1

Introduction

Data Analytics for Cybersecurity

1.1 What Is Cybersecurity?

Cybersecurity refers to securing valuable electronic assets and physical assets, which have electronic access, against unauthorized access. These assets may include personal devices, networked devices, information assets, and infrastructural assets, among others.

Cybersecurity deals with security against threats also referred to as cyber threats or cyberattacks. Cyberattacks are the mechanism by which security is breached to gain access to assets of value.

The first aim of cybersecurity is prevention of cyberattacks against critical assets. The second aim of cybersecurity is detection of threats. The third aim is to respond to threats in the event that they penetrate access to critical assets, and, finally, the fourth aim is to recover and restore the normal state of the system in the event that an attack is successful. Cybersecurity is achieved by addressing each of these three aspects to prevent, detect, and respond to threats against critical assets. Essentially, it deals with securing everything that is in cyberspace so these assets of value are not tampered with.

What really are these assets? Are they the data on a hard drive, a contact list on a cell phone, an Excel sheet with sales numbers, or a program to switch on a device?

It is all of these and more. With the ever-increasing use of electronics for every possible function of life, cybersecurity is becoming a pervasive problem permeating every individual's life. If you look around yourself, there are several electronic devices that you may see or use, and each electronic device may in turn be connected to another device physically or virtually. This connectivity enables us to access additional functionalities on the device.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

However, this also makes the device vulnerable since it has external connectivity and access is fragmented through multiple external applications.

Now the obvious question is, does cybersecurity apply to connected devices only? Although a big part of cybersecurity is a result of the high level of connectivity, it also includes threats resulting from compromised physical security. For instance, a highly secure desktop in a secure room, which is not accessible via the traditional internet but only through a biometric authentication or a card swipe, is also at risk. An unauthorized access to the secure room poses a cybersecurity threat due to the electronic assets at risk and potential risk to the systems providing access to the room. This is primarily because some form of connectivity is always possible in this highly connected world. Another example is that of an insider threat where an authorized user accesses materials with a malicious intent.

1.1.1 Assets Affected

Cybersecurity is a major challenge due to the potential of damage. An empty hard drive that is stolen is a theft, but a hard drive with data accessed in an unauthorized manner poses a much bigger threat due to the value of the information stored on it. A sensor controlling a chemical flow into a vat that breaks down accidentally may lead to a hazmat accident, but an authorized access leading to tampering of the program controlling the sensor is a cybersecurity risk. This is because the intent and extent of the tampering are uncontrolled and may be much more catastrophic than a sensor breakdown. Thus, cybersecurity aims to prevent unauthorized access to electronic assets and physical assets with electronic access.

Let us look at some types of assets – personal, public, and corporate – that can be impacted by cybersecurity risks, summarized in Figure 1.1. We are now referring to the electronic access to physical assets or electronic assets. This list is a small sample of such assets in the highly connected world that we live in. These include the following.

- **Personal assets (mostly used for personal needs):** Phones (home and mobile), tablets, personal computers (desktop and laptops), external physical hard drive, cloud drive, email accounts, fitness trackers, smart watches, smart glasses, media devices (TiVo, Apple TV, cable box), bank accounts, credit cards, personal gaming systems, blogs, vlogs, photos, and videos.
- **Public assets (for managing public utilities and services):** Smart meters, power grid, sewage controls, nuclear power plant, rail lines, air traffic, traffic lights, citizen databases, websites (county, state and federal), space travel programs.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

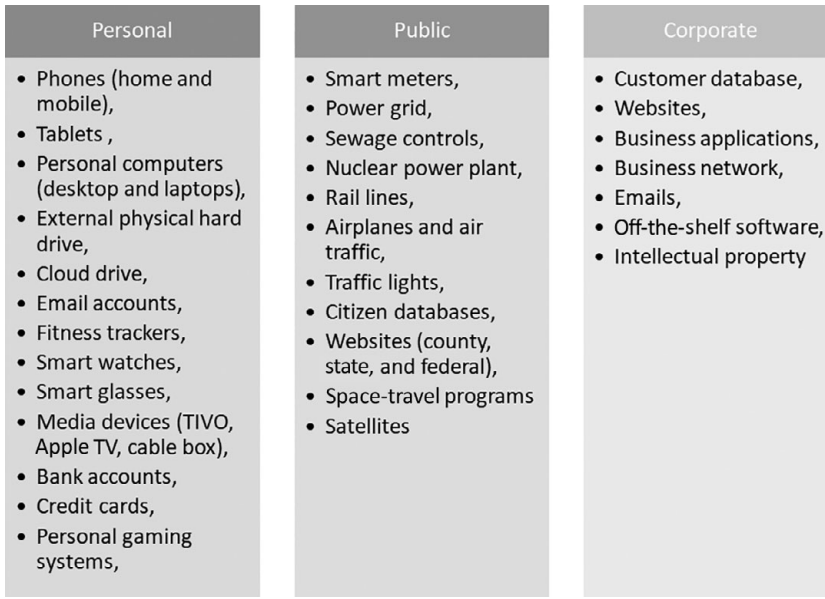

Figure 1.1 Sample assets at risk due to cyberattacks.

