Detecting Cyber-Adversarial Videos in Traditional Social media

Bingyan Du
Dept. of Information Systems
Arizona State University
Tempe, AZ
bingyand@asu.edu

Pranay Singhal
Dept. of Information Systems
Arizona State University
Tempe, AZ
psingha6@asu.edu

Victor Benjamin
Dept. of Information Systems
Arizona State University
Tempe, AZ
Victor.Benjamin@asu.edu

Weifeng Li
Dept. of Information Systems
University of Georgia
Athens, GA
Weifeng.Li@uga.edu

***Abstract*— Cyber-threat intelligence (CTI) has matured and grown into its own industry within recent years. Many CTI efforts involve scrutinizing text-based conversations in DarkNet forums and markets. However, hackers commonly share knowledge and other information through video formats that have been largely ignored. Further, cybercriminals are increasingly making use of mainstream social media to transmit hacking knowledge and assets, but this has gone unexplored in literature. In this research-in-progress, a video classifier to detect cybercriminal content in mainstream social media is designed and implemented. A collection of hacking and non-hacking videos was retrieved from a popular social media website to serve as a testbed. Feature sets included video metadata as well as features engineered from the videos themselves, including object detection and aesthetic qualities. This study demonstrates a methodological proof-of-concept that can enable future research that further investigates cyber-adversarial video contents, which have remained largely unexplored to this day. This study also contributes to literature regarding cyber-adversarial contents in mainstream social media.**

Keywords*— *Cybersecurity, DarkNet, Video Analytics

I. Introduction

In recent years, cyber threat intelligence (CTI) has been established and matured as an industry. A multitude of CTI companies are actively identifying and analyzing cyber threats from various sources to generate operational intelligence for their client firms to prevent and prepare for potential attacks. One major intelligence source is the DarkNet where hackers form communities to exchange stolen data, malicious tools and services, and hacking knowledge. For example, Recorded Future monitors DarkNet Marketplaces to identify emerging cyber threats for their client organizations. Trend Micro investigates the DarkNet to develop threat intelligence reports.

The DarkNet CTI community has been primarily focusing on analyzing the texts posted by hackers, such as forum discussions and malicious product listings, and the social network structure among hackers. Descriptions of malicious product listings on DarkNet Marketplaces are studied to identify emerging cyber threats [Reza 2018]. Customer reviews of malicious product sellers are analyzed to identify key hackers [Li 2014]. Jargons in online hacker language are examined to better understand hacker communication [Benjamin 2015].

Nonetheless, as an emerging form of hacker communication, hacking videos have been rarely analyzed and studied. Hackers are increasingly using videos to transmit hacking information and spread recruitment for hacktivist campaigns. For example, the Anonymous hacktivist group posted videos to recruit members of Operation Green Rights [Benjamin 2019]. The video format affords hackers to communicate procedures and ideas more effectively than text-based communication and are particularly useful for hacking tutorials and soliciting hacking group members. Therefore, analyzing hacking videos could enrich the existing CTI by introducing the hacking content (e.g., attack vectors and hacking campaigns) that is not available in the text-based hacker community content and social networks.

Moreover, many hacking videos are not only disseminated in the dark web but also accessible from the surface web. For example, hackers have been extensively posting hacking tutorials of their malicious tools on YouTube, allowing for hackers worldwide to access the hacking content. The increasing presence of hacking videos transmitted through traditional social media on the surface web and the broader audience these videos can now reach has a profound impact on the cybersecurity landscape. As such, there is an urgent need for hacking video analytics to facilitate research inquiries into understanding the role of hacking videos and developing CTI from these videos.

In this paper, we propose a deep learning-based approach for identifying hacking videos from traditional social media. Specifically, our proposed approach builds upon the state-of-the-art speech recognition, optical character recognition, and image quality assessment techniques to extract useful representations of hacking content and leverages a deep neural network to identify hacking videos from non-hacking related videos based on these representations. In our preliminary experiment, our proposed approach was evaluated using several modeling approaches and achieved an F1 score of 0.8724 with the best performing model. Our proposed approach enables a variety of research opportunities in hacking video analytics.

II. Literature Review

A. Cyber Threat Intelligence

Cyber threat intelligence (CTI) concerns identifying, collecting, and analyzing data about threats to inform the prevention and preparation of potential attacks [Benjamin 2019]. CTI developed from the DarkNet data is of particular value because hackers extensively use the DarkNet for discussing attack vectors, exchanging malicious hacker assets, soliciting hacker group members, and disseminating hacking knowledge. Prior research has primarily focused on two types of data: text and social network structure. The majority of hacker communication is based on texts, including descriptions of malicious products in DarkNet Marketplaces (DNMs), discussions in hacker forums and Internet Relay Chats, and key

attributes of stolen data in carding shops [Benjamin 2015]. Extensive research has been conducted on textual data from the DarkNet. For example, Ebrahimi et al. leverage DNM listing descriptions to detect cyber threats. Benjamin et al. study the lexical semantics of hacker jargons to better understand hacker language. Li et al. examine customer reviews of malicious product listings to identify key hackers in data breaches. Moreover, the social network structure of hacker communities has enabled researchers to better understand the social dynamics of hacker communities and develop useful CTI. For example, Samtani et al. identify key hackers for key logging tools by analyzing the social network structure [Samtani 2016].

