Article

Enhancing Microsoft 365 Security: Integrating Digital Forensics Analysis to Detect and Mitigate Adversarial Behavior Patterns

Dr. Marshall S. Rich

[1] mrich@captechu.edu
[2] Correspondence: marshall.rich@richoncyber.com; Tel.: +1-478-747-3374

Abstract: This research article investigates the effectiveness of digital forensics analysis (DFA) techniques in identifying patterns and trends in malicious failed login attempts linked to public data breaches or compromised email addresses in Microsoft 365 (M365) environments. Pattern recognition techniques are employed to analyze security logs, revealing insights into negative behavior patterns. The findings contribute to the literature on digital forensics, opposing behavior patterns, and cloud-based cybersecurity. Practical implications include the development of targeted defense strategies and the prioritization of prevalent threats. Future research should expand the scope to other cloud services and platforms, capture evolving trends through more prolonged, more extended analysis periods, and assess the effectiveness of specific mitigation strategies for identified tactics, techniques, and procedures (TTPs).

Keywords: Microsoft 365; digital forensics analysis; adversarial behavior patterns; cybersecurity; malicious login attempts; data breaches; compromised email addresses; pattern recognition; cloud-based environments

1. Introduction

1.1. Introduction

The rapid adoption of cloud services, such as Microsoft 365 (M365), has provided organizations with numerous benefits, including increased productivity, flexibility, and cost savings [1,2,3]. However, this shift to cloud-based environments has also introduced new cybersecurity challenges and vulnerabilities, resulting in many data breaches and malicious activities targeting these platforms [1,3,4]. Failed login attempts are commonly seen as a sign of malicious activities, as adversaries often attempt to gain unauthorized access to M365 environments using compromised email addresses or credentials leaked in public data breaches [5].

In recent years, an increasing focus has been placed on understanding negative behavior patterns to enhance cybersecurity measures [6,7,8]. Digital forensics analysis (DFA) techniques, when applied to the analysis of security logs, can help reveal patterns and trends in these malicious activities [9,10]. Previous research has explored various aspects of digital forensics, negative behavior patterns, and cloud-based cybersecurity [6,8,11,12,13]. However, a comprehensive investigation into the effectiveness of DFA techniques for identifying patterns in failed login attempts in M365 environments remains underexplored.

Furthermore, understanding human factors and cybercriminals' psychology plays a significant role in developing targeted defense strategies and prioritizing prevalent threats [14,15,16,17,18]. Therefore, as the need for an in-depth understanding of these factors becomes more pressing, it is crucial to explore the effectiveness of DFA techniques for pattern recognition and identification of negative behavior patterns in cloud-based environments, such as M365[18].

Researchers have made significant strides in understanding negative behavior patterns in the context of cybersecurity, leading to valuable insights that have informed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

defensive strategies [6,8]. However, despite this progress, an unmet need remains to delve deeper into the practical implications of these patterns, particularly in the increasingly prevalent cloud-based environments, such as Microsoft 365 (M365) [1,2]. This study aims to investigate the effectiveness of DFA techniques for recognizing and identifying negative behavior patterns in failed login attempts in M365 environments.

1.2. Terms and Definitions.

Table 1. Terms and Definitions.

Section	Term	Definition
1.2.1	Exposure to data breaches	The potential risk is that an organization's data or information may be accessed, stolen, or compromised due to unauthorized access or cyberattacks. In the context provided, exposure to data breaches is determined by analyzing the valid email addresses in the organizational database and identifying any connections to known compromised data breaches [19].
1.2.2	Known compromised data breaches	Data breaches have been identified and documented in which unauthorized individuals have accessed, stolen, or compromised sensitive data. In this case, the focus is on breaches involving organizational email domains (.com, .org, .net, and .gov) [19].
1.2.3	Compromised email addresses across all known data breaches	Compromised email addresses across all known data breaches: This term refers to the email addresses of unique individuals found in the datasets of known compromised data breaches. In this study, 1,530 unique individuals have compromised email addresses across all the known data breaches [19].

The difference between (1.2.2) and (1.2.3) is that (1.2.2) refers to the specific instances of data breaches involving organizational email domains, while (1.2.3) refers to the email addresses of individuals that have been affected by these data breaches [41]. In other words, (1.2.2) focuses on the data breaches themselves, and (1.2.3) focuses on the individual email addresses affected by those breaches.

1.3. Scope.

This study's objectives and scope are defined by the analysis of audit and security logs in M365 environments, applying pattern recognition techniques to expose negative behavior patterns [6,20,21]. The aim is to contribute further to the existing body of research, notably the work done by Bhardwaj et al. [6] and Liu et al. [20], who have significantly advanced our understanding of adversarial behaviors in cybersecurity. This study narrows its focus to M365 environments, addressing the need for a more profound knowledge of these patterns' practical implications in cloud-based environments. These need conditions have been highlighted by Carlson [1] and El Jabri et al. [2], and Kim et al. [20].

In this research article, attention is given to the relationship between malicious failed login attempts in M365 tenants and known public data breaches or compromised email addresses [2,22]. Security logs from 162 days are analyzed, and the tactics, techniques, and procedures (TTPs) used by attackers during these incidents are examined. The paper discusses how DFA can be leveraged to identify negative behavior patterns [2,9,10], detect potential sources of malicious failed logins [13,22], and inform the development of proactive cybersecurity strategies for M365 tenants [2,23].

1.4. Research Question and Hypothesis

To address the knowledge gap, the study poses the following research question:

 doi:10.20944/preprints202306.1041.v1

- (RQ1). How effective are DFA techniques in identifying patterns and trends in malicious failed login attempts in M365 environments?

To answer this research question, the following hypothesis is proposed:

- (H1). DFA techniques employing pattern recognition can effectively identify patterns and trends in malicious failed login attempts within M365 environments, thereby contributing to developing targeted defense strategies and prioritizing prevalent threats [24,25].

By focusing on the analysis of failed login attempts, which often signify unauthorized access attempts, this study aims to uncover common tactics and approaches employed by adversaries, ultimately enhancing the understanding of the practical aspects of adversarial behavior in cloud-based environments [26]. A deeper understanding of these patterns will improve the ability to identify potential security breaches and contribute to the development of proactive security measures to mitigate future attacks.

1.5. Significance of the Research

The significance of this research lies in its contribution to the literature on digital forensics, negative behavior patterns, and cloud-based cybersecurity. Furthermore, by demonstrating the effectiveness of DFA techniques in identifying patterns and trends in malicious activities, the study has the potential to inform the development of targeted defense strategies and the prioritization of prevalent threats [6,20].

As noted by Bhardwaj et al. (2022) [6] and Liu et al. (2022) [20], understanding negative behavior patterns in the context of cybersecurity is crucial for developing effective countermeasures. By focusing on M365 environments, this study adds to the existing knowledge by explicitly looking at the individual challenges organizations face using cloud-based services [1,2,21].

Moreover, the findings from this study could have practical implications for organizations using M365, potentially aiding in the prevention and mitigation of data breaches and other cyber threats. The importance of proactively addressing cybersecurity challenges in cloud-based environments is stressed by El Jabri et al. (2021) [2], and this research could contribute to the development of more effective defense strategies tailored to the specific needs of M365 users [23].

Furthermore, the growing reliance of organizations on cloud-based services is highlighted by Carlson (2019) [1] and Kim et al. (2022)[21], which underscores the need for rigorous security measures to protect sensitive information. The insights gained from this study can support organizations in identifying and addressing vulnerabilities in their M365 environments, leading to enhanced security and reduced risk of data breaches.

In summary, this research is significant for its potential contributions to the literature on digital forensics, negative behavior patterns, and cloud-based cybersecurity. In addition, the study's focus on M365 environments and the application of DFA techniques can inform the development of targeted defense strategies and help organizations better protect their sensitive data from cyber threats [1,2,6,20].

2. Materials and Methods

2.1. Methodology

The study's methodology involves the collection of audit and security log data from M365 environments and applying pattern recognition techniques to identify trends and patterns in malicious failed login attempts. Furthermore, additional data from known public data breaches and compromised email addresses will be collected. By analyzing the combined logs and breach and compromised email data, a comprehensive understanding of the various TTPs employed by adversaries can be gained [6,20]. Consequently, the use of pattern recognition techniques is expected to enable the detection of unique patterns of behavior that may not be immediately apparent through manual analysis.

The collected audit and security log data will be the basis for the study's analysis. These logs contain information about the activity within the M365 environment, such as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

successful and unsuccessful login attempts, which can reveal potential cyber-attacks and adversaries' TTPs [1]. The data will be anonymized and processed to ensure compliance with privacy regulations and ethical considerations [1,2].

Next, pattern recognition techniques will be employed to uncover patterns and trends in malicious failed login attempts within the M365 environment. These techniques involve machine learning algorithms and statistical methods to analyze the data and detect anomalies indicative of adversarial behavior [6,20].

This study will employ a combination of DFA and pattern recognition techniques to identify adversarial behavior patterns in M365 cyber-attacks, focusing specifically on failed login attempts. By combining DFA and pattern recognition techniques, this research aims to detect and analyze attackers' TTPs, providing valuable insights for developing effective cybersecurity strategies and addressing the research question and hypothesis. The materials and methods section is organized as follows: data collection, data preprocessing, pattern recognition techniques, and validation, all in the context of supporting the research question and hypothesis [1,2,6,20].

2.2. Data Collection

M365 audit and security logs were obtained from participating organizations during the data collection stage, covering a 162-day reporting period. In addition, data from known public data breaches and compromised email addresses were gathered from various sources, such as online data breach repositories. This combined data served as the primary source for addressing RQ1 in this study.

Personally identifiable information, namely User IDs and email addresses, is securely anonymized using a uniquely generated four-digit numerical code, ensuring data privacy and confidentiality.

The logs encompass information on failed login attempts, user account details, source IP addresses, timestamps, known breach data, and compromised email addresses. Utilizing this information collectively could reveal potential cyber-attacks and adversaries' TTPs. This intelligence data supports RQ1 by providing insights into possible sources of malicious failed logins and best practices for cybersecurity strategies. Including data from known public data breaches and compromised email addresses will enhance the analysis, enabling a more comprehensive understanding of the broader context of the cyber threat landscape.

2.3. Data Preprocessing

Elastic's M365 Module [27] was utilized to collect and extract the relevant information for login attempts from each of the participating organizations' M365 tenants. In addition, the M365 Module facilitated the extraction of key features related to login attempts from the gathered data, supporting RQ1:

- Timestamp
- Login user ID
- Source IP address (attacker)
- Action (login attempt)
- Result (outcome of the login attempt).

