# Curating Tweets: A Framework for Using Twitter for Workplace Learning

## ABSTRACT

Cybersecurity is a rapidly evolving field where professionals constantly need to keep up with new technologies and retrain. In this paper, we present a study that analyzed social media data and use the findings to aid professionals and students to learn more effectively using Twitter. We analyzed 23,000 cybersecurity related tweets posted on Twitter across two hashtags #cybersecurity and #infosec. Our analysis created a framework that explains how using descriptive, content, and network analysis can generate information that can help professionals learn. In addition, our analysis provided insights on the tweets and the cybersecurity community that use them. These insights include: Most tweets covered multiple topics and used three or more hashtags. Companies and other organizations had the highest numbers of followers, but individual users, experts in the field, were the most retweeted. Popular users, based on follower counts, were not necessarily the most influential (based on retweets). In terms of content, popular tweets consisted of infographics that packed a lot of information. Tweets were commonly used to announce file dumps of hacks and data leaks. Many highly used hashtags represented current threats and the overall sentiment of cybersecurity tweets are negative. Highly connected users on Twitter served as hubs across the three primary sub communities identified in the data. Insights from his study can assist with improving workforce development by guiding professionals in getting pertinent information and keeping up to date with the latest security threats and news.

## 1 Introduction

Security professionals are required to constantly learn in order to be successful at what they do. This learning is critical as the security landscape constantly changes, with new threats and technologies being introduced on a daily basis. This dynamic landscape means that professionals must be able to keep up with the changes in a way that is effective. For students, traditionally, education programs have served as the foundation source of knowledge in training people to enter the workforce. However with the rise of information technology professions such as cyber security, there have been many challenges in terms of preparing the cyber security workforce. Security education programs have found it difficult to keep up with the fast changing security landscape [1–3]. Security threats evolve over time and tend to change very rapidly. A threat that is prevalent today might not be prevalent tomorrow. Many educators find it difficult to keep their curriculum updated to keep pace with the changes [2]. Outdated security programs discussing obsolete topics can degrade the quality of the security education [4], affecting the instructor's effectiveness in helping students understand threats. This could potentially lead to security breaches and violations that could have been avoided with more relevant curricula.

There has been numerous initiatives by government, academia, and organizations to address these shortcomings in cyber security education. The National Initiative for Cybersecurity Education was created under the National Institutes of Standards and Technology as a partnership between government, academia and the private sector to address workforce challenges through standards and best practices. The Association for Computing Machinery (ACM) publishes the Cybersecurity Curricula that identifies knowledge areas and knowledge units that provide a foundation for cybersecurity education programs [5]. The Pedagogic Cybersecurity Framework extended the Open Systems Interconnection (OSI) model to include three additional layers to help explain the non-technical areas that influence security within an organization [6]. The National Security Agency and Department of Homeland Security sponsors the National Centers of Academic Excellence (CAE) program to certify colleges and universities that meets their requirement by aligning their curriculum to cybersecurity knowledge units that are validated by subject matter experts [7].

A study by IBM to understand cybersecurity academic programs around the world found that less than 60 percent of students and educators surveyed believe that their academic program addresses cybersecurity practices in emerging technology areas such as mobile computing, cloud, and social business [1]. It found that the cybersecurity field has expanded significantly over time, with more security domains to cover and more types of attacks that must be understood. At the same time, courses require cybersecurity education to be integrated to it while having the same number of hours as before. This means that security education programs alone cannot meet all the requirements of the workforce, and continuous education beyond the classroom is vital to the field.

Social media has emerged as a critical platform for accessing information. Unlike traditional resources, the major advantage of social media is that it provides real-time, up-to-date information often from trusted sources. In particular, a microblogging platform such as Twitter serves as a useful resource not only because the information shared is current but also because of its participatory features – users can follow products, services, brands, and topics (hashtags) that interest them. The goal of this study is to deepen our understanding of how Twitter is used within the cybersecurity field including who the users are, what type of information is shared, and how can this information be leveraged to aid professionals and students alike. Using a Twitter analytical framework we use a combination of methodologies and techniques for descriptive analytics, content analytics, and network analytics. The following research questions motivated our study: 1) What type of information can be found in looking at the data from Twitter? What are the characteristics and features of the tweets? 2) Who are the users of these Twitter communities and what are their interactions? 3) And how can these analytical methodologies aid in the learning process?

## 2 Research Study

Twitter is a social media platform that allows people to communicate using short messages called tweets. Tweeting allows a person to publish a message to anyone user on the platform or even publicly about their thoughts, feelings, or opinion of anything that is of interest to them. There are about 330 million active users that send out about 500 million tweets per day [8]. Of those, about 79% are located outside the United States. A tweet is a message that can contain up to a maximum of 280 characters (used to be 140). There are three types of messages on Twitter: original tweets, replies, and retweets. Original tweets are messages that originate with a user and are posted on their timeline. Retweets are tweets that have been shared by other users who find a tweet to be interesting or important to them. Replies are tweets written in response to what someone else has said.

One thing that makes the Twitter platform attractive is that it does not require the user to be proactive in order to obtain information. A user can just follow some influential accounts and be able to obtain all the information they need. These accounts can be experts in the field, or organizations that focus and aggregate security related news. However, the drawback to using this approach is that for many users, it is not easy to know who they should follow or apparent who the influential expert is in the field. A user with high number of followers might not necessary tweet the most relevant information. As discovered in our study, a user's popularity does not determine the value of the information they tweet. Therefore being able to find and know which users to follow is critical in being able to effectively use Twitter as a platform for learning. The findings from this study showed effective ways to find these key users.

Twitter offers an API that allows researchers and organizations to collect the data for analysis purposes (the terms and data shared changes continuously). These Search and Streaming APIs allows data collection different types of queries, such as by using a specific keyword. This open data policy has made Twitter data one of the most popular data source and they have been used widely for analysis. Some examples of practical applications of Twitter data includes predicting flu trends [9], predicting elections [10], and user sentiment analysis [11]. Twitter has also been used in many fields such as healthcare [12], cybersecurity [13], finance [14], amongst many others.

Twitter data for the research reported in this paper were collected using the publically available API that Twitter provides. Data collected through the API was parsed and stored in an Excel spreadsheet and included the actual tweet along with its metadata. Metadata collected included information on the user who posted the tweet, how many followers they have, how many times the tweet was retweeted, date and time stamps, along with other relevant information. The Twitter API only allows acquiring 1% of all the publically available Twitter data, however even with this limitation, we were able to collect a dataset large enough for this research within a period of 45 days. For the initial data collection, we configured the data collector to collect tweets from multiple hashtags, including "infosec", "malware", "breach", "security", among a few others. After analyzing the tweets that were collected, it was discovered that the two primary hashtags used in the cybersecurity field are #cybersecurity and #infosec, as tweets from the other hashtags also included either one of these two primary hashtags. The dataset consists of 23,313 tweets and their metadata from the hashtag #cybersecurity and #infosec in a 45 day period from October 31st to December 14th, 2017. In order to analyze the data from different analytical aspects, the proposed framework in Figure 1 is used. The framework follows previous research [15] of Twitter data on the topic of supply chain and consist of three types of analytics – descriptive, content, and network. Each area focuses on a specific aspect of the data in order to understand the characteristics of the tweets, the users, and the connections between the tweets, and their relevance to real life events. The framework enhances on previous research by improving on aspects that are more relevant to cybersecurity, and improving the way certain research methods are utilized in order to yield more accurate results.

