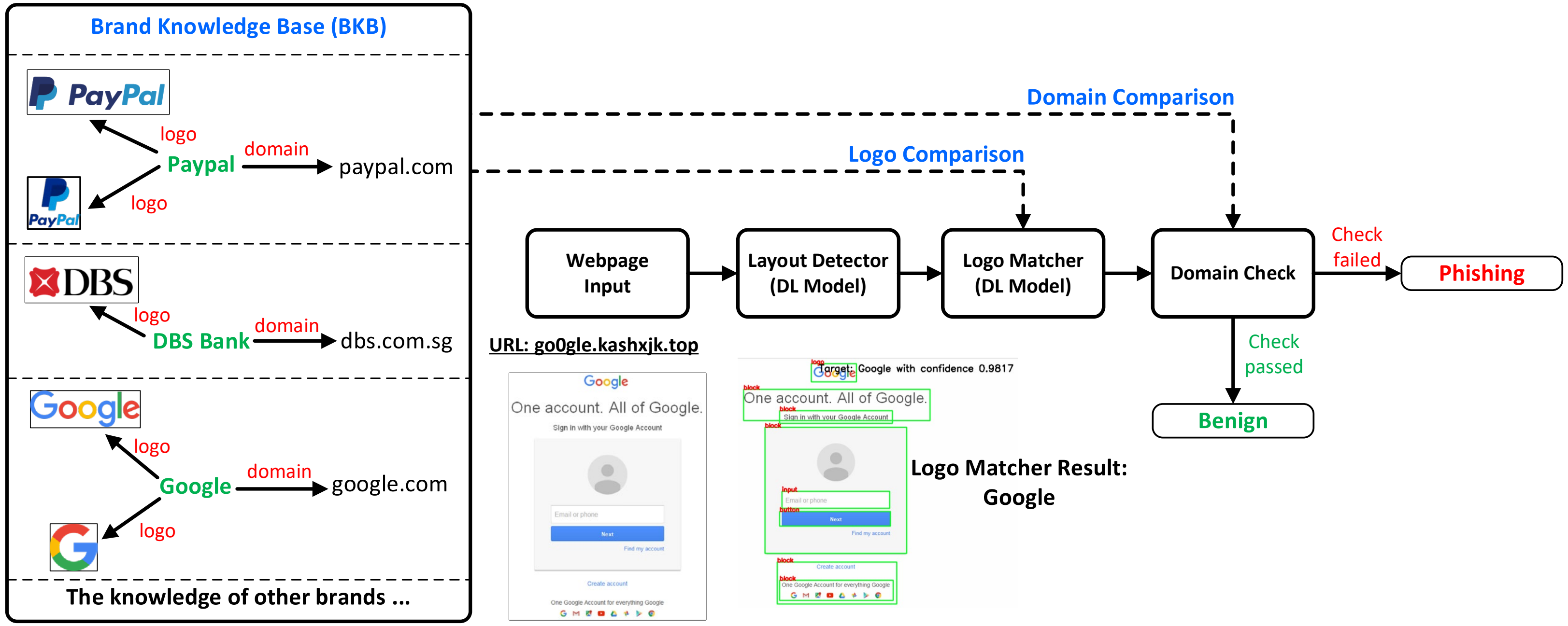
[1] https://www.straitstimes.com/singapore/courts-crime/scam-victims-in-s-pore-lost-6518m-in-2023-with-record-high-of-over-46000-cases-reported
[2] https://www.darkreading.com/endpoint-security/south-african-railways-reports-1m-phishing
[3] https://cyberscoop.com/norfund-hacked-wealth-fund-10-million/

State-of-the-art Solutions

Reference-based phishing detectors (RBPDs) using computer vision

- E.g., **Phishpedia** (*USENIX Security* 2021), **PhishIntention** (*USENIX Security* 2022)
- Utilize deep learning models to analyze the **logo** (from the screenshot) of the webpage
- If the input domain is different from the **brand's legitimate domain**, it is very likely to be phishing

1. *Generalizable*
2. *Explainable*

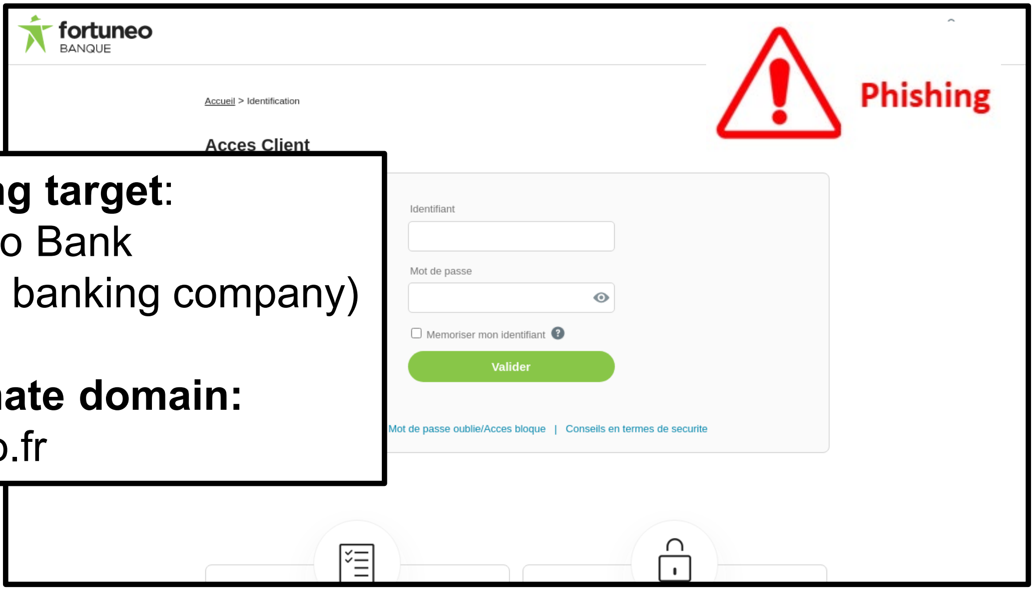


[1] Y Lin, *et al*. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. *USENIX Security* 2021.
[2] R Liu, *et al*. Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach. *USENIX Security* 2022.

Challenge 1: Limited-Scale Brand Knowledge

Existing RBPDs only maintain the knowledge of 277 brands

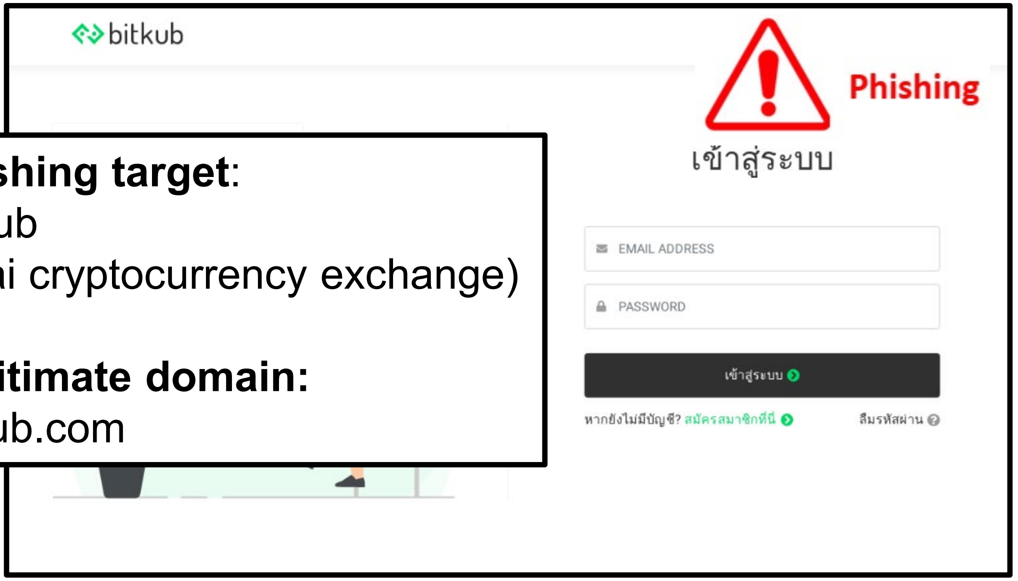
- Real-world **phishing attacks are diverse**, ranging from multinational companies (e.g. Microsoft, Facebook) to local firms
- If we do not have the brand knowledge, we are less likely to detect the phishing webpage targeting that brand



Phishing target:
Fortuneo Bank
(French banking company)

Legitimate domain:
fortuneo.fr

https://fortunneo.nl/



Phishing target:
Bitkub
(Thai cryptocurrency exchange)

