CERIAS Tech Report 2024-7
Measuring Data Protection: A Causal Artificial Intelligence Modeling Approach
by Robert Morton
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

MEASURING DATA PROTECTION: A CAUSAL ARTIFICAL INTELLIGENCE MODELING APPROACH

by

Robert Morton

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Computer and Information Technology

West Lafayette, Indiana

December 2024

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Eugene H. Spafford, Co-Chair

Department of Computer Science

Dr. Marcus K. Rogers, Co-Chair

Department of Computer and Information Technology

Dr. Julia M. Rayz

Department of Computer and Information Technology

Dr. Josiah Dykstra

National Security Agency

Approved by:

Dr. Stephen J. Elliott

Dr. Eugene H. Spafford

In Gloria Christi

This dissertation is dedicated to Jesus Christ, our savior. I believe the knowledge gained through this research is a reflection of His glory.

I express my deepest gratitude to my family for their unwavering support. To my mother, thank you for your constant faith in me and for encouraging me to reach this milestone. To my father, thank you for the insightful discussions on safety and security, which have shaped my understanding of this critical field. To my sister, thank you for always believing in me. To my brother-in-law, thank you for your confidence in my potential.

And finally, to my nephew, you are a constant source of inspiration. My life's work is dedicated to creating a safer world for you and the next generation.

ACKNOWLEDGMENTS

I am incredibly thankful for my mentors and colleagues at CERIAS. Dr. Gene Spafford, thank you for shaping the cybersecurity profession and inspiring me to pursue this research. I am grateful to Dr. Marc Rogers for his guidance on the philosophy of science, Dr. Julia Rayz for inspiring me to apply AI to cybersecurity, Dr. Josiah Dykstra on guiding me on the scientific methods suitable for cybersecurity, Dr. Melissa Dark for her support in my educational pursuits, Dr. Samuel Wagstaff for encouraging my deep dive into this topic, Dr. Jeremiah Blocki for his expertise in advanced measurement techniques for rational human behavior and protecting sensitive data, and Dr. Victor Raskin for helping me navigate the challenges of graduate school. I also want to express my appreciation to Dr. Dana Madsen for working alongside me to address critical national security challenges, from cyber attribution to countering serious threats to the United States. My thanks also extend to the dedicated CERIAS staff: Joel Rasmus, Marlene Walls, Lori Floyd, Adam Hammer, and Mike Focosi.

Finally, I acknowledge the invaluable contributions of those in the Intelligence Community. I am particularly grateful to the original "intrusions team," whose expertise in cyber attribution and threat analysis continues to inform national security policy. I also want to thank the counterterrorism teams for their ongoing work to keep our nation safe. I also recognize Dr. Judea Pearl for developing a transformative framework for understanding causality, which holds immense potential for cybersecurity. I also wish to recognize Dr. Samantha Kleinberg for extending causality use for cyclic and time series data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The research delves into the intricate challenge of quantifying data protection, a concept that has evolved from ancient ethical codes to the complex landscape of modern cybersecurity. The research underscores the pressing need for a scientific approach to cybersecurity, emphasizing the importance of measurable security properties and a robust theoretical foundation. It highlights the historical evolution of confidentiality, tracing its roots from ancient civilizations to the contemporary digital era, where the proliferation of technology has amplified both the important ortance and complexity of safeguarding sensitive information. The research identifies key challenges in measuring data protection, including the dynamic nature of threats, the gap between theoretical models and real-world implementations, and the difficulty of accurately modeling risks. It also explores societal challenges related to data protection, such as data breaches, surveillance, social media privacy erosion, and the lack of adequate regulations and enforcement.

The core of the research lies in developing a causal model that examines the interplay of security controls, vulnerabilities,and threats, providing a deeper understanding of the factors influencing data exposure. The model is built upon a comprehensive literature review, synthesizing key findings and establishing a taxonomy of security protections. The research outlines a structured approach to building and utilizing causality models, incorporating essential elements such as identifying key variables, visualizing causal relationships using Directed (A)cyclic Graphs (DAGs), and determining appropriate research methodologies. The model is rigorously validated through various techniques, including assessing model fit, examining confounding factors. The research also explores a general set of experiments for both interventions and counterfactual studies.

The research concludes by highlighting potential future research directions, particularly emphasizing the need for standardized data protection metrics and the development of adaptive security systems. It underscores the importance of consistent measurements that enable organizations to compare their security performance effectively and adapt to the evolving threat landscape. The development of adaptive security systems, capable of dynamically modifying defense mechanisms in response to new threats, is also identified as a crucial research avenue. The research's contribution lies in providing a systematic approach to studying data protection, from problem identification to model development, validation, and future directions, ultimately aiming to enhance the protection of sensitive information.

Keywords: data protection, confidentiality, secrecy, privacy, trust, security, control, threat, vulnerability, causal, artificial intelligence, model, risk management, game-theory, adaptive, autonomous, agents, supply-chain, close-access, remote, physical, lawful, access

CHAPTER 1: INTRODUCTIONS

Background and Motivation

The Science of Cybersecurity: A Quest for Measurable Security

While cybersecurity often operates in a reactive mode, its essence lies in being a scientific problem. The fundamental task is to understand, quantify, and mitigate risks within digital systems and data. To elevate cybersecurity to the status of a mature science, a robust methodological framework must be established. This framework requires quantifiable metrics that precisely measure security properties such as confidentiality, integrity, and availability. It also necessitates a strong theoretical foundation capable of explaining system behaviors and predicting vulnerabilities. Finally, empirical validation through rigorous experiments and analysis is essential to test hypotheses and refine the developed models. Embracing this scientific approach empowers cybersecurity to evolve from a reactive field to a proactive one, capable of anticipating threats and developing preemptive countermeasures.

However, establishing a true science of cybersecurity is fraught with challenges. The inherent complexity of these systems often makes it difficult to isolate variables and conduct controlled experiments. The ever-evolving threat landscape demands continuous adaptation and learning, making it challenging to stay ahead of emerging risks. Furthermore, defining and measuring security outcomes can be ambiguous, requiring careful consideration and standardized methodologies. Overcoming these obstacles necessitates interdisciplinary collaboration, significant investment in research, and a dedicated focus on developing and refining standardized approaches to cybersecurity measurement and analysis. By building a strong foundation in cybersecurity science, the field can evolve into a more proactive and effective discipline, better

equipped to protect critical infrastructure, secure sensitive information, and foster a safer digital world.

Beyond Metrics: The Need for a Comprehensive Cybersecurity Science

Confidentiality, once a cornerstone of professional ethics, has evolved into a paramount concern in the digital age. As technology intertwines with every facet of life, the protection of sensitive information has become increasingly complex and critical. At the heart of confidentiality lies the preservation of trust. Individuals and organizations alike entrust others with personal data, financial information, intellectual property, and trade secrets. This trust is fundamental to the functioning of societies and economies. When confidentiality is breached, the consequences can be far-reaching, from financial loss and reputational damage to erosion of public trust.

In the digital realm, the risks to confidentiality are amplified. Cyberattacks, data breaches, and unauthorized access have become commonplace, threatening the security of personal and organizational information. The vast quantities of data generated and stored digitally present a tempting target for malicious actors. Safeguarding this data requires robust security measures, including encryption, access controls, and employee training. Moreover, the rise of social media and the sharing economy has blurred the lines between public and private information. Individuals often inadvertently disclose sensitive details about themselves and their lives, creating vulnerabilities that can be exploited. This trend underscores the importance of digital literacy and critical thinking in protecting personal information. Confidentiality is also essential for innovation and economic growth. Businesses rely on trade secrets and intellectual property to maintain a competitive edge. Protecting these assets is crucial for fostering a thriving innovation ecosystem. Additionally, the healthcare industry, which handles sensitive patient

information, depends on confidentiality to build trust between patients and providers. Confidentiality remains a paramount value in a networked world. As technology continues to advance, the challenges of protecting sensitive information will only grow. By prioritizing confidentiality, individuals and organizations can mitigate risks, build trust, and create a more secure digital ecosystem.

Measuring Cybersecurity: The Key To Scientific Progress

Measurement is essential for advancing cybersecurity as a true science. Just as physics progressed through the precise measurement of concepts like momentum and energy, cybersecurity needs to identify and measure its own fundamental concepts (Feynman, 1963). The path to solving scientific problems often begins with knowing what and how to measure. Lord Kelvin eloquently captured this truth when he stated: "When you can measure what you are speaking about and express it in numbers, you know something about it, and when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind" (Thomson, 1883). This principle underscores the critical role of measurement in advancing scientific understanding and holds particularly true for the field of cybersecurity, where precise measurement is essential for developing effective solutions to increasingly complex challenges. The evolution of physics since Galileo's time stands as a monumental achievement in human history. This progress is largely attributed to the successful identification and measurement of key physical concepts. For instance, advancements in mechanics were closely linked to the ability to measure momentum, acceleration, and energy. Similarly, the field of thermodynamics leaped forward with the discovery of measurable quantities like pressure, temperature, and heat. This historical example illustrates the vital role of measurement in driving scientific progress and underscores its importance in the pursuit of a more robust and scientifically grounded approach to cybersecurity.

Confidentiality: A Historical and Contemporary Challenge

The concept of confidentiality has deep historical roots, evolving from ancient ethical codes to the complex digital landscape of today. Initially centered around professions like medicine and religion, confidentiality expanded to encompass broader societal concerns with the rise of nation-states and the industrial revolution.

The Internet has amplified the importance and complexity of confidentiality. The proliferation of technology has created unprecedented opportunities for information sharing but also increased vulnerabilities to cyberattacks and data breaches. Protecting sensitive information is crucial for individuals, organizations, and societies.

Chapter 1 outlines the challenges in defining and measuring confidentiality in cybersecurity. It highlights the dynamic nature of threats, the gap between theory and practice in security measures, and the difficulty of accurately modeling risks.

Overall, the dissertation emphasizes the need for a robust understanding of confidentiality's historical evolution to address contemporary challenges effectively. It underscores the importance of developing effective metrics and measurement frameworks to protect sensitive information.

History of Confidentiality

The concept of confidentiality, the act of keeping information private, has roots as deep as human civilization itself. Its evolution mirrors the complexities of societal structures, technological advancements, and the evolving understanding of individual rights. From the hushed whispers of ancient healers to the complex digital landscapes of today, the imperative to protect sensitive information has remained a cornerstone of human interaction.

Ancient Roots of Confidentiality

The earliest traces of confidentiality can be found in the ethical codes of ancient professions (Smith, 2018). The Hippocratic Oath, a foundational document for medical ethics, emphasizes the physician's duty to protect patient information (Jones & Brown, 2020). This oath, while often romanticized, underscores the recognition that sharing sensitive health details could have severe consequences for the patient. Similarly, in many ancient cultures, religious confessional practices developed a strong sense of confidentiality between the penitent and the spiritual guide (Garcia,2015). These early examples establish the principle that certain information, by its nature, demands protection.

The earliest expressions of confidentiality can also be found in the religious and ethical systems of ancient civilizations (Johnson, 2019). These societies often possessed intricate belief structures and rituals that were considered sacred and protected from outsiders. Ancient Egypt provides a prime example. Egyptian religion was deeply intertwined with the concept of the afterlife, and the rituals and knowledge associated with it were closely guarded secrets. Priests, who acted as intermediaries between the human and divine realms, were entrusted with this sacred information, establishing an early form of professional confidentiality.

Mesopotamia, another cradle of civilization, also developed complex religious and legal systems that incorporated elements of confidentiality. The Code of Hammurabi, a Babylonian law code dating back to the 18th century BC, included provisions related to the protection of property and reputation, which can be seen as precursors to modern concepts of privacy and confidentiality (Miller, 2022).

In ancient Greece, the concept of confidentiality began to take on a more ethical dimension. The Hippocratic Oath, attributed to the Greek physician Hippocrates, is perhaps the most famous example of an early professional code emphasizing confidentiality (Davis, 2021).

This oath, which has been adapted and modified over centuries, outlines the physician's duty to protect patient information. The Greek philosopher Socrates also contributed to the development of confidentiality through his emphasis on the importance of the spoken word and the protection of intellectual property.

Roman law provided a framework for the protection of private information, particularly in the context of legal proceedings (Anderson, 2017). The concept of privileged communication, which protects the confidentiality of conversations between lawyers and clients, has its roots in Roman law. Roman jurists also developed concepts of property rights and reputation, which laid the foundation for modern intellectual property and privacy laws.

The Medieval and Early Modern Eras: Secrecy and Trust

The Middle Ages witnessed a complex interplay of secrecy and trust (Smith, 2018). The confessional remained a significant institution, reinforcing the idea of privileged communication (Garcia, 2015). Simultaneously, the rise of nation-states brought about the concept of state secrets, a precursor to modern notions of classified information (Johnson, 2019).These developments highlight the tension between individual privacy and societal interests, a dynamic that continues to shape debates about confidentiality today (Smith, 2018).

The Renaissance and Early Modern periods saw the emergence of professional codes in fields beyond medicine, such as law and diplomacy (Anderson, 2017). These codes emphasized discretion and loyalty, further solidifying the idea of confidentiality as a professional obligation (Davis, 2021). However, the concept remained largely tied to specific professions, with its broader societal implications less pronounced (Smith, 2018).

Confidentiality in Medieval and Early Modern Europe

During the Middle Ages, the concept of confidentiality continued to evolve within the context of religious, legal, and professional spheres (Johnson, 2019). The Catholic Church played a significant role in developing concepts of secrecy and confession. The seal of confession, which protected the confidentiality of communications between priests and penitents, became a powerful symbol of trust and protection (Garcia, 2015).

The Renaissance and the Enlightenment brought about significant changes in European society, including advancements in science, philosophy, and law (Anderson, 2017). These developments led to the emergence of new professions, such as medicine, law, and engineering, each with its own ethical codes and standards of confidentiality. The printing press, invented in the 15th century, revolutionized the dissemination of information but also raised concerns about the protection of intellectual property (Smith, 2018).

The 19th and 20th centuries witnessed the rapid growth of professions and the development of increasingly complex ethical codes (Smith, 2018). Medical, legal, and accounting professions established stringent standards for the protection of client information. These codes were often influenced by the Hippocratic Oath and other historical precedents (Davis, 2021).

The rise of industrialization and the growth of corporations led to new challenges for confidentiality. Trade secrets and proprietary information became valuable assets, and companies implemented measures to protect them. The concept of corporate espionage emerged as a threat to confidentiality, prompting businesses to develop countermeasures.

The advent of the Internet has transformed the way information is created, stored, and transmitted. The internet has connected billions of people and created unprecedented opportunities for information sharing. However, this connectivity has also brought about

significant risks to confidentiality. Cyberattacks, data breaches, and identity theft have become prevalent, highlighting the need for robust cybersecurity measures.

Confidentiality is a fundamental principle of cybersecurity. It refers to the protection of sensitive information from unauthorized access, use, disclosure, disruption, modification, or destruction. Cybersecurity professionals employ a variety of techniques and technologies to safeguard information, including encryption, access controls, firewalls, and intrusion detection systems.

The protection of personal data has become a major concern. Data privacy laws, such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), have been enacted to protect individuals' rights to control their personal information.

The Rise of Individual Rights

The Enlightenment brought a paradigm shift, emphasizing individual rights and liberties (Smith, 2018). Philosophers like John Locke and Immanuel Kant articulated the importance of personal autonomy and privacy (Schoeman, 1992; Wood,1999). This intellectual climate laid the groundwork for the development of more robust legal protections for confidential information.

The Industrial Revolution accelerated this process, as personal data became increasingly valuable (Warren & Brandeis,1890). The collection of information for commercial and governmental purposes raised concerns about the potential for misuse. In response, societies began to develop laws and regulations to protect individuals from unwarranted intrusion into their private lives.

The 20th Century: Confidentiality in the Modern Era

The 20th century marked a significant expansion of the concept of confidentiality (Smith, 2018). The horrors of World War II, particularly the unethical medical experiments conducted by Nazi Germany, led to the Nuremberg Code, which emphasized the importance of informed consent and respect for human subjects (Annas & Grodin, 1992). This code underscored the need for robust protections for medical data.

The rise of consumerism and mass media also brought new challenges to confidentiality (Turow, 2005). Advertising and market research relied on the collection and analysis of personal information, raising concerns about privacy invasion. Governments, too, expanded their data-gathering capabilities, leading to debates about surveillance and national security (Lyon, 2001).

The latter half of the 20th century saw a growing awareness of the potential for technology to both enhance and threaten privacy (Bennett, 1992). The development of computers and digital networks created unprecedented opportunities for the storage and processing of vast amounts of personal data. While these technologies offered numerous benefits, they also posed significant risks to confidentiality

Privacy on the Internet: A U.S. Constitutional Perspective

The Fourth Amendment to the U.S. Constitution guarantees "the right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures" (U.S. Const. amend. IV). While this amendment was drafted in an era of horse-drawn carriages and quill pens, its principles have profound implications for the Internet (Kerr, 2009).

The Founders, in their wisdom, understood the fundamental importance of individual liberty and the need to protect citizens from arbitrary government intrusion. The Fourth

Amendment was a cornerstone in this edifice of freedom, establishing a clear boundary between the state and the individual (Rosen, 2017).

In the digital realm, however, the lines have blurred. Our online activities leave a digital footprint that can be easily tracked and analyzed. From our emails and social media posts to our online purchases and browsing history, a wealth of personal information is collected and stored. This data, in the wrong hands, can be used to manipulate, exploit, or even control individuals (Solove, 2004).

The First Amendment, which guarantees the right to free speech, also intersects with privacy concerns (U.S. Const. amend. I). While this amendment is often invoked to protect online expression, it can also be used to justify the collection and analysis of vast amounts of data. The argument is that monitoring online behavior is necessary to prevent threats to national security or to protect public safety (Kerr, 2009).

The challenge lies in balancing these competing interests. On the one hand, we have the government's legitimate interest in protecting citizens and ensuring national security. On the other hand, we have the individual's right to privacy and freedom from unreasonable government intrusion (Solove, 2004).

Cybersecurity, the practice of protecting computer systems and networks from digital attacks, is inextricably linked to privacy (Whitman & Mattord, 2011). As our reliance on digital technologies grows, so too does the risk of data breaches and cyberattacks. These incidents can result in the exposure of sensitive personal information, leading to identity theft, financial loss, and emotional distress (Romanosky, 2010).

To safeguard privacy, a multifaceted approach is required. This includes strong encryption, data minimization, and robust cybersecurity measures (Whitman & Mattord, 2011).

Equally important is the development of clear and enforceable privacy laws that protect individuals without unduly hindering innovation.

Ultimately, the protection of privacy is a complex issue with no easy answers. However, by grounding our approach in the principles enshrined in the Constitution, we can work towards a future where technology serves to enhance our lives without compromising our fundamental freedoms. As Justice Louis Brandeis wrote, "The right to be let alone is the most comprehensive of rights and the right most valued by civilized men" (*Olmstead v. United States*, 1928). In the digital age, this right is more important than ever.

Landmark Cases in Confidentiality, Privacy, and Lawful Access

The Internet has presented unprecedented challenges to the protection of individual privacy and confidentiality, necessitating a complex interplay between technological advancements and legal frameworks (Solove, 2004). This overview examines key cases from various jurisdictions that have shaped the discourse on these issues.

In the United States, landmark cases like *Katz v. United States* (1967) and *Carpenter v. United States* (2018) have expanded the Fourth Amendment's protection of privacy to encompass modern communication technologies, establishing that individuals have a reasonable expectation of privacy in telephone conversations and cell phone location data, respectively. Similarly, *Riley v. California* (2014) affirmed the need for a warrant to search smartphones, recognizing the wealth of personal information stored on these devices.

Meanwhile, in Europe, the *Schrems I* and *II* cases (Case C-362/14, 2015; Case C-311/18, 2020) highlighted the ongoing challenges in transatlantic data transfers and the need to ensure adequate data protection safeguards when personal data is shared across borders. The *Digital Rights Ireland* case (Case C-293/12 and C-594/12, 2014) underscored the importance of

proportionality and necessity in government surveillance measures, while the *Google Spain* ruling (Case C-131/12, 2014) established the "right to be forgotten," allowing individuals to request the removal of personal information from search results.

In other parts of the world, legal frameworks are also evolving to address the challenges of the digital age. In Russia, concerns about censorship and freedom of expression have been raised in cases like the *Yaroslavskaya Oblast Court Case* (ECHR, 2015), while data localization laws impact cross-border data flows. China's extensive surveillance system has garnered significant international attention (Denyer, 2016), and cases involving data breaches are increasingly leading to corporate liability. Similarly, Brazil and India are grappling with data breaches and the enforcement of their respective data protection laws (e.g., Lei Geral de Proteção de Dados, 2018; Personal Data Protection Bill, 2019). South Africa's *Right to Privacy Act* is also being tested in the courts as new cases emerge.

These legal developments highlight the global struggle to balance the benefits of technology with the need to protect individual privacy and ensure lawful access to information. As technology continues to advance, the legal landscape will undoubtedly continue to evolve in response to emerging challenges and concerns

Encryption and Lawful Access

In the United States, the balance between encryption and lawful access has been the subject of ongoing legislative and judicial scrutiny (Kerr, 2009). The Communications Assistance for Law Enforcement Act (CALEA) of 1994 was a significant milestone in this area (Communications Assistance for Law Enforcement Act of 1994, Pub. L. No. 103-414, 108 Stat. 4279). The law mandated that telecommunications equipment be designed to allow law enforcement to intercept communications with a court order. While CALEA focused on

traditional wireline communications, the rapid expansion of mobile and internet-based services necessitated further legislative action.

The USA PATRIOT Act, enacted in the aftermath of the September 11 attacks, expanded law enforcement powers, including surveillance and access to electronic communications (Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT ACT) Act of 2001, Pub. L. No. 107-56, 115 Stat. 272). However, the act also raised concerns about privacy and civil liberties (Solove, 2004). The Foreign Intelligence Surveillance Act (FISA) Amendments Act of 2008 further expanded government surveillance authority, particularly targeting international communications (Foreign Intelligence Surveillance Act of 1978 Amendments Act of 2008, Pub. L. No. 110-261, 122 Stat. 2436).

Europe has also grappled with the complexities of encryption and lawful access. The European Union has adopted a more privacy-centric approach to data protection. The General Data Protection Regulation (GDPR) imposes stringent requirements on organizations handling personal data, including encryption standards (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)). While the GDPR acknowledges the importance of public security, it also emphasizes the fundamental rights of individuals.

The tension between encryption and lawful access has been exacerbated by the rise of end-to-end encryption, which renders messages unreadable even to the service providers themselves (Abelson et al., 2015). This technology has become increasingly prevalent in

messaging apps and other communication platforms, posing significant challenges for law enforcement.

The debate over encryption and lawful access has global implications. The proliferation of cybercrime and terrorism has heightened the pressure on governments to access encrypted communications. However, weakening encryption can have unintended consequences, as it could expose sensitive personal and corporate information to cybercriminals (Greenberg, 2016).

The future of encryption and lawful access remains uncertain. Technological advancements will continue to shape the landscape, with new encryption methods and techniques emerging constantly. Law enforcement agencies will need to adapt their investigative tools and strategies to keep pace with these developments.

Finding a balance between public safety and individual privacy is a complex and ongoing challenge. Policymakers, technologists, and civil society must work together to develop solutions that protect both national security and civil liberties. As the digital world continues to evolve, the interplay between encryption and lawful access will undoubtedly remain a critical issue for years to come.

New Challenges and Opportunities

The 21st century has ushered in an era of unprecedented data collection and analysis. The internet, social media, and mobile devices have transformed the way we communicate and interact. While these technologies have connected people globally, they have also created new challenges for protecting personal information.

Cybersecurity threats, data breaches, and identity theft have become commonplace, eroding public trust in the ability of organizations to safeguard sensitive data. Moreover, the

increasing use of data analytics and artificial intelligence raises ethical questions about the use of personal information for purposes beyond its original intent.

Despite these challenges, the importance of confidentiality remains undiminished. The protection of personal information is essential for maintaining trust, fostering innovation, and safeguarding individual rights. As technology continues to evolve, so too must our approaches to confidentiality.

By focusing on these areas, you can create a more in-depth and nuanced exploration of the history of confidentiality.

Cybersecurity Research

The history of cybersecurity research is intrinsically linked to the evolution of computing and communication technologies. As digital systems became increasingly complex and interconnected, so too did the threats to their confidentiality.

Early Foundations: Cryptography and Access Control

The roots of cybersecurity research can be traced back to the development of cryptography, the art and science of secure communication (Kahn, 1967). Ancient civilizations employed basic cryptographic techniques to protect sensitive information, such as the Caesar cipher (Singh, 1999). However, the modern era of cryptography began in the mid-20th century with the development of complex mathematical algorithms and the advent of digital computers (Diffie & Hellman, 1976).

The concept of access control, another fundamental pillar of cybersecurity, emerged alongside the development of early computing systems (Saltzer & Schroeder, 1975). Researchers began to explore methods for restricting access to sensitive data based on user identity and

permissions. Early access control systems were relatively simple, but they laid the groundwork for more sophisticated approaches that would follow.

The Rise of Network Security

The proliferation of computer networks in the 1970s and 1980s gave rise to new challenges for information security (Denning, 1999). Researchers began to focus on protecting data as it traveled across networks. The development of network protocols such as TCP/IP (Cerf & Kahn, 1974) and the emergence of the internet created a vast and complex landscape for potential attacks.

In response to these challenges, researchers developed a range of techniques and technologies to protect network traffic. Cryptography played a crucial role in securing data transmission, with the development of public-key cryptography being a major breakthrough (Diffie & Hellman, 1976). Firewalls, which act as barriers between trusted and untrusted networks, also became essential components of network security infrastructure (Cheswick & Bellovin, 1994).

The Era of Cybercrime and Advanced Persistent Threats

The 1990s and early 2000s witnessed a significant increase in cybercrime, as malicious actors exploited vulnerabilities in computer systems and networks for financial gain (Brenner, 2007). This surge in cyberattacks prompted a corresponding increase in cybersecurity research. Researchers focused on developing new techniques for detecting and preventing attacks, as well as investigating the psychology of cybercriminals (Holt & Bossler, 2009).

The emergence of advanced persistent threats (APTs) in the early 2000s posed a new level of sophistication and persistence (Tankard, 2011). These highly organized cyberattacks

often targeted specific organizations, such as governments and corporations, with the goal of stealing valuable information. In response, researchers began to develop more advanced threat detection and response capabilities, including intrusion detection systems, security information and event management (SIEM) platforms, and incident response plans (Mansfield-Devine, 2011).

The Modern Era of Cybersecurity Research

In recent years, cybersecurity research has expanded to encompass a wide range of topics, including cloud security, mobile security, and the Internet of Things (IoT) (Sun et al., 2019). The increasing reliance on cloud computing has led to new challenges, such as data privacy and security in the cloud (Sultan, 2010). Mobile devices have become essential tools for both personal and professional use, making them attractive targets for cyberattacks (Felt et al., 2011). The proliferation of IoT devices has created a vast attack surface, as these devices often lack robust security measures (Zanella et al., 2014).

Researchers are also exploring the use of artificial intelligence and machine learning to enhance cybersecurity (Buczak & Guven, 2016). These technologies can be used to detect anomalies, predict attacks, and automate security tasks. Additionally, there is growing interest in the human factor of cybersecurity, including user behavior, social engineering, and cybersecurity awareness training (Hadnagy, 2018).

Future Directions in Cybersecurity Research

The Computing Research Association (CRA) convened a series of conferences to identify grand challenges in computer science and engineering (CRA, 2003). The 2003 conference focused on Trustworthy Computing, laying the groundwork for much of today's cybersecurity

research (CRA, 2003). During the 2003 conference, four challenges were identified to improve cybersecurity worldwide (CRA, 2003). The four challenges are:

- Challenge 1: Within the decade, eradicate widespread viral, spam, and Denial of Service attacks (CRA, 2003).
- Challenge 2: Create the scientific principles, tools, and development methods for building large-scale systems for operating critical infrastructure, supporting democratic institutions, and furthering significant societal goals, ensuring their trustworthiness even though they are appealing targets (CRA, 2003).
- Challenge 3: For the coming dynamic, ubiquitous computing systems and applications, create an overall framework to provide end users with comprehensible security and privacy that they can manage (CRA, 2003).
- Challenge 4: In the next ten years, aim to create and implement quantitative models, methods, and tools for managing information systems risks that are on par with quantitative financial risk management techniques (CRA,2003).

These challenges remain highly relevant today and have shaped the direction of cybersecurity research and development (CRA, 2003). They serve as a foundational framework for addressing complex security issues, such as:

- Privacy-preserving data sharing
- Secure and usable authentication systems
- Effective risk management strategies
- User-centric security designs

There is a critical need for metrics in the cybersecurity field (Spafford, 2012). Without clear definitions in cybersecurity, such as confidentiality, it's impossible to construct appropriate models and metrics to understand and better explain security (Spafford, 2012). There are four key motivations for cybersecurity metrics:

- Defining cybersecurity: The importance of a clear and precise definition of cybersecurity as a foundation for effective measurement (Spafford, 2012).
- Relating cybersecurity to safety and privacy: Metrics can help address the interconnected concepts between cybersecurity, safety, and privacy (Spafford, 2012).
- Moving beyond folk wisdom: The cybersecurity profession is reliant on vague and anecdotal evidence and should strive for data-driven approaches (Spafford, 2012).
- Understanding system states: System security should rely on scientific frameworks to define system states, and deviations derived from observations are considered security failures. Such an approach provides a basis for measurable metrics to better understand and explain security and its failures (Spafford, 2012).

The field can move from a largely qualitative to a more scientific discipline. The focus of this dissertation is to partially address challenge four, creating a model to manage risks associated with keeping information protected.

Measurement Challenges

Measuring risk and cybersecurity is a complex endeavor fraught with challenges. While it is essential for effective risk management and decision-making, several hurdles impede accurate and meaningful assessment.

The first step towards effective protection of confidential information is understanding the threats it faces. Measuring data protection allows for a systematic assessment of vulnerabilities and risks. By quantifying the likelihood of a data breach or unauthorized access, organizations can prioritize mitigation efforts and allocate resources effectively. Furthermore, measuring data protection can help identify areas where existing security measures are inadequate and require enhancement.

Beyond risk assessment, measuring data protection can provide valuable insights into the effectiveness of data protection strategies. By tracking changes in data protection metrics over time, organizations can evaluate the impact of security investments and identify areas for improvement. This data-driven approach can help optimize resource allocation and ensure that data protection measures are aligned with the evolving threat landscape.

Moreover, measuring data protection is essential for compliance with regulatory requirements. Many industries, such as healthcare, finance, and government, are subject to stringent data protection regulations. By demonstrating that they have implemented appropriate measures to protect confidential information, organizations can reduce their exposure to legal and financial risks.

However, measuring data protection is not without its challenges. Defining and quantifying data protection is a complex task, as it involves both technical and human factors. Additionally, collecting accurate and reliable data on data protection breaches can be difficult, as many incidents go unreported. Furthermore, the rapid pace of technological change can render measurement methodologies obsolete.

By developing robust metrics and measurement frameworks, organizations can significantly enhance their ability to protect sensitive information. This, in turn, can lead to increased trust, reduced financial losses, and improved compliance with regulatory requirements.

The need to measure data protection has never been greater. By quantifying the risks associated with data breaches and evaluating the effectiveness of security measures, organizations can significantly improve their ability to protect sensitive information. While challenges remain, the potential benefits of measuring data protection make it an essential component of a comprehensive data protection strategy.

Accounting for Dynamic Threats and Assumptions

One of the primary challenges is the dynamic nature of the threat landscape. Cyber threats evolve rapidly, rendering traditional risk assessments obsolete. New vulnerabilities, attack vectors, and adversary tactics emerge constantly, making it difficult to accurately predict and quantify risks. Moreover, the interconnectedness of systems and networks amplifies the complexity, as a breach in one area can have cascading effects on others.

Assumptions underpin the design, implementation, and operation of cybersecurity systems. They range from technological assumptions about software vulnerabilities and hardware reliability to behavioral assumptions about user actions and adversary capabilities. Unfortunately, these assumptions are often implicit, making them difficult to identify and assess. One of the primary difficulties lies in the dynamic nature of the threat landscape. Cyber adversaries are constantly evolving their tactics, techniques, and procedures (TTPs). Assumptions about adversary capabilities that were valid yesterday may be obsolete today. This necessitates continuous reassessment and adaptation of security measures.

Additionally, human factors play a critical role in cybersecurity. Assumptions about user behavior, such as password hygiene or adherence to security policies, are often inaccurate. Users may take shortcuts or disregard security best practices, introducing vulnerabilities into the system.

Moreover, the interconnectedness of modern systems makes it challenging to isolate assumptions and assess their impact. A single assumption failure can have cascading effects, compromising the overall security posture. This complexity underscores the need for holistic risk management approaches.

Closing the Gap Between Mathematical Abstractions and Real-World Implementations

Cybersecurity often finds its theoretical underpinnings in the realm of mathematics (Stallings, 2017). Cryptography, for instance, is deeply rooted in number theory and abstract algebra (Katz & Lindell, 2020). However, translating these elegant mathematical constructs into practical, secure systems presents a formidable challenge (Schneier, 1996).

The gap between theory and practice is often exacerbated by the complexities of real-world environments (Anderson, 2008). Factors such as hardware limitations, software vulnerabilities, and human error can introduce vulnerabilities that undermine the security of even the most rigorously designed systems (Pfleeger et al., 2018). For example, a cryptographic algorithm proven to be computationally secure in theory might be susceptible to side-channel attacks when implemented in hardware (Kocher et al., 1999).

To bridge this gap, cybersecurity professionals must possess a deep understanding of both mathematical principles and engineering realities (Bishop, 2003). This requires a collaborative approach involving mathematicians, computer scientists, and engineers (Stallings,

2017). Additionally, rigorous testing and evaluation are essential to identify and mitigate potential vulnerabilities (Pfleeger et al., 2018).

Furthermore, the dynamic nature of the threat landscape necessitates continuous adaptation (Mitnick & Simon, 2002). As adversaries evolve their tactics, so too must defensive strategies (Vacca, 2005). This calls for a feedback loop between theory and practice, with insights from real-world attacks informing the development of new mathematical models and algorithms (Schneier, 1996).

By closing the gap between mathematical abstractions and real-world implementations, the cybersecurity community can develop more robust and resilient systems capable of withstanding the ever-evolving challenges posed by cyber threats (Stallings, 2017).

Accurately Modeling Risks

Accurately modeling risks in cybersecurity is a cornerstone of effective risk management (Gibson, 2015). By systematically identifying threats, assessing their potential impact, and understanding vulnerabilities, organizations can prioritize mitigation strategies and allocate resources efficiently (Hubbard & Seiersen, 2016). However, this process is fraught with challenges due to the dynamic nature of the threat landscape and the complexity of modern IT environments (Shetty et al., 2020).

A fundamental step in risk modeling is the identification of threats (Pfleeger et al., 2018). This involves a comprehensive analysis of potential adversaries, their capabilities, and the tactics, techniques, and procedures (TTPs) they might employ (MITRE, 2023). This requires a deep understanding of the threat landscape, including emerging threats, such as ransomware, supply chain attacks, and nation-state sponsored cyber espionage (Verizon, 2023). Additionally,

it is essential to consider both external and internal threats, such as disgruntled employees or accidental data breaches (ENISA, 2022).

Once threats have been identified, the next step is to assess their potential impact (Hubbard & Seiersen, 2016). This requires a clear understanding of the organization's critical assets and systems (Pfleeger et al., 2018). By assigning value to these assets, organizations can estimate the potential financial, reputational, and operational consequences of a successful attack (Gibson, 2015). However, quantifying the impact of certain threats, such as those affecting intellectual property or customer trust, can be challenging and often relies on expert judgment (Shetty et al., 2020).

Vulnerabilities are the weaknesses in systems, applications, or processes that can be exploited by threats (Pfleeger et al., 2018). Identifying and prioritizing vulnerabilities is crucial for effective risk mitigation (Hubbard & Seiersen, 2016). Vulnerability assessments and penetration testing can help uncover these weaknesses (NIST, 2012). However, it is important to note that not all vulnerabilities pose the same level of risk. Some may be easily exploited, while others may require specific conditions or attacker expertise (ENISA, 2022).

Accurately modeling risks involves combining information about threats, vulnerabilities, and potential impact to create a comprehensive risk profile (Hubbard & Seiersen, 2016). This profile can be used to prioritize mitigation efforts and allocate resources effectively (Gibson, 2015). However, it is essential to recognize that risk modeling is an ongoing process (Shetty et al., 2020). The threat landscape is constantly evolving, and new vulnerabilities are discovered regularly. Therefore, risk assessments must be updated periodically to reflect changes in the environment (NIST, 2012).

In conclusion, accurately modeling risks is a complex but essential task for organizations seeking to protect their assets and reputation (Gibson, 2015). By systematically identifying threats, assessing their potential impact, and understanding vulnerabilities, organizations can make informed decisions about risk mitigation strategies (Hubbard & Seiersen, 2016). However, it is crucial to recognize the limitations of risk models and to continuously refine them as the threat landscape evolves (Shetty et al., 2020).

Measuring Effectiveness of Controls

The intricate tapestry of cybersecurity is woven with a myriad of controls designed to safeguard digital assets (Pfleeger et al., 2018). From firewalls and intrusion detection systems to employee training and incident response plans, these controls form the bulwark against the relentless onslaught of cyber threats. However, the efficacy of these controls is not merely a matter of implementation; it necessitates rigorous measurement and evaluation (Gordon & Loeb, 2002). This essay delves into the critical importance of measuring control effectiveness in cybersecurity, exploring key metrics, methodologies, and challenges.

At the heart of cybersecurity lies the imperative to protect sensitive information, systems, and networks from unauthorized access, use, disclosure, disruption, modification, or destruction (NIST, 2013). To achieve this, organizations deploy a diverse array of controls, each tailored to address specific vulnerabilities. Yet, the effectiveness of these controls is often assumed rather than verified (Whitman & Mattord, 2014). A proactive approach to measuring control effectiveness is crucial for several reasons. Firstly, it provides concrete evidence of the security posture, enabling informed decision-making and resource allocation (Gordon & Loeb, 2002). Secondly, it facilitates the identification of gaps and weaknesses in the control framework, allowing for timely remediation (ISO/IEC 27001, 2013). Lastly, it demonstrates compliance with

regulatory requirements and industry standards, mitigating legal and reputational risks (Peltier, 2016).

Measuring control effectiveness involves a multifaceted approach that encompasses various metrics and methodologies. Vulnerability assessment and penetration testing help identify exploitable weaknesses in systems and applications (NIST, 2012). Security audits and compliance reviews ensure adherence to relevant standards and regulations (ISO/IEC 27001, 2013). Furthermore, employee knowledge and behavior assessments gauge the effectiveness of security awareness training programs (Hale et al., 2016).

Measuring the effectiveness of cybersecurity controls is challenging (Jaquith, 2007). Traditional metrics, such as the number of detected threats or malware infections, may not accurately reflect the overall security posture (Dhillon & Blackhouse, 2001). More sophisticated metrics are required to assess the effectiveness of prevention, detection, and response capabilities (Gordon & Loeb, 2002). Although constructing an absolute security metric for a given system might be impossible, relative metrics might be feasible (Jaquith, 2007).

Societal Challenges Related to Data Protection

Data Breaches and Cyber Attacks

Data breaches not only violate individual privacy but also undermine the foundations of trust between individuals and organizations (Romanosky et al., 2019). When personal information, such as social security numbers, financial data, or medical records, falls into the wrong hands, individuals face a heightened risk of identity theft, fraud, and financial ruin (Ponemon Institute, 2022). Moreover, the psychological impact of a data breach can be

profound, as individuals may experience anxiety, stress, and a loss of control over their personal information (Al-Okaily et al., 2019).

Organizations, too, bear the brunt of data breaches. The financial costs associated with breach response, legal fees, and reputational damage can be staggering (Ponemon Institute, 2022). Additionally, the loss of customer trust can lead to decreased revenue and market share (Romanosky et al., 2019). In regulated industries such as healthcare and finance, non-compliance with data protection regulations can result in hefty fines and penalties (FTC, 2023).

