EURECOM
Department of Network & Security
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-13-277

The Role of Phone Numbers in Understanding Cyber-Crime Schemes

December 6th, 2012
Last update February 19th, 2013

Andrei Costin, Jelena Isacenkova, Marco Balduzzi, Aurélien Francillon, Davide Balzarotti

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {andrei.costin, isachenk, aurelien.francillon, balzarotti}@eurecom.fr, {marco_balduzzi}@trendmicro.it

[1] EURECOM's research is partially supported by its industrial members: BMW Group Research & Technology, IABG, Monaco Telecom, Orange, SAP, SFR, ST Microelectronics, Swisscom, Symantec.

The Role of Phone Numbers in Understanding Cyber-Crime Schemes

Andrei Costin, Jelena Isacenkova, Marco Balduzzi, Aurélien Francillon, Davide Balzarotti

Abstract

Internet and telephones are part of everyone's modern life. Unfortunately, also several criminal activities rely on these technologies to reach their victims. While the use and importance of the network has been largely studied, previous work overlooked the role that phone numbers can play into understanding online threats.

In this work we aim at determining if leveraging phone numbers analysis can improve our understanding of the underground markets, illegal computer activities, or cyber-crime in general. This knowledge could then be adopted by several defensive mechanisms, including blacklists or advanced spam heuristics. In our study we collected phone numbers from various public or private sources and we designed a framework for mining, analyzing, enriching and, finally, correlating phone numbers to malicious activities.

Our results show that, in scam activities, phones numbers remain often more stable over time than email addresses. Finally, using a combination of graph analysis and geographical HLR lookup, we were able to identify recurrent cyber-criminal business models and to link together scam communities that spread over different countries.

Index Terms

Measurement, Security, Economics, Online crime, Web frauds

Contents

List of Figures

List of Tables

1 Introduction

In the current digital economy, cybercrime is ubiquitous and has become a major security issue. Every year, new attack avenues and business models arise [19, 26]. Criminals use different techniques to trap victims into various schemes and to achieve their, usually financial, goals. The used communication mechanism depends on the abuse scheme, but criminals need to have a form of interaction with their victims; for example a web page (phishing, selling counterfeit goods), an IM contact or a phone number (scams).

In many fraud schemes phone numbers play an important role. For example, criminals have been analyzed by authorities based on their phone numbers on public or underground forums [9]. In other online fraud cases, like one-click fraud [15], usage of a phone number can make the fraud appear more legitimate to a victim. Finally, scammers will often use the phone to defraud victims [36].

While the role of other features in illegal online activities has been extensively studied [29] [38] [28] [18] [16], the role of phone numbers remains relatively uncovered. The existing work is limited to the study of spam over SMS, or to phone number abuses through premium services [35] [34] [25]. However, a recent study of fraud activity in Japan [15] demonstrates that phone numbers play an important role in online fraud and can be used as a way to link and identify criminals. While there are several indications of criminals using phone numbers for their malicious activities [9], we still lack a global understanding to compare the usage and the role of the phone numbers in different criminal schemes.

In this context, our research has three main objectives. First, we want to evaluate the reliability of leveraging an automated phone numbers analysis to improve our understanding of the underground markets, illegal computer activities and cybercriminals in general. Second, by looking at the analyzed data, we try to find various patterns associated to recurrent criminal business models. Finally, we correlate the extracted information and enrich them with a geographical HLR lookup process to automatically identify the communities responsible for Nigerian scam campaigns.

Along these three directions, we can summarize our main findings and contributions as follows:

- We present an approach, its limitations, and possible improvements for extracting phone numbers from unstructured text input.

- We study the use of phone numbers across multiple malicious online activities, with a particular focus on scam attacks. We found that while there are many overlapping numbers *within* each category, we discovered no correlation *between* datasets.

- We show that phone numbers are a good way to detect communities of scammers and to find links between scam campaigns.

- To the best of our knowledge, we are the first to propose and use HLR lookups to verify our findings, and to study the use of phones over time of different and distributed criminal groups.

The rest of the paper is organized as follows: we start by presenting our framework and data processing methodologies in Section 2; subsequently we present our general results along with data analysis and present key findings in Section 3; Section 4 continues on the key findings and presents interesting fraud business models discovered during the experiments; subsequently in Section 5 we analyze criminals behind the fraud business models; we then continue on presenting in Section 6 our analysis of mobile phones used in scam frauds; finally, we discuss similar and related works in Section 7; we conclude with Section 8.

2 Architecture and Data Collection

In this section we describe the datasets we used in our study, and we introduce our data collection, filtering, and analysis methodology along with the challenges we faced and the approaches to solve them.

2.1 Datasets

Our focus was to obtain data from several sources related to illegal online activities, which contains phone numbers. we selected mainly data from scam messages, spam messages, DNS whois registrations and Android malware. We selected those data sources because they are very likely to contain phone numbers, are connected to cyber-crimes or fraud schemes, and we were able to get access to such data sources.

2.1.1 Scam dataset (SCAM)

The SCAM dataset consists of data from user reports. There are several *user reports aggregators* that cover a wide range of fraudulent activities. This information is usually reported in dedicated forums, blogs, and other online media sites. We selected the community-supported site `419scam.org` because it has a large dataset of well formatted scam reports. This data was collected, manually filtered, and pre-processed from January 2009 to August 2012. Additional information about the scam emails is also provided, including the category, the message headers and, for 16% of them, the corresponding original email body.

2.1.2 Spam datasets (SPAM)

Our SPAM dataset included data from two different sources. The first part of it included generic spam messages that an average public mail server receives on a constant basis. The corpus consisted of around 40 thousands messages collected over the period of roughly 5 years by a low traffic mail server.

The second part of the SPAM dataset included over 260 thousands spam messages collected by a commercial anti-spam filter deployed in a number of medium size enterprises in a period of six months.

2.1.3 Android malware datasets (Android ML)

Most of the phone numbers in the malware dataset were extracted from Android malware by mostly manual reverse engineering. Two datasets were provided by well known anti-virus companies. The first Android dataset consists of 5,739 phone numbers extracted by reverse engineering mobile malware mainly found in China. The second one is made of approximately 400 manually extracted phone numbers.

Unfortunately, we faced two major problems with these datasets which made them unusable in our experiments. First, despite the manual analysis, the dataset contains a considerable amount of noise (false positives introduced by the extraction tools). Most importantly, most of the phone numbers in this category turned out to be short numbers used for premium SMS frauds. Short numbers are a commodity provided to users, in general those numbers are translated to a long number by the mobile operator. However, this translation is not public, it is country specific and sometimes operator specific. Therefore, short number analysis and tracking is not trivial, and hence it is left as future work.

2.1.4 DNS datasets (DNS)

When a new domain name is registered, some details needs to be provided by the registrant, e.g., contact name, address, email address, and phone number. Such details are stored and can later be retrieved by performing a Whois database lookup. We decided to collect the phone numbers provided during domain registration that are known to be used for malicious activities. To do so we used two sources of malicious domains. The first is a public list of malicious domains that are flagged by the Exposure [13] malicious domain detection service. We included around 70 thousands domains.

Another list of malicious domains was provided by an anti-virus company that consisted of 1,136 domains. The latter were used to host websites distributing FakeAVs, ransomwares, and other types of malware.

2.2 Phone number extraction and processing framework

Because of the variety of the data sources we used, it quickly became clear that we had to manage data in a flexible and extensible framework. Figure 1 describes the framework we built for data extraction, filtering and enrichment. The overall architecture includes several tasks that can be preformed independently and in parallel. Our framework is easily extensible, both to include new datasets and to add additional data filtering or enrichment modules.

