vb 2023
LONDON

4 - 6 October, 2023 / London, United Kingdom

INTENT-BASED APPROACH TO DETECT EMAIL ACCOUNT COMPROMISE

Abhishek Singh & Fahim Abbasi

Cisco, USA & New Zealand

abhisheksingh245@gmail.com
fahabbas@cisco.com

ABSTRACT

Email account compromise is a sophisticated category of BEC scam in which the threat actor may send phishing or scam emails from a compromised account, which can result in a significant loss to the business. Since the email account is compromised, algorithms that use feature sets such as DMARC checks, SPF checks, the difference between the email addresses used in the From and Reply-To header fields, checking if the email is sent from a free email address, look-alike domains, spoofed domains, etc., to detect malicious emails will be bypassed.

In the first part of this paper we will dive into the details of an intent-based approach to detect email account compromise. The design isolates suspicious emails on internal (east-west) and outbound traffic. As per the threat actor's intent, suspicious emails are separated based on keywords derived from n-gram analysis of the body and the subject of emails. Once the suspicious emails have been isolated, the past 90 days' record of the sender is extracted. Features that map to the sender's behaviour from the past 90-day historical record and the suspicious email are extracted. These features are correlated to detect email account compromise.

In the second part of the paper we will share the results of the intent-based approach on the production traffic. We will conclude by comparing the intent-based approach with other approaches to detect email account compromise.

INTRODUCTION

In 2022 the IC3 received 21,832 complaints about BEC/EAC scams, which resulted in a loss of $2.7 billion, making these scams the number one threat. Many previous reports have used the terms 'phishing', 'scam' and 'BEC' interchangeably. For this paper, we will use the following definitions:

- **BECs** are malicious emails with a conversational payload targeting businesses – for example, fraud emails requesting a change of direct deposit information for an employee; emails impersonating a C-level executive requesting money transfers; requests for W-2 forms; requests for aging reports; gift card requests, etc.
- **Scams** are conversational payloads targeting individuals, rather than businesses, and include as romance scams, advanced fee scams, employment scams, etc.
- **Phishing emails** are unsolicited emails that appear to come from a legitimate company, requesting personal, financial, or login credentials.

In the case of email account compromise (EAC), threat actors compromise the email account and then leverage these compromised corporate accounts to send other phishing, scam, BEC and malware emails, both internally and externally, to partners of the organization and other targets.

Since the emails originate internally from legitimate corporate *O365* email accounts, they easily evade defences and pass through existing detection algorithms or signatures focused on email features on the inbound/external interface only, including email authentication controls like SPF, DKIM and DMARC, the difference between email addresses used in the 'From' and 'Reply-to' header fields, checking if the email is sent from a free email address, look-alike domains, spoofed domains, etc., thus making them challenging to detect.

Features to detect email account compromise can be divided into two types. The first set comprises features that detect compromised accounts without analysing email messages.

- The first class consists of algorithms that detect anomalies in login events by leveraging features such as UserId, UserAgent, ClientIp, Operation from *0365* audit log events [4] to detect accounts that have been compromised.

The second set of features detect compromised accounts by analysing emails.

- Approaches leveraging XDR and retrospective phishing verdicts to detect account compromise will also fall under this category. Many email technologies deliver retrospective verdicts, i.e. after the email has been delivered to the end-user, verdicts will be issued, and emails delivered to the end-user will be pulled out from their respective mailboxes if determined to be malicious. It can happen that the end-user has already entered their credentials in the phishing URLs before the email has been pulled out, leading to the generation of a POST request for the URL. For retrospective phishing verdicts for a URL from email technologies, the XDR orchestration platform can validate if a POST request has been generated by the recipients of emails, by inspecting web gateway or endpoint agent logs. If the condition is found to be true, then it can be concluded that the account has been compromised.
- An intent-based approach, detailed in the next section, detects email account compromise (EAC) by isolating suspicious emails from internal and outbound traffic. For these suspicious emails, the sender's behaviour is computed and correlated with features from emails to detect whether the email account has been compromised. This falls into the category of solutions that detect email account compromise by inspecting email messages.

The following section details the design of a system that uses an intent-based approach to detect email account compromise.

DESIGN OF THE SYSTEM

A system to detect email account compromise based on the intent of the threat actor can be divided into four main parts:

1. Intent-based pre-filter
2. URL analyser
3. Retrospective behaviour engine
   i. Recipient analyser
   ii Volumetric stats analyser
4. Verdict correlation engine

The detailed system design is illustrated in Figure 1.