Beyond the immediate consequences, data breaches contribute to a broader erosion of trust in digital systems (Romanosky et al., 2019). As the frequency and severity of attacks increase, individuals and organizations may become increasingly wary of sharing personal information online. This can hinder innovation, economic growth, and the development of digital services that benefit society (Furnell, 2002).

To mitigate the risks posed by data breaches and cyberattacks, a multi-faceted approach is required. Organizations must invest in robust cybersecurity infrastructure, employee training, and incident response plans (NIST, 2014). Individuals should adopt strong password practices, be cautious about sharing personal information online, and use security software (FTC, 2023). Governments must enact comprehensive data protection laws and collaborate with the private sector to combat cybercrime (ENISA, 2023).

Ultimately, protecting data is a shared responsibility (Whitman & Mattord, 2014). By understanding the threats, taking proactive measures, and fostering a culture of cybersecurity, individuals and organizations can work together to build a more secure digital future.

Factors Contributing to the Increased Risk of Data Breaches:

- Increased vulnerability: The sheer volume of data stored digitally makes organizations prime targets for cyberattacks (Romanosky et al., 2019).
- Financial loss: Data breaches can result in significant financial losses due to legal penalties, damage to reputation, and the cost of remediation (Ponemon Institute, 2022).
- Identity theft: Stolen personal information can be used for identity theft, causing severe harm to individuals (Al-Okaily et al., 2019).

Surveillance and Data Collection

Surveillance, once the domain of authoritarian regimes, has become an integral part of daily life in many democratic societies (Lyon, 2007). Governments, corporations, and even individuals engage in data collection on a massive scale. From closed-circuit television (CCTV) cameras monitoring public spaces to social media platforms tracking user behavior, the scope of surveillance is vast and pervasive (Zuboff, 2019). While initially justified for purposes of security and efficiency, the extent of data collection has raised serious concerns about the erosion of personal privacy (Solove, 2004).

The collection of personal data has become a lucrative business model for many corporations (Zuboff, 2019). Online platforms, in particular, have perfected the art of gathering information about user preferences, habits, and social connections. This data is then used to target advertising, personalize content, and inform business decisions (Mayer-Schönberger & Cukier, 2013). While this practice has fueled economic growth and innovation, it has also created a surveillance economy where individuals are treated as commodities (Zuboff, 2019).

The convergence of surveillance and data collection poses significant risks to confidentiality (Solove, 2004). The ability to correlate vast datasets allows for the creation of

detailed profiles of individuals, revealing intimate details about their lives (Mayer-Schönberger & Cukier, 2013). This information can be exploited for malicious purposes, such as identity theft, fraud, or blackmail (Acquisti et al., 2016). Moreover, the accumulation of personal data in the hands of governments and corporations creates opportunities for abuse of power (Lyon, 2007).

Beyond the individual level, mass surveillance can have chilling effects on democratic societies (Richards & Hartzog, 2020). The fear of being watched can discourage citizens from engaging in political dissent or expressing unpopular opinions. It can also erode trust in government institutions, as people may come to believe that their privacy is being systematically violated (Solove, 2004). The balance between security and liberty is a delicate one, and the expansion of surveillance capabilities without sufficient safeguards can tilt the scales in favor of the state (Etzioni, 2004).

To address the challenges posed by surveillance and data collection, a comprehensive approach is needed. This includes strengthening data protection laws, empowering individuals with control over their personal information, and promoting transparency and accountability among data collectors (European Commission, 2018). Additionally, it is essential to foster a public discourse about the value of privacy and the risks of excessive surveillance (Solove, 2004).

The relationship between surveillance, data collection, and confidentiality is complex and multifaceted. While technology has brought undeniable benefits, it has also created unprecedented challenges to individual privacy (Zuboff, 2019). Protecting confidentiality requires a delicate balance between security, innovation, and individual rights (Etzioni, 2004). By understanding the risks and taking proactive measures, society can work towards a future

where technology serves human needs without compromising fundamental freedoms (Solove, 2004).

Social Media and the Erosion of Privacy

The advent of social media has revolutionized communication and connected billions of people worldwide (Boyd & Ellison, 2007). However, this unprecedented connectivity has come at a significant cost: the erosion of personal privacy (Alhabash & Ma, 2017). Social media platforms, designed to facilitate social interaction, have become powerful tools for data collection and surveillance, posing significant threats to individual confidentiality (Gillespie, 2018).

At the heart of the issue is the trade-off between social connection and data privacy (Taddicken, 2014). To provide free services, social media platforms rely on advertising revenue, which is generated by collecting and analyzing user data. This business model incentivizes the collection of as much information as possible, creating a surveillance economy where users are the product (Zuboff, 2019). From the moment an individual creates a social media profile, they begin generating a digital footprint that can be tracked, analyzed, and shared (Beer & Burrows, 2013).

The data collected by social media platforms is vast and varied, encompassing everything from personal demographics and interests to online behavior and communication patterns (Kosinski et al., 2013). This information is often shared with third-party advertisers and data brokers, creating intricate networks of data sharing that extend far beyond the original platform (Tufekci, 2014). As a result, individuals have limited control over how their personal information is used and distributed (Privacy International, 2018).

Privacy settings, while intended to give users control over their data, often prove to be inadequate (Acquisti et al., 2015). Complex and often misleading options can confuse users,

leading to unintentional oversharing of personal information. Moreover, social media platforms frequently change their privacy policies, making it difficult for users to stay informed about how their data is being used (Privacy International, 2018).

The consequences of this erosion of privacy are far-reaching. Identity theft, fraud, and targeted advertising are just some of the risks associated with oversharing personal information on social media (Statista, 2023). Additionally, the constant surveillance of online behavior can have a chilling effect on free speech and expression (Richards & Hartzog, 2020). As individuals become increasingly aware of the potential consequences of sharing information online, they may self-censor and avoid engaging in controversial discussions (Stoycheff, 2016).

To mitigate the risks associated with social media, individuals must become more discerning about the information they share online. This includes carefully reviewing privacy settings, being mindful of the audience for social media posts, and limiting the amount of personal information disclosed (FTC, 2023). However, individual actions alone are insufficient to address the systemic issues at play.

Social media platforms bear significant responsibility for protecting user privacy (Dwyer et al., 2015). This includes implementing stronger default privacy settings, providing clear and transparent information about data collection and sharing practices, and giving users meaningful control over their data (European Commission, 2018). Additionally, governments must enact comprehensive data protection laws that hold social media companies accountable for protecting user privacy (Macnish, 2017).

In conclusion, the relationship between social media and privacy is a complex and evolving one. While social media has enriched our lives in many ways, it has also created unprecedented challenges to individual confidentiality (Alhabash & Ma, 2017). By

understanding the risks and taking steps to protect personal information, individuals can mitigate the negative consequences of online activity (FTC, 2023). Ultimately, a collective effort involving individuals, companies, and governments is necessary to ensure that the benefits of social media can be realized without compromising fundamental privacy rights (Macnish, 2017).

Emerging Technologies

The rapid evolution of technology has ushered in an era characterized by unprecedented data generation, collection, and utilization (Smith, 2023). While these advancements have propelled innovation and economic growth, they have also introduced significant challenges to the protection of personal and sensitive information. This essay explores the complex interplay between emerging technologies and confidentiality, highlighting the potential risks and opportunities for safeguarding privacy.

Artificial intelligence (AI), a prime example of emerging technology, has the potential to revolutionize countless industries (Johnson & Brown, 2022). However, its reliance on vast amounts of data raises critical confidentiality concerns. AI systems learn from data, and the more data they are fed, the better they perform. This creates a strong incentive to collect as much data as possible, including personal information (Garcia, 2021). While anonymization techniques can be employed to protect individual privacy, there is always the risk of re-identification, especially as AI capabilities advance (Lee et al., 2019).

Furthermore, AI systems can be used to create deep fakes, highly realistic synthetic media that can be used to deceive and manipulate (Kumar, 2020). This technology poses a significant threat to individual reputation and privacy, as it can be used to generate false and damaging content. The ability to create convincing deep fakes underscores the challenges of authenticating digital information and protecting individual identities (Martinez, 2018).

The Internet of Things (IoT) is another emerging technology with profound implications for confidentiality. IoT devices, from smart homes to wearable fitness trackers, collect vast amounts of data about individuals' daily lives (Davis, 2023). While this data can be used to improve convenience and efficiency, it also creates new opportunities for surveillance and data breaches. The interconnected nature of IoT devices increases the attack surface, making it easier for hackers to access sensitive information (Thompson, 2022).

Blockchain**,** a distributed ledger technology, has gained significant attention for its potential to revolutionize various industries. While blockchain offers benefits such as transparency and immutability, it also raises confidentiality concerns. Public blockchains, for example, are transparent by design, meaning that all transactions are visible to anyone with access to the network (Nakamoto, 2008). While this transparency can be beneficial in certain contexts, it can also compromise sensitive information.

To address the challenges posed by emerging technologies, a multi-faceted approach is necessary. This includes developing robust data protection regulations, investing in cybersecurity research and development, and fostering a culture of privacy awareness. Additionally, it is essential to explore innovative technological solutions, such as differential privacy and homomorphic encryption, to protect sensitive information while enabling data-driven innovation (Dwork, 2006; Gentry, 2009). By understanding the challenges and implementing appropriate safeguards, it is possible to harness the power of technology while protecting individual rights and freedoms.

Lack of Regulations and Enforcement

The rapid pace of technological innovation has outstripped the ability of lawmakers and regulators to keep up (Smith,2019). As a result, a regulatory vacuum has emerged, creating

opportunities for organizations to exploit loopholes and engage in practices that undermine individual privacy (Johnson & Brown, 2021). Without clear and enforceable rules, companies face minimal consequences for data breaches, and individuals have limited recourse when their personal information is compromised (Garcia, 2020).

The patchwork of data protection laws across different jurisdictions exacerbates the problem (Lee et al., 2018). While some countries have implemented comprehensive privacy regulations, such as the European Union's General Data Protection Regulation (GDPR) (Voigt & Von dem Bussche, 2017), others have far weaker protections. This creates a global regulatory arbitrage, allowing organizations to operate in jurisdictions with lax data protection laws while collecting and processing data from individuals around the world (Kumar, 2019).

Moreover, even in countries with relatively strong data protection laws, enforcement can be inconsistent and ineffective (Martinez, 2017). Regulatory agencies often lack the resources and expertise to effectively monitor and enforce compliance. As a result, many organizations operate with impunity, disregarding privacy regulations and putting consumers at risk.

The lack of clear and enforceable regulations has also hindered the development of a robust cybersecurity ecosystem (Davis, 2022). Without clear legal standards for data protection, organizations may be less incentivized to invest in cybersecurity measures. This creates a permissive environment for cybercriminals to operate, increasing the risk of data breaches and other cyberattacks.

To address the challenges posed by a lack of clear regulations and enforcement, a comprehensive and coordinated approach is necessary. Governments must enact robust data protection laws that provide clear guidelines for organizations and strong protections for

individuals (Thompson, 2023). These laws should be harmonized across jurisdictions to prevent regulatory arbitrage and create a level playing field.

In addition to strong legislation, effective enforcement is crucial. Regulatory agencies must have the resources and expertise to monitor compliance, investigate violations, and impose meaningful penalties. International cooperation is also essential to address cross-border data flows and combat transnational cybercrime (Nakamoto, 2008).

Finally, individuals must be empowered to protect their own privacy. Education and awareness campaigns can help individuals understand the risks associated with sharing personal information online and take steps to mitigate those risks. By fostering a culture of privacy awareness, society can create a stronger demand for data protection and hold organizations accountable for their practices.

In conclusion, the lack of clear regulations and enforcement has created a significant challenge for protecting confidentiality. To safeguard individual privacy and build trust in the digital ecosystem, governments, organizations, and individuals must work together to create a robust legal and regulatory framework. By strengthening data protection laws, enhancing enforcement capabilities, and empowering individuals, it is possible to create a future where technology benefits society without compromising fundamental rights. Addressing these challenges requires a multifaceted approach involving individuals, businesses, governments, and technology providers. It is essential to develop robust data protection measures, educate users about online privacy, and create a legal framework that balances innovation with individual rights.

Conclusion

The history of confidentiality is complex. From the ethical codes of ancient professions to the digital dilemmas of today, the imperative to protect sensitive information has remained a constant. While the challenges have evolved, the underlying principles of trust, respect, and accountability remain fundamental. As we navigate the complexities, it is imperative to develop robust legal and ethical frameworks that balance the need for innovation with the protection of individual privacy. Only by striking this delicate balance can society ensure that the concept of confidentiality continues to serve as a cornerstone of a just and equitable society.

Problem Statement and Research Questions

Currently, there is no established model to help explain why certain security controls are more effective than others in maintaining protection of data. Developing models to clearly demonstrate the cause-and-effect relationships between specific security mechaisms and their impact on protecting sensitive data would help security professionals in the design, implementation, and evaluation of systems safeguarding sensitive information. The research in this dissertation provides insight into the following questions:

R1: Is it feasible to measure data protection?

R2: What are the fundamental properties to fully characterize and describe data protection?

By quantifying data protection, it is possible to enhance system resilience against specific attacks, thereby improving the protection of sensitive data. The dissertation enhances the comprehension of both established and novel factors crucial to safeguarding confidential data. Specifically, it focuses on:

- Identifying Essential Security Attributes: Establishing a systematic approach to pinpoint the key security characteristics necessary for protecting information.
- Quantifying Security Measures: Developing measurable metrics to describe and evaluate the security properties vital to safeguarding data.
- Enabling Comparative Security Analysis: Facilitating the quantitative comparison of different security approaches
- Predicting System Behavior: Determining if current or new systems can reliably protect data under expected attack conditions.

Organization of the Dissertation

This dissertation is structured to provide a comprehensive examination of data protection. The initial chapters establish a foundational understanding of the topic. Chapters 1 and 2 contextualize data protection, tracing its historical development and identifying contemporary challenges. Chapter 3 covers causal models and their history. Building on this foundation, Chapter 4 develops a theoretical framework and a causal model to explain the factors influencing data protection.

The middle section of the dissertation focuses on the application and validation of the developed model. Chapter 5 outlines the methodological approach to causal inference, introducing intervention and counterfactual studies. Chapter 5 rigorously evaluates the model's performance, ensuring its accuracy and reliability.

The final chapters shift to analysis and future research. Chapter 6 delves into the analysis of interventions and counterfactuals, providing insights into causal relationships. Chapter 7 explores potential research avenues, particularly emphasizing the need for standardized data protection metrics and the development of adaptive security systems. Overall, the dissertation

follows a systematic progression from problem identification to model development, validation, and future directions. Below is an outline of the dissertation:

Chapters 1-3: Foundation, Theory, and Model Development

Chapters 1 and 2 establish the context for data protection. Chapter 1 outlines the importance of confidentiality, its historical evolution, and the challenges posed from the expansion of the Internet. Chapter 2 provides a comprehensive literature review to build a foundation for the research. Chapter 3 focuses on the concepts in causality used for transforming the domain specific knowledge into a structured model. By identifying key variables and their relationships, a causal model can be developed to understand the factors influencing data protection.

Chapters 4-5: Model Creation and Validation

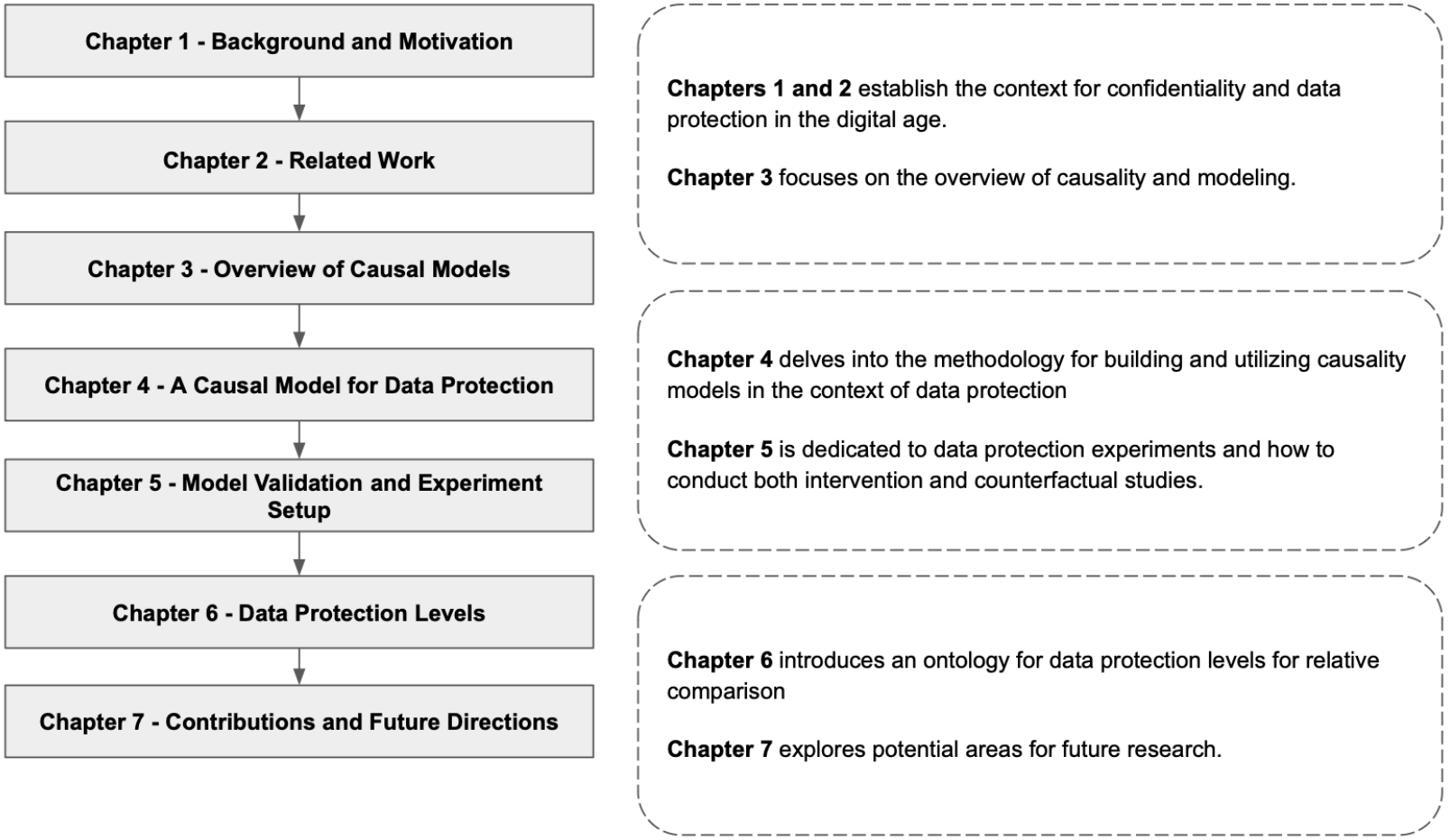
Chapter 4 delves into the methodology for building and utilizing causality models in the context of data protection. Chapter 5 is dedicated to validating the developed causality model. It discusses techniques to assess the model's accuracy, reliability, and generalizability.

Chapters 6-7: Experiments and Future Research Directions

Chapter 6 explores general experiments for conducting intervention and counterfactual studies to understand and better explain data protection. It outlines methods to evaluate internal and external validity, as well as techniques for visualizing and quantifying the results. Chapter 7 explores potential areas for future research.

Figure 1

The organization of the research conducted.



Research Methods

Chapter 2 - Overview of Causality

Chapter 3 provides an overview of causal models, starting with the history of causality and the ladder of causation. It explains the different types of causes (necessary, sufficient, contributory) and explores paradoxes that challenge our understanding of cause-and-effect. The chapter then introduces structural causal models (SCMs), including directed acyclic graphs (DAGs) and causal Bayesian networks, and discusses causal algorithms like the adjustment formula, front-door, and back-door criteria, and d-separation. It also covers causal mediation analysis (CMA) for understanding mediating variables and the concept of transportability for generalizing causal findings. The chapter concludes by presenting a framework for causal model creation in cybersecurity, emphasizing the importance of systemizing knowledge, model validation, and impact analysis.

Chapter 3 - Related Work: Systemization of Knowledge

In Chapter 2, the literature review presents a comprehensive review of emerging research to establish a robust knowledge base. A backward search method was employed, examining top security conferences for relevant research to data protection. Complementarily, a forward search was conducted from seminal works in the field. This iterative process continued until no new significant contributions were identified, reaching a saturation point. Given the volume of research, papers with high relevancy and citations were chosen first. Foundational papers or research with formal proofs without sufficient modeling in real-world threat or vulnerability scenarios were omitted. The research findings were then organized into a structured framework, or taxonomy, to facilitate the measurement of data protection. This taxonomy encompassed five

key categories: access control, system use, information disclosure, data modification, and data destruction, providing a systematic way to evaluate and understand the various aspects of data protection.

Chapter 4 - A Causal Model for Data Protection

Chapter 4 introduces a causal model for data protection, utilizing Judea Pearl's causal inference framework to understand the complex interplay of factors that influence data exposure. The chapter emphasizes quantifying the impact of security solutions and identifying hidden factors. It presents a general model showcasing relationships between threats, security measures, and data exposure, further refined by incorporating specific measurements from existing research. The chapter concludes by offering causal models for various aspects of data protection, including authorized access, system use, information disclosure, data modification, and data destruction, and proposes metrics for measuring data protection.

Chapter 5 - Model Validation and Experiment Setup

Chapter 5 focuses on validating the causal models developed in the previous chapter. It underscores the importance of model testing and causal search in comprehensively understanding the data protection of a system. By rigorously applying these techniques across the pillars of data protection – authorized access, system use, information disclosure, data destruction, and data modification – researchers can gain deeper insights into the complex causal relationships at play. This systematic evaluation not only validates the proposed causal models but also uncovers potential vulnerabilities and areas for improvement, ultimately facilitating the development of more effective security measures to protect sensitive information.

Chapter 6 Data Protection Levels

Chapter 6 introduces a framework for classifying the capabilities and behavior of data protection security mechanisms and systems. The framework introduces levels of data protection system behavior and ability to respond to a variety of threat actor intrusions and cyber attacks. It provides a common language to do relative comparisons and measure progress across the research community in developing robust data protection strategies.

Limitations

Causal models, while offering valuable insights into factors influencing data protection, inherent limitations common to causal inference. Pearl's work highlights the necessity of assumptions about the data-generating process, such as causal sufficiency and correct model specification, to establish true causality. In the context of data protection, this translates to potential challenges in accounting for all relevant factors, including hidden vulnerabilities or subtle interactions between variables. Additionally, the dynamic nature of cybersecurity threats and evolving attack vectors may necessitate frequent model updates to maintain accuracy. Finally, while causal models can guide intervention design, their effectiveness hinges on the validity of the underlying assumptions and the fidelity with which the model represents the real-world system. Specific limitations include the frequent assumption of causal sufficiency, the focus on DAGs to represent causal relationships, the partial treatment of transportability, computational challenges in implementing some methods, and the limited scope of Pearl's framework in fully addressing all aspects of causal reasoning, such as understanding causal mechanisms or incorporating causal knowledge into decision-making.

Disclosures, Biases, and Influences

This work explores the multifaceted nature of cybersecurity research, drawing upon diverse philosophical perspectives to illuminate the challenges and opportunities in measuring and understanding this complex domain. It examines the influence of leading cybersecurity thinkers and the role of interdisciplinary collaboration in shaping a measurement-driven approach to cybersecurity, while also reflecting on how my personal experiences and viewpoints may introduce potential biases into the research process. Furthermore, it delves into the critical role of measurement in cyber attribution and national security, highlighting the importance of rigorous scientific inquiry in addressing real-world cybersecurity challenges. Finally, it acknowledges the potential biases inherent in a Western-centric perspective on data protection measurement and emphasizes the need for transparency and inclusivity in future research.

Philosophy of Science in Cybersecurity Research

Philosophy of science explores how we gain scientific knowledge and what makes that knowledge credible. Several key perspectives offer valuable insights into this question. Logical positivism emphasizes that knowledge comes from observation and logical analysis, requiring testable evidence (Schlick, 1936). Falsificationism, on the other hand, suggests that scientific theories can't be proven true, only disproven through testing, highlighting the importance of seeking evidence that might challenge a theory (Popper, 1959). Instrumentalism takes a more pragmatic view, suggesting that scientific theories are tools for predicting and explaining phenomena, and their value lies in their usefulness rather than their truthfulness (Dewey, 1925). Kuhn's idea of scientific revolutions argues that science advances through paradigm shifts, where old ways of thinking are replaced with new ones, acknowledging the influence of social factors in science (Kuhn, 1962). Finally, social constructivism proposes that scientific knowledge is

shaped by social and political factors, not just objective observation (Latour & Woolgar, 1979). Each of these perspectives offers a different lens for understanding the nature and progress of scientific knowledge.

My research is informed by a philosophy of science that blends logical positivism, social constructivism, and with a primary emphasis on instrumentalism. I believe in the power of observation and testing to understand cybersecurity phenomena, aligning with the core tenets of logical positivism. However, I also recognize that cybersecurity is not a purely natural science; it's a human-made construct influenced by social, cultural, and political factors, acknowledging the perspective of social constructivism. Furthermore, I see immense value in the instrumentalist view that theories are tools for understanding and solving problems, and their worth lies primarily in their usefulness rather than absolute truth. This resonates most strongly with my approach to cybersecurity research, as I prioritize the development of practical solutions and actionable insights. This combination of viewpoints may introduce biases into my research, particularly when it comes to defining what constitutes valid scientific inquiry in this domain. While I value the role of falsification in refining scientific theories, I don't believe that the inability to falsify a theory automatically disqualifies it as science, especially in a field like cybersecurity where the necessary tools and methodologies for testing may still be evolving. Ultimately, I view cybersecurity as an abstract concept that can only be understood by considering the interplay of technical systems, threats, vulnerabilities, and the societal context in which they exist.

A Measurement-Driven Approach: Insights from Academia and Government

My research has been driven by a fundamental conviction: cybersecurity must be grounded in the rigor of scientific inquiry, with measurement as its cornerstone. This belief has

been shaped by the profound insights of leading cybersecurity experts like Dr. Spafford, Dr. Rogers, Dr. Rayz, and Dr. Dykstra, and nurtured within the interdisciplinary environment of CERIAS. Their influence has guided my exploration of how measurement can transform cybersecurity from a reactive practice to a proactive science, capable of anticipating, mitigating, and responding to threats with greater precision and effectiveness. This journey has led me to delve into the intricacies of system integrity, the human factors in security, the potential of artificial intelligence, and the critical role of measurement in cyber attribution, all in pursuit of a more secure and resilient digital world.

My experiences within the US government, specifically working to solve attribution problems by identifying and measuring key aspects, have deeply reinforced my belief in a measurement-driven approach to cybersecurity. Witnessing firsthand how precise measurements can lead to accurate attribution, and ultimately inform effective policy decisions, has solidified my conviction that cybersecurity can and should be treated as a rigorous science. The successes achieved through meticulous data collection and analysis in cases like Agent.btz, and the subsequent development of initiatives like CNCI and the rise of threat intelligence, all underscore the power of measurement in strengthening national cybersecurity.

My cybersecurity research has been significantly shaped by the viewpoints of esteemed researchers like Dr. Spafford, Dr. Rogers, Dr. Rayz, and Dr. Dykstra, along with the interdisciplinary environment fostered by CERIAS. These influences have converged to solidify my belief in establishing cybersecurity as a rigorous science through the power of measurement. Dr. Spafford's pioneering work on system integrity and his emphasis on a holistic approach to security have been instrumental in shaping my research perspective. His focus on going beyond traditional security measures and considering the human element resonates deeply with my own

approach. Inspired by his work, I strive to incorporate measurement-based evaluations of security mechanisms, ensuring they address not just technical vulnerabilities but also the human factors that often contribute to security breaches. This involves quantifying the effectiveness of security controls in the context of real-world scenarios, considering how users interact with systems.

Dr. Rogers' expertise in cyber forensics and the psychology of cybercrime has further enriched my understanding of the human dimension in cybersecurity. His research on cybercriminal motivations and the role of human behavior in security incidents has highlighted the importance of incorporating psychological insights into security research. This has led me to explore how measurement can be used to assess the effectiveness of security awareness training and to develop more robust methods for detecting and preventing social engineering attacks. By quantifying the impact of human behavior on security outcomes, we can develop more effective interventions and mitigate the risks posed by human vulnerabilities.

Dr. Rayz's work in Artificial Intelligence, particularly on natural language processing and human-computer interaction, has broadened my perspective on the potential for applying computational techniques to enhance cybersecurity. Her research on extracting meaning from text and understanding the nuances of human language has inspired me to explore how these techniques can be used to analyze security-related data, such as incident reports and threat intelligence. By applying NLP and machine learning to these data sources, we can potentially identify patterns and insights that would be difficult to discern through manual analysis alone. This could lead to the development of more sophisticated tools for threat detection, incident response, and security automation.

Dr. Dykstra's emphasis on applying scientific methods to cybersecurity and his advocacy for knowledge sharing and collaboration have reinforced my commitment to establishing cybersecurity as a rigorous science. His book, "Essential Cybersecurity Science," serves as a valuable guide for conducting empirical research in cybersecurity and has inspired me to incorporate experimental methods into my own work. This involves designing and conducting experiments to test the effectiveness of security controls, measure the impact of security interventions, and evaluate the performance of security tools. By adopting a scientific approach, we can generate evidence-based insights that contribute to the advancement of cybersecurity as a field.

Dr. Dana Madsen, a seasoned cyber intelligence expert with over 25 years of experience in the US government and military, was the Deputy Director of the Cyber Threat Intelligence Integration Center (CTIIC). He previously served as CTIIC's National Intelligence Manager for Cyber, where he spearheaded significant advancements, including the publication of a unified cyber intelligence strategy and substantial investments in cyber capabilities within the Intelligence Community. Madsen has a proven track record of leadership in various cyber and counterintelligence programs, including roles at the CIA where he developed and revitalized key initiatives such as helping to solve the cyber attribution measurement problem with Robert Morton among others in the Intelligence Community. His expertise spans geopolitical, technical, and policy aspects of cyber threats.

The interdisciplinary environment at CERIAS, with its focus on research, education, and engagement, has provided fertile ground for my research to flourish. The center's commitment to addressing the growing challenges in information security through collaboration and knowledge sharing has fostered a culture of innovation and intellectual curiosity. This has encouraged me to

explore new research directions, collaborate with researchers from diverse backgrounds, and contribute to the development of solutions that address real-world security problems.

National security decisions demand a "reasonable standard," requiring thorough intelligence assessments, consideration of risks and harms, and adherence to legal constraints. This principle, rooted in administrative law and intelligence directives like the National Security Act of 1947 (50 U.S. Code § 3021, 2018), necessitates evidence-based judgments, which rely heavily on accurate cyber attribution. Knowing what to measure in the process of attribution is crucial. It's not simply about collecting data; it's about identifying and analyzing the specific digital artifacts that can reliably link an attack to its source. This understanding has been pivotal in both identifying attackers and empowering policymakers to take action.

Early cases like Moonlight Maze (1996-1998) highlighted the challenges of attribution when knowledge of what to measure was limited. The attackers' sophisticated techniques hindered investigators' ability to definitively trace the attack to its source (Doman, 2018). However, the Agent.btz case (2008) marked a turning point. Investigators successfully linked the attack to Russian actors by identifying specific code similarities and attack infrastructure (FBI & DHS, 2016). This success stemmed partly from a better understanding of which digital artifacts to prioritize for analysis, ultimately enabling policymakers to respond with diplomatic pressure, sanctions, and other measures.

The Comprehensive National Cybersecurity Initiative (CNCI), launched in 2008, recognized the importance of measurement in cyber attribution. By establishing standardized procedures for data collection and analysis, the CNCI aimed to improve the government's ability to identify attackers and understand their tactics ("The Comprehensive National Cybersecurity Initiative," 2011). It also emphasized developing a skilled cybersecurity workforce, including

professionals trained in digital forensics and attribution techniques. This focus on expertise further enhanced the ability to know what to measure and how to interpret the data.

The rise of the threat intelligence profession is directly linked to this need for effective measurement in cyber attribution. Threat intelligence analysts specialize in collecting, analyzing, and interpreting data related to cyber threats, helping policymakers understand the evolving threat landscape and make informed decisions. In essence, knowing what to measure in cyber attribution is fundamental to both identifying attackers and unlocking effective policy decisions. By understanding which digital artifacts are most relevant and how to analyze them, investigators can provide policymakers with the evidence needed to make reasoned judgments and take action to protect national security.

Data Protection Measurement: A Western-Centric Perspective

This research on data protection measurement was conducted with a specific focus on Western legal frameworks. It's important to acknowledge that this emphasis may have inadvertently introduced a bias towards Western philosophical perspectives in both the literature review and the subsequent development of measurement tools. The influence of this bias could potentially limit the applicability or generalizability of the research findings to non-Western contexts, where different cultural, legal, and ethical considerations may be relevant. Furthermore, it is essential to disclose that this research received funding from the US intelligence community and the National Science Foundation. A portion of this funding was explicitly allocated to support research that could yield insights into cybersecurity, with particular emphasis on its applications to US national security. This funding source may have influenced the research direction, priorities, and potentially even the interpretation of findings. While the research aims to contribute broadly to the understanding and measurement of data

protection, it's crucial to recognize the potential influence of national security interests on the research process and outcomes.

The focus on Western legal frameworks in this research may limit the applicability of its findings to other legal and cultural contexts, potentially overlooking non-Western perspectives on data protection and privacy. Furthermore, the selection of literature, research questions, and methodologies could be subtly influenced by Western-centric assumptions.

Funding Influences

The research's funding sources, particularly from the US intelligence community, might also shape the research agenda and priorities, leading to a narrower focus on cybersecurity threats relevant to US national security, and potentially influencing the interpretation of findings. Nonetheless, these disclosures highlight the importance of transparency in scientific inquiry, encouraging critical evaluation and a more comprehensive understanding of data protection measurement. Future research should strive for inclusivity and cultural sensitivity, expanding the literature review to include non-Western perspectives, considering cultural nuances in the concept of data protection, and engaging diverse stakeholders in the research process.

CHAPTER 2: OVERVIEW OF CAUSAL MODELS

History of Causality

In the annals of causal reasoning, ancient philosophers like Aristotle laid the groundwork by proposing four causes – material, formal, efficient, and final – though these focused more on explaining the nature of things than on establishing causality as we understand it today (Pearl & Mackenzie, 2018). Later, Hume's skepticism famously challenged the very notion of causality, arguing that we only observe constant conjunction rather than causation itself, thus fueling further inquiry into cause and effect (Pearl & Mackenzie, 2018).

The Enlightenment brought a shift towards quantifying associations, but often at the cost of conflating correlation with causation. Laplace's deterministic view of the universe left little room for free will or causal complexities (Pearl & Mackenzie, 2018). While Galton's work on heredity led to regression analysis, a powerful tool for quantifying associations, he frequently fell into the trap of equating correlation with causation (Pearl & Mackenzie, 2018). Fisher's contributions to experimental design and statistics were pivotal in establishing causality within controlled settings, yet the limitations of experiments in real-world scenarios persisted (Pearl & Mackenzie, 2018).

The debate surrounding smoking and lung cancer underscored the challenges of using observational data to establish causality. Hill and Doll's groundbreaking studies built a strong case for the causal link, but skeptics argued that other factors could explain the association (Pearl & Mackenzie, 2018). This led to the rise of the randomized controlled trial (RCT) as the gold standard for establishing causality, although RCTs are not always feasible or ethical, nor can they answer questions about past events or hypothetical scenarios (Pearl & Mackenzie, 2018).

The causal revolution brought a paradigm shift. Sewall Wright's path analysis introduced graphical models for representing causal relationships, paving the way for Pearl's later work (Pearl & Mackenzie, 2018). Pearl's development of Bayesian networks and the do-calculus provided a formal language for encoding causal knowledge, reasoning about interventions, and answering counterfactual queries. This marked a significant leap in causal inference, enabling us to tackle complex causal questions previously deemed intractable (Pearl & Mackenzie, 2018).

Ladder of Causation

The Ladder of Causation, a concept introduced by Judea Pearl (2018), is a framework that categorizes causal reasoning into three distinct levels or rungs: association, intervention, and counterfactuals.

Association (Observational)

This is the first and most basic rung, where we observe patterns and correlations in data (Pearl, 2018). It allows us to answer questions like "What does observing X tell me about Y?" but falls short of establishing causal relationships.

Intervention (Doing)

The second rung involves actively intervening in the system to manipulate a variable and observe the effect on another (Pearl, 2018). It addresses questions like "What happens to Y if I do X?" and enables us to estimate causal effects.

Counterfactuals (Imagining)

The top rung deals with hypothetical scenarios and "what if" questions (Pearl, 2018). It allows us to reason about alternate realities and answer questions like "What would have happened to Y had X been different?" Counterfactuals are essential for understanding individual cases and attributing blame or credit.

The Ladder of Causation emphasizes the increasing complexity and power of causal reasoning as we ascend its rungs (Pearl, 2018). While association is limited to observational data, intervention allows for controlled experiments, and counterfactuals enable us to explore hypothetical scenarios. Pearl's framework provides a valuable tool for understanding the different levels of causal inference and their implications for decision-making and scientific inquiry.

Types Of Causes

In the intricate tapestry of cause-and-effect relationships, understanding the nature of causality is paramount. Within this realm, there are various types of causes, each playing a distinct role in shaping outcomes (Pearl & Mackenzie, 2018).

Necessary Cause

A necessary cause is an event or condition that must occur for an effect to happen. Without the necessary cause, the effect will not occur, regardless of the presence of other factors (Mackie, 1965). An example is: Oxygen (X) is a necessary cause for fire (Y). Without oxygen, a fire cannot start or sustain itself.

Let X be the cause and Y be the effect. X is a necessary cause of Y if and only if: $P\ (Y \mid \neg\ X)\ =\ 0$ (2)

The probability of Y happening given that X does not happen is zero.

Sufficient Cause

A sufficient cause is an event or condition that, when present, guarantees the occurrence of the effect. Other factors may also be capable of causing the same effect, but the sufficient cause alone is enough (Mackie, 1965). An example is: A lightning strike (X) is a sufficient cause for a forest fire (Y). If lightning strikes a dry forest, it will ignite a fire, even though other factors (like human carelessness) could also cause a forest fire.

Let X be the cause and Y be the effect. X is a sufficient cause of Y if and only if: $P(Y \mid X)\ =\ 1$ (3)

The probability of Y happening given that X happens is one.

Contributory Cause

A contributory cause is an event or condition that increases the likelihood of an effect occurring, but it is neither necessary nor sufficient on its own. It contributes to the effect in conjunction with other factors (Rothman, Greenland, & Lash, 2008). An example is: Smoking (X) is a contributory cause of lung cancer (Y). While smoking increases the risk of lung cancer, it's not guaranteed to cause it, and lung cancer can also occur in non-smokers due to other factors.

Let X be the cause and Y be the effect. X is a contributory cause of Y if and only if: $0 < P(Y \mid X) > P(Y \mid \neg X)$ (4)

The probability of Y happening given that X happens is greater than the probability of Y happening given that X does not happen.

Paradoxes in Causality

Paradoxes in causality arise when our intuitive understanding of cause-and-effect relationships clashes with the predictions or implications of formal causal models (Pearl & Mackenzie, 2018). These paradoxes can be perplexing and even seem to defy logic, but they often serve as valuable tools for refining our understanding of causality and highlighting the subtleties of causal reasoning. Let's explore some prominent examples:

Simpson's Paradox

This paradox occurs when a trend appears in several different groups of data but disappears or reverses when these groups are combined (Pearl, 2018). It often arises due to confounding variables that influence both the cause and the effect. For instance, a medical treatment might appear effective in two separate hospitals but shows no overall benefit when data from both hospitals is aggregated, due to differences in patient severity between the hospitals. Simpson's Paradox reminds us of the importance of considering all relevant factors and potential confounders when analyzing causal relationships (Pearl, 2018).

