Microsoft Research

CCSLAB Computer & Communication Security Laboratory

Detecting Malicious Web Links and Identifying Their Attack Types

[1]Hyunsang Choi, [2]Bin B. Zhu, [1]Heejo Lee

[1] Korea University,

[2] Microsoft Research Asia

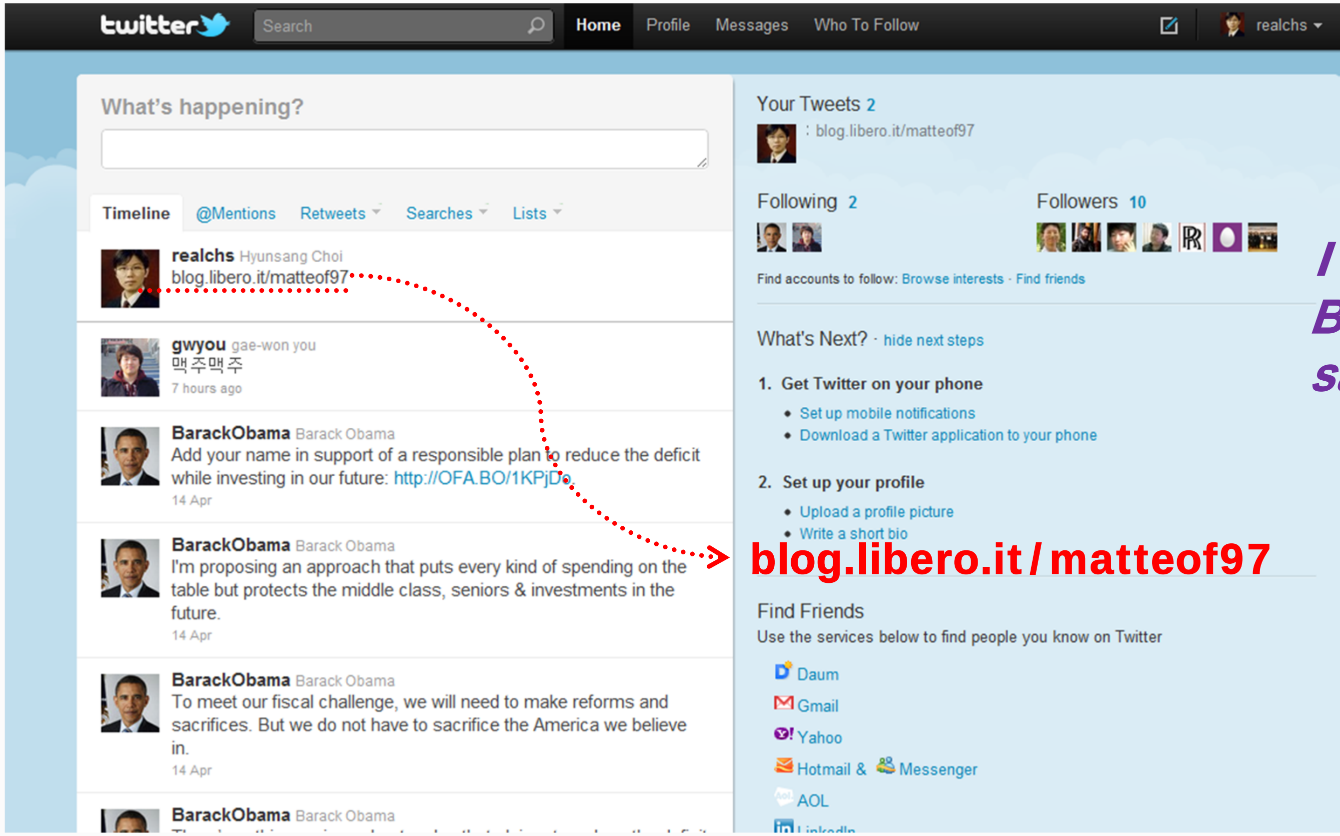
USENIX WebApps 2011

2011-06-21

Outline

- Introduction
- Existing solutions
- Highlights of our approach
- Discriminative features
- Experimental results
- Evadability
- Conclusion

