An Analysis of Research Trends in Computer Security over the Last Decade

Daehwa Rayer Lee and Hyoungshick Kim
Department of Computer Science and Engineering, Sungkyunkwan University
Suwon, Republic of Korea
Email: dhwa1206, hyoung@skku.edu

***Abstract*— Recently, text mining has popularly been used not only to identify important topics but also predict future tends in a research field. In this paper, we apply text mining to analyze the research trends in the field of computer security over the last decade. We collected 2,256 papers published in the four major security conferences (CCS, NDSS, USENIX Security and S&P) between 2008 and 2017. We extracted keywords from the title and abstract in each paper and used them to identify important topics based on word frequency counts for each keyword. We analyzed the rankings and the changes in the most frequent keywords every year. Our observations from the analysis would be useful for understanding the popularity of computer security research, and predict potentially important research topics (e.g., SGX, Fuzzing, Blockchain, IoT and SDN) in the near future.**

I. INTRODUCTION

Every year, various and new papers have been introduced in the field of computer science. Those academic papers would be useful source to study and understand research trends in a specific research area. In this paper, we present an analysis method to identify important research topics in a specific research field and apply the emthod to the computer security field to capture most popular research topics. In general, the main research topics of computer security area have dynamically been changed over time because new attack and defense techniques were continuously invented.

To show the feasibility of the proposed method, we collected 2,256 papers published in the four major security conferences (CCS, NDSS, USENIX Security and S&P) between 2008 and 2017 and extracted key research topics using text mining. The trend analysis has often been studied in various domains (e.g., academic publications in computer science [1]). With the collected papers, we analyzed the rankings and the changes in the most frequent keywords every year. We surmised that the most frequently used keywords can be used to identify most important research topics in those papers. From our analysis results, we found that the key research topics in the field of computer security are greatly changed. The popular topics in early years research are Rootkit, RFID, Botnet and Hash but those topics seem to have disappeared in recent years. Instead, new research topics such as Blockchain and SGX are identified in 2016 and 2017. Our analysis results might be useful for understanding the popularity of computer security research, and predict potentially important research topics (e.g., SGX, Fuzzing, Blockchain, IoT and SDN) in the near future.

II. METHODOLOGY

We briefly summarize how to analyze research trends in the field of computer security from academic publications over the last decade. We first extracted high frequency keywords from the title to abstract of each paper published for the last decade. We aggregated publications from the four top security conferences–"ACM Conference on Computer and Communications Security (CCS)", "Network and Distributed System Security Symposium (NDSS)", "IEEE Symposium on Security and Privacy (S&P)", and "USENIX Security Symposium (Security)". These conferences were selected under the conference rankings of well-known websites. Additionally, we extracted keywords in that years, using the Sketch Engine. Sketch Engine technology configures the term structure using parts of speech tagging and characterization. Based on this, it uses criteria to describe the structure of the terms allowed in the language during the keyword extraction phase. For instance, keywords in english are likely to take the form of the term structure, either noun + noun or adjective + noun. We ruled out common terms that are generally used in extracted keywords based on the level of the dictionary, and we set out to collect only the more rare terms, the more technical terms. We used it as a refined keyword by sorting 50 of them in frequency order. Our results are shown in Table I. The number means that the number of times mentioned, + means that the ranking is higher than last year, and - means that the ranking is lower than last year.