The Grandfather Paradox

This classic time travel paradox arises when a time traveler goes back in time and kills their grandfather before their parents were conceived, seemingly preventing their own existence

(Deutsch & Lockwood, 1994). This creates a logical contradiction: if the time traveler never existed, they couldn't have gone back in time to kill their grandfather. The Grandfather Paradox raises fundamental questions about the nature of time, causality, and the possibility of altering the past. It challenges our linear understanding of time and suggests the existence of multiple timelines or self-correcting mechanisms to prevent such paradoxes.

Newcomb's Paradox

This thought experiment involves a game with two boxes: one transparent containing a visible $1,000, and one opaque that may contain either $1,000,000 or nothing (Nozick, 1969). A superintelligent predictor, who has almost always been right in the past, has already predicted whether you will take only the opaque box or both boxes. If they predict you'll take both, they leave the opaque box empty. If they predict you'll take only the opaque box, they put $1,000,000 in it. The paradox lies in the conflict between two seemingly rational strategies: taking both boxes (maximizing immediate gain) and taking only the opaque box (trusting the predictor and potentially getting the larger reward). Newcomb's Paradox highlights the complexities of decision-making under uncertainty and the potential conflict between free will and determinism.

The Monty Hall Problem

This probability puzzle involves a game show with three doors: behind one is a car, and behind the other two are goats (vos Savant, 1990). You pick a door, and the host, who knows where the car is, opens another door to reveal a goat. You're then given the option to switch to the remaining closed door or stick with your original choice. Counterintuitively, switching doors doubles your chances of winning the car. This paradox challenges our intuition about probability and highlights the importance of updating our beliefs based on new information.

Paradoxes in causality serve as valuable thought experiments that push the boundaries of our understanding. They force us to confront the limitations of our intuitive reasoning and expose the subtleties of causal relationships. By grappling with these paradoxes, we can refine our causal models, develop more robust methods for causal inference, and make more informed decisions in complex and uncertain situations.

Structural Causal Models (SCMs)

Structural Causal Models (SCMs) are a powerful mathematical framework for representing and reasoning causal relationships between variables (Pearl, 2009). They combine graphical models, like Directed Acyclic Graphs (DAGs), with functional relationships to provide a clear and interpretable way to express causal assumptions and derive testable implications (Pearl, 2009). Structural Causal Models provide a powerful and versatile framework for representing and reasoning about causal relationships (Pearl & Mackenzie, 2018). By combining graphical models with functional relationships, they allow for a deeper understanding of complex systems and facilitate evidence-based decision-making (Pearl & Mackenzie, 2018). However, it is essential to remember that the validity of causal inferences depends on the accuracy of the causal assumptions encoded in the model (Pearl, 2009).

Directed Acyclic Graph (DAG)

Structural causal models (SCMs), powerful tools for understanding causal relationships, are built on several core components. At their heart lies the directed acyclic graph (DAG), where nodes symbolize variables and directed edges signify direct causal influences (Pearl, 2009). The absence of cycles guarantees that a cause cannot affect itself through a chain of causal relationships. Each node within the DAG is associated with a function that determines its value

based on its direct causes (parents in the graph) and an exogenous error term capturing unobserved factors (Pearl, 2009). These functions encapsulate the causal mechanisms within the system. Additionally, exogenous variables represent external or background conditions influencing the system, while endogenous variables have values determined by other variables within the model (Pearl, 2009).

Causal Bayesian Networks

Causal Bayesian networks (CBNs), an extension of Bayesian networks, incorporate causal semantics by interpreting directed edges as causal relationships (Pearl, 2009). They enable causal reasoning, including predicting intervention effects and answering counterfactual queries (Pearl, 2009). However, CBNs necessitate strong causal assumptions and may not be suitable for complex systems with feedback loops or unmeasured confounders (Pearl, 2009).

Directed Paths and Causal Effect Estimation

Within SCMs, graphs and paths play a crucial role. Paths, sequences of connected edges regardless of direction, help identify potential causal relationships (Pearl, 2009). Directed paths, where edges point in the same direction, represent causal chains (Pearl, 2009). Backdoor paths, non-causal paths with an arrow pointing into the cause variable, can create spurious correlations and need to be controlled for in causal effect estimation (Pearl, 2009). D-separation, a graphical criterion, determines the independence of two variable sets given a third, crucial for identifying conditional independencies and guiding causal inference (Pearl, 2009).

The applications of SCMs are vast. They allow researchers to estimate causal effects from observational data by identifying and controlling for confounders (Pearl, 2009). SCMs can be used for prediction tasks, incorporating the system's causal structure, and generating

explanations for observed phenomena by identifying causal pathways (Pearl & Mackenzie, 2018). Additionally, they can guide decision-making by simulating intervention effects and evaluating policy options (Pearl & Mackenzie, 2018).

Causal Algorithms

Causal inference seeks to establish cause-and-effect relationships from observational data (Pearl, 2009). A crucial tool in this endeavor is the adjustment formula, which helps identify and control for confounding variables to estimate the causal effect of one variable on another (Pearl, 2009). These causal inference algorithms, along with the adjustment formula and graphical criteria like the backdoor and front door criteria (Pearl, 2009), empower researchers to uncover causal relationships from observational data. By carefully considering confounding variables and utilizing appropriate techniques, we can move beyond mere associations and gain deeper insights into the true causal mechanisms underlying complex phenomena.

Adjustment Formula

The adjustment formula allows us to estimate the causal effect of a treatment (X) on an outcome (Y) by adjusting for a set of confounding variables (Z). It is based on the fundamental idea of controlling for common causes of both the treatment and the outcome to isolate the true causal effect (Pearl, 2009).

$$P(Y = y \mid do(X = x)) = \sum_{z} P(Y = y \mid X = x, Z = z)\, P(Z = z) \quad (5)$$

where:

- $P(Y = y \mid do(X = x))$ is the causal effect of setting X to x on the probability of Y being y.
- $\sum_{z}$ denotes the sum over all possible values of the confounding variables Z.
- $P(Y = y \mid X = x, Z = z)$ is the conditional probability of Y being y given that X is set to x and Z takes on a particular value z.
- $P(Z = z)$ is the probability of Z taking on the value z in the observed data.

Frontdoor Criterion

The front door criterion provides another graphical condition for identifying a causal effect, even when there are unmeasured confounders. It leverages a specific causal structure known as the "front door path" (Pearl, 2009).

Criterion: A set of variables M satisfies the front door criterion relative to (X, Y) if:

- M intercepts all directed paths from X to Y.
- There is no unblocked backdoor path from X to M.
- All backdoor paths from M to Y are blocked by X.

Implication: If M satisfies the front door criterion, then we can estimate the causal effect using the following formula:

$$P(Y = y \mid do(X = x)) = \sum_{m} P(M = m \mid do(X = x)) \sum_{x'} P(Y = y) \mid M = m, X = x')P(X = x') \quad (6)$$

Conditional Interventions

Conditional interventions involve setting a variable to a specific value conditional on the values of other variables. This allows us to explore the causal effects of interventions that depend on the context or state of the system (Pearl, 2009).

$$P(Y = y \mid do(X = x \mid Z = z) \tag{7}$$

This represents the causal effect of setting X to x, given that Z takes on the value z.

Covariate-Specific Effects

Covariate-specific effects examine how the causal effect of a treatment varies across different subgroups defined by the values of covariates (pre-treatment variables) (Pearl, 2009).

$$P(Y = y) \mid do(X = x), Z = z) - P(Y = y) \mid do(X = x'), Z = z) \tag{8}$$

This represents the difference in the causal effect of setting X to x versus x' , for the subgroup where $Z = z$.

Backdoor Criterion

The backdoor criterion provides a graphical condition for identifying a sufficient set of variables (Z) to adjust for in order to eliminate confounding bias and obtain an unbiased estimate of the causal effect (Pearl, 2009).

Criterion: A set of variables Z satisfies the backdoor criterion relative to (X, Y) if:

- No node in Z is a descendant of X.

- Z blocks all backdoor paths from X to Y. (A backdoor path is a path that starts with an arrow pointing into X.)

Implication: If Z satisfies the backdoor criterion, then we can use the adjustment formula to estimate the causal effect:

$$P(Y = y) \mid do(X = x)) = \sum z\, P(Y = y) \mid X = x, Z = z)\, P(Z = z) \quad (9)$$

D-Separation

D-separation, or directional separation, is a graphical criterion used in causal Bayesian networks to determine whether two sets of variables are conditionally independent given a third set of variables [Pearl, 1988]. It plays a fundamental role in identifying causal relationships and making inferences about the effects of interventions in a system. The significance of d-separation in causality models is multifaceted. It acts as a powerful tool for identifying causal effects by precisely pinpointing the variables that become independent after an intervention, thus revealing the true impact of that intervention on the system. Additionally, it simplifies complex models by uncovering conditional independencies, making them more manageable and computationally efficient. Furthermore, d-separation aids in guiding experimental design by identifying the necessary control variables, ensuring that causal inferences drawn from experiments are valid and reliable.

D-Separation Criterion (10)

A path between two nodes X and Y is blocked (and hence X and Y are d-separated) given a set of nodes Z if and only if there is a node W on the path such that either:

1. **Chain or Fork:** W is in Z, and the arrows on the path meet head-to-tail or tail-to-tail at W.
2. **Collider:** Neither W nor any of its descendants are in Z, and the arrows on the path meet head-to-head at W.

Causal Meditation Analysis (CMA)

In the realm of causal inference, it's often not enough to simply establish that a treatment causes an effect (Pearl, 2009). We also want to understand how this effect occurs – the mechanisms through which the treatment influences the outcome. This is where Causal Mediation Analysis (CMA) steps in, providing a framework to identify and quantify the role of intermediate variables (mediators) in the causal pathway (Imai et al., 2010). Causal Mediation Analysis is a valuable tool for understanding the mechanisms through which treatments influence outcomes. By identifying and quantifying the role of mediators, CMA can provide deeper insights into causal relationships and guide the development of more targeted interventions.

Mediators

These are variables that lie on the causal path between the treatment (X) and the outcome (Y) (VanderWeele, 2015). They transmit part or all of the treatment's effect on the outcome. Below are the key quantities in CMA analysis.

Total Effect (TE)

The overall effect of the treatment on the outcome, combining both direct and indirect effects. Mathematically, the total effect of X on Y can be decomposed into:

$$Total\ Effect\ (TE)\ =\ Direct\ Effect\ (DE)\ +\ Indirect\ Effect\ (IE) \quad (11)$$

This quantifies the effect of the treatment on the outcome when the mediator is held constant at its natural value (i.e., the value it would have taken in the absence of the treatment) (Pearl, 2001).

General Mathematical Formulation for CMA Let's consider the following notation:

Y: Outcome variable

X: Treatment variable

M: Mediator variable $Y(x, m)$:

Potential outcome of Y if X is set to x and M is set to m $M(x)$:

Potential value of the mediator MM if XX is set to x 'x'

Represents the overall effect of changing the treatment from x to $x*$ on the outcome.

$$TE = E\,[Y(x*, M(x*)) - Y(x, M(x))] \quad (12)$$

This is often estimated using standard regression or other statistical techniques, comparing the average outcome in the treatment group ($X = x*$) to the control group ($X = x$).

Natural Direct Effect (NDE)

Represents the effect of changing the treatment from x to $x*$ on the outcome, while keeping the mediator at its baseline level when the treatment is x, i.e., $M(x)$

Mathematically: $NDE = E[Y(x*, M(x) - Y(x, M(x))]$ This is typically estimated using the following formula (assuming no unmeasured confounders):

$$NDE = \sum m[EY \mid X = x*, M = m] - E[Y \mid X = x, M = m]] P(M = m \mid X = x) \quad (13)$$

We sum over all possible values of the mediator m. For each m, we compute the difference in the expected outcome between the treatment and control groups, while keeping M at m We weight this difference by the probability of observing that mediator value m in the control group ($X = x$).

Natural Indirect Effect (NIE)

Represents the effect of changing the mediator from its baseline level when the treatment is x 'x' (i.e., $M(x)$ to its level when the treatment is $x*$ (i.e., $Mx*$) ,while keeping the treatment fixed a xt '. Mathematically: $NIE = E[Y(x*, M(x*)) - Y(x*, M(x))]$. This is typically estimated using the following formula (assuming no unmeasured confounders):

$$NIE = \sum m[Y \mid X = x*, M = m][P(m = m \mid X = x*) - P(M = m \mid X = x)] \quad (14)$$

We sum over all possible values of the mediator 'm' For each 'm', we compute the expected outcome in the treatment group ($X = x*$) . We weight this outcome by the difference in the probability of observing 'm' between the treatment and control groups Estimation Methods Several methods exist to estimate these causal effects, including: Baron and Kenny's Approach: This traditional approach involves a series of regression analyses to estimate

the direct and indirect effects (Baron & Kenny, 1986). However, it relies on strong assumptions and may not be suitable in the presence of confounders or interactions. The Causal Mediation Formula and its approach, based on the potential outcomes framework, provides a more general and robust way to estimate the natural direct and indirect effects (Imai et al., 2010). It leverages counterfactual reasoning to quantify the effects of the treatment and mediator under different hypothetical scenarios.

Controlled Direct Effect (CDE)

This measures the effect of the treatment on the outcome when the mediator is fixed at a specific value, regardless of its natural value (Pearl, 2001).

Sensitivity Analysis

This technique assesses the robustness of the mediation analysis results to potential violations of the key assumptions, such as the absence of unmeasured confounding between the mediator and the outcome (Imai et al., 2011).

Example

Consider a study examining the effect of exercise (X) on heart health (Y), with diet (M) as a potential mediator. CMA can help determine how much of the effect of exercise on heart health is mediated through its impact on diet, and how much is a direct effect of exercise itself.

Transportability

Beyond simple binary outcomes, causal inference often deals with scenarios where the treatment and outcome variables can take on multiple values or even be continuous (Imbens &

Rubin, 2015). In such cases, generalized metrics are employed to capture the nuanced and complex relationships between cause and effect. These generalized metrics offer a more nuanced understanding of causal relationships, especially when dealing with complex and heterogeneous treatment effects. By employing these measures, researchers can gain deeper insights into the mechanisms of causality and make more informed decisions about interventions and policies.

Threshold Effects Concept

Threshold effects occur when the causal impact of a treatment is only observed beyond a certain threshold or critical value. Below this threshold, the treatment may have no or minimal effect, while above it, the effect becomes significant. A certain dosage of a drug may be required to achieve a therapeutic effect, while lower dosages might be ineffective. Emissions reductions might only lead to significant improvements in air quality beyond a specific threshold. Threshold effects can be identified and quantified using regression discontinuity designs or dose-response curves (discussed next) (Imbens & Lemieux, 2008).

Dose-Response Curves

Dose-response curves depict the relationship between the level or intensity of a treatment (dose) and the magnitude of the effect (response). These curves can reveal non-linear relationships, saturation points, and optimal treatment levels. Dose-response curves are essential for determining safe and effective dosages of medications. They can be used to assess the impact of different levels of public health programs or policies. Dose-response curves are typically estimated using regression models or other statistical techniques that can capture non-linear relationships.

Fraction of Attributable Risk (FAR)

FAR quantifies the proportion of the risk of an outcome that can be attributed to a specific exposure or treatment. It helps assess the public health impact of an exposure and can guide interventions (Greenland & Robins, 1988). FAR represents the percentage reduction in the risk of the outcome that would be achieved if the exposure were completely eliminated.

$$FAR = P(Y = 1) - P(Y = 1 \mid \neg X)) / P(Y = 1) \quad (15)$$

where

$P(Y = 1)$ is the probability of the outcome occurring in the general population

$P(Y = 1 \mid \neg X)$ is the probability of the outcome occurring in the absence of the exposure

Interpretation:

Average Causal Effect (ACE)

ACE represents the average causal effect of a treatment on an outcome across the entire population. It's a useful summary measure when the treatment effect might vary across individuals (Imbens & Rubin, 2015).

$$ACE = E[Y(1) - Y(0)] \quad (16)$$

where

$Y(1)$ is the potential outcome if the treatment is received

$Y(0)$ is the potential outcome if the treatment is not received

$E[]$ denotes the expectation or average over the entire population

Effect of Treatment on the Treated Population (ETT)

ETT focuses on the average causal effect of the treatment specifically among those who actually received it. This is relevant when there might be selection bias or when the treatment effect is heterogeneous across different groups (Heckman, 1997).

$$ETT = E[Y(1) - Y(0) \mid X = 1] \quad (17)$$

where

The conditioning on $X = 1$ indicates that the average is taken only among those who received the treatment.

Generalizability

In the realm of causal inference, the ability to generalize causal findings from one domain or setting to another is known as transportability. It addresses the fundamental question: "Can we trust the results of a study conducted in one population when applied to a different population?" Elias Bareinboim's research on transportability has made significant contributions to the field of causal inference, enabling researchers to generalize causal findings across different domains and make more informed decisions in complex and diverse settings (Bareinboim & Pearl, 2016). By providing a formal framework and practical tools, his work has expanded the scope and applicability of causal inference, with far-reaching implications for various fields of science and decision-making. Bareinboim's framework for transportability is generalizable to various scenarios where causal findings need to be applied across different domains or populations. It has implications for: Generalizing findings from randomized controlled trials (RCTs) to real-

world settings: RCTs often have strict inclusion/exclusion criteria, leading to selection bias. Transportability can help generalize the findings to broader populations. Combining data from multiple studies or sources: When data comes from different studies of populations, transportability can help integrate and synthesize the findings. Making predictions in new environments: When deploying machine learning models or making policy decisions in new settings, transportability can assess the validity of applying findings from previous contexts.

Cyclical Causality in Time Series Data: A Probabilistic Approach

In the realm of dynamic systems where time and probability intertwine to shape causality, Probabilistic Computation Tree Logic (PCTL) emerges as a powerful framework for understanding cause-and-effect relationships. Unlike traditional approaches that often oversimplify causal relationships, PCTL explicitly incorporates temporal and probabilistic operators, allowing for a more nuanced and precise representation of how events unfold over time (Baier & Katoen, 2008; Hansson & Jonsson, 1994). This capability is particularly crucial in cybersecurity, where understanding the timing and likelihood of events, including complex interactions and feedback loops, is essential for effective risk assessment and mitigation. To facilitate the analysis of such systems, PCTL often employs Kripke structures, a type of mathematical model that represents system behavior in terms of its possible states and transitions between them (Clarke et al., 2018). This combination of PCTL and Kripke structures provides a rigorous and versatile toolset for analyzing and verifying causal properties in dynamic systems, enabling a deeper understanding of the intricate interplay between time, probability, and causality.

Probabilistic Computation Tree Logic (PCTL) offers a robust framework for understanding causality, particularly in dynamic systems where time and probability play crucial

roles (Baier & Katoen, 2008). Unlike traditional causal frameworks, PCTL explicitly incorporates temporal operators, enabling the expression of cause-and-effect relationships with specific time constraints and probabilities (Hansson & Jonsson, 1994). This is essential for capturing the nuances of real-world causality, where effects may not be immediate or guaranteed. For instance, instead of simply stating "A causes B," PCTL allows for more precise expressions like "A causes B within 5 to 10 time units with a probability of at least 0.8" (Huth & Ryan, 2004). This capability is particularly valuable in cybersecurity, where understanding the timing and likelihood of events is critical for effective risk assessment and mitigation.

Furthermore, PCTL facilitates the analysis of time-series data, a common feature in cybersecurity research (Bowman & Lin, 2004). By combining temporal and probabilistic operators, PCTL enables researchers to analyze how events unfold over time and assess the likelihood of specific outcomes. This can be applied to various cybersecurity phenomena, such as investigating the impact of security patches on the probability of successful attacks or predicting the spread of malware within a network. Moreover, PCTL can be used in conjunction with model checking techniques to formally verify causal properties in a system (Clarke et al., 2018). This allows for rigorous analysis of system behavior and ensures that it conforms to expected causal relationships.

Importantly, PCTL can also address cyclical causality models, where A influences B, and B subsequently influences A. This is achieved through the use of temporal operators that can express sequences of events and feedback loops. For example, one could express a cyclical causal relationship as: "A causes B within 5 time units with probability 0.6, and B subsequently causes A within 10 time units with probability 0.8." This ability to model cyclical causality is crucial in cybersecurity, as many security threats involve complex interactions and feedback

loops, such as the interplay between vulnerability exploitation, attacker persistence, and defensive countermeasures.

In conclusion, PCTL provides a powerful and flexible approach to causal reasoning, especially in domains like cybersecurity where time and probability are critical factors. Its ability to handle temporal information, quantify uncertainty, enable formal verification, and address cyclical causality makes it a valuable tool for understanding and explaining complex causal relationships in dynamic systems.

A Kripke structure, named after Saul Kripke, is a type of mathematical model used to represent the behavior of a system in terms of its possible states and transitions between them (Clarke et al., 2018). It's essentially a directed graph where nodes represent states, and edges represent transitions. Here's a breakdown with the formal definition and an example:

Formal Definition (Clarke et al., 2018):

A Kripke structure M over a set of atomic propositions AP is a 4-tuple:

M = (S, I, R, L)

where:

S: A finite set of states.

I: A set of initial states ($I \subseteq S$).

R: A transition relation ($R \subseteq S \times S$) that is left-total, meaning every state has at least one outgoing transition.

L: A labeling function (L: S → 2^{AP}) that assigns to each state a set of atomic propositions that are true in that state.

Explanation:

States: Represent the different configurations or situations the system can be in.

Initial States: The states where the system can start.

Transition Relation: Defines how the system can move from one state to another.

Labeling Function: Assigns properties or characteristics to each state

Advances in Causality Research

Recent years have witnessed significant strides in causal inference, encompassing various facets of causality. Methodological advancements have been made in causal discovery from observational data, with algorithms like PCMCI and extensions of FCI and GES enabling the identification of causal links in complex datasets with hidden confounders (Chickering, 2002; Peters et al., 2016; Runge et al., 2019; Shimizu et al., 2011). Simultaneously, causal inference with interventions has seen a growing focus on "soft" interventions, where the treatment influences but doesn't fully determine the outcome, leading to new methods for analyzing real-world scenarios with nuanced causal relationships (Eberhardt, 2012; Jaber et al., 2018; Jin et al., 2024). Causal representation learning has emerged as a promising approach for enhancing machine learning models, with applications in offline reinforcement learning and few-shot learning demonstrating its potential to improve performance and robustness (Zhang et al., 2022; Yue et al., 2022). Furthermore, specialized techniques have been developed for addressing confounding and feedback in time series data, crucial for understanding complex systems with temporal dependencies (Peters et al., 2017; Wu et al., 2019; Zheng et al., 2018). Finally, the pursuit of explainable AI (XAI) has driven research towards both inherently interpretable models and methods for explaining black-box models, with a growing emphasis on human-centered approaches and domain-specific applications (Letham et al., 2013; Ribeiro et al., 2016; Simonyan et al., 2013; Sundararajan et al., 2017; Lundberg & Lee, 2017; Koh & Liang, 2017; De Cao et al., 2020; Wachter et al., 2017; Mothilal et al., 2020; Karimi et al., 2020; Doshi-Velez & Kim, 2017; Hoffman et al., 2018). The ability to generalize across a wide range

of domains is a hallmark of intelligence (Richens & Everitt, 2024). While it has been hypothesized that causal reasoning is necessary for this, recent work has shown that in some cases strong generalization can be achieved without causal reasoning (Richens & Everitt, 2024). Any agent capable of satisfying a regret bound for a large set of distributional shifts must have learned an approximate causal model of the data generating process, which converges to the true causal model for optimal agents (Richens & Everitt, 2024).

Causal Discovery from Observational Data

Key methods include the PCMCI algorithm, which effectively identifies causal links in high-dimensional time series datasets by combining constraint-based methods with information-theoretic measures to address nonlinear relationships and hidden confounders (Runge et al., 2019). The FCI algorithm has been extended to handle cases with non-independent noise variables, crucial for real-world scenarios with hidden confounders (Peters et al., 2016). Score-based methods like GES, which efficiently searches for causal graphs by optimizing a score function (Chickering, 2002), have also seen advancements. This includes extensions to handle non-Gaussian data, enabling more accurate causal discovery in diverse datasets (Shimizu et al., 2011). Functional causal models like LiNGAM, which assumes linear relationships and non-Gaussian noise (Shimizu et al., 2006), and post-nonlinear causal models, which accommodate nonlinear relationships and measurement distortions (Zhang and Hyvärinen, 2009), have further broadened the scope of causal discovery.

Causal Inference with Interventions

A key focus in causal inference research has been on understanding and leveraging "soft" interventions, where the treatment doesn't fully determine the outcome (Eberhardt, 2012). This is

crucial because many real-world interventions don't completely control a variable but rather shift its distribution. Eberhardt (2012) laid the groundwork for analyzing such interventions, moving beyond the traditional focus on interventions that set variables to fixed values. Building on this, Jaber et al. (2018) explored how to identify causal effects even when the exact causal structure is unknown by leveraging constraints on possible interventions. This is particularly valuable in complex systems where the underlying causal relationships are not fully understood. More recently, Jin et al. (2024) introduced Cladder, a benchmark designed to evaluate the ability of causal discovery algorithms to handle soft interventions and complex causal relationships, further advancing research in this area.

Causal Representation Learning

Recent advances in causal representation learning have shown promising results in improving the performance and robustness of machine learning models. Zhang et al. (2022) utilized this approach to enhance offline reinforcement learning by learning representations that remain invariant across different environments. This leads to more robust policies that can generalize better to unseen situations. In a different application, Yue et al. (2022) employed causal interventions to improve few-shot learning. By simulating various interventions on the training data, they enabled the model to learn more effectively from limited examples. These studies highlight the potential of causal representation learning to address challenges in diverse machine learning tasks by explicitly incorporating causal knowledge into the learning process.

Causality in Time Series

Dealing with confounding and feedback in time series data requires specialized causal inference techniques to disentangle true causal effects from complex temporal dependencies and

feedback loops (Peters, Mooij, Janzing, & Schölkopf, 2017). These techniques are crucial in fields like Earth system sciences, where understanding the causal relationships between variables like temperature, precipitation, and carbon dioxide levels is essential for predicting and mitigating the impacts of climate change (Wu et al., 2019). For instance, Wu et al. (2019) applied causal inference methods to analyze time series data in Earth system sciences, addressing challenges such as confounding variables and feedback loops that can obscure true causal relationships. Furthermore, research has explored methods for automatically identifying causal moderators in time-series data (Zheng, Claassen, & Kleinberg, 2018). These moderators can provide valuable insights into how different factors influence causal relationships, helping to refine our understanding of complex systems.

Explainable Artificial Intelligence (AI)

The pursuit of explainable Artificial Intelligence (AI) (XAI) has been a driving force in machine learning research, motivated by the need to understand and trust the predictions of increasingly complex models. Early work in this area, exemplified by Letham et al. (2013), focused on building inherently interpretable models, such as rule-based classifiers that offer transparent decision-making processes. This approach aimed to provide clear explanations by design, ensuring that the models themselves were readily understandable to humans. However, as the field progressed and more complex models like deep neural networks gained prominence, the focus shifted towards explaining existing black-box models. A pivotal contribution in this direction was LIME (Local Interpretable Model-agnostic Explanations) by Ribeiro et al. (2016), which provided a method for explaining any classifier locally by approximating it with a simpler, interpretable model in the vicinity of a specific prediction.

This era also saw the development of various post-hoc explainability and saliency methods aimed at shedding light on the inner workings of black-box models. Simonyan et al. (2013) laid the groundwork for many saliency map techniques, which highlight the input features most relevant to a model's prediction. Further advancements came with the introduction of Integrated Gradients by Sundararajan et al. (2017), a more theoretically grounded method for attributing feature importance based on integrating gradients along a path from a baseline input to the input of interest. Another influential method, SHAP (SHapley Additive exPlanations) by Lundberg and Lee (2017), connected game theory with local explanations to fairly distribute feature importance according to their contributions to the prediction.

As XAI research matured, there was a growing recognition that explainability needs to be tailored to specific domains and tasks. Koh and Liang (2017) delved into understanding model behavior by analyzing the influence of training data points on predictions, providing insights into how models learn and generalize. In the realm of natural language processing (NLP), De Cao et al. (2020) explored explainability by disentangling different factors in language representations, allowing for a deeper understanding of how language models process and generate text. This trend towards domain-specific XAI solutions highlights the increasing need for explanations that are relevant and meaningful within particular contexts, such as healthcare, finance, and law.

Beyond simply identifying correlations between features and predictions, XAI research has also ventured into the realm of counterfactual explanations and causal reasoning. Wachter et al. (2017) highlighted the importance of counterfactual explanations for understanding how to change a model's prediction, essentially answering "what-if" questions about the input. Mothilal et al. (2020) further advanced this area by focusing on generating diverse counterfactual explanations to provide a more comprehensive understanding of the model's decision boundary.

Karimi et al. (2020) provided a comprehensive survey of algorithmic recourse, a closely related field that explores how to provide actionable recommendations for individuals to change their outcomes based on a model's prediction.

Most recently, the field of XAI has seen a surge in research on evaluation and human-centered approaches. Doshi-Velez and Kim (2017) emphasized the need for rigorous evaluation metrics for XAI methods, arguing that explainability should be assessed based on its impact on human understanding and decision-making. Hoffman et al. (2018) explored the challenges of evaluating XAI and proposed potential metrics, recognizing the subjective nature of human interpretation. This reflects a growing recognition that XAI needs to be human-centered and aligned with human cognitive processes, taking into account how humans perceive, understand, and use explanations. Ongoing research in this area includes user studies, cognitive science perspectives, and ethical considerations in XAI, aiming to bridge the gap between technical explanations and human comprehension.

In parallel to these broader trends in XAI, Cynthia Rudin's research has consistently championed the development and use of interpretable machine learning models, particularly in high-stakes domains like healthcare and criminal justice (Rudin, 2019; Rudin & Bushway, 2021). Her work emphasizes the importance of transparency, accountability, and human understanding in the development and deployment of AI systems (Rudin, 2022; Rudin et al., 2022). She has made significant contributions to the creation of optimal rule lists and scoring systems (Ban & Rudin, 2019; Chen et al., 2021; Rudin & Ustun, 2018; Ustun & Rudin, 2016; Zeng, Ustun, & Rudin, 2016), sparse decision trees (Parikh, Rudin, & Volfovsky, 2022; Xin et al., 2022; Liu et al., 2020; Hu, Rudin, & Seltzer, 2019), and interpretable generalized additive models (Sun et al., 2024; Chen, Zhong, Seltzer, & Rudin, 2023).

While advocating against black-box models, Rudin also investigates making deep learning models more interpretable, particularly in image recognition, through techniques like concept whitening and case-based reasoning (Barnett et al., 2024; Donnelly et al., 2024; Yang et al., 2024; Barnett et al., 2021; Chen et al., 2020; Chen et al., 2019). Her research has found applications in various domains, including healthcare (Struck et al., 2017; Souillard-Mandar et al., 2016; Barnett et al., 2024; Parikh et al., 2024; Parikh et al., 2023; Ustun & Rudin, 2016; Letham, Rudin, McCormick, & Madigan, 2015), criminal justice (Rudin & Ustun, 2018; Garrett & Rudin, 2023; Wang et al., 2022; Rudin, Wang, & Coker, 2020; Zeng, Ustun, & Rudin, 2016), and finance (Chen et al., 2021; Rudin & Shaposhnik, 2023). Rudin's work not only advances the technical frontiers of interpretable machine learning but also emphasizes the ethical and societal implications of AI, advocating for models that are understandable and trustworthy for human users.

It has long been hypothesized that causal reasoning is necessary for robust and general intelligence (Richens & Everitt, 2024). However, recent empirical work has shown that agents can be robustly adaptive without explicitly learning or using causal models (Richens & Everitt, 2024). Any agent capable of satisfying a regret bound for a large set of distributional shifts must have learned an approximate causal model of the data generating process, which converges to the true causal model for optimal agents (Richens & Everitt, 2024). The implications of this result for several fields including transfer learning, causal inference and the practical design of causal and robust artificial agents (Richens & Everitt, 2024).

CHAPTER 3: RELATED WORK

Models for Cybersecurity

Cybersecurity, the practice of protecting computer systems and networks from digital attacks, has evolved into a critical aspect of modern life. As the digital landscape expands, so too does the complexity of the threats it faces. To effectively address these challenges, a robust framework for understanding and managing security risks is essential. Two foundational models in cybersecurity are the C-I-A Triad and the Parkerian Hexad.

The C-I-A Triad, a cornerstone of information security, provides a basic framework for understanding the core principles of protection. It encompasses three fundamental concepts: confidentiality, integrity, and availability. Confidentiality ensures that information is accessed only by authorized individuals. Integrity guarantees the accuracy and completeness of information, preventing unauthorized modification. Availability ensures that information and systems are accessible when needed.

While the C-I-A Triad offers a solid foundation, it has limitations in comprehensively addressing the multifaceted nature of information security. The Parkerian Hexad, proposed by Donn B. Parker, expands upon the C-I-A Triad by introducing three additional elements: possession, utility, and authenticity. Possession or control refers to the rightful ownership of information, emphasizing the importance of access control mechanisms. Utility captures the value of information in fulfilling its intended purpose. Authenticity ensures that information is genuine and can be trusted as being from the claimed source. provides a more holistic perspective on information security by incorporating elements that go beyond the traditional C-I-A Triad. It acknowledges the importance of ownership, value, and trust in safeguarding

information. By considering these additional factors, organizations can develop more robust security strategies.

However, it is essential to recognize that both the C-I-A Triad and the Parkerian Hexad represent foundational models. The cybersecurity landscape is constantly evolving, with new threats and technologies emerging. These models should be considered as starting points rather than definitive frameworks.

Other models and frameworks have been developed to address specific aspects of cybersecurity. For example, the NIST Cybersecurity Framework focuses on identifying, assessing, and managing cybersecurity risks. The ISO/IEC 27000 family of standards provides a comprehensive set of information security management practices.

In conclusion, the C-I-A Triad and Parkerian Hexad serve as valuable tools for understanding the fundamental principles of cybersecurity. While they offer a solid foundation, they do not encompass the entire spectrum of security challenges. A combination of these models, along with other frameworks and best practices, is necessary to develop a comprehensive and effective cybersecurity strategy. As the threat landscape continues to evolve, it is imperative to stay informed about emerging trends and adapt security measures accordingly.

The C-I-A Triad: A Cornerstone of Cybersecurity

The C-I-A Triad, comprising Confidentiality, Integrity, and Availability, is a foundational model in cybersecurity. This framework provides a structured approach to understanding and managing information security risks. Each component is critical for ensuring the protection and reliability of systems and data.

While Saltzer and Schroeder (1975) introduced related concepts—unauthorized disclosure, modification, and denial of use—they didn't explicitly define the triad. The National

Research Council (1991) popularized the C-I-A terminology without citing a specific source. Parker (1992) also used the C-I-A model, with some variations.

Subsequent works, including those by Cherdantseva, Hilton, Knipp, and Parker, attribute the formalization of the C-I-A Triad to the NASA Johnson Space Center's "Pink Book" (1989) and its authors, Leo, Tipton, and Owens. However, Leo himself has claimed authorship of the concept:

> "In early 1986, I discussed with Hal Tipton (my then contractor boss at Rockwell Intl.) and Rich Owen (my then NASA boss at JSC) what the basis of our InfoSec program ought to be. As ISSO, I had been giving much thought to this subject and had focused on the thing we should be trying to protect: the information itself. After much consideration, I hit on a theme that I thought would express to the Management in simple terms. The focus of our program would accurately characterize this vital asset's essential characteristics. It was from this that I came up with C-I-A.
>
> "I suggested this to my colleagues, and it shortly became our theme. I used this term in every presentation, every document, and every report after that to "spread the gospel" and eventually spread it to the other NASA centers. It mushroomed much further from there, as we have since seen, and of course, I had no expectation of how far it would travel or indeed how far it has traveled." (Bellovin, 2021).

Confidentiality is the cornerstone of information security, ensuring that sensitive data is accessed only by authorized individuals. It involves protecting information from unauthorized disclosure, use, modification, or destruction. Implementing robust access controls, encryption, and data classification are essential measures to maintain confidentiality. Access controls restrict access to systems and data based on user roles and permissions, preventing unauthorized

individuals from gaining entry. Encryption converts data into a coded format, making it unreadable to those without the decryption key. Data classification categorizes information according to its sensitivity level, enabling organizations to apply appropriate protection measures.

Confidentiality, one of the core pillars of the C-I-A Triad, ensures that sensitive information is accessed only by authorized individuals. Measuring confidentiality is challenging due to the intangible nature of information and the dynamic nature of threats. However, several methods can be employed to assess the protection of confidential data.

Measuring confidentiality is crucial for safeguarding sensitive information, and several key methods play a vital role in this endeavor. Risk assessments help identify and prioritize potential threats and vulnerabilities that could compromise confidentiality, while data classification allows organizations to categorize data based on sensitivity levels and implement appropriate protection measures. Access controls, including monitoring and auditing, along with encryption strength evaluation and key management practices, ensure that data remains secure both at rest and in transit. Data loss prevention (DLP) systems actively monitor and prevent unauthorized data transfers, while incident response and analysis provide valuable insights into weaknesses in confidentiality protection. Regular employee training and awareness programs help cultivate a security-conscious culture, and third-party risk management ensures that vendors and partners maintain adequate confidentiality practices. However, measuring confidentiality also presents challenges. Defining what constitutes confidential information can be complex, obtaining accurate data on breaches can be difficult, and quantifying the impact of a breach is often subjective. Moreover, the constantly evolving threat landscape demands continuous adaptation and improvement of confidentiality measures.

Integrity guarantees the accuracy and completeness of information, preventing unauthorized modification or corruption. It ensures that data is reliable and trustworthy. To maintain integrity, organizations must implement robust data validation, error detection, and correction mechanisms. Input validation checks data for accuracy and consistency before processing, preventing malicious input from compromising system integrity. Error detection and correction techniques identify and rectify data errors, ensuring data accuracy. Digital signatures can be used to verify the authenticity and integrity of messages and documents.

Several major methods contribute to measuring integrity in various domains. Data integrity checks, employing techniques like hash functions, checksums, and Cyclic Redundancy Checks (CRCs), ensure the accuracy and consistency of data by detecting any unauthorized modifications or corruption. System integrity monitoring, through tools like File Integrity Monitoring (FIM), system call monitoring, and Intrusion Detection Systems (IDS), helps identify anomalies and potential attacks that could compromise the integrity of the system. In the realm of databases, constraints and triggers within database management systems, along with data validation, ensure data integrity during storage and processing. Moreover, regular backups and effective recovery procedures are vital for restoring data integrity in case of unforeseen corruption.

Software integrity is also crucial, and measures like code signing, static and dynamic code analysis, and timely application of software updates and patches help maintain the integrity of software code by verifying authenticity, identifying vulnerabilities, and addressing security issues. In conclusion, a combination of these diverse methods, applied across different layers of a system, is essential for maintaining data and system integrity and safeguarding against potential threats and vulnerabilities.

Availability ensures that information and systems are accessible when needed. It requires reliable hardware, software, and network infrastructure. To maintain availability, organizations must implement redundancy, disaster recovery, and business continuity plans. Redundancy provides backup systems and components to prevent system failures. Disaster recovery plans outline procedures for restoring IT systems and data in case of a disaster. Business continuity plans focus on maintaining critical business functions during and after a disruption.

Availability, one of the three pillars of the C-I-A Triad, ensures that systems and data are accessible when needed. Measuring availability is critical for assessing system performance, identifying potential bottlenecks, and evaluating the effectiveness of disaster recovery plans. Several key metrics and methods are employed to assess availability:

Key metrics for measuring availability include uptime, which is the percentage of time a system is operational, Mean Time Between Failures (MTBF), which represents the average time between system failures, and Mean Time To Repair (MTTR), the average time needed to restore a system after a failure. Service Level Agreements (SLAs) also play a crucial role in defining contractual obligations for service levels and performance metrics. To measure availability, organizations employ various methods such as performance monitoring tools that collect data on system performance, log analysis to gain insights into system behavior and failures, and network monitoring to identify bottlenecks. Application Performance Monitoring (APM) tools focus specifically on application performance, while simulation and modeling can help predict system behavior under different loads. User experience monitoring provides valuable real-world feedback on system availability. However, measuring availability poses several challenges, including defining what constitutes system availability, collecting accurate and comprehensive

data, analyzing large volumes of data, and dealing with the dynamic nature of modern IT environments.

