IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 May 2024, accepted 23 June 2024, date of publication 4 July 2024, date of current version 12 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3423651

RESEARCH ARTICLE

Assessing Data Breach Factors Through Modern Crime Theory: A Structural Equation Modeling Approach

NARJISSE NEJJARI[1,2], KARIM ZKIK[3], HICHAM HAMMOUCHI[1,2], (Member, IEEE), MOUNIR GHOGHO[2,4], (Fellow, IEEE), AND HOUDA BENBRAHIM[1]

[1]IRDA, Rabat IT Center, ENSIAS, Mohammed V University, Rabat 10000, Morocco
[2]TICLab, International University of Rabat, Rabat 11103, Morocco
[3]ESAIP, Ecole d'Ingenieur, CERADE Angers, 49124 Saint-Barthélemy-d'Anjou, France
[4]Faculty of Engineering, University of Leeds, LS2 9JT Leeds, U.K.

Corresponding author: Narjisse Nejjari (nejjari.narjisse@gmail.com)

ABSTRACT Research into data security often emphasizes the need to understand the factors linked to security breaches, aiming to prevent future information security incidents. The advancement of digital technology has made safeguarding an organization's sensitive data more complex. Despite the growth of research in data security, there's currently a shortage of studies that specifically investigate the factors contributing to information security breaches in organizations. Previous studies have primarily examined the security posture of companies and organizations, focusing on the breach type and location. However, few studies have explored external factors that may contribute to organizations' vulnerability to information security breaches. The current study addresses this gap in the literature by integrating modern crime theory (MCT) to investigate the exogenous factors influencing the victimization of public and private organizations to data breach incidents. We use insights from crime theories and information about organizations' technical, organizational, and financial aspects to investigate how attractiveness, visibility, and guardianship affect the likelihood of data breaches. We build a theoretical model to explore the relationship between these factors as independent predictors of data breaches. A covariance-based structural equation modeling (CB-SEM) based framework is developed to conduct a comprehensive examination of the dynamics within the context of cybercrime. Through the examination of collected data from 4,868 organizations, this study demonstrates a good fit of the hypothesized model to the data, supporting the validity of the proposed constructs. The results of this study validate the use of MCT in the study of information security breach, and enable the identification of the major exogenous factors influencing data breaches, such as the attractiveness of valuable data and effectiveness of guardianship measures.

INDEX TERMS Information security breach, data breach, crime theory, covariance-based structural equation modeling.

I. INTRODUCTION

A. BACKGROUND AND CONTEXT

Over the past two decades, organizations and institutions have encountered significant incidents involving data breaches. Data fraud and theft have emerged as prominent global risks, with the year 2023 witnessing a particularly devastating impact on organizations worldwide, especially in the United States. The IBM Cost of a Data Breach Report 2023 [1] reveals that the average cost of a breach during this period amounted to approximately $4.45 million. Additionally, the 2024 Thales Data Threat Report, conducted by the International Data Corporation (IDC) [2], indicates that nearly two-thirds of US companies encountered at least one data breach in recent

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

times. Thus, data breaches are considered as a serious threat that disrupts businesses and make damages to their assets and reputation, since they cost firms and companies huge financial loss. As data breaches become more sophisticated and challenging to manage, organizations face an urgent need to collaborate with academic researchers and experts, as emphasized by the US federal agency [3], in order to assess data breach incidents effectively and mitigate their damaging consequences.

Data breaches occur as a result of many hacking and phishing incidents such as unauthorized access, theft, loss of computer, improper disclosures, targeting laptops, desktop computers, hacking portable devices, emails, network servers, and other IT devices. Numerous studies have explored the security posture of companies and organizations, primarily focusing on the nature and location of breaches. However, limited attention has been given to external factors that may contribute to their susceptibility to victimization. To the best of our knowledge, none of these studies combined the use of modern crime theory to study external factors that may contribute to the victimization of public and private organizations to data breach incidents.

In cybercrime, traditional criminological theories often need to be adapted and extended to capture the unique dynamics and characteristics of the digital environment. The main objective of this research is to advance the understanding of factors associated with organizations' vulnerability to information security breaches. Specifically, we investigate the extent to which constructs derived from crime theory, namely attractiveness, visibility, and guardianship, can serve as predictors of data breaches. This research paper aims to investigate the phenomenon of victimization in the context of cyberspace, specifically focusing on data breaches, by using routine activity theory as theoretical framework. These theoriy provides valuable insights into the situational characteristics that enhance crime opportunities and can guide the development of effective preventive measures.

B. RESEARCH OBJECTIVES

Building upon these theoretical frameworks, we aim to construct a comprehensive theoretical model that examines the relationships between organization's attractiveness, visibility, and guardianship as independent predictors of data breaches. Our study focuses on a sample of 4,868 organizations operating in the United States (US) between 2018 and 2020. The dataset includes measures assessing the risk of victimization for organizations. To achieve this, we collect multivariate data from organizations that have experienced data breaches and those that have not, then analyse them using the lens of crime theory. By adopting this perspective, our aim is to provide organizations with recommendations and support to proactively prevent potential breaches, placing emphasis on the victim rather than on the crime itself.

The key contributions of our manuscript can be outlined as follows:

- We collect multivariate data about victim and non-victim organizations, using web-scraping techniques.
- We design a collection of indicators that capture organizational, financial and technical items.
- We examine the use of crime theories in the study of security breach in cyberspace.
- We propose a theoretical framework that examines the relationship between organization's attractiveness, visibility, and guardianship as independent predictors of data breaches on organizations.
- We examine the use of covariance-based structural equation modeling in the study of factors related to security breach.
- We empirically assess the proposed model's validity and reliability and we highlight the most influential indicators when dealing with information security breaches.

The key objective of our work is to validate the use of MCT in the study of information security breach and enable organizations to comprehend the impact of various factors that may contribute to their victimization. Then, guide their efforts in implementing targeted security measures.

This research paper is structured into six main sections. Section II provides a comprehensive review of pertinent literature. In section III, we present the research problem and establish the research hypothesis. In Section IV, we present the research methodology, detailing our approach to addressing the research questions. Section V, explains the results and findings derived from the data analysis. Section VI, the discussion section, we critically examine and interpret the implications of the findings, offering valuable insights for organizations, we present also the study's limitations and we propose potential avenues for future research. Finally, we conclude the paper.

II. RELATED WORK

This section examines the literature regarding data breach incidents encountered by organizations. We focus on determining the risk of victimization within the contextual framework of crime theories. While modern crime theories lay the foundation for this empirical study, it's noteworthy that prior research often refers to ''lifestyle'' or ''routine activity'' theories without explicitly stating how these behaviors may lead to victimization in data security. To bridge this gap,we explore how crime theories relate to the vulnerability of organizations to experiencing data breaches.

Through an extensive review of pertinent studies, our exploration aims to demonstrate how crime theories can be applied to gain deeper insights and address the challenges posed by cyber threats to data security within organizations.

A. ASSESSING THE FACTORS INFLUENCING DATA BREACH INCIDENTS

A considerable research work has been conducted in the field of information security risk (ISR) literature. Li and Li conducted an insightful analysis using CiteSpace-based

visualization techniques to map knowledge structures and illuminate the ISR landscape [4]. Despite their valuable contributions, their analysis does not explicitly pinpoint research gaps. On a related note, Mayer et al. delve into the intention-behavior gap, shedding light on motivators and obstacles while providing practical insights for interventions [5].

Within regulatory contexts, Ashraf [6] explore the implications of the Securities and Exchange Commission's (SEC) guidance on cyber risk factor disclosure, examining how peer breaches influence the cyber risk disclosures of non-breached firms. Additionally, another study by Bouveret [7], introduces a quantitative framework for cyber risk assessment, providing a holistic perspective that is applied to cross-country data. In [8], authors examine, from a spatio-temporal perspective, the factors and the context associated to data breaches targeting healthcare sector in the United States.

Several studies, including those by Barati and Yankson [9], Fang et al. [10], Sun et al. [11], and Zhang and Chen [12], seek to enhance data breach prediction. Barati and Yankson propose a predictive model utilizing historical data to estimate breach likelihood and size [9]. Addressing sparsity in unstructured data, Huang et al. introduce the Adaptive Weighted Graph Walk model (AGW) [13].

In enterprise-level breach prediction, Fang et al. present a statistical framework leveraging time series interdependencies, outperforming benchmarks [10]. Sun et al. propose an advanced approach combining a hurdle-Poisson model and a mixed non-parametric kernel distribution [11]. Zhang and Chen develop a hybrid model for big data breach prediction, excelling in accuracy and efficiency [12].

Bouveret question the escalation of data breaches, finding stable breach frequency and increased sizes [7]. The study identifies organizational traits predictive of breach size and frequency. Examining the effects of privacy breaches on market value, research indicates a temporary and significant reduction, especially for larger companies [14], emphasizing the importance of both privacy and security measures for maintaining profitability.

B. MODERN CRIME THEORY: BACKGROUND

In our quest to understand cyber-crimes, we turn to theories, like modern crime theory, to guide our efforts. Theory, in research, serves as a set of ideas that explain real-world events. Evaluating theory involves making sure it makes sense, fits the context, is straightforward, testable, supported by evidence, and has practical applications. Modern crime theory is our theoretical framework for understanding factors associated with data breach incidents in organizations.

This section provides an overview of theory, introduces modern crime theory, discusses the concept of victimology, and presents many related concepts including routine activity theory, Deviant Place Theory, deterrence theory, and rational choice theory, all in the context of understanding data violations incidents. These theories offer valuable insights into the motivations and decisions behind criminal actions. Here are some key theories that are particularly relevant to our study:

1) ROUTINE ACTIVITY THEORY

This theory provides insights into the underlying reasons for the occurrence of crimes [15], [16], [17]. It highlights the fact that offenders often make rational choices when deciding to commit a crime. For a criminal act to take place, three critical elements must converge simultaneously: a desirable target, a lack of effective guardianship, and a motivated offender, all within the same time and place.

2) DEVIANT PLACE THEORY

This theory sheds light on the circumstances that increase an individual's vulnerability to becoming a victim of crime [18], [19], [20]. It emphasizes that individuals are more likely to fall prey to criminal activities when they find themselves in environments characterized by risk and criminal behavior.

3) DETERRENCE THEORY

Deterrence theory delves into the factors influencing an individual's decision to comply with or violate the law [21], [22]. It posits that the perceived consequences of an action play a crucial role in deterring or encouraging criminal behavior. People tend to weigh their own experiences and awareness of potential punishments when making these decisions.

