Malicious Website Detection

Effectiveness & Efficiency Issues

Birhanu Eshete
eshete@fbk.eu
Fondazione Bruno Kessler,Trento, Italy

1st SysSec Workshop
July 6th, 2011 - Amsterdam, The Netherlands

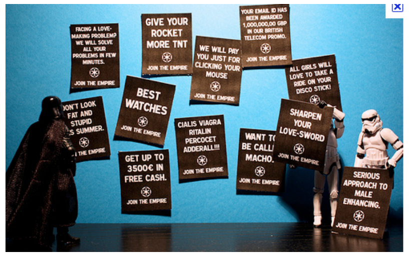
Malicious Websites

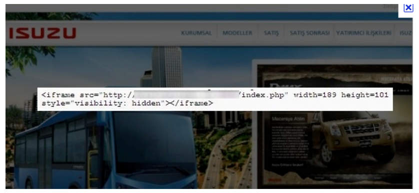
Malicious Websites

- uncover vulnerabilities (browser, plugins, webapp, server), initiate attack
- steal sensitive information, install malware, compromise victim’s machine

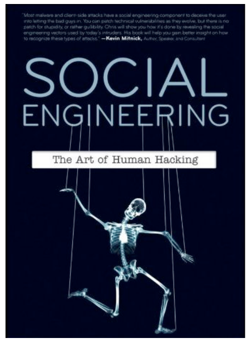
Malicious Websites

- uncover vulnerabilities (browser, plugins, webapp, server), initiate attack
- steal sensitive information, install malware, compromise victim’s machine


FACING A LOVE-MAKING PROBLEM? WE WILL SOLVE ALL YOUR PROBLEMS IN FEW MINUTES. JOIN THE EMPIRE
GIVE YOUR ROCKET MORE TNT JOIN THE EMPIRE
YOUR EMAIL ID HAS BEEN AWARDED 1,000,000,00 GBP IN OUR BRITISH TELECOM PROMO. JOIN THE EMPIRE
WE WILL PAY YOU JUST FOR CLICKING YOUR MOUSE JOIN THE EMPIRE
ALL GIRLS WILL LOVE TO TAKE A RIDE ON YOUR DISCO STICK!
DON'T LOOK FAT AND STUPID THIS SUMMER. JOIN THE EMPIRE
BEST WATCHES JOIN THE EMPIRE
CIALIS VIAGRA RITALIN PERCOCET ADDERALL!!! JOIN THE EMPIRE
WANT TO BE CALLED MACHO. JOIN THE EMPIRE
SHARPEN YOUR LOVE-SWORD JOIN THE EMPIRE
GET UP TO 3500€ IN FREE CASH. JOIN THE EMPIRE
SERIOUS APPROACH TO MALE ENHANCING. JOIN THE EMPIRE

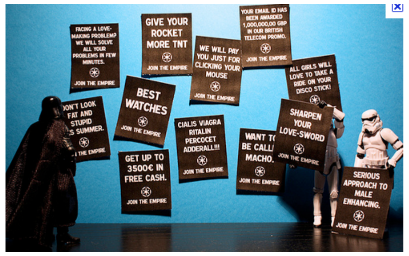


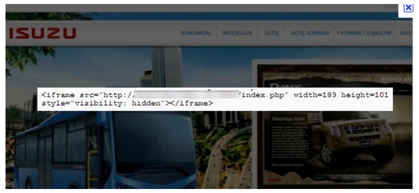
```
<iframe src="http://[redacted]/index.php" width=189 height=101
style="visibility: hidden"></iframe>
```



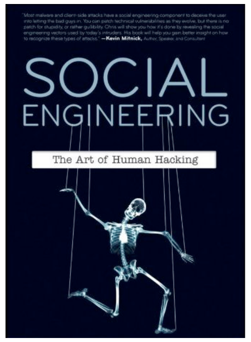
Malicious Websites

- uncover vulnerabilities (browser, plugins, webapp, server), initiate attack
- steal sensitive information, install malware, compromise victim’s machine


GIVE YOUR ROCKET MORE TNT
BEST WATCHES
SHARPEN YOUR LOVE-SWORD
GET UP TO 3500€ IN FREE CASH
SERIOUS APPROACH TO MALE ENHANCING
JOIN THE EMPIRE



```
<iframe src="http:/[redacted]/index.php" width=189 height=101
style="visibility: hidden"></iframe>
```



- 111.4% rise [2009-10], 79.9% malicious legitimate sites [2010], WebSense’10
- 310,000 unique malicious domains, 4.4m average monthly malicious pages, July 2009-June 2010, Symantec’10
- 70 / top 100 reputable websites host malicious content/ have luring redirections to other malicious websites, Symantec’11

Analysis & Detection Approaches

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]
- URL & host information [Canali et. al. 2011], [Ma et. al. 2009]

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]
- URL & host information [Canali et. al. 2011], [Ma et. al. 2009]
- Page content [Canali et.al. 2011], [Tsung et. al. 2010], [Seifert et al. 2008]