By effectively measuring availability, organizations can identify potential issues, prioritize remediation efforts, and improve overall system reliability. Combining multiple measurement methods and using advanced analytics can provide a comprehensive view of system availability and performance

While the C-I-A Triad provides a solid foundation for cybersecurity, it is essential to recognize its limitations. The model focuses primarily on technical aspects of security and may not adequately address organizational and human factors. Additionally, the increasing complexity of the threat landscape necessitates a more comprehensive approach to security.

To address these limitations, organizations can adopt a layered security approach, combining multiple security controls to protect information assets. This involves implementing a defense-in-depth strategy, which incorporates various security measures at different levels of the IT infrastructure. By employing a combination of technical, administrative, and physical controls, organizations can create a more robust security posture.

Furthermore, it is crucial to consider the human element in cybersecurity. Employees play a critical role in preventing security breaches. Security awareness training programs can help employees recognize and respond to potential threats. Social engineering attacks, which exploit human psychology, are a common method used by cybercriminals. Organizations must implement measures to protect employees from these attacks.

The C-I-A Triad is a fundamental concept in cybersecurity, providing a framework for understanding and protecting information. However, it is essential to recognize its limitations and adopt a more comprehensive approach to security.

The Parkerian HEXAD

The Parkerian Hexad, a framework proposed by Donn B. Parker (1998), offers a more comprehensive approach to information security than the traditional C-I-A Triad (Confidentiality, Integrity, Availability). By introducing three additional elements, Parker expanded the scope of security considerations.

The hexad comprises:

- **Confidentiality:** Protecting information from unauthorized disclosure (Parker, 1998).
- **Integrity:** Ensuring information is accurate and complete (Parker, 1998).
- **Availability:** Ensuring information is accessible when needed (Parker, 1998).
- **Possession or Control:** Protecting information from unauthorized use or modification (Parker, 1998).
- **Authenticity:** Verifying the origin and integrity of information (Parker, 1998).
- **Utility:** Ensuring information has value and is usable (Parker, 1998).

While the C-I-A Triad provides a foundational understanding of information security, the Parkerian Hexad offers a more holistic perspective (Schultz, 2003). By incorporating elements such as possession, authenticity, and utility, it addresses a broader range of security challenges.

For example, possession or control emphasizes the importance of physical security and access controls. Authenticity focuses on verifying the origin and integrity of information, which is crucial in preventing fraud and deception. Utility highlights the need for information to be usable and accessible in a meaningful way.

The Parkerian Hexad has become a valuable framework for organizations seeking to develop a comprehensive information security program (Whitman & Mattord, 2017). It provides

a structured approach to identifying and addressing security risks, enabling organizations to protect their assets effectively.

By expanding upon the C-I-A Triad, the Parkerian Hexad offers a more robust and nuanced view of information security. It encourages organizations to consider a wider range of factors when developing and implementing security measures.

The Ware Report and the Trusted Computer System Evaluation Criteria

The Ware Report and the Trusted Computer System Evaluation Criteria (TCSEC), often referred to as the "Orange Book," are seminal works in the history of cybersecurity (Parker, 2002). Both documents emerged in response to the growing reliance on computer systems and the associated security challenges (Pfleeger, Pfleeger, & Margulies, 2018).

Published in 1970, the Ware Report was a pioneering effort to address the security concerns of multi-user computer systems (Ware, 1970). It was commissioned by the U.S. Defense Science Board to provide guidance on the development and operation of these systems, particularly those handling classified information.

Key findings and recommendations of the Ware Report include (Ware, 1970):

- **Need for Comprehensive Security:** The report emphasized the importance of considering both technical and organizational factors in securing computer systems.
- **Flexible Security:** Security measures should adapt to changing conditions and threats.
- **Need-to-Know Principle:** Access to information should be restricted to those with a legitimate need.
- **Layered Security:** Multiple security controls should be implemented to provide defense in depth.

- **Human Factor:** Users play a critical role in security and should receive appropriate training.
- **Security Policy:** Clear and enforceable security policies are essential for effective protection.

The Ware Report laid the groundwork for subsequent cybersecurity efforts by recognizing the complexities of securing computer systems and emphasizing the need for a holistic approach (Anderson, 2008).

The TCSEC, published in the 1980s, built upon the foundations laid by the Ware Report (DoD, 1985). It provided a hierarchical classification system for evaluating the security capabilities of computer systems, ranging from Division D (minimal protection) to Division A (maximum protection).

Key elements of the TCSEC include (DoD, 1985):

- **Discretionary Access Control (DAC):** Users can control access to data they own.
- **Mandatory Access Control (MAC):** System-enforced access controls based on security labels.
- **Identification and Authentication:** Verifying user identities and granting appropriate access.
- **Accountability:** Tracking user actions for auditing and forensic purposes.
- **Documentation:** Providing clear and comprehensive system documentation.
- **Trusted Computing Base (TCB):** Identifying the core security components of the system.

The TCSEC provided a structured approach to assessing system security and promoted the development of security standards. However, it also faced criticism for its rigidity and limited focus on certain security aspects (Pfleeger et al., 2018).

Both the Ware Report and the TCSEC have had a lasting impact on the field of cybersecurity. They established foundational principles that continue to be relevant today (Whitman & Mattord, 2017). While the technological landscape has evolved significantly, the core concepts of these documents remain essential for building secure systems.

While the Ware Report provided a broader perspective on security, the TCSEC offered a more detailed classification framework. Together, they contributed to the development of modern cybersecurity practices, including risk management, access control, and incident response.

It is important to note that the cybersecurity landscape has evolved significantly since the publication of these documents. New threats and technologies have emerged, requiring continuous adaptation and innovation in security practices (ENISA, 2023). However, the fundamental principles outlined in the Ware Report and the TCSEC remain valuable for understanding the core challenges of cybersecurity.

Cyber Security Measurement

The efficacy of a cybersecurity system hinges on two critical aspects first appearing in the TSEC report: adherence to design specifications and actual operational performance. While these concepts might seem interconnected, there's a distinct difference between a system being 'built-to-specification' and 'operating-as-intended'.

A system is considered 'built-to-specification' when it is constructed in alignment with its predefined design parameters. This encompasses a thorough requirements analysis, where

security objectives and functional needs are clearly defined. It further includes a well-structured design and architecture that outlines the system's components and their interactions. Additionally, the implementation phase requires the accurate translation of the design into code and configuration. Finally, rigorous testing and verification procedures are essential to ensure the system's adherence to the specified requirements. Adhering to specifications is crucial for establishing a secure foundation as it helps prevent deviations from the intended design, which could inadvertently introduce vulnerabilities. However, it is important to acknowledge that meeting specifications alone does not guarantee optimal performance or security in the dynamic and unpredictable conditions of the real world.

For a system to truly 'operate-as-intended,' it must consistently fulfill its designed functions even under real-world conditions. This involves correct deployment and configuration, followed by continuous monitoring and management to identify and address any arising issues. Furthermore, effective incident response mechanisms are necessary to handle security breaches and system failures promptly and efficiently. Lastly, the system's success also hinges on user behavior, requiring users to adhere to established security policies and procedures. In essence, building a system to specification lays the groundwork for security, while operating as intended ensures that the system functions effectively and securely in the face of real-world challenges.

While a system might be meticulously built to specification, its performance can be impacted by various factors such as user error, environmental conditions, and evolving threats. For instance, a firewall configured correctly (built-to-specification) might not effectively block sophisticated attacks (operate-as-intended) due to emerging threat vectors.

Scientific Approaches to Cybersecurity

The early days of computer security research were heavily influenced by foundational papers like Saltzer-Schroeder (1975) and the U.S. government's Orange Book (Department of Defense, 1985). The Multics operating system, dating back to the 1960s, also played a significant role in shaping security research (Organick, 1972).

The desire for a more scientific approach to security research has been a constant theme throughout its history. McLean's 1987 critique highlighted the limitations of implicit assumptions in security research (McLean, 1987). This led to a community-wide debate about the Bell-LaPadula model (Bell & LaPadula, 1973) and the very definition of security itself. There have been three major approaches to establishing cybersecurity as a science:

1. **Formal Approaches:** A thread of scientific research focuses on formal approaches in cybersecurity. Schneider (2000) argues that the current state of security systems poses significant risks and advocates for building systems based on first principles. He suggests that cryptography exemplifies the kind of science-based approach that is needed. Krawczyk (2005) further reinforces the importance of formal mathematical models in assessing the security of cryptographic schemes. He notes that empirical evidence cannot definitively prove the security of a design, and that formal reasoning is the only reliable method.
2. **Empirical Work and Data Collection:** Another approach to the science of cybersecurity emphasizes data collection and empirical work. The JASON report (JASON, 2008) highlighted the need for more experimental research in this area. The development of the usable privacy and security community (SOUPS) has contributed to progress in this area. Studies by Whitten and Tygar (1999), and Schechter et al. (2008) have shown that users

often behave differently than expected, highlighting the importance of empirical research. Shostack and Maxion (2006) emphasize the value of data gathering and good experimental methods. Peisert and Bishop (2007) stress the importance of clearly stated hypotheses and experimental design and are more optimistic about the applicability of the scientific method to security.

3. **Quantitative Metrics:** While efforts have been made to develop quantitative metrics, progress has been slow. Pfleeger (2001) suggests that we are learning from our mistakes and that both quantitative and qualitative measurements are valuable. Sanders (2005) advocates for the use of relative metrics. Stolfo et al. (2005) highlights the challenges of developing security metrics and offers suggestions for advancing the field. A survey by Verendel (2009) found that there is limited evidence to support the hypothesis that security can be accurately represented with quantitative information. Many assumptions in formal treatments are not well-supported by empirical data.

The concept of a "Science of Security" has gained increasing attention in recent years, with calls for a more scientific approach to cybersecurity research. However, despite numerous discussions and initiatives, there remains no clear consensus on what this entails. The definition of Science of Security varies widely among experts. Some prioritize purpose, while others emphasize methodology, reproducibility, and clear communication. Some even seek a unifying theory, similar to those found in the physical sciences.

Despite the desire for a more scientific approach, many experts believe that cybersecurity is still a long way from achieving the rigor and standards of traditional physical sciences. Research often fails to adhere to the scientific method, and the reporting of experimental results is inconsistent.

While there is a growing recognition of the need for a more scientific approach to cybersecurity, the exact nature and scope of such an approach remain elusive.

Related Concepts

Confidentiality is a broad concept that encompasses the protection of sensitive information from unauthorized disclosure. It involves ensuring that information is only accessible to those who are authorized to have it.

Confidentiality differs from other similar concepts such as privacy, secrecy, and deception:

Privacy

While privacy also involves the protection of information, it primarily focuses on individuals' control over their personal data (Solove, 2004). Confidentiality, on the other hand, can apply to any type of sensitive information, whether it's personal or not.

Warren and Brandeis (1890) introduced the concept of privacy as a "right to be let alone" in their seminal 1890 Harvard Law Review article. They argued for protection from government and third-party intrusion into personal affairs. Subsequent scholars, such as Stacy Edgar (2002) and Mary B. Williams (2010), expanded on these ideas, highlighting the psychological and emotional impacts of privacy violations.

The term "privacy" lacks a precise definition, leading to diverse interpretations across legal, ethical, and public discourse. While some focus on individual control over personal information (Solove, 2004), others emphasize the protection from surveillance and government intrusion (Lyon, 2001).

Prosser (1960) identified four primary types of privacy violations: intrusion upon seclusion, public disclosure of private facts, false light publicity, and appropriation of name or likeness. However, these categories may not fully capture the complexities of contemporary privacy challenges (Bennett, 1992).

Secrecy

Secrecy is about keeping information unknown (Diffie & Hellman, 1976). It's a narrower concept than confidentiality, as it only focuses on the concealment of the information itself. Confidentiality, in contrast, involves both keeping the information secret and ensuring that only authorized individuals can access it (Pfleeger, Pfleeger, & Margulies, 2018).

Claude Shannon categorized secrecy into three types (Shannon, 1949):

- **Concealment Systems:** These systems hide the very existence of a message, often through steganography.
- **Privacy Systems:** These systems require specialized equipment to decrypt the message, such as voice inversion technology.
- **True Secrecy Systems:** These systems rely on encryption to protect the content of a message, assuming the adversary has access to the communication channel.

The Importance of Secrecy

Secrecy plays a critical role in various aspects of life and society. It can be used to protect individuals, businesses, and governments (Bok, 1982). For instance, secrets about vulnerabilities can be used to prevent harm, while secrets about business strategies can provide a competitive

advantage. However, the misuse of secrecy can also have negative consequences, such as the protection of criminal activities (Simmel, 1906).

Secrecy is a complex concept with implications for individuals, organizations, and society as a whole (Sissela Bok, 1982). Understanding the different types of secrecy and their applications is essential for developing effective security measures.

Deception

Deception involves intentionally misleading others (Bok, 1982). While it can be used to protect information, it's not the primary means of ensuring confidentiality (Pfleeger et al., 2018). Confidentiality is about protecting information from unauthorized access, not about actively misleading others (Whitman & Mattord, 2014).

Deception involves intentionally misrepresenting information (Bok, 1982). It's distinct from secrecy, which is simply the state of information being unknown. Kerckhoff's Principle emphasizes that the security of a system should not rely on the secrecy of its algorithms or implementation (Kerckhoffs, 1883). Instead, security should be derived from the strength of the cryptographic keys. This principle underscores the importance of open design and analysis in security systems. While deception can contribute to confidentiality by misleading adversaries, it should not be the sole reliance for security. A robust security system requires a combination of deception, obscurity, and strong cryptographic techniques (Schneier,1996).

Confidentiality is a broader concept that encompasses both secrecy and the protection of information from unauthorized access (Smith, 2018). While privacy and deception can play a role in ensuring confidentiality, they are not synonymous with it.

Legal Frameworks

Data Protection Laws

The Internet has ushered in an unprecedented era of data collection and utilization, necessitating robust legal frameworks to safeguard individual privacy. The United States and the European Union have adopted divergent approaches to data protection, reflecting distinct cultural, economic, and political contexts (Bennett, 1992).

The European Union has established a comprehensive data protection regime with the General Data Protection Regulation (GDPR) (Voigt & Von dem Bussche, 2017), which came into effect in 2018. The GDPR grants individuals' extensive rights over their personal data, including the right to access, rectify, erase, and restrict processing (Regulation (EU) 2016/679, 2016). It imposes stringent obligations on organizations that collect and process personal data, such as data minimization, accountability, and data breach notification (Regulation (EU) 2016/679, 2016). The GDPR's territorial scope is broad, applying to any organization processing the personal data of EU residents, regardless of the organization's location (Regulation (EU) 2016/679, 2016).

In contrast, the United States has a patchwork of data protection laws, with no overarching federal privacy law comparable to the GDPR (Bennett, 1992). Instead, data protection is primarily regulated through sector-specific laws, such as the Health Insurance Portability and Accountability Act (HIPAA) for healthcare (Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191), the Gramm-Leach-Bliley Act (GLBA) for financial institutions (Gramm-Leach-Bliley Act, Pub. L. 106-102, 1999), and the Children's Online Privacy Protection Act (COPPA) for children's online privacy (Children's Online Privacy

Protection Act, 15 U.S.C. §§ 6501-6506). While these laws provide some level of protection for specific types of data, they leave significant gaps in overall data privacy.

The differing approaches of the United States and Europe reflect fundamental differences in their legal and cultural traditions (Bennett, 1992). The EU's emphasis on individual rights and privacy has led to a more proactive and comprehensive regulatory framework. The United States, with its focus on free markets and limited government intervention, has adopted a more industry-specific and less prescriptive approach.

The divergence between these two major jurisdictions has significant implications for businesses operating on a global scale. Companies must navigate complex and often conflicting legal requirements to protect personal data. The absence of a comprehensive federal data protection law in the United States has created uncertainty and increased compliance costs for businesses.

In recent years, there has been growing momentum for comprehensive data privacy legislation in the United States. California's Consumer Privacy Act (CCPA) (California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100-1798.199) and similar laws in other states represent a step in this direction. However, the patchwork nature of state privacy laws creates challenges for businesses operating nationwide.

The evolving landscape of data protection raises important questions about the future of privacy regulation. As technology continues to advance, new challenges and opportunities will emerge. Finding a balance between protecting individual privacy and facilitating innovation will be a critical task for policymakers and industry stakeholders.

In conclusion, the United States and the European Union have adopted fundamentally different approaches to data protection. While the EU has established a comprehensive

framework, the United States has relied on a patchwork of sector-specific laws. The growing importance of data in the digital economy highlights the need for a more harmonized global approach to data protection.

The GDPR: A Use-Based Approach

The European Union's General Data Protection Regulation (GDPR) represents a landmark in data protection legislation, establishing a comprehensive framework for the processing of personal data (Voigt & Von dem Bussche, 2017). Unlike many other jurisdictions that focus primarily on data security or breach notification, the GDPR adopts a fundamentally different approach: it centers on the use of personal data.

This use-based perspective is evident in several key provisions of the GDPR (Regulation (EU) 2016/679, 2016):

- **Purpose Limitation and Data Minimization**
- **Lawful Basis for Processing**
- **Individual Rights**
- **Accountability**

By focusing on the use of personal data and granting individuals robust rights, the GDPR has set a new global standard for data protection. It represents a significant shift away from a purely security-focused approach towards a more human-centric framework. While challenges remain in implementing and enforcing the GDPR, its impact on data protection practices is undeniable.

The U.S. Data-type Approach

Unlike the European Union's comprehensive General Data Protection Regulation (GDPR), which adopts a principle-based approach centered on individual rights, the United States has historically taken a more sector-specific approach to data protection, often referred to as a "data-type" approach (Bennett, 1992). This means that privacy laws in the U.S. are primarily focused on protecting specific types of data rather than providing a broad framework for data protection.

In the United States, data privacy is governed by a patchwork of sector-specific laws that offer varying levels of protection. Health data falls under the purview of HIPAA, which sets national standards for electronic health records and safeguards patient information, particularly Individually Identifiable Health Information (IIHI) (Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191). Financial data is regulated by the Gramm-Leach-Bliley Act (GLBA), which mandates financial institutions to maintain the confidentiality, integrity, and security of customer information (Gramm-Leach-Bliley Act, Pub. L. 106-102, 1999). The privacy of children under 13 is protected by COPPA, which restricts the collection and use of their personal information online (Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501-6506). In the educational context, FERPA safeguards the privacy of student education records, granting parents and eligible students control over access (Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g).

However, this data-type approach has limitations. The patchwork of laws leads to inconsistent protections and potential loopholes across different sectors. Additionally, many data types remain uncovered, leaving individuals with inadequate safeguards. Enforcement of multiple laws is complex, and the absence of a unified framework hinders the development of a comprehensive privacy culture.

In recent years, there has been growing momentum for more comprehensive data protection in the United States. California's Consumer Privacy Act (CCPA) (California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100-1798.199) represents a significant step towards a broader privacy framework. However, the patchwork of state laws remains a challenge.

Furthermore, technological advancements and the increasing volume of data have raised concerns about data privacy and security. Issues such as data breaches, surveillance, and the use of artificial intelligence have prompted calls for stronger data protection measures.

In conclusion, the United States' data-type approach to data protection offers targeted protection for specific types of data but lacks a comprehensive framework for safeguarding personal information across all sectors. As the digital landscape continues to evolve, the need for a more unified and robust data protection regime becomes increasingly apparent.

Data Breach Disclosure Laws

The increasing prevalence of cyberattacks and data breaches has necessitated robust legal frameworks governing the disclosure of such incidents (Romanosky, 2010). The United States and the European Union have adopted distinct approaches to data breach notification, reflecting their broader philosophies on data protection and privacy (Bennett, 1992).

The European Union's Approach

The European Union's General Data Protection Regulation (GDPR) has established a comprehensive regime for data breach notification (Voigt & Von dem Bussche, 2017). Article 33 of the GDPR mandates that organizations notify competent supervisory authorities without undue delay, and where feasible, within 72 hours of becoming aware of a personal data breach

that is likely to result in a high risk to the rights and freedoms of natural persons (Regulation (EU) 2016/679, 2016).

The GDPR's focus on timely notification is designed to enable individuals to take protective measures and mitigate potential harm (Regulation (EU) 2016/679, 2016). It also empowers supervisory authorities to investigate breaches and hold organizations accountable. The regulation further emphasizes the importance of data protection impact assessments (DPIAs) to identify and mitigate risks to the rights and freedoms of data subjects (Regulation (EU) 2016/679, 2016).

The United States' Approach

In contrast to the EU's unified approach, the United States has a patchwork of data breach notification laws at the state level (Solove, 2004). While there is no federal mandatory data breach notification law, a growing number of states have enacted their own legislation. These laws vary significantly in terms of the types of data covered, the entities required to report breaches, and the notification timelines.

California's Consumer Privacy Act (CCPA) (California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100-1798.199) and the New York Cybersecurity Act (New York State Technology Law § 208) are examples of state laws that include data breach notification requirements. These laws often mandate notification to affected individuals, as well as to government agencies in some cases. However, the lack of a federal standard creates challenges for businesses operating nationwide, as they must comply with multiple state laws.

In the United States, the protection of different data types is governed by various sector-specific laws. Health data finds its safeguard in the Health Insurance Portability and Accountability Act (HIPAA), which establishes national standards for electronic health records

and ensures the protection of individually identifiable health information (IIHI) (Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191). Financial data is regulated by the Gramm-Leach-Bliley Act (GLBA), mandating financial institutions to uphold the confidentiality, integrity, and security of customer information (Gramm-Leach-Bliley Act, Pub. L. 106-102, 1999). The privacy of children under 13 is specifically addressed by the Children's Online Privacy Protection Act (COPPA), which places restrictions on the collection and use of their personal information online (Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501-6506). In the educational sector, the Family Educational Rights and Privacy Act (FERPA) protects the privacy of student education records, granting parents and eligible students the right to review and control access to their records (Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g).

However, this data-type approach to privacy protection in the U.S. has its shortcomings. The existence of multiple sector-specific laws leads to inconsistencies in data protection standards across different industries, creating potential loopholes and opportunities for regulatory arbitrage. Furthermore, the limited scope of these laws leaves many types of data without explicit protection. Enforcing compliance with multiple laws can be a complex and resource-intensive task for both regulators and businesses. The lack of a comprehensive, unified data protection framework in the U.S. also hinders the development of a holistic privacy culture.

Both the EU and the US face challenges in effectively addressing data breaches. The rapid evolution of technology and the increasing sophistication of cyberattacks make it difficult to stay ahead of threats (ENISA, 2023). Additionally, the global nature of data flows creates complexities in determining which jurisdiction's laws apply.

There is a growing consensus that a more unified approach to data breach notification is needed. While the US has been slower to adopt a federal law, recent developments, such as the proposed American Data Privacy and Protection Act (ADPPA) (American Data Privacy and Protection Act, H.R. 8152, 117th Cong.), indicate a potential shift towards a more comprehensive framework.

The EU and the US have adopted different approaches to data breach disclosure. The EU's GDPR provides a more comprehensive and proactive framework, while the US's patchwork of state laws offers varying levels of protection. As the threat landscape continues to evolve, it is likely that both jurisdictions will need to adapt their regulations to ensure effective protection of individuals' rights.

Lawful Data Access Laws

The balance between protecting individual privacy and enabling legitimate access to data for law enforcement, national security, and intelligence purposes has been a complex and contentious issue. The United States and the European Union have adopted distinct approaches to this challenge, reflecting their respective legal and political traditions (Bennett, 1992).

Lawful Access in the United States

In the United States, lawful access to data is governed by a patchwork of laws, primarily prioritizing law enforcement and national security interests (Kerr, 2009). Key legislation such as the Electronic Communications Privacy Act (ECPA) (Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508), the Foreign Intelligence Surveillance Act (FISA) (Foreign Intelligence Surveillance Act of 1978, 50 U.S.C. §§ 1801-1885e), and the USA PATRIOT Act (Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and

Obstruct Terrorism (USA PATRIOT ACT) Act of 2001, Pub. L. No. 107-56) establish standards for government access to communications and data, often with a broader scope of surveillance authorities and lower thresholds for obtaining warrants compared to the European Union (Kerr, 2009).

Lawful Access in the European Union

In contrast, the European Union adopts a more restrictive approach to lawful access, emphasizing individual privacy rights (Voigt & Von dem Bussche, 2017). The General Data Protection Regulation (GDPR) strengthens these rights, potentially limiting government access to data (Regulation (EU) 2016/679, 2016). While the EU recognizes the necessity of law enforcement and national security, it places stricter safeguards on data access. Although currently under review, the e-Privacy Directive provides a framework for protecting electronic communications (Directive 2002/58/EC, 2002). The EU has also grappled with the challenge of balancing privacy rights with counterterrorism efforts, evident in debates over data retention and access to encrypted communications (Case C-293/12 and C-594/12, 2014).

Key differences between the two regions include the broader scope of surveillance powers in the US versus the more restrictive approach in the EU, the EU's stronger emphasis on individual privacy rights, and differing stances on data retention and access to encrypted communications. These contrasting approaches highlight the ongoing global debate on how to balance security needs with the protection of individual privacy in an increasingly digital world.

The balance between lawful access and privacy is a complex and evolving issue. Technological advancements, such as encryption and cloud computing, have made it more difficult for governments to access data (Abelson, Ledeen, & Lewis, 2015). The rise of

cybercrime and terrorism has also increased pressure on law enforcement to obtain access to relevant information (Brenner, 2007).

Both the US and the EU are grappling with these challenges. Finding the right balance between protecting public safety and preserving individual privacy will require ongoing legislative and technological innovation. International cooperation will also be essential to address cross-border data access issues (Bennett, 1992).

The US and the EU have adopted distinct approaches to lawful data access, reflecting different priorities and values. The ongoing evolution of technology and the changing threat landscape will continue to shape the development of data access laws in both jurisdictions.

Desired Security Properties

This section distills desirable security properties from Western legal frameworks, focusing on data protection and access laws. It translates individual, organizational, and societal priorities enshrined in law into high-level technical features, setting aside the ongoing debate on balancing competing priorities like national security and individual privacy.

Instead of framing data protection solely around preventing unauthorized access, use, disclosure, modification, and destruction, this analysis extracts more nuanced properties. **Confidentiality** remains paramount, shielding data from unauthorized access. **Integrity** ensures data accuracy and prevents tampering. **Availability** guarantees accessibility for legitimate purposes. Beyond these core tenets, legal frameworks emphasize **accountability**, ensuring compliance with data protection rules. **Fairness** and **lawfulness** guide ethical and transparent data handling. **Purpose limitation** restricts data use to specified, legitimate ends, while **data minimization** and **storage limitation** curb excessive data collection and retention. Crucially, **data subject rights** empower individuals with control over their personal data.

To bolster these protections, **incident detection and response** mechanisms are essential for swiftly addressing data breaches. **Notification protocols** ensure timely alerts to affected parties and authorities. Proactive **risk assessment identifies** vulnerabilities and guides mitigation. **Privacy by design and default** become guiding principles in system development. **Data protection impact assessments (DPIAs)** scrutinize high-risk activities, and adherence to **international data transfer regulations** safeguards cross-border information flows.

This approach reframes data protection by moving beyond a simple "prevention of unauthorized actions" model. It incorporates legally mandated principles that promote responsible data stewardship and empower individuals, reflecting a more holistic and nuanced understanding of data protection in the digital age.

Definition of Data Protection

Data Protection can be defined as protecting data from unauthorized access, use, disclosure, modification, and destruction. Only authorized individuals (I) should be able to access, use, disclose, modify, or destroy the data. The effectiveness of security measures (E) in upholding this policy (P) determines the overall data protection level.

$$D = (P, I, E) \tag{1}$$

The traditional definition of confidentiality focuses on protecting sensitive information from unauthorized disclosure. Data protection is a more comprehensive definition that emphasizes protecting data from unauthorized access and its dependencies such as unauthorized use, disclosure, modification, and destruction. The traditional notion of the C-I-A triad cannot be measured easily, because the components of C-I-A are not independent. The data protection

definition overcomes this limitation helping connect the dependencies of confidentiality with aspects of integrity and availability. Data protection aims to ensure that data is accurate, reliable, and accessible to authorized individuals while preventing unauthorized access, use, disclosure, modification, or destruction.

Data protection, in the context of evaluating security mechanisms, can be defined as protecting data from unauthorized access, use, disclosure, modification, and destruction. This definition focuses specifically on the role of security measures in preventing these unauthorized actions. It emphasizes that only authorized individuals should be able to interact with data in these ways, and the effectiveness of the implemented security measures determines the overall level of data protection achieved. This perspective acknowledges that data protection involves a broader set of procedures and policies, but for the purpose of measuring the efficacy of security mechanisms, the definition is limited to these key actions.

Systemization of Knowledge

The literature review systematizes researchers' knowledge on measuring aspects of data protection, adhering to the definition established in section 2.4. Building upon this foundation, the review employed a combination of forward and backward research techniques to uncover data protection strategies that directly counter malicious actors. By analyzing top cybersecurity conferences and the oldest computer security journal, the most relevant and highly cited works were selected to pinpoint the most impactful contributions in the field. The literature review highlights the significance of conferences like IEEE Symposium on Security and Privacy (S&P), ACM CCS, and USENIX Security Symposium as leading platforms for cybersecurity research. These conferences, along with the Journal of Computers & Security, collectively cover a broad spectrum of topics within the field, providing valuable insights into the latest advancements and

emerging threats in cybersecurity. Papers that are foundational or provide formal proofs without sufficient modeling in real-world threat or vulnerability scenarios were out-of-scope.

To comprehensively understand the landscape of security protections, a taxonomy was created to organize the literature based on the definition of data protection with the following categories: access, use, disclosure, modification, and destruction. The categorizations then were contextualized with measurements from the literature. This systemization of knowledge provided a structured approach to analyzing and comparing different security measures, identifying strengths and weaknesses, and understanding how each interacts with various threats and vulnerabilities. This taxonomy is the first step in helping build a causal model for data protection later discussed in Chapter 4. This section focuses on understanding the literature, including research approaches, measurements, and empirical results demonstrating security effectiveness. The taxonomy focused on security measurements and omitted performance, reliability, usability, or efficiency measurements unless those directly were needed for security.

Measuring Authorized Access

When measuring authorized access, these metrics offer valuable insights into a system's capacity to maintain data protection. By evaluating factors such as the strength of passwords and encryption keys, the accuracy of access control mechanisms, and the effectiveness of vulnerability detection and management, we can gauge the system's resilience against unauthorized access attempts. These metrics highlight potential weaknesses and areas for improvement, enabling us to better understand and enhance a system's ability to protect sensitive information.

Key Metrics Summary

Security Strength

- Password strength (guessable to unguessable), 2FA/PGP adoption (yes/no), PGP key strength (≤2048 or >2048 bits), key reuse & traceability (yes/no) [van de Laarschot & van Wegberg, 2021].

Security Accuracy

- Relative True Positive Rate (TPR) and Relative Risk Score Relation (RSR), both relative to a baseline (0 to 1) [Wiefling et al., 2021].

Vulnerability Management

- Number of vulnerabilities detected (0-8), false positives (0-2), coverage (80-100%),analysis time (7 seconds - 2 minutes), detection rate (86%-100%), precision (average 89%), forensic success rate (85%-98%), training time (up to 12 minutes), validation/forensic time (microseconds to milliseconds per access),resolved/unresolved offsets (0% to 100%), time consumption (seconds per contract), and memory consumption (megabytes or gigabytes) [Sun et al., 2011; Xiang et al., 2019; Tsankov et al., 2018].

Vulnerability Detection

- Detection time, OSS detection accuracy (precision 82-92%, recall 82-89%), number of vulnerable OSS instances (>100,000 apps), true/false warnings, success rate of key recovery attacks, fraction of violations/warnings/compliances (0-100% each) [Deng et al., 2023; Kim & Lee, 2017; Tsankov et al., 2018].

Table 1

Authorized Access Metrics Taxonomy

Metric	Measurement	Range	Scale	Type	Citation
Security Strength	Password strength	guessable to unguessable	Ordinal	Intrinsic	van de Laarschot & van Wegberg, 2021
	2FA/PGP adoption	yes/no	Nominal	Intrinsic	
	PGP key strength	≤2048 or >2048 bits	Ordinal	Intrinsic	
	Key reuse & traceability	yes/no	Nominal	Intrinsic	
Security Accuracy	Relative True Positive Rate (TPR)	0 to 1	Ratio	Relative	Wiefling et al., 2021
	Relative Risk Score Relation (RSR)	0 to 1	Ratio	Relative	Wiefling et al., 2021
Vulnerability Management	Number of vulnerabilities detected	0-8	Ratio	Intrinsic	Sun et al., 2011; Xiang et al., 2019; Tsankov et al., 2018
	False positives	0-2	Ratio	Intrinsic	
	Coverage	80-100%	Ratio	Intrinsic	
	Analysis time	7 seconds - 2 minutes	Ratio	Intrinsic	
	Detection rate	86%-100%	Ratio	Intrinsic	
	Precision	average 89%	Ratio	Intrinsic	
	Forensic success rate	85%-98%	Ratio	Intrinsic	
	Training time	up to 12 minutes	Ratio	Intrinsic	
	Validation/forensic time	microseconds to milliseconds per access	Ratio	Intrinsic	

Table 1 continued

	Resolved/unresolved offsets	0% to 100%	Ratio	Intrinsic	
	Time consumption	seconds per contract	Ratio	Intrinsic	
	Memory consumption	megabytes or gigabytes	Ratio	Intrinsic	
Vulnerability Detection	Detection time	-	Ratio	Intrinsic	Deng et al., 2023; Kim & Lee, 2017; Tsankov et al., 2018
	OSS detection accuracy	precision 82-92%, recall 82-89%	Ratio	Intrinsic	
	Number of vulnerable OSS instances	>100,000 apps	Ratio	Intrinsic	
	True/false warnings	-	Nominal	Intrinsic	
	Success rate of key recovery attacks	-	Ratio	Intrinsic	
	Fraction of violations/warnings/ compliances	0-100% each	Ratio	Intrinsic	

These measurements, spanning various ranges and categories, provide a comprehensive view of the diverse approaches and metrics employed in evaluating measuring systems protecting from unauthorized access. The research papers employed a range of measurements to assess security aspects.

For security strength, the studies measured password strength (ranging from very guessable to very unguessable), 2FA/PGP adoption (binary), PGP key strength (≤2048 or >2048 bits), and key reuse & traceability (binary) (van de Laarschot & van Wegberg, 2021).

In the context of security accuracy, relative True Positive Rate (TPR) and Relative Risk Score Relation (RSR), both relative to a baseline and ranging from 0 to 1, were used to evaluate the balance between privacy and security in risk-based authentication systems (Wiefling et al., 2021).

Lastly, for vulnerability management, measurements included the number of vulnerabilities detected (0-8), false positives (0-2),coverage (80-100%), analysis time (7 seconds - 2 minutes), detection rate (86%-100%), precision (average 89%), forensic success rate (85%-98%), training time (up to 12 minutes), validation/forensic time (microseconds to milliseconds per access), resolved/unresolved offsets (0% to 100%), time consumption (seconds per contract), and memory consumption (megabytes or gigabytes) (Sun et al., 2011; Xiang et al., 2019; Tsankov et al., 2018). In the realm of vulnerability detection, metrics such as detection time, OSS detection accuracy (precision ranging from 82-92% and recall from 82-89%), the number of vulnerable OSS instances (over 100,000 apps), true/false warnings, success rate of key recovery attacks, and the fraction of violations/warnings/compliances (0-100% for each) were utilized (Deng et al., 2023; Kim & Lee, 2017; Tsankov et al., 2018).

Access Control

The research papers investigated various aspects of access control, code security, and privacy. In terms of access control, studies focused on static detection of vulnerabilities (Sun et al., 2011), continuous validation and forensics (Xiang et al.,2019), automated policy generation for microservices (Li et al., 2021), and context sensing for IoT (He et al., 2021). For code security, the impact of information sources on code security was examined (Acar et al., 2016). Additionally, the detection of license violations and vulnerable code in mobile apps was

addressed (Kim & Lee, 2017). Finally, the research papers explored secure and private data storage, focusing on privacy and access control (Maffei et al., 2015).

The researchers used a variety of metrics to measure these aspects, including the number of vulnerabilities detected, false positives, coverage, analysis time, detection rate, precision, forensic success rate, training time, validation/forensic time, functional correctness, security, resource usage, confidence, security thinking, OSS usage, license violations, vulnerable OSS versions, computation time, communication overhead, and scalability. The specific ranges for these measurements varied depending on the research paper and the focus of the study.

For example, in terms of static detection of vulnerabilities, Sun et al. (2011) found that their approach could detect vulnerabilities with a coverage of 80-100% and an analysis time of 7 seconds to 2 minutes. Xiang et al. (2019) reported a detection rate of 86%-100% and a precision of 89% for their continuous validation and forensics tool. Acar et al. (2016) found that participants in their study who used Stack Overflow for code security tasks were less likely to produce secure solutions than those who used official documentation or books. Kim and Lee (2017) analyzed 1.6 million mobile apps and found that approximately 40,000 cases of license violations and over 100,000 apps with vulnerable OSS were detected. Maffei et al. (2015) evaluated the performance of their group ORAM scheme and found that it could scale to large databases and many users with low communication overhead.

Overall, the research papers provide valuable insights into the challenges and best practices for ensuring the security and privacy of software systems. They highlight the importance of effective access control mechanisms, secure coding practices, and robust privacy protection measures.

Authentication

The research presented addresses various security and privacy challenges in different domains. Acar et al. (2016) explored the impact of information sources on code security, finding that Stack Overflow usage correlated with lower security scores (51.4%) compared to official documentation (85.7%). The authors measured functional correctness (40.4% - 67.3%), security, resource usage, confidence (40.7% - 55.4%), and security thinking (0 - 79.6%) to demonstrate the trade-off between usability and security, with developers often prioritizing the former.

Kim and Lee (2017) developed OSSPolice to detect license violations and vulnerable OSS in mobile apps, achieving high accuracy (precision: 82-92%, recall: 82-89%) and a 65% improvement in version matching over LibScout. Their analysis of 1.6 million apps revealed significant license violations and the widespread use of vulnerable OSS.

Wiefling, Tolsdorf, and Lo Iacono (2021) tackled privacy in risk-based authentication systems. They evaluated the impact of privacy enhancements on Relative True Positive Rate (relative to a baseline) and Relative Risk Score Relation (also relative to baseline). Their findings showed that truncation (removing 0-32 bits from IP addresses) and k-anonymity (grouping users, k=1 to 6) can improve privacy, with 3-bit truncation being optimal. However, increasing k in k-anonymity reduced security.

Gavazzi et al. (2023) studied multi-factor and risk-based authentication adoption, measuring MFA availability (0-100%,with an overall average of 42.31%) and RBA effectiveness (0-100%, with an overall average of 22.12%). They found that Single Sign-On (SSO) significantly increased both MFA and RBA availability but also raised privacy concerns due to third-party tracking.

Wu et al. (2023) proposed ChkUp to address firmware update vulnerabilities. They measured its effectiveness in terms of accuracy of update entry finding (50% to 100%),

correctness of execution path recovery (122/150 UFGs sound and complete), metrics of procedure recognition (overall 461 TPs, 45 FNs, 17 FPs), performance overhead (126s for path recovery, 216.1s for recognition per image), success rate of vulnerability validation (73/119 PoCs created), and scalability of validation (44.1% emulatable, 72.2% repacked, 82.7% re-emulated). Their results showed high accuracy in identifying and addressing firmware update vulnerabilities.