4) RATIONAL CHOICE THEORY

Grounded in the principle of expected utility, this theory suggests that individuals make decisions by carefully assessing the benefits in comparison to the losses [23], [24], [25]. People are more inclined to follow the law when they perceive the advantages as outweighing the disadvantages of engaging in criminal behavior.

These theories collectively provide a comprehensive framework for understanding the dynamics of criminal activities, shedding light on the motivations of offenders, the vulnerabilities of potential victims, and the decision-making processes that drive criminal actions. They are instrumental in our exploration of modern crime theory's application to the study of cyber-crimes, particularly data breaches.

C. BRIDGING MODERN CRIME THEORY WITH CYBERCRIME

Cyber criminologists and cyber security scholars have increasingly applied modern crime theory to gain valuable insights into the intricacies of digital crimes and strategies to safeguard against them. Kennedy et al. [26] highlight the necessity for criminological frameworks in the automotive industry's cybersecurity, introducing a security pattern model. However, their conceptual work lacks empirical research. in [27], Nejjari et al. introduce the use of crime theory in the context of data security. Holt et al. [28] distinguish ideologically motivated cyberattacks, addressing a gap in

online ideological attack research, employing routine activity theory and Subcultural Theory.

Leukfeldt and Yar [29] critically examines the challenges of applying routine activity theory (RAT) to cybercrimes, emphasizing the need for further investigation. Ngo et al. [30] explore challenges in identifying juvenile hacking behaviors globally, revealing associations between factors like gender, age, self-control, and detection of hacking behaviors.

Holt et al. [28] explore the connection between routine activities and malware infection indicators using Routine Activities Theory (RAT). Reyns and Henson [31]delve into factors contributing to online victimization, grounded in routine activity theory and drawing on data from the Canadian General Social Survey.

Shifting focus to college students, Choi [32] investigate the interplay between computer crime victimization and introduce a novel model inspired by lifestyle-exposure theory and routine activities theory, employing a self-report survey and structural equation modeling (SEM) analysis. In malware infections research, Holt et al. [33] investigate broader factors at the national level, using a routine activities framework. In web defacement research, Holt et al. [34] explore motives in the Netherlands, utilizing routine activity theory, and Howell et al. [35] evaluate the predictive power of Routine Activities Theory (RAT) in forecasting the frequency of website defacement across countries.

Jacques and Bonomo [36] offer a unique perspective by exploring crime prevention strategies from the standpoint of offenders. Maimon and Louderback [37] work provides a nuanced understanding of computer-focused crimes against a university computer network. In preventive behaviors research, Reyns et al. [38] investigate factors associated with online victimization. Shifting focus to college students, Reyns et al. [39] contribute empirical evidence by investigating predictors of online victimization. Reyns et al.'s [40]) work examines the concept of guardianship concerning cyberstalking victimization, employing routine activity theory. In the field of cyber terrorism, Holt et al. [41] apply routine activities theory to explore the characteristics of jihadi-associated cyberattacks in the United States. According to Smirnova and Holt [42] examination of the geographical distribution of victim nations in stolen data markets, variations in victim nations are revealed, which sheds light on how actors make decisions in these markets based on perceived rewards and risks.

D. MODERN CRIME THEORY APPLIED TO STUDY INFORMATION SECURITY BREACH

Routine activity theory (RAT) is often considered highly applicable when assessing organizational vulnerability to data breaches. routine activity theory, developed by Felson and Clarke [43], posits that crime occurs when three elements converge in time and space: a motivated offender, a suitable target, and the absence of a capable guardian. These opportunity structures are determined by factors like value, inertia, visibility, and accessibility, collectively defining the appeal of a victim to a motivated offender.

Transitioning to the cyber space, routine activity theory and lifestyle exposure theory, while distinct, share commonalities in their focus on factors facilitating crime and increasing victimization risk. Importantly, both theories do not delve into the motivations of criminals. The application of these theories to cyberspace reveals insights into how technological advancements can create opportunities for cybercrimes and simultaneously empower potential targets to protect themselves.

In cyberspace, the absence of guardianship exposes individuals to elevated risks. Ashalan [44] highlights that frequent internet users face a greater likelihood of encountering motivated cybercriminals, especially during online activities involving personal and financial information. However, not all individuals within cyberspace are equally vulnerable; engaging in risky online activities, such as downloading freeware programs or visiting file-sharing websites, significantly heightens the risk of victimization compared to safer online actions.

Understanding cyber victimization, therefore, requires a nuanced exploration of individuals' online lifestyles. Certain activities, like checking emails or browsing online news channels, are safer, while riskier behaviors, such as downloading freeware or engaging in file-sharing, make individuals more prone to victimization. This nuanced perspective allows for a comprehensive understanding of the interplay between routine activities, lifestyle choices, and the risk of cyber victimization.

Therefore, applying this theory to the context of data breaches, the following elements can be considered:

1) MOTIVATED OFFENDER

In the context of cybercrime, this could be individuals or groups with malicious intent, such as hackers, insiders, or even state-sponsored actors.

2) SUITABLE TARGET

Organizations that handle valuable and sensitive information become suitable targets. The type of data they hold, the industry they operate in, and the perceived value of their information can influence their attractiveness to potential attackers.

3) ABSENCE OF CAPABLE GUARDIAN

The effectiveness of cybersecurity measures, policies, and practices within an organization serves as the capable guardian. If there are lapses in these areas, the risk of a successful data breach increases.

By applying routine activity theory, we can examine how the routine activities and patterns within an organization contribute to or mitigate these three elements, thereby influencing the likelihood of a data breach. This theory allows for a comprehensive analysis of the organizational

environment and the factors that may make it more susceptible to cyber threats.

III. RESEARCH PROBLEM AND HYPOTHESES

A. RESEARCH PROBLEM

The security and integrity of organizational information in the US are seriously threatened by the rising frequency and seriousness of data breaches. Despite advancements in cybersecurity measures, organizations in the United States continue to experience data breaches, raising concerns about the effectiveness of existing protective mechanisms. Consequently, there is a need to systematically examine the vulnerability of organizations to data breaches and understand the underlying factors contributing to their victimization. To address this gap, this study aims to provide a nuanced understanding of the dynamics influencing organizational vulnerability to data breaches by developing and testing hypotheses derived from modern crime theory. Through this, effective cybersecurity and risk mitigation strategies can be developed. The main goal of this research work is to investigate the applicability of routine activity theory (a component of modern crime theory) in understanding the organizational vulnerability to data breaches. By empirically testing hypotheses derived from routine activity theory, this study seeks to identify key determinants and potential areas for intervention to enhance organizational resilience against data breaches, thereby contributing to the development of targeted and preventive cybersecurity strategies.

B. FORMULATION OF RESEARCH HYPOTHESES

In quantitative research, formulating hypotheses is a critical step that bridges theoretical concepts with empirical testing. These hypotheses, grounded in established theories, serve as testable predictions about the relationships between variables. The validation or rejection of these hypotheses through empirical testing can then offer support or challenge the underlying theories.

Our study draws upon existing literature to construct a theoretical framework that examines the role of an organization's characteristics in predicting the likelihood of data breaches. Specifically, we focus on three key organizational attributes: attractiveness, visibility and guardianship. These constructs, derived from the routine activity theory, are posited to be crucial factors in determining an organization's vulnerability to cyber threats. As highlighted in previous studies, tree constructs, built on indicators profiling organizations, are developed on the basis of routine activity theory.

Attractiveness refers to the qualities of an organization that make it an appealing target for cyber attacks. This includes factors like the volume of sensitive data held, the financial worth of the organization, or the strategic value of its data. A higher level of attractiveness increases the perceived gains for potential attackers, thereby elevating the risk of data breaches.

Visibility encompasses the extent to which an organization is exposed or known to potential attackers. This could be influenced by the organization's online presence, media coverage, or its position in the industry. Greater visibility can lead to increased attention from cybercriminals, thereby escalating the likelihood of a data breach.

Guardianship involves the measures and controls in place to prevent or respond to cyber threats. Strong guardianship can deter potential attacks or mitigate their impact, thus playing a crucial role in safeguarding against data breaches.

Based on these theoretical concepts, our proposed theoretical model integrates these three exogenous constructs - attractiveness, visibility and guardianship - to assess their collective impact on the likelihood of a data breach, an endogenous construct. This model enables an in depth examination of how these organizational characteristics interplay to influence cyber vulnerability.

To empirically test this theoretical framework, we propose the following hypotheses:

Attractiveness and Organizational Vulnerability

Null Hypothesis (H_0): There is no significant relationship between the level of data attractiveness and the likelihood of organizational vulnerability to data breaches.

This hypothesis suggests that organizations with higher levels of attractiveness will experience a greater likelihood of being targeted for data breaches.

Visibility and Organizational Vulnerability

Null Hypothesis (H_0): The visibility of an organization is not significantly associated with the likelihood of organizational vulnerability to data breaches.

This hypothesis suggests that organizations with greater visibility, in terms of public awareness and online presence, are more likely to encounter data breaches.

Guardianship and Organizational Vulnerability

Null Hypothesis (H_0): There is no significant relationship between the level of guardianship within an organization and the likelihood of organizational vulnerability to data breaches.

This hypothesis asserts that robust guardianship measures within an organization will reduce the likelihood of data breaches occurring.

These hypotheses provide clear statements about the expected relationships between our latent variables (constructs) and the likelihood of organizational vulnerability to data breaches. These hypotheses will be empirically tested using the data collected, providing insights into the validity of our theoretical model and contributing to a deeper understanding of the factors influencing cybersecurity risks in organizational contexts.

IV. METHODOLOGY

A. RESEARCH DESIGN

Structural Equation Modeling (SEM) is widely utilized across various cyber security research studies due to its effectiveness in exploring intricate connections and patterns within the realm of cyber threats [45], [46], [47], [48]. The empirical

research design employed in this study encompasses the application of CB-SEM to examine the intricate dynamics of organizational victimization to data breaches through the lens of crime theory. This methodological approach allows for the systematic evaluation of hypothesized relationships derived from crime theory within the context of data security and organizational vulnerability. CB-SEM facilitates the simultaneous examination of multiple latent variables and their observable indicators, providing a comprehensive framework to analyze complex relationships and test theoretical propositions derived from crime theory.

