Sec-Buzzers: a Web Service for Exploring Cyber Security Emerging Topics based on Social Network Mining

Chih-Hung Hsieh[*]
Institute of Informaiton Industry
Taipei, Taiwan
chhsieh@iii.org.tw

Kuo-Chen Lee[†]
Institute of Informaiton Industry
Taipei, Taiwan
kclee@iii.org.tw

Ching-Hao Mao[‡]
Institute of Informaiton Industry
Taipei, Taiwan
chmao@iii.org.tw

Chia-Min Lai[§]
Institute of Informaiton Industry
Taipei, Taiwan
senalai@iii.org.tw

Chiun-How Kao
Institute of Informaiton Industry
Taipei, Taiwan
maokao@iii.org.tw

Jyun-Han Dai[¶]
Institute of Informaiton Industry
Taipei, Taiwan
markdai@iii.org.tw

ABSTRACT

Recognition of information threats from social media can give advantages to incident response in very early stage. Previous related studies mostly focus on finding general hot terms instead of specific continuously-changing targets, such that usage of these methods may be limited when given specific theme as default. To our best knowledge so far, the proposed Sec-Buzzers is the first web-based service not only dedicated to finding the various emerging topics of cyber threats (i.e., nearly zero-day attacks) but also providing the possible remedy solutions. Unlike previous works, Sec-Buzzers mainly benefits from the strategy of community-oriented resource filtering and a novel modified topic association graph, such that a set of highly-contributing Twitter users was grouped, and information from that was explored then exploited. Demonstrations show that, by combining several measurements to quantify significance, Sec-Buzzers indeed uncovered emerging (or suddenly appearing) issues which are highly related to real cases about security threats.

[*]who help fine tune the kernel mechanism, organize content and demonstration of this publication, draft and revise the paper.

[†]who develop and implement the kernel mechanism.

[‡]who first innovate the idea of this work, and start up the whole team for research and development.

[§]Additional affiliation: Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology in Taipei, Taiwan

[¶]who is corresponding author and project manager responsible for development progress.

ASE BD&SI 2015, October 07-09, 2015, Kaohsiung, Taiwan

DOI: http://dx.doi.org/10.1145/2818869.2818897

Availability: Sec-Buzzers is freely available over the Internet at http://sec-buzzers.ctti.iii.org.tw
Contact: sec.buzzers@gmail.com

CCS Concepts

•Information systems → Document topic models; Web and social media search;

Keywords

Cyber Security, Information Retrieval, Social Network Mining, Information Security, Web Service, Online Tool, Data Mining

1. INTRODUCTION

With an endless information stream of general purposes, automatic recognition of incoming topic threads from social network (e.g. Twitter or Facebook) will help to accelerate making rapid response in the early stage. Among various kinds of the social media, Twitter has been wildly studied for mining hot topics from microblogs cause its open accessibility [16][7][13][12]. Zubiaga [7] propose a typology to categorize the triggers that leverage trending topics: news, current events, memes, and commemoratives. Furthermore, they define a set of straightforward language-independent features that rely on the social spread of the trends to discriminate among those types of trending topics. However in social network In order to handle divorced and ambiguous semantic meaning, Diakopoulos et al. [13] develop and investigate new methods for filtering and assessing the verity of sources found through social media by journalists. Gnanasambandam et al. [12] intends to derive patterns relating to spatio-temporal traffic flow, visit regularity, content and social ties as they relate to an individual's activities in an urban environment.

On the other hand, compared to other applications of topic finding from social network, such as public events (e.g., Olympics) or business (e.g., the release of a new smartphone), uncovering emerging hot issues of cyber security domains in time is especially worth lots of values. For instance, in March 2014, a Turkish hacker crash Google Play

store twice for the purpose of testing the vulnerability he found after announcing on his microblog [1]. However, due to diversity of cyber threats, finding emerging topic in information security is also quite different compared to applications with specific subjects (e.g., advertisements for the christmas sales, or news for a political campaign). Most of previous related studies or currently available on-line service focus on finding general hot terms instead of specific targets, such that the usage of these works needs given keywords as input in advance and is limited when only category terms (e.g., information security emerging threads) instead of specific keywords(e.g., advanced persistent threat, APT) are available.

