2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), and 2015 IEEE 12th International Conf on Embedded Software and Systems (ICESS)

Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities

Shams Zawoad and Ragib Hasan
{zawoad, ragib}@cis.uab.edu
University of Alabama at Birmingham
Birmingham, Alabama 35294-1170, USA

Abstract—**The age of big data opens new opportunities in various fields. While the availability of a big dataset can be helpful in some scenarios, it introduces new challenges in digital forensics investigations. The existing tools and infrastructures cannot meet the expected response time when we investigate on a big dataset. Forensics investigators will face challenges while identifying necessary pieces of evidence from a big dataset, and collecting and analyzing those evidence. In this article, we propose the first working definition of big data forensics and systematically analyze the big data forensics domain to explore the challenges and issues in this forensics paradigm. We propose a conceptual model for supporting big data forensics investigations and present several use cases, where big data forensics can provide new insights to determine facts about criminal incidents.**

Keywords*—*Big Data Forensics, Forensic Investigation

I. INTRODUCTION

As we enter the age of big data, we are producing massive amounts of data through our everyday activities. For example, Sagiroglu *et al.* showed that the amount of information that was created every two days in 2013 was equal to the total amount of data that was ever created by humans till 2003 (5 exabytes) [1]. The low cost of digital storage, the increasing ubiquity of computing, and the growth of the type and number of the Internet of Things (IoT) drive this massive increase in the amount of digital data. According to a report by Gartner in 2014, there will be 26 billion IoT devices in deployment by 2020, which will create new challenges for all aspects of data management [2]. As reported by Intel, the world of digital data was expanded to 2.72 zettabytes by the end of 2012 and it is predicted to be doubled in every two years, reaching about 8 zettabytes of data by the end of 2015 [3]. The amount of data generated by different services is also getting larger. For example, Facebook reported that its users uploaded 300 million photos per day in early 2015 [4]. Radio frequency identification (RFID) systems used by retailers and others can generate 100 to 1,000 times the data of conventional bar code systems [5]. Besides these, more than 5 billion people are calling, texting, and browsing on mobile phones worldwide as of 2012 [5].

Big data presents a real opportunity in identifying actionable insights from information. However, the massive increase in the size of available digital data also plays a major role behind the increase in the size of digital evidence. According to an annual report of the Federal Bureau of Investigation (FBI), the size of the average digital forensic case is growing at 35% per year in the United States. From 2003 to 2007, it increased from 83GB to 277GB [6]. Another report of the FBI states that during the fiscal year 2012, the Computer Analysis Response Team (CART) of the FBI supported nearly 10,400 investigations and conducted more than 13,300 digital forensic examinations that involved more than 10,500 terabytes of data [7].

The availability of large-scale potential evidence presents new challenges in digital forensics investigations. The more data is available, the harder it is to spot fraudulent activity and malicious users behind those activities. For example, the large number of IoT devices will produce very big datasets of activity logs. Hence, using traditional log correlation or visualization techniques, we cannot identify malicious activities/users from a big dataset of logs. According to a 2014 survey by American Institute of Certified Public Accountants (AICPA), big data is listed as the top issue facing forensic and valuation professionals in between 2014 and 2019 [8].

In this paper, we address the availability of big data in digital forensics and introduce a new branch of digital forensics – *Big data forensics*. We discuss big data forensics from two perspectives: executing digital forensics procedures on a large size of possible evidence and determining new/unknown facts by utilizing the available big data. Based on the characteristics of big data and digital forensics procedures, we identify the challenges of executing each of the processes of digital forensics, where the evidence remains within a big dataset. We also present a real life case study involving big data and propose a conceptual model for big data forensics and opportunities of using big data for forensics investigations.

Contribution: The contributions of this work are as follows:

1) To the best of the authors' knowledge, this is the first work to define big data forensics. The two different perspectives of big data forensics that are introduced here can spawn future research works in this area.
2) We systematically analyze the challenges and opportunities for big data forensics and propose a conceptual model. Our analysis can help researchers to focus on specific research sub-problems of the big data forensics problem domain.
3) We present a real life case study of a cybercrime investigation problem involving a big dataset of spam emails and suspected phishing websites.

Organization: The rest of the paper is organized as follows: Section II provides the background knowledge about digital forensics, big data, big data forensics, and the case study. In section III, we present the challenges of big data forensics.

