The Role of Phone Numbers in Understanding Cyber-Crime Schemes

Andrei Costin*, Jelena Isacenkova*, Marco Balduzzi†, Aurélien Francillon*, Davide Balzarotti*
*Eurecom, Sophia Antipolis, France
{andrei.costin, isachenk, aurelien.francillon, balzarotti}@eurecom.fr
†Trend Micro Research, EMEA
{marco_balduzzi}@trendmicro.it

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

Our results show that, in scam activities, phone numbers remain often more stable over time than email addresses. Using a combination of graph analysis and geographical Home Location Register (HLR) lookups, we identify recurrent cyber-criminal business models and link together scam communities that spread over different countries.

I. Introduction

In the current digital economy, cyber-crime is ubiquitous and has become a major security issue. New attacks and business models appear every year [24], [16] and criminals keep improving their techniques to trap their victims in order to achieve their, usually financial, goals. The adopted communication mechanism depends on the abuse scheme, but criminals often need to have a form of interaction with their victims. For example, the interaction channel can be a web page (phishing, selling counterfeit goods), or an instant messaging 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 [6]. In other online fraud cases, like one-click fraud [12], 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 [34].

While the role of other features in illegal online activities has been extensively studied [27] [36] [26] [15] [13], the role of phone numbers has been largely ignored. Previous work is limited to the study of spam over SMS, or to phone number abuses through premium services [33] [31] [23]. However, a recent study of fraud activity in Japan [12] demonstrates that phone numbers can 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 [6], 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 automated phone numbers analysis to improve our understanding of the underground markets, illegal computer activities and cyber-criminals in general. Second, we aim at finding patterns associated with recurrent criminal activities, in particular we automatically identify the communities responsible for Nigerian scam campaigns. Finally, we correlate the extracted information and enrich it with geographical and phone number life-cycle information from HLR lookups, to validate our hypothesis of phone numbers being actively re-used instead of discarded.

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

- We present a study of the use of phone numbers on Nigerian scam attacks.
- We show that phone numbers are a good way to automatically detect communities of scammers and study their behavior.
- 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 and phone numbers over time by different and distributed criminal groups.

The rest of the paper is organized as follows: we present the problem of automatic phone numbers extraction in Section II; subsequently we present the dataset used in our study in Section III; Section IV continues on with interesting fraud business models discovered during the experiments; subsequently in Section V we analyze criminals behind the fraud business models; we then continue on presenting in Section VI our analysis of mobile phones used in scam frauds; finally, we discuss related work in Section VII; we conclude in Section VIII.

II. Phone Numbers: Extraction and Quality

Phone numbers are often used, both directly and indirectly, in many cyber-criminal activities. For example, they appear in the registration of malicious domains, in the signatures of spam messages, in malware for mobile devices, and as main contact in scam and phishing campaigns. In some cases they are provided just to increase the credibility of some fake information, while in other scenarios they may represent a core component of the malicious activity itself.

At the beginning of our study we collected data from several sources related to illegal online activities. In particular, we focused on scam messages, spam messages, registration information of malicious domains (WHOIS) and Android malware. We selected those data sources because they are very likely to contain phone numbers and they are strictly related to cyber-crimes or fraud schemes.

After a first screening of the data, we observed a great variability in the quality and reliability of the collected information. To better describe this phenomenon, we classified the phone numbers along two directions: how difficult it is to extract them from raw data, and how reliable they are once they are properly extracted.

Extracting Phone Numbers

Properly recognizing and extracting numbers from a raw data stream proved to be quite challenging, which is consistent with results in [32]. The results mainly depend on three orthogonal factors:

How structured and easy to parse the information is: For example, WHOIS records are very easy to process and the phone number is always located inside a known and well defined field. At the other end of the spectrum, phone numbers stored in malicious binaries can be obfuscated and are, in general, very difficult to extract automatically.