In this study, we proposed Sec-Buzzers, an on-line community oriented emerging topic finding service from twitter data, which can assist enterprises or organizations to timely recognize the security threats. To our best knowledge so far, the proposed Sec-Buzzers is the first web-based service not only dedicated to finding the various emerging topics of cyber threats (i.e., nearly zero-day attacks) but also providing the possible remedy solutions. Unlike previous works, Sec-Buzzers mainly benefits from two aspects: 1) the strategy of community-oriented resource filtering, where a set of Twitter users highly-related to cyber security was grouped as expert community; 2) a novel modified topic association graph wherefore information from that was easy explored then exploited. The proposed system consists of three components, i.e., an efficient and scalable social media connector, weighting community-related Twitter users, and emerging topics uncovering. The first part makes routinely and automatically data collecting realized, The second part captures the followships between Twitter users and measures the influence or importances of each Twitter user in expert community. And the last one is responsible for uncovering emerging topic of cyber security according to the weights of Twitter experts, the significances of extracted terms used in information security, and other measurements representing the changing on the trends or lifecycle of those terms.

In order to demonstrate the effectiveness of our framework, two well-known real cases of cyber security issues are found by our system and discussed in this work. One is "Venom": a zero-day security exposure threat in virtualized environment [6], and the other is "Duqu 2.0": the most sophisticated malware ever seen [3]. Both cases show that by 1) collecting Twitter data from a hundred of famous information security experts as domain expert community, 2) maintaining the instant topology describing followships among experts and tweets, and 3) combining several measurements to quantify significance of all retrieved terms, Sec-Buzzers indeed is able to uncover emerging (or suddenly appearing) hot issues with instant remedy solution bundled, if available.

The remainder of this paper is organized as following. In Section 2, we introduce our proposed system Sec-Buzzers, including the collected expert pool, the system framework, and used mining algorithm. Section 3 shows the details about two demonstration case studies mentioned above. Section 4 gives the precision evaluation and proof of concept experiment for the future improvement of Sec-Buzzers. Finally, we concludes this work in Section 5.

2. ARCHITECTURE OF SEC-BUZZERS

In this section, we describe the kernel approach of Sec-Buzzers, which is responsible for information security emerging topics finding, for finding the emerging topics of information security threats. Unlike previous works, the high effectiveness of Sec-Buzzers mainly benefits from three components. The first one is 'social media connector', a system component implement stuffs for handling tweet contents. The second is by forming a weighted highly-contributing expert pool to extract content precisely related to cyber security. And the last one is by exploring then exploiting information from expert pool to recognize emerging threats with an efficient topic detection algorithm proposed in [10]. Details about the components included in Sec-Buzzers can be found in the following sub-sections and in the references [10][8].

2.1 Social Media Connector

For the purpose of gathering Twitter dataset, we built a specific social media connector. The main concept of our connector is that if a Twitter user saw a tweet he was interested in, then he would retweet the tweet into his timeline to share the content with others.

How the social media connector being crafted is shown in Figure 1. Followings are the main steps of how social media connector works.

1) Multiple threads, which are managed with master-slave structure, periodically request storage database for a Twitter access token.

2) The database will return access tokens if available, accompanied with an expert account pending updating to threads.

3) Creating a data channel connection from Twitter to Sec-Buzzers is the third step, "Twitter REST application program interface (API)"[4] - a variant of Twitter RESTful API [17] is used to crawl all needed contents of the given Twitter accounts. The reason for using REST API, instead of alternative Twitter Streaming API, is to facilitate handling exception case as system abnormal shut down happens [5].

4) Once getting new arrival content from Twitter, we formalize and store these to structured information implemented with elastic database. The pre-defined contents to be crawled are tweet, retweet, authorships for tweet or retweet, followships among experts, and corresponding timestamps of each information.

2.2 Expert Authority Weighting

One of the key idea adopted by Sec-Buzzers is the community-oriented strategy for content resource filtering. The hypothesis of this strategy is that a set of Twitter users, who are deemed highly involved in cyber security domain, was grouped as expert community. Therefore, Sec-Buzzers devotes to explore and exploit information from those expert. Currently, the expert pool used in Sec-Buzzers contains one hundred most active cyber security experts, which was labeled as must-follow security experts in Twitter referenced to survey of David Jevans, the Chairman and founder of cyber security enterprise "IronKey" [8][2].