978-1-4799-8937-9/15 $31.00 © 2015 IEEE
DOI 10.1109/HPCC-CSS-ICESS.2015.305

Section IV presents a conceptual model for big data forensics and the opportunities provided by big data in digital forensics investigations. Section V presents the related work and finally, we conclude in Section VI.

II. Background

In this section, we first present an overview of digital forensics and big data. Next, we define big data forensics and present a case study of a digital forensics investigation involving big data.

A. Digital forensics

Federal Rules of Civil Procedure (FRCP) expanded the scope of evidence in the 2006 amendment and stated that electronically stored information (ESI) can be used in civil litigation [9]. According to a definition from the National Institute of Standards and Technology (NIST) [10], digital forensic is *"an applied science to identify an incident, collection, examination, and analysis of evidence data"*. From the above working definitions, we note that digital forensics comprises four main processes [11]:

Identification: The first step of a digital forensics investigation is identification, where an investigator identifies the incidents that are important to prosecute a litigation and identifies the evidence related to those incidents.

Collection: After identifying the evidence, an investigator needs to collect evidence from various digital medias, such as cell phone, hard disk, router, etc.

Organization: Organizing the collected evidence efficiently leads towards the facts of a criminal incident. First, an investigator inspects the data and its characteristics. After that, the investigator interprets and correlates the available data to determine the facts.

Presentation: In the final phase, an investigator prepares an organized report to state his or her findings about the case, which should be admissible to the court.

B. Big Data

Madden defines Big data as the data that is too big, too fast, and too hard for existing tools to process [12]. Big data is also characterized by the five Vs: variety, velocity volume, veracity, and value [13]. .

Variety. Big data can come from different sources, such as web pages, network or process logs, posts from social medias, emails, documents, and data from various sensors. These data can be categorized in three general types: structured, semi structured, and unstructured. Data stored in a regular database are treated as structured data. Unstructured data do not have a fixed format and hence, it is much harder to execute forensics analysis on such data. Semi-structured data may not have fixed fields but can contain tags to expedite the data analyze process [1], [14].

Volume. In the age of big data, organizations deal with terabytes and petabytes of data. For example, Walmart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data [5]. Google processes 20 Petabytes of data per day [15]. AT&T has a database that is 312 terabytes in size including 1.9 trillion phone call records [16].

Velocity. Velocity of new data that comes from different systems makes the big data even bigger everyday. For example, International Data Corporation (IDC) estimates that by 2020, the number of business transactions on the Internet, which includes Business-to-Business (B2B) and Business-to-Consumer (B2C), will reach 450 billion per day [16]. Big data can come at different velocities, such as, real-time, batch, streams, etc. [1]. Velocity does not only refer to the speed of the incoming data but also the speed of data flow inside a system. For time limited processes, big data should be used as it streams into the organization in order to maximize its value [12], [14].

Veracity. Veracity refers to correctness and accuracy of information. Veracity involves data quality, data governance, and metadata management, along with considerations for privacy and legal concerns. With many forms of big data, quality and accuracy are less controllable and the volumes often cause the lack of quality or accuracy.

Value. Value refers to the ability of turning the data into a value. The availability of big data can reveal previously unknown insights from data and this understanding leads to the value.

C. Big data forensics

We define big data forensics as an special branch of digital forensics where the identification, collection, organization, and presentation processes deal with a very large-scale dataset of possible evidence to establish the facts about a crime.

Big data forensics can be discussed from two perspectives. *First*, a small piece of evidence can exist in a big dataset. For example, to investigate a criminal incident, we may need the information of a few call records among the 1.9 trillion call records of AT&T. *Second*, a crucial piece of information can be revealed by analyzing a big dataset or by correlating data of multiple big datasets. For example, a spam email classifier that is trained on small dataset may not perform well in determining a new spam email. However, when the classifier is trained on a big dataset, most likely it will perform better in identifying a new spam email.

Case Study. We collected a big dataset of spam emails from bounce.io[1] for anti-cybercrime research. One of the major threats on Internet users is the phishing attack. Phishing websites resemble legitimate websites and draw users to visit those malicious websites and deceive them to provide their private credentials, such as usernames and passwords, bank account numbers, and credit card numbers with pin codes on these fake websites. Most often, spam emails contain URLs of these phishing websites. Hence, spam emails are great sources of identifying phishing websites.

