Anti-Phishing Security Strategy

Angelo P. E. Rosiello

angelo@rosiello.org

The European House
Ambrosetti

Who am I?

- Angelo P. E. Rosiello received the **B.S. and M.S. degrees in Computer Science Engineering cum laude** from "Politecnico di Milano" in 2004 and 2006, respectively.
- Previously Angelo worked for **Accenture** in the Security Strategy Service Line and collaborates with Prof. Christopher Kruegel and Prof. Engin Kirda (Technical University of Vienna) in the ICT security field.
- Angelo works for "***The European House – Ambrosetti***" in the management consulting field. He owns the *ITIL Service Management Certification* and he is a specialist of IT Strategy&Governance.
- Master in Marketing & Communication Management in progress...

Agenda

- **Brief introduction to phishing**
- Strategic defense techniques
- A new client based solution: DOMAntiPhish
- Conclusions

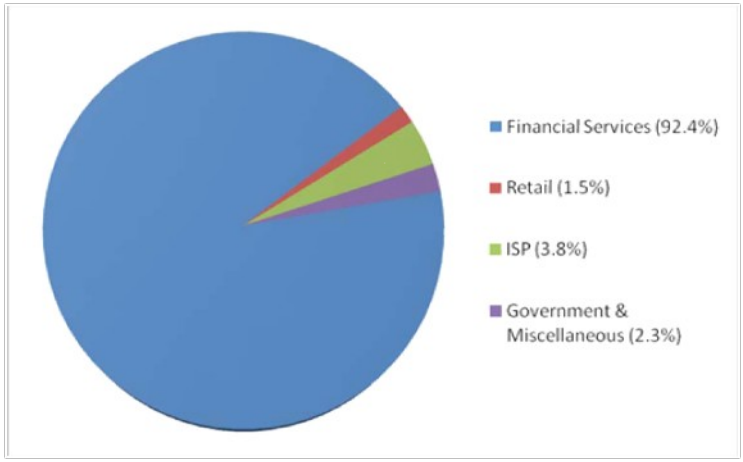
Nature of Phishing

Statistics from the Anti Phishing Working Group (AWPG) confirm the global nature of phishing whose primary target are financial institutions

- List of the main highlights reported for Jan 2008 -

Number of unique reports	29284
Number of unique sites	20305
Number of brands hijacked by phishing campaigns	131
Average time on line for site	3.1 days
Country hosting the most phishing websites	U.S.

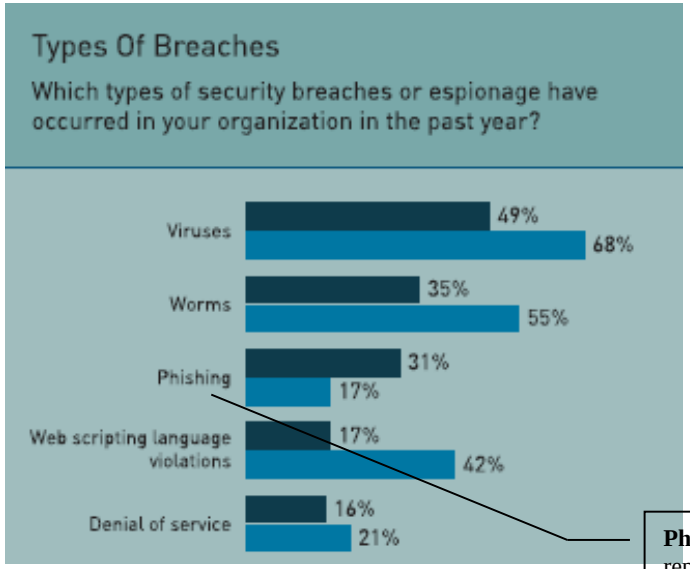
Financial Services continue to be the most targeted industry sector at 92.4% of all attacks in the month of January 2008



Growing Effectiveness and Efficiency of Phishing

Over the last months phishing attacks have become more effective and complex to track and challenge

- The top 5 list of breaches -



InformationWeek Research & Accenture – Information Security Survey 2007

Phishing represents the **third type of successful attack** against enterprises (mainly banks)

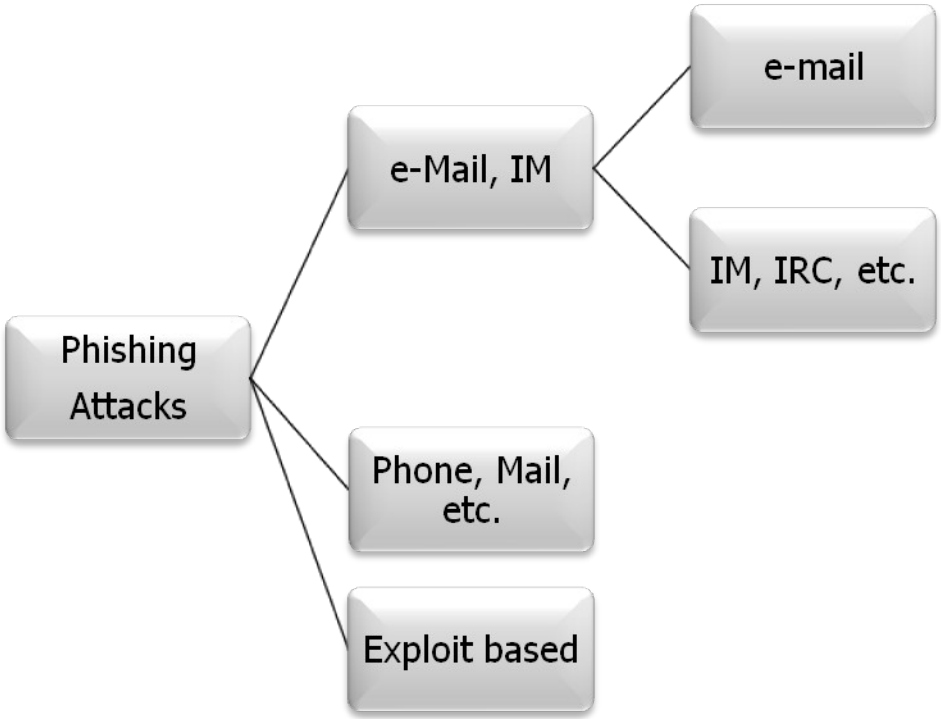
- Improving Phishing quality attacks -

Symantec has detected a number of phishing sites that have been hosted on government URLs over recent months. In June alone (2007), fraudulent sites were identified on sites run by the governments of Thailand, Indonesia, Hungary, Bangladesh, Argentina, Sri Lanka, the Ukraine, China, Brazil, Bosnia and Herzegovina, Colombia, and Malaysia.
"*Hosting a phishing Web page on a government site has a number of advantages for a phisher. Government Web sites often receive a high volume of traffic, so their servers can handle the extra traffic generated by a phishing site*" writes Symantec researcher Nick Sullivan. "*This extra traffic might not be noticed immediately, giving the phishing site a longer lifespan before it is detected and shut down. Perhaps most importantly, hosting a phishing site on an actual government URL gives the phishing site a sense of authenticity that's hard to beat.*"

