International Journal of Cyber Criminology
Vol 10 Issue 2 July – December 2016

Cyber Crime Nation Typologies: K-Means Clustering of Countries Based on Cyber Crime Rates

Alex Kigerl[1]
Washington State University, United States of America

Abstract

Cyber crime is a worldwide problem, with a global reach. Cyber crimes do not respect national boundaries, and so can be sent to and from anywhere in the world. Many reports, mostly by cyber security firms, regularly release information ranking the different nations in terms of top cyber crime output, broken down into varying cyber crime types. However, little has been done to classify nations according to separate cyber crime typologies using any multivariate methods. Instead, reporting is descriptive and unidimensional. The present research sought to fill this gap by conducting K-means clustering analysis on a sample of 190 countries using seven dimensions of cyber crime ranging from malware, fraud, spam, and digital piracy, as well as measures of GDP and internet use. The findings determined that nations can be broken down into four distinct categories based on cyber crime activity: low cyber crime countries, non-serious cyber crime countries, advance fee fraud countries, and phishing scam countries. The implications of these findings and the directions for future research are discussed.

Keywords: Cyber crime, malware, Spam, Digital Piracy, Countries, Typologies.

Introduction

A disproportionate amount of the cyber crime attacks worldwide are accounted for by a minority of countries (Spamhaus, 2016). Many reports, mostly by cyber security firms, regularly conduct rankings of nations based on being top sources of cyber crime activity and attacks[2]. Nations are differentiated based on the particular type of cyber crime or offense that they are associated with (APWG, 2014; Ultrascan, 2014; V.I. Labs, 2014).

However, most of the reporting is univariate and descriptive, offering a simple unidimensional listing of nations by rank via one metric at a time. The reports also offer

[1] Department of Criminal Justice and Criminology, Washington State University, Spokane, WA 99210, United States of America. Email: alex@kigerl.com

[2] Some examples of the online reports include: https://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf, http://www.vilabs.com/news-section/code-confidential/top-20-countries-software-piracy-2014, and http://www.ultrascan-agi.com/public_html/html/pdf_files/Pre-Release-419_Advance_Fee_Fraud_Statistics_2013-July-10-2014-NOT-FINAL-1.pdf.

little in the way of also ranking nations in the context of non-cyber crime variables, such as economic and technological metrics at the national level. There have been some multivariate macro-level cyber crime studies at the national level, which seek to identify inferential predictors of cyber crime rates between nations (Kigerl, 2013; 2016).

Among some of the multivariate studies conducted, the degree of internet connectivity within a nation has been found to consistently predict higher levels of multiple cyber crime types, including fraud, malware, spam, and digital piracy (Kigerl, 2013; 2016). Nations that are the source of spam and digital piracy are also more likely to be wealthier countries, measured in the form of a nation's gross domestic product (GDP). However, the findings are more mixed in regards to GDP's relationship to fraud and malware.

Yet these studies say little about how specific nations specialize in cyber crime, or how nations can be grouped under different typologies with varying cyber crime profiles. Predictors linked to cyber crime outcomes are assumed to be equally predictive across nations. The present research cannot describe the distinct differences between nations, only between variables.

The present research seeks to address this gap by performing K-means clustering analysis at the national level among 190 countries using seven cyber crime variables capturing fraud, malware, spam, and digital piracy, as well as each nation's GDP and internet use per capita. The results are intended to better understand the differences between individual countries by assigning them to discrete categories, rather than identifying the general relationship of various economic, technological, and legal differences between nations on cyber crime outcomes. Via clustering analysis techniques, nations may be grouped together with similar nations, attempting to identify if certain countries specialize in cyber crime, or act as cyber crime generalists.

Cyber crime differences between Countries

The internet is scattered with many reports and posts about the top cyber crime countries, the biggest source of cyber attacks, and the most likely home of residence for cyber criminals themselves. Multiple cyber security firms regularly release detailed reports on the state of cyber crime activities within their networks, often including a section on countries. Symantec has created an index of cyber crime activity that includes hosting malware, botnets, phishing server hosting, and how many botnet command and control (C&C) servers there are within each nation.[3] The top Five nations on this index include the US, China, Brazil, Germany, and India. Russia is seventh (Fossi, Turner, Johnson, Mack, Adams, et al., 2010).

The harms of the various cyber crimes are well advertised in such reports. Email spam is a common attack vector for multiple types of cyber crime (Rao & Reiley, 2012). Today, spam makes up 72% of all emails sent worldwide (Gudkova, 2013). Over half of all internet traffic in general not just that of email traffic, is actually spam (Lachhwani & Ghose, 2012). Loss of human resources due to the nuisance of spam was estimated to be at $22 billion in 2004 (Lachhwani & Ghose, 2012). Some of the top spam sending countries include China, Brazil, the United States, and Russia (Project Honey Pot, 2016).

[3] The report can be read at the following address: http://eval.symantec.com/mktginfo/enterprise/white_papers/b-whitepaper_internet_security_threat_report_xv_04-2010.en-us.pdf.

Spam can be harmful beyond just being a nuisance. Most spam today is sent from a type of malware called a botnet. As much as 76% of all spam is sent from such botnets (Symantec, 2014). A botnet is a network or cluster of malware-infected machines that the cyber criminal can control remotely over the internet. The victim that owns the infected computer is unaware of the botnet installation, and so the botnet master can use as many as thousands of remote PCs running in parallel to flood user inboxes with spam. The number of internet-connected computers that are infected with at least one botnet is estimated to be at 14% (Kindsight, 2012). Among the top countries known to host botnets include India, Vietnam, China, and Russia (Spamhaus, 2016).

The technical ability to write malicious code is not necessary to profit from cyber crime. Many times, an attacker does not have to exploit a technical vulnerability in a computer system to perpetrate his or her schemes. Many times the weakest link is the human target itself, as it is easier to socially engineer a human victim than it is to engineer the breach of a security flaw. There are three common forms of internet fraud, most often perpetrated using email spam. The first is phishing, whereby an attacker attempts to acquire a victim's "credential goods" through deceit. A credential good is any information a person may have that can be converted into cash, such as credit card numbers, internet banking logins, or social network member passwords. Typically, the phisher will link the victim to a website that looks identical to a trusted source the victim uses, such as a bank, whereby the user is requested to fill out a web form to capture the sought-after credential goods.

The United States is targeted the most by phishing attacks, suffering 60% of worldwide phishing volumes (RSA, 2014). When an individual falls victim to a phishing attack, the average amount lost is $1,800 (Cyveillance, 2008). However, businesses are often targeted, which result in a loss of $20,070 per business if successful (Ponemon Institute, 2013). Among some of the top sources of phishing attacks include China, Russia, Ukraine, the United States, and Brazil (APWG, 2014).

The second form of internet fraud is the advance fee fraud scheme, also known as 419 scams. These methods rely on even fewer technical skills to pull off, as the scheme is entirely that of social engineering. Advance fee fraud is a confidence trick whereby the fraudster contacts the victim via spam email with some sort of proposal. The proposal can be anything the victim wants, such as news of lottery winnings, a profitable business deal, or romantic relationship over an online dating website. However, before the deal can be finalized, the victim must wire the fraudster an "advance fee." Of course, there is no deal that the fraudster will deliver, and the scammer will continue to string the victim along making additional advance fees for as long as possible.

