Analyzing the Impact and Implications of COMB: A Comprehensive Study of 3 Billion Breached Credentials

Chloe Stejskal*, Alexander Perminov*, Aaron Lester*, Suman Bhunia*, Mohammad Salman†, Paulo A Regis‡
*Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA 45056
†University of Anbar, Anbar, Iraq
‡ Department of Computer Science, Southeastern Louisiana University–Hammond, Louisiana, USA
Email: {stejskcl, perminaa, lesterar, bhunias}@miamioh.edu, mohammed_salman@uoanbar.edu.iq, pregis@southeastern.edu

***Abstract*—In early 2021, a file containing an organized, interactive dataset of 3 billion email and password combinations was posted on a hacker forum. This dataset, known as the 'Compilation of Many Breaches' (COMB), represents the largest and most recent data breach compilation to spread on the internet, comprising data from various company breaches. This paper aims to dissect the architecture of COMB, explore its contents, and assess the security dangers it poses. In this study, COMB is analyzed in relation to other data breach compilations, uncovering concerning patterns in the lifecycle of these data files. We find that the risk posed by cybersecurity attacks, such as credential stuffing and deep learning password cracking, escalates significantly when these attacks utilize data breach compilations. Despite the absence of prior academic literature on data breach compilations, the evolution of these compilations and the hazards they create highlight the necessity for further investigation. This study emphasizes the importance of implementing defense solutions, such as proper password hygiene, at both individual and organizational levels to mitigate the damage of COMB and prevent similar future incidents.**

***Index Terms*—Data Breach Compilation, Cybersecurity Risks, Credential Stuffing, Deep Learning Password Cracking, COMB Analysis**

I. Introduction

The security event under analysis is a compilation encompassing over 3 billion email and password login credentials, released on a hacking forum in early 2021. This dataset, named the 'Compilation of Many Breaches' (COMB) by its creator, aggregates data from hundreds of individual breaches spanning the past few years. Although some users, who purchased COMB for two dollars, criticized the data as 'nothing new' and 'low quality' [1], it nonetheless garnered attention as a potential security threat. This shift in perception occurred following publications by journals such as Cybernews and ACFE Insights, which detailed specific past breaches included in COMB and encouraged readers to verify if their personal information was compromised [2].

Our research objective is to elucidate the security risks posed by this data breach compilation and propose defensive strategies to counter these risks. Achieving this necessitates a thorough analysis of COMB's structure and content. We intend to dissect its formation by scrutinizing the individual breach compilations that constitute COMB. A historical analysis of prior data breach compilations, in conjunction with an examination of COMB's relationship to these precedents, is pivotal for identifying recurring patterns in the lifecycle of breach compilations. This will enable us to assess the implications of COMB for both individual and organizational security. Subsequently, we aim to delineate defense mechanisms tailored to mitigate the specific security risks engendered by COMB.

Our research indicates that data breach compilations akin to COMB are likely to persist into the future, posing significant threats to digital security. It is imperative to actively defend against these emerging threats. Recognizing that these compilations originate from individual data breaches, we posit that reducing the frequency and severity of such breaches can curtail the proliferation of future compilations. Consequently, our study delves into the major breaches encompassed by COMB, with a focus on identifying the exploited vulnerabilities, the methodologies employed in these attacks, and the subsequent impacts of the breaches.

A data breach represents a critical incident for companies, often resulting in significant reputational damage and financial loss. The leakage of login credentials – typically email and password combinations – onto the internet further compounds these issues by endangering the security of individuals whose information is compromised. The most apparent risk associated with such breaches is the potential compromise of user accounts on the affected website. Beyond this immediate threat, however, are more insidious risks, including credential stuffing and spear phishing attacks, which can also emerge from these data leaks.

The severity of these threats is significantly amplified by the aggregation of breached data into compilations. A data breach compilation is an assemblage of login credentials from multiple breaches, where billions of credentials are consolidated. The architects of these compilations often enhance their potency by converting hashed passwords into plaintext, meticulously organizing the data, and developing scripts for efficient management and access. This phenomenon of compiling and disseminating data breach information is a relatively recent development in cybersecurity, with the first known instance

being the Exploit.in compilation of 2016 [3]. In this paper, we investigate the lifecycle and historical evolution of data breach compilations, contextualizing them within the framework of COMB.

The subject of data breach compilations remains largely unexplored within academic literature. Our research has uncovered evidence indicating that these compilations constitute a significant security hazard, meriting thorough investigation and analysis. This paper is dedicated to elucidating the security risks inherent in data breach compilations. It is crucial to recognize that these risks are distinct and operate independently from those associated with the individual breaches that comprise these compilations. Accordingly, it becomes imperative to consider data breach compilations as holistic security events in their own right.

In summary, the main contributions of this paper are:

- Analysis of the Compilation of Many Breaches (COMB), providing insights into its structure, content, and the implications for cybersecurity.
- Investigation of the historical development and lifecycle of data breach compilations, with a focus on patterns and trends leading up to COMB.
- Examination of various high-profile data breaches included in COMB, revealing common vulnerabilities and attack methodologies.
- Assessment of the impact of data breach compilations on individual and organizational security, emphasizing the amplification of risks such as credential stuffing and deep learning password cracking.
- Presentation of defensive strategies and policies to mitigate the risks posed by data breach compilations, including encryption, password management, and multi-factor authentication.
- Exploration of the role of human error and the advancement of AI in cybersecurity, offering future directions for research and development in data protection.
- Emphasis on the necessity for businesses to adopt proactive cybersecurity measures, highlighting the cost-effectiveness of prevention over reactive responses to breaches.

We commence our exploration with a historical analysis of the individual data leaks and compilations that culminated in COMB, detailed in Section II. This section also includes an overview of the attack methodologies employed in the individual breaches. Section III further elaborates on these methodologies, providing a comprehensive description of how data breach compilations are formulated. In this section, we also present our findings regarding the structure and data of COMB. Section IV delves into the impact of COMB, examining various attack vectors that could exploit this compilation maliciously. Section V is devoted to investigating defensive strategies implemented by companies post-breach and proposes measures for both corporate entities and individuals to prevent and mitigate the ramifications of data breaches. The limitations of our study, as well as avenues for future research, are discussed in SectionVI. Finally, Section VII provides a summation of our findings and addresses the ongoing challenges related to this event.

