Follow the blue bird: A study on threat data published on Twitter*

Fernando Alves[1], Ambrose Andongabo[2], Ilir Gashi[2],
Pedro M. Ferreira[1], and Alysson Bessani[1]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] Centre for Software Reliability, City, University of London, UK

Abstract. Open Source Intelligence (OSINT) has taken the interest of cybersecurity practitioners due to its completeness and timeliness. In particular, Twitter has proven to be a discussion hub regarding the latest vulnerabilities and exploits. In this paper, we present a study comparing vulnerability databases between themselves and against Twitter. Although there is evidence of OSINT advantages, no methodological studies have addressed the quality and benefits of the sources available. We compare the publishing dates of more than nine-thousand vulnerabilities in the sources considered. We show that NVD is not the most timely or the most complete vulnerability database, that Twitter provides timely and impactful security alerts, that using diverse OSINT sources provides better completeness and timeliness of vulnerabilities, and provide insights on how to capture cybersecurity-relevant tweets.

Keywords: OSINT · Twitter · Vulnerabilities

1 Introduction

Cybersecurity has remained a hot research topic due to the increased number of vulnerabilities indexed and to the severe damage caused by recent attacks, from ransomware (*e.g.*, wannacry) to SCADA systems attacks (*e.g.*, the attacks on the Ukrainian power stations). A growing trend for obtaining cybersecurity news is to collect Open Source Intelligence (OSINT) from the Internet [57]. OSINT sources include vulnerability databases (*e.g.*, the National Vulnerability Database—NVD), online forums (*e.g.*, Reddit), social networks (*e.g.*, Twitter), and scientific literature. Although more technical, exploit databases (*e.g.*, ExploitDB) are a useful OSINT source providing code excerpts known as Proofs of Concept (PoC) that show how to exploit a vulnerability. PoCs can be analysed by a specialised audience capable of using the exploit's code to understand and counteract vulnerability exploitation, thereby removing the vulnerability.

The research community has shown many different uses for OSINT, from its collection and processing [26, 29, 36, 37, 40, 43, 48, 54, 59], vulnerability life

* Appears on the *Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS'20)*. September 2020.

cycle analysis [28, 31, 42, 50, 56], to evaluating vulnerability exploitability [24, 31, 32, 38, 45, 51]. There are two predominant OSINT sources in the literature: NVD (*e.g.*, [24, 31, 45, 51, 55]), and Twitter (*e.g.*, [26, 40, 41, 43, 59]). The first provides curated vulnerability data, while the latter is more generic, concise, and covers more topics.

As Twitter's usage grew, various information sources began to link their content on Twitter to increase visibility and attract attention. Twitter's continued growth placed it among the most relevant communication tools used by the vast majority of companies who have a Twitter account to interact with the world. All this activity also caught the attention of the research community. The information flow and interaction graphs meant new research opportunities, such as detecting emerging topics [34, 46], or finding events related to a specific topic, such as riots [25], patients experience with cancer treatment drugs [30], or earthquakes [52]. Twitter popularity instigated the development of tools to collect tweets (*e.g.*, Tweet Attacks Pro [20]), APIs for programming languages (*e.g.*, Tweepy [19]), and many OSINT-collecting tools developed specific plugins to collect tweets (*e.g.*, Elastic Stack [21]), including cybersecurity-oriented ones (*e.g.*, SpiderFoot [18]). The cybersecurity field also found opportunities in using Twitter (discussed in Section 2).

However, to the best of our knowledge, there is no evidence in the literature highlighting one security data source as advantageous over the others. For instance, the following questions are yet to be answered: Why use solely NVD when there are several reputable vulnerability databases? Is NVD the richest (in terms of number of vulnerabilities reported) and timeliest (does it contain the earliest reporting date of a vulnerability) vulnerability database? Why use Twitter to gather cybersecurity OSINT? Does Twitter provide any advantage over vulnerability databases? Is it useful for security practitioners?

In this paper, we present an extensive study on OSINT sources, comparing their timeliness and richness. We analysed the vulnerability OSINT sources indexed on vepRisk [27], which aggregates several vulnerability databases, advisory sites, and their relationships. We compared Twitter against these data sources to understand if there are any advantages in using it as a cybersecurity data source. To explore this topic, we formulated three research questions:

RQ1: Is NVD the richest and timeliest vulnerability database?

RQ2: Does Twitter provide a rich and timely vulnerability coverage?

RQ3: How are vulnerabilities discussed on Twitter?

Our findings show that: vulnerability databases complement one another in richness and timeliness (*i.e.*, no single source contains all the vulnerabilities; no single sources can be relied on as providing the earliest vulnerability reporting date); Twitter is a rich and timely vulnerability information source; and finally, Twitter complements other OSINT sources. In summary, our contributions are:

- A comparison between some of the most reputable and complete vulnerability databases in terms of timeliness and coverage;
- An analysis of the coverage and timeliness of Twitter with respect to vulnerability information;

– An analysis about "early alerts" on Twitter, *i.e.*, vulnerabilities disclosed or discussed on Twitter before their inclusion on vulnerability databases;
– An analysis on how vulnerabilities are discussed on Twitter;
– Insights on how to collect timely tweets;
– Insights regarding OSINT (and in particular Twitter's) usage for cybersecurity threat awareness.

2 Background and Related Work

The following sections present some vulnerability databases and previous research contributions related to this work.

2.1 Vulnerability Databases

MITRE Corporation [11] maintains the Common Vulnerabilities and Exposures list [3] (in short, CVE), a compilation of known vulnerabilities described in a standard format. A global index of known vulnerabilities simplifies complex analyses such as detecting advanced persistent threats. Therefore, indexing known vulnerabilities in CVE became standard practice for all kinds of security practitioners, including software vendors. Each CVE entry has an ID (CVE-ID), a short description, and the creation date.

NIST's National Vulnerability Database [12] mirrors and complements CVE entries on their database. Every hour, NVD contacts CVE to obtain newly disclosed vulnerabilities (we contacted NVD directly to get this information). Each vulnerability indexed in NVD undergoes a thorough analysis, including attributing an impact score based on the Common Vulnerability Scoring System (for both versions 2.0 [5] and 3.0 [4]), and links related to the vulnerability, such as advisory sites or technical discussions. NVD uses the CVE-ID in place of an ID of its own.

There is a significant difference between the dates of NVD and CVE entries. In CVE, it is the date when entries became *reserved*, but not yet *public*. NVD entries are always public, using the date when they were indexed, even prior to their analysis completion. Thus, in practice, a vulnerability has the same public disclosure date on both CVE and NVD, which is NVD creation date. Therefore, this study considers only the NVD vulnerability disclosure date.

