BOWIE STATE UNIVERSITY

INFORMATION SYSTEMS SECURITY

TECHNICAL PAPER

AN EFFECTIVE APPROACH TO CLOSING THE BREACH DETECTION GAP (BDG)

INSTRUCTOR
DR. DAVID ANYIWO

GRADUATE ASSISTANT
JERRY DIABOR

BY

Sydney Raymond
David Tan
Khalil Davis

August 3, 2022

Abstract

This research seeks to establish a framework for reducing the breach detection gap (BDG) and therefore limiting the time span in which leaked data consisting of Personal Identifiable Information (PII) is publicly accessible and available for use in cybercrime activities, such as phishing. BDG reduction is defined as the process of limiting the time between a breach's occurrence and its detection. The BDG is a specific problem facing companies engaged in data management or retention, and companies that fall victim to data breaches can be left with damaged reputations and customer losses. The average BDG exceeds three months, reducing user trust in online entities and leaving users' data vulnerable to malicious actors (Botha et al., 2016 and Deidrich, 2019).

Many factors in the current cyber environment have contributed to widening the BDG. In the interest of limiting the damage caused by intrusions, properly identifying these factors - as well as finding technological solutions that efficiently target the most significant of these factors - is essential. There are three primary factors we have targeted as potential application areas for BDG reduction methods: technology, corporate and government policy, and human error.

The qualitative and quantitative methods adopted in this research concerned the BDG, associated causes, and possible solutions. We included an exploration of the application of content analysis to several case studies as well as a research design that employed secondary data collection, with the overarching goal of proposing an intrusion detection framework for organizations to adopt into their own cybersecurity infrastructure. Our qualitative analysis focused primarily on identifying recurring breach detection gap factors and viable BDG reduction solutions from secondary data, i.e. previous research.

The data process and analysis were based on content analysis of the selected case studies, via an examination of several news articles, press releases, and journal publications related to each of the selected case studies. Based on this analysis, we concluded that technological factors were the most frequent contributors to extended breach detection times and selected these factors as the main vector of focus for the development of BDG solutions.

The program created in this study was therefore developed as a multi-tool framework that addresses these factors via detecting the occurrence of a fault in a network system due to a breach. Through extensive experiments, the program has demonstrated the ability to reduce breach detection time with high accuracy, thus eliminating errors resulting from false positives.

Acknowledgements

We would like to thank Dr. Anyiwo and our graduate mentor, Jerry Diabor. We would also like to thank Bowie State University, the National Science Foundation, and our fellow summer undergraduate research participants for their support and feedback throughout the development of this research.

Table of Contents

Tables

Figures

I. Background

In “Data Breaches,” Diedrich (2019) wrote that over the fifteen years prior to the article’s publication, over ten billion records had been breached as the result of over 9,000 data breaches in the United States. The author defines a data breach as “an unintentional release of secure or personally identifiable information to an unsecure environment.” On an individual level, this can include email addresses, passwords, credit card numbers, and financial information. On a corporate level, it can include business-sensitive trade secrets.

Many victims of these cyberattacks and threats have been left with damaged reputations and loss of customers. Diedrich (2019) asserts that the two most common motivations for conducting data breaches are financial gain and espionage. Pieces of data can gain value on the online black market depending on their ability to provide access to new sources of data or unlock other accounts. The author states that “state-affiliated groups and countries are behind ninety-six percent of espionage-motivated breaches.” There is a considerable cyber-espionage threat from both Chinese and Russian state-backed hacking groups, according to the NSA (2022) and Kang (2022). Furthermore, in this era of the Internet of Things (IoT), the risks and vulnerabilities of network systems are increasing exponentially, as there are over 500,000 new internet users and over 2,000 cyberattacks per day (Santos, Inácio, & Silva, 2021 and Wang & Park, 2017).

However, despite the massive security risks associated with data breaches, breaches are often detected by third-party groups rather than breach victims, and the average breach detection gap (BDG), the period between a cyberattack and the system’s response, exceeds three months (Botha, Eloff, & Swart, 2016). Botha et al. (2016) additionally identified multiple instances in which Personally Identifiable Information (PII) remained publicly available for nearly two years. The authors of the paper “Pro-Active Data Breach Detection: Examining Accuracy and Applicability on Personal Information Detected” argued that a reduction in the BDG would reduce the opportunity for cybercrime, indicating the importance of research into gap reduction (Botha et al., 2016). Nevertheless, conventional solutions for reducing the BDG have proven to be ineffective and inefficient against emerging cyberattacks and threats (Santos, Inácio, & Silva, 2021).

Therefore, the aim of this literature review is to conduct a systematic analysis of past and current research studies concerning BDG minimization in order to find an efficient approach to shortening the gap. The systematic analysis of these research studies consisted of categorizing the studies into three foci, studies concerning the identification of factors that contribute to BDG expansion, studies involving potential solutions for closing the BDG, and studies relating to case studies. This specific grouping of literature allows us to identify and construct the proper research concept and methodology for our particular contribution to the field of information system security.

II. Literature Review

In identifying the main factors that contribute to data breach occurrence, Diedrich (2019) cited figures from the Ponemon Institute (2019), which stated that approximately one-quarter of 2019 data breaches were the result of technology failures. Kraemer et al. (2009) additionally identified technology as one of nine core causes of data breaches. If technology failures are major factors in data breach causation, then one can infer that they may also be major factors in the BDG, as these problems do not cease to impact an organization's technological capabilities after a breach has occurred. Further indicating that poor use of cyber security technology is a major cause of unauthorized data access is the article by CISA (2022), which stated that malicious cyber actors "exploit poor security configurations (either misconfigured or left unsecured), weak controls, and other poor cyber hygiene practices to gain initial access or as part of other tactics to compromise a victim's system." Poor cyber hygiene practices could extend to poor or little usage of intrusion detection systems, lengthening BDG times.

The weak cyber security practices identified by CISA (2022) as objects of frequent exploit by these malicious actors include "incorrectly applied privileges or permissions and errors within access control lists," out-of-date software, remote services (like VPNs) without adequate unauthorized access controls, and "open ports and misconfigured services are exposed to the internet." These practices additionally have the potential to increase breach detection time or prevent breach detection entirely. Errors with access permissions, whether in access control lists or remote service controls, can allow individuals to access private data undetected until the permissions are fixed, as occurred with unauthorized official accounts in the Aadhaar card breach (Jain, 2019). Out-of-date software, or unpatched software, leaves systems vulnerable to known exploits until the ignored patch is noted, as in the Equifax breach (Johnson and Wang, 2018). Situations in which data services are misconfigured and exposed to the internet would leave data vulnerable until either the misconfiguration happened to be noted internally or an external detection system like that proposed in Botha et al. (2016) discovered the public exposure. Furthermore, Dolezel and McLeod (2019) additionally cite technological flaws as sources of data breaches and refer to tools including authentication, backups, and data encryption as existing breach protection methods. However, the authors do not discuss the possibility of these tools being poorly employed and therefore

contributing to a lack of breach detection. Data breaches can also result from technology failure or system glitches, according to Diedrich (2019). They can include "application failures, inadvertent data dumps, [and] logic errors in data transfer." The author provides the First American Financial Corporation data breach as an example of this kind of breach. In conclusion, all of these research studies indicate that technology flaws/failures are a factor in the BDG.

The following studies identified inefficient corporate and federal cybersecurity policies as factors in the BDG. For example, in addressing the Equifax data breach, Johnson and Wang (2018) draw upon the work conducted by Wang and Park (2017) who examined incident handling concerning the Yahoo data breaches. One communication strategy proposed by Wang and Park (2017), "denial," is expanded upon by Johnson and Wang (2018) to the strategy of "scapegoating." These incident communication strategies seek to allow corporations to defer blame and avoid responsibility in response to data breaches, indicating that one cause of the BDG is companies' decision to focus on falsely preserving public image rather than investigating breaches. According to Dolezel and McLeod (2019), "by increasing the number of business processes, an organization also increases the probability of a data breach." It may be possible that these numerous business processes and their affiliated increase of potential breach points contribute to lengthening breach detection time. Kraemer et al. (2009) presented several data breach causes related to corporate cyber security policy: management, organization, performance and resource management, policy issues, and training. Dolezel and McLeod refer to six potential root causes of breaches as identified by Kamoun and Nicho (2014): security culture, governance practices, policies and procedures for handling security, ongoing employee security training, vendor selection, and risk management processes.

Cheng et al. (2017) divided the causes into intentional and inadvertent threats. Most of the intentional threats, such as social engineering and hacking, were further classified as external threats. All of the inadvertent threats, such as configuration error and privilege abuse, were further classified as insider threats. Two examples of intentional threats, cyber espionage, and sabotage, were also further classified as insider threats. The paper also discussed how data leaks could be classified by other attributes, such as industry sector or type of occurrence. The example that was used was that the business and medical/healthcare sectors are where the majority of data leaks

happened (45.2% and 34.5% respectively in 2016). As for the type of occurrence, their graph indicated that the leading cause of enterprise data leak threats is due to insider threats, with more than 40% of breaches penetrated from inside the company.