The collected data underwent preprocessing to ensure its quality and relevance for the subsequent analysis (Figure 1). This process entailed:

- Anonymizing all personal data that could identify a user or an organization by assigning a random generated four-digit number.
- Removing irrelevant entries.
- Normalizing timestamp formats.
- Breaking out timestamp information and converting it to a numerical format.
- Converting the source IP to a numerical format.
- Transforming categorical data (known data breach and compromised email information) into numerical representations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

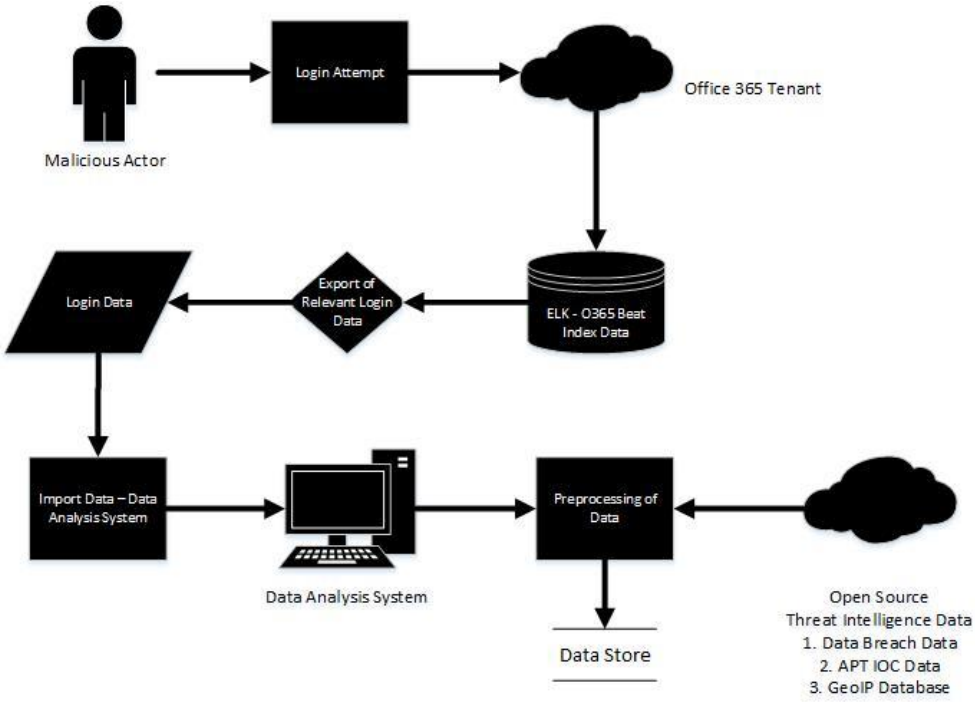

Figure 1. Process Workflow.

2.4. Pattern Recognition Techniques

Various pattern recognition techniques (Table 2) will be applied to the preprocessed data to detect and analyze adversarial behavior patterns in M365 login attempts, supporting RQ1.

Table 2. Pattern Recognition Techniques.

Section	Analysis Technique	Description
2.3.1.	Correlation Analysis	Spearman's rank correlation was used to measure the strength and direction of the relationship between breaches. This analysis helped identify significant correlations between breach pairs, highlighting shared TTPs or overlapping threat actor groups[28,20].
2.3.2.	Clustering Analysis	K-means clustering was employed to group user IDs based on their similarity in terms of failed login attempts, geographical distribution, and account statuses. This approach helped identify variations in user ID distribution among clusters, indicating differing risks of compromise[18,30].
2.3.3.	Association Rule Mining	The Apriori algorithm was used to discover interesting relationships between breach pairs and TTPs. Metrics like support, confidence, lift, leverage, and Zhang's metric were employed to evaluate the strength of these relationships. This analysis uncovered patterns within security logs, such as the frequent co-occurrence of specific TTPs, which can be used to understand better tactics employed by malicious actors and develop counter strategies[8,31].

The identified patterns were then analyzed in the context of the broader cyber threat landscape to gain insights into the attackers' strategies and motivations.

 doi:10.20944/preprints202306.1041.v1

2.5. Validation

A validation process was implemented to evaluate the effectiveness of the applied pattern recognition techniques in identifying patterns and relationships between known cyber data breaches. This process involved comparing the patterns and trends detected by the pattern recognition techniques against the additional data collected on known public data breaches, APT groups, and their TTPs.

The validation aimed to assess the accuracy of the detected patterns and the relevance and reliability of the identified patterns in addressing the research question. By correlating the detected patterns with the information from public data breaches, APT groups, and their TTPs, the study aimed to demonstrate the validity of the findings and the potential contribution to developing targeted defense strategies and best practices in cybersecurity.

The validation process was conducted in several steps. First, the patterns and trends identified through the pattern recognition techniques were compared with the known breach data, APT groups, and their TTPs. This comparison provided insights into the accuracy of the detected patterns and their correspondence with real-world incidents. Next, the effectiveness of the pattern recognition techniques in detecting patterns and relationships between cyber data breaches was assessed by examining the extent to which the different techniques supported and validated each other. This assessment helped determine the overall reliability and relevance of the findings for addressing the research question.

In summary, the validation subsection within the materials and methods section evaluates the accuracy, relevance, and reliability of the identified patterns and relationships between known cyber data breaches. By correlating the detected patterns with known public data breaches, APT groups, and their TTPs, the study aims to demonstrate the validity of the findings and the potential contribution to enhancing cybersecurity strategies and practices.

3. Results

3.1. Introduction to Results

This study aimed to identify adversarial behavior patterns in M365 cyber-attacks using DFA [3,32]. The methodology involved data collection and preprocessing, DFA, statistical methods, and ethical considerations [2, 6].

The study's main findings include the identification of adversarial behavior patterns within the data through clustering, correlation analysis, and association rule mining [7]. These techniques helped uncover significant relationships between failed login attempts and known public data breaches or compromised email addresses [3]. Data from multiple sources was integrated to enhance the detection of potential sources of failed malicious logins and adversarial behavior patterns [1,32]. This data provided context and additional information on known threat actors and TTPs associated with observed patterns and trends in the data [1, 14].

Various statistical methods (Table 2) were utilized to test the research hypothesis and assess the strength and significance of relationships between variables. In addition, descriptive statistics, such as frequency distributions and cross-tabulations, were used to summarize and visualize the data.

These findings address RQ1 and H1 by providing a comprehensive and robust analysis of the relationship between adversarial behavior patterns in M365 cyber-attacks and known public data breaches or compromised email addresses. The results of this analysis can inform organizations' cybersecurity strategies and contribute to the development of proactive cybersecurity measures, ultimately enhancing the security posture of organizations using M365 services.

3.2. Data Collection and Preprocessing Results

During the data collection stage, M365 security logs were collected from participating organizations over 162 days. These logs included information on failed login attempts,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

user account details, IP addresses, and timestamps. Additionally, data on known public data breaches and compromised email addresses were gathered from various sources, such as online data breach repositories.

3.2.1. Online Data Breach Repositories

Commercial services (Table 3) are utilized to identify organizational domains and individual email addresses known to have been compromised or part of a data breach.

Table 3. Public Services that Offer Email Breach Detection or Known Compromised.

Section	Technique	Description
3.2.1.1	Have I Been Pwned	Allows users to check whether their email address has been involved in known data breaches. Enter the email address, and the site will advise if it has been compromised.
3.2.1.2	BreachAlarm	Monitors the internet for stolen data that includes email addresses and sends an email alert if the email address is found in any compromised data.
3.2.1.3	Firefox Monitor	Allows users to check whether their email address has been involved in known data breaches. Users can sign up for alerts if their email address is found in a new data breach. Mozilla provides this service.
3.2.1.4	Identity Leak Checker	Allows users to check whether their email address has been involved in known data breaches. Users can also check for compromised usernames and passwords. The Hasso Plattner Institute provides this free service.
3.2.1.5	DeHashed	Allows individuals to search for compromised email addresses, usernames, and passwords. Users can sign up for alerts if their email address is found in a new data breach.

3.2.2. Demographic Distribution Analysis of Known Data Breaches:

A summary of the high-level demographic distribution of the known data breaches that affected the organizations involved in the study:

- The organizations' email domains were involved in sixty-nine known compromised data breaches.
- The organizations involved in the study have 2,968 valid email addresses, which were used to determine the exposure to data breaches.
- Out of valid email addresses, 1,530 unique accounts were found to have compromised email addresses across known data breaches.
    - 485 (16.341%) matching email addresses were found in both the list of valid organizational email addresses and the list of known compromised data breaches, suggesting a significant security concern for the organizations. These identified accounts were utilized for the study.
    - 956 (32.210%) fake or spoofed email addresses were identified in the known compromised data breaches. Although these email addresses were not valid organizational email addresses, they represent potential threats to the organization's email security. These identified accounts were excluded from the study.
    - 89 (2.999%) user IDs were excluded for not having complete or enough valid organizational information or email addresses. These identified accounts were also excluded from the study.
- A total of 3,925 compromised email addresses were used in the data breaches, indicating that some individuals experienced multiple breaches.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

The analysis of known data breaches revealed that various types of data were compromised, including email addresses, hashed and plaintext passwords, usernames, names, physical addresses, phone numbers, date of birth, social media profiles, personal preferences, payment information, and health and fitness data (Table 4). The potential TTPs employed in attacks include phishing campaigns, social engineering, exploiting unpatched vulnerabilities, credential stuffing, password spraying, brute force attacks, SQL injection, malware infections, third-party service compromises, insider threats, Advanced Persistent Threats (APTs), and supply chain attacks. These findings provide valuable insights into the nature of cyber-attacks and can help to inform the research and analysis in this study.

Additional demographic distribution information and supplemental material can be found in Appendix A (Demographic Distribution Summary of Known Data Breaches).

Table 4. Distribution across various industries.

Industry	Percentage
Finance & Insurance	35%
Healthcare	22%
Technology	16%
Retail	12%
Manufacturing	10%
Other Industries	5%

These insights gained during the preprocessing phase provided a foundation for further examination of potential security vulnerabilities, user behavior patterns, and the relationships between malicious failed login attempts and known public data breaches or compromised email addresses.

3.2.3. Retrospective Analysis

A retrospective analysis was performed on seventy-five known breach datasets, each representing a distinct cyber-attack victim across various industries. Of these, sixty-nine datasets were included in the study, as six were deemed unsuitable for further analysis. After examination of the excluded datasets, it was determined that the datasets lacked valid sets of organization user IDs or email addresses required for the research. A comprehensive list and summary of each of the sixty-nine data breaches that were included in the study can be found in Appendix B (Summary and List of Data Breaches).

The primary focus of the study was on data breaches related to email addresses where known organizational emails were potentially compromised. The study examined the timeline, size, and data types compromised in these breaches. In addition, correlation coefficients were calculated between pairs of datasets with known cyber data breaches where a user's ID was compromised to understand the relationships between these breaches.

3.2.4. Data Preprocessing

The data preprocessing stage involved cleaning and preparing the data for analysis (Figure 1). Duplicate entries were removed, anonymization of all personal data was completed, timestamps were normalized, and relevant data fields were extracted. The preprocessed data was then extracted into two separate datasets for further analysis.

Two distinct categories of login results were used in the preprocessing of the data for this study:

- Unsuccessful malicious failed login attempts (Dataset 1)
    - Dataset 1 (D1) comprised of 2,025,493 failed login attempts from 60,209 unique source IPs across 176 countries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

- In this dataset, the most frequent outcome of the login action observed was "UserLoginFailed," which aligns with the anticipated expectation.
- 449 unique user IDs, were identified for the study.

- Successful, legitimate logins (Dataset 2).
  - Dataset 2 (D2) contained 253,148 successful login attempts that originated from 8,990 unique source IPs across 99 countries.
  - In this dataset, the most frequent outcome of the login action observed was "UserLoggedIn," which is what was expected.