How well formatted the number is: A simple regular expression can be used to extract a fully qualified number with a clearly separated international prefix (e.g., "+1 (805) 403-1234"). Unfortunately, numbers can be written in many different forms, which can be combined thus making automated parsing even harder. Phone numbers can include international prefix '+' or '00' codes, only local prefix codes, or only the phone number digits. After that, phone numbers can be grouped in variable-length groups of 2, 3 or 4 digits. Additionally, the prefixes and groups can be separated by spaces, '.', '-' or other delimiting characters, which can be country specific as well. A number without its international prefix may potentially correspond to many different numbers in different countries. Therefore, a normalization algorithm has to be used to transform the 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 leads to many false positives. This is a common problem in localized cyber-crime (e.g., malicious mobile application targeting the Chinese market) because the lack of an international prefix may force the analyst to try many possibilities, thus decreasing the reliability of the collected information. Finally, short numbers (e.g., 57341) can be very challenging to detect. In fact, since the length and format are country-specific, these numbers can be easily confused with other short sequences of digits.

How noisy the data source is: This is a measure of how often the source data includes strings of digits that can be misinterpreted as phone numbers, such as identification or reference numbers, and IP addresses. This is often a problem when parsing email messages that contain several numbers mixed with text. The presence of many sequences that may resemble valid phone numbers can greatly increase the number of false positives of the automated extraction routine.

A number of heuristics can be used to improve the extraction process. For example, the immediate context of a phone number can 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., *'.', ':'*).

The language used in the text surrounding the extracted number can also be used as a good indication of the geographic areas in which the number is supposed to be used. This is especially true for phone numbers used in scam activities, when the scammer expects the victim to call that number without ambiguity. For example, for a message written in Russian language, that includes a phone number without a full international prefix, one can try to complete the number by considering those countries where the Russian language is widely spoke, e.g., Russia '+7', Ukraine '+380', Belarus '+375', Moldova '+373'.

However, there is always a trade-off between the amount of extracted numbers and the accuracy of the results. Even by applying properly tuned heuristics, the amount of false positives when extracting poorly formatted numbers from noisy sources can be very high.

Phone Number Extraction Reliability

After a set of candidate numbers are extracted from the raw data, it is important to distinguish the real numbers from the fake ones. This is largely dependent on the type of activity and on the reason why the phone number was used by the attacker.

For example, numbers present in spam messages can be randomly-generated or spoofed to mimic existing phone numbers and to deceive anti-spam filters. Also, when registering a domain name there is often no validation of the authenticity of the provided numbers. However, in certain forms of cyber-crime the number has to be real and somehow controlled by the attacker. This is the case of premium numbers used in mobile malware or contact numbers used in scam campaigns.

Since distinguishing a fake or spoofed number from a real one is very hard, we decided to focus our analysis on a data source containing more reliable numbers. Unfortunately, the mobile malware dataset is very small and most of its data consists of short numbers. Therefore, in the rest of the paper we adopt the SCAM dataset for our study.

A potential improvement to relaible extraction could be achieved via *dynamic analysis validation*, i.e. calling the numbers. However, this technique is not feasible for many reasons, ranging from illegality of unsolicited calling or wardialing to financial infeasibility to call so many numbers. It is left as a separate future work.

III. Data Enrichment

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 dataset was manually collected, filtered and pre-processed from January 2009 to August 2012. The dataset includes meta-data on each entry, i.e., the category, message headers and, for 16% of them, the corresponding original email body.

The original dataset was enriched with the service type (e.g., mobile, land line, premium) of each phone number using two different databases (so called *numbering plans* or *NNPC*). 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 [9] which is more complete. We use both sources to cross-check the results and detect possible discrepancies.

In our SCAM dataset, we identified in total 67,244 unique normalized phone numbers. Out of them 34,424 were UK PRS (*Premium Rate Services*) numbers (51% of total) and the rest 32,820 were non UK PRS numbers (49% of total). Out of the 32,820 non UK PRS numbers, there were 29,685 mobile phone numbers.

