IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 22, 2020, accepted October 29, 2020, date of publication November 9, 2020, date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037015

The World of Defacers: Looking Through the Lens of Their Activities on Twitter

ÇAĞRI BURAK ASLAN[1,2], SHUJUN LI[3, 4], (Senior Member, IEEE), FATIH V. ÇELEBI[1], AND HAO TIAN[5]

[1]STM Defense Technologies Engineering and Trade Inc., 06530 Ankara, Turkey
[2]Computer Engineering Department, Ankara Yıldırım Beyazıt University, 06010 Ankara, Turkey
[3]School of Computing, University of Kent, Canterbury CT2 7NZ, U.K.
[4]Kent Interdisciplinary Research Centre in Cyber Security (KirCCS), University of Kent, Canterbury CT2 7NZ, U.K.
[5]School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding authors: Çağri Burak Aslan (cagriaslan@gmail.com) and Shujun Li (hooklee@gmail.com)

This work was supported in part by the Scientific and Technological Research Council of Turkey [Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TUBITAK)] through the Bursary Scheme under Grant 2211-C, in part by the Research Project "PRIvacy-aware personal data management and Value Enhancement for Leisure Travellers" (PriVELT) funded by the Engineering and Physical Sciences Research Council (EPSRC), U.K., under Grant EP/R033749/1, and in part by the National Key Research and Development Program of China under Grant 2017YFB0802704.

ABSTRACT Many web-based attacks have been studied to understand how web hackers behave, but web site defacement attacks (malicious content manipulations of victim web sites) and defacers' behaviors have received less attention from researchers. This paper fills this research gap via a computational data-driven analysis of a public database of defacers and defacement attacks and activities of 96 selected defacers who were active on Twitter. We conducted a comprehensive analysis of the data: an analysis of a friendship graph with 10,360 nodes, an analysis on how sentiments of defacers related to attack patterns, and a topical modelling based analysis to study what defacers discussed publicly on Twitter. Our analysis revealed a number of key findings: a modular and hierarchical clustering method can help discover interesting sub-communities of defacers; sentiment analysis can help categorize behaviors of defacers in terms of attack patterns; and topic modelling revealed some focus topics (politics, country-specific topics, and technical discussions) among defacers on Twitter and also geographic links of defacers sharing similar topics. We believe that these findings are useful for a better understanding of defacers' behaviors, which could help design and development of better solutions for detecting defacers and even preventing impeding defacement attacks.

INDEX TERMS Cyber attacks, defacers, defacement, graph-based analysis, hacking, hackers, online social networks, natural language processing, NLP, OSN, sentiment analysis, social media, topic modeling, Twitter.

I. INTRODUCTION

Cybercrime such as hacking activities of cyber criminals have been causing a significant amount of damage to their victims (organizations and people) [1], and such threats are becoming more and more advanced and severe in recent years [2]. A lot of research has been conducted for a better understanding, detection and prevention of cybercrime and behaviors of cyber criminals and victims. Some studies on cybercrime have a technical focus, but others are more social science research in fields such as criminology, psychology, and economics.

Most technical research work focuses on automatic detection and prevention of cyber attacks and attackers (e.g., intrusion detection [3], anti-malware techniques [4], and anti-phishing solutions [5]), and forensic analysis of cyber security incidents to facilitate investigation of cyber attacks [6]. On the other hand, studies in social sciences mostly focuses on understanding motivations and modus operandi of cyber attackers, legal contexts of cyber attacks, or estimate social effects of cyber attacks to victims and the society at large. For instance, Gandhi *et al.* proposed to categorize cyber attacks into four groups based based on motivations of human attackers [7]: politically motivated attacks, socio-cultural conflict triggered attacks, economically motivated attacks, and espionage related attacks. Another example

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Shen.

is Yang *et al.*'s work [8], which studied social relationships in hacker communities to understand how they acted on online social media, using Twitter as an example platform.

We have seen few studies bridging social and technical aspects of cyber attacks using computational methods to study activities, behaviors and organizational aspects of hacking communities, often using data from open and dark web. Among such research, much has been done by analyzing data from online social networks (OSNs) such as Twitter and underground forums [8]–[15].

Although there has been a lot of research conducted on different cyber attacks, one common type of cyber attack has received relatively less attention from researchers – web site defacement attacks. This type of attacks are among the most common web attacks and frequently reported by the media. For instance, just to give one such example reported recently, a hacking group, Ghost Squad Hackers,[1] defaced the website of European Space Agency (ESA) twice, on 14 and 19 July 2020 [16], [17]. Particularly, to the best of our knowledge, there have no previous attempts on analyzing defacers' behaviors based on OSN data. See Section II for a more detailed literature review on related work on defacement attacks and defacers.

This paper reports our efforts in filling the gap of OSN-based behavioral analysis of defacers to enhance the existing body of knowledge about defacement attacks and defacers, which could help design and development of better solutions for detecting defacers and even preventing impeding defacement attacks.

Our work is based on a computational data-driven analysis of a public database of defacers and defacement attacks and activities of 96 selected defacers who were active on Twitter. Our analysis revealed a number of key findings: a modular and hierarchical clustering method can help discover interesting sub-communities of defacers; sentiment analysis can help categorize behaviors of defacers in terms of attack patterns; and topic modelling revealed some focus topics (politics, country-specific topics, and technical discussions) among defacers on Twitter and also geographic links of defacers sharing similar topics.

The rest of the paper is organized as follows. Related work about defacers and defacement attacks is summarized in Section II. We explain our main research questions (RQs) and the method used to study the RQs in Section III, and the data we used in Section IV. Our analysis methods and the results are elaborated in Section V. Lastly, the paper is concluded by Section VI, with some future work discussed.

II. RELATED WORK

Defacement attacks have been frequently reported, and there have been a number of public archiving and mirroring web sites of actual defacement attacks. One of the most well-known web sites of this kind is Zone-H,[2] which has been widely used in studies of defacement attacks and defacers. Mirror-H[3] is another similar web site, and according to Alexa it was the website with the highest overlap of audience with Zone-H.[4] Yet another such web site is H@CK M1RROR,[5] but this web site has not been updated since 2017, so the data become quite out-dated. Such defacement archiving and mirroring websites mostly depend on self-reported attacks by defacers themselves.

Automatic detection of web site defacement attacks have been studied by many researchers, typically using machine learning techniques [18]–[22]. Most of such studies are based on textual analysis of the web site contents, but there has been some work exploring image-based analysis using computer vision techniques without prior knowledge of the target web site's original content [23].

A lot of socio-technical work in this area focuses on understanding the motivations of defacers. For instance, Woo *et al.* used the content of defaced web sites to infer psychological motivations of defacers [24]. They used several psychological and social theories and concluded with two groups: pranksters and militants. They classified 70% of the defacement activities they examined as done by pranksters. They also argued that although a majority of defacers were pranksters, their defacement activities were not harmless. In addition, they reported that unlike the general belief about hackers often being "sole wolves", they seemed to have existed in a social community of hackers.

