Flipping 419 Cybercrime Scams: Targeting the Weak and the Vulnerable

Gibson Mba
Royal Holloway
University of London

Jeremiah Onaolapo
University College London

Gianluca Stringhini
University College London

Lorenzo Cavallaro
Royal Holloway
University of London

ABSTRACT

Most of cyberscam-related studies focus on threats perpetrated against the Western society, with a particular attention to the USA and Europe. Regrettably, no research has been done on scams targeting African countries, especially Nigeria, where the notorious and (in)famous 419 advanced-fee scam, targeted towards other countries, originated. However, as we know, cybercrime is a global problem affecting all parties. In this study, we investigate a form of advance fee fraud scam unique to Nigeria and targeted at Nigerians, but unknown to the Western world. For the study, we rely substantially on almost two years worth of data harvested from an online discussion forum used by criminals. We complement this dataset with recent data from three other active forums to consolidate and generalize the research. We apply machine learning to the data to understand the criminals' modus operandi. We show that the criminals exploit the socio-political and economic problems prevalent in the country to craft various fraud schemes to defraud vulnerable groups such as secondary school students and unemployed graduates. The result of our research can help potential victims and policy makers to develop measures to counter the activities of these criminal groups.

1. INTRODUCTION

Most computer security research focuses on scams aimed at the West, especially the U.S. and Europe [5, 15, 25, 42], with a few exceptions focused on Asian countries [10, 22, 27, 30, 31, 46] and South America [36]. To date, there has been no detailed study dealing with cybercrime aimed at the African continent, especially Nigeria. Although a considerable number of studies have been conducted on the notorious Nigerian 419 scam [8, 16, 20, 24, 29, 41, 44], none of them revealed attacks focused specifically on Africans or Nigerians living in Nigeria. While the long history and advanced use of Information and Communication Technology (ICT) in the Western societies justify the large body of literature dealing with its misuse, the absence of detailed studies on other regions, such as Africa, creates the impression that these societies are immune from cyber attacks. On the contrary, there is more happening out there that has not been brought to the attention of the concerned public. Our study aims at bridging this gap, i.e., the paucity of African unique contribution to the global cybercrime, by studying in detail a form of 419 (advance fee fraud) scam unknown to the West but unique to Africa—Nigeria in particular—targeted at Nigerians living in Nigeria. This scam can be compared to One-Click Fraud [10] which uses online tools with low-tech contents to target a predominantly Japanese audience.

With a few exceptions [44], Nigerian 419 fraudsters usually have foreigners as their primary target. The reason for the criminals' preference is obvious: difficulty in investigation or prosecution arising from differences in security and legal systems of countries and higher per capita income in the targeted countries, among other factors [18]. In this work, however, we study a different form of 419 fraud targeting economically weak and vulnerable Nigerians (e.g., unemployed youths, secondary school leavers seeking admission into higher education institutions, and other vulnerable online users) as we observed in some online forums being used by the fraudsters. We investigate and highlight their modus operandi and tools of the trade showing examples of scam campaign operations carried out by members of such Internet criminal groups.

In particular, we make the following contributions:

1. We highlight a unique form of scam targeting weak and vulnerable Nigerian students, secondary school leavers, and other online users.
2. We rely on machine learning techniques to automatically identify common themes (e.g., academic, employment, dating) that fraudsters rely on to lure their victims. Our results show, for instance, that unemployment and prolonged periods of strike by university lecturers are the main drivers of this type of cybercrime.

2. THE CYBERCRIME ENVIRONMENT

In this section, we describe the source of our main data, the data extraction process as well as the social and eco-

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053892

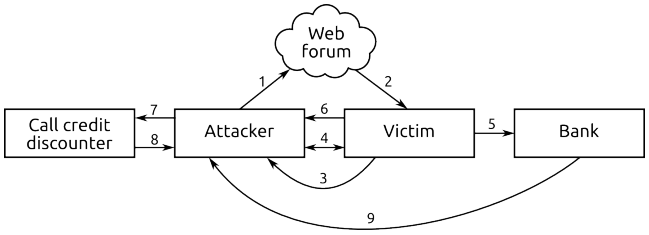

Figure 1: Forum scam structure.

nomic environment promoting the development of the cybercrime.

2.1 The Forums

Every country has its share and variants of the most common cybercriminal activities [1] and various factors driving its growth. In the case of Nigeria, the socio-political and economic problems which afflicted the country since the late 1980s impoverished many of its citizens forcing some of them to engage in crime as a business. Various forms of frauds operated initially by means of letter writing and later by telephone and faxes became widespread. With advent of Internet/email and their inherent economies of scale, the scammers wasted no time in migrating their operations to those new platforms [29].

The forum `www.topix.com` is a news aggregation website based in the United States. The website maintains different forums that cater for the interest of its users. One of those forums maintained by `www.topix.com` is the Nigerian forum[1], an open forum that enables users to login and comment on issues related to Nigeria. At inception in 2005, posts on the forum centered on news and current affairs especially those relating to Nigeria. It appears, however, that as time went on, from 2012–2013, fraudsters found the facilities provided by the website (which enables anybody to post to the forum without any form of verification or stringent registration conditions) as good incentives to exploit the platform. Thus, from 2012 onwards, scam related posts began to appear, unveiling a new, previously unstudied cybercrime, which is discussed in the rest of this paper.

The Nigerian forum hosts valuable easy-to-crawl information which sheds light on 419 scams perpetrated against Nigerians, in particular, the forum mainly hosts plain-text posts promoting different types of scam services as well as contact information—mostly telephone numbers—to reach out to the fraudsters.

Suspecting that the scammers operating on the main forum under investigation (`topix.com/nigeria`) are not likely to restrict their activities to that single forum, we used some sample phone numbers found in the dataset to perform a web engine search which revealed many pages (URLs) containing similar scam posts. We pursued the investigation further and identified some active sites which we crawled using our scraping tools to arrive at the data shown on Table 1.