Besides CVE and NVD, many online databases compile known vulnerabilities and provide unrestricted use of their contents, such as the Security Database [17] and PacketStorm [15]. The complementary information provided by each database differs, but in general, these provide a description, some analysis of the security issues raised by the vulnerabilities, known exploits, and possible fixes or mitigation actions.

2.2 Cybersecurity-Related OSINT Studies

To the best of our knowledge, Sauerwein *et al.* performed the most similar study to the one present on this paper [55]. For two years, the authors collected all

tweets with a CVE-ID in its text. They show a comparison of the tweet publishing dates with the disclosure dates of those CVEs on NVD. The results show that 6232 vulnerabilities (25.7% of their dataset) were discussed on Twitter before their inclusion on NVD. However, this study falls short in some aspects. Firstly, the NVD is not always the first database to report new vulnerabilities, which changes the vulnerabilities first confirmed report date (see Section 4). Secondly, the authors search only for CVE-IDs on Twitter, which will not capture issues that have been disclosed to the public but not (yet) indexed on CVE or NVD. Finally, the analysis is focused solely on the vulnerabilities life cycle and Twitter appearance, overlooking vulnerability characterisation such as their impact.

There is some research work providing evidence that relevant and timely cybersecurity data is available on Twitter [33, 41, 44], *i.e.*, that some vulnerabilities were published on Twitter before their inclusion on vulnerability databases. However, these are case studies concerning a single vulnerability, and compare the tweets referring them solely with NVD. Other Twitter-based contributions include correlating security alerts from tweets with terms found in dark web sources [53], studying the propagation of vulnerabilities on Twitter [58], and finding that exploits are published on Twitter (on average) two days before the corresponding vulnerability is included in NVD [51].

In a similar research line, Rodriguez *et al.* [49] analysed vulnerability publishing delays on NVD when comparing to other OSINT sources: Security Focus, ExploitDB, Cisco, Wireshark, and Microsoft advisories. The authors report that NVD presents publishing delays (from 1 to more than 300 days) from 33% to 100% of the cases when comparing with those databases, *i.e.*, sometimes it publishes after these databases or it always publishes after these databases. However, the authors consider only the year of 2017. Similarly, the Recorded Future company reports that for 75% of the vulnerabilities NVD presents a 7-day disclosure delay [16]. However, the company does not reveal how it obtained these results.

The literature lacks a systematic and thorough analysis regarding the data published on Twitter and on vulnerability databases, including crucial aspects such as coverage, timeliness, and the actionability provided by such OSINT.

3 Methodology

The objective of this study is to compare some aspects of the information present on vulnerability databases with another OSINT source, namely Twitter. Instead of searching, collecting, and parsing a set of databases, we use the vepRisk database [27]. It contains several types of security-related public data, including all entries published on NVD, Security Database, Security Focus, and Packet-Storm databases, from their creation until the end of 2018.

We chose Twitter as an OSINT source, as it is a known aggregator of content posted by all kinds of users (hackers, security analysts, researchers, etc.), news sites, and blogs, among others who tweet about their content to increase visibility [9]. Thus, Twitter became an information hub for almost any kind of content. Unlike vulnerability databases—that contain only security data—

Twitter includes discussions over a vast universe of topics. Since the results of this study are based on tweets mentioning indexed vulnerabilities, we decided to search for tweets mentioning the vulnerabilities indexed on NVD. Finally, to ensure the validity of our results, we opted to *manually* match tweets to vulnerabilities. These decisions raised two questions: 1) what part of the vulnerability description are we going to use as a search term? and 2) How to reduce the number of vulnerabilities to manually inspect?

The NVD description of some vulnerabilities includes a "colloquial" name for which the vulnerability became known. For example, CVE-2014-0160 is known as the "Heartbleed bug". These names fall mostly within two categories: a generic description of the vulnerability class (*e.g.*, "Microsoft Search Service Information Disclosure"), or some "creative" designation related to the vulnerability (*e.g.*, "Heartbleed" is a vulnerability on the "heartbeat" TLS packets which can be exploited to leak or "bleed" information). These colloquial names are easily recognisable since they always appear in the NVD vulnerability description after the "aka" acronym (for *also known as*). Therefore, to guide the search on Twitter, all vulnerabilities with a colloquial name were selected, and the names were used as query terms. This decision also reduced the number of vulnerabilities to analyse to 9,093, an amount of data manually processable. Additionally, vulnerabilities with colloquial names are more likely to be discussed on Twitter since most were "named" due to media attention. The IDs of the 9,093 vulnerabilities with a colloquial name that were used in this study are listed online [1].

We were unable to use the Twitter API to collect the tweets for the study as it only provides access to tweets published in the previous week. However, the Twitter web page allows searching for tweets published at any point in time. To automate the querying process, a library called `GetOldTweets` [7] was employed. It mimics a web browser performing queries on the Twitter page, enabling fast and programmatic retrieval of any number of tweets from any time.

Regarding matching tweets and vulnerabilities, we consider that a tweet t unequivocally refers a specific vulnerability v if and only if (1) t mentions in its text v's CVE-ID even if the vulnerability has not yet been disclosed on NVD, or (2) t contains a link mentioned in v's NVD description, even if the web page pointed by the link is currently down, or (3) t mentions a security advisory that is also referred by v's NVD links about that threat, or (4) t or t's links mention an ID associated with v. Two assumptions are made: 1) if an ID is present on a tweet, then the advisory has been published; and 2) a security analyst that receives a tweet containing a security advisory ID can search for this advisory, thus having the same result as publishing the advisory link on the tweet. If a vulnerability is mentioned by up to a thousand tweets, all tweets were manually inspected. The colloquial name of some vulnerabilities is also a word commonly used on tweets, such as "CRIME" (CVE-2012-4930) or "RESTLESS" (CVE-2018-12907). For those cases, where a search term can return more than 350,000 tweets, the manual inspection was done in two steps. First, the description is analysed to understand the vulnerability characteristics and related terminology. Then, a large set of informed searches were performed on the tweet set in search

of tweets potentially referring the vulnerability. In total, *about a million tweets were manually inspected*, and any links present in potential matches were also examined to confirm the matches. The data labelling was performed solely by a PhD student with a cybersecurity background, and it took roughly eight months to complete. All potential matches were triple checked to ensure their validity.

