EURASIP Journal on Information Security
a SpringerOpen Journal

RESEARCH Open Access

Inside the scam jungle: a closer look at 419 scam email operations

Jelena Isacenkova[1*], Olivier Thonnard[2], Andrei Costin[1], Aurélien Francillon[1] and David Balzarotti[1]

Abstract

419 scam (also referred to as Nigerian scam) is a popular form of fraud in which the fraudster tricks the victim into paying a certain amount of money under the promise of a future, larger payoff.
Using a public dataset, in this paper, we study how these forms of scam campaigns are organized and evolve over time. In particular, we discuss the role of phone numbers as important identifiers to group messages together and depict the way scammers operate their campaigns. In fact, since the victim has to be able to contact the criminal, both email addresses and phone numbers need to be authentic and they are often unchanged and re-used for a long period of time. We also present in detail several examples of 419 scam campaigns, some of which last for several years - representing them in a graphical way and discussing their characteristics.

Keywords: Email spam; Nigerian scam; 419 scam; Cyber crime; Campaigns

1 Introduction

Nigerian scam [1], also called '419 scam' as a reference to the 419 section in the Nigerian penal code [2], has been a known problem for several decades. The name encompasses many variations of this type of scam, like advance fee fraud, fake lottery, black money scam, etc. Originally, the 419 scam phenomenon started by postal mail, and then evolved into a business run via fax first, and email later. 419 scam is a popular form of fraud in which the fraudster tricks the victim into paying a certain amount of money under the promise of a future, larger payoff. The prosecution of such criminal activity is complicated [3] and can often be evaded by criminals. As a result, reports of such crime still appear in the social media and online communities, e.g. *419scam.org* [4], exist to mitigate the risk and help users to identify scam messages.

Nowadays, 419 scam is often perceived as a particular type of *spam*. However, while most of the spam is sent today in bulk through botnets and by compromised machines, 419 scam activities are still largely performed in a manual way. Moreover, the underlying business and operation models differ. Spammers trap their victims through engineering effort, whereas scammers rely on human factors: pity, greed, and social engineering techniques. Scammers use very primitive tools (if any) compared with other forms of spam where operations are often completely automated. A distinctive characteristic of email fraud is the communication channel set up to reach the victim: from this point of view, scammers tend to use emails and/or phone numbers as their main contacts [5], while other forms of spam are more likely to forward their victims to specific URLs. For instance, a previous study of spam campaigns [6] (in which scam was considered a subset of spam) indicates that 59% of spam messages contain a URL. However, even though 419 scam messages got eclipsed by the large amount of spam sent by botnets, they still pose a persistant problem that causes substantial personal financial losses for a number of victims all around the world.

The traditional spam and scam (not 419) scenarios have been already thoroughly studied (e.g. [6,7]), where a big part of existing unsolicited bulk email identification techniques rely on high volumes of similar messages. However, 419 messages are more likely to be sent in lower copies and from webmail accounts. Thus, criminals aim to stay unnoticed by the traditional spam filters and avoid drawing attention to abused webmail accounts. The exact distribution methods of 419 scam messages have not been studied as deeply as, for example, the distribution of botnet spam. However, based on Microsoft Security Intelligence

*Correspondence: jelena.isachenkova@eurecom.fr
[1] Eurecom, Route des Chappes, Sophia Antipolis, France
Full list of author information is available at the end of the article

Reports [8], 419 scam messages constitute on average to 8% of email spam traffic (data over the last 5 years).

A recent study by Costin et al. [5] describes the use of phone numbers in a number of malicious activities. The authors show that the phone numbers used by scammers are often active for a long period of time and are reused over and over in different emails, making them an attractive feature to link together scam messages and identify possible campaigns. In this work, we test this hypothesis by using phone numbers and other email features to automatically detect and study scam campaigns by using a public dataset. To the best of our knowledge, this is the first in-depth study of 419 email campaigns. While a preliminary version of this study was published in [9], this is an extended version of the study.

Our goal is to study how scammers orchestrate their scam campaigns, by looking at the interconnections between email accounts, phone numbers, and email topics used by scammers. To this aim, we use a novel multi-criteria decision algorithm to effectively cluster scam emails that are sharing a minimal number of commonalities, even in the presence of more *volatile* features. Because of this set of commonalities, scam emails originating from the same scammer(s) are likely to be grouped together, enabling us to gain insights into scam campaigns. Additionally, we also evaluate the quality and consistency of our clustering results. To this aim, we perform a threshold sensitivity analysis, as well as evaluate the homogeneity of clusters using the graph *compactness* and Adjusted Rand Index as metrics.

In our analysis, we have identified over 1,000 different campaigns and, for most of them, phone numbers represent the cornerstone that allows us to link the different pieces together. We also discovered some larger-scale campaigns (so-called 'macro-cluster'), which are made of loosely inter-connected scam clusters reflecting different operations of the same scammers. We believe these can be attributed to different scam runs orchestrated by the same criminal groups, as we observe the same phone numbers or email accounts being reused across different sub-campaigns.

As demonstrated by our experiments, our methods and findings could be leveraged to *pro-actively* identify new scam operations (or variants of previous ones) by quickly associating a new scam to ongoing campaigns. We believe that this would facilitate the work of law enforcement agencies in the prosecution of scammers. Our approach could also be leveraged to improve investigations of other cybercrime schemes by logging and investigating various groups of cybercriminals based on their online activities. In this regard, our methodology already proved its utility in the context of other security investigations, such as in the analysis of rogue AV campaigns [10], spam botnets operations [11], and targeted attacks [12].

The rest of the paper is organized as follows: We start by describing the scam dataset (Section 3), to which we apply our cluster analysis technique to extract scam campaigns, and compare the usage of email addresses and phone numbers (Section 4). In Section 5, we focus on a number of individual campaigns to present their characteristics. Finally, we draw our conclusions in Section 6.

2 Related work

Scammers employ various techniques to harvest money from ingenuous victims. Tive [13] introduced the tricks of 419 fee fraud and the philosophy of the tricksters behind. Stajano and Wilson [14] studied a number of scam techniques and demonstrated the importance of security engineering operations. Herley [15] analysed attack decisions as binary classification problems studying the case of 419 scammers. The author looks into the economical aspects of adversaries trying to understand how scammers find viable victims out of millions of users, so that their business still remains profitable. A brief summary of 419 scam schemes was presented by Buchanan and Grant [3] indicating that Internet growth has facilitated the spread of cyber fraud. They also emphasized the difficulties of adversary prosecution - one of the main reasons why 419 scam is still an issue today. A more recent work by Oboh et al. [16] discussed the same problem of prosecution in a more global context taking the Netherlands as an example.

Another work by Goa et al. [17] proposed an ontology model for scam 419 email text mining demonstrating high precision in detection. A work by Pathak et al. [6] analysed email spam campaigns sent by botnets, describing their patterns and characteristics. The authors also showed that 15% of the spam messages contained a phone number. A recent patent has been published by Coomer [18] on a technique that detects scam and spam emails through phone number analysis. This is the first mentioning of phone numbers being used for identifying scam, but with no technical implementation details. Costin et al. [5] studied the role of phone numbers in various online fraud schemes and empirically demonstrated it's significance in 419 scam domain. Our work extends the study by focusing on scam email and campaign characterization that relies on phone numbers and email addresses used by scammers.

3 Dataset

In this section, we describe the dataset we used for analyzing 419 scam campaigns and provide some statistics of the scam messages. There are various sources of scam often reported by users and aggregated afterwards by dedicated communities, forums, and other online activity groups. The data chosen for our analysis come from `419scam.org` - a 419 scam aggregator - as it provides a large set of preprocessed data: email bodies, headers, and

some already extracted emails attributes, like the scam category and the phone numbers. Note that IP addresses data are absent. We downloaded the emails for a period spanning from January 2009 until August 2012.

In our study, we also exploited the fact that the phone numbers can indicate a geographical location, typically the country where the phone is registered. Although it does not prove the origin of the message or the scammer, still it references a country of a scam operation, and improves victim's level of confidence in the received message. For example, receiving a new partnership offer from UK could seem suspicious if the phone contact has a Nigerian prefix, or a fake lottery notification with contact details originating from an African country while the victim being from Europe. Moreover, as shown in a previous study [5], 419 scam mobile phone numbers are precise in indicating the country of residence of the phone owner (scammer) as few roaming cases were found. Therefore, the phone attribute is precise enough to indicate geographical origins.

The resulting dataset consists of 36,761 messages with 11,768 unique phone numbers. The general statistics of the data are shown in Table 1. A first thing to notice is that the number of email addresses is three times bigger than the number of phone numbers, emphasizing the facility to acquire mailboxes for malevolent purposes. However, still the ratio is quite low indicating rather cheap and easy access to the phone numbers. Another specificity of this dataset is that each message can contain several email addresses and phone numbers, where the number of different email addresses can be as high as five per message: *from* email address, *reply* email, and other email addresses indicated in the body of the messages. Hence, although we collected only 36,761 messages, we extracted 112,961 email addresses from them.

In our dataset, we did not notice any significant bursts of scam messages (verified on a monthly basis) during the three year span, suggesting that the email messages were constantly distributed over time. It is also important to note that the dataset is mostly limited to the European and African regions (with only a few Asian samples), which is due to the way the website owners are collecting and classifying the data. Nevertheless, the geographical distribution of the mentioned continents is reflected in our dataset, excluding only some minor actors.

To better understand the dataset, we look at the time during which emails and phones were advertised by scammers in scam messages. Seventy one percent of the email addresses in our dataset were used only during 1 day. The remaining were used for an average duration of 79 days each. Phone numbers have a longer longevity than email addresses: 51% of the phone numbers were used only for 1 day; the rest were used on average for 174 days (around 6 months) hence making it an important feature in our data clustering analysis.