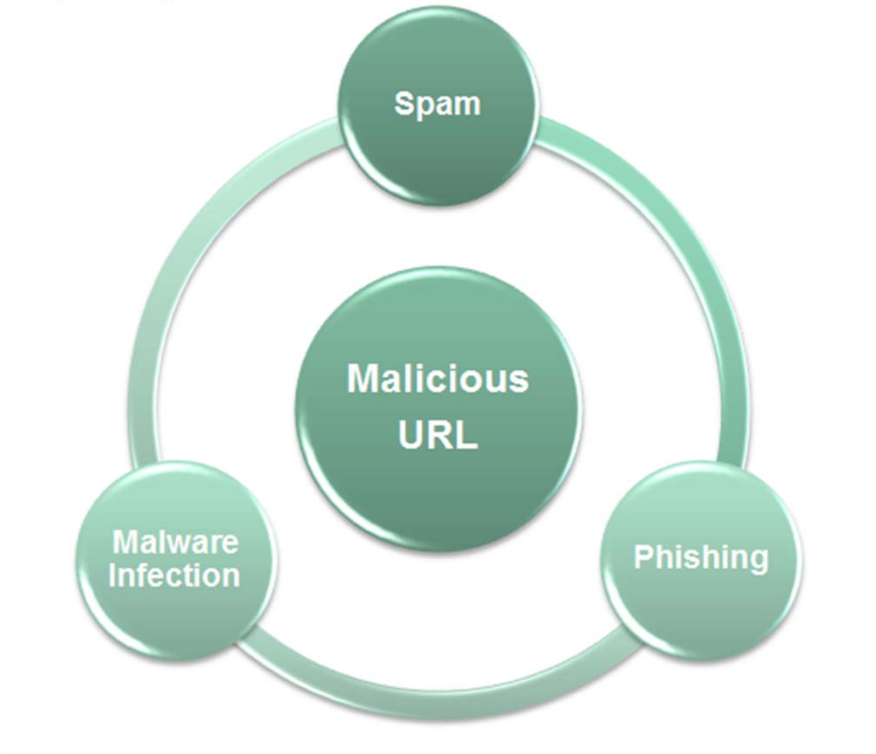
Access or not access, that is a problem



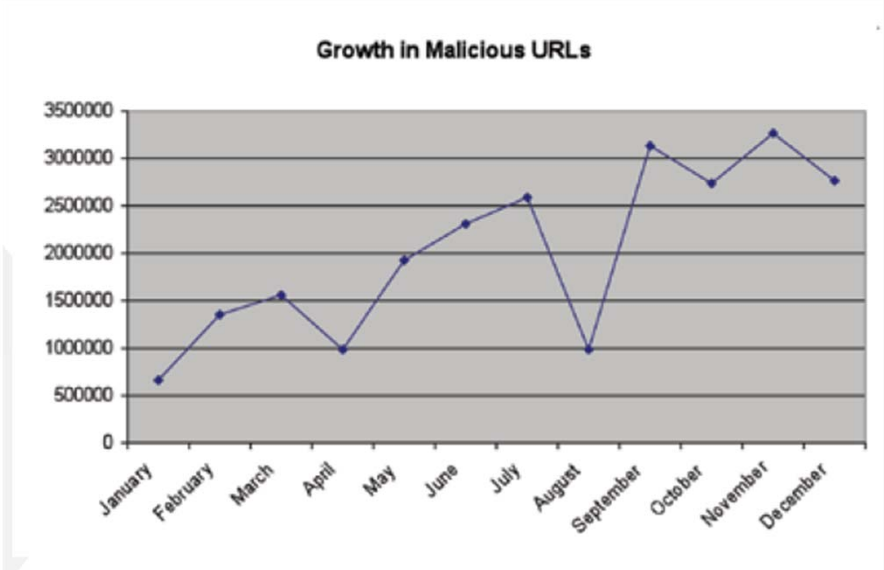
I want to read, …
But is this Webpage safe to read?