The time range considered in this study begins on March 2006 (Twitter's creation date) until the end of 2018. The tweets were collected between early 2017 and the end of 2019. The resulting dataset contains 3,461,098 tweets. The tweets publishing times were adjusted to the day time-scale to match the time granularity provided by the vulnerability databases. Therefore, all time comparisons performed in this study used the publishing day.

4 Vulnerability database comparison

As NVD is considered a standard for consulting vulnerability data, many research works use only it as their vulnerability database (*e.g.*, [42, 47, 50, 55]). This is a natural choice since NVD includes multiple resources for further understanding of the issue at hand. However, other reputable vulnerability databases, with their own disclosure procedures and timings, provide useful information for security practitioners. Therefore, it is interesting to understand if there is evidence that supports using only NVD for practice or research work. To investigate this point, we collected data about two different aspects: the number of entries and their publishing date. The first measures the coverage of the database, while the second is related to its timeliness and practical usefulness.

Table 1 shows the number of entries in each of vepRisk's databases: NVD, PacketStorm (PS), Security Database (SD), and Security Focus (SF). It also shows the number of entries shared between each database pair. Tables 2 and 3 are related to timeliness. Table 2 is divided in two blocks. The first shows the number of occurrences where one database was the *first* to disclose a vulnera-

Table 1: The number of entries in each database (in bold) and the number of shared entries between database pairs.

	NVD	**PS**	**SD**	**SF**
NVD	**110,353**	-	-	-
PS	9,290	**129,130**	-	-
SD	110,353	9,344	**117,098**	-
SF	60,378	8,597	60,843	**98,445**

Table 2: The number of times one database or a group of databases were the first to disclose a vulnerability.

Database(s)	**# Occurrences**	**%**
NVD	0	0.00
PS	853	0.77
SD	0	0.00
SF	40,208	36.44
NVD, SD	51,238	46.43
PS, SF	1,265	1.15
NVD, SD, SF	16,580	15.02
NVD, PS, SD	85	0.08
NVD, PS, SD, SF	124	0.11

bility ahead of other databases. The second block shows the number of occurrences where various groups of databases were *simultaneously first* to disclose a a vulnerability. Table 3 complements the previous table by showing the percentage of time each database was one of the first to disclose a vulnerability.

Table 3: The percentage of times each database was one of the first to disclose a vulnerability.

Database	# Occurrences	%
NVD	68,027	61.64
Security Database	68,027	61.64
Security Focus	58,177	52.72
PacketStorm	2,327	2.11

There are five key takeaways obtained from analysing the tables: (1) NVD is not the most complete vulnerability database, with the Security Database and PacketStorm containing more entries; (2) NVD is not the most timely database. Alone, it was never the first to publish a vulnerability; (3) No database stands out as the most timely; (4) Security Database contains all of NVD's entries (this was manually verified); (5) With the exception of NVD, all databases publish different vulnerabilities. Therefore it is important to follow a set of data sources instead of relying solely on one.

5 Twitter Vulnerability Coverage and Timeliness

Coverage. A first validation on using Twitter for cybersecurity is verifying if vulnerability data reaches Twitter. We searched for tweets mentioning each of the CVE-IDs published on NVD after Twitter's creation. Of the 94,398 CVE-IDs searched, 71,850 (76.11%) were mentioned in tweets. However, by analysing Fig. 1 it is possible to observe that since the beginning of 2010, CVEs became regularly discussed on Twitter. In fact, from 2010 forward, the coverage became above 97.5%, validating the hypothesis that vulnerability data reaches Twitter.

The drastic increase in tweets mentioning CVEs in 2010 may be connected to the sudden growth Twitter underwent in that period [10]. Nevertheless, the turning point on cybersecurity threat awareness and on the importance of coordinated vulnerability disclosure mechanisms *may* have been in the beginning of 2010, when Google publicly disclosed that their infrastructure in China was targeted by an advanced persistent threat codenamed "Operation Aurora" [14]. Later on, it was discovered that other major companies were targeted, such as Adobe Systems, Rackspace, Yahoo, and Symantec. This event *could* have triggered two crucial social phenomena: that companies are attacked and should not be ashamed of it, and should disclose the details of these attacks in a coordinated effort to detect, understand, and prevent them; and that the users prefer transparency in cybersecurity events since when data breaches occur, typically it is the user data that is affected.

Timeliness. Regarding timeliness, we performed the analysis only for the 9,093 vulnerabilities that were manually analysed to ensure the correctness of the results. Fig. 2 shows, for those vulnerabilities, which source discussed them first: either one of the vulnerability databases considered in this study, Twitter,

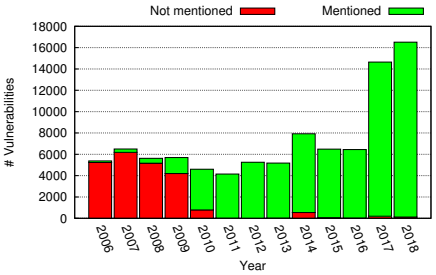

Fig. 1: Twitter's CVE coverage.

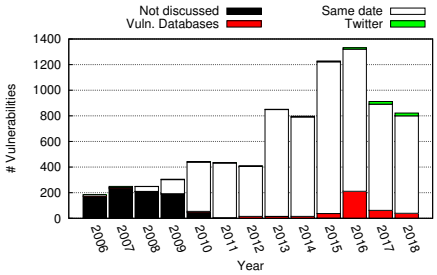

Fig. 2: A timeliness comparison between vulnerability databases and Twitter.

or Twitter and at least one of the databases, simultaneously. There are also the cases where the vulnerability was not discussed on Twitter, which are predominant before 2010. Although we are not evaluating the whole databases, the figure shows a predominance of same day publishing cases (84.56% when considering 2006–2018, and 93.73% in 2010–2018). We consider that these results validate the hypothesis that Twitter is a timely source of vulnerability data. In the next section we present an in-depth study of the cases where the vulnerabilities were discussed on Twitter ahead of vulnerability databases.

6 Early Vulnerability Alerts on Twitter

Of the 9,093 vulnerabilities analysed, 89 were referred by tweets before being published on at least one of the vulnerability databases considered. Even though these vulnerabilities represent a small percentage of the vulnerability sample under study (0.98%) we decided to characterize them to understand if searching for early alerts on Twitter is a worthy endeavour. The most mentioned vendors in the early alerts are the Ethereum blockchain (17 mentions), Microsoft products (5), Debian (5), Oracle (4), Linux (4), and Apple (4), while the most mentioned assets are Javascript (9), SSL/TLS (8), Xen Hypervisor (4), Safari Browser (3), Mercurial version control (3), and Cloud Foundry (3). These mentions provide evidence of the usefulness of these alerts, as both vendors and assets are some of the major players in their respective fields.