Taxonomy of Phishing Attacks

Phishing attacks can be classified according to their nature

- Classification of the Attacks -

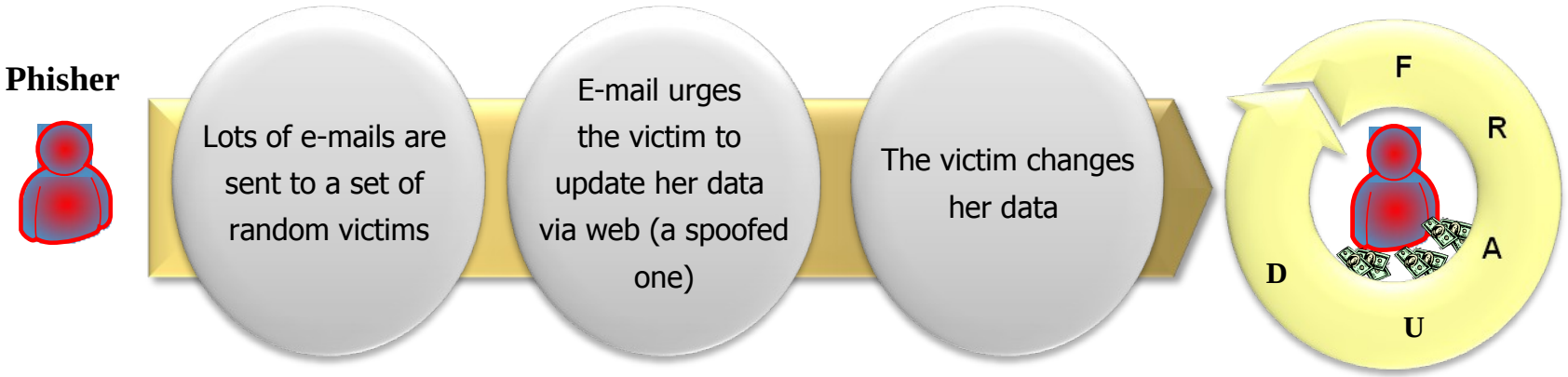


- Description -

- **Spoofed e-mail** are sent to a set of victims asking them (usually) to upgrade their passwords, data account, etc.
- **MSN, ICQ, AOL and other IM channels** are used to reach the victims. Social engineering techniques are used to gain victim's sensitive information
- Calling the victims on **the phone, classic social engineering techniques** are used by phishers
- Another kind of attack is based on the **internet browser vulnerabilities**. This approach is usually adopted to automatically install dialers

A Process of Phishing Attacks

- In a typical attack, **the phisher sends a large number of spoofed** (i.e. fake) **e-mails to random Internet users** that seem to be coming from a legitimate and well-known business organization (e.g. financial institutions, credit card companies, etc)
- The e-mail **urges** the victim **to update his personal information** as a condition to avoid loosing access rights to specific services (e.g. access to online bank account, etc).
- By clicking on the link provided, the **victim is directed to a bogus web site implemented by the attacker**
- The **phishing website is structured as a clone of the original website** so that the victim is not able to distinguish it from that of the service he/she has access to.



New Phishers Skills

To confuse the victim, phishers are devising new tricks

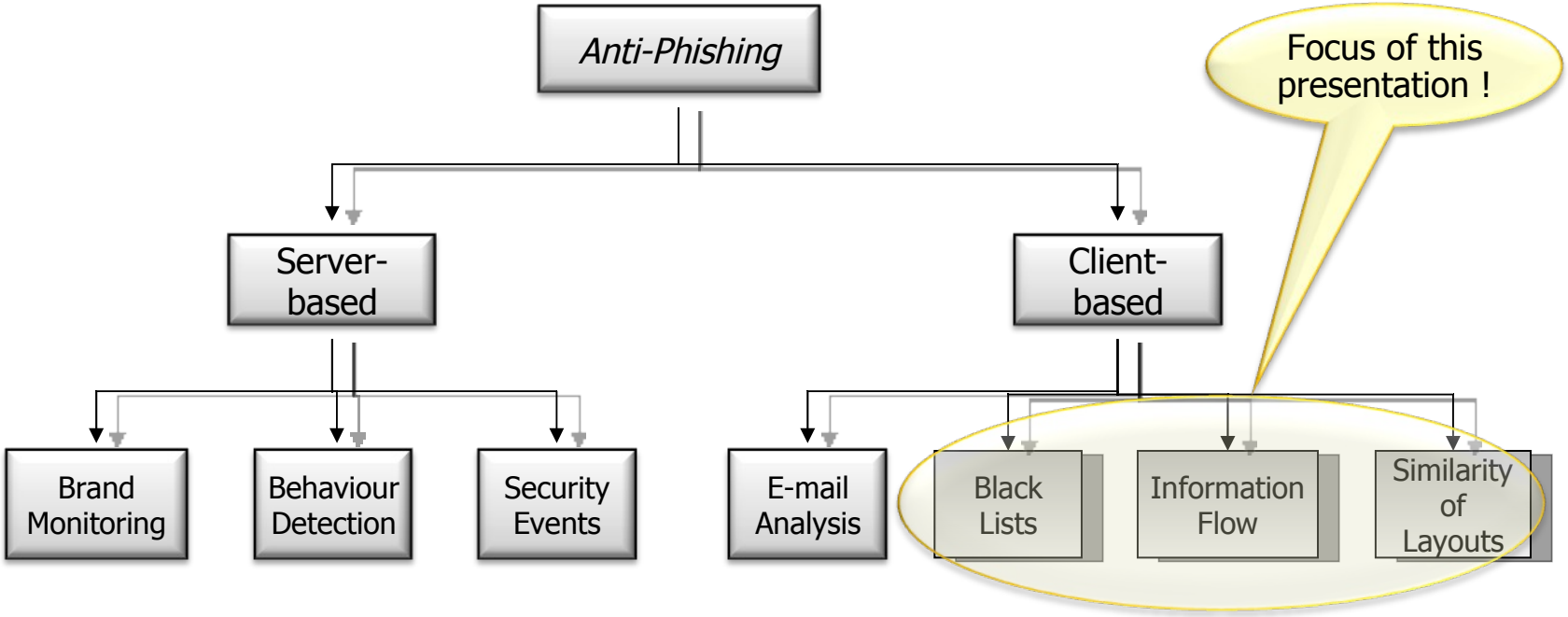
- **Phishing e-mail embed hyperlinks from the original website** so that the users mainly surf on the real web server executing only a small number of connections to the fake web server.
- **Website URL** are **encoded or obfuscated** to not raise suspicious. IDN spoofing, for example, uses Unicode URLs that render URLs in browsers in a way that the address looks like the original web site address but actually link to a fake web site with a different address.
- Victims are redirected to a phishing website by first using **malwares to install a malicious Browser Helper Object** (BHO). BHOs are DLLs that allows developers to customize and control Internet Explorer but also phishers to compromise connections.
- The **hosts file** on the victim's machine is **corrupted**, for example using a malware. The host files maintains local mappings between DNS names and IP addresses. By inserting a fake DNS entry into the user's hosts file, it will appear that their web browser is connecting to a legitimate website when in fact it is connecting to a phishing website.