Authorization

The research papers presented shed light on the crucial role of addressing threats and vulnerabilities in various aspects of security and privacy. Acar et al. (2016) highlighted the potential for insecure coding practices due to developers' reliance on unreliable information sources, particularly Stack Overflow, emphasizing the need for secure coding practices and reliable documentation. This study measured functional correctness (40.4% - 67.3%) and security (51.4% - 85.7%),revealing that while Stack Overflow is convenient, it can lead to less secure code compared to official documentation.

Kim and Lee (2017) tackled license violations and security risks from vulnerable open-source software (OSS) in mobile apps using their OSSPolice tool. Their empirical analysis of 1.6 million apps showed significant license violations and the use of vulnerable OSS, demonstrating the need for improved security practices in the software supply chain. The tool's effectiveness was measured through OSS detection accuracy, with precision ranging from 82-92% and recall from 82-89%, and a 65% improvement in version matching compared to previous methods.

Wiefling, Tolsdorf, and Lo Iacono (2021) focused on mitigating user re-identification and tracking in risk-based authentication (RBA) systems. Their empirical evaluation of privacy enhancements on a real-world dataset measured the relative True Positive Rate (TPR) and

Relative Risk Score Relation (RSR) to assess the balance between privacy and security. They found that truncation (removing bits from IP addresses) and k-anonymity can improve privacy, but the effectiveness is limited to specific parameter choices.

Gavazzi et al. (2023) examined the adoption of multi-factor authentication (MFA) and RBA on popular websites, measuring MFA availability (42.31% overall) and RBA effectiveness (22.12% overall). Their findings revealed low adoption rates, highlighting the need for improved authentication practices to mitigate account hijacking risks.

Wu et al. (2023) addressed vulnerabilities in firmware update procedures by developing ChkUp, a tool that achieved high accuracy (50% to 100%) in identifying update entry points and recovering execution paths. The effectiveness of procedure recognition varied, but overall, it showed promise, especially for integrity checks. The study also demonstrated successful vulnerability validation with proof-of-concept exploits and highlighted the scalability limitations of their approach.

Close Access

The research presented explores diverse threats to security and privacy, highlighting the potential for information leakage through seemingly innocuous channels. Notably, Vuagnoux and Pasini (n.d.) demonstrated how electromagnetic emanations from wired and wireless keyboards can be exploited to recover keystrokes, potentially compromising sensitive information. Their empirical approach involved capturing and analyzing these emanations, revealing successful keystroke recovery rates of up to 95% from a distance of 20 meters. Similarly, Nassi et al. (2023) identified a new class of optical TEMPEST attacks, the "Glow Worm attack", where sound can be recovered from the subtle light fluctuations of a device's power indicator LED. Their experiments showed the feasibility of recovering speech with good

intelligibility at 15 meters and fair intelligibility even at 35 meters, emphasizing the need for hardware-level countermeasures.

Camurati et al. (2018) discovered "screaming channels," a side channel in mixed-signal chips that can inadvertently broadcast sensitive information via radio transmissions. They successfully demonstrated full key recovery attacks from up to 10 meters, underscoring the potential for long-range exploitation of this vulnerability. Chen et al. (n.d.) examined side-channel information leaks in web applications, demonstrating how sensitive user data can be inferred from encrypted web traffic due to inherent design features. They measured the reduction power and network overhead of various mitigation techniques, revealing the challenges in effectively addressing these leaks without significant performance impacts.

Finally, Backes et al. (2009) revisited the threat of compromising reflections, demonstrating the feasibility of reconstructing monitor images from reflections in the human eye and diffuse reflections. Their research utilized measurements like font size, distance, and telescope diameter, and showed that a 36pt font could be read from 10 meters away using a 235mm telescope. They also evaluated countermeasures, finding polarization filters ineffective and suggesting optical notch filters as a potential solution.

Lawful Access

The research presented addresses various threats to privacy and security, emphasizing the need for proactive and comprehensive mitigation strategies. In the context of Android apps, Nguyen et al. (2022) investigated the prevalence of non-compliance with GDPR consent requirements, revealing that a significant portion of apps either lacked consent notices entirely or exhibited violations such as sharing data without explicit consent or disregarding opt-out choices. Wiefling et al. (2021) tackled the challenge of user re-identification and tracking in risk-

based authentication (RBA) systems. Their empirical evaluation of privacy enhancements on a real-world dataset demonstrated that techniques like truncation (reducing the precision of IP addresses) and k-anonymity (grouping users) can effectively enhance privacy while maintaining security and usability.

On a broader scale, Ladisa et al. (2022) proposed a taxonomy of attacks on open-source software supply chains, aiming to mitigate the injection of malicious code into OSS artifacts. They identified 107 unique attack vectors and 33 safeguards, providing a comprehensive framework for understanding and addressing these threats. The study also assessed the utility and cost of these safeguards, offering insights for developers and organizations.

The research also delved into specific vulnerabilities and attacks. Boucher and Anderson (n.d.) introduced "Trojan Source" attacks, which exploit text-encoding subtleties to introduce hidden vulnerabilities in source code, bypassing traditional code review processes. Jia et al. (2021) uncovered "Codema" vulnerabilities in IoT devices due to disjointed device management channels, leading to unauthorized access. Their proposed CGuard framework, evaluated through functionality tests and user studies, proved effective in mitigating these risks.

Finally, Acar et al. (2016) explored the impact of information sources on code security, demonstrating that relying on Stack Overflow often leads to less secure code compared to official documentation. They measured functional correctness (40.4% - 67.3%) and security (51.4% - 85.7%) to underscore the importance of using reliable sources and adopting secure coding practices.

Supply-Chain

Several studies have focused on mitigating threats in the software supply chain and access control. Duan et al. (2017) developed OSSPolice to detect license violations and

vulnerable OSS components in Android apps, achieving high accuracy with precision and recall ranging from 82-92% and 82-89%, respectively. Newman, Meyers, and Torres-Arias (2022) proposed Sigstore to enable widespread software signing, thereby mitigating software supply chain attacks. The system's effectiveness was evaluated through usage frequency (millions of signatures), latency (sub-second for core components), and scalability (handling 347K+ entries/day). Ladisa et al. (2022) presented a comprehensive taxonomy of attacks on open-source software supply chains, along with 33 potential safeguards. Their mixed-method approach involved expert and developer surveys to assess the taxonomy's correctness (75% agreement on structure) and the utility and cost of safeguards (most rated medium to high utility, cost varied).

In addressing access control vulnerabilities, Sun et al. (2011) proposed a static analysis method for web applications, demonstrating its effectiveness with high coverage (80-100%) and low false positives (0-2). Maffei et al. (2015) focused on secure cloud storage using Group ORAM, showcasing its scalability and efficiency in handling large datasets (1GB-1TB) and multiple clients (1-10000). Xiang et al. (2019) introduced a method for continuous access control validation and forensics, achieving high detection rates (86%-100%) and precision (89%) in identifying access control misconfigurations. Li et al. (2021) presented an automated policy generation approach for microservices, demonstrating 100% request extraction effectiveness and significant improvements in policy management performance and scalability.

Lastly, He et al. (2021) evaluated sensors for contextual access control in smart homes, revealing vulnerabilities to physical attacks and privacy concerns. They advocated for sensor redundancy and careful policy design to mitigate these risks. Jia et al. (2021) identified "Codema" vulnerabilities in IoT devices and proposed the CGuard framework to address

unauthorized access. Their evaluation showed CGuard's effectiveness in mitigating these risks with negligible overhead.

These studies collectively highlight the diverse challenges in ensuring security and privacy and demonstrate the potential of various approaches, ranging from technical solutions like static analysis and cryptographic protocols to empirical evaluations and user studies, to effectively mitigate threats and vulnerabilities.

Measuring System Use

Understanding the data protection of a system hinges on its ability to keep data secret and prevent unauthorized access. Security metrics like those related to cost and attacker success rate, security strength, and detection rate offer insights into potential vulnerabilities and the effectiveness of safeguards. By examining the impact of attacks on network performance, the strength of passwords and random number generation, and the accuracy of detection mechanisms like spam filters, we gain a clearer picture of a system's resilience against breaches and its overall ability to maintain data protection.

Key Metrics Summary

Cost and Attacker Success Rate

- Network impact: Measured by network degradation (increased latency), network round trip time, and request latency. Degradation can range from 50% (slowdown) to 300% (significant slowdown) on VMs sharing the same physical server. (Varadarajan et al., 2015)

Security Strength

- Password strength: Evaluated using guesswork, ranging from 0 (easily guessed) to 2^40 (very strong). Probability thresholds were used to assess the likelihood of guessing passwords. (Ma et al., 2014)
- PRNG security: Based on entropy (randomness), the actual entropy collected was often lower than expected, particularly in environments with limited randomness sources. One example showed hard disk events only providing 1.03 bits of entropy per event, much lower than the assumed 32 bits.

Detection Rate

- Accuracy: Used for spam filters, with a high accuracy rate above 90%, specifically 90.78% in this case.

Table 2

System Use Metrics Taxonomy

Metric	Measurement	Range	Scale	Type	Citation
Cost and Attacker Success Rate	Network degradation	50% to 300%	Ratio	Intrinsic	(Varadarajan et al., 2015)
	Network round trip time	-	Ratio	Intrinsic	
	Request latency	-	Ratio	Intrinsic	
Security Strength	Password strength	0 to 2^40	Ratio	Intrinsic	(Ma et al., 2014)
	Probability thresholds	0 to 1 (log scale)	Ratio	Intrinsic	
	PRNG security (entropy)	1.03 bits per event (example)	Ratio	Intrinsic	

Detection Rate	Accuracy	90.78% (example)			(Ma et al., 2014)

Varadarajan et al. (2015) revealed that placement vulnerabilities persist in modern clouds, even with advanced isolation technologies. Their empirical research, quantifying the cost and success rate of various attack strategies, demonstrated that achieving co-location is surprisingly easier and less expensive than it should be. They measured performance degradation, network round trip time, request latency, and chances of co-residency, finding significant performance degradation ranging from 50% to 300% on co-resident VMs (Varadarajan, Zhang, Ristenpart, & Swift, 2015).

In a 2014 study, Ma et al. introduced probability-threshold graphs as a faster alternative to guess-number graphs for evaluating password strength. Their empirical results showed that whole-string Markov models, particularly with backoff and end-symbol normalization, consistently outperform PCFGW models and other approaches in terms of cracking efficiency and accuracy across different scenarios and datasets. The study used guess-number graphs, probability-threshold graphs, and Average-Negative-Log-Likelihood (ANLL) as measurements to evaluate the effectiveness of different password models. The guess number ranged from 0 to 2^40 (approximately 1 trillion), the probability threshold ranged from 0 to 1 (in log scale), and the ANLL0.8 values ranged from 12.8 to 33.8. The ANLL values for the best performing models ranged from 12.8 to 23.5, demonstrating their effectiveness in capturing the underlying password distribution and improving password cracking success rates (Ma, Yang, Luo, & Li, 2014).

Measuring Information Disclosure

These information disclosure metrics offer a multi-faceted view into a system's capacity to maintain data protection. By quantifying aspects such as the accuracy of password/vault identification, the potential for unauthorized PII extraction or reconstruction, and the strength of

passwords and overall system security, these metrics enable us to gauge the effectiveness of safeguards against information leaks. Additionally, performance overheads and vulnerability management metrics shed light on the system's efficiency and its ability to adapt and respond to potential threats, thereby contributing to a comprehensive understanding of its overall data protection posture.

Key Metrics Summary

Security Accuracy

- Classification accuracy: Ranges from 50-97%, indicating the system's ability to correctly identify true passwords or vaults.
- Average rank of true password/vault: Ranges from 0.3-70% (single password) and 0.6-41.4% (vault),representing the system's ability to rank true passwords or vaults higher in search results.
- PII Extractability (Precision/Recall): Precision ranges from 0-35%, indicating the accuracy of extracted PII. Recall ranges from 0-23%, showing the percentage of actual PII that was successfully extracted.
- PII Reconstruction/Inference Accuracy: Ranges from 0-18% and 0-70%, respectively, reflecting the accuracy of reconstructing or inferring PII.

Security Performance

- Runtime overhead: Less than 1%, indicating minimal impact on system performance during execution.

- Load-time overhead: Approximately 1µs, demonstrating a negligible increase in system load time.
- Space overhead: 6% file size and 2MB memory, suggesting a relatively small increase in storage and memory requirements.

Security Strength

- Password strength metrics (bits): H∞ (5.0-9.1), λ˜10 (7.5-10.9), G˜0.25 (17.0-26.6), G˜0.5 (19.7-29.3).These metrics quantify the strength of passwords, with higher values indicating stronger passwords.
- TCB size: 350KB - 1.61MB for TEEs and 19MB for Linux kernel, representing the Trusted Computing Base size, which can impact system security.

Vulnerability Management

- Policy size: 9-43 lines of code, indicating the complexity of security policies.
- Number of rules: 3-7, reflecting the number of rules enforced by the security policy.
- Remediation rate: Varies depending on the vulnerability type, with IPv6 showing up to 18% and ICS up to 11%. This metric shows the percentage of contacts who took action to fix the vulnerability.
- Common Vulnerability Scoring System (CVSS) score: Ranges from Low (0-5) to Critical (>=9), assessing the severity of vulnerabilities.

Attacker Success Rate and Uncertainty

- Attacker Success Rate: Not available (N/A).

- Uncertainty set size (USS): Varies, indicating the level of uncertainty in the system's security state.
- Number of gadgets leaked: Up to 100%, reflecting the potential for attackers to exploit vulnerabilities.

Table 3

Information Disclosure Metrics Taxonomy

Metric	Measurements	Ranges	Measurement Scale	Type	Citation
Security Accuracy	Classification accuracy, average rank of true password/vault	Classification accuracy: 50-97%, Average rank: 0.3-70% (single password), 0.6-41.4% (vault)	Ratio, Ordinal	Relative	Chatterjee, R., et al. (2015)
	Perplexity, MI ROC AUC, PII Extractability (Precision/Recall), PII Reconstruction/Inference Accuracy	Perplexity: 9-18, MI AUC: 0.5-0.96, Extraction Recall: 0-23%, Extraction Precision: 0-35%, Reconstruction Accuracy: 0-18%, Inference Accuracy: 0-70%	Ratio	Relative	Lukas, N., et al. (2023)
	Performance overhead (in percentage), utility (relative error)	N/A	Ratio	Relative	Johnson, N., et al. (2021)
	Runtime overhead, load-time overhead, space overhead, memory snapshot analysis, number of encrypted code locators, average indirect targets	Runtime overhead: <1%, Load-time overhead: ~1µs, Space overhead: 6% file size, 2MB memory	Ratio (for overheads), Nominal (for memory snapshot)	Relative (for overheads), Intrinsic (for memory snapshot and counts)	Gudka, K., et al. (2015)
Attacker Success Rate	Success probability of attacks	N/A	Ratio	Relative	Fábrega, A., et al. (2024)
	Uncertainty set size (USS), number of gadgets leaked (total, distinct, syscalls)	USS: Varies, Gadgets leaked: Up to 100%	Ratio	Intrinsic	Seibert, J., et al. (2014)
Security Strength					
	Frequency of breached credential reuse, rate of users ignoring warnings, strength of passwords (before/after resets)	N/A	Ratio, Ordinal	Relative	Thomas, K., et al. (2019)
	Percentage of vulnerable servers over time	N/A	Ratio	Relative	Rescorla, E. (2003)

Table 3 continued

	Message volume, content, geographic distribution, and code entropy	N/A	Ratio (for volume, entropy), Nominal (for content, geographic distribution)	Intrinsic	Reaves, B., et al. (2016)
	Recovery rate, error rate	Keyword attacks: Recovery rate varies (0-90%), Range attacks: Error rate varies (0-0.5)	Ratio	Relative	Kamara, S., et al. (2022)
	CVSS score, TCB size	CVSS score: Critical (>=9), Severe (7-9), Medium (5-7), Low (0-5). TCB size: 350KB - 1.61MB for TEEs, 19MB for Linux kernel	Ordinal (for CVSS), Ratio (for TCB size)	Relative (for CVSS), Intrinsic (for TCB size)	Cerdeira, D., et al. (2020)
	Password strength metrics (bits): H_∞, λ_{10}, $\tilde{G}_{0.25}$, $\tilde{G}_{0.5}$	H_∞: 5.0-9.1, λ_{10}: 7.5-10.9, $\tilde{G}_{0.25}$: 17.0-26.6, $\tilde{G}_{0.5}$: 19.7-29.3	Ratio	Intrinsic	Bonneau, J. (2012)
Vulnerability Management					
	Policy size, number of rules, effectiveness in mitigating CVE bugs, throughput, latency	Policy size: 9-43 lines of code, Rules: 3-7	Ratio (for policy size, rules, throughput, latency), Nominal (for effectiveness)	Intrinsic (for policy size, rules), Relative (for effectiveness), Intrinsic (for throughput, latency if measuring absolute values, Relative if comparing to a baseline)	Muthukumaran, D., et al. (2015)
	Remediation rate (percentage of contacts who took action to fix the vulnerability)	IPv6: Up to 18%, ICS: Up to 11%, DDoS amplifiers: No significant improvement	Ratio	Relative	Li, F., et al. (2016)

Knittel et al. (2021) focused on qualitative analysis of XS-Leaks, while Li et al. (2016) measured remediation rates, observing improvements with direct notifications (Li et al., 2016). Kamara et al. (2022) used recovery rate and error rate to evaluate leakage attacks, and Cerdeira et al. (2020) employed CVSS scores and TCB size to assess vulnerabilities (Cerdeira et al., 2020). Bonneau (2012) introduced new password strength metrics, and Thomas et al. (2019) measured breached credential reuse and password strength changes (Thomas et al., 2019). Rescorla (2003) tracked the percentage of vulnerable servers over time, and Reaves et al. (2016) measured message characteristics in SMS messages (Reaves et al.,2016). Wang et al. (2012) focused on qualitative vulnerability analysis, while Muthukumaran et al. (2015) evaluated policy size, rule count, and performance impact (Muthukumaran et al., 2015). Seibert et al. (2014) used uncertainty set size and leaked gadgets to assess side-channel attacks, and Rösler et al. (2018) conducted qualitative protocol analysis (Rösler et al., 2018). Venkatadri et al. (2018) demonstrated attacks without quantitative measurements, and Fábrega et al.(2024) measured attack success probabilities (Fábrega et al., 2024). Johnson et al. (2021) measured performance overhead and utility, while Lu et al. (2015) evaluated runtime and space overheads (Lu et al., 2015).

Measuring Data Modification

The current security landscape presents a mixed picture when it comes to protecting data. While tools like ARTISAN offer high accuracy in detecting threats, persistent vulnerabilities in areas such as solution compatibility and control-flow integrity defenses expose potential avenues for unauthorized data modification. This underscores the need for continuous improvement in security measures, particularly in safeguarding against unauthorized data alteration.

While performance remains a concern, solutions like Fidelius and ASAP demonstrate that it's possible to implement robust security measures without significantly impacting system performance. This is crucial, as prioritizing data protection should not come at the cost of hindering overall system functionality.

Furthermore, research highlighting concerns about unpatched vulnerabilities and web communication security emphasizes the ongoing challenge of protecting data from unauthorized modification, especially in the face of evolving threats. In this context, log reduction techniques can play a role in optimizing storage while preserving essential information for forensic analysis and identifying potential data breaches or tampering attempts.

In conclusion, while progress has been made in enhancing data protection, there's still room for improvement, particularly in addressing vulnerabilities that could lead to unauthorized data modification. By prioritizing security enhancements, adopting performance-efficient solutions, and leveraging techniques like log reduction, organizations can better protect sensitive information while maintaining acceptable system performance.

Key Metrics Summary

Security Accuracy

- ARTISAN achieves a high precision of 93.9% and recall of 98.8%.

Vulnerability Management

- 47% of solution-test pairs show incompatible or insecure operation.
- Cross-thread stack-smashing attacks bypass all tested CFI defenses.
- 111 unique exploitable handlers impact 379 sites.
- SPIDER finds 67,408 safe patches and 2,278 security patches lacking CVE entries.

- OSV-Hunter detects vulnerabilities in all 74 tested apps, OSV-Free eliminates vulnerabilities in patched frameworks.

Security Performance

- Fidelius adds acceptable overhead to page load and user interaction for secured pages.
- ASAP can select 87% of sanity checks with less than 5% overhead.
- Combined techniques achieve up to 90.7% log size reduction.

Abadi et al. (2003) analyzed SE Protection, focusing on key results, threats mitigated, generalizability, and empirical measurements, but specific details were not provided. Xu et al. (2019) introduced CONFIRM to evaluate control-flow integrity (CFI) protections, demonstrating generalizability but limitations based on target program complexity. Empirical results showed 47% of solution-test pairs exhibited incompatibility or insecurity, with a cross-thread stack-smashing attack defeating all CFI defenses. Yu et al. (2024) proposed ARTISAN, a cost-effective forensics technique for IoT devices, achieving high precision and recall with low overheads compared to state-of-the-art methods. Lone et al. (2022) conducted a randomized control trial on Source Address Validation (SAV) deployment, finding no significant improvement from notification mechanisms, despite some remediation across all groups. Eskandarian et al. (2018) presented Fidelius, utilizing trusted hardware enclaves to protect user secrets even with compromised browsers, demonstrating acceptable overhead on secured pages. Wagner et al. (2015) developed ASAP, a tool maximizing security within a specified overhead budget, effectively protecting against known vulnerabilities with minimal overhead. Steffens & Stock (2020) introduced PMForce, an automated framework for analyzing post Message handler security, uncovering 111 exploitable handlers affecting 379 sites. Sun et al. (2021) systematized

control logic modification attacks and formal verification defenses, highlighting challenges and future research directions. Inam et al. (2023) surveyed provenance-based system auditing literature, evaluating log reduction techniques and their impact on anomaly detection. Machiry et al. (2020) introduced SPIDER to identify "safe patches," evaluating it on real-world commits and CVE patches, highlighting potential unfixed vulnerabilities. Yang et al. (2018) identified Origin Stripping Vulnerabilities (OSV) and developed OSV-Hunter and OSV-Free APIs, demonstrating OSV prevalence and effective mitigation.

Table 4

Data Modification Metrics Taxonomy

Title	Authors	Year	Measurement	Ranges	Scale	Type
Analyzing SE Protection	Abadi, M., Budiu, M., Erlingsson, Ú., & Ligatti, J.	2003	Measurements used in Analyzing SE Protection	Summary of measurement ranges in Analyzing SE Protection	Nominal	Intrinsic
ConFIRM: Evaluating Compatibility and Relevance of Control-flow Integrity Protections for Modern Software	Xu, X., Ghaffarinia, M., Wang, W., Hamlen, K. W., & Lin, Z.	2019	Compatibility, Performance overhead	Compatibility: Pass/Fail, Performance overhead: Percentage increase in execution time	Ratio	Relative
Cost-effective Attack Forensics by Recording and Correlating File System Changes	Yu, L., Ye, Y., Zhang, Z., & Zhang, X.	2024	Precision, Recall, Runtime overhead, Space overhead	Precision and Recall: 0-100%, Runtime overhead: Percentage increase in CPU time, Space overhead: Megabytes of storage consumed per day	Ratio	Relative
Deployment of Source Address Validation by Network Operators: A Randomized Control Trial	Lone, Q. B., Frik, A., Luckie, M., Korczyński, M., van Eeten, M. J. G., & Hernandez Ganan, C.	2022	Remediation rates, Relative risk ratio	Remediation rates: 0-100%, Relative risk ratio: Factor by which one group is different from another in terms of remediation rate	Ratio	Relative

Table 4 continued

Fidelius: Protecting User Secrets from Compromised Browsers	Eskandarian, S., Cogan, J., Birnbaum, S., Brandon, P. C. W., Franke, D., Fraser, F., Garcia, G., Gong, E., Nguyen, H. T., Sethi, T. K., Subbiah, V., Backes, M., Pellegrino, G., & Boneh, D.	2018	Performance overhead on page load and user interaction	Acceptable overhead for secured pages, no impact on unsecured pages	Nominal	Relative
High System-Code Security with Low Overhead	Wagner, J., Kuznetsov, V., Candea, G., & Kinder, J.	2015	Overhead, Sanity level (fraction of protected instructions)	Overhead: 0-100%, Sanity level: 0-100%	Ratio	Relative
PMForce: Systematically Analyzing postMessage Handlers at Scale	Steffens, M., & Stock, B.	2020	Number of vulnerable handlers, Exploitability	Number of handlers: Count, Exploitability: Yes/No	Ratio	Intrinsic
SoK: Attacks on Industrial Control Logic and Formal Verification-Based Defenses	Sun, R., Mera, A., Lu, L., & Choffnes, D.	2021	Not directly applicable (survey paper)	Not directly applicable (survey paper)	Nominal	Intrinsic
SoK: History is a Vast Early Warning System: Auditing the Provenance of System Intrusions	Inam, M. A., Chen, Y., Goyal, A., Liu, J., Mink, J., Michael, N., Gaur, S., Bates, A., & Hassan, W. U.	2023	Log size reduction, anomaly detection performance	Log reduction: Up to 90.7% with combined techniques, Anomaly detection: Varies depending on reduction technique	Ratio	Intrinsic
SPIDER: Enabling Fast Patch Propagation in Related Software Repositories	Machiry, A., Redini, N., Camellini, E., Kruegel, C., & Vigna, G.	2020	Number of safe patches identified, missing patches in active forks	Safe patches: 19.72% of commits, 55.37% of CVE patches. Missing patches: Varies depending on project.	Ratio	Intrinsic
Study and Mitigation of Origin Stripping Vulnerabilities in Hybrid-postMessage Enabled Mobile Application	Yang, G., Huang, J., Gu, G., & Mendoza, A.	2018	Prevalence of hybrid postMessage, effectiveness of OSV-Hunter and OSV-Free	Prevalence: 74 out of 1104 apps, Effectiveness: OSV-Hunter detected vulnerabilities in all 74 apps, OSV-Free eliminated vulnerabilities in patched frameworks.	Nominal	Intrinsic

Measuring Data Destruction

These papers collectively contribute to the understanding and measurement of destruction as it pertains to data protection in various ways. They explore the techniques used by malware and ransomware to destroy data protection, propose methods to detect and prevent such attacks, and develop tools to recover from the damage caused by these attacks. The research approaches range from empirical studies analyzing real-world malware and ransomware samples to formal methods developing and proving the security of protocols. The measurements used include detection rates, false positives, recovery times, performance overheads, and various other metrics that quantify the effectiveness of the proposed solutions in mitigating the destruction of data protection.

Key Metrics Summary

Security Accuracy & Performance

- **Wong et al. (2021)**: This paper primarily used qualitative analysis, focusing on understanding malware analysis workflows and challenges rather than quantitative metrics.
- **Xu et al. (2016)**: The accuracy of crash point identification, stack trace recovery, and vulnerability localization were used to evaluate the effectiveness of the CREDAL tool in identifying and addressing memory corruption vulnerabilities.
- **Anthoine et al. (2021)**: This paper combined formal proofs with empirical experiments, measuring the performance of their PoR protocols, particularly focusing on the tradeoff between persistent storage and audit computation time.

- **Huang et al. (2017)**: The effectiveness of FlashGuard was assessed using recovery time, storage latency, throughput, wear balance, and write amplification factor (WAF) to demonstrate its ability to recover from ransomware attacks while maintaining SSD performance and lifetime.
- **Kharraz et al. (2016)**: The UNVEIL system's performance was evaluated using detection rate, false positives, and new detections, showcasing its ability to identify ransomware threats.

Detection

- **Feng et al. (2016)**: While not explicitly stating metrics, the effectiveness of their approach was demonstrated through its ability to detect and stop ransomware in real-time.
- **Oz et al. (2023)**: Impact analysis was used to evaluate the reach of the R\u00f8B ransomware, demonstrating its ability to encrypt files in various locations, highlighting the potential for confidentiality breaches.
- **Tekiner et al. (2021)** and **Szekeres et al. (2013)**: These survey papers did not directly employ metrics but provided qualitative insights into the techniques and challenges associated with crypto jacking malware and memory corruption attacks, respectively.

Table 5

Data Destruction Metrics Taxonomy

Measurement	Range	Scale	Type	Citation (author, date, citation)
Krippendorff's alpha intercoder reliability	0 to 1	Ratio	Intrinsic	Wong et al., 2021
Accuracy of crash point identification	0% to 100%	Ratio	Intrinsic	Xu et al., 2016
Stack trace recovery (Percentage)	0% to 100%	Ratio	Intrinsic	Xu et al., 2016
Vulnerability localization (Percentage)	0% to 100%	Ratio	Intrinsic	Xu et al., 2016
Extra storage size	N/A	Ratio	Relative	Anthoine et al., 2021
Audit cost	N/A	Ratio	Relative	Anthoine et al., 2021
Data size	N/A	Ratio	Intrinsic	Anthoine et al., 2021
Recovery time	Seconds	Ratio	Intrinsic	Huang et al., 2017
Storage latency overhead	Percentage increase	Ratio	Relative	Huang et al., 2017
Throughput overhead	Percentage decrease	Ratio	Relative	Huang et al., 2017
Wear balance	Standard deviation of remaining lifetime	Ratio	Intrinsic	Huang et al., 2017
Write amplification factor (WAF)	Ratio	Ratio	Relative	Huang et al., 2017
Entropy change	Bits	Interval	Relative	Oz et al., 2023
File size change	Percentage	Ratio	Relative	Oz et al., 2023
Accuracy	Percentage	Ratio	Intrinsic	Oz et al., 2023
Recall	Percentage	Ratio	Intrinsic	Oz et al., 2023
Precision	Percentage	Ratio	Intrinsic	Oz et al., 2023
F1 score	0 to 1	Ratio	Intrinsic	Oz et al., 2023
True Positives (TP)	Count	Ratio	Intrinsic	Oz et al., 2023
True Negatives (TN)	Count	Ratio	Intrinsic	Oz et al., 2023
False Negatives (FN)	Count	Ratio	Intrinsic	Oz et al., 2023
False Positives (FP)	Count	Ratio	Intrinsic	Oz et al., 2023
Detection rate	Percentage	Ratio	Intrinsic	Kharraz et al., 2016
False positives	Percentage	Ratio	Intrinsic	Kharraz et al., 2016

Wong et al. (2021) examined the practices of malware analysts, highlighting the challenges in understanding and combating the destructive capabilities of malware, including those that compromise data through data exfiltration or encryption. Xu et al. (2016) presented CREDAL, a tool designed to locate memory corruption vulnerabilities that can lead to unauthorized data modification or leakage, impacting data protection. Anthoine et al. (2021) introduced new protocols for dynamic Proof of Retrievability (PoR), ensuring data integrity and availability in remote storage. While not directly addressing destruction, their work is crucial in preventing unauthorized data modification or deletion, which are forms of destruction that impact data protection. Huang et al. (2017) proposed FlashGuard, a ransomware-tolerant SSD that enables recovery from encryption ransomware attacks, a direct countermeasure against the destruction of data protection. Feng et al. (2016) presented an approach to detect crypto-ransomware in real-time using deception and behavior monitoring, aiming to prevent data loss before the ransomware's destructive encryption takes effect. Oz et al. (2023) introduced R\u00f8B, a browser-based ransomware that encrypts user files, highlighting a novel attack vector that directly threatens data protection. Tekiner et al. (2021) provided a systematic overview of cryptojacking malware, which, while not directly destructive, compromises data by using a victim's resources without their knowledge. Szekeres et al. (2013) offered a comprehensive overview of memory corruption attacks and defenses, including techniques that can lead to unauthorized data modification or leakage, thus affecting data protection. Kharraz et al. (2016) presented UNVEIL, a system designed to detect ransomware by focusing on its destructive behavior, such as tampering with user files or the desktop.

CHAPTER 4: A CAUSAL MODEL FOR DATA PROTECTION

Chapter 4 establishes a general causal model for data protection, recognizing that data protection is a complex concept influenced by various factors. The chapter leverages Judea Pearl's causal inference framework, adapting it to the cybersecurity domain to enable a more nuanced understanding of how different elements contribute to data protection breaches. It emphasizes the importance of quantifying the impact of security solutions and identifying hidden factors that may influence their effectiveness. The chapter outlines a general model that illustrates the relationships between threats, security measures, and data exposure, and then refines this model by incorporating specific measurements from existing research to evaluate the impact of particular security solutions. By doing so, the chapter aims to provide a structured approach to understanding and measuring data protection, enabling researchers to make more informed decisions about security solution design and implementation.

A Framework for Causal Model Creation

The framework outlines a research approach to understand system data protection. It starts by organizing expert knowledge and mapping system behavior to security properties. Then, a Structural Causal Model (SCM) is built using a directed acyclic graph (DAG) to represent causal pathways and overlay data. The model is validated by focusing on causal relationships and testing it with conditional distributions. Finally, impact analysis is performed through interventions, model adjustments, and counterfactual studies to assess the system's data protection under various scenarios and its applicability to other systems.

Systemize Domain Knowledge

Utilize the following framework as a blueprint to design a research study aimed at comprehending and explaining a system's capacity to maintain data protection. The first step is to systemize expert knowledge, carefully identifying the security properties that will safeguard your system. Next, the researcher maps out the system's behavior in relation to these security properties, similar to charting the flow of people within a building to ensure efficient evacuation routes.

Structure Causal Model (SCM) Creation

Now comes the creation of your Structural Causal Model (SCM). You pinpoint the critical system actions that directly influence security, alongside their quantifiable data. A directed acyclic graph (DAG) is then constructed to visually represent the proposed causal pathways, sketching out the structural framework of the system. With the DAG as the foundation, you overlay data onto the nodes and edges, adding detailed specifications to model. The explicit assumptions about the relationships within the system, graphically representing them with formal analysis.

Model Validation

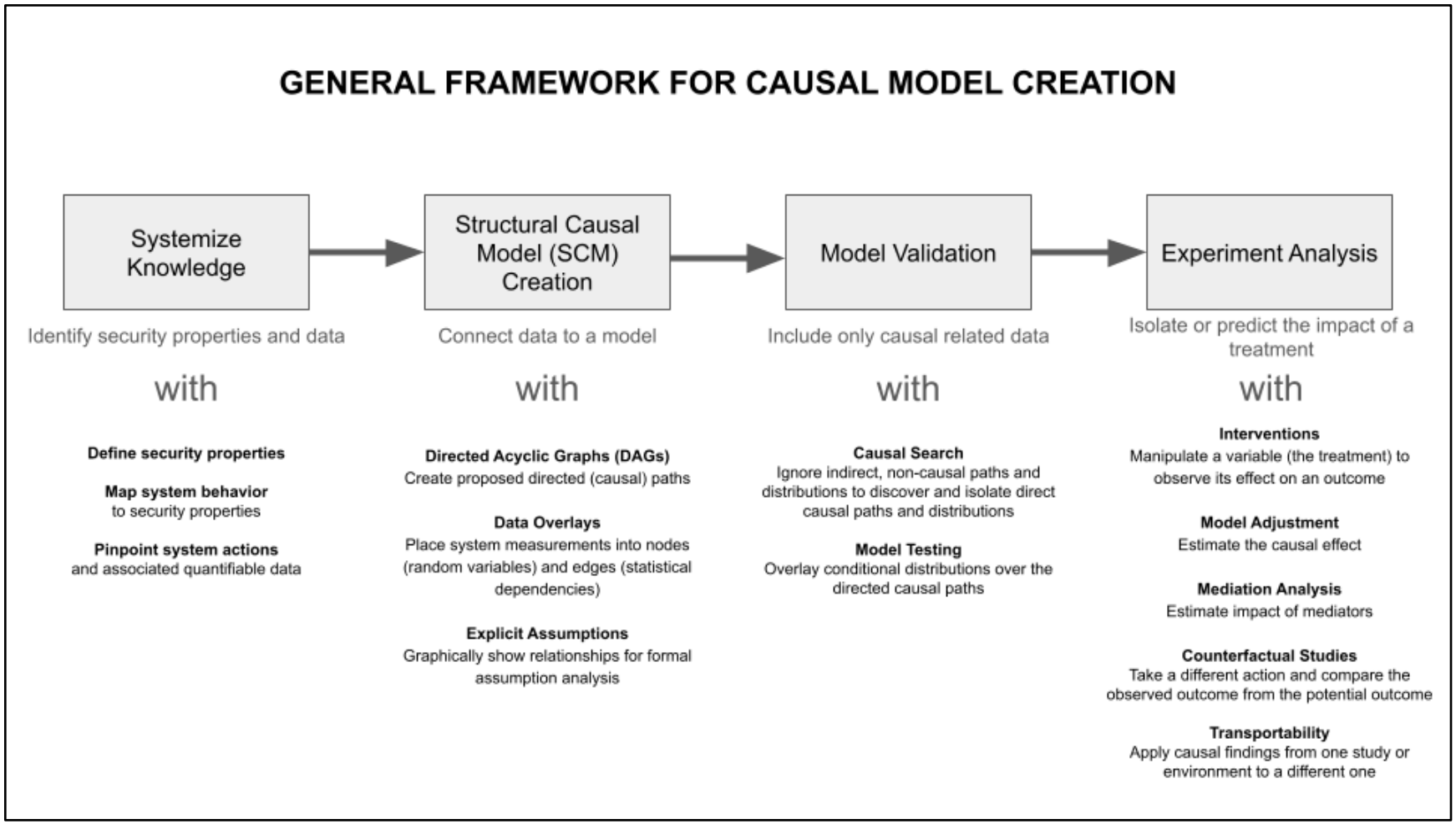
The next crucial phase is model validation. Here, you focus solely on causal relationships, discarding any non-causal data. The model is tested by overlaying conditional distributions over the causal paths, simulating stress tests on your design. Causal search helps isolate the direct causal relationships.

Experiment Analysis

With a refined model, a researcher can now conduct interventions, manipulating variables to observe their impact on outcomes. The research should employ model adjustment to estimate the causal effect of these interventions. Counterfactual studies allow you to explore "what-if" scenarios, assessing the potential impact of alternative actions. Lastly, the final parts of analysis include transportability, examining whether the model's insights can be applied to other systems or environments. This comprehensive framework provides a robust methodology for constructing and evaluating causal models, ensuring the security of system design.

Figure 2

The general framework for causal model creation is illustrated in figure 2.



A Causal Model for Data Protection

Section 4.2 of the dissertation focuses on developing a causal model for data protection, drawing upon the theoretical framework and literature review presented in earlier chapters. It recognizes that data protection is a multifaceted concept influenced by various factors, and thus, a causal modeling approach is adopted to understand the complex interplay of these factors. The chapter leverages Judea Pearl's causal inference framework, adapting it to the cybersecurity domain to enable a more nuanced understanding of how different elements contribute to data breaches. It emphasizes the importance of quantifying the impact of security solutions and identifying hidden factors that may influence their effectiveness.

The chapter outlines a general model that illustrates the relationships between threats, security measures, and data exposure. It then refines this model by incorporating specific measurements from existing research to evaluate the impact of particular security solutions. The chapter provides causal models for various aspects of data protection, including:

- **Authorized Access:** This model explores the factors influencing whether access to data is granted only to authorized individuals. It considers metrics such as security strength, accuracy, vulnerability management, and detection.
- **System Use:** This model examines how the use of a system can impact data protection, focusing on factors like the cost and attacker success rate, security strength, and detection rate.
- **Information Disclosure:** This model delves into the unauthorized release of sensitive information, considering metrics related to security accuracy, performance, strength, vulnerability management, and attacker success rate and uncertainty.

- **Data Modification:** This model investigates the unauthorized alteration of data, incorporating metrics related to security accuracy, vulnerability management, and the trade-off between security and performance.
- **Data Destruction:** This model focuses on the threats of data destruction, primarily from malware and ransomware attacks, and considers metrics such as detection rates, false positives, recovery times, and performance overheads.

