Proactive and automated techniques for detecting sophisticated phishing by Kaspersky Security for Microsoft Office 365

kaspersky

How phishing, and its detection, has evolved

Phishing remains one of the most common forms of email traffic-based fraud. You yourself may have experienced an attacker trying to wrangle your personal information under the guise of a well-known brand. The problem is far from new –so why have people still not learned how to deal with it? Because attack mechanisms and defense methods are evolving constantly, and the struggle will continue for as long as attackers keep profiting from these attacks.

Spam and phishing statistics, including common attack scripts, are published by us at Kaspersky on a quarterly basis, and can be found at securelist.com

1. Initially, dictionaries were created manually to detect phishing in email traffic. The dictionaries described the content of messages from attackers. Then heuristic detection methods began to appear - for example, spotting an unprotected protocol (http) in an email link or an email address in the referral part of the URL,

Message format resembling those of well-known brands, etc.

Another area of development has been and remains the analysis of technical message headers that contain information about the sender's IP address, domain, email agent and other signs that help us identify fraudulent senders. Over time, we then started to combine these signs in order to create rules that successfully block phishing.

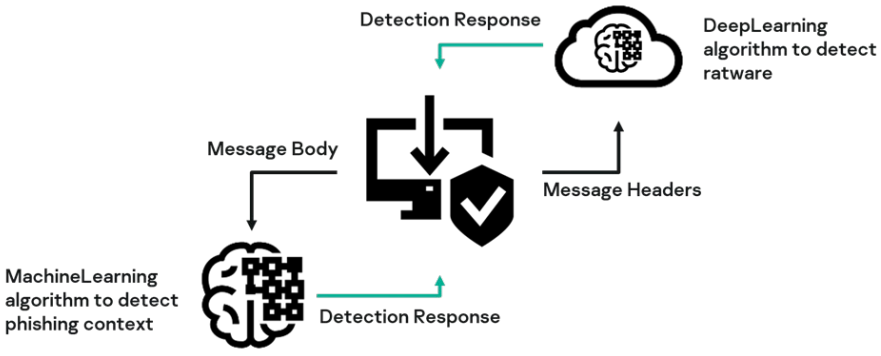
However, with the rapid development of information technology, and its increased importance in people's daily lives, phishing attacks have also continued to develop and evolve. The number of attacks has increased, thanks to a constant stream of new online services, under the guise of which malicious actions can be carried out. The latest hooks (for example, the release of a new TV show a large sports/music event, etc.) can be used to create attack script In addition to new pretexts for creating scripts, the construction of malicious email contents has become more sophisticated, and methods of disguising phishing attempts have become more elaborate. The text and headers have become more varied, phishing can now be easily disguised even on pages with a secure protocol, and emails can be sent using botnets.

We continuously improve our products so we can keep protecting users from even the most cunning and underhanded attacks. We've been accumulating data on phishing mailing lists for a long time, and continue to do so. Thanks to this data, we're able to automatically extract the information needed to block phishing. Machine learning methods enable us to speed up our response time and be more proactive.

Let's now focus on one technology designed to detect email phishing, based on analyses of technical headers and the text of the message.

Technology operating procedure

The diagram below shows the general principle of the technology, which is a combination of two machine learning algorithms. The first classifier is a 'deep learning' algorithm located in a cloud service, where it processes technical message headers in emails using deep neural networks to detect 'ratware' (software for generating and sending messages). The second 'machine learning' based classifier works on the customer's device and identifies phishing vocabulary in the emails. This combination of classifiers provides a solution that can automatically detect and block phishing mailing lists, without creating false positives.



Operating procedure of technology for detecting email phishing.

What's the advantage of this approach?

Firstly, this is a **proactive solution**. The technology's able to detect phishing that has never been seen before. This differs from popular methods based on signature approaches, which must recognize the mailing list as familiar before blocking it.

Secondly, it's a **continuously learning solution**. The technology independently and automatically learns from newly collected statistical data. This enables us to detect phishing mailing lists faster and more accurately, while minimizing false positives.

Thirdly, it **identifies sophisticated patterns**. Thanks to the architecture of the model and the large amount of learning data used, it can extract complex patterns that a human expert would struggle to spot.

How does it work, and why are two classifiers needed?

Before describing the technology in detail, let's look at how sending a message works, and what attackers do to circumvent security systems.

SMTP email transfer scheme.

In the illustration above, the user initially creates a message via a Mail User Agent (MUA) such as Microsoft Outlook. The MUA is responsible for generating the message and sending it to the Mail Transfer Agent (MTA) for further routing. In addition to fields completed manually by the user, such as the message body, subject and recipient addresses, the MUA enters the necessary technical headers, which the email writer and reader won't generally be aware of. These headers contain a lot of information, both about the message itself (date of creation, identification number, type of encoding, etc.) and about its sender (email address, IP address, etc.) and are widely used by email protection technology providers to detect malicious mailing lists.

Ratware software typically obtains legitimate email lists through illegal methods and then uses them to spam the recipient, using a spoofed email address. The final objective is usually to illicitly obtain money or data.

