IET Information Security

Research Article

Detecting lateral spear phishing attacks in organisations

ISSN 1751-8709
Received on 17th March 2018
Revised 7th October 2018
Accepted on 30th November 2018
E-First on 21st January 2019
doi: 10.1049/iet-ifs.2018.5090
www.ietdl.org

Aniket Bhadane[1] ✉, *Sunil B. Mane*[1]

[1]*Department of Computer Engineering and Information Technology, College of Engineering Pune, India*

✉ *E-mail: aniketbhadane93@gmail.com*

Abstract: Lateral spear phishing attack is a powerful type of social engineering attack carried out using compromised email account(s) within the target organisation. Spear phishing attacks are difficult to detect due to the nature of these attacks. The inclusion of a lateral attack vector makes detection more challenging. The authors present an approach to detect lateral spear phishing attacks in organisations in real-time. Their approach uses features derived from domain knowledge and analysis of characteristics pertaining to such attacks, combined with their scoring technique which works on non-labelled dataset. They evaluate the approach on several years' worth of real-world email dataset collected from volunteers in their institute. They were able to achieve false positive rate of below 1%, and also detected two instances of compromised accounts which were not known earlier. A comparison of their scoring technique with machine learning based anomaly detection techniques shows the proposed technique to be more suited for practical use. The proposed approach is primarily aimed at complementing existing detection techniques on email servers. However, they also developed a Chrome browser extension to demonstrate that such a system can also be used independently by organisations within their network.

1 Introduction

Spear phishing attacks have been in limelight of cyber-attacks, costing organisations and individuals' huge tangible and intangible losses [1]. Lazarus group known for its involvement in WannaCry ransomware attack recently launched a malicious spear phishing campaign with lure as a job opening for CFO role at a cryptocurrency company [2]. With sharp increase in prices of cryptocurrencies, cypto-phishing has seen steep increase since 2017 [3]. Also, recently, sensitive data of up to 30,000 Medicaid recipients was compromised when an employee of Florida's Agency for Health Care Administration fell for a phishing attack. Organisations, be it government or corporate or non-profit, have been compromised by these types of attacks [4]. Such emails also take advantage of current events such as Equifax hack [5].

Spear phishing emails are targeted towards specific individuals or organisations and trick target(s) using social engineering techniques into performing actions that are of benefit to the attacker. The actions could be clicking on a URL in the email and entering credentials on the resulting phishing page; revealing sensitive information about self or organisation; doing wire transfer etc. Adversaries can easily launch spear phishing campaigns with availability of many toolkits to aid the tasks. Phishing-as-a-service [6] has made conducting attacks even more convenient for them. Adversaries also use various mechanisms for extending the lifespan of phishing sites by limiting the visibility of the attack [7]. Defending from such attacks is difficult as they are crafted to appear legitimate. 70% of the spear phishing emails are opened by their targets and 50% of these users are found to click on URLs in the email [1, 8].

Current detection methods often depend on users' reporting the attack. Works dealing with spear phishing attacks have their false positive rates not suitable for organisations [9–11]. Classical learning techniques for anomaly detection face various limitations in detecting such attacks. The low rate of these attacks causes high imbalance in the dataset. Supervised learning techniques do not perform well with such highly imbalanced datasets [9]. Classical unsupervised or semi-supervised techniques do not consider directionality of features' values, treat an event as anomalous even if one or few of the features give anomalous values, and assume underlying distribution of dataset or need parameters to be set correctly to give good performance [9]. Due to these limitations, classical learning techniques do not perform well in detecting such attacks.

A lateral attack involves compromising a node in the target organisation, which gives attackers a foothold in the network. Attackers can then move laterally within the organisation targeting sensitive data such as high level personnel credentials, financial data, intellectual properties etc. Lateral spear phishing is a powerful type of attack in which spear phishing emails are sent to people in the organisation from a compromised email account within the organisation itself. Such types of attack are particularly rare than other types of spear phishing attacks and can prove to be highly expensive for organisations. There is no generic solution for detecting these types of attacks and solutions need to incorporate the domain knowledge of their problem space to get best results [9].

We collected several years of emails from volunteers in College of Engineering Pune (COEP), an autonomous engineering institute in the state of Maharashtra in India. The institute has faced various lateral spear phishing attacks within the last few years. COEP uses corporate Outlook as its email service provider. The lateral spear phishing attacks faced by COEP are primarily credential seeking or trying to establish a conversation to extract sensitive information. We made effort to collect emails such that the dataset contains all known attack instances.

Our goal is to create a practical and deployable technique, which can detect a fair number of lateral spear phishing attacks in real-time, keeping a low false positive rate. Standard machine learning techniques suffer due to the high imbalance of dataset and have their false positive rate impractical for use by organisations. Importantly, intra-domain emails (emails to and from the same domain) do not tend to get blocked as they have a much higher '*trust*' score [12]. We derive a compact set of features from the analysis of lateral spear phishing attacks. Furthermore, our scoring technique leverages domain knowledge of the attacks on the organisation. This also implies that administrators can customise the technique to integrate their domain knowledge pertaining to their organisation. We compare our technique with machine learning-based anomaly detection techniques and show that our technique gives better accuracy and false positive rate than the compared machine learning techniques.

Our technique is mainly aimed to be implemented in email servers providing service to organisations, to complement existing detection techniques. If that is not possible due to plausible reasons, an organisation can implement such a technique within their network, with the help of plugins, to aid users in decision-making. We also developed a Chrome Extension, which uses the API provided by our scoring system, to aid COEP's users in taking better informed decisions when dealing with emails on COEP's Outlook email domain. The major contributions of this work are:

- Usable real-time approach for detection of lateral spear phishing attacks.
- Use of domain knowledge and characteristic features of these attacks, along with our scoring technique, which requires a non-labelled dataset.
- Suitable for use by organisations where intra-domain emails have a much higher trust score, and aggressive heuristics are not used to avoid false positives and false negatives.
- System which is primarily aimed to be implemented on email servers complementing existing detection mechanisms, but can also be implemented independently by organisations if they do not manage their email server.

In Section 2, we discuss about some concepts surrounding these attacks and challenges in detection of such attacks. In Section 3, we provide the characterisation of the dataset that we collected. We describe the features we use and our scoring technique in Section 4. In Section 5, we present the evaluation of our scoring technique and compare its performance with anomaly detection techniques. In Section 6, we provide discussion, limitations of the system, and some details on the Chrome Extension we developed for use by COEP. We describe some of the related work in Section 7 before concluding in Section 8.