The choice to utilize Structural Equation Modeling in the investigation of crime theory and data breaches is justified by its capacity to comprehensively examine complex relationships among multiple variables concurrently. CB-SEM facilitates the integration of various constructs such as organizational attractiveness, visibility, and guardianship into a unified framework, aligning with the multifaceted nature of the phenomenon under study. By employing CB-SEM, we can rigorously test hypotheses derived from crime theory, validating theoretical propositions and empirically assessing the proposed relationships. Furthermore, CB-SEM enables quantitative analysis, allowing for the quantification of the impact of different factors on the likelihood of a data breach, hypothesis testing, and model comparison. Overall, CB-SEM offers a robust statistical approach to understanding the dynamics of crime theory in the context of data breaches, providing empirical support for theoretical frameworks while capturing the intricate interplay of factors influencing organizational vulnerability to data breaches.

In the following, we provide our research scenario that serves as the foundation for our study (see figure 1).

B. DATA COLLECTION AND PROCESSING

This section outlines the comprehensive processes undertaken for data collection, processing, and the operationalization of key constructs in our study.

It involves gathering a comprehensive dataset comprising financial, organizational, and technical information pertaining to both victim and non-victim public and private organizations affected by data breach incidents. To conduct a detailed analysis of such incidents, we expand our data collection efforts to encompass a wide range of publicly available information concerning organizations, enabling us to offer a more nuanced examination of various types of data breaches.

1) WEB-BASED DATA ABOUT ORGANIZATIONS

In order to gather the necessary data, web scraping techniques were employed, which involve extracting structured and unstructured data from websites. Numerous studies have developed automated solutions, approaches, and tools for web scraping, some of which are discussed in [49]. Techniques such as the Document Object Model (DOM) [50], query languages for semi-structured data [51], and API computer languages are commonly used in web scraping. This approach offers the advantage of saving time due to its speed and automation capabilities, allowing for the collection of web data in a structured format.

In our research, we utilized a scraping approach that employed predefined customized rules for each website, configuring the scraper to locate and extract specific data. This technique relies on DOM selectors and leverages programming language libraries, particularly Python libraries such as Beautiful Soup and Selenium.

To gather diverse types of data, we relied on several sources. Crunchbase [52] served as a primary source for obtaining various data, including financial data, and is widely used by professionals and scholars seeking detailed business information on a wide range of entities. BuiltWith [53] is a web technology information profiler tool that provides datasets containing information on Internet technologies from 2011 to the present, offering valuable insights for researchers in the fields of investment and technology. CuteStat [54] is a web service used to retrieve information related to domain names, IP addresses, web servers, and search engine optimization (SEO). It provides statistical reports on various aspects of a website, such as valuation, search engine reports, traffic, and safety. Since 2005, personal data breaches have been systematically collected and reported by Privacy Rights Clearinghouse [55], offering access to data breach records and allowing users to search the database based on criteria such as year, company, organization type, and breach type. The database of Privacy Rights Clearinghouse is compiled from various sources including media reports, security bulletins, and other reputable organizations.

In the following, we provide a description of the pre-processing manipulations carried out from the initial scraping process to the subsequent steps of sample assembly and data storage

2) DATA SCRAPING PROCESS

The data utilized in this study is sourced from websites and public datasets. Web scraping techniques were employed to extract the data, followed by text preprocessing techniques using natural language processing (NLP) to prepare it for analysis. The dataset encompasses five sub-data sources, each containing a range of features related to breached organizations, such as information system details, network infrastructure, financial data, and organizational information.

In this subsection, we will explain how we transformed the scraped HTML files from our information resources into a structured format that contains the relevant insights for our study (see figure 3). The majority of the files in our collected data are raw HTML snippets from various websites. Each website has its own unique HTML page structure, necessitating the development of a customized parser for each site. The semi-structured nature of HTML and the specific tags associated with each field facilitated the extraction of the required information. However, certain parsed pages, particularly those containing advertisements, were noisy and

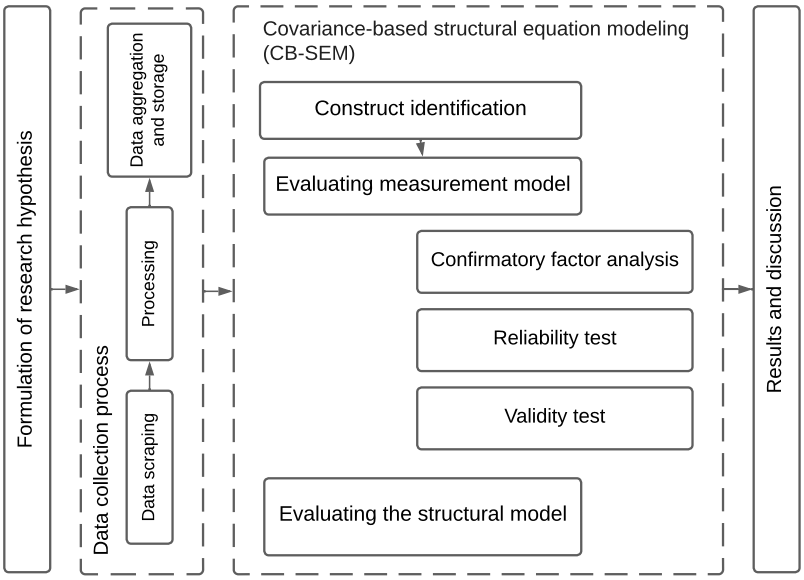

FIGURE 1. Scenario of the research methodology.

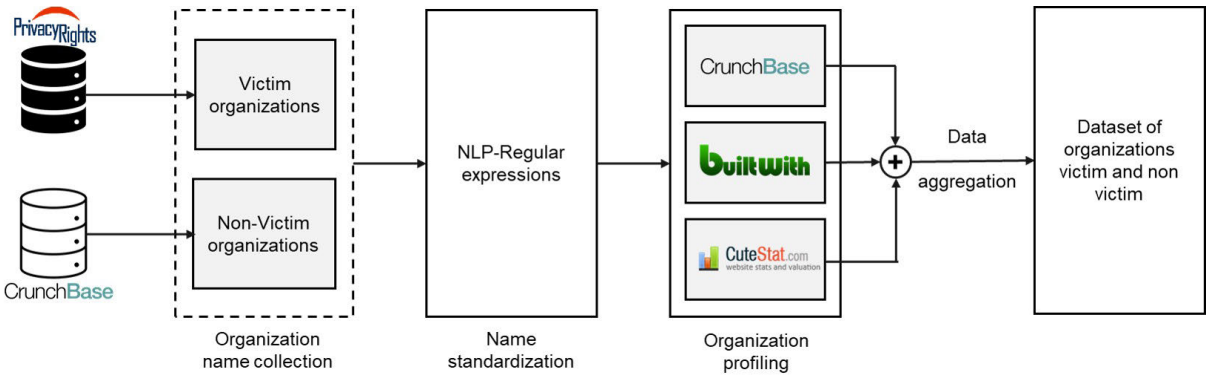

FIGURE 2. Data aggregation process.

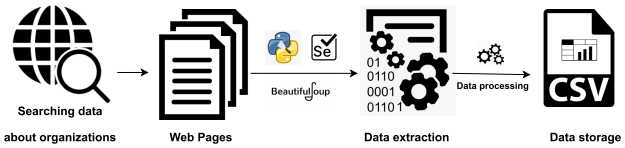

FIGURE 3. Web scraping process.

contained extraneous symbols and numbers. To address this, we applied web and data preprocessing techniques to clean the data.

We utilized the Python programming language for the scraping process, leveraging open-source libraries such as Beautiful Soup and Selenium. These libraries facilitated the extraction of text from different types of web pages. Beautiful Soup is a Python library specifically designed for extracting data from HTML and XML files, while Selenium is a package used to automate web browser interaction. By utilizing these tools, we developed a robust web scraper that enabled efficient text extraction.

NLP and Text Mining techniques were then applied to preprocess the collected data. Our study relied on data mining from multiple reliable sources, forming the foundation of our research (see figure 2).

Following the data preprocessing step, the data was stored in CSV format, where each row represents a public or private organization that has experienced a breach, and each column represents an attribute of that organization.

To assemble the sample of non-victim organizations, a random selection was made from the CrunchBase dataset. The same profiling process used for victim organizations was then applied to profile the non-victim organizations.

The resulting multivariate dataset contains 4868 organizations, covering the period from January 2018 to November 2020. Our final sample consisted of over 60 variables that captured various characteristics of each organizational data breach (refer to Table 1). Based on the literature review,

TABLE 1. Dataset statistics.

Period	Between 2018 and 2020
Total Sample Size	4868 Organizations
Victim Records	2616 Organizations
Non-Victim Records	2252 Organizations

the independent variables were categorized as related to attractiveness, level of guardianship, and visibility.

In our work, we used 16 items to measure the constructs of our research framework. A total of nine items derived from previous studies were used to assess the attractiveness level of an organization to data breach. The visibility was assessed using three items that include the sub-constructs presence in news/media, web traffic. The guardianship capabilities were evaluated using four items related to domain name security, technology expenditures.

3) DATA LABELING

The process of data labeling was conducted manually, where victim organizations were identified based on their presence in the Privacy Rights Dataset, while non-victim organizations were selected randomly. To ensure the validity of our research results, it was crucial to confirm that the non-victim organizations had not experienced any data breaches during our study period. This involved manually examining each randomly chosen organization to verify their breach-free status. This careful inspection aimed to address the issue of noisy labels, which refers to organizations that may have been victims of data breaches but were not reported in the Privacy Rights Dataset. By performing this manual check, we aimed to ensure that our non-victim sample only included organizations that had not encountered any data breach incidents throughout the study period. Henceforth, we will proceed with the assumption that our non-victim organization sample is ''noise-free'' and consists exclusively of organizations that remained breach-free during the study period.