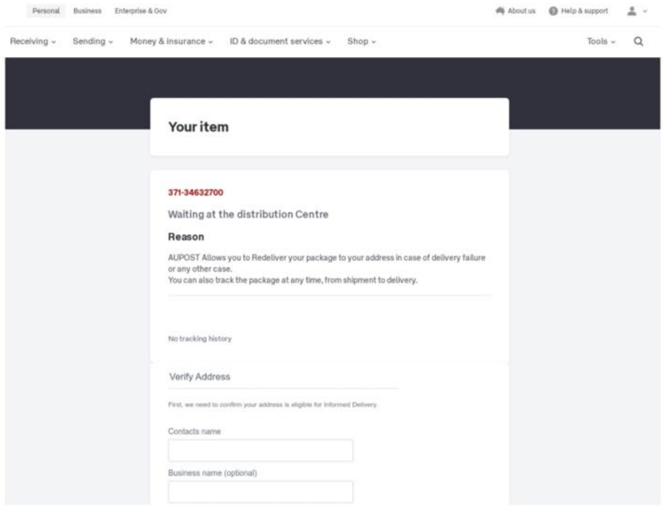
Legitimate domain:
bitkub.com

https://bitkub-th.app/wallet/

Challenge 2: Logo-less Phishing Webpage

Logo-less phishing webpage with textual brand intention

- Phishing webpage may not always convey their brand intention via logos
- Instead, they can show such intention via HTML texts
- Existing image-based RBPDs completely fails in such cases because they solely relies on logos to identify brand intention



(a) Screenshot

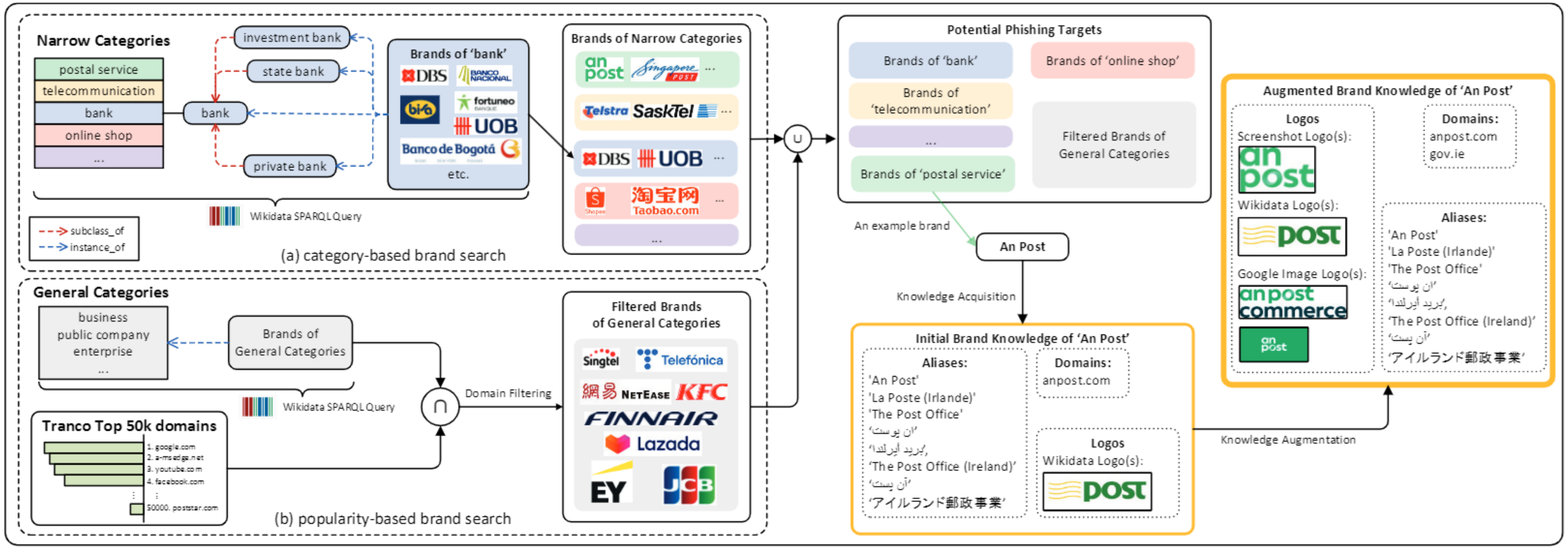
```
<head class="at-element-marker">
  <meta http-equiv="Content-Type" content="text/html">

  <title>Track your items - Australia Post </title>

  <!--<base href="/mypost/track/">-->
  <meta name="viewport" content="width=device-width">
  <meta property="og:title" content="MyPost Deliveries">
  <meta property="og:type" content="website">

  .
  .
  .

</head>
```

(b) HTML

Solution 1: KnowPhish

KnowPhish: A large-scale multimodal brand knowledge base

- Covering more than **20k potential phishing targets** worldwide
- Comprehensive **multimodal brand knowledge** (e.g., brand names and aliases, logos, and legitimate domains)



KnowPhish Construction: Motivation

"What indicates a potential phishing target?"

- **Question 1**: Do phishing targets differ across different phishing feeds?
- **Question 2**: What are the enduring characteristics shared by phishing feeds across different sources and periods?

We used two datasets for this empirical study:

Dataset	Source	Sample Size	Collection Time
D_1	Phishpedia paper	30k	2021
D_2	APWG	5k	2023

[1] Y Lin, *et al*. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. *USENIX Security* 2021.
[2] https://apwg.org/

KnowPhish Construction: Motivation

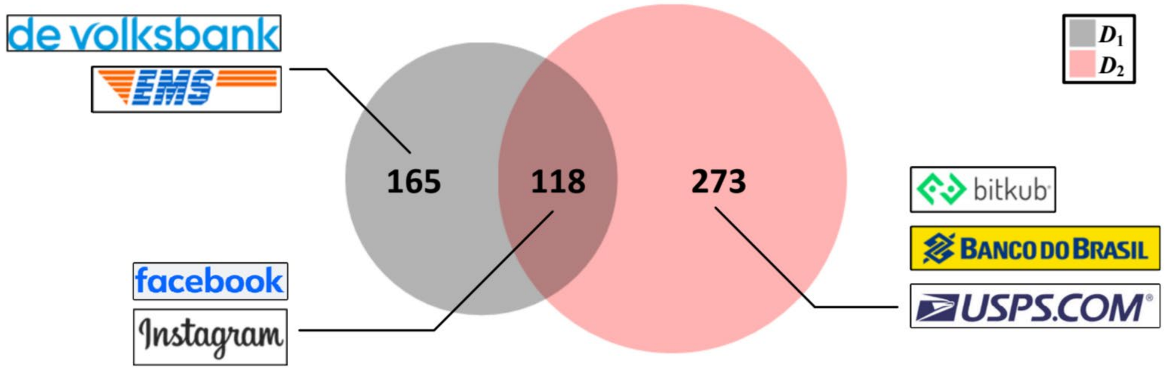
“What indicates a potential phishing target?”

- **Question 1**: Do phishing targets differ across different phishing feeds?
- **Question 2**: What are the enduring characteristics shared by phishing feeds across different sources and periods?

Observation 1:

Yes. The difference can be affected by

- Collection time
- Collection source
- Collection methodology
  - Proprietary Detectors (automated) or Human Report (manual)