Agenda

- Brief introduction to phishing
- **Strategic defense techniques**
- A new client based solution: DOMAntiPhish
- Conclusions

Strategic Defense Techniques

Antiphising defenses can be server and client based solutions



Server-based Solutions

Server based techniques are implemented by service providers (e.g. ISP, e-commerce stores, financial institutions, etc…)

Brand Monitoring	Crawling on-line websites to identify "clones" (looking for legitimate brands), which are considered phishing pages. Suspected websites are added to a centralized "black-list".
Behaviour Detection	For each customer a profile is identified (after a training period) which is used to detect anomalies in the behaviour of users
Security Event Monitoring	Security event analysis and correlation using registered events provided by several sources (OS, application, network device) to identify anomalous activity or for post mortem analysis following an attack or a fraud
Strong Authentication	Using more than one identification factor is called strong authentication. There are three universally recognized factors for authenticating individuals: something you know (e.g. password); something you have (e.g. hw security token); something you are (e.g. fingerprint)
New Authentication Techniques	New techniques of authentication are under research, such as using an image during the registration phase which is shown during every login process

Client-based Solutions

Client-based techniques are implemented on users' end point through browser plug-ins or e-mail clients

E-mail Analysis	E-mail-based approaches typically use filters and content analysis. If trained regularly Bayesian filters are actually quite effective in intercepting both spamming and phishing e-mails.
Black-Lists	Blacklists are collections of URLs identified as malicious. The blacklist is queried by the browser run-time whenever a page is loaded. If the currently visited URL is included in the blacklist, the user is advised of the danger, otherwise the page is considered legitimate.
Information Flow	Information flow solutions are based on the premise that while a user may be easily fooled by URL obfuscation or a fake domain name, a program will not. AntiPhish is an example of this type of defense technique which keeps track of the sensitive information that the user enters into web forms, raising an alert if something is considered unsafe
Similarity of Layouts	Most advanced techniques try to distinguish a phishing webpage from the legitimate one comparing their visual similarity [[Wenyin, Huang, Xiaoyue, Min, Deng], [Rosiello, Kirda, Kruegel, Ferrandi]

Trends on client-based Market Solutions

In the last months the major browsers (e.g. IE7 and Mozilla Firefox) have integrated specific anti-phishing functionalities (black-lists and static page analysis)

- In October 2006, **a Microsoft-commissioned report** on various anti-phishing solutions was released. **The testers found that Microsoft Internet Explorer (IE) 7.0 has better anti-phishing technology than competing solutions**. The products tested included IE 7.0 Beta 3, EarthLink ScamBlocker, eBay Toolbar with Account Guard, GeoTrust TrustWatch, Google Toolbar for Firefox with Safe Browsing, McAfee SiteAdvisor Plus, Netcraft Toolbar, and Netscape Browser with built-in antiphishing technology
- **The Mozilla Foundation commissioned its own study** to gauge the effectiveness of Mozilla Firefox 2.0's anti-phishing technology as compared with IE 7.0's. **This study found that Firefox's anti-phishing technology was better than IE**'s by a considerable margin
- It seems evident that **we cannot trust both above studies** and for this reason **we consider a third independent evaluation** realized by the Security Lab of the Technical University of Vienna

Analysis of the Black-Lists

Over a period of three weeks the Technical University of Vienna (TUWIEN) has collected 10,000 URLs to benchmark Microsoft and Google's black-lists. Based on three indicators, the research shows that Google performs better than Microsoft

- Experimental Results -

	Google	Microsoft
Sites	3,595 (100%)	3,592 (100%)
BL initally **BL delayed**	3,157 (87.89%) 84 (2.34%)	2,139 (59.55%) 274 (7.63%)
BL Total	3,241 (90.23%)	2,413 (67.18%)
ART	9.3 h	6.4 h

- KPI -

Coverage: percentage of phishing URLs already included in the list

Quality: percentage of legitimate URLs incorrectly included in the list

Average Response Time (ART): average time required to insert not initially included URLs

