Short Paper: Addressing Sophisticated Email Attacks

Markus Jakobsson

Agari
San Mateo, CA
mjakobsson@agari.com

Abstract. We argue that as email attacks continue to increase in sophistication, error rates and filter processing times are both likely to increase. We address the problem at its root by introducing the notion of *open quarantine*, an approach that avoids tradeoffs between filtering precision and delivery delays. This is achieved using a multi-phase filtering approach, combined with the neutralization of messages with undetermined security posture.

Keywords: email, error rate, nation-state attacks, social engineering

1 Introduction

Just ten years ago, Internet security abuses were almost synonymous with small-time crime, whether involving poorly spelled email messages used in attempts to steal banking credentials or computer viruses used to send Viagra spam to millions of consumers.

The threat is very different these days, and points in the direction of dramatically increased attacker sophistication. This increase can be tracked and predicted by observing techniques used in nation-state sponsored attacks, such as recent politically motivated attacks, as techniques developed for or perfected in nation-state attacks are commonly re-used to attack enterprise targets and—in some cases—individuals.

While early politically motivated cyberattacks focused on *disruption*—whether related to the Internet [2], the power grid [5] or the operation of uranium centrifuges [14]—a more recent breed of politically motivated attacks have instead aimed at extraction of sensitive information [1, 3, 6, 19]. Another form of attack based on extraction focuses on *funds* instead of *information*; an example of this is the 2016 attacks on banks using Swift, epitomized by the heist on Bangladesh Bank [17]. This attack straddled the fence between politics and profit by transferring massive amounts of funds to a politically ostracized regime.

Interestingly, while the sophistication of attacks has shot through the roof as groups sponsored by nation states have entered the playing field, the principal attack vectors have remained much the same. Namely, most of the attacks described above involved malware, and most used deceptive emails—commonly

for delivering Trojans, sometimes for stealing credentials. This paper focuses on the use of email as an attack vector.

Deceptive emails are used by cyberattackers to carry out three different types of attacks: (1) to coerce the recipient to follow a hyperlink to a website masquerading as a trusted site, where the recipient's login credentials are requested; (2) to compel the recipient to install malware – whether by opening a malicious attachment or visiting a malicious website; and (3) to convince the recipient to surrender sensitive information or willingly transmit money to the attacker. To succeed with their deception, attackers masquerade as parties trusted by their intended victims; use social engineering laden messages; and, occasionally, hyperlinks or attachments that pose danger to users.

In contrast to traditional phishing attacks and typical spam, the detection of deceptive emails cannot be done in ways that leverage large volumes of identical or near-identical unwanted messages, disreputable senders, or keywords indicative of abuse. This is because cyberattacks typically are *targeted*. They use customized messages, senders and hyperlinks without bad reputation, and—to the extent that they contain malware attachments— individually repacked malware instances that avoid triggering signature-based anti-virus filters.

The analysis of messages with the goal of identifying targeted attacks, accordingly, is time consuming. Diligent scrutiny can easily take minutes of computational effort for difficult emails, and the time is expected to increase as more rules are added to address the mushrooming of new attacks and the increased sophistication likely to be seen onwards. Particularly subtle forms of deceit may require human-assisted review to detect, further adding to the worst-case delivery delays. Without meticulous screening, of course, we expect to see either false positives or false negatives to increase—or, potentially, both of these.

The delays caused by filtering—and the associated fears of lost messages—may very well become the greatest liability when it comes to deploying strong security against targeted attacks. This is due to the resistance among decision makers to accept security methods that have the potential of introducing noticeable delivery delays or, worse still, causing false positives. Given the relatively low commonality of targeted attacks and a widespread hubris among end users as it comes to being able to identify threats, this reluctance is understandable.

This paper addresses the intrinsic tradeoffs between false positives, false negatives and delivery delays by introducing a new filtering paradigm that we refer to as *open quarantine*. Open quarantine balances the needs of security and usability using a two-phase filter process. In the first phase, a risk score is computed for each incoming message. Messages with a risk score corresponding to near-certainty *malice* (e.g., those containing known malware attachments) are *blocked*, and messages with a risk score corresponding to a near-certainty *benevolence* (e.g., messages from trusted parties, with no risky contents) are *delivered*. The remainder—which comprises on the order of 1% of the traffic volume for typical organizations—will be subject to careful scrutiny carried out in a second phase. The power of open quarantine is that the undetermined emails will not be kept out of the inbox of the recipient as they are being

subjected to additional scrutiny. Instead, they will be neutralized and delivered. The neutralization limits the functionality of the email but allows the recipient to access non-risky components while the second-phase filtering is performed. After the second phase of filtering concludes, the neutralization will be reverted (for safe emails) or a blocking action will be carried out.