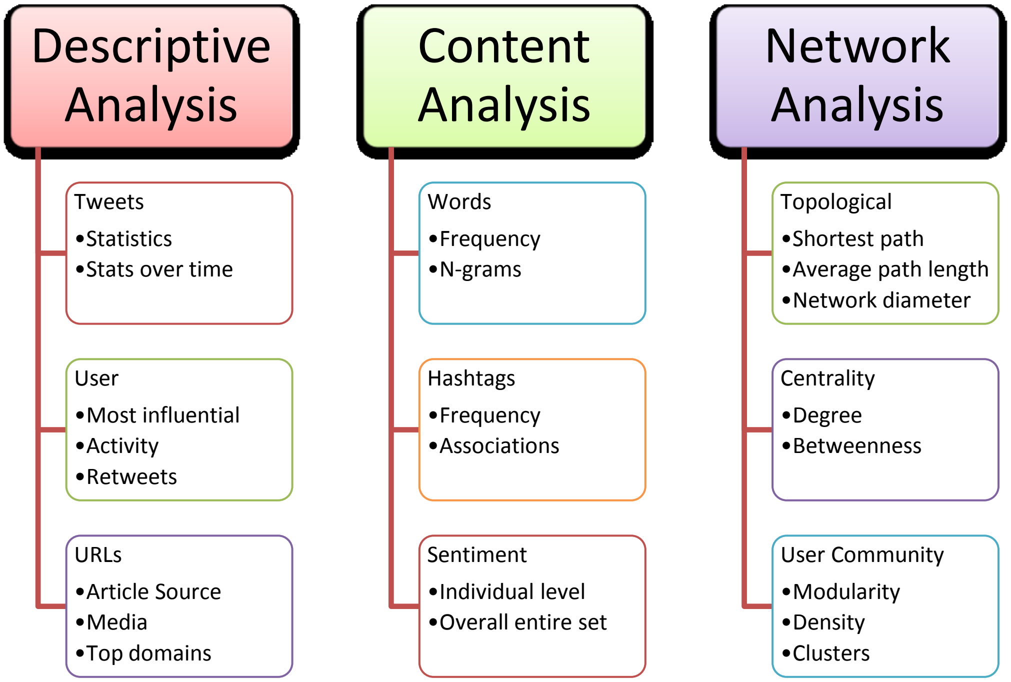


Figure 1 - Analytical framework for cybersecurity tweets

## 2.1 Descriptive Analysis

The data collected from Twitter contains a large number of information that must be analyzed. Descriptive analysis is used to describe basic information and feature of the data in the study. It provides summaries of the data including the sample and the measures. In quantitative analysis, it serves as the basis for this type of research and is also known as descriptive statistics [16–18]. It is important to understand the features of the dataset in order understand how and why certain events or pattern occurs within the dataset. It provides quantitative insights across a large data and break it down into smaller manageable pieces of information. The benefits of using Twitter data for this type of analysis is that the whole dataset can be used, rather than having to select a small sample from the dataset.

In this study, descriptive analysis will look at information and metrics in three main areas of the dataset: tweets, users, and URLs. The outcome of this analysis will provide a picture into the data and provide metrics about the tweets. Analyzing the tweets, the study will look at word counts, hashtags that are used, how tweets are produced over time, and the overall statistics of the tweets themselves. For the users, the study looks at who write the tweets, and who response to it. In addition, it identify the key players and characteristics that makes them important in the community, and whether there are correlations between certain tweet statistics such as followers and retweets. And lastly, URLs are an important aspect as most tweets include a hyperlink that leads to news articles, pictures, or other resources. The URL analysis will look at the types of links that are tweeted and where they come from. This could show topics and resources that are trusted, and whether there are domains that stand out for certain types of tweets.

With descriptive analysis, there are many other methods and metrics that can be used to show interesting facts about the dataset. However, the goal of this research is to show methods that would be useful to security professionals and researchers. Methods used are also geared towards analyzing cybersecurity tweets, with variations to be expected if it were to be used in other fields.

## 2.2 Content Analysis

The second part of the framework looks at the actual content of the tweets. It looks at the actual text and hashtags that is in the dataset in order to understand the content of the tweets and the dataset as a whole. Text data is considered to be “unstructured” data [19], thus it requires content analysis techniques such as text mining and natural language processing (NLP) in order to analyze the data and gain meaningful information from it  [18]. Although each tweet has a specific purpose and meaning in of itself, content analysis looks at all the tweets together as whole, rather than the individual pieces. Each tweet contains three component; a list of words, the hashtags that are associated with the tweet, and the URL that typically accompany the tweet. In order for the tweets to be analyzed, it must go through pre-processing steps to prepare the unstructured text to be analyzed [20]. These steps include tokenization, stemming, and the removal of stop words.

Term frequency shows all the keywords that are frequently discussed within the tweets, which shows key topics that are important to users. Topics that are more discussed, are more likely to be an area that are of interest to users, and can be a helpful indicator of trending or upcoming topics that are important. As many important topics consist of multiple words, n-grams are often used with term frequency in order to discover key phrases that are important within the text corpus. The unique aspect of Twitter data is that it allows the user to personally label and categorize their tweets using hashtags. This provide an important piece of information about each tweet that can be useful in understanding the overall subject of the tweet. Other data sources previously such as articles and documents, do not have this critical piece of information. Therefore, part of content analysis involves analyzing the hashtags themselves as it provides an important piece of information about the tweet in addition to the tweet itself.

Analyzing terms and hashtags only looks at the overall message and content of the tweet, it does not look at the actual opinion and the overall feeling of the user writing the message. Sentiment analysis [11, 21] extract the emotion of the user from the words that are used in the tweet. Most sentiment analysis techniques look at the words that are used in the sentence in order to determine whether the overall tone of the message is positive, negative or neutral. The words and its associated sentiment score is based on a dictionary that has been defined by experts in the field. There are many different methods for performing this analysis, and typically it is done on larger content of text such as a long paragraph or even an entire book, as the more words there are within the corpus, the more accurate the results. As tweets are much shorter in length, it will be important to choose a method that will account for that while at the same time be able to provide meaningful insights into the user's emotions.

### 2.3 Network Analysis

One major aspect of social media such as Twitter is the fact that there are a large number of users and interactions between them. Users don't just tweet messages to others; they respond to them and retweets the ones that they believe are important. These interactions between users can create communities of users within the social network. Community of users can show the dynamics between different groups of users, whether they are professionals, organizations, large corporations, or media companies.

In network theory [22], nodes and edges are used to represent these users and their relationships. With Twitter data, nodes represent the user, while their interactions (@reply or user mentions) with another user is the edge (relationship). This allows visualizations to be created that can show these relationships and provide network metrics on these interactions. Relationships can be visualized to determine community of users and how they connect to one another. Network theory also provide metrics such as betweenness centrality, which shows important key players within the network, and degree centrality which shows which user has the most connections (degrees) to others [23]. Betweenness centrality is a measure of centrality based on the shortest paths within the network. The shortest path between two users in the Twitter network shows how connected they are to each other, and the node (user) with higher betweenness centrality means multiple users connect to others through them, signifying their importance within the network. Degree centrality measure the number of connections a user have to others, and its simplest form can represent the popularity of that user [24]. Centrality analysis focuses on each individual user and their connectiveness to others.

Communities of users are an importance aspect of network analysis. On Twitter, users can use the @reply in their tweet to respond to another user's tweet. They can also use the @ symbol along with the name of a specific user to mention them within their tweet. These reply and user mentions form connections between users and generate communities. Modularity is a measure of the structure of networks and how it can be divided into modules (communities). Communities are groups of highly interconnected nodes that are only sparsely connected to the rest of the network [25]. Therefore, it is interesting to identify these communities and determine who the users are and how they interact with one another.

## 3 Findings

### 3.1 Descriptive Analysis

The 23,313 tweets collected contained at least two or more hashtags, including the #cybersecurity tag that was used to collect the data. This indicated that cybersecurity tweets typically touches multiple domains and validate how the field is multidisciplinary. The hashtags used were widely varied, with 5424 unique hashtags that were used, ranging from security related areas such as "malware" and "ransomware", to surprising areas such as "bitcoin" and "blockchain".

There are 4290 unique users in the dataset determined by their User ID. There is an average of 5.66 tweets per user. Table 1 shows the top 10 users with the most followers and retweets. Users with the most followers were mostly large organizations, highlighted in blue, such as BlackBerry, ZeeNews, EconomicTimes, and Nokia. Whereas users with the most retweets are mostly individual accounts, highlighted in yellow. This suggest that while companies and organizations typically get the most followers, it is the individual users that make the most influential and important tweets. Companies typically get followers due to their brand recognition, and people might follow them to get updates and news. Users with the most retweets are usually experts in the field and gets followers from other professionals.

|  | Most Followed |  | Most Retweeted |  |
|---|---|---|---|---|
|  | User | Followers | User | Retweets |
| 1 | BlackBerry | 4,529,498 | Fisher85M | 5,741 |
| 2 | ZeeNews | 3,758,453 | reach2ratan | 4,505 |
| 3 | policia | 3,080,679 | kennethholley | 4,382 |
| 4 | EconomicTimes | 2,866,978 | MikeQuindazzi | 3,328 |
| 5 | RT_com | 2,667,001 | JacBurns_Comext | 2,242 |
| 6 | nokia | 2,241,954 | quttera | 2,239 |
| 7 | bsindia | 1,555,483 | WiseCrowdGlobal | 1,554 |
| 8 | dez_blanchfield | 1,051,293 | jblefevre60 | 1,527 |
| 9 | PoliciaColombia | 1,041,438 | x0rz | 1,493 |
| 10 | 2morrowknight | 953,005 | CyberHitchhiker | 1,475 |

Table 1 - List of users with the most followers and retweets. Organizations are highlighted in blue, users in yellow

Looking at another perspective, these statistics about a user can be representative of a user's popularity and overall influence. Popularity can be defined as the number of followers a user may have, while influence can be defined as how often their tweets are retweeted. Is there a correlation between a user's popularity and their influence? Figure 2 shows the top 25 users with the most retweets and the number of followers they have. The figure shows that users who has a high number of followers (blue line), doesn't necessary means that they get a lot of retweets (orange line). One user "TheHackerNews" has a high number of followers but does not get a lot of retweets, compared to "Fisher85M" who have a lower number of followers but a lot of retweets. This suggests that there is no correlation between a user's popularity and their overall influence. This influence is also reinforced through the use of network graphs that will be discussed later on in this paper.