Dolezel and McLeod (2019) stated that funding derived from the Health Information Technology for Economic and Clinical Health Act led to the proliferation of electronic health records via mobile devices, further leading to an increase in data sharing and remote collaboration. The authors hold that the use of mobile devices to transmit medical data causes an increase in security vulnerabilities as it places excessive responsibility on the user. This increased data sharing resembles both the shared responsibility model of cloud computing referred to in Lane (2017), which could also cause failures in breach detection via unequal participation and blaming, and the proliferation of IoT devices as described by Isaak and Hanna (2018). Meanwhile, through their research, Botha et al. (2016) determined that South African privacy and data breach legislation had little impact on PII disclosures. Although this discovery is particular to a singular country, it raises further questions about the effectiveness of technological legislation globally in actually reducing the quantity of private data available for malicious use. Therefore, these studies supported the assumption from previous studies that, in addition to technology failures, misguided and insufficient corporate and federal cybersecurity policies also influenced the BDG.

The following studies identified human error as a factor in the BDG. Figures from the Ponemon Institute (2019) additionally stated that approximately one-quarter of 2019 data breaches were the result of human error. Kraemer et al. (2009) also viewed human error as a core cause of data breaches. Diedrich (2019) identified human error as the most "preventable" factor and provides "failing to fix known vulnerabilities," such as those that caused the Equifax data breach, according to Wang and Johnson (2018), as an example of human error. This category additionally includes the mistaken publication of private data in public locations, such as incidents that occurred with Aadhaar data as noted by Jain (2019). Human error can lengthen the BDG as the time it takes for an organization to note that a required patch was not performed lengthens the BDG for breaches that resulted from the unpatched vulnerability. Inadvertent public exposure of private data can also incur an extended BDG as it is not the result of a malicious intrusion that could be detected by IDS. Furthermore, the high accuracy rates of current IDS technologies indicated by Saranya et al.

(2020) contrast sharply with the current state of intrusion detection and raise questions about why such a large BDG exists for data leaks when accurate IDS are available. Put into the context of questions raised about the human response (i.e. human administrator's dismissal of IDS alerts) to the false positives mentioned by Chen and Aickelin (2006) and Ahmad et al. (2020), it seems possible that the BDG is primarily a result of human error, as it could take longer for administrators to officially "detect" breaches that have already been technologically detected. As the result of the perspectives offered in these studies, technology failures, inefficient corporate and federal cybersecurity policies, and human error were identified as factors that contribute most significantly to expanding the BDG.

Once the factors that contribute to the widening of the BDG were identified, possible solutions to shortening the BDG were selected from the analyzed research studies. For example, Botha et al. (2016) proposed pro-active automated breach detection as a method of reducing detection times and therefore limiting the period in which Personal Identifiable Information (PII) is publicly accessible and available for use in cybercrime activities. Botha et al. (2016) suggested an application that scans the internet for documents containing leaked PII, collects the PII, and obtains an IP address and approximate geographical location for websites responsible for the dissemination of personal data. The current application is limited in that it can only extract specific forms of PII: ID numbers, landline numbers, cell phone numbers, email addresses, and card numbers and addresses. However, combining a lexical analysis AI similar to that described by Perotto et al. (2020) with this framework may allow for the additional identification of names, job locations, and religious beliefs. The application is also limited by the fact that it requires manual or human intervention to classify collected data as leaked PII and that it only addresses data breaches in which leaked data has been made public rather than kept private by online criminal groups. However, organizations could potentially use the system to identify instances of their private data being exposed publicly and could improve the system's flaws by giving it the capability to additionally scan dark websites for maliciously used data.

The authors Jamil et al. (2022), concur that in the era of big data, the current functionality of the technology, such as big data analytics, security, connectivity, centralization, hardware capabilities, user data privacy, and GIS visualization, etc., is limited in terms of keeping up with the current

demand. Jamil et al. (2022) focused on the issue of trickiness when it comes to providing data integrity, nonrepudiation, and event management due to the centralized system caused by sophisticated devices such as gateways in a simple IoT network. The article confirmed three main components when looking at the QoS (Quality of service) assessment which are the response time, cost, and availability; the response time was between twelve and fourteen seconds, which the researchers guaranteed could decrease. The cost is not as certain since the price varies per application. The availability was at 99.9% and resulted in neither packet loss nor outrage when doing the tests. However, due to the constraints placed on the test environment, an alternate approach may have to be sought out when dealing with a bigger and/or more complex environment. Finally, even though limited right now, the article does reveal a possible approach to take on if the environment is ideal; however, it does reveal that cost efficiency may not be ideal due to the varying price based on the application. Lei Hang and Do-Hyeun Kim (2019) discussed the integration of IoT-based technologies to everyday electronic appliances such as smartphones along with the challenges of such an integration. The challenges stem from the inferior development speed of embedded devices, low computing powers, and limited data storage in comparison to desktop systems that operate transactions properly using the current blockchain-based systems.

The article presented by Cymulate (2022) on cybersecurity testing solutions focuses primarily on Breach & Attack Simulation (BAS) technology, which simulates a "threat actor's hostile activities with some level of automation." The authors identify three BAS categories: agent-based vulnerability scanning solutions, "malicious" traffic-based testing solutions, and blackbox multi-vector testing solutions. Cymulate (2022) described one category of BAS technology, "malicious" traffic-based testing solutions, as a system of setting up Virtual Machines within an organization's network and using a database of attack scenarios to send attacks between the Virtual Machines. The test is then monitored to determine if an organization's cyber security systems can detect malicious traffic. The organization can then assess which attacks did not trigger alerts and use the list of rules and alerts provided by the test to block future such attacks. A system like this would be able to test the efficacy of suspicious traffic detection systems like those which eventually uncovered the Equifax breach (Johnson and Wang, 2018). Honing systems in this manner would reduce the impact of poorly-equipped technology, ideally reducing one factor in the BDG.

According to Cymulate (2022), blackbox multi-vector testing solutions, another category of BAS technology, are capable of detecting vulnerabilities in both an organization's perimeter and internal network and approaching an organization's cybersecurity capabilities in a manner more closely resembling that of a cybercriminal or malicious hacker. These solutions are mostly cloud-based and conduct assessments consisting "of multi-step tests utilizing distinct adversary types of attack techniques, tactics and practices (TTPs) and payloads in an attempt to bypass security measures both internal and external to an organization's local area network. A report is then generated, revealing discovered security vulnerabilities. Detecting vulnerabilities in this manner may be helpful for gap reduction as it would allow organizations to detect vulnerabilities that may be actively allowing a breach to occur, such as the unpatched software in the Equifax breach (Wang and Johnson, 2018). However, even this most complex model only includes simulations covering "the newest threats detected in-the-wild," according to Cymulate (2022), and therefore is incapable of testing for threat types that have not yet occurred on other systems.

The article presented by Saranya et al. (2020) on Intrusion Detection Systems (IDS) primarily focuses on conducting a review of existing Machine Learning (ML) algorithms being used for intrusion detection. In "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," the authors additionally seek to classify intrusions using ML algorithms with the KDD-CUP dataset. Saranya et al. (2020) referred to the "7v's" of big data: volume, velocity, variety, value, veracity, variability, and visualization. The authors establish that IDS can be either hardware systems or software systems and cite three categories of IDS: anomaly- or behavior-based detection, signature- or knowledge-based detection, and hybrid-based detection. The authors' reference to a "hybrid-based" detection system expands on the two categories written about by Chen and Aickelin (2006), who established their two categories as "misuse detectors" and "anomaly detectors." Their use of the term "misuse detectors" seems to operate synonymously with the "signature-based" category of Saranya et al. (2020).

Anomaly or behavior-based IDS generates alerts when user, network, and host system behavior differs from normal behavior. However, as stated by Ahmad et al. (2020), these alerts often come with impractical and inefficient false alarm rates. These false alarms can then distract from genuine alerts, leaving breaches undetected. Saranya et al. (2020) wrote that signature or knowledge-based

detection systems rely on a database of attack signatures and known vulnerabilities. The weakness of this system is that, as stated by Chen and Aickelin (2006), it cannot detect day zero attacks. According to Saranya et al. (2020), hybrid-based IDS combines anomaly- and signature-based detection. The authors additionally classify IDS into active IDS and passive IDS. Passive IDS monitors and analyzes traffic and produces alerts about attacks and vulnerabilities. Active IDS performs the same functions as passive IDS and additionally takes action to block suspicious traffic.

