Constructing and Analyzing Criminal Networks

Hamed Sarvari, Ehab Abozinadah, Alex Mbaziira, Damon McCoy
George Mason University

***Abstract*—Analysis of criminal social graph structures can enable us to gain valuable insights into how these communities are organized. Such as, how large scale and centralized these criminal communities are currently? While these types of analysis have been completed in the past, we wanted to explore how to construct a large scale social graph from a smaller set of leaked data that included only the criminal's email addresses.**

We begin our analysis by constructing a 43 thousand node social graph from one thousand publicly leaked criminals' email addresses. This is done by locating Facebook profiles that are linked to these same email addresses and scraping the public social graph from these profiles. We then perform a large scale analysis of this social graph to identify profiles of high rank criminals, criminal organizations and large scale communities of criminals. Finally, we perform a manual analysis of these profiles that results in the identification of many criminally focused public groups on Facebook. This analysis demonstrates the amount of information that can be gathered by using limited data leaks.

I. Introduction

Previous work has shown that cybercriminals often do not work alone and that they often form underground online communities, such as forums, that enable them to communicate with fellow criminals to engage in information sharing and criminal to criminal commerce [17]. Gaining a deeper understanding of how cybercriminals operate and communicate can enable defenders to craft more effective interventions to disrupt their illicit activities [12], [14]. Our perception of cybercriminals is also furthered by using the leaked data that can be analyzed to understand the economics of different criminal operations [15], [24].

In spite of our increasing understanding of the business models of cybercrime, we still do not have a firm grasp on the social structure of the cybercriminal ecosystem. Underground forums primarily include inferred links between actors rather than explicit social links.

In this paper, we focus on a publicly leaked data set from a data theft service that contained slightly over one thousand email addresses of predominantly Nigerian advanced fee fraud scammers [4]. We present a method based on the facts that profiles on Facebook are often searchable by email addresses and cybercriminals sometimes reuse a single email address for multiple services. In this case, criminals reused their email address when registering for an account at the data theft service and their Facebook profile. By using this technique, we are able to link 262 profiles and scrape their friends lists to build up a large scale social graph of over 40 thousand profiles.

Our analysis of this social graph and criminal profiles enables us to address questions, such as "Who are key people within this Nigerian scammer community?", "Are there smaller more connected groups within the profiles?", "What is the potential scale of this scammer community?" Our analysis of the social graph is able to at least partially answer these questions for the scammer community that we have identified. For instance, we identified the top-10 most central profiles by using graph centrality measures. Also, we can find tightly connected groups of profiles that might be working together. In addition, we can evaluate if PageRank is able to efficiently isolate additional scammer profiles out of the set of friends of these scammers. Finally, our manual analysis of these scammers' and friends' profiles led us to discover public Facebook groups focused on criminal activity.

Our key contributions include:

1. Constructing a large scale social graph from a leaked set of email addresses. Our technique of linking other profiles from social networking sites to an actor can significantly amplify the amount of data and analysis that can be conducted on a limited data set that includes email addresses.

2. Evaluated techniques to identify additional criminal profiles from social graph. We manually evaluate the effectiveness of the PageRank algorithm for identifying other potential criminal profiles and find that it is more effective at isolating criminal profiles than picking random profiles from our data set.

3. A manual analysis of criminal profiles on Facebook. Our manual analysis of criminal Facebook profiles reveals many interesting findings, such as public groups that are focused on criminal activities. It also allows us to gain additional insights into the culture and methods used by this criminal community.

II. Related work

In this section, we first present an overview of previous work in the domain of social networks in general that is by no means complete. We then present a more comprehensive set of previous studies focused on criminal social networks.

A. Social Networks

Mislove et al. [16] analyzed several large scale online social networks with the goal of improving the design of online social networks. Kumar et al. [9] performed an analysis of Flicker photo sharing and Yahoo 360 social networks and focused on understanding how they change over time. Kwak et al. [10] gathered data by crawling Twitter and did a comparison among different ranking criteria along with an analysis of the impact of retweets in this network. More recently, Ugander et al. [26] did a complete analysis of Facebook's social graph and computed different features of that graph, such as quantifying

DOI 10.1109/SPW.2014.22

IEEE computer society

the "your friends have more friends than you" phenomenon. Our work uses similar methods, but focus on a smaller subset of the network that we believe is densely populated with cybercriminals. We also concentrate on pursuing the goal of identifying key members of criminal organizations.

