Creating noise: The emerging obfuscation technique designed to evade email security NLP detection capabilities

Our Threat Intelligence team has observed an emerging obfuscation technique, specifically used to make Natural Language Processing (NLP) detection capabilities less effective. Broadly, malicious actors are adding additional characters, break lines, and legitimate links to the end of a phishing email in an attempt to disguise their malicious payloads amongst the noise and evade NLP detection.

For this threat alert, our team analysed 40 emerging attacks identified by Egress Defend that used this technique to understand how it works, why attackers are employing it, and its potential for success.

Of the analysed attacks, the most common legitimate part of an email appended to the attack was the Bank of America email signature, while 'Uber.com' and 'Bofa.com' were the most frequently used legitimate links.

Quick attack summary

- Vector and type: Email phishing
- Technique: Natural Language Processing (NLP) obfuscation
- Targets: Organisations in North America
- Platform: Microsoft 365, layered with Integrated Cloud Email Security (ICES) solutions

In an effort to obfuscate malicious payloads such as links and attachments, threat actors are appending an additional email body to their malicious phishing email. The appended emails are usually harmless and often include benign language that will not trigger NLP detection and legitimate links that are not present on any block list.

The phishing email can be divided into two key components: the malicious content immediately visible to the recipient at the top of the email and the obfuscation element appended at the bottom. This obfuscation element is usually where the benign links and language are present. These two parts are often separated by numerous HTML break lines (empty whitespace) that aim to deter the recipient from scrolling all the way down to notice the obfuscation element. Our Threat Intelligence team has identified only eight instances where fewer than 100 break lines were used, with an average of 157 break lines observed. In addition, cybercriminals are significantly increasing the character count in this part of the email to give NLP more data to process.

Average percentage split of the benign elements present in an email using this technique:

- **Randomised tex**t: 5.93%
- **Graymail:** 62.6%
- **Legitimate email chain**: 31.47%

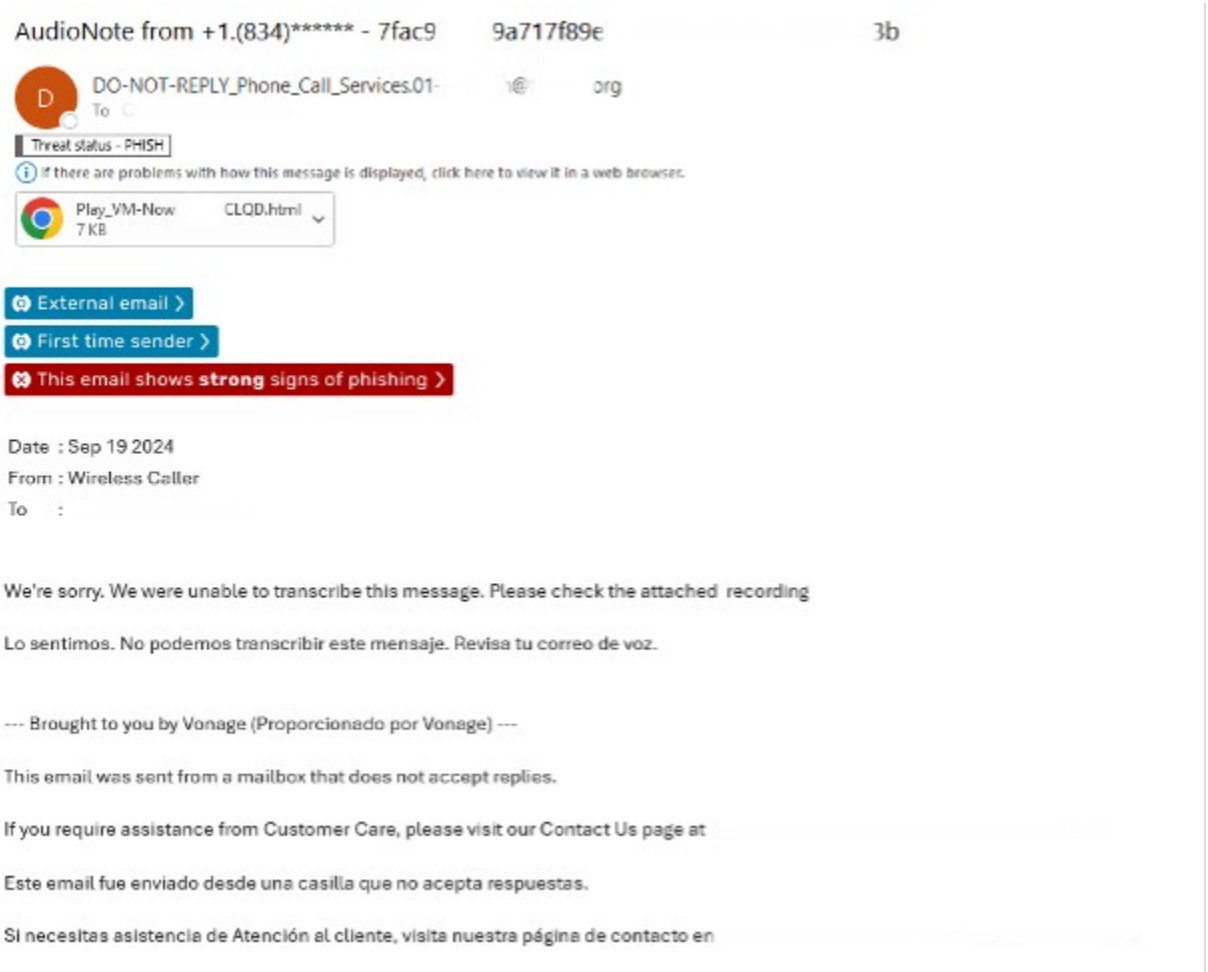
The most common legitimate obfuscation element appended to one of these attacks was the Bank of America email signature.