Using metrics including accuracy, precision, recall, and F-Score, Saranya et al. (2020) reviewed the performance of ML algorithms including Modified K- Means, J.48, Support Vector Machine (SVM), decision table, Principal Component Analysis (PCA), Logistic regression, decision tree, and Artificial Neural Network (ANN) for IDS and other algorithms like Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART) and Random Forest (RF) algorithms to classify intrusion detection. According to the authors, the detection rate, false positive rate, and accuracy depend both on the type of algorithm used and the application area. The high rate of ML IDS accuracy, between 92.7 percent and 99.81 percent for the technologies included by Saranya et al. (2020), indicates that core intrusion detection problems that lengthen the BDG may not necessarily come as a result of IDS technology itself but rather as a result of human error and poor use of the technology. This type of behavior would include the dismissal of legitimate alerts due to the occasional issuance of false alarms, as mentioned by Ahmad et al. (2020).

The section of improvement strategies proposed by CISA (2022) that is most valuable for research concerning the BDG is that which discusses employing detection tools and searching for vulnerabilities. Recommendations under this category include employing an intrusion detection system (IDS), such as those discussed by Ahmad et al. (2020), or an intrusion prevention system. CISA (2022) recommends conducting vulnerability scanning, such as that discussed by Cymulate (2022), “to detect and address application vulnerabilities.” The authors additionally recommend implementing “endpoint and detection response tools,” conducting “penetration testing to identify misconfigurations,” and using “cloud service provider tools to detect overshared cloud storage and monitor for abnormal access.” Although some of these strategies relate primarily to breach prevention, IDS and using cloud service provider tools can be utilized for breach detection. CISA’s

(2022) final suggestion is the implementation of a "software and patch management program." Routinely identifying software in need of patching may have helped to prevent or detect the Equifax data breach, as the breach occurred via an unpatched vulnerability, according to Wang and Johnson (2018). Therefore, the possible solutions proposed for shortening the BDG ranged from proactive automated breach detection to adopting corporate patch management programs.

Once the factors and solutions relating to the BDG were identified, case studies were then analyzed and identified in order to identify the most efficient approach to shortening the BDG. For example, Johnson and Wang (2018) stated that the Equifax breach included social security numbers, birth dates, addresses, driver's license numbers, credit card information, and financial dispute documents for over 44% of the U.S. population. On March 9, 2017, the United States Computer Emergency Readiness Team (US-CERT) announced a patch resolving "a vulnerability in Apache's Struts 2 software." That same day, Equifax issued an internal request that the software needed to be patched. On March 10, 2017, the breach began via this as yet unpatched vulnerability. On May 13, 2017, those conducting the breach began accessing files with PII. On July 29, 2017, those conducting the breach were detected. The following day, the vulnerability was eliminated. According to Equifax (2017), system intruders were detected internally via the eventual identification of suspicious traffic. However, the time gap between when the intrusion occurred and when the intruders were detected indicates that it may be beneficial for companies like Equifax to consider additional solutions. For example, a system like that presented in Botha et al. (2016) could have potentially detected the breach externally.

Johnson and Wang (2018) cited several examples of poor data security behavior from Equifax following the data breach, indicating that being the victim of a massive data breach does not necessarily improve company security behaviors. Equifax's official Twitter account repeatedly tweeted a phishing link in weeks following the breach, and it was discovered that the username and password combination of "admin" and "admin" was used to secure an Equifax web portal. Therefore, in examining BDG factors, it is important to note that previous data breaches experienced by a company may do little to improve company behavior or accelerate threat detection. The approaches recommended to businesses within the article by Johnson and Wang (2018) illustrate that organizations that fall victim to data breaches often devote more time and

energy to repairing their public image than to improving their cyber security capabilities. Additionally, the authors write that the Equifax data breach was the result of a failure to patch a known vulnerability. Although inadvertently, the authors indicate that the greatest factor in broadening the BDG is human error.

The article presented by Jain (2019) on the Aadhaar data breach primarily focuses on the cybersecurity problems with the Aadhaar Card, the resulting privacy issues, and the larger impacts of the data breach. In "The Aadhaar Card: Cybersecurity Issues with India's Biometric Experiment," the author additionally seeks to argue against the existence of Aadhaar by presenting an analysis of its shortcomings. Jain (2019) wrote that Aadhaar was created in 2009 as a "tool to standardize the process of data collection and ease the dispersal of money from government schemes to the citizens of the country, especially the poor." According to Jain (2019), Aadhaar became one of the world's largest databases by associating a 12-digit unique ID number with a citizen's fingerprints, retina scans, and face photos and garnering about 1.2 billion enrollments. The card is additionally linked to individuals' driving licenses, school scholarships, cooking gas subsidies, passports, pensions, provident fund accounts, and ATM services. Jain (2019) asserts that approximately 200 official government websites inadvertently made private Aadhaar data publicly available. The author additionally states that thousands of government databases containing private data could be accessed via Google. This mirrors the process described by Botha et al. (2016), in which the researchers created an application that discovered publicly available leaked data simply by searching the internet. It is evident that a system like the one described by Botha et al. (2016) would have been able to externally detect leaked Aadhaar data that was available via Google. This, therefore, increases the viability of Botha et al. (2016) as a data breach detection mechanism. Jain (2019) continues by saying that the Indian government was forced to block 5,000 government officials because unauthorized government personnel were accessing the Aadhaar data.

Jain (2019) states that estimates of illegal Aadhaar data access exceeded 100,000 prior to the closure of a vulnerability in the Aadhaar system that allowed an anonymous group on WhatsApp to sell the Aadhaar data of specific individuals. Careless mistakes on the part of the Indian government were similar to the poor cyber security practices of Experian in the weeks following

their data breach as described by Wang and Johnson (2018). According to Jain (2019), "a Jharkhand state website accidentally released the data of 1.6 million pension beneficiaries, including their addresses and bank account details," and about 130 million Aadhaar numbers and affiliated private data were also inadvertently made public. These incidents point to the role of negligence and human error in data leaks. Jain (2019) further states that the Aadhaar card is vulnerable to duplication and falsification when used as a photo-ID as it lacks traditional photo-ID security features like microchips, holograms, or official seals. After RS Sharma tweeted his Aadhaar number, people found his private data, created a fake Aadhaar card in his name, and opened ad services in his name. According to the author, "most private and public entities now ask for photocopies of Aadhaar as valid identity proofs which are then stored on unprotected networks, worsening the potential for abuse of this information." This indicates that the involvement of third-party organizations increases the likelihood of data leaks.

The article presented by Isaak and Hanna (2018) on user data privacy primarily focuses on the Facebook data breach, specifically referred to as the Cambridge Analytica scandal. In "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection," the authors additionally advocated for increased privacy policy laws. Isaak and Hanna (2018) report that Facebook gave the data firm Cambridge Analytica access to the personally identifiable information (PII) of over 87 million users. Isaak and Hanna (2018) wrote that in 2013, researchers from the University of Cambridge's Psychometrics Centre "analyzed the results of volunteers who took a personality test on Facebook" evaluating their psychological profile "and correlated it with their Facebook activity." Global Science Research, in cooperation with Cambridge Analytica, then undertook a related research project using a personality quiz on Amazon's Mechanical Turk platform and Qualtrics. The quiz required participants to grant Global Science Research access to their Facebook profiles.

Isaak and Hanna (2018) stated that, via the Facebook Open API (until May 2015), Global Science Research additionally had access to Facebook users' friends' data. According to the authors, Cambridge Analytica used this and other data to develop its ability "to 'micro-target' individual consumers or voters with messages most likely to influence their behavior." This demonstrates that data leaks can not only be used to support dangerous phishing schemes as mentioned by Botha

et al. (2016) but can also be used to support alarming social engineering operations. Isaak and Hanna (2018) further stated that "every website with the Facebook logo" is connected to Facebook, indicating that individuals who may not even consider themselves to be Facebook users could have had their data exposed through Facebook. The authors proceed to advocate for data protection legislation that includes public transparency, disclosure for users, control, and notification. Although this article focused primarily on legislative rather than technical solutions to data breaches, it provided helpful insights on the Facebook data breach as a case study as well as the social engineering risks associated with data breaches. In conclusion, the most efficient approach to shortening the BDG was identified by pinpointing the factors that contribute the most to expanding the BDG, identifying the possible solutions to shorten the BDG, and analyzing case studies concerning the BDG.

III. Methods

Introduction

In this section of our paper, we present the methods used to conduct our research into the breach detection gap (BDG), associated causes, and possible solutions. We include here an exploration of the application of content analysis of several case studies as well as the research techniques utilized throughout our research.

Research Design

We employed secondary data collection as well as both qualitative and quantitative methods in our research with the overarching goal of proposing an intrusion detection framework or model for organizations to adopt into their cybersecurity infrastructure. Our qualitative analysis focused primarily on identifying recurring breach detection gap factors and viable BDG reduction solutions from secondary data, i.e. previous research. We then used a deductive method to draw hypotheses based on these recurring BDG factors and to test their validity. Our quantitative analysis focused on the validity testing of these potential factors as well as selecting corresponding solutions.