B. Criminal Social Networks

There is a large body of previous work on using social graph analysis to gain a better understanding of criminal social networks structure and to identify key members of criminal groups. Xu and Chen [30] introduced a framework called CrimeNet for automated network analysis and visualization and claim they can identify central members and interaction patterns between groups significantly faster. Xu et al. [29] proposed a link analysis technique that uses shortest-path algorithms to identify the strongest association paths between entities in a criminal network. Qin et al. [22] applied Web structural mining techniques with the goal of terrorist network analysis. Harper and Harris [6] used link analysis in an experiment involving 29 law enforcement analysts to portray relationships of a criminal organization. Sparrow [23] used network analytic techniques to analyze criminal networks with focused on identification of vulnerabilities in criminal organizations. Kerbs [8] uses public data to analyze the tragic events of September 2001. Xu and Chen [28] analyzed several networks consisting of criminals based on the crime incident data provided by the Tucson Police Department. Lu et al.[13] also used social graph analysis methods to analyze a hacker community based on textual data obtained from newspapers, court proceedings and trial transcripts.

In this set of studies on criminal social networks, they either used news stories or information from media or data collected through empirical studies, which results in small scale networks with inconsistencies and inaccuracies. They had to use this data to speculate about relationships among members of that criminal community.

Another set of studies have performed larger scale analysis of criminal networks. Yang et al. [31] did an analysis of criminal community on Twitter. They used twitter profiles which posted malicious URLs identified by Google safe browsing as their initial community and introduced an algorithm similar to PageRank in order to infer criminal accounts. Stringhini et al. [25] worked on detecting spammers in different social networks by using honey-profiles.

In our research, we have started with a set of emails associated with criminals. We then constructed a large social graph from this limited set of emails by linking these to public social network profiles, Facebook, in order to scrape friend's list of these criminals. This enabled us to find relationships among the members of the criminal community. We have also employed community detection technique based on the modularity in order to discover communities inside the criminal network which to our knowledge has not been done before in this area. Finally, we perform a manual analysis of these profiles to provide some evidence of criminal activity and supporting our ranking of the members.

TABLE I
DATA SET DESCRIPTION

Emails	Profiles found	Public profiles	Private profiles	Total friends' URLs scraped
1036	262	183	79	43125

III. DATA SET

Our analysis is based on a publicly leaked set of 1036 customer email addresses from BestRecovery, which is an online data theft service that was primarily used by Nigerian cybercriminals that focused on advanced fee fraud (more commonly referred to as 419 Scams) and online dating scams [3]. We searched Facebook for profiles linked to these email addresses and found 262 profiles of which 183 made their friends lists public. We then scrapped these 43,125 friends profiles for their social links that also partially compensates for the 79 scammers that did not make their friends lists public [2]. A description of the data set is summarized in Table I.

We were unable to scrape friends of friends of the actors due to the large number of profiles and limitations of our scraping abilities. Therefore, the "friend" nodes have degree one, unless they are friends of multiple actors. In this case their degree is greater than 1 and sometimes up to 20 (we have friends in common with up to 20 actors). This characteristic of our graph affects all of the measurements calculated throughout this paper in the sense that "friend" nodes do not rank high in centrality measures (see section V-A). We provide a separate ranking and analysis for friends in section VIII.

We limited all data collection to publicly available information and throughout the paper we only refer to profiles by their first 5 characters in order to protect their privacy. The data we collected from 183 public Facebook profiles was comprised of: actors' IDs, actor's names, actor's Facebook URLs, number of friends, URLs of friend's Facebook profiles and URLs of Facebook groups that each actor joined. During the data collection stage, we did not engaged the actors or their friends by sending friend requests or communicate with them through direct messages.