As retrieving hot terms from contents of pre-defined experts, it will be better if we can measure importance of

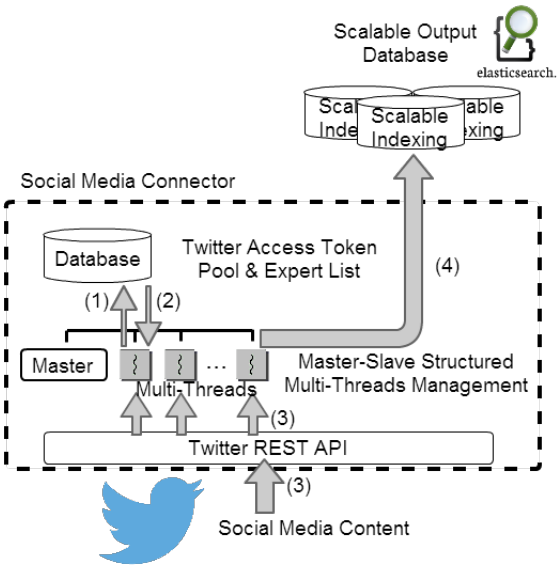

Figure 1: Working flow of social media connector with four major steps.

terms according to the authority of each experts who mentioned these terms. Two naive but very useful hypotheses are that 1) if an expert is tracked by another expert with great impact, the weight of expert being followed should be increased; and 2) the more people's attention an expert can attracts, the more credit or authority this expert should be gave. A directed graph, $G(A, F)$, can be used to describe the followship among Twitter accounts crawled by social media connector and to realize the above assumptions. The vertex set A of $G(A, F)$ represents the expert accounts included in Sec-Buzzers. The edge set F of $G(V, F)$ is used to describe the followship among experts, where $\forall a_1, a_2 \in A$ and $f(a_1, a_2) \in F$ means that expert a_1 follows expert a_2.

In Sec-Buzzers, the initial setting of expert weight/authority is set as $weight^0(a_i) = 1/|A|$. The subsequent updates of each weight, according to Page Rank Algorithm [10][14], is as following:

$$\forall a_i, a_j \in A, \text{ and } a_j \in follower(a_i)$$
$$weight^t(a_i) = d \times \sum_{a_j} \frac{weight^{t-1}(a_j)}{|followee(a_j)|} + (1-d), \quad (1)$$

where d is a learing factor used for two purposes: 1) giving each expert a minimum weight, $(1-d)$; and 2) adjusting the proportion of consideration between minimum weight $(1-d)$ and weights of followees.

2.3 Emerging Topic Recognition

In this component, Sec-Buzzers incorporate and fine tune the algorithms of [10] to detect the emerging **topic**. The first step to recognize emerging topic is to recognize the emerging **terms** from extracted tweets. In regular case, an article or short comment contain several common or stop words such as "is", "am", or "OK", etc. These terms appear with high frequency, however, usually irrelevant to specific topics. For this reason, the definition of "emerging" term are **keywords** which appear less frequently but could become hot because its highly relevance to one topic. Here, Sec-Buzzers leverage **Nutrition Calculation Algorithm** and **Energy Calculation Algorithm** to identify the hot terms for building up hot topics. The nutrition of one term is a kind of weighted score considering frequencies and authorities of eq. (1) corresponding to each tweet, while energy is used to describe the extents of how suddenly a term emerging. Details about these two algorithm can be found in [10].

After processes mentioned above, a modified variant of emerging topic detection algorithm proposed in [10] is now adopted here. For each time interval t, Sec-Buzzers rank and select top n^t_{top} terms/keywords in terms of their estimated energies to form a set S^t_{top}, where $|S^t_{top}| = n^t_{top}$. Our goal is to realize whether there exists common concepts and meanings among subsets of terms, such that we can group them as so called "topics". The first step is to investigate the co-occurrence situations of terms. An idea of correlation vector referenced from [18] leverages a set of weigthed terms describing the co-occurrence relationships that exist from one term to all the others in the considered time interval. The correlation weight $c^t_{k,z}$ respecting to keywords $k, z \in S^t_{top}$ given time interval t can be formalized as following:

$$\begin{aligned}
&\forall k, z \in S^t_{top} \\
&c^t_{k,z} = \alpha^t_{k,z} \times \beta^t_{k,z} \\
&\alpha^t_{k,z} = \left| \frac{|TW^t_k \cap TW^t_z|}{|TW^t_k|} - \frac{|TW^t_z - (TW^t_k \cap TW^t_z)|}{|TW^t - TW^t_k|} \right| \\
&\beta^t_{k,z} = log \frac{\gamma^t_{k,z}}{\delta^t_{k,z}} \\
&\gamma^t_{k,z} = |TW^t_k \cap TW^t_z| \cdot |TW^t - (TW^t_k \cup TW^t_z)| \\
&\delta^t_{k,z} = |TW^t_k \cap (TW^t - TW^t_z)| \cdot \\
&\qquad |TW^t_z \cap (TW^t - TW^t_k)|,
\end{aligned} \quad (2)$$

where

TW^t: the set of all tweets in time interval t.

TW^t_k: the set of all tweets containing keyword k in time interval t.

TW^t_z: the set of all tweets containing keyword z in time interval t.

$TW^t_{k,z}$: the set of all tweets containing both keywords k and z in time interval t.

By explaining equation (2), it can be observed that:

1) $\alpha^t_{k,z}$ compares the posterior probabilities of keyword z's occurrence, given events of keyword k's occurrence or non-occurrence, respectively. If $\alpha^t_{k,z} \neq 0$ means occurrences of k and z is not independent, otherwise they are independent.

2) $\beta^t_{k,z}$ considers the numbers of tweets that k and z are co-occurring or co-absent (i.e., $\gamma^t_{k,z}$), and numbers of tweets that containing one and only one keyword (i.e., $\delta^t_{k,z}$).

After forming a set C^t_{top} of correlation weights for all pairs of terms in S^t_{top}, instead of original topic graph used in [2], a modified topic graph $MTG^t = (V^t_{MTG}, E^t_{MTG}, C^t_{MTG})$ was proposed as following:

V^t_{MTG}: the vertex set $V^t_{MTG} = S^t_{top} \cap \tilde{S}^t_{top}$ of MTG^t includes the set S^t_{top} as well as a set $\tilde{S}^t_{top}$ containing antonym terms against to all terms in S^t_{top}, e.g., "not malware" $\in \tilde{S}^t_{top}$ against to "malware" $\in S^t_{top}$.

E^t_{MTG}: E^t_{MTG} is the edge set of MTG^t. For any pair (k, z) of terms in S^t_{top} with correlation weight $c^t_{k,z} > 0$, there is one edge linking from term k to term v with weight $c^t_{k,z}$. While, for case of $c^t_{k,z} < 0$, there also exists an edge connecting from term k to term "not v" with positive weight $(-1 \cdot c^t_{k,z})$.

C^t_{MTG}: This contains all weights annotating edges in E^t_{MTG}. It should be noted that all weights in C^t_{MTG} are positive.

The strongly connected components contained in MTG^t will group related terms together and can be viewed as topics. The major advantages of using MTG^t are twofold. 1) All edge weights in MTG^t are positive such that the inconvenience and confusion when finding strongly connected components in topic graph with negative-weighted edges can be alleviated. 2) The created term "not v", in fact, help users to understand the appropriate topics from the resulted strongly connected components. For example, if one strongly connected component generated from MTG^t is {"Apple", "on-sale", "not fruit"}, it is much easier for users to know that the corresponding topic may be "apple - the famous company established by Steve Jobs". Sec-Buzzers uses an efficient algorithm to find strongly connected components [15] which is implemented in well known library "SciPy"[9].

3. CASES DEMONSTRATIONS

To demonstrate the effectiveness of Sec-Buzzers, this section introduces two real cases founded by Sec-Buzzers: 1) "Venom" and 2) "Duqu 2.0". The scopes of selected cases are from severe system vulnerabilities to novel malware virus never seen. Both of them may cause lots of personal or corporate damage and are worth studying. The parameter settings of current Sec-Buzzers system are as followings: 1) a time interval for updating uses setting of three days; 2) the learning factor d in equation (1) is set to 0.85.