The qualitative and quantitative aspects of our research process were joined via our case study analysis, through which we sought to gain data allowing us to quantify the occurrence of qualitative factors in breach detection time, another quantifiable metric. Therefore, our research provides both statistical data and organizational cybersecurity recommendations.

a. Identification of Problem

Since its creation in 1983, the internet has presented its exponentially increasing users with a multitude of both opportunities and challenges in terms of cybersecurity, user privacy, and data protection. At present, many users are being critically impacted by cyber threats. However, the period between a data breach's occurrence and its discovery, known as the Breach Detection Gap (BDG), can span months, damaging user trust in online entities and leaving users' data vulnerable to malicious actors, who can then conduct identity theft,

phishing, and extortion. Many factors in the current cyber environment have contributed to widening the BDG. In the interest of limiting the damage caused by intrusions, properly identifying these factors - as well as finding technological solutions that efficiently target the most significant of these factors - is essential.

b. Definition of Problem

The breach detection gap (BDG) is a specific problem facing companies engaged in any degree of Personally Identifiable Information (PII) data management or retention. Several prominent data breaches featured months-long BDGs, indicating a need for BDG reduction (Johnson and Wang, 2018). BDG reduction is defined as the process of limiting the time between a breach's occurrence (via malicious actors, incompetence, technological failure, etc.) and its detection (via intrusion detection systems, third-party discovery, external PII detection systems, etc.). There are three primary factors we have targeted as potential application areas for BDG reduction methods: technology, corporate, and government policy, and human error.

i. Technology is a broad factor group encompassing intrusion detection systems (IDS), cyber hygiene practices, application performance, system stability, and a range of other characteristics of cyberinfrastructure. Technological failures include system glitches such as "application failures, inadvertent data dumps, [and] logic errors in data transfer" (Diedrich, 2019).

ii. Corporate and government policies are the communication strategies, requirements, management structures, and response frameworks established by companies or organizations and state or federal governments, respectively. They are the prescriptive or proscriptive means by which security incidents such as data breaches are addressed.

iii. The human error in this context refers to the unintended negative effects of an action or inaction taken by an individual or individuals. In the context of our research, this individual or these individuals are those who have some control over or impact on the security or management of PII.

c. Selection of Solution

In this element of the research model, we examine various frameworks for addressing BDG reduction. We critically evaluate each and select appropriate solutions that best suit the issue of BDG reduction. Some of the solutions examined are IDS machine learning (ML) algorithms, external detection systems, and patch management programs.

d. Solution Architecture

In this element of the research model, we examine the formulation of solutions into a cohesive whole. The result is the assembly of solutions into a framework or model to be applied as a cybersecurity solution for organizations involved in the handling of PII.

e. Solution Analysis

In this element of the research model, we analyze the potential solutions in terms of their efficacy, efficiency, and correspondence to the most significant BDG expansion factors. This process allows us to select the most appropriate solution based on the problem requirements.

f. Solution and Recommendation

In this element of the research model, we propose recommendations and specific solutions following the analysis and evaluation of the data collected. The data is analyzed to determine the correlation between each identified factor (technology, corporate and government policy, and human error) and breach detection time. If there is no correlation between one or more of the identified factors, this factor or these factors are removed as valid and viable application areas for BDG reduction solutions.

g. Review of Problem and Solution

In this element of the research model, we review the problem addressed in our solution framework to ensure that it is that of the current need for BDG reduction. The limitations of this research are also addressed and reviewed for further opportunities for improvement. Recommendations for these improvements are proposed.

Data Collection Method

The research instrument used in data collection was a content analysis of several case studies. The following table shows a sample of other research studies in which similar methods were used for data collection by researchers examining data security. The table below illustrates that similar studies have used case studies to support their conclusions and that the proposed research model, therefore, aligns with those of other researchers. Our research and its associated method will be further validated via hypothesis testing (see Figure 1).

a. Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, *51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268	Case study of Facebook data breach/Cambridge Analytica scandal
b. Wang, P., & Johnson, C. (2018). Cybersecurity Incident Handling: A Case Study of the Equifax Data Breach. Issues in Information Systems, 19(3), 150-159.	Case study of Equifax data breach
c. Wang, P., & Park, S.-A. (2017). Communication in Cybersecurity: A Public Communication Model for Business Data Breach Incident Handling. Issues in Information Systems, 136-147.	Case study of the Yahoo! data breaches

Table 1. shows methods applied in similar studies.

Sampling Technique Adopted

This study involved an investigation into breach detection time and affiliated technologies and policies. The organizations chosen for case study analysis were selected based on the timing of their respective data breaches. We selected organizations that experienced recent data breaches for two reasons. First, we wanted to ensure that available sources remained, at present, as familiar as possible with the state of technology, policy, and human error prior to the data breach. Second, we wanted to ensure that our study pinpointed the most modern BDG reduction problems.

We identified ten organizations that had experienced significant (compared to others that took place in their same year) data breaches within the past three years (from the time of this research). These organizations were Service Employees International Union: Local 32BJ, South Shore Hospital Corporation, Logan Health Medical Center, Ethos Technologies, SolarWinds, Syniverse, Power Apps from Microsoft, Amazon Vendor Central, MGM hotels, and Marriott International (Komnenic, 2022). We then commenced a case study analysis of these ten organizations as well as the three case studies discussed in our literature review.

Data Process and Analysis

To conduct our content analysis of the selected case studies, we first gathered several news articles, press releases, and journal publications related to each of the selected case studies. We then determined breach detection time (or approximate breach detection time, depending on the organization) based on available timelines of each breach incident. Case studies for which even approximate data was not available were removed from the data pool due to concerns that they would impede the accuracy of our results.

We then reviewed information detailing the events of each data breach to determine the presence of flawed technology, flawed policy, and human error in each breach incident. Only those factors that potentially contributed to extended breach detection time were considered. Other flaws, such as flawed intrusion prevention systems, were excluded from our analysis as we intended to focus exclusively on breach detection.

These factors were then graphically examined in several forms. Factor frequency was determined by comparing a summation of the total instances of a particular factor throughout the selected case studies to a summation of all factors present throughout the selected case studies. The case studies were then divided into groups by breach detection time. These three groups were “less than two months,” “two months to one year,” and “greater than one year.” The frequency of each factor within each of these breach detection time groups was then recorded. We additionally plotted the number of present factors against breach detection time for each particular data breach incident.

Research Model and Hypothesis

Based on the identified potential factors, we generated hypotheses regarding the relationship between the dependent variable (i.e. breach detection time) and the independent variables (i.e. various technological, political, and human factors). These hypotheses are listed below and represented in Figure 2.

i. H1a: An association exists between technology and the breach detection gap.
   H1b: There is a negative correlation between correct technology application and breach detection time.
ii. H2a: An association exists between corporate and government policies and the breach detection gap.
   H2b: There is a negative correlation between effective policies and breach detection time.
iii. H3a: An association exists between human error and the breach detection gap.
   H3b: There is a positive correlation between human error and breach detection time.
iv. H4a: The external environment influences breach detection time.
   H4b: The external influences always act to delay breach detection.

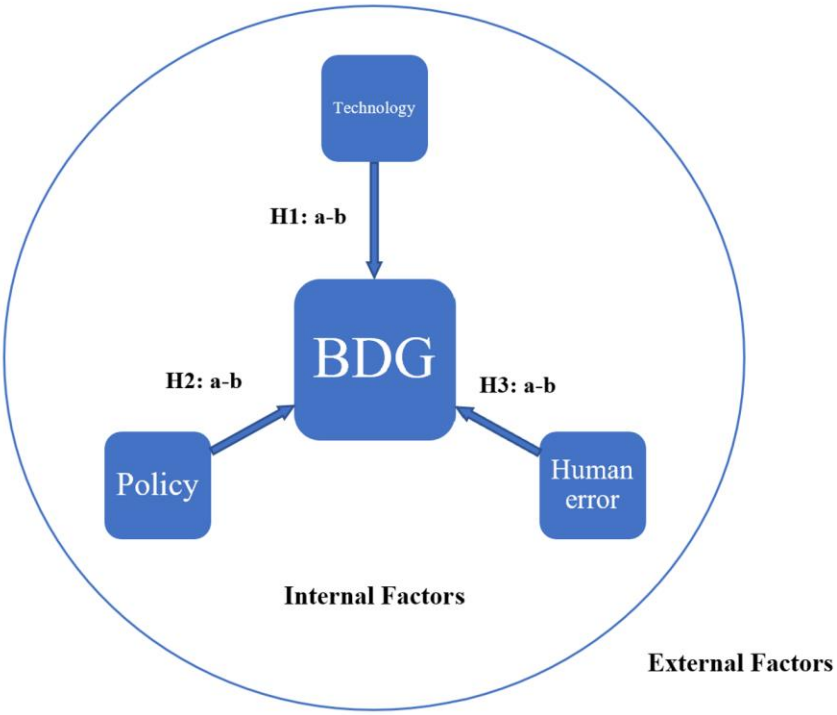

Figure 1 depicts the link between BDG factors and hypotheses.