IV. NIGERIAN SCAMMERS SOCIAL NETWORK

In this section we analyze and interpret the data by creating a social graph in which nodes are the Nigerian criminals and their friends and edges are their Facebook relationship. Two nodes are adjacent if they are friends on Facebook.

Visualization:

The method used for visualizing the graph is Force Atlas 2. Force Atlas 2 [7] is a visualization algorithm which tries to produce a layout that gives the best interpretation of the data. It simulates a physical system in which nodes repulse each other and edges attract nodes they connect.

Having scraped friends list of 262 actors, the whole graph consists of more than 43 thousand nodes. Since It would be visually difficult to interpret this huge graph, we pruned the

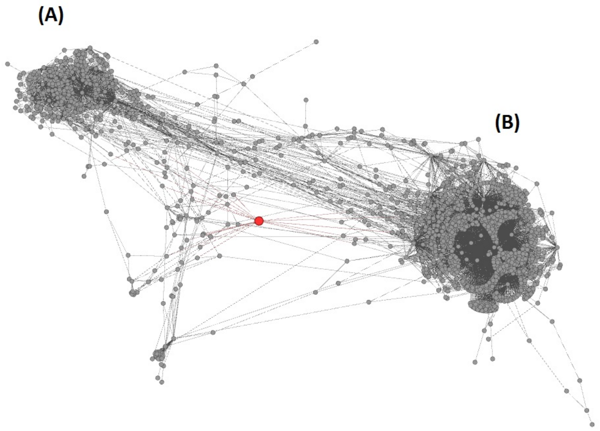

Fig. 1. Graph of Nigerians, discarding the friend nodes connected only to one actor. Two densely connected components of the graph are labeled as A and B

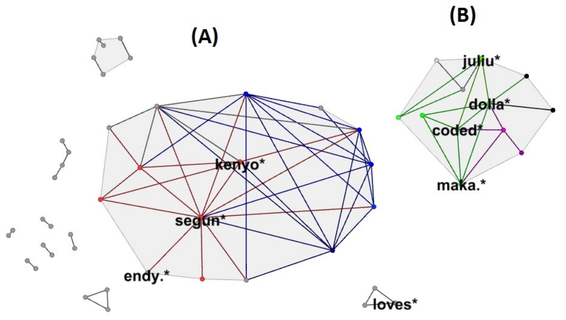

Fig. 2. Graph of connections among Nigerian actors. Actors without any connection to other actors are discarded. Nodes among top-10 actors in centrality measures (introduced in section V and Table II) are labeled by their names. Nodes within the same community (introduced in section VII) are colored the same. Components labeled A and B are subsets of two densely connected components with the same labels in Figure 1.

graph by removing friends who were connected to only one criminal actor in the graph. The result is the graph of the Nigerian community that only includes friends connected to two or more actors, which has 1740 nodes and is depicted in Figure 1.

An important aspect of this network is to see the interactions among the main actors whose email addresses were used as the starting point of building the network. Connections among the original actors is graphed in Figure 2. This graph has 53 nodes meaning that out of 262 actors, 53 have direct connection with each other.

Looking at the graph of actors' interactions in Figure 2, we can see two densely connected components (labeled as A and B) and several other components. The same two dense subsets can be seen in Figure 1 with sparse connections caused by mutual friends of the actors. This emphasizes the fact that although actors in the two subsets are not directly connected to each other they do have mutual friends which produce connections among these two subsets.

V. Social Network Analysis

In analyzing the network of Nigerians, it would be very interesting to be able to find out:

- which criminal has a central position in the graph.
- which subgroups and communities can be found in the network.
- which criminals are acting as brokers of collaboration and information in the network.
- what is the ranking of criminals based on their importance and influence on the network.

The answers to the above questions lies in calculating graph's centrality measures and using community detection techniques.

A. Centrality Measures

The most commonly used centrality measures are degree, betweenness and closeness, which were first introduced by Freeman [5]. We have also considered the eigenvector centrality and PageRank of the nodes, which can help us gain a better understanding of a node's centrality.

Degree Centrality:

The first measure of distinguishing an important node is number of its neighbors. It is believed that the node that has the most neighbors has the most activity and influence in it's local neighborhood and hence is a key member.