Table 2 summarizes the phone number geographical distribution. UK numbers are twice as common as Nigerian, and three times more common than the ones from Benin, the third biggest group. The Netherlands and Spain are the leading countries in Europe. Note that UK should be considered as a special case. As reported by *419scam.org* and Costin et al. [5], all UK phone numbers in this dataset belong to *personal numbering services* - services used for forwarding phone calls to other phone numbers and serving as a masking service of the real destination for the callee. In our dataset there are 44% of such phone numbers (all with UK prefix), another 44% are mobile phone numbers, 12% are fixed lines [5], and only less than 1% of the phones are non-existent.

The initial 419 scam messages are also labeled by the *419scam.org* [4] with a category. Around 64% of the emails are assigned to the category named '419 scam', which is a subcategory of 419 scam and refers specifically to financial fraud types, e.g., transactions, lost funds, etc. As reported by [4], most of the remaining emails (24%) belong to 'Fake lottery' category. However, this distribution has been changing over time as shown in Figure 1. Especially, a big difference can be observed between 2009

Table 1 General statistics table

Description	Numbers
Scam messages	36,761
Unique messages	26,250
Total email addresses	112,961
Unique email addresses	34,723
Total phone numbers	41,320
Total unique phone numbers	11,768
Number of countries	12

Table 2 Phones by countries

Country	Total phones	Total (%)
United Kingdom	4,499	43
Nigeria	3,121	30
Benin	1,448	14
South Africa	562	5
Spain	372	4
Netherlands	263	3
Ivory Coast	89	1
China	68	1
Senegal	47	0.5
Togo	11	0.1
Indonesia	1	0.01

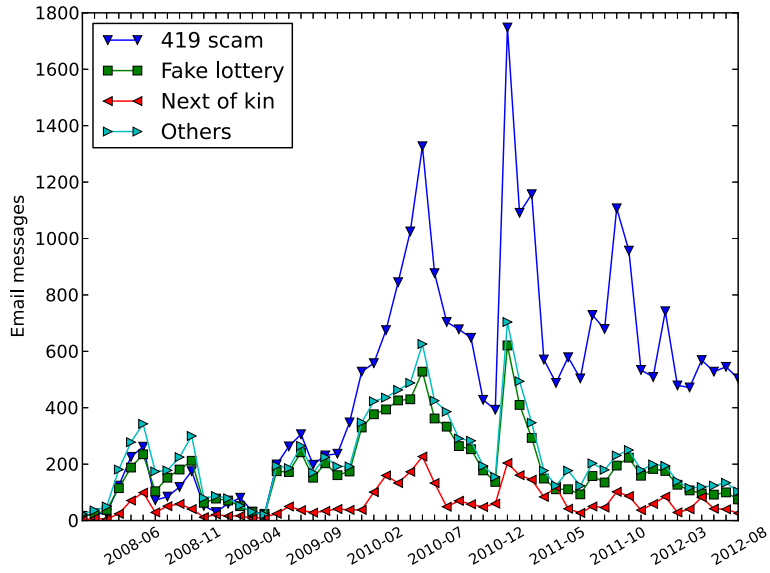

Figure 1 Scam email categories over time.

and 2011, where in 2011, the '419 scam' became a dominant category. As of August 2012, there was five times more emails of '419 scam' than of 'fake lottery' letters. This might be due to an outdated categorization process, as scam topics - like spam - may change and evolve over time. For this reason, in the next section, we describe our process to automatically identify the scam topics based on the word frequencies in the messages. We also observe that most of the 'fake lottery' scams are associated with European phone numbers suggesting that this category is sent to a targeted audience. In the majority of '419 scam' cases, scammers use as many Nigerian phone numbers (Figure 2) as of UK ones. African phone numbers (Figure 2) with UK share being equivalent to Nigerian. Also notice that Benin becomes a much bigger player starting from the beginning of 2011. Hence, the geographical targets may vary with the topics and objectives persuaded by the accomplices.

4 Data analysis methods

In this section, we describe methods used for identifying groups of similar 419 scam emails that are believed to be part of the campaigns and then present the results. We use two different metrics to evaluate the quality and consistency of the created clusters (campaigns). Finally, we extract the most repetitive keywords from the body of the scam messages in order to improve their topic categorization.

4.1 Scam email clustering: the TRIAGE approach

To identify groups of scam emails that are likely part of a campaign orchestrated by the same group of people, we have clustered all scam messages using TRIAGE - a software framework for security data mining that takes advantage of multi-criteria data analysis to group events based on subsets of common elements (subsequently called *features*). Thanks to this multi-criteria clustering approach, TRIAGE identifies complex patterns in data, unveiling sometimes *varying* relationships among series of connected or disparate events. TRIAGE is best described as a security tool designed for intelligence extraction helping to determine the patterns and behaviors of the intruders (*i.e.*, the *tactics*, *techniques* and *procedures*, or TTPs), highlighting 'how' they operate rather than 'what' they do. The framework [19] has already demonstrated its utility in the context of other security investigations, *e.g.*, rogue AV campaigns [10], spam botnets [11] and targeted attacks [12].

Figure 3 illustrates the TRIAGE workflow, as applied to our scam dataset. In step 1, a number of email characteristics (or *features*) are selected and defined as decision criteria for linking the emails. Such characteristics include the sender email address (the *from*), the email *subject*, the sending *date*, the *reply* address (as found in the email header), the *phone* number and any other *email address* found in the message itself (*email body*). In step 2, TRIAGE builds relationships among all email samples with respect to the selected features using appropriate similarity metrics. More specifically, we used various string-oriented similarity measures commonly used in information retrieval, such as the Levenshtein similarity (for the *subject*) and the *N-gram* similarity (for features as *from*, *reply*, *email body*). As shown in [20], N-gram similarity can be seen as a generalization of Levenshtein and

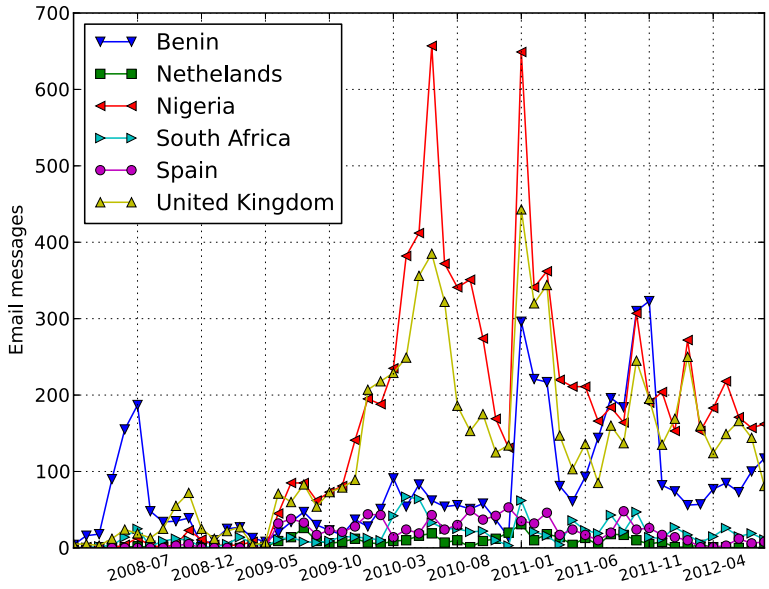

Figure 2 '419 scam' category phone numbers over time by countries.

has proved to work better with short strings. However, N-gram can be computationally expensive. For this reason, we decided to use it solely for email addresses, for which it is somehow more critical to have a reliable comparison method, than for email *subjects* which tend also to be longer and thus more expensive to compare. For features as *phone* and *date*, we simply used the equality comparison method - as this method appeared to us to be the most appropriate to capture the semantics between two different feature values in this particular case.

At step 3, the individual feature similarities are fused using an *aggregation model* reflecting a high-level behavior defined by the analyst, who can impose, *e.g.*, that *at least k* highly similar email features (out of *n*) are required to attribute different samples to the same campaign. The tool allows to assign different *weights* to the features, so as to give higher or lower importance to certain features. Table 3 shows the particular set of weights used for this analysis, in which we emphasize the importance of the phone numbers and the email subjects. The features related to the sender email addresses were given a medium importance, whereas the sending *date* was given a much lower importance. The fused similarity value (or *aggregated* value) is then used as input to a classical *graph clustering* algorithm, such as minimal cut or connected components, which are able to recover clusters of arbitrary size and shape.

At this stage, it is important to stress that, except maybe for the *phone number* used by the scammer (unless this number is fake), *none* of the other features included in the clustering analysis can be considered *alone* to be sufficient to attribute scam emails to the same criminals. For example, the fact that the same (or similar) sender email address was used in two scam messages, even sent on the same

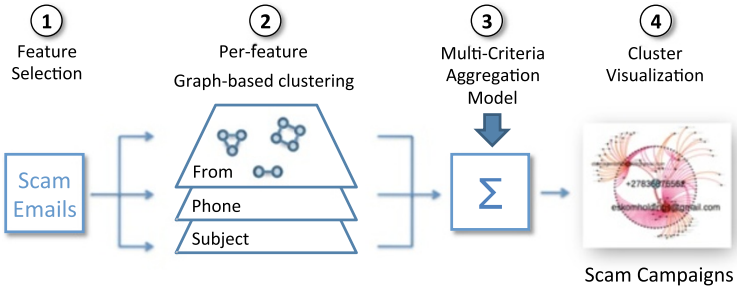

Figure 3 TRIAGE workflow on scam dataset.

