1-1-2020

Unstructured data for cybersecurity and internal control

Jesus Canelón
California State University, San Bernardino

Esperanza Huerta
San Jose State University, esperanza.huerta@sjsu.edu

Normal Leal
Instituto Tecnologico Autonoma de Mexico

Terry Ryan
Claremont Graduate University

Recommended Citation

Jesus Canelón, Esperanza Huerta, Normal Leal, and Terry Ryan. "Unstructured data for cybersecurity and internal control" *Proceedings of the 53rd Hawaii International Conference on System Sciences* (2020): 5411-5420.

Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020

Unstructured Data for Cybersecurity and Internal Control

Jesus Canelón
California State University,
San Bernardino
jesus.canelon@csusb.edu

Esperanza Huerta
San José State University
esperanza.huerta@sjsu.edu

Normal Leal
Instituto Tecnológico
Autónomo de México
nleal@itam.mc

Terry Ryan
Claremont Graduate
University
terry.ryan@cgu.edu

Abstract

This paper proposes a research framework for studying the connections--realized and potential--between unstructured data (UD) and cybersecurity and internal controls. In the framework, cybersecurity and internal control goals determine the tasks to be conducted. The task influences the types of UD to be accessed and the types of analysis to be done, which in turn influences the outcomes that can be achieved. Patterns in UD are relevant for cybersecurity and internal control, but UD poses unique challenges for its analysis and management. This paper discusses some of these challenges including veracity, structuralizing, bias, and explainability.

1. Introduction

Organizations analyze many sources and kinds of data to manage cybersecurity and internal control risks. Data analytics, with its ability to identify patterns, is relevant to the implementation of cybersecurity and internal controls. Data analytics can not only draw from the large amount of structured data produced by accounting information systems, but also from the even larger amount of unstructured data (UD) created inside and outside an organization. Still, there is limited research on how the use of UD can support cybersecurity and internal control. This paper proposes a research framework for studying the connections--realized and potential--between UD and cybersecurity and internal controls.

For purposes of this paper, we define cybersecurity as organizational activities designed to protect systems, networks, and programs from digital threat [1]. We define internal control as organizational activities designed to safeguard assets, minimize errors, and approved occurrence of operations [2].

In cybersecurity, real-world investigations and academic research are fueled by the consequences and the number of attacks perpetrated. [3] reports that close to $600 billion was lost to cybercrime worldwide in 2018. [4] estimates that cybercriminals will steal around 33 billion data records by 2023; it also estimates that nearly 60 million Americans have been impacted by identity theft (with 15 million cases in 2017). The cost of the average data breach to a U.S. company is $7.91 million, while for companies worldwide it is $3.86 million [4]. Because the U.S. has been the foremost target for such attacks, the U.S. government will spend $15 billion on cybersecurity in the 2019 fiscal year.

In internal control, real-world investigations and academic research are fueled by the consequences and the number of frauds and material misstatements. Although it is not possible to provide a completely accurate estimate of the cost, the projected total global fraud loss in 2017 is close to $4 trillion [5]. In addition to fraud prevention and detection, internal and external auditors work to detect and respond to evidence of policies, procedures, regulations, and laws being violated or ignored.

Cybersecurity and internal control must make use of all relevant data: internal or external; current or historical; targeted or generic; private or public; and (importantly) more or less structured.

Cybersecurity has a long history of using UD for detecting attacks, as evidenced by its early use of text filters to identify viruses or phishing attacks. Internal controls, on the other hand, have traditionally relied on the analysis of structured data--identifying unusual patterns in amounts or dates of transactions, for instance. Although UD has been used to support fraud investigations—the Enron investigation, for example, analyzed emails for evidence of fraudulent intent—it is only recently that advances in technology have made it easier to exploit UD.

Analyzing UD imposes many challenges. Auditors, for instance, need to access data beyond what traditionally is used to confirm the existence of events. They need to model markets, operational

URI: https://hdl.handle.net/10125/64407
978-0-9981331-3-3

data, sales, and post-sales activities to provide a basis for evaluating what they observe. They need to adopt new mechanisms to assure the objectivity of their investigations and recommendations. They need to expand the scope of their data collection and retention practices. Finally, they need to complement or replace traditional audit evidence with new forms [6].

The growth of data, especially UD, has implications for those who are charged with investigating cybersecurity and internal control matters. Appreciating these implications can help them carry out their assigned tasks better. UD is vital to consider. "Most of the data that move markets are inherently unstructured—central bank announcements; geopolitical developments; product releases; research breakthroughs; droughts, hurricanes, and other weather-related phenomena; and natural disasters” [7:114]. There are "no reasonable limits on sources of data, but there are great limits on what data an organization can actually store and make useful" [7:22].