Degree Centrality is defined as:

$$D_i = \frac{k_i}{N-1} = \frac{\sum\limits_{j \in G} a_{ij}}{N-1} \tag{1}$$

where k_i is degree of the node, a_{ij} is the ij^{th} element of the adjacency matrix and $N-1$ is the normalization factor (N is the number of nodes of the graph). Therefore D_i will be independent of network size and $0 \leq D_i \leq 1$

Betweenness Centrality:

Betweenness of a node is the number of shortest paths in graph which passes through that node. A node with high betweenness has a key role in flowing information. It usually connects two densely connected parts of the graph so acts as the broker of messages between those communities. Removal of such node can lead to major shortcomings in message passing and communications in the network. Betweenness Centrality is defined as:

$$B_i = \frac{\sum\limits_{j<k \in G} n_{jk}(i)/n_{jk}}{(N-1)(N-2)} \tag{2}$$

Where n_{jk} is the number of shortest paths between j and k and $n_{jk}(i)$ the number of such paths which pass through node i. $(N-1)(N-2)$ is the normalization factor.

Closeness Centrality:

Distance(farness) of a node from other nodes in a graph is defined as the sum of shortest paths between that node and all other nodes in the graph. A node with low distance with other nodes can reach other nodes easier and faster [27].

Closeness of a point is defined as:

$$C_i = (L_i)^{-1} = \frac{N-1}{\sum\limits_{j \in G} d_{ij}} \tag{3}$$

Where d_{ij} is the distance between nodes i and j. L_i is the normalized distance of a node from other nodes in a graph.

Eigenvector Centrality:

Determines to what extent a node is connected to other well-connected nodes.

It is defined as:

$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j \in G} a_{ij} x_j \tag{4}$$

where $M(i)$ is the set of neighbors of i and λ is a constant. and a_{ij} is the ij^{th} element of the adjacency matrix.

B. PageRank:

PageRank [20] can also be used as a ranking among nodes of a graph, giving us a chance to compare relative "importance" of the nodes. The reason behind choosing PageRank is that there is a clear similarity between web pages and links among them and social networks. PageRank is designed to produce a global "importance" for web pages and we are trying to find an overall importance of the criminal actors based on their graph position. PageRank is introduced in this section so that we can make a comparison between centrality measures and PageRank.

PageRank is defined recursively as:

$$PR(A) = (1-d) + d * \sum_{B \in M(A)} \frac{PA(B)}{L(B)} \tag{5}$$

where M(A) are the nodes neighboring A, L(B) is the number of outgoing links in B and d is the damping factor.

Analysis Results and Interpretations:

Node centrality measures and PageRank values are calculated for the graph of the Nigerians using equations (1) to (5). Top 10 actors in each of the categories is summarized in Table II. Taking a closer look at Table II, we can see that 9 out of 10 in each category are the same with slight difference in ranking. Coded* comes first in every category.

The fact that in our graph, centrality measures are correlated and they have the same top actors shows that in this criminal network, highly connected members (high Degree) are located in central position of the graph connecting dense communities, where a big portion of shortest paths pass through them (high Betweenness) and they are also connected to other well-connected members (high Eigenvector centrality). They are also ranked top in the PageRank meaning that they are relatively influential and important nodes of the graph.

TABLE II
TOP10 ACTORS IN DIFFERENT CENTRALITY MEASURES AND PAGERANK

Betweenness	Degree	Eigenvector	PageRank
coded*	coded*	coded*	coded*
kenyo*	dolla*	dolla*	loves*
loves*	loves*	juliu*	dolla*
dolla*	juliu*	maka.*	kenyo*
nkemd*	maka.*	loves*	juliu*
adefo*	kenyo*	kenyo*	maka.*
juliu*	devoe*	segun*	nkemd*
maka.*	segun*	profi*	devoe*
devoe*	nkemd*	devoe*	endy.*
segun*	endy.*	nkemd*	segun*

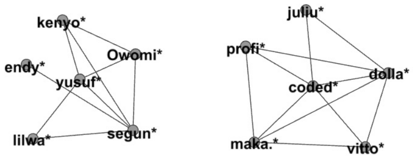

Fig. 3. Interactions of top actors

Actors ranked top 10 in different measures are labeled in Figure 2. The reason why not all of the top 10 actors in different measures can be found in Figure 2 is that this figure contains only actors that have direct relationship with other actors and those actors were not directly connected to another actors.

