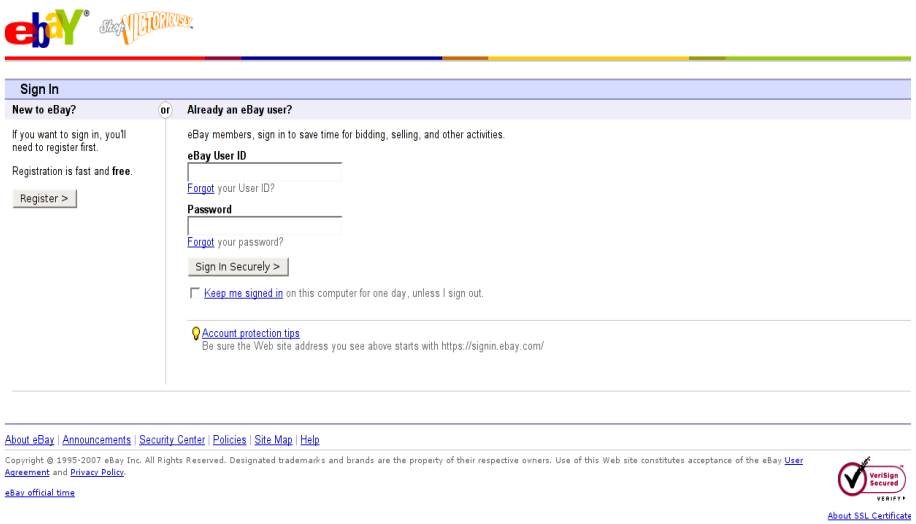
Visual-Similarity-Based Phishing Detection

Eric Medvet, University of Trieste (IT)
Engin Kirda, Christopher Kruegel, Technical University of Vienna (AT)

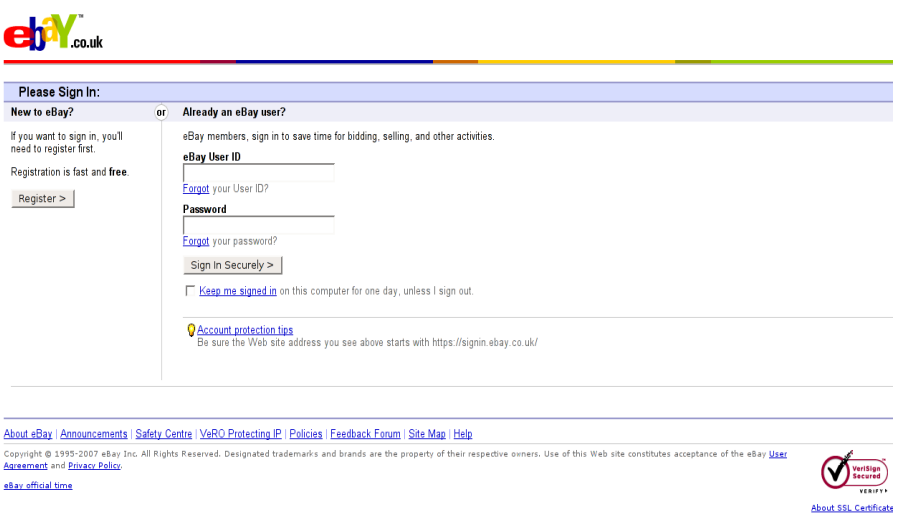
What is phishing?

- A form of **online fraud** aimed at stealing user's sensitive information
- The attacker crafts a web page that mimicks the original page and spreads its URL using emails
- An important and current problem:
  - More than 25000 phishing attacks in Dec. 2007
  - "Do yourself" phishing kits
  - Can be combined with other online attacks

Phishing: an example



The genuine page



The phishing page

Against phishing: state of the art

- At the email level
  - Blocking suspected emails (similarly to antispam techniques)
  - Sender authentication (*not widely used*)
- Black lists (*uneffective vs. zero day attack*)
- Browser-integrated solutions
  - Domain-specific passwords
  - Detect suspected URLs
  - AntiPhish, prevents sending credentials to untrusted unknown websites...