Static Page Analysis

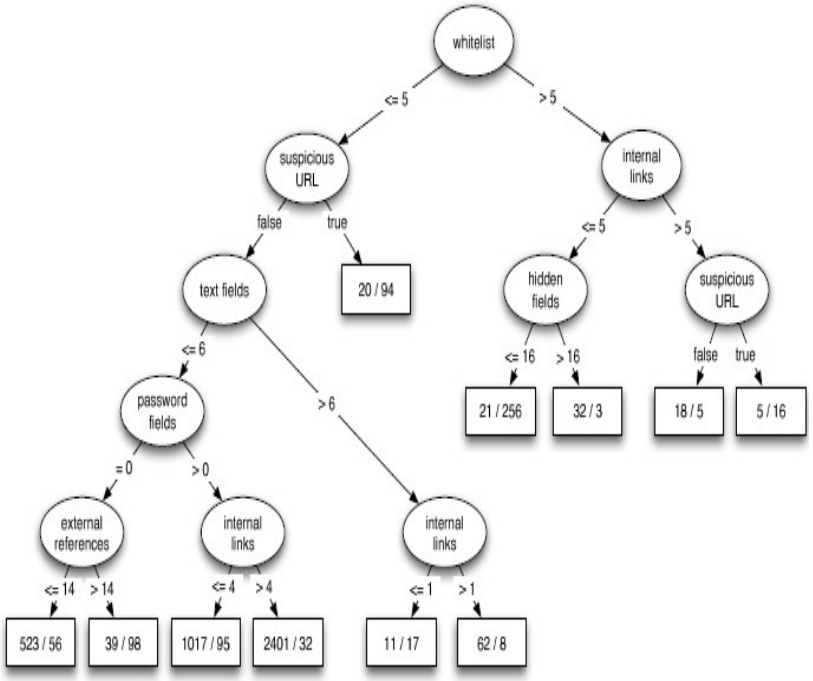
TUWIEN has demonstrated that a set of page properties actually allows to differentiate between malicious (phishing) and legitimate (benign) ones

Select a set of page properties	Collect web pages to be analyzed	Extract the classification model	Infer about phishing
• **18 properties are considered** mainly extracted from the HTML source code (e.g. forms, input fields, links, script tags, etc.)	• A **set of legitimate and phishing web pages** are collected to extract the classification model	• The **C4.2 algorithm** is executed **to identify the classification model** (i.e. the decion-tree)	• An **automatic tool** that uses the extracted classification model **can distinguish phishing from legitimate web pages**

Static Page Analysis: Experimental Results

The decision-tree is extracted using the Weka package (algorithm J48) on a set of 4,829 web pages

- Reduced Tree extracted using the Weka package -



- Confusion Matrix -

	Classified as Legitimate	Classified as Phishing
Legitimate Pages	4,131	18
Phishing Pages	115	565

The qualifier is quite successful in identifying phishing pages (more than 80% are correctly recognized), raising only a very small number of false alerts (18 out of 4,149 pages are incorrectly classified as phishing)

Static Page Analysis: Demo

Starting from the training data-set, a real time demonstration is provided

- Steps to be executed -

- Install the Weka Package
- Load the input ".arf" or ".csv" file
- Select the J48 algorithm
- Run the application
- Check the extracted tree

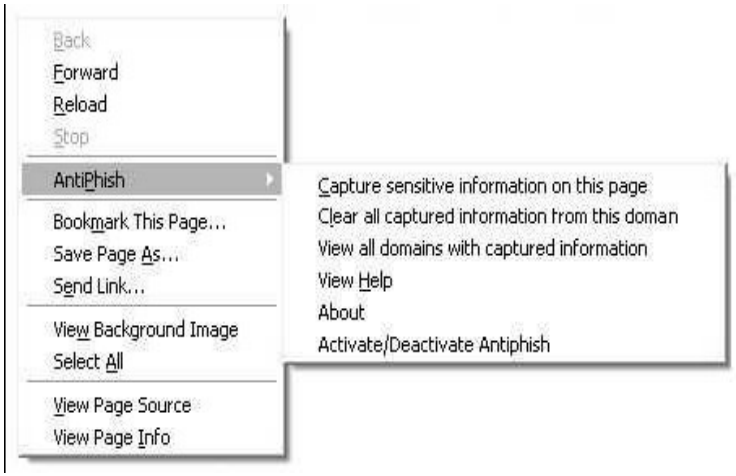
Information Flow Solutions: AntiPhish (1/2)

A limited number of information flow based solutions were realized. The objective is to protect users by checking where the information is sent to

- General description -

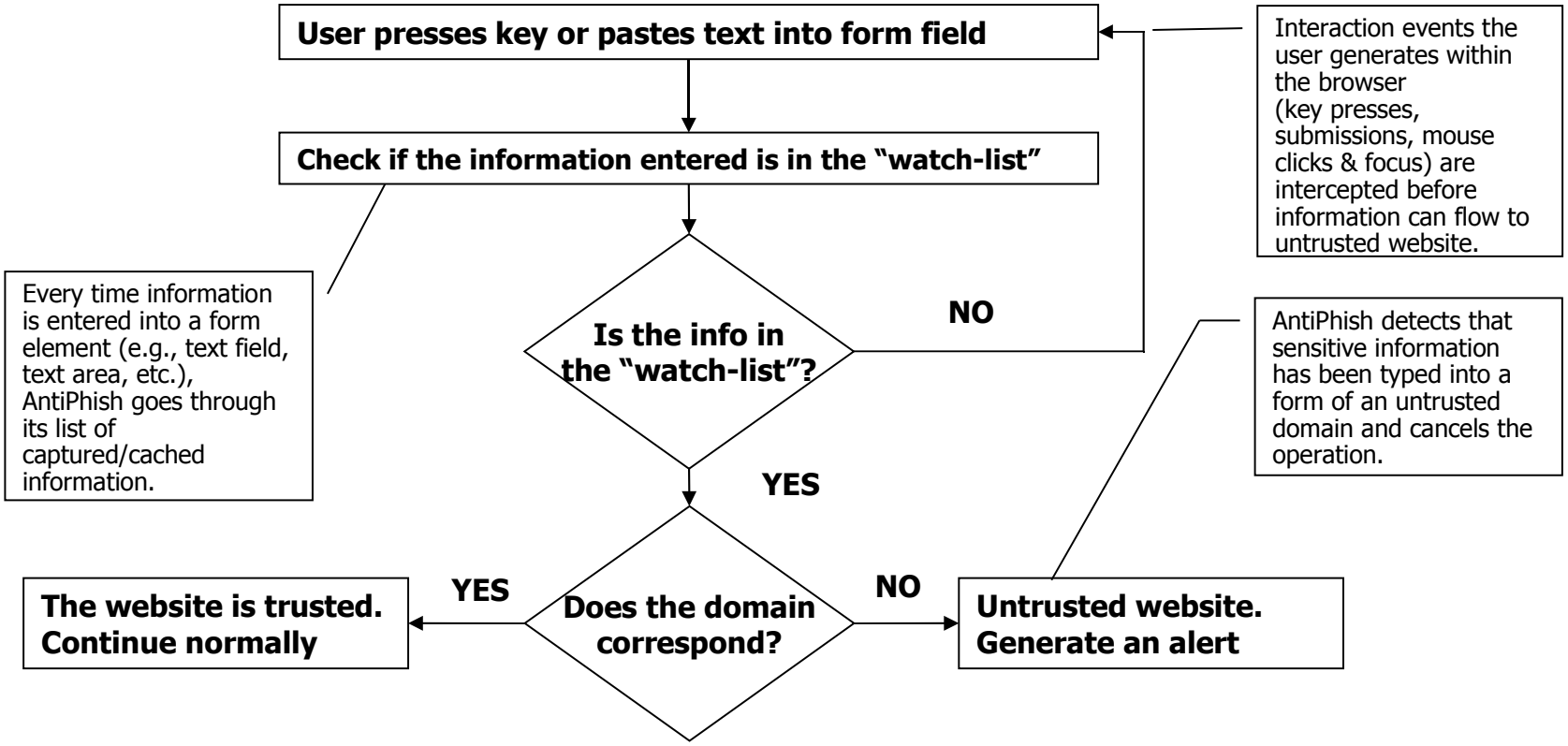
- **AntiPhish** is an application that is integrated into the browser as an **external plug-in**
- After AntiPhish is installed**, the browser prompts a request for a new master password** when the user enters input into a form for the first time
- The **master password** is used **to encrypt the sensitive information** before it is stored (using DES)
- After the user enters sensitive information such as a password, the **AntiPhish menu is used to scan the page and to capture and store** this **information** with the domain of the website, too