To ensure a robust and transparent identification of victim and non-victim organizations, we employed a two-step verification process. First, victim organizations were identified based on their documented presence in the Privacy Rights Dataset, a reliable and comprehensive repository of reported data breaches. This dataset provided a solid foundation for selecting victim organizations with a known history of data breaches. In contrast, the non-victim organizations were initially selected randomly from a pool of entities that were not listed in the Privacy Rights Dataset. However, to address the potential issue of unreported breaches, we conducted an additional layer of verification. This involved a thorough examination of each non-victim organization through multiple sources, including public records, company reports, and industry-specific cybersecurity databases. This comprehensive review aimed to confirm the absence of any data breaches, reported, during our study period. This careful and systematic approach was designed to minimize the risk of including organizations with unreported breaches in the non-victim group, thereby enhancing the validity of our research results.'' In the following, we assume that:

$$\{X_v\} \cap \{X_n\} = \emptyset \tag{1}$$

where X_v is victim organizations sample and X_n is non-victim organizations sample

4) FROM DATASET VARIABLES TO CONSTRUCT-BASED MEASUREMENT

In our research, we transitioned from raw dataset variables to construct a comprehensive organization data security questionnaire. This questionnaire, summarized in Table 2, serves as a structured instrument to measure the latent constructs of attractiveness, visibility, and guardianship pertaining to organizational data security. Each construct is represented by multiple indicators, with corresponding abbreviations, indicating specific facets of data security. For instance, under the construct of attractiveness (abbreviated as ATT), indicators such as location, number of employees, and IPO status are measured on scales ranging from ordinal to nominal. These indicators capture diverse aspects of organizational appeal to potential cyber threats. Similarly, the constructs of visibility (abbreviated as VIZ) and guardianship (abbreviated as GRD) are assessed through indicators like web traffic, technological stack, and security posture, each with its own scale reflecting varying levels of exposure and protective measures. This transition from dataset variables to construct-based questionnaire design ensures a nuanced and comprehensive assessment of organizational data security, enabling us to use the CB-SEM and delve deeper into the underlying factors influencing data breach risks.

C. MODEL IDENTIFICATION

1) SELECTION OF LATENT VARIABLES AND INDICATORS

Structural Equation Modeling (SEM) serves as a powerful tool for analyzing complex relationships among latent and observed variables. In our case study, we apply CB-SEM to investigate the likelihood of data breaches in organizational contexts, incorporating key constructs such as Attractiveness, Visibility, and Guardianship, along with carefully chosen measurement items. Identification is a crucial step before estimating parameters in CB-SEM. It ensures that the model's parameters can be uniquely determined from the observed data. Identification stands as a critical prerequisite in CB-SEM, ensuring that the model's parameters can be uniquely determined from the available observed data.

In the context of our study, identification plays a pivotal role in establishing the reliability and accuracy of the relationships between the latent and observed variables shaping the likelihood of data breaches. The population covariance matrix encompasses the variances and covariances of the variables assumed to follow a specific model characterized by a set of parameters. For our data breach likelihood model, the

population covariance matrix incorporates the interplay of Attractiveness, Visibility, and Guardianship constructs. While variances and covariances in the population covariance matrix can be estimated from sample data, the focus shifts to determining whether the unknown parameters in the sample parameters can be uniquely identified from the elements of the population covariance matrix. This step is vital in ensuring the robustness and validity of our data breach likelihood model. For our data breach likelihood model, identification will validate the relationships between each indicator and the respective construct. It ensures that the unique impact of each on its construct can be discerned, providing a solid foundation for subsequent parameter estimation.

Therefore, we specify constraints on model parameters based on theoretical or empirical considerations [56]. Based on theory, this study uses three constructs, Attractiveness, Visibility, and Guardianship, as predictor variables. Constructs are theoretical concepts that cannot be directly measured but are assessed through a collection of indicators to evaluate their validity. Confirmatory factor analysis (CFA) is employed to assess how well the data align with the hypothesized measurement model and to study the relationships between observed variables(indicators) and their underlying latent constructs. CFA verifies the factor structure of the observed variables and statistically tests the suggested relationship pattern. In this study, we propose an over-identified model, where the number of constraints exceeds the number of free parameters. Therefore, the proposed model can be estimated, and its fit can be evaluated. The following path diagram illustrates the graphical representation of possible cause-and-effect relationships based on the theory (see figure 4).

For the Attractiveness construct, we included nine specific indicators: location, number of employees, ipo status, number of investors, industries, operating status, funding type, company type, and hosted category. The chosen indicators collectively reflect an organization's profile and resources, factors that may influence its appeal to potential attackers. For instance, the number of employees and the industry in which the organization operates can provide insights into the size, value of its assets and it's data sensitivity, while its funding type and operating status may indicate its financial stability and attractiveness as a target.

Visibility, another key construct, was assessed using three measurement items that capture an organization's digital presence, including web traffic, number of articles, and website rank. These indicators were selected based on the understanding that a higher digital footprint often correlates with greater exposure to cyber threats. Online presence is a relevant indicator for assessing how easily an organization can be discovered and targeted in the digital. A strong online presence can make an organization more visible to potential attackers, thereby increasing its vulnerability to cyber attacks. Network traffic analysis provides insights into the organization's digital interactions, indicating the extent of its online activities and potential vulnerabilities. By including these indicators, we aimed to capture the extent to which an organization's online activities may enhance its visibility as a target for cybercriminals.

TABLE 2. Indicators for for the reflectively measured constructs of our model.

Construct	Abbreviation	Indicator
Attractiveness (ATT)	Att_1	Location
	Att_2	Number of Employees
	Att_3	IPO Status
	Att_4	Number of Investors
	Att_5	Industries
	Att_6	Operating Status
	Att_7	Funding Type
	Att_8	Company Type
	Att_9	Hosted Category
Visibility (VIZ)	Viz_1	Web Traffic
	Viz_2	Number of Articles
	Viz_3	Website Rank
Guardianship (GRD)	Grd_1	Software Data
	Grd_2	Technological Stack
	Grd_3	Financial Allocation
	Grd_4	Security Posture

Lastly, the Guardianship construct was measured through four items: software data, technological stack, financial allocation towards technological infrastructure, and security posture of the technological infrastructure. These indicators were chosen to represent the organization's defensive capacity against cyber threats. They not only reflect the technological safeguards in place but also the level of investment and strategic importance placed on cybersecurity within the organization. By assessing these aspects, we sought to gain insights into the organization's readiness to defend against potential data breaches and mitigate cybersecurity risks effectively. Overall, the chosen indicators for each construct are justified based on crime theory literature, as they align with the theoretical principles of offender rationality, opportunity, and guardianship.

This following table 2outlines the specific indicators for each construct (attractiveness, visibility, and guardianship) in the context of organizational data breach prediction. These indicators are reflective measures that capture different aspects of each construct, allowing for a comprehensive assessment of the organization's attractiveness, visibility, and guardianship against cyber threats.

2) MODEL SPECIFICATION

This section provides a comprehensive overview of the general structural equation model including the structural part and the measurement part.

Structural Part

According to [57], the structural part establishes connections among latent variables through systems of simultaneous equations. This structural aspect is mathematically expressed as

$$\eta = \mathbf{B}\eta + \Gamma\xi + \zeta$$

where η is a vector of endogenous latent variables that are influenced by other variables within the model, ξ is a

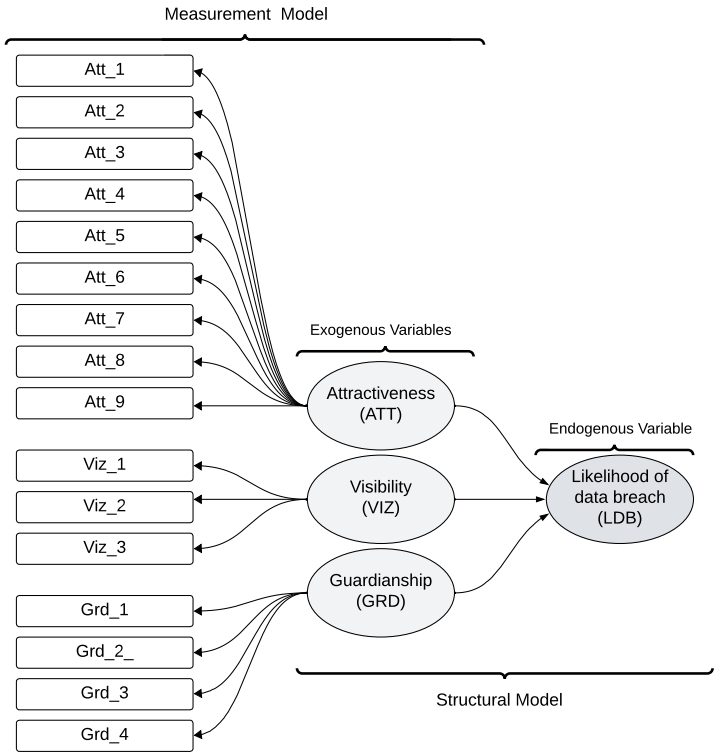

FIGURE 4. Theoretical model: Likelihood of data breach model.

vector of exogenous latent variables that are not influenced by other variables within the model, **B** is a matrix of regression coefficients relating the latent endogenous variables to each other, Γ is a matrix of regression coefficients relating endogenous variables to exogenous variables, and ζ is a vector of disturbance terms.

Measurement Part

As outlined by [57], the measurement part of a SEM links latent variables,endogenous variables and exogenous variables, to observed variables via a restricted factor model. These equations are defined as Endogenous varible measurement equation and Exogenous variable measurement equation.

Endogenous varible measurement equation

$$\mathbf{y} = \Lambda_{\mathrm{y}}\boldsymbol{\eta} + \varepsilon$$

Where y represents the observed variables related to the endogenous latent variable, $\mathbf{\Lambda}_{\mathrm{y}}$ is the matrix of vector loadings, representing the strength and direction of the relationship between latent variable $\boldsymbol{\eta}$ and the observed variable y, and $\boldsymbol{\varepsilon}$ is the vector of uniqueness capturing the unobserved factors specific to each observed variable.

Exogenous variable measurement equation

$$\mathbf{x} = \Lambda_{\mathrm{x}}\xi + \boldsymbol{\delta}$$

Where x represents the observed variables related to the exogenous latent variable, $\mathbf{\Lambda}_{\mathrm{x}}$ is the matrix of vector loadings, representing the relationship between latent variable $\boldsymbol{\xi}$ and the observed variable x, and $\boldsymbol{\delta}$ is the vector of uniqueness capturing the unobserved factors specific to each observed variable.

Practically, in our study, the structural part of the SEM model establishes connections among latent variables to examine the likelihood of data breaches within organizational contexts. Mathematically, this structural aspect is expressed as:

$$\boldsymbol{\eta} = \beta_1 \times \text{Attractiveness} + \beta_2 \times \text{Visibility} + \beta_3 \times \text{Guardianship} + \boldsymbol{\zeta}$$

Here, $\boldsymbol{\eta}$ represents a vector of endogenous latent variables influenced by other variables in the model, Attractiveness, Visibility, and Guardianship are latent variables, β_1, β_2, and β_3 are regression coefficients linking endogenous latent variables, and $\boldsymbol{\zeta}$ is a vector of disturbance terms capturing unobserved factors.