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]
- URL & host information [Canali et. al. 2011], [Ma et. al. 2009]
- Page content [Canali et.al. 2011], [Tsung et. al. 2010], [Seifert et al. 2008]
- Execution trace [Qassrawi et al. 2011], [Kim et al. 2011], [Dewald et al. 2010], [Cova et al. 2010], [Iknici et al. 2008], [Alexander et al. 2008]

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]
- URL & host information [Canali et. al. 2011], [Ma et. al. 2009]
- Page content [Canali et.al. 2011], [Tsung et. al. 2010], [Seifert et al. 2008]
- Execution trace [Qassrawi et al. 2011], [Kim et al. 2011], [Dewald et al. 2010], [Cova et al. 2010], [Iknici et al. 2008], [Alexander et al. 2008]

1. Malicious websites are increasing and attack payloads are getting sophisticated (zero-day exploits!)

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]
- URL & host information [Canali et. al. 2011], [Ma et. al. 2009]
- Page content [Canali et.al. 2011], [Tsung et. al. 2010], [Seifert et al. 2008]
- Execution trace [Qassrawi et al. 2011], [Kim et al. 2011], [Dewald et al. 2010], [Cova et al. 2010], [Iknici et al. 2008], [Alexander et al. 2008]

1. Malicious websites are increasing and attack payloads are getting sophisticated (zero-day exploits!)
2. Current approaches are biased to a single prominent attack (partial snapshot=>false signals!)

Analysis & Detection Approaches

- Blacklist-based[Google Safe Browsing]
- URL & host information [Canali et. al. 2011], [Ma et. al. 2009]
- Page content [Canali et.al. 2011], [Tsung et. al. 2010], [Seifert et al. 2008]
- Execution trace [Qassrawi et al. 2011], [Kim et al. 2011], [Dewald et al. 2010], [Cova et al. 2010], [Iknici et al. 2008], [Alexander et al. 2008]

1. Malicious websites are increasing and attack payloads are getting sophisticated (zero-day exploits!)
2. Current approaches are biased to a single prominent attack (partial snapshot=>false signals!)
3. Page features are evolving continously(completeness, semantics, selection => outdated models!)

Effectiveness & Efficiency Issues

Effectiveness & Efficiency Issues

- Which machine learning technique is effective(false signals) and efficient(time to analyze single page) for detecting malicious websites and why?

Effectiveness & Efficiency Issues

- Which machine learning technique is effective(false signals) and efficient(time to analyze single page) for detecting malicious websites and why?
- How and when to update models when page features change?

Effectiveness & Efficiency Issues

- Which machine learning technique is effective(false signals) and efficient(time to analyze single page) for detecting malicious websites and why?
- How and when to update models when page features change?
- Which features to select when there are many candidate feature sets?

Our Approach & Progress

Our Approach & Progress

- A holistic approach that:

Our Approach & Progress

- A holistic approach that:
  - combines URL tokens, host information, page content & execution-trace features (to capture a more comprehensive snapshot of a page), SVMs & HMMs

Our Approach & Progress

- A holistic approach that:
  - combines URL tokens, host information, page content & execution-trace features (to capture a more comprehensive snapshot of a page), SVMs & HMMs
  - incorporates feature evolution (feature-diff monitoring to catch zero-day exploits), GAs

Our Approach & Progress

- A holistic approach that:
  - combines URL tokens, host information, page content & execution-trace features (to capture a more comprehensive snapshot of a page), SVMs & HMMs
  - incorporates feature evolution (feature-diff monitoring to catch zero-day exploits), GAs
  - continuously updates models (fast re-training on selected features), Online LAs

Our Approach & Progress

- A holistic approach that:
  - combines URL tokens, host information, page content & execution-trace features (to capture a more comprehensive snapshot of a page), SVMs & HMMs
  - incorporates feature evolution (feature-diff monitoring to catch zero-day exploits), GAs
  - continuously updates models (fast re-training on selected features), Online LAs
- Work in progress:

Our Approach & Progress

- A holistic approach that:
    - combines URL tokens, host information, page content & execution-trace features (to capture a more comprehensive snapshot of a page), SVMs & HMMs
    - incorporates feature evolution (feature-diff monitoring to catch zero-day exploits), GAs
    - continuously updates models (fast re-training on selected features), Online LAs
- Work in progress:
    - Feature enhancement using Support Vector Machines, preliminary SVM binary model

Our Approach & Progress

- A holistic approach that:
  - combines URL tokens, host information, page content & execution-trace features (to capture a more comprehensive snapshot of a page), SVMs & HMMs
  - incorporates feature evolution (feature-diff monitoring to catch zero-day exploits), GAs
  - continuously updates models (fast re-training on selected features), Online LAs
- Work in progress:
  - Feature enhancement using Support Vector Machines, preliminary SVM binary model
  - Genetic Algorithms for feature evolution (cross-over and mutation)

Thank You!