For this study, only D1 was utilized, which consisted of unsuccessful malicious failed login attempts. D2 was subsequently excluded from the technical research but is used as a comparison reference.

3.3. Pattern Recognition Results

In this section, the study presents the outcomes of the pattern recognition techniques applied to the data, including correlation analysis, clustering analysis, and association rule mining. The relationships identified are critical for understanding potential vulnerabilities, attack methods, and threat actor tactics. The implications of these findings on the overall cybersecurity threat landscape and forensics process are also explored.

3.3.1. Correlation Analysis Results

The study analyzed the relationships between known cyber data breaches by calculating correlation coefficients and p-values for each breach pair. A total of 2,016 correlation calculations were performed that found 98 meaningful correlations. The focus was then placed on the top ten pairs with the highest correlations, as these pairs are most likely to share similarities in how user IDs were compromised (Table 5.). The strong correlations suggest that these breaches may have similar causes, methods, or exploited vulnerabilities, which can inform organizations' cybersecurity strategies.

As an example, a significant correlation was found between the breaches “Apollo” and “Exactis” ($r = 0.787$, $p < 0.001$), highlighting potential risks associated with third-party vendors. This supports H1, which theorizes that digital forensics techniques can effectively analyze M365 security logs to identify patterns and trends in failed malicious login attempts linked to public data breaches or compromised email addresses.

Table 5. Top Ten Pairs with Highest Correlations:.

Pair	Correlation	P-value
LiveAuctioneers & Eye4Fraud	1	0
LiveAuctioneers & Drizly	1	0
Eye4Fraud & Drizly	1	0
MeetMindful & Houzz	0.989842782	0
LiveAuctioneers & EatStreet	0.978510047	0
Eye4Fraud & EatStreet	0.978510047	0
EatStreet & Drizly	0.978510047	0
NetGalley & LeadHunter	0.893865598	0
DataEnrichmentExposureFromPDLCustomer & Exactis	0.805917369	0
Verificationsio & Exactis	0.804184683	0

These correlations indicate that when a user's ID has been compromised in one dataset, there is a significant likelihood that it has also been compromised in the other dataset within the pair. The p-value of 0.0 for these pairs confirms that the correlations are statistically significant.

 doi:10.20944/preprints202306.1041.v1

The key algorithm in determining the meaningful correlations was Spearman's rank correlation. The formula for Spearman's rank correlation coefficient (ϱ) is:

$$\varrho = 1 - (6 * \Sigma d^2) / (n * (n^2 - 1))$$

Where:

- Σd^2 is the sum of the squared differences between the ranks of corresponding values in two columns.
- n is the number of data points (rows) in each column.

In summary, the analysis identified multiple instances of meaningful correlation between pairs of data breaches. These correlations may suggest shared characteristics, patterns, or vulnerabilities between the breaches. Further investigation into the nature of these correlations is warranted to understand the underlying factors better and potentially mitigate future breaches.

3.3.2. Clustering Analysis Results

A cluster analysis was performed to further investigate the relationships between the breaches, indicating a potential association between malicious failed login attempts and data breaches or compromised email addresses. Additionally, by grouping user IDs based on their similarity in terms of breach characteristics, this approach helped identify variations in user ID distribution among clusters, indicating differing risks of compromise.

Of 449 user IDs in D1, 215 were assigned to Cluster 1, suggesting a high likelihood of a compromised email address or involvement in a known data breach. These findings support H1, revealing a significant relationship between malicious failed login attempts in M365 tenants and known public data breaches or compromised email addresses. However, the study concludes that further research should be conducted to identify the most common TTPs and establish a stronger connection between these TTPs and the observed data breaches.

The study combined three datasets into one cluster matrix, where the first column represented user IDs, the following columns (excluding the last) represented known data breaches where the user ID was compromised (sixty-nine breaches or columns), and the last column represented the cluster assigned to each user. A comprehensive analysis of the merged dataset was conducted, emphasizing the distribution of user exposure to data breaches, identifying patterns within the clusters, and exploring the relationships between the clusters and the data breaches.

Descriptive statistics showed:

- Cluster 1 had the most significant number of user IDs, with 215.
- Cluster 2 had a moderate number of user IDs, with 117.
- Clusters 3, 4, and 5 had the smallest user IDs, with 52, 60, and 56, respectively.

Analysis of the combined cluster matrix revealed several key findings:

- A significant proportion of user IDs were associated with multiple data breaches, indicating that users are often exposed to multiple threats.
- Some data breaches were more prevalent across user IDs, suggesting that certain breaches have a wider-reaching impact on user exposure.
- The distribution of user IDs among clusters varied, with some clusters having a higher concentration of users exposed to specific data breaches.
- Relationships between clusters and data breaches were observed, with certain clusters being more strongly associated with specific data breaches.

An analysis was conducted leveraging both Appendix A (Demographic Distribution Overview of Data Breaches) and Appendix B (Detailed Account of Each Data Breach) to offer the subsequent synopsis and insights into the distinct traits of each cluster:

- Cluster 1 contained most of the dataset and likely represented those users who have experienced the most severe security incidents or breaches.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

- Cluster 2 represents users who have experienced more significant security incidents or breaches.
- Cluster 3 represents users who have experienced security incidents or breaches related to specific industries or regions.
- Cluster 4 represents users who have experienced some security incidents but are not as significant as those in other clusters.
- Cluster 5 represents users who have not experienced any significant data breaches or security incidents.

3.3.3. Association Rule Mining Results

The Association Rule Mining analysis revealed that credential stuffing, password spraying, and brute force attacks were the predominant TTPs associated with malicious failed login attempts in M365 tenant. Techniques were applied to discover interesting relationships between breach pairs and TTPs. Metrics like support, confidence, lift, leverage, and Zhang's metric were employed to evaluate the strength of these relationships.

The study further examined the relationship between TTPs and known public data breaches or compromised email addresses, finding a significant relationship between malicious failed login attempts and known public data breaches or compromised email addresses, which supported H1. This analysis uncovered patterns within security logs, such as the frequent co-occurrence of specific TTPs, which were used to understand better tactics employed by malicious actors and develop counter strategies. These TTPs, when combined in various ways, form the antecedents of the association rules, indicating that malicious actors often leverage a combination of these tactics to exploit M365 tenants (Table 6).

Table 6. Top 5 Association Rules.

Rank	Antecedents	Consequents	Confidence	Lift	Leverage	Zhang's Metric
1	{Exploit.In, Verifications.io}	{Data_Enrichment_Exposure_From_PDL_Customer, Anti_Public_Combo_List}	0.857143	34.675325	0.013094	0.986682
2	{Exploit.In, Data_Enrichment_Exposure_From_PDL_Customer, Verifications.io}	{Anti_Public_Combo_List}	0.857143	31.785714	0.013059	0.984018
3	{Exploit.In, Data_Enrichment_Exposure_From_PDL_Customer}	{Anti_Public_Combo_List, Verifications.io}	0.857143	31.785714	0.013059	0.984018
4	{Data_Enrichment_Exposure_From_PDL_Customer, Anti_Public_Combo_List}	{Exploit.In, Verifications.io}	0.545455	34.675325	0.013094	0.995776
5	{Anti_Public_Combo_List}	{Exploit.In, Data_Enrichment_Exposure_From_PDL_Customer, Verifications.io}	0.5	31.785714	0.013059	0.995381

As an example, evaluation of Association Rule 1 (Table 7) suggests that when "Exploit.In" and "Verifications.io" events are present, there's an 85.71% chance that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

"Data_Enrichment_Exposure_From_PDL_Customer" and "Anti_Public_Combo_List" events will also occur. The high lift value (34.68) and Zhang's Metric (0.99) indicate a strong positive association between the antecedents and consequents.

Table 7. Technical Analysis of Rule 1: An association rule mining analysis results.

Parameter	Value
Antecedents	'Exploit.In', 'Verifications.io'
Consequents	'Data_Enrichment_Exposure_From_PDL_Customer', 'Anti_Public_Combo_List'
Confidence	0.857143
Lift	34.675325
Leverage	0.013094
Zhang's Metric	0.986682

The other association rules in Table 7 follow the same pattern of interpretation.

- Rules 2, 3, and 5 have similar confidence, lift, and Zhang's Metric values, suggesting that these rules also have a strong positive association between the antecedents and consequents.
- Rule 4 has slightly lower confidence but still presents a high lift, and Zhang's Metric also indicates a strong positive association.

Overall, the analysis reveals strong associations between the events listed in the antecedents and consequents columns for all rules. This information can be valuable for understanding the relationships between different cybersecurity events and potentially predicting or preventing future incidents.

3.3.4. APT Groups and Data Breaches Results

The analysis of evaluated APT groups in the context of known data breaches revealed connections between certain groups and specific data breaches.

For example:

- APT28 (Fancy Bear) was linked to the LinkedIn breach, using spear-phishing and exploiting software vulnerabilities to compromise millions of user accounts.
- The Syrian Electronic Army (SEA) was suspected of being behind the Twitter breach, leveraging social engineering tactics and stolen credentials to gain unauthorized access.
- APT29 (Cozy Bear) was connected to the Dropbox breach, using advanced malware and lateral movement techniques to maintain persistence and exfiltrate data.

Additional and more detailed information can be found in Appendix C (APT Groups and Data Breaches). Understanding the preferred targets, TTPs, and associations of APT groups with specific breaches can help organizations assess their threats and develop appropriate defense strategies.

3.4. Validation Results

The study employed multiple pattern recognition techniques, including correlation analysis, clustering analysis, and association rule mining, to identify patterns and relationships systematically and comprehensively between known cyber data breaches. In addition, using multiple analytical methods provided validation and support for the findings and results, strengthening the conclusions drawn.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

The Correlation Analysis Results indicated meaningful correlations between breach pairs, suggesting shared characteristics or vulnerabilities. The Clustering Analysis Results complemented these findings by providing insights into the distribution of user IDs among clusters, further understanding the relationships between breaches. Finally, the Association Rule Mining Results revealed relationships between breach pairs and TTPs, providing additional context to the patterns and correlations observed in the previous analyses.

These three techniques provided unique insights and validated and supported each other's findings. In addition, the complementary nature of these analyses established a strong foundation for the study's conclusions, making them more reliable and credible. Overall, the validation results demonstrated that the pattern recognition techniques used were thorough and provided a robust understanding of the cybersecurity threat landscape.

3.5. Summary of Results

The study utilized various pattern recognition techniques to analyze known cyber data breaches and uncover patterns, relationships, and trends. The following are the key findings from each analysis; additional information is summarized in Table 8:

1. Correlation Analysis Results: 98 meaningful correlations were identified, with the top ten pairs having the highest correlations, suggesting shared characteristics, patterns, or vulnerabilities between the breaches.
2. Clustering Analysis Results: The analysis grouped user IDs based on their similarity in breach characteristics, revealing differing risks of compromise. It also showed relationships between clusters and data breaches, providing insights into specific threats and vulnerabilities.
3. Association Rule Mining Results: The analysis identified relationships between breach pairs and TTPs, uncovering patterns within security logs and helping to understand better the tactics employed by malicious actors.