However, the number of URLs that we collect from the spam email source of bounce.io is very large. Table I shows the number and volume of URLs collected in seven consecutive days (October 23rd to October 29th 2014). The average number of URLs collected in these seven days is 84 million and the average size of storage to preserve those URLs is 28 GB.

[1] http://bounce.io/

Date	Total No. of URLs (Million)	Size (GB)
10/23/2014	97.98	34
10/24/2014	88.74	31
10/25/2014	55.08	19
10/26/2014	52.06	18
10/27/2014	92.44	29
10/28/2014	115.04	39
10/29/2014	93.42	30

TABLE I: Statistics of seven days of URLs collected from a big spam source

Using traditional phishing detection techniques [17], [18], [19], [20] and regular computers, it is not possible to identify a phishing website from this big data set of possible phishing URLs. On the other hand, if we had an ideal system to process this big dataset, we could identify several issues, such as phishing and spam campaigns and the criminals who are behind the strongest phishing/spam campaigns.

III. Challenges in Big Data Forensics

According to Gantz *et al.*, big data requires a new generation of technologies and architectures, designed to efficiently extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis [21]. Unfortunately, the traditional tools and technologies of digital forensics are not designed to handle the big data. In this section, we identify the challenges in each of the steps of digital forensic investigations that deal with big data.

A. *Identification*

It can be difficult to identify the important pieces of evidence to determine the facts when the amount of possible evidence is very large. Finding evidence in a big data source may turn out to be finding a needle in the haystack. Network logs are important to detect any network intrusion. However, identification and isolation of the logs for a compromised instance can be very challenging for a large Infrastructure-as-a-Service (IaaS) cloud, considering the high velocity and volume of network logs produced in such infrastructures.

B. *Collection*

After identifying the evidence, investigators need to collect the evidence to analyze and find the facts. Any errors that have occurred in the collection phase will propagate to the evidence organization and reporting phase, which will eventually affect the whole investigation process. Hence, this is one of the most crucial steps of the digital forensic procedure.

Some of the factors that make the data acquisition process in big data forensics harder than traditional computer forensics are discussed below:

In traditional computer forensics, investigators have full control over the evidence (e.g., a hard drive confiscated by a police). However, in the age of big data, investigators need to depend on organizations to collect data. The existing forensics tools for data extraction and preservation are mostly designed for personal owned computing system, where the volume of hard drive can be a few terabytes at most. The law enforcement agencies are not equipped to collect evidence from a data center running thousands of machines hosting petabytes/zetabytes of data. Evidence residing in a large data center are decentralized in hundreds of computers. Therefore, it is not possible to collect evidence from a vast decentralized infrastructure using the current digital forensics tools.

Since clouds are among of the main sources of big data, some of the problems of collecting evidence from clouds, such as the physical inaccessibility of storage, also apply to the big data forensics domain. The established digital forensic procedures and tools assume that we have physical access to the computing resources, e.g., hard disks, network routers, etc. However, in clouds, sometimes we do not even know where the data is physically located at as it is distributed among many hosts in multiple data centers.

The availability of big data introduces the possibility of leaking confidential information [22]. Hence, the evidence collection process in the case of big data forensics may violate the privacy and data protection laws if the process cannot protect the confidential information of a suspect. Since such laws can vary depending on the jurisdiction, it may happen that a forensic investigator is in one jurisdiction and the data reside in another jurisdiction, where the data protection laws are different. Therefor, it is important for the organizations to carefully choose the location of data storage and processing to comply with various regulatory compliance issues [22].

In Section I, we have seen that the amount of digital evidence is rapidly increasing and so is the amount of data. Guo *et al.* pointed out the requirement of large bandwidth issue for time critical investigations [23].

In the case study presented in Section II-C, it takes more than 24 hours to collect and preserve one day of spam email information from bounce.io's Amazon S3 storage. Because of the delay in the collection phase, it is not possible to create a live feed of phishing websites from this big dataset. Preservation of this large volume of dataset is also challenging since it grows above terabytes within a month.