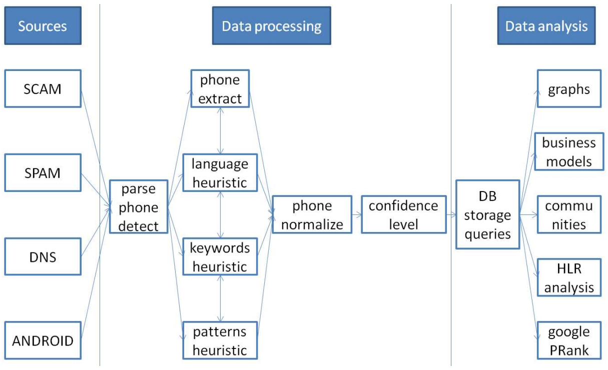

Figure 1: Overall processing architecture

As it is described in Figure 1, the framework relies on three main components: the data sources, the data processing, and the data analysis. In particular, the data processing phase includes phone number matching, extraction, and normalization. Matching and extraction is performed by using Google's public library for parsing phone numbers (*libphonenumber* [2]). Then, normalization transforms the number in a standard, international, format.

Afterwards, we derive the service type (e.g., mobile, land line, premium) of a phone number using two different databases (so called "numbering plans"). The first one is a free and open source XML-based database included in *libphonenumber* which derives the service type during the extraction and normalization process. The second one, is a commercial database [4] that provides a wider coverage. We use both sources to cross-check the results and find potential discrepancies.

2.3 Matching, normalization and filtering

One of the main challenges when trying to automatically recognize and extract phone numbers comes from many different representations that are used to write them in a textual form. For example, they can include international codes ('+' and '00'), only local codes, or only digits. They can be grouped by 2 or 3 or 4 digits, and separated by spaces, '.', '-' or other characters. In addition to this, source data often includes strings of digits that can be misinterpreted as phone numbers, e.g., ID numbers, IP addresses.

A number without its international prefix may potentially correspond to many different numbers in different countries. Therefore, a normalization algorithm transforms an extracted number into a non ambiguous fully qualified E.164 number. When adding a country code to a candidate phone number, a numbering plan can be used to check if the resulting number is a valid number or not (e.g., the range is allocated and it has the correct number of digits). Unfortunately, repeating this step with too many possible country codes would lead to many false positives.

Therefore, our goal is twofold: in the case of multiple normalized phone numbers we want to distinguish which one (country code wise) is the most probable match; in the case of noise, determine that the phone match by *libphonenumber* library is a false-positive and is not a phone number.

Though a simple normalization and filtering step is preformed by the *libphonenumber* library, we try to improve the accuracy of *libphonenumber* normalization by using additional processing steps to distinguish the most likely country codes and/or false-positives. We introduce several heuristics allowing to identify the most confident normalized numbers for a given non-normalized number and to create a weighted list of the most promising country codes.

2.3.1 Language heuristic

The language used in the surrounding text from which a phone number is extracted is good indication of the geographic areas in which the number is supposed to be used. This is especially true for SCAM numbers, in which the sender expects the victim to call that number without ambiguity.

For example, for a message written in Russian, that includes a phone number without a full international prefix, we try to normalize by testing few target countries where the Russian language is widely spoke, e.g., Russia ’+7’, Ukraine ’+380’, Belarus ’+375’, Moldova ’+373’.

Based on this observation, we adopted an automatic language detection technique (provided by the *guess-language* [7] python module) to infer the language and thus derive likely phone country codes to improve the normalization accuracy. For this we maintain a dictionary that associates languages to promising country codes.

2.3.2 Keywords dictionary match heuristic

The immediate context of a phone number can also be very useful to detect the presence of a phone number. Such context may include abbreviations or words to indicate a phone number is following (e.g., *phone, mobile, tel, fax, mobile, call, contact, line, dial, direct, ext*), combined with punctuation marks (e.g., ’.’, ’:’,).

Based on this observation, we use the local text context (i.e. a 50-byte prefix and suffix) of any sequence of digits detected as a possible phone number. Matching this context to known dictionary words or abbreviations helps to improve the confidence level of detected numbers. We also maintain a dictionary of such words patterns.

2.3.3 Format pattern match heuristic

Depending on the locales, phone numbers take different form. Therefore, when written in local formats, a phone number in France can appear as ’04.73.33.49.30’, while a local format phone number in the U.S. can appear as ’(803) 951-4544’.

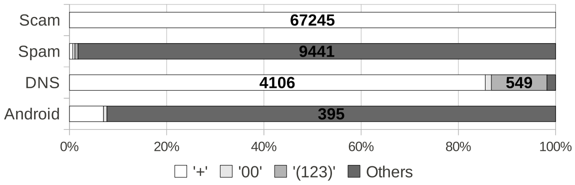

Figure 2: Format patterns of the original phone data. ’(xxx)’ is used to represent a country prefix.

Table 1: General phone numbers table

	Extraction phase		Normalization phase				
Data sources	Total extracted	Unique (unique from total)	International format	Need normalization	After normalization	Total numbers nomarlized (normalized from extracted)	Countries
Android ML	428	388 (91%)	33	355	95	127 (33%)	19
DNS	76,836	4,649 (6%)	4,003	646	161	4,060 (87%)	92
SPAM	52,231	8,677 (17%)	64	8,613	897	945 (11%)	51
SCAM	128,291	67,242 (52%)	67,242	0	0	67,242 (100%)	15
Total	257,786	80,956	71,342	9,614	1,153	**72,374**	107

We therefore designed specific pattern matching rules on the detected digits of possible phone numbers, in order to increase our confidence and improve our normalization in “guessing” the proper international prefix. We also maintain a dictionary of such rules for the most common country code derivation.

2.3.4 Confidence level

To deal with the variety of our datasets and our detection heuristics, we introduced a *confidence level* index. It is a composite, weight-oriented, metric that indicates the chances that a given extracted and normalized phone number is indeed a real phone number.

For a given sequence of digits we increase the confidence index for each submodules (*phonenumber, language heuristic, dictionary heuristic, patterns heuristic*) that confirms the detected sequence. We then lower the index according to the number of normalized numbers resulting from the initial sequence of digits. For example, if the normalization process returns only one candidate this increase our confidence level, while if the process returns several dozens of possible numbers we lower their confidence index to reflect the higher probability of these numbers to be false positives.

The confidence level algorithm outputs a value between 0.0 and 1.0, where 1.0 is 100% match. In most of the cases a 100% match corresponds to matching a phone number that was already including the international prefixes and country codes (like for example ’+1 695 123456’).

2.4 Assessment of results validity

As discussed in the previous section, there is a trade-off between the amount of extracted numbers and the accuracy of the results. Even by applying the heuristics discussed above, the normalization phase often leads to several possible numbers for each candidate entry. With the exception of the correct one (assuming that it is found by our method), all the other are false positives.

This introduce some uncertainty in the resulting datasets. To mitigate this problem, for the rest of the study we discard all numbers with a confidence value below *0.6*. Table 1 shows the impact of this filtering step on the different datasets. The column *International Format* contains the numbers we extracted in their full form (e.g. '+1 (805) 403 8813'), while the *Likely Number* column reports those numbers that, given the context in which they appeared, they were most likely phone numbers, but for which a normalization step was required.