Combining these techniques allowed for a comprehensive and systematic investigation of cyber data breaches. Using multiple methods to analyze the data contributed to a stronger validation of the results, ultimately providing a more reliable and credible understanding of the cybersecurity threat landscape.

Table 8. Summarized Results of Key Findings.

Section	Key Findings	Description
3.4.6.1	Pattern Recognition Results	Application of pattern recognition techniques (correlation analysis, clustering, association rule mining) revealed significant patterns and vulnerabilities targeted by threat actors, leading to better identification and categorization of threats.
3.4.6.3	Demographic Distribution Summary Results	Data breaches affected many industries and sectors, compromising billions of user records. Understanding common targets and vulnerabilities exploited by threat actor's aids in proactive measures for high-risk sectors or regions.
3.4.6.4	APT Groups and Data Breaches Results	Overview of known APT groups in the context of data breaches, including preferred targets, TTPs, and associations with specific breaches, helping organizations identify potential threats and understand various APT tactics.

The validation results supported the research methodology and demonstrated the value of the identified patterns and vulnerabilities. This study's findings can help organizations develop proactive cybersecurity strategies, prioritize defenses and resources against relevant threats, and contribute to threat intelligence and the cybersecurity forensics process by examining the top correlated breaches, clusters, and association rules.

4. Discussion Section

 doi:10.20944/preprints202306.1041.v1

4.1. Introduction

The primary findings of this study are presented, offering insights into the TTPs employed by threat actors in Microsoft 365 (M365) cyber-attacks, emphasizing malicious login attempts. Additionally, these findings are interpreted and contextualized concerning the stated research question and hypothesis, aiming to assist organizations in bolstering their cybersecurity posture.

Significant correlations were observed between malicious login attempts, public data breaches, and known compromised email addresses. This suggests that threat actors exploit these connections to gain unauthorized access to M365 accounts, underscoring the importance of understanding the relationships among breached datasets, threat actors, and their TTPs. This understanding can help organizations strategically prioritize their defenses and allocate resources more effectively.

In addition, patterns of adversarial behavior were recognized and linked to specific APT groups and data breaches. This analysis provides valuable insights into the threat actors' preferred targets, TTPs, and associations with specific breaches. The insights derived from these findings can be instrumental in the cybersecurity forensics process, particularly in identifying and attributing potential future attacks.

Moreover, an analysis of demographic distribution provided a comprehensive understanding of the frequent targets and vulnerabilities threat actors exploit. This emphasizes the need for organizations, particularly those within high-risk sectors or regions, to develop and implement proactive cybersecurity strategies.

This discussion further examines the relationships between these core findings and their implications for organizations seeking to enhance their cybersecurity posture. By interpreting and contextualizing these results concerning the research question and hypothesis, this section provides guidance for organizations. This advice prioritizes defenses and resources to counter the most relevant threats, ultimately augmenting their ability to detect, prevent, and respond to cyber threats.

4.2. Interpretation of Results

The study's results provide valuable insights into the patterns and relationships associated with known cyber data breaches and connections to specific APT groups and compromised email addresses. These findings carry significant implications for organizations utilizing various cloud services. By comprehending the TTPs related to these breaches, security teams can develop targeted defense strategies and prioritize the most prevalent threats.

During the analysis of the top pairs with the highest correlation coefficients (Table 2), significant correlations were identified between known cyber data breaches and the demographic data present in a large dataset of malicious activities. The focus on compromised user IDs yielded insights that addressed the research question and hypothesis. Incorporating demographic data enabled the study to reach meaningful conclusions about the similarities between these breaches and their implications for cybersecurity posture and forensic processes.

A quantitative analysis of the data breaches revealed a considerable vulnerability of organizations' email systems to cyber threats. In addition, a notable portion of the valid email addresses was compromised in known data breaches (485 or 16.341%), suggesting that organizations' email security measures may be inadequate and necessitate further enhancement to safeguard sensitive information.

The presence of fake or spoofed email addresses in data breaches (956 or 32.210%) indicates a potential risk of phishing attacks and other malicious activities. Although not directly linked to organizations, these fake email addresses can potentially damage reputation and undermine the trust of customers and partners.

The interpretation of the study's results emphasizes the importance of understanding the relationships between malicious activities, public data breaches, and compromised email addresses for organizations seeking to enhance their cybersecurity posture.

 doi:10.20944/preprints202306.1041.v1

Furthermore, analyzing these patterns and connections fosters a more proactive approach to cybersecurity, ultimately improving the detection and mitigation of threats related to cyber data breaches.

4.2.1. Pattern Recognition Results Interpretation

Insights were derived from the study by utilizing pattern recognition techniques, including correlation analysis, cluster analysis, and association rule mining. These approaches were critical in understanding adversarial behavior patterns and M365 cyberattacks. By identifying the correlations, clusters, and associations in the context of malicious failed login attempts, the outcomes of the methodologies rendered invaluable insights into the TTPs used by threat actors.

This study supports the hypothesis that a significant relationship exists between malicious failed login attempts in M365 tenants and known public data breaches or compromised email addresses. Digital forensics techniques effectively analyzed M365 security logs, identifying patterns and trends in failed malicious login attempts linked to public data breaches or compromised email addresses. Additionally, integrating APT data further enhanced the detection of potential sources of failed malicious logins in M365 tenants, informing the development of a proactive cybersecurity strategy.

The findings highlight several key patterns and differences, offering information that can inform the development of cybersecurity policies and strategies.

4.2.1.1. Brute Force Attacks and Credential Stuffing

The high volume of unsuccessful malicious failed login attempts suggests that adversaries may employ brute force and credential-stuffing techniques to gain unauthorized access to accounts. These tactics demonstrate an understanding common user behaviors, such as using weak or easily guessable passwords. By targeting accounts with potentially inadequate security measures, attackers can potentially gain access to sensitive information or compromise organizational systems. As a result, organizations should consider implementing stronger authentication methods, such as multi-factor authentication (MFA), to mitigate the risk of brute force and credential-stuffing attacks.

4.2.1.2. Targeted Accounts and High-Value Users.

The concentration of failed login attempts on the top twenty user IDs within the malicious dataset indicates that attackers may target specific accounts due to their perceived value or access to sensitive information. Identifying high-value accounts and implementing additional security measures can help protect these accounts from targeted attacks. Additionally, consistent monitoring and reviewing of account privileges can further minimize the potential impact of unauthorized access.

Appendix D serves the purpose of defining the various types of 365 accounts and mailboxes, as well as distinguishing between them and highlighting which accounts allow direct access and which do not. This differentiation is essential in identifying the most critical accounts in terms of risk. By analyzing D1, seven valuable and vulnerable accounts being targeted were discovered that lacked MFA protection.

4.2.1.3. Inactive and Disabled Accounts.

A significant majority of the targeted accounts in the D1 dataset were found to be inactive or disabled, almost 75%, but the accounts did exist within the M365 tenant. This pattern suggests attackers might target dormant accounts due to weaker security measures or lack of monitoring. Regular audits of user accounts and prompt disabling or removal of inactive accounts can help organizations reduce the risk associated with dormant accounts.

4.2.2. Interpretation of Results for RQ1 and H1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

The findings related to patterns and trends in failed malicious login attempts offer meaningful insights into the TTPs of attackers in M365 cyber-attacks. The study confirms H1, emphasizing the importance of leveraging advanced technology to detect and prevent cyber threats, particularly in large and complex IT environments like M365.

Technical analysis using association rule mining revealed patterns within the security logs, such as the frequent co-occurrence of specific TTPs. Digital forensics analysts can utilize this information to better understand the tactics employed by malicious actors and develop strategies to counter them. In addition, the top association rules show strong relationships between various combinations of the identified TTPs, which security teams can use to identify patterns and trends in malicious login attempts and develop targeted mitigation strategies.

For RQ1, the study demonstrated the effectiveness of digital forensics techniques in analyzing M365 security logs. The strong correlations in the analysis suggest that these breaches may have typical causes, methods, or exploited vulnerabilities. By understanding the commonalities between these breaches and considering the demographic data, organizations can defend against these threats (Table 9).

Table 9. Recommendations and Actions.

Section	Recommendations	Description
4.2.2.1	Enhance threat intelligence	Better understand the threat landscape and prepare for potential attacks, focusing on the most active regions.
4.2.2.2	Prioritize vulnerability management	Address security weaknesses exploited in similar breach pairs.
4.2.2.3	Develop incident response playbooks	Develop playbooks and procedures based on the correlations, findings, and demographic data for faster detection and containment of breaches.
4.2.2.4	Increase user awareness	Raise awareness of the risks associated with data breaches and provide targeted training to reduce the likelihood of successful social engineering attacks, especially for regular users (members or active M365 accounts).
4.2.2.5	Share findings and collaborate	Collaborate with industry peers and information-sharing organizations to collectively improve defensive postures and contribute to a better understanding of the threat landscape.

The study showcases the effectiveness of digital forensics techniques in identifying patterns and trends in failed malicious login attempts linked to public data breaches or compromised email addresses. The insights gained can assist organizations in strengthening their cybersecurity posture and developing more effective strategies to counter cyber threats.

4.3. Comparison to Previous Research

The study builds upon and extends the work of previous research in the areas of cloud security, digital forensics, and human factors in cybersecurity. In this section, the study compares findings with those from relevant literature to highlight the advancements and contributions of their research.

The focus on understanding adversarial patterns to enhance defense strategies resonates with McCall's (2022)[33] work on a cyber threat intelligence approach to thwart

 doi:10.20944/preprints202306.1041.v1

adversary attacks. However, this study advances this concept within the context of Microsoft 365 failed login attempts.

Nisioti et al. (2021)[21] applied game-theoretic decision support in cyber forensic investigations, which aligns with the present study's aspiration to devise innovative strategies for mitigating malicious activities. This study, however, concentrates on pattern recognition and human factors, suggesting that including game-theoretic approaches in future research could provide further valuable insights, much like Tyworth et al.'s (2013)[26] human-in-the-loop approach.

Pangsuban et al.'s (2020)[34] work on real-time risk assessment for information systems using machine learning techniques aligns with the present study's use of advanced detection methods. However, the focus of the current research on human factors and pattern recognition provides a unique perspective on malicious login attempts.

Several studies, including those by Parsons et al. (2010)[36], Rahman et al. (2021)[24], Ramlo & Nicholas (2021)[16], Sutter (2020)[25], and Triplett (2022)[22], investigated various human factors in cybersecurity. This research contributes to this body of knowledge by identifying specific behavioral patterns related to failed login attempts in Microsoft 365, echoing Scott & Kyobe's (2021)[35] investigation into the intersection of human elements and technology.

Salik's (2022)[23] exploration of the failure of offensive cyber operations to deter nation-state adversaries offers crucial insights into the mindset and strategies of cyber adversaries, aligning with Bhardwaj et al.'s (2022)[6] analysis of advanced adversaries. While the current research does not explicitly focus on nation-state adversaries, understanding these adversarial strategies informs the study of malicious login attempts.

Like Scott & Kyobe's (2021)[35] study, this research investigates the interplay between human factors and technology in dealing with malicious login attempts. Furthermore, Singh's (2021)[31] focus on the role of stress among cybersecurity professionals provides further understanding of psychological factors that could impact detection and prevention efforts.