Open quarantine enables additional security measures that were not practically meaningful in a world where filtering decisions need to be made within milliseconds. For example, consider an email received from a trusted sender, e.g., a party with whom the recipient has communicated extensively in the past. Under normal circumstances, this would be considered safe. However, if the email contains high-risk content, such as apparent wiring instructions, and the sender does not have a DMARC reject policy, then this poses an uncomfortable risk since the email may have been spoofed. To address this potential threat, the receiver's system can send an automated message to the apparent sender[1], asking this party to confirm having sent the email by clicking on a link or replying to the message. If an affirmative user response is received then this is evidence that the email was not spoofed, as an attacker that spoofs emails would not receive the confirmation request.

Outline. After reviewing the related work (section 2), we describe open quarantine, providing detailed examples of the filters to be used in the two phases (section 3). We then turn to the user experience, describing example neutralization methods (section 4).

2 Related Work

Our focus is on addressing fraudulent email. DMARC [20], which combines DKIM and SPF, has done a terrific job addressing one type of fraudulent mail, namely *spoofed* emails. However, it does not address abuse using look-alike domains, display name attacks or corrupted accounts, nor does it protect an organization against malicious incoming email as much as it protects it against abuse of its brand. This paper considers the threat of fraudulent email from the perspective of the *receiving* organization, as opposed to the *impersonated* organization.

The use of social engineering in cyberattacks is on the rise [7,10], and has long been known that the use of targeting increases an attacker's yield dramatically [9]. Publicly available resources, including social network services, can be used by criminals to improve the success of targeting [8]. In addition to being part of the recipe of many of the attacks described in the introduction, the confluence of email-borne social engineering and Trojans has recently resulted in a rapid rise of ransomware attacks [15].

A problem of growing importance is the attack of personal accounts of users belonging to targeted organizations; this is known to have taken place, for example, in the attacks on the DNC [1, 6]. This is made easier as a result of large-scale

[1] Note, however, that the confirmation request would *not* be sent to a potential reply-to address.

breaches (e.g., [18]) and using clustering of identities [4]. One of the reasons for the increasing prominence of this attack is that it is mounted outside the security perimeter of the targeted organization, and as such, circumvents traditional detection methods. We show how open quarantine enables the validation of high-risk messages coming from personal accounts.

Another problem is that, increasingly, sophisticated attacks rely on custom messages and, to the extent malware is employed, custom-packed Trojans. This complicates automated analysis, sometimes requiring manual review of contents to make security determinations. This is an approach that has been started to be tested in a handful organizations (e.g., [16]). While promising, it is an approach that causes longer processing times. Consequently, manual review is impractical for the traditional email delivery paradigm, as it requires quarantine in order to offer security improvements. The use of open quarantine enables increased use of manual review without imposing delays.

Traditional wisdom has that there is a tradeoff between false positives and false negatives where ROC curves are defined in the context of a limited amount of processing. This means that the maximum tolerable to delivery delay defines the ROC curve in the context of a particular problem and filter technology. Our approach shows that these constraints can be escaped by the introduction of temporary neutralization methods applied to messages of uncertain security posture, and a user experience designed to convey potential risk.

3 Open Quarantine

The notion of open quarantine depends on being able to perform a tripartite classification of messages into *good*, *bad* and *undetermined*, where the two first categories have a close to negligible probability of containing misclassified messages. For email delivery, this classification can be done *in flow*, i.e., without any notable delay. One approach uses a scoring, of each incoming email, in terms of its measured *authenticity* (determining the likelihood that it was not spoofed, based on the infrastructure that it originated from); *reputation* (a measure of the past behavior of the sending infrastructure) and *trust* (a measure of previous engagement between the sender and the recipient, and their organizations). More details can be found in the extended version of this paper [12].

The second phase filtering depends on the outcome of the first phase filtering, and may involve in-depth database lookups; manual review; automated messaging to the apparent sender; and more. We will provide details around three of these filtering actions to clarify the approach:

High Risk of Spoofing. While DMARC deployment is on the rise, there is far from universal deployment of this de-factor standard. As a result, email spoofing is still a reality organizations have to deal with . Roughly half of all attempts to pose as somebody else involve spoofing. For emails that the first-phase review identify as undetermined due to a low authenticity score, more thorough scrutiny should be performed.