Because the fraudster can continue to victimize the same person many times before it is realized to be a scam, losses due to advance fee fraud can be greater than that of a phishing attack. Small losses are considered to be $200 to $30,000 per victim over the course of half a year (Ultrascan, 2013). Higher losses can reach as much as $210,000 over a period of one and a half years.

The loss of money is not the only risk that advance fee fraud imposes. In some cases, victims are lured to the home countries of the fraudster as part of the scheme, such as Nigeria where these scams are highly prevalent. If successful, the victim is kidnapped and held for ransom for more money (Ultrascan, 2013). Nigeria tends to be the most well-known as a top source of advance fee fraud schemes, but others include South Africa, Ghana, and the United States (Ultrascan, 2014).

The third form of internet fraud would include ransomware related schemes, which tend to be more targeted than either phishing or advance fee fraud. The offense is carried out only after having compromised a victim's computer with malware, where it searches for files important to the user and encrypts them, making them inaccessible to the victim without the decryption key. The fraudster then contacts the victim and demands a payment in order to be given the key so that the user can recover their inaccessible data (Gazet, 2010).

While the cyber crimes described tend to be serious, and reserved for career offenders who specialize in committing their given offenses, more common and minor cyber crimes would also include that of digital piracy. Unlike other forms of cyber crime which show volatility and have slowed in their expansion over time, digital piracy continues to rise to this day (Steele, 2015). Among some of the top pirating countries include the United States, China, Russia, and the Ukraine (V.I. Labs, 2014).

While there are many sources attempting to measure where the bulk of the cyber crimes are originating from in the world, few multivariate methods to assign nations to different categories based on cyber crime output have been attempted. There are also few sources that attempt to ascertain the causes behind such assignments as well, such as that of economic and technological differences between countries that are also part of their assigned typologies.

The research question is also a macro-level cyber crime question that involving cyber crime rates at the national level. Existing research has examined cyber crime at the national level, but only in terms of inference, rather than description via categorization. Nations highest in internet connectivity tend to be responsible for the most spam, malware, fraud, and digital piracy (Kigerl, 2013; 2016). Wealthier nations also send more spam but are not necessarily responsible for more fraud and malware.

While the predictors of cyber crime at the national level have been examined (Kigerl, 2012; 2016), less has been said about how nations specialize in cyber crime. Predictors are assumed to be equally predictive across nations, but what are the differences between nations? Clustering analyses would help to answer this question.

The existing typologies research in the context of cyber crime has mostly focused on categorizing the different types of cyber crime offenses themselves, such as distinguishing technical from non-technical cyber crimes, traditional crimes moved to cyberspace vs. entirely new crimes that require cyberspace, as well as classifying different crime and internet fraud types (Choo, 2008; Stabek, Watters, & Layto, 2010). Most of the research attempting to cluster countries into categories takes place in the marketing research domain (Cavusgil, Kiyak, & Yeniyurt, 2004), but nothing addresses cyber crime at this level.

The present research seeks to address this gap by performing a cluster analysis on 190 countries from around the world. The results will allocate each country to a typological category distinguished from nations of other types based on seven measures of cyber crime involving spam, malware, fraud, and piracy, as well as two other variables, GDP and internet connectivity, confirmed in prior research to relate strongly to cyber crime (Kigerl, 2012; 2013; 2016). The analyses should facilitate a better understanding of the different specializations and differences in the context of cyber crime at the national level.

Methods

Data

The population from which the sample was drawn from includes all sovereign countries in the world. The exact number of countries in the world varies depending on how the number is estimated[4] (Rosenberg, 2010). The count of countries can range from between 193 and 200. The Bureau of Intelligence and Research (2009) estimates the count to be 194[5]. The measures used to build the dataset were pulled from multiple reports, websites, spam email archives, and other sources, detailed below in the Measures section. Most variables constructed were retrieved from different sources from each other.

Sample

Initial construction of the national dataset from combining the various sources yielded a sample of 264 countries total. Different methods were used to construct each variable. Some sources only contained data on a smaller set of countries; other sources included data on territories as well as countries. Methods to build each variable from each source were conducted individually for each piece of source data. Following the construction of each variable, all measures were merged into a single dataset of nine variables, with the unit of analysis representing a country. Following a list wise deletion of all cases with one or more missing values for the variables selected, a final sample of 190 countries remained to be included in subsequent analyses.

Measures

Nine different measures were used for the analysis. The nine measures related to five different categories. Four of those categories were cyber crime related and included two digital piracy measures, three fraud measures, one spam variable, and one malware variable, totaling seven cyber crime related measures. An additional two variables were also included, relating to non-cyber crime characteristics of a country: gross domestic product and internet users per capita. These two measures have been confirmed in prior research to be highly predictive of cyber crime at the national level. Caution was exercised in the number of measures chosen as the clustering method being employed, that of K-means clustering would not be effective with highly dimensional data with only a sample size of 190. While a sample of 190 countries captures the majority of the nations on Earth, in absolute terms it is low enough that some selectivity is required for variable inclusion.

Four of the seven measures of cyber crime were acquired from spam email archives: phishing scams, advance fee fraud, malware distribution, and non-serious spam. The data were retrieved from the Untroubled Software website (http://untroubled.org/spam) on December 18, 2013. The available spam archives were collected by posting multiple "honey net" email addresses publicly online for spam crawlers to harvest. The honey net

[4] Some estimates are biased downward, such as those including only members of the United Nations, resulting in 193 countries. Other lists might reflect political agendas and thus exclude certain countries from the list of recognized independent states. Others are highly biased upwards as they include territories, which are not independent, as well as sovereign nations. There are as many as 60 territories.

[5] The definitions used for a country is a people politically organized into a sovereign state with a definite territory recognized as independent by the US.

approach is intended to bait spammers to add a given email address to a spam listserv, with the goal of intentionally receiving spam emails. When an email address is posted on internet websites such as forums, message boards, and on personally hosted web pages, web bots may scan and identify them as email addresses, extracting and storing them in a spam list or database for subsequent spam targeting. Spam messages in the archives received during 2012 were selected, totaling a sample of 871,146 messages to be processed.

The spam sample was then scanned with software extracting the originating IP address from each email (which is assumed to be the first IP address appended in the message's headers), looking up the IP address in a database of world addresses[6] to save each message's country code. It should be mentioned that the IP address contained in spam does not always represent where the spammers themselves reside. Seventy-six percent of spam messages are sent from botnets (Symantec, 2014), so the dataset should capture where cyber criminals choose to build their spam-sending botnets. Aggregated by nation, there were 183 countries that had sent one or more spam messages in 2012. All other nations were assigned a value of zero on all email cyber crime measures. The four email cyber crime measures were operationalized as rates based on the total amount of spam sent per country.

