On the Effectiveness of Risk Prediction Based on Users Browsing Behavior

*Davide Canali**, *Leyla Bilge*, *Davide Balzarotti*

EURECOM Software and System Security Group, France

Symantec Research Labs, France

* now at Lastline, Inc.

Motivations

EURECOM
Sophia Antipolis

Understanding the reasons **why certain users are safer than others** on the web

Is there any **correlation** between **browsing behaviors** and **user risk**?

- Previous studies used survey-like approaches, and studied infections on end-user laptops (Lévesque et al, 2013)
- Simple indicators given by the study of the Australian threat landscape by TrendMicro and Deakin University

Can we build **risk profiles** for web users?

- User profiling has been mostly studied in the area of recommender systems
- Think of Cyber-insurance schemes...

Cyber Insurance Scenario

EURECOM
Sophia Antipolis

The concept of “cyber insurance” has been around for several years, however

- Very **little empirical data on incidents**
- Companies **do not want to reveal** their **security breaches**
- No standardized cyber insurance prices and policies

Little has been done to know which factors affect risk

- Unlike traditional insurance (car, house, etc.)

Dataset

Telemetry data from Symantec

EURECOM
Sophia Antipolis

3 months of browsing data (August 1 - October 31, 2013)

- HTTP requests only
  - Performed voluntarily, within a browser (no automatic requests)
- Anonymized user information

202M URL hits (38M distinct)

from **160K users**, who:

- opted-in to share their browsing histories
- visited at least 100 pages during the observation period

User Risk Categories

EURECOM
Sophia Antipolis

Based on URL labeling from:

- Norton Safe Web
- Google SafeBrowsing
- Public domain blacklists

Following a classical insurance approach, users are **categorized** based on their **past experiences**:

Safe *Uncertain* *At Risk*

User Risk Categories

EURECOM
Sophia Antipolis

Based on URL labeling from:

- Norton Safe Web
- Google SafeBrowsing
- Public domain blacklists

Following a classical insurance approach, users are **categorized** based on their **past experiences**:

Safe
50%

Uncertain

At Risk

User Risk Categories

EURECOM
Sophia Antipolis

Based on URL labeling from:

- Norton Safe Web
- Google SafeBrowsing
- Public domain blacklists

Following a classical insurance approach, users are **categorized** based on their **past experiences**:

Safe

Uncertain

At Risk
19%

Analysis

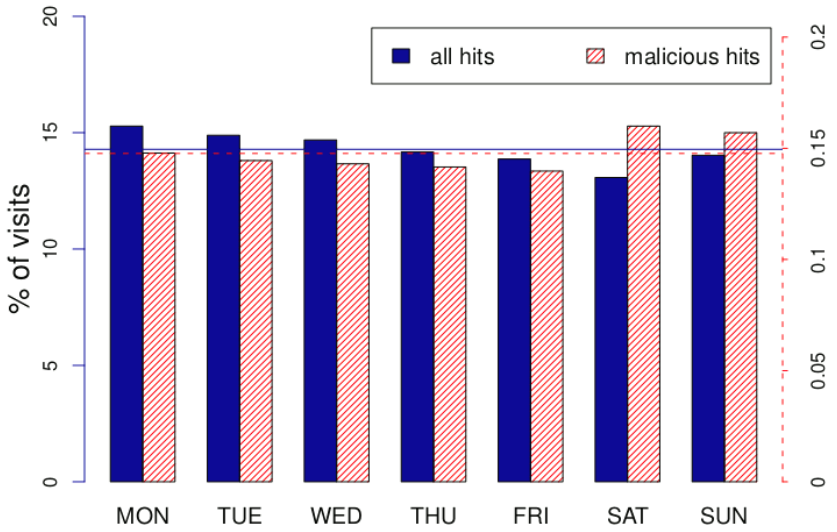
A quick look at average values...