Malicious Webpages

Webpages have been widely used for malicious purposes



3 Major types of malicious URLs



Growth of malicious URLs in 2010, Trend Micro Annual Threat Report, 2010

Existing Solutions: Blacklisting

The Achilles' heel of blacklisting

Popular URL analysis tools

- Not work for new/unknown URLs
- Evadable easily

Existing Solutions: Anomaly-Based Detection

- **Other existing solutions:**
  - **VM execution**
  - **Rule-based detectors**
  - **Machine learning based detectors**
- **Detecting typically a single type of an attack**
- **Critical issues in machine learning based approach**
  - **What are highly effective discriminative features?**
  - **Are the discriminative features en masse hard to evade?**

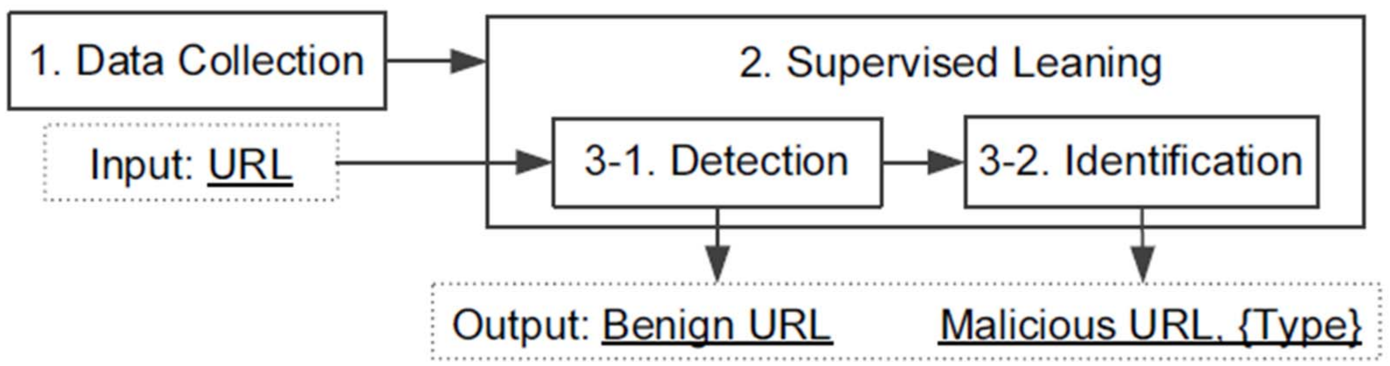
Highlights of Our Research Project

- **Research Goals:**
  - **Detect all major malicious types of URLs**
  - **Identify attack types of a malicious URL**
    - **Much harder than detection due to ambiguity**
  - **Develop effective & hard to evading discriminative features**
- **Methodology: machine learning based approach**
  - **SVM for detecting malicious URLs**
  - **RAkEL & ML-kNN for identifying attack types of a malicious URL**