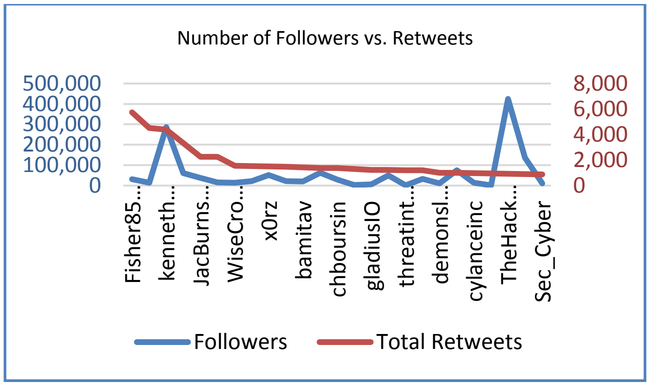
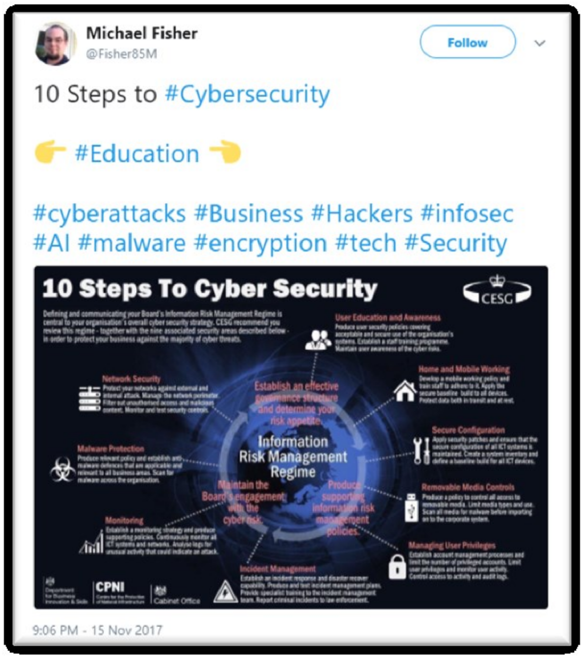


Figure 2 - Users' number of followers versus retweets

### 3.1.1 Positive Tweet Factors

So what makes a tweet "good" so that users are more likely to retweet it? We looked at Fisher85M, the user with the most retweets in all of our dataset to find the answer. Fisher85M's profile stated that he was a "Full-Time Analyst, Technology Evangelist, #CyberSecurity and #VR Influencer: @DZone & @Onalytica. Made out of 100% Geek". His most retweeted tweets had an average of 150 retweets. The top 25 most retweeted tweet all had some type of graphic within the tweet. This shows that users tend to "like" or "retweet" tweets that has graphics that shows accompanying information in addition to the text itself. However, this doesn't mean that every tweet with a graphic is popular. Fisher85M's infographics contained a lot of information all presented within one graphic, and this seems to be the reason why he is the most retweeted user. His 10 most retweeted tweets all contained a very detailed infographic relating to cybersecurity. Figure 3 shows an example of such tweet. The infographic is very dense and full of information. What makes this interesting is that in other media such as presentations, a slide with too much text or information packed in one page is typically frowned upon or discouraged. The opposite seems to be true for tweets based on our analysis. As tweets are very short in nature, the ability to present a lot of information at once seems to be the most useful for users ingesting the information. In a way, this seems to be contradicting with Twitter's main appeal, which is to limit the length of the message in order to keep it short and precise. In the context of learning, this finding suggests that users like tweets that contain useful information, such as a graphic that shows things that could benefit and apply to them. These infographics could be presented to security professionals to help them learn and stay abreast of the latest changes that are occurring within the profession. As there are thousands of cybersecurity tweets everyday, being able to identify and discover these infographics and presenting it to those who are interested would help them learn in an efficient and effective manner.

Tweets can be a very useful tool for learning and staying up to date with the latest information. But for most professionals and students, the challenge often comes to knowing who to follow in order to most effectively obtain the relevant information. Following too many users and there is information overload, where relevant information is buried beneath all the junk and irrelevant tweets. Many people often follow users who has the most followers, as the number of indicator often seems to be an important indicator of relevancy. However, the findings show that this is not often the case. Users with high number of followers are often due to the fact that they have a name that people recognize, rather than from the quality of tweets they produce. It is more important to follow users that are more influential in the community, as they tweet out information that are highly regarded by others in the community as indicated by the retweets.



10 Steps to #Cybersecurity

👉 #Education 👈

#cyberattacks #Business #Hackers #infosec #AI #malware #encryption #tech #Security

10 Steps To Cyber Security

Defining and communicating your Board's Information Risk Management Regime is central to your organisation's overall cyber security strategy. CESG recommend you review this regime - together with the nine associated security areas described below - in order to protect your business against the majority of cyber threats.

Information Risk Management Regime

Establish an effective governance structure and determine your risk appetite.

Maintain the Board's engagement with the cyber risk.

Produce supporting information risk management policies.

User Education and Awareness
Produce user security policies covering acceptable and secure use of the organisation's systems. Establish a staff training programme. Maintain user awareness of the cyber risks.

Home and Mobile Working
Develop a mobile working policy and train staff to adhere to it. Apply the secure baseline build to all devices. Protect data both in transit and at rest.

Secure Configuration
Apply security patches and ensure that the secure configuration of all ICT systems is maintained. Create a system inventory and define a baseline build for all ICT devices.

Removable Media Controls
Produce a policy to control all access to removable media. Limit media types and use. Scan all media for malware before importing on to the corporate system.

Managing User Privileges
Establish account management processes and limit the number of privileged accounts. Limit user privileges and monitor user activity. Control access to activity and audit logs.

Incident Management
Establish an incident response and disaster recovery capability. Produce and test incident management plans. Provide specialist training to the incident management team. Report criminal incidents to law enforcement.

Monitoring
Establish a monitoring strategy and produce supporting policies. Continuously monitor all ICT systems and networks. Analyse logs for unusual activity that could indicate an attack.

Malware Protection
Produce relevant policy and establish anti-malware defences that are applicable and relevant to all business areas. Scan for malware across the organisation.

Network Security
Protect your networks against external and internal attack. Manage the network perimeter. Filter out unauthorised access and malicious content. Monitor and test security controls.

CESG
Department for Business Innovation & Skills
CPNI
Centre for the Protection of National Infrastructure
Cabinet Office

9:06 PM - 15 Nov 2017

Figure 3 - Tweet with the highest number of "retweets" from Fisher85M

### 3.1.2 URL analysis

There are two URLs that are present in almost all the tweets. The first is the link to the media for the tweet, the second is the external URL that links to an article or another page on the internet. There are three possible media types; photos, animated GIFs, and videos. About half the tweets have no media (12869), with 46% having photos (11224) and 1% having GIFs (155) and videos (65). This suggests that users still predominantly photos or images within their tweets over flash animations or videos.

The majority of tweets that are highly retweeted all have external URLs within the tweet. Most of the URLs are in shortened formats using the domains such as bit.ly, goo.gl, buff.ly, among others. In order to determine the actual domains that are linked to, all shortened URLs were expanded to reveal the actual URL. Popular domains include zdnet.com, lastline.com, socialhub.com, darkread.com, csoonline.com, bleepingcomputer.com, thecyberwire.com, scmagazine.com, infosecurity-magazine.com, and thehackernews.com. These are software companies, social media analytic platforms, and security news websites. These make up the top domains that were linked to in the tweets. As many tweets are related to current news, the finding reinforce this fact as most of the external links are pointing to news websites. These top domains also identify the major news resources that professional use for security news and updates. The finding also discovered one interesting domain that shows up on a regular basis, PasteBin. Pastebin is an online text storage site that allows users to post up anonymous “pastes”. It was found that the website was used to post dumps of controversial files such as passwords, configurations, system settings, and anything else that is obtained through a hack or leak. This suggest that Twitter is widely used to announce and distribute files related to security hacks. There is a Twitter bot named “Dumpmon” that monitors PasteBin for these types of text paste and automatically tweets this information when it is available.