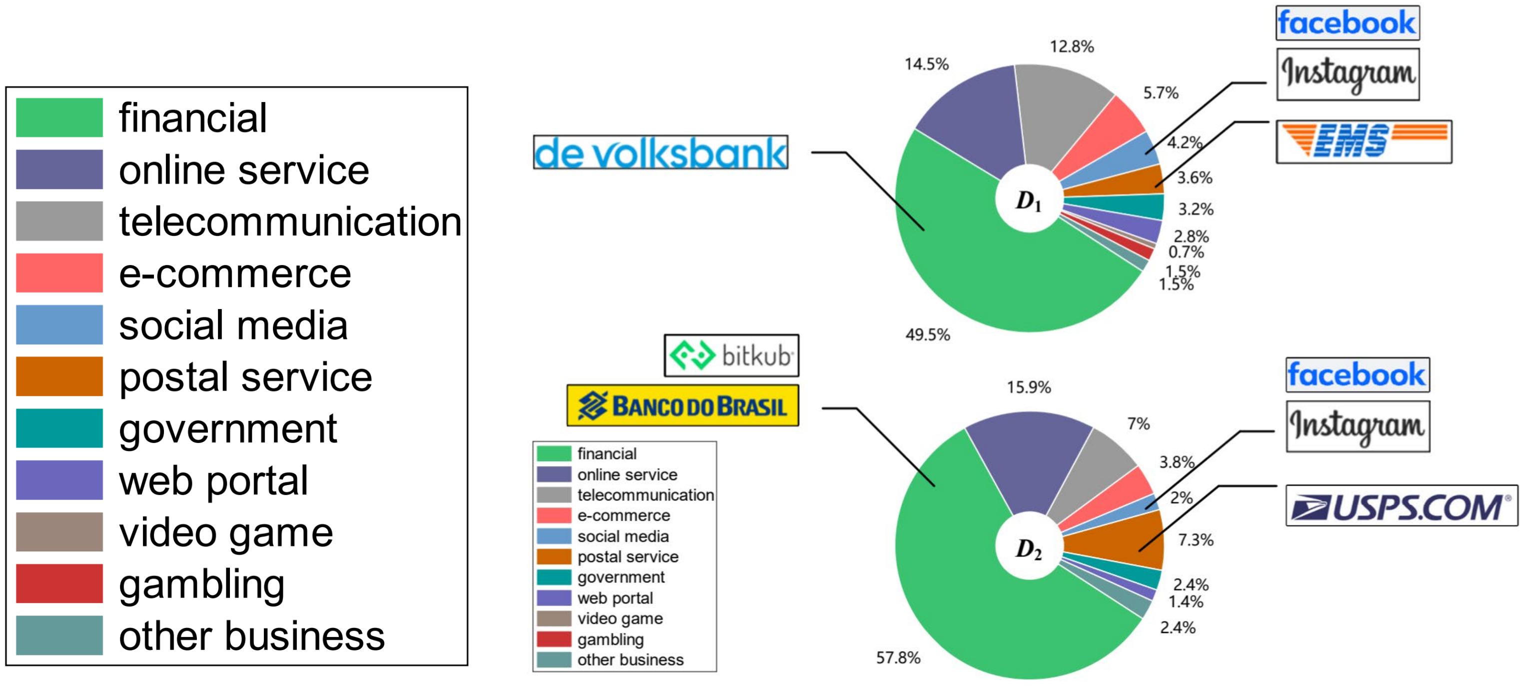
KnowPhish Construction: Motivation

"What indicates a potential phishing target?" **High-value industries!**

- **Question 1**: Do phishing targets differ across different phishing feeds?
- <u>**Question 2**: What are the enduring characteristics shared by phishing feeds across different sources and periods?</u>

<u>Observation 2:</u>

The <u>**industries**</u> of those phishing targets remain mostly consistent.



KnowPhish Construction: Motivation

High-value industries usually indicates phishing targets

- We search for potential Wikidata categories (c) of phishing targets (b) to represent the 10 high-value industries

$$(b, instance\_of, c)$$

Knowledge Graph

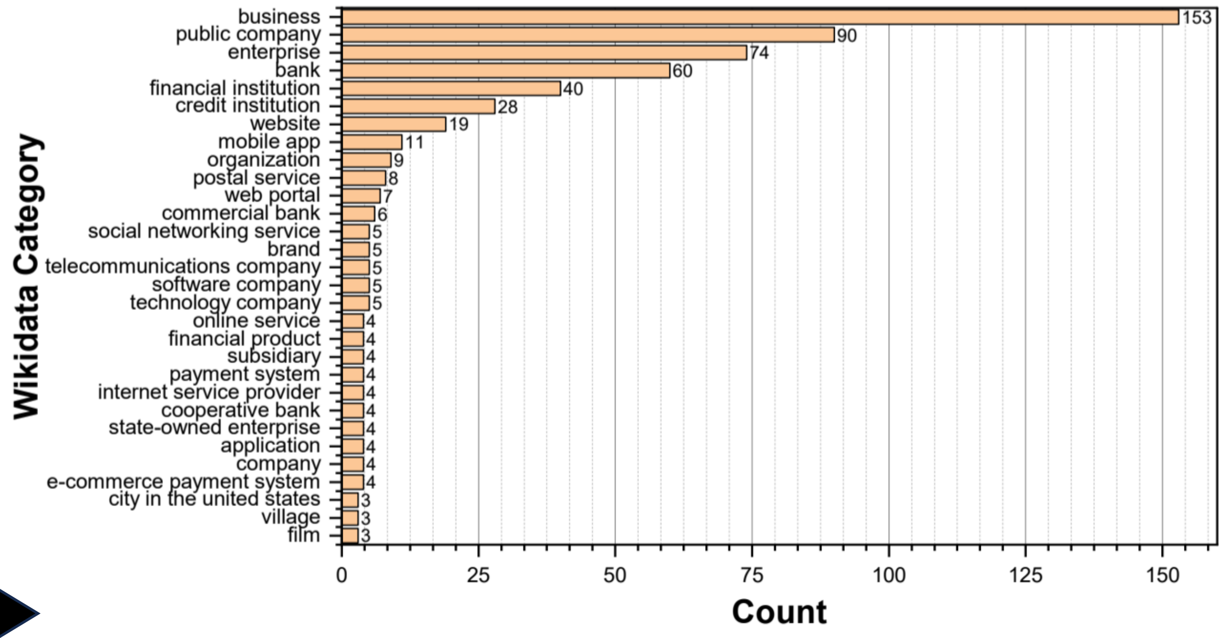


Figure 4: Distribution of the top 30 Wikidata categories of the phishing targets in D_2.

[1] Denny V, *et al*. Wikidata: A free collaborative knowledgebase. *Communications of the ACM* 2014.

KnowPhish Construction: Motivation

High-value industries usually indicates phishing targets

- **Narrow Categories C_n**: directly referring to specific high-value industries
- **General Categories C_g**: representing a wider range of potential phishing targets
- The selected Wikidata categories can further guide us to search for potential phishing targets in knowledge graph G

Industries	Wikidata Category	Wikidata ID
other business	business public company enterprise online service government organization	Q4830453 Q891723 Q6881511 Q19967801 Q2659904

Table 10: Full list of General Categories C_g

Industries	Wikidata Category	Wikidata ID
financial	bank financial institution credit institution federal credit union payment system digital wallet cryptocurrency exchange	Q22687 Q650241 Q730038 Q116763799 Q986008 Q1147226 Q25401607
online service	webmail web service mobile app office suite	Q327618 Q193424 Q620615 Q207170
telecommunication	telecommunication company mobile network mobile network operator internet service provider	Q2401749 Q15360302 Q1941618 Q11371
e-commerce	online shop online marketplace	Q4382945 Q3390477
social media	social media social networking service online video platform	Q202833 Q3220391 Q559856
postal service	postal service package delivery	Q1529128 Q1447463
government	government	Q7188
web portal	web portal web search engine	Q186165 Q4182287
video game	video game distribution platform	Q81989119
gambling	gambling	Q11416

Table 9: Full list of Narrow Categories C_n

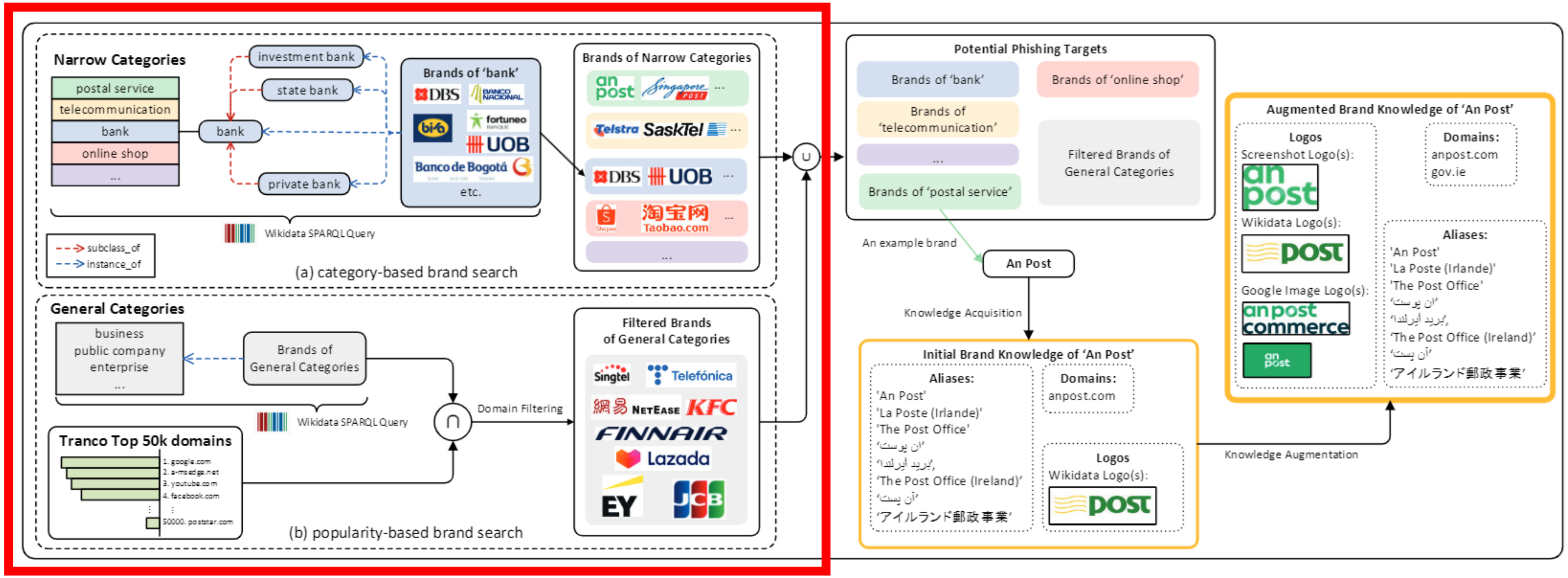
[1] Denny V, *et al*. Wikidata: A free collaborative knowledgebase. *Communications of the ACM* 2014.