An emerging type of data in the DarkNet is videos. Facilitated by videos, hackers are able to communicate complex procedures and ideas with much higher efficiency than traditional text-based communication. In particular, videos are extensively utilized for demonstrating hacking tools, presenting hacking skills, and soliciting hacker group members. These videos contain much detail that is not available in textual data. Figure 1 shows a screen shot of a video for soliciting members of 'Operation Green Rights', a hacktivist campaign with environmentalist motives. The video sets forth an argument to justify the attacks and encourages individuals to participate. Notably, the video was posted on YouTube; many hacking videos as such are easily accessible from the surface web, which enables these videos to address a much larger audience. Such videos could help recruit hackers into the greater hacker community and expedite their learning. While the analysis of hacking videos could facilitate important research, little CTI research has studied hacking videos.

Fig. 1. 'Operation Green Rights' Hacktivism Recruiting Video

B. Video Analytics

Computational analysis of video content is a rapidly growing field. Some techniques seek to capture image data from video frames. Deeper video content analyses generally include object recognition from video frames, such as detecting humans [Abu-El-Haija 2016].

For example, the Neural Image Assessment (NIMA) pre-trained model uses established object recognition algorithms to learn from human annotations regarding the aesthetics of various images [Talebi 2018]. Some aesthetic features include image blurriness/clarity, brightness, and attractiveness. NIMA is then able to evaluate new, previously unseen images and provide 'aesthetic scores' that closely resemble human perception on new images. Overall video analytics is an increasingly complex topic with new advancements being achieved rapidly.

Along with image analyses, it is also common to analyze audio streams embedded within video files. Literature describes several analyses. One common analysis includes speech detection, where classifiers are built to detect the presence of spoken words. Speech is a particularly interesting characteristic in cyber-adversarial content as hackers generally attempt to maximize their anonymity, and thus may not want to imprint their hacking identity with their real-world voice.

III. Research Design

A first step to hacking video analytics is the identification of hacking videos on the internet. Due to their profound impact on the internet beyond the DarkNet, hacking videos accessible to the surface web are of significant research value. However, identifying hacking videos from the surface web is nontrivial because there are many non-hacking videos containing security-related content, many of which are posted for cybersecurity education purposes. Manual efforts to search for hacking-related videos may be fruitful, but this approach would yield many false positives and is also not scalable across time. Motivated by this challenge and the great potential of hacking video analytics, we propose a computational system capable of detecting hacking videos from the surface web.

Utilizing previous Darknet forum collections retrieved by procedures outlined in Benjamin et al., 2019, some simple text analyses were performed (e.g., frequency counts, tf-idf) to identify a subset of terms appearing specific to hacker jargon and are not generalizable to other contexts. These keywords served as a seed list of search terms to identify the first subset of hacking-related videos. After identifying this initial subset of videos, the author of the videos and contents that appear in them (e.g., tool names, mentions of hacking communities) were used to further enhance the search. Thus a snowball collection procedure was performed. Further, along with hacking-related videos, a subset of non-hacking videos was also developed. This was done by identifying videos targeting security professionals and 'white hack' hackers, as these videos were still related to security but contained legitimate content not designed for cybercriminals. The data collection effort resulted in 2,340 videos (1,170 samples of hacking videos and 1,170 videos of non-hacking videos).

After data collection of hacking and non-hacking related videos was completed, several methods were taken to extract features from the videos. Some features consisted of metadata, including the "likes/dislikes" users submitted for a video, how many views the video received, the age of the video, how many comments the video generated, and the duration of the video. In this study we treat these features as generalized metadata, but in future analyses it would be interesting to develop deeper insight as to why some videos receive higher levels of user interaction (e.g., comments, likes) than other videos.

After metadata was extracted, features extracted from the video files themselves, specifically, two visual features and one audio feature. The visual features consist of NIMA scores and of human object detection (binary feature). The logic behind including these two features is that videos geared towards

security professionals or for legitimate purposes are more likely to have higher-quality production value and feature a human within the video. Conversely, cyber-adversarial videos are typically produced by individuals wishing to maintain anonymity, and so such videos may lack the presence of a human visible on-screen and may also suffer from lower aesthetics/production quality.

The audio feature engineered for this study is speech detection; that is, the presence of whether a human speaks or not during the duration of a given video (binary feature). This feature was included for similar reasons as the human object detection; hackers who wish to remain anonymous are less likely to include their voice in videos promoting hacking content when compared against legitimate cybersecurity videos. After feature engineering, several classification experiments were performed to classify between cyber-adversarial videos and non-adversarial videos.