2 Concepts and challenges

Lateral spear phishing attack first involves compromise of an email account within the target organisation. This serves as an entry point and foothold in the organisation's network for an attacker. The compromise can be due to several reasons such as brute force, theft etc. On setting the foothold, the attacker can move laterally within the organisation seeking sensitive information. The attacker can also choose to stay passive, with the objective of just scanning the email conversations of the victim to get further sensitive information. The presence of the attacker within the organisation will remain until any suspicious activity is detected. Although remote access is essential for workflow of organisations, it also provides chances for the attackers to get into the organisation's network. Stolen credentials are also sold on dark web from an average of \$8 to maximum of \$190 each [13]. Zhang *et al.* [14] also found that compromised credentials of university accounts were used to access scholarly articles and bypass censorship.

Email account compromise can be detected at two stages:

i. During login: by analysing properties of login activity.
ii. Post-login: by analysing email sending activity.

Our research detects attacks in the second category. We look for suspicious email sending activity to detect account compromise regardless of the way of compromise. Note that we refer to lateral spear phishing attacks as those which are conducted from compromised account within the organisation, not from user's personal email account.

One challenge in dealing with these attacks is that such incidences occur infrequently, which makes this problem similar to finding needle in haystack. However, a successful compromise from these attacks could be devastating for the organisation. Such infrequent occurrence of attacks leads to base-rate issues with imbalanced dataset having a small training set. Due to these issues, the success of the standard machine learning techniques is improbable. Other issues include limited history of users and benign churn in header values, which further degrades the results of standard machine learning techniques [9]. We provide a comparison of our technique with machine learning algorithms and discuss more about their results in Section 5.1.

Table 1 Number of emails in our dataset

total number of emails	113,726
unique emails	19,087
unique intra-domain emails	12,064

Another challenge is that intra-domain emails have a higher trust score and hence an anomalous intra-domain email is not filtered [12]. Phishing emails, which should normally get detected when sent from a personal email account, do not get detected when sent from an email account within the domain, mainly due to the higher trust score. Aggressive heuristics are not used largely as they tend to have higher false positives and false negatives for organisational settings [15] which hamper productivity of the organisation. Legitimate employee-to-employee emails can look a lot like spam (e.g. got my discount, follow this referral link [URL]). This is an issue of ease of use. The sensitivity level of intra-domain filters can be modified by the administrators in the organisation. However, it involves the risk of false positives and needs sufficient testing to actually provide better results without hampering normal flow of business.

Multi-factor authentication can increase cost to attackers for infiltrating the network. However, these authentication techniques suffer from deployability and usability issues. The success of these techniques depends largely on its implementation and is not difficult for attackers to get past weaker implementations [16]. Organisations including universities and national laboratories find it difficult to adopt these authentication techniques in practice [17].

Lateral spear phishing attack can also have an avalanche effect within the organisation, which further amplifies the damage caused by the attack. Taking example of one of our case, a staff's compromised account led to compromise of multiple students' accounts using spear phishing emails. These accounts in turn were used to conduct further attacks and compromise more number of accounts. Furthermore, in the current scenario, organisations largely depend on users to report spear phishing attacks [9]. In the next section, we discuss about our dataset and provide its characterising details.

3 Data characterisation

Our work draws upon the emails collected from volunteers in COEP. COEP uses Office365 Outlook for its emailing needs. Due to privacy and confidentiality reasons, we do not have access to the entire emailing history of our organisation. However, we collected emails from 40 volunteers in the organisation. The volunteers' emails were downloaded and converted to a processable format. The emails contain both their respective header and body. The dataset contains 3.5 years' worth of emails, ranging from July 2014 to January 2018. The total number of emails we collected from 40 volunteers is 113,726. On merging the emails from individual volunteers, our dataset had 19,087 unique emails. Out of these, 12,064 are intra-domain emails. This is recapped in Table 1.

We made effort to have all known attack instances in the dataset even though we do not have all the benign emails of the organisation. The dataset contains entire history of emails of student volunteers from various academic years (First Year BTech, Second Year BTech, etc.) and also from some of the staff members. The types of spear phishing attacks that the attacker(s) carried out after compromising an account in the organisation were either to steal credentials using a phishing site impersonating COEP's site or to establish communication with another target within the organisation to get sensitive information.

Our dataset contains a total of 27 account compromise instances, 25 of which were known earlier and 2 were detected after we test our system with the dataset. The 27 compromise instances correspond to 323 spear phishing emails. Once an account was compromised, it was used to send spear phishing emails to other targets in the organisation. The compromised account access was revoked only after users reported the attack to the administration team. Since the administration team is not aware

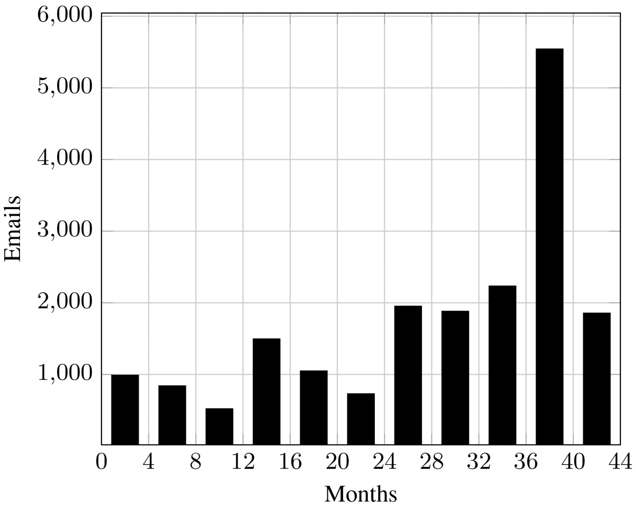


Fig. 1 *Number of unique emails in our dataset with intervals of 4 months*

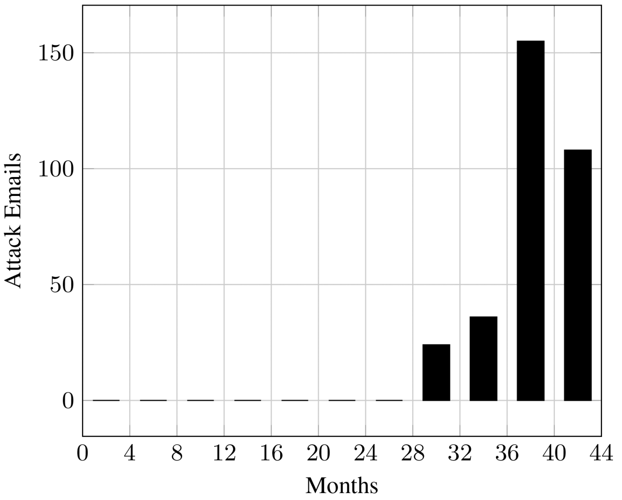


Fig. 2 *Number of attack emails in our dataset with intervals of 4 months*

of any on-going attack, it requires reporting by the users. Some of these spear phishing emails were even sent in bulk to others in the organisation impersonating as helpdesk of the organisation. It is not known how many users have responded to the spear phishing emails given the large number of users that the spear phishing emails were sent to.