Finally, we collected additional information about the mobile numbers by performing an HLR lookup. HLRs are databases maintained by mobile operators containing information about the current status of a phone number – i.e., the International Mobile Subscriber Identity (IMSI), roaming status, and roaming operator. This can be very useful for our study, because this allows 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 [2] 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 performing an HLR lookup periodically for a given mobile phone number, we can get insight on the evolution of it's network status. Such status information can be used to draw conclusions about activities related to a mobile phone number. We describe the use and results of this technique in Section VI.

IV. Fraud business models

In this section we summarize some of the fraud business models we observed in this work. 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 not well identified or were lacking empirical evidence.

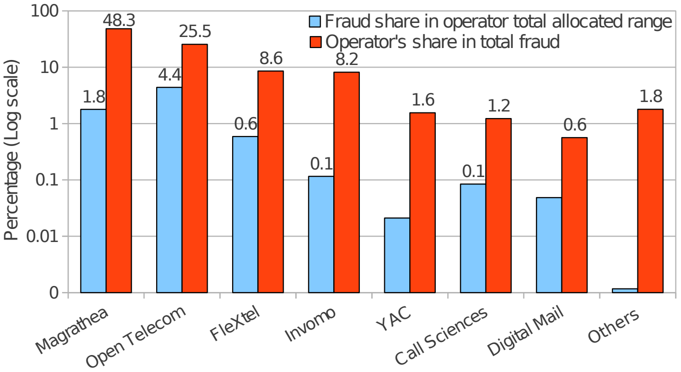

Fig. 1: UK 07x fraud-share and fraud-vs-range allocation ratio.

A. Premium Phone Numbers

Premium phone numbers can be categorized as follows:

National Short Premium: numbers can provide high profit but are difficult to set up. However, some third party businesses offer simple point-and-click interfaces to register and configure such services.

National Premium: numbers can provide moderate to high profit, with low operational costs, and quick set up.

International Premium: numbers are complex to set up and have high operational costs. Moreover, they are blocked by some telecom operators.

UK Personal Numbering Services: UK's number ranges 070/075/076 are associated with the so called *personal numbers* allocations [7]. We detail this specific category in the next section.

B. UK Personal Numbering Services

Personal Numbering Services (PRS) (also known as *international call forwarding services [12], [1]*) are premium numbers commonly used in information services or hospital lines. However, these numbers are often abused by fraudsters as part of scams or by deceiving a victim to call a number that charges higher cost than expected. As mentioned in III, there were 34,424 unique phone numbers in UK range of `07x` PRS numbers, which were consistent with the allocation range of UK operators [8].

Many telecom operators, some of which are only virtual operators, offer the possibility to register such numbers online. These are often offered for free: the price of communications is shared between the registrant and the operator (often retaining between 30% and 50%). In addition to this, operators can forward incoming calls to international phone numbers. This can be used as anonymization service to hide the actual geographic location of the scammer.

An interesting observation is that certain operators are used more often than others to register scam numbers. Figure 1 shows the distribution of phone numbers used by scammers among the providers. We observe that, in our dataset, the top

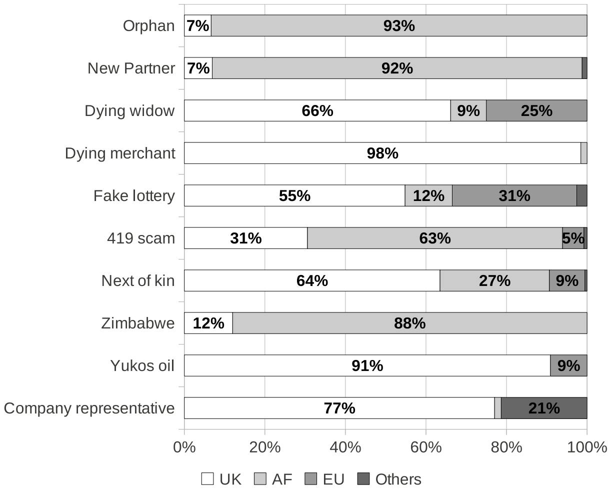

Fig. 2: Scam email category preferences by phone number country codes.