- How does it look like? -



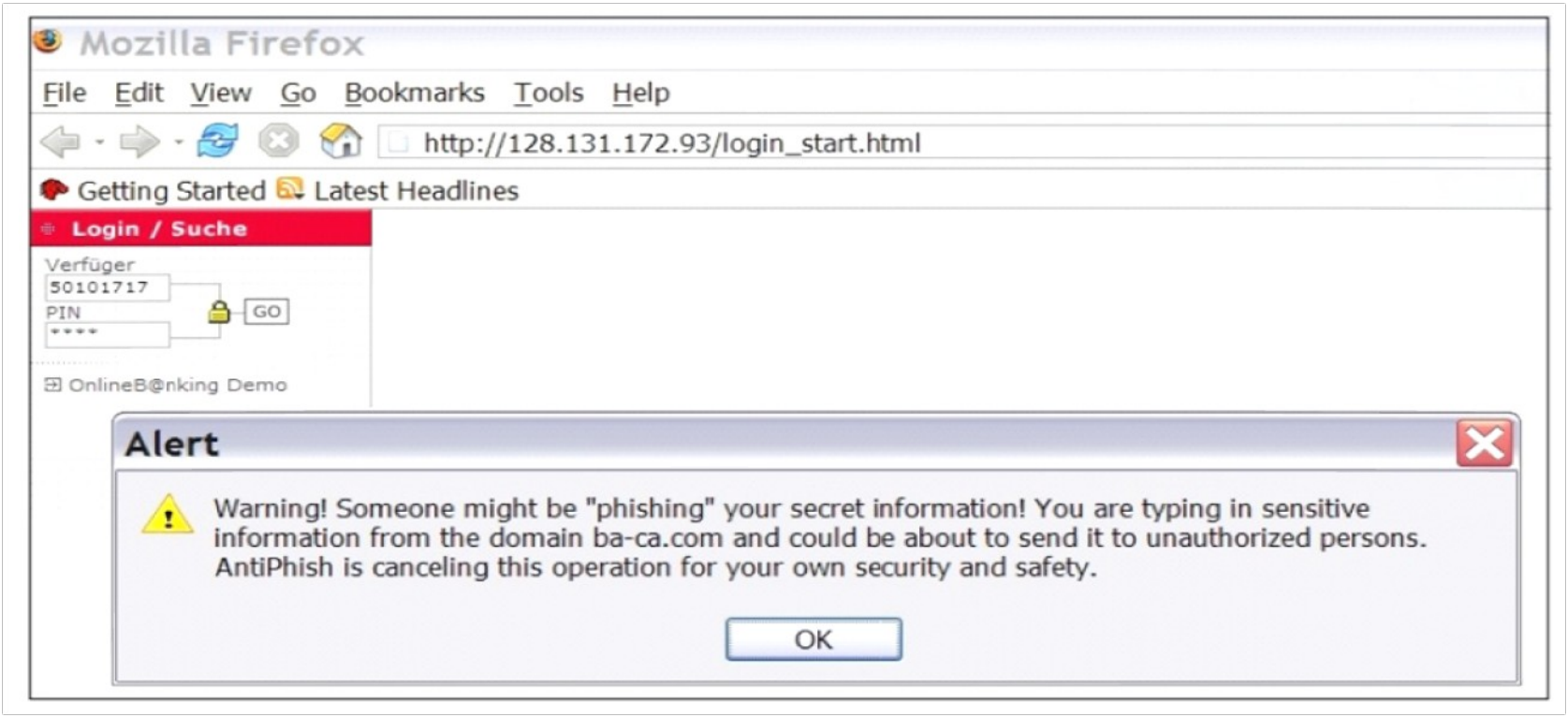
Information Flow Solutions: AntiPhish (2/2)

The execution flow chart of AntiPhish indicates how this tool allow to protect potential victims



AntiPhish in Action

When the victim inserts his username and password to an untrusted web site, an alert is raised before sensitive information are sent to the phisher