Table 2 shows the number of attack instances that contained phishing URLs and the number of attack instances which involved display name spoof. All the phishing websites were impersonating as COEP's site to steal target's credentials and also all display name spoof instances involved impersonation of someone within the organisation.

Our dataset contains more number of emails towards the end of the observation period as the dataset contains emails from all volunteers in that period; which is not necessarily true at the start of the observation period since we have volunteers from different academic years. Fig. 1 shows a plot of number of emails in our dataset with intervals of 4 months. Also, Fig. 2 shows the plot of the number of attack emails with intervals of 4 months. The number of such attacks increased tremendously from the latter half of 2016.

4 Design of scoring system

The main goal of the scoring technique is to have a low false positive rate and detecting fair number of true positives. Another goal is to be real-time in order to process emails immediately as they are received. We leverage the structured information in email headers and one element in email body. The scoring technique determines the score for every received email and decides actions to take correspondingly. The following two sections describe the features that are feed into our scoring technique, and then the scoring technique itself.

Table 2 Attack instances with phishing URLs and display name spoof

attack instances containing phishing URLs	22/27
attack instances involving display name spoof	04/27

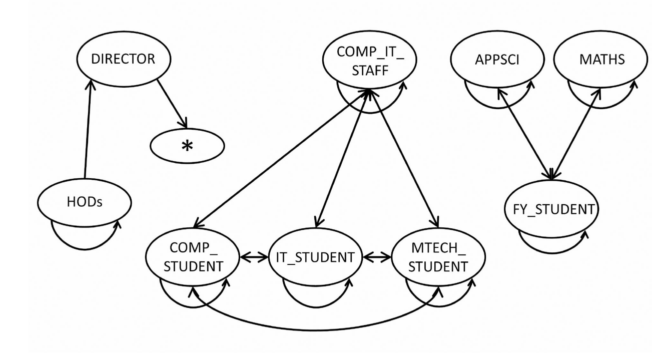


Fig. 3 *Example of a hierarchy related to computer department in an institute. Note: the figure is not representative of any hierarchy and is meant only for illustration purpose to explain the concept of hierarchy. Organisations can decide the format in which they want to represent their hierarchy*

4.1 Feature characterisation

We derive four feature categories from the characteristics of the lateral spear phishing emails we observed in our organisation. Each feature category contains one or more features. Based on its nature, a feature could be either context-based or history-based. A context-based feature analyses a specific property/context of the email, whereas history-based feature compares itself with a historically learned profile.

4.1.1 Behavioural context feature category: Once an account within the organisation is compromised, the attacker can send spear phishing emails to other users in the organisation who are not related to the original user of compromised account. We try to identify this behaviour. A natural solution is to find frequent email sending patterns from the email history. However, this does not suit to our case as we explain later in this section. The idea is to check whether there is a '*meaningful*' relationship between the sender and recipients. Note the use of the term 'meaningful' relationship, and not just relationship. An example of a meaningful relationship is the staff of computer department sending emails to students of computer department, whereas the staff of computer department sending email to a metallurgy department student is not meaningful. We discuss more on this after we describe how we establish meaningful relationships.

We introduce the concept of organisation email sending hierarchy. This hierarchy is a graph made of nodes and links between them, which shows the meaningful email sending relationships in the organisation. The nodes can be individual users or groups of users, and are connected using links. Links are directional. A link from node A to node B means that node A has a meaningful email sending relationship with node B. The same is not applicable vice-versa unless there is a link from node B to node A.

Though the nodes in the hierarchy can be individual users, to reduce the complexity of the hierarchy, the nodes could also be 'groups' within the organisation (e.g. teams, classes, departments etc.). Users who have more responsibility/rights than what the group has (e.g. team leader, HoD etc.), can be considered as separate nodes along with the group node. In such cases, a user would be part of multiple nodes and would have different links with other nodes. Various levels/subsets of groups can be used in the hierarchy. A simple example of hierarchy related to a department in an organisation is shown in Fig. 3.

We call this email sending hierarchy and not email sending patterns, since we do not consider only frequent patterns per se, e.g. Director of an institute may not be sending emails very frequently to students in the institute and the pattern of sending could be arbitrary, but it is a meaningful relationship. Due to this, we cannot judge anomalous behaviour by looking at factors such as how many previous days the sender has sent emails to the recipient. Hence, we do not just use email sending patterns or number of previous days the sender has sent emails to the recipients, but look for meaningful relationships in email sending for construction of hierarchy. Administrators can embed the domain knowledge of their problem space to construct such hierarchy.

Administrators can leverage email history of the organisation using learning techniques to find email sending patterns. As mentioned above, we look for 'meaningful' patterns and not just frequent patterns. Therefore, administrators can use learning techniques to get a preliminary view of the organisation's patterns, and then remove and add relationships depending on their needs. The hierarchy can be designed easily and saved to database using JavaScript libraries such as GoJS, SigmaJS, VisJS etc. (or libraries available in any other language of choice). This is a context-based feature, which checks whether it is 'meaningful' for a sender to send email to its recipients. There are legitimate exceptions where email sending might not have meaningful relationships, but these are presumably rare.

More features can be added to this category depending on the organisation. For example, in our case, we consider an email to be more suspicious if there is no link in the hierarchy for the particular email and the recipients are included in blind carbon copy (BCC).

4.1.2 IP feature category: An account's compromise can take place from anywhere in the world. For efficiency and ease of use, organisational users are allowed to access their accounts remotely. This also provides attackers with opportunities to infiltrate the network. This feature category checks whether the email is sent from a known geographic location. The check is both network-wide as well as per-user. The first feature in this category is based on sender's city obtained from Geo-IP mapping. This feature further has two sub-features. These give

i. the number of times the sender has sent emails from the city, and
ii. the number of other users that have sent emails from the city.