Looking at Figure 2 we can see that most of the top actors are located in the two densely connected components of the graph (labeled by A and B) while the actor "loves*" is not inside these two subsets. It is interesting to track the position of this node in Figure 1 where a portion of "friend" nodes are also added. "loves*" has a bigger size than other nodes and is also color coded with red in Figure 1. You can see that this node has an important role in connecting parts A and B and is connected to a big number of high connected nodes in these two subsets. That's why this node also ranks high in centrality measures.

C. Top actors interactions

A main topic to investigate in a criminal network would be analyzing patterns of interactions among top actors. The top 20 actors in the PageRank were taken and the induced graph on these people is represented in Figure 3. The graph has 12 nodes meaning that 12 out of the top 20 have direct connection with each other. It is interesting to see that top actors are also densely connected to each other and form the same two disjoint components.

TABLE III
CLIQUE FINDING RESULTS

communities	cliques	vertices	edges
7-clique community	18	17	88
6-clique community	57	42	227
5-clique community 1	180	108	544
5-clique community 2	228	238	996

VI. CLIQUES

An important aspect of each graph is finding its highly connected subunits. A k-clique is a complete subset of size k of a graph. A k-clique community is defined as union of all k-cliques that can be reached from each other through a series of adjacent k-cliques (adjacent k-cliques are k-cliques that share k-1 nodes). The intuition behind importance of k-clique communities is that, it is a subset of graph in which nodes can communicate through series of nodes that are all members of well connected subsets of graph. k-cliques and k-clique communities are extracted for Nigerian network using CFinder [21].

Number of k-clique and k-clique communities for $k = 5, 6, 7$ are reported in Table III. There are actually a large number of 3-cliques and 4-cliques in this graph which are less important. The large number of 7-cliques and 6-cliques found, shows strong inner connection in the network. The fact that we have two 5-clique communities shows that 5-cliques can be found on both densely connected parts of the graph (A and B) while having only one community of 6-cliques and one community of 7-cliques shows that they can be found only in part B (see Figure 1).

Clustering Coefficient:

Clustering coefficient of a vertex of a graph is the probability that any two randomly chosen neighbors (friends) of that vertex are connected (have a friendship) themselves. It is computed by dividing the number of triangles that contain that vertex [11] by the number of possible edges between its neighbors. The clustering coefficient of a graph is calculated as the average of the clustering coefficients of each of its nodes. A higher clustering coefficient indicates a greater “cliquishness”. The clustering coefficient for the Nigerian criminals network is 0.657 with total number of 8793 triangles. High clustering coefficient in this network is correlated with large number of cliques found in the graph and both state high connectivity in the network.

VII. COMMUNITY DETECTION

The next important topic in analyzing each network is finding patterns and substructures of that network. In our analysis, subsets of the criminal network could be interpreted as smaller groups collaborating and involved in the same malicious activity. One commonly used community detection method is clique communities which was analyzed in section VI.

In this section we provide a community detection technique based on the modularity concept. In this method each community is assigned a modularity value. The modularity of a partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities [19]. The algorithm finds partitions of a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected. We found communities within the Nigerian network using a heuristic method based on modularity optimization introduced by Blondel et al. [1].

TABLE IV
2 MAIN COMMUNITIES EXTRACTED FROM THE NETWORK

community	contains actors	contains top-10	no. of members
community 2	9	4	8962
community 1	8	2	3144

The network of 43 thousand nodes including the main actors and their friends is divided into 108 communities. One way of determining which communities are of the highest importance to us, is to check how many main actors reside in each community and if these actors are central as ranked by PageRank. Also, which communities have the biggest overlap with the two main components of the graph where most of the top actors are placed. For example, as summarized in Table IV communities 1 and 2 have the highest number of actors and also 6 out of top 10 central actors reside in these two communities. Therefore, from an investigative point of view, these two communities might be the top priority to pursue, since they are potentially larger groups of criminals involved in the same activity.