The measurement part encompasses the following equations:

Attractiveness Measurement Equation:

$$\text{Attractiveness} = \sum_{i=1}^{9} \mathbf{\Lambda}_{\mathrm{y_i}} \times \text{Item}_i + \epsilon_{\text{Attractiveness}}$$

Here, Attractiveness represents the observed variables related to the endogenous latent variable, $\mathbf{\Lambda}_{\mathrm{y_i}}$ is the vector of factor loadings indicating the relationship strength and direction between the latent variable Attractiveness and the observed variable Item_i, and $\epsilon_{\text{Attractiveness}}$ is the error

term capturing unobserved factors specific to each observed variable.

Visibility Measurement Equation:

$$\text{Visibility} = \sum_{i=1}^{3} \mathbf{\Lambda}_{\text{x}_\text{i}} \times \text{Item}_i + \epsilon_{\text{Visibility}}$$

Here, Visibility represents the observed variables related to the exogenous latent variable, $\mathbf{\Lambda}_{\text{x}_\text{i}}$ is the vector of factor loadings representing the relationship between the latent variable Visibility and the observed variable Item_i, and $\epsilon_{\text{Visibility}}$ is the error term capturing unobserved factors specific to each observed variable.

Guardianship Measurement Equation:

$$\text{Guardianship} = \sum_{i=1}^{4} \mathbf{\Lambda}_{\text{z}_\text{i}} \times \text{Item}_i + \epsilon_{\text{Guardianship}}$$

Here, Guardianship represents the observed variables related to the Guardianship latent variable, $\mathbf{\Lambda}_{\text{z}_\text{i}}$ is the vector of factor loadings indicating the relationship strength and direction between the latent variable Guardianship and the observed variable Item_i, and $\epsilon_{\text{Guardianship}}$ is the error term capturing unobserved factors specific to each observed variable. This equation encapsulates the relationship between the Guardianship construct and its observed indicators, such as software data, technological stack, financial allocation towards technological infrastructure, and security posture of the technological infrastructure.

D. MODEL ESTIMATION

Once the structural model is clearly identified, the next step is parameter estimation, which aims to accurately determine the values of the model's free parameters while adhering to constraints imposed by fixed parameters. The objective is to ensure that these estimated parameters effectively capture the complex structure inherent in the data breach likelihood model.

As Sarstedt et al. [58] note, "[t]hese constraints often contradict theoretical considerations, and the question arises whether model design should guide theory or vice versa.

CB-SEM, a widely used method in structural equation modeling (SEM), treats constructs as common factors that explain the relationships among their associated indicators. This approach aligns with reflective measurement philosophy, where indicators and their covariations reflect the underlying construct. While CB-SEM can accommodate formative measurement models, specific constraints are necessary to ensure model identification, which may sometimes conflict with theoretical considerations.

During parameter estimation, the model is applied rigorously to the sample covariance matrix, aiming to minimize the discrepancy between estimated and observed covariance matrices. Various fitting functions are employed to assess the goodness of fit, with criteria such as non-negativity and continuity. Maximum likelihood estimation, assuming multivariate normality, is commonly used in SEM, as it yields unbiased and efficient estimates, particularly suitable for large samples.

Identification guarantees that model parameters are uniquely determined, and parameter estimation seeks to find values that best fit the observed data [58], [59], [60].

V. MODEL EVALUATION AND VALIDATION

A. EVALUATING MEASUREMENT MODEL

As stated by Anderson and Gerbing [61], the evaluation of Covariance-based SEM results follows a two-step approach. The initial stage involves analyzing the measurement model through confirmatory factor analysis (CFA) to assess the validity and reliability of its components. Once the measurement models have been confirmed, the subsequent step entails evaluating the structural model.

Among the three multi-indicator constructs, namely attractiveness and level of guardianship, two displayed favorable indicator loadings. However, concerns arose regarding the visibility construct due to two items. In comparison to the recommended threshold, item viz_1 exhibited a slightly lower loading (0.691, $p = 0.001$), while item viz_3 displayed a significantly lower loading (0.622, $p = 0.201$) and lacked statistical significance. Retaining both items would result in the internal consistency reliability (Cronbach's alpha), composite reliability (CR), and convergent validity (average variance extracted (AVE)) measures falling below acceptable criteria. To address this issue, we made the decision to eliminate the weakest items from the visibility construct. The refinement of the visibility construct's indicators was undertaken with a dual emphasis on statistical and theoretical validation. Initially, indicators that exhibited weak factor loadings during the confirmatory factor analysis (CFA) phase were identified. These low loadings suggested a relatively weak correlation with the visibility construct, which warranted a closer examination. Beyond statistical criteria, we also engaged in a theoretical review of each indicator to ensure its relevance and alignment with the established cybersecurity literature. This process was crucial in determining whether the removal of certain indicators would create a conceptual gap in our understanding of an organization's visibility and its susceptibility to data breaches. In our approach, we ensure that the final set of indicators for the visibility construct not only met empirical robustness but also adhered to theoretical soundness.

1) CONFIRMATORY FACTOR ANALYSIS

Factor loadings in Structural Equation Modeling (SEM) denote the correlation between observed variables and latent factors [62]. These estimated parameters indicate the strength of the association between each observed variable and a specific latent variable, offering insights into the underlying structure of the data and aiding in the explanation of the variability in the observed variables.

The estimation of factor loadings in SEM can be accomplished through various techniques, such as Maximum

TABLE 3. Confirmatory composite analysis (CCA) results: reliability tests and average variance extracted (AVE).

Constructs	Cronbach's Alphas	Composite Reliability	AVE
Attractiveness	0.877	0.902	0.715
Visibility	0.602	0.631	0.352
Level of guardianship	0.817	0.827	0.503

Likelihood Estimation (MLE), Weighted Least Squares (WLS), and the Bayesian method [63]. In this study, we opted for Maximum Likelihood Estimation (MLE) due to its capability to provide more accurate parameter estimates and robustness against deviations from normality [64].

2) RELIABILITY TESTS

When assessing the internal consistency of a set of observed items, two commonly used measures are Cronbach's alpha and composite reliability [65], [66]. These measures evaluate the interrelationships among the observed item variables and provide a robust assessment of the data's reliability. The outcomes of these measures can guide decisions regarding the consistency and stability of the measures employed in the study.

In this study, the internal consistency of a set of items was evaluated using two tests: Cronbach's alpha and composite reliability [67], [68]. Cronbach's alpha is a well-established measure of scale reliability that assesses how closely related the items are as a group. Additionally, the composite reliability test was employed to gauge the internal consistency of the items, providing an indication of construct reliability. The evaluation of convergent validity involves assessing the level of correlation among multiple indicators of the same construct [69]. To establish convergent validity, three key factors need to be considered: the factor loading of each indicator, the composite reliability (CR), and the average variance extracted (AVE) [70]. These factors offer insights into the correlation among the indicators and the overall consistency and stability of the measures employed in the study.

The results presented in Table 3 demonstrate acceptable reliability for the items within the three constructs [71].

3) VALIDITY TESTS

The adequacy of our over-identified model can be assessed using various fit indices, including the chi-square test, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Squared Residual (SRMR) [72], [73], [74]. The corresponding results are presented in Table 4.

To evaluate the fit of our model to the data, multiple fit indices were employed. The chi-square test provided a statistically significant result ($p > 0.05$), indicating reasonable fit. However, the chi-square test can be influenced by sample size and non-normal distribution of variables. To mitigate these concerns, additional fit indices were computed.

TABLE 4. CFA model fit measures: Comparative Fit Index (CFI), Tucker-Lewis index (TLI), Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Squared Residual (SRMR).

	CFI	TLI	RMSEA	SRMR
Value	0.911	0.934	0.075	0.079
Threshold value	>0.90	>0.90	<0.08	<0.08

The comparative fit index (CFI), which is less sensitive to sample size, was calculated. A CFI value greater than 0.90 suggests a good fit. Furthermore, the Tucker-Lewis Index (TLI), which is appropriate for smaller sample sizes, was determined. A TLI value of at least 0.90 indicates a good fit.

To assess the parsimony of our model, the root mean square error of approximation (RMSEA) was used. RMSEA takes into account model complexity and rewards parsimony, with a value less than or equal to 0.08 indicating a good fit [75].

Overall, these fit indices were utilized to evaluate the adequacy of our model and assess its goodness of fit to the data.

B. EVALUATING THE STRUCTURAL MODEL

Once all parameters of the measurement model for Confirmatory composite analysis (CCA) have been met, the next step is to evaluate the relevance and predictive capability of the structural model. The results of the assessment metrics for the structural model are presented in Table 5.

This evaluation involves analyzing the explanatory and predictive elements of the model. During the testing of our model, we perform the following steps:

(a) To assess multi-collinearity among the endogenous components, we calculate the Variance Inflation Factor (VIF).

(b) Using p-values less than 0.05 as a threshold, we examine the magnitude and statistical significance of the path coefficients in the structural model. This analysis confirms the meaningfulness and relevance of all hypothesized links or predicted paths.

(c) We evaluate the in-sample predictive validity of all endogenous constructs by examining the coefficient of determination (R2) measures.

The collinearity among the constructs was found to be acceptable. In the second step of CCA, where the size and significance of the path coefficients are assessed, all paths were found to be significant ($p < 0.05$) with effect sizes (f2) ranging from medium to strong. Although the path coefficient of the relationship between Visibility and Likelihood of data breach was relatively low ($\beta = 0.152$), it was still deemed acceptable and significant based on the sample size.

VI. RESULTS DISCUSSION AND IMPLICATIONS

A. INTERPRETATION OF FINDINGS

In this study, we explore the application of a modern crime theory, namely routine activity theory, in understanding the dynamics of crime and victimization within the field of cyberspace. These theories shift the focus towards

TABLE 5. Structural model results.

Path Direct effects	Standardized path Coefficient (β)	Effect size (f_2)	p-value(p)	Variance inflation factor (VIF)
Attractiveness –>Likelihood of data breach	0.409	0.302	0.000	1.021
Visibility –>Likelihood of data breach	0.152	0.210	0.000	1.044
Level of guardianship –>Likelihood of data breach	0.472	0.268	0.002	1.197

situational factors that create opportunities for crime, rather than solely examining offender motivations. routine activity theory emphasizes the spatial and temporal aspects of crime, considering the presence of motivated offenders, suitable targets, and a lack of guardianship.

Drawing from the insights provided by this theoretical framework, we investigate the victimization experiences of organizations in relation to security breaches. By applying routine activity theory to the context of cyberspace, we aim to gain a deeper understanding of the factors that contribute to victimization to data breach.