KnowPhish Construction: Approach

KnowPhish constructs brand knowledge through a 2-step process

(1) Brand Search

(2) Knowledge Acquisition and Augmentation



KnowPhish Construction: Approach

Brand Search

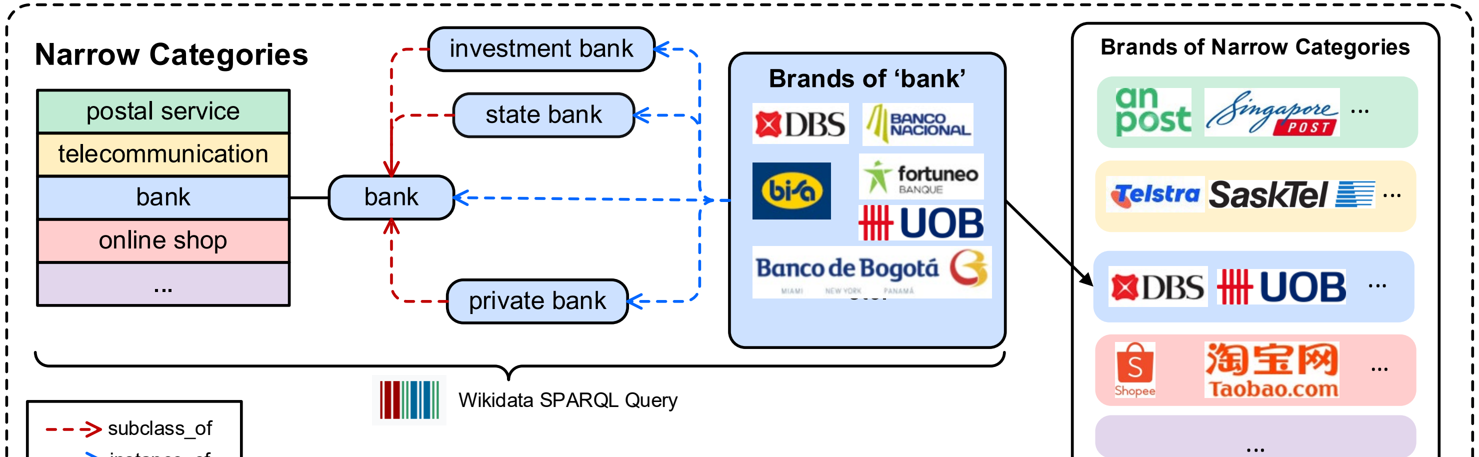
(1) Category-based Brand Search

- Search for brands that belong to Narrow Categories C_n and their subcategories, directly identifying potential phishing targets

$$\mathcal{B}_n(c_n) = \{b \mid (b, \texttt{instance\_of}, c) \in \mathcal{G}, c \in \{c_n\} \cup C'_n\}$$

Industries	Wikidata Category	Wikidata ID
financial	bank	Q22687
	financial institution	Q650241
	credit institution	Q730038
	federal credit union	Q116763799
	payment system	Q986008
	digital wallet	Q1147226
	cryptocurrency exchange	Q25401607

Examples of Narrow Categories



(a) category-based brand search

KnowPhish Construction: Approach

Brand Search

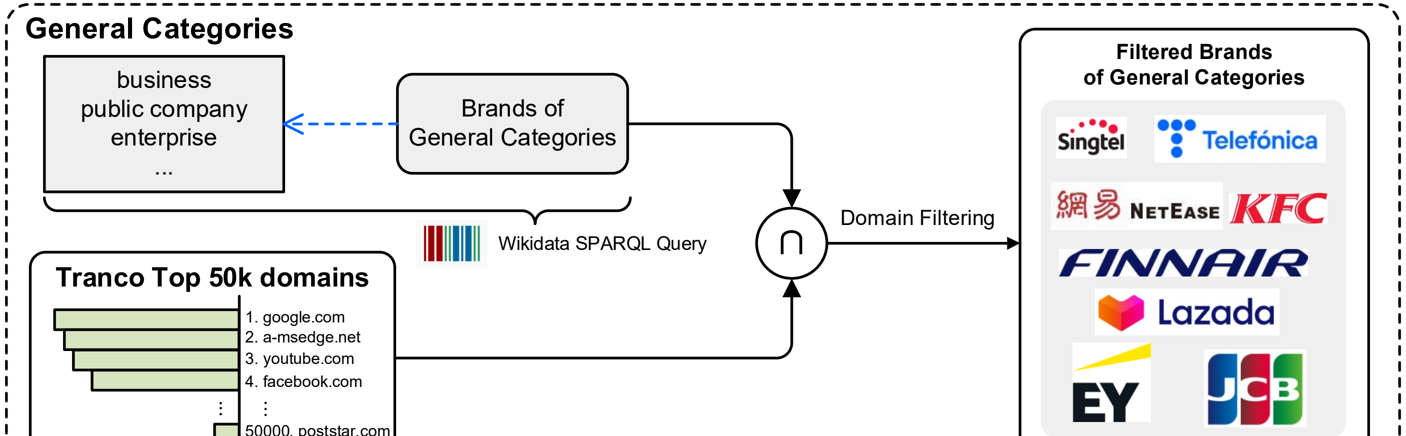
(2) Popularity-based Brand Search

- Search for brands that belong to General Categories C_g and are popular, augmenting the set of potential phishing targets

$$\boxed{\mathcal{B}_g(c_g) = \{b \mid (b, \texttt{instance\_of}, c_g) \in \mathcal{G}, r_{\mathcal{D}}(b.domains) \leq \eta\}}$$

Industries	Wikidata Category	Wikidata ID
other business	business public company enterprise online service government organization	Q4830453 Q891723 Q6881511 Q19967801 Q2659904

Table 10: Full list of General Categories C_g



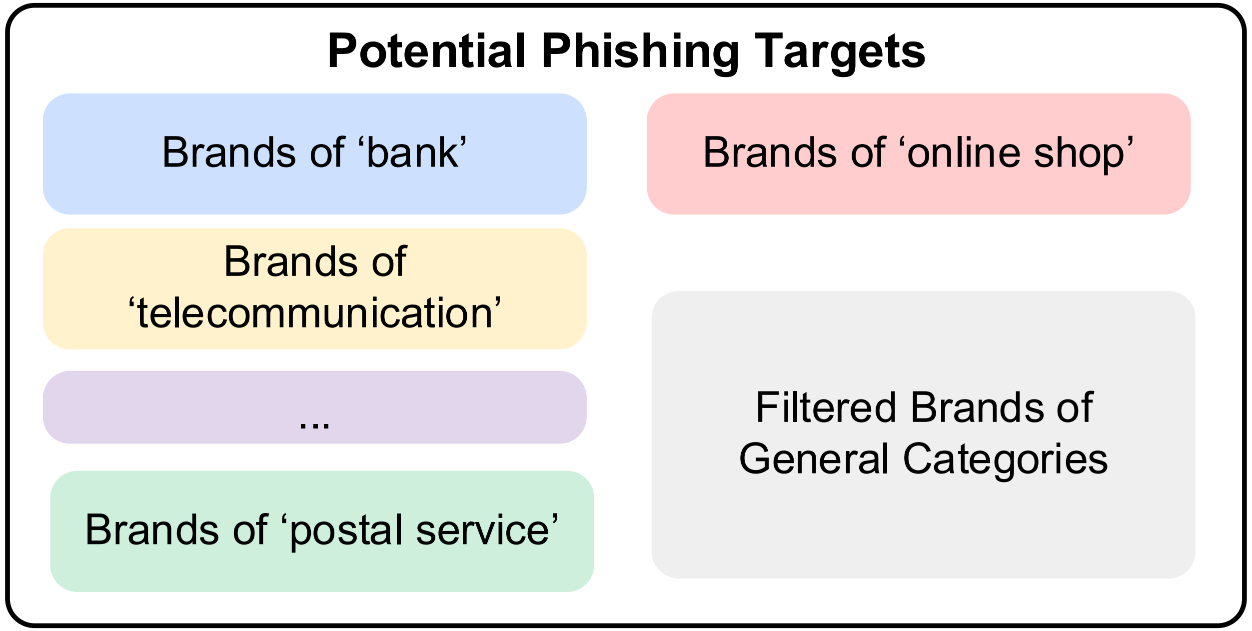
(b) popularity-based brand search

KnowPhish Construction: Approach

Brand Search

The two brand search components return a list of potential phishing targets

- Ready for brand knowledge collection
- Necessary to enhance RBPDs in terms of identifying brand intention