C. *Organization*

The characteristics of big data: volume, variety, velocity, veracity, and value do not comply with the existing analysis techniques [12], [14]. Therefore, organizing a big dataset to identify the facts about a criminal incident can be challenging. For a big dataset of evidence, manual review and decision-making cannot work at such large scale. Often times, it requires special data mining techniques and tools to identify the facts [24], [25].

The wide variety of structures of big data makes the examination and analysis phase challenging. Analyzing various logs, such as network logs, process logs, and application logs is one of the most vital tasks in digital forensics investigation to identify malicious activities and the users behind those malicious activities. However, gathering this crucial information in the big data paradigm is not as simple as it is in a small-scale homogeneous environment.

For example, the data format for IoT devices depends on the manufacturer of the devices since there are no common standards. Because of the heterogeneous log formats, it can be very difficult to analyze logs collected from various IoT devices. In the worst case, it may happen that some IoT devices do not even store very crucial forensics information, such as who, when, where, and why some incident were executed [26]. A common format for logs for different system would provide an investigator a homogeneous dataset of logs with all the crucial information. This could also enable the investigator to correlate logs collected from various systems. The low quality and accuracy of big data also make the analyzing task challenging.

Since law enforcement agencies do not have the capability of downloading and analyzing a big dataset of possible evidence, they can distribute the forensics procedures to the data provider, computation provider, and a computational platform (such as a MapReduce framework where the code will be run on the evidence). However, since a big dataset may contain confidential information of end users, it is challenging to prevent unauthorized leaks of private information by the third party computation provider [27].

Because of the large volume and velocity of big data, we require especial search and data mining techniques. For example, in our case study, even though we were able to collect the spam email information in real time, we could not create live phishing feed or identify a new live spam campaign because of the volume and velocity of the data.

D. Presentation

The final step of a digital forensic investigation is presentation, where an investigator accumulates his/her findings and presents to the court as the evidence of a case. Challenges also lie in this step of big data forensics. Providing the evidence in front of the jury for traditional computer forensics is relatively easy compared to the complexity of managing big data. Jury members possibly have basic knowledge of personal computers or at most privately owned local storage. But the technicalities behind filtering, analyzing big data, and identifying a value can be far too complex for them to understand.

IV. Big Data Forensics Model and Opportunities

In this section, we first propose a conceptual model of big data forensics. Then, we present several opportunities of big data forensics, given that the ideal system for managing big data forensics presents.

A. Conceptual Model for Big Data Forensics

Based on the challenges of big data forensics, we propose a Hadoop Distributed File System (HDFS) and cloud based conceptual model to support reliable forensics investigation on big data. The proposed model is presented in Figure 1.

Hadoop[2] is an open source implementation of Google's proprietary MapReduce framework. The Hadoop Distributed File System (HDFS) is the file system component of Hadoop, which is designed to store very large data sets reliably, and to stream those datasets at high bandwidth to user applications [28]. Therefore, HDFS will perform better than the traditional database management systems when retrieving a small piece of information from a big dataset of possible evidence.

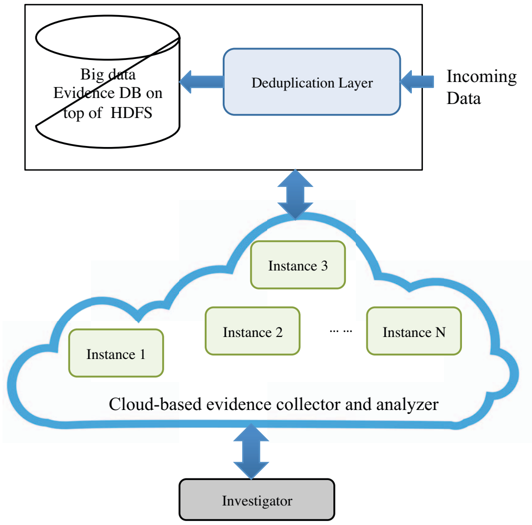

Fig. 1: A conceptual model for big data forensics

Since data duplication is a common characteristic of a big dataset, we propose a deduplication layer to remove redundant data from the incoming data stream. For example, deduplicaiton can save a great amount of storage cost for any content-based phishing identification since it is required to download the content of the suspected phishing website. Table II presents two random hours' statistics of our case study regarding the duplicate URLs and the extra amount of storage caused by the duplicate URLs. The results suggest that deduplication can save a large amount of computing and storage resources.