EURECOM
Sophia Antipolis

- Number of visited URLs
  - *safe* users: 743 (daily avg: 17)
  - *at risk* users: 2411 (daily avg: 37)
- Distinct visited URLs
  - *safe* users: 231 (daily avg: 6)
  - *at risk* users: 874 (daily avg: 14)
- Percentage of visited malicious URLs
  - *uncertain* users: 0.14%
  - *at risk* users: 0.71%

Analysis

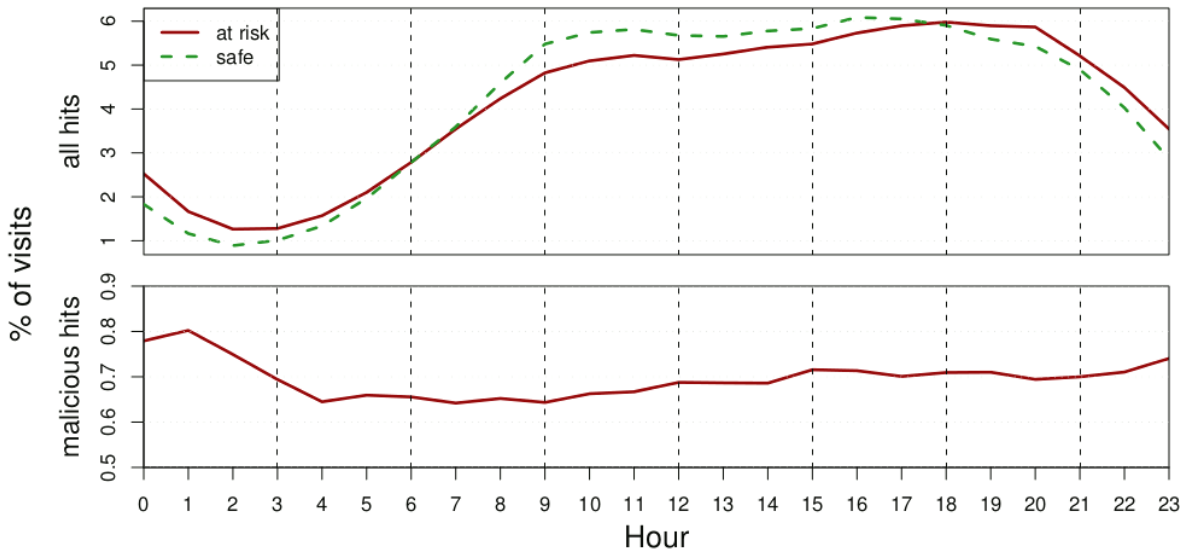
Daily trends



- Less web hits during weekends
- **Increase** in the percentage of **malicious URL visits during weekends** (+10%)

Analysis

Hourly trends



- People surf less at night
  - But percentages of **malicious hits at night are higher** (+6.5%)
- ***At risk* users** are less active in the morning and **more active at night**, compared to safe ones

Geographical Trends

EURECOM
Sophia Antipolis

Country	Users	% users *at risk*	Average hits on		Visited Pages			# languages
			malicious URLs	blacklisted domains	total	distinct	domains	
US	67967	20.8	2.2 (0.22%)	2.0 (0.15%)	1250	422	194	3.6
UK	26204	17.8	1.5 (0.16%)	2.0 (0.16%)	1097	379	183	4.2
JP	16556	**10.0**	1.1 (**0.05**%)	3.1 (0.14%)	**1989**	**641**	205	3.8
CA	6798	20.9	2.0 (0.22%)	2.4 (0.17%)	1214	387	186	3.8
AU	6107	16.4	1.5 (0.17%)	1.5 (0.15%)	1007	343	173	3.7
DE	5606	22.3	2.0 (0.20%)	2.6 (0.23%)	1042	366	192	4.9
FR	4566	**29.1**	2.8 (0.27%)	3.3 (0.27%)	1127	390	209	4.5
NL	3415	15.9	1.1 (0.12%)	2.3 (0.21%)	1009	361	195	5.2
ES	1842	**28.3**	2.4 (0.23%)	3.9 (**0.33**%)	1121	391	200	**5.7**
SE	1755	15.3	1.9 (0.17%)	1.9 (0.14%)	1049	327	167	6.4
IT	1665	**27.4**	1.8 (0.18%)	**7.0** (**0.69**%)	1097	350	186	5.4
BE	1454	21.3	2.2 (0.21%)	2.5 (0.20%)	1126	396	208	5.5
NO	1208	**11.8**	1.1 (**0.10**%)	2.5 (0.11%)	1219	341	166	**6.1**