In 2017 Romagna *et al.* studied the connection between hacktivism and defacement attacks [25], where they investigated the reasons (or motivations) behind defacement attacks. For that, they collected data from Zone-H for 12 months. Interviews with some of the defacers from that period were also conducted. They noticed that Zone-H data included some tags referring to the reasons of the recorded defacement attacks, and analyzed such tags statistically. Most of the defacement attacks were tagged using "Heh…just for fun". Romagna *et al.* observed that hacktivists had generally displayed socio-political contents on the defaced web sites. They argued that most hacktivists had been motivated by socio-political matters, often linked with regional or international tensions. Another conclusion drawn from their study is that defacers sought attention and therefore used archiving web sites like Zone-H to make their activities public.

Romagna *et al.*'s work depended on tags on Zone-H to analyze the motivations of defacers. This approach was challenged by Maggi *et al.* in 2018 [26]. They argued that it is untrustworthy to rely on such tags because they can be forged easily. Following a different approach, Maggi *et al.* studied defacement campaigns by examining the contents of the actual defaced web sites. They stated that using actual content of defaced web sites is more reliable than using metadata only. They used latent Dirichlet allocation (LDA) [27] as

[1] This particular hacking group is captured in our later analysis.
[2] http://www.zone-h.org/
[3] http://www.Mirror-H.org/
[4] https://www.alexa.com/siteinfo/Zone-H.org
[5] http://www.hack-mirror.com/

a topical modeling tool to study the shift of defacement topics over time. They conducted both dynamic (in web browser) and static (on hard disk) analysis on the defaced web sites to extract features, which were then normalization, clustered and visualized for studying different defacement campaigns over time. They studied both lone defacers and hackers who cooperated with each other. One interesting observation they reported is that the use of Twitter as a contact information was on the rise, whereas e-mail usage was decreasing.

In 2019 Howell *et al.* [28] studied hackers' valuation of potential targets, particularly focusing on how cyber attacks targeted different nations. They tried to understand why hackers chose their targets and if it has something to do with the guardianship level of the victim. They based their work on the routine activity theory, which introduces three aspects for victimization: convergence of motivated offenders, suitable targets, and the lack of adequate guardianship. Based on information collected from various sources including Zone-H (for defacement attacks), statistics about 114 nations and their development levels, they reported that there is a strong connection between the level of a nation's overall guardianship and the rate of attacks targeting that nation: more capable nations received less attacks.

Surprisingly, although a lot of work has been done on using OSN data to analyze activities and behaviors of hackers, cyber criminals and other types of cyber attackers [8]–[15], [29], [30], there has been very limited research on using OSN data to study defacement attacks and defacers. The only work we are aware of was done by Maimon *et al.* in 2017 [31]. They used data from Zone-H and 187 accounts on multiple OSN platforms including Twitter, Facebook and YouTube to study if defacement attacks and defacers' activities on those OSNs are related. They found that the active use of OSNs was correlated with a high frequency of defacement attacks. They also reported that if an attacker used both Twitter and Facebook this increased defacement attacks on non-USA targets while Twitter-only account usage yielded a high number of attacks against USA-based web sites. This study however was based on a relatively small dataset on Zone-H (May-July 2017) and only some limited aspect of defacers' activities on OSNs were studied. Our work reported in this paper is more comprehensive in both breadth and depth.

III. RESEARCH QUESTIONS AND METHOD

Considering the lack of past studies investigating connections between defacement attacks and activities of defacers on OSNs, we decided to conduct a more comprehensive analysis of defacers' activities on Twitter and how they are connected with actual defacement attacks. We chose to focus on Twitter because it was the OSN platform increasingly used by defacers as reported by Maggi *et al.*'s research in 2018 [26].

This work's overall aim is to find out how defacers' activities on Twitter can help us better understand behaviors of defacers. We set three separate RQs for our study:

- RQ1: Can we study the social structure of defacers based on the data on Twitter?
- RQ2: Can we connect publicly visible sentiment of defacers on Twitter with defacement attacks they launched?
- RQ3: What topics did defacers talk about on Twitter and how are they related to the motivation of defacement attacks?

To study all RQs, we needed to identify a number defacers who were / are active on Twitter. We explain how such defacers and their Twitter accounts were identified in the next section. After identifying selected defacers' Twitter accounts, we proceeded to collect data about their activities on Twitter and also defacement attacks attributed to them.

Based on the data we collected, we studied RQ1 using the friendship graphs of selected defacers' Twitter accounts, and applied unsupervised clustering methods to identify sub-communities within those defacers and their friends. Social structures were then studied around such sub-communities. For RQ2, we applied natural language processing (NLP) based sentiment analysis of defacers' activities on Twitter, and compared the results with the actual attacks launched, which led to the discovery of different sentiment-attack patterns. Such patterns can be explained by different motivations of defacers. For RQ3, we used the NLP based topical modeling method LDA to analyze discussions of defacers in order to understand what topics they were discussing, focusing on common topics shared among different sub-communities of defacers.

IV. DATA USED

To study the research questions listed in the previous section, we first collected defacement attacks and their details such as dates of occurrence and the defacers responsible for such attacks. Such data are needed to identify defacers and their Twitter accounts. Once we had Twitter accounts of some defacers, we collected those accounts' profiles and timelines, and constructed their friendship graphs, for further analysis. In the following, we explain the two different sets of data used in greater details.

A. DEFACEMENT DATA

We used the defacement archiving web sites Zone-H and Mirror-H to collect data needed for our analysis. The reason for choosing Zone-H is its popularity among the hacker community and researchers who have studied defacement attacks and defacers [18], [21], [23]–[26], [28], [31]. Mirror-H was used to increase the size and diversity of our data. We decided not to use H@CK M1RROR since the data on that website is mostly out-dated.

These sources were used in two different ways. Firstly, we crawled defacement attacks (including the mirror web pages of the defaced web sites) from both web sites to collect names of the likely responsible defacers (notifier in Zone-H, attacker in Mirror-H). For Mirror-H and Zone-H we used different ways to crawl the data. For Mirror-H, the data was collected crawled by a cyber security firm. They crawled the ''Archive'' page of Mirror-H[6] from January to July in 2019. They kindly shared their data for academic purposes. For Zone-H, we could not find a collaborative firm so we crawled the website ourselves by collecting defacement attacks on the ''Special Defacement Archive'' page of the website[7] between December 2018 to March 2019. As neither web sites required user registration to post defacement attacks, there were a lot of unusable hacker names in our initial collection. A cleaning process was therefore applied to the initial collection by eliminating some obviously unreliable names such as too long entries, entries with many punctuation marks, entries with incremental names (e.g., WSO-01, WSO-02). At the end of the cleaning process, we obtained a list of around 2,002 defacers, which we considered more reliable.

Later on, after the extraction of defacers who existed on Twitter (see below), corresponding defacers' profiles on Zone-H and Mirror-H were collected. We only collected their timelines and crawling took place between 15th of August to 15th of October 2019. The timelines were collected up to the 3,200 tweets limit set by the Twitter API, going back from the crawling date. This data was used to analyze the attack patterns of those defacers and to compare them with defacers' sentiments on Twitter (for RQ2).

B. TWITTER DATA

After having a more sensible list of defacers, we moved on to check if these defacers were active on Twitter using the Twitter API. Several approaches were attempted to check if Twitter accounts used by those defacers existed.

