Accurate Spear Phishing Campaign Attribution and Early Detection

YuFei Han
Symantec Research Labs
Sophia Antipolis
France
Yufei_Han@symantec.com

Yun Shen
Symantec Research Labs
Dublin
Ireland
Yun_Shen@symantec.com

ABSTRACT

There is growing evidence that spear phishing campaigns are increasingly pervasive, sophisticated, and remain the starting points of more advanced attacks. Current campaign identification and attribution process heavily relies on manual efforts and is inefficient in gathering intelligence in a timely manner. It is ideal that we can automatically attribute spear phishing emails to known campaigns and achieve early detection of new campaigns using limited labelled emails as the seeds. In this paper, we introduce four categories of email profiling features that capture various characteristics of spear phishing emails. Building on these features, we implement and evaluate an affinity graph based semi-supervised learning model for campaign attribution and detection. We demonstrate that our system, using only 25 labelled emails, achieves 0.9 F1 score with a 0.01 false positive rate in known campaign attribution, and is able to detect previously unknown spear phishing campaigns, achieving 100% 'darkmoon', over 97% of 'samkams' and 91% of 'bisrala' campaign detection using 246 labelled emails in our experiments.

CCS Concepts

•**Security and privacy** → **Phishing;**

Keywords

spear phishing emails, semi-supervised learning

1. INTRODUCTION

Spear phishing emails refer to the emails, appearing legitimate, sent to targeted individuals using relevant contextual information to trick them into disclosing sensitive information to the attackers or installing malware on their computers. After thorough reconnaissance profiling where the victims work, what their rankings are within the organisation, what they are interested in, etc., attackers usually sent out a small number of spear phishing emails to target a small number of carefully chosen individuals or groups to gain stealthy access. These emails can even be sent from legitimate email addresses from compromised machines. Attachments of these spear phishing emails usually contain previously unknown exploits that evade AV engines detection and are handcrafted for specific organisations.

There are growing evidences that these spear phishing campaigns are increasingly pervasive [25], sophisticated [10], and remain the starting points to more sophisticated attacks leading to damaging losses in terms of identity theft, sensitive intellectual property and customer information, and national-security secrets [16]. Conventional spam detection techniques, such as filtering emails sent by botnets [19, 11], behavioural blacklist [20, 12], reputation-based methods [9, 18], linguistic stylometry attribution [17, 2], are less effective to detect spear phishing emails. For example, spear phishing emails sent from compromised machines render botnet or reputation-based methods ineffective. Research on phishing has covered various aspects of phishing attack - social engineering [13], psychology [26], economics [1], awareness [5] and counter measures [14, 27, 15]. Those techniques mainly focused on preventing deceptive phishing from redirecting users to fake Websites through an embedded link within the email, and can not be easily adapted to campaign attribution and identification.

It is important to identify series of attack campaigns that are likely performed by the same organisations as early as possible to understand their TTP (Tactics, Techniques and Procedures), and devise countermeasures accordingly. However, in the real world, spear phishing campaign attribution tasks require considerable time and manual efforts to incorporate various aspects such as backchannels, static/dynamic malware analysis results, shared intelligence, etc. Consequently only a small number of suspicious emails are attributed and majority of them are left not investigated. With limited data accumulated in the manual investigation process, it is difficult to apply off-the-shelf machine learning techniques (e.g. SVM [8], Naive Bayes [3]) to achieve automated spear phishing campaign identification since they require a large amount of labelled data to train the models so as to reach high accuracy.

We are particularly interested in two challenges relating to

SAC 2016,April 04-08, 2016, Pisa, Italy

http://dx.doi.org/10.1145/2851613.2851801

spear phishing campaign attribution in this paper: 1) can we build a campaign attribution system that requires limited data to train and can achieve high classification accuracy? 2) at the same time, can we use this model to identify previously unknown campaigns from these unlabelled spear phishing emails? To tackle these challenges, we first identify four categories of features profiling the characteristics of spear phishing emails in a holistic way. They are not only devised to capture various aspects of a spear phishing email but also are robust to the evolution of spear phishing campaigns. We then design and implement an attribute graph based semi-supervised learning framework to effectively identify and attribute campaigns.

In this paper, we demonstrate that our proposed method, using only 11 spear phishing emails from 5 different campaigns as the training data (1% of manually labelled emails), can effectively identify 90% of spear phishing emails belonging to these campaigns in the unlabelled data, while keeping a low FP less than 0.02%. It is important to note that our system is able to achieve early detection since limited labelled emails are required. We also demonstrate that our proposed method can accurately detect previously unknown spear phishing campaigns (e.g. over 97% of 'samkams' campaign and 91% of 'bisrala' campaign) from unlabelled data requiring only 336 emails (i.e. 25% of manually labelled emails), effectively reducing 75% manual labelling efforts.