4 operators (out of 88) provide more than 90% of fraud-related UK PRS numbers. In one case, fraud-related numbers represent almost 5% of an operator allocated numbers range.

By manually comparing those and other six operators [5], we found that scammers preferred operators that:

- Have an online registration and configuration service.
- Provide an API to automate the registration process.
- Offer cheap or free international call forwarding.
- Offer a cash back program to pay the registrant for each incoming call.

Indeed, these features are appealing to scammers and, in general, cyber-criminals that perform illegal activities.

V. 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. Scammers are known to provide real phone numbers, at which they can be reached by their victims. Therefore, this dataset is less polluted with fake or spoofed numbers, which makes our results and conclusions more reliable.

A. *The SCAM Dataset*

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 the original email that was used to perpetrate the scam. These emails are classified in 10 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 scams are not evenly distributed geographically. As shown in Figure 2, certain types of scams rely mainly on African numbers (e.g., *new partner, orphan* scams), while others (e.g., *fake lottery, dying merchant, next of kin* scams) are almost always perpetrated by hiding behind a UK *personal number*.

B. *Scam Communities*

We first aimed at establishing relationships between phone numbers and email addresses used by scammers.

For this, we built 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 her scam emails. The initial graph has 34,740 nodes and 27,409 edges – 66% of nodes are emails and 34% are phone numbers. We then removed the smallest subgraphs (below 20 nodes) as they are less representative. We obtained 3,681 nodes (10.6%) and 4,360 edges (16%), consisting of 699 nodes as phone numbers and 2,982 nodes as email addresses. Globally, we identified 102 communities and 79 subgraphs.

The graph, a portion of which is shown in Figure 3, shows some interesting relationships. First, scammers seem to reuse a given email address to send scam messages, each message containing different phone numbers. Second, a given phone number seems to be reused in multiple scam messages or in combination with multiple 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 than the email addresses. In fact, email addresses are easily blacklisted and accounts are blocked when their connection with criminal activities is discovered. Also, while email addresses are virtually free, phone numbers are usually not. This forces the scammers to continually register fresh emails for new scam campaigns. Our analysis shows that phone numbers used in scams are more stable than emails 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%), originating 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 5 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 a character that the scammers tried to impersonate.

Looking at the largest communities - densely connected sets of nodes - we see that some groups are geographically distributed over several countries. For example, Figure 4 shows how the eight largest communities are organized. All these communities rely on UK premium numbers (for at least 29% of their phone numbers) and on numbers from Nigerian operators. Also, these communities use cellphone numbers in several European and African countries.

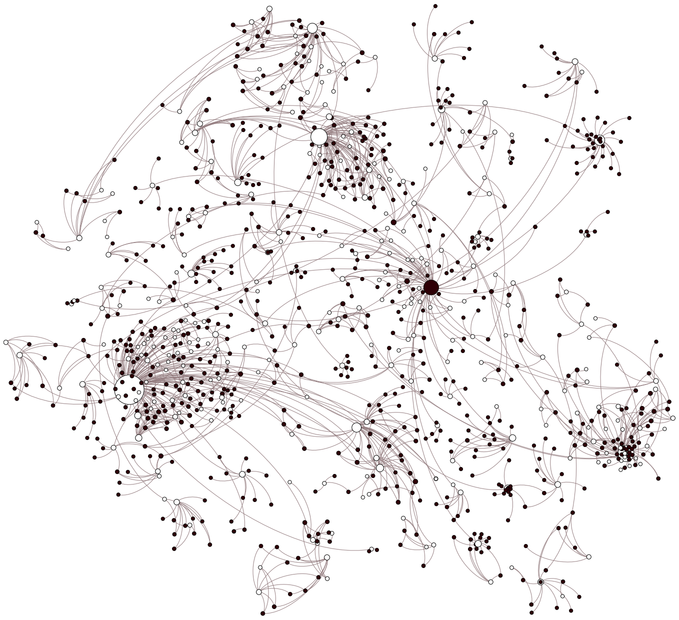

Fig. 3: Visual relationships between phone numbers (white nodes) and email addresses (black nodes) that are used as point of contact in scam messages. The size of nodes is proportional to the number of edges.