II. Background

COMB is an aggregation of hundreds of data leaks [4], each originating from an attack that inflicted harm on both the affected companies and their users. In our analysis, we will examine select incidents among these, focusing on the attack strategies employed to acquire sensitive credential information. Initially, this section will highlight some of the major companies that have suffered data leaks included in COMB. Subsequently, it will delve into the background and architecture of COMB itself, providing a comprehensive understanding of its composition and the nature of the collected data.

A. Yahoo

Yahoo, a widely-used email platform, experienced a significant breach in 2014, resulting in the exposure of credentials from all of its 3 billion user accounts. The perpetrators of the Yahoo breach were identified as four individuals with affiliations to the Russian state service. Their primary objective was to infiltrate the Yahoo accounts of targeted individuals, including U.S. government officials and Russian journalists [5]. This breach not only highlights the scale of potential data leaks but also underscores the diverse motives behind such cyberattacks.

B. Dropbox

Dropbox, a widely-utilized cloud storage service, is employed for storing a variety of information, ranging from personal photographs to sensitive files. Despite employing robust encryption algorithms and protocols to safeguard users' personal data, Dropbox was not immune to cyber threats. In 2016, a significant security breach occurred, leading to the compromise of credentials for approximately 68 million Dropbox user accounts [6]. This incident serves as a stark reminder that even platforms with stringent security measures can be vulnerable to sophisticated cyberattacks.

C. Canva

Canva, a popular graphic design platform, experienced a data breach in 2019, compromising the emails and hashed passwords of 139 million users. A few months following the breach, approximately 4 million of these passwords were decrypted and subsequently leaked online [7]. The entity responsible for both the initial breach and the decryption of the leaked passwords is identified as GnosticPlayers, an anonymous hacker group. This group has been implicated in orchestrating a series of data breaches and leaks affecting numerous companies since 2019 [8]. The Canva incident exemplifies the ongoing threats posed by organized cybercriminal groups in the digital domain.

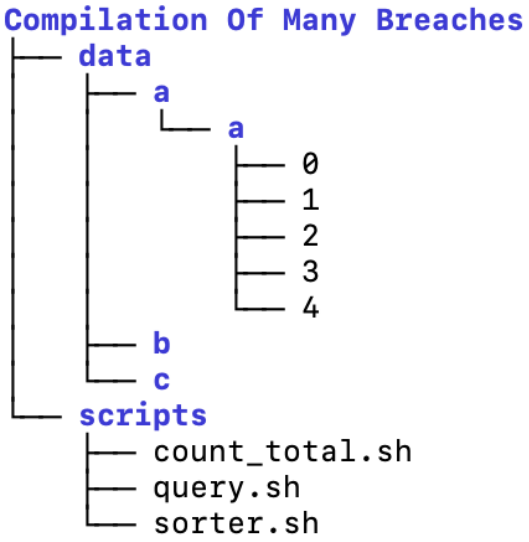

Figure 1: COMB is a directory containing the data directory and shell scripts. The data directory contains subdirectories 0-9 and a-z, each of which contain more levels of subdirectories. Within the data directory are text files containing each email and password combination. For example, the file a/a/0 contains all login credentials with an email beginning with "aa0." This alphabetical tree structure allows for quick querying of entries, using the query.sh script. The sorter.sh script manages the structure by sorting entries.

D. Netflix

Netflix, a prominent streaming service, encountered an incident in 2015 where user credentials were leaked online, contradicting the company's claims of never having been breached [9]. The credentials were compromised through an attack known as credential stuffing. This method of attack specifically targets poor security practices among individual users, highlighting that vulnerabilities can exist irrespective of the robustness of a company's system. The Netflix incident underscores the critical need for strong personal security measures in the digital era.

E. Zoom

Zoom, a widely-used video and audio communication platform, is primarily utilized for web conferencing. Its key features include the ability to interact through a chat window and communicate via video. Prior to the COVID-19 pandemic, Zoom served as an alternative communication tool, attracting comparatively less internet traffic. However, with the onset of the pandemic and the consequent surge in user base, several issues emerged, highlighting the platform's growing importance and the potential challenges associated with its expanded usage.

F. COMB

COMB was disseminated on a hacking forum by a user identified as Singularity0x01 on February 2, 2021 [10]. Distributed in the form of a ZIP file named 'Compilation of Many Breaches.7z', the unzipped directory reveals a data subdirectory accompanied by three shell scripts. The data subdirectory is further organized into subdirectories labeled from 'a-z' and '0-9', facilitating alphabetical sorting of the data, as illustrated in Figure 1. The actual data resides in text files, with each email and corresponding password separated by a colon, as depicted in Figure 2. The shell scripts provided enable querying, counting, and sorting of the data. This alphabetically structured tree arrangement of COMB's data significantly enhances the efficiency of sorting and querying processes [4].

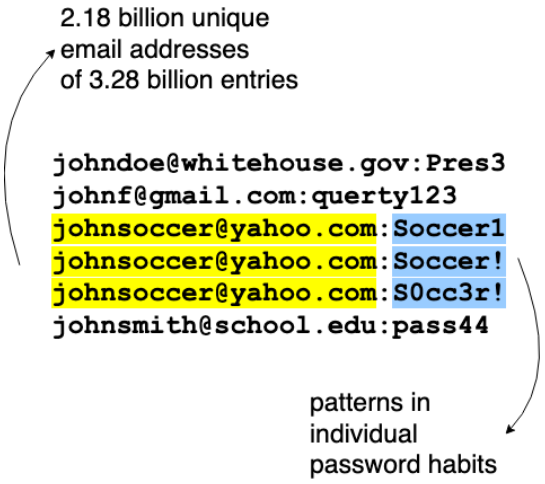

Figure 2: Analysis of COMB's data showed only 2.18 billion unique email address of 3.28 billion entries, indicating that users were included in COMB through mulitple previous breaches, with different passwords for the same email. COMB could be further analyzed to understand the password habits and patterns of these individuals [10].