In summary, this paper makes the following contributions:

- We propose four categories of email profiling features covering meta-information about origin, recipient, content and attachment of a spear phishing email. They form a holistic description of email characteristics, and provide a solid base for automated campaign attribution and identification.
- We propose an attribute graph based semi-supervised learning (SSL) framework to improve the applicability of machine learning based methods in spear phishing campaign attribution with limited labelled emails.
- We use the proposed SSL framework to gain insights on spear phishing emails from different campaigns, and demonstrate how previous unknown spear phishing campaign can be detected via a detailed case study.

The rest of the paper is organised as follows. In Section 2 we state our goals and present the overall threat intelligence and attribution system. Section 3 formally presents the challenges and details our methodology. Section 4 presents experimental results. We provide a case study on how to identify newly emerged campaign using the proposed method. We provide a detailed case study in Section 5 and conclude our work in Section 6.

2. MODEL OVERVIEW

In this section, we articulate the goals of our research and present an overview of the spear phishing campaign attribution and identification model.

Problem Statement. Current spear phishing investigation process is hampered by considerable manual efforts from security experts to incorporate various aspects such as backchannels, static/dynamic malware analysis, shared intelligence, etc. In this paper, we focus on using the features extractable from spear phishing emails to build an automated spear phishing campaign attribution model to effectively and automatically attribute unlabelled suspicious emails to known campaigns or identify newly emerged spear phishing campaigns.

We also have some non-goals: in this paper, we do not aim to analyse the attachments and the payloads further dropped by these attachments, or network-level activity information. Rather, our main goal is to use an attribute graph based semi-supervised learning framework to accurately and automatically attribute unlabelled suspicious emails to either known or newly emerged campaigns.

System Architecture. Fig. 1 shows the overall work flow of campaign attribution system. Given a group of labelled spear phishing emails (from known campaigns) and unlabelled suspicious emails mixed together, *email profiling* module extracts features from each input email and generates vectorised emails. *Benign email filtering* module is employed at first to filter out benign emails and identify spear phishing emails from the input data. Detected spear phishing emails are then fed into the following *unknown campaigns identification* module to verify whether they emerge from any known campaigns or brand-new campaigns. If they belong to previously unknown campaigns, new campaign labels are assigned. If classified as known campaigns, *known campaign attribution* module is later employed to attribute the spear phishing emails into known campaigns.

For each analytical module in the flowchart, semi-supervised learning is performed following the same process. A K-Nearest-Neighbouring (KNN) attribute graph is constructed based on the email profiling features. Each node represents an email, and an edges represents similarity between nodes. The topological structure of the attribute graph represents distribution of email samples in email profiling feature space. Instead of selecting labelled seeds in emails by pure random sampling, a heuristic sampling scheme is applied, allowing automatically locate key emails representing typical campaign profiles. The system propagates label information within the attribute graph, and attribute emails to respective campaigns.

3. METHDOLOGY

Affinity relation between email instances helps to model classes' distribution in the feature space given limited labelled email samples. Similar email samples tend to be categorised into the same class. In our work, we propose a affinity graph based semi-supervised learning method to build classifier for spear-phishing analysis.

3.1 Building spear phishing email profiles

Four categories of email profiling features: origin features, text features, attachment features and recipient features, are used to characterise a spear phishing email.

Origin features. These features are *from domain*, *source ip*, *Autonomous System (AS) number*, *origin country*, *organ-*

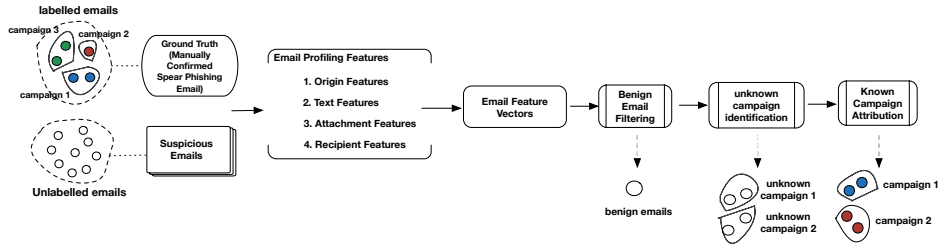


Figure 1: Flowchart of spear phishing email analysis.

isation maintains the AS and *sent date* of the email. Note that Stringhini et al. [23] didn't use origin features as their approach is user centric and did not have access to such information; in contrast, our proposed method assumes that the system has a global view of different spear phishing campaigns. Additionally, time related features are proven useful to identify candidates of spear phishing attacks. Attackers maximise the chance that the victims would read emails during working hours. We extract *day of week* from the *sent date* as an additional categorical feature.