The first one we attempted is using the GET users/search Twitter API to search for possible accounts that may be related to each defacer in our list. This Twitter API returns up to 20 accounts per search. This approach has a downside on the large number of candidate accounts that may have to be checked manually. This means that we may have to manually check up to nearly 40,000 Twitter accounts. After evaluating the amount of manual work needed and potential human errors that can occur in the manual process, we decided that such a manual process is not ideal so we skipped this method.

Next, we simply assumed that some defacers used the same name on the defacement archiving web sites and on Twitter, i.e., the names in our list are the corresponding Twitter handlers. By this approach 557 Twitter accounts were detected and 56 of them were protected accounts. The remaining 449 accounts were subjected to timeline filtering using keywords (hacker, zone-h, mirror-h, hack, deface, defaced). The resulting 87 Twitter accounts were checked manually to see if they indeed correspond to a defacer on the defacement archiving web sites. This was done by looking for references to original links to defacement pages in timelines. This method's main drawback is that it can miss some defacers' Twitter accounts if they use different names on the defacement indexing web sites and on Twitter.

To further extend our list of defacers on Twitter we applied a third method. On defacement indexing web sites there are several information items about a specific defacement:

- Date
- Defacer (the notifier, may not be the attacker)
- Country
- URL (of the defaced web page)
- IP
- Mirror of the defaced state of URL

We looked for Twitter links in mirrors of the defaced web sites. Indexing services takes a snapshot of the defaced state as an HTML file. We leveraged these HTML files to look for Twitter links. Then we checked the Twitter handles to see if they were actually related to one or more defacers from the defacement indexing web sites. Similar methods have been reported in the literature, e.g., for extracting defacers' contact information (e-mail or Twitter) from defaced web sites [32].

After combining results from the second and the third methods, we obtained in total 100 Twitter accounts we believed to belong to defacers. After that, we crawled the following information for each of the 100 Twitter accounts for further analysis: timelines, account information, friends (other Twitter accounts they followed). We also translated all tweets to English using the Translate API from Yandex.[8] This API was used because most other alternative solutions we found were using Google's previous API which was no longer valid. Google had switched to a paid service, which we decided to avoid using because the cost will be too high and the performance of Yandex API is sufficiently good.

V. RESULTS

A. GRAPH-BASED SOCIAL STRUCTURE ANALYSIS

Among the 100 Twitter accounts of defacers, 4 did not have any friends, so for this analysis we looked at the remaining 96 accounts only. An initial list of the 96 accounts' friends on Twitter was crawled, where ''friends of a Twitter account A'' means other Twitter accounts A follows. We chose only friends but not followers of the defacers' accounts because one cannot control who would follow himself/herself. In addition, we also recorded friendship relationships between friends of all defacers, including friendships among nodes who do not share the same defacer(s) as friends. The reason of including the friendships of defacers' friends is to allow a better coverage of the social structure of the defacers community. At the end, we obtained a directed friendship graph with 10,360 nodes (Twitter accounts) and 1,188,360 edges (friendships).

The resulting graph is visualized in Fig. 1.[9] While the graph shows some clear visual patterns, it is too large to allow extraction of more concrete semantic insights.

[6] https://mirror-h.org/archive

[7] http://www.zone-h.org/archive/special=1

[8] https://tech.yandex.com/translate/

[9] For all figures representing graphs, the ForceAtlas2 layout algorithm [33] was used with the LinLog mode, Dissuade Hubs and Prevent Overlap selections activated.

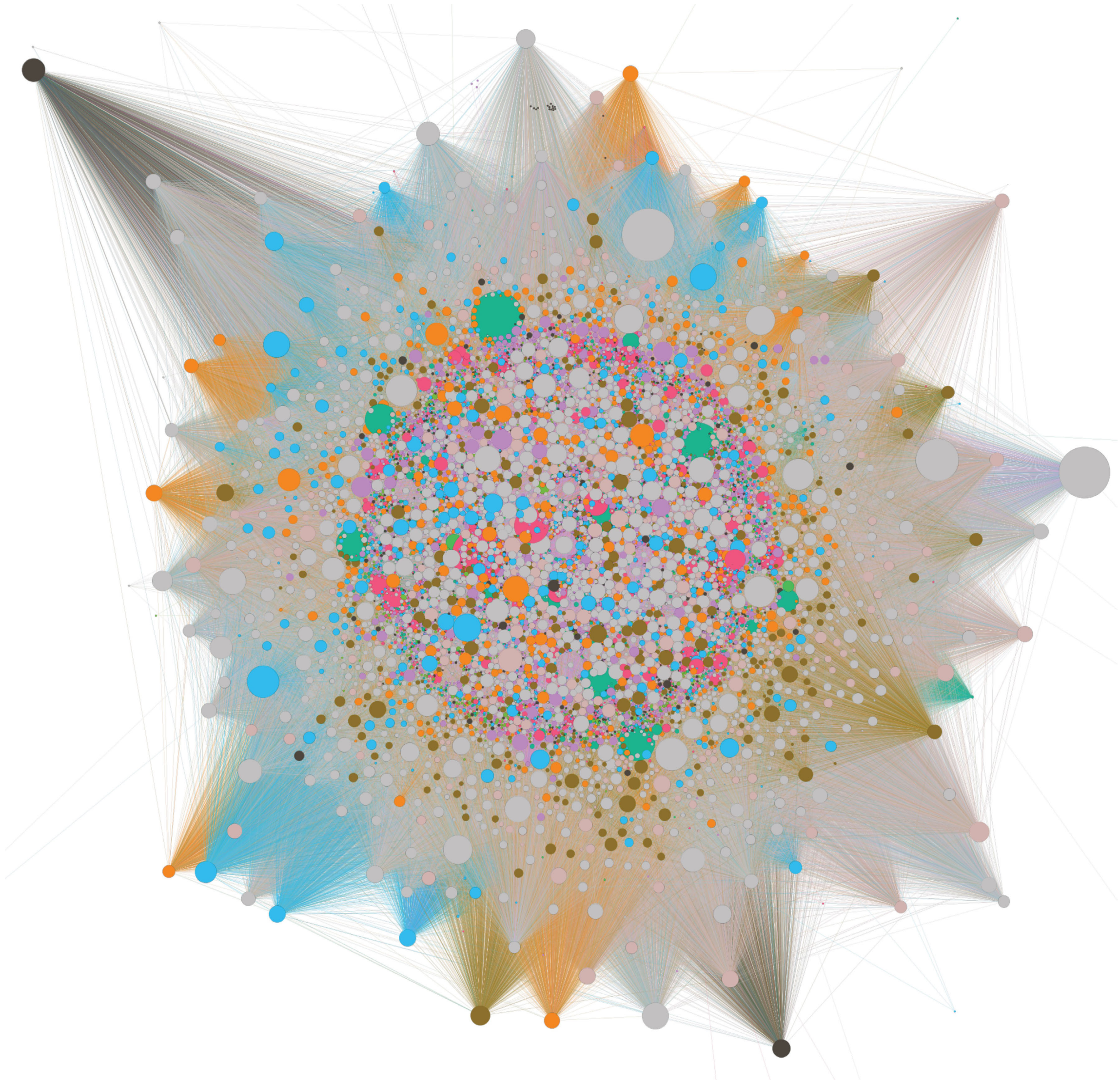


FIGURE 1. **The whole graph with 10,360 nodes. Node names are not shown since the graph is too large.**