Data Reliability and Validity

The specific previously stated policies (i.e. case study selection and factor identification) were adopted to ensure the reasonable validity of collected data. Additionally, data from each case study was checked against several news sources, incident reports, and company publications on the data breach incident to ensure consistency and validity.

Data Validity: Questionnaire

We additionally intended to validate our data through the use of a questionnaire. Unfortunately, due to sample interest and time constraints, this was not possible to accomplish within the span of our research. However, we would like to include further information about this questionnaire, as well as recommend its usage in further studies.

The questionnaire included both open- and close-ended questions. The questionnaire is to be distributed to individuals working in the Information Technology (IT) departments of companies that have experienced a data breach. The questionnaire included a series of questions presented to respondents for the purpose of collecting data on BDG factors.

a. The questionnaire included two open-ended questions. This type of question enables a respondent to provide an answer that does not originate from a series of options but rather is composed of the respondent's own words.
b. The questionnaire was comprised primarily of close-ended questions. This type of question requires a respondent to select an answer from a series of predefined options.
c. The majority of the close-ended questions in the distributed questionnaires were based on a Likert scale, which enabled respondents to select options on a scale from "Disagree" to "Agree". Each scale option corresponded to a numeric value, ranging from one to five.

The following table shows a sample of other research studies in which similar methods were used for data collection by researchers examining data security. The table below illustrates that similar studies have used and proposed scales and surveys and that the proposed research model, therefore, aligns with those of other researchers (see Table 2).

d. Ponemon Institute. (2019). *Cost of a Data Breach Report.* IBM Security.	Number line scale and company selection
e. Wang, P., & Park, S.-A. (2017). Communication in Cybersecurity: A Public Communication Model for Business Data Breach Incident Handling. Issues in Information Systems, 136-147.	Survey of corporate organizations (proposed)
f. Ponemon Institute. (2012). *The Human Factor in Data Protection.* Trend Micro.	Questionnaire with Likert scale

Table 2. depicts the methods applied in similar studies.

The questionnaire was composed of four sections: general information, questions concerning technology, questions concerning policy, and questions concerning human error. The sections concerning technology, policy, and human error included subsections which separated positive statements from negative statements and statements involving negative characteristics. The questionnaire was clear and included only those technical terms which we believed would be easily understood by an individual working in an IT department. Essentially, we only included technical jargon related to technology. The majority of questions were presented on a Likert scale, and respondents were given the option to select "Agree," "Mostly agree," "Neither agree nor disagree," "Mostly disagree," or "Disagree." We included two open-ended questions in the general information section, the first of which asked respondents to quantify breach detection time for their organization, and the second of which asked respondents to identify what they believed to be the primary BDG cause or causes. The questionnaire is to be delivered in digital form to respondents by their employer organizations. We request that respondents be given reasonable time to complete the questionnaire. We used the Microsoft Forms platform to create the survey, which lends it a readable format. These steps were taken to ensure that respondents would not encounter any difficulty in completing the questionnaire.

Multiple linear regressions can then be plotted for each factor group (technology, policy, and human error) against breach detection time. Scores determined by the cumulative Likert scale values should be assigned to each factor group, and breach detection time should be measured in days. The outcome of each regression then depicts the correlation (or lack thereof) between each factor group and breach detection time.

Limitations of Research

Future research is needed to collect data via questionnaires and more clearly determine the quantitative correlation between factors and breach detection time. Additionally, further studies must be taken to more clearly analyze the relationship between external factors and breach detection time, as our research focused primarily on internal factors.

IV. Results

Introduction

In this section of our paper, we discuss the presentation and analysis of the data collected from the case studies analyzed for this research as well as the results produced by our program. This analysis sought to examine the potential factors discussed in the literature review and enable us to create the solutions that would best address the issue of breach detection.

Data Collection Sources

The following comprises a list of sources from which data was collected for the purpose of this research project:

a. Journals and articles written on topics related to data breaches and breach detection
b. Online lists of breached organizations
c. News articles related to data breach incidents at the selected organizations

Data Collection Process

The purpose of this data collection was to understand the underlying factors in breach detection time. Understanding these underlying factors would then provide us with insight into the core causes of the breach detection gap. Journals and articles were used to further comprehend possible factors and their associated solutions. Several case studies were selected to examine the correlation between potential factors and breach detection time.

The examination of factors and selection of solutions allowed us to formulate a system of recommendations for organizations interested in data breach detection. The data collection employed for this purpose consisted of content analysis of several case studies, which confirmed that technology, policy, and human error all impact breach detection time.

Content Analysis

The analysis tool used to understand the relationship between the factors shown in the literature review and the BDG was content analysis. The outcome of the analysis reveals the correlation between each of the three factor groups (technology, policy, and human error) and breach detection time.

Critical Analysis of Breached Companies' Detection Time Factors

The three primary factors selected as potential application areas for BDG reduction methods were technology, human error, and corporate and government policy. We selected these as likely BDG factors based on their repeated occurrence in several previous studies as breach causation factors and their possible impact on detection time. Unlike factors like vulnerability to phishing attacks, which only impact whether or not a breach occurs, we believed that these factors could also affect the progression of a data breach, i.e. the BDG.

Using content analysis of several case studies, we analyzed news reports, press releases from breached companies, and academic articles that featured data breach events to test for the presence of each identified factor in previous data breaches.

Breach Incident	**Present Factors**	**BDG**
Logan Health	**None**, as the hospital detected suspicious activity within only four days (Kordenbrock, 2022).	**4 days** (November 18, 2021 – November 22, 2021) (Kordenbrock, 2022)
Service Employees International Union, Local 32BJ	**Technological flaws**, as security tools required upgrades. (Young, 2022)	**11 days** (October 21, 2021 – November 1, 2021) (Young, 2022)
Marriott International	**Technological flaws,** as	**~45 days** (mid Jan – late Feb

	suspicious logins were not immediately detected (DPM, 2022). **Flaws in organizational policy,** as adequate cybersecurity policies were not adopted following an earlier breach of the same company (DPM, 2022).	2020) (Lyles, 2020)
Equifax breach	**Human error**, evidenced by the fact that the internal patch request was ignored by the individual(s) responsible for conducting patching (Johson and Wang, 2018). **Flaws in organizational policy**, as no policy safeties were in place to double-check vulnerability closure (Johson and Wang, 2018). **Technological flaws**, as the company's IDS took over two months to detect the breach (Johson and Wang, 2018).	**~75 days** (Mid-May through July 2017) (Scipioni, 2017)
Ethos	**Technological flaws**, as systems in place to detect	**182 days** (July 15, 2021 – January 12, 2022) ("Data

	attacks against Ethos's Online Flow failed (Demas).	Breach Alert: Ethos Technologies, Inc.", 2022)
Facebook breach	**Flaws in government policy,** as a lack of adequate privacy policy laws requiring public transparency allowed the breach to continue unnoticed for some time (Isaak and Hanna, 2018). **Technological and Organizational policy flaws,** as Facebook's protocols allowed surveyors to access respondents' friends data continuously (Isaak and Hanna, 2018).	**~365 days** (2013-2014) (Meredith, 2018)
SolarWinds	**Technological flaws,** as antivirus software failed to detect the malware (Oladimeji and Kerner, 2022). **Flaws in organizational policy**, as the infected software was distributed to SolarWinds clients (Oladimeji and Kerner, 2022). **Flaws in government policy,**	**~700 days** (January 2019 – December 11, 2020) (Panettieri, 2021)

	as the federal government was also unable to detect the hack (Oladimeji and Kerner, 2022).	
Syniverse	**Human error,** as an employee described it as the result of laziness (Alleven, 2021). **Technological flaws,** as cybersecurity structures allowed the breach to go undetected for five years (Alleven, 2021). **Flaws in government policy**, as an official stated that the FCC needed to set mandatory cybersecurity standards (Alleven, 2021).	**~1,825 days** (May 2016 – May 2021) (Alleven, 2021)
Aadhaar breach	**Human error**, as government officials inadvertently made private Aadhaar data publicly available for periods of time (Jain, 2019). **Flaws in organizational and government policy**, as the Aadhaar card technology was applied nationally despite its cybersecurity flaws (Jain,	Included several incidents, most notably one of **~1825 days** (2014-2019) (Whittaker, 2019)

	2019). **Technological flaws**, as an unclosed vulnerability allowed 100,000 illegal Aadhaar data accesses (Jain, 2019).	

Table 3. lists the selected case studies* alongside the present factors and breach detection time.