2.2 Scam Structure and Scheme Monetization

Cybercrime has shifted from proof-of-concept and show-off of technical skills to monetary pursuits [21, 23]. To achieve this, every economically motivated crime must have a way of converting its scheme to cash — its ultimate target. For our dataset, we investigated the economic model and monetization schemes used by the criminals involved. For instance, the criminal would, at first contact with the victim (usually by the victim responding to an advertisement in one of the forum posts), request them to purchase mobile phone prepaid credit of 3,000 Naira (approx. USD 81.50) and send the code together with their (victim's) credentials for the required service to the fraudster's mobile phones. The fraudster would claim to use the prepaid credit to call the officers in the applicable institution who would effect the required service (e.g., an "upgrade" of examination results). The catch here is that the fraudsters would receive the prepaid credit and convert it to cash by discounting it to friends, relations or those operating paid call services at business centers. Each case would end up with the victim parting with prepaid credit without receiving the promised or advertised service(s).

Figure 1 is a schematic representation of scam operations on the forum as revealed by our study; the cybercriminal posts scam-related ads on the forum (1) to lure victims (2), who contact the fraudster via email or mobile phone calls, as advertised in the post (3). A series of interactions take place between the fraudster and the victims during which social engineering attacks are carried out to convince victims that the services are genuine (4). This phase of the discussions would result in the fraudster requesting for an advance payment in form of mobile phone prepaid credit or cash. Victims then pay the amount agreed to the fraudster's bank account (5) or as prepaid mobile phone credit (6). The fraudster then sells the prepaid credit (at a discount) received at (6) to third parties, such as friends, relations, business bureau (7). The scammer receives cash for the prepaid credit sold at (7) and withdraws cash (directly or through a mole) paid into their bank account by victims at (5).

2.2.1 Typical Prices of Scam Goods and Services

We depend on the four forums and websites mentioned in Section 2 for our data extraction and analysis. However, a simple web search using combinations of keywords usually found in scam messages (WAEC, JAMB, etc.) would reveal many of those sites. Narrowing the investigation further would yield some pages that contain scam products and their prices in plain text. Samples of such sites containing scam products and their prices can be seen on [2] and [3]. From these, we see fraudsters operating these sites advertising fake examination questions and answers for the May/June 2016 West African Examination Council, O'Level examination. Prices range from 400 Naira (USD 2) to 10,000 Naira (USD 50) depending on the number of subjects and subject combinations sought.

2.2.2 Payment and Settlement Methods

Unlike sophisticated cybercriminals that operate underground, employing different techniques to hide their activities, fraudsters operating on our forums of interest appear not to be bothered about such concealment. Scam posts are made in plain text and posted on open forums and websites

[1] `www.topix.com/forum/nigeria`

[2] `http://blastexam.com/site_43.xhtml`

[3] `http://expowap.com/site_123.xhtml`

	http://www.adsafrica.om	http://www.123nigeria.com/	http://forumng.com/
Total Posts	115,038	96,038	95,778
Unique Phones	6,204	373	651
Crime posts	50,157	75,168	94,242
% of Crime posts [+]	43.60%	78.27%	98.40%

[+] Obtained by applying the ML classification procedure described in Section 3.1 to the datasets

Table 1: Summary statistics of supplementary scam data.

with weak security controls. While the sophisticated cybercriminals would opt for monetization and payment systems that give them some elements of anonymity e.g. bitcoins, the class of scammers under our study go for simple options of receiving their payments in the form of assets that are easily converted to cash (mobile phone top-up credit) or direct receipt of cash into advertised bank accounts. A visit to the sample scam pages mentioned in Section 2.2.1 shows both payment methods being advertised.

3. GENERAL STATISTICS

In this section we present general statistics and trends of the data in the dataset explaining factors that may be responsible for such statistical occurrences. Shown on Table 2 are the general characteristics of the extracted crime data. The forum under study is a very active one, used by scammers to advertise various types of schemes. It has a total of 711,861 posts recorded between 01-01-2012 and 20-11-2013, an average of 1,033 posts per day.

The total number of posts with phone numbers is 598,572. The appearance of phone numbers in 598,572 posts out of a total of 711,861 (84.09%), most of them Nigerian registered phone numbers, suggests that the mobile phone is a tool of contact preferred by the fraudsters. We will rely on analysis of this attribute (phone numbers), relying on insight gained from [13], to further understand the criminals' mode of operations. Furthermore, careful review of the topics of the posts and reference to companies and institutions based in Nigeria in most of them further suggests that this is a crime targeted at mostly Nigerians or Nigerian residents.

3.1 Identifying Crime Posts

Not all posts found in the forum are necessarily crime related, but the majority of them are. This raises the important question, how do we automatically determine if a post is a scam or not? We treat this as a typical binary class classification problem and depend on supervised machine learning techniques (data classification) to answer it. The approach we used is explained in this section.

Dataset. We use random sampling without replacement to select 663 posts (samples) from the total of 711,861 posts in our dataset at a confidence level of 95% and error rate of 5%, based on the statistical method described in [34], [14]. We augmented the data with additional 372 random samples chosen from the data to address the "Imbalanced Dataset" problem discussed in [9].

Addressing the "Imbalanced Dataset" Problem. A phenomenon known as the "Imbalanced Dataset" [9] occurs when the y-label of one class in an n-class classification problem overwhelms the labels in other classes. The tendency in such cases is for the learning algorithm to generalize and categorize (predict) all cases in favour of the majority class. The solution to the Imbalanced Dataset problem is to balance it artificially through up-sampling or under-sampling. In the up-sampling method, we craft more data of the type of the minority and inject it into the training set to bring the minority at par or near par with the majority class. The under-sampling approach does the balancing by removing some of the data items in the majority category to bring it at par to the minority - this approach can result in loss of some features.