Phishing Scam Emails

Phishing scam emails were measured from the sample of spam emails detailed above. A re-purposed spam filter was written to calculate the probability a message was a phishing scam. Spam filters are used to calculate the probability a message received is spam, as opposed to the legitimate email that the users wish to receive/read (termed "ham" emails). Spam filters typically are trained on a sample of two data sources, a collection of spam emails that are already known to be spam, as well as a collection of ham emails (legitimate messages) a coder has already identified as legitimate. Spam filters scan the two data sources and calculate base probabilities that a given keyword found in an email is spam based on the keyword frequencies found in these two training data samples. The spam filter can then predict the probability a new set of emails are spam based on these joint keyword probabilities.

The spam filter employed for this research was a naïve Bayes classifier, which attempts to classify an instance of text as either being in one of two dichotomous categories based on trained probabilities associated with the text's keyword frequencies (Conway & White, 2012). To calculate the probability that the keyword is associated with a category, the probabilities that a keyword is found in phishing emails must first be analyzed.

There is a two-step process in order to employ fraud classification on the spam archives data sample: training the base keyword probabilities and then incorporating those probabilities into the software that can calculate whether an email instance is fraudulent. The training of the classifier was performed using two spam email samples, one taken from the spam archives itself, and the other pulled from a separate phishing fraud data source from online. One thousand spam emails were taken from the existing spam archives used in this sample that were confirmed via a human rater to be non-fraudulent in nature. The sample was used as the non-fraudulent training sample and was excluded from the final

[6] Downloaded from http://software77.net/geo-ip/

dataset building process by the software to avoid classifying messages the software was trained on. The second sample, consisting of phishing scam training emails, was acquired from two sources. The first source was acquired from a training corpus website intended to build spam filters,[7] which consisted of 2,275 example phishing emails. The second source was downloaded via a web crawler which extracted all phishing emails from the Anti-Fraud International web forum,[8] consisting of 2,936 example phishing scam emails. There was a total of 5,211 phishing scam emails used to train the classifier.

The accuracy of the measure of phishing used required testing to determine its predictive power. 2-fold cross validation was employed to assess predictive performance. 2-fold cross validation randomly splits the dataset into two halves, each used for both training the classifier and testing it, switching the halves each time (Kohavi, 1995). The keyword probabilities are trained based on one-half and tested on the second half to evaluate the number of correct predictions on the new data. These steps are performed both ways on the two halves, producing two scores which are averaged.

The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) (Fawcett, 2004) was selected as the metric to be used to determine the performance of the naïve Bayes classifier. The AUC typically ranges from .5 to 1.0, with a score of .5 being no better than a coin toss at correctly classifying a case and 1.0 being perfect predictive accuracy. The AUC yielded an average score of .94, suggesting high predictive performance for the measure of phishing. While a score of .94 may seem extraordinarily high, it is not unusual to achieve this level of predictive power in the context of spam filters (Cormack & Lynam, 2006). The classifier was then incorporated into the software used to code the spam sample, coding each spam email as '1' for being identified as a phishing scam, and '0' for being identified as non-fraudulent.

Advance Fee Fraud Emails

The above steps for constructing the measure of phishing were also repeated to build a measure of advance fee fraud. The training corpus of example advance fee fraud emails was pulled via a custom written web crawler to scrape the Anti-Fraud International website.[9] The website provides a web forum where users can submit example advance fee fraud emails they have received. There are 48 separate categories of advance fee fraud where users can submit their posts. The web crawler was written to download a random subset of messages from the site, yielding 5,213 messages total. The same sample of 1,000 non-fraudulent messages as was used for the phishing variable was used again for training the classifier. The 2-fold cross validation test was performed, yielding an average AUC of .97, suggesting very high predictive performance.

The classifier was then incorporated into the software used to code the spam data. Any message identified to be an advance fee fraud email was coded '1', otherwise it was coded '0'. The software was written so that classifying a message as phishing or as an advance fee fraud email would be mutually exclusive. If one message was identified to be both phishing or advance fee fraud, the type of fraud with the highest probability would be chosen. If both types of fraud were of equal probability, one type was chosen at random so as not to bias the data.

[7] Retrieved from https://monkey.org/~jose/wiki/doku.php?id=PhishingCorpus
[8] Retrieved from http://antifraudintl.org/
[9] Ibid

Malicious Spam Emails

Malicious spam was defined as any spam message distributing malware. Four measures of malware distribution were combined to capture this concept. The measure included any spam message containing a URL that was present on a blacklist of known malicious website links. Many types of security software will incorporate a list of malicious websites to block for security purposes. The blacklist was created from combining the publicly available lists from four different websites.[10]

Many times malicious URLs are hosted by legitimate websites that have been compromised by hackers. Eventually, many of these URLs will be cleaned. Where possible, cleaned website addresses that were taken off the present lists were also included in the blacklist. The reasoning for this was that the spam emails in the sample were exclusively unsolicited messages. It should be expected that if a formerly malicious URL is included in an unsolicited spam message, the message is not being sent on behalf of the legitimate website owner but is instead the compromised version.

There was a total of 234,273 unique URLs in the blacklist. The blacklist was added to an indexed database for speed. The software extracted all URLs contained in the body of each spam message in 2012 and that were identified in the blacklist database. If there was a positive match, the spam message was coded as malicious.

A second measure of malicious URLs was also constructed, attempting to measure malicious direct-download links. Some less sophisticated cyber criminals may link a recipient directly to an executable file download, after which it is hoped the user will open it. The software extracted all URLs contained in each email message that began with "http://" and ended with a dot (.) followed by a string that matches a list of executable file extensions (such as ".exe", ".pif", "docm", etc.). If the match was found, the message was coded as malicious.

A third measure of malware included executable attachments. Although no longer a common attack vector for distributing malware, if an unsolicited spam email includes an attachment to an executable file, the file is highly likely to be malicious. The software identified if a given message contained a file attachment and extracted the file extension from it. The extension was then matched with a known list of executable file extensions. If the file was found to be executable, the message was marked as malicious.

The fourth and final measure of malware distribution was the inclusion of malicious scripts. Executable scripts, such as JavaScript, can be embedded in the email itself which the email client, not the operating system or web browser per se, executes. Emails can be formatted with HTML, the same language used to design web pages. HTML can also include script tags, which an HTML interpreter or script engine, such as that in a web browser or email client, can execute. Most email clients disable running scripts in email because of the possibility for abuse, so it is rare for an email sender to include script tags in email. When an email does include such tags, it is likely for malicious purposes, such as installing malware.

The software matched any opening script tag embedded in the email body, or if applicable, any attached HTML files in the email. If the software matched "<script",

[10] Retrieved from https://isc.sans.edu/suspicious_domains.html, http://cyber crime-tracker.net, https://zeustracker.abuse.ch/monitor.php?browse=binaries, and http://hosts-file.net/?s=Download

followed by zero or more of any character of any length so long as there was no line break, ending with a closing bracket (">"), the software coded the message as malicious.[11]

Non-Serious Spam Emails

Both the fraud and malware measures were aggregated as a count by country of origin. Total spam sent in 2012 was also similarly aggregated and counted. A unique count of phishing, advance fee fraud, and malware was also computed, meaning if a message was coded as fraudulent or malicious or both, it would be counted as '1'. This count was subtracted from the total count of all messages to yield all non-fraudulent and non-malicious emails. The intent of this measure was to capture less serious spam, most likely a majority of which were spamvertizements selling some product, such as pharmaceutical drugs or counterfeit merchandise.