Geographical Trends

EURECOM
Sophia Antipolis

Country	Users	% users *at risk*	Average hits on		Visited Pages			# languages
			malicious URLs	blacklisted domains	total	distinct	domains	
US	67967	20.8	2.2 (0.22%)	2.0 (0.15%)	1250	422	194	3.6
UK	26204	17.8	1.5 (0.16%)	2.0 (0.16%)	1097	379	183	4.2
JP	16556	**10.0**	1.1 (**0.05**%)	3.1 (0.14%)	**1989**	**641**	205	3.8
CA	6798	20.9	2.0 (0.22%)	2.4 (0.17%)	1214	387	186	3.8
AU	6107	16.4	1.5 (0.17%)	1.5 (0.15%)	1007	343	173	3.7
DE	5606	22.3	2.0 (0.20%)	2.6 (0.23%)	1042	366	192	4.9
FR	4566	**29.1**	2.8 (0.27%)	3.3 (0.27%)	1127	390	209	4.5
NL	3415	15.9	1.1 (0.12%)	2.3 (0.21%)	1009	361	195	5.2
ES	1842	**28.3**	2.4 (0.23%)	3.9 (**0.33**%)	1121	391	200	**5.7**
SE	1755	15.3	1.9 (0.17%)	1.9 (0.14%)	1049	327	167	6.4
IT	1665	**27.4**	1.8 (0.18%)	**7.0** (**0.69**%)	1097	350	186	5.4
BE	1454	21.3	2.2 (0.21%)	2.5 (0.20%)	1126	396	208	5.5
NO	1208	**11.8**	1.1 (**0.10**%)	2.5 (0.11%)	1219	341	166	**6.1**

Japan: lowest percentage of malicious hits and at risk users

Geographical Trends

EURECOM
Sophia Antipolis

Country	Users	% users *at risk*	Average hits on		Visited Pages			# languages
			malicious URLs	blacklisted domains	total	distinct	domains	
US	67967	20.8	2.2 (0.22%)	2.0 (0.15%)	1250	422	194	3.6
UK	26204	17.8	1.5 (0.16%)	2.0 (0.16%)	1097	379	183	4.2
JP	16556	**10.0**	1.1 (**0.05**%)	3.1 (0.14%)	**1989**	**641**	205	3.8
CA	6798	20.9	2.0 (0.22%)	2.4 (0.17%)	1214	387	186	3.8
AU	6107	16.4	1.5 (0.17%)	1.5 (0.15%)	1007	343	173	3.7
DE	5606	22.3	2.0 (0.20%)	2.6 (0.23%)	1042	366	192	4.9
FR	4566	**29.1**	2.8 (0.27%)	3.3 (0.27%)	1127	390	209	4.5
NL	3415	15.9	1.1 (0.12%)	2.3 (0.21%)	1009	361	195	5.2
ES	1842	**28.3**	2.4 (0.23%)	3.9 (**0.33**%)	1121	391	200	**5.7**
SE	1755	15.3	1.9 (0.17%)	1.9 (0.14%)	1049	327	167	6.4
IT	1665	**27.4**	1.8 (0.18%)	**7.0** (**0.69**%)	1097	350	186	5.4
BE	1454	21.3	2.2 (0.21%)	2.5 (0.20%)	1126	396	208	5.5
NO	1208	**11.8**	1.1 (**0.10**%)	2.5 (0.11%)	1219	341	166	**6.1**

France, Spain, Italy: percentages of at risk users almost 3x higher than Japan

Feature Extraction

for user profiling

EURECOM
Sophia Antipolis

More than 70 features extracted from the data

- How much a user surfs the web
- In which period of the day a user is more active
- How diversified is the set of visited websites
- Computer type
- Which website categories the user is interested in
- Popularity of visited websites
- How stable is the set of visited pages

Feature Extraction

for user profiling

EURECOM
Sophia Antipolis

How much does a user **surf** the web?