**South Shore Hospital, Power Apps from Microsoft, Amazon Vendor Central, and MGM Resorts did not have adequate information available publicly regarding their data breach incidents.*

Graphic Representations of Results

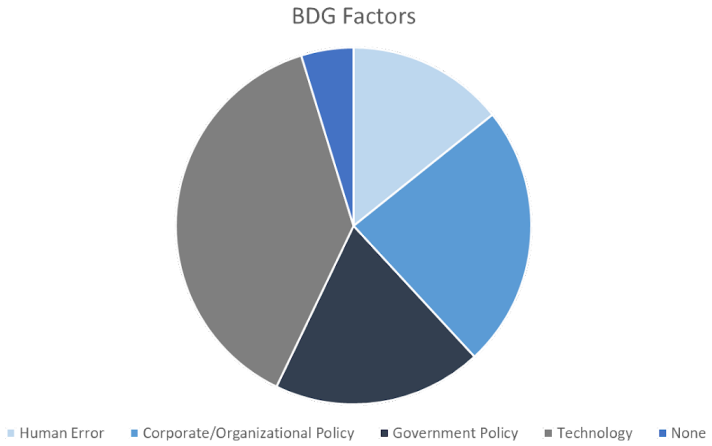

Fig. 2. shows the breakdown of factor occurrence.

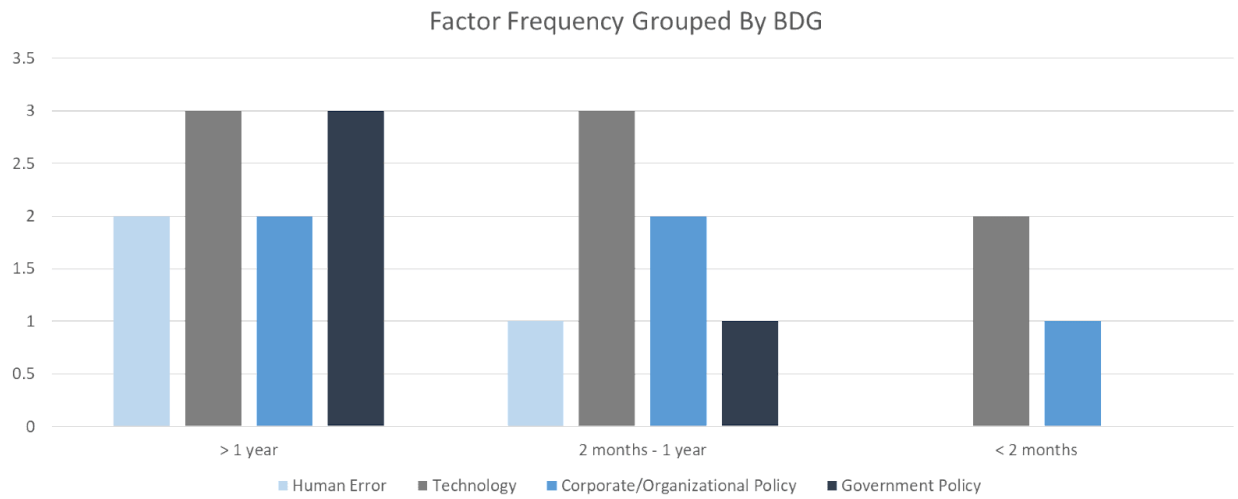

Fig. 3. illustrates factor frequency.

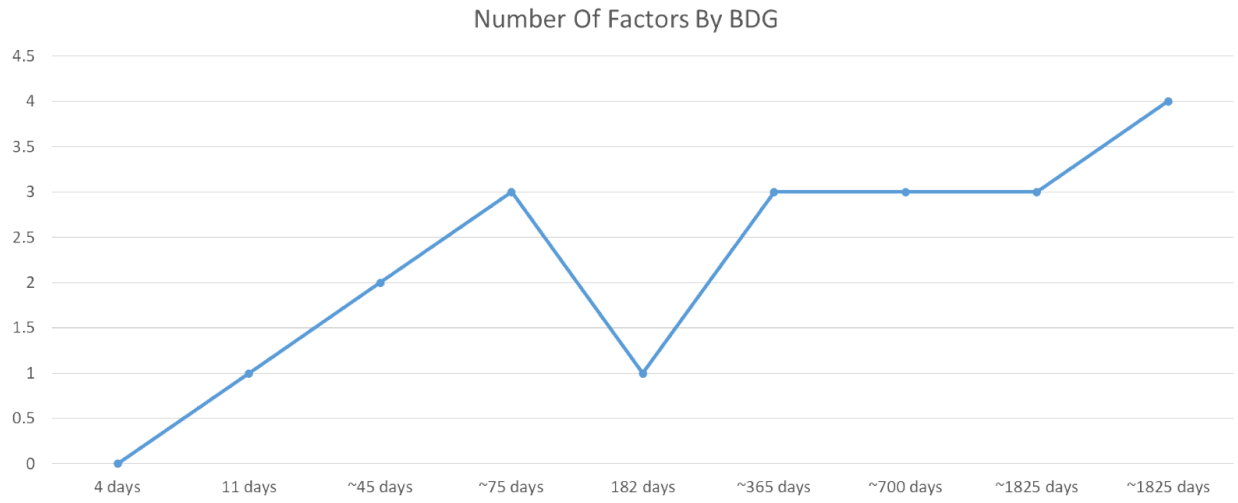

Fig. 4. shows the relationship between BDG times and number of factors.

Hypothesis Testing

The hypotheses of this study were developed to encompass potential factors discussed in the literature review as linked to or associated with the breach detection time, such as technology, policy, and human error (see Table 4).

Fig. 1 demonstrates the relationship of the hypotheses and factors to the breach detection time/gap. Using content analysis of several case studies, the hypotheses involving the association between each factor were verified (excluding external factors) and the breach detection gap (see Table 5).

Hypothesis #1	H1a: An association exists between technology and the breach detection gap.
	H1b: There is a negative correlation between correct technology application and breach detection time.
Hypothesis #2	H2a: An association exists between corporate and government policies and the breach detection gap.
	H2b: There is a negative correlation between effective policies and breach detection time.
Hypothesis #3	H3a: An association exists between human error and the breach detection gap.
	H3b: There is a positive correlation between human error and breach detection time.
Hypothesis #4	H4a: The external environment influences breach detection time.
	H4b: The internal influence always acts to delay breach detection.

Table 4. describes our hypotheses.

Hypothesis #1	H1a: Verified via content analysis
	H1b: Requires further analysis
Hypothesis #2	H2a: Verified via content analysis

	H2b: Requires further analysis
Hypothesis #3	H3a: Verified via content analysis
	H3b: Requires further analysis
Hypothesis #4	H4a: Requires further analysis
	H4b: Requires further analysis

Table 5. describes our hypothesis verification.

Results of Program Development

The program created in this study was developed as a multi-tool framework that addresses these factors and the occurrence of a fault in a network system due to a breach. The occurrence of a fault could be catastrophic in certain systems and conditions, and it can be very expensive to fix the fault. Sometimes it could be almost impossible to manually rectify the fault, and single faults are much more likely to occur compared to multiple failures. Containing single failures is more important these days because of improved system reliability. To make sure that non-faulty processes stay mostly unaffected by such local faults and thereby allow for a quicker detection and stabilization, our program permits only a small section of the network around the faulty node to make state changes (see Fig. 5 and Fig. 6).

Through extensive experiments, the program has demonstrated the ability to reduce the breach detection time with high accuracy; thus eliminating errors resulting from false positives. In Table 6, the average stabilization time as the node number increases is depicted, and it seems to depict a linear correlation between stabilization time and average node number. The program addresses the link between technology and the BDG by producing an applicable fault containment, which is a critical feature of stabilizing breached systems, that allows organizations to be able to detect corrupted nodes under a shorter time period and to stabilize the network under a shorter period of time.

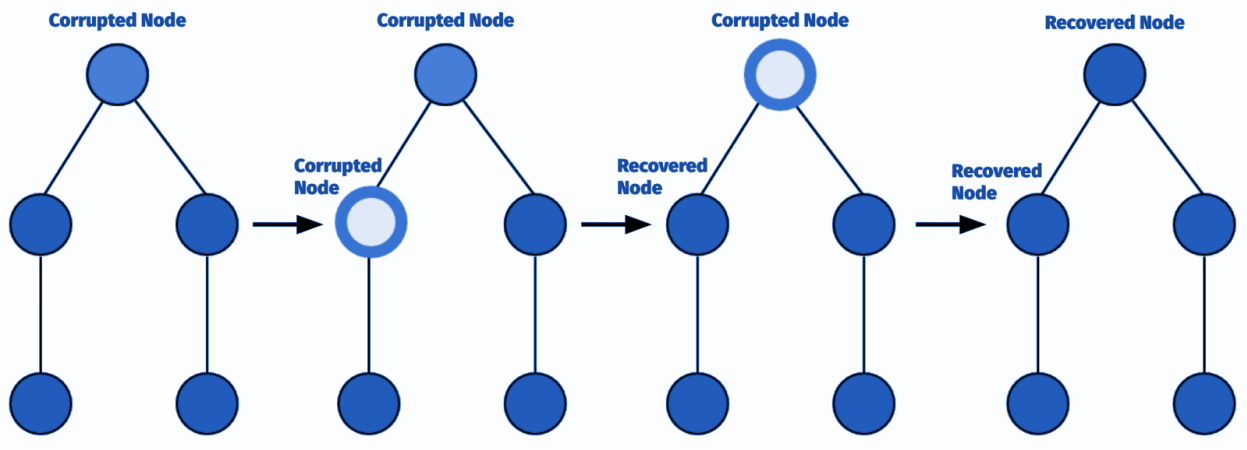

Fig. 5. demonstrates the program concept.