3.1 "Venom" - a zeroday security exposure threat in virtualized environment

The term "venom" is the abbreviation for "virtualized environment neglected operations manipulation", which first appeared on May/13/2015, then was officialy named as that [6]. The venom, CVE-2015-3456, is a security vulnerability in the virtual floppy drive code used by many computer virtualization platforms. Hackers may make use of venom to transcend the guest limitation of an affected virtual machine (VM) and potentially obtain code-execution access right to the host. Without fixing this VM vulnerability of a host, hacker could even further to access other VMs on the same host or all other adjacent systems, by elevating access right. Venom vulnerability can leads to severe damage of exposing access to corporate intellectual property (IP), as well as sensitive and personally identifiable information (PII), potentially impacting the thousands of organizations and millions of end users that rely on affected VMs to allocate shared computing resources, connectivity, storage, security, and privacy.

Figure 2 shows the topic discovery corresponding to venom vulnerability by Sec-Buzzers. The histogram shows the number of tweets that Sec-Buzzers collected on different dates. The different stages are labeled and shown as:

a) With efficient social media connector, precise content sourcing, and then topic discovery, Sec-Buzzers realized that there had been a set of emerging tweets all related to a system vulnerability hiding in flopy disk code of VM. The date that Sec-Buzzers finding this topic is as soon as on May/13/2015.

b) Still on May/13/2015, this vulnerability was officially renamed as "venom" by most of experts on Internet, while Sec-Buzzers was able to find and point out that the vulnerability on flopy disk code of VM and the term of "venom" share the same concept and meanings.

c) On May/14/2015, the next day of VM vulnerability being noticed, Sec-Buzzers subsequently detected another hot sub-topic belonging to venom: a proof-of-concept exploit code was released.

d) On May/15/2015, the 3^{rd} day from the beginning, Sec-Buzzers found some highly-cited feature articles of venom.

e) After 5 days later, Sec-Buzzers recognized another emerging trend of sub-topic that the Oracle company released official patch for fixing venom problem. Sec-Buzzers also collected related information and provided assoicated download links on our web site.

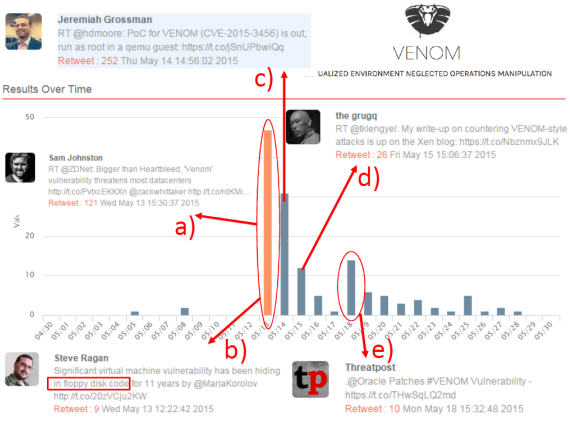

Figure 2: The topic discovery of Sec-Buzzers for venom vulnerability

3.2 "Duqu 2.0": the most sophisticated malware ever seen

As a highly sophisticated malware which is a variant of the notorious Duqu malware appeared in 2012, Duqu 2.0 successfully compromised the virus protection giant Kaspersky and shocked all the IT security in 2015 by report from famous Infosec institute [3]. Duqu 2.0 was described by security researchers as a malware of advanced persistent threat (APT) that exploited lots of zero-days vulnerabilities listed in at least three common vulnerabilities and exposures (CVE) reports: 1) CVE-2015-2360, 2) CVE-2014-4148, and 3) CVE-2014-4148.

According to experts' investigation, the developer of Duqu 2.0 adopts sophisticated evasion techniques. First, it resides in memory, making detecting itself difficult. Second, experts at Symantec explained that Duqu 2.0 comes in two variants, one is a backdoor be used to gain persistence in the targeted entity by infecting multiple computers. Another variant represents its evolution, and implements more sophisticated features. And the most important, authors of Duqu 2.0 used a stolen certificate from the Foxconn company to implement a persistence mechanism and remain under the radar.

Figure 3 is an illustration of how Sec-Buzzers detected the topic about Duqu 2.0 malware. There are two peaks in the trend of associated tweets number depicted in Figure 3:

a) The 1^{st} peak suddenly arose on June/10/2015 when the news of that famous Kaspersky being hacked are spread out in the Internet. Sec-Buzzers detected this emerging and grouped these tweets/retweets as a new urgent topic. Moreover, Sec-Buzzers also extracted a detail report of the Duqu 2.0 because of its large number of being retweeted.

b) The 2^{nd} peak on June/15/2015 was made due to the several latest reports pointing out that attackers stolen certificate from Foxconn to hack into Kaspersky with Duqu 2.0.