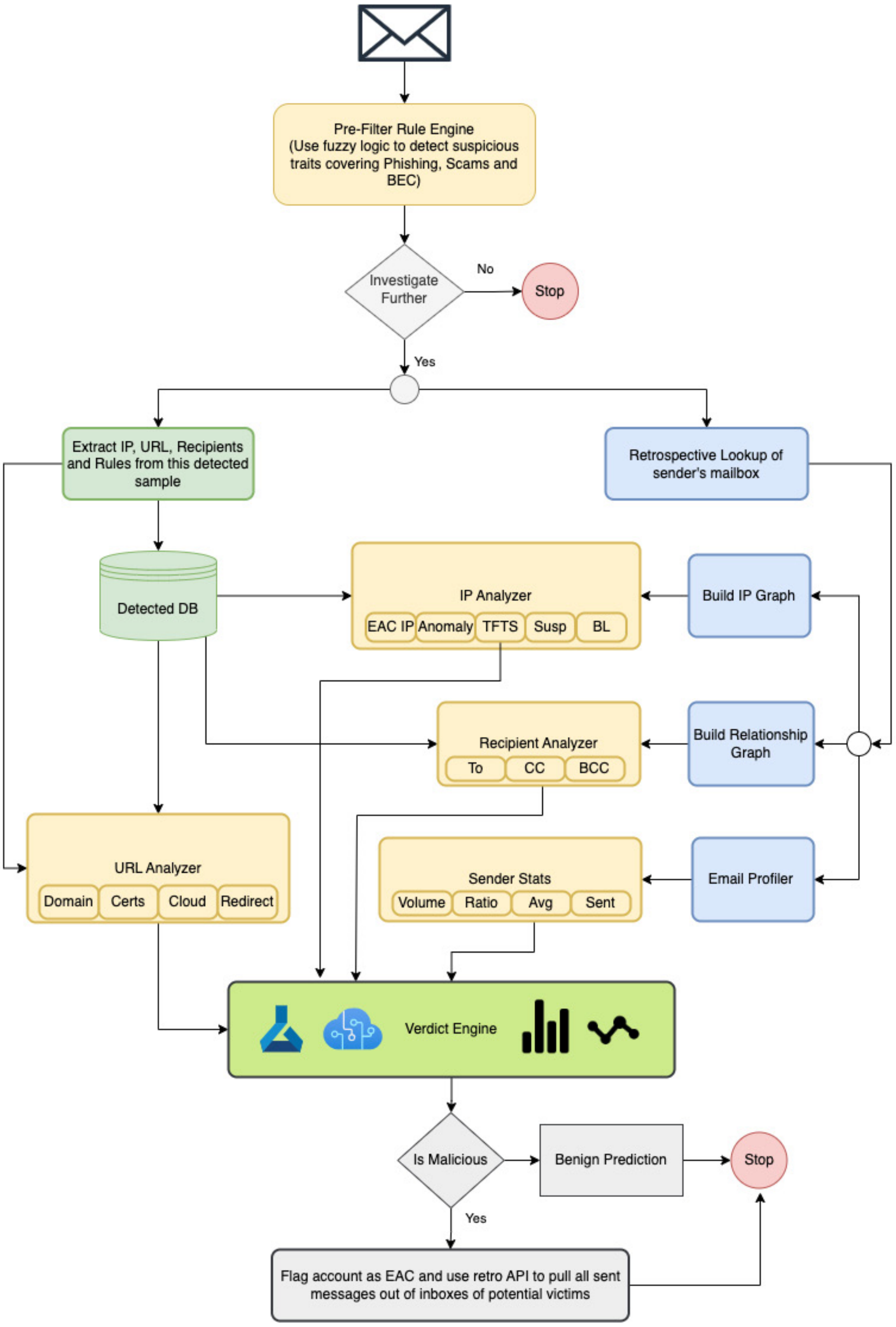

Pre-Filter Rule Engine (Use fuzzy logic to detect suspicious traits covering Phishing, Scams and BEC)
Investigate Further
No
Stop
Yes
Extract IP, URL, Recipients and Rules from this detected sample
Retrospective Lookup of sender's mailbox
Detected DB
IP Analyzer
EAC IP
Anomaly
TFTS
Susp
BL
Build IP Graph
Recipient Analyzer
To
CC
BCC
Build Relationship Graph
URL Analyzer
Domain
Certs
Cloud
Redirect
Sender Stats
Volume
Ratio
Avg
Sent
Email Profiler
Verdict Engine
Is Malicious
Benign Prediction
Stop
Yes
Flag account as EAC and use retro API to pull all sent messages out of inboxes of potential victims

Figure 1: System to detect email account compromise.