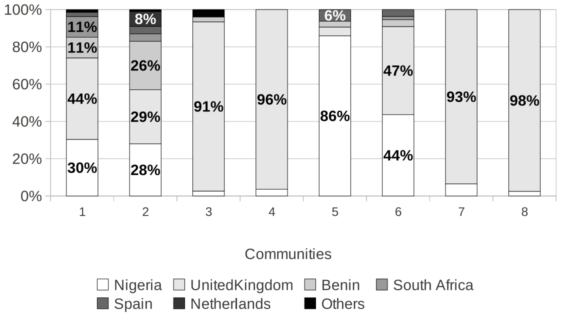

Fig. 4: Top 8 largest communities in SCAM dataset, ordered by decreasing size from left to right.

C. Reusing Phone Numbers

We further tackle the question of reused phone numbers from a different angle. By looking at the SCAM 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 over long time periods.

Table I shows that 4% of the numbers that were in use in 2009 are still active in 2012. Figure 6 shows 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 reappears yearly from 2009 to 2012. These figures do not include a detailed analysis of numbers reuse split by their type (e.g., UK PRS, mobile).

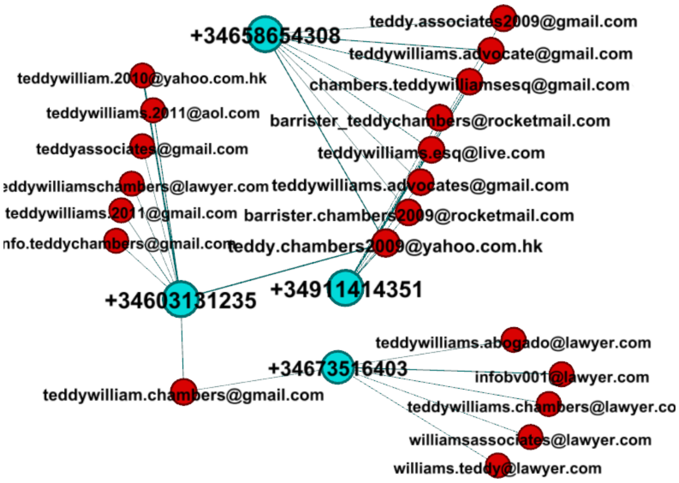
+34658654308
teddy.associates2009@gmail.com
teddywilliams.advocate@gmail.com
teddywilliam.2010@yahoo.com.hk
chambers.teddywilliamsesq@gmail.com
teddywilliams.2011@aol.com
barrister_teddychambers@rocketmail.com
teddyassociates@gmail.com
teddywilliams.esq@live.com
eddywilliamschambers@lawyer.com
teddywilliams.advocates@gmail.com
teddywilliams.2011@gmail.com
barrister.chambers2009@rocketmail.com
nfo.teddychambers@gmail.com
teddy.chambers2009@yahoo.com.hk
+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

Fig. 5: Example of links between phone numbers and email addresses.

TABLE I: Count of SCAM phone numbers encountered in 2009-2011, reused in 2012. Includes all types of numbers.

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

D. 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, phone numbers may constitute a better detection feature for the discussed threat categories. Second, even though the majority of scammers seem to operate in small groups, few communities appear to be spread over multiple countries.

However, this analysis alone is not enough to draw complete conclusions. For instance, we are still unsure how common is the phone number reuse habbit: given that 48% of phone numbers are reused within 12 months, does it mean that the remaining ones are discarded or does it mean 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 when traveling abroad). In the next section, we introduce a dynamic phone analysis technique that helps answering these questions.