Time	No. Of Unique URLs	No. Of Duplicate URLs	Redundant Storage (GB)
10/28/2014 6:00 am	22445	12552	4.85
10/30/2014 8:00 am	24318	42846	23.05

TABLE II: Deduplication statistics for two random hours of URL data

Using the current tools and techniques of digital forensics, it is not possible to collect a big dataset of evidence and process it using local infrastructures. For a very large volume of evidence, we need an extremely costly local infrastructure to provide the desired functionality in a reasonable time period. Moreover, local infrastructures suffer from low scalability. Hence, we propose a cloud-based distributed evidence collection and analyzing system. A cloud-based architecture can be cost effective and easily scalable. The evidence collection and analysis can be provided as Forensics-as-a-Service to investigators.

B. Opportunities in Big Data Forensics

The value property of big data refers to the fact that previously unknown insights can be revealed from big data. This is very applicable in terms of digital forensics and opens new

[2] http://hadoop.apache.org/

opportunities given that we have ideal tools and technologies to identify, collect, and analyze a big data set.

Correlating different data sets to identify cyber criminals. zone-h[3] is an online forum, where hackers post their hacking incidents and take pride of their hacking accomplishments. Very often, hackers reveal some sort of personal information, such as email address, name, hacking group, or location. Sometimes, we can identify the possible real identify of a hacker by searching the personal information on different online hacking community websites, Facebook, and Twitter. However, manual searching is cumbersome in this scenario, where we need to correlate the participation of a suspect among different groups, friend list, and Facebook/Twitter activities of that suspect. Because of the large volume of data in different systems, we are yet to come up with an efficient technology for the law enforcement agencies to automate the above search procedure.

Authorship attribution [29] can be used to identify the characteristics of the creator of comments or posts in a hacking forum. Using the unique characteristics, we can build a profile of every user of different hacking forums. Later, we can map the characteristics found on the hacking forums with the Facebook or Twitter posts and identify the hacker's Facebook/Twitter profile. However, the data on both hacking forums and social media are too big for such analysis. As reported in [4], there are over 1.39 billion monthly active Facebook users in the world (as of early 2015), where five new profiles are created in every second and 293,000 statuses are update in every minute. If we had efficient tools and procedures to handle data of such high velocity and volume, we could easily identify an individual or members of a hacking group behind a hacking incident or even can correlate these data with credit card fraud or money laundering cases.

Live feed of an ideal phishing black list. Availability of a live phishing feed can protect users to visit malicious phishing websites. Internet service providers (ISPs) can use such feed to block the phishing websites, or a web browser can use the feed to stop the users from visiting those websites. However, Sheng *et al.*showed that there is a need to better protect victims from phishing attacks, as blacklists are not working [30]. Phishing attacks continue to exist since victims are not appropriately warned or blocked from the attacks [31]. These works indicate that we still do not have an ideal black list of phishing websites to protect users.

One of the main reasons behind the unavailability of an ideal black list is the lack of technology to identify phishing websites from a massive amount of potential phishing websites. The case study presented in the Section II-C gives an idea about the amount of potential phishing websites coming from spam emails. In our case study, bounce.io publish their spam data in every fifteen minutes, which contains more than millions of URLs of potential phishing websites. If we had efficient technology to identify phishing websites from this data stream in real-time, we could make a robust live feed of phishing websites.

Utilizing IoT devices in criminal case investigation. By 2020, there will be 10 connected IoT devices for every person of the world and 40 to 80 billion IoT devices in total [32]. The availability of the large number of IoT devices can be used effectively in a digital forensics investigation. For example, most of IoT devices are equipped with various sensors and thus can report important information of the surroundings of the devices. This in turn can help an investigator to determine facts about a criminal incident. For example, some IoT devices (such as a smart air conditioner) located at the suspect's home or office building log the presence of the user nearby the devices. Therefore, by analyzing the logs of such IoT devices, an investigator can identify the location of a suspect at a particular time. However, the amount of sensors or log data collected from the IoT devices will be very big. Hence, an ideal solution could correlate high volume of logs efficiently and help an investigator to rapidly identify crucial pieces of information.