For our problem, we chose the over-sampling approach by randomly searching through our dataset to select non-crime data items and include them in the training / development set to achieve a reasonable balance in the classes of posts contained in the training dataset. This resulted in a rebalanced dataset of 1,035 posts (samples) from the total of 711,861 posts. This sample set was then randomized and adopted as the training set after checking that it captures most relevant information about the main dataset [38].

Feature Sets. The body text of posts contained in the development set was adopted as the feature set after stemming and removal of stop words. The words frequency based on bag of words approach commonly used in NLP was used to extract words from the texts.

Training Set and Class Labels. As explained earlier, we randomly sample 1,035 messages out of 711,861 at a specific confidence level (95%) and error rate (5%). The reason is that we needed to build ground truth. To this end, we manually vetted the messages. It is not feasible to carry out such an activity on the whole corpus (711,861 messages), but this becomes a reasonable task on a smaller dataset: we then rely on statistical sampling theory to provide statistical bounds on errors and confidence by looking at 1,035 instead of 711,861. Each line of the training set was manually inspected and assigned a class label of "Yes" if the post suggests it is crime-related or "No" otherwise. The following heuristics were developed and used to decide if a post is crime-related or not:

1. Any post advertising that JAMB[4], WAEC[5], or GCE[6] form is out and asking to be contacted is a scam; 2. Any post offering to give assistance to any candidate to gain admission into any institution of learning is a scam; 3. Any post offering any form of fun services e.g., sex for money is a fraud; 4. Any post offering any form of assistance for job or employment is a scam; 5. Offers of spiritual / religious assistance e.g., prayers, illuminati membership, magical powers, healing etc.; 6. Offering products e.g., Dangote Cement below its market price; 7. Advertising forbidden or illegal services e.g., sale of kidneys or sperm, or porn services; 8. Offering any form of introduction e.g., recruitment

[4]Joint Admissions and Matriculation Board
[5]West African Examination Council
[6]General Certificate of Education

Item Description	Number
Total Threads	711,861
Posts with Phone numbers	598,572
Total Unique Posts	589,956
Total Unique Authors	37,948
Distinct Locations	613
Distinct Phone Numbers	12,425

Table 2: General statistics of the crime data.

Provider	No.	Percentage
MTN	7,566	60.89%
GLO	2,695	21.69%
ETISALAT	932	7.50%
AIRTEL	920	7.40%
OTHERS	312	2.51%
TOTAL	**12,425**	**100.00%**

Table 3: Distribution of phone no. by telcos.

Metric	SGD	SVM
Accuracy	94.56%	95.17%
Precision (Sensitivity)	96.01%	96.54%
Recall	95.41%	95.80%
Specificity	93.18%	94.14%
F1	95.68%	96.16%
Error	5.44%	4.83%

Table 4: k-fold (k = 5) validation on SGD and SVM on dataset.

to the Nigerian music or movie industry or a popular artist, audition, football academy, etc.; 9. Offering financial loan services; 10. Assistance for procurement of visa services; 11. Assistance with pool betting and casinos including the Nigerian version – "Baba Ijebu"; 12. Any form of assistance to get rich quick; 13. Offers of scholarships; 14. Any offer for "EXPO" services – selling exam questions (and answers) in advance of the examinations; 15. Any post advertising upgrades of WAEC or UTME[7] grades; 16. Any post advertising telecommunication cheats codes involving entry of prepaid credit.

Choice of Machine Learning Algorithm. Since our experiment involves classification of text data, we chose Support Virtual Machine (SVM) – a commonly used algorithm that has been shown to perform well on text classification tasks [45] [32].

ML Model Development. The training (X-labels) set was fed into an ML pre-processor for vectorization and conversion to Term Frequency – Inverse Document Frequency (TF-IDF) using the bag of words contained in the posts after stemming and removal of stop words. The output of the vectorization and the class labels (Y-labels) were then fitted into the Support Vector Machine (SVM), the purpose being to evaluate the utility of the algorithm on our dataset.

Model Validation. To hinder overfitting, we evaluate our approach in a 5-fold cross validation setting. Table 4 shows the results, with SVM slightly outperforming SGD with 96.16% F1-measure.

The Support Vector Machine and Choice of Kernel Parameters. Support Vector Machine (SVM) is a machine learning algorithm suited for supervised learning problems (classification and regression). Developed in 1992 [7], the algorithm has found a wide range of applications such as text classification, image recognition and written digits recognition. The algorithm works by computing an optimal hyperplane that separates a given set of data into two distinct classes. The algorithm is versed in statistical learning and mathematical theories whose details are beyond the scope of this work. The algorithm works with some parameters that need to be tuned to obtain optimal performance. In this work, we depend on the implementation of the algorithm contained in the Scikit-learn package [39] with its default parameters (kernel = rbf) which we consider adequate for our relatively small sample dataset. However, in high-dimension data, the default parameters may not be adequate and would need to be set to appropriate values.

Classification of the Entire Dataset. Based on the satisfactory performance of the SVM model on the development set, we applied the model on the entire dataset (711,861 less 1,035 posts used for training). The classification resulted in 679,222 (95.55%) of the posts being classified as crime (Yes) and the balance of 31,604 (4.45%) posts classified as non-crime at the chosen confidence level of 95% and error rate of 5%.

Conclusion. The forum is a crime hub used by scammers to advertise various forms of schemes aimed at deceiving and exploiting their victims. In Section 4.2, we will attempt to group the posts by crime category.

3.2 Other Characteristics of the Dataset

In the following, we discuss the characteristics of the data in more details.

Distinct Phone Numbers and their Distribution. The total distinct phone numbers found in the posts is 12,425, clearly showing that this is a crime riding mostly on the back of Internet and mobile phones (see Table 3).