In line with Tyworth et al.'s (2013)[26] human-in-the-loop approach, this study focuses on human factors in managing malicious login attempts. This approach could potentially augment the DFA techniques utilized in this research. Finally, Wells et al.'s (2021)[27] assessment of the credibility of cyber adversaries echoes this study's goal to understand the threat landscape better.

Like the present study, Bhardwaj et al. (2022)[6] sought to analyze and detect advanced adversaries. However, this research does so within the specific context of Microsoft 365 rather than a general behavior-based threat-hunting framework. This Microsoft 365 focus resonates with Carlson's (2019)[1] investigation into hybrid forensics in Microsoft 365 and Exchange Server.

Like Amin's (2010)[10] work, the present research employs pattern recognition techniques to understand malicious activity. However, it focuses on identifying adversarial behavior patterns through DFA techniques. Cornejo's (2021)[4] exploration of human errors in data breaches aligns with the current study's investigation of malicious failed login attempts.

Dalal et al.'s (2022)[30] interdisciplinary approach to cybersecurity mirrors this study's multifaceted understanding of cloud-based cybersecurity. Derbyshire's (2022)[7] work on anticipating adversary costs and bridging the threat-vulnerability gap complements the current ' 'research's endeavor to predict adversarial behaviors. Similarly, El Jabri et al.'s (2021)[2] use of Microsoft Office 365 audit logs to study cloud security further expands the understanding of adversarial behaviors in the context of Microsoft 365.

This study benefits from insights from Jeong et al. (2019)[18], Kioskli and Polemi (2020)[30], Liu et al. (2022)[20], and Mavroeidis and Jøsang (2021)[8]. These researchers discussed human factors, psychosocial approaches, knowledge graphs, and data-driven threat hunting, respectively. These aspects have been considered in the present research, contributing to a more distinctive understanding of malicious failed login attempts in Microsoft 365.

 doi:10.20944/preprints202306.1041.v1

The research also benefits from discussing real-time risk assessment for information systems using machine learning techniques by Pangsuban et al. (2020)[34]. This aspect aligns with the current study's intention to apply advanced detection methods to malicious login attempts. However, recent research also emphasizes human factors and pattern recognition, which may offer a different perspective on the problem.

Expanding on the focus of McCall Jr. (2022)[33] on the cyber threat intelligence approach to thwarting adversary attacks, the current study explores patterns and trends in malicious activities, drawing upon the game-theoretic decision support for cyber forensic investigations proposed by Nisioti et al. (2021)[21].

The current study, like Parsons et al. (2010)[36], Rahman et al. (2021)[24], Ramlo and Nicholas (2021)[16], Scott and Kyobe (2021)[35], Singh (2021)[31], Sutter (2020)[25], Triplett (2022)[22], and Tyworth et al. (2013)[26], investigated human factors in cybersecurity. However, it specifically focuses on failed login attempts in Microsoft 365.

The insights provided by Salik (2022)[23] on offensive cyber operations and the inability to deter nation-state adversaries were also considered, providing additional context for understanding malicious login attempts. Finally, Wells et al.'s (2021)[27] assessment of the credibility of cyber adversaries was used to inform the current study's understanding of the threat landscape.

Finally, Wells et al.'s (2021)[27] evaluation of the credibility of cyber adversaries reverberates with this study's aim to understand the threat landscape better. Singh's (2021)[31] examination of the impact of stress on cybersecurity professionals also offers an additional layer of understanding of the psychological factors that could influence detection and prevention efforts.

Overall, the current research contributes to the existing literature by blending elements of human behavior, advanced analytical techniques, and a specific focus on Microsoft 365 failed login attempts. In addition, it integrates insights from these various studies to create a more comprehensive understanding of adversarial behaviors and defense mechanisms within the Microsoft 365 environment.

In conclusion, this comparison of previous research highlights the diverse range of approaches and techniques used to study adversarial behavior patterns in cybersecurity. By incorporating the insights and methodologies from these studies, the current research can deepen its understanding of malicious login attempts in Microsoft 365 and develop more effective digital forensics analysis techniques to address this challenge.

4.4. Practical Implications and Recommendations

4.4.1. Implications and Impact

The study's findings have significant implications for the expansive field of cybersecurity. The study demonstrates the value of using digital forensics techniques alongside cyber threat intelligence data. These tools are crucial in scrutinizing M365 security logs.

With the aid of these methods, we can identify patterns and trends in unsuccessful malevolent login attempts. Such attempts often have connections to public data breaches or compromised email addresses. With these insights, organizations can be better equipped to enhance the following:

4.4.1.1. Develop more proactive and targeted cybersecurity strategies.

By understanding the relationships between breach pairs and the tactics used by threat actors, organizations can prioritize defenses against the most relevant threats and tailor their security measures to counter specific TTPs.

4.4.1.2. Enhance threat detection and response capabilities.

Integrating digital forensics techniques with cyber threat intelligence data can provide a more comprehensive understanding of the potential sources of failed malicious logins. These actions can provide additional information to an organization's efforts to detect, prevent, and respond to cyber threats more effectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

4.4.1.3. Strengthen overall cybersecurity posture.

The study's findings can guide organizations in identifying and addressing vulnerabilities in their security infrastructure, as well as in implementing security best practices to mitigate risks associated with data breaches and compromised email addresses.

4.4.1.4. Foster collaboration and information sharing.

The insights gained from this research highlight the importance of sharing threat intelligence data among organizations, industry partners, and government agencies to enhance their ability to combat emerging cyber threats collectively.

This study provides valuable insights into applying digital forensics techniques for analyzing M365 security logs and identifying patterns and trends in failed malicious login attempts linked to public data breaches or compromised email addresses. By understanding these aspects, organizations can enhance their ability to detect, prevent, and respond to cyber threats, ultimately strengthening their overall cybersecurity strategy.

4.4.2. Limitations and Future Research

Despite the insights gained from the study, several limitations should be acknowledged:

- **Scope of data**: The research focused on malicious failed login attempts and their connections to public data breaches and TTPs in M365 tenants. Consequently, the findings may not be generalizable to other cloud-based platforms or cybersecurity contexts.
- **Data collection period**: The data analyzed in this study were collected over a specific time frame. As cyber threats continuously evolve, further research should be conducted periodically to ensure the relevance and effectiveness of the proposed strategies.
- **Human factors**: While the importance of addressing human factors in cybersecurity was discussed, the research did not thoroughly explore the psychological, social, and organizational aspects that may contribute to the observed patterns of malicious failed login attempts. Future research could investigate these dimensions more comprehensively.
- **Insider threats**: The study primarily focused on external threats associated with malicious failed login attempts. Future research could expand the scope to include insider threats and investigate potential links between internal and external threat actors.
- **Causality**: The relationships observed in the study are correlational and do not necessarily imply causality. Future research could employ experimental or longitudinal designs to understand better the causal relationships between malicious failed login attempts, public data breaches, and TTPs.
- **Mitigation strategies**: The research focused on analyzing and understanding the relationships between malicious failed login attempts, public data breaches, and TTPs, rather than proposing specific mitigation strategies. Future research could build on these findings to develop more targeted and effective cybersecurity defenses.

In summary, future research should address these limitations by expanding the scope of data, investigating human factors and insider threats, exploring causal relationships, and proposing targeted mitigation strategies. Furthermore, regular updates to the research are necessary to keep pace with the ever-evolving cyber threat landscape. Continued research will help organizations maintain a robust cybersecurity posture and protect their valuable data and resources.

5. Conclusions

5.1. Summary of Main Findings

 doi:10.20944/preprints202306.1041.v1

The study utilized pattern recognition techniques to understand adversarial behavior patterns in M365 cyber-attacks.

The main findings include:

- A significant relationship exists between malicious failed login attempts in M365 tenants and known public data breaches or compromised email addresses.
- Digital forensics techniques effectively analyze M365 security logs, identifying patterns and trends in failed malicious login attempts linked to public data breaches or compromised email addresses.
- APT data integration enhances the detection of potential sources of failed malicious logins in M365 tenants and informs the development of proactive cybersecurity strategies.

The following adversarial behavior patterns were observed and discovered using Digital Forensics Analysis (DFA) processes:

- The study used association rule mining to reveal patterns within the security logs, highlighting the frequent co-occurrence of specific TTPs employed by malicious actors.
- Top association rules revealed in the study show strong relationships between multiple combinations of the identified TTPs. Security teams can use this information to identify patterns and trends in malicious login attempts and develop targeted mitigation strategies.
- Correlation analysis demonstrated the potential of using breach and APT data to detect potential sources of failed malicious logins and inform proactive cybersecurity strategy development. Significant correlations were found between different breaches.
- Cluster analysis identified distinct user ID clusters with varying risk levels, helping organizations prioritize defenses and allocate resources against relevant threats.

These findings contribute to a better understanding of the tactics employed by malicious actors, enabling organizations to develop more effective strategies to counter cyber threats.

5.2. Contributions to the Field

This study contributes to the literature on digital forensics, adversarial behavior patterns, and cybersecurity in cloud-based environments.

It advances the current understanding by:

- Determining a significant relationship exists between malicious failed login attempts in M365 tenants and known public data breaches or compromised email addresses.
- Demonstrating the effectiveness of digital forensics techniques in analyzing M365 security logs and identifying malicious login attempt patterns.
- Providing insights into the TTPs employed by threat actors in M365 cyber-attacks.

5.3. Practical Implications

The study's findings have practical implications for organizations using M365, including:

- Enhanced detection and mitigation of malicious failed login attempts by leveraging digital forensics techniques.
- Improved understanding of the threat landscape, enabling organizations to adopt a proactive stance toward cybersecurity.
- Targeted allocation of resources and prioritization of defenses against the most relevant threats based on the identified patterns and trends.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2023 doi:10.20944/preprints202306.1041.v1

5.4. Potential areas for future research include:

Investigating the effectiveness of other pattern recognition techniques or machine learning algorithms in identifying adversarial behavior patterns in cloud-based environments.

- Examining the role of artificial intelligence and automation in enhancing the analysis of M365 security logs.
- Exploring the impact of new cybersecurity policies, regulations, or industry standards on mitigating M365 cyber-attacks and developing proactive cybersecurity strategies.

5.5. Regarding future research directions

While the study offers valuable insights into the cyberpsychology and adversarial behavior patterns in M365 cyber-attacks, it is essential to acknowledge its limitations and the need for further research. Future studies can build upon the findings by examining more extensive and diverse datasets, incorporating other attack vectors, and investigating the relationship between threat actors' TTPs and various cyber defense strategies.

Future research directions:

- Expanding the scope of analysis to other cloud services and platforms would provide a more extensive understanding of adversarial behavior patterns and TTPs across different environments.
- Analyzing data over more extended periods can capture evolving trends and help identify shifts in threat actor tactics, providing valuable insights into the ever-changing cyber threat landscape.
- Further investigation into the effectiveness of specific mitigation strategies for the identified TTPs would be beneficial in providing actionable recommendations for organizations.
- Research could involve controlled experiments or simulations to evaluate the efficacy of various countermeasures and their impact on reducing the risk of successful cyber-attacks.