In the context of crime theory, especially when investigating data breaches and the constructs associated with attractiveness, visibility, and guardianship levels, the concept of target attractiveness holds significant importance.

Target attractiveness refers to the desirability or appeal that an individual or entity possesses, making them more enticing to potential offenders. Miethe and Meier [76] argue that the level of target attractiveness directly impacts the risk of victimization, illustrating this concept with examples such as portable electronic devices and jewelry. Various factors contribute to target attractiveness, including income level, social class, and ownership of valuable goods [77]. For instance, individuals carrying valuable possessions or publicly displaying wealth increase their likelihood of becoming victims.

In the realm of cyberspace, data becomes the primary target, encompassing private and personal information, intellectual property, and other digitally stored assets. Unlike physical objects, data is intangible and often perceived as ''weightless'' in terms of measuring inertia. Consequently, it becomes highly vulnerable and appealing to cybercriminals. The approach differs from targeting specific individuals or entities. Instead, cybercriminals exploit millions of potential targets by spreading malware through networks. They patiently wait for criminal activities to align with suitable targets, leveraging the widespread nature of their attacks.

In the case of data breaches targeting organizations, factors such as the organization's activity type and potential financial gain play important roles in explaining attractiveness. Cybercriminals often target organizations that store significant amounts of confidential information, including financial records, personal data, and intellectual property. Hence, the type and classification of an organization indicate the sensitivity of its data. Financial gain is another motivating factor for offenders. They are attracted to cyberattacks that offer a high potential payoff, focusing their efforts on organizations with a likelihood of financial reward. This includes organizations that have raised substantial funds, possess a specific funding type, have a large number of investors, or hold a particular IPO status. High-profile targets also attract cybercriminals due to the potential for widespread impact or as stepping stones to other targets, particularly in the public sector or organizations with a significant presence (higher number of exits).

According to Cohen et al. [78], exposure to crime refers to the physical visibility and accessibility of individuals or objects to potential offenders at any given time or place. An increase in exposure is associated with a higher risk of victimization, as individuals who spend more time in public spaces become more accessible to offenders. In cyberspace, visibility and accessibility take on a different dimension. Networks in cyberspace transcend borders, connecting and making all entities and individuals visible to each other and to motivated offenders. The online lifestyle and activities of entities and individuals in cyberspace can increase the likelihood of exposure and contact with offenders. Engaging in risky online activities and spending more time connected to the internet have been found to be associated with a higher risk of cybercrime victimization [32], [44], [79], [80]. In the context of cyberspace, the timing of online activity may exert little influence on the level of victimization risk. Unlike in the physical world where timing can be consequential, cyber offenders can generate automated threats like malware, which can be unleashed remotely without necessitating proximity or adherence to the same time zone as their targets. This underscores the distinctive dynamics of victimization in the cyber domain.

Cohen et al. [78], provide a comprehensive understanding of guardianship, which encompasses both social and physical dimensions. Guardianship can be defined as the effectiveness of individuals or objects in deterring violations through their mere presence or by taking direct or indirect actions.

Having a strong level of guardianship is associated with a reduced risk of victimization. In the physical world, social guardianship is measured through indicators such as neighbors, friends, relatives, bystanders, or property owners [81]. Clarke and Felson argue that these indicators can enhance crime prevention efforts. On the other hand, physical guardianship refers to target-hardening measures such as barriers, door locks, theft alarms, and the like.

However, it is worth noting that cyberspace can be seen as an extension of the physical world. The social, political,

and economic relationships observed in the physical realm are projected into cyberspace through virtual spaces, websites, forums, social networks, and more. Castells [82] and Yar [83] suggest that the use of the virtual world is closely linked to social and economic class, thereby reflecting the disparities found in the physical world. Nevertheless, empirical studies have yet to establish the effectiveness of guardianship, despite its significance in theory.

When examining the factors influencing the level of guardianship against security breaches, we identified that the absence of security measures within organizations can be emphasized through an exploration of their security practices. Organizations with inadequate security measures in place, such as low Web of Trust trustworthiness and safety and low Web of Trust privacy, are more susceptible to being targeted by security breaches.

Drawing on previous research, we formulated a theoretical model to investigate the associations between different constructs and the probability of data breaches in organizations. The model consists of three primary constructs: organization's attractiveness, organization's visibility, and organization's guardianship. These constructs were examined as independent predictors of data breaches.

Given the importance of evaluating the likelihood of data breaches in organizations, we developed a comprehensive model based on CB-SEM (Covariance-Based Structural Equation Modeling) to integrate all the constructs. The theoretical model incorporates three exogenous constructs: attractiveness, visibility, and guardianship. Our hypothesis suggests that these constructs exert a positive influence on the endogenous construct, namely, the likelihood of data breach.

The integration of CB-SEM provides us with a powerful toolkit for exploring intricate relationships and dynamics within the domain of crime theory. This integration enables the identification of key factors that contribute to the occurrence of data breaches, such as the appeal of valuable data, the visibility of potential vulnerabilities, and the effectiveness of guardianship measures.

The results of the Structural Equation Modeling (SEM) indicated a good fit of the hypothesized model to the data, as evidenced by the following fit indices: Comparative Fit Index (CFI) = 0.91, Tucker-Lewis Index (TLI) = 0.93, Root Mean Square Error of Approximation (RMSEA) = 0.07, and Standardized Root Mean Square Residual (SRMR) = 0.07. These indices suggest that the model fits the data well and provides a satisfactory representation of the underlying relationships.

Furthermore, all factor loadings were found to be statistically significant at a significance level of $p < 0.05$. This indicates that the observed indicators effectively captured the underlying constructs, providing support for the validity of the measurement model. The results also demonstrated the relevance and predictive capability of the structural model, further supporting the hypothesized factor structure and the validity of the measures used in the overall model proposed in this study.

B. IMPLICATIONS FOR THEORY AND PRACTICE

The results of the CB-SEM analyses have important practical implications for cybersecurity and crime prevention strategies. Policymakers and organizations, for example, can use the knowledge acquired to prioritize resources and conduct targeted actions to improve guardianship while decreasing the attractiveness and exposure of potential targets. CB-SEM provides evidence-based decision-making by identifying the underlying causes that contribute to data breaches, ultimately leading to more effective prevention and mitigation solutions.

1) COLLABORATIVE EFFORTS

Since cyberspace is an interconnected field, organizations can benefit from collaborative efforts with peers in the industry, regulatory bodies, law enforcement agencies and criminal research specialist. Sharing best practices, threat intelligence, and collaborating on incident response can help organizations improve their overall cybersecurity posture.

2) ENHANCED RISK ASSESSMENT

Organizations can conduct more in depth risk assessments if they understand the factors that contribute to the likelihood of data breaches. Organizations can assess their vulnerabilities and prioritize their cybersecurity activities by considering these aspects.

3) TARGETED SECURITY MEASURES

Organizations can create targeted security measures to protect their sensitive data by recognizing the importance of attractiveness as a factor in data breaches. This includes putting in place strict access controls, encryption measures, and safe storage procedures. In addition, organizations should update and patch their software systems on a regular basis to reduce vulnerabilities.

4) CONTINUOUS EVALUATION AND ADAPTATION

Because the cybersecurity landscape is constantly changing, businesses must constantly evaluate and adjust their security procedures to handle emerging threats. Organizations that do regular reviews of their cybersecurity policies, risk assessments, and security controls can stay ahead of possible data breaches and enhance their overall security posture.

C. LIMITATIONS AND FUTURE RESEARCH DIRECTION

While our research gives useful insights into the factors that influence data breaches in organizations, it is crucial to recognize some limitations:

1) GENERALIZABILITY

Because our research was conducted in a particular context, it may not be fully representative of all organizations and industries. Features and conditions unique to the study's sample, limiting the conclusions' generalizability to larger populations, may influence the findings.

2) DATA AVAILABILITY

Data breach analysis is dependent on the availability and accuracy of relevant data. Due to issues such as underreporting, inconsistent reporting standards, and limited access to sensitive information, obtaining comprehensive and credible statistics on security breaches can be difficult. These data constraints may have an impact on the accuracy and completeness of our findings.

3) CAUSALITY

While our theoretical model investigates the relationships between attractiveness, visibility, guardianship, and the possibility of data breaches, it is critical to stress that the model is based on correlations rather than causative linkages. The results do not indicate a clear cause-and-effect relationship between the constructs.

4) CONSTRUCT SUBJECTIVITY

The measurement and operationalization of constructs like attractiveness, visibility, and guardianship rely on subjective judgments and interpretations. Different scholars may define and measure these characteristics differently, adding variability and subjectivity to the analysis.

5) ANALYTICAL TECHNIQUE LIMITS

While CB-SEM provides useful insights,it also has limits. This method relies on assumptions and simplifications that may not fully convey the complexities of varied linkages and interactions. Alternative analytical methodologies and further validation studies could bring new insights into the subject issue.

Despite these limitations, our research contributes to the understanding of data breaches in organizations and provides a framework for future studies that will further explore and develop the theoretical model.

In future research, we plan to explore alternative methodologies to enhance the depth and breadth of our analyses. Specifically, we propose incorporating Partial Least Squares Structural Equation Modeling (PLS-SEM) alongside covariance-based SEM (CB-SEM) to provide a more comprehensive understanding of the likelihood of data breaches in organizational contexts. PLS-SEM offers advantages in handling small sample sizes and non-normal data distributions, complementing the capabilities of CB-SEM. Additionally, we aim to integrate qualitative methods such as grounded theory or case study analysis to enrich our insights into the contextual factors influencing organizational vulnerability to data breaches. This multi-method approach will enable us to advance our understanding of cybersecurity risks and contribute to the development of targeted risk mitigation strategies.

VII. CONCLUSION

In this study, we investigated how routine activity theory can help us understand victimization in cyberspace, focusing specifically on data breaches. This theory highlights the importance of factors like target attractiveness, visibility, and guardianship in creating opportunities for cybercrime. We developed a theoretical model to explore how these factors influence the likelihood of data breaches. Using a CB-SEM framework, we proposed that these factors positively affect the likelihood of data breaches. CB-SEM allows for a comprehensive analysis of security breaches factors and the underlying constructs of criminal theory, offering valuable insights for developing effective crime prevention strategies. This study seeks to advance the existing literature on cybercrime by incorporating crime theory into the analysis of data breaches. Through this integration, the research aims to enhance our understanding of cyber threats and assist industries in identifying and addressing vulnerabilities within digital systems.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] IBM/Ponemon. (2023). *Cost of a Data Breach Report*. [Online]. Available: https://www.ibm.com/security/data-breach

[2] (2024). *Thales Data Threat Report*. [Online]. Available: https://go.thalesesecurity.com/rs/480-LWA-970/images/2024-DTR-Global-A4-Web-ar.pdf

[3] J. Straub, "Evaluating the use of technology readiness levels (TRLs) for cybersecurity systems," in *Proc. IEEE Int. Syst. Conf.*, Apr. 2021, pp. 1–6.