To further analyze the graph, we need to focus on smaller sub-graphs, each of which representing a sub-community. To this end, we decided to apply unsupervised clustering methods because we did not have ground truth labels to train a supervised classifier for detecting sub-communities. Several clustering algorithms were tested, including DBSCAN [34], Girvan-Newman clustering [35], Leiden algorithm [36], and the modularity score optimization based clustering [37]. After we got the clustering results, we used a prominent Turkish hacking group, Turk Hack Team (THT), and manually identified Twitter accounts of this group as a validation dataset, in order to estimate the performance of all the clustering methods in terms of clustering related hackers together. This led to the discovery that the modularity optimization based algorithm worked the best because all the other clustering methods failed to cluster most THT related accounts into a single cluster. To further validate the modularity optimization based algorithm indeed performed well, we looked at other clusters and identified a new hacking group, Ghost Squad Hack (GSH), as a separate cluster. This provides further evidence that the modularity optimization based algorithm worked well. We acknowledge that this performance evaluation method is not ideal, but given the lack of ground truth knowledge of defacers and their Twitter accounts, there was not an alternative approach we could use.

The first application of the modularity optimization based clustering algorithm resulted with 46 clusters (subgraphs). Showing all the 46 clusters and reporting the analysis results are impossible given the space limit, so we use two representative example sub-graphs to illustrate how the modularity based clustering helped us gain some useful insights about social structures of defacers. The first selected subgraph includes some of the defacers from the initial list we obtained from Zone-H and Mirror-H. It has 412 nodes and 4,713 edges,

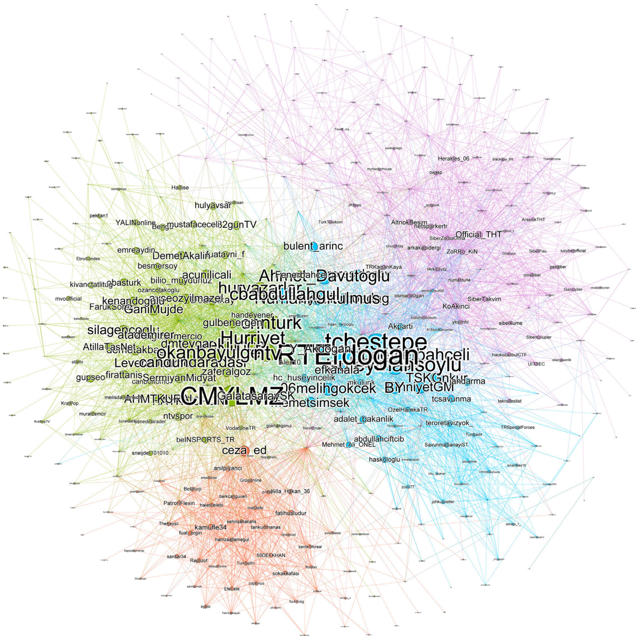


FIGURE 2. A Turkish defacers community extracted from the initial graph shown in Fig. 1.

as shown in Fig. 2. Examining Twitter accounts belonging to this subgraph, we observed that this is a sub-community whose members are mostly Turkish-speaking. We further applied the same clustering method on this subgraph in order to explore finer structures within this cluster, and got 5 sub-clusters (sub-subgraphs). At this second level we started to see more semantically interesting information about how this cluster is formed: except one obviously uninteresting sub-cluster with just 3 nodes, the other four sub-clusters represent four different sub-communities (colored differently in Fig. 2):

- Celebrities (green sub-cluster)
- Cyber security community (purple sub-cluster)
- Rapper community (orange sub-cluster)
- Politics community (blue sub-cluster)

Since we are more interested in defacers, we decided to apply a third level of clustering to the cyber security sub-subgraph, which we expected to include more defacers. This sub-subgraph has 148 nodes and 1,085 edges (see Fig. 3). After applying the same modularity optimization based clustering method, we obtained a finer structure of this sub-community, splitting into different groups of cyber security related people and organisations. For instance, the blue cluster in Fig. 3 includes members of the Turkish group of hackers THT. Another example is the orange cluster, which includes mostly cyber security professionals and organizations.

Similarly, the second selected subgraph is about another hacking group GSH. After applying two levels of clustering like we did for the first selected subgraph, hackers who claimed to have been hacking on behalf of GSH are grouped together, as shown in Fig. 4. The cluster colored green includes six accounts attributed to GSH (based on the corresponding nodes' Twitter profiles). Interestingly, most accounts in this subgraph are not in our original list of defacers, implying that the graph analysis could be used to discover unknown defacers and hackers. The fact that THT and GSH are in two different subgraphs in the first level of clustering indicates that the clustering-based graph analysis has the potential to separate different hacking groups and identifying members of hacking groups.

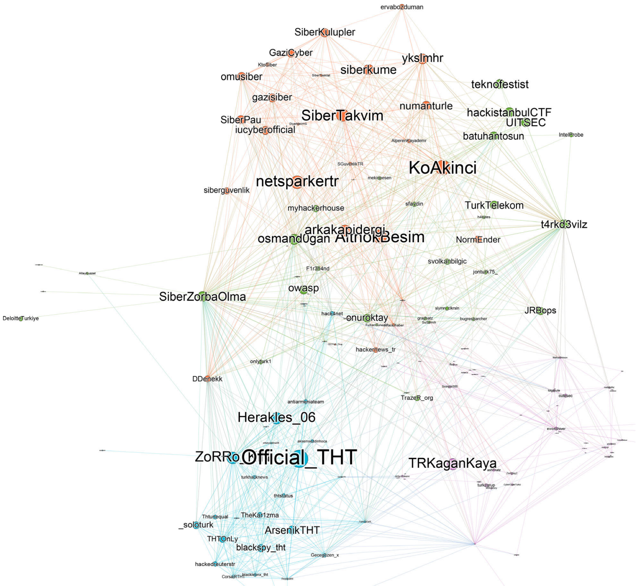


FIGURE 3. A Turkish cyber community extracted from the graph shown in Fig. 2.

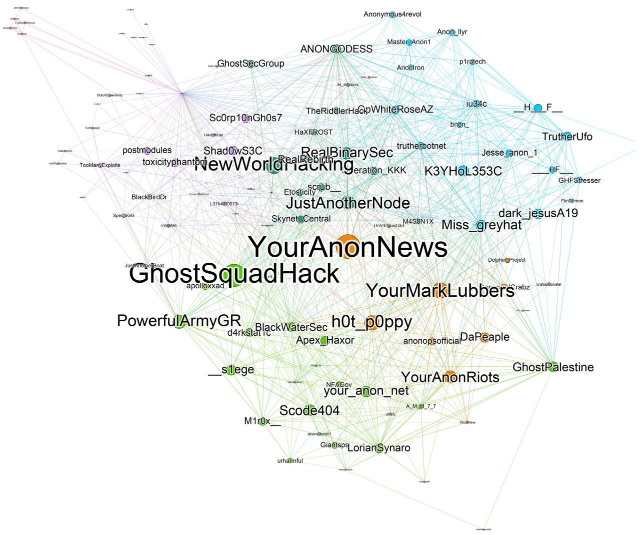


FIGURE 4. The Ghost Squad Hack (GSH) community extracted from the initial graph shown in Fig. 1.