IV. Preliminary Results

Several classification models were selected for trial in this study, including (1) logistic regression, (2) decision tree, (3) support vector machine, (4) linear discriminant analysis, (5) quadratic discriminant analysis, (6), random forest, (7) k-nearest neighbors, and (8) naïve bayes. Further, several variations of our feature set were experimented with to observe which variation produced the best classification results. These feature set variations include: (V1) all of the metadata: likes, dislikes, view, age of the video, human object, speech, video duration, number of comments; without NIMA score; (V2) removed collinearity between some metadata features by removing the 'likes' feature; (V3) all of the metadata, including the NIMA score, but no features engineered from the videos themselves; (V4) includes human object detection, NIMA scores, and view counts, but no other metadata features in order to assess how much video-centric features may influence video popularity; (V5) contains only human object and speech to investigate whether human presence is a reliable predictor of DarkNet/hacking-related videos in the security context; (V6) video duration as many hacking videos are of tutorial nature, and thus much longer than other forms of cybersecurity video content; (V7) video age and popularity as DarkNet/hacking-related videos tend to get reported and purged overtime thus leaving only legitimate videos to age. Classification results for detecting cyber-adversarial videos from non-hacking videos is displayed in Table I for all 8 models. The displayed results are the average of all feature sets described, and Random Forest is the best performer. The results for each specific feature set variation using Random Forest can be found in Table II.

TABLE I. CLASSIFICATION RESULTS USING DIFFERENT MODELS

Model	Precision	Recall	F1_score	AUC_ROC
Random Forest	0.8761	0.8718	0.8716	0.9451
Decision Tree	0.8358	0.8240	0.8227	0.8240
K-Nearest Neighbors	0.6737	0.6712	0.6701	0.7398
Linear Discriminant Analysis	0.6118	0.6048	0.5986	0.6415
Logistic Regression	0.5888	0.5841	0.5792	0.6165
Quad. Disc. Analysis	0.5781	0.5383	0.4456	0.6739
Bayes	0.5544	0.5276	0.4246	0.6363
Support Vector Machine	0.550142	0.545901	0.477072	0.602626

TABLE II. FEATURE VARIATIONS USING RANDOM FOREST

Variations	Precision	Recall	F1_score	AUC_ROC
V3	0.8802	0.8722	0.8724	0.9523
V1	0.8767	0.8716	0.8712	0.9477
V2	0.8726	0.8700	0.8696	0.9421
V7	0.8250	0.8179	0.8162	0.9044
V4	0.8359	0.8193	0.8206	0.8975
V6	0.6906	0.6840	0.6804	0.722989
V5	0.6257	0.5993	0.576965	0.626894

V. Conclusion

The CTI industry is continuing to grow and incorporate new data sources as evidenced by academic literature involving text analyses of DarkNet communities, as well as an increasing number of products/services offered by businesses to perform DarkNet monitoring. However, one common missing element among these previous efforts is analysis of hacking-related videos. Further, previous emphasis has been placed on the DarkNet while ignoring cybercriminal activity in traditional social media. This research is a step towards developing better understanding of cyber-adversarial behaviors on traditional social media, particularly within video formats. Future research will continue exploring this space by developing more robust video classification models, and also by analyzing what types of hacking videos become the most widely disseminated and drive community engagement among cyber-adversaries.

Acknowledgment

This research was supported in part by NSF CNS-1936370.

References

W. Li and H. Chen, "Identifying Top Sellers in Underground Economy Using Deep Learning-based Sentiment Analysis," ISI 2014, Proceedings of 2014 IEEE International Conference on Intelligence and Security Informatics, The Netherlands, September 2014.

V. Benjamin, J. Valacich, and H. Chen. DICE-E A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics. MIS Quarterly. 43(1). 2019.

V. Benjamin, W. Li, T. Holt, and H. Chen. Exploring Threats and Vulnerabilities in Hacker Web: Forums, IRC and Carding Shops. ISI 2015, Proceedings of 2015 IEEE International Conference on Intelligence and Security Informatics, Baltimore, Md., May 27-29, 2015.

S. Samtani, and H. Chen, "Using Social Network Analysis to Identify Key Hackers for Keylogging Tools in Hacker Forums," ISI 2016, Proceedings of 2016 IEEE International Conference on Intelligence and Security Informatics, Tucson, Arizona, September 2016.

M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, "Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach: An Exploratory Study," Proceedings of 2018 IEEE International Conference on Intelligence and Security Informatics (IEEE ISI 2018), Miami, Florida, November 2018.

S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv:1609.08675*. 2016.

H. Talebi, and P. Milanfar. NIMA: Neural Image Assessment. IEEE Transactions on Image Processing, 27(8), 3998-4011. 2018