When attackers have included links in the attack, an average of 4.68 legitimate links were included per attack, compared to just 1.87 malicious links. Our Threat Intelligence team identified ‘Uber.com’ and ‘Bofa.com’ were the most used legitimate links.

What the attack looks like

Example 1

The malicious part

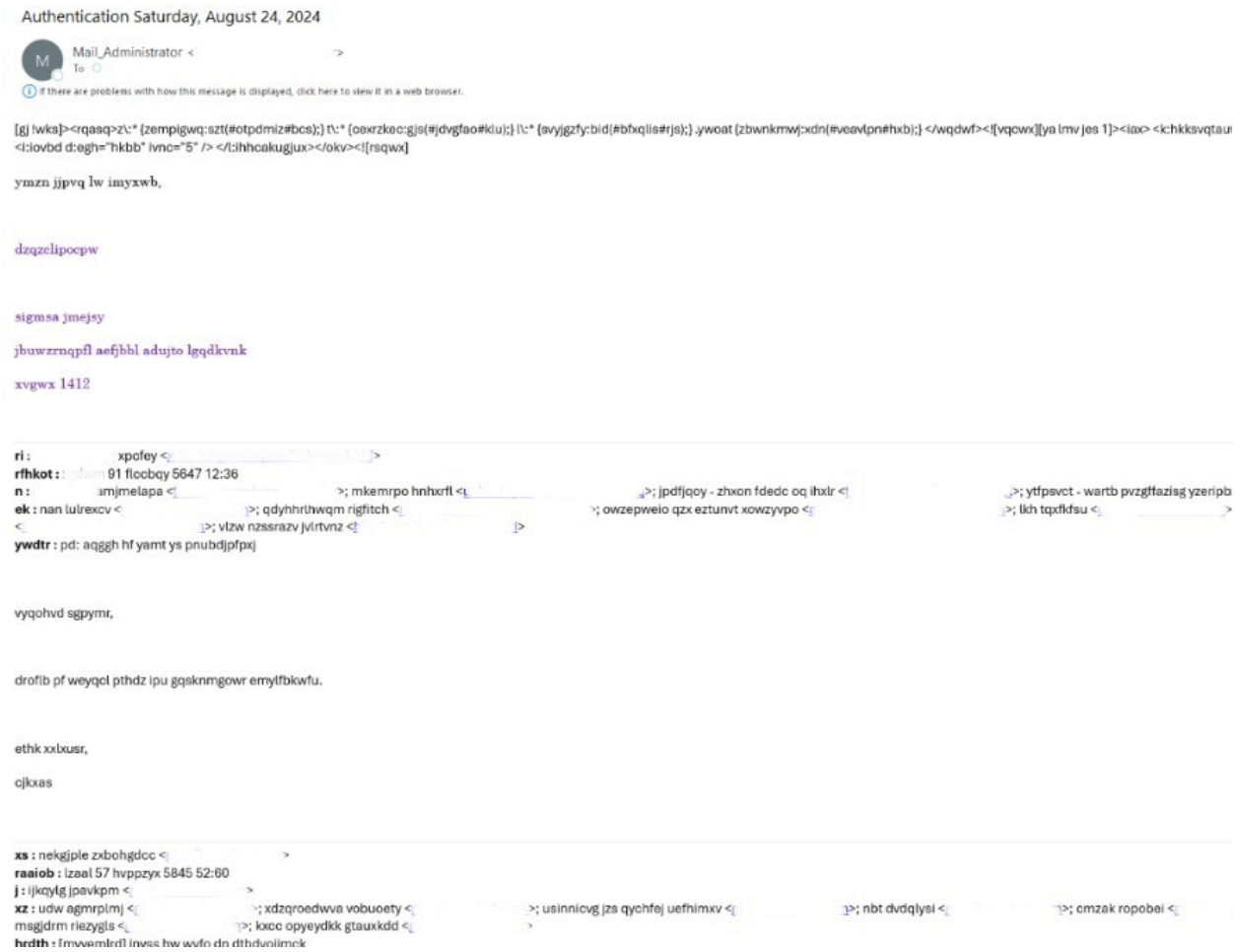


Screenshot of a phishing email impersonating a voicemail service with Egress Defend anti-phishing banners applied.

In the above screenshot, the malicious part of the attack is present. In this case, the email impersonates a voicemail service, prompting the recipient to click on a malicious HTML attachment to listen to the supposed message. Sent as part of a wider phishing campaign,