VI. Dynamic Analysis of Scam Phone Numbers

In order to understand the organization and the dynamics behind the scam communities identified in the previous sections, we performed periodic HLR lookups (Section III) of the mobile phone numbers extracted previously. With this experiment, we aim at understanding how often mobile numbers are used in other countries (i.e., roaming) and over time.

As we discussed previously, UK premium numbers (PRS) are often used by scammers to redirect calls, hiding the final

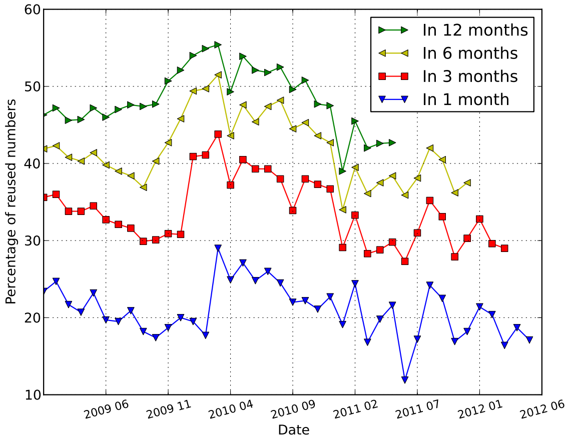

Fig. 6: Accumulated shares of reused cellphones of scammers over time.

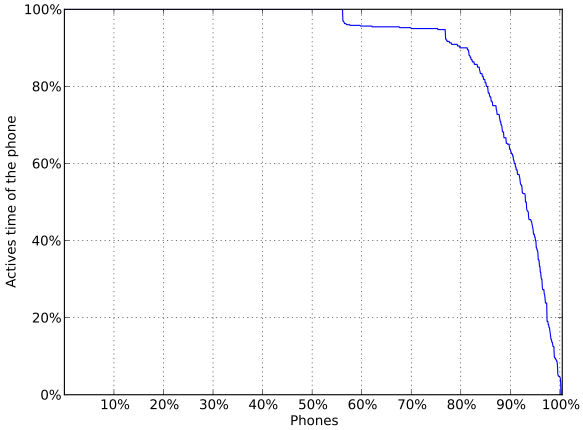

Fig. 7: Mobile phone numbers sorted by frequency of `OK` status.

TABLE II: 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%

call destination. We therefore had to exclude this category. We are left with 32,820 unique non-UK-PRS numbers out of which 29,685 are mobile phone numbers. Moreover, old numbers may be taken offline or assigned to a different customer. Therefore, we eventually selected the 1,333 phone numbers that were collected recently (July-August 2012).

We verified that the selected two months period is representative of the general picture. To verify this, we performed a lookup on August 2nd, 2012 and compared the phone numbers reported in month of July 2012 with the phone numbers reported between January 2012 and June 2012. Table II shows that the population of mobile phones that were either reachable, roaming, or turned off is comparable in the two datasets, but more recently used phone numbers are more likely to be online at the time of our HLR query. This supports the fact that after a certain amount of time some phone numbers might be either discarded or replaced. Interestingly, very few numbers (only 9 in fact) were roaming in a foreign country. A first consideration is that mobile phone numbers are normally operated by criminals residing within their own countries, and not used while abroad or roaming.

That is, our first experiment consisted of doing HLR lookups for the dataset of 1,333 recently used mobile numbers. We did queries every three days and for a period of two months. In order to appropriately choose this query window, we looked at how often the network status of a phone number is updated on 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 (e.g., from `OK` to `OFF`) as being 30 hours. Thus, a three days window seemed to be appropriate 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 phone numbers. In fact, only 97 phones appeared to be unregistered from the network for a long time (status `Absent Subscriber`). The overall distribution of the phone availability on the network is drawn in Figure 7. The average scammer keeps the phone switched `ON` most of the time and only 89 numbers were `OFF` more than 75% of the time. This appears to be in-line with the business model since scammers are interested in being reached by their victims.

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 8. 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 for 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 coordinated and collaboratively from multiple countries.