In both cases, lesser the value more is the suspiciousness. There might be a case where the sender has sufficient number of emails from the geo-located city, but there are not enough users who have sent emails from that city. This can happen when a particular user has to visit the city many number of times and requires sending emails from that location. Hence, if the first sub-feature gives a sufficiently high value, we disregard the second sub-feature value. Yet, the second sub-feature serves a major purpose, since many students in our organisation send emails rarely. So the per-user historical profile is not built sufficiently. In such cases, this network-wide sub-feature plays a major role, which gives the count of other users who have sent emails from the city. We tested various Geo-IP databases and found that IP2Location gives more accurate city mapping than other services for our case, followed by MaxMind.

IP-City mapping is not always reliable and has inaccuracies in some cases, hence we use a second feature in this category, which is based on the autonomous system number (ASN) obtained from the IP-ASN mapping. An ASN can be finer grained than geo-located city, and sometimes vice-versa. However, it is a good enough substitute in practice. Similar to the city-based sub-features, this feature also has two sub-features. These give

i. the number of times the sender has sent emails from the ASN, and
ii. the number of other users that have sent emails from the ASN.

Similar to the city-based sub-features mentioned above, in this case too, if the first feature gives a sufficiently high value for the first sub-feature, we disregard the second sub-feature value. For the IP-ASN mapping, we use MaxMind's ASN database.

Both the features in this category are history-based.

A problem arises when a legitimate user sends email during travel. To determine whether the user is travelling, Javed [17] has used ASNs of airports to check whether user is accessing the account from an airport using the public Wi-Fi. However, this can also be exploited by the attacker if a large number of such ASNs corresponding to public places are whitelisted. Also, in our case, there are many possibilities such as public Wi-Fi in train, bus etc. So, we decided not to consider the case of travelling of user.

Due to the dynamically changing IPs within our organisation and also for users accessing accounts remotely, we cannot check whether the user has used the IP address before. A user's machine can be allocated different IP on different occasions within our organisation.

4.1.3 Display name feature category: Lateral spear phishing attacks can also be accompanied by display name spoof, as is found in four such attack instances in our dataset. As shown in various previous studies, even technically sound people may fall for display name spoof [18, 19]. For example, a student's compromised account has been given the display name of a lab assistant (Computer lab assistant <studentemailid@coep.ac.in>). The receivers would respond to the email, perceiving that the sender is from the real lab assistant.

Authentication techniques such as Sender Policy Framework/ DomainKeys Identified Mail/Domain-based Message Authentication, Reporting & Conformance [20] do not work in this case since the email is intra-domain and is being sent from a legitimate account (which is compromised).

The first feature in this category gives the number of times the sender has sent emails with the current display name. This is a history-based feature.

Many students in the institute send emails rarely. Hence, sufficient display name records are not available for such students. The second feature in this category checks whether display name is lexically similar to email address, if there are insufficient number of current display name records for the sender. Many times, organisations set email addresses of employees based on their real name. Hence, lexicographic comparison can work even if employees do not use exact name (e.g. using nicknames). For example, email addresses in our institute have a part of the name (first name and/or surname) of the user. We use this structure to lexicographically compare the sender's email address and display name. This is a context-based feature.

We do not consider the trustworthiness/reputation of the spoofed display name, since we have instances where people have fallen for these attacks without the use of authoritative/known display names.

4.1.4 Fully qualified domain name (FQDN) feature category: As seen from Table 2, 22 attack instances in our dataset used a phishing URL in the email. URLs can have many variations, such as various parameters in the URL. Instead of using the full URL, we use its FQDN to deal with fine granularity and yet not be too coarse. We consider a FQDN which is newer to the organisation to be more risky. The first feature checks whether the FQDN is seen for the first time. We notice that most of the FQDNs in our dataset appeared only once in the entire observation period. In Fig. 4, it can be seen that only 35% of the FQDNs were seen more than once or in other words, 65% of the FQDNs were seen only once.

Due to this fact, we cannot use familiarity of FQDNs to our organisation alone as a feature. This can also be abused by attackers by sending a large number of emails containing a phishing FQDN to increase the familiarity of the FQDN to the organisation. We keep track of the *reputation* of the FQDNs *within* the organisation. If the FQDN is seen before, then we use our second feature, which checks for the reputation of the FQDN

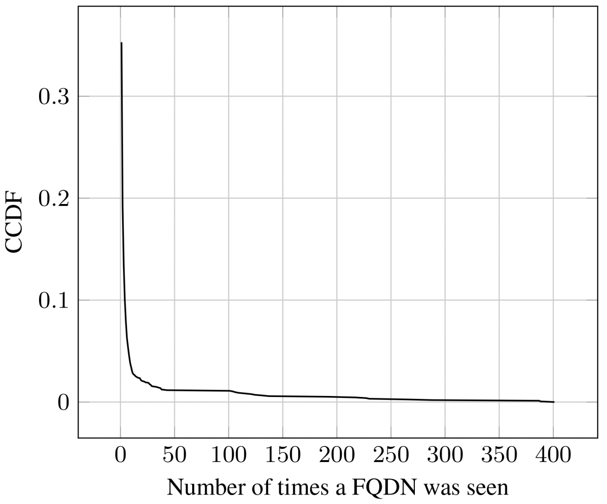


Fig. 4 *Complementary cumulative distribution function (CCDF) of number of times a FQDN was seen in our dataset*

Table 3 Summary of features fed into our scoring system

Feature	Nature
hierarchy feature category	
link between the nodes in hierarchy	context-based
receivers in BCC	context-based
IP feature category	
city features	history-based
ASN features	history-based
display name feature category	
display name history	history-based
lexicographical matching	context-based
FQDN feature category	
familiarity to organisation	history-based
reputation within organisation	history-based

within the organisation. We define the reputation of a FQDN as the suspiciousness of the emails it has appeared in.

We do not use domain reputation such as PageRank etc., since adversaries can avoid detection by hosting their pages on reputed domains (e.g. weebly, wordpress etc.). This second feature gives the past reputation of the FQDN within the organisation. The reputation of a FQDN is dependent on the scores of the emails in which it has been seen. The calculation of the reputation of the FQDN is discussed in the next section when we describe the scoring technique. If a phishing FQDN is seen more number of times, the first feature would fail, but since we maintain the reputation of the FQDN, the second feature would aid us due to the historical low reputation of the FQDN in our system.

Both features in the category are history based. We do not use rest of the text in the email body. We only use the FQDN of any URL that is present in the email body.