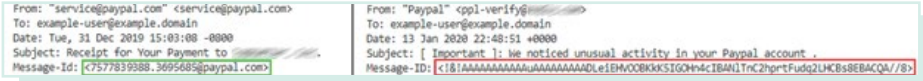
In order to circumvent security systems, attackers often use their own MUAs, which are a type of ratware, so that they can freely create their own emails without encountering any problems. By constantly varying the details in the technical headers, they can achieve maximum diversity in their mailing lists, which makes grouping and blocking them using signatures a complex task.

Classifier No. 1

Based on attacker behavior, the required task is to create a tool that can detect different abnormal indicators in the metadata or technical headers, rather than describing a particular type of mailing list by its signatures. The first classifier focusses on this form of detection.

The classifier is based on deep neural networks that are regularly trained on hundreds of millions of marked-up metadata in order to extract sometimes subtle anomalies. This enables the detection of suspicious technical headers in the message.

This example illustrates how the algorithm works:



On the left is a message from PayPal, and on the right is a fake message. One of the required headers for sending messages is Message-Id, which is a unique letter identifier that has a specific look for different MUAs. The absence of a domain and the random sequence of characters and registers is clear when you look at the differences between the original message from the payment service and the fake one.

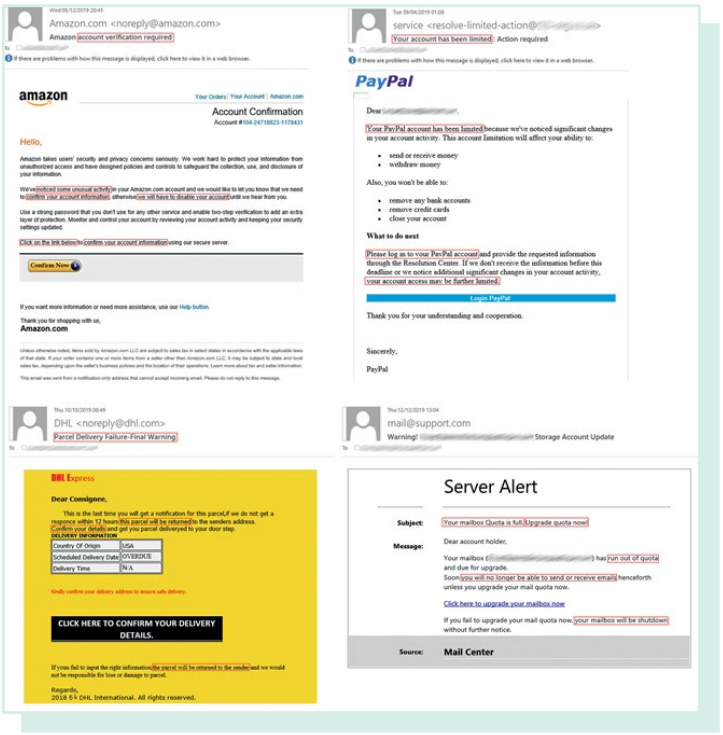
These and other traces can be left by scammers in the technical headers of emails, which the model is able to recognize through its learning, but for which would be difficult to create analytical rules.

Implementing this classifier in a cloud service enables high levels of computing capabilities to be used on the server, reduces the volume of calculations undertaken on the user's computer, and instantly updates the model based on new data.

Classifier No. 2

Phishing attacks play on human emotions. Someone who's concerned about the state of their bank account and waiting for a monthly statement, or for a long-awaited parcel from the foreign online shop is clearly more likely to follow an apparently related link, and fill in all the proposed fields. And most of us are guilty of a tendency to act fast and without too much thought if we're panicked into thinking we're about to lose out on something.

Attackers' text is carefully crafted to achieve the desired emotional effect. For example, calls to action ('Download attachment'' 'Click this link'), notification of a serious problem ('Your parcel couldn't be delivered', 'Your account will be suspended indefinitely') and phrases related to finance ('Reverse this payment', 'View invoice') are often seen in phishing emails. Below are some examples of messages with phishing phrases highlighted:



Unlike spam, the text in phishing messages doesn't contain explicit triggers (for example, phrases from the pharmaceutical industry). The text varies greatly and changes over time, making it difficult to identify them using traditional signature methods, or through recognition of the text alone.

The second classifier – logistic regression – is used to detect these phishing phrases. Logistic regression is a conceptually simple but highly effective method that offers two important benefits: high interpretability and fast learning. The latter allows us to constantly update the model and maintain optimum performance as phishing techniques and gambits change over time.

Conclusion

By combining classifiers that inspect the text content of messages for phishing phrases and those that trawl its metadata in search of unwanted mailing lists, it's possible to create a technology that can detect and confidently block phishing messages in real time. The main advantages of this approach include high and consistent performance quality, the ability to prevent new, previously unseen attacks, and full automation at every stage, leading to optimum efficiency and fast response times.

All the technology described above has been implemented and is already operational in applications included in Kaspersky Security for Mail Server and Kaspersky Security for Microsoft Office 365.