We then looked at the mobile operators, in order to evaluate if some of them are preferred over others. We analyzed the market share of the major four countries, which contain more than 700 numbers related to scam activities: Nigeria, Benin, South Africa and Senegal. Figure 9 shows the difference in distribution between the market share of each operator and the "scam share" between criminals (dataset from December 2009 to December 2011). We can see that some operators seem to be less preferred by scammers (e.g., Cell-C in South Africa, Teracel in Benin), while others are clearly favored (e.g., GloBenin in Benin). The reason behind this might be due to pricing (e.g., for international calls) or stricter registration policies (e.g., strict ID checks). Like with UK PRS numbers

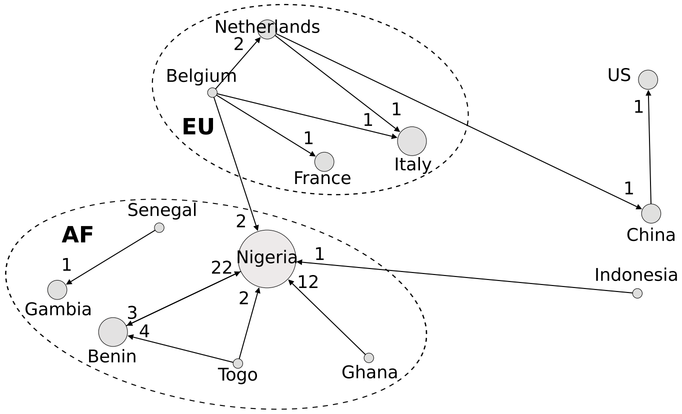

Fig. 8: Mobile phones roaming per country. The arrow goes from the originating country to the roaming country. Edge labels indicate the number of roaming phones. The size of the node reflects the number of roaming phones in that country.

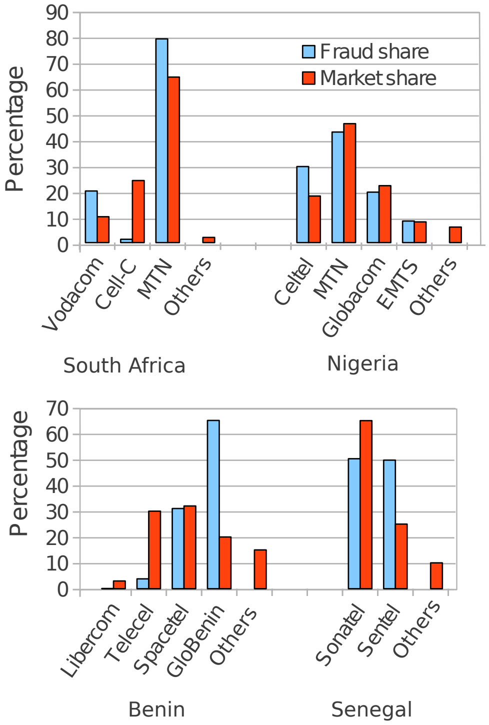

Fig. 9: Distribution of mobile phone operators in Top 4 leading countries - market share vs. scam share.

we compared market-share and fraud-share of mobile network operators, however we did not notice any discrepancy between the two.

VII. Related Work

Cybercrime has become economically significant since around 2004 [29], and several research works have been conducted ever since. To this need, Fallmann et al. [18] 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 [35] by taking over 16 C&C servers. Similarly, Holz et al. [22] 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 [34], raised the need of understanding "human factors" vulnerabilities and to take them into account in security engineering operations. One of the most popular category of scam, that goes under the name of *Nigerian/419* scam, has been extensively studied and reported, for example in [11] and [20]. Herley [21] looks into economical aspects of adversaries by trying to understand how scammers find viable victims out of millions of users, so that their business would be still profitable. Coomer [4] 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 [12] Christin et al. made a study on the entire business model behind these operations by analyzing over 2,000 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 [14] covers several other varieties of scams over phone numbers.