The four GSM mobile phone companies operating in Nigeria (MTN, GLO, Airtel, and Etisalat) account for 97.49% of the unique phone numbers found in the dataset. This information strongly suggests that any major attempt at preventing or reducing this type of crime must involve the mobile phone companies.

In April 2013, Nigeria launched the Mobile Number Portability (MNP) scheme to allow users to switch from their mobile service operator to another, without changing their phone numbers. We wanted to see if the MNP scheme would adversely affect the values in Table 2. According to the 2013 Year End Subscriber Network Data Report [12] released by the Nigerian Communications Commission (NCC), the total number of ported lines was only 0.09% of the total connected lines, as of December 2013. This shows that there was minimal migration of users across mobile operators, and Table 2 remains valid.

4. DETAILED ANALYSIS

We present results of the analysis performed on our dataset, complementing the information presented above.

[7]Unified Tertiary Matriculation Examination

Specifically, we show scam growth patterns and the activity of top actors.

Days and Posts Collection. Each day encompasses business activities for the criminals, totaling 689 days worth of cybercrime-related posts. The top ten posters are responsible for 5.71% of the total posts with phone numbers. In fact, the most prolific poster accounts for 1.12% of the total posts. Cybercrime disruption intervention should start by targeting top players.

Pareto Principle (80-20 rule). The Pareto principle states that for most events, 20% of causes are responsible for 80% of outcomes. We tested this principle on the dataset and found that 20% of posters (phone numbers) are indeed responsible for 89.89% of crime posts containing phone numbers, thus confirming our dataset as being in conformance to the Pareto rule. This information strongly suggests that any operation that disrupts the activity of this 20% will almost eradicate—even if temporarily—criminal activities on the forum.

Monthly Growth of Posts. We observe a steady increase in the number of posts from a low value of 218 in January 2012 to 142,344 posts in September 2013 (Figure 2). One noticeable point from this data (Figure 2) is a sharp drop from October 2013. This corresponds to the period during which lecturers in the universities called off their six-month industrial action. We attribute the drop to the fact that most of the fraudsters are students who need to temporarily step down their activities to return to their studies. However, we expect the growth to continue as soon as they are fully settled in school. This finding further supports the observations in [2] and [43] that Nigerian youths and undergraduates are actively involved in cybercrime, a trend that if not checked will continue to contribute to the declining standard of education in the country.

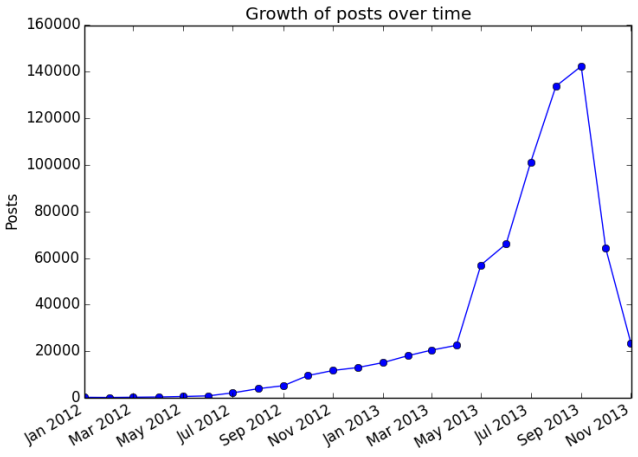

Figure 2: Growth of posts.

4.1 Data Analysis

We first aim at classifying forum posts according to the type of scams they relate to. The purpose of this classification is to gain deeper insight into the various schemes that are in common use by the fraudsters and the reason such schemes are in use.

To infer the type of scams (e.g., job, academic), we rely on simple yet effective features, such as keywords contained in the post topics. Thus, posts bearing words, such as *job*, *employment*, and *recruitment* would be classified as *Employment scams*; those relating to *exam*, *examination*, *WAEC* (West African Examinations Council), and *JAMB* (Joint Admissions and Matriculations Board) would instead be classified as *Academic scams*. Using a bag-of-words model used in natural language processing and information retrieval, we extracted 200 most frequent words that appeared in the forum posts. We then (manually) grouped them into five distinct sets to derive following categories: Employment, Academic, Dating, Spirituality, and Others (not in any of the previous categories).

4.2 Automatic Data Classification

We treat the grouping of forum posts into various (related) crime categories as a multi-class classification task and adopt supervised machine learning procedures to solve it as follows:
Dataset. We adopt the 679,222 posts identified as crime-related in our dataset in Section 3.1 as our dataset for the multi-class problem.
Training Set. The 655 crime posts out of the 1,035 samples selected in Section 3.1.
Training Set Y-labels (Class labels). Academic, Employment, Dating, Spirituality, and Other.
Feature Sets. Words frequency from bag of words contained in forum posts after stemming and removal of stop words.
Choice of ML Algorithm. SVM based on its good performance on our dataset from the experiment on Section 3.1.
Model Validation. We applied 5-fold cross validation on the resulting model using the training set as input and obtained the performance shown on the confusion matrix (Table 5).
Multi-Class Dataset Classification. The result of classification of the crime subset of the dataset using the SVM model (Table 6) shows that posts exploiting academic fraud are the most significant, accounting for 68.32% of the posts. We are not surprised by this result. For example, the results of the May/June 2014 West African Examinations Council (WAEC) exams released in August 2014, shows that over 68% of candidates that sat for the exam failed to achieve the five required minimum credits (including English and Mathematics) required to gain entrance into institutions of higher learning in Nigeria. Performance has continued to dwindle over the years - being at abysmal levels of between 23 to 38 percent since 2008, except in 2011 when 42 percent was achieved.

WAEC is not the only examination body that is experiencing dwindling performance. In fact, cheating and academic fraud are common occurrences in most public examination bodies including the Joint Admissions and Matriculation Board (responsible for entrance examinations into institutions of higher learning in Nigeria), GCE, academic departments of institutions of higher learning, among others.