Table 3 Weights of individual features $(\sum = 1)$

Feature	Importance
Phone	0.30
From	0.12
Reply	0.18
Subject	0.25
Email body	0.10
Date	0.05

date, does not necessarily mean that these two messages originate from the same individual(s). Only one or two common features can appear merely due to a coincidence, to a commonly seen pattern, or perhaps to some commonly used technique shared by scammers. This motivates our choice for a multi-criteria fusion model that incorporates various weights defined according to a fuzzy *linguistic quantifier* (also referred to as *Regular Increasing Monotone* or RIM quantifier [21,22]), which allows us to model fusion strategies such as *at least k* strong correlations are required to link two scam emails. Note that other fuzzy linguistic models may be considered to model the portion of criteria to be satisfied in the aggregation step, *e.g.*, *few*, *some*, *half*, *many*, or *most* of the features can be required to be strongly correlated.

The TRIAGE framework provides advanced aggregation modelling capabilities, such as the *Choquet* integral - a fuzzy integral that aggregates a set of scores by not only taking into account importance factors assigned to individual criteria, but also *interactions* among subsets of criteria [23,24]. This enables us to also include interactions among *groups of criteria* (*i.e.*, email features), like synergies and redundancies. For this analysis, we have assigned synergies to coalitions of features involving at least the *phone* number of the scammer, so as to boost the overall similarity between emails having the same phone number, plus one additional feature in common. Inversely, some redundancy was put on certain combinations of email address-related features, such as (*reply*, *email body*), in order to diminish their redundancy effect on the overall similarity score. The definition of this aggregation model and its parameters must be guided by domain expert knowledge; in our case, we leveraged the insights gained previously by analyzing the role of phone numbers and email addresses in such scam operations in [5].

As an outcome (step 4), TRIAGE identifies *multi-dimensional clusters* (MDCs), which in this analysis are clusters of scam emails in which any pair of emails is linked by a *number of* common traits. As explained in [19], a decision threshold can be chosen such that undesired linkage between attacks are eliminated, *i.e.*, to drop any irrelevant connection that could result from a combination of small values or an insufficient number of correlated features. The result of this sensitivity analysis is shown in Figure 4, which represents the total number of clusters (MDCs) found by the algorithm for increasing values of the decision threshold. The best trade-off between quality and completeness of the clustering process is usually obtained for threshold values corresponding to the maximum number of clusters [19], *i.e.*, we chose here to set the threshold at 0.35. Given the set of importances and interactions defined above, we can easily verify that the outcome of the Choquet aggregation will exceed this threshold for combinations of any two features involving similarities for the *phone* number and at least one other feature (besides the *date*). Any coalition of three (or more) similar features will also exceed the threshold and will lead to the formation of a cluster.

4.1.1 k-additive Choquet integral

The Choquet integral is defined with respect to a so-called *fuzzy measure*. Given a set of criteria $\mathcal{N}$ (*e.g.*, email features), a fuzzy measure is simply a set function used to define, in some sense, the importance (or strength) of any subset belonging to the power set of $\mathcal{N}$. More formally, a fuzzy measure (alternatively called a *capacity* in the literature) is defined as follows:

Definition. ***(Fuzzy measure):*** Let $\mathcal{N} = \{1, 2, \ldots, n\}$ be the index set of n criteria. A *fuzzy measure* [25] (or *capacity* [26]) is a set function $\nu : 2^{\mathcal{N}} \rightarrow [0, 1]$ which is monotonic (i.e., $\nu(\mathcal{A}) \leq \nu(\mathcal{B})$ whenever $\mathcal{A} \subset \mathcal{B}$) and satisfies $\nu(\emptyset) = 0$. The measure is normalized if in addition $\nu(\mathcal{N}) = 1$. □

In multi-criteria decision making, a fuzzy measure is thus a set of 2^n real values where each value can be viewed as the degree of importance of a combination of criteria (also called a *coalition*, in particular in game theory).

However, since the exponential complexity of fuzzy measures makes them impractical for a decision maker to define them manually, Grabisch proposed a sub-model called k-order additive fuzzy measures, or shorter *k-additive* fuzzy measures [27]. The idea is to construct a fuzzy measure where the interaction among criteria is limited to groups of size k (or less). For example, in a 2-additive fuzzy measure, we can only model pairwise interactions among criteria, but no interactions in groups of 3 or more.

In MCDA, k-additive fuzzy measures have proved to provide a good trade-off between complexity and flexibility of the model. Instead of $2^n - 2$ values, they require only $\sum_{i=1}^{k} \binom{n}{i}$ values to be defined. 1-additive fuzzy measures are just ordinary additive measures (for which only n values are needed), but they are usually too restrictive for an accurate representation of complex problems[1]. In practice, 2-additivity seems to be the best compromise

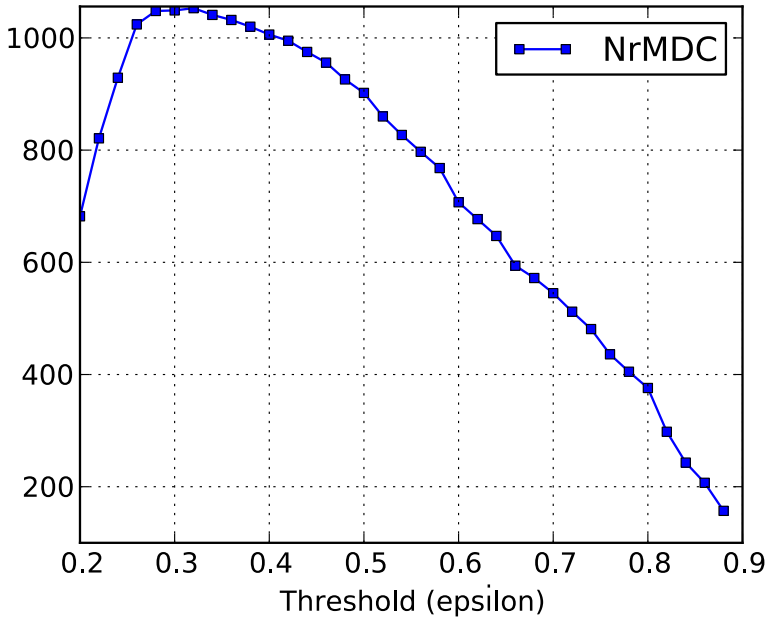

Figure 4 Sensitivity analysis on the decision threshold used in the TRIAGE clustering.

between low complexity and richness of the model [27]. In this case, only $n(n+1)/2$ values need to be defined.

From a practical viewpoint, a decision maker may thus consider defining a *2-additive* fuzzy measure since it appears to be a good trade-off between complexity of the model (number of values to determine) and its effectiveness and flexibility to include interactions (synergies or redundancies) among pairs of features. To do this, a commonly used approach consists to calculate the Choquet integral directly from the values provided by the decision maker for the *interaction indices* $I_2(\mathcal{A})$, and by using the expression of C_v given by Grabisch in [28]:

$$C_v(\mathbf{z}) = \sum_{i,j\in N|I_{ij}>0} (z_i \wedge z_j) I_{ij} + \sum_{i,j\in N|I_{ij}<0} (z_i \vee z_j)|I_{ij}| + \sum_{i\in N} z_i \left[\phi_i - \frac{1}{2}\sum_{j\neq i} |I_{ij}|\right], \quad (1)$$

where ϕ_i is the *Shapley* value of v as defined in [29] (which are directly related to the *importances* of individual features), I_{ij} is the interaction index between criteria i and j (as defined in [30]), and $\mathbf{z}$ is the vector of pairwise similarities obtained from the comparison of email features for any pair of scam messages.

As pointed out by Grabisch in [28], the expression (1) here above is remarkable for two reasons:

- It explains clearly the meaning of the interaction index and Shapley value: a positive interaction induces a *conjunctive* aggregation of scores, while a negative interaction induces a *disjunctive* aggregation (it is *sufficient* that one score is high). Clearly, the Shapley value is the linear part of the model, while interaction is the nonlinear part.
- Coefficients are non-negative, and if the capacity is normalized, they sum up to 1. In other words, it means that the Choquet integral is a convex combination of the scores z_i on all criteria, and of all disjunctive and conjunctive combinations of scores on pairs of criteria. Hence, the coefficient of a given term can be interpreted as the *percentage of contribution* of such term to the overall score. This aspect is highly appreciated in practice because it allows an analyst to understand each decision behavior made by the algorithm when calculating a global score between two entities.

Note that when defining interaction indices for pairs of criteria, the only conditions one must verify are the additivity and monotonicity of the resulting measure. That is, it is sufficient to verify that the following quantity is always positive $\forall i \in N$:

$$\left[\phi_i - \frac{1}{2}\sum_{j\neq i} |I_{ij}|\right] \geq 0 \quad (2)$$

4.2 Clustering results and experimental validation

The TRIAGE clustering tool identified 1,040 clusters that consist of at least 5 scam emails correlated by various

combinations of features. Because of the way these clusters are generated (*i.e.*, the multi-criteria aggregation), we anticipate that these email clusters represent different *campaigns*, potentially organized by the same individuals - as emails within the same cluster share several common traits.

Table 4 provides some global statistics computed across the top-250 largest scam campaigns. In over half of these campaigns, scammers are using only two distinct phone numbers, but they still make use of more than five different mailboxes to get the answers from their victims. Most scam campaigns are rather *long-lived* (lasting on average about a year). We note that cluster sizes are small on average, indicating that there are many small, isolated campaigns and only a few dozens of messages belong to the same campaign. This might be also an artefact of the data collection process; nevertheless, we anticipate that this could also reflect the scammers' behavior who may want to stay 'under the radar'. Indeed, bulk amounts of the same emails would have more potential to compromise their scamming operations, as this would become visible for content-based spam filters and, hence, would get blocked on the earlier stages of email filtering.

4.2.1 Assessment of clustering results

As data clustering is essentially an *unsupervised classification* approach, it is important to assess clustering results by means of objective criteria, so as to validate the soundness of the results. There are mainly two approaches to perform this validation: external or internal evaluation.