Phone numbers are often used in email scams, as *premium-rate* numbers, part of fraud operations against mobile users. Porter et al. [19] 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 [23]). A more extensive study has been conducted by Niemel [30] 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 [28] 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. [25] proposed a Markov clustering-based method for detecting suspicious calls, while Enck et al. [17] used lightweight certification of applications to mitigate mobile malware at install time. Finally, Prakasam et al. [10] 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.

Last but not least, [32] describes a fully automated process of address book enrichment by means of information extraction in e-mail signature blocks. This work also confirms and emphasizes the difficulties in automated parsing of email blocks for contact details, and in particular for phone numbers.

VIII. Conclusions

We analyzed the role of phone numbers in cyber-crime schemes. We collected a number of datasets and designed a technique to identify and extract phone numbers out of them. A first result 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 actors and analyzed their geographic distributions.

While a phone number appears to be a weak metric for identifying spam messages, on scams messages it proved to be a good identification mechanism when compared to email addresses. We showed that this may be helpful in analyzing scammers operations, possibly supporting investigations in order to reduce future scam messages. The reuse of phone numbers is vital in certain business models where trust must be established over a long period of time (e.g., wire funds transfer fraud). For other business models, changing the phone numbers for cyber criminals might be more vital for their untraceability. One option is to change the SIM cards, but it requires operational risks (e.g., ID checks) and other overheads. Another option is to use virtual mobile numbers (VMN). VMNs are most inviting, with competitive or free pricing, laxed ID checks, and most importantly with remote operation and high-level API automation.

In addition, we discussed common business models found during our experiments. Our results show that a restricted number of mobile operators are used to deliver the majority of fraud related numbers. This suggests that some operators are preferred over others by fraudsters.

Acknowledgment: The research leading to these results was partially funded by the European Union Seventh Framework Programme (contract N 257007).

References

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

[2] Bulk SMS services and HLR lookups. http://routomessaging.com/.

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

[4] Patent US7917655: Method and system for employing phone number analysis to detect and prevent spam and e-mail scams.

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

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

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

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

[9] Worldwide National Numbering Plans Collection. http://bsmilano.it/.

[10] 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.

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

[12] N. Christin, S. S. Yanagihara, and K. Kamataki. Dissecting one click frauds. CCS '10. ACM, 2010.

[13] 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 of *ACSW Frontiers '06*, 2006.

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

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

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

[17] W. Enck, M. Ongtang, and P. McDaniel. On lightweight mobile phone application certification. CCS '09. ACM, 2009.

[18] H. Fallmann, G. Wondracek, and C. Platzer. Covertly probing underground economy marketplaces. DIMVA'10. Springer-Verlag, 2010.

[19] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner. A survey of mobile malware in the wild. SPSM '11. ACM, 2011.

[20] Y. Gao and G. Zhao. Knowledge-based information extraction: a case study of recognizing emails of nigerian frauds. NLDB'05. Springer-Verlag, 2005.

[21] C. Herley. Why do nigerian scammers say they are from nigeria? In *WEIS*, 2012.

[22] T. Holz, M. Engelberth, and F. Freiling. Learning more about the underground economy: a case-study of keyloggers and dropzones. ESORICS'09. Springer-Verlag, 2009.

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

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

[25] 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. ACM, 2012.

[26] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. On the spam campaign trail. LEET'08, 2008.

[27] 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.

[28] F. Maggi. Are the con artists back? a preliminary analysis of modern phone frauds. CIT '10. IEEE Computer Society, 2010.

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

[30] J. Niemelä. Mobile Malware And Monetizing 2011.

[31] C. Pollard. Telecom fraud: Telecom fraud: the cost of doing nothing just went up. *Network Security*, 2005(2), Feb. 2005.

[32] G. Recourcé. Interpreting contact details out of e-mail signature blocks. In *Proceedings of the 21st international conference companion on WWW*. ACM, 2012.

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

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

[35] 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, 2011.

[36] 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*, 2011.