Second to the academic fraud category is Employment fraud with a score of 19.11%. This can be linked to high unemployment rate in Nigeria which is stated as 23.9% in 2011 according to Nigerian Bureau of Statistics [37]. Fraudsters exploit this development by advertising non-existent jobs on different media in order to defraud unsuspecting victims. The Dating/Romance Category constituting of 3.37%, targets men/women in search of partners or sexual relation-

	Academic	Employment	Spirituality	Other	Dating	**Total**
Academic	**413**	7	1	9	1	**431**
Employment	2	**136**	0	2	0	**140**
Spirituality	0	0	**16**	1	0	**17**
Other	1	5	1	**23**	5	**35**
Dating	1	0	0	0	**31**	**32**
Total	**417**	**148**	**18**	**35**	**37**	**655**

Table 5: Confusion matrix: k-fold validation (k=5).

Class	**Posts**	**Percentage**
Academic	464,069	68.32%
Employment	129,811	19.11%
Other	48,228	7.10%
Dating	22,897	3.37%
Spirituality	14,217	2.09%
Total	**679,222**	**100.00%**

Table 6: Inter-class data classification.

ships. Confronted with various social and economic problems prevalent in the country, many people would embrace religion/spirituality in search of solutions to their problems. The Spirituality category made up of 2.09% targets victims in this category. The *Other category* of 7.10% refers to posts that do not belong to any of the previous categories. More insight on this class is given next.

Inside the "Other" Category. As stated earlier, this category constitutes 7.10% of the dataset while the themes used by the fraudsters are varied. One of those schemes which we observed in this set is the use of the Dangote brand. Alhaji Aliko Dangote is a Nigerian billionaire and Africa's richest man, ranked number 24 in Forbes World's Billionaires [19]. Some of the schemes impersonate his company by purporting to sell his products at far below market prices while others claim that the company is offering loans on generous terms to Nigerians. We found a total of 2,188 posts and 719 unique phone numbers pertaining to this brand's impersonation. Other brands being impersonated include telecommunication service providers and banks.

4.3 Clustering and Visualization

Text clustering aims to organize a given set of text documents into related groups called clusters. The process of finding relationships among the documents and grouping them together is referred to as clustering. Clustering aims to give order to otherwise disorderly data with the objective of aiding analysis and knowledge discovery [26]. The clustering (grouping) is done in a way that documents within a given cluster are related (similar) and different from documents in other clusters with respect to some measurement parameters. Text document clustering is a well-researched and active application area for Machine Learning covering tasks like document organization and corpus summarization [3], web mining, and search engines [40].

Our corpus being mostly text-based, we saw it as a good candidate for the application of unsupervised learning (clustering) for the discovery of knowledge within the data. Specifically, we set out to cluster a selected sample of the data (for convenience) to see if there are groups of criminals coordinating their activity. By grouping together some related posts, we hope to discover some posts linked to the same phone number(s) probably indicating coordination of activities among the fraudsters (existence of criminal gangs).

Using the random sampling without replacement method [34], [14], we selected 16,194 posts (samples) from the total of 679,222 posts in our dataset categorized as crime-related posts at a confidence level of 99% and an error rate of 1%. We then applied ML pre-processing steps (count vectorization, scaling, dimension reduction and transformation) [35] to the data to get it in a form amenable to ML.

Choice of Clustering Algorithm. For the clustering task, we chose Density-based spatial clustering of applications with noise (DBSCAN) [17] The algorithm works by grouping together data items that are closely located (with many nearby neighbors) and treating the rest(items with few nearby neighbors) as outliers. Because of its great impact in the data mining community, the algorithm was awarded the 2014 SIGKDD Test of Time Award reserved for KDD Conference papers that have significantly impacted the data mining research community beyond the last decade [33].

Our sample data was fitted into the DBSCAN algorithm yielding 197 clusters. Some of the clusters were fed into Gephi [6], a network manipulation and visualization package. Figures 3 and 4 show the same post topics being used by different phone numbers, an indication of possible coordination of activities among scammers. It could also be as a result of some fraudsters trying to copy post topics of other scammers. Figure 5 shows the owner of the phone number 07033589XX (shown as a node) specializing in scams that target candidates seeking recruitment into military and paramilitary agencies (the army, air-force, navy, police, etc.). The scammer's (07033589XX) activities are further linked to those of other scammers via related post topics. Figure 6 is an example of a more elaborate scheme showing many scammers identified by phone numbers, coordinating campaigns, sharing post topics and scam schemes. This type of information is likely to be of interest to security researchers and law enforcement.

5. DISCUSSION

In this section, we discuss a countermeasure instituted by the Nigerian telecommunications regulatory authorities (aimed at reducing crimes perpetrated using mobile phones such as the one seen in this study) and note the ineffectiveness of the measure in addressing the type of cybercrime we studied, especially during the period of our study. In addition, we give recommendations on future work to address this problem.

Figure 3: Sample cluster.

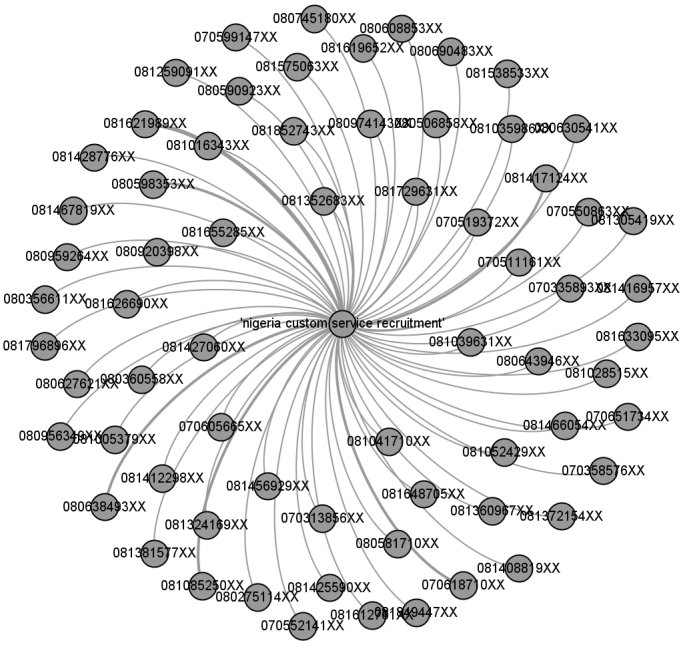

Figure 4: Sample cluster.

