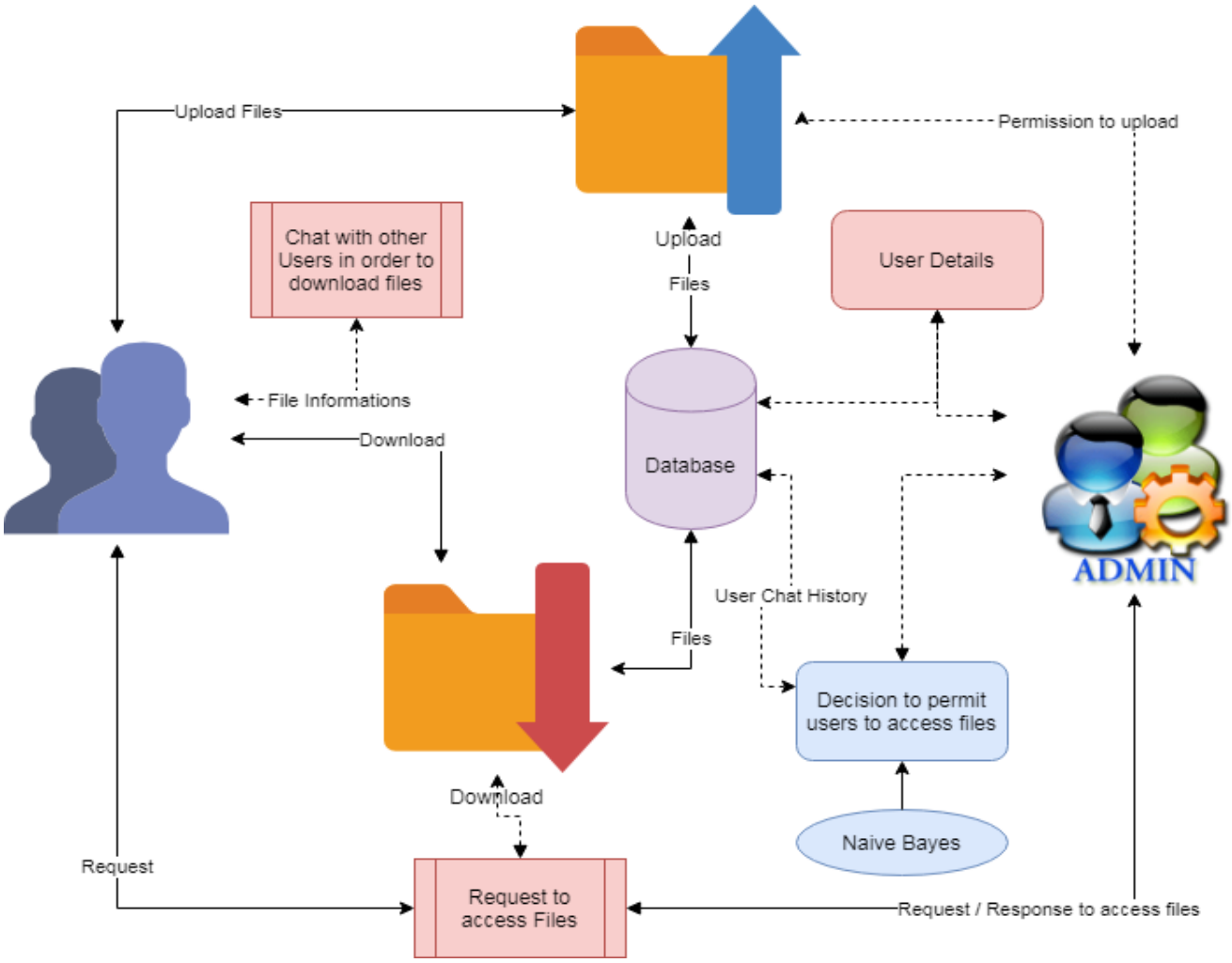
A DATA ANALYTICS APPROACH TO THE CYBERCRIME UNDERGROUND ECONOMY

ABSTRACT:

Despite the rapid escalation of cyber threats, there has still been little research into the foundations of the subject or methodologies that could serve to guide Information Systems researchers and practitioners who deal with cyber security. In addition, little is known about Crime-as-a-Service (CaaS), a criminal business model that underpins the cybercrime underground. This research gap and the practical cybercrime problems we face have motivated us to investigate the cybercrime underground economy by taking a data analytics approach from a design science perspective. To achieve this goal, we propose (1) a data analysis framework for analyzing the cybercrime underground, (2) CaaS and crime ware definitions, and (3) an associated classification model. In addition, we (4) develop an example application to demonstrate how the proposed framework and classification model could be implemented in practice. We then use this application to investigate the cybercrime underground economy by analyzing a large dataset obtained from the online hacking community. By taking a design science research approach, this study contributes to the design artifacts, foundations, and methodologies in this area. Moreover, it provides useful practical insights to practitioners by suggesting guidelines as to how governments and organizations in all industries can prepare for attacks by the cybercrime underground.