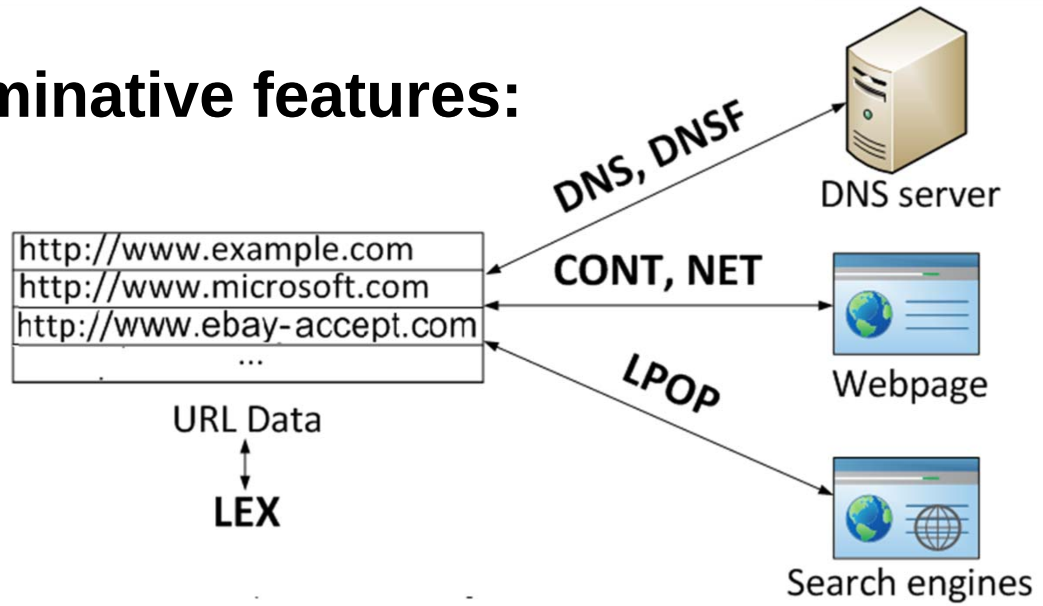
Key Properties of Our Detector and Major Contributions

- **First study to classify multiple types of malicious URLs**
- **A rich set of highly effective discriminative features**
  - **Many features are novel and unique**
  - ***Same* discriminative features for both detection and classification tasks**
  - **Robust against known evadsion techniques**
- **A systematical study of the effectiveness of each feature group**

Overview of Our System



- 6 groups of 53 discriminative features:
  - Lexical
  - Link popularity
  - Webpage content
  - DNS
  - DNS fluxiness
  - Network



- 31 out of the 53 features are novel or modified from prior arts

1. Lexical Features

- **Lexical features**
  - **Most are targeted to detect phishing attack (phishing attack has discriminate lexical property to deceive users)**
  - **Discriminative features effective on some attack types but not on other are desirable to distinguish different types**

No.	Feature	Type	Targeted types
1	Domain token count	Integer	Phishing
2	Path token count	Integer	Phishing
3	Average domain token length	Real	Phishing
4	Average path token length	Real	Phishing
5	Longest domain token length	Integer	Phishing
6	Longest path token length	Integer	Phishing
7∼9	Spam, phishing and malware SLD hit ratio	Real	All types
10	Brand name presence	Binary	Phishing