5.1 Curbing Crime through SIM Card Registration

The Nigerian Communications Commission (NCC) enacted a policy in 2011 requiring all GSM phone owners to register their SIMs with their telecommunication service providers [11]. The registration involves capturing some essential personal data and biometric information of the subscriber into a database. The NCC explained that one of the key objectives of the SIM registration was to assist law enforcement agencies during the investigation of crimes employing such phone numbers. Since our work concerns investigation of the use of mobile phones for cybercriminal activities, we decided to further analyze our data to see whether the SIM registration introduced by the NCC actually helped in reducing cybercrime on the forum, or influenced attacker behavior in any way.

Our study of activities on the forum from inception to the end of 2011 (the year that SIM registration started) shows that the number of posts on the forum were very few and phone numbers were rarely present. However, events (posts) on the forum took a different dimension from 2012 (a few months after the commencement of SIM registration), with the number of posts shooting up from 47,875 in 2012 to 663,986 in the first eleven months of 2013 and the number of recorded phone numbers rising from 1,008 to 11,763 in the same period. From this, it appears as if the SIM registration policy, instead of curbing crime activities on the forum, has encouraged its growth instead. But that is not the situation. We think the real issue is the absence of cybercrime law in the country up till the end of 2014 – one was enacted in November 2014; and weak investigation and prosecution capabilities on the part of the law enforcement agencies. Various laws and enactments exist in the Nigerian law books that deal with fraud and other financial crimes (e.g., Section 419 of the Nigerian criminal code from which the notorious 419 scam got its name and the economic and financial crimes bill), but none of them was strong and comprehensive enough to deal with modern day intricate and sophisticated cybercriminal activities.

Furthermore, the law enforcement agencies – the police, Economic and Financial Crimes Commission (EFCC) and Code of Conduct Bureau are not well equipped to investigate and prosecute this type of crime.

In conclusion, the enforcement of GSM registration in the country in 2011 would not be a serious bother to the cybercriminals involved in this forum. For instance, the NCC reported that the telecommunication companies have not been cooperative in the implementation of the SIM registration policy resulting in the unprecedented fine of USD 5.2bn imposed on MTN by NCC and the loss of 5.1 million subscribers by MTN in August 2015 as a result of the NCC sanction [4].

5.2 Recommendation for Future Work

Opportunity exists to improve this work by carrying out interdisciplinary research, meshing together the economics of institutions with the economics of security to understand the social, political, and economic factors influencing the development of this type of cybercrime in its environment. Another possible direction to extend this work is to carry out interdisciplinary studies exploring whether more people will fall for those scams in Nigeria than the UK, for instance. Yet another possible research direction is to build a browser plugin to flag posts on forums, and "phone home" to us, to report the geo-location of victims that click through scam posts despite warnings from the plugin. Another possible way to extend this work is to carry out a measurement study searching for forums containing scam services and their prices with the aim of quantifying the size of this

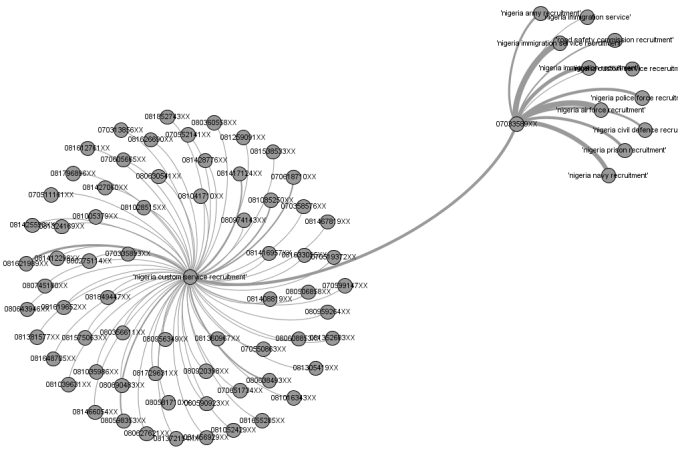

Figure 5: Sample cluster.

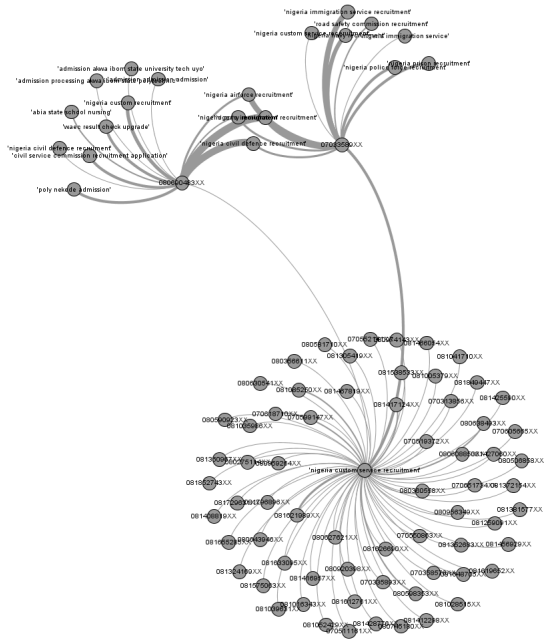

Figure 6: Sample cluster.

scam. These would allow us to have a more comprehensive view of what is happening out there.