V. Related Work

Tankard identified several security issues of big data [22]. According to Tankard, the security threat from big data is the aggregation and analysis of large amount of sensitive information regarding customers and employees, as well as intellectual property, trade secrets and financial information. Since this big sensitive data is now centralized, it becomes an attractive target for external attackers to violate the privacy of users, which in turn can damage the trust and reputation of an organization. Another problem addressed by Tankard is related to the regulatory compliance, especially with privacy and data protection laws.

In [33], Zawoad *et al.* proposed a Cloud-based spam URL Analyzer (CURLA), built on top of Amazon Elastic Computer Cloud (EC2) and Amazon Simple Queue Service (SQS), which can process a large number of spam-based URLs in parallel. By leveraging the power of clouds, CURLA reduces the cost of establishing equally capable local infrastructure and can be scaled up depending on amount of URLs that needs to be analyzed. Huang *et al.* proposed Cumulon, which can be used to rapidly develop and intelligently deploy matrix-based big-data analysis programs in the cloud [34]. The proposed system is build on top of Hadoop and performs better than the existing Hadoop-based systems for statistical data analysis.

Mall *et al.* [24] showed the feasibility of utilizing the Kernel Spectral Clustering (KSC) method for the purpose of community detection in large-scale synthetic networks and real world networks like the YouTube network, a road network of California and the Livejournal network. The kernel matrix for the KSC method cannot fit into memory for a big data set. Therefore, they proposed to select a smaller subgraph that fits into memory and preserves the overall community structure to construct the model. In [25], researchers proposed a multi-view K-means clustering method for a big dataset. They proposed a robust large-scale multi-view K-means clustering approach, which can be easily parallelized and performed on multi-core processors for big visual data clustering.

Das *et al.* developed Ricardo on top of Hadoop data management system and a statistical analysis software, R [35]. The goal is to support analysts to work on a large scale datasets

[3] http://www.zone-h.org/

from within a popular, well supported, and powerful analysis environment by merging the capability of Hadoop and R. Chang *et al.* proposed HAWQ – a parallel processing SQL engine built on top of HDFS [36]. HAWQ adopts a layered architecture and relies on the distributed file system for data replication and fault tolerance.

VI. Conclusion

The ubiquitous nature of current computing devices is driving the massive increase in the amount of digital data generated by humans and machines. With the increase in the amount of digital data, there is a growing need for providing support for digital forensics in big data application domains. However, the natural consequence of this is the need for increased efforts to derive meaningful information from the accumulated data. The amount of digital information is now growing beyond the capacity of current digital forensics tools and procedures. In this article, we define the term *big data forensics* and identify the challenges of executing reliable forensics in the big data paradigm. We propose a conceptual model of handling big data for digital forensics and present some opportunities in digital forensics for the availability of big data. Solving all the challenges of big data forensics can open the opportunity of identifying many new insights that were not possible before.

Acknowledgment

This research was supported by the National Science Foundation CAREER Award CNS-1351038, a Google Faculty Research Award, and the Department of Homeland Security Grant FA8750-12-2-0254.

References

[1] S. Sagiroglu and D. Sinanc, "Big data: A review," in *International Conference on Collaboration Technologies and Systems (CTS)*. IEEE, 2013, pp. 42–47.

[2] www.gartner.com, "Gartner Says the Internet of Things Will Transform the Data Center," http://www.gartner.com/newsroom/id/2684616, 2014.

[3] Intel, "Planning guide: Getting started with big data," http://goo.gl/jyU73v, Tech. Rep., 2013.

[4] D. Noyes, "The Top 20 Valuable Facebook Statistics Updated February 2015," http://goo.gl/P1iUB2, 2015.

[5] SAS, "Big data meets big data analytics," http://www.sas.com/resources/whitepaper/wp_46345.pdf, Tech. Rep.

[6] FBI, "Annualreport for fiscal year 2007," 2008 Regional Computer Forensics Laboratory Program, 2008.

[7] ——, "Piecing together digital evidence," https://goo.gl/5cBWDb, 2013.

[8] AICPA, ""Big Data" Listed as Top Issue Facing Forensic and Valuation Professionals in Next Two to Five Years: AICPA Survey," http://goo.gl/1BgdWB, 2014.

[9] Federal Rules of Civil Procedure, "Rule 34," http://goo.gl/NfL61.

[10] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response," *NIST Special Publication*, pp. 800–86, 2006.