All vulnerabilities with Twitter early alerts can be found online [1], together with their publishing dates on the vulnerability databases and some extra notes. In the following sections these early alerts are further analysed on their impact and usefulness. We conclude the section with a discussion of the significance of these results.

6.1 Timeliness

Twitter versus vulnerability databases. Fig. 3 presents the distribution of early alerts over the years considered. The number of early alerts increased in

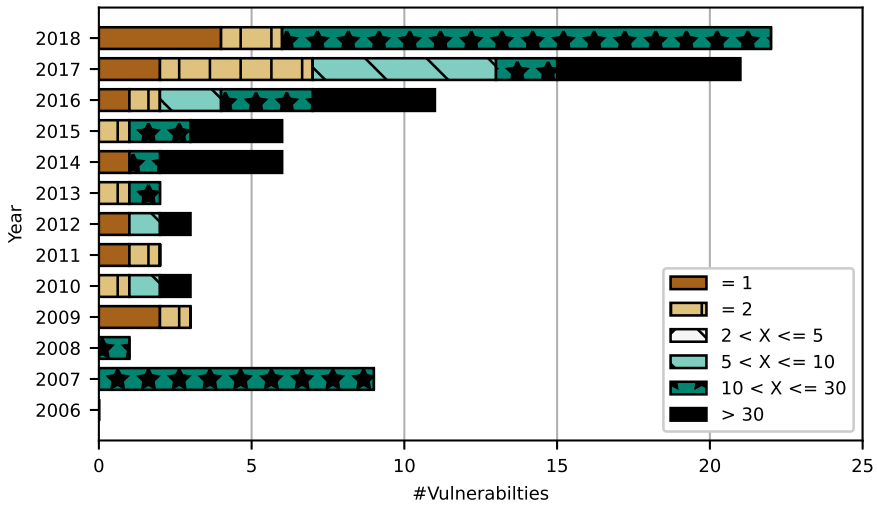

Fig. 3: The number of early alerts found per year.

the last two years, which matches the increase of vulnerabilities published on databases since the beginning of 2016.

Concerning publishing timing, the majority of early alerts were available up to thirty days ahead of vulnerability databases (78.65%–70 cases). Notably, four early alerts appeared between 31 and 50 days ahead, four between 51 and 100 days ahead, nine between 101 and 200 days ahead, and finally, the two cases with the highest antecedence were 371 and 528 days ahead. The number of days Twitter is ahead of vulnerability databases increased continuously since 2008, but only after 2016 we found more than 10 cases. No relevant patterns were discovered in the early alerts due to the small number of occurrences.

Twitter versus advisory sites. Besides vulnerability databases and social media, advisory sites are an essential source of vulnerability information. Many companies use websites to announce software patches, along with which vulnerabilities are fixed. Therefore, we compare Twitter with advisory sites as these are specialized OSINT sources directly connected to the software vendors.

We manually searched for advisory notices for each of the early alerts, obtaining only 33 advisories (approximately of early alerts). Table 4 presents the number of times either Twitter or the advisory was the first publisher, or when both published on the same day.

The majority of early alerts are not paired with an advisory, but the tweets referring them contain links that describe these vulnerabilities. This observation reinforces the idea that Twitter is a useful cybersecurity discussion hub by connecting various knowledge resources in a single place.

Table 4: The number of times Twitter, advisory site, or simultaneously both, were the first to publish an advisory notice.

1st publisher	#	%
Twitter	11	12.36
Same date	13	14.61
Advisory site	9	10.11
No Advisory	56	62.92

Table 5: The CVSS 2.0/3.0 impact of the early alert vulnerabilities.

(a) CVSS 2.0.

CVSS 2.0	#	%
Low	5	5.62
Medium	64	71.91
High	20	22.47

(b) CVSS 3.0.

CVSS 3.0	#	%
Low	0	0.00
Medium	16	17.98
High	37	41.57
Critical	5	5.62
N/A	31	34.83

Table 6: The exploitation status of the early alerts.

Exploitation status	Twitter publishing		At disclosure	
	#	%	#	%
Exploited	21	23.60	22	24.72
PoC	11	12.36	11	12.36
No data	57	64.04	56	62.92

6.2 Vulnerability impact

Although the existence of early alerts is relevant by itself, it is essential to assess the impact of the vulnerabilities. Table 5 presents how many early alerts have low, medium, high, and critical (CVSS 3.0 only) CVSS scores according to the CVSS 2.0 and 3.0 scoring systems. As the CVSS 3.0 was released in 2015, 31 early alerts are ranked only according to CVSS 2.0 (the N/A line in Table 5b).

Almost all early alerts are ranked by CVSS 2.0 as having a medium or high impact (about 94%). When considering the CVSS 3.0, no alerts are ranked with low impact, and five are graded with a critical score.

Despite the small number of early alerts, the CVSS score points out that these are relevant vulnerabilities and should not be disregarded. For example, CVE-2016-7089 is a WatchGuard firewall vulnerability that allows privilege escalation via code injection. This vulnerability belongs to the set of issues disclosed by the "Shadow Brokers" [8], and has a public exploit on ExploitDB [23].

6.3 Vulnerabilities exploited at disclosure time

A vulnerability only has an actual impact once it is exploited. Table 6 shows the exploitation status of the early alert vulnerabilities, both at the Twitter publishing and disclosure dates. The majority of vulnerabilities are not paired with observations of their exploits in the wild (64% or 57). A quarter of these cases (23.6% or 21) are known to be exploited. In a few cases (12% or 11), a PoC was referred by the vulnerability notice, describing how to exploit the vulnerability. As it is impossible to know if that PoC was used, we categorised these separately from the cases where the exploitation was confirmed.

We matched the early alerts with CVE-mentioning exploits present in ExploitDB [6] to complement the previous result. Only one case was found, published before the earliest vulnerability database and after the disclosing tweet. This information was used to update Table 6, adding the "At disclosure" column.

Considering vulnerabilities known to have been exploited and those with a PoC, the total amounts to about 34%. Current studies estimate that the percentage of vulnerabilities that are exploited in the wild is 5.5% [38], meaning that these early alerts include many appealing targets for hackers.