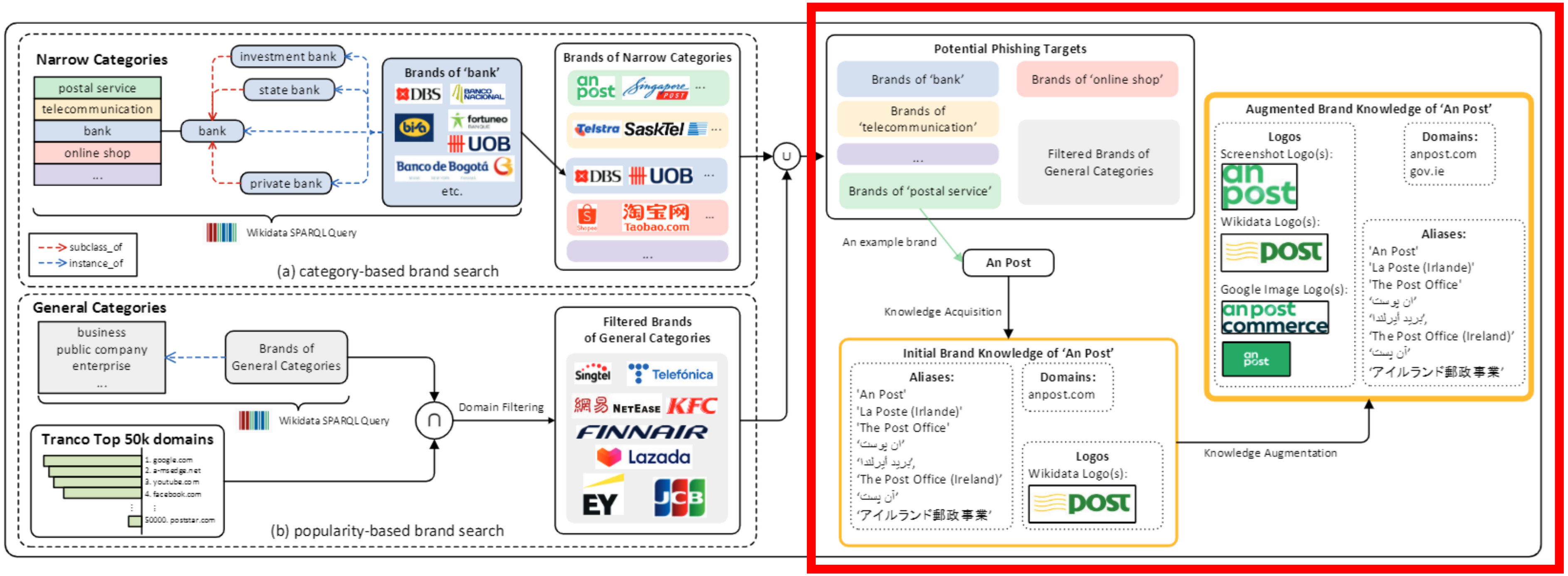


KnowPhish Construction: Approach

KnowPhish constructs brand knowledge through a 2-step process

(1) Brand Search

(2) Knowledge Acquisition and Augmentation



KnowPhish Construction: Approach

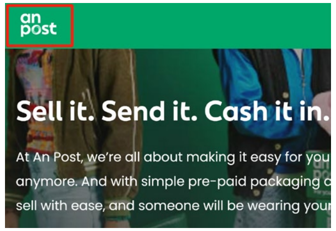
Knowledge Acquisition and Augmentation

- We collect brand knowledge with
  1. Logos
  2. Aliases (alternative names)
  3. Legitimate domains

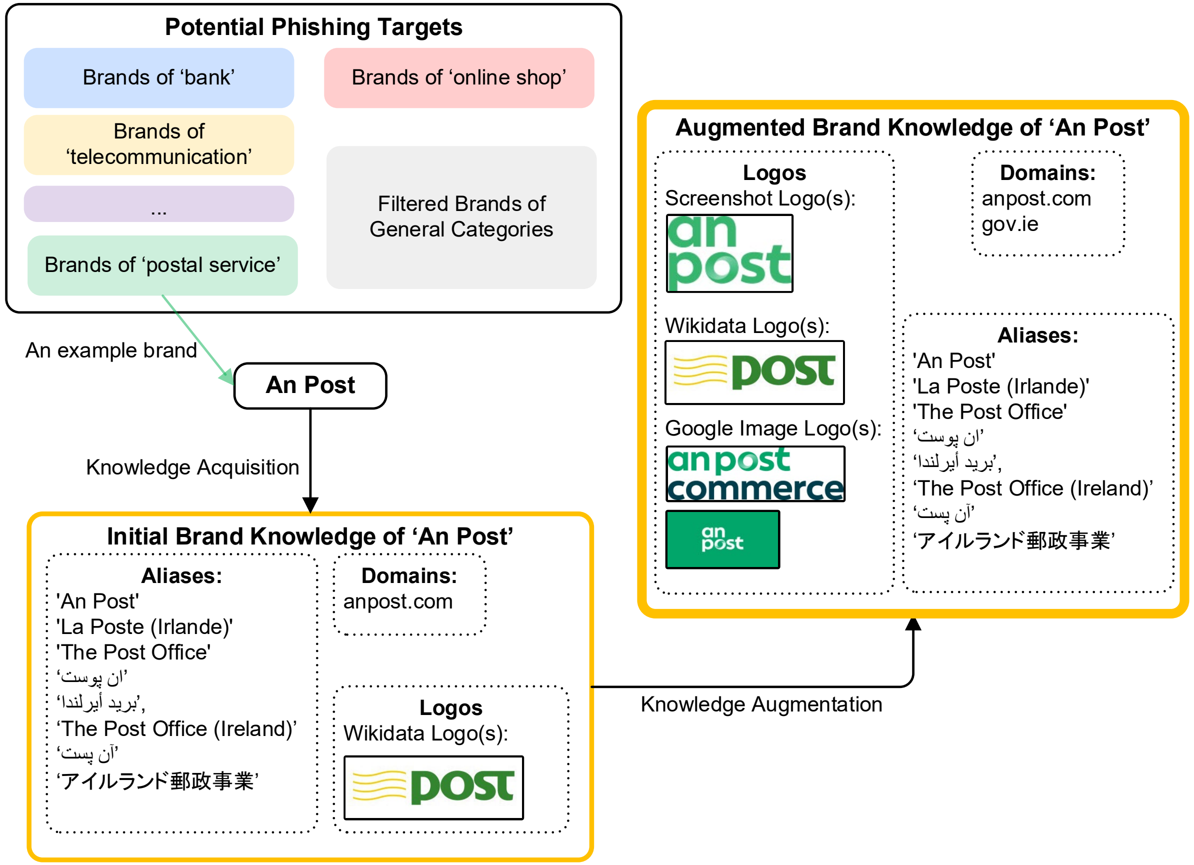
Wikidata

Top-ranking Domain List

Google Image Search



Official Webpage of Each Brand



[1] D Vrandečić, *et al*. Wikidata: A free collaborative knowledgebase. *Communications of the ACM* 2014.
[2] V Pochat, *et al*. Tranco: A research-oriented top sites ranking hardened against manipulation. *NDSS* 2019.

Solution 1: KnowPhish

KnowPhish can be equipped with any RBPDs to enhance their phishing detection performance

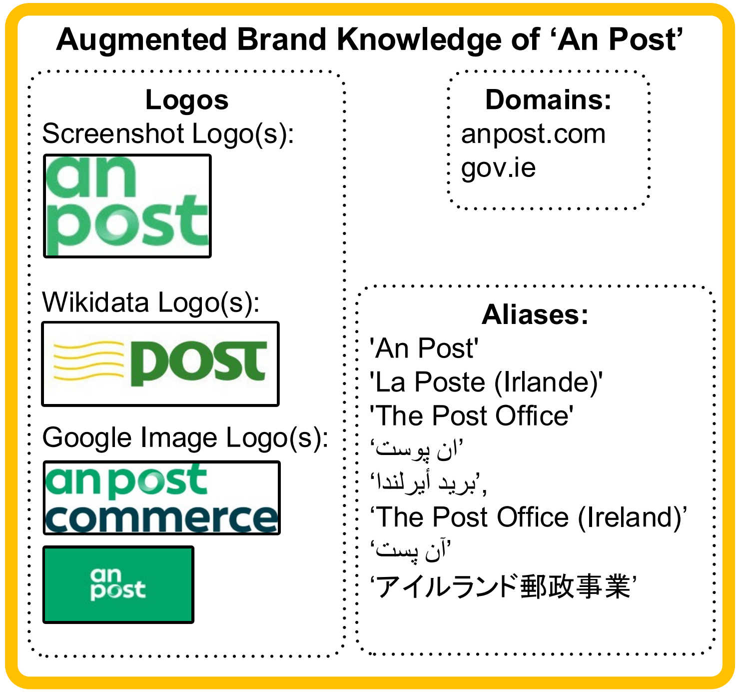


Image-based RBPDs

- Phishpedia
- PhishIntention

Multimodal RBPDs

- Our proposed KPD (discuss soon)

[1] Y Lin, *et al*. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. *USENIX Security* 2021.
[2] R Liu, *et al*. Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach. *USENIX Security* 2022.