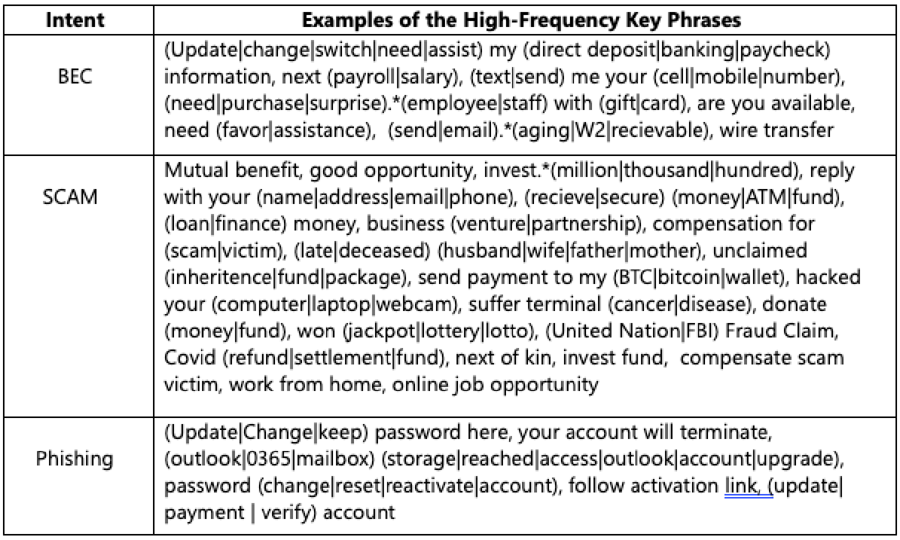
The following subsections describe each of the parts in further detail.

Intent-based pre-filter

Intent-based pre-filters scan every email sent internally or externally (outbound) and match against high-frequency phrases and heuristics used in phishing, BEC, scams (such as fake company, adult and job scams), and emails requesting money transfers with banking details.

Email messages are collected and categorized into their respective datasets for phishing, BEC and scams to create intent-based filters. All messages are analysed for each dataset, and critical phrases are extracted from them using NLP techniques implemented via the Python NLTK library.

All messages in the dataset are read to extract these high-frequency phrases from each spam category, and the body text from these messages is parsed. The text is cleaned by removing characters, stop words, Unicode code, and extra spaces. An email body can comprise multiple phrases. Phrases from each email body are extracted and written out to a file containing all phrases. A 3-gram and 4-gram approach is used to build an n-gram sequence on these phrases using the NLTK library, and a frequency distribution is run on these n-grams to build a sorted list of high-frequency words. Top high-frequency phrases per category are then considered for the intent-based pre-filter. Some high-frequency phrases are illustrated in pseudo-regexes in Figure 2.



Intent	Examples of the High-Frequency Key Phrases
BEC	(Update|change|switch|need|assist) my (direct deposit|banking|paycheck) information, next (payroll|salary), (text|send) me your (cell|mobile|number), (need|purchase|surprise).*(employee|staff) with (gift|card), are you available, need (favor|assistance), (send|email).*(aging|W2|recievable), wire transfer
SCAM	Mutual benefit, good opportunity, invest.*(million|thousand|hundred), reply with your (name|address|email|phone), (recieve|secure) (money|ATM|fund), (loan|finance) money, business (venture|partnership), compensation for (scam|victim), (late|deceased) (husband|wife|father|mother), unclaimed (inheritence|fund|package), send payment to my (BTC|bitcoin|wallet), hacked your (computer|laptop|webcam), suffer terminal (cancer|disease), donate (money|fund), won (jackpot|lottery|lotto), (United Nation|FBI) Fraud Claim, Covid (refund|settlement|fund), next of kin, invest fund, compensate scam victim, work from home, online job opportunity
Phishing	(Update|Change|keep) password here, your account will terminate, (outlook|0365|mailbox) (storage|reached|access|outlook|account|upgrade), password (change|reset|reactivate|account), follow activation link, (update| payment | verify) account

Figure 2: High-frequency critical phrases used in BEC, SCAM, and phishing emails.

With a daily traffic average of around 20 million messages, our intent-based pre-filters successfully select and isolate a very small percentage of highly suspicious emails and forward them to the retrospective behaviour engine and URL analyser. Our pre-filters select approximately 0.000005% (4,000 emails out of 20M) of the traffic daily for further analysis.