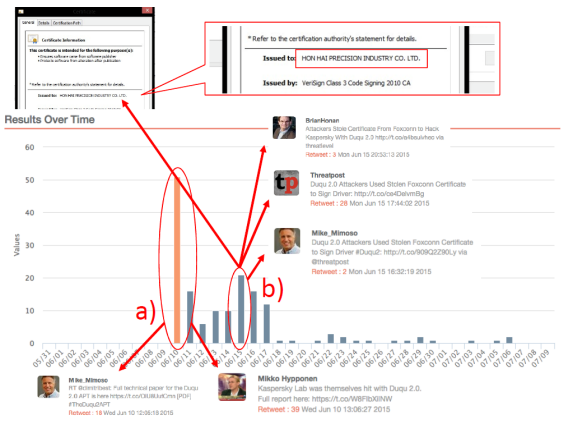

Figure 3: The topic discovery of Sec-Buzzers for Duqu 2.0 malware

4. EVALUATION & DISCUSSION

Table 1 lists the collected data in Sec-Buzzers system so far. There are 333,263 tweets and 334,560 retweets from January/01/2014 to August/31/2015. For those 100 selected experts, there totally exsits 2,145 followship relations among them. During the last two months, July and August, Sec-Buzzers successfully identified 361 cyber security related topics among totally 423 identified ones. Therefore, a corresponding precision rate of 85.34% was derived. Figure 4 shows the distribution of 361 identified cyber security emerging topics among different pre-defined categories.

Another goal for this section is dedicated to a proof of concept experiment for the future improvement of Sec-Buzzers. Our goal is to further refine the precision of content sourcing by partitioning/assigning 100 experts into several subcommunities where different information security disciplines

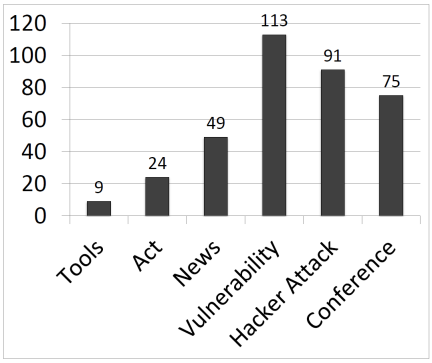

Figure 4: Distribution of 361 cyber security emerging topics identified by Sec-Buzzers

and techniques are emphasized. For this reason, it is very worth realizing whether there exist sub-communities or subtypes among the 100 experts. Based on the retweeting condition above, total 57,704 Twitter accounts who ever retweeted the 100 security experts are then extracted and identified.

Table 1: Description for data collected by Sec-Buzzers

Description	Amount
Followships among 100 security experts	2,145
People who retweeted	57,704
Original tweets	333,263
Retweets	334,560
Total topics identified	423
Cyber security related topics	361
Precision for cyber security topics	85.34%

A matrix $\mathbf{M} \in \{0,1\}^{row \times col}$ can then be defined, where *row* is the number of security experts, and *col* is the number of Twitter accounts who ever retweeted the security experts. In this study, $row = 100$ and $col = 57,704$. If the r^{th} experts was ever retweeted by c^{th} Twitter account, the element (r, c) of matrix $\mathbf{M}$ is set to 1, otherwise $(r, c) = 0$. Therefore, for each Twitter account taken as an instance for clustering, there is a *col*-dimensional binary vector representing this instance. Due to purposes of feature reduction and data visualization, singular value decomposition (SVD) algorithm [11] selecting major two components is applied to $\mathbf{M}$. K-means algorithm [11]using cosine similarity are used with $K = 3$ on the 2-dimensional vector space from SVD, subsequently.