Phishing Server Top Level Domain (TLD) Rate

A third measure of fraud was intended to measure which countries hosted more phishing servers. Phishing TLD is the count of unique phishing servers that use a top level domain (TLD) representing a certain nation. Top level domains (such as .com and .net) can also represent a country code (.us, .uk). For example, in 2010, 435 phishing domain names had the form "http://www.{DomainName}.us," resulting in the United States being coded as being the source of 435 phishing servers. The TLD count was used as a rate based on population size. The data were taken from a report authored by the APWG (Aaron & Rasmussen, 2013). The TLD may or may not represent the physical location of which the phishing server was actually hosted, however. Some nation's TLDs are only able to be registered by its citizens, so it is possible that this variable is a better measure of cyber criminal location than are the fraudulent spam measures.

Count of BitTorrent Trackers

BitTorrent is a decentralized P2P protocol used for file sharing. BitTorrent networks are managed by servers called trackers, which users of the network must connect to in order to begin downloading a chosen file (Cuevas, Kryczka, Cuevas, Kaune, Guerrero, & Rejaie, 2010). Unlike previous P2P networks, BitTorrent trackers are less centralized and tracking servers can be hosted anywhere in the world where there is an internet connection. There also exist many BitTorrent tracker lists users may download to add to their chosen file sharing client; the intended purpose of which is to facilitate faster and more efficient downloads.

Eight BitTorrent tracker lists were downloaded for this research from multiple BitTorrent indexes and file locker websites. The lists included the URL of each tracker, the total of which was 2,476 trackers from the 8 lists. A script was written to automatically run DNS lookups of each URL to retrieve its IP address. At the time the script was run, 576 trackers could not be resolved on January 3, 2012. Of the IP addresses acquired, 998 were unique. The addresses were geolocated by country of origin on January 3, 2012 (via the following service: http://software77.net/geo-ip/multi-lookup). Six IP addresses were reserved and could not be geolocated, totaling 992 unique trackers from 51 nations that

[11] Scripts in emails tend to be of the form "<script language='JavaScript'>", and so the pattern matcher would identify such text.

had at least one tracker. All remaining countries not included were assigned a zero indicating no trackers present.

BitTorrent trackers are intended to capture piracy facilitating-countries. While those that use such trackers may reside anywhere in the world, the intent of this measure is to capture countries that are more apt to help with such infringing traffic. While it is true that not all BitTorrent traffic is infringing, present reports indicate that most of it is, with 78.1% of music traffic and 92.9% of television traffic being identified as infringing (Price, 2013).

File Sharing Client Downloads per Capita

The number of downloads of four file sharing clients available from SourceForge.net in 2012was used for this variable. SourceForge is a free online software repository for developers to maintain and share open source software that can be downloaded. File sharing clients (such as that of BitTorrent) are available for download at this website. A search for "p2p" on the SourceForge website was run on March 2, 2016 and the first four results were used for subsequent analysis (all file sharing clients). Top results are the most frequently downloaded (tens to hundreds of thousands of downloads per week).

SourceForge has a statistical reporting Application Programming Interface (API) that users can utilize to compute download statistics for each software application present on the site. For each of the four clients, the total number of downloads for the year 2012 per country were computed. There were over 39 million downloads in 200 countries and sovereign states in 2012 among the four file-sharing clients. The number of downloads per country was used as a rate based on country population size.

It is the intent of this research that this measure captures some of the variation in infringing behavior, as P2P clients are often necessary to infringe copyright via distributional networks. It is expected that many of these P2P clients that were downloaded were subsequently used for infringing purposes. While P2P networks can be used for non-illegal activities, such as sharing non-copyrighted works, much of the evidence surveilling these types of networks indicates the majority of P2P traffic is, in fact, infringing (Envisional, 2011).

Gross Domestic Product

GDP has been linked to all four types of cyber crime used for this research in previous studies (Kigerl, 2012; 2013; 2016). Wealthier nations are thought to be both more attractive targets for cyber criminals as the financial payoff for success is greater, as well as being home to more cyber criminal individuals. The impact of wealth on cyber crime is thought to be mediated by technological development. That is, wealthier nations lead to greater information and communications technology, which are necessary technologies to have access to in order to become involved in cyber crime. The data were retrieved from the World Bank website.[12] The data pertains to 192 nations in 2012.

Internet Users Per Capita

The number of internet users in a country is also considered to be one of the most important variables associated with cyber crime. Cyber criminals must often target existing

[12] Retrieved from http://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

internet users to send spam, malware, or fraudulent messages to. Cyber criminals themselves must also use the internet in order to perpetrate their crimes. The data on internet users were downloaded from the World Bank website[13] and details 201 countries in 2012.

Analytic Plan

The purpose of this research was to separate countries into distinct typologies based on their cyber crime and related activity. The research problem is a clustering problem, and the methodology chosen was a K-means clustering routine. K-means clustering groups cases within a dataset into distinct categories based on specified input variables (Hartigan, 1975). Cluster category assignment is based on the proximity of a given case's variable values to the mean or centroid of the cluster center in the hyperdimensional space that the variables occupy. Basically K-means groups cases together into categories based on how close those case are to each other. Cases with joint variable values that are closer to one another will tend to be assigned the same cluster category.

Prior to entry into the K-means algorithm, all measures were standardized so as to ensure their values were comparable. The different variables in their untransformed raw format are of different scale. K-means uses distances, so the measures were scaled to eliminate any bias the model would have on measures that are scaled with more units (Terrill, Paoline, & Manning, 2003).

A limitation with the K-means algorithm is that it does not select the optimum number of K clusters automatically. In order to determine the correct number of clusters to specify, the Flexclust package provided in the R language was used (Leisch, 2005). Flexclust offers a subroutine that performs K-means clustering on multiple separate bootstrapped samples of the original dataset. For this research, 100 pairs of bootstrapped datasets were produced seven times, once for seven different cluster sizes of two through eight.

For each pair of bootstrapped samples, K-means using the selected cluster size was conducted for each pair, and the agreement between the two models was assessed via the Rand index (Rand, 1971). The Rand index is similar to accuracy in measuring predictive performance. The Rand index measures the proportion of the cases which were assigned to the same category between the two models. The 100 Rand index scores are then averaged per each cluster size test.

The bootstrapping not only allows tests of generalizability of the model, but also assesses stability in assignment. In order to determine cluster centers, each cluster center must start at a given seed location within the input variable space. An iterative model will then move the initial values to a cluster center such that the distances between each cluster member to the center is minimized. This process makes K-means sensitive to initial conditions. Therefore, assessments of stability are important.

In order to determine the appropriate number of clusters, the elbow method was employed, whereby the moment an increase in the number of K clusters experiences diminished returns in improvements to the Rand index score, the optimum K size has been reached (Aalsma & Lapsley, 2001).