The SCAM data is obviously the most reliable source, since it was already manually preprocessed by the users/community. The DNS dataset is of a medium quality. In fact, the data is well structured and relatively easy to parse in an automated fashion. Unfortunately, both in the Android and SPAM datasets most of the numbers do not reach the required threshold, indicating that there is a lot of noise in the input data. The problem with the Android dataset is that it includes many short numbers (49% out of total 388 uniquely extracted from Android ML) for which our approach is not able to compute the normalized number. The SPAM dataset is even more noisy, mainly because of random-looking content inserted into the messages to make spam detection more challenging.

Our framework was able to extract a candidate phone number in 17% of the emails. This is consistent to the 15% that was measured by Pathak et al. [33]. However, most of the candidates turned out to be false positives. In fact, when the extracted sequences were then fed into the normalization process, the confidence-based filtering discarded most of them, thus reducing the initial set to only 945 phone numbers. Even worse, by performing a manual check of the extracted data, we were able to confirm only 106 as authentic phone numbers. These underline the fact that the process of automatically extracting phone numbers (beyond the one already present in an international format) from free text is hard and error prone.

Finally, Figure 2 presents the three most common prefix format of phone numbers for each data source. As we notice from the table, the less structured data is found in the SPAM and Android datasets, which confirms our findings from the filtering process.

2.5 HLR lookups

Home Location Registers (HLR) are databases maintained by mobile operators containing information about the current status of a phone number (e.g., IMSI, roaming status, and roaming operator). This can be very useful for our study, because they allow to know if a mobile phone number is still active and if it is roaming to a foreign country. However, HLRs are only accessible from within the SS7 telecommunication network, and therefore we had to rely on a third party commercial service [8] to query this information.

A detailed description of how HLR lookups are performed can be found in [3]. The basic idea is to contact the homing operator of a phone number pretending to be interested in initiating either an SMS or a voice call (e.g., by sending a `MAP_SEND_ROUTING_INFORMATION` message). At this point, the homing operator of the subscriber number checks the status of the mobile number and returns the details.

By periodically doing a query for a given number, we can get insight on the evolution of a number status. Such status can be used to draw conclusions about activities related to a mobile phone number. We use this technique in Section 6.

3 Data Analysis

In this section we analyze the phone numbers we extracted from our datasets, we report some general statistics, and we study the correlation and reuse of phone numbers both between and within datasets.

3.1 General statistics

As we explained in the previous section, our original dataset consists of four different sources covering phone numbers used in android malware (Android ML), malicious and phishing DNS registrations (DNS), spam messages (SPAM), and scam (SCAM).

A first interesting finding is that in certain datasets the unique numbers are a small percentage of the total. For example, we observed that over 96% of the malicious domains are registered with overlapping phone numbers. This shows a strong correlation withing single datasets, and it seems to suggest that a blacklist of phone numbers could be a potentially useful feature to identify suspicious domain registrations. However, it is important to be careful before drawing any conclusions. For example, by manually reviewing the phone numbers repeating more often in the DNS dataset, we found that the majority of them belong to organizations and law enforcement agencies (CERTs, FBI, Microsoft, . . .) that were actively trying to take down large botnets. Other numbers belong to anonymization services used to hide the real information of the person registering the domain. However, by removing them from the list, the majority of the domains are still registered with overlapping numbers.

The next step consisted in acquiring the information about the type of service. Figure 3 shows the repartition for each dataset, including duplicate numbers to more precisely represent the global picture. The graph clearly demonstrates a different dominant service in each source. For example, in the DNS set we notice

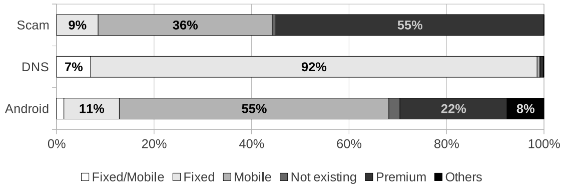

Figure 3: Distribution of phone service types

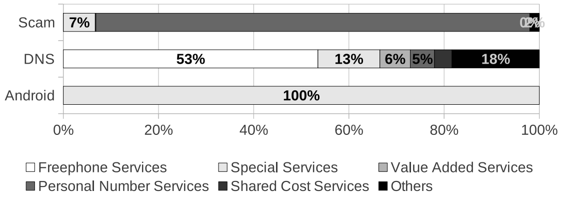

Figure 4: Premium services types in data sources (scam data omitted)

that fixed lines are used more often in domain name registrations, while most of Android ML samples are mobile numbers, and scammers (SCAM) are mostly interested in premium numbers.

Regarding the use of premium numbers, Figure 4 breaks down this category in different types. Again, we can see how different malicious activities often rely on different services. The vast majority (over 90%) of scammers' premium numbers are *personal number* services (ref Section 4), while the totality of the ones used in mobile malware belong to the more traditional *special services* category.

Finally, looking at countries, we can see that the majority of the DNS phone numbers are located in the US (59%), most of the phone numbers in Android dataset are either in China (33%) or in Hungary (20%), while most of the SCAM ones are in UK (54%) or Nigeria (21%). These results seem to confirm that different type of malicious activities are usually carried out by separate groups.

3.2 Correlations between datasets

Our first aim was to find phone numbers that are shared within different groups of criminals. For example, if a scammer registers a malicious domain for phishing purposes, does she simply re-use her phone contacts in the registration form? Unfortunately, in our collected datasets we were not able to find any case like that. This shows that such events, if they exist, are rare and may require larger datasets to be observed.

However, as reported in the general statistics, we observed overlapping inside each category. In fact, scammers often reuse the same phone numbers across different campaigns, and criminals register several different domains using the same phone contact. The lack of links between the two categories may also be due to a lack of overlapping between these groups.

4 Fraud business models

In this section we try to summarize some of the fraud business models we observed in our research. Such models were identified using information from various sources (e.g., forums, and abused users complaints) as well as the observations we made while analyzing our datasets. While some of those business models are known, many were lacking documentation or were not backed with empirical evidence.

4.1 UK Personal Numbering Services

The number ranges 070/075/076 in the United Kingdom national numbering plan are associated with *Personal Numbers* allocations [10]. Such numbers usually carry a special connection and/or a per-minute rate. There are many legitimate uses of those numbers, such as information service or hospital lines. However, as we mentioned in the previous section, those numbers are often abused by fraudsters as part of scams or by deceiving a victim to call a number and be charged an higher cost than expected.

These type of numbers are also identified as *international call forwarding service* on `419scam.org` resource, and have also been used for the study of frauds models in [15].

Such numbers can be registered online on many telecom operators, some of which are only virtual operators. These premium rate services (PRS) numbers are often offered for free, since the price of each communication is then shared between the registrant and the operator (often taking between 30 and 50%). In addition to this, operators can then forward all incoming calls to any other international phone number, thus providing a perfect anonymization service for scam or illegal activities.

In our scam dataset, we have identified 34,424 unique numbers which belong to the UK range of `07x` PRS numbers and were consistent with the allocation range of UK operators [11],

Interestingly, certain operators are used more often than other to register scam numbers (the overall distribution is presented in Figure 5). For example, we can observe that top 4 operators (out of 88) provide more than 90% of fraud-related UK PRS 07x numbers in our dataset. In particular, for one of the operator, fraud-related numbers represent almost almost 5% of it's entire 07x range - and for the next three operators, fraud-related numbers represented respectively 1.79%, 0.59%, and 0.12% of their respective 07x ranges.