Community detection gives us the insight that the two disjoint components of the graph (see Figure 1 and Figure 2) which visually seemed to be densely connected subsets of the graph are themselves divided into substructures. Each of these substructures has higher modularity and is more densely connected. Overlap of those communities with the graph of main actors' interactions is color coded in the Figure 2. Red nodes are members of community 1 which includes segun* and kenyo*, two of the top-10 actors. Green nodes in the other component are members of community 2 which includes juio* ,dolla* ,coded* and maka.*, four of the top-10 central actors. Other nodes with the same color are also members of the same communities. All other communities can be ranked based on their importance which is defined above.

VIII. MANUAL ANALYSIS AND DETECTING CRIMINAL FRIENDS

Recall that the main graph of Nigerian's community is built by scraping the friend's list of the 262 criminals' profiles. Looking at the social graph we find some friends that have social links to many of the criminals' profiles and it is reasonable to suspect that some are involved in similar criminal activities. From an investigative perspective, it would be useful to efficiently locate additional criminal profiles from among friends of the criminals. A previously proposed method of accomplishing this task is to use the PageRanking algorithm to

TABLE V
CRITERIA FOR ASSESSING PROFILES

Criteria	Description
Groups	Members of groups focused on hacking and scamming activity
Likes	Events or activities related to hacking, carding and scamming
Pictures	Self portraits of the actors showing off large amounts of money, fake identification cards, multiple phones
Posts	Informational posts about scamming techniques or bragging about exploited victims

TABLE VI
POSSIBLE OUTCOMES OF MANUAL ANALYSIS

Category	Criteria
Probable Scammer	• Membership in scamming groups • Comments about underground activity • Possession of more than one smart phone, holding hundreds of dollars • Pictures of online bank account or credit cards
Scammer Community Member	• Involvement in social groups affiliated with the scammer community • Pictures of guns, money, drug • Displaying “misplaced wealth”
Unclear	• No signs of criminal activity found in their profile

rank profiles that are more connected to our original criminal actors.

In this section we evaluate how well PageRank performs at identifying potential criminals by performing a case study based on manual analysis of profiles. In these profiles, many people choose to enable a high level of privacy settings or are cautious to not leave any overt signs of criminal activity on their Facebook profiles. Thus, we do not have any strong ground truth information for which profiles are linked to criminals. However, we manually analyzed some of the accounts and found that at least some of the profiles do contain signs of criminal activity that are publicly visible.

A. Methodology

Table V shows the criteria we used to assess which profiles were connected to scammers. The first criteria was to see if the profile has joined any groups that are focused on hacking, scamming or any other criminal activity. Our second criteria was to check if the profile has any likes for events or activities related to hacking, carding and scamming. We also looked for self portraits that include fake identification cards, cash, and other signs of “misplaced wealth,” along with posts in which they talk about scamming or exploiting victims.

Our first data set was composed of the top 100 direct actors’ profiles ranked using the PageRank algorithm. This data set is useful for understanding how many of these profiles exhibited our criteria and was useful for developing the initial set of criteria. Next, we manually analyzed the top 100 friends of the actors ranked by using the PageRank algorithm. This is an interesting data set that enables us to expand the set of criminal profiles from the social graph produced by the direct actor’s profiles. Finally, we selected 100 profiles at random from the complete set of actors and their friends profiles. This set gives us a good baseline to compare how much PageRank improves the density of criminal profiles.

For each Facebook profile, we manually analyzed its news feeds, Likes, Groups and Photos for evidence indicating criminal activity. Within the timelines we checked for posts indicating scamming or bragging about exploited victims. These criminals use slang words like ”Mugun”, ”Maga”, ”Magan don pay” which means fool for victims who have been scammed. We also analyzed Facebook Likes for pages, events or activities on computer hacking, carding and scamming. In pictorial evidence, we analyzed photos for self portraits of the actors showing off money, drugs like marijuana, luxury cars parked in ghettos since this group is a youthful slum dwelling population, ostentatious jewelry and watches, expensive scotch whiskeys, wines and champagnes, guns, possession of at least 2 smart phones and excessive partying. We examined Facebook groups which actors were subscribed to, for evidence of scamming activity. All the evidence we found was annotated against each profile for purposes of classifying each underground actor to level of exhibited criminal activity.