Automated analysis can identify senders that are particularly vulnerable to spoofing attacks, as DMARC records are publicly available. This corresponds to email from senders whose organizations do not have a DMARC reject policy in place. Messages that are at high risk of having been spoofed can be validated by generating an automated message for the apparent sender, requesting a confirmation that he or she sent the message. If an affirmative reaction to this message is observed, the initial message is classified as good; if a negative reaction is received, it is classified as bad. Heuristics can be used how to classify messages resulting in no response after a set time has elapsed; for example, a message with a reply-to address not previously associated with the sender, or containing high-risk content, could be classified as spoofed if there is no affirmative reaction within ten minutes of the transmission of the automated validation request.

High Risk of Impersonation. The first phase filtering may indicate a higher than normal risk for impersonation. Consider, for example, an email is received from a sender that is neither trusted by the recipient or her organization, nor has a good reputation in general, but for which the display name is similar to the display name of a trusted party or a party with high reputation (see, e.g., [10]). This, by itself, is not a guarantee that the email is malicious, of course. Therefore, additional scrutiny of the message is beneficial.

Automated analysis can be used to identify some common benevolent and malicious cases. One common benevolent case involves a sender for which the display name and user name match[2], and where the sender's domain is one for which account creation is controlled[3]. A common malevolent case corresponds to a newly created domain, and especially if the domain is similar to the domain of the trusted user to which the sender's display name is similar. There are additional heuristic rules that are useful to identify likely benevolent and malevolent cases. However, a large portion of display names and user names do not match any of these common cases—whether the message is good or bad—for these, manual review of the message contents can be used to help make a determination.

Another helpful approach is to send an automated request to the trusted party whose name is matches the sender's name, asking to confirm whether the email from the new identity was sent by him or her. For example, the request may say *"Recently, <recipient> received an email from a sender with a similar name to yours. If you just sent that email, please click on the link below and copy in the subject line of the email and click submit. Doing this will cause your email to be immediately delivered, and fast-track the delivery of future emails sent from the account."*

High Risk of Account Take-Over. The first phase filtering may indicate a higher than normal risk for an account take-over of the account of the sender.

[2] This does not mean a character-by-character equivalence, but rather, a match according to one of the common user name conventions.

[3] This corresponds to typical enterprise, government and university accounts, for example, but not to typical webmail accounts or domains that may have been created by a potential attacker.

For example, one such indication is an email with high trust, authenticity and risk scores—this is an email likely to be sent from the account of a trusted party, but whose content indicates potential danger.

If the source of potential danger is an attachment then this can be scrutinized, including both an anti-virus scan and processing of potential text contents of the attachment to identify high-risk storylines (see, e.g., [13]). Similarly, a suspect URL can be analyzed by automatically visit the site and determine whether it causes automated software downloads, or has a structure indicative of a phishing webpage. The system can also attempt to identify additional indications of risk; for example, by determining if the sender of the suspect email is associated with a recent traffic anomaly: if the sender has communication relationships with a large number of users protected by the system, and an unusual number of these received emails from the sender in the recent past, then this increases the probability of an ATO having taken place. A second-phase risk score is computed using methods like this. If the cumulative risk score falls below a low-risk threshold, then the message is deemed safe, and the second phase concludes. If the cumulative score exceeds a high-risk threshold, then the message is determined to be dangerous, and a protective filter action is taken. If the score is inbetween these two thresholds then additional analysis may be performed. For example, the message can be sent for manual review, potentially after being partially redacted to protect the privacy of the communication. An another approach involves automatically contacting the sender using a second channel (such as SMS) to request a confirmation that the sender intended to send the message. Based on the results of the manual review, the potential response of the sender, and other related results, a filtering decision is made.

4 Recipient User Experience

The user experience of the recipient is closely related to the method of neutralization of messages that are classified as *undetermined*. As soon as a message is identified as undetermined, its primary risk(s) are also identified, and one or more neutralization actions are taken accordingly. Generally speaking, the neutralization may involve a *degradation or modification of functionality* and the *inclusion of warnings*. We provide details on the same three cases described in section 3:

High Risk of Spoofing. A message that is identified in the first phase as being at a higher-than-normal risk of being spoofed can be modified by rewriting the the display name associated with the email with a subtle warning—e.g., replacing "Pat Peterson" with "Claims to be Pat Peterson"—and by inclusion of a warning. An example warning may state *"This email has been identified as potentially being forged, and is currently scrutinized in further detail. This will take no more than 30 minutes. If you need to respond to the message before the scrutiny has completed, please proceed with caution."*. In addition, any potential reply-to address can be rewritten by the system, e.g., by a string that is not an email

address but which acts as a warning: *"You cannot respond to this email until the scrutiny has completed. If you know that this email is legitimate, please ask the sender to confirm its legitimacy by responding to the automatically generated validation message he/she has received. You will then be able to reply."*

High Risk of Impersonation. Emails appearing to be display name attacks can be modified by removing or rewriting the display name, and by adding warnings. These warnings would be different from those for a high-risk spoof message; an example warning is "This sender has a similar name to somebody you have interacted with in the past, but may not be the same person". Alternatively, the recipient can be challenged to classify the source of the email [11] in order to identify situations in which the recipient believes an email comes from a trusted party, whereas it does not.