III. RESEARCH TRENDS IN COMPUTER SECURITY OVER THE LAST DECADE

A. *"Rootkit" in 2008*

Rootkit is a malware program that attackers install to sneak into the computer, making it highly unlikely to be detected by users. In the 2008, the issue was raised when a Chinese hacker deployed a rootkit that caused Microsoft Windows 2008 to execute random code without authentication to receive an RPC request. In the 2008, the CCS paper "Rootkit Resistant Disks" [2], a study was conducted on rootkit resistant disks to prevent such attacks. The NDSS Paper "HookFinder: Identifying and Understanding Malware Hooking Behaviors" [3], a study was analyzed on behavior of rootkit. Recently, the Kaspersky Lab analyzed the sophisticated malware called slingshot and found a rootkit called Cahnader. It is known that this rootkit can control

TABLE I

TOP 5 HIGH FREQUENCY KEYWORDS IN COMPUTER SECURITY BETWEEN 2008 AND 2017.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
1	Rootkit (16, +)	RFID (32, +1)	CAPTCHA (26, +)	Cloud computing (16, +)	Covert channel (13, +)	Botnet (36, +)	Botnet (26, =)	CFI (45, +2)	Blockchain (43, +)	SGX (90, +4)
2	RFID (13, +)	Botnet (22, +1)	Hypervisor (22, +)	CAPTCHA (15, -1)	Hijacking (12, +)	SGX (23, +)	ROP (25, +)	Botnet (27, -1)	Sandbox (26, +)	Fuzzing (52, +)
3	Botnet (11, +)	Rootkit (13, -2)	Sandbox (19, +)	Sandbox (15, =)	CAPTCHA (8, -1)	PRNG (23, +)	CFI (23, +2)	SDN (23, +2)	Hijacking (24, +1)	Blockchain (32, -2)
4	Hash (10, +)	Location privacy (10, +)	Botnet (13, -2)	Location privacy (11, +)	Cloud computing (7, -3)	Hash (22, +)	Security analysis (19, +)	Hijacking (19, +)	Key exchange (23, +1)	IoT (27, +)
5	Automata (9, +)	Hash (9, +)	Key exchange (8, +)	Key exchange (8, =)	Sandbox (7, -2)	CFI (17, +)	DDoS (12, +)	Key exchange (18, +)	SGX (21, +)	SDN (24, =)

system processes to monitor network traffic, clipboard and keystrokes.

B. "RFID" in 2009

RFID is a technology used by the reader to invoke tag information using radio frequencies. It became an issue in the 2009 when the Chinese government purchased three billion units to issue resident registration cards using RFID. In the 2009, the CCS paper "RFID Privacy: Relation between two notions, minimal condition, and efficient construction" [4], a study was conducted on EPC RFID security vulnerabilities and defenses. The Security paper "Physical layer Identification of RFID Devices" [5], a study was conducted on RFID Transfonder's physical layer identification to use for detecting document forgery. Lately, a company called Sprima unveiled its approach control system, which provides a basic set of Multi-Band RFID security features that are used simultaneously with the UHF band and the HFU band.

C. "CAPTCHA" in 2010

CAPTCHA is used to determine whether a website is accessed by a person or not. There are many different types, such as reading a text, listening to a sound. In the 2010, the issue was raised when the spam ratio rapidly evolved as a result of the easier way to disable the capture phase with the development of OCR technology. In the 2010, the CCS paper "Attacks and design of image recognition CAPTCHAs" [6], a study was conducted on Cortcha, a framework that is designed to complement this and meet the requirements of large scale applications. The S&P paper "How Good Are Humans at Solving CAPTCHAs?" [7], a study was conducted in which we evaluated CAPTCHA from a user's point of view and found that Mechanical Turk is better at attacking CAPTCHA than it is at underground service. Recently, a company called Vicarious developed an artificial technology to solve CAPTCHA by duplicating the structure of the visual cortex in the human brain.

D. "Cloud computing" in 2011

Cloud computing is the use of services through the Internet. There are IaaS that provide infrastructure such as servers and operating systems, and PaaS that provide a platform to develop services, and SaaS that provide software such as applications. It became an issue in the 2011 when IBM released its SmartCloud framework to support its Smarter Planet. In the 2011, the CCS paper "Eliminating the hypervisor attack surface for a more secure cloud" [8], a study was conducted to remove targets from hypervisor attacks by allowing guest VMs to run on the underlying hardware while maintaining the ability to perform multiple VMs simultaneously. The S&P paper "HomeAlone: Co-residency Detection in the Cloud via Side-channel analysis" [9], a study was conducted to using the Side channel as a new detection tool for defense. Lately, google unveiled its platform AGONES, a cloud compute-based game service.