[4] X. Li and H. Li, "A visual analysis of research on information security risk by using CiteSpace," *IEEE Access*, vol. 6, pp. 63243–63257, 2018.

[5] P. Mayer, Y. Zou, B. M. Lowens, H. A. Dyer, K. Le, F. Schaub, and A. J. Aviv, "Awareness, intention, (In)action: Individuals' reactions to data breaches," *ACM Trans. Comput.-Hum. Interact.*, vol. 30, no. 5, pp. 1–53, Oct. 2023.

[6] M. Ashraf, *Potentially Unintended Consequences of the Sec Restricting Managerial Discretion: Evidence From Peer Data Breaches and Cyber Risk Factors*, document SSRN 3807487, 2021.

[7] A. Bouveret, "Cyber risk for the financial sector: A framework for quantitative assessment," *IMF Work. Papers*, vol. 18, no. 143, p. 1, 2018.

[8] N. Nejjari, K. Zkik, and H. Benbrahim, "The breach is dead, long live the breach: A spatial temporal study of healthcare data breaches," in *Proc. Int. Conf. Sci., Eng. Manag. Inf. Technol.*, 2022, pp. 287–303.

[9] M. Barati and B. Yankson, "Predicting the occurrence of a data breach," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 2, Nov. 2022, Art. no. 100128.

[10] Z. Fang, M. Xu, S. Xu, and T. Hu, "A framework for predicting data breach risk: Leveraging dependence to cope with sparsity," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2186–2201, 2021.

[11] H. Sun, M. Xu, and P. Zhao, "Modeling malicious hacking data breach risks," *North Amer. Actuarial J.*, vol. 25, no. 4, pp. 484–502, Oct. 2021.

[12] X. Zhang and X. Chen, "Research on breach prediction for big data through hybrid ensemble learning and logistic regression," *J. Phys. Conf. Ser.*, vol. 1982, no. 1, Jul. 2021, Art. no. 012049.

[13] X. Huang, Y. Lu, D. Li, and M. Ma, "A novel mechanism for fast detection of transformed data leakage," *IEEE Access*, vol. 6, pp. 35926–35936, 2018.

[14] A. Acquisti, A. Friedman, and R. Telang, "Is there a cost to privacy breaches? An event study," in *Proc. ICIS*, 2006, p. 94.

[15] S. Cook, L. Giommoni, N. Trajtenberg Pareja, M. Levi, and M. L. Williams, "Fear of economic cybercrime across Europe: A multilevel application of routine activity theory," *Brit. J. Criminology*, vol. 63, no. 2, pp. 384–406, Mar. 2023.

[16] Z. I. Vakhitova, C. L. Alston-Knox, and R. I. Mawby, "Online routine activities and self-guardianship against cyber abuse," *Victims Offenders*, vol. 18, no. 4, pp. 623–645, May 2023.

[17] D. Maimon, C. J. Howell, R. C. Perkins, C. N. Muniz, and T. Berenblum, "A routine activities approach to evidence-based risk assessment: Findings from two simulated phishing attacks," *Social Sci. Comput. Rev.*, vol. 41, no. 1, pp. 286–304, Feb. 2023.

[18] R. Stark, "Deviant places: A theory of the ecology of crime," *Criminology*, vol. 25, no. 4, pp. 893–910, Nov. 1987.

[19] J. C. Cross and A. H. Hernández, "Place, identity, and deviance: A community-based approach to understanding the relationship between deviance and place," *Deviant Behav.*, vol. 32, no. 6, pp. 503–537, Jul. 2011.

[20] P. Puente Guerrero, "Lifestyle-exposure theory as a framework to analyze victimization of people experiencing homelessness," *Deviant Behav.*, vol. 44, no. 10, pp. 1549–1569, Oct. 2023.

[21] R. Jervis, "Deterrence theory revisited," *World Politics*, vol. 31, no. 2, pp. 289–324, Jan. 1979.

[22] J. D'Arcy and T. Herath, "A review and analysis of deterrence theory in the IS security literature: Making sense of the disparate findings," *Eur. J. Inf. Syst.*, vol. 20, no. 6, pp. 643–658, Nov. 2011.

[23] S. L. Green, "Rational choice theory: An overview," Fac. Develop., Baylor Univ., Waco, TX, USA, 2002, pp. 1–72.

[24] S. Sattler, F. van Veen, F. Hasselhorn, G. Mehlkop, and C. Sauer, "An experimental test of situational action theory of crime causation: Investigating the perception-choice process," *Social Sci. Res.*, vol. 106, Aug. 2022, Art. no. 102693.

[25] J. Scott, "Rational choice theory, understanding contemporary society: Theories of the present," *Int. Encyclopedia Social Sci.*, vol. 129, pp. 126–138, Jun. 2000.

[26] J. Kennedy, T. Holt, and B. Cheng, "Automotive cybersecurity: Assessing a new platform for cybercrime and malicious hacking," *J. Crime Justice*, vol. 42, no. 5, pp. 632–645, Oct. 2019.

[27] N. Nejjari, K. Zkik, H. Benbrahim, and M. Ghogho, "Beyond locks and keys: Structural equation modeling based framework to explore security breaches through the lens of crime theories," in *Proc. Int. Conf. Networked Syst.*, 2023, pp. 96–101.

[28] T. J. Holt, J. R. Lee, J. D. Freilich, S. M. Chermak, J. M. Bauer, R. Shillair, and A. Ross, "An exploratory analysis of the characteristics of ideologically motivated cyberattacks," *Terrorism Political Violence*, vol. 34, no. 7, pp. 1305–1320, Oct. 2022.

[29] E. R. Leukfeldt and M. Yar, "Applying routine activity theory to cybercrime: A theoretical and empirical analysis," *Deviant Behav.*, vol. 37, no. 3, pp. 263–280, Mar. 2016.

[30] F. T. Ngo, A. R. Piquero, J. LaPrade, and B. Duong, "Victimization in cyberspace: Is it how long we spend online, what we do online, or what we post online?" *Criminal Justice Rev.*, vol. 45, no. 4, pp. 430–451, Dec. 2020.

[31] B. W. Reyns and B. Henson, "The thief with a thousand faces and the victim with none: Identifying determinants for online identity theft victimization with routine activity theory," *Int. J. Offender Therapy Comparative Criminology*, vol. 60, no. 10, pp. 1119–1139, Aug. 2016.

[32] K.-S. Choi, "Computer crime victimization and integrated theory: An empirical assessment," *Int. J. Cyber Criminology*, vol. 2, no. 1, pp. 1–12, Aug. 2008.

[33] T. J. Holt, G. W. Burruss, and A. M. Bossler, "Assessing the macro-level correlates of malware infections using a routine activities framework," *Int. J. Offender Therapy Comparative Criminology*, vol. 62, no. 6, pp. 1720–1741, May 2018.

[34] T. J. Holt, R. Leukfeldt, and S. van de Weijer, "An examination of motivation and routine activity theory to account for cyberattacks against Dutch web sites," *Criminal Justice Behav.*, vol. 47, no. 4, pp. 487–505, Apr. 2020.

[35] C. J. Howell, G. W. Burruss, D. Maimon, and S. Sahani, "Website defacement and routine activities: Considering the importance of hackers' valuations of potential targets," *J. Crime Justice*, vol. 42, no. 5, pp. 536–550, Oct. 2019.

[36] S. Jacques and E. Bonomo, "Learning from the offenders? Perspective on crime prevention," in *Crime Prevention in the 21st Century: Insightful Approaches for Crime Prevention Initiatives*. Cham, Switzerland: Springer, 2017, pp. 9–17.

[37] D. Maimon, "'Louderback' cyber-dependent crimes: An interdisciplinary review," *Annu. Review Criminology*, vol. 2, pp. 191–216, Aug. 2019.

[38] B. W. Reyns, R. Randa, and B. Henson, "Preventing crime online: Identifying determinants of online preventive behaviors using structural equation modeling and canonical correlation analysis," *Crime Prevention Community Saf.*, vol. 18, no. 1, pp. 38–59, Feb. 2016.

[39] B. W. Reyns, B. S. Fisher, A. M. Bossler, and T. J. Holt, "Opportunity and self-control: Do they predict multiple forms of online victimization?" *Amer. J. Criminal Justice*, vol. 44, no. 1, pp. 63–82, Feb. 2019.

[40] B. W. Reyns, B. Henson, and B. S. Fisher, "Guardians of the cyber galaxy: An empirical and theoretical analysis of the guardianship concept from routine activity theory as it applies to online forms of victimization," *J. Contemp. Criminal Justice*, vol. 32, no. 2, pp. 148–168, May 2016.

[41] T. J. Holt, N. D. Turner, J. D. Freilich, and S. M. Chermak, "Examining the characteristics that differentiate Jihadi-associated cyberattacks using routine activities theory," *Social Sci. Comput. Rev.*, vol. 40, no. 6, pp. 1614–1630, Dec. 2022.

[42] O. Smirnova and T. J. Holt, "Examining the geographic distribution of victim nations in stolen data markets," *Amer. Behav. Scientist*, vol. 61, no. 11, pp. 1403–1426, Oct. 2017.

[43] M. Felson and R. V. Clarke, "Opportunity makes the thief," *Police Res. Ser.*, vol. 98, pp. 1–36, Jul. 1998.

[44] A. Al-Shalan, "Cyber-crime fear and victimization: An analysis of a national survey," Theses and Dissertations 1244, 2006.

[45] S. D. M. Wahid, A. G. Buja, M. N. H. H. Jono, and A. A. Aziz, "Assessing the influential factors of cybersecurity awareness in Malaysia during the pandemic outbreak: A structural equation modeling," *Int. J. Adv. Technol. Eng. Explor.*, vol. 8, no. 74, pp. 73–81, Jan. 2021.

[46] S. A. A. Bokhari and S. Myeong, "The influence of artificial intelligence on E-governance and cybersecurity in smart cities: A stakeholder's perspective," *IEEE Access*, vol. 11, pp. 69783–69797, 2023.