Our figures are only a rough estimation, and probably a lower bound, of the real values. The percentages are in fact computed against the total allocated numbers

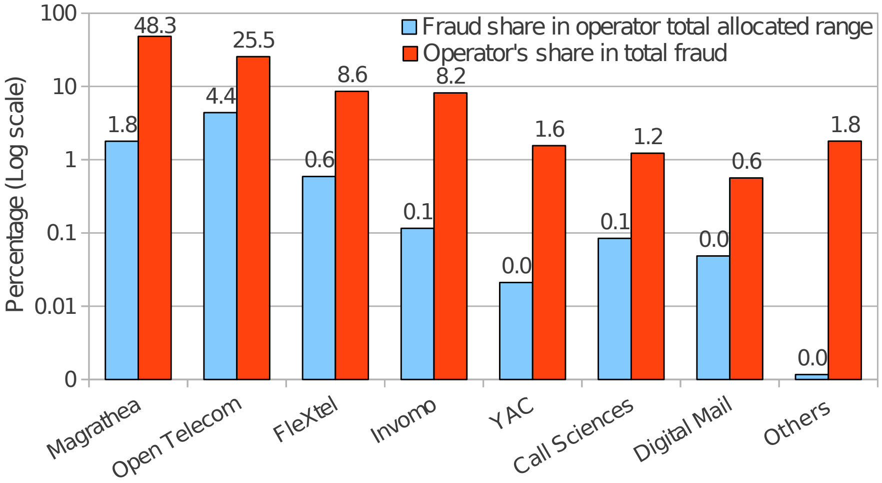

Figure 5: UK 07x fraud-share and fraud-vs-range allocation ratio.

for the operator, while in practice all the numbers may not be currently in use. Second, our data-set is limited and those percentages are based only on the numbers collected in our database.

The fact that four operators cover 90% of the fraud is surprising. Therefore, we manually compared the offers of ten operators from the list of 88. These 10 operators include the top 4 mentioned, and the other 6 were randomly chosen for the operators which had a web-site and which provided the information on their service setups. The contacts and web-sites of operators where taken from [6].

What we noticed is that operators more often associated to scam numbers normally provides three important things: 1) an online registration and configuration service, with available APIs to script and automate the process; 2) a cheap or free international call forwarding; and 3) a cash back program to pay the registrant for each incoming call.

4.2 Other Premium Phone Numbers

Figure 4 shows several other premium phone number categories, beside the one already explained in the previous section. In particular, we observed the following three cases:

- *National Short Premium*
  Those numbers can provide high profit but they are generally more difficult to set up. However, third parties businesses provide such services as simple point-and-click interfaces and also provide quick operational set-up which can be easily changed at the end of the month. These numbers are usually found in the mobile malware dataset.

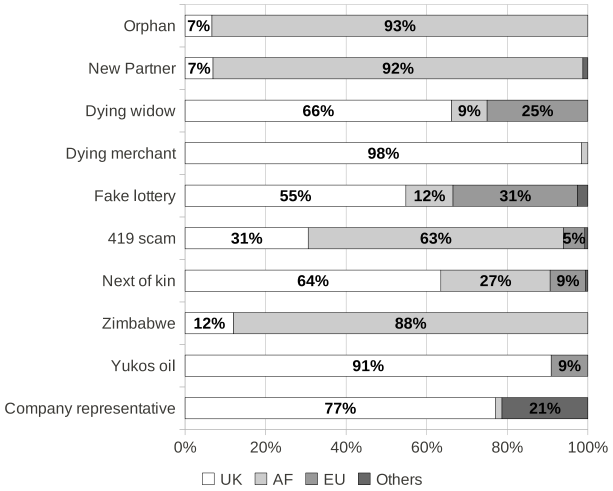

Figure 6: Scam email category preferences by phone number country codes

- *National Premium*
  Such numbers can provide moderate to high profit, with moderate to no operational costs, and quick set-up. These are found in all of our datasets.

- *International Premium*
  Those numbers are quite complex to set-up, and have high operational costs. Moreover, they are blocked by some of the operators. For these reasons, we only observed few of these numbers in the mobile malware dataset.

5 Criminals Behind the Phone

In this section, we used the SCAM dataset to evaluate the use of phone numbers to identify criminals, study their behavior, and unfold the structure and the size of their networks. Since scammers are known to provide real phone numbers, at which they pretend to be reached by their victims, this dataset is less polluted with fake or spoofed numbers, by making our results and conclusions more reliable.

The SCAM dataset covers the period from January 2009 to August 2012 (with the exception of August 2011, which is missing from our dataset [1]). For 16% of the phone numbers, we have information on the emails that were used to perpetrate the scam. These emails are classified into categories, three of which cover over 90% of the data: general scam (62%), fake lottery (25%) and next of kin (inheritance) (8%).

A first look at the relation between phone numbers and scam categories shows that this threat has a strong geographical component. As it's shown in Figure 6, certain types of scams rely prevalently on African numbers like *new partner, orphan* scams, while others like *fake lottery, dying merchant, next of kin* are almost always perpetrated by hiding behind a UK premium *forward number*. The first question we try to answer is the relationship between phone numbers and email addresses that are used by scammers as their main point of contact.

5.1 Scam communities

We start our analysis by building a graph where the nodes represent either a phone number or an email address that is used as point of contact in a scam message. The edges connecting the two types of nodes indicate that the owner of the address used that phone number in one of his scam emails. The graph has 34,740 nodes and 27,409 edges where 66% of nodes are emails and 34% are phone numbers. We remove the smallest subgraphs that are less representative by filtering out the ones smaller than 20 nodes. This leaves us with 3,681 nodes (10.6%) and 4,360 edges (16%), which consist of 699 nodes as phone numbers and 2,982 nodes as email addresses. Globally, we identify 102 communities and 79 subgraphs.

The graph, a portion of which is shown in Figure 7, shows many interesting relationships: Apparently scammers reuse the same email address with different phones, and the same phone number in multiple scam messages or in combination with different email addresses.

In particular, we observe that 37% of the phone numbers were reused by more than one scammer. Most of the largest nodes are white (phone numbers) and surrounded by several small black nodes (email addresses). This suggests that phone numbers play an important role in the activities of scammers. The set of phone numbers used by scammers in their campaigns is less diverse compared to the email addresses. In fact, email addresses are easily blacklisted and accounts blocked when their connection with criminal activities is discovered. Also, while email addresses are free, phone numbers are not. This forces the scammers to continually register fresh emails for new scam campaigns. But our analysis shows that scam phone numbers are most stable and tend to be reused over time.

By looking at the smallest subgraphs, we notice that most of them contain phone numbers registered in a single country (76%), or a country combined with UK premium numbers (10%), mostly from UK, Benin or Nigeria. This indicates that most of the scammers work alone, or in small groups located in a particular country. Figure 9 shows a real example of how scammers used four Spanish mobile phone numbers in the same campaign. All the email addresses are small variations of the same person's name, probably the victim that the scammers tried to impersonate.

However, by looking at the largest communities we see that some groups are geographically distributed over several countries. For example, Figure 8 shows how the eight largest communities are organized. They all rely on UK premium numbers (for at least 29% of their phones) and on Nigerian operators; they also own cellphone numbers in several European and African countries.