Solution 2: KnowPhish Detector (KPD)

(Image + Text)

KPD: A multimodal reference-based phishing detector

- Leveraging Large Language Models (LLMs) to analyze text information in HTML (e.g., extracting textual brand intention), breaking the limit of existing image-based RBPDs that only analyze logos

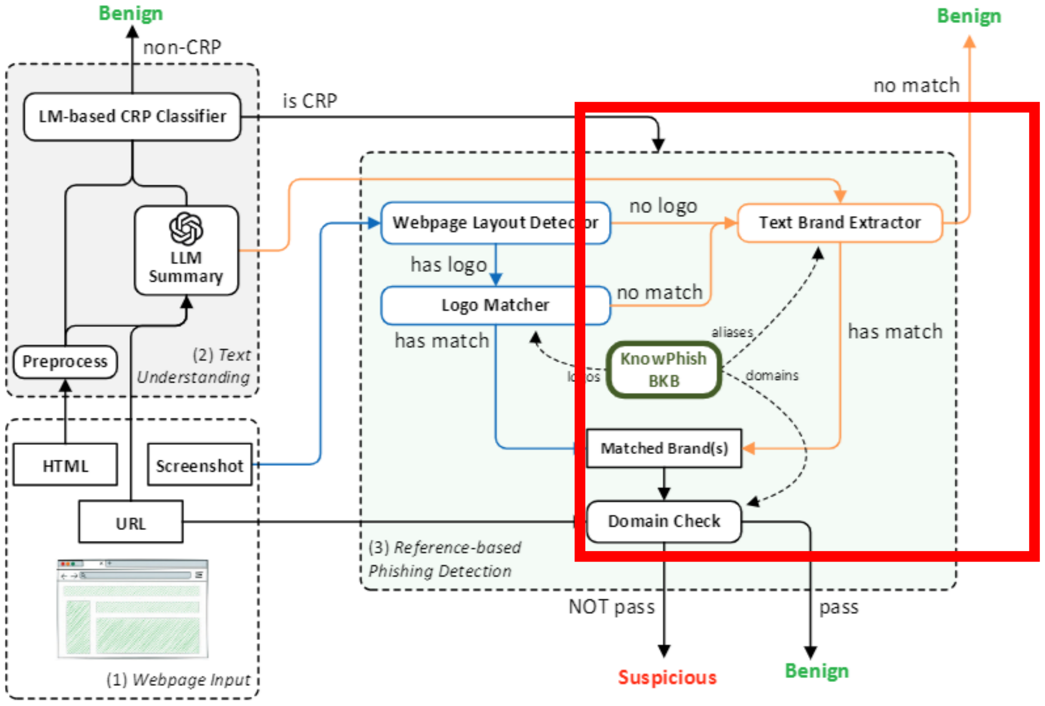
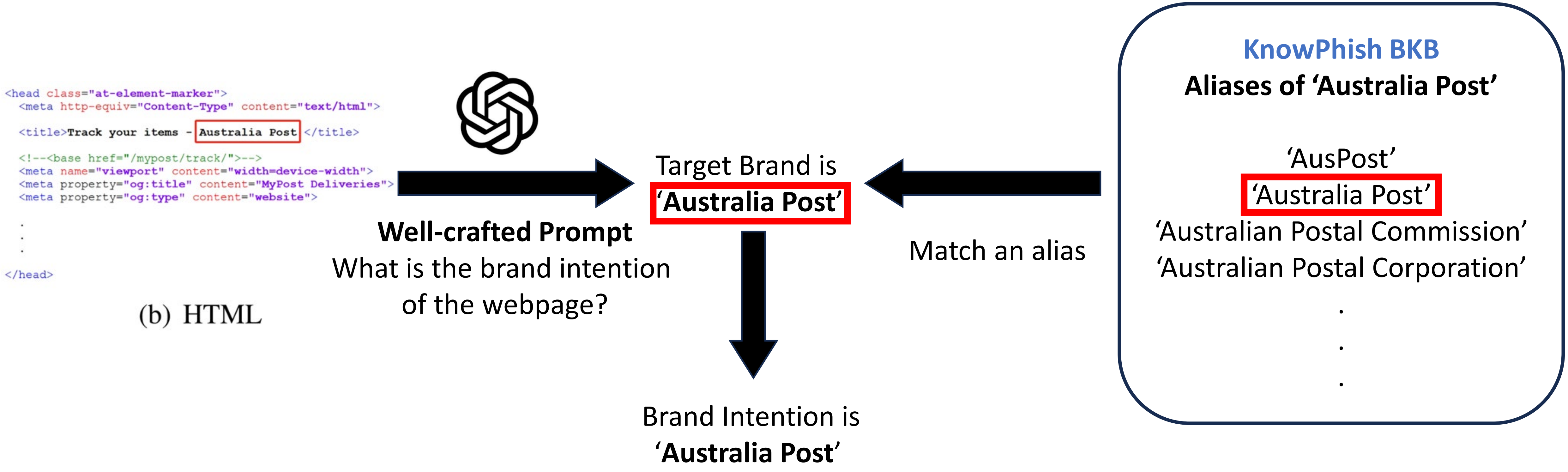


Figure 6: An overview of our phishing detector KPD.

KPD: Text Brand Extractor

Text Brand Extractor identifies textual brand intention through

1. LLM predictions and
2. Brand aliases in KnowPhish



KPD: Text-based CRP Classifier

CRP = Credential Requiring Page

Our text-based CRP Classifier can detect both explicit and implicit CRPs

- Explicit CRP has credential submission field
- Implicit CRP only contains buttons that redirect to explicit CRP pages, and cannot be detected by existing solution because they solely look at credential submission field
- Our text-based CRP classifier can analyze HTML texts and LLM summaries to recognize potential CRP signals encoded in HTML elements, regardless whether they have credential submission field

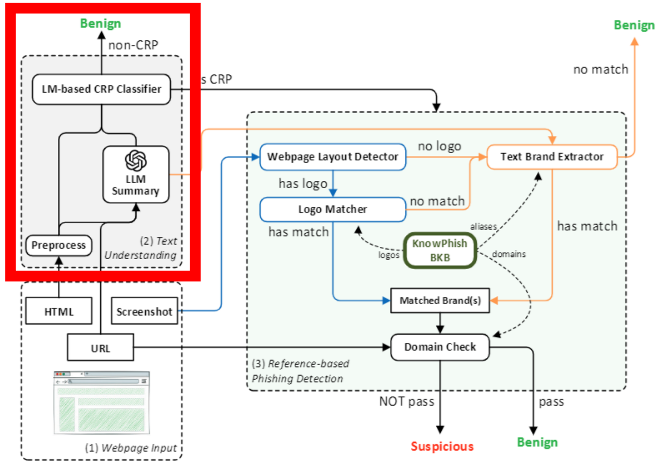
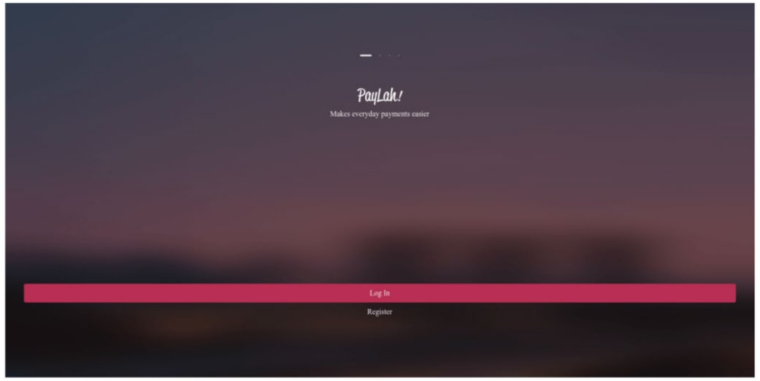


Figure 6: An overview of our phishing detector KPD.