By the early 2010s, people produced approximately 2.5 exabytes (2.5 quintillion bytes or 2.5 billion gigabytes) of data every day. By 2020, the data produced daily will reach 40 zettabytes (40 trillion gigabytes), more than 5,200 gigabytes for every person [8]. Determining the amount of UD would be difficult and imprecise. However, because of the volume and speed at which UD is generated, it is reasonable to assume that structured data comprises only a small portion of the overall data, with UD being the majority of it. It has been estimated that 90% of the data now being created is unstructured [9].

This paper contributes to the literature in three different ways. First, it discusses current and potential applications of UD to cybersecurity and internal control. In this, it supports auditors, regulators, and technology vendors, giving guidance on incorporating the analysis of UD into business and audit procedures. Second, it outlines the challenges accountants face to adopt and exploit UD, providing them with guidance for its analysis. Third, it proposes a research framework to foster a research agenda on UD for cybersecurity and internal control.

The remainder of the paper is organized as follows. The first section defines UD and discusses its characteristics. The second provides an overview of the opportunities that the analysis of UD affords to cybersecurity and internal control investigators. The third highlights the challenges that the analysis of UD poses for cybersecurity and internal control investigations. The fourth puts forward a research framework on the use of UD for cybersecurity and internal control investigations, and it discusses how the framework can guide different types of research that accounting information systems scholars may pursue.

2. Unstructured data

Data can be classified using different criteria. The classifications are not mutually exclusive; each provides a lens through which data may be viewed. One categorization is based on representation: digital vs. non-digital. Outside the computer, data can be represented in many forms; inside it, all data (e.g., numbers, text, images, audio, video, or sensor readings) must be represented in binary form.

Data also can be classified on the basis of who created it; some data is created by people, and some is created by machines. Humans invent data as an abstraction of the reality they experience, but they can design machines to create data automatically. Human-created data is produced by people when they use their devices to conduct transactions (e.g., purchasing movie tickets). Most data is machine-created—produced without human intervention. Apps in phones, for instance, continuously report their location, and sensors in wearable devices and machines report the physical signal they are capturing, like temperature, pressure, light, or sound.

Data can be classified based on its relationship to the event that triggered its creation, resulting in either event data or circumstantial data. Event data has to do with the main objective of an event. When someone purchases a movie ticket (main event), the event data includes the movie, showtime, movie theater, and so on. Circumstantial data, on the other hand, is indirectly associated with an event; it is data related to the circumstances under which the event happened. When someone purchases a movie ticket, a large amount of circumstantial data can be collected, such as the device used, the location, the starting time and duration, and many other characteristics of the situation in which the event occurred.

Finally, data can be classified based on the nature of its arrangement. Some data is structured, and some is unstructured. Structured data has a predefined arrangement. That is, how the data is arranged (or organized) is established in advance, and the predefined arrangement determines (at least in part) the meaning of the data. For instance, a list of employees in an organization might be arranged to consist of employee id, last name, first name, middle name, and other related information. Unstructured data, as the name implies, does not have a predefined arrangement.

Images are considered UD because the meaning of an image is not inferred merely from the bits it holds. Text is also regarded as UD because the content expressed is more than a collection of characters. Audio and video are also considered UD, the meaning of the recording must be gathered from "seeing" and "listening" to the recording--making sense of it--not from noting the structure of the file in which they are stored.

Dichotomizing data into structured and unstructured is useful to describe its characteristics. However, the pattern of organization is more a continuum than a dichotomy—UD usually includes some degree of organization. Images, for instance, include some data organized in a predefined pattern, such as file name and extension, date, size, and other metadata associated with the image. To facilitate our discussion, we acknowledge the continuum of structure in the data, but classify data in structured and unstructured, as it is customary in data analytics.

UD is multifaceted; that is, it contains multiple concurrent pieces of unique information in a single data point [10]. An individual image, for instance, can be described based on facts (location or the number and type of objects in the picture), or on inferred meanings (happy, sad, or neutral feeling), in addition to the more mundane descriptors of file size and set of pixels. The richness of UD is what makes it powerful; we can gain insights that cannot be obtained from structured data. Although humans process UD naturally, computers require the transformation of UD into a set of structured descriptors (also called dimensions, labels, or features) before the data can be analyzed.

It can be helpful to think about the many forms of data that can be accessed and used by companies as an ecosystem [6]: traditional data from ERPs and legacy computer systems; data captured by scanners; data mined from the Internet (e.g., URLs, click paths, Website content, emails, social media postings, and online news), and data from cell phone usage (e.g., mobility data). The ecosystem can include two additional data domains that are much larger, although not as easily analyzable: audio data (utterances, telephone recordings, media audio streams, and audio surveillance streams) and video data (video surveillance, news-piece videos, cell phone video recordings, and media programming videos). Audio streams can contain not only semantic content but also vocalic content like pitch and intonation. The ecosystem can include the analysis of audio and video data with tools that include vocalic analysis, automatic face recognition, video threat assessment, and others not yet fully developed [6].