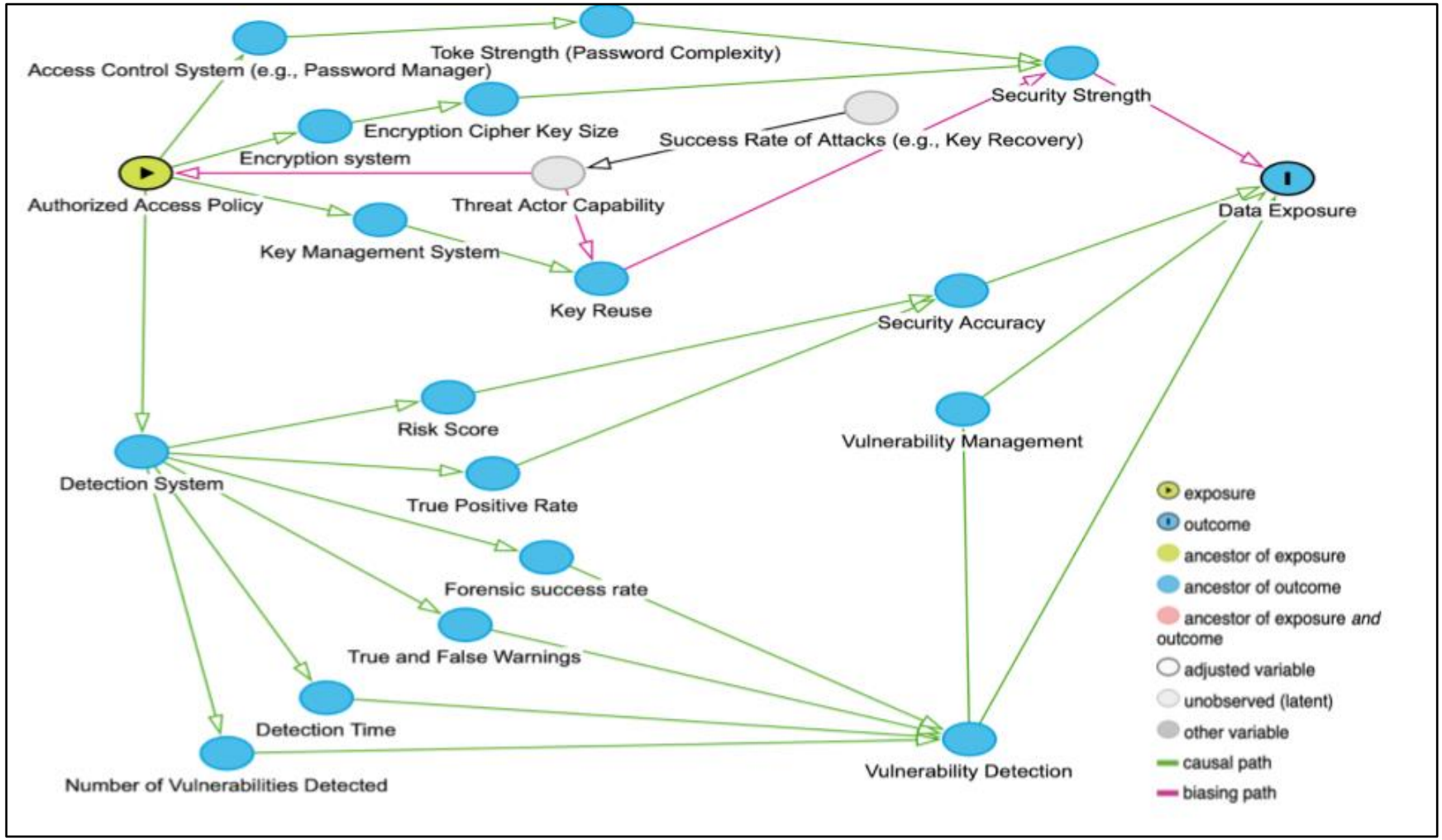
Finally, the chapter integrates these individual models into a general causal model for data protection, providing a holistic view of the factors influencing data exposure. It concludes by proposing a set of metrics for measuring data protection, emphasizing the importance of quantifiable measures for effective security assessment and improvement.

Authorized Access Causal Model

Chapter 4 .2.1 introduces a causal model for authorized access, a critical aspect of data protection. It constructs a Directed Acyclic Graph (DAG) to visually represent the relationships between various metrics that contribute to authorized access. The root node, "Authorized Access," is influenced by four key factors: Security Strength, Security Accuracy, Vulnerability Management, and Vulnerability Detection. Each of these factors is further broken down into specific metrics, such as password strength, detection rates, and the number of vulnerabilities detected. The DAG serves as a framework for understanding the complex interplay of these metrics and their impact on ensuring that only authorized entities can access sensitive information.

Figure 3

The causal model for authorized access overlays the security properties and corresponding measurements from the literature review.



The Authorized Access DAG will be structured hierarchically, starting with a root node representing the overarching "Authorized Access Policy." From this root, the first level branches into four key categories: Security Strength, Security Accuracy, Vulnerability Management, and Vulnerability Detection. Each of these Level 1 nodes further expands into a series of specific metrics that define and measure their respective areas. For instance, Security Strength encompasses metrics like password strength and 2FA adoption, while Vulnerability Management includes aspects such as the number of vulnerabilities detected and false positives. The connections, or edges, in this DAG flow downwards, linking the root to the Level 1 categories and then each category to its specific metrics. While potential relationships between individual metrics might exist, the current information doesn't explicitly outline them. The visual representation of this DAG would use nodes of varying shapes or colors to differentiate between the hierarchical levels and directed edges to illustrate the dependencies and flow of influence within the Authorized Access framework.

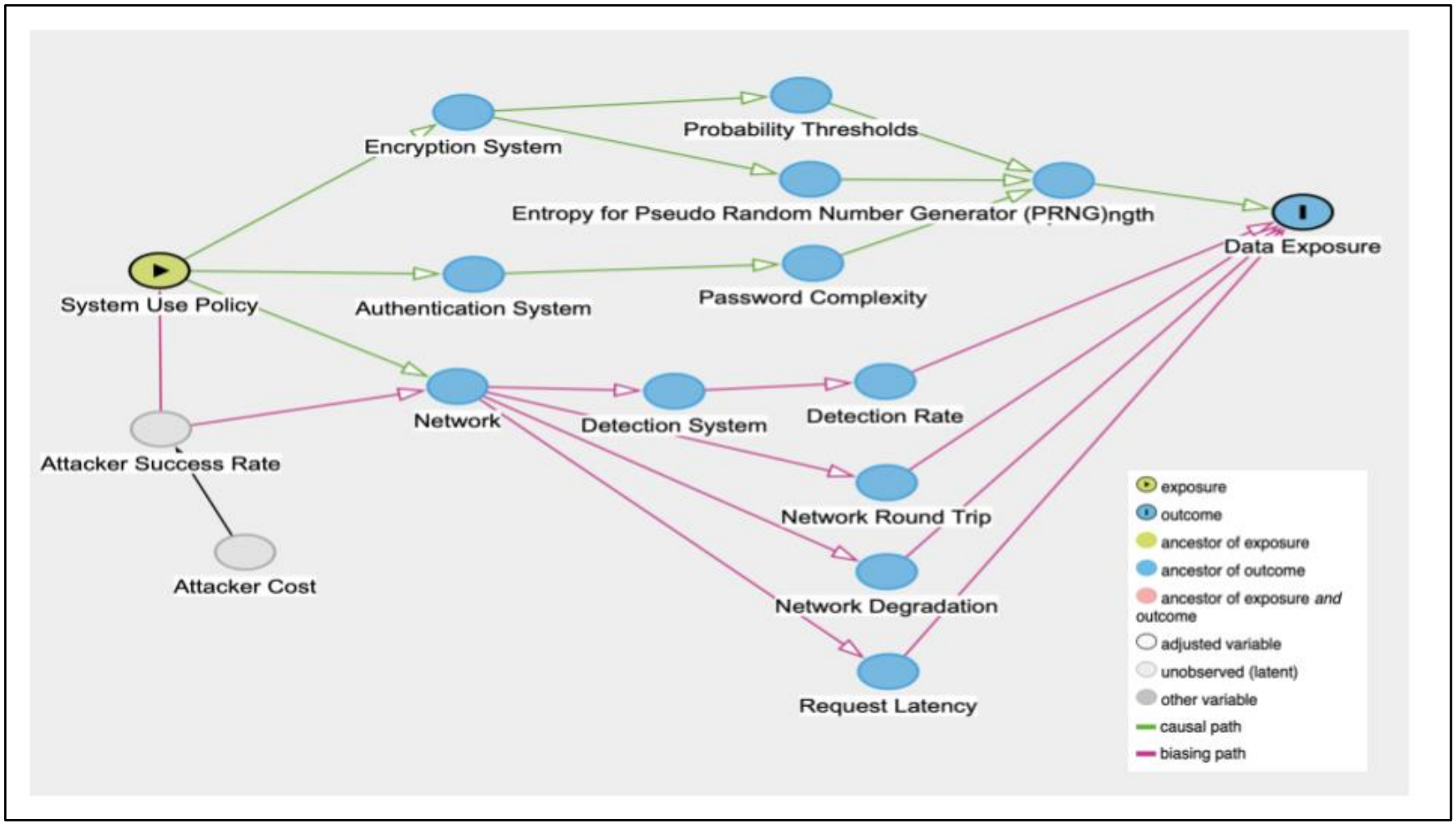
System Use Causal Model

The overall data protection of a system’s use is influenced by three primary factors: the cost and potential success rate for attackers, the inherent strength of the system's security measures, and its ability to detect threats. These factors are further broken down into specific metrics. For instance, the cost and success rate for attackers might be influenced by network performance metrics like degradation, round trip time, and request latency. Security strength is evaluated through password strength, probability thresholds, and the security of random number generation. Meanwhile, the detection rate hinges on the accuracy of the system's threat identification. These relationships are visualized in a Directed Acyclic Graph (DAG), where the

root node "System data protection" branches out into these three primary factors, which then further connect to their respective metrics. While the provided information outlines a basic framework, it's important to recognize that there might be additional, more complex relationships between these metrics that could further refine our understanding of system data protection.

Figure 4

The causal model for the system overlays the security properties and corresponding measurements from the literature review.

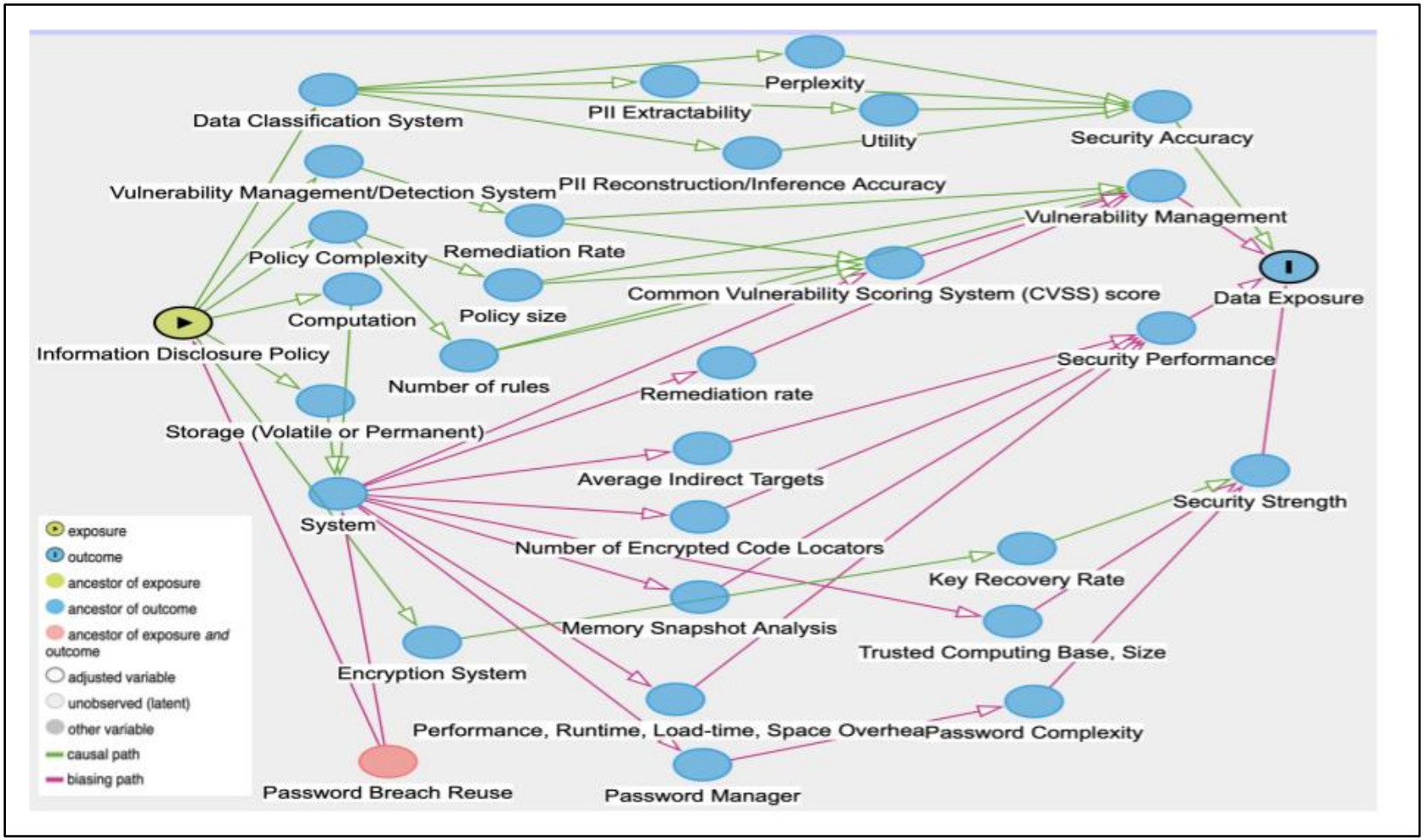


Information Disclosure Causal Model

The risk of information disclosure is a complex issue influenced by several interconnected factors. The effectiveness of a system's security hinges on its accuracy in handling sensitive information, the performance impact of its security measures, the strength of its defenses, and its ability to manage vulnerabilities. Additionally, understanding the likelihood of a successful attack and the uncertainties within the system's security posture is vital for a comprehensive risk assessment. This intricate relationship is illustrated in a Directed Acyclic Graph (DAG), where "Information Disclosure" is the central concern, branching out into these key factors which are further detailed by specific metrics. It's important to note that this DAG offers a structural overview and that in reality, there are likely interdependencies between these metrics across different categories. Some metrics might even have a more direct impact on information disclosure than others. As more insights into the system and its specific security challenges become available, this DAG can be further refined and expanded to provide a more nuanced understanding of the risk landscape.

Figure 5

The causal model for information disclosure overlays the security properties and corresponding measurements from the literature review.

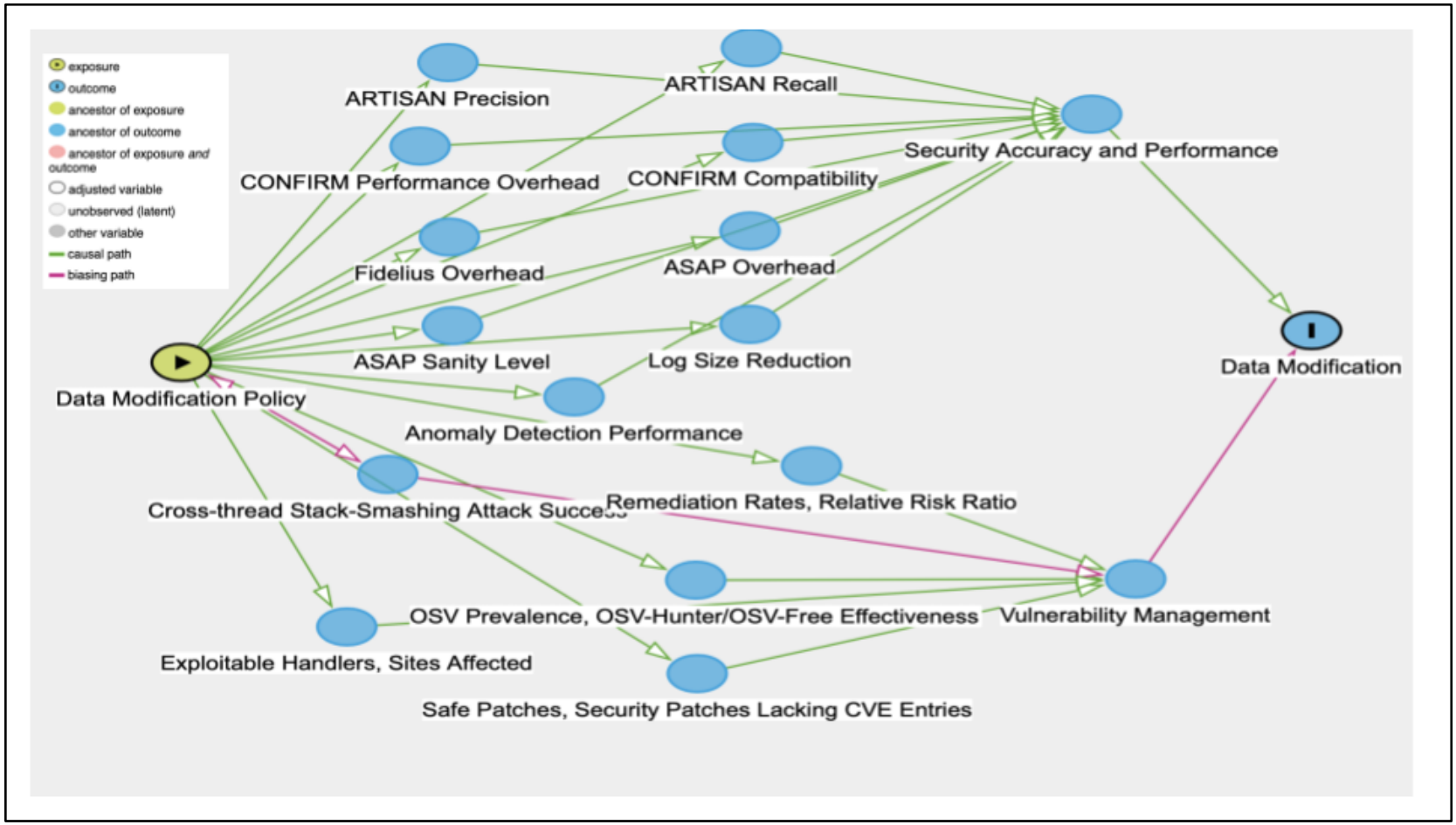


Data Modification Causal Model

The causal model for data modification focuses on understanding and preventing unauthorized changes to data, which is a critical aspect of maintaining data protection. The risk of unauthorized data modification is a core concern that is influenced by a combination of factors, visualized in a Directed Acyclic Graph (DAG). At the heart of this DAG is the "Data Modification" root node, which branches into three primary areas: Security Accuracy, Vulnerability Management, and the interplay between Security Accuracy and Performance. Security Accuracy delves into the precision and recall of detection mechanisms, while Vulnerability Management assesses the system's ability to handle potential exploits and apply necessary patches. Meanwhile, the Security Accuracy & Performance node examines the balance between robust security measures and their impact on system efficiency. Each of these areas is further elaborated through specific metrics, such as ARTISAN Precision for Security Accuracy or Remediation Rates for Vulnerability Management. While this DAG provides a valuable structural overview, it's important to acknowledge that the relationships between these metrics might be more intricate than a simple hierarchical representation, and the DAG itself can be refined and expanded as more information about the system and its security landscape becomes available.

Figure 6

The causal model for data modification overlays the security properties and corresponding measurements from the literature review.

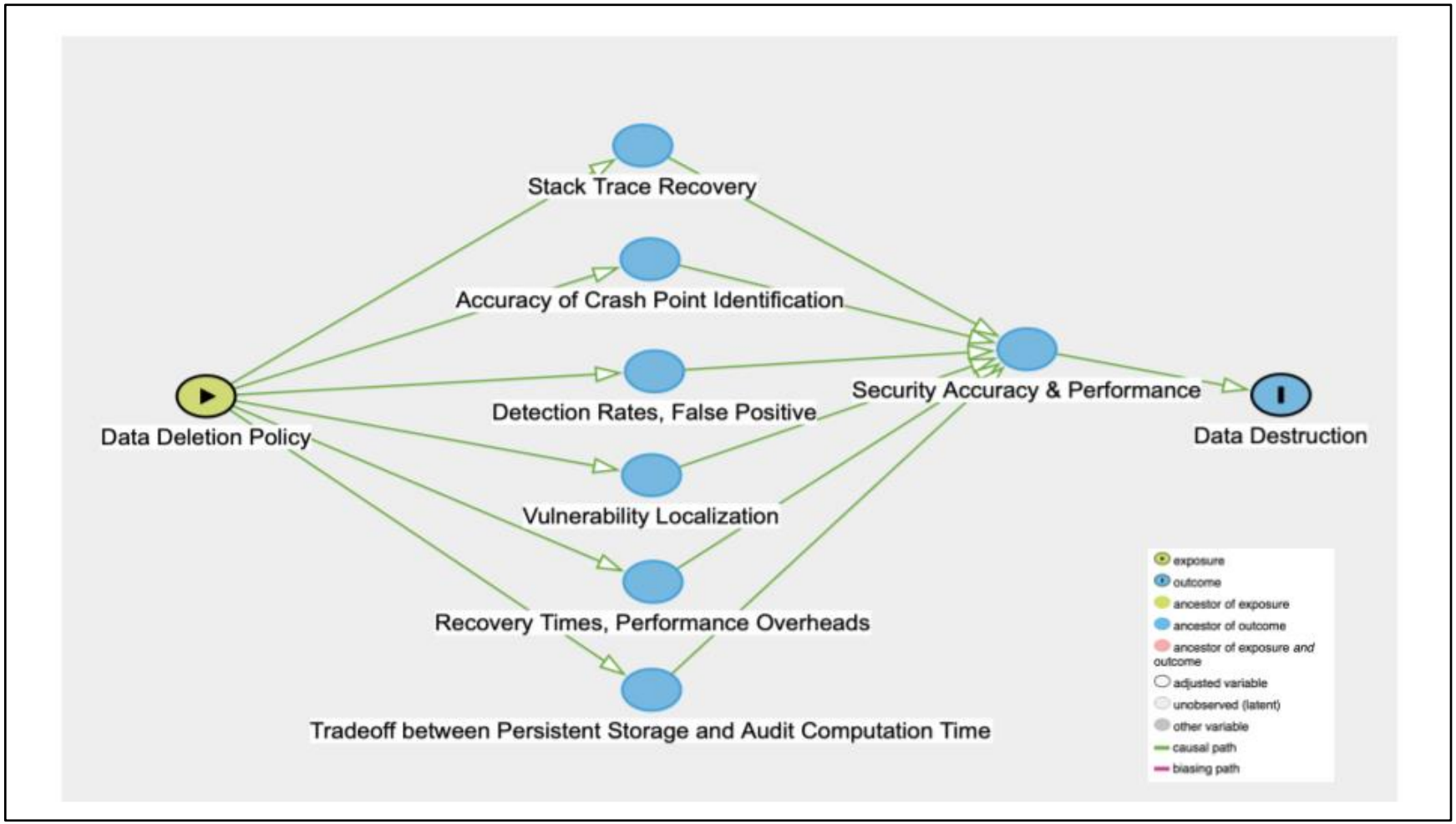


Data Destruction Causal Model

The causal model for data destruction emphasizes the role of malware and ransomware attacks as the primary source of threats to data protection. These attacks can lead to the compromise of data protection through data exfiltration, encryption, or deletion, with the latter two also potentially causing a loss of data availability. The model highlights the importance of detection rates and false positives in evaluating the effectiveness of security mechanisms against these attacks. Additionally, it underscores the significance of recovery times and performance overheads in assessing the efficiency of data recovery solutions, such as FlashGuard, which aim to restore data after a destructive attack. Finally, the model recognizes the importance of identifying and addressing memory corruption vulnerabilities using tools like CREDAL, as these vulnerabilities can be exploited to facilitate data destruction. By considering these various factors and metrics, the causal model provides a comprehensive framework for understanding and mitigating the risks associated with data destruction in the context of data protection.

Figure 7

The causal model for data destruction overlays the security properties and corresponding measurements from the literature review.



The primary threat to data integrity can manifest in two main forms: "Data Exfiltration" and "Data Encryption & Deletion." Both of these attacks directly compromise data protection, but the latter can also lead to a loss of data availability, making it inaccessible even to legitimate users. To combat these threats, various metrics and concepts come into play. Detection rates and false positives are crucial in evaluating the effectiveness of preventive measures, while recovery times and performance overheads assess the efficiency of recovery tools and techniques. Furthermore, the accuracy of identifying crash points, recovering stack traces, and localizing vulnerabilities is critical in addressing potential exploits that could lead to data destruction. Finally, there's an inherent tradeoff between persistent storage and the computational cost of maintaining data integrity and availability, highlighting the need for balance in security solutions. This interconnected web of attack vectors and defense mechanisms underscores the importance of both proactive and reactive measures in protecting sensitive information.

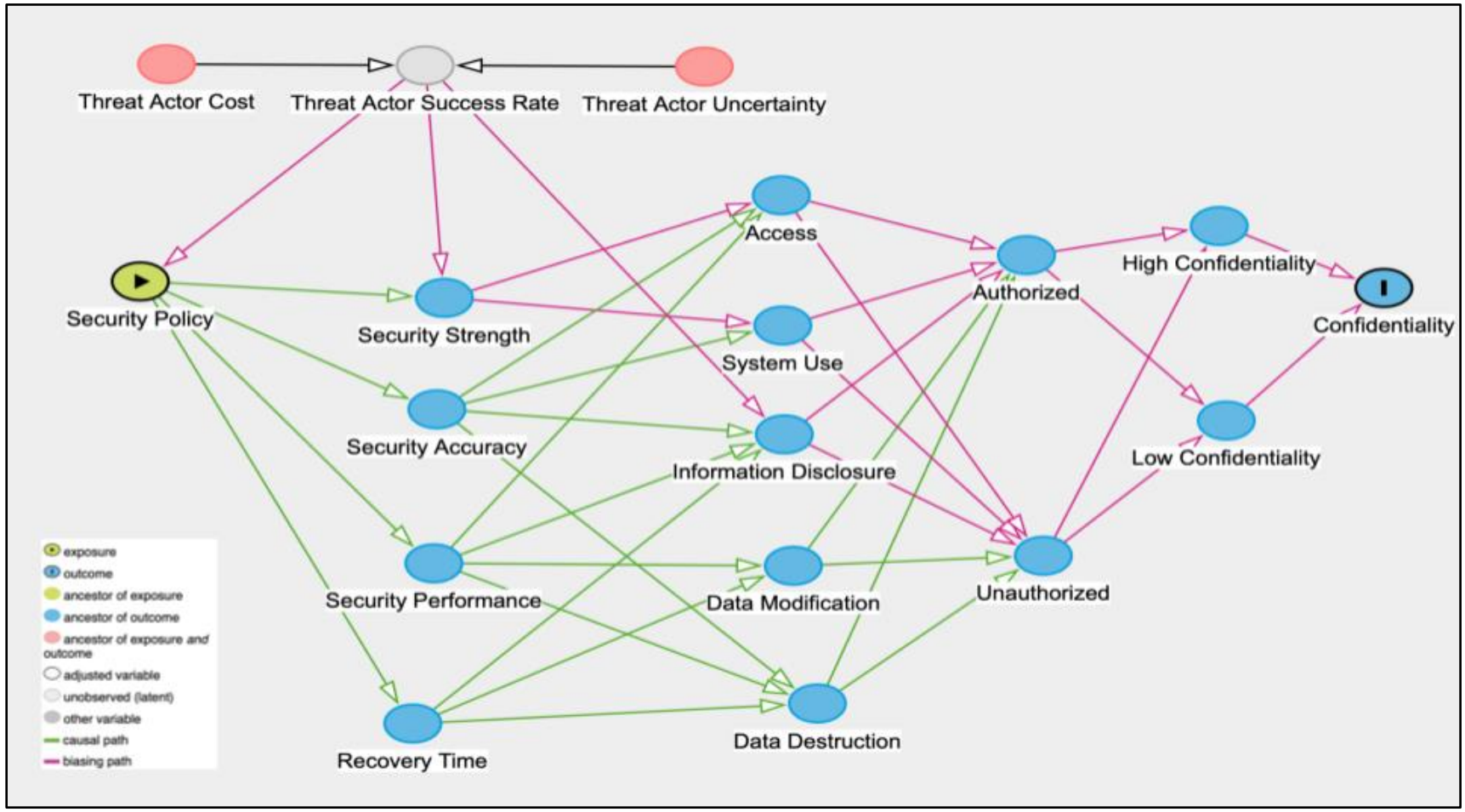
General Causal Model for Data Protection

A general causal model for data protection is crucial for researchers looking to conduct rigorous experiments and ensure the validity of their scientific studies. A general causal model for data protection, grounded in thorough literature review and established causal frameworks, offers researchers a systematic understanding of the complex factors influencing data protection. By explicitly mapping causal relationships, the model empowers researchers to pinpoint key factors and their interactions, leading to rigorous experiment designs that control for confounding variables and isolate the true causal effects of security measures. This understanding also allows for predicting the impact of security interventions, guiding informed decision-making. Ultimately, the model's strong theoretical foundation and comprehensive

approach enhance the validity and generalizability of scientific conclusions, boosting the trustworthiness and impact of data protection research.

Figure 8

The causal model for data protection overlays the threat actor, security properties, system capabilities, and data protection levels security properties from the literature review.



The risk of data exposure and subsequent data breaches is a complex issue influenced by various factors. Threats like malware and vulnerabilities constantly challenge the security measures put in place, which include mechanisms to control authorized access, system use, data modification, and destruction. The success of these measures depends on their strength, accuracy, and the system's ability to detect threats. Moreover, the attacker's potential for success and any uncertainties in the system's security posture also play a significant role. Though not directly tied to data protection, cost and performance overhead can indirectly influence it by impacting the feasibility of implementing robust security measures. In the event of a breach, the ability to recover data quickly is crucial in mitigating the damage. This interconnected web of threats, defenses, and potential outcomes underscores the need for a multi-faceted approach to protect sensitive information, highlighting the dynamic and evolving nature of data protection.

Metrics for Data Protection

Establishing well-defined data protection metrics is crucial for several reasons. First, metrics allow for relative comparisons between different security measures or systems, helping to make informed choices about the most effective approaches for protecting sensitive information. Second, they enable the replication and validation of research studies by providing quantifiable measures, promoting transparency and ensuring findings can be verified in various environments. Finally, metrics offer a standardized language for assessing and communicating a system's capacity to safeguard sensitive data, leading to a wider understanding of data protection across different fields. In short, data protection metrics are essential for evaluating, comparing, and enhancing security measures. They improve decision-making, support research reproducibility, and foster a shared understanding of how well systems protect sensitive

information. The following generally key metrics for data protection have been identified with a definition for each:

- **Security Strength**: Measures the robustness of the security measures, influencing their ability to withstand attacks and prevent data exposure. Stronger security measures generally lead to lower chances of data exposure and data compromise.
- **Security Accuracy**: Evaluates the precision and effectiveness of security measures in correctly identifying and handling sensitive data, contributing to the prevention of an intrusion or cyber-attack.
- **Security Performance** - Security performance measures how effectively threats and vulnerabilities are detected, enabling timely response and mitigation to prevent data exposure and data breaches.
- **Recovery Time -** Recovery time refers to the duration it takes for an organization to restore its systems, applications, and data to their normal operational state after an intrusion or cyberattack.
- **Threat Actor Success Rate -** Threat actor success rate quantifies how often attackers are able to breach defenses and compromise systems or data.
- **Threat Actor Uncertainty**: Threat actor uncertainty quantifies the lack of complete knowledge or predictability that attackers face when planning and executing cyberattacks.
- **Threat Actor Cost**: Threat actor cost refers to the resources, effort, and expenses incurred by malicious actors to plan, execute, and succeed in a cyberattack.

CHAPTER 5: DATA PROTECTION EXPERIMENTATION

Chapter 5 explores how to assess the data protection of a system, especially when sensitive information might be at risk. It introduces a visual modeling approach using Causal Bayesian Networks (CBNs) to simplify this complex task. A CBN allows for flexible analysis and comparison of different situations, ultimately helping researchers make informed decisions to protect sensitive data.

Essentially, this chapter emphasizes the importance of thoroughly testing these models and using them to uncover the cause-and-effect relationships that influence data protection. By rigorously examining how threat actors, security properties, system capabilities, and volume of individuals with access to confidential data interact, researchers can identify better methods to protect sensitive information.

To ensure these models accurately reflect reality, the chapter highlights the use of "d-separation" and "causal search." D-separation helps distinguish true cause-and-effect relationships from mere correlations in the data (Pearl, 2009). Causal search algorithms then use this information to build and refine models to ultimately accurately represent how the system data protection works (Spirtes, Glymour, & Scheines, 2000). This rigorous process ensures that the causal models are reliable and can be used to make effective predictions and interventions to strengthen data protection.

The focus shifts to a comprehensive impact analysis of the causal models for data protection. This analysis will encompass a range of critical factors that influence the protection of sensitive information, including authorized access, system use, information disclosure, data destruction, and data modification. By employing mediation analysis, researchers will gain a deeper understanding of the pathways through which these factors affect data protection,

uncovering the underlying mechanisms and potential points of vulnerability. Additionally, the average data exposure rate will serve as a crucial metric to evaluate the overall effectiveness of data protection measures and provide a holistic assessment of a system's ability to safeguard sensitive data.

Interventions, adjustments, counterfactuals, mediation analysis, and transportability are essential components of Pearl's causal inference framework. They empower researchers to move beyond observational studies, estimate causal effects, understand causal mechanisms, and generalize findings across different contexts. Embracing these concepts is key to advancing scientific knowledge and making informed decisions based on a sound understanding of cause-and-effect relationships.

Judea Pearl's causal inference framework revolutionizes how we approach research questions, enabling us to move beyond mere associations and delve into the realm of cause-and-effect relationships. Central to this framework are the concepts of interventions, adjustments using the adjustment formula, counterfactual reasoning, mediation analysis, and transportability. Each plays a crucial role in enhancing the rigor and applicability of research findings.

Interventions represent deliberate actions taken to manipulate a specific variable in a causal system. They are fundamental for establishing causal effects because they disrupt the natural flow of events, allowing researchers to isolate the impact of a particular factor. By performing interventions, we can answer "what if" questions and estimate the causal effect of an intervention on an outcome of interest.

In observational studies, where interventions are often not feasible or ethical, the adjustment formula provides a powerful tool for estimating causal effects. It allows researchers to control for confounding variables - factors that influence both the treatment and the outcome -

by mathematically adjusting for their impact. This adjustment helps to isolate the true causal effect of the treatment, even in the presence of confounders.

Counterfactual reasoning involves imagining alternative scenarios and asking, "what would have happened if?" questions. It helps researchers estimate the causal effect of an intervention by comparing the observed outcome under the actual treatment with the hypothetical outcome that would have occurred had a different treatment been applied. Counterfactuals provide a powerful framework for understanding causal effects in situations where controlled experiments are not possible.

Together, these concepts form a powerful toolkit for causal inference. They allow researchers to go beyond simple associations and uncover the true causal relationships that underlie observed phenomena. By understanding the mechanisms of causation, researchers can design more effective interventions, make informed policy decisions, and contribute to a deeper understanding of the world around us.

A Causal Bayesian Network for Data Protection

Section 5.1 explores the dynamics of data protection using Causal Bayesian Networks (CBNs), a powerful tool for visualizing and analyzing complex cause-and-effect relationships. CBNs enable the mapping of various factors, including security policies, security properties, system capabilities, and threat actors, to affect the overall data protection of a system. By examining different scenarios through the lens of CBNs, a researcher now has a general way to examine data sets to understand and better explain data protection and mitigate the risks of intrusions and cyber-attacks.

Assume a dataset corresponding to a scenario where a security policy is implemented on a system and its ability to keep information confidential is based on the security effectiveness,

system capabilities, and number of individuals with system access. A rational threat actor's success to perform an intrusion or cyber-attack is based on the intruder cost and knowledge (i.e., or its uncertainty) of a targeted system.

Imagine creating a CBN where:

- ***P*:** Security Policy (a set of rules, guidelines, and procedures to protect confidential information)
- ***E*:** Security Effectiveness (how well the system protects data with security strength, accuracy, performance, recovery time)
- ***S*:** System capability (the degree to which a system can access, use, disclose, modify, or delete confidential data)
- ***I*:** People (the number of people with authorized access to confidential data)
- ***D*:** Data Protection Level (the degree to which data could be exposed, altered, or destroyed)
- ***T*:** Success Rate of a Threat Actor
- ***U*:** Threat Actor Knowledge and Uncertainty
- ***K*:** Threat Actor Cost

In the data protection CBN shown in figure 10, a path from node P to node ***D*** is defined as a sequence of linked nodes starting at ***P*** and ending at ***D***. ***P*** is the security policy and cause of data protection level ***D*** if there exists a causal path from ***P*** to ***D***, a path whose links are pointing from preceding nodes toward the following nodes in a sequence. For example, the path $\boldsymbol{P \rightarrow E \rightarrow S \rightarrow I \rightarrow D}$ is causal path and the path $\boldsymbol{T \rightarrow P \leftarrow E}$ is non-causal. The security policy ***P*** has a

direct influence on the data protection level (***D***) through the causal path of security property effectiveness ***E***, system capability (***S***), Individual users (***I***). The data exposure can be influenced by the threat actor's success rate. The success rate of a threat actor is dependent on its capability, cost of the intrusion or attack, and knowledge or level of uncertainty of a targeted system.

Kripke Structure for Data Protection

A Kripke structure is used as a generic set of states for a system and not dependent on a specific implementation or platform. The model is a subset of system states to meet Western legal requirements to protect data. The Kripke structure for data protection is dynamic, and therefore, needs to be mapped to both authorized and unauthorized user access events with corresponding data protection levels.

Data protection can be generally defined as:

T is defined to be less than or equal to the smallest duration between state changes where only a single event could occur between t and t+1 t .

(t) symbol represents a general system for time t T.

E(t) is the set of all authorized and unauthorized events and represents a subset of all system at time t T from equation 3.5.

D(t, e) function represents the specific data protection state for an initial subset set of all system events for t T or $E=(E_t : t \ T)$.

D is the set of data protection states occurring after an event e.

DE(t,e,s) is the transition function of data protection for both authorized and unauthorized events e E(t) corresponding to a specific data protection system state d D(t).

Necessity for the Kripke Structure

Theorem 1 - The 16 events represented in Equation 2.1 are necessary to represent an FSM for system confidentiality for all authorized and unauthorized time periods in E(t).

Proof - The necessity condition requires a confidentiality event occurred during the times in E(t)only if it is represented in the confidentiality FSM model. The necessity condition is proved with contradiction. Let e1be a trusted event. If e1 occurred in E(t) and E(t+1) ande1SE(t,e,s), then either multiple events occurred between t+1or e1E(t). The definition of T does not allow multiple events to occur during the exact same time and if e1SE(t,e,s), e1 could not have occurred in E(t) as defined.

5.2.2 Sufficiency for the Kripke Structure

Theorem 2 -The 16 variables defined in Equation 3.3 and 3.4 are sufficient to represent the system confidentiality FSM for the times in E(t) .

Proof -The sufficiency condition requires that an event be represented in the model only if it existed or occurred. Let e1 be a trusted event If trusted event e1 occurred in E(t) and SE(t,e,s) but e1 did not occur at t, t+1, or t+n, then e1E(t). e1must have occurred in , because all

authorized access and unauthorized access events are included in E(t) , SE(t,e,s) includes all transition states for E(t) and C(t, e) includes both events and states derived from E(t) and SE(t,e,s).

Completeness for the Kripke Structure

Theorem 3 - The Kripke structure completely includes the desired security properties and events for system data protection.

Proof - The completeness of the data protection is proved through exhaustion in the table below. The data protection system properties are transformed into security events.