We have summarised the features in Table 3.

4.2 Lateral spear phishing (LSP) scoring

In the previous section, we described the features used to characterise individual emails. We now describe how the individual feature categories are aggregated to provide a score for an email and take actions correspondingly. Hence, the scoring technique is called as LSP scoring.

The scoring technique gives a final score in the range of 0.0–1.0. A higher score corresponds to email being more suspicious. We consider a score >0.5 to be a lateral spear phishing email. All feature category scores are combined together using simple summation to give the final LSP score for the email. A feature category gets its score from the scores of its individual features. Each feature category has a maximum allowed score such that it does not cross 0.5. This is done to ensure that only one feature category does not get to classify an email as lateral spear phishing. Hence, for an email to be classified as lateral spear phishing, a non-zero score is required from at least two feature categories. Individual feature scores in the categories are normalised accordingly.

Each history-based feature has its own threshold, decided by the administrators in the organisation. If the value given by a history-based feature exceeds its threshold, the feature does not give any score. The score increases linearly with lesser historical values.

Once the LSP score for the email is calculated, the reputation of any FQDNs in the email bodies is re-calculated. This reputation is calculated using the weighted average of the LSP score of the current email and past reputation of the FQDN, with the number of previous days the FQDN has been received. This new reputation is used the next time this FQDN is seen.

Features such as IP and display name require prior history to build historical profiles and give accurate results. A bootstrapping period is required for these history-based features. Although more history would increase the chances of the system to give more accurate results, it also means that the system would require a large amount of history to give good results. A trade-off needs to be obtained in this case such that a sufficient amount of history is used, and yet it is not too much. Considering this, our training dataset contains initial four months of emails.

In this bootstrap period, though IP and display name history will be built, the FQDN reputations of URLs in the emails also need to be determined. FQDN reputations are derived from the LSP scores of the emails. During the bootstrap period, if we simply determine LSP scores of emails directly using all the features without the pre-existence of historical profiles, we get a large number of false positives. This is mainly due to the lack of history for IP and display name features. Also, benign FQDNs get bad reputation if they are present in these earlier emails and further contribute to the false positives. To overcome this issue of building history (for IP and display name features) as well as finding LSP scores of the same emails (in order to calculate FQDN reputations), we use two steps in the training phase. In the first step, we use the initial four months of emails to build IP and display name history. Then in the second step, we use the same four months of emails to find LSP scores of the emails using all features (IP and display names; tables are only referred to and not updated in this step, since they are filled with data of the same emails in the first step). As the LSP scores of emails are determined, FQDN reputations of any URLs in the email bodies are updated in the database. Once the database is filled with IP, display name and FQDN reputation data from training emails, we run the algorithm on the emails after the bootstrap period.

There were inaccuracies even in the best performing Geo-IP database in our case. Due to this, city feature in the IP feature category gives a non-zero score owing to the absence of sufficient history of the incorrect city. There were considerably lesser number of inaccuracies in the IP-ASN database. Hence, if city feature gives a non-zero score due to insufficient history, but ASN feature gives a zero score due to sufficient history, we consider the IP feature category as a whole to have zero score. We assume that if the system contains sufficient historical profiles of an ASN, the system should also contain sufficient historical profiles of the corresponding city.

An attacker may try to contaminate our historically built profiles by sending a large number of emails from an anomalous IP and/or using spoofed display name. Therefore, if an email gets a score >0.5, we do not add its history based values to our database. Algorithm 1 (see Fig. 5) shows the procedure of LSP scoring in short.

5 Evaluation

We evaluate our scoring technique on the dataset, which we collected from 40 volunteers in COEP, amounting to 3.5 years' worth of emails. The dataset contains 27 attack instances, which correspond to 323 lateral spear phishing emails. Using our scoring system, we were able to detect 22 attack instances and 1 attack instance partially (some emails in the attack instance were

```
process_email(email $E$):
    $F \leftarrow$ parse_fields($E$)
    $s_1 \leftarrow$ behav_score($F[emailid, receivers]$)
    $s_2 \leftarrow$ ip_score($F[ip]$)
    $s_3 \leftarrow$ dn_score($F[emailid, displayname]$)
    $s_4 \leftarrow$ fqdn_score($F[fqdns]$)
    $lsp\_score \leftarrow s_1 + s_2 + s_3 + s_4$
    if $lsp\_score > 0.5$ then
        flag email
        update tables except history-based tables
    else
        update all tables
    return lsp_score
```

Fig. 5 *Algorithm 1: LSP scoring*

Table 4 Comparison of performance of our scoring technique with that of anomaly detection techniques

Algorithm	True positive, %	False negative, %	True negative, %	False positive, %	Accuracy, %	F1 score, %
LSP scoring	86.69	13.31	99.12	0.88	98.79	79.32
local outlier factor	96.28	3.72	2.22	97.78	4.74	5.14
one-class SVM	99.07	0.93	45.32	54.68	46.76	9.07
robust covariance	91.95	8.05	96.3	3.7	96.18	56.3

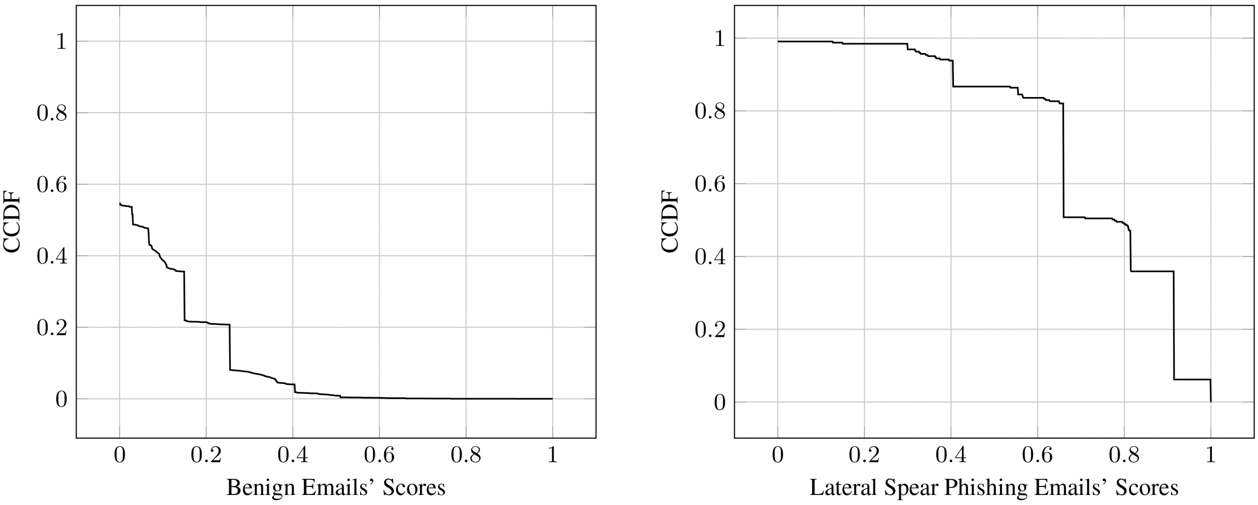


Fig. 6 *CCDF of LSP scores for benign and lateral spear phishing emails given by the LSP scoring system*