6.4 Actionability

Perhaps even more important than knowing the impact or exploitation status of a vulnerability, is to avoid exploitation. This can be achieved by applying patches or configurations to protect the vulnerable system. Table 7 shows which vulnerability mitigation measures can be reached by following the hyperlinks found in early alert tweets. In almost 40% of the cases, the tweet includes a link pointing to a patch that solves the vulnerability. For another 40% of vulnerabilities there is no patch available ("None"), or that information is not clear or the topic is not discussed ("No data"). The unreadable case is due to a page not written in English, where some parts of the text were not clear even after translation. The N/A entries are due to dead links, which blocked the analysis.

In the majority of cases (57%), the early alerts provided some information on how to protect the vulnerable systems from exploitation, either by patch or configuration. If the cases where we could not get more information (the "No data" cases) provided some solution, then the protection rate would increase to more than 70%. Therefore, we conclude that besides impact and exploitation relevance, early alerts are also useful due to the actionability they enable, as they inform security practitioners of possible actions to protect their systems.

Table 7: The actionability provided by the early alert tweets.

Action types	#	%
Patch	34	38.20
Configuration/patch	5	5.62
Configuration	12	13.48
None	23	25.84
No data	13	14.61
Unreadable	1	1.12
N/A	1	1.12

7 How Vulnerabilities are Discussed on Twitter

In this section we characterize some aspects of how vulnerabilities are discussed on Twitter. By identifying these aspects we provide guidelines for topic detection techniques oriented at capturing cybersecurity events. The following results are based on the analysis of the 9,093 vulnerabilities considered in this study.

7.1 Duration and Number of Tweets

Fig. 4 (left) presents the discussion duration. We observed that half of the vulnerabilities were discussed during up to eight days. However, it is interesting to see that the other half is middling spread across to up to 2,000 days. In some cases, the discussion can continue to up to almost 3,800 days. Discussion periods are extensive on some vulnerabilities mainly due to three different reasons: being used as comparative examples when discussing new events (*e.g.*, CVE-2014-0160, the "Heartbleed" bug); being (partly) reused on new attacks or as a

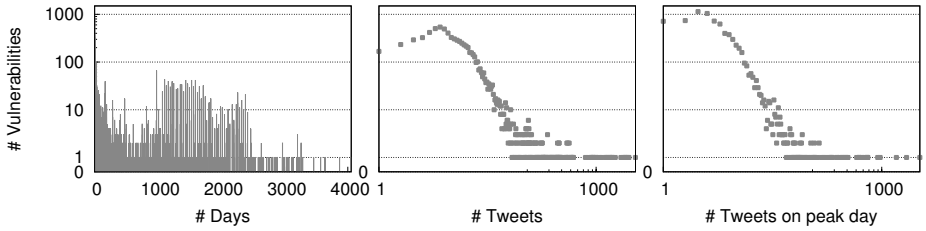

Fig. 4: Vulnerability discussion analysis: vulnerability discussion duration in days on Twitter (left), the total number of tweets discussing a vulnerability (middle), and the peak number of tweets in a single day (right).

part of a campaign, (*e.g.*, CVE-2017-11882 [13]); as case studies, therefore being remembered by their impact, specificity, or technical details.

Fig. 4 (middle) presents how many tweets discuss the vulnerabilities. From the graph we are omitting two outliers: one vulnerability discussed by 7,749 tweets, and another discussed by 15,733. Half of these vulnerabilities are discussed by two to thirteen tweets. These results are not surprising considering that most vulnerabilities are uneventful. The large majority of vulnerabilities are described, patched, and forgotten. Also, only a small percentage of vulnerabilities is exploited in the wild [38], which are the ones more likely to attract more attention. Only 351 vulnerabilities were discussed by more than 50 tweets, showing that although this content is posted on social media, relatively few vulnerabilities attract attention. However, taking a closer look at those 351 vulnerabilities, 14 of them have low severity rating according to CVSS 2.0, 124 have medium severity, and 213 have high severity. Although it is not implied that vulnerabilities with medium and high severity are going to be widely discussed, these results indicate that those referred by more tweets tend to have higher severity ratings.

Fig. 4 (right) presents the daily peak discussion, *i.e.*, the maximum number of tweets discussing the vulnerability in a single day. Three-quarters of the vulnerabilities have daily peaks between one and ten tweets. This is an important factor for topic detection techniques, as most of these identify new trends based on detecting bursts of tweets discussing the same event [60, 61].

7.2 Accounts

The tweets discussing security content used in this study were posted by 194,016 different accounts. We performed a quick analysis to understand if any account(s) stand out as sources of cybersecurity tweets. Out of the 194,016, only 5,863 of them published more that one relevant tweet, and only 228 posted more than 5. The highest tweet count for a single account is 73 tweets. Therefore, in this study, we did not find any best accounts to follow for cybersecurity content.

Table 8: Percentage of correctly detected tweets according to the various datasets and methods. The header row includes the dataset size between brackets.

	Early (89)	**Early post 2010** (76)	**Colloquial** (9101)	**Colloquial post 2010** (7923)	**All CVES** (94,398)	**All CVES post 2010** (77409)
Heuristic	56.18%	63.16%	57.01%	63.98%	71.14%	70.99%
Heuristic + CVE	62.92%	71.05%	88.46%	99.24%	84.56%	93.73%
CNN	57.30%	45.68%	89.51%	87.76%	87.74%	87.59%

8 How to find timely tweets

The results presented so far on this paper are a forensic analysis of the content posted on vulnerability databases and on Twitter. However, security practitioners are interested in capturing these posts live. Therefore, this section provides insights for data collection methods based on the aforementioned knowledge.

Systems that collect threat intelligence are designed to detect relevant news items while discarding non-relevant ones. The various systems proposed in the literature vary in terms of the complexity of the data selection approach, and as it was infeasible to test all of them, we selected three approaches to test against our data. The first is a simple heuristic-based approach. A tweet is considered relevant if it mentions a software element and a threat word from the VERIS [22] or ENISA [22] cybersecurity taxonomies. The second one is equal to the first but also detects the word "CVE". The third is a more sophisticated approach. We used a convolutional neural network-based approach (CNN) from a previous Twitter-based cyberthreat detection work developed by the authors [36, 37]. The test simply measures if the approach correctly detects the target tweets.

Table 8 shows the results of these approaches. We distinguish pre- and post-2010 periods as the coverage differs significantly. The percentages on first four columns in the table are obtained against the labelled data used in Sections 6 and 7, respectively. The last two columns are obtained by running the techniques mentioned above against Section 5. These are speculative results as the data is not labelled, but should transpose from the relatively large labelled data.