G. Analyzing COMB's Data

An initial examination of COMB's data revealed a significant number of email accounts from major providers, including approximately 450 million Yahoo emails and 200 million Gmail accounts. Notably, a considerable number of .gov emails, linked to the US government, were also present. Further in-depth analysis indicated that out of the 3.28 billion records in COMB, there were only 2.18 billion unique email addresses [10]. This finding suggests that numerous users had multiple passwords associated with the same email address in the compilation. An intriguing avenue for future research within COMB is to investigate the evolution of individual password habits over time. A visual representation of the data entries and a summary of this analysis are provided in Figure 2.

III. Attack Methodology

Each company discussed in the Background section exemplifies a distinct type of cyber attack. This section will delve into the various vulnerabilities that were exploited and describe the execution of each attack. Additionally, we will elucidate

Table I: Attacks on Different Companies

Attack Type	Description of Attack	Companies Affected
Poor Password Management	This type of attack involves having a poor practice of either storing passwords in a unsafe location or using the same password on other login systems. So if your password is discovered, those other systems are now in danger.	Dropbox, Deep Root Analytics, eBay
Phishing	This type of attack involves sending an email disguised as a valid email from a trusted party. The link or attachment that the user is convinced to click/download is usually malware or a tracking software to record passwords.	Yahoo
Outdated Software	This type of attack involves exploiting a vulnerability already recorded by other malicious hackers. The vulnerability is already explained and detailed on how to further hack into the target system.	Canva, MyFitnessPal, T.J. Maxx
Exploited Bug	This type of attack involves exploiting a feature that has a bug which can lead to an elevation of privilege or an easier breaching point.	Facebook, Zoom, First American Corporation
Credential Stuffing	Botnets put leaked login credentials into many sites across the internet, accessing accounts of those who reuse passwords.	Uber

the process of creating data breach compilations by analyzing the development and structure of COMB.

A. Yahoo: Phishing

Yahoo's user information database, containing names, phone numbers, password challenge questions and answers, password recovery emails, and cryptographic values for each account, presented a centralized target for attackers [11]. A significant breach, subsequently utilized in further compilations, was the Yahoo data breach. The primary methodology employed in this attack was phishing [11]. An attacker sent a phishing email containing a malicious link to a Yahoo employee [12]. Upon clicking the link, a script was activated that transmitted the cookie and session data to the attacker. This data enabled the attacker to infiltrate the Yahoo network. Once inside, the attacker located the network management tool and the user database. Utilizing the account management tool, the attacker modified the database, installed a backdoor in the server, and extracted a copy of the database for themselves [5].

B. Dropbox: Password Reuse

In 2012, a breach of LinkedIn led to the leakage of login credentials, including those belonging to a Dropbox employee who used the same credentials for both their Dropbox and LinkedIn accounts. The attacker, exploiting this overlap, accessed the Dropbox employee's account using the compromised LinkedIn credentials. Within this account, the attacker discovered a file containing Dropbox user emails and passwords, most of which were hashed and salted. Dropbox acknowledged the compromise of their employee's account at the time of the incident in 2012. However, it was not until four years later, in 2016, that 68 million Dropbox credentials began to circulate on the dark web. Dropbox later confirmed that this wider breach was indeed a consequence of the 2012 incident. This case exemplifies two critical lapses in employee security: the reuse of passwords across multiple accounts and the storage of sensitive corporate data in a personal account [13].

C. Canva: Credential Cracking

Canva's system was compromised by an attack known as credential cracking, executed by the group GnosticPlayers [14]. In this type of attack, the attacker employs brute-force methods to access accounts, aiming specifically to infiltrate an admin account. Once gaining control of an admin account, they can then access the company's database containing user passwords. In this incident, GnosticPlayers successfully downloaded usernames, emails, and bcrypt-salted and hashed passwords from Canva's systems. The group initially posted this data online in May 2019. Subsequently, they began the process of decrypting the passwords, and by January 2020, had successfully decrypted and published 4 million of these passwords on the internet [7].

D. Netflix: Credential Stuffing

Netflix users whose credentials were leaked fell victim to an attack known as credential stuffing. In such attacks, a botnet is employed, armed with a list of potential username and password combinations, often sourced from previous data leaks or compilations. The botnet systematically brute-forces these combinations across multiple websites. This methodology enables attackers to generate data leaks for companies without directly infiltrating their systems [6]. The attackers can configure the botnet to log into various services, capture the credentials that successfully authenticate, and then either release or sell this information. In these scenarios, the vulnerability stems from poor individual password practices. Despite robust security measures at the organizational level, companies can still suffer the adverse effects of public data leaks, including damage to their reputation and financial loss. Thus, credential stuffing attacks highlight the indirect yet significant risks companies face due to inadequate personal security habits.

E. Zoom: Metalinks

Amid the COVID-19 pandemic in early 2020, Zoom witnessed a significant surge in new users. Concurrently, several security issues emerged on the platform, culminating in a data breach that resulted in over 500 million usernames and passwords being leaked, sold, and subsequently published on the dark web [15]. This breach was largely attributed to inadequate security measures within Zoom. Attackers exploited a vulnerability in the Zoom-Windows client's group chat feature, which allowed malicious links to be shared and leaked beyond the confines of the application. A major security flaw involved the conversion of Windows Universal Naming Convention (UNC) paths into clickable links within Zoom. Attackers manipulated this feature to redirect users to harmful

websites or to execute malicious software on their devices [16]. This incident underscores the critical need for robust security controls, especially in applications experiencing rapid user growth.

F. COMB

Singularity0x01 disclosed that the construction of COMB involved integrating data from the 'Collection 1-5' series and various smaller breaches into the 'Breach Compilation' of 2017 [10]. The 2017 Breach Compilation is a significant data breach compilation, comprising 1.7 billion entries organized in a tree structure and sorted alphabetically, complete with shell scripts for querying and sorting [4]. To create COMB, Singularity0x01 initially downloaded the 2017 Breach Compilation and then utilized its provided shell scripts to incorporate additional data from other breaches. This supplementary data included the extensive 'Collection 1-5', which itself contains 2.2 billion entries. A visual representation of the process involved in creating data breach compilations is depicted in Figure 3.