Antiphish

1. Store user's credentials and link to corresponding sites:
    - <username1, passwd1, url1>
    - <username2, passwd2, url2>
    - . . .
2. Whenever credentials are to be sent to a website, check if it is trusted

- Problem: what if the user uses the **same credentials** on many websites? False positives
- Solution: check if the website is **suspiciously similar** to the trusted one

Contribution of this work

- A novel visual comparison technique for phishing detection
- **Compare** the suspected web page with the corresponding genuine one: if they are too similar, raise an alert
  - Thought to be part of AntiPhish
  - ... but can also be integrated in other solutions: e.g., in a mail server to block suspected emails

Visual comparison: overview

- Why "visual"?
- Users get convinced to be on a legitimate page mainly by perceived page content
- We compare page features that **affect user perception**:
  - Text blocks
  - Images
  - Overall image (the page **as rendered by the browser**)
- We obtain a visual similarity index

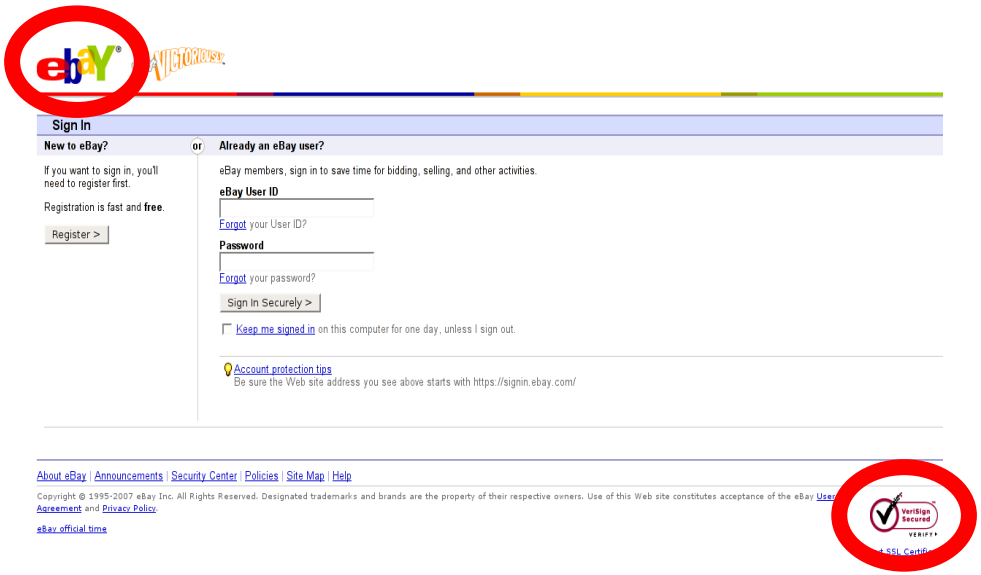
What we consider

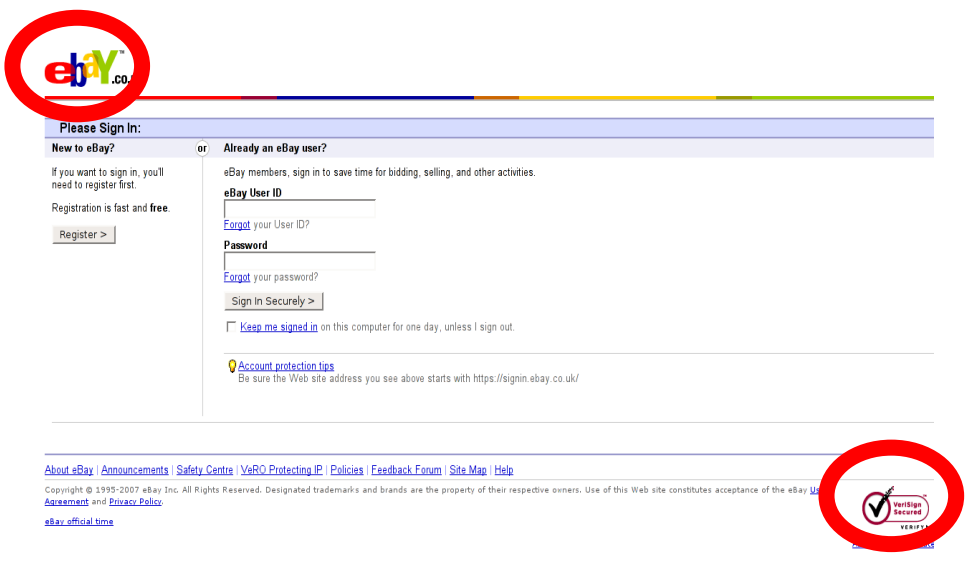
- Text blocks:
  - textual content, foreground color, background color, font size, font family name, position
- Images:
  - `src` attribute, area (w x h), color histograms, 2D Haar transform, position
- Overall image:
  - color histograms, 2D Haar transform