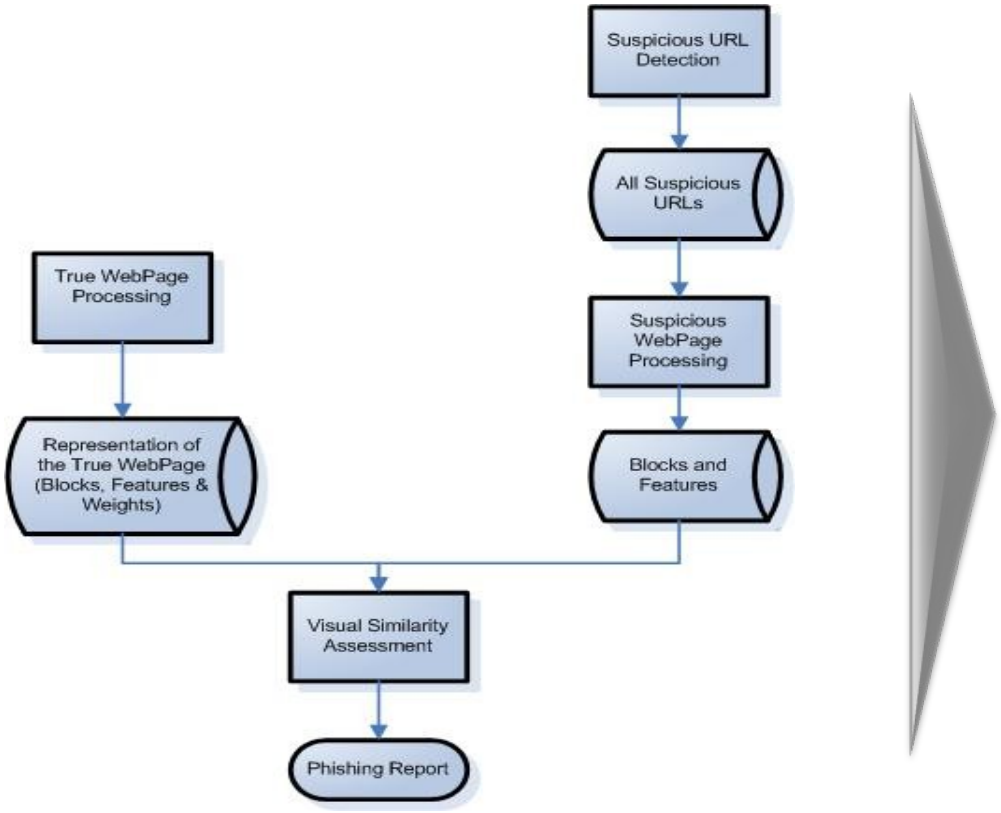


Agenda

- Brief introduction to phishing
- Strategic defense techniques
- **A new client based solution: DOMAntiPhish**
- Conclusions

Layout-Similarity-based Solutions (1/2)

Layout-similarity-based approaches classify a web page as a phishing page if its "visual" similarity value is above a predefined threshold



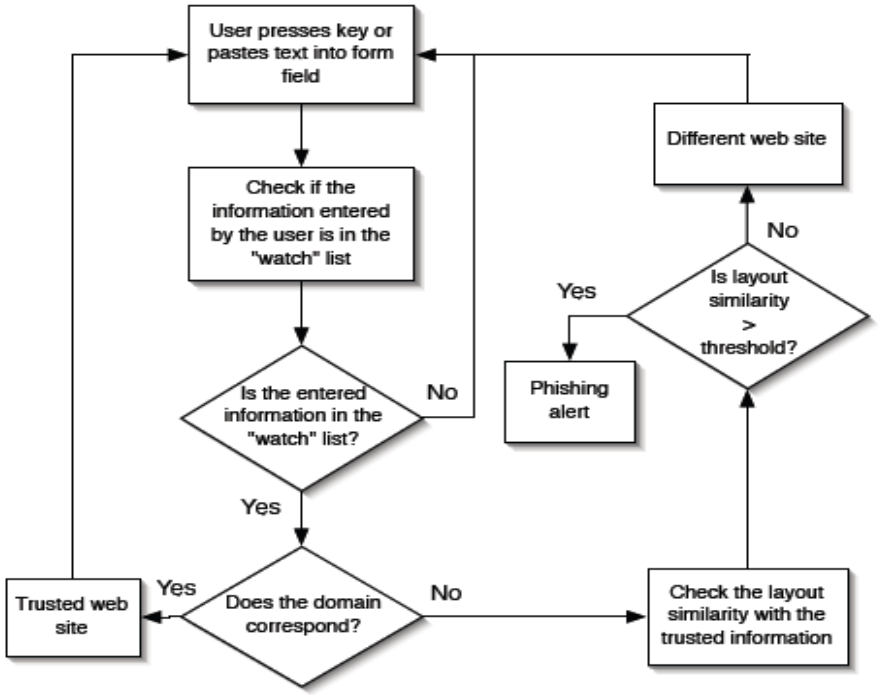
- Wenyin et al. Approach -

- The webpage is decomposed into **salient blocks** according to "**visual cues**".
- The **visual similarity** between two web pages is **measured**.
- A web page is considered a **phishing page if the similarity** to the legitimate web page is **higher than a threshold**.

Layout-Similarity-based Solutions (2/2)

DOMAntiPhish [Rosiello, Kirda, Kruegel, Ferrandi] computes the similarity value extracting the DOM-Tree of the considered webpages

- DOMAntiPhish Flowchart -



- DOMAntiPhish description -

- When a password associated with a certain domain is reused on another domain the system compares the layout of the current page with the page where the sensitive information was originally entered.
- For the comparison the DOM-Tree of the original webpage and the new one are checked.
- If the system determines that these pages have a similar appearance, a phishing attack is assumed