the attack includes a polymorphic element, with the subject line and attachment name being randomised. This tactic hinders security teams from performing mass manual remediation of emails with the same subject line or conducting a direct search for a specific attachment, as each has been personalised to the recipient.

The obfuscation part



Screenshot showing the lower section of the same phishing email, where the attacker has inserted random characters to mask a malicious attachment from NLP detection.

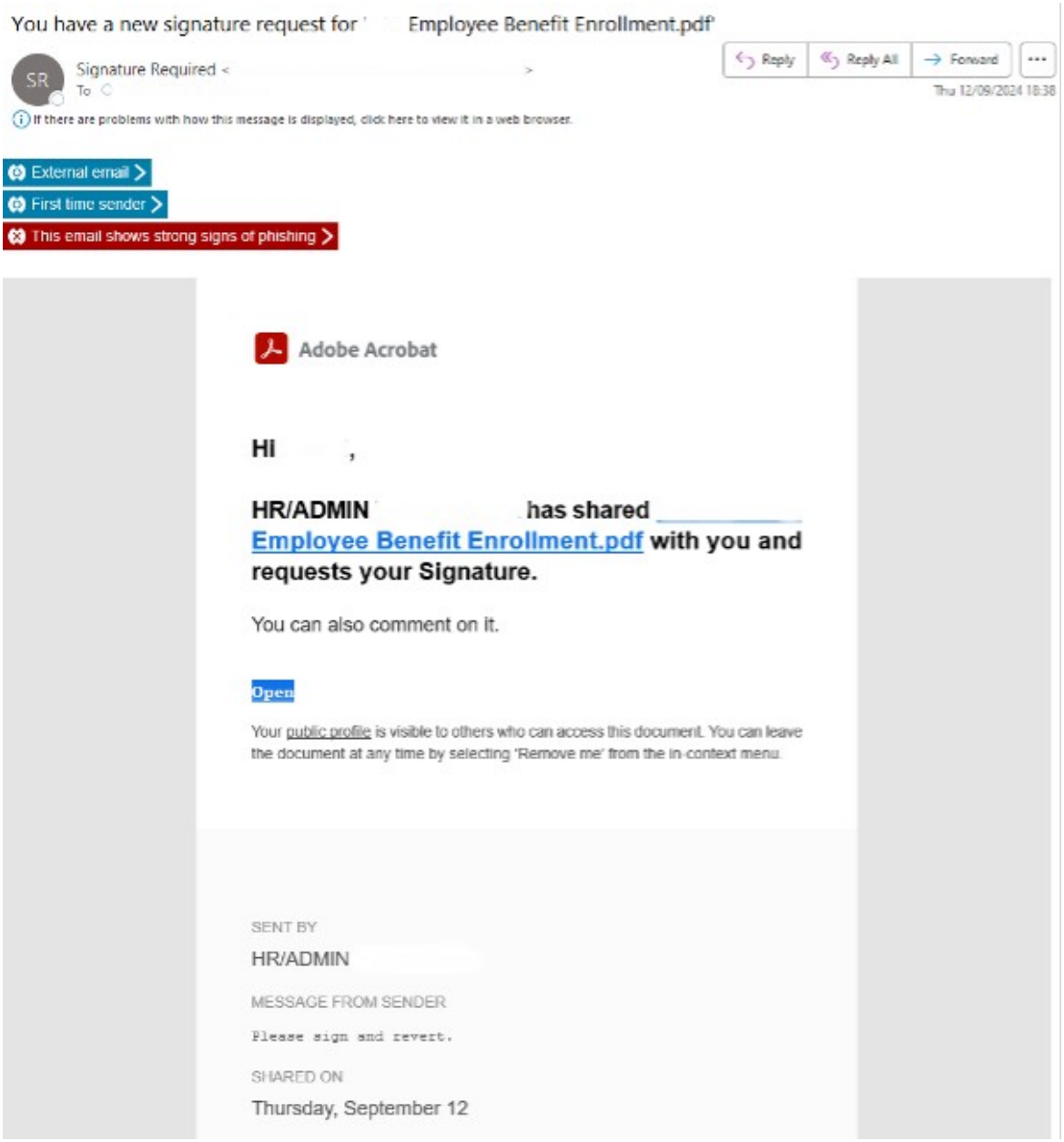
If a recipient scrolls down, they will eventually encounter the second part: the obfuscation text. In this example shown in the screenshot above, the attacker has included random characters in the form of an email chain to increase the character count of the email, ensuring two things:

1) Overall, there are more benign elements for NLP detection to pick up on

[2)] The length of the email has increased. For some email security tools, if an email takes too long to scan, it will be released before the scan is complete, so phishing email can get through without classified as malicious.

Example 2

The malicious part



Screenshot of a phishing email impersonating Adobe with Egress Defend anti-phishing banners applied.

In this second example, the attack impersonates Adobe, a well-known organisation whose software has a high adoption rate among professionals. The email is sent from a

compromised account, posing as the recipient's HR team and urging them to click a malicious link to learn more about employee benefits.

The obfuscation part

Hi Joe,

To ring in the New Year, we're kicking off the winter travel season with savings.

Use Uber Car Hire for your winter escape, and get $15 in Uber Credits when you book any rental vehicle - or if you're going on a longer winter getaway get $30 for 4 days or more. Offer valid on booking made 01/01 through 31/01.

It's that easy to get away this winter.

Rent a vehicle›

*Offer valid for Uber Car Hire reserved before 01/01/2024 and vehicle pickup completed before 30/01/2024, for users who receive this email, have an Uber account, and have a valid United States Driver's License. Uber Car Hire Terms and Conditions apply. $15 in Uber Credits for less than 4 day booking durations and $30 for bookings with a duration of 4 days or more will be applied to your account in up to seven (7) days after the rental company confirms payment and the vehicle is returned. Uber Credits will be issued in Dollars to the user's Uber account and cannot be used in countries with different currencies. Uber Credits cannot be sold, transferred, exchanged, or converted to Credits. This offer cannot be combined with other offers, is non-transferable, and is subject to change or withdrawal at any time and without notice.

Screenshot showing the lower section of the same phishing email, where the attacker has inserted an uber advertisement with legitimate links to mask a malicious link from NLP detection.

Upon scrolling, the recipient would come across what appears to be an Uber advertisement, where the attacks are once again impersonating a well-known brand.

There will be a number of legitimate links such as the ‘rent a vehicle’ hyperlink, and the various pages linked in the sign off.

Egress analysis

A balance of probabilities

The primary objective of this obfuscation technique is to bypass NLP detection. But how do extra characters, benign language, and legitimate links facilitate this? Some NLP solutions operate on a probability scale; if enough benign elements are present to outweigh a single suspicious link or attachment, tools may not classify it as a phishing email with high confidence, and others may not flag it at all.

It is the attackers' hope that, by stacking enough benign elements at the bottom of an email, an NLP tool will generate a general conclusion that the email is safer than it is malicious and deliver it to the recipient’s inbox.

Can attackers outsmart an ICES?

Our Threat Intelligence team suspects that these types of attacks are designed to bypass advanced tools like integrated cloud email security (ICES) anti-phishing solutions, as secure email gateway (SEG) systems do not typically utilise NLP functionality. This suggests that attackers are aware of the shift toward cloud-based email security and are tailoring their tactics based on the technology stack used by their targets.

Identifying advanced phishing threats

While these attacks seem to have been created to bypass ICES solutions, Cybersecurity leaders shouldn’t start reverting to the SEG just yet. As can be seen in both examples above, the attacks have been flagged as high confidence phishing emails by Egress Defend. For the Egress solution specifically, one of the main reasons is because we utilise a zero-trust approach to detect and neutralise emerging threats.

Ultimately, effectively identifying and preventing this type of threat requires a sophisticated tool capable of detecting and neutralising the various techniques employed in each attack, including polymorphic subject lines, impersonation, and account compromise. Egress Defend takes a holistic approach to detection, using AI and a zero-trust approach to detect and neutralise emerging threats like impersonation and zero-day attacks.

-ENDS-