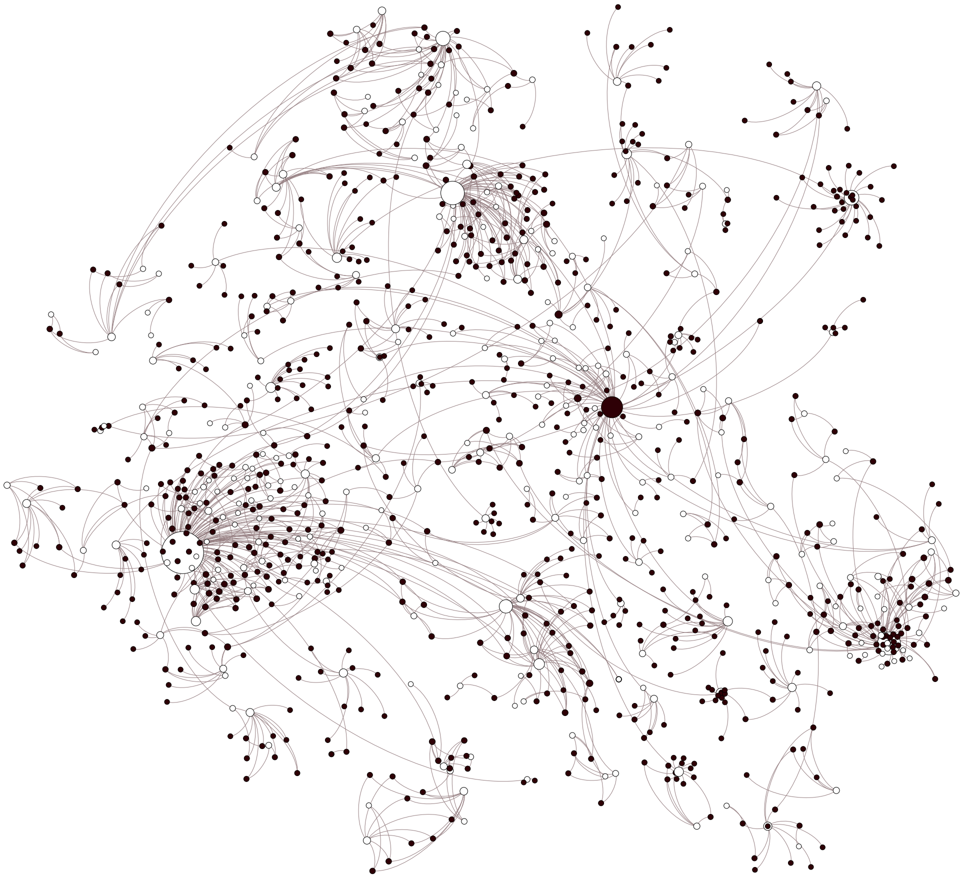

Figure 7: Visual relationships between phone numbers (white nodes) and email addresses (black nodes) that are used as point of contact in SCAM messages. The nodes' size is proportional to the number of edges.

5.2 Reusing phone numbers

We further tackle the question of reused phone numbers from a different angle. By looking at our dataset, which contains information on when these phone numbers have been used by the scammers (year and month), we understand that several of them were reused through time for long periods.

Table 2 shows that 4% of the numbers that were in use 3 years ago are still active in 2012, while from Figure 10 we notice that, as the period of time gets longer, the amount of numbers being reused grows from 21% (1 month) to 34% (3 months) and 48% over a year. In addition, a group of 307 phone numbers reappear yearly through 2009 to 2012.

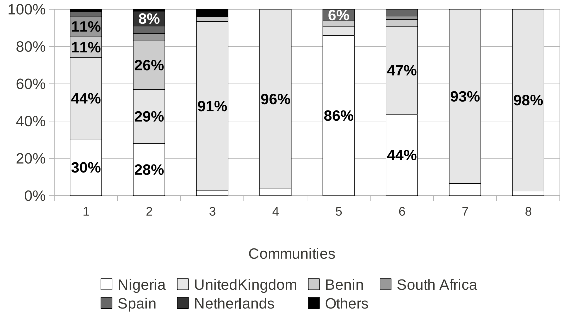

Figure 8: TOP 8 largest SCAM communities. Ordered in size from left to right.

Table 2: Phones from 2009-2011 reused in 2012

Year	Total numbers	Reused in 2012	%
2009	20,517	829	4%
2010	26,785	1,922	7%
2011	23,450	3,795	16%

5.3 Discussion

The relationship between phone numbers and email addresses suggests two interesting findings. First, phones are more stable than emails and they are reused for longer periods. Therefore, they are a better feature to detect this kind of threat. Second, even though the majority of scammers seem to operate in small groups, some communities spread over multiple countries.

However, this analysis alone is not enough to draw complete conclusions. For instance, we are still unsure how common is to reuse phone numbers: If 48% of them are reused within 12 months, does it mean that the remaining ones are discarded or that they are simply not reported by the website? Moreover, the fact that phones registered in different countries are used in conjunction with the same email address might be the consequence of individuals owning multiple SIM cards (e.g. collected by traveling abroad). In the next section, we introduce a dynamic phone analysis technique that we use to answer these questions.

6 Live Analysis of Scam Numbers

To understand the organization and the dynamics behind the scam communities identified in the previous sections, we performed periodic HLR lookups of the mobile phone numbers extracted previously. With this experiment, we aim at un-

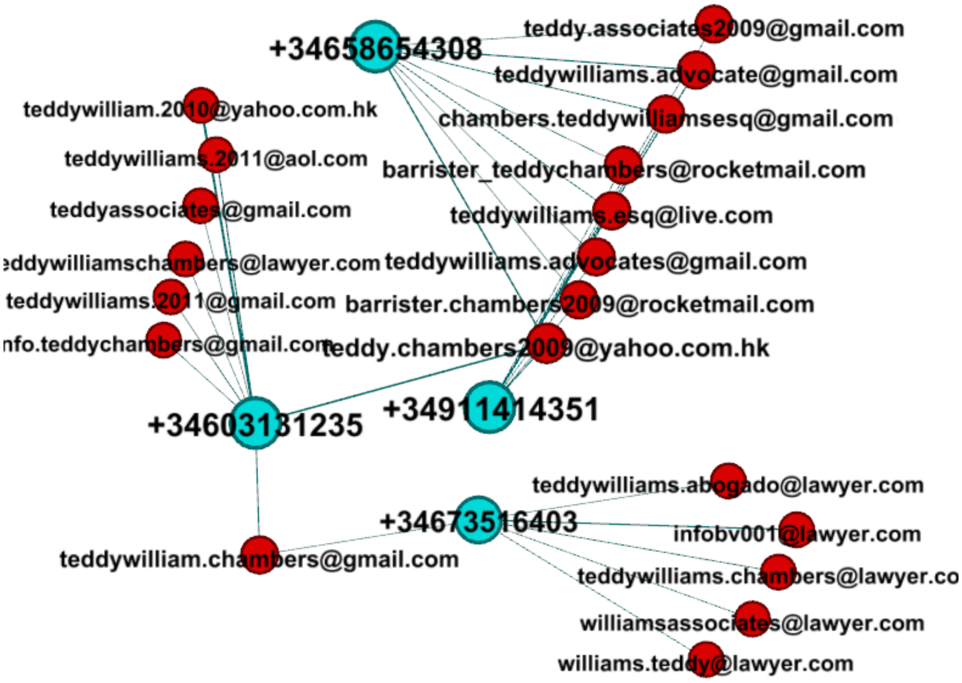
+34658654308
teddy.associates2009@gmail.com
teddywilliams.advocate@gmail.com
chambers.teddywilliamsesq@gmail.com
barrister_teddychambers@rocketmail.com
teddywilliams.esq@live.com
teddywilliams.advocates@gmail.com
barrister.chambers2009@rocketmail.com
teddy.chambers2009@yahoo.com.hk
teddywilliam.2010@yahoo.com.hk
teddywilliams.2011@aol.com
teddyassociates@gmail.com
eddywilliamschambers@lawyer.com
teddywilliams.2011@gmail.com
nfo.teddychambers@gmail.com
+34603131235
+34911414351
teddywilliams.abogado@lawyer.com
+34673516403
infobv001@lawyer.com
teddywilliam.chambers@gmail.com
teddywilliams.chambers@lawyer.co
williamsassociates@lawyer.com
williams.teddy@lawyer.com

Figure 9: Example of links between phone numbers and email addresses

derstanding how often mobile numbers are used in other countries (i.e. roaming) and over time.