Using the simplest heuristic method is the worst method of the three except in one case. This means that the trivial approach works but lacks expressiveness regarding how vulnerabilities are discussed. Adding the word "CVE" to the detection mechanism enables it to detect tweets about already indexed vulnerabilities that do not follow the "software name and threat type" rule, drastically increasing the detection rates. Therefore, this is a suitable detection technique.

Finally, the CNN presents rather poor results regarding detecting early alerts but otherwise is consistently close to 90% accuracy. Although sometimes the CNN has a lower accuracy rate than the heuristics, this test does not cover false positive rates, where the CNN is expected to largely outperform heuristics.

We performed a follow-up analysis to the early alert tweets in an effort to understand the CNN results. We used BERT [35] to obtain semantic-rich feature vectors from the tweets, and cosine similarity [62] as a similarity measure. The

tweets were grouped by similarity, where each tweet was grouped with its most similar peers, as long as each group had an average similarity rating above 0.8. In almost all cases, the CNN gave the same classification for all members of each group. By observing these groups we can assert that the CNN accurately classified as relevant tweets with a more cybersecurity-oriented speech ("*New "Lucky Thirteen" attack on TLS CBC...*" or "*Misfortune Cookie: The Hole in Your Internet Gateway...*"), while incorrectly classifying less structured tweets ("*Only in the IT world can you say things like "header smuggling" (...) or in regex "Did you escape the caret?*" or "*The text for TBE-01-002 references TBE-01-004 - which does not seem to be included in the report. Is that intentional?*"). As our CNN was trained mostly using tweets directly discussing cybersecurity, any tweets not conforming to the pattern are likely to be discarded. Thereby we conclude that a diverse training set for neural networks is required towards a complete detection coverage.

9 Summary of Findings

In the following, we summarize the findings associated with each of the research questions formulated in this study.

RQ1: Is NVD the richest and timeliest vulnerability database? The NVD does not stand out as the most complete vulnerability database, as there are others that index more vulnerabilities. Also, NVD is not the most timely database. In fact, it was never the first database to publish a new vulnerability ahead of the others. However, NVD is known to have a strict publishing policy, allowing for consultation and comments from product vendors, which means that vulnerability publication may be delayed.

RQ2: Does Twitter provide a rich and timely vulnerability coverage? Since the beginning of 2010 Twitter provides a timely and rich coverage of known vulnerabilities. Moreover, there is a small subset of vulnerabilities (less than 1% of those we inspected) that are discussed on Twitter before their inclusion on vulnerability databases. Although these are very few cases, our analysis shows that they are relevant, impactful, and in many cases provide useful security recommendations. Overall, we consider Twitter as a useful cybersecurity news feed that should be taken into account by security practitioners.

RQ3: How are vulnerabilities discussed on Twitter? Vulnerability discussion on Twitter is carried out mostly in small bursts of two to thirteen tweets. Most vulnerabilities stopped being discussed within eight days, although tweets about them can appear for several years. Vulnerabilities discussed by a higher-than-usual volume of tweets (more than 50) tend to have higher impacts.

10 Insights for Practical Usage

Beyond the comparative analysis presented in this paper, there are a set of insights that we gathered while analysing the tweets collected. Below we present practical takeaways related to OSINT usage and its advantages.

No vulnerability database stands out as the best. NVD is an essential OSINT source, especially due to its thorough analysis and important link aggregation, but other reputable databases should be considered as a complement for four main reasons: *Timeliness*—NVD is not the most timely database (see Section 4); *Actionability*—NVD does not directly provide suggestions to mitigate or avoid vulnerabilities, unlike other databases (*e.g.*, PacketStorm, Security Focus); *Known exploits*—NVD does not collect information about known exploits, unlike other databases (*e.g.*, PacketStorm, Security Focus); *Completeness*—NVD is not the most complete database (see Table 1). Therefore security practitioners should use a database ensemble to collect security events.

Twitter is relevant. OSINT is provided by many reputable sources and should be taken seriously. Besides the significant research efforts (*e.g.*, [26, 40, 41, 43, 59]), there are companies and tools dedicated to OSINT sharing and enrichment. Sections 5 and 6 demonstrate that tweets provide timely, relevant, and useful cybersecurity news.

Twitter is a natural data aggregator. Another clear advantage of using Twitter to gather information is its natural data aggregation capability. The 89 early alerts mention 73 different products from 59 different vendors. When considering the 9,093 vulnerabilities analysed, these numbers extend to 1,153 products from 346 vendors. Forty-two CVEs are not indexed in the Common Platform Enumeration—a database of standard machine-readable names of IT products and platforms [2] used by CVE.

Security advisories may not be provided by smaller companies. The majority of vendors mentioned in the 89 early alerts do not provide an advisory site or news blog, while the vendors who provide advisories may not provide an API or a feed subscription. Since advisory sites link their content to Twitter at publication time, one can receive security updates by following the advisory accounts or by accessing Twitter's stream API and applying appropriate filters.

Twitter is important but not omniscient. We believe that a plausible trend for OSINT is to use Twitter as a front-end of the latest events. Since tweets have a relatively small size, messages tend to be concise, efficiently summarising the content of the associated links. This is one of Twitter's characteristics that made it so popular: reading a set of tweets is much faster than inspecting a collection of web sites. Therefore, Twitter naturally provides an almost standardised summary, quick and straightforward to process, which is very attractive for Security Operation Centres.

Tweets will not replace the current security publishing mechanisms in place. Once a security-related tweet is received, a visit to the associated site is practically mandatory to understand the issue at hand or to search for patches, among other relevant data. It is also arguable that a similar feed can be obtained by using an RSS feed. However, through Twitter, it is possible to monitor multiple accounts and to gather additional information not provided by RSS, such as timely breakthroughs or further discussion concerning the issue.

Collecting OSINT is a continuous process. Another takeaway for security practitioners is that it is essential to follow news about all layers of the

software stack by including keywords related to network protocols (*e.g.*, SSL, HTTP) or purchased web services (*e.g.*, cloud services, issue tracking services). This may seem obvious to the reader given this paper's discussion. However, as part of a research work unrelated to this paper, when we asked security analysts of three industrial partners (nation-wide and global companies with dedicated Security Operations Centres) for keywords to describe their infrastructure (to guide our tweet collection), they did not include network protocols or hardware elements.