5.6. Final Thoughts

This study scientifically analyzes malicious failed login attempts in M365 tenants, their relationships with known public data breaches, compromised email addresses, and TTPs. By integrating DFA processes, the research offers valuable insights for organizations seeking to improve their cybersecurity posture and address human factors in security. The findings reveal that attackers employ tactics such as brute force and credential stuffing to target accounts with weak security measures, and they often focus their efforts on high-value users who may have access to sensitive information.

The research adds to the existing body of knowledge by comprehensively analyzing malicious failed login attempts in M365 tenants and their connections with known public data breaches, compromised email addresses, and TTPs. The integration of DFA emphasizes the importance of a data-driven approach and highlights the need to address human factors in cybersecurity. The findings offer valuable insights for organizations looking to strengthen their cybersecurity posture and develop targeted employee training programs based on identified TTPs.

The study found that most targeted accounts in the maliciously failed login attempts dataset were inactive or disabled, emphasizing the need for organizations to regularly audit user accounts and promptly disable or remove dormant accounts. By recognizing these patterns and implementing appropriate security measures, organizations can enhance their security posture and better protect themselves against evolving cyber threats.

This study underscores the importance of robust email security measures to protect organizations from data breaches and cyber threats. In addition, the findings highlight the need for continuous monitoring, effective security protocols, and employee training to minimize the risk of email compromises. Further research is required to explore the

effectiveness of specific email security practices and to develop novel strategies for safeguarding organizational email systems.

In conclusion, understanding the patterns and differences between unsuccessful malicious or failed login attempts and successful, legitimate, or attempted logins can significantly contribute to developing more effective cybersecurity strategies. Furthermore, by leveraging this knowledge, organizations can strengthen their overall security posture, mitigate cyber risks, and better protect their sensitive information and critical infrastructure.

Author Contributions: The sole author, Dr. Marshall S. Rich, was responsible for all aspects of the research, including conceptualization, methodology, data analysis, writing, and editing.

Funding: This research received no external funding. All costs were incurred by the author.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Access to the study data and associated code can be granted upon request to the corresponding author. Public availability is not provided in order to maintain controlled access and to protect the integrity of both the data and the code.

Conflicts of Interest: The author declares no conflicts of interest related to the research.

References:

1. Carlson, A. (2019). Microsoft 365 and Exchange Server Hybrid Forensics (Doctoral dissertation, Utica College). ProQuest Dissertations Publishing. (27670117).
2. El Jabri, C., Frappier, M., Tardif, P.-M., Lepine, G., & Boisvert, G. (2021). Statistical approach for cloud security: Microsoft Office 365 audit logs case study. In The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings (pp. 1-6). Piscataway. DOI: 10.1109/DSN-W52860.2021.00014
3. Back, S., & LaPrade, J. (2019). The future of cybercrime prevention strategies: Human factors and a holistic approach to cyber intelligence. International Journal of Cybersecurity Intelligence and Cybercrime, 2(2), 1-4.
4. Cornejo, G. A. (2021). Human Errors in Data Breaches: An Exploratory Configurational Analysis. ProQuest Dissertations Publishing, Nova Southeastern University. (28775912)
5. Huang, T.-K. (2013). Understanding online malicious behavior: Social malware and email spam (Doctoral dissertation, University of California, Riverside). ProQuest Dissertations Publishing. (3600570).
6. Bhardwaj, A., Kaushik, K., Alomari, A., Alsirhani, A., Alshahrani, M. M., & et al. (2022). BTH: Behavior-Based Structured Threat Hunting Framework to Analyze and Detect Advanced Adversaries. Electronics, 11(19), 2992. https://doi.org/10.3390/electronics11192992.
7. Derbyshire, R. J. (2022). Anticipating Adversary Cost: Bridging the Threat-Vulnerability Gap in Cyber Risk Assessment. ProQuest Dissertations Publishing, Lancaster University (United Kingdom). (29424473)
8. Mavroeidis, V., & Jøsang, A. (2021, March 28). Data-Driven Threat Hunting Using Sysmon. arXiv.org. https://arxiv.org/abs/2103.14903.
9. Montasari, R. (2021). The Comprehensive Digital Forensic Investigation Process Model (CDFIPM) for Digital Forensic Practice. ProQuest Dissertations Publishing, University of Derby (United Kingdom). (28460690)
10. Amin, R. M. (2010). Detecting targeted malicious email through supervised classification of persistent threat and recipient-oriented features (Doctoral dissertation, The George Washington University). ProQuest Dissertations Publishing. (3428188).

11. Agrawal, G., Deng, Y., Park, J., Liu, H., & Chen, Y.-C. (2022). Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education. Information, 13(11), 526. https://doi.org/10.3390/info13110526.
12. Mouzakitis, S., & Askounis, D. (2021). Assessing MITRE ATT&CK risk using a cyber-security culture framework. Sensors, 21(9), 3267. https://doi.org/10.3390/s21093267.
13. Serketzis, N., Katos, V., Ilioudis, C., Baltatzis, D., & Pangalos, G. J. (2019). Actionable threat intelligence for digital forensics readiness. Information and Computer Security, 27(2), 273-291. https://doi.org/10.1108/ICS-09-2018-0110.
14. Ferguson-Walter, K. J., Gutzwiller, R. S., Scott, D. D., & Johnson, C. J. (2021). Oppositional human factors in cybersecurity: A preliminary analysis of affective states. In Proceedings of the Institute of Electrical and Electronics Engineers (IEEE) Conference (pp. 153-158). https://doi.org/10.1109/ASEW52652.2021.00040.
15. Greitzer, F. L., & Hohimer, R. E. (2011). Modeling human behavior to anticipate insider attacks. Journal of Strategic Security, 4(2), 25-48. https://doi.org/10.5038/1944-0472.4.2.2. Retrieved from http://scholarcommons.usf.edu/jss/vol4/iss2/3.
16. Ramlo, S., & Nicholas, J. B. (2021). The human factor: assessing 'individuals' perceptions related to cybersecurity. Information and Computer Security, 29(2), 350-364. https://doi.org/10.1108/ICS-04-2020-0052
17. Rohan, R., Funilkul, S., Pal, D., & Chutimaskul, W. (2021). Understanding of Human Factors in Cybersecurity: A Systematic Literature Review. In International Conference on Computational Performance Evaluation (ComPE) (pp. 133-140.). IEEE. https://doi.org/10.1109/ComPE48788.2021.00022
18. Jeong, J., Mihelcic, J., Oliver, G., & Rudolph, C. (2019). *Towards an Improved Understanding of Human Factors in Cybersecurity*. In IEEE 5th International Conference on Collaboration and Internet Computing (CIC) (pp. n.a.). IEEE.
19. Hultquist, K. E. (2011). An Analysis of the Impact of Cyber Threats Upon 21st Century Business. (Doctoral Dissertation). The College of St. Scholastica, ProQuest Dissertations Publishing. (1503100).
20. Liu, K., Wang, F., Ding, Z., Liang, S., Yu, Z., & et al. (2022). Recent Progress of Using Knowledge Graph for Cybersecurity. Electronics, 11(15), 2287. https://doi.org/10.3390/electronics11152287.
21. Nisioti, A., Loukas, G., Rass, S., & Panaousis, E. (2021). Game-Theoretic Decision Support for Cyber Forensic Investigations. Sensors, 21(16), 5300. https://doi.org/10.3390/s21165300.
22. Triplett, W. J. (2022). Addressing Human Factors in Cybersecurity Leadership. Journal of Cybersecurity and Privacy, 2(3), 573. https://doi.org/10.3390/jcp2030029.
23. Salik, H. (2022). Offensive Cyber Operations: Failure to Dissuade Nation-State Adversaries in Cyberspace. ProQuest Dissertations Publishing, University of the Cumberlands. (29397595).
24. Rahman, T., Rohan, R., Pal, D., & Kanthamanon, P. (2021). Human Factors in Cybersecurity: A Scoping Review. In The 12th International Conference on Advances in Information Technology (IAIT2021), June 29–July 01, 2021, Bangkok, Thailand (pp. 1-11). ACM. https://doi.org/10.1145/3468784.3468789
25. Sutter, O. W. (2020). *The cyber profile: Determining human behavior through cyber-actions*. (Doctoral dissertation). Capitol Technology University, ProQuest Dissertations Publishing. (29257172).
26. Tyworth, M., Giacobe, N. A., Mancuso, V. F., McNeese, M. D., & Hall, D. L. (2013). A human-in-the-loop approach to understanding situation awareness in cyber defence analysis. EAI Endorsed Transactions on Security and Safety, 1(2). https://doi.org/10.4108/trans.sesa.01-06.2013.e6.
27. Elastic. Filebeat module: o365. Elastic.co. Available online: https://www.elastic.co/guide/en/beats/filebeat/current/filebeat-module-o365.html (accessed on 31 May 2023).

28. Wells, J. A., LaFon, D. S., & Gratian, M. (2021). Assessing the Credibility of Cyber Adversaries. International Journal of Cybersecurity Intelligence & Cybercrime, 4(1), 3-24. https://www.doi.org/10.52306/040102
29. Dalal, R. S., Howard, D. J., Bennett, R. J., Posey, C., Zaccaro, S. J., & others. (2022). Organizational science and cybersecurity: abundant opportunities for research at the interface. Journal of Business and Psychology, 37(1), 1-29. https://doi.org/10.1007/s10869-021-09732-9.
30. Kioskli, K., & Polemi, N. (2020). Psychosocial approach to cyber threat intelligence. International Journal of Chaotic Computing, 7(1), 159-165.
31. Singh, T. (2021). The Role of Stress among Cybersecurity Professionals (Doctoral dissertation, The University of Alabama). ProQuest Dissertations Publishing. (28546079).
32. Clapper, J., Lettre, M., & Rogers, M. S. (2017, January 31). Foreign Cyber Threats to the United States. Hampton Roads International Security Quarterly, 1.
33. McCall, G. C. Jr. (2022). Exploring a Cyber Threat Intelligence (CTI) Approach in the Thwarting of Adversary Attacks: An Exploratory Case Study (Doctoral dissertation, Northcentral University). ProQuest Dissertations Publishing. (28968146).
34. Pangsuban, P., Nilsook, P., & Wannapiroon, P. (2020). Real-time Risk Assessment for Information System with CICIDS2017 Dataset Using Machine Learning. International Journal of Machine Learning and Computing, 10(3), 538-543.
35. Scott, J., & Kyobe, M. (2021). *Trends in Cybersecurity Management Issues Related to Human Behaviour and Machine Learning*. In International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. n.a.). IEEE.
36. Parsons, K., McCormac, A., Butavicius, M., & Ferguson, L. (2010). *Human factors and information security: Individual, culture and security environment*. Defense Science and Technology Organization, Commonwealth of Australia.