As we already discussed previously, UK premium forward numbers are often choose by scammers to redirect incoming calls and anonymize final call destinations. If we exclude this category, we are left with 32,165 unique numbers, 22,537 of which are mobile phones. However, old numbers may not be used any longer or being assigned to different customers. Therefore, we finally selected the 1,333 phone numbers that were collected in July and August 2012.

Table 3: Mobile phone network status query results on 2012/08/02

Status	2012/01-06	%	2012/07	%
On the network	3,122	73%	984	84%
Replied with error	416	10%	67	6%
Turned off	734	17%	127	11%
Roaming	6	0.14%	3	0.26%

We verified that this two months period is representative of the general picture by performing a lookup on August 2nd and comparing the month of July with the numbers reported between January and June. Table 3 shows that the amount of

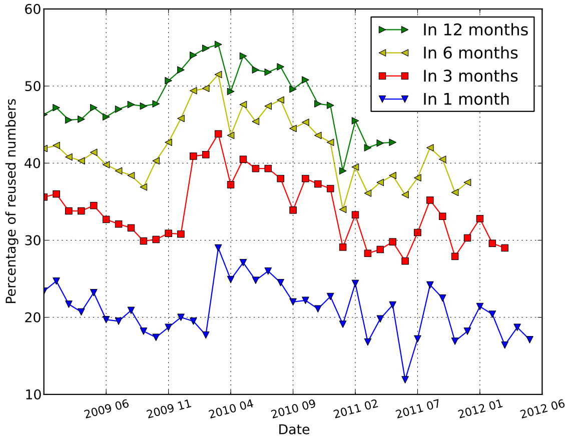

Figure 10: Accumulated shares of reused cellphones of scammers over time

mobile phones that were either reachable, roaming, or turned off is comparable in the two datasets, but more recently used numbers are more likely to be online at the time of our query. This supports the fact that after a certain amount of time some numbers might be either discarded or replaced. Interestingly, very few numbers (only 9) were roaming in a foreign country.

A first consideration is that mobile phone numbers are normally operated by criminals residing within the same country, and not used from abroad.

That is, our first experiment consisted of doing HLR lookups for the dataset of 1,333 recent mobile numbers. We did queries every three days and for a period of two months. In order to appropriate choose this query window, we looked at how often the network status of a phone number is updated, in average. A phone number first gets registered on the network and the HLR is updated instantly. When a phone gets turned off, the status is not updated, by default, but only when a call is received.

By using one of our personal phone numbers we determined the delay in a status change (i.e. from `ok` to `off`), as being 30 hours. Thus, a three days window seemed to be appropriate enough for our analysis.

By looking at changes in the network status attribute, we noticed that about half of the numbers have a constant `ok` status. This shows that scammers use phone numbers for long time periods by keeping them *online* most of the time.

It also means that they rarely switch to new numbers. In fact, only 97 phones showed to be unregistered from the network for a long time (status `absent subscriber`).

The overall distribution of the phone availability on the network is draw in Figure 12. The average scammer keeps the phone up most of the time, probably because is interested in being reached by his victims, and only 89 numbers were off more than 75% of the time.

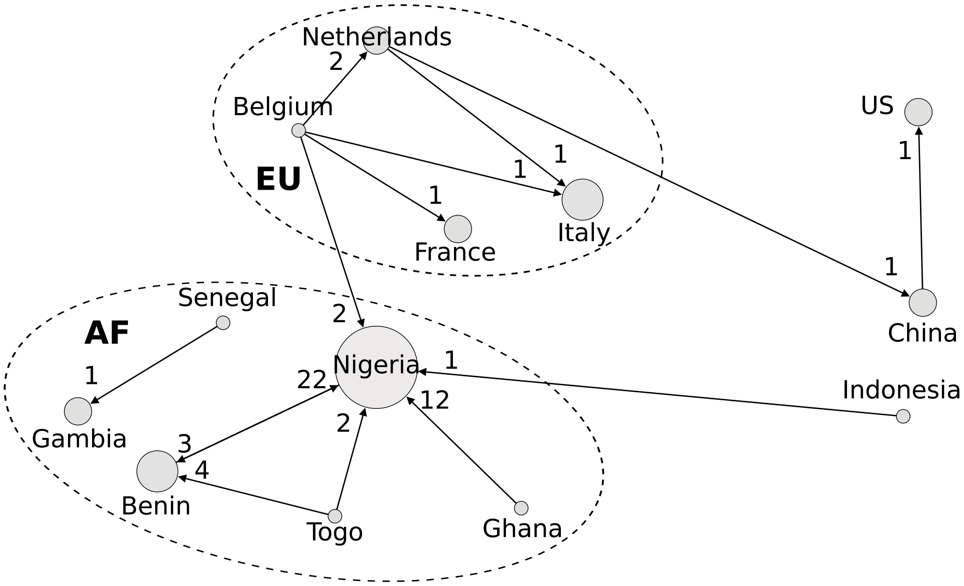

Figure 11: Mobile phones roaming per country. The arrow goes from the originating country to the roaming country. Edge labels indicate the number of roaming phones. The size of the node reflects the number of roaming phones in that country.

Finally, according to the roaming status attribute, only 50 phones were used in a different country during our evaluation (i.e. roaming). The exact roaming locations are summarized in Figure 11. The Figure clearly shows two clusters – one in Africa and one in Europe – with a small intersection of the two. Nigeria is still a key country of this type of business, with about 80% of the roaming belonging to it. This again supports our hypothesis that distributed groups exist and that they operate in conjunction from multiple countries.

We then looked at mobile operators to evaluate if some of them are preferred over others. We analyzed the market shares of the major four countries, which contain more than 700 numbers related to scam activities: Nigeria, Benin, South Africa and Senegal. Figure 13 shows the difference in distribution between the market share of each operator (data from December 2009 to December 2011) and the “scam share” between criminals.

We can see that some operators seem ignored by scammers (Cell-C in South Africa, Teracel in Benin), while others are clearly favored (GloBenin in Benin). The reason behind this might be that they have less advantageous pricing (e.g., for international calls) or stricter registration policies (e.g., strict ID checks). While Figure 13 does not explain discrepancies between operators market share and fraud share they could be used as an additional heuristic in a scam detection mechanism.