The analysis described above shows that a hierarchical application of the modularity optimization based clustering method is useful to reveal social structures of the network of defacers and their friends on Twitter. This is not surprising since social structures are often hierarchical, therefore a one-layered clustering cannot work well. This also suggests that a similar approach can be used to group like-minded people and organizations on OSNs and other online platforms. It deserves noting that the clustering-based graph analysis relies on active involvement of a human analyst in the process, therefore following the human-machine teaming paradigm, which is necessary for such complicated tasks that neither humans or machines alone can do well. In addition to the clustering method, graph visualization is another automated tool helping human analysts to conduct the graph analysis. As a whole, we can see the answer to RQ1 is positive: graph analysis using OSN data does help us study social structure of defacers.

B. SENTIMENT AND ATTACK FREQUENCY ANALYSIS

In this part we investigate the correlation between a defacer's sentiment and his/her attack frequency. For this purpose, we crawled the timelines of Twitter accounts of the 100 defacers identified. Crawling results showed that 5 of them did not have any tweets at the time of the crawling, leaving 95 Twitter accounts and timelines for further analysis. The crawled tweets were analyzed using TextBlob,[10] a popular NLP library written in Python on top of another widely used NLP toolkit NLTK.[11] The sentiment analysis algorithm in TextBlob was run on each tweet. Before running the algorithm all of the tweets are translated to English using the Yandex API. This sentiment analysis implementation uses a naive Bayes analyzer trained on a movie reviews dataset. This may not give the best results and a more specific training process may be more suitable, but after manually evaluating the results on a subset of tweets we considered it good enough for our purpose here. For each defacer, the sentiment scores for all tweets were then grouped together to form an average value for each day. Then such daily values and the defacer's daily attack frequencies obtained from Zone-H and Mirror-H were compared and the overlapping time periods were then extracted. As a consequence, for each day in those time periods we have a sentiment score and an attack frequency, which can be checked for possible correlation. Since the sentiment scores are between -1 and 1, to facilitate comparison and visual presentation, we normalized attack frequencies so that the maximum frequency per defacer (across the whole timeline of defacement attacks attributed to each defacer) is mapped to 1.

After getting the daily sentiment scores and attack frequencies for each defacer, we examined how they are correlated manually in a grouped bar chart. For some defacers, the overlap between the sentiment scores and the attack frequencies is insufficient to make any conclusive judgments.

Out of the 95 defacers, 46 were considered to have a sufficient level of overlap. For half (23) of the 46 defacers, some level of correlation between the sentiment scores and the attack frequencies were observed (see Fig. 5 for six representative examples). The correlation typically has a lagging effect, i.e., the sentiment appeared shortly (one or a few days) before or after the actual defacement attack(s), and it can be either negative or positive. A sentiment status appearing before an attack reflects the motivation (e.g., being angry about some political news, which drives a defacer to launch an attack) and one afterwards corresponds to the consequence (e.g., feeling happier or less angry after launching a successful attack). Combining the temporal pattern, the positivity and the strength of the correlation, we can infer useful information about motivation and attitude of defacers. What is the most interesting is that such a correlation pattern can be repeatedly observed for some defacers (e.g., Defacers #2 and #3 in Fig. 5), implying that it would be possible to monitor their Twitter accounts to predict potential future attacks.

Although the sentiment-attack correlation analysis does not cover all defacers, the fact that a significant portion of defacers (24%) demonstrated such a correlation indicates that the analysis is a useful tool for some defacers. For other defacers, the major reason of the lack of observation of such a correlation is due to insufficient data. If we can collect more data about them from other platforms such as underground forums/chat rooms and hacker-oriented OSN and instance messaging groups, it may be possible to discover a similar correlation.

C. TOPICAL ANALYSIS

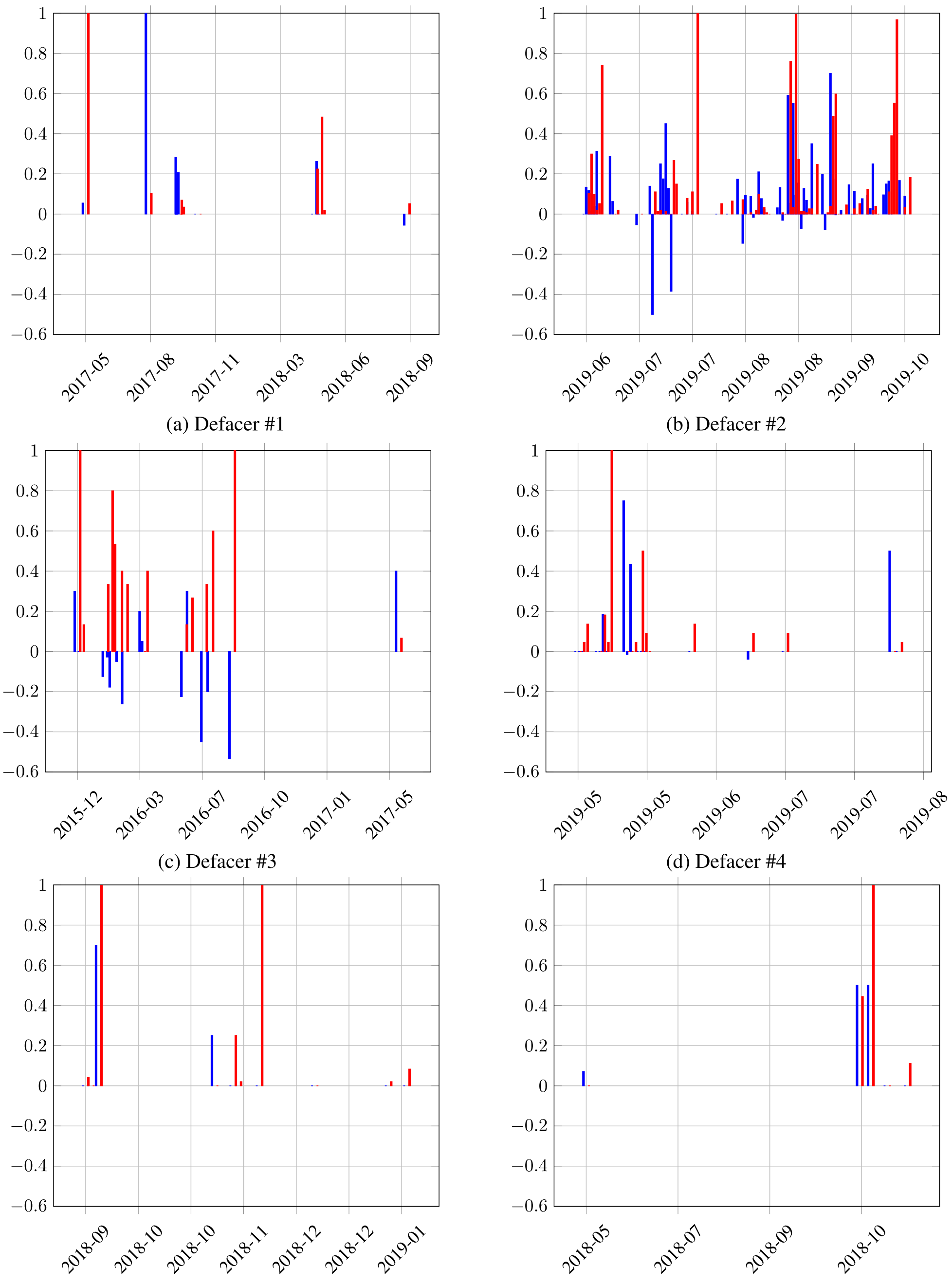
Motivation behind defacement attacks can potentially be revealed by topics defacers discuss with each other or a defacer talks with others. For instance, if a defacer's main motivation of launching defacement attacks is about to demonstrate his/her disagreement with some political movement, then it is likely he/she will talk about this specific political movement on Twitter.