Text features. Spear phishing emails are tuned to simulate casual email communication so as to enable the recipient to "infer the true nature of the email, based on a 'lens' of information (or information cues) that intercedes between the email and the internal perceptions" [26]. Inspired by this, we propose the following features to capture the text characteristics of a spear phishing email.

i) Layout features. We extract length of subject and body text of each email as layout features. These two are the essential visual triggers of a spear phishing email.

ii) Topic features. We merge subject and body text of a spear phishing email and treat the combined text as a document. Latent semantic indexing (LSI) [22] is then applied on the derived documents to find out email topics. We empirically choose top 10 topics and used them as *topic* features for each email in this paper.

iii) Readability features. Emails from different spear phishing campaigns usually have different styles of text writing and organising habits. In light of this, 8 text readability features are used to describe quantitatively the text writing and organisation style . They involve *Count of function words*[21], *Count of complex and simple words*, *Average word length*, *Fog readability index, inverse Fog Index, SMOG index and Flesph-Kincard index (FKRI)*[21].

In addition, we also use categories of *character encoding* of email texts to extend description of email texts' characteristics.

Attachment features. Email attachment is also an import information source. We measure the *size* in bytes and extract the *type* of attachment as features. Additionally, we *malware family* as an additional categorical attribute. In most cases, malware family is not presented. We cluster attachments based on their fuzzy hash similarity, and assign a label to each cluster as a malware family.

Recipient features. These categorical features include *recipient's domain* and *organisation information*. We use them to validate the targeted characteristics of a spear phishing email since the contextual information provided in such emails should match the recipient's organisational interests. Note that recipient's detailed domain information is anonymised for privacy and security reasons.

3.2 Attribute graph propagation based semi-supervised learning

We firstly construct affinity graph of training data instances. Each node of the graph represent an email i ($i = 1, 2, 3..., n$). The attribute of the node is email profiling feature vector x_i as stated in Section 3.1. The edge weight in the graph reflects the similarity between two emails. Each node can be labelled or unlabelled. Note that *labelled* means a spear phishing email has been attributed to a campaign, otherwise it is *unlabelled*. Let $\mathbf{L}$ and $\mathbf{U}$ denote the labelled and unlabelled emails, and importantly, $|\mathbf{L}| \ll |\mathbf{U}|$. y_i denotes the true class label of an email (i.e. to which campaign the email belongs). It is encoded as a M-dimensional 0/1 vector for M-class classification. That is, if the email sample i belongs to the m-th class, $y_{i,j=m} = 1$ and $y_{i,j/m} = 0$. For emails in $\mathbf{L}$, $\{x_i, y_i\}_{i \in L}$, y_i is provided either from human experts or any other sources. Class labels of emails in $\mathbf{U}$ are unknown and treated as the learning target of the proposed model. $\hat{y}_i$ is the estimated class label of each email sample i and is also encoded as a M-dimensional 0/1 vector.

We use K-nearest-neighbour (KNN) graph for learning purpose in our work. That is, each node is connected only to its K nearest neighbouring node in the graph. Similarity measure between linked nodes is calculated using Eq.1 and Eq.2, in order to handle both categorical and numerical attributes in email profiling features:

$$S_{i,j} = \exp\left(-\frac{{D_{i,j}}^2}{\sigma^2}\right) \tag{1}$$

$$D_{i,j} = \gamma \sum_{t \in C} h(x_{i,t}, x_{j,t}) + \sum_{t \in R} d(x_{i,t}, x_{j,t}) \tag{2}$$

where $D_{i,j}$ is the weighted distance between email profiling feature vector x_i and x_j of two emails i and j. $\mathbf{C}$ and $\mathbf{R}$ denote the set of categorical and numerical attributes respectively. $h(\cdot)$ is hamming distance between categorical variables. $d(\cdot)$ is cosine distance between numerical variales. γ is a user-defined weight balancing variable between categorical and numerical attributes in email profiling features.

As shown in Eq.1, the distance $D_{i,j}$ between the two emails is mapped to a similarity score $S_{i,j}$ between 0 and 1 through a gaussian function kernel. Once we have the KNN graph built, class label information is propagated in the graph

Our objective function to solve the class label estimation problem in KNN graph is shown in Eq. 3:

$$\hat{Y} = \overbrace{\min_{\hat{Y}} \sum_{i \in L, m=\{1,2,3..,M\}} \|y_{i,m} - \hat{y}_{i,m}\|^2}^{\text{first term}} + \overbrace{\mathrm{w}_1 \sum_{i \in \{L \bigcup U\}} \sum_{j \in N_i} S_{i,j} \|\hat{y}_{i,m} - \hat{y}_{j,m}\|^2}^{\text{second term}} + \overbrace{\mathrm{w}_2 \sum_{i \in \{L \bigcup U\}} \|\hat{y}_{i,m} - u\|^2}^{\text{third term}} \quad (3)$$