When there is no 'ground truth' information, we usually perform an *internal* evaluation of the clusters validity to confirm that the obtained results cannot have reasonably occurred 'by chance', or as an artefact of the clustering algorithm. Various cluster validity indices have been proposed to assess the quality and consistency of clustering results [31]. In graph clustering, most indices are based on the comparison of intra-cluster connectivity (*i.e.*, the compactness of clusters) and the inter-cluster variability (*i.e.*, the separability between clusters).

Table 4 Global statistics of the top 250 clusters

Statistic	Average	Median	Maximum
Number of emails	38	28	376
Number of from	13.9	9	181
Number of replies	6.2	5	56
Number of subjects	9.9	7	114
Number of phones	2.5	2	34
Duration (in days)	396	340	1,454
Number of dates (distinct)	27.9	22	259
Compactness	2.5	2.4	5.0

To evaluate the quality of our clustering results, we have examined their overall *compactness*, broken down by individual features. The graph compactness (C_p) is a cluster validity index that indicates how 'compact' (or homogeneous) the clusters are, based on their intra-connectivity characteristics. C_p reflects the average edge similarity between two objects of the cluster. More formally, for any cluster C_k, we calculate a normalized compactness index, as proposed in [32]:

$$C_{p_k} = \frac{\sum_{i=1}^{N_k-1} \sum_{j=i+1}^{N_k} \omega(i,j)}{N_k(N_k-1)/2}, \tag{3}$$

where $\omega(i,j)$ is a positive weight function reflecting node similarities in the graph, and N_k is the number of members within cluster k.

Figure 5 depicts graphically the overall C_p for the top 50 clusters. Besides a few exceptions, most clusters have an average compactness value above 1.5, which in most cases is associated with a combination of at least three strongly correlated features.

Since the TRIAGE tool is keeping track of all individual links in the similarity graphs, it is also possible to compute the proportion of emails that are linked by specific combinations of features within clusters. This can be very useful to understand the reasons behind the formation of the clusters, and hence provide insights into 'stable' (less *volatile*) features used by scammers when performing new campaigns.

From Table 5, we can observe that the top combination of features that tend to link scam emails (in 13% of the cases) involves the *phone* number, the *subject*, and also all three email addresses (*from, reply, email body*) used by the scammers. To confirm our intuition about the importance of certain features (phone numbers and, to a lesser extent, email addresses) and their effective role in identifying campaigns, we look at all similarity links within clusters. We observe that the features mainly responsible for linking scam messages in the clusters involve phone numbers (in 88% cases), followed by the *reply* email address (for 66% of the links). Not surprisingly, the *from* address (which can be easily spoofed) changes much more often and is used as linking feature in only 46% of cluster formations.

On the other hand, *external* validation techniques aim at comparing the recovered clusters to an *a priori* known structure. Given the knowledge of the ground truth class assignments and the clustering algorithm assignments of the same samples, external validation metrics measure the similarity of the two assignments based solely on clusters

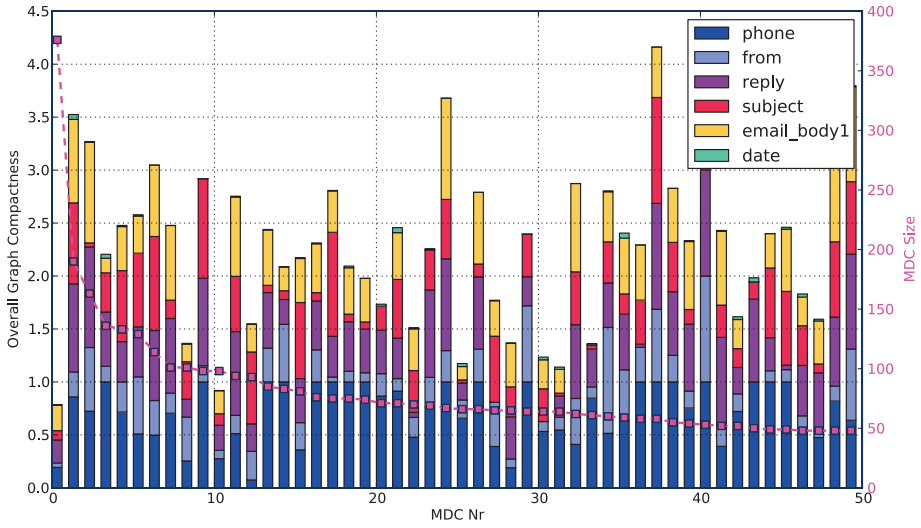

Figure 5 Overall compactness of the top 50 clusters (broken down by feature).

labels, and by ignoring permutations. Examples of well-known external validation metrics include:

- The *(Adjusted) Rand Index*, which measures the similarity of two clustering assignments, ignoring permutations and with chance normalization;
- The *Mutual Information* based scores, which compares two partitions based on (normalized) entropy measurements;
- The *Homogeneity*, *Completeness* and *V-measure*, which all measure some degree of agreement between two clustering assignments by defining an intuitive metric using conditional entropy analysis.

To further evaluate our clustering results, we have used here the Adjusted Rand Index (ARI) to compare different sets of clusters obtained by varying the parameters of the data fusion model used as input to the TRIAGE algorithm. The results of this comparison is shown in Table 6. Note that ARI has

Table 5 Top coalitions of features across all clusters

Coalition	Percentage (%)
(phone, subject, from, reply, email body)	13
(phone, reply, email body)	12
(phone, subject, reply, email body)	11
(phone, from, reply, email body)	7
(phone, subject)	6
(phone, from)	5
(phone, reply)	4
(phone, reply, subject)	4
(phone, reply, subject, from)	4
others	33

- A bounded range in $[-1, 1]$: negative values are bad (independent labellings), similar clusterings have a positive ARI, 1.0 is the perfect match score.
- Random (uniform) label assignments have an ARI score close to 0.0, thanks to the 'corrected-for-chance' adjustment made in order to reduce the effect of random labellings.

Overall, ARI and compactness validation indices indicate that most clustering results are similar, although they have obviously some matching differences with each other. Still, we observe that removing one or the other feature can greatly impact the data partitioning results, leading in general to fewer, coarse-grained MDClusters.

Further, we compare our results to results achieved by applying a different aggregation function. Weighted averaging functions, such as Ordered Weighted Averaging (OWA) or the weighted mean, can be quite convenient aggregation functions when we deal with data fusion tasks, in which criteria of interest are expressed with numerical values (usually, in $[0, 1]^n$). However, the weights used in weighted mean (WM) and the ones defined for the OWA operator play very different roles. Torra proposed in [33] a generalization of both WM and OWA, called *Weighted OWA* (WOWA). This aggregation function combines the advantages of both types of averaging functions by allowing the user to quantify the reliability of the information sources with a vector $\mathbf{p}$ (as the weighted mean does), and at the same time, to weight the values in relation to their relative position with a second vector $\mathbf{w}$ (as the OWA operator).

The two analyses that include all features (Table 6), Choquet and WOWA, are more similar - with an ARI of 0.68 - which seems to validate our selection of the features set. Here we used a different method in order to compare

Table 6 Experimental results of different feature sets and clustering algorithms

	Threshold	ARI	Compactness	MDC	Clusters	Emails (% of total)	Features
Choquet	0.35	-	3.08	1,036	19,866	13,779 (39%)	6
Choquet, no phone	0.35	0.53	2.64	912	16,745	17,753 (50%)	5
Choquet, no emails	0.35	0.52	2.2	868	13,458	14,884 (42%)	4
Choquet, no subj.	0.35	0.55	2.63	1,034	17,918	17,287 (49%)	5
WOWA	0.5	-	3.3	1,012	21,775	16,920 (48%)	6
WOWA, no phone	0.6	0.56	2.98	1,393	18,032	11,103 (31%)	5

our analysis with another one as there is no other existent previous work to which we could compare.

Since we have no ground truth, it is hard to reach a final conclusion on which analysis is providing the best results. However, the various quality measures presented above (Cp and ARI), combined with the data set *coverage* (proportion of clustered emails) provides a good indication of the results quality and a good confidence in the parameters chosen.

4.3 On the role of emails and phone numbers

One could wonder about the longevity of these key scam features; hence, we also looked at phone numbers and email addresses from a time perspective. Figure 6 represents the usage of the same email addresses and phone numbers over time. The Y-axis is density of the features that indicates their distribution in time on a 100% scale. As mentioned before, a larger part of them is used for only 1 day, so there is a slight concentration on the left side of the plot. However, the phone numbers are more often reused over time than email addresses. This could be explained by an easy access to new mailboxes offered by many free email providers. As for the *phone*, they probably still require some financial investment compared with emails. We checked the domain names of email addresses used in our scam dataset and found that the top 100 belong to webmail providers from all over the world. This finding suggests that email messages sent from such accounts would overpass sender-based anti-spam filters that are widely deployed today.

If we represent a scatter plot of *from* email addresses against phone numbers, on a per cluster basis (Figure 7), we find that these two parameters are uncorrelated. There are more changes performed with email addresses by scammers than with phone numbers, and even larger clusters of scammers sometimes maintain few email addresses.

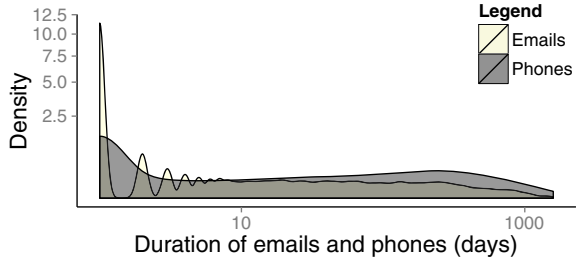

Figure 6 Duration of phone numbers and emails used by scammers, in days.