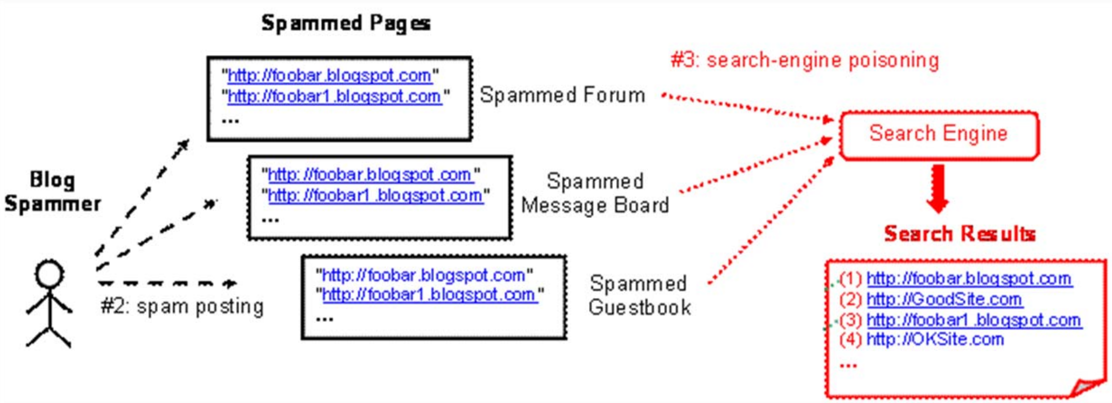
2. Link Popularity Features

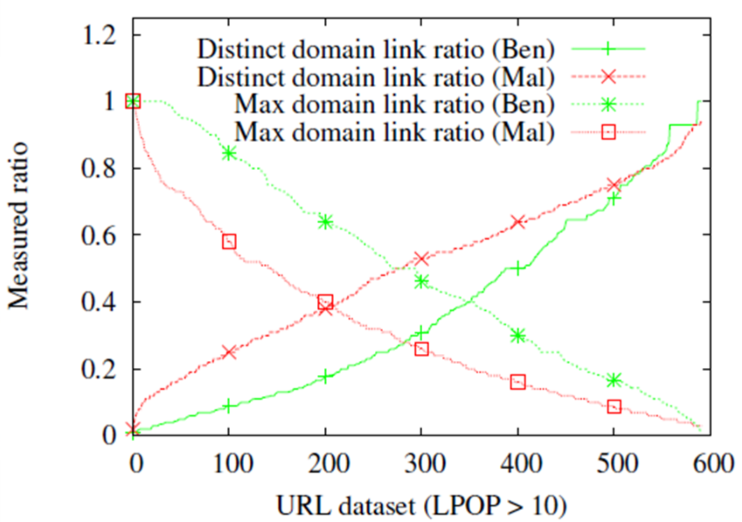
- **Link popularity features**
  - **Intuition: Malicious URLs are hardly indexed by normal users**
  - **Methodology: Get inlink (incoming link) count from search engines**
  - **Search engines: AlltheWeb, Astalavista, Google, Yahoo, Ask**

No.	Feature	Type	Targeted types
1∼5	5 LPOPs of the URL	Integer	All types
6∼10	5 LPOPs of the domain	Integer	All types
11	Distinct domain link ratio	Real	All types (SEO)
12	Max domain link ratio	Real	All types (SEO)
13∼15	Spam, phishing and malware link ratio	Real	All types (SEO)

2. Link Popularity Features (cont.)

- **Blackhat SEO & link farming**
  - **Blackhat Search Engine Optimization (SEO) is used to get unethically higher search rankings**
    - **Link farming: link manipulation using a group of webpages to link together**
  - **5 features for detecting link manipulated URLs by Blackhat SEO**
    - **Distinct domain link ratio, max domain link ratio**
    - **Spam, phishing, and malware link ratio**





3. Webpage Content Features

- **Webpage content features**
  - **Features used by** Hou et al., “Malicious web content detection by machine learning”, Expert Systems with Applications, 2010

No.	Feature	Type	Targeted types
1	HTML tag count	Integer	Malware, phishing
2	Iframe count	Integer	Malware
3	Zero size iframe count	Integer	Malware
4	Line count	Integer	All types
5	Hyperlink count	Integer	Malware, spam
6∼12	Count of each suspicious JavaScript function	Integer	Malware
13	Total count of suspicious JavaScript functions	Integer	Malware

4. DNS Features

- **DNS features**
  - **Features from the DNS server**
  - **Methodology: Use DNS answer data from DNS server**

No.	Feature	Type	Targeted types
1	Resolved IP count	Integer	All types
2	Name server count	Integer	All types
3	Name server IP count	Integer	All types
4	Malicious ASN ratio of resolved IPs	Real	All types
5	Malicious ASN ratio of name server IPs	Real	All types

5. DNS Fluxiness Features

- **DNS fluxiness features**
  - **Features to detect fast-fluxing URLs**
  - **Fast-flux: DNS technique to hide malicious websites behind an ever-changing network of compromised hosts acting as proxies**
  - **Methodology: Send queries to DNS server (first and consecutive lookups)**
  - **Features by** Holz et al., “Detection and mitigation of fast-flux service networks”, NDSS 2008