[47] B. AlGhanboosi, S. Ali, and A. Tarhini, "Examining the effect of regulatory factors on avoiding online blackmail threats on social media: A structural equation modeling approach," *Comput. Hum. Behav.*, vol. 144, Jul. 2023, Art. no. 107702.

[48] W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. B. Kamin, and F. A. Moafa, "A model of factors affecting cyber bullying behaviors among university students," *IEEE Access*, vol. 7, pp. 2978–2985, 2019.

[49] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," *Briefings Bioinf.*, vol. 15, no. 5, pp. 788–797, Sep. 2014.

[50] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "DOM-based content extraction of HTML documents," in *Proc. 12th Int. Conf. World Wide Web*, 2003, pp. 207–214.

[51] J. Siméon, D. Chamberlin, D. Florescu, S. Boag, M. F. Fernández, and J. Robie, *XQuery 1.0: An XML Query Language*, W3C Rec., Jan. 2007.

[52] (2021). *Crunchbase: Search Companies*. Accessed: Jan. 20, 2021. [Online]. Available: https://www.crunchbase.com/

[53] (2021). *Builtwith: Datasets*. Accessed: Jan. 20, 2021. [Online]. Available: https://builtwith.com/

[54] (2021). *Cutestat: Website Stats and Valuation*. Accessed: Jan. 20, 2021. [Online]. Available: https://www.cutestat.com/

[55] (2021). *Privacyrights: Data Breaches*. Accessed: Jan. 20, 2021. [Online]. Available: https://privacyrights.org/data-breaches

[56] R. B. Kline, *Principles and Practice of Structural Equation Modeling*. New York, NY, USA: Guilford publications, 2023.

[57] C. E. Werts, K. G. Joreskog, and R. L. Linn, "Identification and estimation in path analysis with unmeasured variables," *Amer. J. Sociology*, vol. 78, no. 6, pp. 1469–1484, May 1973.

[58] M. Sarstedt, J. F. Hair, C. M. Ringle, K. O. Thiele, and S. P. Gudergan, "Estimation issues with PLS and CBSEM: Where the bias lies!" *J. Bus. Res.*, vol. 69, no. 10, pp. 3998–4010, Oct. 2016.

[59] P. Schury, C. Bachelet, M. Block, G. Bollen, D. A. Davies, M. Facina, C. M. Folden III, C. Guénaut, J. Huikari, E. Kwan, A. A. Kwiatkowski, D. J. Morrissey, R. Ringle, G. K. Pang, A. Prinke, J. Savory, H. Schatz, S. Schwarz, C. S. Sumithrarachchi, and T. Sun, "Erratum: Precision mass measurements of rare isotopes nearN=Z=33produced by fast beam fragmentation," *Phys. Rev. C*, vol. 80, no. 2, Aug. 2009, Art. no. 055801.

[60] A. Diamantopoulos and P. Riefler, "Using formative measures in international marketing models: A cautionary tale using consumer animosity as an example," in *Advances in International Marketing*. Bingley, U.K.: Emerald Group Publishing Limited, 2011, pp. 11–30.

[61] J. C. Anderson and D. W. Gerbing, "Structural equation modeling in practice: A review and recommended two-step approach," *Psychol. Bull.*, vol. 103, no. 3, pp. 411–423, 1988.

[62] R. P. Bagozzi, "Causal modeling: A general method for developing and testing theories in consumer research," *Adv. Consum. Res.*, vol. 8, no. 1, 1981.

[63] L. A. Hayduk and D. N. Glaser, "Jiving the four-step, waltzing around factor analysis, and other serious fun," *Struct. Equation Model., A Multidisciplinary J.*, vol. 7, no. 1, pp. 1–35, Jan. 2000.
[64] J. L. Arbuckle, G. A. Marcoulides, and R. E. Schumacker, "Full information estimation in the presence of incomplete data," *Adv. Struct. Equation Model.*, vol. 243, p. 277, Jul. 1996.
[65] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951.
[66] T. Raykov, "Estimation of composite reliability for congeneric measures," *Appl. Psychol. Meas.*, vol. 21, no. 2, pp. 173–184, Jun. 1997.
[67] J. R. Lewis, "Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ," *ACM SIGCHI Bull.*, vol. 23, no. 1, pp. 78–81, Jan. 1991.
[68] C. Feldmann, "These small leaks that make the big releases; CES petites fuites QUI font les grands rejets," *Clim Pratique*, no. 2, Apr. 1997.
[69] D. T. Campbell and D. W. Fiske, "Convergent and discriminant validation by the multitrait-multimethod matrix," *Psychol. Bull.*, vol. 56, no. 2, pp. 81–105, 1959.
[70] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *J. Marketing Res.*, vol. 18, no. 1, p. 39, Feb. 1981.
[71] V. A. Vieira, "Experimental designs using ANOVA," *Revista De Administração Contempornea*, vol. 15, no. 2, pp. 363–365, Apr. 2011.
[72] P. M. Bentler, "Comparative fit indexes in structural models," *Psychol. Bull.*, vol. 107, no. 2, pp. 238–246, 1990.
[73] L. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Struct. Equation Model., A Multidisciplinary J.*, vol. 6, no. 1, pp. 1–55, Jan. 1999.
[74] D. Hooper, J. Coughlan, and M. Mullen, "Evaluating model fit: A synthesis of the structural equation modelling literature," in *Proc. 7th Eur. Conf. Res. Methodol. Bus. Manag. Stud.*, 2008, pp. 195–200.
[75] J. H. Steiger, "Understanding the limitations of global fit assessment in structural equation modeling," *Personality Individual Differences*, vol. 42, no. 5, pp. 893–898, May 2007.
[76] T. D. Miethe and R. F. Meier, *Crime and Its Social Context: Toward an Integrated Theory of Offenders, Victims, and Situations*. Albany, NY, USA: Suny Press, 1994.
[77] T. D. Miethe and R. F. Meier, "Opportunity, choice, and criminal victimization: A test of a theoretical model," *J. Res. Crime Delinquency*, vol. 27, no. 3, pp. 243–266, Aug. 1990.
[78] L. E. Cohen, J. R. Kluegel, and K. C. Land, "Social inequality and predatory criminal victimization: An exposition and test of a formal theory," *Amer. Sociol. Rev.*, vol. 46, no. 5, p. 505, Oct. 1981.
[79] N. Nejjari, S. Lahlou, O. Fadi, K. Zkik, M. Oudani, and H. Benbrahim, "Conflict spectrum: An empirical study of geopolitical cyber threats from a social network perspective," in *Proc. 8th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Dec. 2021, pp. 01–07.
[80] J. François, F. Beck, G. Mezzour, K. M. Carley, A. Lahmadi, M. Ghogho, A. Houmz, H. Hammouchi, M. Zakroum, N. Nejjari, and O. Cherqi, "Threatpredict: From global social and technical big data to cyber threat forecast," in *Advanced Technologies for Security Applications*. Cham, Switzerland: Springer, 2020, pp. 45–54.
[81] R. V. G. Clarke and M. Felson, *Routine Activity and Rational Choice*, vol. 5. Piscataway, NJ, USA: Transaction Publishers, 1993.
[82] M. Castells, *The Internet Galaxy: Reflections on the Internet, Business, and Society*. London, U.K.: Oxford Univ. Press, 2002.
[83] M. Yar, "The novelty of 'cybercrime': An assessment in light of routine activity theory," *Eur. J. Criminology*, vol. 2, no. 4, pp. 407–427, Oct. 2005.

NARJISSE NEJJARI received the M.Sc. degree, in 2018. She is currently pursuing the Ph.D. degree with the National School of Computer Science and Systems Analysis, Rabat. She joined the International University of Rabat, in 2019, as a Research Fellow, and received a research grant to join the Thread predict project. She has published research papers in conferences. Her research interests include AI-enabled cybersecurity analytics, machine learning, and data science.

KARIM ZKIK received the Engineering degree in network and telecommunications engineering from the National School of Applied Sciences of Safi (ENSA-Safi), in 2013, and the Ph.D. degree in computer science from Mohammed V University, Rabat, in 2018. He is currently a Research Professor of computer science and cybersecurity with the ESAIP Engineering School and a member of the CERADE Research Laboratory. He also holds the position of the Head of the "CyberSecurity Innovation Hub by ESAIP," where he supervises innovative projects in cybersecurity. His research interests include information systems security, cybersecurity of connected systems, cyber resilience, and the security of industrial control systems and manufacturing processes. He has contributed to numerous research works in renowned scientific journals, such as: *Computers and Industrial Engineering*, *International Journal of Logistics Research and Applications*, *Journal of Network and Computer Applications*, *Production Planning and Control*, *International Journal of Production Research*, and *Journal of Cleaner Production*.

HICHAM HAMMOUCHI (Member, IEEE) received the master's degree in big data analytics and smart systems from Sidi Mohammed Ben Abdellah University, Fez, Morocco, in 2017. He is currently pursuing the Ph.D. degree with Mohammed V University of Rabat in co-direction with the International University of Rabat. His research interests include machine learning and cybersecurity. He was a recipient of the Kambule Masters Award of Deep Learning Indaba for his master's thesis on automatic lip reading.

MOUNIR GHOGHO (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the National Polytechnic Institute of Toulouse, France, in 1993 and 1997, respectively. He was an EPSRC Research Fellow with the University of Strathclyde, Scotland, from September 1997 to November 2001. In December 2001, he joined the School of Electronic and Electrical Engineering with the University of Leeds, U.K., where he was promoted to a Full Professor, in 2008. While still affiliated with the University of Leeds, in 2010, he joined the International University of Rabat, where he is currently the Dean of the College of Doctoral Studies and Director of the ICT Research Laboratory (TICLab). He has coordinated around 20 research projects and supervised more than 30 Ph.D. students in the U.K. and Morocco. His research interests include machine learning, signal processing, and wireless communications. He is a fellow of Asia-Pacific AI Association (AAIA), and a recipient of the 2013 IBM Faculty Award and the 2000 U.K. Royal Academy of Engineering Research Fellowship. He is the Co-Founder and Co-Director of the CNRS-Associated International Research Laboratory DataNet, in the field of big data and artificial intelligence. He served as an Associate Editor for many journals, including the *IEEE Signal Processing Magazine* and IEEE TRANSACTIONS ON SIGNAL PROCESSING.

HOUDA BENBRAHIM, photograph and biography not available at the time of publication.