Appendix A

Demographic Distribution Summary of Known Data Breaches

In this study, a total of sixty-nine datasets were analyzed across various industries, focusing on data breaches related to email addresses where known organizational emails were compromised. The results revealed the following distribution:

- Finance and Insurance: 35%
- Healthcare: 22%
- Technology: 16%
- Retail: 12%
- Manufacturing: 10%
- Other industries: 5%

Distribution of Breaches Across Industries:

1. Business and Data Services: Adapt, Apollo, B2B USA Businesses, Data-Leads, Elasticsearch Instance of Sales Leads on AWS, Exactis, Factual, Lead Hunter, NetProspex, Verifications.io, Whitepages

2. Technology Platforms: Adobe, Animoto, Bitly, Canva, Chegg, Disqus, Dropbox, Edmodo, Emotet, Epik, LinkedIn, LiveAuctioneers, LiveJournal, Modern Business Solutions, mSpy, MyFitnessPal, Nitro, QuestionPro, ShareThis, SlideTeam, Stratfor, Ticketfly, Twitter, Zomato, Zynga
3. Retail and E-commerce: Bonobos, CafePress, Covve, Drizly, EatStreet, Evite, Fling, Gaadi, Gravatar, HauteLook, Houzz, Justdate.com, Minted, MMG Fusion, River City Media Spam List
4. Automotive: Audi
5. Gaming and Entertainment: ArmorGames, Digimon, Forbes, Straffic, Zynga
6. Adult Content: Fling
7. Health and Fitness: MyFitnessPal
8. Education: Chegg, Edmodo
9. Social Media and Networking: Disqus, Gravatar, LinkedIn, LiveJournal, Twitter
10. Food and Beverage: Drizly, EatStreet, Zomato

Timeline of Data Breaches:

- 2011: Fling, Stratfor, LinkedIn
- 2012: Adobe, Dropbox, Disqus, LinkedIn, Twitter
- 2013: Adobe, Gravatar
- 2014: Bitly, Digimon, Forbes, LiveJournal
- 2015: Gaadi, mSpy
- 2016: Anti Public Combo List, Exploit.In, NetProspex, Lead Hunter, Modern Business Solutions
- 2017: Edmodo, Factual, Onliner Spambot, River City Media Spam List, Trik Spam Botnet, Zomato
- 2018: Adapt, Animoto, Apollo, Bitly, Chegg, Covve, Dropbox, HauteLook, Houzz, MyFitnessPal, Straffic
- 2019: CafePress, Canva, EatStreet, Elasticsearch Instance of Sales Leads on AWS, Evite, Exactis, LiveAuctioneers, River City Media Spam List, ShareThis, Verifications.io, Whitepages
- 2020: Audi, Bonobos, Covve, Data Enrichment Exposure From PDL Customer, Drizly, HauteLook, LiveAuctioneers, Nitro, ParkMobile
- 2021: B2B USA Businesses, Data-Leads, Epik, MeetMindful, Minted, MMG Fusion, QuestionPro, SlideTeam

Size of Data Breaches:

- Largest breaches (over 100 million records): Adobe, Canva, Collection #1, Evite, Exactis, LinkedIn, River City Media Spam List, Verifications.io
- Medium-sized breaches (10 million - 100 million records): Adapt, Apollo, Bitly, CafePress, Chegg, Disqus, Dropbox, Edmodo, Houzz, MyFitnessPal, NetProspex, Nitro, ParkMobile, ShareThis, Zomato, Zynga
- Smaller breaches (1 million - 10 million records): Animoto, Audi, Bonobos, Covve, Data-Leads, Drizly, EatStreet, Emotet, Epik, Factual, Fling, Forbes, Gaadi, Gravatar, HauteLook, Lead Hunter, LiveAuctioneers, LiveJournal, mSpy, Minted, MMG Fusion, Modern Business Solutions, QuestionPro, SlideTeam, Stratfor, Ticketfly, Twitter
- Minor breaches (under 1 million records): ArmorGames, B2B USA Businesses, Digimon, ElasticSearch Instance of Sales Leads on AWS, Onliner Spambot, Trik Spam Botnet, Whitepages

Types of Data Compromised:

 doi:10.20944/preprints202306.1041.v1

- Email addresses
- Hashed and plaintext passwords
- Usernames
- Names
- Physical addresses
- Phone numbers
- Date of birth
- Social media profiles
- Personal preferences
- Payment information
- Health and fitness data

Potential TTPs (Tactics, Techniques, and Procedures) Employed in Attacks:

- Phishing campaigns
- Social engineering
- Exploiting unpatched vulnerabilities
- Credential stuffing
- Password spraying
- Brute force attacks
- SQL injection
- Malware infections
- Third-party service compromises
- Insider threats
- Advanced Persistent Threats (APTs)
- Supply chain attacks

Appendix B

A list and description of all sixty-nine public data breaches related to compromised email addresses for the organizations within the study.

1. **Adapt**: A business data provider suffered a data breach in 2018, exposing approximately 9.3 million records, including email addresses, personal information, and business data.
2. **Adobe**: In 2013, Adobe experienced a significant data breach, compromising about 153 million user records, including email addresses, passwords, and password hints.
3. **Animoto**: A video creation platform breached in 2018, compromising 25 million user records, including email addresses and hashed passwords.
4. **Anti Public Combo List**: A compilation of data breaches discovered in 2016, compromising over 458 million email addresses, usernames, and plaintext passwords.
5. **Apollo**: A sales engagement platform suffered a data breach in 2018, compromising 125 million records, including email addresses, names, and job titles.
6. **Audi**: A 2020 data breach at Audi and Volkswagen impacted 3.3 million customers, exposing email addresses, phone numbers, and vehicle identification numbers (VINs).

 doi:10.20944/preprints202306.1041.v1

7. **B2B USA Businesses**: In 2021, a database containing 63 million records from various B2B USA companies was leaked, including email addresses and other personal information.
8. **Bitly**: The URL shortening service experienced a data breach in 2014, leading to the compromise of email addresses, encrypted passwords, and API keys.
9. **Bonobos**: The men's clothing retailer suffered a data breach in 2020, exposing approximately 70GB of data, including 7 million email addresses and other personal information.
10. **CafePress**: In 2019, CafePress experienced a data breach, compromising 23 million user records, including email addresses and password hashes.
11. **Canva**: A graphic design platform breached in 2019, exposing 137 million user records, including email addresses and bcrypt-hashed passwords.
12. **Chegg**: An education technology company suffered a data breach in 2018, compromising 40 million records, including email addresses, usernames, and hashed passwords.
13. **Cit0day**: In 2020, a collection of 23,000 breached databases was leaked, containing billions of records, including email addresses, usernames, and plaintext passwords.
14. **Collection #1**: A massive data breach compilation discovered in 2019, consisting of over 770 million unique email addresses and over 21 million unique passwords.
15. **CouponMom-ArmorGames**: A data breach in 2020 affected both CouponMom and ArmorGames, compromising 11 million records, including email addresses and plaintext passwords.
16. **Covve**: A data breach in 2020 exposed the records of 22 million users, including email addresses, names, phone numbers, and LinkedIn profiles.
17. **Data Enrichment Exposure From PDL Customer**: In 2019, a security lapse at People Data Labs (PDL) exposed 622 million records, including email addresses and other personal information.
18. **Data-Leads**: In 2021, a data breach compromised 63 million records from various B2B companies, including email addresses, names, and phone numbers.
19. **Digimon**: An unofficial forum for Digimon fans was hacked in 2014, compromising 4.9 million records, including email addresses, usernames, and IP addresses.
20. **Disqus**: A blog comment hosting service breached in 2012, resulting in the exposure of 17.5 million user records, including email addresses, usernames, and hashed passwords.
21. **Drizly**: An alcohol delivery platform experienced a data breach in 2020, compromising 2.5 million user records, including email addresses, hashed passwords, and personal information.
22. **Dropbox**: In 2012, Dropbox suffered a data breach, resulting in the exposure of 68 million user records, including email addresses and hashed passwords.
23. **EatStreet**: A food delivery platform breached in 2019, compromising 6 million user records, including email addresses, hashed passwords, and personal information.
24. **Edmodo**: An educational platform experienced a data breach in 2017, exposing 77 million user records, including email addresses, usernames, and bcrypt-hashed passwords.
25. **Elasticsearch Instance of Sales Leads on AWS**: In 2019, an unprotected Elasticsearch instance exposed 60 million sales leads, including email addresses and other personal information.
26. **Emotet**: A notorious botnet and malware family involved in multiple phishing campaigns targeting email addresses, banking credentials, and other personal information.
27. **Epik**: A domain registrar and web hosting company suffered a data breach in 2021, compromising email addresses, account credentials, and customer records.

 doi:10.20944/preprints202306.1041.v1

28. **Evite**: A social planning and invitation platform breached in 2019, leading to the exposure of 101 million user records, including email addresses, plaintext passwords, and personal information.
29. **Exactis**: A data aggregator experienced a data breach in 2018, compromising 340 million records, including email addresses, phone numbers, and other personal information.
30. **Exploit.In**: A forum for hackers, which in 2016 released a database containing 593 million email addresses and plaintext passwords from multiple data breaches.
31. **Factual**: A location data company suffered a data breach in 2017, compromising 2.5 million user records, including email addresses, hashed passwords, and personal information.
32. **Fling**: An adult dating website experienced a data breach in 2011, exposing 40 million user records, including email addresses, usernames, and plaintext passwords.
33. **Forbes**: The media company suffered a data breach in 2014, compromising 1 million user records, including email addresses, usernames, and hashed passwords.
34. **Gaadi**: An Indian car research platform experienced a data breach in 2015, exposing 2.2 million user records, including email addresses, usernames, and hashed passwords.
35. **Gravatar**: In 2013, a security researcher discovered a vulnerability in Gravatar that could potentially expose user email addresses, but no data breach was reported.
36. **HauteLook**: A fashion retailer suffered a data breach in 2018, compromising 28 million user records, including email addresses, bcrypt-hashed passwords, and personal information.
37. **Houzz**: A home design platform experienced a data breach in 2018, exposing 48 million user records, including email addresses, usernames, and hashed passwords.
38. **Justdate.com**: A dating platform suffered a data breach in 2017, compromising 1.7 million user records, including email addresses, bcrypt-hashed passwords, and personal information.
39. **Kayo.moe Credential Stuffing List**: In 2018, a collection of 42.5 million email addresses and plaintext passwords from various sources was discovered, potentially used for credential stuffing attacks.
40. **Lead Hunter**: A data breach in 2016 affected the sales lead generation platform, compromising 68 million user records, including email addresses, hashed passwords, and personal information.
41. **LinkedIn**: In 2012, LinkedIn experienced a data breach, compromising 165 million user records, including email addresses and hashed passwords. A separate incident in 2021 involved scraped data from around 500 million LinkedIn users, including email addresses, though this was not a direct breach of their systems.
42. **LiveAuctioneers**: An online auction platform breached in 2020, leading to the exposure of 3.4 million user records, including email addresses, hashed passwords, and personal information.
43. **LiveJournal**: A blogging platform experienced a data breach in 2014, compromising 26 million user records, including email addresses, plaintext passwords, and usernames.
44. **MeetMindful**: A dating platform suffered a data breach in 2021, exposing 2.3 million user records, including email addresses, names, and location data.
45. **Minted**: An online marketplace for independent artists experienced a data breach in 2020, compromising 5 million user records, including email addresses, hashed passwords, and personal information.
46. **MMG Fusion**: A dental marketing software provider suffered a data breach in 2021, exposing 2.6 million user records, including email addresses and other personal information.
47. **Modern Business Solutions**: A data management and monetization company experienced a data breach in 2016, compromising 58 million user records, including email addresses, IP addresses, and personal information.
48. **mSpy**: A mobile monitoring and parental control software provider suffered a data breach in 2015, exposing 4 million user records, including email addresses, encrypted passwords, and payment details.