[11] S. Zawoad and R. Hasan, "Cloud forensics: a meta-study of challenges, approaches, and open problems," *arXiv preprint arXiv:1302.6312*, 2013.

[12] S. Madden, "From databases to big data," *IEEE Internet Computing*, vol. 16, no. 3, pp. 4–6, 2012.

[13] B. Marr, "Why only one of the 5 vs of big data really matters," http://goo.gl/azsnse, 2015.

[14] P. Zikopoulos, C. Eaton *et al.*, *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[15] S. Gunelius, "The Data Explosion in 2014 Minute by Minute Infographic," http://goo.gl/1Xkxv4, 2014.

[16] B. Davis, "How Much Data We Create Daily," http://goo.gl/a0ImFT, 2013.

[17] B. Wardman and G. Warner, "Automating phishing website identification through deep MD5 matching," in *eCrime Researchers Summit, 2008*. IEEE, 2008, pp. 1–7.

[18] Y. Pan and X. Ding, "Anomaly based web phishing page detection," in *22nd Annual Computer Security Applications Conference (ACSAC)*. IEEE, 2006, pp. 381–392.

[19] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: a content-based approach to detecting phishing web sites," in *16th international conference on World Wide Web*. ACM, 2007, pp. 639–648.

[20] M. Dunlop, S. Groat, and D. Shelly, "Goldphish: Using images for content-based phishing analysis," in *5th International Conference on Internet Monitoring and Protection (ICIMP)*. IEEE, 2010, pp. 123–128.

[21] Gantz, John and Reinsel, David, "Extracting value from chaos," Available at http://goo.gl/PxPc2, Tech. Rep., 2011.

[22] C. Tankard, "Big data security," *Network security*, vol. 2012, no. 7, pp. 5–8, 2012.

[23] H. Guo, B. Jin, and T. Shang, "Forensic investigations in cloud environments," in *International Conference on Computer Science and Information Processing (CSIP)*. IEEE, 2012, pp. 248–251.

[24] R. Mall, R. Langone, and J. A. Suykens, "Kernel spectral clustering for big data networks," *Entropy*, vol. 15, no. 5, pp. 1567–1586, 2013.

[25] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *23rd international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 2598–2604.

[26] R. Marty, "Cloud application logging for forensics," in *ACM Symposium on Applied Computing*. ACM, 2011, pp. 178–184.

[27] R. Hasan, *Security and Privacy in Big Data and Cloud Computing: Challenges, Solutions, and Open Problems in Advances in data processing techniques in the era of Big Data"*. CRC Press, 2013.

[28] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *26th IEEE Symposium on Mass Storage Systems and Technologies (MSST)*. IEEE, 2010, pp. 1–10.

[29] P. Juola, "Authorship attribution," *Foundations and Trends in information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.

[30] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists," in *6th Conference on Email and Anti-Spam (CEAS)*, 2009.

[31] T. Ronda, S. Saroiu, and A. Wolman, "Itrustpage: a user-assisted anti-phishing tool," in *ACM SIGOPS Operating Systems Review*, vol. 42, no. 4. ACM, 2008, pp. 261–272.

[32] blog.xively.com, "Infographic: The Future of the Internet of Things," http://goo.gl/xym4so, 2014.

[33] S. Zawoad, R. Hasan, M. M. Haque, and G. Warner, "Curla: Cloud-based spam url analyzer for very large datasets," in *7th IEEE International Conference on Cloud Computing (CLOUD)*, 2014, pp. 729–736.

[34] B. Huang, S. Babu, and J. Yang, "Cumulon: Optimizing statistical data analysis in the cloud," in *ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 1–12.

[35] S. Das, Y. Sismanis, K. S. Beyer, R. Gemulla, P. J. Haas, and J. McPherson, "Ricardo: integrating r and hadoop," in *ACM SIGMOD International Conference on Management of data*. ACM, 2010, pp. 987–998.

[36] L. Chang, Z. Wang, T. Ma, L. Jian, L. Ma, A. Goldshuv, L. Lonergan, J. Cohen, C. Welton, G. Sherry *et al.*, "Hawq: a massively parallel processing sql engine in hadoop," in *ACM SIGMOD international conference on Management of data*. ACM, 2014, pp. 1223–1234.