where $\hat{Y} = \{y_i\}, i \in \{L \bigcup U\}$ represents the estimated class labels of both labelled and unlabelled emails. N_i represents all nearest neighbours connected to x_i in the KNN attribute graph. $\|u - v\|^2$ denotes euclidean distance. The *first term* of Eq. 3 requires the estimated class labels $\hat{y}_i$ of the labelled data to be consistent with the corresponding true label y_i of **L**. The *second term*, a graph Laplacian constraint [28], forces similar emails to have same class labels. The *third term* represents uncertainty of class label estimate u is valued as $\frac{1}{M}$. It encourages $\hat{y}_i$ to be assigned to each class with equal probability if not preferred to the contrary by the first two terms. This regularisation term is specially necessary for sparsely connected KNN graph, which prunes connection with low similarity, the side effect is disconnecting some unlabelled nodes from labelled nodes and their neighbours. The third term ensures the class label estimation of such isolated nodes follow uniform distribution. Empirically, these disconnected emails correspond to the emails emerging from previously unknown campaigns. Preserving uncertainty of class assignment is helpful to detect these spear phishing emails of unknown campaigns. w_1 and w_2 are regularisation coefficients, balancing trade-off between the last two regularisation constraints on the class label estimates and the supervised term. Eq.4 gives the solution to Eq.3, iteratively updating class label estimate of the unlabelled emails.

$$\hat{y}_i = \frac{y_i \delta_i + \mathrm{w}_2 u + \mathrm{w}_1 \sum_{j \in N_i} S_{i,j} \hat{y}_j}{\delta_i + \mathrm{w}_2 + \mathrm{w}_1 \sum_{j \in N_i} S_{i.j}} \quad (4)$$

where $\delta_i = 1$ if $i \in \mathbf{L}$ (0, otherwise). Class label of each unlabelled node x_i is finally decided using Eq.5:

$$m = \arg \max_{m=1,2,..,M} \hat{y}_{i,m} \quad (5)$$

4. EXPERIMENTS

4.1 Dataset

Datasets. We obtain two datasets from Symantec's enterprise email scanning service[1]. Spear phishing email dataset, denoted as **S**, contains 1,467 spear phishing emails from 8 campaigns. Each email is manually examined by security experts and attributed to a specific campaign with strong confidence. Campaigns vary from large persistent campaign like 'Elderwood' to small scale 'darkmoon' campaign. Benign email dataset, denoted as **B**, contains 14,043 emails. These emails were also sent between 2011 and 2013, and have attachments. It is important to note we preserve *class imbalance* between spear phishing email dataset and benign email dataset, and among different campaigns within spear phishing email dataset as well. For example, 'Elderwood' campaign contains more than 60 times of emails than that of 'bisrala' campaign in our data set (as shown in Table 1), and benign emails are 10 times the size of spear phishing emails. It is purposely designed to simulate imbalanced data, a common issue in practical security applications, in order to test real world robustness of the proposed semi-supervised learning framework besides its learning accuracy.

Campaign name	Number of emails
krast	157
CommentCrew/APT1	125
layork	139
Elderwood	770
nitro	153
bisrala	12
darkmoon	33
samkams	78

Table 1: Summary of collected spear phishing email samples

4.2 Experiments Overview

Organisation of the experiments. We design three experiments to verify the validity of the proposed semi-supervised learning approach.

- **Experiment 1** - Spear phishing email identification (Section 4.3) aims at differentiating spear phishing email from benign emails. That is, given a set of emails mixed with both benign emails and spear phishing emails, the learning framework identifies spear phishing emails while filtering out benign ones.
- **Experiment 2** - Unknown campaign identification (Section 4.4) is designed to verify whether the detected spear phishing emails are from previously unknown campaigns. Newly emerging campaigns indicate potential zero-day exploits, and they are of high interests for further security analysis.
- **Experiment 3** - Known campaign attribution (Section 4.5) focuses on analysing spear phishing at finer granularity. The learning framework attributes spear phishing emails into specific known campaigns.

We employ random forest, a popular supervised learning method as baseline for comparison purpose in Experiment 1 and Experiment 3. It is important to note that supervised

[1]All PII and customer information are anonymised in these two datasets.

learning methods, such as random forest, can not identify unknown campaigns. We provide detailed explanation in Experiment 2.

Evaluation Methodology. For each experiment, we have a pre-defined percentage list $P = \{p_1, p_2, ..., p_k\}$, $p_i \in (0, 1)$ and $p_i < p_j$ if $i < j$. Each p_i controls the number of labelled emails extracted from **S** and **B**. A smaller p_i indicates less manual effort required to label spear phishing emails by security experts. We sample **S** and/or **B** with replacement for a given percentage p_i to form **L**, and treat the rest of emails as **U**. We then train the proposed method and the baseline method using **L**, generate the performance metrics with respect to the test data **U**, and repeat this process 20 times. We average their performance metrics and use them for comparison study.