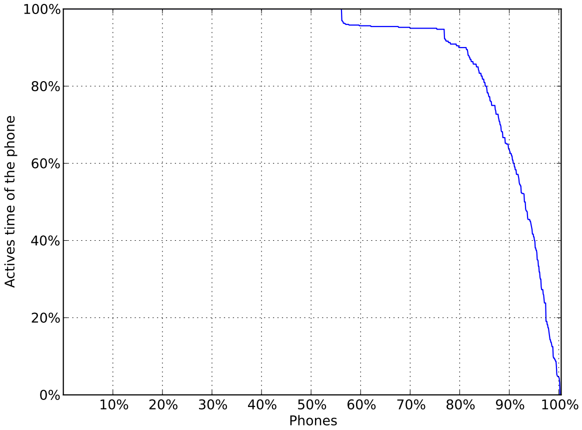

Figure 12: Mobile phone numbers sorted by frequency of `ok` status.

7 Related work

Cybercrime has become economically significant since around 2004 [31], and several research works have been conducted ever since. To this need, Fallmann et al. [21] proposed and deployed a stealthy monitoring system to capture and analyze trading information exchanged over underground Internet channels, in particular IRC and web forum marketplaces. Private forums, such as `Spamdot.biz`, are often used to conduct large-scale spam operations as Stone-Gross et al. have described in [37] by taking over 16 C&C servers.

Similarly, Holz et al. [24] monitored over a period of seven-months a dropzone used to collect keylogger-based stolen credentials. These works investigated the motivations and nature of these emerging underground marketplaces.

Scam is another popular technique employed by online criminals to harvest money from ingenuous victims. Stajano and Wilson, after analyzing a variety of scam techniques [36], raised the need of understanding "human factors" vulnerabilities and to take them into account in security engineering operations. One of the most popular scam operation, that goes under the name of *Nigerian/419* scam, has been extensively studied and reported, for example in [14] and [23]. Coomer [5] has recently patented a technique to use phone numbers to flag suspicious emails as either scam or spam. In comparison, our method takes an empirical approach and tries to correlate phone numbers to identify relationships between scammers and evaluate the role of phones in criminal activities. Also, it is unclear whether the patent is actually implemented in any real product.

In another scam variant, the so called "one-click" fraud, the victims click on a link presented to them, only to be informed that they just entered a binding contract and are required to pay a registration fee for a service. In [15] Christin et al. made a study on the entire business model behind these operations by analyzing over 2,000

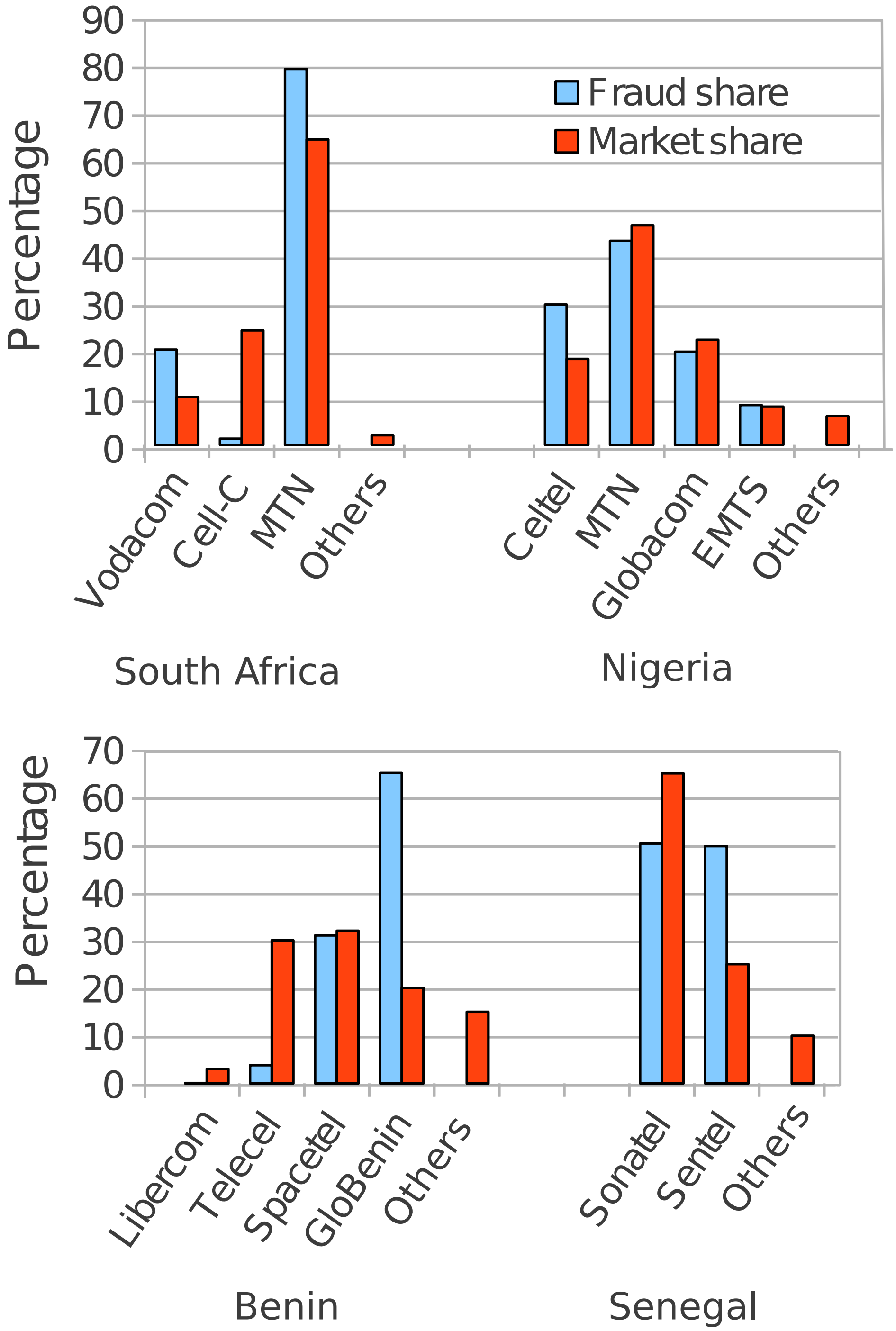

Figure 13: Distribution of mobile phone operators in Top 4 leading countries - market share vs. scam share

reported incidents and correlating them using different attributes such as whois data, bank accounts, and *phone numbers*. In particular, phone numbers have been used to analyze and cluster the actors involved in the same campaign, in a similar way as we performed in our study. Dodge [17] covers other several varieties of scams over phone numbers.

Phone numbers are often used in email scams, as *premium-rate* numbers are in fraud operations against mobile users. Porter et al. [22] analyzed 56 iOS, Android, and Symbian malware and showed that 52% of them send SMS messages to premium-rate numbers while two place *phone calls*. For example, *RedBrowser* (discovered February 2006) sends a stream of text messages, at a premium rate of $5 each to a *phone number* in Russia (as Hypponen reported in [25]). A more extensive study has been contacted by Niemel [32] who analyzed different "trojanized" and fake mobile applications that call and send SMSes to premium-rate numbers belonging to Globalstar satellite or Antarctica operators among others.

Another recent fraud that exploits telephone services for the purpose of financial rewards is *vishing* (voice phishing). Maggi [30] recently published an analysis on a real-world database of vishing attacks reported by victims through a publicly-available web application.Some papers have proposed methodologies for detecting and preventing voice-related fraud activities. Jiang et al. [27] proposed a Markov clustering-based method for detecting suspicious call, while Enck et al. [20] used lightweight certification of applications to mitigate mobile malware at install time. Finally, Prakasam et al. [12] proposed a three step approach that first identifies emerging popular international terminating numbers, then identifies correlated foreign numbers which are contacted by the same group of mobile users, and then correlates billing information to confirm the detection results.