Finally, after a K size was selected, the model was cross-validated to assess the confidence in cluster category assignment. The dataset was randomly divided in to two

[13] Retrieved from http://data.worldbank.org/indicator/IT.NET.USER.P2.

halves, and K-means clustering was performed on one-half to assign categories to each case. A random forest model was then trained to predict the K categories. Random forests is a machine learning technique that can be used to predict categorical-level outcomes (Liaw & Wiener, 2002). The K-means centers trained on the first dataset half were then used to classify cases in the second dataset half based on case proximity to the given K center. The random forest model was also used to predict category assignment on the second dataset half. Lastly, Cohen's kappa was computed to determine the agreement between the two predictive models to assess the confidence in assignment.

Results

Descriptive Statistics

Table 1 contains the results of the descriptives of the nine unstandardized input variables used for clustering. There are seven cyber crime measures capturing malware, fraud, spam, and digital piracy, as well as a measure of GDP and internet use per capita. There is an average of five BitTorrent trackers hosted per nation, but with a very high amount of variability between countries (*SD* = 26.69). There is less variability for file sharing client downloads per capita, although there is still a much higher frequency of this type of measure (*M* = 914, *SD* = 2,565).

Table 1. Country Descriptives (*n* = 190)

Measure	Mean/%	Range	SD
BitTorrent Tracker Server Hosting Count	5.17	0 – 312	26.69
File Sharing Client Downloads per Capita	913.86	0 – 30,600	2564.63
Phishing TLDs Hosted Per Capita	1.98	0 – 69.65	8.11
Email Spam Percentage Breakdown★			
Percent Phishing Scams	7.57	0 – 100	13.56
Percent Advance Fee Fraud	14.35	0 – 100	18.9
Percent Malware	2.91	0 – 100	8.24
Percent Other Spam	56.31	0 – 100	34.56
Gross Domestic Product (in billions USD)	386	0.039 – 1,616	1,460
Percent Internet Users	39.72	0.8 – 96.21	28.62

★ Percentages do not total 100 as some countries sent zero spam messages

Email spam cyber crime types are all measured as an average percentage of total spam sent per country. The types include phishing scams, advance fee fraud scams, malware distributing spam, and other, non-serious spam. The average percentages contained in Table 1 do not sum to 100% as some countries were not identified to have sent any spam emails, resulting in the assignment of zero for all four measures of email spam.

The largest proportion of spam sent was non-serious spam (*M* = 56), as would be expected a priori. Promoting a spamvertised product relies more heavily on volume and requires less human attention and involvement as something such as fraud. Advance fee fraud was found to be almost twice as prevalent as phishing scams (*M* = 14 vs *M* = 8 for

phishing). Advance fee fraud requires less technical sophistication as it relies more on social engineering. It is also more directly profitable as the misappropriated funds are wired directly to the fraudster. In the case of phishing, stolen credential goods must then be converted into cash. Therefore, there may be greater motivation in the world to pursue advance fee fraud as the internet fraud scheme of choice.

Average GDP in the sample is 386 billion USD, with a standard deviation in the trillions (*SD* = 1,460). The result suggests a high degree of positive skew, with many countries clustered around the low and average end of the spectrum with a few highly wealthy countries well above the average. Finally, average internet access is a populace that is 40% connected (*SD* = 29%).

Cluster Analysis

In order to identify the number of K clusters to select, seven bootstrapping models were conducted, once for each of 2-8 cluster size models. All input variables were standardized prior to being entered in subsequent K-means models. Per each cluster size test, 100 bootstrapped pairs of K-means models were constructed, and their agreement was measured via a Rand index score. Index scores were averaged within each of the seven models, the results of which are presented in Table 2.

Table 2. Bootstrap Flexclust Algorithm Rand Index Scores to Select Cluster Number

Clusters	Mean Rand Index Score	Difference
2	.214	
3	.378	.164
4	.5191	.1411
5	.5645	.0454
6	.5836	.0191
7	.6236	.04
8	.5955	-.0281

Table 2 contains the mean Rand index score per each of the 2-8 cluster sizes, along with their differences with each cluster size increase. Rand index scores are a measure of cluster assignment agreement between two bootstrapped pairs of datasets with clustering performed on them. Larger differences are seen as cluster number size increases starting from two, decelerating at the point which a cluster size of four increases to five, after which a diminishing return on improvement to the Rand index is witnessed. A cluster number of four was concluded to be the most appropriate number of categories of which to divide countries into.

The K-means model of K = 4 was then validated to assess the confidence in cluster member assignment. A split-sampling approach was taken, training a K-means model as well as a random forest model on one random half of the 190 countries, and creating the 4-categtory estimates for the second half for both the trained K-means model and the random forest algorithm. Cohen's kappa was then computed to measure the rater agreement in category assignment between these two methods. The result yielded a kappa statistic of .9 ($p < .001$), suggesting high confidence in cluster category assignment.

A K-means model of K size 4 was then applied to the full dataset. The results of the cluster analysis and the centroid scores for each input variable are presented in Table 3. The names of the countries classified and their cluster assignment can be found in the Appendix. The first cluster appears to score low to average on every measure of cyber crime. Cluster 1 contains 90 countries (47.37% of the total). Cluster 1 presents consistently moderately low average scores for both measures of piracy (BitTorrent trackers = -.167 and file sharing client downloads = -.313), both measures of phishing (phishing server TLDs = -.216 and phishing emails = -.272), as well as on advance fee fraud scam emails (M = -.358). The cluster appears to be approximately average for both malware distribution (M =.104) and non-serious spam (M = .057).

Table 3. Country Clustering Centroids

Measure	Cluster1 Low Cyber crime Countries n = 90 M (SD)	Cluster2 Advance Fee Fraud Specialists n = 16 M (SD)	Cluster3 Non-serious Cyber crime Countries n = 8 M (SD)	Cluster4 Phishing Specialists n = 76 M (SD)	F
BitTorrent Trackers	-0.167 (0.151)	-0.161 (0.103)	3.099 (3.818)	-0.094 (0.24)	1.74
File Sharing Downloads	-0.313 (0.098)	-0.127 (0.344)	1.538 (4.07)	0.235 (0.714)	16.75★★
Phishing Servers	-0.216 (0.08)	0.22 (1.228)	-0.102 (0.24)	0.22 (1.445)	7.42★★
Phishing Emails	-0.272 (0.603)	-0.214 (0.506)	-0.042 (0.31)	0.372 (1.338)	18.94★★
Advance Fee Fraud	-0.358 (0.527)	2.642 (1.012)	0.013 (0.755)	-0.133 (0.521)	0.39
Malicious Emails	0.104 (1.407)	-0.201 (0.219)	-0.007 (0.284)	-0.08 (0.356)	1.23
Non-serious Spam	0.057 (1.118)	-0.767 (0.499)	0.198 (0.882)	0.074 (0.88)	0.16
Gross Domestic Product	-0.209 (0.157)	-0.135 (0.356)	3.118 (3.641)	-0.052 (0.363)	3.48†
Percent Internet Users	-0.801 (0.466)	-0.2 (1.044)	1.237 (0.561)	0.86 (0.569)	329.44★★★

★ < .05, ★★ < .01, ★★★ < .001, † < .1

Cluster 1 has been designated as the "low cyber crime countries" cluster, and the reason why such crimes may be low could be due to the similarly low centroids for GDP (M = -.209) and internet users per capita (M = -.801). For both of these economic measures, Cluster 1 scores the lowest out of any of the four clusters, especially for internet users.