Parameters. For Experiment 1, we set the number of class M to be 2. For Experiment 3, M is set to the number of spear phishing campaigns. For Experiment 2, since there is only one class 'known campaign' existing in the labelled data, the true class label y_i of each labelled email sample i in **L** is a scalar value in this case and set to be 1 for computational convenience in our work. The estimated class label $\hat{y}_j$ of each unlabelled email sample j is also a scalar value. The closer the derived $\hat{y}_j$ is to 1, the more likely that the corresponding unlabelled email j comes from one of the known campaigns. For decision purpose, we choose the best threshold on the derived $\hat{y}_j$ to achieve optimal trade-off between precision and false positive rate. In all experimental analysis, γ (see Eq.2) is set to be 0.7 and the number of nearest neighbours K is set to be 10.

Evaluation metrics. Since the true labels of emails in both **S** and **B** are known, we are able to use quite a few evaluation metrics to measure the performance. We report recall and false positive rate (FPR) in Experiment 1 and Experiment 2. Recall directly evaluates identification/detection sensitivity, while FPR measures reliability of the identification/detection results. For Experiment 3, since campaign attribution is a multi-class classification test, in addition to recall and FPR, we also apply F1 score [7] to form comprehensive evaluation of classification performance.

4.3 Spear Phishing Email Identification

Experimental Setup. We uniformly sample a given percentage **p** from **S** and **B** (Section 4.1) to generate the initial labelled data **L**. Spear phishing emails are labelled with '+1' and benign emails are labelled with '-1' in **L**. The rest of the data in **S** and **B** are treated as unlabelled data **U** and used as our test data. The purpose of using a percentage **p** is simulating limited labelled training data. We ensure that at least one email is sampled for each class (i.e. benign and spear phishing) in **L**. Random forest is used as the baseline method to evaluate the merits of the proposed semi-supervised learning method. In this task, a 2000-tree random forest[2] is trained on **L** and applied on **U** to identify spear phishing emails.

Experimental Results. We report the experimental results in Table. 2 as p is chosen from 1%, 2% 3%,4%,5% and 6% respectively. When the number of the labelled data **L** is limited, e.g. when p is equal to 1% or 2%, recall of random forest is distinctively lower than 80%. The major cause is the class imbalance between spear phishing emails and benign emails in the data set. There are 9 times more benign emails than spear phishing emails in our data. Especially when limited labelled emails (i.e. **L**) are available, class imbalance issue becomes even more problematic to supervised methods, i.e. random forest in this case, since they are trained directly on the imbalance set easily biased towards benign class, consequently giving low identification precision for spear phishing class.

Notably, our proposed method achieves superior identification accuracy for all 6 levels of labelled data percentage. Especially, when p is fixed to 1% (that is, using only 11 out of original 1,467 spear phishing emails), our method manages to achieves a high recall of 90%, while keeping a low FP less than 0.02%. As shown in Table. 2, our method has a stable recall level when p is greater than 4%, with recall larger than 96%. The results demonstrate that our proposed method can effectively identify spear phishing emails using *limited* and *imbalanced* training data.

4.4 Unknown Campaign Detection

Few supervised learning methods handle this one-class density estimate problem in previous machine learning research. One-class SVM [4] and one-class random forest [6] are the most popular variants of supervised learning applied to attack this issue. However, one-class SVM can not handle categorical types of input attributes due to the limitation of SVM design. One-class random forest needs strong prior knowledge about data distribution of unseen class samples, which is not presented in unseen spear phishing campaign detection. Both methods don't fit our requirements. Therefore, we don't involve a supervised learning baseline in the experiment.

Experimental Setup. We choose 'krast', 'CommentCrew/ APT1', 'layork', 'Elderwood' and 'nitro' campaigns in **S** as *known campaigns*, denoted as **KC** and treat the other three campaigns as *unknown campaigns* to be detected, denoted as **UC**. **KC** is randomly divided into two parts, $\mathbf{KC}_L$ and $\mathbf{KC}_U$. $\mathbf{KC}_L$ forms our labelled data **L** simulating limited attributed spear phishing emails in threat intelligence investigation. $\mathbf{KC}_U$ together with **UC** forms our unlabelled data **U**, simulating *to-be-investigated* spear phishing emails containing both known and unknown campaigns. Each email sample in **L** is labelled as '+1' (i.e. *known*) regardless specific campaigns they belong to.