Moreover, beyond receiving updates about selected assets, it is vital to obtain trending security news. It is hard to describe all relevant elements of a large company thoroughly, and maybe not all software in use is indexed or known. By extending the collection elements with (for example) topic detection techniques (*e.g.*, [34, 39, 46, 60]), one is more likely to cover all software in use. As Twitter can provide all these types of news and the research community has studied thoroughly topic detection on this platform, having trend detection might be mandatory for effective OSINT collection.

Diverse sources complement each other. Finally, and to complement the previous insight, it is important to follow a diverse set of accounts to observe the broad universe of software vendors. The early alerts were posted by 53 different accounts (for 89 alerts), demonstrating that diversity of sources is crucial for awareness. Moreover, during this study we collected tweets posted by about 194,000 accounts, reinforcing the idea that Twitter is a cybsersecurity discussion hub. It is also important to discuss critical cases like the exploitation of CVE-2017-0144, which became known as "wannacry". The vulnerability was published on CVE/NVD and Microsoft's security advisory, and patched a few months before the wannacry crisis. Therefore, by following Microsoft or CVE/NVD one would be aware of the issue and could avoid the ransomware.

Once the vulnerability started being exploited, several online discussions suggested a set of configurations that blocked the exploit. Therefore, those that did not patch their systems (and for the Windows versions that were not patched by Microsoft) could benefit from OSINT once the attacks began. The wannacry ransomware generated a massive discussion on Twitter: describing the issue, how to avoid it, and informing about the kill switch that eventually disabled it.

11 Conclusions and Limitations

In this paper we provide an analysis of the richness of coverage of vulnerabilities and timeliness (in terms of reporting dates of vulnerabilities) of some of the most important OSINT sources, namely Twitter and several vulnerability databases. Our key findings are the following: no source could be considered clearly better than others and therefore diverse OSINT sources should be used as they complement each other; when considering only confirmed vulnerabilities, NVD should not be the unique vulnerability database subscribed; since 2010, Twitter provides an almost perfect vulnerability coverage; Twitter discusses vulnerabilities ahead of databases for very few cases (about 1% for the vulnerabilities examined), and

is as timely as the vulnerability databases for the remaining cases; and finally, most of the vulnerabilities reported early on Twitter have a high or critical impact, with the tweet leading to usable mitigation measures. Beyond the collected facts, we provide a set of insights for the security practitioner interested in using OSINT for cybersecurity. These insight are based on our experience of manual inspection of almost one million tweets, and analysing many thousands of vulnerabilities. We believe this knowledge should be valuable for security analysis and research both in industry and academia.

Limitations. The results presented in this paper are somewhat pessimistic in terms of the number of vulnerabilities that were found to have early alerts on Twitter. There could be more cases with media attention or early alerts that were not captured by our methodology since we cover a reduced amount of vulnerabilities. There is also the possibility of human error, as manual processing of tweets can lead to mistakes and missing some matches—all early alerts were triple checked to avoid false positives.

Another factor that we cannot control (since we are performing a forensic analysis) is that some early alert tweets could have been deleted before this study, and thus not captured. Many dead links also invalidated possible matches, especially when the tweet links used some shortening system, such as "dlvr.it", "hrbt.us", "url4.eu", "bit.ly", or "ow.ly".

Funding

This work was supported by the H2020 European Project DiSIEM (H2020-700692) and by the Fundação para a Ciência e a Tecnologia (FCT) through project ThreatAdapt (FCT-FNR/0002/2018) and the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020).

Bibliography

[1] Additional paper data. `https://github.com/fernandoblalves/Follow-the-Blue-Bird-Paper-Additional-Data`, [Accessed 10-07-2020]

[2] Common platform enumeration. `https://cpe.mitre.org/about/`, [Accessed 15-04-2020]

[3] Common vulnerabilities and exposures (cve). `http://cve.mitre.org/`, [Accessed 15-04-2020]

[4] Common vulnerability scoring system version 3.0. `https://www.first.org/cvss/v3-0/`, [Accessed 15-04-2020]

[5] Cvss v2 archive. `https://www.first.org/cvss/v2/`, [Accessed 15-04-2020]

[6] Exploit database. `www.exploit-db.com/`, [Accessed 15-04-2020]

[7] Get old tweets programatically. `https://github.com/Jefferson-Henrique/GetOldTweets-java`, [Accessed 15-04-2020]

[8] Hackers say they hacked nsa-linked group, want 1 million bitcoins to share more. https://www.vice.com/en_us/article/ezpa9p/hackers-hack-nsa-linked-equation-group, [Accessed 15-04-2020]
[9] How people use Twitter in general. https://www.americanpressinstitute.org/publications/reports/survey-research/how-people-use-twitter-in-general/, [Accessed 15-04-2020]
[10] How Twitter evolved from 2006 to 2011. https://buffer.com/resources/how-twitter-evolved-from-2006-to-2011, [Accessed 15-04-2020]
[11] The mitre corporation. https://www.mitre.org/, [Accessed 15-04-2020]
[12] National vulnerability database. https://nvd.nist.gov/, [Accessed 15-04-2020]
[13] New targeted attack in the middle east by APT34, a suspected iranian threat group, using cve-2017-11882 exploit. https://www.fireeye.com/blog/threat-research/2017/12/targeted-attack-in-middle-east-by-apt34.html, [Accessed 15-04-2020]
[14] Operation aurora. https://en.wikipedia.org/wiki/Operation_Aurora, [Accessed 15-04-2020]
[15] Packet storm. https://packetstormsecurity.com/, [Accessed 15-04-2020]
[16] The race between security professionals and adversaries. https://www.recordedfuture.com/vulnerability-disclosure-delay/, [Accessed 15-04-2020]
[17] Security database. https://www.security-database.com/, [Accessed 15-04-2020]
[18] Spiderfoot. https://www.spiderfoot.net/documentation/, [Accessed 15-04-2020]
[19] Tweepy. https://www.tweepy.org/, [Accessed 15-04-2020]
[20] Tweet attacks pro. http://www.tweetattackspro.com/, [Accessed 15-04-2020]
[21] Twitter input plugin. https://www.elastic.co/guide/en/logstash/current/plugins-inputs-twitter.html, [Accessed 15-04-2020]
[22] Veris taxonomy. http://veriscommunity.net/enums.html#section-incident_desc, [Accessed 13-06-2018]
[23] Watchguard firewalls - 'escalateplowman' ifconfig privilege escalation. https://www.exploit-db.com/exploits/40270, [Accessed 15-04-2020]
[24] Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., Shakarian, P.: Proactive identification of exploits in the wild through vulnerability mentions online. In: 2017 CyCon US (2017)
[25] Alsaedi, N., Burnap, P., Rana, O.: Can we predict a riot? Disruptive event detection using Twitter. ACM TOIT **17**(2) (2017)
[26] Alves, F., Bettini, A., Ferreira, P.M., Bessani, A.: Processing tweets for cybersecurity threat awareness. Information Systems (2020)
[27] Andongabo, A., Gashi, I.: vepRisk - A Web Based Analysis Tool for Public Security Data. In: 13th EDCC (2017)