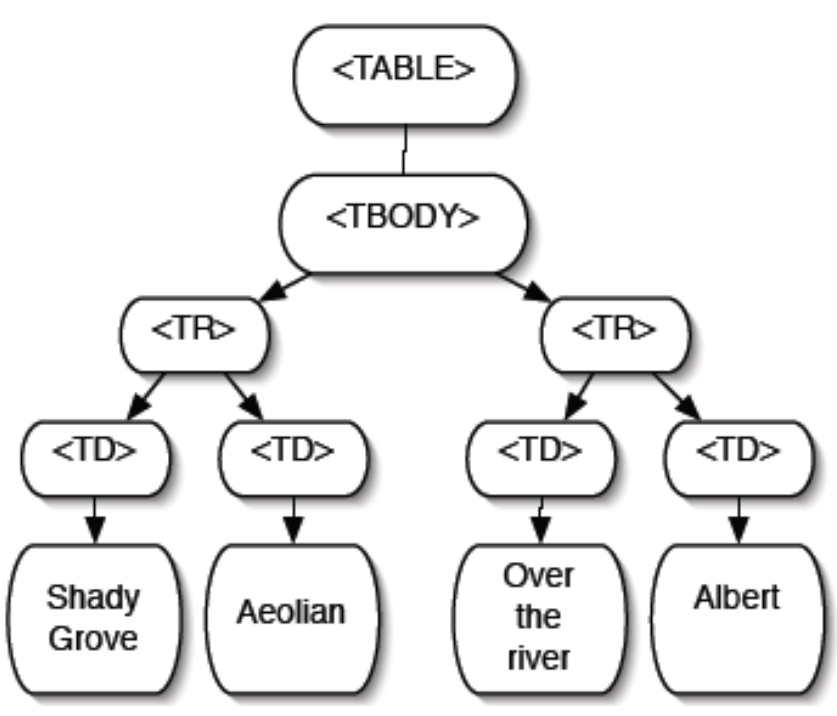
DOMAntiPhish: DOM-Tree Extraction

The Document Object Model (DOM)-Tree is an internal representation used by browsers to represent a web page

- HTML source code -

```
<TABLE>
<TBODY>
<TR>
<TD> Shady Grove </TD>
<TD> Aeolian </TD>
</TR>
<TD> Over the river </TD>
<TD> Albert </TD>
</TR>
</TBODY>
</TABLE>
```

- DOM-Tree representation -



DOMAntiPhish: Similarity Computation

DOM-Trees reduce the problem of computing the layout similarity of two webpages to the problem of establishing if two trees are isomorphic

- Templates computation algorithm -

INPUTS: vertex v, vertex u, firstSubTree Φ, secondSubTree Φ

WHILE continue_while exists equivalent_subTrees_branches **DO**

firstSubTree = **getSubTree**(u, firstSubTree);

secondSubTree = **getSubTree**(v, secondSubTree);

IF are similar(firstSubTree, secondSubTree) **THEN**

float penalty=**compute_similarity_penalty**( );

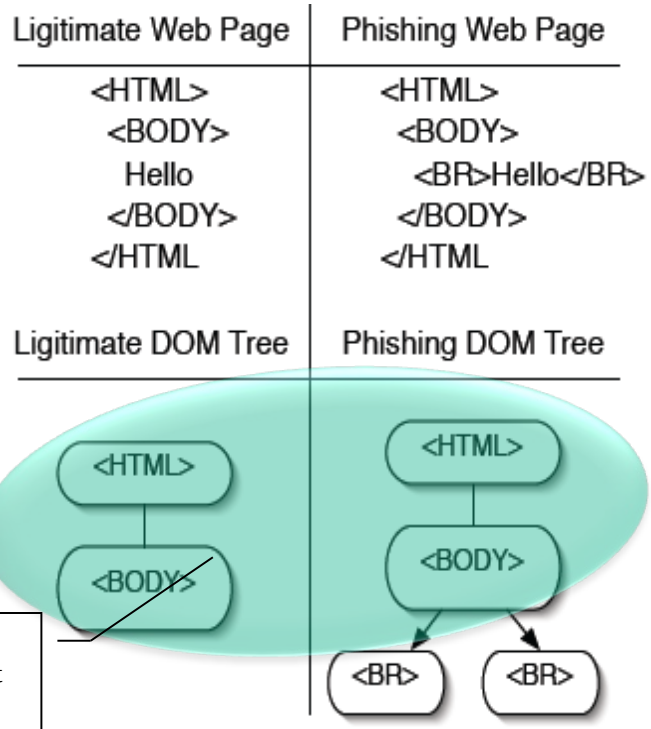
store subTrees(u, v, firstSubTree, secondSubTree, penalty);

END IF

END WHILE

Equal templates extracted by the algorithm. To cover the trees, the best set of templates are selected (minimizing the similarity penalties)

- Phishing Example -



DOMAntiPhish: Implementation Process

DOMAntiPhish prototype is implemented as a Javascript plug-in for Mozilla Firefox 2.0 which invokes a Java software to compute the layout similarity

- The **Javascript plug-in** for Mozilla Firefox 2.0 **extracts the DOM-Tree** representation of each stored webpage and browsing one

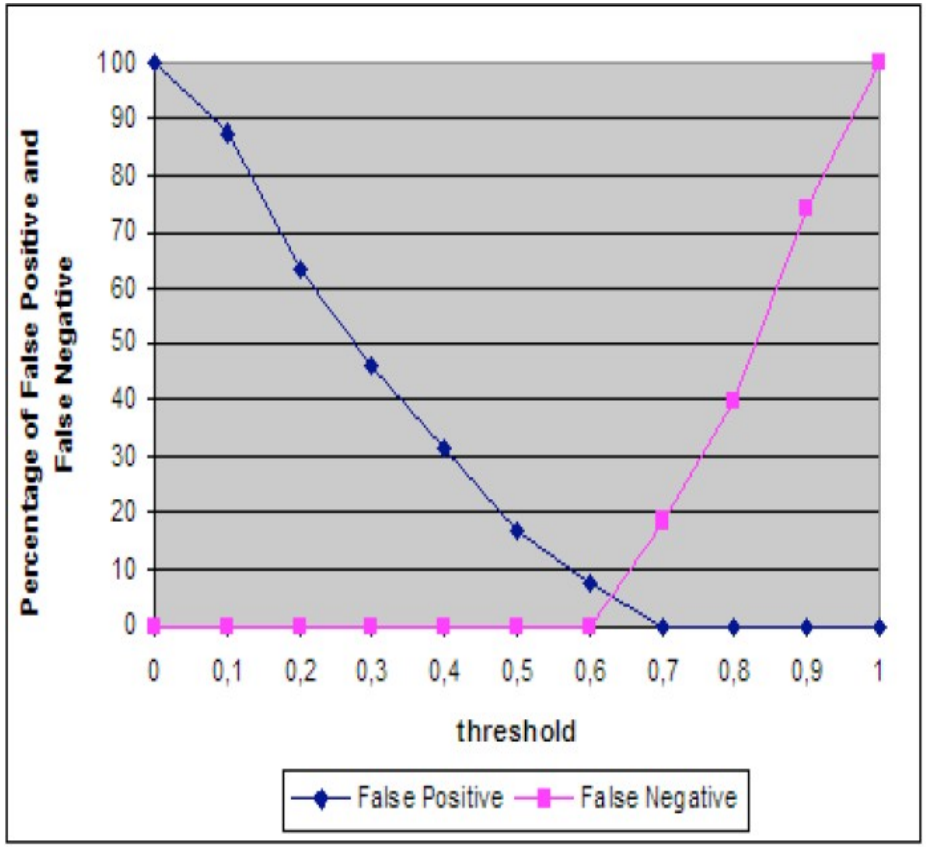
- The **Javascript plug-in writes** down two text files that contain the **extracted DOM-Trees**
- The Javascript **plug-in invokes** the **Java software**