IV. Impact

To comprehensively grasp the ramifications of COMB, it is essential to first examine the impacts of the individual breaches discussed in Sections II and III. Such an analysis will illuminate the various challenges and repercussions faced by companies in the wake of public data leaks. In the context of data breach compilations, the most significant impact lies in the continued propagation and intensification of large-scale unauthorized access to accounts and data leaks. These risks primarily emanate from attacks leveraging techniques like credential stuffing and advanced deep learning for password cracking. Despite the original post containing COMB being removed, the compilation had already been downloaded and widely disseminated across the internet. With relative ease, one could locate and download COMB from sources like Google or GitHub. Consequently, the likelihood of an attack exploiting the data within COMB remains substantially high.

A. Reputation and Revenue Loss Due to Breaches

Over time, as attackers build a reputation for executing multiple successful data breaches, affected companies concurrently develop a public reputation of vulnerability. This evolving perception among users, adjacent businesses, and investors can erode confidence in the integrity of these companies' systems, particularly when they have been compromised repeatedly. As a result, these companies begin to experience various forms of revenue loss. Initially, the most evident impact is a significant decline in their stock prices [17]. As news of the breach disseminates and persists over time, the company's reputation as a reliable and secure entity diminishes, leading to reduced usage of their services and, consequently, further long-term revenue loss. Perhaps most critically, these companies often lose valuable data and assets, which are fundamental to their success and business operations.

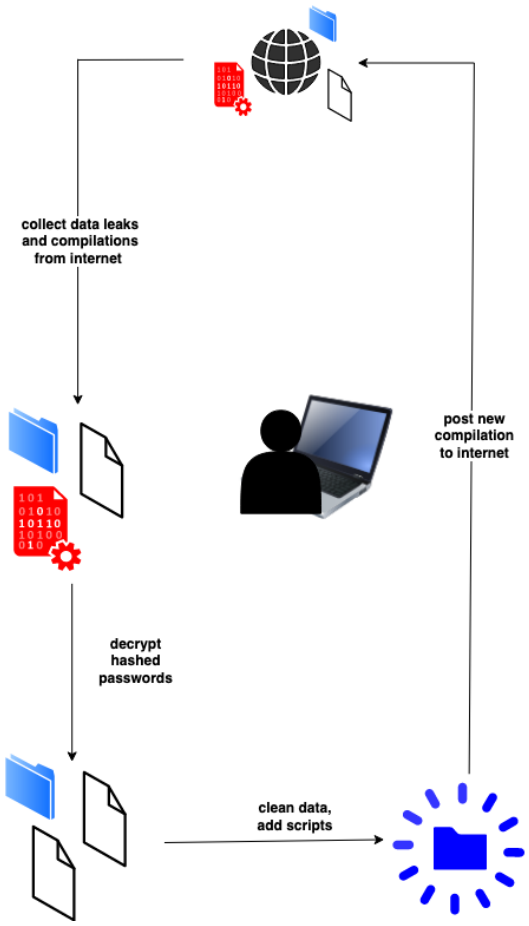

Figure 3: The creation of data breach compilations can be viewed as a cycle. An attacker first downloads existing data leaks and compilations from the internet. They may clean the data, decrypt hashed passwords, and add scripts to manage the data. They will then post the new compilation on the internet, where other attackers may download it and continue the cycle.

B. Growing Data Leak Compilations

COMB is a culmination of data from two major breach compilations: the Breach Compilation and Collection 1-5. Notably, the Breach Compilation itself is derived from earlier compilations, namely Anti Public and Exploit.in [10]. The interconnectedness of COMB with these prior compilations is depicted in Figure 4. This illustration highlights a trend of successive generations of compilations building upon and enlarging their predecessors, as further evidenced in Figure 5. The ever-increasing scale of these compilations is significant for two primary reasons. Firstly, the larger the dataset, the greater the number of users potentially impacted by attacks leveraging these compilations. Secondly, certain attack methodologies, detailed subsequently, become increasingly effective as the volume of data escalates. Consequently, this incentivizes malicious actors to utilize expansive datasets, rendering users more susceptible to successful attacks. Given that COMB represents the largest known data breach compilation to date,

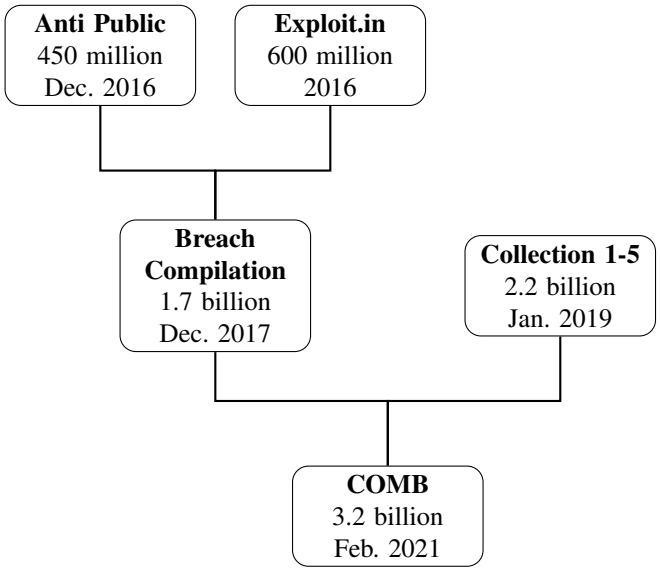

Figure 4: Data leak compilations with name, size, and date, showing COMB in 3 known levels of data breach compilation cycles. [18] [1] [19] [3].