Once the pre-filter identifies a suspicious email, it is sent to the retrospective behaviour engine and URL analyser.

URL analyser

As a part of the first step, the URL analyser extracts all the URLs in the suspicious email and checks them against *Cisco*'s 'Umbrella popularity list' of the top 1 million most queried domains based on passive DNS usage [2]. If the domain is present on the list, it is considered to be benign. If the URL is not on the list, its Whois and certificate information is fetched to help determine if it is suspicious. Some of the features from Whois data that can indicate a suspicious URL include the creation time being less than six months ago, the domain of the URL being due to expire in less than a year, the name server being missing or there being no Whois server. Similarly, some of the features from the certificate information that can indicate a suspicious URL include the certificate being due to expire in less than six months, the Host name not matching, or the certificate being issued by a certificate provider known to be used extensively by malware, such as cPanel, Let's encrypt, etc.

Besides the Whois and certificate information, the URL structure is inspected. The structure is examined to check if there is a redirect in the URL or if it is shortened using a URL-shortening service such as *bit[.]ly*, *TinyURL*, *goo.Gl*, etc., or if it is hosted on a file-sharing service like *Google Forms*, *Google Draws*, *DocuSign*, *JotForm*, *Square*, etc., or on a cloud provider. If any of these conditions are found to be true, then the URL is considered suspicious.

Retrospective behaviour engine

The retrospective behaviour engine extracts the sender's email address from the suspicious email selected by the pre-filter email. The sender's email address is used to pull the past x days' historical record of the sender; this information is used for volumetric analysis, recipient analysis, and analysis of the sender's IP address. X has been set to 90 days for each analysis value in the following.

Volumetric analysis

Under this behavioural analysis of the sender, the volume of emails sent by the sender on the day the suspicious email was detected is computed. The past *x* days' record of the sender is extracted and used to calculate the volume of emails sent daily by the sender. From this volumetric data, a volumetric ratio is calculated. Below are the equations which are used to compute the volumetric ratio. For *x*, we use a 90-day window.

$$\boldsymbol{Vratio = \frac{Total\ emails\ sent\ on\ the\ day\ suspicious\ email\ was\ detected}{Average\ number\ of\ emails\ sent}}$$

$$\boldsymbol{Average\ number\ of\ emails\ sent = \frac{Total\ number\ of\ emails\ sent\ in\ the\ past\ X\ days}{X}}$$

Figure 3: Calculation of volumetric analysis.

Recipient analysis

Under this behavioural analysis of the sender, a list of recipients specified in the 'CC:', 'BCC' and 'To:' email header fields in the suspicious email is computed. The past *x* days' record of the sender is extracted and used to calculate the recipient list. The difference between the list containing recipients specified in the suspicious email and the list of recipients with which a person communicates is computed to identify the number of unique new recipients to which the suspicious email was sent.

IP analyser

Under this approach, the originating client IP address in the suspicious email identified by the pre-filter is extracted by parsing `X-Originating-IP` or `X-MS-Exchange-Organization-OriginalClientIPAddress` and then used in anomaly detection, IP profiling, and submitted to a Too Fast Too Soon algorithm to determine if the sender's IP is anomalous. The anomalous IP then acts as a feature set and is correlated with other conditions from the retrospective behavioural engine and URL analysers to detect compromised accounts.

Anomaly detection using GMM clustering

This approach to detecting anomalies in the sender's IP uses the Gaussian Mixture Models (GMM) unsupervised learning algorithm. The sender's IP from the past *x* (= 90) days is extracted from the emails sent in the past *x* days. Six features from the IP data are used as input to the GMM algorithm:

- Each octet of the IP address is used as an individual feature, thus resulting in four features. For example, the IP address 192.168.1.1 would be converted into four features, namely, octet1: 192, octet2: 168, octet3: 1, octet4:1.
- Since disparate IPs can belong to the same ASN, ASN is used as another feature, calculated using the MaxMind database.
- Geo-location of an IP adds some contextual information about an IP address. Hence, the country code is used as the sixth and final feature.

```
features = {"oct1": oct1,
       "oct2": oct2,
       "oct3": oct3,
       "oct4": oct4,
       "ASN": asn_codes,
       "CountryCode": country_codes}
```

The 90-day historical IP dataset is converted into the six features and analysed via PCA. After plotting PCA, the cluster patterns were observed to resemble an ellipsoidal shape; hence Gaussian Mixture Model (GMM) was used since it suits ellipsoidal-shaped clusters. Clusters are formed on this dataset. Similar IPs are grouped in the same cluster.