There are many websites out there that provide information security news. The vast number of available sites can be challenging to students and those new to the field, as they might not know which sites are good and are highly used by the community. The URL analysis identifies some of the top sites that are used by the community to disseminate news and updates in the cyber security domain. This information is useful in narrowing down some of the best sites that provide security news and information.

### 3.1.3 User activity over time

This analysis looks at the overall activity of users and when they tweet. Figure 4 shows the time in which tweets are created over the course of a 24 hour day in the UTC time zone. Activity starts off slow and gradually builds up to the highest levels at 4 PM, then tapers off as each hour passes. As the tweets come from different countries around the world, this would translate to roughly around the morning hours for the Americas, and the evening hours for Asia. Figure 5 shows the activity level over the 45 days that we collected our data. The number of tweets being sent each week fluctuate up and down depending on the day of the week. The number of tweets is at its lowest on Sundays, picks up drastically on Mondays then increasing to its maximum on Wednesday and Thursday, and then declines again. This cycle repeats every week. The study looked to see if there are correlations between the number of tweets and current security events but was not able to conclusively determine this factor as the time period of our data collection is too short.

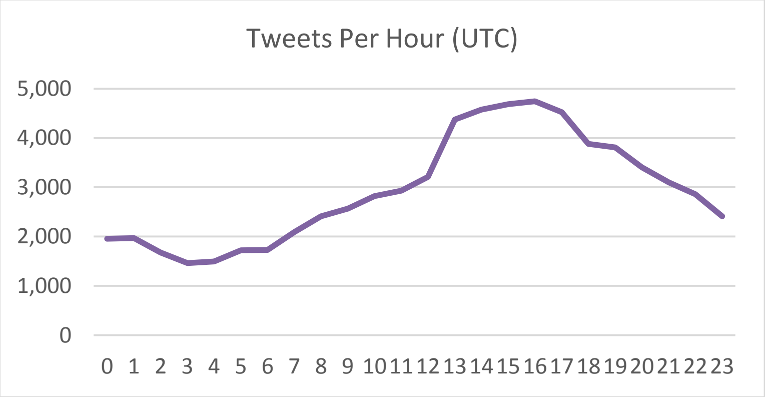


Figure 4 - Tweets per hour in UTC time zone

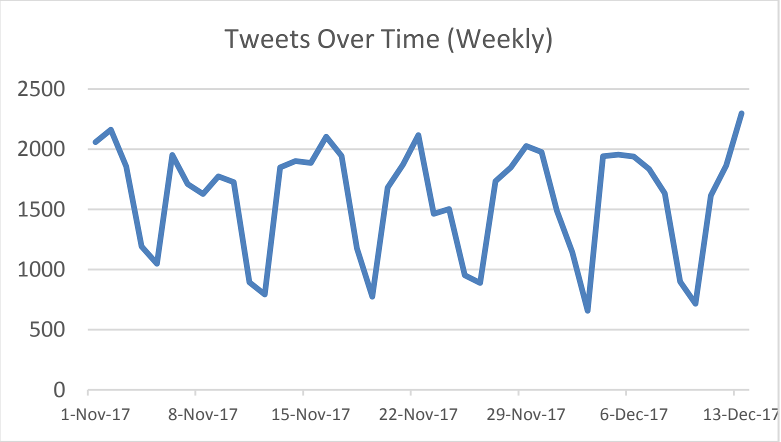


Figure 5 - Tweets per week

## 3.2 Content Analysis

This part of the study looked at the actual text content of the tweets of the dataset. Text mining methodologies were employed to analyze the terms, hashtags, and the overall sentiments of the tweets. In word analysis, the study looked at the words that are used in the tweets to better understand the topics that are being discussed and what the overall discussion is about within the tweets. Hashtags are an important part of every tweet, as it allows the user to label their tweets and associate it with specific categories that is relevant to users. The goal is to understand the overall content of the tweets and how it can be used to improve cybersecurity education.

### 3.2.1 Word analysis

The most popular words found in the tweets were malware (8566), ransomware (8156), cyberattack (3506), security (3356), data (1971), cybercrime (1909), attacks (1740), iot (1576), breach (1554), hackers (1518). This shows that tweets related to cybersecurity are mostly discussing about malware and attacks from hackers. Manual analysis found that many of these tweets were warnings about new malware and discussions about incidents of cyber-attacks. This suggested that the tweets are a good source of information for current events and could be potentially used as a data source for security news and tracking.

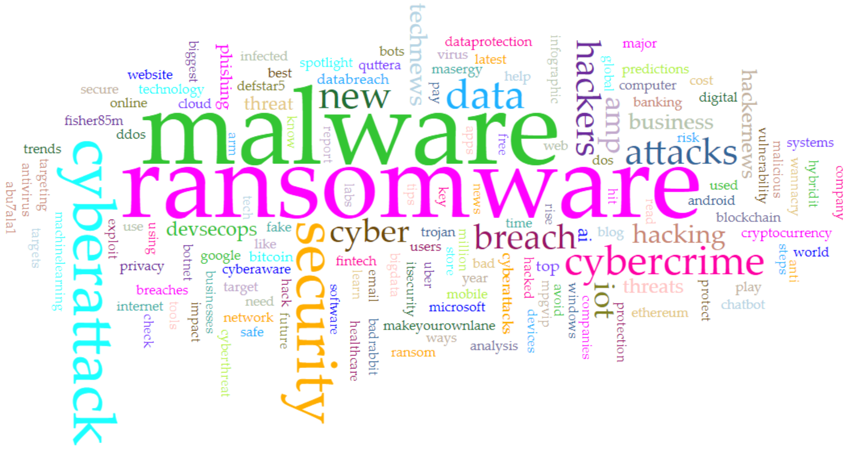


Figure 6 – Word cloud of the most popular words in tweets

In order to determine if certain words appear together with one another, n-grams were generated. As most of the tweets uses hashtags as a word in of itself, these hashtags were converted to regular terms for this analysis. Table 2 lists the top 10 bigrams and trigrams that appears within all the tweets. The findings show that the terms cybersecurity and cyberattack appears very frequently together. This suggests that both terms are synonomous with one another, which shows that tweets relating to cybersecurity are often about cyberattacks. Another top bigram has the terms bigram and ransomware, suggesting that the most common malware that occurs is a ransomware. Technews and hacker is another bigram that seems to appear often together, suggesting that these two words appear often together. The overall results however shows that many of the terms that appear together are repetitive due to the nature of the tweets. Tweets are short in nature and many of the words that appear in them are similar throughout each tweet. Therefore our analysis showed that the use of n-grams can provide some useful information but are limited in nature.

| Rank | Frequency | N-Grams (2 and 3) |
|---|---|---|
| 1 | 1575 | cybersecurity cyberattack |
| 2 | 1163 | cybersecurity infosec |
| 3 | 1017 | malware ransomware |
| 4 | 972 | cyberattack technews |
| 5 | 968 | cybersecurity cyberattack technews |
| 6 | 923 | technews hacker |
| 7 | 921 | cyberattack technews hacker |
| 8 | 871 | hacker cybercrime |
| 9 | 870 | technews hacker cybercrime |
| 10 | 830 | cybercrime hackernews |

Table 2 - Top 10 bigrams and trigrams

### 3.2.2 Hashtag analysis

About 86% of all tweets included three or more hashtags. The most popular hashtags are #cybersecurity, #infosec, #malware, #ransomware, #cyberattack, #security, amongst others.  The top two hashtags were used to collect data for this study as they were the most widely used by security professionals. The widely used #malware and #ransomware hashtags is very telling in terms of the major threats in cybersecurity during that timeframe. Malware and ransomware are malicious software that are designed to cause harm to the system that it infects, often leaving them in a state that is non-functional for use or in a severely degraded state. Ransomware is a type of malware that extort the user or organization by encrypting the system and demanding a monetary payment in order to restore it.  Both are some of the biggest threat facing computer users and organizations today. The hashtags that are highly used in tweets can serve as an important piece of information that can highlight current trends and issues in cybersecurity. This can be used to help professionals realize the current threats and understand trending areas that could be of concern.

Cybersecurity is an inter-disciplinary field that encompass a large number of subject areas that must be taught and learnt. Overtime, the number of topics that must be taught to students has grown, while the number of hours in the curriculum remained the same. This dynamic means educators must choose what topic areas they must focus on based on its importance and relevance in the always changing security landscape. Being able to see the topic areas that are important during a specific time period can help educators focus on these topics in their courses, helping them focus on the areas that are currently trending and important in the security field. The methods used in the word and hashtags analysis can be used to generate the list of the top areas within the cybersecurity field to help educators see what is happening. For example, the results from our dataset shows that malware and specifically ransomware are the most highly discussed within the tweets. Educators knowing this knowledge can adjust their course to focus a bit more about these areas as they represent some of the ongoing security threats professionals and companies are facing. By doing this, students are able to learn about topics that are relevant to ongoing events and apply it within their studies and educators are able to keep their courses more relevant.