3. Cybersecurity and unstructured data

Cybersecurity systems aim to detect, prevent, and protect computers from threats such as computer viruses, Trojans, worms, spam, and botnets, to mention a few [11]. Cybersecurity systems are traditionally designed to fight those threats by collecting data at the network and host level. Data sources include event logs from hosts (desktops, laptops, tablets, and mobile phones) and servers (including active directory servers); network flow logs from routers; domain name server (DNS) lookup records; web proxy logs; antivirus logs; cyber-incident response tickets; and intrusion detection and prevention systems.

A recent area of concern are threats from the Internet of Things (IoT)--connected devices. It is estimated that by 2020 there will be 32 billion IoT devices connected to the internet [12]. Data created by IoT devices, although already in digital format, is to a large degree unstructured, because communication patterns can be at irregular intervals of time and transmit images, video, or audio, and sensor-data. The number of attacks enabled by IoT devices has increased due to their ubiquity in businesses and homes [13]. Connected devices can be used to gain control over or attack a network, or as bots in a botnet-based distributed denial of service attack. These threats are the result of poor security of IoT devices that have reduced processing power for encryption and insecure communication protocols [12]. Traditional methods to identify compromised IoT devices include the analysis of UD to detection of unusual activity such as spikes in internet usage and cost, slow devices and connections and unusual Domain Name Service queries [14]

Another area of concern is social engineering, in which criminals exploit human psychology to deceive users and gain illegal access to computer systems and networks [15]. Social engineering attacks are increasing because they are an inexpensive, yet efficient, method of reaching large pools of potential victims [16]. Preventing social engineering attacks have largely relied on analyzing text data from emails. Similar to the scanning for viruses in which code is compared to code from known viruses, email content is compared to content known to be from social engineering attacks. Social engineering, however, has moved beyond text data to include more convincing ways of delivering content (audios and videos) and communication channels other than email (text and social media). Detecting social engineering attacks from this type requires the creation of new datasets of known social engineering attacks and the ability to analyze UD in real time.

UD can also be used after criminals have obtained the credentials of a legitimate user to access the systems. An increasing number of biometric techniques, which analyze UD, can be used to verify the identity of users. Beyond iris and fingerprint scanners, biometrics like keyboard typing patterns can be used to compare the expected typing pattern of the legitimate user against a given pattern. Using typing patterns, a criminal can be identified because typing patterns do not match. As most cybersecurity techniques, this technique requires the creation of biometrics data for benchmark and real time analysis of UD.

An additional area of concern is insider threats, in which a “malicious insider ... intentionally exploits his or her privileged access to the organization’s network, system and data, [to] ... negatively affect the confidentiality, integrity, and availability of the organization information” [17:1397]. In 2014, approximately 92% of organizations reported data security incidents, where 74% of those incidents were originated by insiders [18]. Insider threats are commonly detected by identifying unusual activity. However, this analysis can be supplemented with the analysis of UD publicly available. For instance, a relational analysis of an employee can determine whether he or she has links to competitors or suspicious entities. Also, the analysis of the employee’s social media postings can signal whether the employee has expressed disgruntlement with the company, because disgruntlement can lead to illegitimate actions against the company.

Although cybersecurity has traditionally focused on data created within an organization, opportunities arise from analyzing UD generated outside the organization. For instance, the severity of vulnerabilities can be forecasted analyzing tweets [19]. Tweets can also be analyzed to extract topics, opinions, and knowledge related to security breaches from consumers. Social media can be a valuable tool for tracking security breaches, and sentiment score and impact factors are good predictors of public opinions and attitudes towards security breaches. Beyond text, images can also be used for identifying malware variants with accuracy of over 89% [20].

UD can also be used to monitor communication among criminals coordinating their attacks. Knowing that some communication channels are monitored, criminals have moved away from email--a highly monitored channel--to communication channels with limited or null monitoring such as video games [21]. Criminals could also coordinate with audio and video, highlighting the need to monitor this type of UD.

4. Internal control and unstructured data

Internal controls are the policies and procedures implemented to provide reasonable assurance of the reliability of the information, safeguarding of assets, and compliance with laws and regulations. Until now, auditors have relied almost exclusively on transactional data to evaluate the reliability of the information and compliance with laws and regulations. Data is typically drawn from the structured databases of accounting information systems. Analyses of this data attempt to identify unusual patterns of transactions--anomalies-- that can be the result of errors, as well as fraud, bribery, money laundering, or other illegal activities.

To identify anomalies, auditors must first establish a benchmark pattern, in terms of quantities, prices, dates, and potentially other pieces of structured information. They then compare the results of their analyses with the established benchmark to identify anomalies; any found are investigated further. Not all anomalies are necessarily fraud or illegal activities; they may be due to unusual but legal events.

The use of UD for internal control is not new, as it has been used before as evidence supporting fraud cases. In 2017, the Securities and Exchange Commission (SEC) used satellite images to demonstrate that a construction company recognized revenue for buildings that had not been built at all. However, it is until recently that UD can be systematically analyzed for internal control. “Auditors should seek to verify transactions, not with just an invoice and receipt, but multi-modal evidence that a transaction took place. Photo, video, GPS location, and other meta data could accompany transaction data” [6:9].