Cluster 2 has been designated the "advance fee fraud specialist" cluster, as it scores substantially higher than any of the other three clusters on advance fee fraud emails (M = 2.642). Sixteen countries make up Cluster 2, accounting for eight percent of the sample. Unsurprisingly, Cluster 2 contains the nation of Nigeria, well known for its 419 scams.

Refer to the Appendix for further details on which nations were allocated to this group. Cluster 2 also scores modestly above average for phishing server TLDs (M = .22), but not phishing emails (M = -.214), perhaps indicating these countries have a small hand in this other categories of fraud as well. However, Cluster 2 scores somewhat low on BitTorrent trackers (M = -.161), file sharing client downloads (M = -.127), malicious emails (M = -.201), and very low on non-serious spam (M = -.767), perhaps because so much of the cluster's spam is that of advance fee fraud.

Cluster 2 appears to be second only to Cluster 1 in being the lowest in GDP (M = -.135) and internet connectivity (M = -.2). Advance fee fraud scams tend to be the least technically sophisticated types of cyber crimes. The low internet connectivity of this type of country may explain the tendency towards this form of serious, but less technically involved, cyber crime.

Cluster 3 scores the highest out of any of the four clusters on non-serious forms of cyber crime and hence has been labeled the "non-serious cyber crime countries" cluster. Non-serious cyber crimes here refer to digital piracy and regular email spam that is not otherwise classified as fraudulent or malicious. There are eight countries allocated to this cluster, and make up just four percent of the sample. Nations in this cluster include the United States, Canada, China, and Japan.

Cluster 3 is especially high in digital piracy activity, and by far has the most BitTorrent trackers (M = 3.099) and file sharing client downloads (M = 1.538). The cluster also scores the highest out of the other three on non-serious spam specialization (M = .198), although the effect is not as pronounced as it is for piracy. The cluster scores about average for the other remaining measures of cyber crime, including phishing servers (M = -.102), phishing emails (M = -.042), advance fee fraud (M = .013), and malware distribution (M = -.007).

Nations within this group are by far the wealthiest (M = 3.118) and most connected (M = 1.237). With so many internet users, it makes sense that these nations would be a greater source of digital piracy. Wealthier nations would also make better targets for which to send spam to, in hopes of selling a product.

The last cluster is Cluster 4, which has tentatively been labeled the "phishing specialist countries" cluster. The cluster includes 76 nations, totaling forty percent of all countries tested. The cluster scores among the highest on both measures of phishing attacks, scoring equally as high as Cluster 2, the advance fee fraud cluster, in terms of phishing servers (M = .22, SD = 1.445) and easily the highest in terms of being the source of the most phishing email attacks (M = .372, SD = 1.338). Although it should be noted that there is a large degree of variability among these phishing measures within the cluster that is not present as much in the other clusters, suggesting not all nations assigned to the Cluster 4 group are especially high in phishing. Nations assigned to this cluster include Russia and Brazil.

Cluster 4 scores about average in terms of BitTorrent trackers (M = -.094), advance fee fraud (M = -.133), malicious emails (M = -.08), and spam (M = .074). However, these nations do indicate modestly high file sharing client downloads per capita (M = .235). Nations within this group have average GDP (M = -.052) and high internet connectivity (M = .86), a high internet connectivity that is second only to the non-serious cyber crime countries of Cluster 3. Phishing schemes require more technical sophistication to carry out, and so it is intuitive that this cluster scores high on this measure.

ANOVA testing was also conducted to determine significant differences on each of the input variables between the 4 cluster categories. File sharing downloads (F = 16.75, p<

.001), phishing servers ($F = 7.42$, $p < .01$) and phishing emails ($F = 18.94$, $p < .001$) were all found to significantly differ across the four clusters. The remaining cyber crime measures did not achieve significance. It is therefore stressed that caution should be exercised when interpreting the means for the remaining cyber crime measures. However, it should also be remembered that the purpose of significance is to generalize findings in a sample to the population at large, and with 190 countries in the sample and thus 95% of the entire population of nations on Earth, significance becomes less useful.

Discussion and Conclusion

The present research sought to build on prior reporting attempting to rank nations into categories based on cyber crime output. Previous research has investigated how the predictors at the national level can be used to explain cyber crime. The current findings expanded on this by partitioning individual nations into four different cluster categories.

K-means clustering was performed on approximately 95% of all the countries worldwide, or 190 out of 200 total. Nations were assigned to one of four clusters, including low cyber crime countries with low GDP and internet connectivity, advance fee fraud specialist nations, with modestly low connectivity; non-serious cyber crime nations, high in piracy and email spam and that were the wealthiest with the most internet users, and phishing specialist countries, also with high internet connectivity but average wealth.

The analyses for the current research were performed on data sources that have been used in previous studies on cyber crime. Some of the previous research used the same data sources only for different years (Kigerl, 2012; 2013), another used the same year of 2012 (Kigerl, 2016). The results of the clustering analysis help shed some light on the implications of the findings from prior studies using these datasets.

GDP and internet users per capita were non-cyber crime measures utilized for the present study because they had been highlighted as among the most important predictors of cyber crime in the context of these data sources. Internet use has been shown to consistently positively predict all cyber crime types. In the case of the present clustering analysis, the low cyber crime nations cluster possessed the lowest average internet use, which was not surprising.

However, the average internet connectivity score for advance fee fraud specialist cluster nations was slightly below average. Nations determined to have higher average phishing scam scores predictably also had higher average internet user scores. Previous research using these data sources did not seek to break fraudulent spam emails into these two distinct types of fraud, hence why the finding is new. Future research might seek to further examine how technological development can be used to predict differences in both advance fee fraud and phishing at the national level.

The centroid scores for GDP had similar implications as that of internet use. The low cyber crime cluster also had the lowest wealth. A previous investigation into this dataset, however, found GDP to be a negative predictor of fraud in general (Kigerl, 2016). Advance fee fraud nations presented with the second lowest GDP of the four clusters, although the average was very modestly low. Phishing specialist countries had average GDP scores. Again, some differences between these two types of fraud emerge once fraud is broken down into these two categories in such a way.

Advance fee fraud nations tended to be lower in GDP and internet use, as contrasted with phishing countries which tended to be higher on these dimensions. Again, the

differences in technical sophistication required to carry out such crimes might explain the variation. Phishing usually relies on linking recipients to a website hosted on a server set up by the scammer, with a fillable form to capture credential goods presented on a webpage also written by the scammer or an accomplice. Advance fee fraud relies on almost exclusively social engineering tactics, rather than technical engineering.