Algorithm for BDG

Attack structure classification algorithm

Input: Detection and reporting attack structures

Output: Malware Attack category

1. if FGd3mmd=FGHmmd=FGlocalgen=MJd3mmd=ABSmmd=MJlocalgen=no then

2. malware ← Node-1

3. else

4. if delShdCpy=ovrFile=no then

malware ← Node-2

6. else

7. if FGd3mmd = FGHmmd = MJlocalgen=no then

8. malware ← Node-5

9. else

10. if FGd3sym = FGHmmdsym = MJlocalgensym = yes then

11. malware ← Node-3

12. else

13. malware ← Node-4

14. end if

15. end if

16. end if

17. end if=0

```java
public class Graph {

    public static void main(String[] args) {
        ArrayList<Integer> x_values = new ArrayList<>();
        int size = 21;
        for (int i = 0; i < size; i++) {
            x_values.add(i);
        }
        Collections.shuffle(x_values);
        ArrayList<Node> nodes = new ArrayList<>();
        // generate 20 nodes
        for (int i = 0; i < size; i++) {
            nodes.add(new Node(x_values.get(i)));
        }
        // link them randomly
        linking(nodes);

        long start = System.currentTimeMillis();
        String firstTime = java.time.LocalTime.now().toString();

        Node finalNode = processing(nodes);

        System.out.println("Time start: " + firstTime);
        System.out.println("Time end: " + java.time.LocalTime.now());
        System.out.println("Total time taken for execution: "
                + (System.currentTimeMillis() - start) + " milli second");

        System.out.println("node " + finalNode.getNumber() + " was selected.");
        System.out.println(finalNode.connection());

    }

```

```
Node:  ->  { 14, 1 } -> 9 { true }
X-value: 248, 238 234
Node:  ->  { 2 } -> 10 { true }
X-value: 244 227
Node:  ->  { 5 } -> 11 { true }
X-value: 214 213
Node:  ->  { 3, 13, 4, 16, 6 } -> 12 { true }
X-value: 233, 241, 212, 236, 227 237
Node:  ->  { 11, 17, 1 } -> 13 { true }
X-value: 213, 221, 238 241
Node:  ->  { 8, 15, 2 } -> 14 { true }
X-value: 242, 240, 244 248
Node:  ->  { 7, 16, 9 } -> 15 { true }
X-value: 247, 236, 234 240
Node:  ->  { 9 } -> 16 { true }
X-value: 234 236
Node:  ->  { 12 } -> 17 { true }
X-value: 237 221
Node:  ->  { 4, 0 } -> 18 { true }
X-value: 212, 227 230
Node:  ->  { 5 } -> 19 { true }
X-value: 214 218
Node:  ->  { 4, 14 } -> 20 { true }
X-value: 212, 248 237

Time start: 02:27:43.568833
Time end: 02:27:46.475394
Total time taken for execution: 2916 milli second
node 8 was selected.
node 8 { 18, 14, 4, 17 } --> true
the new x value for node 8: 230, 248, 212, 221 252
```

Fig. 6. shows a sample of the code developed to address technological BDG factors.

Node Number	20	40	60	80	100	120	140	160	180	200
	26579	55572	72356	110259	132021	153226	210065	221569	243962	277231
	4207	41235	55465	63569	124436	110363	196362	206654	223651	256213
	3477	22578	52365	72569	105569	123554	153265	215565	241623	263316
	8791	32152	12258	82246	113659	151235	206543	213326	235663	271621
	9274	38426	36542	100698	98625	136854	192236	196354	243316	243165
	80	42315	39569	95562	105378	110369	206321	205698	233656	269563
	853	21589	71125	73256	121369	146559	194563	220364	237128	275363
	819	37856	46589	83465	115639	142396	186336	199856	231236	265436
	17971	51582	44569	103356	126963	152646	201136	200656	229633	256312
	29	26539	25349	756233	110639	134886	204563	211566	240361	273165
Total	72080	369844	456187	1541213	1154298	1362088	1951390	2091608	2360229	2651385
Average	7208	36984.4	45618.7	154121.3	115429.8	136208.8	195139	209160.8	236022.9	265138.5

Table 6. depicts the average stabilization time as the node number increases up to node number 200.

Our second coding element assists in intrusion detection by identifying open ports, i.e. detecting intrusion points. The program begins by generating an empty list that will be filled with any open ports that it loops over. It then loops over all the ports, for which there can be a maximum of 65535. For each port, a connection must be made, allowing the user to observe if the port will close or not. If the port is open, it is added to the list that was generated. The program then prints the list of all open ports, a piece of data that can be used by an organization's breach detection analysts to identify possible previous points of intrusion.

Next, the program takes the list of open ports and loops through them to check if they have closed between the time of the first observation and the current observation period. If a port is still open, the program will force it to close, so that any potential vulnerabilities cannot be exploited easily, and the port is then removed from the list. If a port is closed, then it is removed from the list without any further use.

```
public class PortScanner {
   public static void main(String []args) {
       long startTime = System.currentTimeMillis();
       ArrayList<Integer> openPorts = new ArrayList<>();
       for (int port = 1; port <= 65535; port++) {
         try {
            Socket socket = new Socket();
            socket.connect(new InetSocketAddress("localhost", port), 1000);
            openPorts.add(port);
            socket.close();
            System.out.println("Port " + port + " is open");
            }
          catch (Exception ex) {
         }
       }
       long endTime = System.currentTimeMillis();
       System.out.println(System.lineSeparator() + "Time to scan all possible locations: "
                          + (endTime - startTime) + " milliseconds");
       long startTime2 = System.currentTimeMillis();
       for(int i = 0; i < openPorts.size(); i++) {
           Socket s = new Socket();
           if(s.isClosed() == false) {
               try {
                 s.close();
                 openPorts.remove(i);
                 }
               catch (IOException e) {
                 e.printStackTrace();
               }
           }
           else {
               openPorts.remove(i);
           }
       }
       long endTime2 = System.currentTimeMillis();
       System.out.println("Time to close all previous open ports: " +
                          (endTime2 - startTime2) + " milliseconds");
   }
}
```

Fig. 7. shows a sample of the code developed to further aid in breach detection.

V. Conclusions

Introduction

In our literature review, we identified three potential BDG factors. Using content analysis of several case studies, we verified our hypotheses involving the association between each factor (excluding external factors) and the BDG. We then developed a program that addressed the most significant of these factors and therefore aims to reduce the BDG. The program developed by our group reduces breach detection time and is highly accurate, eliminating errors resulting from false positives.

Observations

From our data, we noted that technology occurred the most frequently among the examined factors in our selected case studies (Fig. 2). We, therefore, focused primarily on solving technological flaws that lead to increased breach detection times. This aim led to the development of our breach detection algorithm and code (Fig. 6).

We additionally observed what we believe to be a positive correlation between breach detection time and the number of present factors, which can be seen in Fig. 3 and Fig. 4, although we recognize that a full analysis of this possible correlation requires further study. We additionally noted that the frequency of *each* factor tended to increase with rising breach detection time (Fig. 4). We believe that these results may signify that factors may compound, leading to longer breach detection times and that each of these factors may indeed directly correlate with longer breach detection times.

Readers may note an inconsistency in Fig. 4 at 182 days. Other breaches within this time range (2 months to 1 year) tended to show a wider variety of factors than the 182-day Ethos breach, which only involved technological flaws, according to our content analysis. We have identified two possible explanations for this. It is possible that there were other factors at play in this breach that was not disclosed to the public, causing us to unintentionally under-represent the number of factors

present. It is additionally possible, however, that flawed technology simply had a larger impact in this breach than in other breach incidents. Upon closer examination of the facts of the breach, we learned that the hackers involved in the breach took measures deliberately crafted to circumvent the specific IDS that Ethos had in place for its Online Flow (Demas). We believe that these targeted strategies could have resulted in an unusually extended breach detection time.