To analyze what topics defacers were talking about on Twitter, we applied the LDA-based topic modeling algorithm [27] to the corpus of timelines of all the 96 defacers. A topical modeling algorithm like LDA sees a document as a bag of words and assumes that k topics (which is a set of words) spread across all m documents in a corpus. In our context, a document is a timeline of a defacer's Twitter account (so $m = 100$), and k is a parameter we need to set manually as the input of the LDA algorithm. The main output of the algorithm is a document to topic matrix, which shows what topics each document includes. The LDA is an iterative algorithm that requires another parameter r to be set: the maximum number of iterations.

The topic number k was set to be 10 after several attempts with different candidate values. For the parameter r, we varied its value from 100 to 2,000 (incremented by 100), and 200 turned out to be the best value. The best values of k and r were determined by examining the output topics and words in these topics and comparing them with manual topical modeling analysis results of some selected defacers.

[10] https://textblob.readthedocs.io/en/dev/

[11] https://www.nltk.org/



(a) Defacer #1

(b) Defacer #2

(c) Defacer #3

(d) Defacer #4

(e) Defacer #5

(f) Defacer #6

FIGURE 5. Some examples of comparison between the sentiment score and the attack frequency. Legend: blue bars – sentiment score, red bars – attack frequency.

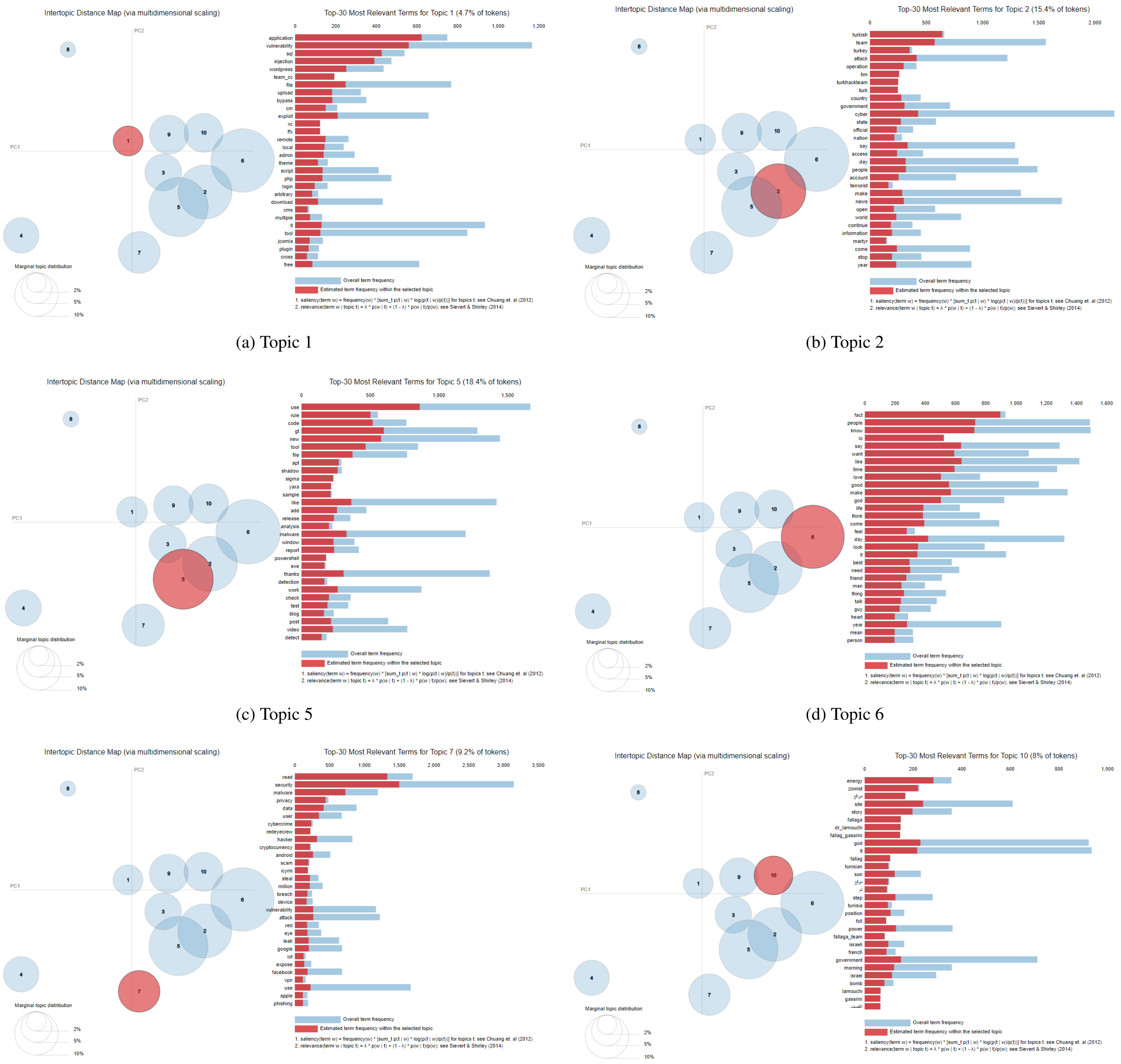
The LDA implementation in scikit-learn[12] was used in this experiment because of its wide usage in the machine learning community.

The timelines were first preprocessed following a pipeline including the steps below, and after that the results were feed to the LDA algorithm:

- All tweets are translated to English using the Yandex API in order to homogenize the dataset.
- All URLs are removed because they can negatively affect the topic analysis algorithm
- All words are converted to lowercase.
- The `tokenize` algorithm from gensim[13] is applied because of its ease of use. We also removed all punctuation marks because they are irrelevant in term-based topical modelling.
- Stopwords are eliminated using NLTK's stopwords list.
- The lemmatization algorithm in TextBlob is applied.

[12] https://scikit-learn.org/stable/index.html

[13] https://radimrehurek.com/gensim/



(a) Topic 1

(b) Topic 2

(c) Topic 5

(d) Topic 6

(e) Topic 7

(f) Topic 10

FIGURE 6. Example topics and words belonging to the topics.

When examining the topical memberships of defacers, we noticed that it is necessary to focus on top $k' \leq k$ topics with a high probability. We observed that in general we do not need to look at all k topics especially those with a small probability, and that for some defacers using more than one topics (e.g., the top 2 ones) can give a lot more useful information than using just one. Two example are vurkacteam and MahkuM_BeY, two hackers from Turkey with mild political intentions. Their most dominating topic is one of the general (i.e., less distinctive) topics (6c) but their second dominating topic reflects their interest in Turkish politics (number 2). Details about Topic 6 can be seen in Fig. 6b. Several examples of the topics and their terms can be seen in Fig. 6.