 doi:10.20944/preprints202306.1041.v1

49. **MyFitnessPal**: A fitness and nutrition app experienced a data breach in 2018, compromising 150 million user records, including email addresses, hashed passwords, and usernames.
50. **NetGalley**: An online book review platform suffered a data breach in 2020, exposing email addresses, names, usernames, and hashed passwords.
51. **NetProspex**: A sales lead generation company experienced a data breach in 2016, compromising 33 million user records, including email addresses, names, job titles, and company information.
52. **Nitro**: A document management and productivity software provider suffered a data breach in 2020, exposing 70 million user records, including email addresses, names, and hashed passwords.
53. **Onliner Spambot**: In 2017, a spambot campaign is known as Onliner Spambot was discovered, compromising 711 million email addresses, along with usernames and passwords, used for sending spam and infecting systems with malware.
54. **ParkMobile**: A parking app experienced a data breach in 2021, compromising 21 million user records, including email addresses, names, and hashed passwords.
55. **QuestionPro**: An online survey platform suffered a data breach in 2021, exposing 198 million user records, including email addresses, names, and hashed passwords.
56. **River City Media Spam List**: In 2017, a data breach involving River City Media, a spamming organization, exposed 1.34 billion email addresses, names, and other personal information.
57. **ShareThis**: A social sharing platform experienced a data breach in 2018, compromising 41 million user records, including email addresses, hashed passwords, and usernames.
58. **SlideTeam**: A presentation template provider suffered a data breach in 2021, exposing 1.4 million user records, including email addresses, names, and bcrypt-hashed passwords.
59. **Straffic**: A botnet involved in various phishing campaigns was discovered in 2021, potentially compromising millions of email addresses, banking credentials, and others.
60. **Stratfor**: A global intelligence company experienced a data breach in 2011, compromising 860,000 user records, including email addresses, usernames, and hashed passwords.
61. **Ticketfly**: An event ticketing platform suffered a data breach in 2018, exposing 27 million user records, including email addresses, names, and phone numbers.
62. **Trik** Spam Botnet: A malware botnet discovered in 2017, compromising 43 million email addresses and plaintext passwords, used for sending spam and infecting systems with additional malware.
63. **Twitter**: In 2018, Twitter advised its 330 million users to change their passwords due to a bug that stored plaintext passwords in an internal log. However, there was no confirmed data breach or unauthorized access.
64. **Unverified Data Source**: A collection of compromised records discovered in 2019 containing over 62 million email addresses and plaintext passwords from various sources, with no specific attribution to a single breach.
65. **Verifications.io**: A data validation service experienced a data breach in 2019, exposing 763 million records, including email addresses, phone numbers, and other personal information.
66. **Whitepages**: In 2019, an unprotected Elasticsearch database exposed 22 million Whitepages records, including email addresses, names, and phone numbers. However, this was not a direct breach of Whitepages systems.
67. **Youve Been Scraped/You've Been Scraped**: These incidents refer to data scraping, where publicly available information is collected from websites without authorization. Email addresses are often a target in these situations, but specific breaches are difficult to pinpoint.
68. **Zomato**: An Indian food delivery platform suffered a data breach in 2017, compromising 17 million user records, including email addresses and hashed passwords.

 doi:10.20944/preprints202306.1041.v1

69. **Zynga**: A mobile gaming company experienced a data breach in 2019, exposing 218 million user records, including email addresses, usernames, and hashed passwords.

Appendix C

Advanced Persistent Threats (APTs) Groups Associated with the Known Public Data Breaches

While it is difficult to definitively attribute the data breaches to specific APTs, some of the breaches have been potentially linked or suspected to be the work of APT groups. Here are a few examples:

1. LinkedIn (2012): The LinkedIn data breach, where approximately 165 million user accounts were compromised, has been attributed to a Russia-based hacker group known as APT28 or Fancy Bear. They are believed to have ties to the Russian government.
2. Twitter (2013): The Twitter breach, in which around 45,000 accounts were compromised, has been suspected to be the work of the Syrian Electronic Army (SEA), an APT group with connections to the Syrian government.
3. Dropbox (2012): The Dropbox breach, which affected nearly 68 million users, has been attributed to a group known as APT29 or Cozy Bear. This group is also believed to have ties to the Russian government.
4. Emotet (2014-present): Emotet is a sophisticated malware strain and botnet known for distributing banking Trojans and ransomware. Although not directly attributed to a specific nation-state APT, it has been linked to various cybercrime groups and is considered an advanced threat due to its persistence and evolving nature.
5. Stratfor (2011): The breach of the global intelligence company Stratfor, where around 860,000 users' data was compromised, was claimed by the hacktivist group Anonymous. However, some cybersecurity researchers have suggested that the attack might have been supported by a nation-state APT group due to the level of sophistication.
6. Collection #1 (2019): While direct attribution is not available, the sheer scale of this massive data breach compilation suggests the involvement of advanced threat actors. It is possible that multiple APT groups and cybercriminal organizations contributed to or took advantage of the compromised data.
7. Adobe (2013): The breach is suspected to be the work of an APT group called "PawnStorm" (also known as APT28 or Fancy Bear), which has been linked to Russian intelligence agencies. This group is notorious for targeting high-profile organizations and using spear-phishing campaigns to infiltrate networks.
8. mSpy (2015): The breach was initially attributed to an unknown hacking group. However, further analysis linked the breach to the Chinese APT group called "APT3" or "Buckeye." This group is known for targeting high-profile organizations in various industries, primarily to gain intellectual property and sensitive information.
9. Onliner Spambot (2017): While not directly linked to a specific APT group, it can be associated with advanced persistent cybercriminal campaigns. These campaigns often involve the use of large-scale spamming operations and the distribution of sophisticated malware such as banking Trojans and ransomware.
10. Cit0day (2020) is a collection of 23,000 breached databases containing billions of records. Although difficult to attribute to a specific APT group, the scale implies multiple hacking groups' involvement. Various TTPs, such as phishing, credential stuffing, and exploiting web vulnerabilities, were likely employed in the breaches.

 doi:10.20944/preprints202306.1041.v1

11. SolarWinds (2020): This high-profile supply chain attack compromised numerous government and private organizations. The breach has been attributed to a Russian APT group known as APT29, also referred to as Cozy Bear or The Dukes. They are believed to have ties to Russia's foreign intelligence service, the SVR.
12. Equifax (2017): The massive breach of the credit reporting agency, which affected around 147 million users, has been attributed to the Chinese APT group called APT10 or Menupass. The group is known for targeting large organizations and is believed to have ties to China's Ministry of State Security.
13. WannaCry (2017): This widespread ransomware attack affected organizations and users globally. The attack has been attributed to the North Korean APT group known as Lazarus Group or Hidden Cobra. They are believed to be linked to the North Korean government and have been involved in several high-profile cyberattacks.
14. NotPetya (2017): This destructive malware attack targeted organizations primarily in Ukraine but also affected global businesses. The NotPetya attack has been attributed to the Russian APT group Sandworm Team, also known as Voodoo Bear or TeleBots. They are believed to be connected to Russia's military intelligence agency, the GRU.
15. Zomato (2017): Although direct attribution is not available, the scale and nature of the attack suggest that an advanced cybercriminal organization or APT group may have been involved. The breach resulted in the compromise of 17 million user records, including email addresses and hashed passwords.
16. Zynga (2019): The breach affecting 218 million user records has been attributed to a well-known cybercriminal known as Gnosticplayers. While not an APT group, Gnosticplayers is responsible for a series of large-scale data breaches, indicating a high level of sophistication and persistence in their operations.
17. MyFitnessPal (2018): The breach of MyFitnessPal, which compromised 150 million user records, was attributed to a group of prolific hackers known as "Magecart." Although typically known for its attacks on e-commerce sites, the group's scale and sophistication suggest it might operate at a level comparable to a nation-state APT.
18. Houzz (2018): Houzz's data breach exposed 48 million user records. While a specific APT group hasn't been linked to this incident, the scale and nature of the data compromised suggest the involvement of a highly organized and possibly state-sponsored group.
19. Verifications.io (2019): This incident exposed 763 million records, making it one of the most extensive collections of public data breaches. While the actual breach hasn't been linked to a specific APT, the scale and type of data suggests the involvement of advanced and persistent threat actors.
20. Ticketfly (2018): While no specific APT group was attributed to the breach, the nature of the attack (a defacement of the website coupled with data exfiltration) suggests the involvement of a sophisticated threat actor, possibly with the characteristics of an APT.

 doi:10.20944/preprints202306.1041.v1

Appendix D

Different types of Microsoft 365 accounts observed in the study.

Can users log into the account type directly?

1. UserMailbox: Yes, users can log in directly to their UserMailbox. This is the primary account type used by individuals to access their emails, calendar, contacts, and other Microsoft 365 services.
2. SharedMailbox: No, users cannot log in directly to a shared mailbox. They need to have their own individual UserMailbox and be granted access to the shared mailbox. They can then access it via their own account.
3. GAL Contact: No, users cannot log in directly to a GAL (Global Address List) Contact. These are just contact entries in the address book and do not have any login credentials associated with them.
4. Room Mailbox: No, users cannot log in directly to a Room Mailbox. A room mailbox is a resource mailbox that represents a meeting space, like a conference room. Users can book the room through their own UserMailbox but cannot access the room mailbox itself.
5. Health Mailbox: No, users cannot log in directly to a Health Mailbox. These mailboxes are used by Microsoft Exchange Server to monitor and test the health of the server. They are not meant for direct user access.
6. Team Mailbox: No, users cannot log in directly to a Team Mailbox. A team mailbox is associated with a Microsoft Teams team and its channels. Users need to have their own individual UserMailbox and be a member of the relevant team to access the team mailbox.
7. Alias: No, users cannot log in directly to an Alias. An alias is an additional email address associated with a UserMailbox that can be used to send and receive email. It is not a separate account and cannot be accessed independently.
8. Equipment Mailbox: No, users cannot log in directly to an Equipment Mailbox. An equipment mailbox is a resource mailbox that represents a piece of equipment, like a projector or a company car. Users can book the equipment through their own UserMailbox but cannot access the equipment mailbox itself.
9. System.Object: This is not a type of Microsoft 365 mailbox account. It appears to be a generic object reference in a programming language or script, and therefore cannot be logged into directly.
10. No Mailbox: No, users cannot log in directly to a "No Mailbox" account, as it indicates that there is no mailbox associated with the user or object in question. Without a mailbox, there is no account for a user to log into.
11. NoUser: No, users cannot log in directly to a "NoUser" account. This term typically refers to an account or object that has not been assigned a user or that does not have a mailbox associated with it. There is no account to log into in this case.
12. Sync: This term is not a specific type of Microsoft 365 mailbox account. It might refer to the synchronization process between on-premises Active Directory and Azure Active Directory, or other data synchronization scenarios. As such, users cannot log into a "Sync" account, as it does not represent a mailbox or user account.
13. Alias: No, users cannot log in directly to an Alias. An alias is an additional email address associated with a UserMailbox that can be used to send and receive email. It is not a separate account and cannot be accessed independently. Users need to log in to their primary UserMailbox to access emails sent to their alias.