ARCHITECTURE:



EXISTING SYSTEM:

Cybercrime has undergone a revolutionary change, going from being product-oriented to service-oriented because the fact it operates in the virtual world, with different spatial and temporal constraints, differentiates it from other crime taking place in the physical world. As part of this change, the cybercrime underground has emerged as a secret cybercrime marketplace because emerging technological changes have provided organized cybercriminal groups with unprecedented opportunities for exploitation. The cybercrime underground has a highly professional business model that supports its own underground economy. This business model, known as CaaS, is “a business model used in the

underground market where illegal services are provided to help underground buyers conduct cybercrimes, such as attacks, infections, and money laundering in an automated manner,". Thus, CaaS is referred to as a do-it-for-me service, unlike crimeware which is a do-it-yourself product. Because CaaS is designed for novices, its customers do not need to run a hacking server or have high-level hacking skills. Consequently, the CaaS business model can involve the following roles: writing a hacking program, performing an attack, commissioning an attack, providing an attack server (infrastructure), and laundering the proceeds. Sood and Enbody have suggested that crimeware marketplaces have three key elements, namely actors (e.g., coders, operators, or buyers), value chains, and modes of operation (e.g., CaaS, pay-per-install, crimeware toolkits, brokerage, or supplying data). Periodic monitoring and analysis of the content of cybercrime marketplaces could help predict future cyber threats.

PROPOSED SYSTEM:

The goal of our data analysis framework is to conduct a big-picture investigation of the cybercrime underground by covering all phases of data analysis from the beginning to the end. This framework comprises four steps: (1) defining goals; (2) identifying sources; (3) selecting analytical methods; and (4) implementing an application. Because this study emphasizes the importance of RAT for analyzing the cybercrime underground, the proposed RAT-based definitions are critical to this framework: Steps 1–4 all contain the RAT elements A. **Step 1:** Defining Goals The first step is to identify the conceptual scope of the analysis. Specifically, this step identifies the analysis context, namely the objectives and goals. To gain an in-depth understanding of the current CaaS research, we investigated the cybercrime underground, which operates as a closed

community. Thus, the goal of the proposed framework is to “investigate the cybercrime underground economy.” B. **Step 2:** Identifying Sources the second step is to identify the data sources, based on the goals defined by Step 1. This step should consider what data is needed and where it can be obtained. Since the goal of this study is to investigate the cybercrime underground, we consider data on the cybercrime underground community. We therefore collected such data from the community itself and obtained a malware database from a leading global cyber security research firm. Because cybercriminals often change their IP addresses and use anti-crawling scripts to conceal their communications, we used a self-developed crawler that can resolve captchas and anti-crawling scripts to gather the necessary data. We collected a total of 2,672,091 posts selling CaaS or crimeware, made between August 2008 and October 2017, from a large hacking community site (www.hackforums.net) with over 578,000 members and more than 40 million posts. We also collected 16,172 user profiles of sellers and potential buyers, based on their communication histories, as well as prices and questions and answers about the transactions. The black market uses traditional forum threads (e.g., bulletin boards) instead of typical e-commerce platforms (e.g., eBay, and Amazon). For example, sellers create threads in marketplace forums to sell items, and potential buyers comment on these threads. One of the most significant challenges was therefore converting this unstructured data into structured data. Since the product features, prices, and descriptions were explained within longer texts, we used a variety of text mining techniques to extract the important features: for example, we used named entity recognition to extract company names (see Section IV-C(2)). Since these texts included many typographic errors and jargon terms, we had to create a dictionary for use during

a preprocessing step. In addition, we obtained a malware database from a cybersecurity firm containing over 53,815 entries covering cybercrimes between May 11, 2010 and January 13, 2014. This unique dataset strengthened our study by providing real-world evidence from a different viewpoint.