Looking at the topics generated, three topical themes are clearly visible: politics, country related, and technical discussions. In total there are 5 politics-related topics. One of them is about anti-Zionist movement whereas another one is about Turkish politics. One such topic is about anti-government discussions, and another is mainly following the hacker group Anonymous on political matters. A third one includes "opsudan" keyword, which is the hashtag for the Sudan operation started by Anonymous, so a sub-topic about Anonymous.

There are also topics relevant for specific countries. We spotted 6 of them and some overlap with political topics. Interestingly, defacers sharing similar scores of the dominating topics tend to be from the same country. Five countries standout according to the topics generated by LDA: Turkey, Brazil, Argentina, Iran, Tunisia. All the political topics also include country-related words, which can be interpreted by the fact that people from the same country tend to have shared interests in political matters of their own country.

The last topical theme include discussions on technologies. One 6e is about general cyber security terms, and some example words are: ''malware'', ''phishing'', ''breach'' and ''privacy''. The other (Topic 1) has a web application security flavor with example words such as ''joomla'', ''wordpress'', ''cms'', ''injection'', ''plugin'', ''sql''. Details about Topic 1 can be seen at Fig. 6a.

There are also two topics whose semantic meaning cannot be easily determined. Example words from these topics include ''people'', ''love'', ''heart'', ''like'', ''make'', ''time'', ''feel'', ''know'', ''friend'', ''person''. The words in these two topics seem to have a flavor of showing negative and positive feelings. They may be signs of emotions that lead to defacement attacks or the result of successful attacks. In Section V-B, we showed that different sentiment feelings deduced from text may be related to attack frequency.

Another interesting finding is that the term ''team'' is mostly used in country-related topics. This may indicate that hacker groups are generally formed by people from the same country. In total 9 out of the 10 topics include ''team'' as a topical word.

The results shown above demonstrate that topic modeling is a useful technique for analyzing defacers' topical interests on Twitter, which can provide useful insights on understanding their motivations of launching defacement attacks and may even provide clues for predicting impeding future attacks. The analysis can be extended to other textual data describing defacers' discussions online, e.g., on other OSNs or underground forums. It is also possible to use the same technique to study victims of defacement attacks and other types of hackers.

D. SUMMARY

In this section, we briefly summarize all the results we reported above, echoing the three research questions listed in Section III.

The answer to the first one is positive. Our results demonstrated that graph analysis based on OSN data is a good method to gain a richer insight into defacers' social structures online. Particularly, our results revealed that it may be possible to detect unknown hacker groups or previously unknown members of a known hacker group using the OSN graph analysis.

The answer to the second research questions is also positive. Our results showed that at least for some defacers their online sentiment status are indeed correlated with the frequency of actual defacement attacks launched, which offers a possible approach to generate early warnings or even pre-attack alerts.

The last research question is about topics defacers discussed on Twitter so that we can achieve a better understanding on their motivation. Our results demonstrated that topical modelling is a useful technique to study defacers' topical interests and politics emerged as one of main topical themes (which was not a surprise since many reported defacement attacks were driven by political tension between nations).

While this paper focuses on defacers' activities on Twitter, we believe some of the results are generalizable to other types of hackers and some could be used to study victims of cyber attacks, too.

VI. CONCLUSION AND FUTURE WORK

This paper reports the first OSN-based behavioral analysis of defacers using their activities on Twitter and data on defacement attack archiving websites. A range of computational data analytic techniques were used to conduct the analysis, including graph-based analysis of defacers and their friends on Twitter, NLP-based sentiment analysis and correlation analysis between defacers' sentiment statuses and their attack frequencies, and NLP-based topical modelling analysis of defacers' discussions on Twitter. The results proved that such computational analysis can help reveal important insights about defacers' motivations and behavioral patterns, and even providing hints for predicting future defacement attacks (e.g., by monitoring known defacers' sentiment status on Twitter) and unknown defacers (e.g., via clustering analysis of known defacers' networks of friends).

The work reported in this paper can be improved in a number of ways, which will be our future work. The number of defacers studied in this paper could be further increased by looking at more data from more defacement archiving websites and online attack monitoring services. We can also extend the OSN platform covered from Twitter only to other platforms including Facebook, underground online forums and chat rooms, hacker-oriented OSN and instance messaging groups. With more data, we could investigate more factors that can influence defacers' behaviors to gain deeper and broader insights, e.g., to reveal correlation between sentiment scores and attack frequencies for more defacers. In addition, we will explore a more automated method for the sentiment-attack correlation analysis, e.g., using a mathematical definition of the correlation and a supervised machine learning algorithm to automatic profile defacers and predict future defacement attacks from a specific defacer or a specific group of defacers. More advanced sentiment analysis and topical modelling techniques could be tried to achieve a better understanding of defacers' behaviors. Our work focused more on English textual analysis by translating non-English tweets to English using an automated translation service. Using NLP tools that can directly process non-English texts will help reduce errors introduced by automated translation and cover semantic meanings that otherwise will be lost during the translation process. In addition, some empirical studies such as interviews and surveys of defacers could also help validate some of the findings in the papers, in order to provide some level of "ground truth".

REFERENCES

[1] C. M. M. Reep-van den Bergh and M. Junger, "Victims of cybercrime in europe: A review of victim surveys," *Crime Sci.*, vol. 7, no. 1, p. 5, Dec. 2018.

[2] A. Zaharia. (2020). *300+ Terrifying Cybercrime and Cybersecurity Statistics & Trends*. [Online]. Available: https://www.comparitech.com/vpn/cybersecurity-cyber-crime-statistics-facts-trends/

[3] A.-S. K. Pathan, Ed., *The State of the Art in Intrusion Prevention and Detection*. Boca Raton, FL, USA: CRC Press, 2014.

[4] J. Saxe and H. Sanders, Eds., *Malware Data Science: Attack Detection and Attribution*. San Francisco, CA, USA: No Starch Press, 2018.

[5] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Syst. Appl.*, vol. 106, pp. 1–20, Sep. 2018.

[6] R. C. Joshi and E. S. Pilli, Eds., *Fundamentals of Network Forensics: A Research Perspective*. London, U.K.: Springer, 2016.

[7] R. Gandhi, A. Sharma, W. Mahoney, W. Sousan, Q. Zhu, and P. Laplante, "Dimensions of cyber-attacks: Cultural, social, economic, and political," *IEEE Technol. Soc. Mag.*, vol. 30, no. 1, pp. 28–38, Spring 2011.

[8] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 71–80.

[9] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2010, pp. 435–442.

[10] R. Y. K. Lau, Y. Xia, and Y. Ye, "A probabilistic generative model for mining cybercriminal networks from online social media," *IEEE Comput. Intell. Mag.*, vol. 9, no. 1, pp. 31–43, Feb. 2014.

[11] B. Wang, A. Zubiaga, M. Liakata, and R. Procter, "Making the most of tweet-inherent features for social spam detection on Twitter," in *Proc. 5th Workshop Making Sense Microposts (CEUR-WS)*, 2015, pp. 10–16. [Online]. Available: http://ceur-ws.org/Vol-1395/paper_07.pdf

[12] R. Aswani, A. K. Kar, and P. Vigneswara Ilavarasan, "Detection of spammers in Twitter marketing: A hybrid approach using social media analytics and bio inspired computing," *Inf. Syst. Frontiers*, vol. 20, no. 3, pp. 515–530, Jun. 2018.