Text data can be processed “to extract textual features such as part of speech, readability, cohesion, tone, certainty, tf-idf scores, and other statistical measures” [6:5]. The SEC, for instance, analyzes text disclosures, computing “tonality” indexes, which reflect the positive or negative tone used in the written discussion of the results. Tonality indexes are then compared with the analysis of the structured data (data from the financial statements). The expectation is that the tonality of text disclosures and the analysis of structured data should match, unfavorable results should align with negative tonality; favorable results should align with positive tonality. Divergence between the analysis of structured and unstructured data would raise a flag for further investigation.

Similarly, text data from transactions can be analyzed along their structured data. Internal control policies commonly require accounting entries to have a written description of the concept originating the transaction. Anti-corruption investigations have analyzed the text on the accounting entries to identify unusual patterns that may reflect bribes, Beyond text data, audio and video conversations can also be used to identify collusion or bribery. Audio and video provide richer information than text data because subtle features, like irony or jokes, could be inferred from the pace and tone of the conversation. In addition to single conversations, a relational analysis of who is related to who can help uncover unknown patterns. A well-known application of relational analysis is the Panama papers, in which, among other things, relational patterns were used to identify players in money laundering and tax evasion schemes.

Safeguarding assets includes installing protections and continuous verification of their existence to prevent theft. Radio frequency identification chips (rfid) attached to inventory items have allowed the tracking of inventory items in real time, providing data not only about their existence but also about their movement. Videos also provide information about movement of inventory items, and because they record the entire environment, videos can provide information about the person handling the items. Amazon self-service stores, for instance, use video to track consumers in their stores and determine the items that consumers place in the baskets for automatic check out. Video does not need to come from fixed cameras. Drones have allowed the use of video to automatically scan inventory items in warehouses or outdoor locations. Audio has also been used for protecting assets. Budweiser for instance, compares the audio of its equipment to benchmark of equipment functioning normally to determine when maintenance is needed before the equipment breaks down, thus preventing factory downtimes.

5. Analyzing unstructured data

The techniques used to analyze UD vary depending on the type of data. Some techniques are well developed, others are still emerging. By far, the largest number of machine-based approaches to understanding UD involve textual data [10]. A bag of words approach treats a text as a collection of words; it does not attend to grammar or the ordering of words. A computational linguistics approach employs rules and statistical procedures to identify linguistic aspects of a text. A custom dictionary approach relies on a list of words and phrases put together by the researcher for a particular purpose. Lexicon-based sentiment analysis also makes use of a list of words, but this kind includes emotional valences attached to each. Linguistic style matching compares the word choices of people known to have contributed to a text, determining how similar their contributions are. Natural language processing, including speech recognition (discussed below), makes use of syntax, semantics, and discourse to assign meaning to naturally-occurring text. Ontology learning-based text mining examines terms, attributes, values, and relationships in a text to identify domain-specific concepts. A pre-existing dictionary approach uses a list of words and phrases not created by the researcher; the list is not context- or purpose-specific. Semantic text analysis considers the relationships between various parts of a text (i.e., phrases, clauses, sentences, paragraphs) and the overall text in attempting to derive a language-independent meaning. A sentiment analysis approach seeks to determine the affective state (negative, neutral, positive) of a communicator from analysis of a text; it is also known as opinion mining and voice of the customer analysis. Finally, text mining attempts to determine meaning through text categorization, clustering, summarization, and extraction of concepts.

In addition to the textually-oriented UD-analysis methods just discussed, some machine-based approaches analyze non-textual data. Among the most familiar of these are speech recognition (aka, voice recognition), which "enables computers to interpret human speech and transcribe that speech to text, and vice versa" [22]. Beyond this, several non-textual UD approaches are available or developing rapidly [10].

Image analysis extracts information from images through a variety of techniques, some tied to specific tasks. Image classification groups images based on patterns or proximity of pixels in the data. Computer-assisted voice analysis identifies non-verbal content, such as prosody, pitch, and speech rate, for extra-linguistic purposes. Computer vision is an emerging area of approaches that is aimed at developing computers that will be able to understand images and videos, representing their meanings as numbers and other symbolic outputs [10].

Additionally, UD analysts apply some well-developed and developing techniques both within and outside the approaches discussed above [10]. A number of techniques are tied to general machine learning. Some of these, like neural networks (or ANNs), are non-deterministic [23]; others can be deterministic, including deep learning, supervised learning, semi-supervised learning, and unsupervised learning. Other task-specific machine learning algorithms include naïve Bayes classifiers, support vector machines, latent Dirichlet allocation (aka LDA), the Viola-Jones algorithm, the conditional random field algorithm, and the Girvan-Newman community clustering algorithm. At least two other algorithms exist, as well: the Porter-Stemmer algorithm, and the pointwise mutual information algorithm [10].