### 3.2.3 Sentiment analysis.

This part of the study looks at the overall sentiment of the tweets. Previous research has shown that tweet sentiments can be an important indicator of current events within the cybersecurity field [21, 26, 27]. There are different algorithms that can be used for sentiment analysis, however most such as LIWC are tailored towards text that are longer in nature such as documents and full articles. The challenge with analyzing sentiments in tweets is that the tweets themselves are very short with few number of words. Sentiment classification requires as many words as possible in order to increase in accuracy. For this analysis SentiStrength is used as it supports short sentences that are common in tweets. Sentistrength uses a dictionary of words and associate those words with a point value on how negative or positive it is. Words such as “stealing”, “threat”, “attack”, are associated with negative sentiments whereas words such as “success”, “great”, and “smile” are given positive points. Figure 7 shows the breakdown of the sentiment across the entire tweet dataset of 23000 tweets.

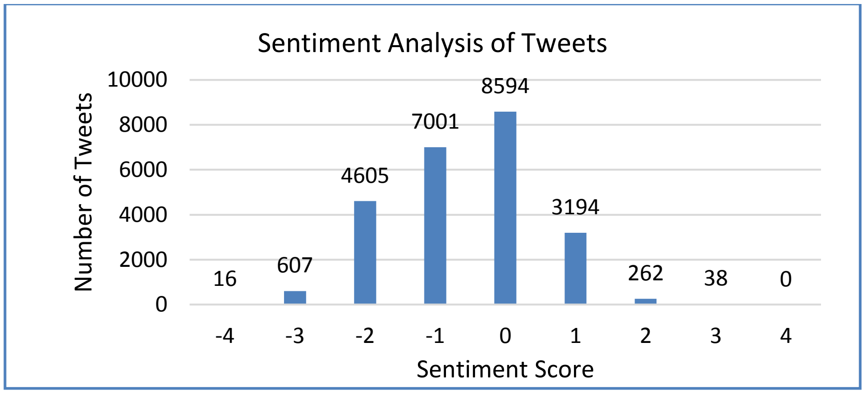


Figure 7 - Sentiment score of tweets

SentiStrength provides a score for both negative and positive sentiment in the tweet. The overall score was calculated by cominbing both scores, shown in Figure 7. The result shows that about half of the tweets had a score of 0, which means they were neutral in sentiment, in which the tweet shows neither positive or negative sentiment. The overall results showed that the tweets had more negative sentiments than positive based on the distribution of the number of tweets and their overall sentiment score, with 7001, 4605, 701, 16 tweets has a score of -1, -2, -3, -4 respectively. The lower the number, the stronger the negative sentiment is portrayed in the tweet. Comparatively, a total of 3494 tweets had a positive sentiment score. The scores indicated that most tweets relating to cybersecurity had negative sentiment. This confirms the fact that many tweets are about attacks, breaches, or warnings about security vulnerabilities. The analysis found that negative tweets are typically about hackers breaking into companies or organizations, or about those organizations being breached. Highly positive tweets (score=3) are typically announcement of awards that companies are receiving or giving. Although Sentistrength provide an overall accurate sentiment of all the tweets as a whole, individually it does not provide an accurate score for each tweet due to the way certain words are scored. For example, a tweet that says "Hackers successfully downloaded data from Citibank", could be scored as positive sentiment, even though when we read it we can see clearly it is not exactly good news. The word "success" is scored as positive, while many other words in the tweets are not scored as they are not in the dictionary. To better score sentiments for cybersecurity tweets, certain domain specific terms need to be added to the dictionary in order to more accurately portray their sentiments.

Our analysis reinforced the finding that sentiments are important indicator of the current events that are occurring. Although cybersecurity related tweets tends to be negative in nature, strong negative scores could indicate that there are major breaches and attacks occurring due to some new type of malware or vulnerability. Seeing the fluctuation in sentiment scores can provide insights on events that are occurring and enable the professional to spend more time to look deeper to see what is happening that could threaten their systems.

## 3.3 Network Analysis

### 3.3.1 Topological analysis

In order to understand the social construct of the users and the relationships between them, a network graph was created with 5459 nodes and 3533 edges. Nodes are users that mention another user as a part of their tweet or reply, and could be a single individual, an organization, or company. Edges are the relationship between these users through their mentions. The network diameter is 15, which represent the longest path between two nodes. The average path length is 5.22, showing that on average each user is 5 nodes away from each other within the entire network.

### 3.3.2 Centrality analysis

There are a few nodes that are more connected than others within the network. The degree of a node represent the number of connections that node has to other nodes. Companies such as @Masergy, @Forbes, @AppKnow, and others are the most connected to other users, with online media companies (@CSOOnline, @ZDNet, @TechRepublic, etc.) dominating these connections. Some cybersecurity experts (@Fisher85M, @reach2ratan) are also highly connected. These users serve as the central node within their communities. Many connections between users and organizations go through these central nodes, showing the importance of these nodes and their role within the cybersecurity Twitter community. These well-connected users enable them to quickly spread information every time they tweet, as their connectiveness means every time someone retweets their tweet, the information is passed on to the larger community. A user with a larger number of followers does not necessary give them the same reach compared to those who have less followers but are central nodes within the community.

### 3.3.3 Community analysis

The overall network has a low graph density (0.011), which is calculated as the number of edges divided by the number of possible edges. This suggests that the entire network is sparsely populated and that most users aren't as highly connected to each other. In order to identify communities within the network, the modularity algorithm [28] was used. This algorithm looks for nodes that are more densely connected together compared to the rest of the network. Using this algorithm, we identified two major communities and one minor one. Figure 8 shows the network graph of the community. The orange nodes represent the companies and organizations community, and the light blue nodes represent the individual user community. Organizations and companies tend to be connected to each other, as they frequently mentions other organizations in their tweets. Two companies stand out within the graph, CSOOnline is a cybersecurity focused news company, while ZDNET is a business technology news website. Both are major nodes within their respective community. CSOOnline is connected more to security organizations, whereas ZDNET have more connections to large companies and corporations. They each serve similar purposes but to a slightly different audience. Within the individual user community, @Fisher85M and @reach2ratan are both highly connected security professionals and serve as the connection point to the organization community. Many individuals within the user community do not have many degrees (connections), as highlighted by their sparse placement within the network graph. This makes the two central nodes even more important as they serve as the hub for connecting to the other communities.

There is a third minor community (dark green) that sits in between the two major ones. They include companies such as @kaspersky and @Malwarebytes. They seem to represent companies that are more connected to individual users rather than other organizations. Kaspersky and MalwareBytes are both highly regarded security software companies that are typically praised by security experts. This suggests that companies within this small community are highly trusted by individuals and explain their closer placement on the network graph, compared to other companies and organizations.

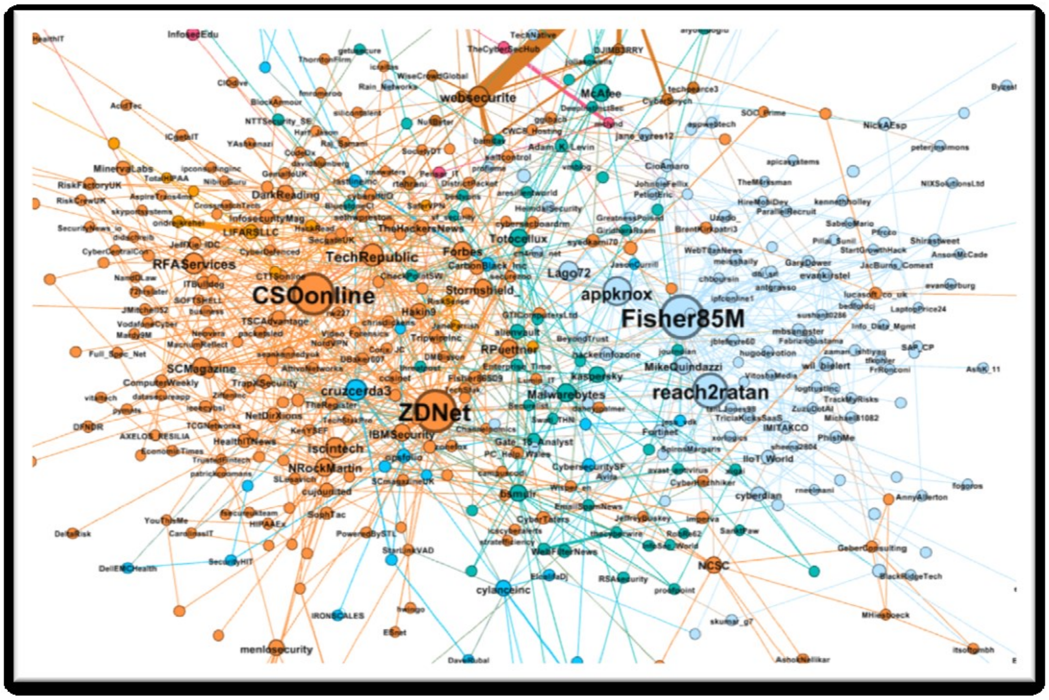


Figure 8 – Network graph showing the community of users

Although tweets are fluid in nature and users are constantly producing new tweets on a daily basis, these community of users are more stable and change at a much slower pace. Key players within the communities do not change as quickly. Students can use these information to discover influential users and companies and follow them in order to gain access to useful information. Students and professionals who are new to the security community on Twitter might not know who to follow, and there is really no easy way to find these users. Oftentimes they would look for users who has a lot of followers, but the findings from this research shows that this might not be the best approach in identifying influential and key players. Users who follow influential users rely on them for posting useful information and enable them to learn new information, which helps in ensuring that they keep up to date with the latest security news and updates.