No.	Feature	Type	Targeted types
1	φ of N_{IP}	Real	All types
2	φ of N_{AS}	Real	All types
3	φ of N_{NS}	Real	All types
4	φ of N_{NSIP}	Real	All types
5	φ of N_{NSAS}	Real	All types

$$\varphi = N_{IP}/N_{single}.$$

6. Network Features

- **Network features**
  - **Detect redirected URLs (URL shortening, iframe redirections)**
  - **Methodology: Use web crawler**

Table 6: Network feature (NET)

No.	Feature	Type	Targeted types
1	Redirection count	Integer	All types
2	Downloaded bytes from content-length	Real	All types
3	Actual downloaded bytes	Real	All types
4	Domain lookup time	Real	All types
5	Average download speed	Real	All types

Experimental Datasets

Single Label URL Type	Single Label Dataset	Amount
Benign	Randomly selected 20K URLs from DMOZ open directory	20K
	Randomly selected URLs from Yahoo directory	20K
Spam	jwSpamSpy list	11K
Phishing	PhishTank list	4K
Malware	DNS-BH list	17K

Evaluation Result – Detection Accuracy

- **Detection accuracy**
  - **98.2% accuracy, 98.9% true positive rate, 1.1% false positive rate, and 0.8% false negative rate**

Dataset	Metric	Feature group					
		LEX	LPOP	CONT	DNS	DNSF	NET
Spam	ACC	73.0	97.2	82.8	77.4	**87.7**	72.1
	TP	72.4	97.4	74.2	75.9	**86.3**	77.4
Phishing	ACC	**91.6**	98.1	77.3	76.3	71.8	77.2
	TP	**86.1**	95.1	82.8	76.9	70.1	78.2
Malware	ACC	70.3	96.2	**86.2**	78.6	68.1	73.3
	TP	74.5	93.2	**88.4**	75.1	74.2	78.2

Evaluation Result – Link Popularity

- **Link popularity**
  - **Google reports a partial list of inlink information**
  - **Without link popularity feature: 91.2% accuracy, 4.0% false positive rate, and 4.8% false negative rate**
  - **90.03% accuracy in detecting link-manipulated malicious URLs**

Metric	AllTheWeb	Altavista	Ask	Google	Yahoo!
ACC	95.1	95.6	84.0	85.7	95.9
TP	95.3	96.3	85.7	86.7	95.7
FP	2.7	2.7	8.4	**12.3**	2.1
FN	2.2	1.6	7.6	**2.1**	2.1

Datasets for Multi-Labels

- **Datasets – Multi labels**
  - **Use two website to crawl the ‘exact’ malicious type of URLs (McAfee SiteAdvisor and Web Of Trust)**
  - **About half of URLs in the data set have multiple labels**

Label	Attribute	L_{SAd}	L_{WOT}	L_{Both}
λ_1	spam	6020	6432	5835
λ_2	phishing	1119	1067	899
λ_3	malware	9478	8664	8105
$\lambda_{1,2}$	spam, phishing	4076	4261	3860
$\lambda_{1,3}$	spam, malware	2391	2541	2183
$\lambda_{2,3}$	phishing, malware	4729	4801	4225
$\lambda_{1,2,3}$	spam, phishing, malware	2219	2170	2080