detected), which correspond to 280 emails. Additionally, out of the 27 attack instances, 25 instances were known previously. We detected two attack instances in the dataset which were not known to us earlier. Out of these two, one attack instance was highly targeted and corresponds to only one lateral spear phishing email. We obtained a false positive rate of 0.88% with a true positive rate of 86.69%. The overall accuracy is 98.79%. The results of various performance metrics are shown in Table 4.

The most common reason of false positives is incorrect city result given by the Geo-IP database. On manual check, we found that the senders were actually sending from the same city as the organisation, but the Geo-IP database had incorrect IP-City mapping. This error along with smaller values given by other feature categories led to the LSP score cross 0.5. Although none of the benign emails got a score of above 0.65.

Fig. 6 shows how the LSP scoring system scores benign and lateral spear phishing emails. 99.1% of benign emails have score <0.5. Also, 86.6% of lateral spear phishing emails score >0.54. This indicates that the compact feature set provides enough separation for correctly identifying benign and lateral spear phishing emails.

5.1 Comparison with anomaly detection techniques

We compare our system's performance with that of the anomaly detection techniques based on machine learning from the literature [21, 22], viz. local outlier factor, one-class Support Vector Machine (SVM), and robust covariance. For each email, we constructed a feature vector from the scores given by each of the feature categories and normalising them. We have shown the comparison of the results in Table 4. Although one-class SVM gives the highest true positive rate, it also gives a very unacceptable false positive rate. Similar results can be observed in other techniques. All the compared anomaly detection techniques have false positive rates impractical for use in detection of such attacks. Our false positive rate, true negative rate, accuracy and F1 score are much better than other algorithms. Even though the true positive and false negative rates of other algorithms are better than ours, it is mainly due to our effort to keep false positive rate to the minimum which is a very important factor for the use of a system in practice.

6 Discussion

Cyber space is in a continuous state of tug of war between attackers and defenders. Our motive behind this research is to make it difficult for attackers to conduct successful lateral spear phishing attacks. Our scoring system uses context-based and history-based features to detect lateral spear phishing emails. The system matures and continuously refines itself as it processes more number of emails. Administrators can take action in real-time and stop attack from causing further damage by revoking access of compromised account. As shown in the previous section, we are able to obtain high accuracy with the compact feature set. Additionally, due to the compact feature set, the decision boundary is less complex.

Depending on the problem space, administrators can add required features to the system. Since these attacks are highly specialised, domain knowledge can help to a large extent in detecting them. As we do not update the history based values if an email has been marked as lateral spear phishing, this would not allow the attacker to easily contaminate the historically built profiles.

6.1 Limitations

An obvious limitation of the system is that it does not detect spear phishing attacks from email addresses outside the organisation. Since our system does not consider spear phishing emails coming from email addresses outside COEP, such attacks would be ignored by our system.

To build historical profiles, the system needs prior training. We have given the detailed results in previous section with the training period of 4 months. We also tested the system with lesser training periods. These tests gave higher false positive rates than the test with 4 month training period. This shows that a minimum required amount of training is needed to give better results. This period depends on the density of the users' activity.

From the email bodies, we consider only FQDN of any present URLs, and not any textual content. This decision was mainly taken due to privacy and confidentiality reasons. A detector may use the textual content to build stylometric models and use Natural Language Processing (NLP) techniques.

Our detection system cannot detect attacks if the attacker chooses to stay passive (in order to get access to sensitive information by scanning conversations), as fundamentally our system works on detection of attack emails sent by the attacker.

As explained in Section 4.1.2, we do not consider the case of travelling user. This would result in the system giving inaccurate score in the IP feature category which would change the total LSP score and hence cause misclassification in such cases.

6.2 Chrome extension – COEP Kumpan

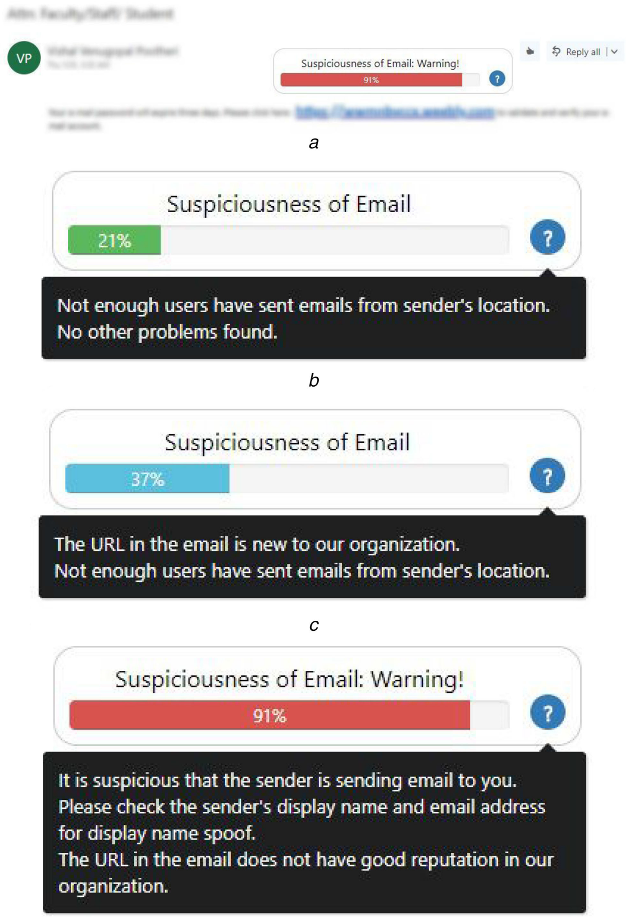
The proposed technique is primarily aimed to be implemented at the server side by email providers who provide email services to organisations, or by organisations that manage their own email server. However, this is not always possible. Depending on the organisation's requirements, the organisation may decide to allow all intra-domain emails to reach the recipients and let the end user decide what action to take on the emails or the organisation might be using an email service which does not support such detection mechanism. In such cases, to aid users in their decision while dealing with an email, organisations can implement a technique such as ours, within their organisation with the help of plugins.