4.4 Clouds of words

419scam.org [4], as mentioned before, categorizes the scam emails into 10 categories. We presented their shares in the dataset in Section 3. Since the provided categorization is rather general, we wanted to evaluate by ourselves the scam categories present in our dataset by measuring the word frequencies in the body of the scam messages. Hence, to extract some additional knowledge from the clustered data, we create a list of the most repetitive keywords (after removing all stop words) and group them into meaningful categories. Stop words are 'words which are filtered out prior to, or after, processing of natural language data' [34]. Examples of such words in English could be short function words as the following: a, which, this, was, etc. By removing such words from the text, we took into account only words containing some knowledge about the topic discussed in the message.

As a result, we identified three big categories within the clusters: money transfer and bank-related fraud schemes (54%), fake lottery scam (22%), and fake delivery services (11%). The rest is uncategorized and refers to 13% of the clusters. The distribution is quite similar to the one provided by the data source, except that the delivery services are separated into other categories. The so-called general 419 scam category corresponds to messages about lost bank payments, compensations, and investment proposals. We grouped them together as it is challenging to clearly separate them due to a large number of shared keywords.

5 Characterization of campaigns

This section provides deeper insights into 419 scam campaign orchestration. We present several typical scam campaigns and show the connections between clusters, which are possibly run by the same group of scammers due to multiple strong interconnections among scam emails belonging to the same cluster.

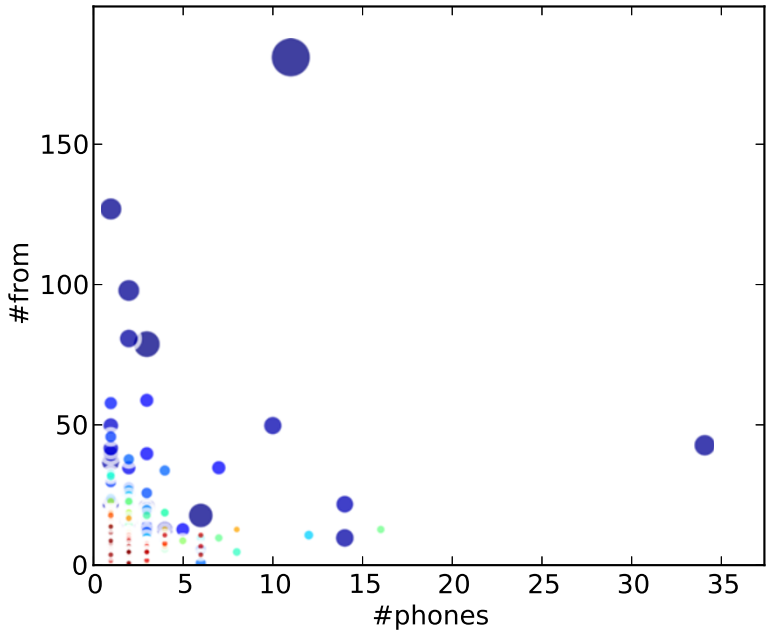

Figure 7 Number of distinct From addresses versus the number of phones used in the clusters. Each node represents a cluster. Node size indicates the number of emails.

5.1 Scam campaign examples

Here we characterize 419 scam campaigns by looking at how they are operated. For this purpose we use data visualization tools to plot the clustered data in an organized manner and look at the 'big picture'. Campaign graphs are likely reflecting the organization of campaigns and their maintenance over time. Interestingly, various campaigns have different operational structures and manage resources differently, as depicted by the examples in Figure 8.

Figures 8, 9, 10 and 11 show examples of different scam campaigns identified by TRIAGE. These diagrams were created using graph visualization tools developed in the VIS-SENSE project[2]. The graph diagrams are drawn using a circular layout that represents the various dates on which scam messages were sent. The dates are laid out starting from 9 o'clock (far left in the graph) and are growing clockwise. The other cluster nodes, which highlight other email features and their relationships, are drawn using a force-directed node placement algorithm. The big nodes in the graphs refer to *phone* numbers and *from* addresses. The smaller nodes represent mostly *subjects* and email addresses found in the *reply* and *from* fields, or in the message content.

5.1.1 ESKOM campaign

Figure 9 is an example of a 419 scam campaign quite likely orchestrated by the same cyber criminals. This campaign actually consists of two sub-campaigns: firstly, a 1-year *fake lottery* campaign located in the upper-left part of the graph (Figure 9); secondly, a 1.5-year campaign impersonating *ESKOM Holdings*, an electricity company in South Africa.

Even though scammers changed the topic of their scam, they kept re-using the very same phone number (represented in the center of the diagram). A noteworthy aspect of this campaign, shared with other campaigns we found, is that it relies on a few *from* email addresses (*i.e.*, the bigger nodes in the figure). A set of email addresses for *reply* and *body* was used in this campaign, however, since the switch of the scam topic a set of mailboxes and subjects has also changed. Also, we observe that the load of the scam campaign is well distributed over time, and does not exhibit very high peaks on specific dates, hence keeping very low volumes of emails sent. Finally, the *from* email accounts used by scammers in this case are mostly Gmail accounts. As we have no sender IP information, we could not verify if these were spoofed or not. However, in case these are genuine email accounts, this suggests that scammers use such webmail accounts for long periods of time while staying unnoticed by the email providers.

5.1.2 Sify-Rolex campaign

A similar campaign, presented in Figure 8A, illustrates the roles of email addresses and phone numbers in 419 scam. This campaign, which lasted for 1.5 years, changed topic five times at a frequency of 1 to 2 months, which is visible in the Figure by looking at

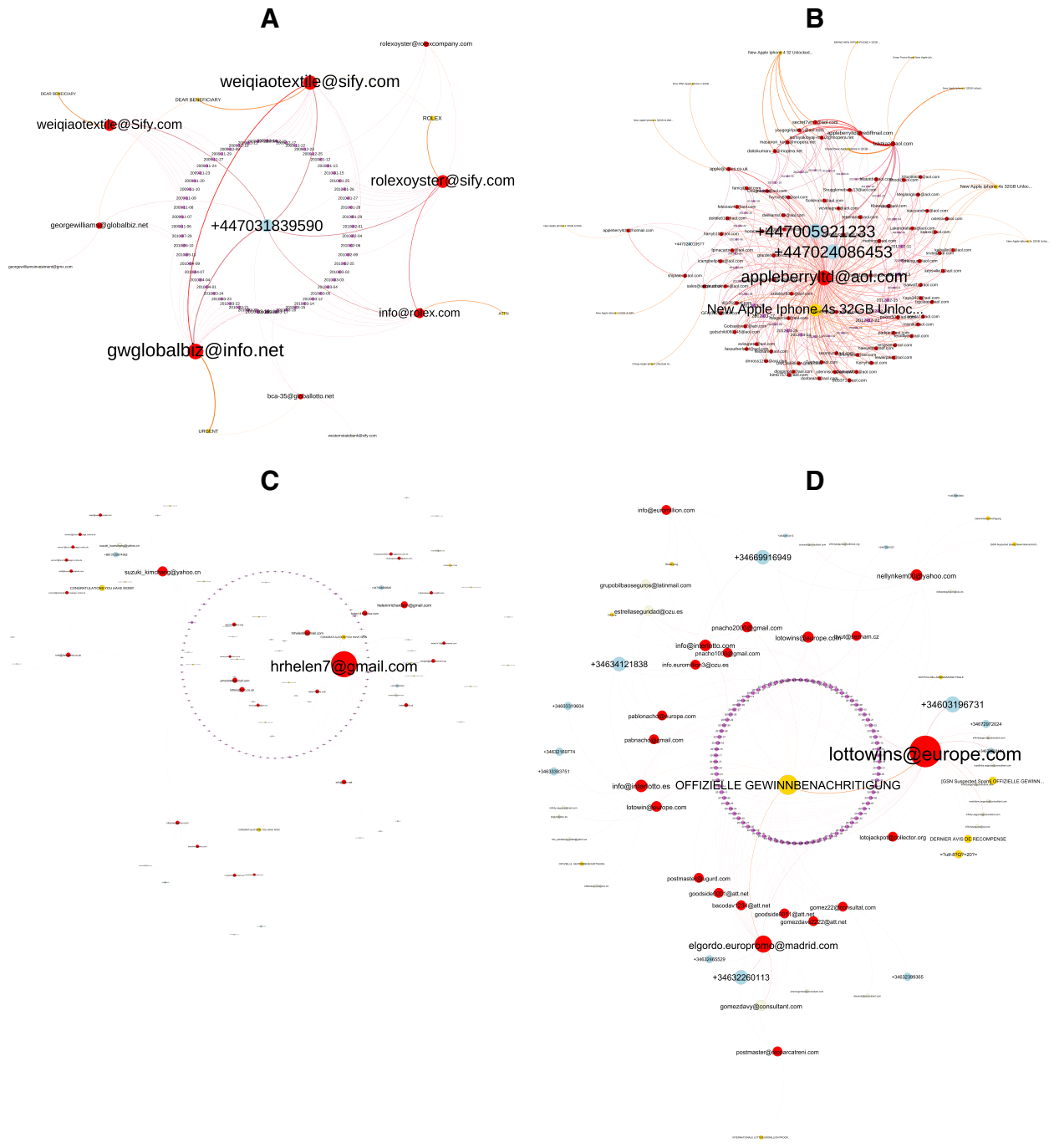

Figure 8 Examples of other scam campaign structures. (A) Distinct sub-campaigns, connected through a phone number. **(B)** A diverse iPhone scam campaign. **(C)** International campaign operated in China, UK, and the Netherlands. **(D)** Spanish lottery campaign often changing phone numbers and email addresses.

the larger subgroups placed around the circle. These shorter sub-campaigns were most probably run by the same group of scammers as the same phone number was reused over all campaigns. Inversely, we observe that the email addresses and subjects were completely changed as scammers were moving from one campaign to another.

Moreover, these email addresses were often selected to match the campaign topic and subjects, probably to make the scam messages appear authentic.