Let's assume we have four sets of IPs. Three of these IP sets are extracted from historical emails, making up 30 IPs. These IPs are 165.225.8.182, 165.225.62.14, and 178.176.175.23, while the suspicious IP extracted from the suspicious email is 188.162.43.102. The clusters formed by the GMM algorithm on the historical data are shown in the table below:

octet1	octet2	octet3	octet4	ASN	CountryCode	Cluster Number	# of IPs
165	225	8	182	22616	840	1	10
165	225	62	14	22616	840	1	15
178	176	175	23	31133	643	2	5

Since the suspicious IP doesn't share the same features as the clusters above, it resides several percentiles away from the densities or centroids of any of these clusters, thus it is flagged as an anomaly.

The IP address of the suspicious email extracted earlier is converted into a feature set and tested against these clusters to detect whether the tested IP is an anomaly or belongs to a known prior cluster. An anomalous IP contributes to the final verdict to detect a compromised account.

IP profiler

The IP profiler uses the Jaccard similarity score to determine a suspicious IP. The Jaccard similarity coefficient is sent to the verdict correlation engine to detect compromised email accounts.

The IP profiler takes two inputs. The first input is `senders_historical_IPs`, a list structure comprising historical IPs extracted from 90-day historical emails sent by the sender. The second input is `suspicious_email_IP`, the IP of the suspicious email identified by the pre-filter.

For each IP, a lookup against the MaxMind database, like the GeoLite2 City database, is performed to extract country and subdivision information, which is written out as an IP 3-tuple:

`suspicious_email_IP` = {IP_{susp}, $\text{Country}_{\text{susp}}$, $\text{Subdivision}_{\text{susp}}$}

`senders_historical_IPs` = [{IP_{hist1}, $\text{Country}_{\text{hist1}}$, $\text{Subdivision}_{\text{hist1}}$}, {$\text{IP}_{\text{hist2}}$, $\text{Country}_{\text{hist2}}$, $\text{Subdivision}_{\text{hist1}}$}, {IPhist3, $\text{Country}_{\text{hist3}}$, $\text{Subdivision}_{\text{hist3}}$}]

Next, the {IP, Country, Subdivision} suspicious set is compared with the historical background to gauge similarity and diversity in the given set. This is done using the Jaccard similarity coefficient, as illustrated here:

$$JaccardSimilarity_{(susp,\, histx)} = \frac{(susp \cap histx)}{(susp \cup histx)}$$

Jaccard similarity results in a score between 0 and 1, where 0 means highly dissimilar, while 1 means the same. Dissimilar observations suggest an unknown or malicious sender.

The working of this heuristic can be explained further with an example. Let’s assume a corporate user has historically sent emails from the IP address 216.248.X.Y. For this IP, a MaxMind database lookup is performed to determine the country and subdivision of this IP. This IP is a US IP and shows Texas as its subdivision. Based on this, our historical IP 3-tuple for this user would be ‘John.doe@acme.edu’: (‘216.248.X.1’,’ US,’ ‘TX’).

Similarly, a 3-tuple for the sender’s IP for the suspicious email is analysed. If it is a Russian IP from Moscow, the IP 3-tuple for ‘John.doe@acme.edu’ will be (‘188.162.43.102’, ‘RU,’ ‘ME’). The Jaccard similarity score is calculated by comparing each instance of the historical IP 3-tuple with the suspicious email’s IP 3-tuple.

An example of such a comparison is illustrated here:

JaccardSimilarity = (‘188.162.43.102’, ‘RU’,’ ME’)∩(‘216.248.X.1’,’ US’, ‘TX’) /
(‘188.162.43.102’, ‘RU’,’ ME’)∪(‘216.248.X.1’,’ US’, ‘TX’)

JaccardSimilarity = 0.0

The final verdict is calculated by applying thresholds to the Jaccard similarity output. After thresholding, the Jaccard similarity score is classed as follows:

Jaccard score	**Inference**
jaccard_similarity_score == 1.0	Benign user has been seen in the past using same IP
jaccard_similarity_score == 0.5	Suspicious user from the same country and subdivision but different IP
jaccard_similarity_score == 0.2	Suspicious user from the same country but different subdivision and IP
jaccard_similarity_score == 0.0	Malicious user, possibly compromised account