By investigating the three derived clusters in figure 5, we found that there are three communities representing the corresponding clusters. Experts in the 1^{st} cluster are given the appellation of "routine security information provider", tweeting about security issues into their personal page regularly. The 2^{nd} cluster includes those experts also known as "geek" who enthusiastically post what interesting technology they find, what conference they attended or participated in, and what they are devoting to. The rest experts in the 3^{rd} cluster not only posted cyber security information or information technology as we expected, but also sports, movies, and so forth. they share things around their life everyday in Twitter. From above result, it can be demonstrated that

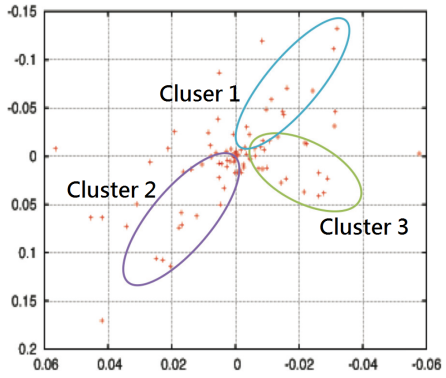

Figure 5: The clustering result or 100 experts on 2-dimensional SVD space

there indeed exist sub-communities in the collected experts, hence by refining and taking advantage of this idea, it is highly prospective to improve precision of content sourcing in Sec-Buzzers.

5. CONCLUSIONS

We developed and released Sec-Buzzers, a web service for exploring cyber security emerging topics based on social network mining for free. With information retrieval from experts on Twitter, Sec-Buzzers quickly recognize the emerging threats of information security then publishes related news, technical reports, and solutions, in time. Our works made following contributions. 1) **design** of the Sec-Buzzers system: A social media connector was carefully designed where integrates Twitter API for routine and automatic data collection, and restore data into an efficiency-scalable database for future analytics. 2) **content sourcing** of the Sec-Buzzers system: To adopt a community-oriented strategy, a set of experts who highly impacts information security domain on social media is grouped and included as the source of content in Sec-Buzzers. Sec-Buzzers leverages information of social network community to find precise and highly-impact context. 3) **discoveries** with Sec-Buzzers: Demonstration shows that by means of a modified topic graph as well as efficient emerging topic finding algorithm, Sec-Buzzers successfully depicts skeleton and outline of security threats that close to what happens in realistic environment in time.

Evaluation result shows that Sec-Buzzers successfully identified emerging cyber security topics among various categories with satisfied precision rate up to 85%. Moreover, from a proof of concept experiment, it also can be observed that several sub-communities can be identified among selected experts by clustering analysis. Sec-Buzzers can take advantage of this great potential as future work to further improve the precision of threats uncovering.

6. ADDITIONAL AUTHORS

Additional authors: Yu-Ting Kuang (Institute of Informaiton Industry, email: `yutingkuang@iii.org.tw`) and Wan-Ching Lin (Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, email: `m10315019@mail.ntust.edu.tw`).

7. REFERENCES

[1] Turkish hacker crashes google play store twice while testing vulnerability, 2014. [Online; accessed: 9-July-2015].

[2] David jevans on wikipedia, 2015. [Online; accessed: 15-July-2015].

[3] Duqu 2.0: The most sophisticated malware ever seen, 2015. [Online; accessed: 16-July-2015].

[4] Twitter rest api, 2015. [Online; accessed: 15-July-2015].

[5] Twitter streaming api, 2015. [Online; accessed: 15-July-2015].

[6] Venom vulnerability: Virtualized environment neglected operations manipulation, 2015. [Online; accessed: 16-July-2015].

[7] V. F. R. M. Arkaitz Zubiaga, Damiano Spina. Classifying trending topics: A typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11*, pages 2461–2464, 2011.

[8] D. Jevans. 100 security experts to follow on twitter, 2013. [Online; accessed: 13-July-2015].

[9] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2015-07-23].

[10] C. S. Mario Cataldi, Luigi Di Caro. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD '10*, 2010.

[11] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[12] I. F. H. Nathan Gnanasambandam, Keith Thompson. Towards situational pattern mining from microblogging activity. In *Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion*, pages 661–666, 2012.

[13] M. N. Nicholas Diakopoulos, Munmun De Choudhury. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, 2012.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[15] D. J. Pearce. An improved algorithm for finding the strongly connected components of a directed graph. Technical report, Technical report, Victoria University, Wellington, NZ, 2005.

[16] M. T. Phuvipadawat, S. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Volume:3)*, pages 120 – 123, 2010.

[17] L. Richardson and S. Ruby. *RESTful web services*. " O'Reilly Media, Inc.", 2008.

[18] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(02):95–145, 2003.