Given the reality of UD, cybersecurity experts and auditors need to be knowledgeable about a variety of statistical and technological topics that they may not have learned yet, including exploratory data analysis [24], NoSQL databases (like Cassandra and HBase), MapReduce (or Hadoop or other tools for processing parallelizable problems across large datasets), and cloud services. Investigators also need to develop skills with emerging audit analytics, such as continuity equations [25], cluster analysis [26], and process mining [27].

[28] discusses the use of topic models created by using Bayesian statistics and machine learning to identify the "thematic content of unlabeled documents, provide application-specific roadmaps through them, and predict the nature of future documents in a collection." [29:16]. Topic models can be applied to text, images, music, DNA sequences, and other kinds of info. They can identify links between documents and latent/hidden structures. In the creation of topic models, analysts use algorithms like LDA (latent Dirichlet allocation) to discover topics and their distributions in documents. Recent variations on LDA-based topic modeling tools need not be told in advance what the topics are and can evolve the number and relationships among them. Some of the tools can find correlations among topics. Some can handle vocabulary changes in topics over time. Some can connect mention of entities, such as organizations and people, in a document, based on topics identified [28].

6. Challenges for using unstructured data

At a high level of abstraction, analyzing UD follows the same general information systems model as any other system: input, process, and output. However, the characteristics of UD add unique challenges in each step. [29] conducted a literature review for papers published between 1996 and 2015 to identify the challenges for exploiting big data. Although their review does not focus on UD, their findings are applicable because a large portion of big data is unstructured. They classified challenges in three categories: data, process, and management.

In the category of data, [29] identified seven challenges (volume, velocity, variety, variability, veracity, visualization, and value) that largely overlap with the "six V's" noted by [30]. These challenges are due to the data itself. Most of these challenges (volume, velocity, variety, and variability) cannot be manipulated by investigators. People, sensors, and machines will continue to create data at an even larger volume and speed and investigators are limited to ensure that the technology they use stores and processes the data. Investigators, however, must tackle the challenge of veracity.

Veracity is a key element for any insight that can be gained from analyzing UD. As the old computer acronym GIGO (garbage-in, garbage-out) indicates, summarizing data and reporting them as facts without using verified data leads to misinformation. The recent Google blunder reporting people death exemplifies the errors of automatically processing large amounts of UD without a verification process. Google's algorithm automatically responds to queries for public figures with 'knowledge panels' that summarize information from the web; the ability of information owners to correct Google's incorrect results is limited. Although the veracity of all data should be verified, UD generated for business purposes inside an organization is more likely to have higher quality than UD generated outside an organization, where there is no vetting process.

Cybersecurity experts and auditors relying on UD generated outside organizations should devise mechanisms to evaluate the veracity of the data. For instance, an auditor could analyze social media to determine whether an increase in revenue is explained by consumers' acceptance of a new product. The expectation is that good performance would be aligned with positive reviews. However, social media postings can be manipulated; companies sell reviews, likes, tweets, that can be purchased to manipulate people's perceptions. Before accepting social media postings as legitimate, auditors would need to conduct a more thorough review, for instance, conducting a relational analysis to determine whether the company being audited has links to companies selling positive data in social media. The SEC has long been aware of the ability of social media to manipulate opinions and continuously investigates whether postings are legitimate from people with no conflict of interest or are intentionally created to manipulate share prices.

The veracity of data is not limited to text data. As the recently emerged fake videos from public figures demonstrate, the ability to create high quality false videos and images makes almost impossible to distinguish fake from legitimate videos. Cybersecurity experts analyzing public videos posted in social media to identify potential threats should also evaluate the veracity of the videos, as criminals could create fake videos to mislead investigations.

Perhaps the most available way to verify the veracity of UD is using multiple verification factors, similar to the multifactor authentication used for access control, in which the identity of a person is verified with two factors, commonly a password and a token. For instance, the content of UD can be triangulated with geographical (location) and relational (how the person is related to the content) data. Data triangulation can support the verification process at the cost of adding complexity but is needed to ensure high quality data.

Data triangulation cannot only be useful to evaluate the veracity of the data, but to enrich the analysis. [17], for instance, proposed adding contextual data to the traditional host and network data to detect threats from people within an organization. Contextual data provides supplemental information about a person. Employment data, for instance, can be obtained from the Human Resource department of the organization, and psychological data can be estimated based on a person's social media posts and activities, or dynamic of social connections. In addition to supporting data verification, contextual data can reduce the rate of false positives (flagging an event as a threat when it is not).