- The **Java software calculates the similarity** of the analyzed DOM-Trees choosing the set of templates which minimize the similarity penalty and maximize the coverage

- The **Javascript plug-in** reads the similarity value from a text file and **returns the phishing report** to the user

DOMAntiPhish: Experimental Results

DOMAntiPhish was tested on a set of over 200 websites proving that our approach is feasible in practice



- Experimental results description -

- During the similarity computation process, for the isomorphic sub-trees identification algorithm, we added a penalty of 0.3 if two corresponding tags had different types or if a tag did not have children and its matched counterpart did.
- If two attributes of matched tags were different, a penalty of 0.1 was added. Moreover, if the attributes had different values, then a penalty of 0.05 was added, too.
- The penalty values were determined empirically by having as objective function the minimization of false positive and negative results for low and high threshold values respectively.

DOMAntiPhish: Limitations

As every security solution, also DOMAntiPhish is not perfect and we can identify the following main limitations:

- Potential attacks -

It could be possible for attackers to use a **combination of images** to create a **spoofed web page** that looks visually similar to a legitimate web page. Hence, the **DOM** of the spoofed web page would be **different** and detection would be evaded.

Another possible problem could be **DOM obfuscation** attempts that would make the visual look similar to the legitimate web page while at the same time evading detection.

- Defensive solutions -

One possibility of dealing with this limitation could be to take a **conservative approach** and to tag web pages as being **suspicious** that contain a **large number of images** or that mainly consist of images.

Our approach raises the difficulty bar for creating phishing pages. Furthermore, one can always **take a more conservative approach** by reducing the phishing alert threshold. Also, if phishers are forced to alter the look and feel of their phishing pages, these **pages will become less convincing** and more suspicious to the victims.

DOMAntiPhish: Demo

Browsing some webpages we show how DOMAntiPhish works against phishing attacks

- Steps to be executed -

- Install DOMAntiPhish plug-in
- Log into a trusted website
- Try to log into a phishing website
- Check the phishing report

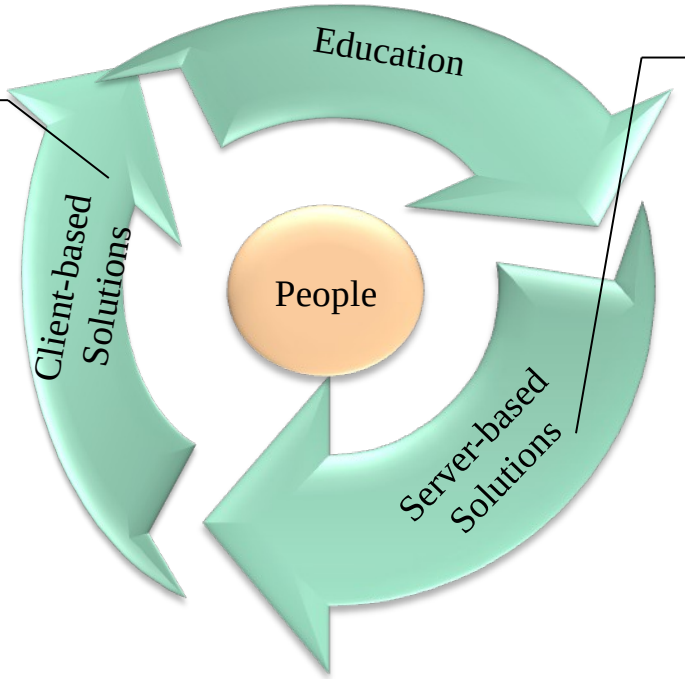
Agenda

- Brief introduction to phishing
- Strategic defense techniques
- A new client based solution: DOMAntiPhish
- **Conclusions**

Conclusions

As for every IT attack, phishing can be prevented, detected and mitigated through server-based and client-based approaches, supported by education and awareness

Client-based techniques try to protect users implementing local solutions, such as browser plug-ins or e-mail clients



Server based techniques are applied on servers or providers that offer services to customers

References

- Angelo P. E. Rosiello, Engin Kirda, Christopher Kruegel, and Fabrizio Ferrandi. *"A Layout-Similarity-Based Approach for Detecting Phishing Pages"*. IEEE International Conference on Security and Privacy in Communication Networks (SecureComm), Nice, France, September 2007
- Christian Ludl, Sean McAllister, Engin Kirda, and Christopher Kruegel. *"On the Effectiveness of Techniques to Detect Phishing Sites"*. Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA) 2007 Conference, Lucerne, Switzerland, July 2007
- Engin Kirda and Christopher Kruegel. *"Protecting Users against Phishing Attacks"*. The Computer Journal, 2006.
- Neil Chou, Robert Ledesma, Yuka Teraguchi, Dan Boneh, and John Mitchell. *"Client-side defense against web-based identity theft"*. In 11th Annual Network and Distributed System Security Symposium (NDSS '04), San Diego, 2005.
- Anti-Phishing Working Group (APWG). APWG Homepage. http://www.antiphishing.org/, 2007.
- Information Security Survey 2007 – *InformationWeek Research & Accenture*
- Google. Google Whitelist. http://sb.google.com/safebrowsing/update?version=goog-white-domain:1:-1, 2007.
- Mozilla. Firefox 2 Phishing Protection Effectiveness Testing. http://www.mozilla.org/security/phishing-test.html, 2006.
- Verisign. Anti-Phishing Solution. http://www.verisign.com/ verisign-business-solutions/anti-phishing-solutions/, 2005.
- Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. Phinding Phish: Evaluating Anti-Phishing Tools. In Network and IT Security Conference: NDSS 2007, San Diego, California, 2007.
- Weka. http://www.cs.waikato.ac.nz/ml/weka/