There are three possible outcomes of our manual analysis: *probable scammer*, *scammer community member* and *unclear*. *Probable scammer* indicated that the profile exhibits signs that they are actively engaged in scamming activity. We identify this category based on one of the following evidence: pictures with the profile in possession of more than one smart phone, holding hundreds of dollars, pictures of online bank account or credit cards, active member in one of scamming groups, a member of more than two scamming groups, or having comments about underground activity either on timeline or group messages. *Scammer community member* indicates that the profile shows signs of being involved in social groups affiliated with the scammer community. Pictures of guns, money, drug, or living in ghetto area while displaying “misplaced wealth” such as upscale partying and possessing luxury cars and clothing. *Unclear* indicated that the profile exhibits little public information or no signs were found in their profile of criminal activity. These profiles can’t be classified clearly, because part of the profiles are not public, which make it hard to observe enough information to categorize the profile to be one of the above categories, or the profile contains regular activates. For example, having a suspicious friend while we cannot see the member pictures or groups, or posting a suspicious picture without the profile owner in the picture. These types of photos could have been reposted from the Internet or from other friend’s profile. The criteria for categorizing profiles into these three categories is summarized in Table VI.

TABLE VII
MANUAL ANALYSIS RESULTS.

Data Set	Probable scammer	Scammer community member	Unclear
Top 100 actors	12	14	74
Top 100 friends	8	19	73
Random 100 profiles	5	15	80

B. Results

Our manual analysis of the top 100 friends' profiles ranked show that 8% of them are probable scammers and 19% are scammer community members. In order to validate how effective the PageRank algorithm is at isolating suspicious friends profiles we also pick a random sample of size 100 from the whole community of Nigerian actors and friends composed of more than 43 thousand members and do the same manual analysis for them. The result show a 5% signs of being a probable scammer and 15% signs of being a scammer community member. The results are summarized in Table VII. Finally, for comparison we performed an analysis of the top 100 direct actors ranked using the PageRanking algorithm. As you can see the top 100 friends set is more densely populated with probable scammer profiles, but is less densely populated than the set of top 100 direct actors. From this analysis, we can see that PageRank is indeed at least slightly effective for the task of isolating additional criminal profiles from the set of friends profiles.

IX. DISCUSSION

A. Lessons learned about scammers

By linking this leaked set of email addresses to Facebook profiles we were able to create a social graph. Our analysis of this social graph and our manual analysis of the profiles improves our understanding of how these scammer communities are connected and their interactions on Facebook.

Communities. We find that these scammers do communicate with one another and they tend to form smaller groups of tightly connected scammers that might be working together. We also find that these tightly connected groups are also loosely connected with other groups of scammers. This indicate that these scammers might be forming organizations and that there are potentially "leaders" of each organizations. These leaders might be effective targets to pursue for legal interventions. Our analysis can separate these groups and help better target important individuals.

Criminal Facebook Groups. From our manual investigation of these profiles we find a number of public criminally focused Facebook groups with names, such as "MoolahGroup Nigeria" and "Unscrupulous Buccaneer." These groups would be an interesting place to locate additional potentially criminal profiles and the communications within these groups shed light on the techniques used by these scammers.

B. Methods to evade our analysis

A simple method of evading our analysis is to use two separate email account for scamming activities and another from personal social networking accounts. Another method would be to tightly lock down the privacy settings of their profiles. While our analysis is easily evaded, we find a large number of scammers that currently leak a large amount of information. This information should be collected before these analysis techniques become more familiar to scammers.

X. CONCLUSIONS AND FUTURE WORK

Future Work. Our research focused on Facebook, but there are many other online social networking sites, such as Twitter and Google+ that allow look up of profiles by email address. As future work we plan on linking the social graph from Facebook with those of profiles on these other services to build an even more complete graph of this scammer community.