8 Conclusions

In this paper, we have analyzed the role of *phone numbers* in cyber-crime schemes. We collected a number of datasets and designed a technique to mine phone numbers from them. A first thing we noticed is that extracting phone numbers from unstructured text is challenging and inaccurate with current tools.

We then focused on analyzing the role of phone numbers in scam related frauds. We identified different groups, created strong links between apparently unrelated email addresses and also analyzed geographic distribution of the groups' activities. A key finding was that while a phone number appeared to be a weak metric for flagging spam messages, it proved to be a much *stronger identification mechanism* in scam when compared to email addresses. This may allow to better analyze scammers operations and for example help investigations to stop such scams.

In addition to this, we discussed common business models found during our experiments. Our results show that in certain cases numbers of a few telecom operators are used to deliver majority of the phone numbers used in fraud campaigns. This also shows that some operators are preferred by fraudsters.

We conclude that phone numbers appear to be a promising metric in a handful of scenarios, ranging from scam to malicious domains registration.

References

[1] 419 Scam Fake Lottery Fraud Phone Directory. http://www.419scam.org/419-by-phone.htm.

[2] Google's common library for parsing, formatting, storing and validating international phone numbers. http://code.google.com/p/libphonenumber/.

[3] Locating mobile phones. http://events.ccc.de/congress/2008/Fahrplan/attachments/1262_25c3-locating-mobile-phones.pdf.

[4] NNPC - Worldwide National Numbering Plans Collection. http://bsmilano.it/aspx/ENG/MainFrameSet_ENG.aspx?Page=NumberingPlans_ENG.aspx.

[5] Patent US7917655: Method and system for employing phone number analysis to detect and prevent spam and e-mail scams. http://www.patentlens.net/patentlens/patent/US_7917655/en/.

[6] Premium Rate Services Network Operators Contact. http://www.phonepayplus.org.uk/For-Business/Setting-up-a-premium-rate-service/Network-operator-contacts.aspx.

[7] Python implementation of natural language guessing of a text. http://pypi.python.org/pypi/guess-language.

[8] Routo Messaging Bulk SMS services and HLR lookups. http://www.routomessaging.com/.

[9] The Koobface malware gang exposed. http://www.sophos.com/medialibrary/PDFs/other/sophoskoobfacearticle_rev_na.pdf.

[10] UK Ofcom Numbering Site. http://www.ofcom.org.uk/static/numbering/index.htm.

[11] UK Phone Info Codes Allocations Lookup. http://www.ukphoneinfo.com/s7_code_allocations.php?GNG=70.

[12] P. Appavu Siva, J. Yu, S. Ann, J. Nan, H. Wen-Ling, and J. Guy. Increased Smart Device Penetration Brings Malware Vulnerability: Methods for Detecting Malware in a Large Cellular Network. 2011.

[13] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. Exposure: Finding malicious domains using passive dns analysis. In *NDSS*. The Internet Society, 2011.

[14] J. Buchanan and A. J. Grant. Investigating and Prosecuting Nigerian Fraud. *High Tech and Investment Fraud*, 2001.

[15] N. Christin, S. S. Yanagihara, and K. Kamataki. Dissecting one click frauds. CCS '10, pages 15–26, New York, NY, USA, 2010. ACM.

[16] D. Cook, J. Hartnett, K. Manderson, and J. Scanlan. Catching spam before it arrives: domain specific dynamic blacklists. In *Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54*, ACSW Frontiers '06, pages 193–202, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

[17] M. Dodge. Slams, crams, jams, and other phone scams. *Journal of Contemporary Criminal Justice*, 17:358–368, 2001.

[18] E. Edelson. The 419 scam: information warfare on the spam front and a proposal for local filtering. *Computers & Security*, 22(5):392–401, 2003.

[19] A. Emigh. The crimeware landscape: Malware, phishing, identity theft and beyond. *J. Digital Forensic Practice*, 1(3):245–260, 2006.

[20] W. Enck, M. Ongtang, and P. McDaniel. On lightweight mobile phone application certification. CCS '09, pages 235–245, New York, NY, USA, 2009. ACM.

[21] H. Fallmann, G. Wondracek, and C. Platzer. Covertly probing underground economy marketplaces. DIMVA'10, pages 101–110, Berlin, Heidelberg, 2010. Springer-Verlag.

[22] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner. A survey of mobile malware in the wild. SPSM '11, pages 3–14, New York, NY, USA, 2011. ACM.

[23] Y. Gao and G. Zhao. Knowledge-based information extraction: a case study of recognizing emails of nigerian frauds. NLDB'05, pages 161–172, Berlin, Heidelberg, 2005. Springer-Verlag.

[24] T. Holz, M. Engelberth, and F. Freiling. Learning more about the underground economy: a case-study of keyloggers and dropzones. ESORICS'09, pages 1–18, Berlin, Heidelberg, 2009. Springer-Verlag.

[25] M. Hypponen. Malware Goes Mobile. `http://www.cs.virginia.edu/~robins/Malware_Goes_Mobile.pdf`.

[26] M. Jakobsson and Z. Ramzan. *Crimeware: Understanding New Attacks and Defenses*. Symantec Press Series. Prentice Hall, 2008.

[27] N. Jiang, Y. Jin, A. Skudlark, W.-L. Hsu, G. Jacobson, S. Prakasam, and Z.-L. Zhang. Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis. MobiSys '12, pages 253–266, New York, NY, USA, 2012. ACM.

[28] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. On the spam campaign trail. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, LEET'08, pages 1:1–1:9, Berkeley, CA, USA, 2008. USENIX Association.

[29] O. B. Longe, V. Mbarika, M. Kourouma, F. Wada, and R. Isabalija. Seeing beyond the surface, understanding and tracking fraudulent cyber activities. *CoRR*, abs/1001.1993, 2010.

[30] F. Maggi. Are the con artists back? a preliminary analysis of modern phone frauds. CIT '10, pages 824–831, Washington, DC, USA, 2010. IEEE Computer Society.

[31] T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *Journal of Economic Perspectives*, 23(3):3–20, Summer 2009.

[32] J. Niemel. Mobile Malware And Monetizing 2011. `http://www.cse.tkk.fi/fi/opinnot/T-110.6220/2011_Spring_Malware_Analysis_and_Antivirus_Technologies/luennot-files/Mobile%20Malware%20And%20Monetizing%20HUT.pdf`.

[33] A. Pathak, F. Qian, Y. C. Hu, Z. M. Mao, and S. Ranjan. Botnet spam campaigns can be long lasting: evidence, implications, and analysis. SIGMETRICS '09, pages 13–24, New York, NY, USA, 2009. ACM.

[34] C. Pollard. Telecom fraud: Telecom fraud: the cost of doing nothing just went up. *Netw. Secur.*, 2005(2):17–19, Feb. 2005.

[35] J. Shawe-Taylor, K. Howker, and P. Burge. Detection of fraud in mobile telecommunications. *Information Security Technical Report*, 4(1):16–28, 1999.

[36] F. Stajano and P. Wilson. Understanding scam victims: seven principles for systems security. *Commun. ACM*, 54(3):70–75, Mar. 2011.

[37] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: a botmaster's perspective of coordinating large-scale spam campaigns. LEET'11, pages 4–4, Berkeley, CA, USA, 2011. USENIX Association.

[38] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time url spam filtering service. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pages 447–462, Washington, DC, USA, 2011. IEEE Computer Society.