Consistent with prior research on these data sources, GDP and internet use was very high in the non-serious cyber crimes clusters, where piracy and email spam scores were higher than the other cluster groups. Advance fee fraud may be more unusual a form of cyber crime than phishing, piracy, and spam. However, it should be noted that malware did not particularly stand out amongst any of the clusters found. Malware never scored any higher than .1 standard deviations above the average for any cluster over the rest. It cannot be concluded that there were any malware specialist countries.

This could be due to a lack of specialization for this type of cyber crime among nation states, or it could be a limitation of the measure of malware constructed for this analysis. Only 2.9% of the total email spam messages in the sample were identified to be malicious, although this number is not substantially lower than separate reporting suggesting 4.5% of all spam emails are malicious (Alt-N Technologies, 2013). While there was no convincing evidence that any nations specialize in malicious spam, it may be that some nations specialize in other types of malware distribution not accounted for here, such as botnet hosting and port scanning source countries.

It should also be mentioned that there was a relatively small number of non-serious cyber crime and advance fee fraud specialist nations, amounting to 4.21% and 8.42% of the total in the sample, respectively. It may be that a small number of countries account for a disproportionate amount of these forms of crime. There were almost twice as many advance fee fraud emails as phishing scam emails worldwide, yet only a small number of nations appear to account for the high advance fee fraud volume. As many as 40% of the nations were assigned to the cluster with above average phishing attacks, suggesting a broader number of countries might rely on this tactic.

Limitations

There are some limitations of the data sources that should be mentioned. Each of the spam email sources of cyber crime in the sample were geolocated to a country using that message's IP address. An IP address contained in an email can represent where the email was sent from. However, the originating IP address is rarely indicative of where the cyber criminals themselves reside. Instead, most spam is sent from botnet infected machines, managed by a botmaster or the spammer. It is estimated that 76% of spam originates from a botnet (Symantec, 2014). The nations that are identified to be the sources of spam are not intended to capture spammer residence, but rather where spammers choose to host their botnets.

The measure of malware also may be somewhat sparse. The most common attack vector for distributing malware via spam email is through a link to a website that performs a drive-by download and install of a malicious payload. The most common method for identifying an email as malicious in this way is to use a blacklist of known bad links. A number of blacklists were acquired for this research to construct the malware measure. However, not all sources that maintain blacklists retain de-listed websites. When a website is identified to be infected, it is either hosted by the cyber criminal or it is a legitimate website compromised by the cyber criminal. Once a website is no longer online, or once

an infected website is cleaned, it is taken off the blacklist. It was therefore not possible to be completely certain that a link found in an email was formerly on a blacklist but was removed by the time this research retrieved the blacklist database. There thus could be some sparseness in how many emails were identified to be malicious in this way, inflating the number of false negatives.

Finally, the two measures of piracy, that of BitTorrent tracking servers and file sharing client downloads, would only partially capture infringing activity. If a user engages in file sharing that is not infringing, then the activity cannot be classified as cyber crime. Present estimates of file sharing traffic indicate that most of it is infringing, with 78.1% of music traffic and 92.9% of television traffic being identified as unauthorized distribution of copyrighted goods (Price, 2013). It may be that the extra 10-20% of file sharing traffic that is not infringing systematically skews results that would otherwise be different. However, the two measures mostly capture infringing activity.

Policy Implications

Similar to how clustering techniques have been used for market segmentation of countries for better targeted advertising based on nation, so too can such techniques be used for other types of targeting, namely legal and law enforcement related pursuits. One of the greatest difficulties for law enforcement against cyber criminals is the international reach of a cyber offense, where attacks are not slowed by physical barriers such as national borders. Increased international cooperation is a solution to this problem.

However, international treaties are slow to be established and are also slow to be enforced. Broad sweeping cyber crime legislation might be more difficult to establish between nations and more complicated to write. Instead, targeted international treaties can focus only on certain nations based on their cluster category assignment and the type of cyber crime they are associated with. If the specific laws negotiated only pertained to the types of offenses the nation is assigned, then there would be a shorter list of laws to agree to, accelerating drafting of any bills and also increasing the chances of ratification. The targeted nature of the bills would also make them the most cost effective in terms of pursuing offenders.

While some of the offenses that are assigned to a nation are not committed by residents of those countries themselves, but instead can be routed through them, international cooperation is still important. The nation serving as the conduit for a cyber attack can still furnish help in tracking the offender down and identifying their true geographic location. As the attack is routed through the nation's own infrastructure and resources, acquiring their assistance with evidence would be beneficial.

Future Research

Future research might seek to explore the differences between advance fee fraud and phishing scams at the national level further. Some important differences were discovered in the present study between the two types of fraud that might suggest some different implications for how economic and technological related predictors could relate to the two in different ways. The website where the advance fee fraud training sample was acquired separated all messages into 48 different categories of fraud. Phishing scams could also be separated further into different categories. Future studies might break fraud down

into more than two categories to examine how the different fraud types relate to the attributes of the given country of which such fraud originates from.

Malware could further be explored as well. Machine learning techniques similar to that used to classify fraud could be employed to create additional measures of malware. Metrics of malware present in contexts outside of email spam, such as botnet hosting and hacking attack source countries, could also be considered. The present research did not find enough convincing evidence that any nations specialize in malware. Alternate measures of malware could yield something different.

References

Aalsma, M. C., & Lapsley, D. K. (2001). A typology of adolescent delinquency: Sex differences and implications for treatment. *Criminal Behaviour and Mental Health*, *11*(3), 173-191.

Aaron & Rasmussen (2013). Global Phishing Survey: Trends and Domain Name Use in 2H2012. APWG, July-December 2012 report.

Alt-N Technologies (2013). Internet threats trend report, April 2013. Retrieved from: http://static.altn.com/Collateral/Security-Threat-Trend-Reports/2013-Q1_Email-Threat-Trend-Report.pdf.

APWG (2014). Phishing activity trends report. APWG, 2nd quarter 2014. Retrieved from: https://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf

Bureau of Intelligence and Research (July 29, 2009). Independent states in the world. U.S. Department of State. Retrieved from http://www.state.gov/s/inr/rls/4250.htm.

Cavusgil, S. T., Kiyak, T., & Yeniyurt, S. (2004). Complementary approaches to preliminary foreign market opportunity assessment: Country clustering and country ranking. *Industrial Marketing Management*, *33*(7), 607-617.

Choo, K. K. R. (2008). Organized crime groups in cyberspace: a typology. *Trends in organized crime*, *11*(3), 270-295.

Conway, D. & White, J. M. (2012). Machine Learning for Email. Sebastopol, CA: O'Reilly Media.

Cormack, G. V., & Lynam, T. R. (2007). Online supervised spam filter evaluation. *ACM Transactions on Information Systems (TOIS)*, *25*(3), 11.

Cuevas, R., Kryczka, M., Cuevas, A., Kaune, S., Guerrero, C., & Rejaie, R. (2010). Is content publishing in BitTorrent altruistic or profit-driven? Co-NEXT '10 Proceedings of the 6th International Conference. ACM New York, NY, USA. ISBN: 978-1-4503-0448-1.