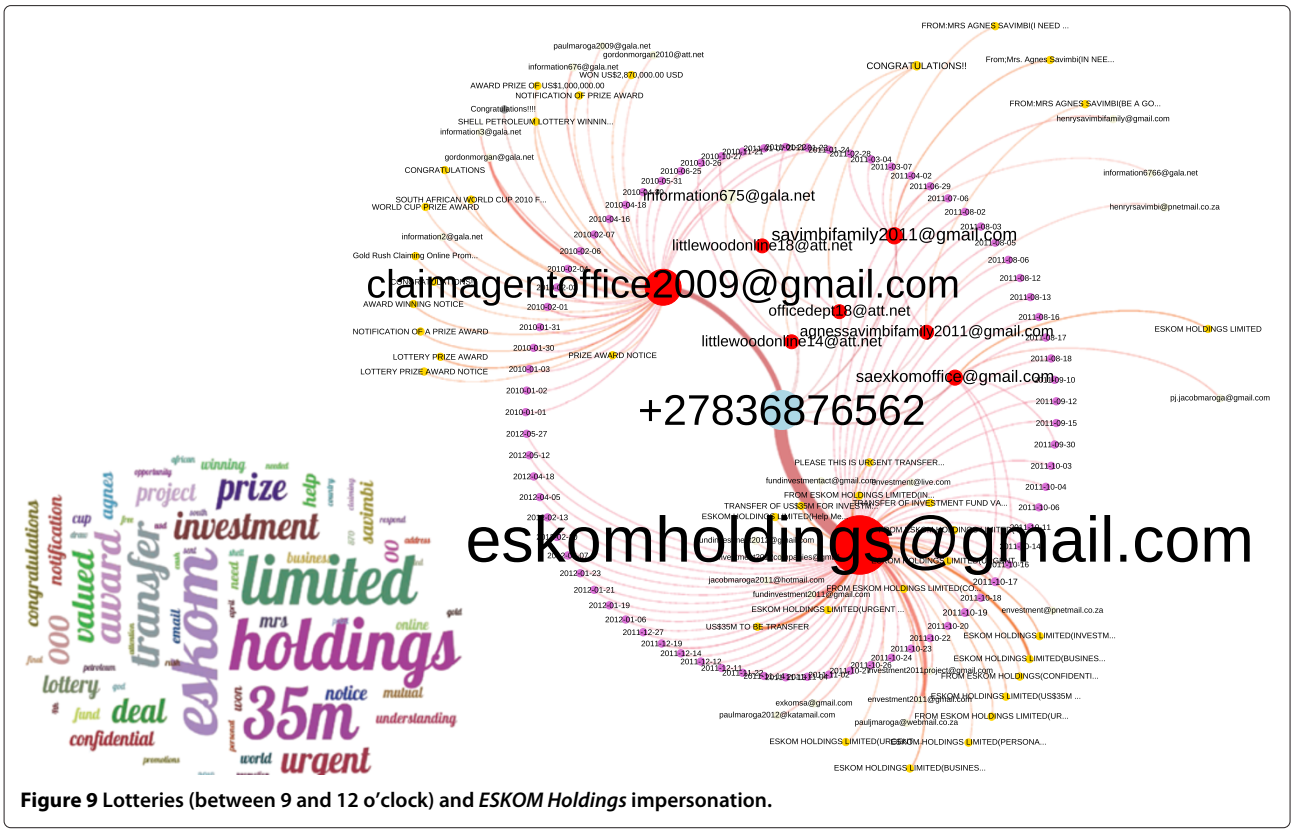

Figure 9 Lotteries (between 9 and 12 o'clock) and *ESKOM Holdings* impersonation.

5.1.3 iPhone campaign

While we observed a large number of such easily distinguishable campaigns, we also identified a very different, more 'chaotic' type of patterns reflecting a very different *modus operandi*, as demonstrated with the graphical illustration in Figure 8B. This diagram shows a cluster representing a recent campaign of iPhone-related scams, which lasted for over 1.5 years. The communication infrastructure of the scammers operating this campaign is much more diverse - around 85 unique email accounts were used in this campaign. Moreover, it relies on a large number of 'disposable' email addresses, which are typically used only once and seldom reused for a long period of time. As opposed to previous examples, however, the same or quite similar subjects and *from* email address were often reused, as well as the very same two phone numbers.

5.1.4 Internationally-operated campaigns campaign

Some scam campaigns still lack organization and do not always exhibit very clearly separated patterns, as illustrated by two campaigns depicted in Figures 8C and 8D. Both seem to use fewer email addresses but many different phone numbers are changing over time. The first one is an international campaign that started with Chinese phone numbers (top-left part), then moved to UK-based anonymous proxy numbers (top-right part), and ended with Dutch-based phone numbers. Almost all analyzed features changed over time, except a single *from* address. Interestingly, most of the phone numbers were regularly switched over time.

We also note that both campaigns exploit *fake lottery* topics. For example, the second one represents a Spanish fake lottery campaign that uses topic-related email addresses (again, in order to look more legitimate), and scammers leverage up to 11 different Spanish phone numbers, which are also regularly changed over time in the campaign. In the middle of the diagram, we can see a larger node representing an email *subject* that has been reused in a large number of scam emails during this campaign, showing that scammers were probably reusing the same fake lottery email template for all these emails. However, this type of scam cluster illustrates well the challenge of identifying such dynamic campaigns, in which the links between scam emails originating from the same criminal group are constantly changing over time. This supports our choice of using a multi-dimensional clustering tool, which can not only take into account multiple features but can also identify groups of emails that are linked by *varying* sets of commonalities (*i.e.*, more *volatile* features). These complex patterns and this volatility in email attributes can also suggest that cyber criminals operate in separate groups, where each group manages its own set of mailboxes and phone number(s); however, these groups are somehow federated and are collaborating

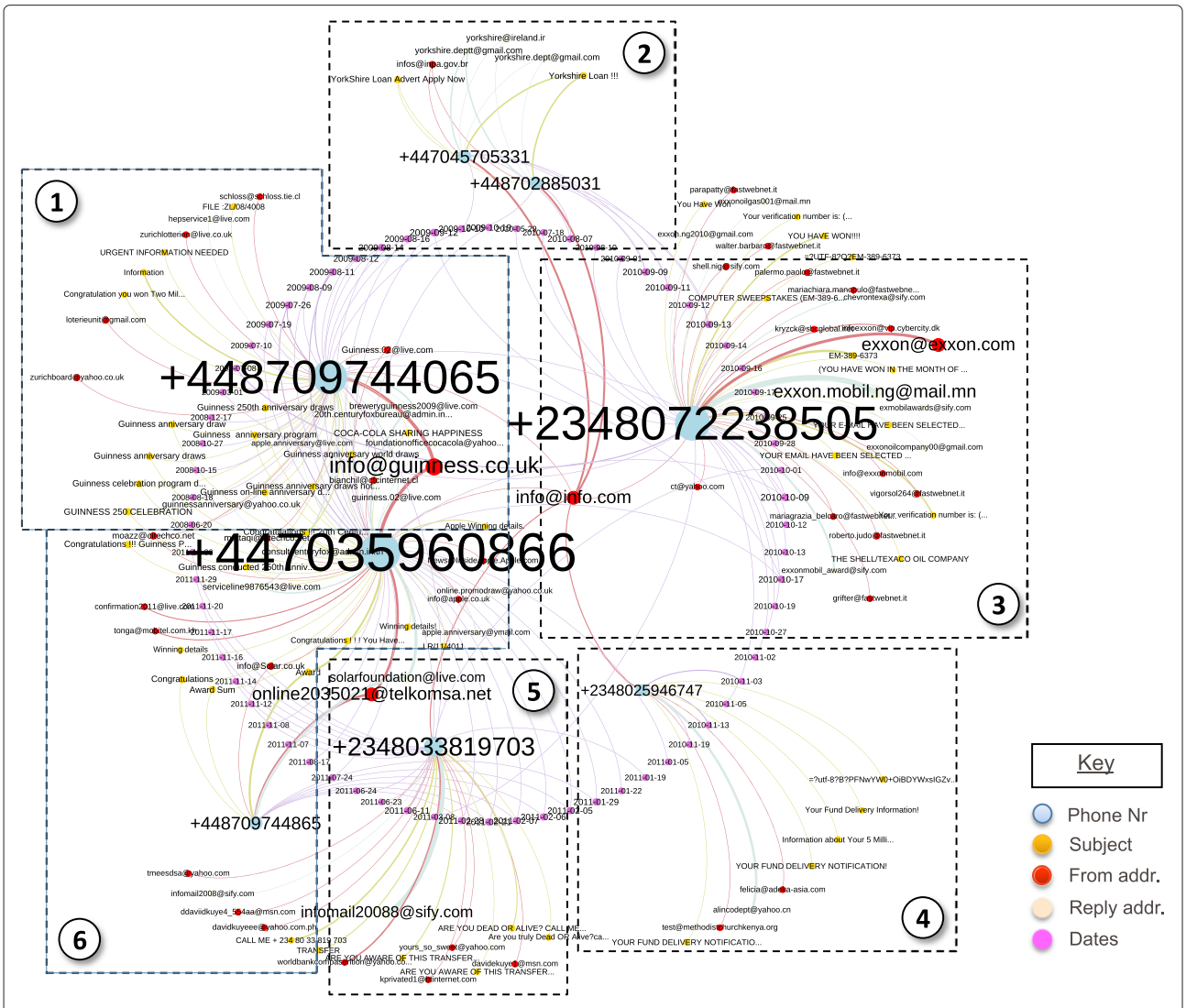

Figure 10 An example of macro-cluster, which consists of 6 different scam campaigns. The nodes laid in clockwise fashion reflect the timeline of the campaigns.

with each other, for example by sharing the same email templates, same distribution lists, or exchanging new scam topics.