We developed a Chrome Extension, called COEP Kumpan, which displays a UI element on Outlook Web (used by a majority of COEP users) to guide users in their decision making. The extension uses an API we provided for our scoring system. Kumpan, in Marathi language, means fence – in our case, to safeguard the organisation.

The extension is developed from the point of view of preventative defence to provide users with warning messages. The design of the UI element displayed by the extension is aimed to nudge with relevant information of the email [23], in order to capture users' attention and guide them to take informed decision. Fig. 7 shows screenshots of the UI element shown to user. Since we do not work on detecting spear phishing emails coming from email addresses not belonging to COEP domain, we show a generic warning message for all emails not from COEP domain, and give instructions to user to proceed carefully for such emails. However, the evaluation of the extension is out of scope of this study.

7 Related work

The closest to our work is [9], where the authors proposed an approach for detecting credential spear phishing attacks in settings where attack instances are as few as one to millions of benign emails. Their work draws on the Simple Mail Transfer Protocol (SMTP), Network Intrusion Detection System (NIDS) and Lightweight Directory Access Protocol (LDAP) logs of 370 million anonymised emails from their organisation. They rank top *n* suspicious events every nightly for an analyst to review. The *n* is their daily budget of alerts that the analyst has to review (budget of ten alerts mentioned in the study). They detect attacks of three impersonation types with three different sub-detectors. Also, each sub-detector detecting attacks at two different stages of attack (lure and exploit). At their daily budget, they achieved a very low false positive rate of 0.005%. However, since they do not mean to alert/detect all the attacks in the given day, the actual false positive rate in this case is not known. Our system would enhance capabilities of such a detector, for detecting lateral spear phishing attacks in particular.



a

Suspiciousness of Email

21%

Not enough users have sent emails from sender's location.
No other problems found.

b

Suspiciousness of Email

37%

The URL in the email is new to our organization.
Not enough users have sent emails from sender's location.

c

Suspiciousness of Email: Warning!

91%

It is suspicious that the sender is sending email to you.
Please check the sender's display name and email address for display name spoof.
The URL in the email does not have good reputation in our organization.

d

Fig. 7 *Example of Chrome Extension's UI element*
(a) Example of the UI element in an email in Outlook Web, ***(b)–(d)*** Examples of UI element with messages, either nudged or on mouse hover based on user preferences

IdentityMailer [11] builds historical profiles for users based on their habits of email writing, composition and recipient interaction. Working at sender's side, it compares the sending user's historical profile with the current email being sent. Similar to this, EmailProfiler [10] works on the recipient's side and builds historical profiles from incoming emails. They use email's metadata and stylometric features with a SVM classifier. Given the organisation's conditions on privacy and confidentiality, such techniques could be of aid to our system by considering textual content in the email bodies.

Pecchia *et al.* [24] worked on detecting compromise of account based on the activities from the account post-login. They used various security indicators such as suspicious command-line activity, file downloads etc. to train a naive Bayesian network. Although they detected all attacks in their dataset, it also cost them huge number of false positives in doing so. The authors suggest that the approach is more suited to guide the administrators in dealing with potential compromises.

Freeman *et al.* [25] developed a statistical Bayesian framework to detect suspicious login attempts. They propose use of what they call reinforced authentication, which checks for additional information beyond credentials during login attempts. They extract this information from the HTTP session logs of authentication sessions. Their results show a false positive rate of 10% which they reason as a business decision. This false positive rate is quite high for other organisational settings. Also, working on authentication logs, Zhang *et al.* [14] employed a logistic regression classifier on features extracted from the authentication logs in their academic institution.

Laszka *et al.* [26] worked on the problem of strategic email threshold selection for filtering spear-phishing e-mails using concepts of game theory. The authors consider the case where the attacker selectively targets users considering various factors which maximise the value gained. They theoretically prove that defenders may reach Nash equilibrium on the long term. However, they do not prove the performance of such a system on any dataset.

On the grounds of social networks, the authors of [27, 28] conducted large-scale experiments in which they construct clusters based on content and URL similarity of Facebook posts/Twitter tweets. Egele *et al.* [27] look for deviation in the users' behaviour as opposed to their historical profiles, whereas Thomas *et al.* [28] use a logistic regression classifier to categorise the clusters for various types of accounts. These studies show that account compromise on social networks is largely due to phishing and the compromised accounts are in turn used to conduct large-scale phishing campaigns. Dewan *et al.* [29] trained four machine learning classifiers on the stylometric features of emails and social features from LinkedIn profiles. However, their results show that their chosen features from LinkedIn did not help in identifying spear phishing emails.

8 Conclusion

In this study, we proposed a scoring technique to detect lateral spear phishing emails using combination of various features. Our aim was to create a practical, deployable and real-time detection system for such attacks. We evaluated our scoring technique on 3.5 years' worth of email dataset collected from 40 volunteers in our organisation. The results show that the scoring technique achieved an accuracy of 98.79% and a false positive rate of 0.88%. Also, we were able to detect two attack instances which were not known to us earlier. We compared the performance of our scoring technique with that of anomaly detection techniques from the literature, and show that our scoring technique gives better overall results. We have also developed a Chrome extension – COEP Kumpan – for use by COEP to help its users in taking better informed decisions. Domain knowledge can play a crucial role in detection of such specialised attacks. Since organisations currently depend mostly on users for reporting spear phishing attacks, our research can serve as an improvement in detecting fair number of true positives, keeping low false positive rate, and also identifying previously unknown attacks.

9 Acknowledgments

The authors would like to thank all the volunteers from our organisation who shared their emails and all those who directly or indirectly helped in this work.