IP reputation service

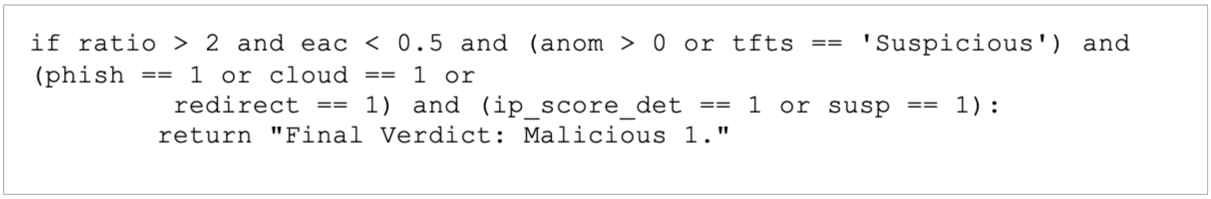
The IP address of the email sender is extracted from the suspicious email and tested against an internal IP reputation service. Bad reputation of the IP address contributes to a malicious signal for the final verdict.

Too Fast Too Soon

Another algorithm that is part of the IP analysers is the Too Fast Too Soon algorithm. This algorithm takes as input the email sender’s historical IP and time information for all emails sent in the past until the suspicious email. The algorithm then loads the data in a time series and determines how frequently the sender changed IPs. If the IP changes during short intervals, the algorithm looks for other metrics like the geographical distance between the IPs and the time it takes to commute between these locations. If a sender’s IP is frequently changing over short time intervals, the switched IP does not belong to the same ASN, and the distance between the IP locations is greater than what could conceivably be covered by a road or air commute, then a malicious flag is raised, suggesting a spambot-like activity.

Verdict correlation engine

A verdict correlation engine is a rule-based expert system where the results from the volumetric analysis, recipient analysis, IP analysis, and URL analysis are correlated to produce a verdict as malicious or benign. Figure 4 shows an example of one such rule.



```
if ratio > 2 and eac < 0.5 and (anom > 0 or tfts == 'Suspicious') and
(phish == 1 or cloud == 1 or
         redirect == 1) and (ip_score_det == 1 or susp == 1):
        return "Final Verdict: Malicious 1."
```

Figure 4: Example of a correlation rule.

The first condition of this rule checks if the email sender's ratio is more significant than two, indicating that today the user has sent twice as many emails as in the past. The second condition checks whether the sender's geo-location and IP address from the suspicious email differs from the past 90 days. This value is the Jaccard score from the IP profiler.

The third condition is whether the IP's anomalous score from GMM is greater than 0 or the IP is fast fluxing between subsequent emails.

The fourth condition stems from whether the extracted URL in the email is a phishing link or belongs to a cloud service or a redirect service. The fifth and final condition checks whether the reputation of the sender's IP address is suspicious. If there is a match in all five conditions a verdict of a compromised email account is given.

Similarly, the rule-based expert system has many rules that correlate the outputs from the URL analyser and behavioural engine to detect a compromised account. An AI model, such as a decision tree, can be used instead of a rule-based expert system to determine a compromised email account. This is ongoing work.

RESULT

Our intent-based approach to detect email account compromise is currently detecting compromised accounts in production. The results and data discussed in this section are based on a dataset and observations from March 2023 to June 2023. This dataset contains 20 million emails processed per day from around 1 million mailboxes for the tenure between March 2023 to June 2023. Geographically, the traffic is US and Canada-centric. The organizations represented range from government/public sector to manufacturing, real estate, construction, technology, and education. These organizations use *Microsoft Office 365* as their email provider.

Traffic originating from compromised accounts is an exceedingly small fraction of an organization's overall email traffic; this dataset reflects this skewed and unbalanced property, with traffic originating from compromised accounts measuring approximately 0.000005% (4,000/20M) of the daily email traffic. It is important to note that our pre-filters select about 4,000 suspicious emails daily for further investigation. These emails are chosen from all the email traffic sent internally and externally.

The following definitions have been used for false positives, false negatives, and true positives:

- A true positive observation is a compromised email account detected by the emails it sends to targets, internal or external to the organization.
- A false positive observation is a benign email account flagged as compromised by the preventive feature based on their emails.
- A false negative observation is an email account that is missed or failed to be marked as compromised.

Our dataset had 119 positive compromised accounts. Preventive features alerted 150 individual accounts across 1 million mailboxes. Out of these 150 accounts, a total of 104 accounts were true positives ((104/150) *100 = 70% TP), and 46 were false positives ((46/150) *100 = 30% FP). False positive detections were fixed by updating pre-filter rules and fine-tuning rules in the verdict correlation engine. Around 1.6 alerts were raised daily, and approximately 0.92 true positive compromised accounts were detected across customers. The low value of false positives per day from the system makes it easy for SOC analysts to validate alerts. During the four months from March 2023 to June 2023, emails from compromised accounts from customer submissions showed that 15 individual compromised accounts were missed and were flagged as false negatives ((15/(104 + 15)) *100)= 12% FN). Here, it is assumed that customer-reported false negatives, i.e. 15 accounts, were the only compromised accounts missed by the preventive feature.