In this experiment, we require that **L** is minimum 25% of **KC**. There are two major reasons to justify such requirement. First of all, different from binary or multi-class classification, more labelled data are needed to describe distribution of the known campaigns. False positive rate becomes too high for detection use when emails from known campaigns are extremely limited. Secondly, unknown campaign detection is used as a successive step following the identification phase (Experiment 1). Its purpose is to verify the existence of novel campaigns. Hence it doesn't require extra labelling efforts from domain experts. Nevertheless, we assume that $|\mathbf{L}|<|\mathbf{KC}|$ and **UC** are less than 8% of the whole

[2]2000-tree random forest is chosen as it empirically as yield the best classification result.

p	Avg. no. of labelled emails	Avg. recall (RF)	Avg. FPR (RF)	Avg. recall (SSL)	Avg. FPR (SSL)
1%	11	0.7987	3.6247e-05	**0.8700**	4.3950e-05
2%	21	0.8000	4.3632e-05	**0.8910**	4.4910e-05
3%	32	0.8323	4.7608e-05	**0.9106**	5.1077e-05
4%	43	0.9251	6.3137e-05	**0.9460**	7.0652e-05
5%	53	0.9399	9.0109e-05	**0.9607**	9.0109e-05
6%	65	0.9483	8.7266e-05	**0.9707**	8.7266e-05

Table 2: Average spear phishing email identification performance metrics

Unknown campaign	Percentage of unknown emails in U	Avg. Recall	Avg. FP
darkmoon	3.27%	**100%**	1.34%
samkams	7.73%	**97.44%**	0.87%
bisrala	1.19%	**91.67%**	5.65%

Table 3: Average unknown campaign detection performance of the semi-supervised learning method

p	Average F1 score (RF)	Average F1 score (SSL)
1%	0.6829	**0.8169**
2%	0.7928	**0.9031**
3%	0.8153	**0.9084**
4%	0.8748	**0.9276**
5%	0.9008	**0.9331**
6%	0.9038	**0.9326**

Table 4: Average F1 score of campaign classification

unlabelled email set **U**. The purpose is to demonstrate that the proposed method reduces considerable labelling efforts while preserving high precision for the unknown campaign detection problem and observe how it performs given class imbalance between the known (1,344 emails) and unknown campaigns (132 emails).

Experimental Results. Table. 3 illustrates the experimental results. Using only 336 known campaign emails, the proposed semi-supervised learning method provides excellent detection accuracy for all three campaigns. No sample of 'darkmoon' escapes from detection. Over 97% of 'samkams' and 91% of 'bisrala' are detected correctly. The false positive rates are limited. For 'darkmoon' campaign, only 13 email samples among 1,008 spear phishing emails of known campaigns are misclassified as 'unknown'. For 'samkams' campaign detection, only 9 emails among 1,008 emails from known campaigns are misclassified as 'unknown'. For 'bisrala' campaign detection, false positive rate is relatively higher that the other two, while still at a ver low ratio. 56 out of 1,008 email samples of known campaigns are misclassified. Notably, 'bisrala' campaign has only 12 samples. 91.67% recall means only one false rejection out of 12 email samples in this campaign.

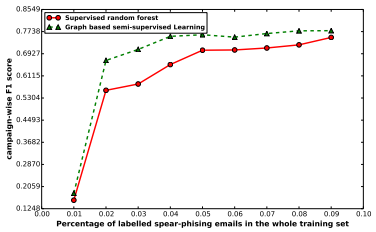
4.5 Campaign Attribution

Experimental Setup. Following the setup in experiment 2, we apply the proposed semi-supervised learning framework on 1,344 emails from the 5 largest spear phishing campaigns, denoted as **KC**. The rationale behind this setup is that all input emails experiment 3 shall be spear phishing emails from known campaigns (see Fig. 1). Initial labelled data **L** are sampled from **KC** uniformly with a given percentage p. **L** must contain at least one sample from each campaign and their campaign names are used as the labels. The rest of **KC** (i.e. 1,344 *(1-p)) form the unlabelled set **U**. The learning target is to estimate campaign labels for emails in **U**. The supervised baseline solution remains a 2000-tree random forest built on the labelled data **L**.

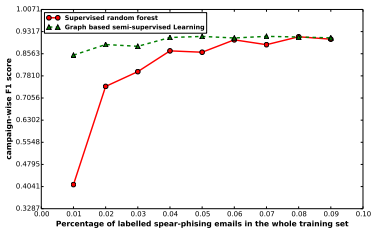
Experimental Results. Table. 4 lists average of the overall multi-class F1 score with respect to all 5 campaigns at each percentage level. From a global viewpoint, our system performs consistently better than its supervised counterpart for all given **p**. Especially, with merely 25 out of 1,344 emails, (2% of the whole dataset), our semi-supervised learning method can improve F1-score from random forest's 0.79 to 0.90. And our method keeps producing a stably high classification accuracy when the number of the labelled seeds is over 4% of the whole dataset, around 53 labelled emails out of 1,344 emails.