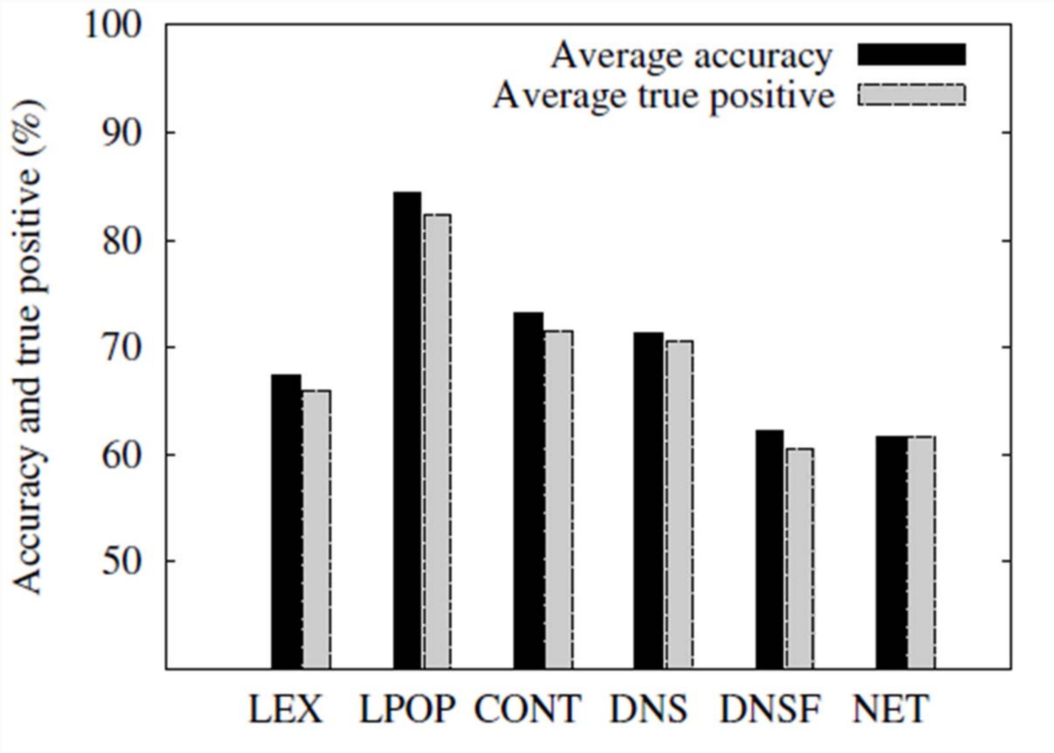
Evaluation Result – Multi-label Classification (1)

- **Metrics**
  - **Micro-averaged and macro-averaged metrics: Micro-average gives equal weight to every data sets, while the macro-average gives equal weight to every category**
  - **Ranking-based metrics: Average precision and ranking loss**
- **Multi-label classification result**
  - **93% averaged accuracy and 98% ranking-based precision**

	Label	Averaged			Ranking-based	
		ACC	micro TP	macro TP	R_{loss}	P_{avg}
RA*k*EL	L_{SAd}	90.70	87.55	88.51	3.45	96.87
	L_{WOT}	90.38	88.45	89.59	4.68	93.52
	L_{Both}	**92.79**	**91.23**	**89.04**	**2.88**	**97.66**
ML-*k*NN	L_{SAd}	91.34	86.45	87.93	3.42	95.85
	L_{WOT}	91.04	88.96	89.77	3.77	96.12
	L_{Both}	**93.11**	**91.02**	**89.33**	**2.61**	**97.85**

Evaluation Result – Multi-label Classification (2)

- **Performance for each feature group**
  - **No single feature group can effectively classify malicious URL types**



Evadability Analysis

- **Robust to known evasion techniques**
  - **Redirection: Network features**
  - **Link manipulation: Link popularity features**
  - **Fast-flux: DNS fluxiness features**
- **URL obfuscation**
  - **IDN (Internationalized Domain Names) spoofing (e.g., www.pаypal.com = www.paypal.com)**
- **JavaScript obfuscation**
  - **Deobfuscator**
- **Social network sites**

Conclusion

- **Goal**
  - **Proposed a machine learning approach to detect malicious URLs and to identify attack types.**
- **Method**
  - **Collect various types of discriminative features, detecting malicious URLs using SVM and identifying malicious URL types using RAkEL and ML-kNN**
- **Result**
  - **Achieved an accuracy of over 98% in detecting malicious URLs and an accuracy of over 93% in identifying attack types.**
- **Contribution**
  - **Proposed several novel and highly discriminative features which provide a superior performance and a much larger coverage**
  - **First study to classify multiple types of malicious URLs, known as a multi-label classification**

Q&A

Thank you!