Threat actor's intent

Scam emails accounted for most of the follow-on activity from the captured data set. They were usually sent out in large volumes internally, or externally to other institutions. Most of the follow-on activity from compromised accounts involved unsophisticated 'spray-and-pray' attacks (nearly 80 per cent), where a threat actor sends many emails.

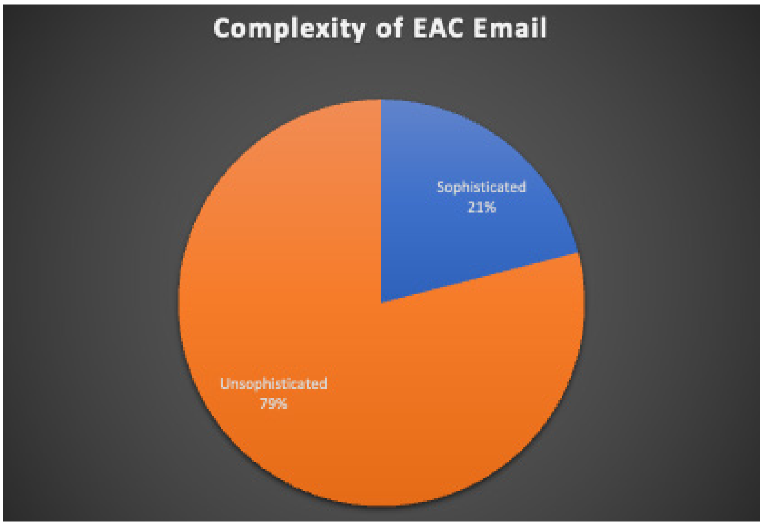
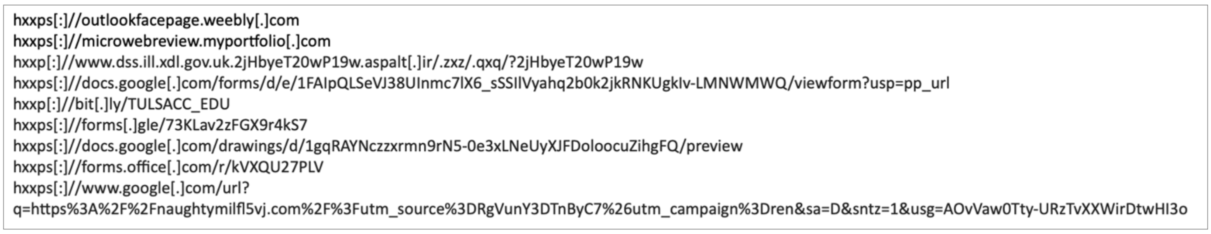


Figure 5: Distribution of attacks from compromised accounts.

The next most common activity was internal phishing, which was used to move laterally by compromising additional accounts. Many of the phishing emails included links designed to evade detection by using legitimate file-sharing services like *DocuSign* and *JotForm*, URL shorteners like *bit[.]ly*, redirects through web browsers like *Google* or *Bing*, data collection services like *Microsoft Forms*, and free web-hosting services like *Weebly*.



```
hxxps[:]//outlookfacepage.weebly[.]com
hxxps[:]//microwebreview.myportfolio[.]com
hxxp[:]//www.dss.ill.xdl.gov.uk.2jHbyeT20wP19w.aspalt[.]ir/.zxz/.qxq/?2jHbyeT20wP19w
hxxps[:]//docs.google[.]com/forms/d/e/1FAIpQLSeVJ38UInmc7lX6_sSSIlVyahq2b0k2jkRNKUgklv-LMNWMWQ/viewform?usp=pp_url
hxxp[:]//bit[.]ly/TULSACC_EDU
hxxps[:]//forms[.]gle/73KLav2zFGX9r4kS7
hxxps[:]//docs.google[.]com/drawings/d/1gqRAYNczzxrmn9rN5-0e3xLNeUyXJFDoloocuZihgFQ/preview
hxxps[:]//forms.office[.]com/r/kVXQU27PLV
hxxps[:]//www.google[.]com/url?
q=https%3A%2F%2Fnaughtymilfl5vj.com%2F%3Futm_source%3DRgVunY3DTnByC7%26utm_campaign%3Dren&sa=D&sntz=1&usg=AOvVaw0Tty-URzTvXXWirDtwHI3o
```

Figure 6: Type of URLs in the emails sent by threat actors from the compromised accounts.