In the category of process, [29] identified four challenges (acquisition & warehousing, mining & cleansing, aggregation & integration, and analysis & modeling), similar to the taxonomy of processes presented by [30]. These challenges are related to processing the data. Among these challenges, data cleaning and integration, described as extract-transform-load (ETL) in data analytics terminology, is of special relevance. UD shares all the challenges that structured data has, from unformatted form to unexpected missing and noisy data, but UD has unique challenges that make the cleaning and integration more difficult. For instance, UD from networks or hosts includes multiple dissimilar files and formats used for logging events [31], requiring the reformatting and merging of the data before it can be used. In addition, UD needs to be structuralized before it can be analyzed. That is, the richness of UD is reduced to numerical values than can be manipulated by computers. After widely publicized blunders on image recognition, like Google's algorithm identifying black people as gorillas, the accuracy of algorithms for image recognition is increasing. However, there are many features than can potentially be extracted from images beyond objects. Features on the mood depicted on the image, like aggressiveness, friendliness, or excitement, can be highly informative and yet, difficult to detect with algorithms.

In the category of management,. [29] identified six challenges (privacy, security, data governance, data & information sharing, cost/operational expenditures, and data ownership). Although these challenges are common to all types of data, privacy is a key concern for the use of UD. Regulations like Europe's General Data Protection Regulation (GDPR) establishes limits on the use of people's data. Even without regulation, scandals of data misuse, like mobile apps sending users' data to Facebook, have increased public awareness and disconformity with the undisclosed and lack of control of persona data. Even within the workplace and beyond regulations, there are ethical concerns on an individual's right to privacy that need to be taken into account. A key element to use UD without compromising the analysis has been data anonymization. Thasos Group, for instance, used number of cell phone signals going in and out of Tesla's factory to determine whether the company was indeed ramping up production as promised. In cybersecurity, the data collected should follow security protections for removing IP addresses, hostnames, and usernames. The anonymization should make it difficult to correlate the data with other external data.

In addition to the challenges identified by [29] and [30], we identified three challenges: digitization of non-digital UD, bias and explainability of algorithms, and availability of UD. Digitization of non-digital UD converts non-digital data to digital format. The extraction of features is what gives power to UD. For instance, employees scan pictures of the New York Times' archive, and Google's algorithms extract information beyond identifying objects, dates, and location. The algorithms aim to infer meaningful content beyond facts. In addition to archival data, non-digital UD is still used. For instance, people may distrust reporting financial wrongdoing through phone, email, or chat, for fear of being identified; an old fashion paper note dropped in a secured box may provide the necessary anonymity. Eliminating the report of financial wrongdoing through secure boxes just for the sake of eliminating non-digital data may deter the reporting of financial wrongdoing.

Bias and explainability of algorithms refer to the inability of algorithms analyzing UD to define the criteria used to reach a decision, which can be inadvertently biased [32]. Amazon, for instance, developed an algorithm to analyze text data--resumes--to identify promising applicants. The initial data set contained a larger proportion of men than women, as it is common in technology jobs, resulting in a biased algorithm that evaluated more favorably men than women. After unsuccessful attempts to debias the algorithm, Amazon discontinued the project. Biases can creep not only on text data but other types of UD. Regulations might limit the use of UD for the lack of explainability. The GDPR, for instance, requires explicit explanations on how decisions are reached when the decisions have a significant impact on people's lives. Denials of mortgage loans or jobs must explicitly indicate the factors for denial, so people have an opportunity to improve.

Availability of UD refers to the limited availability of organizational data for academic research. Data for actual cybersecurity breaches may be omnipresent, as investigators are given access to all data, but data for cybersecurity research is scarce at best [31]. In the case of dynamic network research, the lack of data exists because the majority of computer event logs are created to monitor operations and are formatted to be processed by humans instead of data analytics. The availability of anonymized UD for research purposes would enable the reproducibility of cybersecurity research [33]. The same is true for internal control investigations: data available to auditors in practice is ubiquitous and abundant, but data obtainable by academic researchers is less common or plentiful. UD available for research, like the Enron corpus, is an exception.

Although we discuss these challenges independently, it is reasonable to expect that complex interactions among the challenges will be observed. For example, low veracity (a data challenge) would make data mining & cleansing (a process challenge) more difficult, as well as reduce opportunities for effective data & information sharing (a management challenge).

7. Research framework

We propose the research framework in Figure 1 for the scholarly study of UD for cybersecurity and internal control. Although the constructs in the framework are high-level, they represent the crucial elements that ought to be considered in most studies of how cybersecurity and internal control investigators should (and actually do) use UD.

The constructs in the framework include much of the preceding discussion: 1) the goals of a cybersecurity investigation and the goals of an internal control investigation, which determine the tasks that the investigator must do; 2) the types of UD that must be accessed and the kinds of analysis that must be done, which are determined by the tasks to be done; and 3) the outcomes achieved, which are determined by the analysis done on the UD accessed. Although not depicted in the figure, it is likely that some elements in the framework would be moderated by organizational characteristics and by the knowledge and skills of the investigator.

Perhaps obviously, any individual research project may emphasize some of these, but not others, if that serves the research questions to be addressed. In general, though, this framework is useful as a guide for what categories of variables/factors ought to be addressed. A researcher should have a good reason to omit entirely all concepts that are comprised by any of the framework categories. In the following section, we describe how researchers can design studies that make use of ideas from the framework to address each of the main goals of academic research in this area of interest.