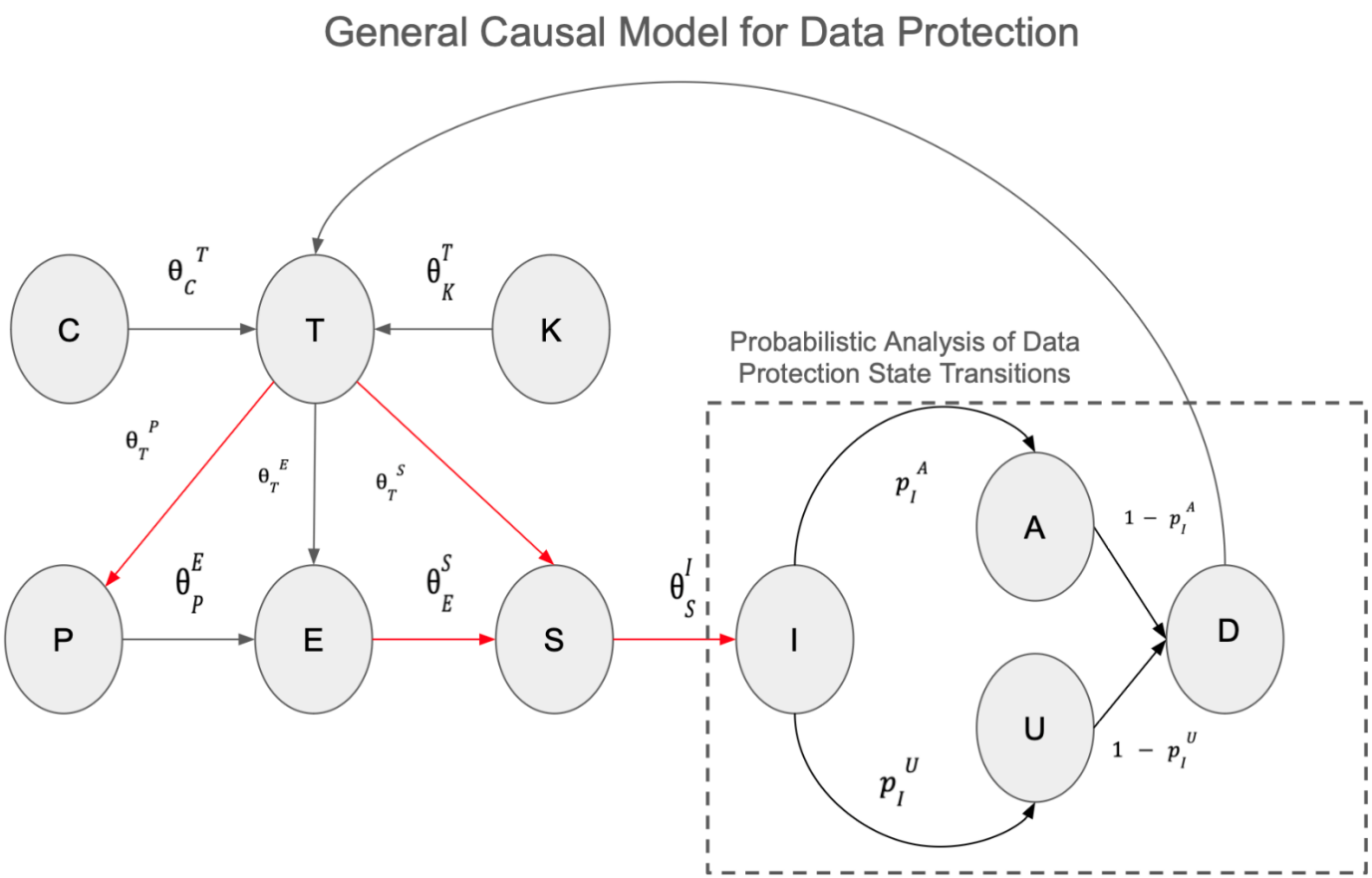
Table 6

Completeness for the Kripke Structure

System Property	Security Event (State)
Access	Authorized
Access	Unauthorized
Use	Authorized
Use	Unauthorized
Disclosure	Authorized
Disclosure	Unauthorized
Modification	Authorized
Modification	Unauthorized
Destruction	Authorized
Destruction	Unauthorized

Figure 9

The general causal model for data protection overlays the security properties and corresponding observational measurements from the literature review.



Expected Scenarios Impacting the Data Protection Level

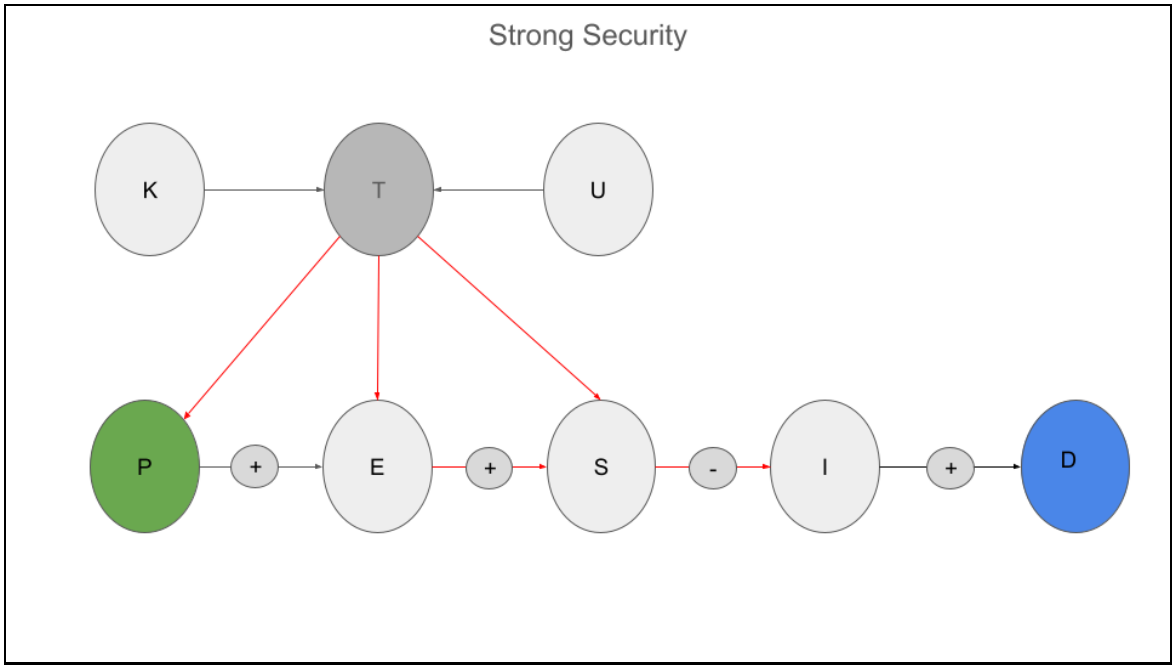
The following are four general expected scenarios that illustrate how security policies, effective security, system capabilities, and threat actors are thought to influence the level of data protection within a system. By analyzing these scenarios, researchers can test whether the expected scenarios are true and gain valuable insights into what constitutes a resilient system—a system capable of minimizing the risk of data intrusions and attacks. The scenario list is not comprehensive but rather serves as a minimum set of scenarios to test the security of a system.

Scenario 1 - Strong Security

In a strong security scenario, a robust security policy ($\boldsymbol{P}+$) would lead to better security effectiveness ($\boldsymbol{E}+$), which limit's a user's capability ($\boldsymbol{S}-$) on using, disclosing, modifying, or deleting confidential data, and also decreases the total ($\boldsymbol{I}-$), ultimately increasing the data protection level ($\boldsymbol{D}+$).

Figure 10

The figure shows how strong security measures can either strengthen or weaken the relationships between different system properties.

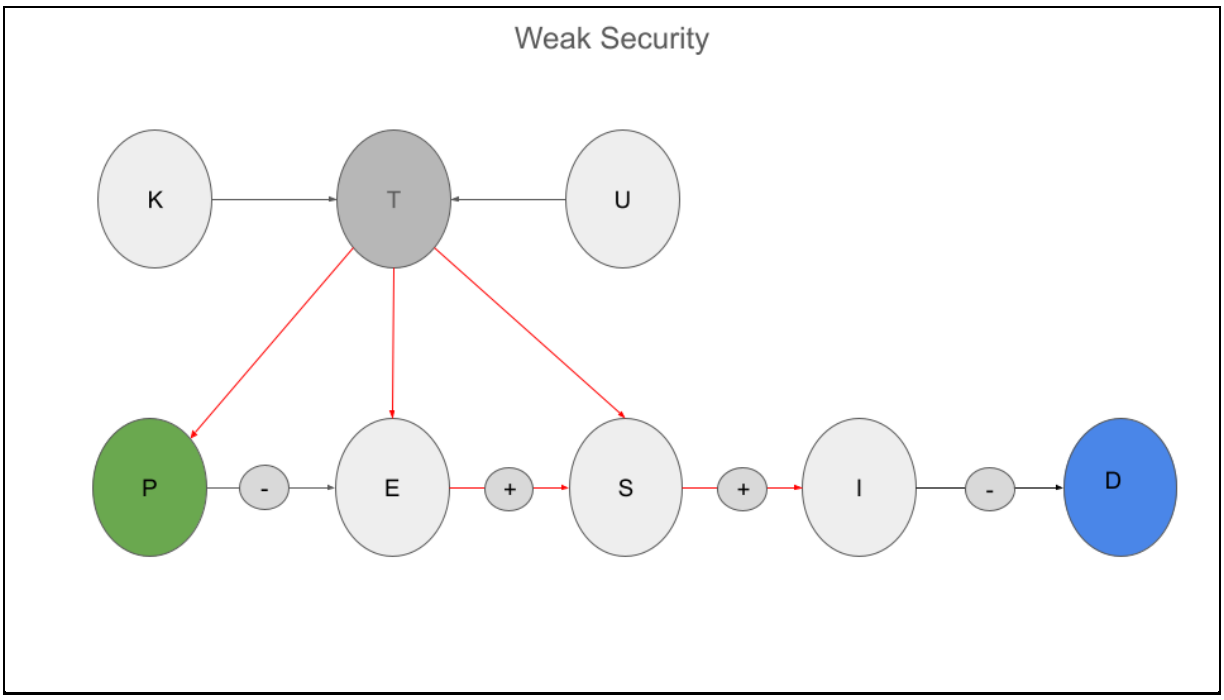


Scenario 2 - Weak Security

In a weak security scenario, a fragile security policy ($\boldsymbol{P}-$) would lead to worse security effectiveness ($\boldsymbol{E}-$), which increases a system's capability ($S+$) to access, use, disclose, modify, or delete confidential information and also increases the total number of individuals with authorized and unauthorized access ($\boldsymbol{I}+$), ultimately decreasing the data protection level ($\boldsymbol{D}-$).

Figure 11

The figure shows how weak security measures can either strengthen or weaken the relationships between different system properties.

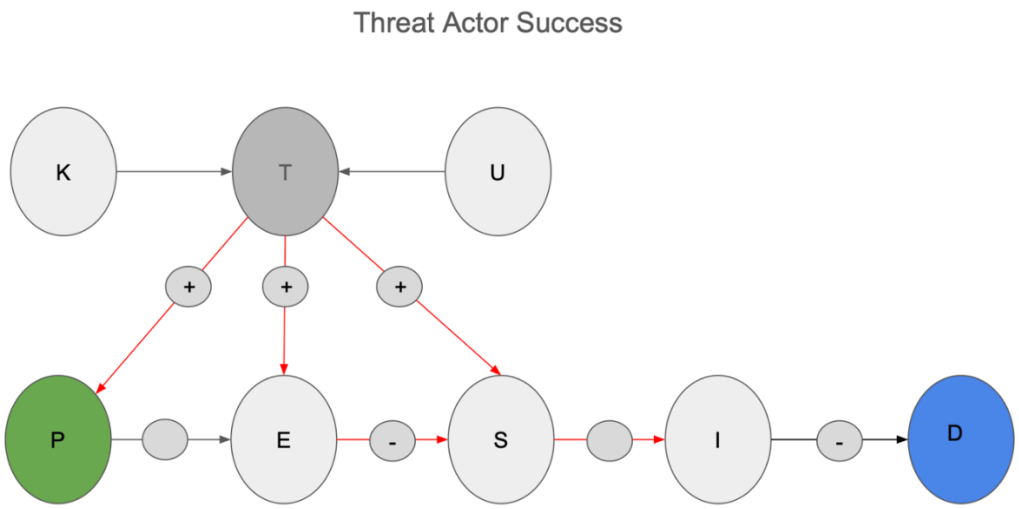


Scenario 3 - Threat Actor Success

In the threat actor success scenario, the data protection level is ultimately lowered ($\boldsymbol{D}-$) when a threat actor's success rate increases ($T+$). A threat actor's success rate decreases the security effectiveness ($\boldsymbol{E}-$) and increases a system's capability ($\boldsymbol{S}+$) to access, use, disclose, modify, or delete confidential information, ultimately decreasing the ($\boldsymbol{D}+$). The success rate of a threat actor is dependent on its capability, cost (i.e., and its intent) of the intrusion or attack, and knowledge or level of uncertainty of a targeted system.

Figure 12

The figure shows how threat actor success can either strengthen or weaken the relationships between different system properties.

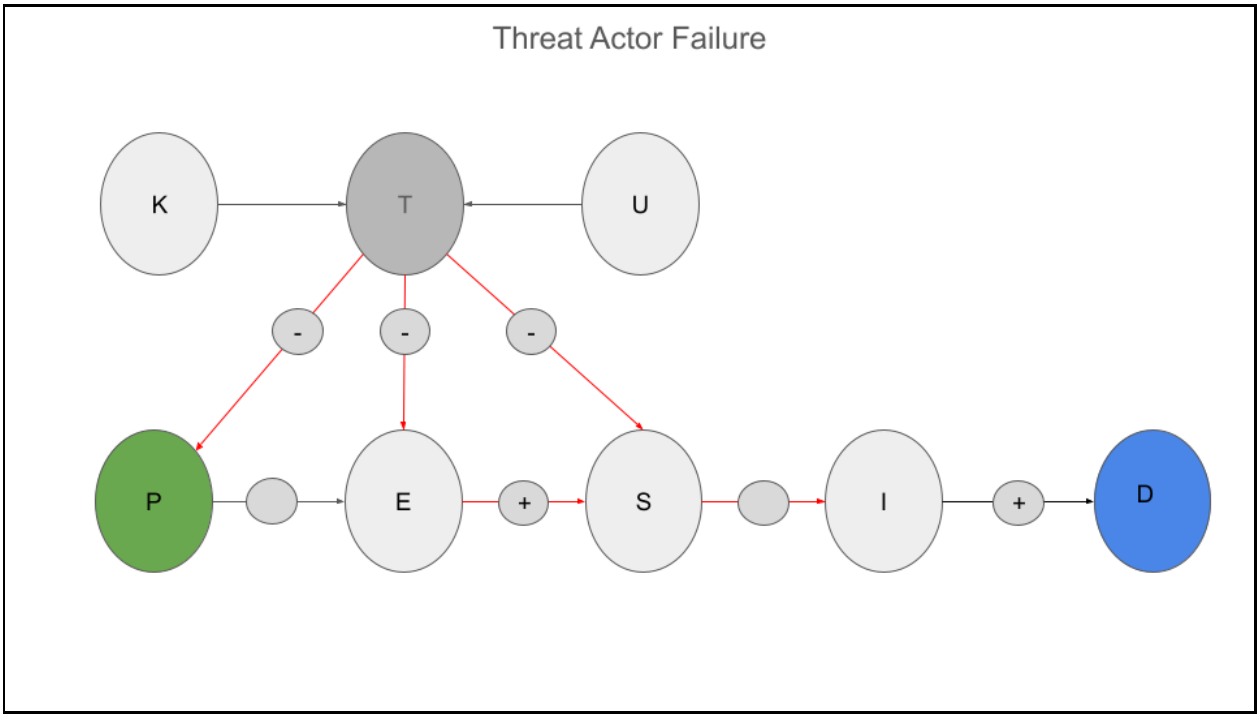


Scenario 4 - Threat Actor Failure

In the threat actor failure scenario, the data protection level is ultimately increased ($\boldsymbol{D}+$) when a threat actor's success rate decreases ($\boldsymbol{S}-$). As the threat actor's success rate decreases, this increases the security effectiveness ($\boldsymbol{E}+$) and is independent of a system's capability ($\boldsymbol{S}$) to access, use, disclose, modify, or delete confidential information, ultimately increasing the data protection level ($\boldsymbol{D}+$). The failure of a threat actor is dependent on its capability, cost (i.e., intent) of the intrusion or attack, and knowledge or level of uncertainty of a targeted system.

Figure 13

The figure shows how threat actor failure can either strengthen or weaken the relationships between different system properties.



Estimating Influence of System Effectiveness, System Capability, and Individual Access

A research is studying a scenario of the effects of a security policy (P) on data protection level (i.e., data exposure) (D). The researcher believes the security policy can influence data protection levels through two primary paths:

1. Direct Path: A security policy (P) and its security effectiveness (E) (security protections) directly influences data exposure (D).
2. Indirect Path: A security policy (P) influences the system capabilities and total number of individuals with access to the confidential data, changing the likelihood of a data breach, influencing the data protection level (D).

Path-Specific Effects:

- Direct Effect of (P) on (D): This quantifies data exposure solely due to the change in security policy, regardless of whether system capabilities or total number of users change.
- Indirect Effect of (P) on (D) (through E, S, I): Indirect effect quantifies data exposure specifically, because the security policy led to changes in security effectiveness, system capabilities, and access to authorized and unauthorized users, ultimately influencing the data protection levels.

To calculate the indirect effect of $\boldsymbol{P}$ on C through the path $\boldsymbol{P} \rightarrow \boldsymbol{E} \rightarrow S \rightarrow I \rightarrow D$, we can use the following formula:

$$IE(P \rightarrow E \rightarrow S \rightarrow I \rightarrow D) = \sum_{E,S,I} \left[E_{E,S,I}[D \mid P = 1, E = e, S = s, I = i] - E_{E,S,I}[D \mid P = 0, E = e, S = s, I = i]\right] P(E = e, S = s, I = i \mid P = 1) \quad (18)$$

Controlling for Threat Actor Success

A research is studying a scenario of the effects of certain security policy (***P***) that causes a change in confidential levels (i.e., data exposure) (***D***). The observational data shows those who implement the security policy tend to have fewer data breaches. However, the researcher suspects that a common cause (co-founder) might influence the security of the systems and data breach frequency, due to threat actor (***T***) success rates. The threat actor success rate (***T***) could be a confounder, because ***T*** can influence both the likelihood of a data breach attempt and the frequency of successful data exposures changing the data protection level (***D***). The backdoor path is a non-causal path between the security Policy (***P***) and data protection level (***D***) that has an arrow pointing into the treatment. In this case $\boldsymbol{P} \leftarrow \top \rightarrow \boldsymbol{D}$ is a backdoor path. To isolate the true causal effect of the security policy (***P***) on data protection level (***D***), the researcher needs to block the backdoor path through conditioning on the confounder (***T***), which accounts for the influence of the threat actor success rate. Analyzing the data while controlling for threat actor success rate (***T***), the researcher can determine if the security policy (P) truly causes a reduction in data exposure (***D***), independent of the influence of a threat actor success rate. The following is the adjustment formula to control for the threat actor success rate:

$$P(D = d) \mid do(P = p)) = \sum z\, P(T = t) \mid P = p, D = d)\, P(D = d) \qquad (19)$$

Estimating Data Protection Levels

A research studies a security policy (P) and inherently its security effectiveness (E) and its impact on data protection levels (D). The researcher believes the security effectiveness might influence data protection levels both directly (by improving the defense of the system) and

indirectly (by limiting system capabilities and reducing the total number of individuals with access to confidential data).

Variables:

- ***P*****:** Security Policy (a set of rules, guidelines, and procedures to protect confidential information)
- ***E*****:** Security Effectiveness (how well the system protects data with security strength, accuracy, performance, recovery time)
- ***S*****:** System capability (the degree to which a system can access, use, disclose, modify, or delete confidential data)
- ***I*****:** People (the number of people with authorized access to confidential data)
- ***D*** Data Protection Level (the degree to which data protection data could be exposed, altered, or destroyed)
- ***T*****:** Success Rate of a Threat Actor
- ***U*****:** Threat Actor Knowledge and Uncertainty
- ***K*****:** Threat Actor Cost

3. Average Causal Effect (ACE): The average causal effect (ACE) is the average difference in the potential outcomes between units assigned to the security effectiveness and units assigned to the system without the security mechanisms.

$$ACE = E[Y(1) - Y(0)]\, E[D \mid do(P = 1)] - E[D \mid do(P = 0)] \quad (20)$$

where

$D(1)$ is the potential outcome if the system has security effectiveness (protections with a security policy)

$D(0)$ is the potential outcome if the system lacks security effectiveness (absent protections with a security policy)

$E[]$ denotes the expectation or average over the entire population

Intervention Experiments

Intervention Study: Imagine a system experiencing a high rate of data breaches or low data protection level (***D***). The security team is considering implementing a new security policy (P) to increase security effectiveness (E) and increase the data protection level ***D*** (i.e., reduce the frequency of data breaches). However, they also know that simply limiting system capabilities and total number of users with access could also contribute to reducing data breaches.

Intervention: The security team decides to implement a new security policy to increase the security effectiveness of the system (P = 1).

Causal Effects:

- **Total Effect:** The overall impact of the new security system on data breaches, considering both its direct effects and any indirect effects through changes in the system.

- **Direct Effect:** The impact of the new security system on data breach frequency, assuming no changes to system capabilities or total number of users with access to confidential data. This isolates the effect of user and system capabilities.

- **Indirect Effect:** The impact of the security system on the frequency of data breaches that occurs *because* of changes in in the system capabilities and the total number of users accessing the system. For example, if the new security system leads to restricted capabilities and access to confidential data.

The following variables:

- $\boldsymbol{E}$: Security Effectiveness (1 = implemented, 0 = not implemented)
- $\boldsymbol{D}$: Data Protection Levels (i.e., Frequency of Data Breach Exposure)
- $\boldsymbol{S}, \boldsymbol{I}$: System Capabilities and Total Number of Users with Access to Confidential Data

Total Effect (TE):

$$TE = E[D \mid do(E = 1)] - E[D \mid do(E = 0)]] \quad (21)$$

This compares the expected data breach rate when the security system is implemented versus not implemented.

Direct Effect (DE):

$$DE = E[D \mid do(E = 1), S, I] - E[D \mid do(E = 0), S, I]] \quad (22)$$

This compares the expected data protection level (i.e, data breach rate) when the security system is implemented versus not implemented, *while holding the system capabilities and total number of individuals constant*.

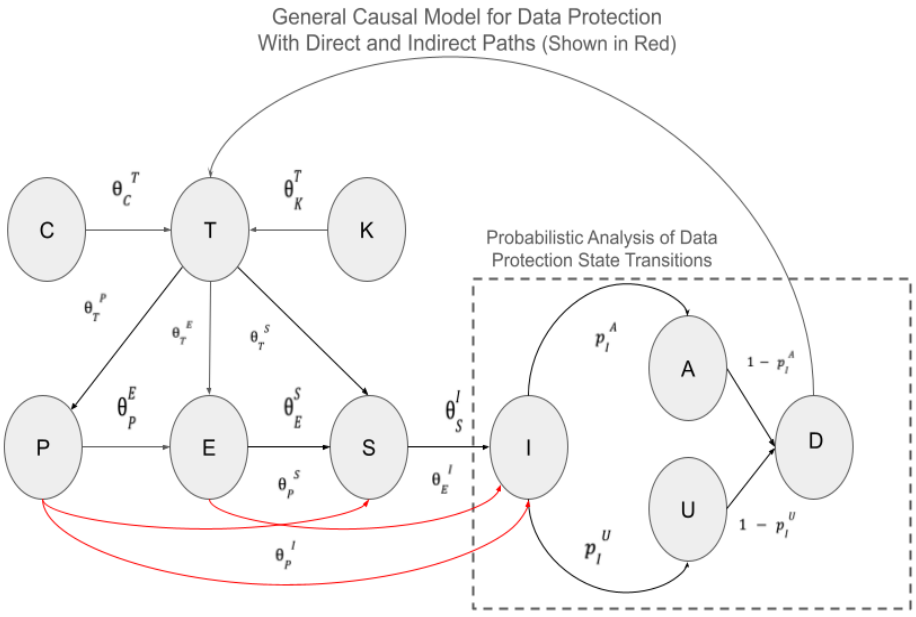
Indirect Effect (IE):

$$IE = [TE - DE] \tag{23}$$

The indirect effect is calculated by subtracting the direct effect from the total effect.

Figure 14

The figure shows the direct and indirect paths (in red) for the general causal model for data protection.



Counterfactual Experiments

Scenario: A company is concerned about data breaches (Y) and is considering offering a new security policy (P) that includes security training for employees (E) (such as avoiding email phishing). The security team believes the security training could directly reduce data breaches. However, the security team also know that limiting users access and system capabilities (S, I), such as allowing only a small amount of data to be read each hour (stopping bulk data exfiltration), could also contribute to reducing exposure of confidential data (D).

Counterfactual Question: Imagine an employee who is currently experiencing high levels of phishing emails leading to lower confidential levels (i.e., data breaches) (D = 1) and did not participate in the security training (E = 0). The counterfactual question is: What would this employee's susceptibility to phishing emails (and in turn contributing to data breaches) if they had participated in the security training (E = 1), assuming the user had the same system capabilities and access to confidential data (S, I)?

The following variables:

- $\boldsymbol{E}$: Security Effectiveness (Security Training) (1 = implemented, 0 = not implemented)
- $\boldsymbol{D}$: Data Protection Levels (i.e., Frequency of Data Breach Exposure)
- $\boldsymbol{S, I}$: System Capabilities and Total Number of Users with Access to Confidential Data
- $\boldsymbol{m}$: constant level for the mediator

Counterfactual Total Effect

$$CTE = E[D(1, E(1)) - D(0, E(0))] \quad (24)$$

This compares the expected confidential level (i.e., data breaches) if tåhe employee had participated in the security training (E) versus their observed data protection level D given they did not participate in the security training , while keeping the same system capabilities and access to confidential data (S, I).

Counterfactual Direct Effect

$$CDE(m) = E[D(1, m) - D(0, m)] \quad (25)$$

D(1, m): The potential outcome (D) when the treatment (E) is set to 1 (treatment condition), and the mediator (S, I) is fixed at a specific level 'm'.

D(0, m): The potential outcome (C) when the treatment (E) is set to 0 (control condition), and the mediator (S, I) is again fixed at the same level 'm'.

E[...]: The expected value (average) over the population.

Counterfactual Indirect Effect

$$CIE(a) = E[D(a, M(a)) - D(a, M(0))] \quad (26)$$

D(a, S, I(a)): The potential outcome (D) when the treatment (E) is set to a specific level 'a' (this could be 1 for treatment or 0 for control), and the mediator (S, I) takes on the value it would naturally have when E=a.

D(a, S, I(0)): The potential outcome (D) when the treatment (E) is set to '1', but the mediator (S, I) is fixed at the value it would have taken under the control condition (E=0).

E[...]: The expected value (average) over the population.

CHAPTER 6: LEVELS OF DATA PROTECTION

A framework for classifying the capabilities and behavior of data protection security mechanisms and systems. The framework introduces levels of data protection system behavior and ability to respond to a variety of threat actor intrusions and cyber-attacks, providing a common language to do relative comparisons and measure progress across the research community in developing robust data protection strategies. There is a brief review for the data protection definition and metrics from chapter 2 and introduces three principles key to understanding data protection strategies. The research proposes "Levels of Data Protection" based on three key principles of system behavior, threat actor capabilities, and security metrics.

Data Protection and Metrics

Data protection is the safeguarding of sensitive information from unauthorized access, use, disclosure, modification, and destruction. It involves implementing security measures to ensure that only authorized individuals can interact with data appropriately. The effectiveness of these measures determines the overall level of data protection. Key metrics for understanding data protection include:

- **Security Strength:** Measures the robustness of security mechanisms against attacks, influencing the likelihood of data exposure or compromise. Stronger security measures generally lead to better protection.
- **Security Accuracy:** Evaluates the precision and effectiveness of security measures in identifying and handling sensitive data, contributing to the prevention of intrusions and cyberattacks.

- **Security Performance:** Measures how efficiently threats and vulnerabilities are detected, enabling timely responses to prevent data exposure and breaches.
- **Recovery Time:** Assesses the duration required to restore systems and data to normal operation after an attack, reflecting the ability to recover from security incidents.
- **Threat Actor Success Rate:** Quantifies how often attackers successfully breach defenses and compromise systems or data, providing insights into the effectiveness of security measures.
- **Threat Actor Uncertainty:** Quantifies the lack of complete knowledge or predictability attackers face, influencing their ability to plan and execute attacks.
- **Threat Actor Cost:** Refers to the resources and expenses incurred by attackers, which can influence their motivation and sophistication.

These metrics offer a comprehensive view of data protection by considering the strength, accuracy, and performance of security mechanisms, the recovery capabilities of the system, and the factors influencing attacker behavior. By analyzing these metrics, organizations can gain valuable insights into their overall data protection posture and identify areas for improvement.

Key Principles for Data Protection

Three principles—system behavior, threat actor access methods, and data protection metrics—are interconnected and crucial for understanding and evaluating the security of a system. System behavior provides insights into how a system operates under various conditions, including normal operation and during or after a cyberattack. This understanding helps establish baselines and identify anomalies that could indicate malicious activity. Threat actor access methods consider the diverse tactics, techniques, and procedures employed by attackers to gain

unauthorized access to systems. Analyzing these methods allows for the development of comprehensive security measures that can defend against a wide range of threats. Finally, data protection metrics provide quantifiable measures to assess the effectiveness of security mechanisms in safeguarding data. These metrics offer insights into the strengths and weaknesses of security controls, enabling informed decision-making and continuous improvement in data protection strategies. By considering these three principles together, a holistic and robust approach to cybersecurity can be achieved, ensuring that systems are protected from various threats and vulnerabilities.

Levels of Data Protection

Data protection is a multifaceted challenge that requires a comprehensive approach to safeguard sensitive information from unauthorized access, use, disclosure, modification, and destruction. This involves implementing a layered security strategy that addresses various levels of protection, from ensuring conformity to established standards to building resilience against evolving threats. This section explores five distinct levels of data protection—conformity, correctness, effectiveness, resistance, and resilience—and highlights research contributions that exemplify each level. By understanding these levels and their associated security considerations, organizations can develop a robust data protection posture that effectively mitigates risks and ensures the confidentiality, integrity, and availability of critical information.

Level 1 - Conformity

Conformity in data protection focuses on adhering to predefined security requirements and specifications, ensuring that systems meet established standards and best practices. This involves designing and implementing security mechanisms that align with industry standards,

regulatory frameworks, and organizational policies. Examples of research contributing to conformity include static analysis methods for web applications (Sun et al., 2011), automated policy generation for microservices (Li et al., 2021), and tools like OSSPolice that ensure compliance with open-source license requirements and security vulnerability databases (Kim & Lee, 2017). Furthermore, evaluating password strength (Ma et al., 2014) and ensuring sufficient entropy in random number generation (Ma et al., 2014) also contribute to conformity by meeting the requirements of various security standards.

Level 2 - Correctness

Correctness emphasizes the accurate implementation and functioning of security mechanisms, ensuring they behave exactly as intended and are free from errors or vulnerabilities that could undermine their purpose. This involves rigorous testing and validation to identify and address any deviations from expected behavior. Research focusing on correctness includes measuring the functional correctness of code (Acar et al., 2016), evaluating the accuracy of execution path recovery (Wu et al., 2023), and accurately detecting access control misconfigurations (Xiang et al., 2019). Additionally, ensuring that random number generators produce sufficient entropy (Ma et al., 2014) and verifying the proper functioning of password/vault classification mechanisms (Chatterjee et al., 2015) are also crucial for achieving correctness in data protection.

Level 3 - Effectiveness

Effectiveness in data protection centers on the ability of security mechanisms to protect data while maintaining operational efficiency under normal circumstances. This requires a balance between robust security and minimal impact on system performance and usability.

Research demonstrating effectiveness includes evaluating the performance and scalability of secure data storage techniques like Group ORAM (Maffei et al., 2015), balancing privacy enhancements with security in risk-based authentication (Wiefling et al., 2021), and measuring the effectiveness of multi-factor authentication (MFA) and risk-based authentication (RBA) (Gavazzi et al., 2023). Furthermore, using accuracy as a metric for spam filters, demonstrating low performance overheads in security mechanisms (Gudka et al., 2015), and showcasing effective data protection with acceptable performance overhead (Eskandarian et al., 2018; Wagner et al., 2015) all contribute to demonstrating the effectiveness of security measures.

Level 4 - Resistance

Resistance focuses on the ability of security mechanisms to withstand intrusions and cyberattacks, preventing unauthorized access and data breaches. This involves implementing strong defenses and employing proactive measures to mitigate potential threats. Research contributing to resistance includes detecting vulnerable open-source software (OSS) versions in mobile apps (Kim & Lee, 2017), addressing firmware update vulnerabilities (Wu et al., 2023), and mitigating vulnerabilities in IoT devices (Jia et al., 2021). Evaluating the impact of attacks on network performance (Varadarajan et al., 2015), ensuring strong password security (Ma et al., 2014), and assessing vulnerabilities using metrics like TCB size and CVSS scores (Cerdeira et al., 2020) are also crucial for enhancing resistance.

Level 5 - Resilience

Resilience emphasizes the ability of security mechanisms to adapt and recover from adverse events that could compromise data, ensuring business continuity and minimizing the impact of security incidents. This involves implementing recovery mechanisms, continuous

monitoring, and adaptive security measures. Research related to resilience includes continuous validation and forensics for faster recovery (Xiang et al., 2019), improving privacy in risk-based authentication systems (Wiefling et al., 2021), and providing a taxonomy of attacks and safeguards for open-source software supply chains (Ladisa et al., 2022). Furthermore, recovering from network performance degradation caused by attacks (Varadarajan et al., 2015), demonstrating a high remediation rate for vulnerabilities (Li et al., 2016), and utilizing log reduction techniques for efficient incident analysis (Inam et al., 2023) all contribute to building resilience in data protection.

Data Protection Level Examples

The data protection framework uses a level, matrixed approach toward classifying systems with in-depth security. This section illustrates the data protection level framework, providing concrete examples from various research studies. Table 6 presents a framework for classifying data protection systems based on five levels: conformity, correctness, effectiveness, resistance, and resilience. Each level builds upon the previous one, representing increasing security capabilities. Conformity focuses on adhering to predefined security standards, while correctness ensures accurate implementation and error-free operation. Effectiveness balances strong security with efficient performance, and resistance emphasizes withstanding attacks and preventing breaches. Finally, resilience enables systems to adapt and recover from security incidents. The table provides examples of research that exemplify each level, illustrating how different security mechanisms and strategies contribute to a comprehensive data protection posture.

Table 6

Data Protection Levels

System Behavior (Internal Validity) *How a system acts and responds under various conditions, including normal operation and during or after an intrusion or cyber attack*	**Threat Actor Access Method (External Validity)** *How a system is able to against wide set of threat actor tactics, techniques, procedures and access methods*	**Data Protection Metrics**
L1 - Conformity *Conformity - The security mechanism is designed according to a predefined set of security requirements and specifications.*	**<u>Access</u>** **Sun et al. (2011):** Their static analysis method for web applications likely conforms to specific security standards and coding practices. **Li et al. (2021):** The automated policy generation for microservices likely adheres to predefined security and access control policies. **OSSPolice (Kim & Lee, 2017):** The tool is likely built to conform to open-source license requirements and security vulnerability databases. **<u>Use</u>** **Password strength:** The evaluation of password strength using guesswork and probability thresholds (Ma et al., 2014) can be seen as conforming to predefined security requirements and specifications. Strong passwords are a fundamental requirement for many security standards and frameworks, and the research provides a method for assessing their strength in relation to these standards. **PRNG security:** Ensuring sufficient entropy in random number generation is crucial for cryptographic operations and security protocols. Evaluating PRNG security based on entropy (Ma et al., 2014) aligns with conformity because secure systems must adhere to specific randomness requirements to be considered compliant with security standards **<u>Disclosure</u>** Password strength metrics (Bonneau, 2012) ensure compliance with predefined security requirements. **<u>Destruction</u>** Anthoine et al. (2021) indirectly address conformity by developing Proof of Retrievability (PoR) protocols that likely adhere to specific security and privacy standards for data storage and retrieval.	Security Strength

Table 6 continued

L2 - Correctness *Correctness - The security mechanism, when implemented, behaves exactly as intended and is free from errors that could undermine their purpose.*	**<u>Access</u>** **Acar et al. (2016):** Measuring functional correctness of code relates directly to this level. **Wu et al. (2023):** Evaluating the correctness of execution path recovery in ChkUp aligns with this level. **Xiang et al. (2019):** Focus on accurate detection of access control misconfigurations related to correctness. **<u>Use</u>** **PRNG security:** While the focus on entropy relates to conformity, the observation that actual entropy collected is often lower than expected (Ma et al., 2014) points to a potential correctness issue. This suggests that the PRNG implementation may not be functioning as intended, leading to weaker randomness than required. **<u>Disclosure</u>** Classification accuracy of passwords/vaults (Chatterjee et al., 2015) verifies proper functioning. **<u>Modification</u>** ARTISAN's high accuracy (Yu et al., 2024) in threat detection suggests correct implementation and functioning as intended. However, the prevalence of solution-test incompatibilities (Xu et al., 2019) indicates potential correctness issues in some security mechanisms. **<u>Destruction</u>** Xu et al. (2016) focus on correctness by evaluating the accuracy of their CREDAL tool in identifying and addressing memory corruption vulnerabilities, which are crucial for ensuring software behaves as intended.	Security Mechanism Accuracy, Error Rate

Table 6 continued

L3 - Effectiveness *Effectiveness - The extent security measures protect data while maintaining operational efficiency under "normal" circumstances.*	**<u>Access</u>** **Maffei et al. (2015):** Evaluating the performance and scalability of Group ORAM relates to effectiveness in secure data storage. **Wiefling et al. (2021):** Balancing privacy enhancements with security in risk-based authentication addresses effectiveness. **Gavazzi et al. (2023):** Measuring MFA availability and RBA effectiveness directly relates to this level. **<u>Use</u>** **Detection Rate (Accuracy):** The use of accuracy as a metric for spam filters directly relates to effectiveness. A high accuracy rate (above 90%) indicates that the spam filter is effectively identifying and blocking spam emails under normal circumstances, contributing to the overall effectiveness of email security. **Disclosure** Low performance overheads (Gudka et al., 2015) demonstrate efficient security with minimal impact. **<u>Modification</u>** Fidelius (Eskandarian et al., 2018) and ASAP (Wagner et al., 2015) demonstrate effective data protection with acceptable performance overhead, highlighting the balance between security and efficiency. **<u>Destruction</u>** Huang et al. (2017) demonstrate the effectiveness of their FlashGuard solution by measuring its ability to recover from ransomware attacks with minimal impact on SSD performance (recovery time, latency, throughput).	Security Mechanism Performance, System Efficiency

Table 6 continued

L4 - Resistance *Resistance - The ability for a security mechanism to withstand an intrusion or cyber-attack.*	**<u>Access</u>** **Kim & Lee (2017):** Detecting vulnerable OSS versions in mobile apps contributes to resistance against exploits. **Wu et al. (2023):** Addressing firmware update vulnerabilities strengthens resistance against attacks exploiting those vulnerabilities. **Jia et al. (2021):** Mitigating "Codema" vulnerabilities in IoT devices enhances resistance to unauthorized access. **<u>Use</u>** **Network impact:** The research by Varadarajan et al. (2015) on placement vulnerabilities in cloud environments directly relates to resistance. Their findings demonstrate that co-location attacks can significantly degrade network performance (50% to 300% slowdown). This highlights the need for stronger resistance against such attacks to maintain network stability and prevent disruptions. **Security Strength (Password strength):** Evaluating password strength (Ma et al., 2014) contributes to resistance by ensuring that passwords are sufficiently strong to withstand guessing attacks. Strong passwords increase the difficulty for attackers to gain unauthorized access, thereby enhancing the system's resistance to breaches. **<u>Disclosure</u>** TCB size (Cerdeira et al., 2020) and CVSS scores (Cerdeira et al., 2020) assess vulnerability and attack surface. **<u>Modification</u>** The existence of cross-thread stack-smashing attacks bypassing CFI defenses (Xu et al., 2019) reveals weaknesses in resistance to certain attack types. However, tools like SPIDER (Machiry et al., 2020) and OSV-Hunter (Yang et al., 2018) contribute to improved resistance by identifying and mitigating vulnerabilities. **<u>Destruction</u>** Feng et al. (2016) contribute to resistance by developing a real-time crypto-ransomware detection approach, aiming to prevent data destruction before it occurs.	Threat Actor Success Rate, Cost, and Knowledge (Uncertainty)

Table 6 continued

L5 - Resiliency *Resiliency - The ability for a security mechanism to adapt and recover from adverse events that could compromise data.*	**<u>Access</u>** **Xiang et al. (2019):** Continuous validation and forensics contribute to resilience by enabling faster recovery and adaptation. **Wiefling et al. (2021):** Improving privacy in risk-based authentication systems enhances resilience against user re-identification and tracking. **Ladisa et al. (2022):** Providing a taxonomy of attacks and safeguards for open-source software supply chains supports resilience by enabling better anticipation and mitigation of threats. **<u>Use</u>** **Network impact:** While Varadarajan et al. (2015) primarily focus on resistance, the impact of co-location attacks on network performance also has implications for resilience. A system's ability to recover from the performance degradation caused by these attacks contributes to its overall resilience. This might involve implementing measures to mitigate the impact of co-residency or having mechanisms for quickly restoring network performance after an attack. **<u>Disclosure</u>** Remediation rate (Li et al., 2016) reflects the ability to recover from vulnerabilities. **<u>Modification</u>** Log reduction techniques (Inam et al., 2023) support resilience by enabling efficient storage and analysis of security logs, crucial for identifying and recovering from data breaches or tampering attempts. **<u>Destruction</u>** Huang et al. (2017) also address resilience by enabling recovery from ransomware attacks, allowing the system to adapt and bounce back from a data destruction event.	Recovery Time, Adaptability

This chapter has presented a framework for classifying the capabilities and behaviors of data protection systems based on five distinct levels: conformity, correctness, effectiveness, resistance, and resilience. These levels provide a common language for evaluating and comparing data protection strategies, enabling researchers and practitioners to assess progress and identify areas for improvement. By considering the key principles of system behavior, threat actor capabilities, and data protection metrics, organizations can develop a comprehensive and robust approach to safeguarding sensitive information. Future research can build upon this framework by further refining the levels, developing standardized assessment tools, and investigating the interplay between different levels in complex systems. Ultimately, the goal is to enhance data protection strategies and ensure the confidentiality, integrity, and availability of critical information in the face of evolving threats.