MODULES:

1. Upload Files

Users are allowed to upload the files with the tags given. Once the file is uploaded, then it is sent to approval from admin to publish or make view to other users. These uploaded files can be in any form document, audio or video but not allowed to upload the executable (.exe) files.

2. Conversation Monitoring

Users are allowed to communicate among the other users. This could be monitor by the admin. The malicious conversion likes to threaten the data. In order to protect the cybercrime and prevents from forming cybercrime community. This can be achieved by the help of classification algorithm named naïve Bayes classification.

3. Download Files

The files can be downloading by requesting for the file and once admin approved the files then can be downloadable. The decision to approve files can be taken from the conversation between users. Admin takes the action on download files and approvable status of users. The users are allowed further actions based on the users.

4. Graphical Representations

The analyses of proposed systems are calculated based on the approvals and disapprovals. This can be measured with the help of graphical notations such as pie chart, bar chart and line chart. The data can be given in a dynamical data.

ALGORITHM:

Naive Bayes Classifier

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called *naive Bayes* or *idiot Bayes* because the calculation of the probabilities for each hypothesis is simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

Make Predictions with a Naive Bayes Model

Given a naive Bayes model, you can make predictions for new data using Bayes theorem.

MAP (h) = max(P(d|h) * P(h))

Using our example above, if we had a new instance with the *weather* of *sunny*, we can calculate:

go-out = P(weather=sunny|class=go-out) * P(class=go-out)
stay-home = P(weather=sunny|class=stay-home) * P(class=stay-home)

We can choose the class that has the largest calculated value. We can turn these values into probabilities by normalizing them as follows:

P(go-out|weather=sunny) = go-out / (go-out + stay-home)
P(stay-home|weather=sunny) = stay-home / (go-out + stay-home)

If we had more input variables we could extend the above example. For example, pretend we have a "*car*" attribute with the values "*working*" and "*broken*". We can multiply this probability into the equation.
For example below is the calculation for the "go-out" class label with the addition of the car input variable set to "working":

go-out = P(weather=sunny|class=go-out) * P(car=working|class=go-out) * P(class=go-out)

FUTURE WORK:

Although our study has made several significant findings, it nevertheless has several limitations that will need to be addressed in future studies. These will be able to add more analysis and significant further insights. First, we only collected data from the largest hacking community and did not consider other hacking communities. Future studies will therefore need to generalize our findings by investigating a wider range of hacking communities. Second, this study has focused on the CaaS and crimeware available in the cybercrime underground, but much in-depth analysis remains to be done on the configurations of cybercrime networks. Future research could cluster keywords and threats by industry to provide a deeper understanding of the potential vulnerabilities, and it could attempt to discover the network effects involved or the leaders of the cybercrime underground.

CONCLUSION:

Because this study takes a DSR approach, we have focused mainly on building and evaluating artifacts rather than on developing and justifying theory: actions are usually considered to be the main focus of behavioral science. We have therefore proposed two artifacts: a data analysis framework and a classification model. We have also conducted an ex-ante evaluation of our classification model's accuracy and an ex-post evaluation of its implementation using example applications. In line with the initiation perspective of DSR, these four example applications demonstrate the range of potential practical applications available to future researchers and practitioners. Unlike previous studies that have presented general discussions of a broad range of cybercrime; our study has focused primarily on CaaS and crime ware from an RAT perspective. We have also proposed sets of definitions for different types of CaaS (phishing, brute force attack, DDoS attack, and spamming, crypting, and VPN services) and crime ware (drive-by download, botnets, exploits, ransomware, rootkits, Trojans, crypters, and proxies) based on definitions taken from both the academic and business practice literature. Based on these, we have built an RAT-based classification model. This study emphasizes the importance of RAT for investigating the cybercrime underground, so these RAT-based definitions are critically important parts of our framework. In addition, unlike prior research that discussed the cybercrime underground economy without attempting to analyze the data, we have analyzed large-scale datasets obtained from the underground community. Looking at the CaaS and crimeware trends, our results show that the prevalence of botnets (attack-related crimeware) and VPNs (preventive measures, related to CaaS) has increased in 2017. This indicates that attackers consider both

the preventive measures taken by organizations and their vulnerabilities. The most common potential target organizations are technology companies (28%), followed by content (22%), finance (20%), e-commerce (12%), and telecommunication (10%) companies. This indicates that a wide variety of companies in a range of industries are becoming potential targets for attackers, having become more vulnerable due to their greater reliance on technology.