(a) Explicit CRP



(b) Implicit CRP

Results: Closed-World Study

TR-OP Dataset
#Samples: 10k (benign 5k + phishing 5k)

KPD+KnowPhish is effective and efficient

- **KPD+KnowPhish** yields the highest accuracy, F1 score, and recall
- **KnowPhish** enhances different RBPDs to detect more phishing webpages (higher recall)
- **KnowPhish** significantly outperforms DynaPhish (*USENIX Sec* '23) in terms of inference time

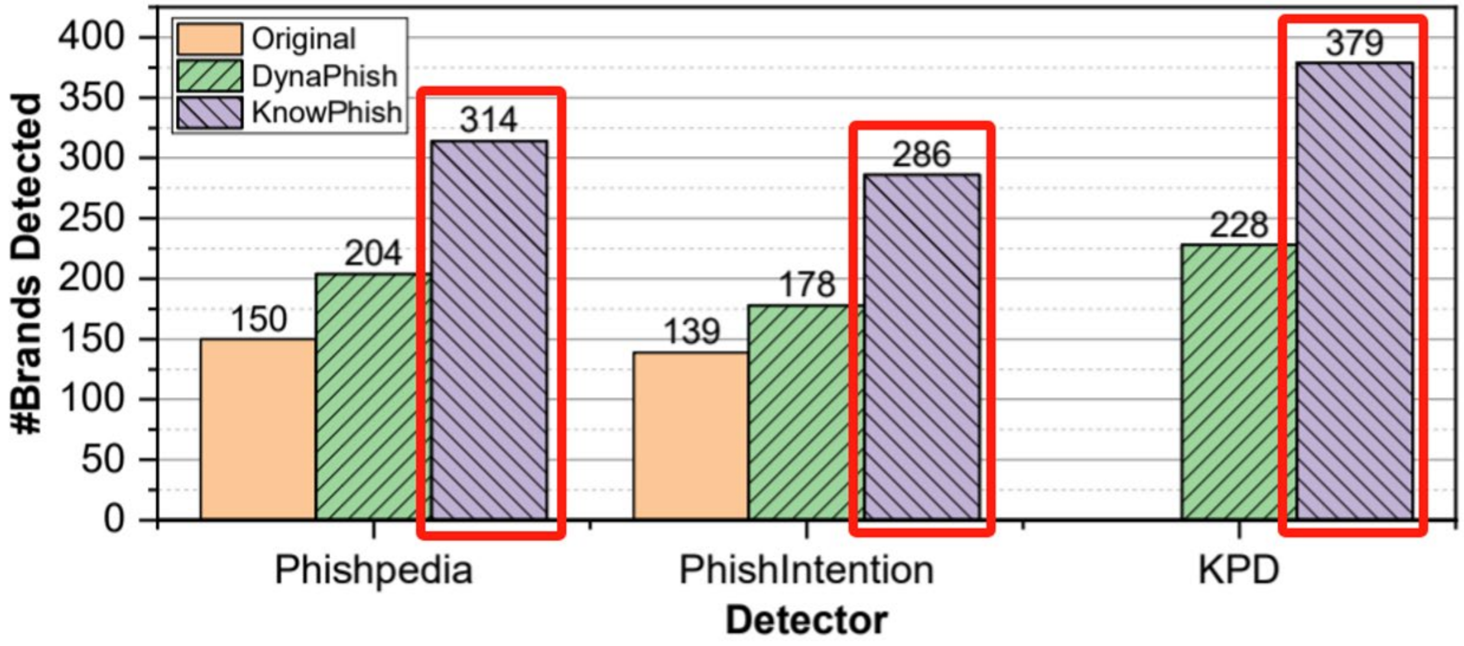
Detector	BKB	ACC↑	F1↑	Precision↑	Recall↑	Time↓
Phishpedia	Original	69.91	57.17	99.16	40.16	0.25s
	DynaPhish	66.40	52.52	89.50	37.16	10.92s
	KnowPhish	85.79	83.67	98.27	72.80	0.22s
PhishIntention	Original	66.62	49.96	**99.76**	33.32	0.28s
	DynaPhish	62.51	41.16	95.62	26.22	10.67s
	KnowPhish	77.84	71.60	99.67	55.84	0.26s
KPD	DynaPhish	76.10	69.71	95.16	55.00	12.18s
	KnowPhish	**92.49**	**92.05**	97.84	**86.90**	2.02s

[1] R Liu, *et al*. Knowledge Expansion and Counterfactual Interaction for Reference-Based Phishing Detection. *USENIX Security* 2023.

Results: Closed-World Study

KPD+KnowPhish detects the most phishing targets (379/440)

- KnowPhish enhances RBPDs better than DynaPhish does

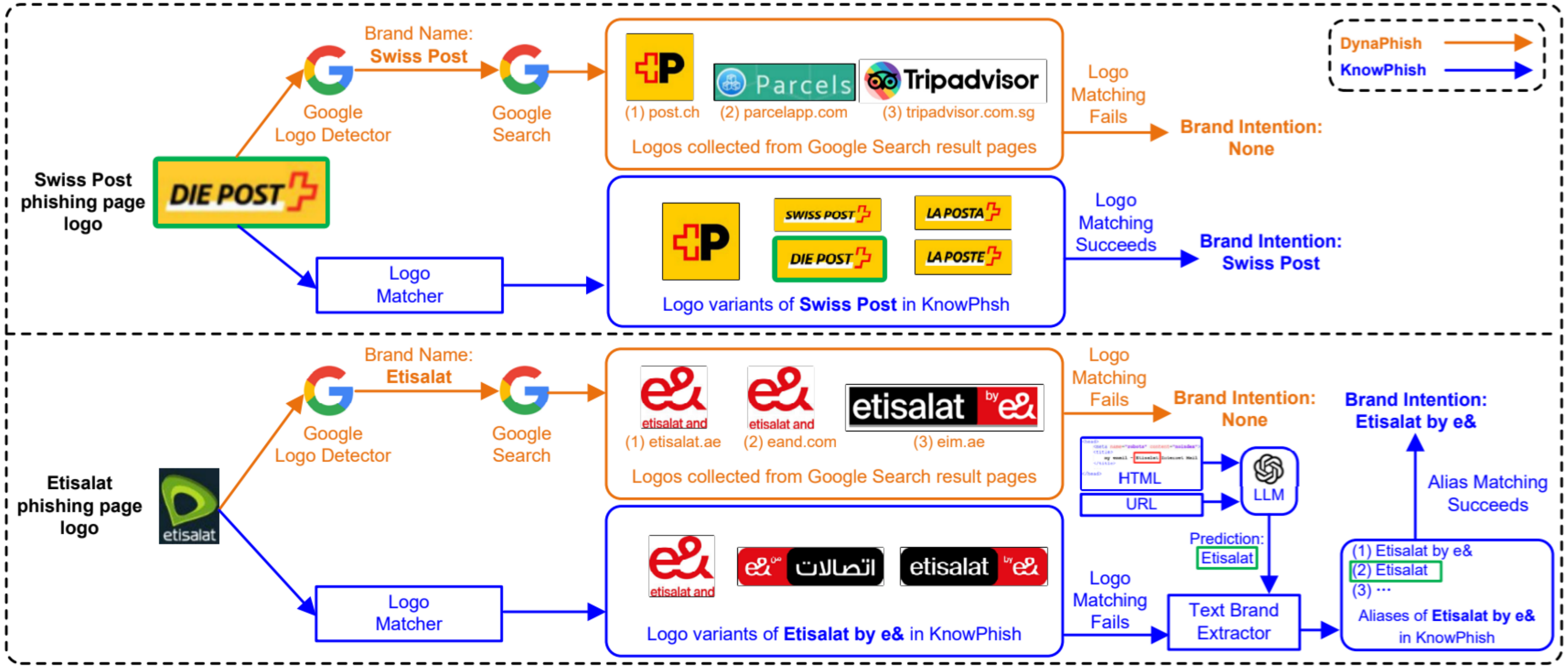


[1] R Liu, *et al*. Knowledge Expansion and Counterfactual Interaction for Reference-Based Phishing Detection. *USENIX Security* 2023.

Comparison with DynaPhish (USENIX Sec '23)

KnowPhish outperforms DynaPhish (static) in terms of

- the diversity of brand knowledge (e.g., logo variants) and
- the ability to detect textual brand intention through KPD when logo-analysis fails



[1] R Liu, *et al*. Knowledge Expansion and Counterfactual Interaction for Reference-Based Phishing Detection. *USENIX Security* 2023.