Key players from each community tweets important but different types of information. Students and professionals can learn a lot by following specific key players from these communities. Within the companies and organizations community, CSOOline is a cybersecurity news organization that provides security news along with their analysis of the situation and how manage it. Anyone following them can learn about security events and threats as well as how to mitigate the risks. Their tweets are more geared towards security professionals and what is relevant to them. The key player within the individual community, Fisher85M, does not tweet out about security events and news. Instead, his tweets are mostly infographics that teaches about a specific topic area. The infographics are about a wide range of topics from artificial intelligence, internet of things, and to workplace topics. The infographics are simple to read and provides short lessons to users who looks at them. Users who follows him can get short informative lessons.

## 3.4 Leveraging Twitter for Cybersecurity Learning Framework

The findings from our analysis shows that Twitter data can be transformed to information that could be used to help professionals and students use Twitter more effectively to learn. Interviews with professionals from our previous research finds that many professionals want to continuously learn in order to keep up with the latest changes in the cyber security landscape, but often lack a method to do so that isn't overwhelming. Some don't even know where to even start or what website to use due to the sheer amount of sites that are available out there. Our study address this gap by proposing the use of Twitter, as information from tweets are short and concise, and can provide a high level overview of news and events.

We propose a framework to enhance learning through Twitter. The framework outlines four areas that can provide valuable information to the professional and students about the cybersecurity field. Although this research is based on cybersecurity tweets, this framework can be similarity applied to tweets from other fields as well. The four areas are topics; people, organizations, and resources.

Topics provide insights into current words and hashtags that are currently popular or trending. These are areas that are heavily discussed throughout the community regarding a specific topic or technology. Identifying these topics can help the user understand the current security landscape and the areas that they should focus on, whether for work or in their studies. Cybersecurity curriculum can often slow to change due to the large number of areas that must be taught, as well as the academic processes that are required to go through in order to update it. Students might learn about many different topics throughout the courses they take, but they might lack a specific focus on the areas that are most important or needed at the time. This part of the framework aim to address this gap by providing valuable insights to topic areas that are currently important.

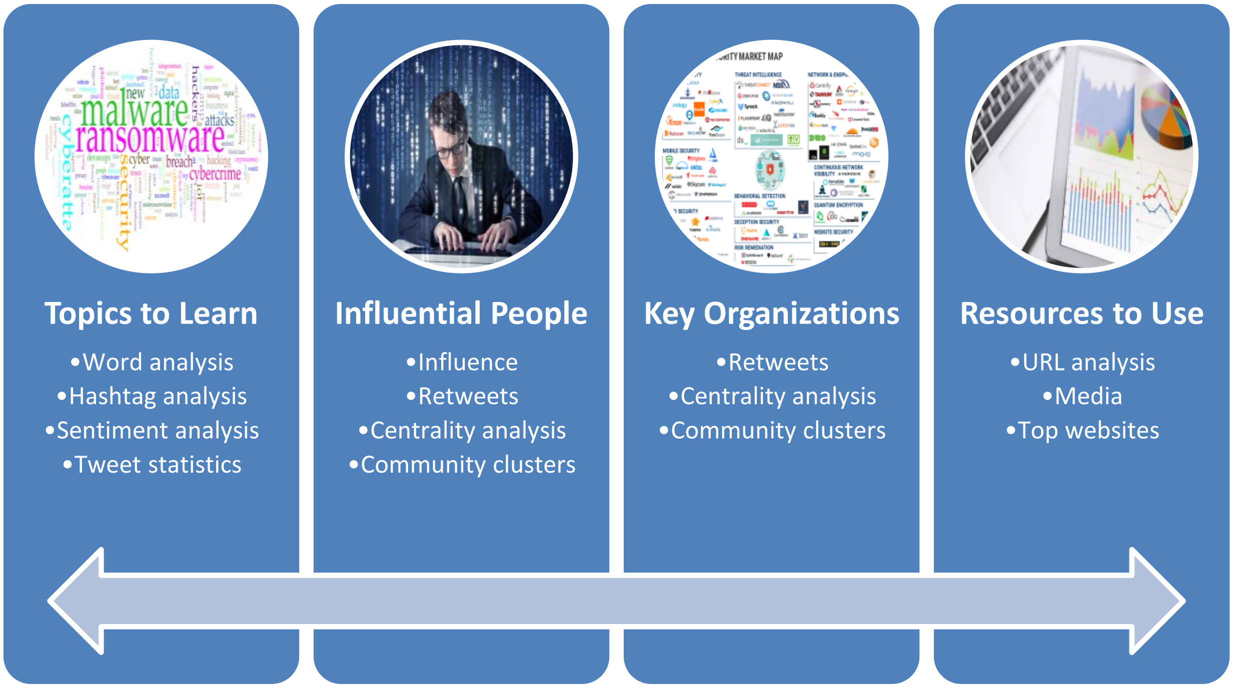


Figure 9 – Leveraging Twitter for Cybersecurity Learning Framework

Professionals use twitter to announce new findings or provide useful information for others when there are new threats. These tweets are then retweeted by other professionals and the information is passed on quickly to relevant interested parties. Highly connected professionals often acts as the information disseminator, passing on security news or informative infographics in order to educate their followers or helping them stay up to date with the latest security news. The challenge here is being able to find and identify these influential users. Professionals and students may want to follow these people, but they don't know who to follow and how to even find these people. Users typically follow people that has a lot of followers, thinking that if they have a lot of followers, it must mean they tweet out important information. Our findings show that this is not the case, and people with the most followers are typically those who are famous or have name recognition. Our research use tweet data in order to find influential people who produce high quality original tweets that are informative and can be used for learning. By following these influential people, professionals and students learn relevant information that is useful to advance their knowledge of the field.

Organizations can play a key role in helping professionals and student learn. The challenge is once again being able to find key organizations that produce high quality original content that contribute to cybersecurity community. Our finding shows that organizations with a high number of followers does not necessary mean they tweet high quality content. Organizations with the most followers tend to be high recognized companies and news organizations. Their high number of followers are typically due to their brand recognition. The research discovered that organizations that are highly connected and most retweeted have one thing in common; and that is they create original content. Many organizations often retweet article and news that are from other organizations. Whereas key organizations actually produce their own articles and tweets out information based on their own discovery. These organizations not only write about current security events, but provide in-depth analysis of the situation and provide information on how to remediate the issue or mitigate the risks related to it. So although these organizations does not necessary have the most followers, they produce high quality content that professionals and students can learn much from.

Previous research has shown that professionals use online resources extensively throughout the work day in order to solve problems and learn [29]. As there are vast amount of websites available on the internet, it can be difficult to know what websites are trustworthy and contain high quality information. Our analysis has discovered websites that are highly referenced to by the security community. These websites can be used by professionals and students who wishes to continuously learn about the field and update their knowledge. It can be especially useful for cybersecurity students, who are new to the field and often feel overwhelmed by all the available resources that are out there. The findings from our study can narrow all these resources into something that is more manageable to help aid the learning process.

Security news and events can occur at a rapid pace, and while news articles can be a reliable source of information, they require extensive writing and editing before they can be published. This makes tweets an alternative source for current events as it is more informal and can be disseminating more quickly. For example, during the WannaCry ransomware incident which spreads quickly throughout many organizations, Twitter was an effective platform for disseminating this type of information. This suggest that Twitter can be an effective tool for learning new information that are more fluid and change at a rapid pace.

### 3.4.1. Framework Validation

In order to validate the framework and get feedback on its effectiveness, a qualitative interview study was done. We interviewed ten cyber security professionals who were willing to take part in the study. Most of these professionals have been previously interviewed for a previous related work regarding how they use online and interpersonal resources to learn and seek information.