High Risk of Account Take-Over. Account Take-Overs (ATOs) are often used by attackers to send requests, instructions and attachments to parties who have a trust relationship with the user whose account was compromised. Accordingly, when an email suspected of being the result of an ATO contains any element of this type, the email recipient needs to be protected. One traditional way to do this is to rewrite any URL to point to a proxy; this allows the system to alert the user of risk and to block access without having to rewrite the message. Attachments can be secured in a similar way—namely, by replacing the attachment with an attachment of a proxy website that, when loaded, provides the recipient with a warning and the attachment. Text that is considered high-risk can be partially redacted or augmented with warnings, such as instructions to verify the validity of the message in person, by phone or SMS before acting on it.

In addition, emails with an undetermined security posture can be augmented by control of access to associated material – whether websites, attachments, or *aspects of* attachments (such as a macro for an excel file). All emails with an undetermined security posture can also be visually modified, e.g., by changing the background color of the text. As soon as the second-phase classification of an email has made a determination—whether identifying an email as good or bad—any modifications can be undone and limitations lifted by a replacement of the modified message with an unmodified version in the inbox of the recipient.

References

1. D. Alperovitch. Bears in the Midst: Intrusion into the Democratic National Committee, CrowdStrike Blog, June 15, 2016.
2. N. Anderson. Massive DDoS attacks target Estonia; Russia accused, Arstechnica, May 14, 2007.
3. D. Barrett, D. Yadron, and D. Paletta. U.S. Suspects Hackers in China Breached About four (4) Million People's Records, Officials Say, Wall Street Journal, June 5, 2015.

4. C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan. Mining email social networks. In *Proceedings of the 2006 International Workshop on Mining Software Repositories*, MSR '06, pages 137–143, New York, NY, USA, 2006. ACM.
5. E-ISAC and SANS. Analysis of the Cyber Attack on the Ukrainian Power Grid Defense, March 18, 2016.
6. L. Franceshi-Bicchierai. How Hackers Broke Into John Podesta and Colin Powell's Gmail Accounts, Motherboard, Oct 20, 2016.
7. C. Hadnagy. Social Engineering: The Art of Human Hacking. ISBN-13: 978-0470639535. Wiley, 2010.
8. D. Irani, M. Balduzzi, D. Balzarotti, E. Kirda, and C. Pu. Reverse social engineering attacks in online social networks. In *Proceedings of the 8th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, DIMVA'11, pages 55–74, Berlin, Heidelberg, 2011. Springer-Verlag.
9. T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, Oct. 2007.
10. M. Jakobsson. Understanding Social Engineering Based Scams, ISBN 978-1-4939-6457-4. Springer Verlag, 2016.
11. M. Jakobsson. User Trust Assessment: A New Approach to Combat Deception. In *STAST*, 2016.
12. M. Jakobsson. Addressing sophisticated email attacks. In *Proceedings of Financial Cryptography*, 2017, Full version of paper at `http://www.markus-jakobsson.com/publications`.
13. M. Jakobsson and W. Leddy. Fighting Today's Targeted Email Scams, IEEE Spectrum, April 2016.
14. D. Kushner. The Real Story of Stuxnet—How Kaspersky Lab tracked down the malware that stymied Iran's nuclear-fuel enrichment program, IEEE Spectrum, Feb 26, 2013.
15. L. Manly, M. Salvador, and A. Maglalang. From RAR to JavaScript: Ransomware Figures in the Fluctuations of Email Attachments, Trendmicro blog, September 22, 2016.
16. N. Olivarez-Giles. To Fight Trolls, Periscope Puts Users in Flash Juries, Wall Street Journal, May 31, 2016.
17. S. Shevchenko. Two Bytes To $951M, BAE Systems Threat Research Blog, April 25, 2016.
18. M. Snider and E. Weise. 500 Million Yahoo accounts breached, USA Today, September 22, 2016.
19. W. Turton. YahooMail Is So Bad That Congress Just Banned It, Gizmodo, May 10, 2016.
20. E. Zwicky, F. Martin, E. Lear, T. Draegen, and K. Andersen. Interoperability Issues Between DMARC and Indirect Email Flows. Internet-Draft draft-ietf-dmarc-interoperability-18, Internet Engineering Task Force, Sept. 2016. Work in Progress.