E. "Covert channel" in 2012

The covert channel is a way to transmit commands over data channels that are not commonly used. It is commonly used for attack, but it is also used for virus prevention and digital signature. In the 2012, the issue was raised when a malicious code, named by Kaspersky as Flame, is mentioned as a dangerous cyber weapon, like stuxnet. It captured instant messaging and email from this malicious code, then sent it to Covert SSL channel. In the 2012, the CCS paper "A covert channel construction in a virtualized environment" [10], a study was conducted to implement the Covert channel in the virtualized environment using memory reduction. The Security paper "Whispers in the hyper-space: high speed covert channel attacks in the cloud?" [11], a study presented a new Covert channel attack, which enables high bandwidth and reliable data transfer to cloud. Recently, malware author has reached the stage of switching from the DNS protocol to the Covert channel.

F. "Botnet" in 2013

A botnet is a number of Internet-connected devices, each of which is running one or more bots. It became an issue in the 2013 when the word press server infected by malware and formed a giant botnet. In the 2013, the CCS paper "Beheading Hydras: Performing Effective Botnet Takedowns" [12], a study was conducted on the sequence, methods, and reasons for removing the botnet from the paper. The S&P paper "SoK: P2PWNED - Modeling and Evaluating the Resilience of Peer to Peer Botnets" [13], a study was conducted how to implement a P2P network-based Botnet infrastructure as it could easily be detected and punished by relevant enforcement agencies. Lately, a company called Dyn was attacked on a DDoS based on the MiraI botnet, and this massive attack paralyzed major Internet companies and government organizations.

TABLE II
CORRELATION BETWEEN KEYWORD RANKINGS IN COMPUTER SECURITY BETWEEN 2008 AND 2017.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2008	1.0000									
2009	0.8317	1.0000								
2010	0.8494	0.8362	1.0000							
2011	0.8138	0.8134	0.8639	1.0000						
2012	0.8151	0.8309	0.8618	0.8591	1.0000					
2013	0.8094	0.8298	0.8459	0.8388	0.8866	1.0000				
2014	0.8013	0.8153	0.8465	0.8544	0.9023	0.9004	1.0000			
2015	0.7833	0.7970	0.8323	0.8347	0.8762	0.8951	0.9196	1.0000		
2016	0.7924	0.7974	0.8220	0.8340	0.8771	0.9125	0.9146	0.9120	1.0000	
2017	0.7963	0.8048	0.8322	0.8167	0.8692	0.9260	0.8935	0.9001	0.9246	1.0000

G. *"ROP" in 2014*

ROP is a method to execute a particular command during a buffer over flow attack using a weak program internal machine code. In the 2014, the issue was raised when Microsoft announced the index list of damage by using ROP attacks to affect various Office versions. In the 2014, the S&P paper "Framing Signals - a return to portable shellcode" [14], a study was conducted on Sigreturn oriented programming, a new method of modulating attacks, such as ROP attack. The Security paper "ROP is still Dangerous: Breaking modern defenses" [15], a study was conducted on the fact that it is still dangerous by introduced the method to break Kbouncer and ROPecker, which is a method used to protect ROP. Recently, ROP evolved into a technology called Counterfeit Object Oriented Programming through code reuse.

H. *"Control flow integrity" in 2015*

Control flow integrity makes the flow to proceed along the designated Control flow graph. It is also used to prevent return oriented programming. It became an issue in the 2015 when the method used to prevent ROP attacks, the keyword of the previous year, was mentioned. In the 2015, the CCS paper "CCFI: Cryptographically enforced control flow integrity" [16], a study was conducted on the technology to protect return addresses, function pointers, based on the encryption message authentication code, using CCFI which is new CFI technology. The NDSS paper "VfGuard: Strict protection for virtual function calls in cots C++ binaries" [17], a study was conducted to assess the efficiency using C ++ binary to configure the correct CFI policy and to improve the correctness of the policy. Lately, this technique is used to understand whether the function name has changed over time.