[28] Arora, A., Krishnan, R., Nandkumar, A., Telang, R., Yang, Y.: Impact of vulnerability disclosure and patch availability-an empirical analysis. In: Third Workshop on the Economics of Information Security (2004)
[29] Behzadan, V., Aguirre, C., Bose, A., Hsu, W.: Corpus and deep learning classifier for collection of cyber threat indicators in Twitter stream. In: IEEE Big Data 2018 (2018)
[30] Bian, J., Topaloglu, U., Yu, F.: Towards large-scale Twitter mining for drug-related adverse events. In: Proc. of the SHB (2012)
[31] Bozorgi, M., Saul, L.K., Savage, S., Voelker, G.M.: Beyond heuristics: learning to classify vulnerabilities and predict exploits. In: 16th ACM SIGKDD (2010)
[32] Bullough, B.L., Yanchenko, A.K., Smith, C.L., Zipkin, J.R.: Predicting exploitation of disclosed software vulnerabilities using open-source data. In: 3rd ACM IWSPA (2017)
[33] Campiolo, R., Santos, L.A.F., Batista, D.M., Gerosa, M.A.: Evaluating the utilization of Twitter messages as a source of security alerts. In: SAC 13 (2013)
[34] Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on Twitter based on temporal and social terms evaluation. In: 10th MDM/KDD (2010)
[35] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018), arXiv:1810.04805
[36] Dionísio, N., Alves, F., Ferreira, P.M., Bessani, A.: Cyberthreat detection from Twitter using deep neural networks. In: IJCNN 2019 (2019)
[37] Dionísio, N., Alves, F., Ferreira, P.M., Bessani, A.: Towards end-to-end cyberthreat detection from Twitter using multi-task learning. In: IJCNN 2020 (2020)
[38] Edkrantz, M., Truvé, S., Said, A.: Predicting vulnerability exploits in the wild. In: 2nd IEEE CSCloud (2015)
[39] Fedoryszak, M., Frederick, B., Rajaram, V., Zhong, C.: Real-time event detection on social data streams. 25th ACM SIGKDD - KDD 19 (2019)
[40] Le Sceller, Q., Karbab, E.B., Debbabi, M., Iqbal, F.: Sonar: Automatic detection of cyber security events over the Twitter stream. In: 12th ARES (2017)
[41] McNeil, N., Bridges, R.A., Iannacone, M.D., Czejdo, B., Perez, N., Goodall, J.R.: Pace: Pattern accurate computationally efficient bootstrapping for timely discovery of cyber-security concepts. In: 12th ICMLA (2013)
[42] McQueen, M.A., McQueen, T.A., Boyer, W.F., Chaffin, M.R.: Empirical estimates and observations of 0Day vulnerabilities. In: 42nd HICSS (2009)
[43] Mittal, S., Das, P.K., Mulwad, V., Joshi, A., Finin, T.: Cybertwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities. In: 2016 ASONAM (2016)
[44] Moholth, O.C., Juric, R., McClenaghan, K.M.: Detecting cyber security vulnerabilities through reactive programming. In: HICSS 2019 (2019)

[45] Nayak, K., Marino, D., Efstathopoulos, P., Dumitraş, T.: Some vulnerabilities are different than others. In: 17th RAID (2014)
[46] Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: 11th NAACL HLT (2010)
[47] Reinthal, A., Filippakis, E.L., Almgren, M.: Data modelling for predicting exploits. In: NordSec (2018)
[48] Ritter, A., Wright, E., Casey, W., Mitchell, T.: Weakly supervised extraction of computer security events from Twitter. In: 24th WWW (2015)
[49] Rodriguez, L.G.A., Trazzi, J.S., Fossaluza, V., Campiolo, R., Batista, D.M.: Analysis of vulnerability disclosure delays from the national vulnerability database. In: WSCDC-SBRC 2018 (2018)
[50] Roumani, Y., Nwankpa, J.K., Roumani, Y.F.: Time series modeling of vulnerabilities. Computers & Security **51** (2015)
[51] Sabottke, C., Suciu, O., Dumitra, T.: Vulnerability disclosure in the age of social media: exploiting Twitter for predicting real-world exploits. In: 24th USENIX Security Symposium (2015)
[52] Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors. In: 19th WWW (2010)
[53] Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., Ferrara, E.: Early warnings of cyber threats in online discussions. In: 2017 ICDMW (2017)
[54] Sapienza, A., Ernala, S.K., Bessi, A., Lerman, K., Ferrara, E.: Discover: Mining online chatter for emerging cyber threats. In: WWW18 Companion (2018)
[55] Sauerwein, C., Sillaber, C., Huber, M.M., Mussmann, A., Breu, R.: The tweet advantage: An empirical analysis of 0-day vulnerability information shared on Twitter. In: 33rd IFIP SEC (2018)
[56] Shahzad, M., Shafiq, M.Z., Liu, A.X.: A large scale exploratory analysis of software vulnerability life cycles. In: 34th ICSE (2012)
[57] Steele, R.D.: Open source intelligence: What is it? why is it important to the military. American Intelligence Journal **17**(1) (1996)
[58] Syed, R., Rahafrooz, M., Keisler, J.M.: What it takes to get retweeted: An analysis of software vulnerability messages. Computers in Human Behavior **80** (2018)
[59] Trabelsi, S., Plate, H., Abida, A., Aoun, M.M.B., Zouaoui, A., et al.: Mining social networks for software vulnerabilities monitoring. In: 7th NTMS (2015)
[60] Xie, W., Zhu, F., Jiang, J., Lim, E.P., Wang, K.: Topicsketch: Real-time bursty topic detection from Twitter. IEEE TKDE **28**(8) (2016)
[61] Yan, X., Guo, J., Lan, Y., Xu, J., Cheng, X.: A probabilistic model for bursty topic discovery in microblogs. In: 29th AAAI (2015)
[62] Zaki, M.J., et al.: Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press (2014)