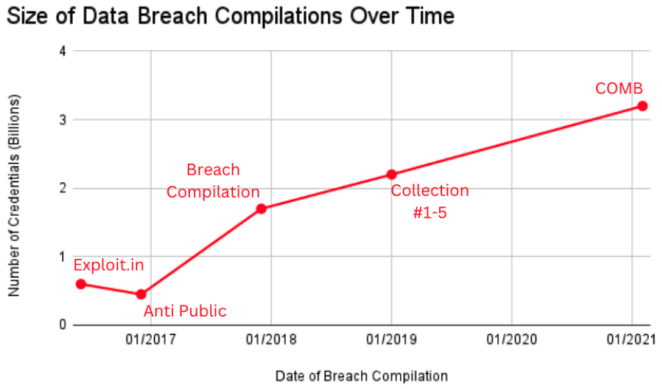

Figure 5: A trend of increasing size of data breach compilations from 2016-2021 [18] [1] [19] [3].

encompassing over 3 billion credentials [10], it stands as a particularly attractive target for attackers.

To construct a data leak compilation, as depicted in Figure 3, a threat actor first gathers various data leaks from common sharing points on the internet, such as hacker forums or dark web marketplaces. These individual datasets are then aggregated into a single compilation. Subsequent to this aggregation, the threat actor may undertake several measures to enhance the value and usability of the data. These measures include decrypting hashed passwords, eliminating obsolete and redundant data, and systematically organizing the entries. Additionally, scripts that facilitate querying and sorting of the data are often integrated at this stage. The cycle of creating breach compilations reaches its culmination when the threat actor either sells or releases the compiled dataset onto the internet, making it accessible to a broader audience.

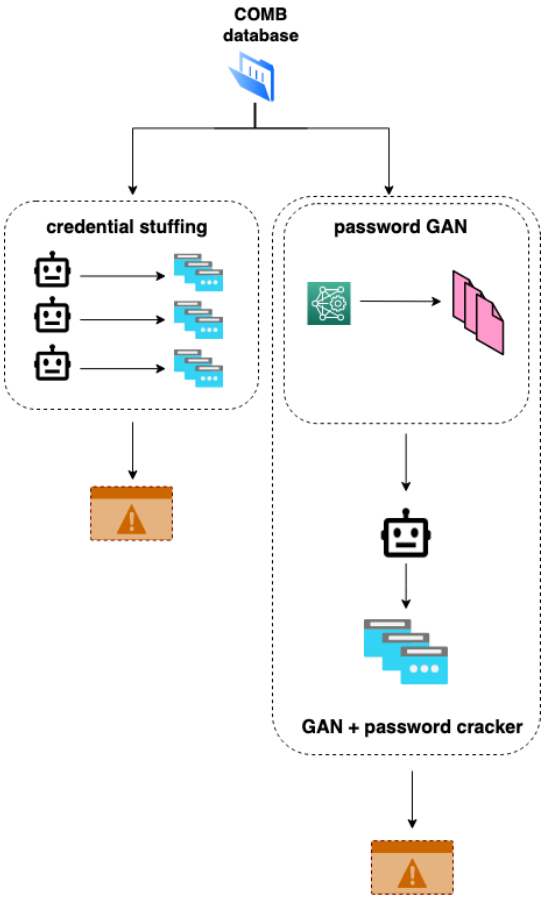

Figure 6: Two attacks, credential stuffing and password cracking, are more effective when given larger amounts of data to work with. Thus, the large size of COMB increases the risk of account compromise across the internet.

C. Credential Stuffing

To construct a data leak compilation, a threat actor first gathers various data leaks from common sharing points on the internet, such as hacker forums or dark web marketplaces. These individual datasets are then aggregated into a single compilation. Subsequent to this aggregation, the threat actor may undertake several measures to enhance the value and usability of the data. These measures include decrypting hashed passwords, eliminating obsolete and redundant data, and systematically organizing the entries. Additionally, scripts that facilitate querying and sorting of the data are often integrated at this stage. The cycle of creating breach compilations reaches its culmination when the threat actor either sells or releases the compiled dataset onto the internet, making it accessible to a broader audience.

D. Deep Learning Password Cracking

Deep learning tools, when utilized in conjunction with COMB, represent an additional cybersecurity threat. Generative adversarial networks (GANs), a type of deep learning network, are designed to create realistic data, such as passwords.

The effectiveness of these networks improves with the volume of data available for training; more data leads to more accurate generation of realistic passwords [20]. For instance, PassGAN, a GAN model specifically developed for password generation, has demonstrated remarkable efficiency. When trained on 20 million records, PassGAN was able to surpass traditional password cracking tools in both the number of passwords cracked and the speed of cracking. This efficiency was further enhanced when PassGAN was used in conjunction with these conventional tools [21]. Given COMB's extensive collection of real passwords, it can serve as an ideal training dataset for GAN models like PassGAN. Due to its sheer size, COMB could significantly improve the model's ability to generate accurate passwords. The resulting risk is twofold: firstly, the GAN-generated passwords could be employed in attacks akin to credential stuffing; secondly, as illustrated in Figure 6, they can be input into password crackers like HashCat, thereby increasing the cracker's accuracy and the likelihood of account compromises [21]. This scenario exemplifies how the aggregation of data breach information can lead to sophisticated advancements in cybersecurity attacks through the application of machine learning techniques.

V. Defense Solution

This section is dedicated to exploring various defense methodologies and strategies aimed at either preventing or mitigating the occurrence of future data breaches. Given the multitude of methods by which hackers can acquire confidential information, our coverage will span a broad spectrum of threats and vulnerabilities. For each identified potential breach method, we will propose a corresponding solution.

Each recommended solution is meticulously crafted to thwart malicious hackers from acquiring usernames and passwords, targeting both individual users and organizations. By implementing these defensive techniques, there is potential to impede the further expansion of data breach compilations like COMB and to diminish their effectiveness in attacks, such as credential stuffing. The defense methodologies and strategies are depicted in Figure 7.

A. Encryption

Implementing encryption is a critical strategy for both preventing and mitigating the impact of cyberattacks by malicious hackers. Once intruders breach a system, easily accessible information becomes a significant vulnerability. It is a misconception to rely solely on firewalls and defensive software for protection; data within the system should not be left easily accessible. Hence, enforcing robust encryption algorithms on data is vital, as it limits the extent of compromise, rather than leaving all data readily available. While encryption may not entirely prevent breaches, it can significantly reduce the volume of usernames and passwords that are compromised. By adding this additional layer of security, the number of accounts vulnerable to being incorporated into data breach compilations like COMB is lessened.