I. *"Blockchain" in 2016*

Blockchain is a distributed ledger technology characterized by integrity, reliability, safety, and transparency. In the 2016, the issue was raised when the blockchain was introduced as the safest technology for storing various data from hacking alongside the financial technology craze. In the 2016, the CCS paper "On the security and performance of proof of work blockchains" [18], a study was conducted on the framework for analyzing the security and performance of different blocks and consensus. The S&P paper "Hawk: the blockchain model of cryptography and privacy-preserving smart contracts" [19], a study allowed to create an intuitive personal smart contract without implementing a secret, and the compiler automatically created an advanced encryption protocol that interacts with the blockchain. Recently, dAPP is evolving based on blockchain.

J. *"SGX" in 2017*

SGX is a CPU-based technology that enhances the security of applications. It became an issue in the 2017, when developers were able to safely protect software from privileged software application. In the 2017, the CCS paper "Iron: Functional encryption using intel SGX" [20], a study was conducted on the values encryption using SGX technology for its function and practicality are much better. The NDSS paper "SGX-Shield: enabling address space layout randomization for SGX programs" [21], a study was conducted to design SGX-Shield, the ASLR framework, for the SGX environment based on the in-enclave loader, it is based on a technique to conduct bootstrap with a sophisticated randomization. Lately, a vulnerability was discovered to use SGX technology to hide malware or to exploit the technology.

IV. CORRELATION BETWEEN KEYWORD RANKINGS FOR THE TEN YEARS

We performed a rank correlation to see the correlation between all the annual keywords. Our results are shown in Table II. We can not predict future keywords through rank correlation, but we can predict how similar future keywords will be in the past. Rank correlation is a method of grasping the linear relationship between two variables in statistics, and the population correlation coefficient p is used as the unit of degree of correlation. In order to find the relationship between the two variables using the population correlation coefficient p, we used the Pearson correlation coefficient. The result value is +1 if the x and y values are completely equal, 0 if they are completely different, and -1 if they are completely equal in the opposite direction. Analyzing the results, we can see that the correlation in 2008 and 2009 are 0.83, and the correlation in 2016 and 2017 are 0.92, so that the keyword is slightly more similar every year than in the previous year.

V. CONCLUSIONS

We used the text mining technique to analyze the research trends in computer security over the last decade and identify key research topics in the field of computer security. We collected 2,256 academic papers from the four major computer security conferences (CCS, NDSS, USENIX Security and S&P) and counted the frequency of the words extracted from the title and abstract of each paper. From our analysis results, we found that the research topics in the field of computer security are dynamically changed. The popular topics (e.g., Rootkit, RFID, Botnet and Hash) in 2008 or 2009 seem to have been less prevalent, or at least had attracted less attention in recent years. Instead, the popularity of new research topics such as Blockchain and SGX seem to have been growing in 2016 and 2017. Our analysis results could be used to predict potentially important research topics in the near future. To show this possibility, we also analyzed the rank correlation between keyword rankings bewteen 2008 and 2017 and found that the keyword rankings in a few years were highly correlated.

As part of future work, we plan to identify relationships between those topics and apply machine learning algorithms to categorize the topics into more structured research fields.

REFERENCES

[1] Apirak Hoonlor, Boleslaw K. Szymanski, and Mohammed J.Zaki. Trends in computer science research. Communications of the ACM, 56(10):7483, 2013.

[2] Kevin R. B. Butler, Stephen McLaughlin, and Patrick D. McDaniel. Rootkit resistant disks. ACM Conference on Computer and Communications Security, 2008.

[3] Heng Yin, Zhenkai Liang, and Dawn Song. Hookfinder: Identifying and under-standing malware hooking behaviors. Network and Distributed System Security Symposium, 2008.