5.2 Macro-clusters: connecting sub-campaigns

At the next step, we looked at scam campaigns from a broader perspective: by searching for loosely interconnected clusters. The goal was to pinpoint possibly larger-scale campaigns, which are made of weakly interconnected scam operations (*i.e.* different scam *runs*). For this purpose, we only used email addresses and phone numbers, since the other attributes are not considered as personally identifiable information. In fact, we looked for clusters that share at least one email address and/or phone number, and use this information to build so-called *macro-clusters*.

As a result, we identified a set 845 isolated clusters, and another set of 195 connected clusters, where the latter consists of 62 macro-clusters. The characteristics of the top six macro-campaigns are shown in Table 7. These macro-clusters are particularly interesting as they consist of a set of scam campaigns that appear to be loosely interconnected and therefore could be also orchestrated by the same cybercriminals. In fact, the links between different scam clusters were considered too weak by the clustering algorithm, because of the decision scheme and thresholds set as parameters, and thus these various scam runs were eventually grouped into separate clusters. However, these weak links can be easily recovered, and it is then up to the analyst to investigate how meaningful these interconnections really are. Indeed, we believe that it is much easier for a cyber investigator to start

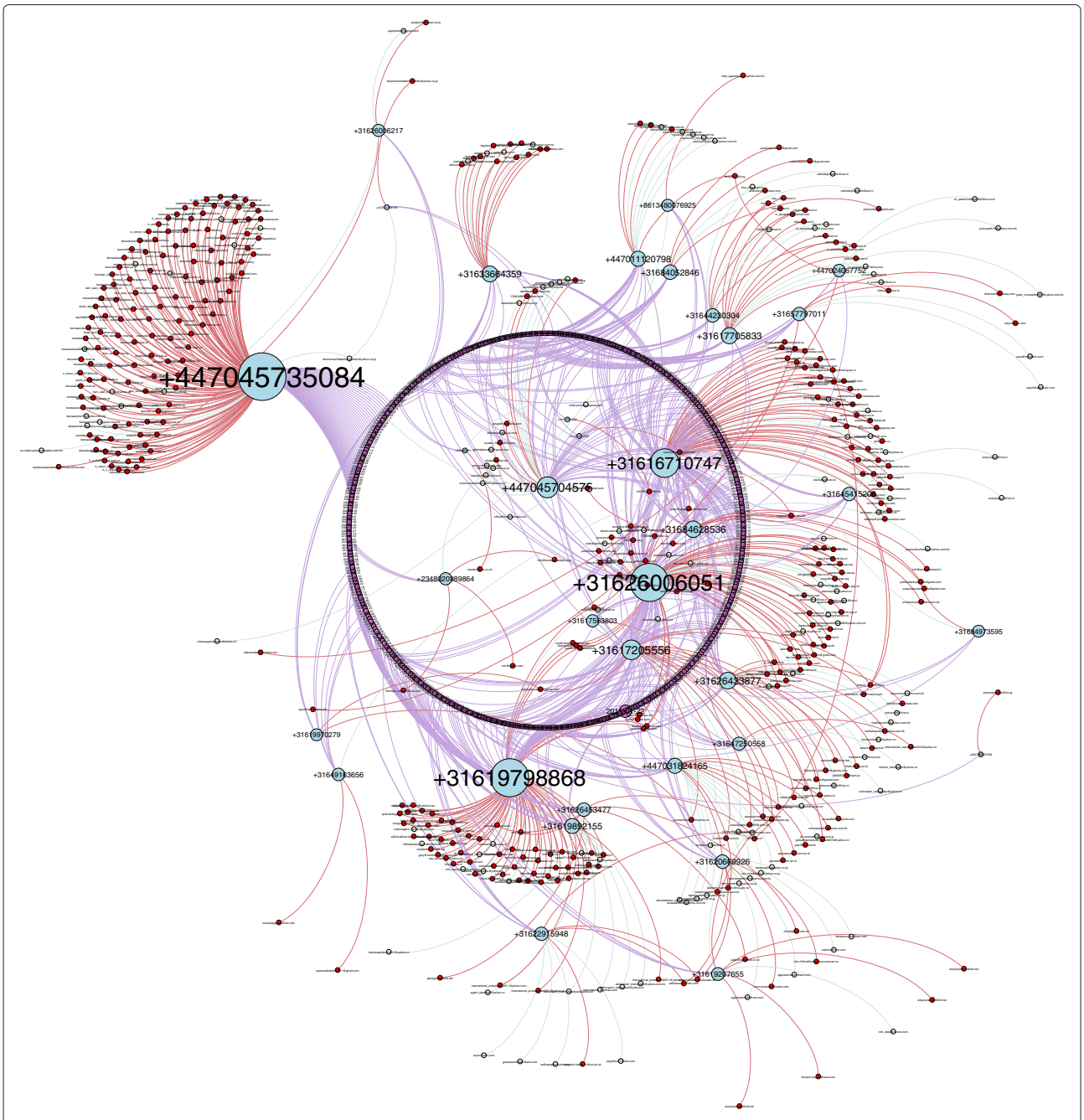

Figure 11 An example of macro-cluster, which consists of 14 sub-campaigns. The nodes laid in clockwise fashion reflect the timeline of the campaigns.

from a set of really meaningful scam clusters, to gradually increase the decision thresholds up to the point where she can decide herself to stop merging data clusters, as it might not be meaningful any more to attribute further different campaigns to the same group due to a lack of evidence.

Macro-clusters usually span across long time periods and exhibit various bursts of emails reflecting different campaigns, which use various topics and can even be operated in different countries. An example of a macro-campaign is illustrated in Figure 10, where it consists of six different scam campaigns of various sizes that include UK and Nigerian phone numbers.

We can easily distinguish them in the diagram as they appear as separate subgroups, each one having one or two bigger nodes (representing phone numbers reused multiple times) and a tail of connected nodes representing a series of *from* email addresses.

Table 7 Macro-clusters, mean values of attributes

Number	Number of campaigns	Phones	Mailboxes	Subjects	Duration (years)	Countries	Topics
1	14	44	677	223	4	4	Lottery, lost funds, investments
2	43	163	1,127	463	4	7	Lottery, banks, diplomats, FBI
3	6	18	128	80	4	4	Lottery
4	5	8	111	51	3.5	2	Packaging, lottery, loans
5	6	7	201	96	1	1	Lottery, UPS & WU delivery
6	4	7	82	33	2	1	Diplomats, monetary and payments scam

Notice that campaigns in this case are well separated with respect to phone numbers and emails, which are dedicated to each campaign (or operation), and the overlaps between campaigns are quite limited. However, there is a small node just in the center that indicates how these are interconnected (through a common *from* email address).

Some contact details were also reused and we used that for grouping them together. All together, these campaigns lasted for almost 3.5 years. Over this rather long time period, scammers have sent emails using 51 distinct subjects and 8 different phone numbers. This diversity of the topics suggests that there might be some competition among them, as they try to cover different online trick schemes instead of specializing in a single one.

Another example of a macro-campaign is illustrated in Figure 11, which consists of 14 sub-campaigns that can be more or less identified in the diagram as separate groups revolving around different phone numbers. Each one has a few dedicated phone numbers (44 in total) and its own set of *from*, *reply*, and embedded email addresses. However, in this case it appears that scammers were operating these different scam runs sequentially, sometimes reusing certain resources of previous campaigns. Hence, in forensic investigations, it might be necessary to look sometimes at weaker links that may possibly connect together some individuals or criminal groups that could be crime associates.

5.3 Geographical distribution of campaigns

To better understand how scammers operate geographically, we look at the data from a different angle. We have represented the scam email distribution per country for three subsets of our original data in Figure 12: (i) for the complete dataset (light grey), (ii) for scam clusters (dark grey), and (iii) for macro-clusters (black). As we can see, most campaigns identified through scam clusters originate either from African countries or from anonymized UK numbers. The difference between the light grey and dark grey bars in Figure 12 probably indicates a large number of stealthier or isolated scammers, as they do not form any cluster. Those quite likely refer to unorganized, opportunistic scammers, or maybe smaller gangs that operate in a loosely organised fashion.

Another interesting point is that macro-clusters (black bars) cover African and most of the European campaigns, forming bigger clusters potentially pointing to large organized groups of accomplices.

Organizing such macro-campaigns might be more expensive and difficult to operate, requiring more people to coordinate in various locations and using different

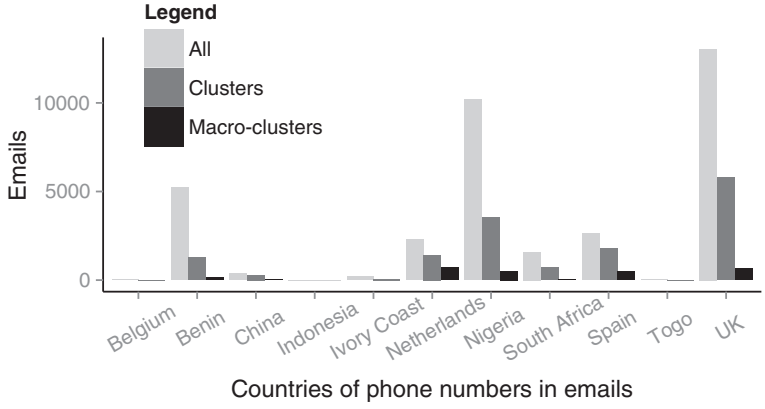

Figure 12 Scam email distribution by country for the largest scam clusters.