6. RELATED WORK

Cybercrime in general and 419 in particular, have attracted investigative interest of many authors and researchers in recent times. Costin et al. [13] investigated cybercriminals' use of phone numbers in various fraud schemes. The study shows that phone number analysis can assist investigators and crime fighters in understanding criminals' mode of operations in an online setting. Knowledge of such schemes can be vital in assisting stakeholders to develop defense strategies against such crimes. As the study found out, phone numbers are stable attributes which are used over a long time by the criminals to reach their victims unlike other volatile attributes like email addresses. In this study we rely on the analysis of phone numbers contained in our dataset to understand the activities of the criminals using the forum as a platform.

Isacenkova et al. [28] studied Nigerian 419 operations using phone number and email address attribute analysis. The authors applied machine learning and graph visualization techniques to analyze attributes of crime data contained in an online repository and revealed the importance of phone numbers and email address attributes in linking and grouping together crime events attributable to the same set of criminals—knowledge vital in understanding crime operations and evolution over time. Their work is closely related to ours in the form of techniques applied—analysis of crime data attributes (phone numbers, email, others) using machine learning techniques and depiction of the actors' operations using graph visualization tools—but differs from it with respect to the data sources used and period covered. Our main data was extracted from an active online discussion forum[1] and covers the period Jan 2012–Nov 2013 complemented with recent data from other online webpages, while Isacenkova et al. [28]'s data consists of aggregated email data stored on `www.419scam.org`, a *scam aggregator* website, covering the period Jan 2009–August 2012. Another major difference between the two studies is the victims being targeted by the criminals responsible for the two data sources. For the `www.419scam.org`, the targets seem to be email users from all over the world while the Nigerian forum[1] focuses on Nigerians resident in Nigeria especially students and unemployed youths, allowing us to shed light on this previously unstudied cybercrime. While the scam and targeted audience are different, the methodology used is similar to that in [10].

7. CONCLUSION

In this paper, we made the point that existing computer security research is not balanced in the coverage of the global online community. Most existing cybercrime research focuses on the US, EU, and some Asian countries to a little extent while some regions especially Nigeria and other African countries are completely neglected. This imbalance in study coverage leaves a lot of room for cybercriminals to hide in under-researched regions to practice unique types of cybercrimes in those regions unknown to the rest of the global online community, thus contributing to the insecurity of the Internet. We underscore this point by investigating a type of cybercrime peculiar to Nigeria and unknown to the West. The main targets of this category of cybercriminals are unemployed youths and secondary school leavers while the high rate of youths' unemployment and dysfunctional education system exacerbated by prolonged strikes by university lecturers were the main catalysts propelling the crime. We conclude by proposing some potential future work.

Acknowledgements

We wish to thank the anonymous reviewers for their comments. Gibson Mba and Lorenzo Cavallaro were partially funded from the European Union Seventh Framework Programme (FP7-SEC-2013) under grant agreement number 607642. Gianluca Stringhini was supported by the EPSRC under grant number EP/N028112/2. Jeremiah Onaolapo was supported by the Petroleum Technology Development Fund (PTDF), Nigeria.

8. REFERENCES

[1] L. Ablon, M. C. Libicki, and A. A. Golay. Markets for Cybercrime Tools and Stolen Data: Hacker's Bazzar. http://www.rand.org/content/dam/rand/pubs/research_reports/RR600/RR610/RAND_RR610.pdf, 2014. Accessed: 2014-07-24.

[2] A. Adeniran. The Internet and Emergence of Yahooboys sub-Culcture in Nigeria. *International Journal of Cyber Criminology*, 2(2):368–381, July-December 2008.

[3] C. C. Aggarwal and C. Zhai. A Survey of Text Clustering Algorithms. https://pdfs.semanticscholar.org/88c2/5e2481ba49cbac75575485cba1759fa4ebcc.pdf, 2012. Accessed: 2016-05-03.

[4] C. Akwaja. NCC Fines MTN Nigeria N1.04trn for SIM Deactivation Default. *The Leadership*, 26 October 2015. http://www.leadership.ng/news/469815/ncc-fines-mtn-nigeria-\n1-04trn-for-sim-deactivation-default.

[5] R. Anderson, C. Barton, R. Bohme, R. Clayton, M. Eeten, M. Levi, R. Clayton, and S. Savage. Measuring the Cost of Cybercrime. http://weis2012.econinfosec.org/papers/Anderson_WEIS2012.pdf, 2012. Accessed: 2012-11-05.

[6] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[7] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory 5*, 1992.

[8] J. Buchanan and A. J. Grant. Investigating and Prosecuting Nigerian Fraud. *United States Attorney's Bulletin*, 49(6):39–47, 2001.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[10] N. Christin, S. S. Yanagihara, and K. Kamataki. Dissecting one click frauds. *Carnegie Mellon University Technical Report CMU-CyLab-10-011*, 2010. https://www.andrew.cmu.edu/user/nicolasc/publications/TR-CMU-CyLab-10-011.pdf.

[11] N. C. Commission. Sim Registration. http://www.ncc.gov.ng/index.php?option=com_content&view=article&id=122&Itemid=113, 2011. Accessed: 2013-07-24.

[12] N. C. Commission. 2013 Year End Subscriber/Network Data Report for Telecommunications Operating Companies in Nigeria. http://www.ncc.gov.ng/index.php?option=com_docman&task=doc_download&gid=563&Itemid=, 2013. Accessed: 2015-11-17.

[13] A. Costin, J. Isachenkova, M. Balduzzi, A. Francillon, and D. Balzarotti. The Role of Phone Numbers in Understanding Cyber-Crime Schemes. In *Annual Conference on Privacy, Security, and Trust (PST)*, PST 13, July 2013.