Comparison

1. We compute distances between elements of each text block pairs, image pairs and the overall image pair:
    - a similarity index for each pair

Comparison





1. ebay logo vs ebay logo
2. ebay logo vs verisign logo
3. verisign logo vs ebay logo
4. verisign logo vs verisign logo
5. ...

Comparison

1. We compute distances between elements of each text block pairs, image pairs and the overall image pair:
    - a similarity index for each pair
2. We consider only the first 10 **best-matching** text block pairs, the 5 best-matching image pairs, the 2 overall images:
    - three similarity indexes for text, images, overall image
3. We obtain the final similarity index averaging indexes of text blocks, images and overall image

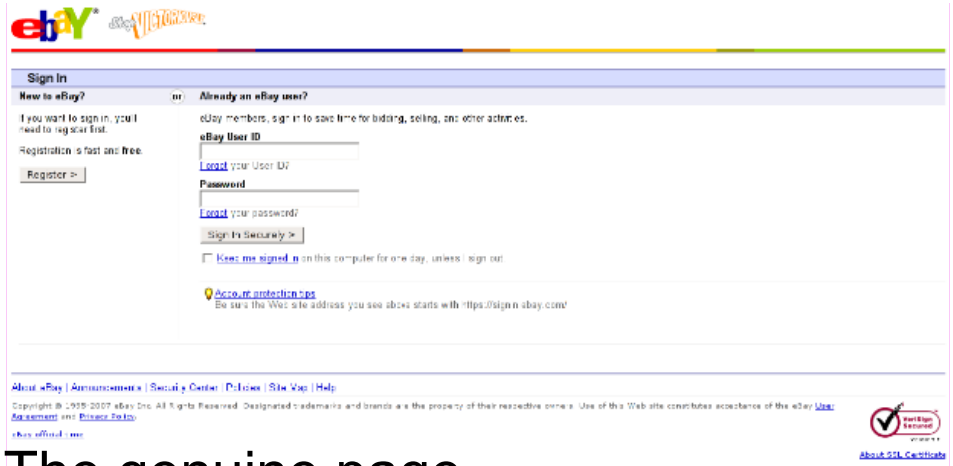
- Some weights (coefficients) are set by domain knowledge, the others by optimization on a small training set
- Some algorithm optimization in order to reduce computational effort