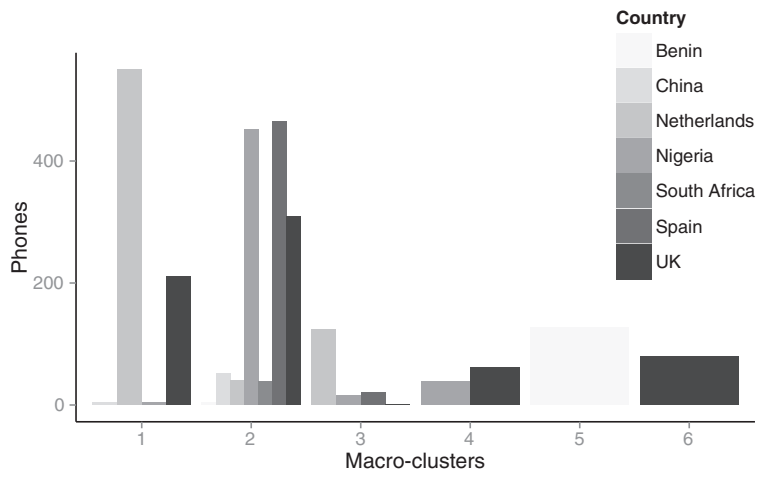

Figure 13 Top six macro-clusters: country distribution versus phone number count.

languages. Yet, these macro-campaigns are likely to be much more profitable, especially for the top-level leaders of these gangs.

We next look in more detail into the specific origins of macro-campaigns. Figure 13 shows the country distribution, versus phone number count, for the top 6 macro-campaigns. The last three campaigns are almost exclusively based in Africa, furthermore in only one or two countries, assuming that anonymised UK phone numbers are most probably used by scammers located in Africa and hiding behind these European phone numbers. The first three campaigns are more biased towards Europe, yet with strong connections to Nigeria and Benin. From an in-depth analysis, we conclude that these groups are competing in several 'fake lottery'-related scam, with the second group leading the pack and covering most of the countries. In comparison to the previous study of scam campaign [5], we observed much less UK and Nigerian numbers per group, and confirm that large-scale scam campaigns can be distributed over several continents. Indeed, the largest macro-campaign identified (#2) seems to be orchestrated by people distributed in many countries, if we assume also that that mobile phones are rarely used outside its country of origin (as highlighted by previous research [5]).

6 Conclusions

In this study, we empirically demonstrated the existence of 419 scam campaigns and a crucial role the phone numbers and email addresses play in 419 email scam business, in contrast to other cybercriminal schemes where email addresses may be often spoofed and phone numbers rarely used. With the help of a multi-dimensional clustering technique for grouping similar emails, we identified around 1,000 of *419 scam* campaigns.

We presented in detail several examples of 419 scam campaigns, some of which lasted for several years - representing them in a graphical way and discussing their characteristics. Finally, we uncovered the existence of *macro-campaigns*, groups of loosely linked together campaigns that are probably run by the same people. We found that some of these macro-campaigns are geographically spread over several countries, both African and European. We showed that infrastructure, orchestration, and *modus operandi* of such campaigns differ from traditional spam campaigns: campaigns are long and scarce, they extensively use anonymization tools like webmail accounts for hiding IPs and anonymous proxy phone numbers. Our analysis has unveiled a high diversity in scam orchestration methods, showing that scammer(s) can work on various topics within a campaign, thus probably competing with each other over trendy scam topics.

We believe that our methods and findings could be leveraged to improve investigations of various cyber crime schemes - other than scam campaigns as well.

Competing interests
The authors do not have any competing interests.

Acknowledgements
This research was partly supported by the European Commission's Seventh Framework Programme (FP7 2007-2013) under grant agreement no. 257495 (VIS-SENSE) and no. 257007 (SysSec). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

Author details
[1]Eurecom, Route des Chappes, Sophia Antipolis, France. [2]Symantec Research Labs, Sophia Antipolis, France.

Received: 30 August 2013 Accepted: 16 December 2013
Published: 22 January 2014

References

1. Definition of Nigerian Scam. http://en.wikipedia.org/wiki/Nigerian_scam. Accessed 3 January 2014
2. Definition of 419 Scam. http://www.419scam.org/419scam.htm. Accessed 3 January 2014
3. J Buchanan, AJ Grant, Investigating and prosecuting Nigerian Fraud. United States Attorneys' Bulletin. **49**(6) (2001)
4. 419 Scam – fake lottery Fraud phone directory. http://www.419scam.org/419-by-phone.htm. Accessed 3 January 2014
5. A Costin, J Isacenkova, M Balduzzi, A Francillon, D Balzarotti, The role of phone numbers in understanding cyber-crime, in *11th International Conference on Privacy, Security and Trust (PST 2013)* (Tarragona, Catalonia, 10-12 July, 2013)
6. A Pathak, F Qian, YC Hu, ZM Mao, S Ranjan, Botnet spam campaigns can be long lasting: evidence, implications, and analysis, in *SIGMETRICS '09 Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems* Seattle, WA, 15-19 June (ACM, New York, 2009), pp. 13–24
7. DS Anderson, C Fleizach, S Savage, GM Voelker, Spamscatter: Characterizing internet scam hosting infrastructure. PhD thesis, University of California, San Diego 2007
8. Microsoft Security Intelligence Report (2008-2012), http://www.microsoft.com/security/sir/archive/default.aspx. Accessed 3 January 2014
9. J Isacenkova, O Thonnard, A Costin, D Balzarotti, A Francillon, F Eurecom, Inside the SCAM jungle: a closer look at 419 scam email operations, in *International Workshop on Cyber Crime (IWCC 2013)* (The Westin Hotel, San Francisco, CA, 24 May 2013)
10. M Cova, C Leita, O Thonnard, AD Keromytis, M Dacier, An analysis of rogue AV campaigns, in *Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection* (Springer-Verlag, RAID'10, Berlin, Heidelberg, 2010), pp. 442–463. [http://portal.acm.org/citation.cfm?id=1894166.1894196]
11. O Thonnard, M Dacier, A strategic analysis of spam botnets operations, in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference,* CEAS '11 (ACM, New York, 2011), pp. 162–171
12. O Thonnard, L Bilge, G O'Gorman, S Kiernan, M Lee, Industrial espionage and targeted attacks: understanding the characteristics of an escalating threat, in *RAID,* Amsterdam, The Netherlands, 12-14 September (Springer, New York, 2012), pp. 64–85
13. C Tive, *419 Scam: Exploits of the Nigerian Con Man*. (iUniverse, New York, 2006)
14. F Stajano, P Wilson, Understanding scam victims: seven principles for systems security. Commun. ACM. **54**(3), 70–75 (2011)
15. C Herley, Why do Nigerian Scammers say they are from Nigeria?, in *Proceedings of the Workshop on the Economics of Information Security* (Berlin, Germany, 25-26 June 2012)
16. J Oboh, Y Schoenmakers, Nigerian advance fee Fraud in transnational perspective. Policing Multiple Commun. **15**, 235 (2010)
17. Y Gao, G Zhao, Knowledge-based information extraction: a case study of recognizing emails of Nigerian frauds, in *Policing multiple communities,* NLDB'05 (Springer-Verlag, Berlin, Heidelberg, 2005), pp. 161–172
18. C Graham, J U. S. Patent 7917655 Nicholas, Method and system for employing phone number analysis to detect and prevent spam and e-mail scams. 29 March 2011 http://www.patentlens.net/patentlens/patent/US_7917655/en/
19. O Thonnard, A multi-criteria clustering approach to support attack attribution in cyberspace. PhD thesis, École Doctorale d'Informatique, Télécommunications et Électronique de Paris, 2010
20. G Kondrak, N-gram similarity and distance, in *Proceedings of the 12th Conference on String Processing and Information Retrieval* Buenos Aires, 2-4 November (Springer-Verlag Berlin, Heidelberg, 2005), pp. 115–126
21. LA Zadeh, A computational approach to fuzzy quantifiers in natural languages. Comput. & Math. Appl. **9**, 149–184 (1983)
22. RR Yager, Quantifier guided aggregation using OWA operators. Int. J. Intell. Syst. **11**, 49–73 (1996). http://dx.doi.org/10.1002/(SICI)1098-111X(199601)11:1<49::AID-INT3>3.0.CO;2-Z
23. M Grabisch, C Labreuche, A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. Ann. Operations Res (2009). http://dx.doi.org/10.1007/s10479-009-0655-8
24. V Torra, Y Narukawa, *Modeling Decisions: Information Fusion and Aggregation Operators*. (Springer, Berlin, 2007)
25. M Sugeno, Theory of fuzzy integrals and its applications. PhD thesis, Tokyo Institute of Technology, 1974
26. G Choquet, Theory of capacities. Annales de l'Institut Fourier. **5**, 131–295 (1953)
27. M Grabisch, k-order additive discrete fuzzy measures and their representation. Fuzzy Sets Syst. **92**(2), 167–189 (1997)
28. M Grabisch, C Labreuche, A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. Ann. Operations Res. **175**(1) (2010)
29. L Shapley, A value for n-person games, in *Contributions to the Theory of Games, Vol. II, Volume 28 of Annals of Mathematics Studies*, ed. by H Kuhn, A Tucker (Princeton University Press, Princeton, NJ, 1953), pp. 307–317
30. T Murofushi, S Soneda, Techniques for reading fuzzy measures (iii) Interaction index, in *Proceedings of the 9th Fuzzy Systems Symposium* (Sapporo, Japan, May 1993), pp. 693–696
31. M Halkidi, Y Batistakis, M Vazirgiannis, On clustering validation techniques. J. Intell. Inf. Syst. **17**(2-3), 107–145 (2001)
32. F Boutin, M Hascoet, Cluster validity indices for graph partitioning, in *8th International Conference on Information Visualization (IV)* (London, 14-16 July 2004)
33. V Torra, The weighted {OWA} operator. Int. J. Intell. Syst. **12**(2), 153–166 (1997)
34. Stop Words. http://en.wikipedia.org/wiki/Stop_words. Accessed 3 January 2014

doi:10.1186/1687-417X-2014-4
Cite this article as: Isacenkova *et al.*: **Inside the scam jungle: a closer look at 419 scam email operations.** *EURASIP Journal on Information Security* 2014 **2014**:4.