A primary step in adopting encryption is to secure laptop hard drives. Given their portability, laptops are particularly susceptible to hacking, either through physical theft or social engineering tactics. They often contain sensitive information, such as cached data, remembered passwords, and contact lists, posing a risk not only to individuals but also to the organizations they are affiliated with. The most effective approach is full-disk encryption of the laptop [22]. By encrypting the entire hard drive, unauthorized individuals are prevented from accessing any data without the requisite password or key card to decrypt it.

Another critical area for implementing encryption is on websites. A significant vulnerability arises when websites do not use encrypted data transmission protocols like SSL (Secure Sockets Layer). Without SSL, hackers can potentially intercept data transmissions and access confidential information. The solution is straightforward: acquire an SSL certificate for the website and utilize the HTTPS (Hypertext Transfer Protocol Secure) protocol for data transmission. This approach ensures that data streams are encrypted, making it extremely challenging for malicious actors to intercept and decipher the data without the corresponding decryption key, thereby preventing potential data breaches.

The third recommended encryption practice involves securing email communications. Similar to website encryption, the primary vulnerability with emails is the risk of interception from the email server, leading to the potential leak of their contents. While not all emails may contain sensitive information, many do, ranging from Private Personal Information (PPI) to crucial documents. To safeguard against data breaches, it is essential to encrypt emails. Email services like Outlook and Gmail provide options to encrypt messages. Additional features include setting a time limit for the recipient to view the information before the email self-destructs. IT departments can further reinforce security by enforcing policies that automatically apply encryption to all outgoing emails. This proactive approach ensures another potential breach point is secured, enhancing overall data protection for both individuals and organizations.

B. Proper Disposal and Archiving

Effective management of sensitive and confidential information is crucial in mitigating the impacts of a data breach. While it is challenging for companies to completely prevent breaches, there are numerous instances where hackers have exploited outdated or irrelevant company data. Such data, if not properly disposed of, can be obtained by malicious actors and subsequently sold on the Dark Web or used to breach other systems. To address this risk, the implementation of specific policies is essential. These policies can limit the accumulation of old data and account information, thereby reducing the potential for this information to be used in augmenting compilations like COMB.

A key policy is the principle of data minimization: 'keep only what you need' [23]. Retaining old and unnecessary documents poses a security risk, particularly if they contain

confidential information. By limiting data retention to only what is necessary and for a predefined duration, the volume of information available for hackers to exploit is significantly reduced. Many companies already have policies dictating where documents should be stored, but incorporating a policy that sets a timeframe for document retention – after which they are either archived or disposed of – can further decrease the accessibility of such data to malicious actors.

A second critical policy involves the thorough destruction of any information or document prior to its disposal [23]. Merely deleting files or reformatting a drive is insufficient, as remnants of data can remain in various locations within a computer system, such as in cached memory or backup file folders. To address this issue, it is advisable to employ specialized software capable of locating and erasing all instances of the targeted file. This ensures that when a document is deleted, it is done comprehensively, preventing any possibility of the data being recovered and exploited by hackers. Such a thorough approach to data destruction is essential to ensure that sensitive information does not become a resource for malicious actors seeking to sell or leverage it for further breaches.

C. Third-party Vendors Cooperation

Beyond the cyber practices of individual employees, the security protocols of third-party vendors represent another significant threat vector. Even if a company excels in training its employees and deploying state-of-the-art cyber protection measures, third-party vendors can still be a vulnerability. A breach of a vendor's credentials to access the company's systems can create an easy entry point for hackers, effectively bypassing the company's internal security policies. Vendors are often more susceptible to social engineering attacks due to their comparatively relaxed guidelines and practices. Therefore, it is imperative for companies to establish stringent security policies for their vendors to adhere to, thus reducing their potential as targets or access points for cyber attacks. A notable example of a breach involving a third-party vendor was the incident with Fazio Mechanical Services, which led to unauthorized access to Target's network and the subsequent leakage of sensitive data [24]. By minimizing vulnerabilities associated with third-party vendors, companies can significantly reduce the likelihood of breaches that contribute to the accumulation of data in compilations like COMB.

The foremost policy recommendation is to establish a comprehensive inventory of all files that a vendor requires access to [25]. By restricting the scope of access granted to vendors, a company can significantly mitigate the impact of a breach in the event of compromised vendor credentials. This approach enables companies to swiftly implement countermeasures in the aftermath of information theft. Furthermore, restricting file access ensures compliance with privacy regulations, safeguarding not only the company's data but also that of its users and employees. Implementing such access limitations serves as a proactive measure in preserving data integrity and preventing the unauthorized dissemination of sensitive information.

D. Employee Security Awareness

Before launching an attack, malicious hackers often engage in both passive and active reconnaissance to gather information about a company and identify potential entry points into its systems. A major vulnerability that employees frequently encounter is social engineering, particularly through phishing and pharming attacks. Phishing involves deceptive emails that either request confidential information or redirect employees to seemingly legitimate websites which may harbor malware or trick them into divulging their login credentials. Pharming, on the other hand, involves creating misleading links that appear to direct to a legitimate site but instead lead to a different, often malicious, website. Consequently, it is imperative for employees to undergo annual training to recognize and thwart social engineering tactics. More informed employees can significantly reduce the number of usernames and passwords compromised and, thus, limit the data available for inclusion in compilations like COMB.

Implementing a comprehensive cybersecurity training module for employees is the most effective way to ensure their preparedness. While annual training might seem repetitive, its importance cannot be overstated – a single employee falling for a phishing scam could jeopardize the entire company's system. A certification process should be established to confirm that employees not only complete the training but also understand and are informed about the latest cybersecurity practices. While such training may not completely prevent sophisticated schemes, it significantly reduces the likelihood of successful attacks in the future.

E. Update Software and Algorithms

Many hackers have access to online resources that detail specific weaknesses in various software types and their versions. These resources often include information on common web servers like Apache and database software like MySQL, outlining each potential exploit and methods to attack these vulnerabilities. Consequently, it is crucial for businesses to regularly update their software and promptly apply patches as they become available to address these known security flaws [26].