The interviews were carried out twice in a period of two weeks. In the first interview, we explained to them the basis and objective of our study and our current findings so far. We then introduced them to our "Leveraging Twitter for Cybersecurity Learning Framework" and provided the results from our study and explained to them how they could use the information to assist them in learning. Half of the participants did not have a Twitter account so we worked with them to create one. We provided results for each of the four pillars of the framework and provided them information that they could use. We provided them a list of users for them to follow based on the "Influential People" and "Organizations to Follow" pillars of the framework. The goal was to update their Twitter feed to receive high quality tweets from the users we have identified. We asked that they check Twitter on a regular basis for a period of two weeks and provide us feedback on what they saw. After two weeks, we met with them again for the second interview which was to collect the feedback.

Out of the ten professionals we interviewed, seven said that they looked at their Twitter feed once every 2-3 days, while three said they only looked at it once during that time period. The three that looked at it once explained that they were just too busy and didn't really get a chance to look at it, and the fact that they normally don't use Twitter on a regular basis. The following table shows a summary of the key findings based on the interviews.

| Positives | Negatives | Other |
|---|---|---|
| All the tweets and related news and information are all in one place | Not all tweets are related to cyber security, especially from companies | Individual users post more useful and relevant cyber security tweets |
| You can learn a lot from tweets with infographics | Some tweets shown are "liked" by the user, and not their original tweet | Some news organizations post a lot of non-relevant articles |
| Security events and news are current and relevant | Not all influential users produce relevant security tweets | Need a way to only show tweets with a specific hashtag from a user |
|  |  | A lot of tweets are related to future technology that might not be relevant today |

Table 3 - Feedback from interview study

The overall feedback from the participants were positive. They liked how using the framework allowed them to see who the influential users and organizations are that they can follow. They mentioned that as professionals, they do not know how to fully use Twitter. There is a somewhat of skepticism on using the platform professionally due to the negative perception on using social media. In cyber security, social media are often frowned upon as it could be seen as a security risk for both the user and the organization they work for. In addition, participants expressed how they feel Twitter can be overwhelming, as there are millions of potential users that they could follow. Common sense would tell them to follow users with the most followers, but this only increased the amount of content that would fill up their feed, making it hard to see relevant content that they cared about. By providing them with the list of users based on data analysis methods, they are able to identify key users that could provide relevant information for them.

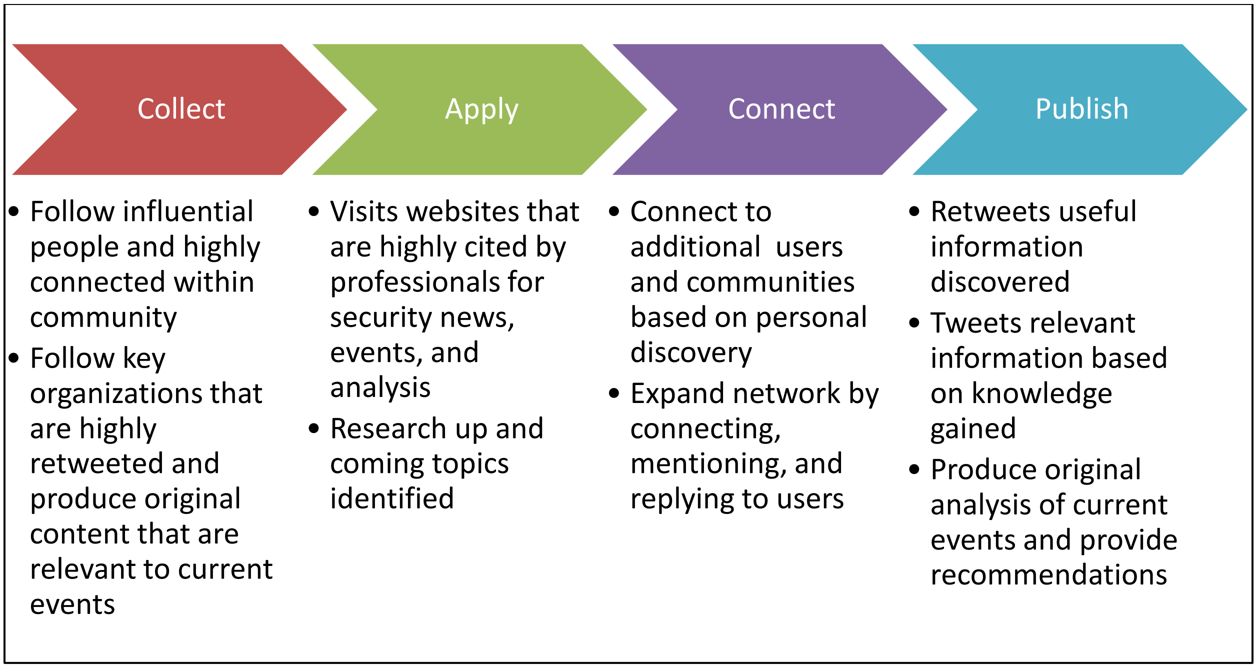
On the negative side, the participants had complaints on the feed they were reading. Many of these are limitations on Twitter itself in the way it display the tweets. A common complaint was the fact that they would often see tweets that the user has "liked", or tweets that are not related to cyber security. Even though the majority of the tweets are cyber security related, users would often tweet out other non-related tweets from other topics. Within the Twitter feed, there is no way to filter out these types of tweets that are not relevant to the user. So although influential users would often tweet about cyber security, there were times when they did not. The participants would like to have a method to be able to view only tweets from those users with the proper cyber security hashtags, and ignore all other tweets without the relevant hashtags. There are third party websites that helps narrow down tweets based on certain criteria, but there is no method to do these types of filtering in an easy and intuitive manner. The negative feedback has more to do with the issues of the Twitter platform and its limitation, rather than from our framework and the information we have provided.

These findings show that our framework can help provide a useful method for cyber security professionals to learn and keep up to date with the latest security news and information. It shows that using data analytical methods on Twitter data can provide valuable information that can be given to professionals in order to assist them to be able to use Twitter more effectively as a learning tool. In addition, the feedback on the limitation of Twitter itself can provide valuable insights on how the platform can be improved in order to be more useful for professionals, and thereby increase its usage amongst them.

## 3.5. Personal Learning Environment

The framework described can be utilized as a part of a personal learning environment (PLE). Dabbagh & Kitsantas looked at using social media in order to facilitate learning and created a framework for self-regulated learning in personal learning environments [30]. The framework uses scaffolding to help students learn by engaging and interacting with social media such as blogs, wikis, and social networking sites. The framework contains three levels of interactivity between the student and the social media tool. Each level contains a list of instructions that students are asked to perform in order to "enable the degree of interaction and sharing desired and/or required for learning". We expand upon this framework by providing a methodology for using Twitter as a personal learning environment for the cyber security field.

The following methodology outlines the four steps that students can follow in order to effectively use Twitter as their personal learning environment. It follows Hiebert's model of PLE [31] by providing a four step process that students can follow information generated from the "Leveraging Twitter for Cybersecurity Learning Framework". The four steps uses scaffolding to help students learn and be engaged. This enables them to not only learn new information and content that is relevant to the field, but also helps connect them to key players, organizations, and the cyber security community as a whole. As networking is an important aspect of the profession, this two-pronged approach helps the student prepare by gaining knowledge and connecting to those in the community.



Collect

- Follow influential people and highly connected within community
- Follow key organizations that are highly retweeted and produce original content that are relevant to current events

Apply

- Visits websites that are highly cited by professionals for security news, events, and analysis
- Research up and coming topics identified

Connect

- Connect to additional users and communities based on personal discovery
- Expand network by connecting, mentioning, and replying to users

Publish

- Retweets useful information discovered
- Tweets relevant information based on knowledge gained
- Produce original analysis of current events and provide recommendations

Figure 10 –Methodology applying the framework for personal learning environment

Personal learning environments are typically geared towards each individual based on their personal goals and motivations, and in this case it is specially geared towards students and those who are in the cyber security field. More research will need to be done in order to test and validate the methodology on students in order to determine its effectiveness. This will be part of a future work that we will look at in order to expand the use of the framework for students as they prepare to enter the workforce.

## 4 Discussion

The findings from the research shows that Twitter can be a very useful tool to help security professionals stay up to date with the latest events and changes. Cybersecurity users typically tweet about information about current events or announce about new findings within the security domain. These tweets almost always contain an URL that links to an external website that contains more details about the information that is presented. Compared to tweets from other domains, these tweets are less conversational and more information focused. Hashtags and words that are used within the tweets can identify important topics that are currently relevant or becoming the topics that are important or soon to be important. For example, the highly used #ransomware hashtag shows it is the prominent threat that many individuals and companies are facing today. Ransomware has hit organizations hard by taking hostage of their most important asset, their data. Being able to see trends in the hashtags could give organization a heads up on potential security threats and help them prepare for the threats that are coming.