10 References

[1] Fireye: 'Best defense against spear-phishing attacks', January 2018. Available at https://www.fireeye.com/current-threats/best-defense-againstspear-phishing-attacks.html

[2] Mak, A.: 'Oh great, a hacking group linked to North Korea is getting very good at targeting bitcoin owners', February 2018. Available at https://slate.com/technology/2018/02/the-lazarus-group-isinvading-bitcoin-wallets-a-mcafee-study-finds.html

[3] K. Lab: 'Phishing for cryptocurrencies: how bitcoins are stolen', January 2018. Available at https://www.kaspersky.com/blog/crypto-phishing/20765/

[4] Fingas, J.: 'Florida phishing attack exposes data for 30,000 Medicaid recipients', January 2018. Available at https://www.engadget.com/2018/01/07/florida-phishing-attack-exposes-data-for-30-000-medicaid-recipie/

[5] B. Krebs: 'Equifax or equiphish? — Krebs on security', September 2017. Available at https://krebsonsecurity.com/2017/09/equifax-or-equiphish/

[6] Imperva: 'Phishing made easy: time to rethink your prevention strategy?', December 2016. Available at https://www.imperva.com/docs/Imperva-HII-phishing-made-easy.pdf

[7] Meyer, S.: 'Phishing attacks: insights from more than 1,000 free phishing kits – page 2 of 2', January 2018. Available at https://www.cpomagazine.com/2018/01/10/phishing-attacks-insights-from-more-than-1000-free-phishing-kits/2/

[8] C. T. R. Labs: 'Phishing got darker and smarter', January 2018. Available at https://www.comodo.com/lab/pdf/phishing-got-darker-and-smarter.pdf

[9] Ho, G., Sharma, A., Javed, M., *et al.*: 'Detecting credential spearphishing in enterprise settings'. 26th USENIX Security Symp. (USENIX Security 17), Vancouver, BC, 2017, pp. 469–485. Available at https://www.usenix.org/conference/usenixsecurity17/technicalsessions/presentation/ho

[10] Duman, S., Kalkan-Cakmakci, K., Egele, M., *et al.*: 'Email-Profiler: spear phishing filtering with header and stylometric features of emails'. 2016 IEEE 40th Annual Computer Software and Applications Conf. (COMPSAC), Atlanta, Georgia, USA, June 2016, vol. 1, pp. 408–416

[11] Stringhini, G., Thonnard, O.: 'That ain't you: blocking spear phishing through behavioral modelling', in '*Detection of intrusions and malware, and vulnerability assessment*', Series. Lecture Notes in Computer Science (Springer, Cham, 2015), pp. 78–97. Available at https://link.springer.com/chapter/10.1007/978-3-319-20550-2_5

[12] I. StackExchange: 'Email – why does outlook not block spam sent by employees? – Information security stack exchange', December 2017. Available at https://security.stackexchange.com/questions/176129/whydoes-outlook-not-block-spam-sent-by-employees

[13] B. Krebs: 'The market for stolen account credentials — Krebs on security', December 2017. Available at https://krebsonsecurity.com/2017/12/the-market-for-stolen-account-credentials/

[14] Zhang, J., Berthier, R., Rhee, W., *et al.*: 'Safeguarding academic accounts and resources with the University Credential Abuse Auditing System'. IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN 2012), Boston, Massachusetts, USA, June 2012, pp. 1–8

[15] E. Online: 'Configure your spam filter policies: exchange online protection help', December 2017. Available at https://technet.microsoft.com/library/jj200684

[16] Brandom, R.: 'Two-factor authentication is a mess', July 2017. Available at https://www.theverge.com/2017/7/10/15946642/two-factorauthentication-online-security-mess

[17] Javed, M.: 'Detecting credential compromise in enterprise networks'. PhD thesis, Electrical Engineering and Computer Sciences University of California at Berkeley, 2016. Available at https://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-216.pdf

[18] Alsharnouby, M., Alaca, F., Chiasson, S.: 'Why phishing still works: user strategies for combating phishing attacks', *Int. J. Hum.-Comput. Stud.*, 2015, **82**, pp. 69–82. Available at http://www.sciencedirect.com/science/article/pii/S1071581915000993

[19] Benenson, Z., Gassmann, F., Landwirth, R.: 'Unpacking spear phishing susceptibility', in Mueller, R. (Ed.): '*Financial cryptography and data security*', Series. Lecture Notes in Computer Science (Springer, Cham, 2017), pp. 610–627. Available at https://fastmail.blog/2016/12/24/spf-dkim-dmarc/

[20] Mueller, R.: 'SPF, DKIM & DMARC: email anti-spoofing technology history and future', December 2016. Available at https://blog.fastmail.com/2016/12/24/spfdkim-dmarc/

[21] Chandola, V., Banerjee, A., Kumar, V.: 'Anomaly detection: a survey', *ACM Comput. Surv.*, 2009, **41**, (3), p. 15

[22] Goldstein, M., Uchida, S.: 'A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data', *PLOS ONE*, 2016, **11**, (4), p. e0152173. Available at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152173

[23] Nicholson, J., Coventry, L., Briggs, P.: 'Can we fight social engineering attacks by social means? Assessing social salience as a means to improve phish detection'. Thirteenth Symp. on Usable Privacy and Security (SOUPS 2017), Santa Clara, CA, 2017, pp. 285–298. Available at https://www.usenix.org/conference/soups2017/technicalsessions/presentation/nicholson

[24] Pecchia, A., Sharma, A., Kalbarczyk, Z., *et al.*: 'Identifying compromised users in shared computing infrastructures: a data-driven Bayesian network approach'. 2011 IEEE 30th Int. Symp. on Reliable Distributed Systems, Madrid, Spain, October 2011, pp. 127–136

[25] Freeman, D.M., Jain, S., Dürmuth, M., *et al.*: 'Who are you? A statistical approach to measuring user authenticity'. Network & Distributed System Security (NDSS), San Diego, California, USA, 2016

[26] Laszka, A., Lou, J., Vorobeychik, Y.: 'Multi-defender strategic filtering against spear-phishing attacks'. Proc. 13th AAAI Conf. on Artificial Intelligence, Ser. AAAI'16, Phoenix, Arizona, 2016, pp. 537–543. Available at http://dl.acm.org/citation.cfm?id=3015812.3015893

[27] Egele, M., Stringhini, G., Krügel, C., *et al.*: 'COMPA: detecting compromised accounts on social networks'. Network & Distributed System Security (NDSS), San Diego, California, USA, 2013

[28] Thomas, K., Li, F., Grier, C., *et al.*: 'Consequences of connectivity: characterizing account hijacking on twitter'. ACM Conf. on Computer and Communications Security, Scottsdale, Arizona, USA, 2014

[29] Dewan, P., Kashyap, A., Kumaraguru, P.: 'Analyzing social and stylometric features to identify spear phishing emails'. 2014 APWG Symp. on Electronic Crime Research (eCrime), Birmingham, Alabama, USA, September 2014, pp. 1–13