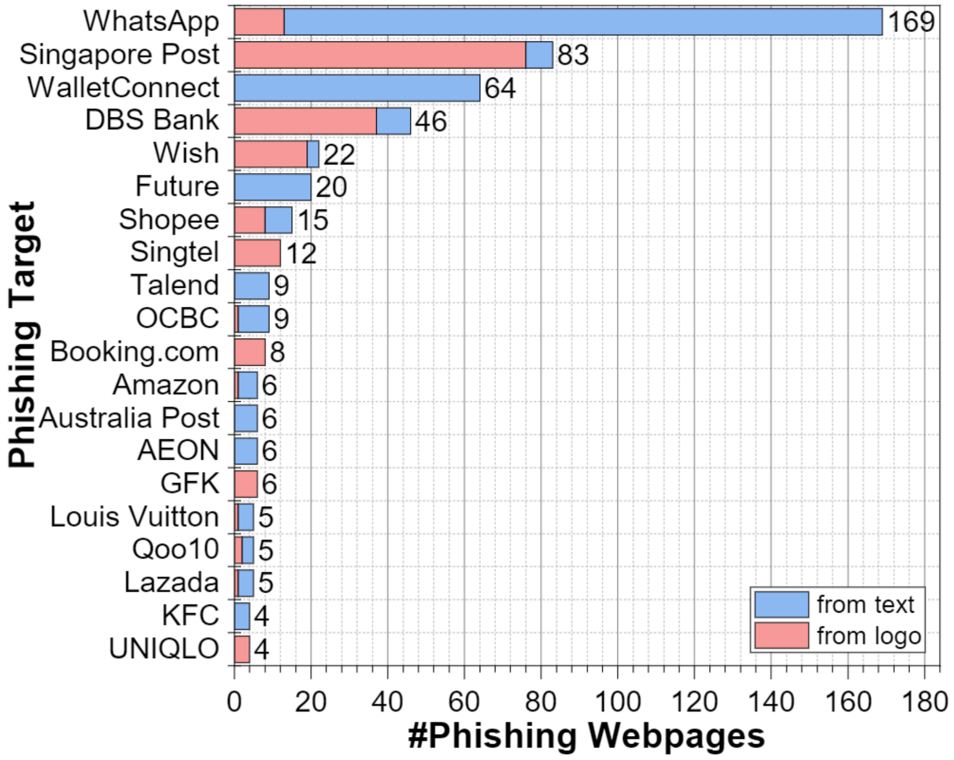
Results: Field Study

Detector	BKB	#P	#TP↑	Precision↑	Time↓
Phishpedia	Original	54	17	31.48	0.16s
	DynaPhish	583	481	82.67	5.98s
	KnowPhish	353	333	94.33	0.16s
PhishIntention	Original	25	8	32.00	0.18s
	DynaPhish	163	140	85.89	5.91s
	KnowPhish	138	133	96.37	0.19s
KPD	DynaPhish	628	581	92.52	7.83s
	KnowPhish	699	681	97.42	1.64s

KPD+KnowPhish identifies many local phishing targets in Singapore

- Detects phishing websites targeting local brands
  - Singapore Post
  - DBS Bank
  - Shopee
  - OCBC
  - Qoo10
  - Lazada
  - …

SG-SCAN Dataset
#Samples: 10k
Imbalanced and unlabelled



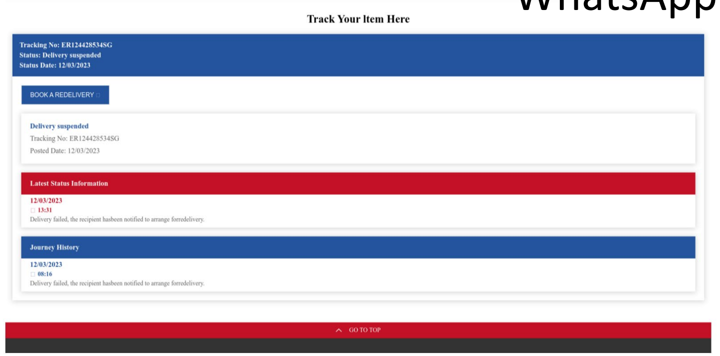
- Further validates our empirical insights
  - high-value industries usually indicate phishing targets

Results: Field Study

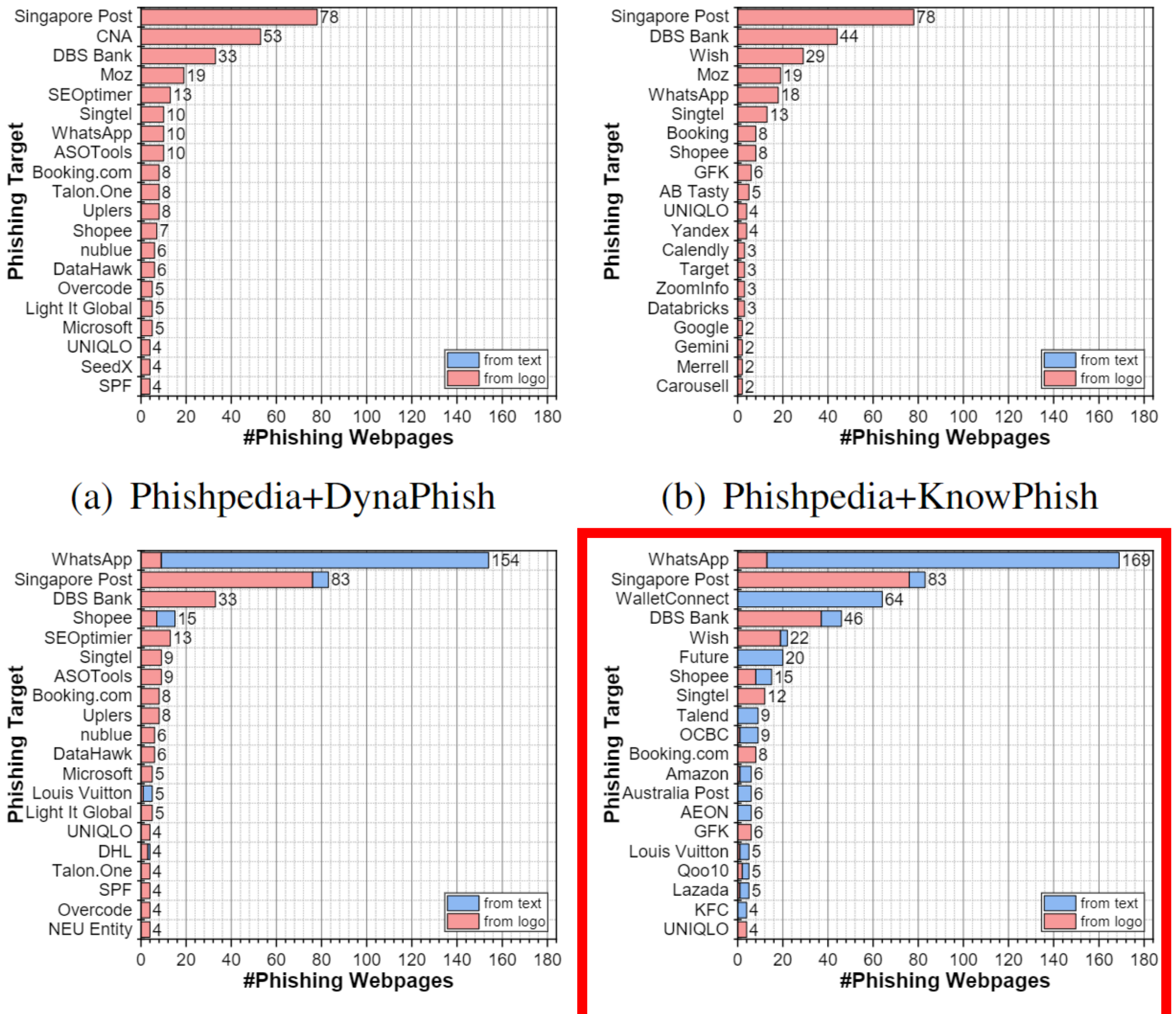
Logo-less phishing webpages are common in real-world

- Image-based RBPDs are not able to detect logo-less phishing in static environment

WhatsApp Phishing



Singapore Post Phishing



(a) Phishpedia+DynaPhish

(b) Phishpedia+KnowPhish

(c) KPD+DynaPhish

(d) KPD+KnowPhish

Takeaway

- **KnowPhish: Large-scale Multimodal Brand Knowledge Base**
  - The industries of phishing targets remain mostly consistent, despite the dynamic nature of phishing targets across different datasets
  - Based on Wikidata, we constructed a large-scale multimodal brand knowledge base covering more than 20k potential phishing targets
  - Can directly enhance any RBPDs without additional runtime maintenance cost

- **KPD: Multimodal Reference-based Phishing Detector:**
  - A multimodal RBPD operating in static environment that can detect phishing webpages with or without logos

Thanks for your listening!

Presenter: Yuexin Li

yuexinli@nus.edu.sg