Experimental evaluation: dataset

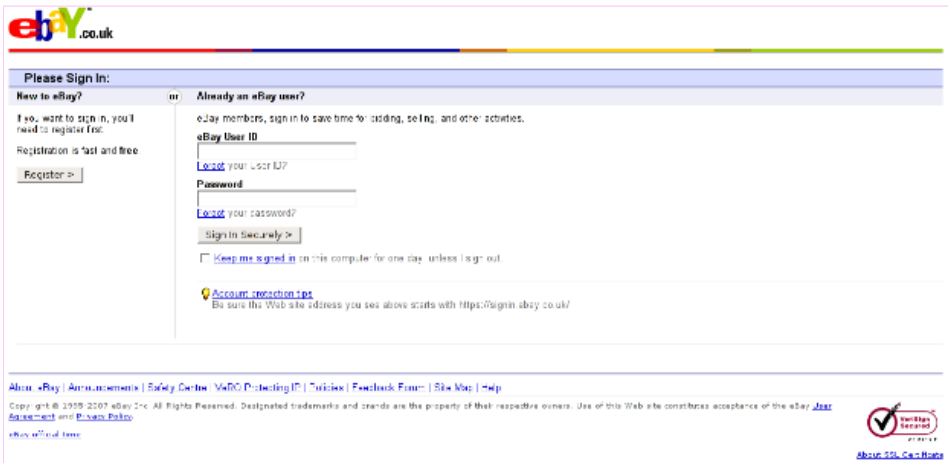
Dataset

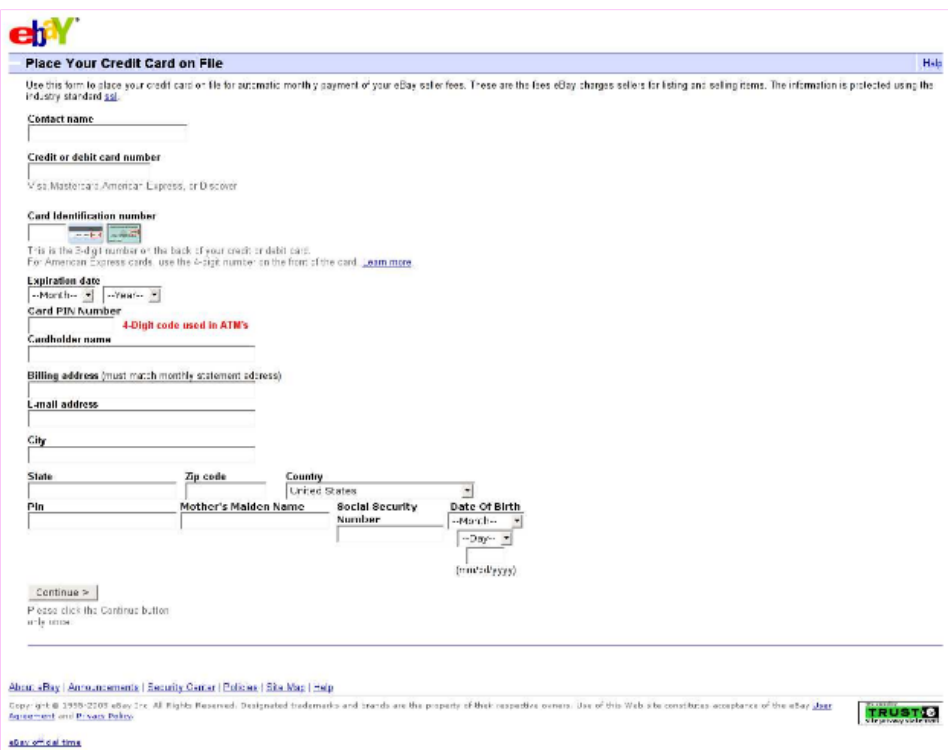
- 41 **positive pairs** <genuine page, phishing page>, partitioned according to visual similarity
  - 20 level 0: almost perfect visual match
  - 14 level 1: small differences
  - 7 level 2: noticeable differences
- 161 **negative pairs** <genuine page, different page>
  - 115 with a login form
  - 46 without a login form

Positive pairs: example



The genuine page





Testing methodology

- Portion of the dataset used for training
  - 14 on 41 positive pairs
  - 21 on 161 negative pairs
- FPR and FNR computed on the remaining part

Results

- Satisfacting results:

Levels	t	f_p	n_n	FPR	f_n	n_p	FNR
All (0, 1 and 2)	0.956	0	140	0%	2	27	7.4%
Only 0 and 1	0.956	0	140	0%	0	20	0.0%
Only 0	0.956	0	140	0%	0	11	0.0%

- **No false positives**
- Only 2 false negatives (level 2, rather difficult to detect)

False negative: a sample

Thanks

Questions?