[4] Changshe Ma, Yingjiu Li, Robert H. Deng, and Tieyan Li. Rfid privacy: Relation between two notions, minimal condition, and efficient construction. ACM Conference on Computer and Communications Security, 2009.

[5] Boris Danev, Thomas S, Heydt-Benjamin, and Srdjan Capkun. Physical layer identification of rfid devices. USENIX Security Symposium, 2009.

[6] Bin B.Zhu, Jeff Yan, Qiujie Li, Chao Yang, Jia Liu, Ning Xu, Meng Yi, and Kaiwei Cai. Attacks and design of image recognition captchas. ACM Conference on Computer and Communications Security, 2010.

[7] Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell, and Dan Jurafsky. How good are humans at solving captchas? IEEE Symposium on Security and Privacy, 2010.

[8] Jakub Szefer, Eric Keller, Ruby B. Lee, and Jennifer Rexford. Eliminating the hypervisor attack surface for a more secure cloud. ACM Conference on Computer and Communications Security, 2011.

[9] Yinqian Zhang, Ari Juels, Alina Oprea, and Michael K. Reiter. Homealone: Co-residency detection in the cloud via side-channel analysis. IEEE Symposium on Security and Privacy, 2011.

[10] Jidong Xiao, Zhang Xu, Hai Huang, and Haining Wang. A covert channel construction in a virtualized enviornment. ACM Conference on Computer and Communications Security, 2012.

[11] Zhenyu Wu, Zhang Xu, and Haining Wang. Whispers in the hyper-space: high speed covert channel attacks in the cloud? USENIX Security Symposium, 2012.

[12] Yacin Nadji, Roberto Perdisci, Manos Antonakakis, David Dagon, and Wenke Lee. Beheading hydras: Performing effective botnet takedowns. ACM Conference on Computer and Communications Security, 2013.

[13] Christian Rossow, Dennis Andriesse, Tillmann Werner, BrettStone-Gross,Daniel Plohmann, Christian J. Dietrich, and Herbert Bos. Sok: P2pwned - modeling and evaluating the resilience of peer to peer botnets. IEEE Symposium on Security and Privacy, 2013.

[14] Erik Bosman and Herbert Bos. Framing signals - a return to portable shellcode. IEEE Symposium on Security and Privacy, 2014.

[15] Nicholas Carlini and David Wagner. Rop is still dangerous: Breaking modern defenses. USENIX Security Symposium, 2014.

[16] Ali Jos Mashtizadeh, Andrea Bittau, and Dan Boneh. Ccfi: Cryptograph- ically enforced control flow integrity. ACM Conference on Computer and Communications Security, 2015.

[17] Aravind Prakash, Xunchao Hu, and Heng Yin. Vfguard: Strict protection for virtual function calls in cots c++ binaries. Network and Distributed System Security Symposium, 2015.

[18] Arthur Gervais, Ghassan O. Karame, Karl Wijst, Vasileios Glykantzis, Hubert Ritzdorf, and Srdjan Capkun. On the security and performance of proof of work blockchains. ACM Conference on Computer and Communications Security, 2016.

[19] Ahmed Kosba, Andrew Miller, Elaine Shi, Zikai Wen, and Charalampos Papamanthou. Hawk: the blockchain model of cryptography and privacy-preserving smart contracts. IEEE Symposium on Security and Privacy, 2016.

[20] Ben Fisch, Dhinakaran Vinayagamurthy, Dan Boneh, and Sergey Gorbunov. Iron: Functional encryption using intel SGX. ACM Conference on Computer and Communications Security, 2017.

[21] Jaebaek Seo, Byoungyoung Lee, Seongmin Kim, Ming-Wei Shih, Insik Shin, Dongsu Han, and Taesoo Kim. SGX-Shield: enabling address space layout randomization for SGX programs. Network and Distributed System Security Symposium, 2017.