A pertinent example illustrating the importance of this practice can be seen in Yahoo's breach. At the time, Yahoo was utilizing the SHA-1 encryption algorithm instead of the more secure SHA-256. This older encryption method allowed hackers to decrypt passwords more easily, leading to their subsequent leak on the Dark Web. Had Yahoo been employing the most advanced encryption algorithm available, such as SHA-256, the breach might have been significantly more difficult to execute, thereby mitigating the extent of the damage.

F. Develop Cyber Breach Plan

Businesses must acknowledge that a security breach is a real possibility, regardless of their size or the advancements in technology. As malicious hackers evolve their techniques, the likelihood of being targeted increases, especially for larger organizations. In the event of a security breach, the execution

of a well-structured Cyber Breach Plan becomes crucial [27]. The key components of this plan, which will be detailed in the following paragraphs, are critical to both addressing the breach effectively and minimizing its impact.

The first component involves assembling an incident response team. Often, companies either fail to inform affected users promptly or deny the breach altogether. By forming a team that can professionally and timely report the breach to its users, a company can maintain or even gain user respect for its transparency and professionalism. This team should not only communicate about the breach but also work actively to limit access and halt the attack in progress. Addressing both the attack and its aftermath is essential for comprehensive incident management.

The second component is the identification of vulnerabilities and critical assets. By understanding potential attack vectors, a company can focus on employing penetration testers or enhancing cybersecurity software to safeguard these weak points. Recognizing where the most critical assets lie helps prioritize protection efforts.

The third component involves identifying backup resources. In a worst-case scenario, where a company loses all its primary resources, having redundant storage and backup systems is vital. This enables the rebuilding of databases and infrastructure in a timely manner, thereby minimizing harm, financial losses, and damage to customer and client trust.

The fourth component is developing a detailed response plan checklist/report. When an attack occurs, having a predefined checklist allows specialists to immediately and effectively address the breach. Post-attack, compiling a report helps analyze the breach's nature, identify the exploited vulnerability, and formulate strategies to prevent future attacks targeting the same weakness.

G. Proper Password Policies

The primary vulnerability in a company's database often lies not in the firewall or data transmission systems, but in password security. Once a malicious hacker acquires login credentials, they can attack from within the system, bypassing other security measures. Regular employee training can mitigate this risk, but human error remains a factor. Therefore, implementing robust password policies is essential for enhancing security and preventing breaches.

The first element of an effective password policy is establishing a minimum password length and mandating the use of various character types. Many websites and browsers, like Chrome, already enforce such restrictions, often suggesting strong passwords that meet these criteria. The inclusion of a mix of uppercase and lowercase characters, numerical digits, and special characters significantly complicates password cracking attempts through brute force [25]. While this doesn't make passwords impervious to hacking, it does provide a substantial layer of difficulty.

The second addition to the password policy involves an automatic lockout system and mandatory password resets. Setting a limit on the number of login attempts before an account lockout can thwart bots employing brute force methods [28]. In cases of lockout, either due to failed attempts or forgotten passwords, users would need to contact their IT department to reset their passwords. This ensures that both the user and the Cyber Security Department are alerted to potential security issues. Complementing this system with mandatory password changes every 60 to 90 days can further enhance security.

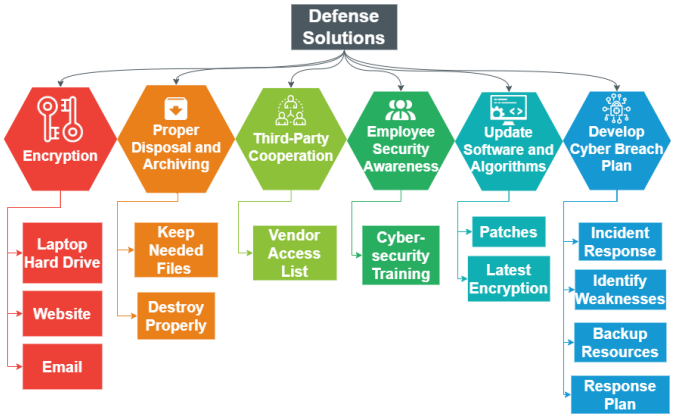

Figure 7: Defensive Solutions Against Breaches

The third enhancement is the implementation of multi-factor authentication (MFA). MFA requires users to provide two or more verification factors to gain account access. This typically includes a password coupled with a code sent via text, email, or an authentication app. MFA is particularly effective because it adds an additional layer of security; even if a password is compromised, unauthorized access is still prevented without the additional verification code.

H. Password Services

Some organizations have capitalized on the data from COMB to develop or refine password security tools aimed at individual users. A notable example is the 'Personal Data Leak Checker' offered by CyberNews. This service maintains a database comprising over 500GB of email addresses that have been compromised in various data leaks or breaches. Individuals can utilize this tool to check whether their email addresses are included in any known leaks. In addition to this leak checker, CyberNews provides other personal security services. One such service is a data leak alert system, which notifies users via email if their address is detected in any new data breaches [1]. These tools represent proactive measures in personal cybersecurity, allowing individuals to stay informed and take necessary steps to protect their online presence.

VI. Limitations

The research presented in this paper faced several constraints. Primarily, our analysis was based on secondary sources, relying on descriptions and analyses of COMB provided by organizations and individuals with direct access to the data. Additionally, the scarcity of academic literature on data breach compilations limited our ability to draw on similar

studies for comparative insights. One avenue for future research is a direct analysis of COMB. This could be approached in various ways. For instance, investigating email addresses that appear multiple times in COMB with different passwords could yield insights into patterns of individual password habits. Another approach is to compare COMB with other major breach compilations to discern similarities, differences, and trace its development timeline more accurately. Such analysis would be instrumental in uncovering the motivations behind the creation of breach compilations.

Another research recommendation involves a deeper examination of the specific exploits used in the breaches. While the companies we studied reported the consequences of their respective data leaks, they did not detail the specific exploits responsible. A thorough analysis utilizing an exploit database could provide more targeted insights into prevention techniques, addressing each type of breach. Although our current research scope and timeframe allowed only a preliminary exploration, delving into an exploit database could reveal critical vulnerabilities in system designs, offering a strong foundation for future preventative strategies.

