Anticipating Hidden Text Salting in Emails (Extended Abstract)

Christina Lioma[1], Marie-Francine Moens[1], Juan-Carlos Gomez[1], Jan De Beer[1*], Andre Bergholz[2], Gerhard Paass[2], and Patrick Horkan[3]

[1] Katholieke Universiteit Leuven, Belgium
{christina.lioma,sien.moens,juancarlos.gomez}@cs.kuleuven.be
[2] Fraunhofer IAIS, Germany {andre.bergholz,gerhard.paass}@ais.fraunhofer.de
[3] Symantec, Ireland patrick_horkan@symantec.com

Abstract. Salting is the intentional addition or distortion of content, aimed to evade automatic filtering. Salting is usually found in spam emails. Salting can also be hidden in phishing emails, which aim to steal personal information from users. We present a novel method that detects hidden salting tricks as visual anomalies in text. We solely use these salting tricks to successfully classify emails as phishing (F-measure >90%).

1 Introduction

Given a text and a user who reads this text, *hidden text salting* is any modification of text content that cannot be seen by the user, e.g., text written in invisible colour, or in zero font size. Hidden text salting can be applied to any medium and content genre, e.g. emails or MMS messages, and can be common in fraudulent *phishing* emails [3]. We present a novel method for detecting hidden text salting and using it to recognise phishing emails.

Related research has focused on filtering email spam. Early spam filters used human-coded ad-hoc rules, often optimized by machine learning, e.g. spamassassin. Such filters were easy to fool, hard to maintain, and outperformed by filters using visual features, e.g. embedded text in images [4]. Recently, statistical data compression has been used to build separate models for compressed ham and spam, and then classify emails according to which model they fit better when compressed [2]. None of these studies addresses hidden salting directly.

2 Hidden Text Salting Detection

Given an email as input, a text production process, e.g. a Web browser, creates an internal parsed representation of the email text and drives the rendering of that representation onto some output medium, e.g. a browser window. We tap into this rendering process to detect hidden content (= manifestations of salting).

[*] Jan De Beer is no longer at K.U.Leuven (jan.debeer@be.ibm.com).

Methodology: We intercept requests for drawing text primitives, and build an internal representation of the characters that appear on the screen. This representation is a list of attributed *glyphs* (positioned shapes of individual characters). Then, we test for *glyph visibility* (are glyphs seen by the user?) according to these conditions: (1) *clipping:* glyph drawn within the bounds of the drawing clip, which is a type of 'spatial mask'; (2) *concealment:* glyph not concealed by other shapes; (3) *font colour:* glyph's colour contrasts with the background colour; (4) *glyph size:* large enough glyph size and shape. We compute a visibility score for each feature and consolidate their product into a single confidence score, parameterised by an empirically-tuned penalty factor. The lower the final glyph visibility score, the stronger the indication of hidden text salting.
Evaluation: We use the above salting tricks as features for classifying emails as ham or phishing in a real-life corpus that contains 16,364 ham and 3,636 phishing emails from 04/2007-11/2007. The corpus is protected by non-disclosure privacy-preserving terms. We use a standard Support Vector Machine (SVM) classifier with 10-fold cross validation. We obtain 96.46% precision, 86.26% recall, and 91.07% F-measure. The best classification feature, found in 86% of all phishing emails, is font colour. State-of-the-art phishing classification reaches F-measures of 97.6% (with random forests [3]) up to 99.4% (with SVMs [1]) when using known discriminative features, such as url length & longevity, HTML & Javascript information on the 2002-2003 spamassassin corpus & a public phishing corpus (these corpora are described in [3]). We use **only salting features**. If we also combine these known discriminative features, performance may improve.

3 Conclusions

We detect hidden text salting in emails as hidden visual anomalies in text, unlike existing methods which target spam in general. We show that hidden text salting is used in phishing emails, and that phishing emails can be identified based on hidden text salting features alone. Our method can be used as improved content representation in filtering, retrieval or mining.

Acknowledgments Partly funded by project EU-FP6-IST-027600-ANTIPHISH.

References

1. A. Bergholz, G. Paass, F. Reichartz, S. Strobel, and J.-H. Chang. Improved phishing detection using model-based features. *Conf. on Email and Anti-Spam (CEAS)*, 2008.
2. A. Bratko, G. Cormack, B. Filipic, T. Lynam, and B. Zupan. Spam filtering using statistical data compression models. *J. of Mach. Learn. Res.*, 7:2673–2698, 2006.
3. I. Fette, N. Sadeh, and A. Tomasic. Learning to detect phishing emails. In *International World Wide Web Conference (WWW)*, pages 649–656, 2007.
4. G. Fumera, I. Pillai, and F. Roli. Spam filtering based on the analysis of text information embedded into images. *J. of Mach. Learn. Res.*, 7:2699–2720, 2006.
5. E. Kirda and C. Kruegel. Protecting users against phishing attacks. *The Computer Journal*, 49(5):554–561, 2006.