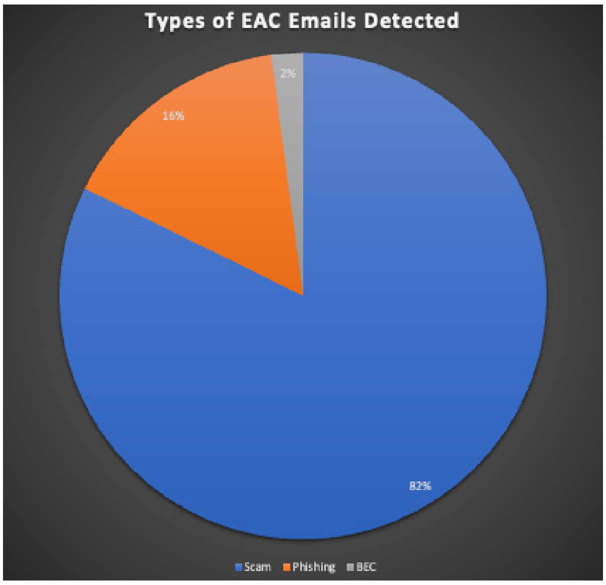
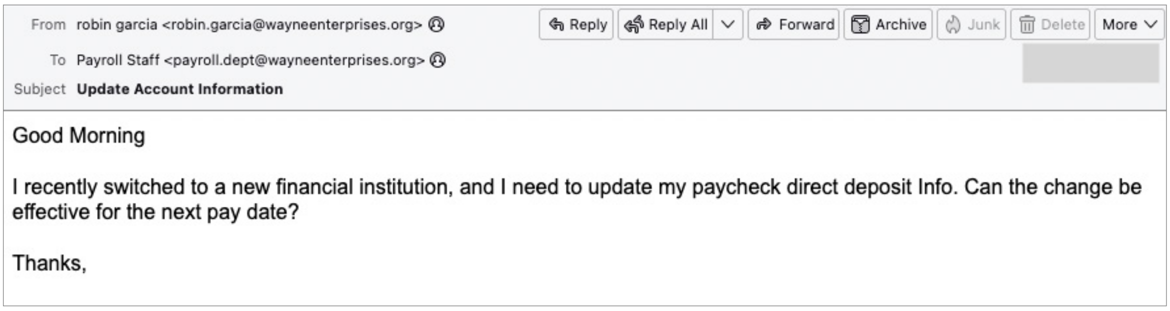


Figure 7: Distribution of attacks from the compromised account per the threat actor's intent.

A few sophisticated and highly targeted BEC emails mainly included payroll-themed lures and were typically sent to business departments involved in finance, such as employee payroll.



From robin garcia <robin.garcia@wayneenterprises.org>

To Payroll Staff <payroll.dept@wayneenterprises.org>

Subject **Update Account Information**

Good Morning

I recently switched to a new financial institution, and I need to update my paycheck direct deposit Info. Can the change be effective for the next pay date?

Thanks,

Figure 8: Example of BEC email.

CONCLUSION

Email account compromise is one of the challenging problems faced by the industry today. There are two sets of victims in email account compromise. The first is the person whose account has been compromised. The second set of victims is the people to whom compromised emails have been sent from the compromised accounts. If the victims have been sent phishing emails, then the passwords of both the compromised account and the recipients to whom the emails have been sent must be reset.

Our intent-based approach to detect email account compromise isolates the suspicious emails based on the threat actor's intent. For these suspicious emails, behavioural analytics determine if the email is sent by the threat actor or from the compromised account. Since the intent-based approach can identify both the compromised account and the set of victims to whom emails have been sent, leading to appropriate remediation, it is the recommended approach.

REFERENCES

[1] FBI Internet Crime report. https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf.

[2] Cisco. Umbrella Popularity List. https://s3-us-west-1.amazonaws.com/umbrella-static/index.html.

[3] Microsoft. Microsoft Graph: message resource type. https: //developer.microsoft.com/en-us/graph/ docs/api-reference/v1.0/resources/message.

[4] Microsoft. 0365 Audit Log Activities. https://learn.microsoft.com/en-us/microsoft-365/compliance/audit-log-detailed-properties?view=o365-worldwide.