- Basic stats
  - » Total number of web requests
  - » Number of distinct URLs
  - » Number of requests per day
  - » Number of distinct URLs per day

In **which period of the day** is the user more active?

- Percentage of hits during night, day, and evening
  - » Night: 00 am – 06 am
  - » Day: 06am – 7pm
  - » Evening : 7pm – 00 am

Feature Extraction

for user profiling

EURECOM
Sophia Antipolis

How diversified are the visited **web sites**?

- Number of distinct domain names
- Number of distinct TLDs
- Number of languages of the visited web pages
  - Coverage: 77% overall

In **which** web **categories** is the user more interested?

- Websites categorized in 11 categories
  - Heuristics: Business websites, Adult, Communications and information search, General interest, Hacking, Entertainment and leisure, Multimedia and downloading, Uncategorized
  - Blacklists: One-click hosting, Porn sites, Bittorrent websites
  - Coverage: 76% overall, 96% of Alexa top 10,000

Feature Extraction

for user profiling

EURECOM
Sophia Antipolis

What are the **computer characteristics**?

- Office computers or home computers
  - Profiles that browse only during week days are likely to be office computers
- Is the computer mobile?
  - Number of different IP addresses the user is browsing the Internet from
  - Number of different ISPs
  - Number of different countries

How **popular** are the visited **web sites**?

- Percentage of domains whose TLD is .com, .org, .net
- Percentage of domains in the Alexa Top 100
- Percentage of domains in the Alexa Top 1M

Feature Extraction

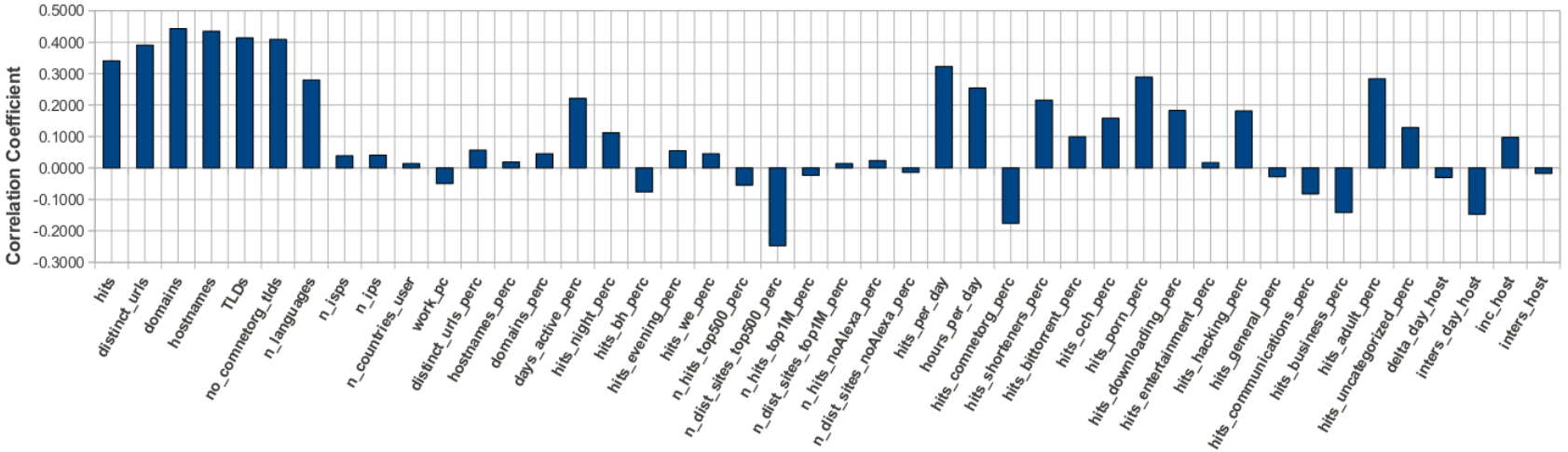
for user profiling

EURECOM
Sophia Antipolis

How **stable** is the set of visited **web pages**?