Recommendations

In light of this research, the authors highly recommend that organizations involved in any degree of PII data management adopt the coding standards established in this project, train and coach employees/staff in proper breach detection practices, improve their IT policies related to breach detection, perform a regular risk assessment (RRA), obtain a certified secure session layer (SSL), and properly secure their networks to avoid malicious attacks and ransomware.

Future Work

The researchers additionally intended to supplement this content analysis with primary data from breached organizations. Although ten breached organizations were selected as a sample and were sent questionnaires, we were limited by both a lack of response and time constraints. Consequently, the portion of our research focused on determining the exact correlation between these factors and breach detection time is left inconclusive, but open to future research. Future research could focus on collecting data to determine the quantitative correlation between factors and breach detection time, and organizations could contribute their individual survey data to a joint effort to understand this relationship. This survey can be used by companies to determine the presence of specific BDG risk elements (BDG factor subgroups) in their organizational structure. The survey allows companies to reformulate their policies and obviate some human error risks.

VI. References

Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2020). Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches. *Transactions on Emerging Telecommunications Technologies*. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/ett.4150

Alleven, Monica. "Syniverse Quietly Reveals 5-Year Data Breach." *Fierce Wireless*, 5 Oct. 2021, https://www.fiercewireless.com/wireless/syniverse-quietly-reveals-hackers-had-access-to-database-over-5-years.

Botha, J., Eloff, M., & Swart, I. (2016). Pro-Active Data Breach Detection: Examining Accuracy and Applicability on Personal Information Detected. ICCWS 11th International Conference on Cyber Warfare and Security, 47-55.

Chen, Q., & Aickelin, U. (2006). Anomaly Detection Using the Dempster-Shafer Method. Proceedings of the 2006 International Conference on Data Mining. Retrieved from https://arxiv.org/ftp/arxiv/papers/0803/0803.1568.pdf

Cheng, Long & Liu, Fang & Yao, Danfeng. (2017). Enterprise data breach: causes, challenges, prevention, and future directions: Enterprise data breach. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 7. e1211. 10.1002/widm.1211.

"Company That Routes Billions of Text Messages Quietly Says It Was Hacked." *VICE*, 4 Oct. 2021,https://www.vice.com/en/article/z3xpm8/company-that-routes-billions-of-text-messages-quietly-says-it-was-hacked.

Cybersecurity & Infrastructure Security Agency. (2022, May 17). *Alert (AA22-137A) Weak Security Controls and Practices Routinely Exploited for Initial Access.* Retrieved from Cybersecurity & Infrastructure Security Agency: https://www.cisa.gov/uscert/ncas/alerts/aa22-137a

Cymulate. (2022). *The 3 Approaches of Breach & Attack Simulation Technologies.* Dallas: Cymulate.

"Data Breach Alert: Ethos Technologies, Inc.." *Console & Associates Accident Injury Lawyers, P.C.*, 10 May 2022, https://www.myinjuryattorney.com/data-breach-alert-ethos-technologies-inc/.

Deidrich, D. (2019). Data Breaches. *Georgetown Law Technology Review*, 315-324.

Demas, Tiana. “Data Breach Notifications.” *Office of the Maine AG: Consumer Protection: Privacy, Identity Theft and Data Security Breaches*, https://apps.web.maine.gov/online/aeviewer/ME/40/7ca66e2b-be43-44fc-b011-fa287a38bd84.shtml.

Dolezel, D., & McLeod, A. (2019). Managing Security Risk: Modeling the Root Causes of Data Breaches. *The Health Care Manager, 38*(4), 322-330. Retrieved June 9, 2022

DPM. “Marriott Security Data Breach- What Really Happened?” *Data Privacy Manager*, 17 Mar. 2022, https://dataprivacymanager.net/new-marriott-breach-2020-what-is-going-on/.

Equifax. (2017). Equifax Releases Details on Cybersecurity Incident, Announces Personnel Changes. Atlanta: Equifax.

Hang, L., & Kim, D.-H. (2019). Design and Implementation of an Integrated IoT Blockchain Platform for Sensing Data Integrity. Sensors, 19(10), 2228. MDPIAG. Retrieved from http://dx.doi.org/10.3390/s19102228

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, *51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268

Jain, M. (2019). *The Aadhaar Card: Cybersecurity Issues with India's Biometric Experiment.* Retrieved from The Henry M. Jackson School of International Studies: https://jsis.washington.edu/news/the-aadhaar-card-cybersecurity-issues-with-indias-biometric-experiment/

Jamil F, Kahng HK, Kim S, Kim DH. (2021). Towards Secure Fitness Framework Based on IoT-Enabled Blockchain Network Integrated with Machine Learning Algorithms. Sensors (Basel), 21(5):1640. doi: 10.3390/s21051640

Kamoun, F., & Nicho, M. (2014). Human and Organizational Factors of Healthcare Data Breaches: The Swiss Cheese Model of Data Breach Causation And Prevention. *International Journal of Healthcare Information Systems and Informatics*, 42-60.

Komnenic, M. (2022, March 25). *75 Biggest Data Breaches, Hacks, and Exposures [2022 Update]*. Retrieved from Termly: https://termly.io/resources/articles/biggest-data-breaches/#biggest-data-breaches-in-2022

Kordenbrock, Mike. “Logan Health Notifies Patients of Data Breach That Affected Thousands of Montanans.” *Flathead Beacon*, 8 Mar. 2022, https://flatheadbeacon.com/2022/03/08/logan-health-notifies-patients-of-data-breach-

that-affected-thousands-of-montanans/.

Kraemer, S., Carayon, P., & Clem, J. (2009). Human and Organizational Factors in Computer and Information Security: Pathways to Vulnerabilities. *Computer Security*, 509-520.

Lane, M. S. (2017). Managing the Risks of Data Security and Privacy in the Cloud: a Shared Responsibility between the Cloud Service Provider and the Client Organization.

Lyles, Taylor. "Marriott Discloses Another Security Breach That May Impact over 5 Million Guests." *The Verge*, The Verge, 1 Apr. 2020, https://www.theverge.com/2020/4/1/21203313/marriott-database-security-breach-5-million-guests.

"Marriott Data Breach 2020: 5.2 Mn Guest Records Were Stolen: Loginradius: Loginradius Blog." *Loginradius*, https://www.loginradius.com/blog/identity/marriott-data-breach-2020/.

Meredith, Sam. "Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal." *CNBC*, CNBC, 10 Apr. 2018, https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html.

Panettieri, Joe. "Solarwinds Orion Security Breach: Cyberattack Timeline and Hacking Incident Details - Page 4 of 4 - ChannelE2E: Technology News for Msps & Channel Partners." *ChannelE2E*, 8 Oct. 2021, https://www.channele2e.com/technology/security/solarwinds-orion-breach-hacking-incident-timeline-and-updated-details/4/.

Perotto, Filipo & Verstaevel, Nicolas & Trabelsi, Imen & Vercouter, Laurent. (2020). Combining Bandits and Lexical Analysis for Document Retrieval in a Juridical Corpora.

Ponemon Institute. (2019). *Cost of a Data Breach Report.* IBM Security.

Saheed Oladimeji, Sean Michael Kerner. "Solarwinds Hack Explained: Everything You Need to Know." *WhatIs.com*, TechTarget, 29 June 2022, https://www.techtarget.com/whatis/feature/SolarWinds-hack-explained-Everything-you-need-to-know.

Santos, J. A., Inácio, P. R. M., & Silva, B. M. C. (2021). Towards the Use of Blockchain in Mobile Health Services and Applications. *Journal of Medical Systems*, *45*(2). https://doi.org/10.1007/s10916-020-01680-w

Saranya, T., Sridevi, S., Deisy, C., Chung, T. D., & Khan, M. A. (2020). Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review. *Procedia*

Computer Science, 1251-1260. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050920311121

Scipioni, Jade. "Equifax Hack: A Timeline of Events." *Fox Business*, Fox Business, 17 Oct. 2017, https://www.foxbusiness.com/features/equifax-hack-a-timeline-of-events.

Wang, P., & Johnson, C. (2018). Cybersecurity Incident Handling: A Case Study of the Equifax Data Breach. Issues in Information Systems, 19(3), 150-159.

Wang, P., & Park, S.-A. (2017). Communication in Cybersecurity: A Public Communication Model for Business Data Breach Incident Handling. Issues in Information Systems, 136-147.

Whittaker, Zack. "Indian State Government Leaks Thousands of Aadhaar Numbers." *TechCrunch*, TechCrunch, 1 Feb. 2019, https://techcrunch.com/2019/01/31/aadhaar-data-leak/.

Young, Susanne. "Service Employees International Union Local 32BJ Data Breach Notice to Consumers." *Office of the Vermont Attorney General*, 17 Feb. 2022, https://ago.vermont.gov/blog/2022/02/11/service-employees-international-union-local-32bj-data-breach-notice-to-consumers/.