Twitter data can be used to generate valuable information that can help professionals and students learn and stay up to date with the latest security news and events. Descriptive analytics can be used to find important tweets and statistics about those tweets. Network analytics can be used to find key players within the community and identify important individuals who are knowledgeable and is influential within the community. Content analytics can be used to discover trending security topics that might affect the security landscape for companies and organizations, or discover upcoming threats that requires proactive or additional learning. The findings from this study led to the creation of a framework that can be used by professionals and students learn and advance their knowledge in the cyber security field. Although the framework is created based on cybersecurity tweets, the same can be used and applied to other fields as well.

The initial feedback from the interview study that was done to validate the framework and findings were positive. Overall the participants found that the framework helped made Twitter more useful in terms of using it as a tool for continuous learning. They were only able to test three of the four areas that were identified in the framework. Further research is needed to determine the overall benefits and effectiveness of the knowledge that is gained through the use of the framework. In addition, more research is needed on whether or not the top websites identified in the "Resource to Use" pillar are better and more effective than other websites when used as a resource for learning.

Professionals can use Twitter to build up a reputation for themselves in order to be noticed or recognized by companies and organizations. A professional who has a sizable number of followers can show that they are an expert in the field, and advertise such status to potential employers. These professionals can build up followers by posting useful and relevant information and be engaged to their peers through the use of replies and retweets. These engagements allow professionals to communicate with people who might not be directly within their network, and allow the expansion of their overall network and connect to others around the world.

In relation to the workplace practices related literature in engineering education, the findings from this work extend prior work that has established how professionals engage with technical communities and learn from them [32] by showcasing that similar learning through engagement with social media also takes place. This research extends the largely organization and placed-based research on workplaces that have previously been reported in the engineering education literature [33-37].It adds to research that has argue for a more 'open organizing' perspective that takes into account that engineering and technology professionals work in organizational fields – which consist of professionals from different organizations – and not just in a specific company or organization [38-39]. This means that they contribute to a larger community by sharing expertise and knowledge.

# 5 Conclusion

This study has shown that data from Twitter can be used to generate valuable information that can be used to help cybersecurity professionals and students learn beyond the classroom. It addresses one of the major gap within cybersecurity education program by providing a continuous learning method that is relevant based on the current events that are taking place. Beyond looking at each individual tweet, analyzing the data collectively as a whole can yield valuable information for learning and gaining knowledge.

# Acknowledgements

This work was partially funded by NSF Awards#1707837 and 1712129. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

# REFERENCES

[1] IBM Center for Applied Insights. *Cybersecurity education for the next generation*.

[2] McGettrick AD, Cassel LN, Dark M, et al. Toward curricular guidelines for cybersecurity. In: *SIGCSE*, pp. 81–82.

[3] Woodward BS, Young T. Redesigning an information system security curriculum through application of traditional pedagogy and modern business trends. *Inf Syst Educ J* 2007; 5: 1–11.

[4] McDuffie E. *National Initiative for Cybersecurity Education*. NIST, http://csrc.nist.gov/nice/documents/nicestratplan/nice-strategic-plan_sep2012.pdf (2012, accessed 29 October 2014).

[5] Association for Computing Machinery (ACM), IEEE Computer Society (IEEE-CS), Association for Information Systems Special Interest Group on Information, et al. Cybersecurity Curricula 2017, https://www.acm.org/binaries/content/assets/education/curricula-recommendations/csec2017.pdf (accessed 22 April 2019).

[6] Swire P. A pedagogic cybersecurity framework. *Commun ACM* 2018; 61: 23–26.

[7] National Centers of Academic Excellence (CAE) | National Initiative for Cybersecurity Careers and Studies, https://niccs.us-cert.gov/formal-education/national-centers-academic-excellence-cae (accessed 22 April 2019).

[8] Aslam S. • Twitter by the Numbers (2018): Stats, Demographics & Fun Facts, https://www.omnicoreagency.com/twitter-statistics/ (2018, accessed 10 May 2018).

[9] Achrekar H, Gandhe A, Lazarus R, et al. Predicting flu trends using twitter data. In: *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 2011, pp. 702–707.

[10] Gayo Avello D, Metaxas PT, Mustafaraj E. Limits of electoral predictions using twitter. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence, 2011.

[11] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of twitter data. In: *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, 2011, pp. 30–38.

[12] Widmer RJ, Engler NB, Geske JB, et al. An Academic Healthcare Twitter Account: The Mayo Clinic Experience. *Cyberpsychology Behav Soc Netw* 2016; 19: 360–366.

[13] Mittal S, Das PK, Mulwad V, et al. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In: *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 860–867.

[14] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. *J Comput Sci* 2011; 2: 1–8.

[15] Chae B (Kevin). Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *Int J Prod Econ* 2015; 165: 247–259.

[16] Ben-Dov M, Feldman R. Text Mining and Information Extraction. In: Maimon O, Rokach L (eds) *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, pp. 809–835.

[17] Feldman R, Fresko M, Kinar Y, et al. Text mining at the term level. In: Żytkow JM, Quafafou M (eds) *Principles of Data Mining and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 65–73.

[18] Mooney RJ, Bunescu R. Mining knowledge from text using information extraction. *ACM SIGKDD Explor Newsl* 2005; 7: 3–10.

[19] Hearst MA. Untangling text data mining. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, pp. 3–10.

[20] Tan A-H, others. Text mining: The state of the art and the challenges. In: *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, p. 65.

[21] Patodkar VN, I.R S. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *IJARCCE* 2016; 5: 320–322.

[22] Gilbert E, Karahalios K. Predicting tie strength with social media. ACM Press, p. 211.

[23] Knoke D, Yang S. *Social Network Analysis*. SAGE, 2008.

[24] Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc Netw* 2010; 32: 245–251.

[25] Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci* 2006; 103: 8577–8582.

[26] Gupta B, Sharma S, Chennamaneni A. Twitter Sentiment Analysis: An Examination of Cybersecurity Attitudes and Behavior. 11.

[27] Jurek A, Bi Y, Mulvenna M. Twitter Sentiment Analysis for Security-Related Information Gathering. In: *2014 IEEE Joint Intelligence and Security Informatics Conference*. The Hague, Netherlands: IEEE, pp. 48–55.

[28] Blondel VD, Guillaume J-L, Lambiotte R, et al. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008; 2008: P10008.

[29] Le H-T, Johri A, Malik A. Situated Information Seeking for Learning: A Case Study of Engineering Workplace Cognition among Cybersecurity Professionals. *Proceedings of ASEE 2018*.

[30] Dabbagh N, Kitsantas A. Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning. *Internet High Educ* 2012; 15: 3–8.

[31] Nejdl W, Tochtermann K. *Innovative Approaches for Learning and Knowledge Sharing: First European Conference on Technology Enhanced Learning, EC-TEL 2006, Crete, Greece, October 1-4, 2006, Proceedings*. Springer, 2006.

[32] Teo, H., Johri, A. & Lohani, V. Analytics and Patterns of Knowledge Creation: Experts at Work in an Online Engineering Community. *Computers & Education*, 2017, Vol. 112, pp. 18-36.

[33] Johri, A. Impressions in Action: The Socially Situated Construction of Expertise Impressions in the Workplace. *Journal of Organizational Ethnography*, 2015, 4(1): 44-63.

[34] Johri, A.Engineers' Knowing in Practice: Reconciling Sociality and Materiality through Action. Fenwick, T. & Nerland, M (Eds.). *Reconceptualising Professional Learning in Turbulent Times: changing knowledges, practices, and responsibilities*. Routledge. 2014.

[35] Stevens, R., Johri, A. & O'Connor, K. Professional Engineering Work. Johri, A. & Olds, B. (Eds). *The Cambridge Handbook of Engineering Education Research*, Cambridge University Press, New York, NY. 2014.

[36] Johri, A. Learning to Demo: The Sociomateriality of Newcomer Participation in Engineering Research Practices. *Engineering Studies*, 2012, Vol. 4, Issue 3, pp. 249-269.

[37] Johri, A.  Situated Engineering and the Workplace. *Engineering Studies*, Dec. 2010.

[38] Pipek, V., Wulf, V. & Johri, A. Bridging Artifacts and Actors: Expertise Sharing in Organizational Ecosystems. *Journal of Computer Supported Cooperative Work*. 2012, 21(2-3):261-282.

[39] Johri, A. Open Organizing: Designing Sustainable Work Practices for the Engineering Workforce. *International Journal of Engineering Education*, 2010, 26(2):278-286.