- To model the variability of the user's browsing activity
  - Are users who browse always the same web pages less at risk than others?
- Measures of:
  - the daily and overall increment in the number of websites visited by the user
  - the daily and overall percentage of websites visited, which had been visited by the user in the past

Feature Correlations



- **Correlation** with being at risk varies from **very weak to moderate**
- Some of the features showing the highest correlation:
  - Number of visited TLDs that are not .org, .net, .com
  - Number of URLs, domains, and hostnames visited by a user
  - Percentage of visited adult websites

Predictive Analysis

EURECOM
Sophia Antipolis

- Can we **predict** whether a user is ***at risk*** or not?

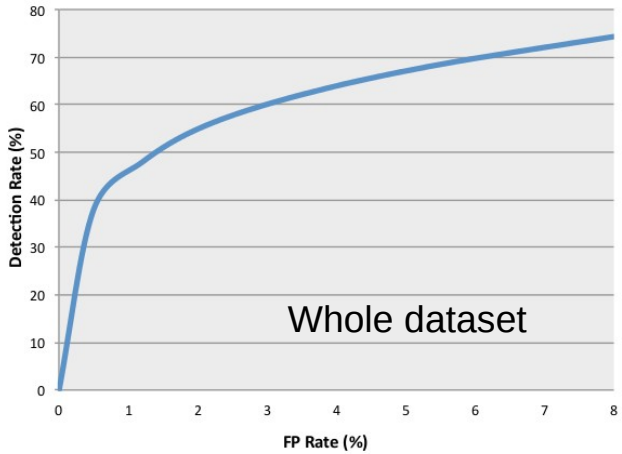
- Experimented with a range of prediction models (SVM, Bayesian classifiers, decision trees, logistic regression)
  - Chosen Logistic Regression
    - » Good for features with continuous or discrete values
    - » Does not explicitly require uncorrelated features
    - » Achieved the best accuracy and FP rates in our tests

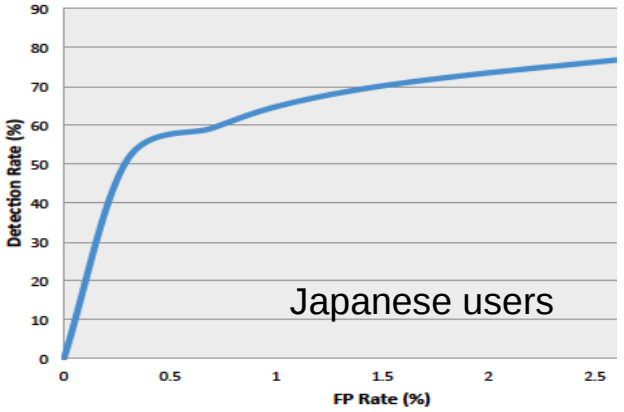
Predictive Analysis

EURECOM
Sophia Antipolis

Logistic Regression classifier

- Area under ROC=0.919
- **74% detection** with **8% FP** (safe users misclassified as at risk)
  - Applied to Japanese users only: 73% detection, **1.9% FP**
- Performances in line with classification algorithms for financial risk prediction





Interesting Result

EURECOM
Sophia Antipolis

- Ability to **predict the users *at risk*** by means of machine learning, by
  - looking **only** at **HTTP requests**
  - without any an access to the user's computer
- Could allow companies or ISPs to silently profile their users
  - ...and calculate aggregated risk factors at a company level
- The **accuracy** of the system is **sufficient** to be used in a **risk prediction** scenario
  - **Simple** but effective way to implement a **cyber-insurance mechanism**
    - rewarding users who show a safe browsing profile

Conclusions

EURECOM
Sophia Antipolis

- The study confirmed some **known trends**:
  - The **more** a user **surfs** the Internet, the **higher** her **risk** of being exposed to cyber attacks
  - The category of the visited web sites does not seem to matter much
    - Few categories are however associated to higher risk (e.g., adult web sites)

- Novel findings:
  - Although not perfect, users' **web browsing profiles** can be used to **predict users** that are more likely to be ***at risk***
    - Having access to users' “social features” could help strengthening the profiles
  - **Cyber Insurance** is a new, attractive area to be researched in depth

Thank you

EURECOM
Sophia Antipolis

?

For further questions, suggestions, comments:

canali@eurecom.fr