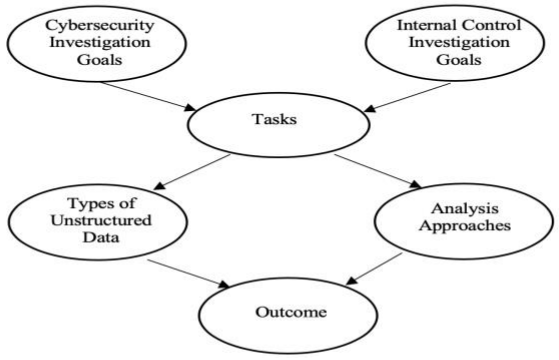
Cybersecurity Investigation Goals
Internal Control Investigation Goals
Tasks
Types of Unstructured Data
Analysis Approaches
Outcome

Figure 1: Proposed research framework for the study of UD, cybersecurity and internal control investigations

8. Utility of the framework

Research about information systems seeks to describe, explain, predict, and effect changes. Research about UD in cybersecurity and internal control pursue all of these ends.

The proposed framework can support research for description through aiding in the selection of relevant case studies. How can/do organizations use UD in their investigations of cybersecurity and internal control? How successful are they? What limits/challenges do they face? What opportunities and next steps do they see?

Researchers who are interested in conducting investigations of relationships among latent and manifest variables—to pursue research for explanation—can use the framework to develop structural equation model studies. What are the factors that influence and are influenced by UD in investigations of cybersecurity and internal control? What causal paths exist? How can one measure the factors? What are alternative models for the variables of interest?

Researchers who wish to advance explanation through the testing of theoretical propositions will find the framework useful in the design of experiments. The framework helps researchers posit and test hypotheses, common in both true experiments and quasi-experiments. The categories provide suggestions concerning how to avoid confounding factors and what to control to make tests as powerful as possible.

Researchers who seek to be able to predict the values of variables associated with cybersecurity or internal control could use the framework to aid in specifying a variety of regression studies. The framework helps researchers who want to use traditional models for prediction that can be assessed with regression and related techniques.

The framework can also be of use by investigators who wish to do research using data-analytics methods. What does the use of UD in investigations of cybersecurity and internal control tell us? There must exist one or more large corpuses of information about how organizations have been investigating cybersecurity and internal control, and much of it must be UD.

Researchers who are interested in advancing what can be known about design in this area can also find the framework to be helpful. Both action research and design research benefit from being pursued with theory in mind. The framework can help one consider critical questions like: What is the best way to initiate the use of UD into these processes? How can existing process be improved? Can research come up with “solutions” that would work for adopting organizations in different contexts? What artifacts (i.e., tools, methods, models) related to the use of UD for cybersecurity and internal control can be designed?

9. Conclusion

Data analytics can identify patterns on UD relevant for cybersecurity and internal control. The characteristics of UD add unique challenges for its analysis and management. This paper highlights some of the challenges for the use of UD and proposes a research framework for studying the connections--realized and potential--between UD and cybersecurity and internal controls.

The framework includes cybersecurity and internal control goals that determine the tasks that the investigator must do. The task influences the types of UD that must be accessed and the types of analysis that must be done, which in turn determine the outcomes achieved. Although the constructs in the framework are high-level, they represent crucial elements to be considered in studies of how cybersecurity and internal could use UD.

10. References

[1] Cisco, “What is Cybersecurity?” https://www.cisco.com/ c/en/us/products/security/what-is-cybersecurity.html, accessed Sep 2019.

[2] AccountingTools, “Internal Control”, https://www.accountingtools.com/articles/internal-control.html, published Apr 2018, accessed Sep 2019.

[3] Center for Strategic & International Studies, “Economic impact of cybercrime”, https://www.csis.org/ analysis/economic-impact-cybercrime, published Feb 2018, accessed Jun 2019.

[4] Symantec, “10 cyber security facts and statistics for 2018”, https://us.norton.com/internetsecurity- emerging-threats-10-facts-about-todays-cybersecurity-landscape-that-you-should-know.html, accessed Jun 2019.

[5] Association of Certified Fraud Examiners, “Report to the Nations: 2018 global study on occupational fraud and abuse”, https://www.acfe.com/report-to-the-nations/2018/, published 2018, accessed Jun 2019.

[6] K.C. Moffitt and M.A. Vasarhelyi, “AIS in an age of big data”, *Journal of Information Systems*, *27*(2), 2013, pp. 1–19.

[7] N. Bumgarner and M.A. Vasarhelyi, “Continuous auditing—a new view”, in AICPA (Eds.) *Audit Analytics and Continuous Audit: Looking Toward the Future*, New York, NY: American Institute of Certified Public Accountants, Inc., 2015, pp. 3-52.