Figure. 2 illustrates campaign-wise F1 scores. Our system achieves significant improvement of classification accuracy at all campaigns given labelled emails **L** using no more than 54 emails (p=4%) from 5 campaign as seeds. Especially, given p=2%, our system achieves 100% better F1 score than the supervised random forest for 'CommentCrew/APT1' and 'Layork'. For 'Elderwood' campagin, our system achieves almost perfect classification with F1 score close to 1 using merely 14 emails (p=1%) from 5 campaigns. Both learning methods provide similar accuracy for 'nitro' campaign. However, the improvement of F1 score for 'nitro' is more significant when p ranges from 2% to 4% (e.g. size of **L** is between 26 and 54 emails). It demonstrates that our affinity graph based semi-supervised learning model can achieve effective campaign attribution given only limited labelled emails, which in turn, reduces the overheads of manual labelling efforts.

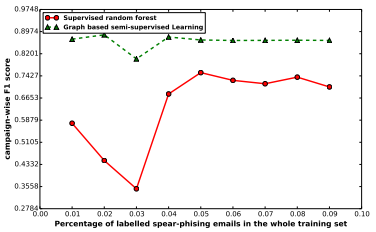
We demonstrate average confusion matrices of the semi-supervised learning method and random forest in Table. 5, given p=4% (e.g. **L** has 54 emails). For both methods, the campaign 'layork' is often misclassified as the campaign 'krast', which leads to deterioration of attribution performance. Consistently, as shown in Figure. 2, these two campaigns are the most difficult classes to distinguish in our work. The emails from these two campaigns share considerable similarity such as 'body text', readability indices,



(a) Krast campaign.



(b) CommenCrew/APT1 campaign.



(c) Layork campaign.

Figure 2: Average campaign-wise F1 score with different number of labelled emails.

Campaign class	Krast	CommentCrew/APT1	Layork	Elderwood	nitro
Krast	68.75%	12.32%	2.85%	5.95%	10.13%
CommentCrew/APT1	19.87%	72.90%	0.69%	4.60%	1.94%
Layork	21.66%	17.61%	49.35%	5.86%	5.52%
Elderwood	0.1%	0%	0%	99.90%	0%
nitro	0.52%	0%	0%	1.15%	98.33%

(a) Confusion matrix of the supervised random forest

Campaign class	Krast	CommentCrew/APT1	Layork	Elderwood	nitro
Krast	**75.87%**	10.35%	5.45%	4.17%	4.16%
CommentCrew/APT1	7.55%	**90.66%**	1.74%	0%	0.05%
Layork	16.01%	8.23%	**71.21%**	0.10%	4.45%
Elderwood	0.25%	0.01%	0%	99.74%	0%
nitro	0.55%	0%	0%	0.82%	**98.63%**

(b) Confusion matrix of the semi-supervised learning method

Table 5: Confusion matrix based campaign-wise classification performance measurement

etc. When two spear phishing emails from two different campaigns share similar features, e.g. topic, readability features, malware related features are overwhelmed by the aforementioned text features. Furthermore, when emails are extremely short, like one sentence or several phrases, the text features become much less stable and less informative for classification. One potential solution is to weight complementary email features to build classifiers. This is part of our future work.

5. ONLINE DETECTION RESULTS AND DISCUSSION

Online Detection. We put our system into an operational test to detect unknown spear phishing campaigns. We trained our model using 2% of **S** and run it in an online mode to filter incoming suspicious emails. Since we are simulating online detection, we use a collection of 1,534 emails (collected in January 2014) from Symantec enterprise services as the test data. These emails have been investigated by security experts but the true labels are withheld during the test phase. Our system identifies an unknown campaign with 5 emails, which are later confirmed as the 'waterbug' campaign [24]. In terms of 'waterbug' emails, they have both distinct email contents and different attachments compared with the other campaigns (e.g. the campaigns of **S** that we used to train the model). Surprisingly, one of the 5 'waterbug' emails have a different type of attachments (in RAR format) from the other 4 (in PDF format). However, features relating to layouts and semantic characteristics are able to compensate the difference of email attachments during the affinity graph propagation process. As a result, all 'waterbug' emails present high intra-class similarity and are well separated from the other campaigns.

Discussion. In general, our system are able to identify the most informative email profiling features to attribute campaigns. These feature include *length of email subject*, *length of email body text*, *readability features of email body text*, *sender's IP address and the corresponding Autonomous System number*, *character encoding*, *file size*, *file type* and *malware family of email attachments*. Attachment related features, especially malware families and file types of the email attachments, are highly correlated to the concrete exploiting strategies of the spear-phishing attacks. For example, a spear phishing campaign is likely to use a specific exploit and pack the code in a generic format (e.g. PDF, Excel) available to the users. Since our system uses fuzzy hashes to group email attachments, not surprisingly, these features are thus good indicators of spear-phishing campaigns. Moreover, spear phishing campaigns disguise themselves as emails from an individual or organisation that the recipients should know or be interested in. Layouts and semantic characteristics of email text, including email readability, email topics, length of email texts and the character encoding schemes, describe psychological pertinence of spear phishing attacks to the targeted users [26]. These features are also ranked high in our proposed model. Interestingly, sender's IP ad-

dress and AS number features also contribute to campaign attribution. We partially speculate that the attackers are likely sent a small batch of emails from the same address due to the targeted nature of such campaigns.