[13] L. Wu, X. Hu, F. Morstatter, and H. Liu, "Adaptive spammer detection with sparse group modeling," in *Proc. 11th Int. Conf. Web Social Media*. Palo Alto, CA, USA: AAAI Press, 2017, pp. 319–326. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15636

[14] J. Grisham, S. Samtani, M. Patton, and H. Chen, "Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 13–18.

[15] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 1049–1057.

[16] P. Paganini. *Exclusive, Ghost Squad Hackers Defaced European Space Agency (ESA) Site*. Accessed: 2020. [Online]. Available: https://securityaffairs.co/wordpress/105918/hacktivism/european-space-agency-esa-site-defacement.html

[17] P. Paganini. *Ghost Squad Hackers Defaced a Second European Space Agency (ESA) Site in a Week*. Accessed: 2020. [Online]. Available: https://securityaffairs.co/wordpress/106111/hacking/esa-site-defaced-again.html

[18] F. Bergadano, F. Carretto, F. Cogno, and D. Ragno, "Defacement detection with passive adversaries," *Algorithms*, vol. 12, no. 8, p. 150, Jul. 2019.

[19] A. Bartoli, G. Davanzo, and E. Medvet, "A framework for large-scale detection of Web site defacements," *ACM Trans. Internet Technol.*, vol. 10, no. 3, pp. 1–37, Oct. 2010.

[20] S. Wu, X. Tong, W. Wang, G. Xin, B. Wang, and Q. Zhou, "Website defacements detection based on support vector machine classification method," in *Proc. Int. Conf. Comput. Data Eng. (ICCDE)*, 2018, pp. 62–66.

[21] X. D. Hoang, "A website defacement detection method based on machine learning techniques," in *Proc. 9th Int. Symp. Inf. Commun. Technol. (SoICT)* in Lecture Notes in Networks and Systems, vol. 63, 2018, pp. 116–124.

[22] X. D. Hoang and N. T. Nguyen, "Detecting website defacements based on machine learning techniques and attack signatures," *Computers*, vol. 8, no. 2, p. 35, May 2019.

[23] K. Borgolte, C. Kruegel, and G. Vigna, "Meerkat: Detecting website defacements through image-based object recognition," in *Proc. 24th USENIX Secur. Symp.*, 2015, pp. 595–610. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/borgolte

[24] H.-J. Woo, Y. Kim, and J. Dominick, "Hackers: Militants or merry pranksters? A content analysis of defaced Web pages," *Media Psychol.*, vol. 6, no. 1, pp. 63–82, Feb. 2004.

[25] M. Romagna and N. J. van den Hout, "Hacktivism and Website defacement: Motivations, capabilities and potential threats," in *Proc. 27th Virus Bull. Int. Conf.*, 2017, pp. 1–10. [Online]. Available: https://www.researchgate.net/publication/320330579_Hacktivism_and_Website_Defacement_Motivations_Capabilities_and_Potential_Threats

[26] F. Maggi, M. Balduzzi, R. Flores, L. Gu, and V. Ciancaglini, "Investigating Web defacement campaigns at large," in *Proc. Asia Conf. Comput. Commun. Secur. (ASIACCS)*, 2018, pp. 443–456.

[27] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://www.jmlr.org/papers/v3/blei03a.html

[28] C. J. Howell, G. W. Burruss, D. Maimon, and S. Sahani, "Website defacement and routine activities: Considering the importance of hackers' valuations of potential targets," *J. Crime Justice*, vol. 42, no. 5, pp. 536–550, Oct. 2019.

[29] K. Jones, J. R. C. Nurse, and S. Li, "Behind the mask: A computational study of Anonymous' presence on Twitter," in *Proc. 14th Int. Conf. Web Social Media*. Palo Alto, CA, USA: AAAI Press, 2020, pp. 327–338.

[30] Ç. B. Aslan, R. B. Sağlam, and S. Li, "Automatic detection of cyber security related accounts on online social networks: Twitter as an example," in *Proc. 9th Int. Conf. Social Media Soc.*, Jul. 2018, pp. 236–240.

[31] D. Maimon, A. Fukuda, S. Hinton, O. Babko-Malaya, and R. Cathey, "On the relevance of social media platforms in predicting the volume and patterns of Web defacement attacks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 4668–4673.

[32] M. Balduzzi, R. Flores, L. Gu, F. Maggi, V. Ciancaglini, R. Reyes, and A. Urano, "A deep dive into defacement: How geopolitical events trigger Web attacks," Trend Micro Forward-Looking Threat Research (FTR) Team, Tech. Rep., 2018. [Online]. Available: https://documents.trendmicro.com/assets/white_papers/wp-a-deep-dive-into-defacement.pdf

[33] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software," *PLoS ONE*, vol. 9, no. 6, pp. 1–12, 2014.

[34] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (Still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Aug. 2017.

[35] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[36] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to Leiden: Guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, no. 1, pp. 5233:1–5233:12, Dec. 2019.

[37] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008.

ÇAĞRI BURAK ASLAN received the bachelor's degree in electrical and electronics engineering from TOBB ETU, in 2010, and the master's degree in computer engineering from Ankara Yıldırım Beyazıt University, in 2015, where he is currently pursuing the Ph.D. degree in computer engineering with the focus on cyber security. He worked for eight years at Ankara Yıldırım Beyazıt University. He currently works with STM Defense Technologies Engineering and Trade Inc. He also visited the Kent Interdisciplinary Research Centre in Cyber Security (KirCCS), University of Kent, U.K., for a part of his Ph.D. research, in 2018. His research interests mainly include cyber threat intelligence, social media usage in cyber security, machine learning applications, and data mining.

SHUJUN LI (Senior Member, IEEE) received the B.E. degree in information science and engineering and the Ph.D. degree in information and communication engineering from the Xi'an Jiaotong University, China, in 1997 and 2003, respectively. Since November 2017, he has been a Professor of Cyber Security with the University of Kent, U.K., leading the university wide Kent Interdisciplinary Research Centre in Cyber Security (KirCCS), a U.K. Government recognized Academic Centre of Excellence in Cyber Security Research (ACE-CSR). He has published over 100 scientific articles with two best paper awards, and his research interests mainly include interfaces between cyber security, human–computer interface, multimedia computing, digital forensics, and cybercrime.

FATIH V. ÇELEBI received the B.Sc. degree from Middle East Technical University, in 1988, the M.Sc. degree from Gaziantep University, in 1996, and the Ph.D. degree from Erciyes University, in 2002, all in electrical-electronics engineering. He is currently the Dean of the Faculty of Engineering and Applied Sciences, Ankara Yıldırım Beyazıt University (AYBU). He has published so many scientific articles, and his current research interests include cyber security, artificial intelligence, machine learning, and optoelectronics.

HAO TIAN received the bachelor's degree in information security from Shanghai Jiao Tong University, China, in 2018. He is currently pursuing the master's degree in electronic and information engineering with the School of Cyberspace Security, Shanghai Jiao Tong University. His research interests include applied cryptography, performance optimization of hashing algorithms, and password security.