In addition, our manual analysis revealed the existence of public Facebook groups that are focused on criminal activities. These groups can also be an interesting topic of future studies. We hope that our methods can be further scaled to other social networking sites and larger sets of leaked criminal email addresses to give us a broader understanding of how criminals organize online.

Conclusions. In this paper, we demonstrate the magnitude of social graph information that can be collected from a small set of criminal email addresses. We collect this information by linking these email addresses with profiles from, Facebook, an online social networking site. Our analysis of the resulting large scale social graph shows that these scammers are organized into tightly connected groups of scammers along with larger communities of loosely connected scammers. By using graph analysis techniques we can identify key members of these criminal communities that might be targeted to disrupt these communities.

Our study shows that key members of this criminal network, have high ranks in all centrality measures and also in PageRank. In other words, we can see that highly connected members are located in central position of the graph and they are also connected to other well-connected members. This feature can be validated in other studies about criminal networks.

ACKNOWLEDGMENT

This work was supported by National Science Foundation grant NSF-1237076, and a Google Research Award. Ehab Abozinadah would like to thank King Abdulaziz University (KAU) - Saudi Arabia Culture Mission (SACM) for funding his PhD scholarship.

REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] J. Bonneau, J. Anderson, R. Anderson, and F. Stajano. Eight friends are enough: Social graph approximation via public listings. In *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, SNS '09, pages 13–18. ACM, 2009.

[3] Brain Krebs. Spy service exposes nigerian yahoo boys. http://krebsonsecurity.com/2013/09/spy-service-exposes-nigerian-yahoo-boys/.

[4] Brain Krebs. "Yahoo Boys Have" 419 Facebook Friends. http://krebsonsecurity.com/2013/09/yahoo-boys-have-419-facebook-friends/.

[5] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[6] W. R. Harper and D. H. Harris. The application of link analysis to police intelligence. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 17(2):157–164, 1975.

[7] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian. Forceatlas2, a graph layout algorithm for handy network visualization. *Paris http://www. medialab. sciences-po. fr/fr/publications-fr*, 2011.

[8] V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[9] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Link Mining: Models, Algorithms, and Applications*, pages 337–357. Springer, 2010.

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[11] M. Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, 407(1):458–473, 2008.

[12] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium and Security and Privacy*, pages 431–446, Oakland, CA, May 2011.

[13] Y. Lu, M. Polgar, X. Luo, and Y. Cao. Social network analysis of a criminal hacker community. *Journal of Computer Information Systems*, 51(2):31–41, 2010.

[14] D. McCoy, H. Dharmdasani, C. Kreibich, G. M. Voelker, and S. Savage. Priceless: The Role of Payments in Abuse-advertised Goods. In *Proceedings of the ACM Conference on Computer and Communications Security*, Raleigh, NC, Oct. 2012.

[15] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. In *Proceedings of the 21st USENIX conference on Security symposium*, pages 1–1. USENIX Association, 2012.

[16] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[17] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. An Analysis of Underground Forums. In *Proceedings of the ACM Internet Measurement Conference*, Berlin, CA, Nov. 2011.

[18] M. E. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[19] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[21] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[22] J. Qin, J. J. Xu, D. Hu, M. Sageman, and H. Chen. Analyzing terrorist networks: A case study of the global salafi jihad network. In *Intelligence and security informatics*, pages 287–304. Springer, 2005.

[23] M. K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, 13(3):251–274, 1991.

[24] B. Stone-Gross, R. Abman, R. Kemmerer, C. Kruegel, D. Steigerwald, and G. Vigna. The Underground Economy of Fake Antivirus Software. In *Proceedings of the Workshop on Economics of Information Security (WEIS)*, 2011.

[25] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9. ACM, 2010.

[26] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[27] S. Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[28] J. Xu and H. Chen. Criminal network analysis and visualization. *Communications of the ACM*, 48(6):100–107, 2005.

[29] J. J. Xu and H. Chen. Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. *Decision Support Systems*, 38(3):473–487, 2004.

[30] J. J. Xu and H. Chen. Crimenet explorer: a framework for criminal network knowledge discovery. *ACM Transactions on Information Systems (TOIS)*, 23(2):201–226, 2005.

[31] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80. ACM, 2012.