[8] J. Gantz and D. Reinsel, “The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the Far East”, IDC – EMC Corporation. http://www.emc-technology.com/collateral/analyst-reports/ idc-the-digital-universe-in-2020.pdf, published Dec 2012, accessed Jun 2019.

[9] C. Dobre and F. Xhafa, “Intelligent services for big data science”, *Future Generation Computer Systems, 37*, 2014, pp. 267–281.

[10] B. Balducci and D. Marinova, “Unstructured data in marketing”, *Journal of the Academy of Marketing Science*, *46*, 2018, pp. 557–590.

[11] S. Dua and X. Du, *Data Mining and Machine Learning in Cybersecurity*, CRC Press, 2011.

[12] B.B. Gupta, A. Tewari, A.K. Jain, and D.P. Agrawal, “Fighting against phishing attacks: state of the art and future challenges”, *Neural Comput & Applic*, *28*, 2017, pp. 3629-3654.

[13] A. Alshammari and M.A. Zohdy. "Internet of things attacks detection and classification using tiered hidden Markov model", in *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, 2019, pp. 550-554.

[14] Z. Zeljka, “The FBI warns about compromised IoT devices”, https://www.helpnetsecurity.com/ 2018/08/06/ spot-compromised-iot-devices/, accessed Jun 2019.

[15] S. Abraham and I. Chengalur-Smith, "An overview of social engineering malware: trends, tactics, and implications", *Technology in Society*, *32*(3), 2010, pp. 183-196.

[16] N.L. Muscanell, R.E. Guadagno, and S. Murphy, "Weapons of influence misused: a social influence analysis of why people fall prey to internet scams", *Social and Personality Psychology Compass*, *8*(7), 2014, pp. 388-396.

[17] L. Liu, D.O. De Vel, Q. Han, J. Zhang, Y. Xiang, “Detecting and preventing cyber insider threat: a survey,” *IEEE Communications Surveys and Tutorials*, *20*(2), 2018, pp. 1397-1417.

[18] Clearswift, “Insider threat: 74% of security incidents come from the extended enterprise, not hacking groups,” https://www.clearswift.com/about-us/pr/press-releases/ insider-threat-74-security-incidents-come-extended-enterprise-not-hacking-groups, published Sep 2017, accessed Jun 2019.

[19] S. Zong, A. Ritter, G. Mueller, and E. Wright, “Analyzing the perceived severity of cybersecurity threats reported on social media, https://arxiv.org/pdf/1902. 10680.pdf, published May 2019, accessed Jun 2019.

[20] J. Hao and H. Dai, "Social media content and sentiment analysis on consumer security breaches", *Journal of Financial Crime*, *23*(4), 2016, pp.855-869, https://doi.org/10.1108/ JFC-01-2016-0001.

[21] A. Makandar and A. Patrot, "Malware analysis and classification using artificial neural network", IIn *2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)*, 2015, pp. 1-6.

[22] Capterra, “Speech recognition software buyers' guide”, https://www.capterra.com/speech-recognition-software/#buyers-guide, accessed Jun 2019.

[23] V. Gavrilov, “What is the difference between machine learning and neural networks?”, https://www.quora.com/What-is-the-difference-between-machine-learning-and-neural-networks, published Nov 2017, accessed Jun 2019.

[24] Tukey, J. W., *Exploratory Data Analysis*, Reading, MA: Addison-Wesley, 1977.

[25] A. Kogan, M.G. Alles, M.A. Vasarhelyi, and J. Wu, “Analytical procedures for continuous data level auditing: continuity equations”, Rutgers Accounting Research Center working paper, 2011.

[26] S. Thiprungsri and M. Vasarhelyi, “Cluster analysis for anomaly detection in accounting data: an audit approach”, *International Journal of Digital Accounting Research*, 2011.

[27] M. Jans, M. Alles, and M.A. Vasarhelyi, Process mining of event logs in auditing: opportunities and challenges, Hasselt University working paper, 2010.

[28] G. Anthes, “Topic models vs. unstructured data”, *Communications of the ACM*, *53*(12), 2010, pp. 16-18.

[29] U. Sivarajah, M.M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of big data challenges and analytical methods”, *Journal of Business Research*, *70*, 2017, pp. 263-286.

[30] A. Gandomi and M. Haider, “Beyond the hype: big data concepts, methods, and analytics”, *International Journal of Information Management*, *35*, 2015, pp. 137–144.

[31] A. Kent, “Cyber security data sources for dynamic network research” in Adams, N., and N. Heard (Eds.), *Dynamic Networks and Cyber-Security*,” Singapore: World Scientific Publishing, 2016, pp. 37-65.

[32] W. Knight, “The dark secret at the heart of AI”, *MIT Technology Review*, *120*(3), 2017, p 55-63.

[33] D. Shou, “Ethical considerations of sharing data for cybersecurity research, in Danezis G., S. Dietrich, and K. Sako (Eds.), *Financial Cryptography and Data Security FC 2011 Lecture Notes in Computer Science*, *7126*, Berlin: Springer, 2012.