Cyveillance (October, 2008). The cost of phishing: Understanding the true cost dynamics behind phishing attacks. A Cyveillance report.

Envisional (January, 2011). Technical report: An estimate of infringing use of the internet. Technical report.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 31, 1-38.

Fossi, M., Turner, D., Johnson, E., Mack, T., Adams, T., Blackbird, J., Entwisle, S., Graveland, B., McKinney, D., Mulcahy, J. & Wueest, C. (April, 2010). Symantec global internet security threat report. Symantec, 15.

Gazet, A. (2010). Comparative analysis of various ransomware virii. *Journal in computer virology*, *6*(1), 77-90.

Gudkova, D. (January, 2013). Kaspersky security bulletin: Spam evolution 2012. Securelist.

Hartigan, J. A. (1975). *Clustering algorithms*. New York, NY: Wiley, 1975.

Kigerl, A. (2012). Routine activity theory and the determinants of high cyber crime countries. *Social Science Computer Review*, *30*(4), 470-486.

Kigerl, A. (2013). Infringing nations: Predicting software piracy rates, BitTorrent tracker hosting, and p2p file sharing client downloads between countries. *International Journal of Cyber Criminology*, 7(1), 62.

Kigerl, A. (2016). *Routine Activity Theory and Malware, Fraud, and Spam at the National Level*. Unpublished manuscript.

Kindsight (2012). Malware report Q2 2012. Kindsight Security Labs. Retrieved from https://www.kindsight.net/sites/default/files/Kindsight_Security_Labs-Q212_Malware_Report-final.pdf.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai,* 14(2), 1137-1145.

Lachhwani, V., & Ghose, S. (2012). Online information seeking for prescription drugs. *International Journal of Business and Systems Research*, *6*(1), 1-17.

Leisch, F. (2005). flexclust: Flexible Cluster Algorithms. R package.

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R news*, *2*(3), 18-22.

Ponemon (2013). *2013 Cost of cyber crime study*. United States. Ponemon Institute.

Price, D. (2013). Sizing the piracy universe. *NetNames Envisional*, September, 2013.

Project honey pot statistics (2016). Project Honey Pot. Retrieved from http://www.projecthoneypot.org/statistics.php on March 13, 2016.

Rao, J. M., & Reiley, D. H. (2012). The economics of spam. *The Journal of Economic Perspectives, 26*(3), 87-110.

Rosenberg, M. (June 14, 2010). The number of countries in the world. About.com. Retrieved from http://geography.about.com/cs/countries/a/numbercountries.htm.

RSA (2014). 2013 A year in review: Fraud report.

Spamhaus (2016). The World's Worst Botnet Countries. Spamhaus Blocklist database. Retrieved from https://www.spamhaus.org/statistics/botnet-cc.

Stabek, A., Watters, P., & Layto, R. (2010, July). The seven scam types: mapping the terrain of cyber crime. In Cyber crime and Trustworthy Computing Workshop (CTC), 2010 Second (pp. 41-51). IEEE.

Symantec (2014). Internet security threat report. 2013 Trends, 19.

Terrill, W., Paoline, E. A., & Manning, P. K. (2003). Police culture and coercion. *Criminology*, *41*(4), 1003-1034.

Ultrascan (2013). 419 advance fee fraud statistics 2013. Ultrascan Advanced Global Investigations.

Ultrascan (2014). 419 advance fee fraud statistics 2013. Ultrascan Advanced Global Investigations. PRE-Release 1.5 Amsterdam, 23 July, 2014. Retrieved from: http://www.ultrascan-agi.com/public_html/html/pdf_files/Pre-Release-419_Advance_Fee_Fraud_Statistics_2013-July-10-2014-NOT-FINAL-1.pdf.

V.I. Labs (2014). Top 20 Countries for Software Piracy and License Misuse. Code Confidential: The V.i. Labs Blog. Retrieved from http://www.vilabs.com/news-section/code-confidential/top-20-countries-software-piracy-2014.

Appendix

List of Countries by Cluster Assignment			
Cluster 1 Low Cyber crime Countries *n* = 90	Cluster 2 Advance Fee Fraud Specialists *n* = 16	Cluster 3 Non-serious Cyber crime Countries *n* = 8	Cluster 4 Phishing Specialists *n* = 76
Afghanistan	Barbados	Brunei	Albania
Algeria	Benin	Canada	Andorra
Angola	Burkina Faso	China	Antigua and Barbuda
Bangladesh	Côte d'Ivoire	France	Argentina
Belize	Estonia	Germany	Armenia
Bhutan	Ghana	Japan	Australia
Bolivia	Iceland	Netherlands	Austria
Botswana	Italy	United States	Azerbaijan
Burundi	Jamaica		Bahamas
Cambodia	Kyrgyzstan		Bahrain
Cameroon	Malaysia		Belarus
Cape Verde	Nigeria		Belgium
Central African Republic	Paraguay		Bermuda
Chad	Samoa		Bosnia and Herzegovina
Comoros	Tajikistan		Brazil
Congo - Brazzaville	Vanuatu		Bulgaria
Congo - Kinshasa			Chile
Cuba			Colombia
Djibouti			Costa Rica
Dominican Republic			Croatia
Egypt			Cyprus
El Salvador			Czech Republic
Equatorial Guinea			Denmark
Eritrea			Dominica
Ethiopia			Ecuador
Fiji			Faroe Islands
Gabon			Finland
Gambia			Greece
Georgia			Grenada
Guatemala			Guyana
Guinea			Hong Kong SAR China

Guinea-Bissau
Haiti
Honduras
India
Indonesia
Iran
Iraq
Jordan
Kenya
Kiribati
Laos
Lesotho
Liberia
Madagascar
Malawi
Maldives
Mali
Marshall Islands
Mauritania
Mauritius
Micronesia
Moldova
Mongolia
Mozambique
Myanmar (Burma)
Namibia
Nepal
Nicaragua
Niger
Pakistan
Palestinian Territories
Papua New Guinea
Peru
Philippines
Rwanda
Senegal
Sierra Leone
Solomon Islands
South Africa
Sri Lanka
Hungary
Ireland
Israel
Kazakhstan
Kuwait
Latvia
Lebanon
Liechtenstein
Lithuania
Luxembourg
Macau SAR China
Macedonia
Malta
Mexico
Montenegro
Morocco
New Zealand
Norway
Oman
Panama
Poland
Portugal
Puerto Rico
Qatar
Romania
Russia
Saint Kitts and Nevis
Saint Lucia
Saint Vincent and the Grenadines
Saudi Arabia
Serbia
Seychelles
Singapore
Slovakia
Slovenia
South Korea
Spain
Sweden
Switzerland
Trinidad and

	Tobago
Sudan	Turkey
	United Arab
Suriname	Emirates
Swaziland	United Kingdom
São Tomé and Príncipe	Uruguay
Tanzania	Venezuela
Thailand	
Timor-Leste	
Togo	
Tonga	
Tunisia	
Turkmenistan	
Tuvalu	
Uganda	
Ukraine	
Uzbekistan	
Vietnam	
Yemen	
Zambia	
Zimbabwe	