6. CONCLUSIONS

In this paper, we developed an affinity graph based semi-supervised learning approach based on a well designed email profiling features. The proposed model can effectively identify spear phishing emails, detect emails of previously unknown campaigns and attribute them to known campaigns accurately. Extensive experiments show the proposed semi-supervised learning performs better than the state-of-the-art supervised learning based solution by reducing manual labelling overheads to much extent and preserving high classification accuracy at the same time. Our method is insensitive to concrete distribution forms of campaigns in email profiling feature space, which makes the proposed method robust against evolution of spear phishing campaigns.

7. REFERENCES

[1] C. Abad. The economy of phishing: A survey of the operations of the phishing market. *First Monday*, 10(9), 2005.

[2] S. Afroz, M. Brennan, and R. Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *S&P*, pages 461–475. IEEE, 2012.

[3] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artif. Intell. Rev.*, 29(1):63–92, Mar. 2008.

[4] A. J. B.Scholkopf, R.C.Williamson and J.C.Platt. Support vector method for novelty detection. In *NIPS*, pages 582–588, 1996.

[5] D. Caputo, S. Pfleeger, J. Freeman, and M. Johnson. Going spear phishing: Exploring embedded training and awareness. *IEEE S&P*, 12(1):28–38, Jan 2014.

[6] C. C.Desir, S.Bernard and L.Heutte. One class random forests. *Pattern Recognition*, pages 3490–3506, 2013.

[7] D.Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *J. of Machine Learning Technologies*, pages 37–63, 2011.

[8] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, 1999.

[9] N. Feamster, A. G. Gray, S. Krasser, and N. A. Syed. Snare: Spatio-temporal network-level automatic reputation engine. 2008.

[10] J. Hong. The state of phishing attacks. *Commun. ACM*, 55(1):74–81, Jan. 2012.

[11] J. Iedemska, G. Stringhini, R. Kemmerer, C. Kruegel, and G. Vigna. The tricks of the trade: What makes spam campaigns successful? In *Security and Privacy Workshops (SPW)*, pages 77–83. IEEE, 2014.

[12] R. Jabeur Ben Chikha, T. Abbes, W. Ben Chikha, and A. Bouhoula. Behavior-based approach to detect spam over ip telephony attacks. *International Journal of Information Security*, pages 1–13, 2015.

[13] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, Oct. 2007.

[14] M. Jakobsson and S. Myers. *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft.* Wiley-Interscience, 2006.

[15] M. Khonji, Y. Iraqi, and A. Jones. Phishing detection: A literature survey. *Communications Surveys Tutorials, IEEE*, 15(4):2091–2121, Fourth 2013.

[16] T. Micro. Spear-phishing email: Most favored apt attack bait. http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-spear-phishing-email-most-favored-apt-attack-bait.pdf.

[17] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song. On the feasibility of internet-scale author identification. In *S&P*, pages 300–314. IEEE, 2012.

[18] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich. A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise. *Computer Networks*, 59:101–121, 2014.

[19] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. M. Voelker, V. Paxson, N. Weaver, and S. Savage. Botnet judo: Fighting spam with itself. In *NDSS*, 2010.

[20] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *CCS*, pages 342–351. ACM, 2007.

[21] R.Shams. Classifying spam emails using text and readability features. In *ICDM*, pages 657–666. IEEE, 2013.

[22] S.Deerwester. Improving information retrieval with latent semantic indexing. In *ASIS&T*, pages 36–40, 1988.

[23] G. Stringhini and O. Thonnard. That ain't you: Blocking spearphishing through behavioral modelling. In *DIMVA*, 2015.

[24] Symantec. The waterbug attack group. http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/waterbug-attack-group.pdf.

[25] O. Thonnard, L. Bilge, G. O'Gorman, S. Kiernan, and M. Lee. Industrial espionage and targeted attacks: Understanding the characteristics of an escalating threat. In *RAID*, pages 64–85, 2012.

[26] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao. Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE Trans. Prof. Communication*, 55(4):345–362, 2012.

[27] G. Xiang, J. Hong, C. P. Rose, and L. Cranor. Cantina+: A feature-rich machine learning framework for detecting phishing web sites. *ACM Trans. Inf. Syst. Secur.*, 14(2):21:1–21:28, Sept. 2011.

[28] J. X.Zhu and Z.Ghahramani. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.