CHAPTER 7: CONTRIBUTIONS AND FUTURE DIRECTIONS

Contributions

This dissertation advances cybersecurity research by providing a structured approach to building and analyzing causal models, enabling a deeper understanding of the factors that influence data protection. It offers a comprehensive framework for constructing robust causal models, starting with a thorough understanding of the security domain and culminating in rigorous experiment analysis. This framework emphasizes the importance of considering both observable and hidden variables that affect data protection. Additionally, the dissertation establishes a broader definition of data protection, which encompasses unauthorized access, use, modification, and destruction of data, in addition to unauthorized disclosure. This expanded definition provides a more holistic view of data protection in the context of modern cybersecurity challenges. The research also develops a causal model using Causal Bayesian Networks (CBNs) to visually represent and quantify the complex interplay of factors contributing to data exposure. By identifying and analyzing various system and threat scenarios, the dissertation further highlights the dynamic nature of data protection and the need to consider both system-level factors and attacker capabilities. Finally, it provides a set of general experiments, including intervention and counterfactual studies, to empirically evaluate data protection and gain insights into causal relationships.

Created A Definition for Data Protection Through an Exhaustive Literature Review

Recognizing the limitations of the traditional definition of data protection, which primarily focuses on preventing unauthorized disclosure, this dissertation undertakes a comprehensive review of cybersecurity literature to formulate a more nuanced and encompassing

definition: data protection (Chapter 2). This new definition broadens the scope of data protection by including not only unauthorized disclosure but also unauthorized access, use, modification, and destruction of data. This expanded definition of data protection provides a more holistic understanding of security of confidential data within the context of modern cybersecurity challenges. It recognizes that data breaches can occur through various means beyond mere disclosure and emphasizes the importance of protecting data throughout its lifecycle. By incorporating these additional aspects, Data Protection definition offers a more robust and comprehensive framework for assessing and improving security measures.

Provided a Structured Approach to Building Causality Models in Cybersecurity

This dissertation provides a comprehensive guide for building robust causal models in cybersecurity research. Acknowledging the intricate nature of these systems, the framework leads researchers through a systematic process. It begins with a deep dive into the specific security domain, identifying key properties and potential causal factors through literature review and expert consultation. Next, researchers analyze the system's behavior concerning those security properties, mapping interactions and dependencies between components, particularly in relation to data breaches. This understanding is then formalized into a Structural Causal Model (SCM), where variables are selected, relationships defined, and a Directed Acyclic Graph (DAG) is constructed to visually represent the causal pathways. Data is overlaid onto the DAG to provide concrete measurements and assumptions about causal relationships are explicitly stated for transparency and analysis. The model then undergoes rigorous validation, focusing on causal relationships and discarding non-causal data. This involves testing the model by overlaying conditional distributions and employing causal search algorithms to pinpoint direct causal links. The validated model is then used for experiments and analysis, where researchers manipulate

variables to observe their effects, employ model adjustment techniques for accurate causal effect estimation, and conduct counterfactual studies to explore hypothetical scenarios. Finally, the transportability of the model is examined to determine if its insights can be generalized to other systems. This structured approach not only strengthens the rigor of cybersecurity research but also fosters transparency and reproducibility, ultimately leading to more reliable conclusions about cause-and-effect in this critical field.

Identified Observable and Hidden Variables for Data Protection

This dissertation delves into the critical factors influencing data protection by identifying and categorizing key variables, distinguishing between those that are directly measurable and those that remain hidden. Observable variables, encompassing aspects like the strength of passwords and encryption keys, the number of login attempts, the frequency of security incidents, adoption rates of security measures like two-factor authentication, and the performance of security tools, can be directly measured or observed through empirical data collection. However, the research also acknowledges the significant role of hidden variables, which are not directly measurable and often represent underlying or latent constructs. These include user behavior and security practices, such as adherence to policies and susceptibility to social engineering, as well as attacker motivations like financial gain or espionage. Furthermore, organizational factors like culture and security awareness, along with external factors like the regulatory environment and emerging threats, are also considered as hidden variables impacting data protection. By considering both observable and hidden variables, this research provides a more complete and nuanced understanding of the complex interplay of factors affecting data protection, emphasizing the importance of not only measuring what is readily apparent but also considering the often-unseen influences that can significantly impact data protection strategies.

Created a Causal Model for Data Protection

This research constructs a causal model for data protection using Causal Bayesian Networks (CBNs) to represent the complex interplay of factors leading to data exposure. This model provides a valuable tool for visualizing, analyzing, and predicting how various elements interact to affect a system's data protection. The model uses a Directed Acyclic Graph (DAG) to visually map the causal relationships between variables, offering a clear picture of how factors like security measures, threats, vulnerabilities, and user behavior influence data protection outcomes. Importantly, the model goes beyond simply identifying these relationships by incorporating specific measurements from existing research, allowing for the quantification of the impact of particular security solutions and the identification of hidden factors affecting their efficacy. Encompassing a wide range of data protection aspects, including authorized access, system use, information disclosure, data modification, and destruction, the model provides a holistic perspective on data exposure. Furthermore, its modular design allows for the analysis of individual components and their interactions within the larger system, enabling researchers to focus on specific aspects of data protection while understanding their interconnectedness.

Identified System and Threat Scenarios Impacting Data Protection

This dissertation recognizes that data protection is not a static concept but rather a dynamic interplay of various factors, and therefore identifies and analyzes a range of system and threat scenarios that can significantly influence the level of data protection. It explores scenarios with strong security implementations, characterized by robust policies, effective controls, and a well-defined security posture, demonstrating how these positively influence data protection by limiting unauthorized access and minimizing data exposure. Conversely, it examines scenarios with weak or inadequate security measures, analyzing how vulnerabilities, ineffective controls,

and a lack of security awareness can heighten the risk of breaches and data protection violations. Furthermore, the research delves into threat scenarios, considering situations where threat actors successfully execute attacks due to their capabilities, resources, and system knowledge, leading to data exposure. It also analyzes situations where threat actors fail to achieve their objectives due to robust defenses, effective detection mechanisms, or other hindering factors. By exploring these diverse scenarios, the research provides a comprehensive understanding of the dynamic nature of data protection, emphasizing the importance of considering both system-level factors and threat actor capabilities when assessing and improving data protection measures.

Provided A General Set of Experiments for Studying Data Protection

This dissertation outlines a set of general experiments designed to systematically study and evaluate data protection, providing a structured framework for empirical research that allows for hypothesis testing and evidence gathering to draw meaningful conclusions about the effectiveness of various data protection measures. The proposed experiments fall into two main categories: intervention studies and counterfactual studies. Intervention studies aim to assess the causal impact of specific interventions or security measures on data protection outcomes. This involves defining the intervention, such as implementing a new access control system or conducting security awareness training, and selecting appropriate outcome measures, such as the number of successful breaches or data exposure rate. The study design may involve control groups, randomization, or other techniques to isolate the effect of the intervention. Data is collected before and after the intervention to measure its impact and analyzed using appropriate statistical methods to determine the causal effect. Counterfactual studies, on the other hand, explore hypothetical scenarios to understand what would have happened under different conditions. This involves defining the counterfactual question, such as "What would the data

exposure rate have been if we had not implemented this security measure?", and selecting a suitable causal model that can accurately represent the relationships between variables and enable the estimation of counterfactual outcomes. This approach allows researchers to explore alternative scenarios and gain a deeper understanding of the factors influencing data protection.

Future Research Directions

This research endeavor embarks on the formulation of a comprehensive agenda dedicated to the application of causal inference within the realm of cybersecurity. The overarching ambition of this undertaking is to transcend the limitations of conventional correlational analysis, which merely identifies associations between events, and to delve into the intricate causal mechanisms that underpin security occurrences. By attaining a profound comprehension of cause-and-effect relationships, the research community can embark on the development of more robust and resilient data protection strategies, attain the capacity to anticipate the repercussions of security interventions, and architect adaptive security systems that exhibit the agility to respond effectively to the perpetually evolving landscape of cyber threats. This ambitious agenda encompasses a multifaceted approach, entailing the reevaluation of established security models, the formulation of novel metrics, the bridging of the chasm between theoretical constructs and real-world implementations, and the exploration of innovative tools and techniques.

Rethinking Security Models

Venerable security models, such as the Harrison-Ruzzo-Ullman (HRU) model, the Bell-LaPadula model, and the Biba model, have long served as invaluable frameworks for the management of access control and the regulation of information flow. However, these models frequently rely upon static rules and assumptions that may falter in the face of dynamic

environments characterized by incessant change and unpredictable adversarial behavior. By infusing these models with a causal inference perspective, we can acquire a more nuanced and comprehensive understanding of the intricate interplay among the constituent elements of these models, thereby empowering us to discern potential vulnerabilities that may elude conventional analysis. This endeavor necessitates a tripartite approach:

- **Systematic Comparison:** A rigorous and methodical comparative analysis of traditional security models must be conducted through the discerning lens of causal inference, with the objective of unraveling the intricate causal relationships that govern the interactions between entities and actions within these models.
- **Framework Development:** The establishment of a coherent and universally applicable framework is indispensable for the systematic comparison of diverse security models. This framework should accentuate the strengths and limitations inherent in each model, elucidate their respective domains of applicability, and provide guidance on the selection of the most suitable model for addressing specific security challenges.
- **Causal Model Integration:** A critical inquiry lies in determining the precise junctures where causal models can augment, supplant, or bestow unique value upon traditional security models. This may involve harnessing the power of causal models to conduct intervention studies, which assess the efficacy of security measures, and counterfactual studies, which explore hypothetical scenarios and alternative outcomes. Such investigations could encompass the modeling of policy modifications, the simulation of diverse attack strategies, and the evaluation of the resilience of security architectures to unforeseen disruptions.

Data Protection Metrics for Causal Inference

To fully harness the analytical prowess of causal inference in the cybersecurity domain, we must equip ourselves with robust and informative metrics that faithfully capture the dynamic and ever-shifting nature of cyber threats and vulnerabilities. This pursuit entails a multi-pronged strategy:

- **Refining Security Goals:** A prerequisite for effective security management is the meticulous articulation of security objectives in the explicit language of causal relationships. This necessitates moving beyond simplistic metrics, such as the mere enumeration of successful attacks, and embracing more sophisticated measures that reflect the intricate causal pathways through which security compromises occur. For instance, a refined security goal might be to curtail the causal influence of phishing attacks on the occurrence of data breaches, thereby addressing the root cause of such incidents.
- **Validating Assumptions:** The foundations of security designs often rest upon a bedrock of assumptions, and the validity of these assumptions is paramount to the efficacy of the resulting security measures. Consequently, we must devise rigorous methodologies to systematically validate these assumptions, ensuring their alignment with the realities of the operational environment. This could involve quantifying the uncertainty associated with each assumption and judiciously allocating resources for validation endeavors based on their potential impact on security posture.
- **Dynamic Threat Analysis:** The landscape of cyber threats is in a perpetual state of flux, with new attack vectors and vulnerabilities emerging relentlessly. To maintain a robust security stance, we must integrate dynamic threat modeling into our metric framework,

thereby accounting for the evolutionary nature of cyberattacks. This could involve leveraging causal models to prognosticate the ramifications of nascent threats on existing defenses, enabling proactive adaptation and mitigation strategies.

- **Comprehensive Metrics:** A holistic understanding of cybersecurity necessitates the development of a comprehensive suite of metrics that encompasses a diverse array of factors, including:
    - **System observations:** These encompass readily observable indicators of system behavior, such as network traffic patterns, user activity logs, and resource consumption metrics.
    - **Hidden variables:** These represent latent factors that are not directly observable but exert a significant influence on security outcomes, such as the intent, capabilities, and resources of malicious actors.
    - **Security strength and fragility:** These metrics gauge the resilience of systems to specific attack types, providing insights into the robustness of defenses and the potential points of failure.
    - **Asymptotic upper and lower data protection limits:** These theoretical bounds, supported by rigorous mathematical theorems and proofs, delineate the ultimate limits of data protection achievable within a given system or environment.

Bridging Theoretical and Empirical Limits

A formidable challenge that confronts the cybersecurity community is the bridging of the chasm between abstract theoretical models and the concrete realities of real-world implementations. Causal inference offers a potent instrument for spanning this divide by:

- **Understanding Limitations:** A prerequisite for effective security engineering is a lucid understanding of the limitations inherent in data protection mechanisms. This entails a critical examination of these limitations from both theoretical and empirical vantage points. Theoretical limitations may stem from computational complexity constraints, information-theoretic bounds, or the inherent limitations of cryptographic primitives. Empirical limitations may arise from software vulnerabilities, human fallibility, or the intricacies of sociotechnical systems.
- **Explaining Security Failures:** Causal inference techniques empower us to transcend the superficiality of mere correlations and delve into the underlying causal factors that contribute to security breaches. By meticulously analyzing attack patterns, we can unearth the root causes of security failures, pinpoint vulnerabilities that were exploited, and glean insights that inform the design of more resilient systems.
- **Predicting Impact:** Causal models furnish us with the capacity to anticipate the consequences of security measures before their deployment, enabling more informed decision-making and resource allocation. By simulating the effects of various security interventions, we can evaluate their efficacy, identify potential unintended consequences, and optimize their implementation.
- **Measuring Resilience:** The ultimate measure of a security system lies in its resilience—its ability to withstand attacks, recover from disruptions, and maintain essential functionality. To assess this critical attribute, we must develop quantitative metrics that gauge vulnerability exposure, implementation effectiveness, and the extent of the gap between theoretical guarantees and real-world performance.

Tools, Techniques, and Methods

The effective application of causal inference within the cybersecurity domain necessitates the development and deployment of specialized tools and techniques tailored to the unique challenges posed by this field. This pursuit encompasses:

- **Tailored Methods:** The exploration and development of causal inference methods specifically designed to address the intricacies of cybersecurity are of paramount importance. These methods must accommodate the temporal dependencies inherent in cyberattacks, the presence of hidden variables that obscure causal relationships, and the dynamic behavior of systems under attack.
- **Automated Discovery:** The sheer volume and complexity of cybersecurity data often preclude manual analysis, underscoring the need for automated causal discovery techniques. These techniques should enable the extraction of causal insights from large-scale observational data across diverse cybersecurity domains, such as password studies, intrusion detection systems, and malware analysis platforms.
- **Addressing Unique Dimensions:** Cybersecurity presents unique challenges to causal inference that necessitate the development of specialized techniques. These challenges include the pervasive influence of human behavior on security outcomes, the presence of adversarial actors who actively seek to conceal their actions and deceive defenders, and the intricate interactions between software, hardware, and network components that constitute complex sociotechnical systems.

Intervention and Counterfactual Studies

Empirical evidence serves as the bedrock upon which sound security practices are built. To amass such evidence, we must engage in rigorous intervention and counterfactual studies:

- **Intervention Studies:** Controlled experiments, meticulously designed and executed, are essential for evaluating the efficacy of data protection techniques under a variety of conditions. These conditions should encompass diverse systems and architectures, various attack scenarios ranging from remote intrusions to data deletion, and different types of threat actors, including nation-states, criminal syndicates, and individual hackers.
- **Counterfactual Studies:** Counterfactual analysis enables us to explore hypothetical scenarios and assess the impact of interventions that were not implemented. This can involve evaluating the effectiveness of user training programs, comparing alternative security configurations, and identifying at-risk populations that may be particularly susceptible to cyberattacks. Such studies can inform resource allocation decisions, prioritize security investments, and guide the development of targeted interventions.

Generalized Autonomous and Adaptable Systems (GASS)

Causal inference holds the key to unlocking the potential of intelligent, adaptive security systems that can autonomously respond to evolving threats and dynamically adjust their defenses. This ambitious endeavor involves:

- **Challenges of GASS:** The development of generalized, autonomous, and adaptable security agents for data protection is fraught with challenges, including scalability concerns, explainability requirements, and ethical considerations. A thorough

examination of these challenges is necessary to ensure the responsible and beneficial deployment of such systems.

- **Causal Inference Engine:** At the heart of GASS is a causal inference engine, responsible for reasoning about cause-and-effect relationships, learning from data, and making informed decisions based on causal insights. The design and implementation of such an engine require the development of sophisticated algorithms for causal reasoning, probabilistic modeling, and decision-making under uncertainty.
- **Performance Enhancement:** The performance of GASS hinges on a multitude of factors, including the speed and accuracy of threat assessment, the effectiveness of automated response mechanisms, and the agility of adaptation to changing circumstances. Identifying key areas for performance enhancement is crucial for realizing the full potential of GASS and ensuring their operational effectiveness.
- **Implementation and Formal Models:** The realization of GASS necessitates the exploration of diverse implementation methodologies and the utilization of formal models, such as finite-state machines, to guide their development. This may involve defining the architecture of GASS, specifying communication protocols between agents, and delineating the decision-making processes that govern their behavior.

Meta-Analysis and Replication

To cultivate a robust body of knowledge in the nascent field of causal inference for cybersecurity, we must embrace the principles of meta-analysis and replication:

- **Meta-Analysis:** Meta-analysis techniques enable the synthesis of findings from multiple independent studies, facilitating the identification of consistent patterns, the assessment

of the overall effectiveness of different security interventions, and the generation of generalizable insights.

- **Replication Research:** Replication studies serve as a cornerstone of scientific rigor, ensuring the reproducibility and reliability of research findings. Encouraging and prioritizing replication efforts in the cybersecurity domain will bolster the credibility of causal claims and foster confidence in the efficacy of security solutions derived from causal inference.

By steadfastly pursuing this comprehensive agenda, we can harness the transformative power of causal inference to elevate data protection and cybersecurity to unprecedented heights. This endeavor will necessitate close collaboration among researchers, practitioners, and policymakers, fostering a synergistic ecosystem dedicated to the development and deployment of effective, adaptive security solutions that address the evolving threat landscape.

REFERENCES

Abelson, H., Ledeen, K., & Lewis, H. (2015, July 7). Keys under doormats: Mandating insecurity by requiring government access to all data and communications. Lawfare.

Acar, Y., Backes, J., Fahl, S., Kim, D., & van Deursen, A. (2016). You get where you're looking for: The impact of information sources on code security. Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 291-302.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. Science, 347(6221), 509-514.

Al-Okaily, M., Kallail, K. M., & Al-Dmour, R. H. (2019). The impact of data breach on customers' trust and loyalty. International Journal of Information Management, 49, 135-149.

Alhabash, S., & Ma, M. (2017). A tale of four platforms: Motivations and uses of Facebook, Twitter, Instagram, and Snapchat among college students? Social Media + Society, 3(1), 2056305117691544.

American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022).

Anderson, J. (2017). Roman law and the origins of privacy. Oxford University Press.

Anderson, R. (2008). Security engineering: A guide to building dependable distributed systems (2nd ed.). Wiley.

Annas, G. J., & Grodin, M. A. (Eds.). (1992). The Nazi doctors and the Nuremberg Code: Human rights in human experimentation. Oxford University Press.

Backes, M., Chen, T., Dürmuth, M., & Pierre, D. (2009). Compromising reflections—or—how to read LCD monitors around the corner. In 2009 IEEE Symposium on Security and Privacy (pp. 158-172). IEEE.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27), 7345-7352.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173–1182.

Beer, D., & Burrows, R. (2013). Popular culture, digital archives and the new personal. International Journal of Cultural Studies, 16(4), 425-443.

Bell, D. E., & LaPadula, L. J. (1973). Secure computer systems: Mathematical foundations. MITRE Corporation, MTR-2547.

Bellovin, Steve (personal communication, 2021).

Bennett, C. J. (1992). Regulating privacy: Data protection and public policy in Europe and the United States. Cornell University Press

Bishop, M. (2003). Computer security: Art and science. Addison-Wesley.

Bok, S. (1982). Secrets: On the ethics of concealment and revelation. Pantheon Books.

Boucher, P., & Anderson, R. (n.d.). Trojan Source: Invisible Vulnerabilities.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210-230.

Brenner, S. W. (2007). Cybercrime: Criminal threat or business risk? Information Systems Security, 16(6), 339-344.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.

California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100-1798.199

Camurati, G., Poeplau, S., Muench, M., Hayes, T., Francillon, A., & Bruneau, W. (2018). Screaming channels: When electromagnetic side channels meet radio transceivers. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 163-177).

Carpenter v. United States, 138 S. Ct. 2206 (2018).

Case C-131/12, Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González , ECLI:EU:C:2014:317 (May 13, 2014).

Cerf, V., & Kahn, R. (1974). A protocol for packet network intercommunication. IEEE Transactions on Communications,22(5), 637-648.

Chen, S., Wang, Z., & Zhang, X. (n.d.). Peeking into your encrypted web traffic: A comprehensive study of side-channel information leaks in web applications.

Cheswick, W. R., & Bellovin, S. M. (1994). Firewalls and Internet security: Repelling the wily hacker. Addison-Wesley Professional.

Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501-6506.

Communications Assistance for Law Enforcement Act of 1994, Pub. L. No. 103-414, 108 Stat. 4279.

CRA. (2003). Grand challenges in trustworthy computing research. Computing Research Association

Davis, A. (2023). The Internet of Things and privacy: Challenges and solutions. Journal of Privacy and Confidentiality, 5(2), 123-145.

Davis, K. (2021). The Hippocratic oath: A historical and ethical perspective. Routledge.

Denning, D. E. (1999). Information warfare and security. Addison-Wesley Professional.

Denyer, S. (2016, December 12). China's all-seeing surveillance state is reading your posts, tracking your movements. The Washington Post.

Department of Defense. (1985). Trusted Computer System Evaluation Criteria (DOD 5200.28-STD).

Deutsch, D., & Lockwood, M. (1994). The quantum physics of time travel. Scientific American, 270(3), 68-74.

Dhillon, G., & Blackhouse, J. (2001). Current directions in IS security research: Towards socio-organizational perspectives. Information Systems Journal, 11(2), 127-153.

Diffie, W., & Hellman, M. (1976). New directions in cryptography. IEEE Transactions on Information Theory, 22(6),644-654

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

DoD (Department of Defense). (1985). Trusted Computer System Evaluation Criteria (DOD 5200.28-STD).

Duan, R., Kim, D., & Lee, W. (2017). OSSPolice: Detecting license violations and vulnerable OSS components in Android apps. 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE), 622–633.

Dwork, C. (2006). Differential privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (pp. 1-12). Springer.

Dwyer, C., Hiltz, S. R., & Passerini, K. (2015). Trust and privacy concern within social networking sites: A comparison of Facebook and LinkedIn. AMCIS 2007 Proceedings, 351.

Edgar, S. L. (2002). Morality and Machines: Perspectives on Computer Ethics (2nd ed.). Jones & Bartlett Learning.

Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848

ENISA (European Union Agency for Cybersecurity). (2023). ENISA Threat Landscape 2023. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

Etzioni, A. (2004). The limits of privacy. Basic Books.

European Commission. (2018). General Data Protection Regulation. https://gdpr-info.eu/

Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g

Felt, A. P., Finifter, M., Chin, E., Hanna, S., & Wagner, D. (2011, October). A survey of mobile malware in the wild. In Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices (pp. 3-14).

Foreign Intelligence Surveillance Act of 1978 Amendments Act of 2008, Pub. L. No. 110-261, 122 Stat. 2436

FTC (Federal Trade Commission). (2023). Data Breach Response: A Guide for Business.

Furnell, S. (2002). Cybercrime: Vandalizing the information society. Addison-Wesley.

Garcia, M. (2015). Confession and confidentiality in ancient religions. Brill.

Gavazzi, D., Arnaboldi, M., Cozza, A., Iacono, G. L., & Sartiani, C. (2023). The devil is in the (third-party) details: Measuring the adoption and implementation of multi-factor and risk-based authentication schemes in 5,000 websites.2023 IEEE Symposium on Security and Privacy (SP), 1536–1554.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of the 41st Annual ACM Symposium on Theory of Computing (pp. 169-178). ACM.

Gibson, D. (2015). Managing risk in information systems. Jones & Bartlett Learning.

Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gordon, L. A., & Loeb, M. P. (2002). The economics of information security investment. ACM Transactions on Information and System Security (TISSEC), 5(4), 438-457.

Gramm-Leach-Bliley Act, Pub. L. 106-102 (1999)

Greenberg, A. (2016, April 19). The FBI's war on encryption: It's not going well. Wired.

Hadnagy, C. (2018). Social engineering: The science of human hacking (2nd ed.). John Wiley & Sons.

Hale, J., Zhang, Z., & Greene, E. (2016). Measuring cybersecurity behaviors: A review and research agenda. Computers in Human Behavior, 65, 318-329.

He, D., Chan, S., Guizani, M., & Shi, W. (2021). Access control for the internet of things: A survey. IEEE Communications Surveys & Tutorials, 23(2), 1135-1160.

Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. Journal of Human Resources, 32(3), 441-462.

Holt, T. J., & Bossler, A. M. (2009). Examining the applicability of lifestyle-routine activity theory for cybercrime victimization. Deviant Behavior, 30(1), 1-25.

Hubbard, D. W., & Seiersen, R. (2016). How to measure anything in cybersecurity risk. Wiley.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. Psychological Methods, 15(4), 309–334.

Imai, K., Keele, L., & Yamamoto, T. (2011). Sensitivity analysis for causal mediation effects.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2), 615-635.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

ISO/IEC 27001. (2013). Information technology — Security techniques — Information security management systems — Requirements. International Organization for Standardization.

Jaquith, A. (2007). Security metrics: Replacing fear, uncertainty, and doubt. Addison-Wesley.

JASON. (2008). Science of Cyber-Security (JSR-08-506). The MITRE Corporation.

Jia, Z., Mao, Z. M., Wang, C., Wu, H., & Song, D. (2021). Codema: Disjointed in-device code updates enable cross-application code reuse attacks on IoT devices. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(4), 1-26.

Johnson, P. (2019). Secrecy and confidentiality in ancient civilizations. Cambridge University Press.

Johnson, P., & Brown, S. (2022). The impact of artificial intelligence on society: Opportunities and challenges. Future of Humanity Institute, 3(1), 56-78.

Jones, A., & Brown, B. (2020). The Hippocratic oath in modern medicine. Springer.

Kahn, D. (1967). The codebreakers: The story of secret writing. Macmillan.

Katz v. United States, 389 U.S. 347 (1967).

Katz, J., & Lindell, Y. (2020). Introduction to modern cryptography (3rd ed.). Chapman and Hall/CRC.

Kerckhoffs, A. (1883). La cryptographie militaire. Journal des sciences militaires, IX, 5-38, 161-191.

Kerr, O. S. (2009). The Fourth Amendment and new technologies: Constitutional myths and the case for caution. Michigan Law Review, 102(5), 801-888.

Kim, D., & Lee, W. (2017). Detecting license violations and vulnerable code in mobile apps. IEEE Transactions on Software Engineering, 44(10), 963-981.

Kocher, P., Jaffe, J., & Jun, B. (1999). Differential power analysis. In M. Wiener (Ed.), Advances in Cryptology - CRYPTO '99 (pp. 388–397). Springer.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Krawczyk, H. (2005). Cryptographic protocols, and how to analyze them: The computational model. In R. Cramer (Ed.), Advances in Cryptology - EUROCRYPT 2005 (pp. 35-52). Springer.

Kumar, R. (2020). Deepfakes: A threat to privacy and democracy. International Journal of Cyber Criminology, 14(2), 234-256.

Ladisa, A., Chaloupka, J., Rivera, J. D., & Scanniello, G. (2022). Unveiling the open-source software supply chain attacks surface: A comprehensive survey. 2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR), 378–389.

Lee, J., Kim, Y., & Park, H. (2019). Anonymization techniques for privacy-preserving data mining: A survey. Expert Systems with Applications, 117, 111-137.

Lei Geral de Proteção de Dados [General Data Protection Law], Law No. 13.709 (Aug. 14, 2018).

Li, Z., Zhang, W., Ma, D., Ma, C., & Zhang, L. (2021). APG: Automated policy generation for microservices. IEEE Transactions on Services Computing, 14(4), 1272-1285.

Lyon, D. (2001). Surveillance society: Monitoring everyday life. Open University Press.

Lyon, D. (2007). Surveillance studies: An overview. Polity.

Mackie, J. L. (1965). Causes and conditions. American Philosophical Quarterly, 2(4), 245-264.

Macnish, K. (2017). Privacy, surveillance and social media: A role for virtue ethics?. Science and Engineering Ethics, 23, 1013-1030.

Maffei, M., Pecoraro, G., Rønne, M., & Scafuro, A. (2015). ORAM: A practical solution for secure and private cloud storage. ACM Transactions on Storage (TOS), 11(3), 1-26.

Mansfield-Devine, S. (2011). Advanced persistent threats and how to monitor and deter them. Network Security, 2011(7),13-16.

Martinez, L. (2018). The ethics of deepfakes: Protecting individual rights in the age of synthetic media. Journal of Media Ethics, 33(4), 234-248.

Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

McLean, J. (1987). Reasoning about security models. In Proceedings of the 1987 IEEE Symposium on Security and Privacy (pp. 123-131). IEEE.

Miller, L. (2022). The code of Hammurabi: A new translation. Yale University Press.

Mitnick, K. D., & Simon, W. L. (2002). The art of deception: Controlling the human element of security. Wiley.

MITRE. (2023). ATT&CK® .

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Nassi, B., Wool, A., & Elovici, Y. (2023). Glow worm attack: Electromagnetic eavesdropping via power indicator LEDs. 2023 IEEE Symposium on Security and Privacy (SP), 1515–1535.

New York State Technology Law § 208

Newman, L., Meyers, J., & Torres-Arias, S. (2022). Sigstore: Software signing for everyone. 2022 IEEE Symposium on Security and Privacy (SP), 1292–1309.

Nguyen, T. D., Acar, G., & De Cristofaro, E. (2022). An empirical assessment of gdpr compliance in android apps' consent notices. arXiv preprint arXiv:2206.09638.

NIST (National Institute of Standards and Technology). (2012). Guide for conducting risk assessments. Special Publication 800-30 Revision 1.

NIST (National Institute of Standards and Technology). (2014). Framework for Improving Critical Infrastructure Cybersecurity. https://www.nist.gov/cyberframework

NIST. (2012). Guide for conducting risk assessments. Special Publication 800-30 Revision 1.

NIST. (2013). Framework for improving critical infrastructure cybersecurity.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher et al. (Eds.), Essays in honor of Carl G. Hempel (pp. 114-146). Springer.

Olmstead v. United States, 277 U.S. 438 (1928).

Organick, E. I. (1972). The Multics system: An examination of its structure. MIT Press.

Parker, D. B. (1998). Fighting computer crime: A new framework for protecting information. John Wiley & Sons.

Parker, D. B. (2002). Fighting computer crime: A new framework for protecting information. John Wiley & Sons.

Pearl, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (pp. 411-420). Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

Pearl, J. (2018). The book of why: The new science of cause and effect. Basic Books

Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books

Peisert, S., & Bishop, M. (2007). How to design computer security experiments. In Proceedings of the 1st USENIX Workshop on Hot Topics in Security (HotSec '07). USENIX Association.

Peltier, T. R. (2016). Information security risk analysis (2nd ed.). CRC Press.

Personal Data Protection Bill, 2019, Bill No. 373 of 2019 (India).

Pfleeger, C. P., Pfleeger, S. L., & Margulies, J. (2018). Security in computing (6th ed.). Pearson.

Pfleeger, S. L. (2001). Software metrics: A rigorous and practical approach (2nd ed.). Prentice Hall PTR.

Ponemon Institute. (2022). Cost of a Data Breach Report 2022. https://www.ibm.com/security/data-breach

Privacy International. (2018). A chilling effect: NSA mass surveillance impacts on journalists and lawyers.

Prosser, W. L. (1960). Privacy. California Law Review, 48(3), 383-423.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Richards, N. M., & Hartzog, W. (2020). The Pathologies of Digital Surveillance. Harvard Law Review, 133(5), 1272-1342.

Riley v. California, 573 U.S. 373 (2014).

Romanosky, S. (2010). Examining the costs and causes of cybercrime. Journal of Urban Technology, 17(3), 91-107.

Rosen, J. (2017). The unwanted gaze: The destruction of privacy in America. Random House.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Modern epidemiology (3rd ed.). Lippincott Williams & Wilkins.

Saltzer, J. H., & Schroeder, M. D. (1975). The protection of information in computer systems. Proceedings of the IEEE, 63(9), 1278-1308.

Sanders, W. H. (2005). Computer security: A hands-on approach. Addison-Wesley.

Schechter, S. E., Dhamija, R., Ozment, A., & Fischer, I. (2008). The emperor's new security indicators. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (pp. 51-65). IEEE.

Schneider, F. B. (2000). Enforceable security policies. ACM Transactions on Information and System Security (TISSEC), 3(1), 30-50.

Schneier, B. (1996). Applied cryptography: Protocols, algorithms, and source code in C (2nd ed.). Wiley.

Schoeman, F. D. (1992). Privacy and social freedom. Cambridge University Press.

Schultz, E. (2003). Security engineering: A guide to building dependable distributed systems (2nd ed.). John Wiley & Sons.

Shetty, S., Adepu, S., & Vijayalakshmi Pai, G. A. (2020). A comprehensive survey on risk assessment and risk management in cybersecurity. Journal of Ambient Intelligence and Humanized Computing, 13(3), 1279–1303.

Shostack, A., & Maxion, R. A. (2006). The new school of information security. Addison-Wesley Professional.

Singh, S. (1999). The code book: The science of secrecy from ancient Egypt to quantum cryptography. Doubleday.

Smith, J. (2023). The data revolution: Challenges and opportunities for privacy protection. Data & Society Research Institute.

Smith, R. (2018). The ethics of confidentiality. Wiley.

Solove, D. J. (2004). The digital person: Technology and privacy in the information age. NYU Press.

Spafford, E. H. (2012). Cybersecurity: We need metrics! In 2012 Cybersecurity Summit (pp. 1-2). IEEE.

Stallings, W. (2017). Cryptography and network security: Principles and practice (7th ed.). Pearson.

Statista. (2023). Number of data breaches and exposed records in the United States from 2005 to 2022. https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/

Stolfo, S. J., Bellovin, S. M., Keromytis, A. D., Hershkop, S., Guha, A., Wang, K., ... & Sinclair, S. (2005). A quantitative framework for measuring security. IEEE Security & Privacy, 3(6), 20-29.

Stoycheff, E. (2016). Under surveillance: Examining Facebook's spiral of silence effects in the wake of NSA Internet monitoring. Journalism & Mass Communication Quarterly, 93(2), 296-311.

Sultan, N. (2010). Cloud computing for education: A new dawn? International Journal of Information Management, 30(2), 109-116.

Sun, Q., Tan, L., & Xie, T. (2011). Static detection of vulnerabilities due to improper handling of exceptional conditions in C++ programs. Proceedings of the 33rd International Conference on Software Engineering, 802-811.

Sun, Y., Song, H., Jara, A. J., & Bie, R. (2019). Internet of things and big data analytics for smart and connected communities. IEEE Access, 7, 175143-175153.

Taddicken, M. (2014). The 'privacy paradox' in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure. Journal of Computer-Mediated Communication, 19(2), 248-273.

Tankard, C. (2011). The advanced persistent threat. Network Security, 2011(9), 16-18.

Thompson, K. (2022). Cybersecurity in the Internet of Things: Protecting sensitive data from breaches. IEEE Security & Privacy, 20(3), 45-53.

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. First Monday, 19(7).

Turow, J. (2005). The daily you: How the new advertising industry is defining your identity and your worth. Yale University Press.

U.S. Const. amend. I.

U.S. Const. amend. IV.

Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT ACT) Act of 2001, Pub. L. No. 107-56, 115 Stat. 272

Vacca, J. R. (2005). Computer and information security handbook. Morgan Kaufmann.

VanderWeele, T. J. (2015). Explanation in causal inference: Methods for mediation and interaction. Oxford University Press.

Verendel, V. (2009). Quantitative security metrics for software. University of Twente.

Verizon. (2023). 2023 Data Breach Investigations Report.

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer.

vos Savant, M. (1990). Ask Marilyn. Parade Magazine, September 9, 1990.

Vuagnoux, M., & Pasini, S. (n.d.). Compromising electromagnetic emanations of wired and wireless keyboards.

Ware, W. H. (Ed.). (1970). Security controls for computer systems: Report of Defense Science Board Task Force on Computer Security (RAND Report R-609). RAND Corporation.

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard Law Review, 4(5), 193-220.

Whitman, M. E., & Mattord, H. J. (2011). Principles of information security (4th ed.). Cengage Learning.

Whitten, A., & Tygar, J. D. (1999). Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In Proceedings of the 8th USENIX Security Symposium (pp. 169-184). USENIX Association.

Wiefling, B., Tolsdorf, K., & Lo Iacono, G. (2021). Evaluating privacy-enhancing technologies in risk-based authentication systems. Proceedings on Privacy Enhancing Technologies, 2021(4), 599–620.

Williams, M. B. (2010). The human right to privacy. Routledge.

Wood, A. W. (1999). Kant's ethical thought. Cambridge University Press.

Wu, H., Chen, H., Chen, Z., Xing, X., Mao, Z. M., & Lee, W. (2023). ChkUp: Detecting and validating firmware update vulnerabilities with formal modeling and hybrid fuzzing. 2023 IEEE Symposium on Security and Privacy (SP), 1645–1663.

Xiang, X., Tan, L., & Zhou, Y. (2019). Rethinking access control: From static detection of vulnerabilities to continuous validation and forensics. IEEE Transactions on Dependable and Secure Computing, 16(4), 661-674

Yaroslavskaya Oblast Court Case, Application no. 41438/10 (Eur. Ct. H.R. Sept. 10, 2015).

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. IEEE Internet of Things Journal, 1(1), 22-32.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.