[14] Custominsight.com. Random Samples and Statistical Accuracy. http://www.custominsight.com/articles/random-sampling.asp, 2014. Accessed: 2014-03-03.

[15] Detica. The Cost of Cybercrime. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60942/THE-COST-OF-CYBER-CRIME-SUMMARY-FINAL.pdf, 2011. Accessed: 2013-03-03.

[16] M. A. Dyrud. I brought you a good news: An analysis of Nigerian 419 letters. In *Proceedings of the 2005 Association for Business Communication Annual Convention*, 2005.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD-96, 1996.

[18] M. C. for Strategic and I. Studies. Net Losses: Estimating the Cost of Cybercrime - Economic Impact of Cybercrime ii. http://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime2.pdf, 2014. Accessed: 2014-07-23.

[19] Forbes. Africa's 50 Richest. http://www.forbes.com/profile/aliko-dangote/, 2014. Accessed: 2013-07-24.

[20] Y. Gao and G. Zhao. Knowledge-based Information Extraction: a Case Study of Recognizing Emails of Nigerian Frauds. In *Natural Language Processing and Information Systems*, pages 161–172. Springer, 2005.

[21] P. Grabosky. The evolution of cybercrime, 2004-2014. *RegNet Working Paper, No. 58, Regulatory Institutions Network*, 2014.

[22] L. Gu. The mobile cybercriminal underground market in china. *Trend Micro Research Paper - Cybercriminal Underground Economy Series*, 2014. http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-the-mobile-cybercriminal-underground-market-in-china.pdf.

[23] J. Herhalt. Cyber crime - a growing challenge for governments. *KPMG INTERNATIONAL: Issues Monitor - Government on Cyber Crime*, 8, 2011.

[24] C. Herley. Why do Nigerian Scammers Say They are from Nigeria? http://research.microsoft.com/pubs/167719/whyfromnigeria.pdf, 2012. Accessed: 2014-07-24.

[25] T. Holz, M. Engelberth, and F. Freiling. Learning more about the underground economy: A case-study of keyloggers and dropzones. In *14th European Symposium on Research in Computer Security*, ESORICS 2009, September 2009.

[26] A. Huang. Similarity Measures for Text Document Clustering. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.4480&rep=rep1&type=pdf, 2008. Accessed: 2016-05-03.

[27] J. Huang, G. Stringhini, and P. Yong. Quit playing games with my heart; understanding online dating scams. In *DIMVA 2015*, July 2015.

[28] J. Isacenkova, O. Thonnard, A. Costin, D. Balzarotti, and A. Francillon. Inside the Scam Jungle: A Closer Look at 419 Scam Email Operations. In *Proceedings of the International Workshop on Cyber Crime (co-located with S&P)*, IWCC 13. IEEE, May 2013.

[29] J. Isacenkova, O. Thonnard, A. Costin, A. Francillon, and D. Balzarotti. Inside the Scam Jungle: a Closer Look at 419 Scam Email Operations. http://jis.eurasipjournals.com/content/pdf/1687-417X-2014-4.pdf, 2014. Accessed: 2014-07-24.

[30] Z. Jianwei, T. Holz, C. Song, J. Guo, X. Han, and W. Zou. Studying malicious websites and the underground economy on the chinese web. In *In Proceedings of the 7th Workshop on the Economics of Information Security (WEIS'08)*, WEIS'08, June 2008.

[31] Z. Jianwei, G. Liang, and D. Haixin. Investigating China's Online Underground Economy. In *Conference on the Political Economy of Information Security in China*, April 2012.

[32] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf, 1998. Accessed: 2013-01-24.

[33] KDD. 2014 SIGKDD Test of Time Award. http://www.kdd.org/News/view/2014-sigkdd-test-of-time-award, 2014. Accessed: 2016-05-03.

[34] R. V. Krejcie and D. W. Morgan. Determining Sample Size for Research Activities. *Educational and Psychological Measurement*, 30:607–610, 1970.

[35] A. A. Kumar and S. Chandrasekhar. Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering. *International Journal of Engineering Research and Technology (IJERT)*, 1, 2012.

[36] F. Mercês. The brazilian underground market: The market for cybercriminal wannabes? *Trend Micro Research Paper - Cybercriminal Underground Economy Series*, 2014. http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-the-brazilian-underground-market.pdf.

[37] N. B. of Statistics. 2011 Annual Socio-Economic Report. http://nigerianstat.gov.ng/pages/download/38, 2011. Accessed: 2013-07-24.

[38] F. Pan, W. Wang, A. K. H. Tung, and J. Yang. Finding Representative Set from Massive Data. In *Fifth IEEE International Conference on Data Mining*, ICDM'05. IEEE, November 2005.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[40] N. Shah and S. Mahajan. Document Clustering: A Detailed Review. *International Journal of Applied Information Systems (IJAIS)*, 4, 2012.

[41] A. Smith. Nigerian scam e-mails and the charms of capital. *Cultural Studies*, 3(2):27–47, 2009.

[42] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna. Your botnet is my botnet: Analysis of a botnet takeover. In *16th ACM Conference on Computer and Communications Security*, CCS 2009, November 2009.

[43] O. Tade and I. Aliyu. Social Organization of Internet Fraud among University Undergraduates in Nigeria. *International Journal of Cyber Criminology*, 5(2):860–875, July-December 2011.

[44] C. Tive. 419 Scam: Exploits of the Nigerian Con Man. iUniverse, 2002.

[45] Y. Yang and X. Liu. A re-examination of text categorization methods. http://www2.hawaii.edu/~chin/702/sigir99.pdf, 1999. Accessed: 2013-01-24.

[46] M. Yip. An investigation into chinese cybercrime and the applicability of social network analysis. In *In ACM WebSci '11*, June 2011.