VII. Conclusion

As breach compilations continue to expand annually, so too does the urgency to address our vulnerabilities. The increasing volume of data we accumulate heightens the risk of breaches, where even minor oversights can lead to significant data loss. However, implementing strategic defenses, such as multi-factor authentication, can effectively mitigate these risks. While the attack methodologies and defensive strategies discussed in this paper are comprehensive, they represent only a fraction of the considerations necessary for robust cybersecurity.

The future of data security presents two critical challenges: the inevitable human error and the advancement of AI and deep learning in password cracking. While human error cannot be completely eliminated, its impact can be minimized through stringent security protocols. Adapting to the evolving landscape of AI-assisted cyber threats will require continuous refinement of our cybersecurity measures.

The proliferation of internet services and users, coupled with the escalating frequency of breach incidents, ensures the ongoing growth of breach compilations. These compilations pose threats at both individual and organizational levels, often exploiting poor security practices. The gap between the widespread use of internet services and the adoption of fundamental security measures remains a significant concern. It is imperative for companies to recognize that investing in proactive cybersecurity is far more cost-effective than reacting to the aftermath of a data breach.

References

[1] B. Meyer, "Comb: Over 3.2 billion Email/Password Combinations Leaked." https://cybernews.com/news/largest-compilation-of-emails-and-passwords-leaked-free/, Jul 2022.

[2] R. Cresswell, "What You Should Know About The Comb Data Leak." https://www.acfeinsights.com/acfe-insights/what-you-should-know-about-comb-data-leak, Jun 2021.

[3] "Database Of Breaches." https://breachdirectory.com/breaches?lang=en/.

[4] J. Casal, "File With 1.4 Billion Hacked And Leaked Passwords Found On The Dark Web." https://medium.com/4iqdelvedeep/1-4-billion-clear-text-credentials-discovered-in-a-single-database-3131d0a1ae14, Dec 2017.

[5] M. Williams, "Inside The Russian Hack Of Yahoo: How They Did It." https://www.csoonline.com/article/3180762/inside-the-russian-hack-of-yahoo-how-they-did-it.html, Oct 2017.

[6] S. Ikeda, "The Data Dump Of 2.2 Billion Breached Accounts: What You Need To Know." https://www.cpomagazine.com/cyber-security/the-data-dump-of-2-2-billion-breached-accounts-what-you-need-to-know/, May 2019.

[7] "Canva Security Incident – May 24 FAQs." https://www.canva.com/help/incident-may24/, Jan 2020.

[8] "The Dark Overlord Cyber Investigation Report." https://nightlion.com/wp-content/uploads/2020/12/The-Dark-Overlord-Investigation-Report-Night-Lion_v1.01.pdf/.

[9] Boulton, "Netflix: How to check if your account has been hacked - and how to fix it." https://www.independent.co.uk/tech/netflix-hacked-recently-watched-fix-a6759336.html.

[10] S. Taylor, "Comb Data Leak: Everything You Need To Know." https://restoreprivacy.com/comb-breach-leak-compilation-of-many-breaches/, Jun 2021.

[11] L. J. Trautman and P. C. Ormerod. https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=2199&context=aulr.

[12] D. D. C. Dell Cameron, "The Great Data Breach Disasters of 2017." https://gizmodo.com/the-great-data-breach-disasters-of-2017-1821582178, Dec 2017.

[13] "Massive Hack Alert! 68 Million Dropbox Credentials Leaked Online." https://www.bitdefender.com/blog/hotforsecurity/massive-hack-alert-68-million-dropbox-credentials-leaked-online/.

[14] M. H. N. Ba, J. Bennett, M. Gallagher, and S. Bhunia, "A case study of credential stuffing attack: Canva data breach," *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 735–740, 2021.

[15] N. Krenz, "An Analysis Of The 2020 Zoom Breach." https://cloudsecurityalliance.org/blog/2022/03/13/an-analysis-of-the-2020-zoom-breach/.

[16] A. Spadafora, "Zoom Security Issues: What's Gone Wrong And What's Been Fixed." https://www.tomsguide.com/news/zoom-security-privacy-woes, Sep 2022.

[17] comparitech, "Data breach effects on stock market." https://www.comparitech.com/blog/information-security/data-breach-share-price-analysis.

[18] Waqas, "Anti public combo list with billions of accounts leaked." https://www.hackread.com/anti-public-combo-list-with-billions-of-accounts-leaked/, May 2017.

[19] A. Greenberg, "Hackers are passing around a megaleak of 2.2 billion records." https://www.wired.com/story/collection-leak-usernames-passwords-billions/, Jan 2019.

[20] F. Daragon, "Comb: The Big Password Leak." https://www.syhunt.com/en/?n=Articles.COMBPasswordLeak2021, Apr 2021.

[21] B. Hitaj, P. Gasti, G. Ateniese, and F. Perez-Cruz, "Passgan: A deep learning approach for password guessing," *Applied Cryptography and Network Security*, p. 217–237, 2019.

[22] N. Drager, "5 Ways To Use Encryption At Your Business To Prevent A Data Breach." https://quantumpc.com/encryption-prevent-data-breach/, May 2021.

[23] "How To Prevent Data Breaches: 12 Best practices — Paysimple." https://paysimple.com/blog/how-to-prevent-data-breach/.

[24] CardConnect, "Case study: What we've learned from the target data breach of 2013." https://cardconnect.com/launchpointe/payment-trends/target-data-breach.

[25] TSMNadmin, "6 Ways To Prevent Cybersecurity Breaches." https://techsupportofmn.com/6-ways-to-prevent-cybersecurity-breaches, Feb 2018.

[26] Kaspersky, "How Data Breaches Happen." https://www.kaspersky.com/resource-center/definitions/data-breach, Aug 2021.

[27] "How To Design A Cyber Incident Response Plan." https://www.embroker.com/blog/cyber-incident-response-plan/, Sep 2022.

[28] A. Froehlich, "How To Prevent A Data Breach: 10 Best Practices and Tactics." https://www.techtarget.com/searchsecurity/tip/How-to-prevent-a-data-breach-10-best-practices-and-tactics, Jul 2022.