- **Corporate assets (for managing business needs):** Customer database, websites, business applications, business network, emails, software, intellectual property.

Some overlap may occur as corporate assets may link to public assets or personal assets may be connected to corporate assets or public assets. As such, these assets should be clearly separated through well-defined security policies. However, with the level of connectivity through multiple devices, it is becoming more and more difficult to keep assets completely discrete and disconnected.

Thus, cybersecurity is securing everything of value that exists in cyberspace, which includes computer networks, mobile devices, cloud data storage, sensor networks, industrial control systems, emergency devices, railway lines, and air traffic controls, to name a few. It is not simply limited to our own private data or banking information but goes beyond to critical infrastructure and corporate assets.

Certain geological regions are more prone to attacks than others due to the availability and richness of such assets. Many tools are available to depict attack maps of popular cyberattacks. Examples include ThreatCloud showing real-time attacks based on monitoring sensors across the world (Check Point

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

2020), Digital Attack Map depicting top distributed denial of service (DDoS) attacks worldwide (Netscout 2020), and Kaspersky Lab's 3D interactive map of cyberattacks (Kaspersky 2020). These are helpful in visualizing the spread of types of attacks and level of activity in a region.

1.1.2 Motivation, Risks, and Attaining Security

What is the motivation behind such attacks? An unauthorized access to secured resources may be due to various factors. Some such factors are summarized in Figure 1.2 and listed as follows:

- Stealing intellectual property: The majority of corporate product development strategies are stored on highly secure infrastructure. Small companies and large corporations alike devise novel products, and their intellectual property holds the key to revenue in the future. Loss of such information could be devastating for the sales and profits for an organization.
- Gaining access to customer data: Banks and retailers are constantly at the receiving end of cyberattacks due to valuable customer data, which can then be sold. Recent notable examples include Target (Washington Post 2013).

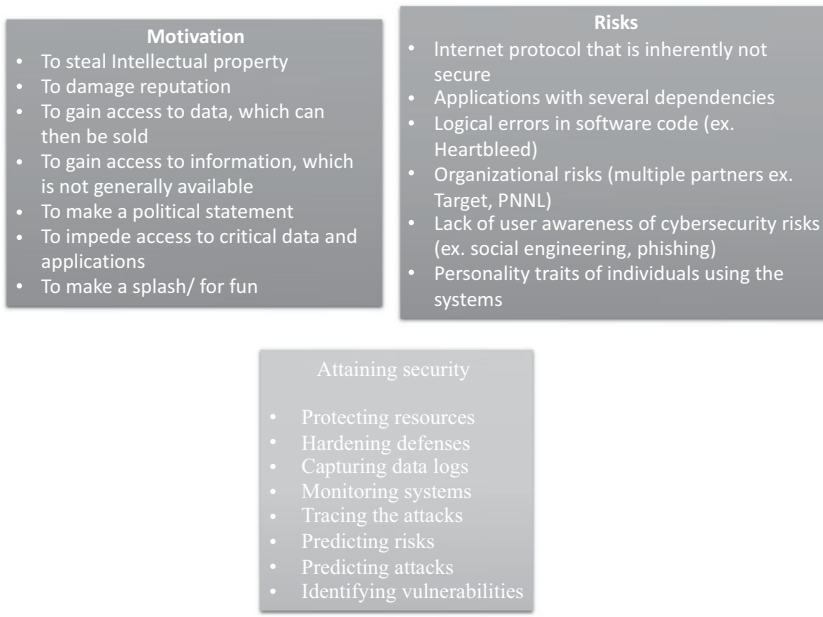

Figure 1.2 Unauthorized access motivations, risks and attaining security.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

- Making a political statement: Hacktivism, or hacking for activism (Samuel 2004), is an act carried out primarily to send a political message. An example included the cyber war in Estonia in 2007 (CSMonitor–Estonia 2007), where a series of DDoS attacks were directed toward official websites and even mobile phone networks. Estonia has several e-government initiatives, which had to be shut down for access from foreign locations due to the massive scale of the attacks. The Estonian foreign minister has also been quoted blaming the Russian government for facilitating these attacks.
- Performing cyber espionage: These attacks originate from government-supported cells, which target state secrets. Unlike theft of intellectual property, cyber espionage is geared toward gaining access to sensitive information. An example is the Titan Rain attack (Lewis 2005, Chi 2014), a type of advanced persistent threat (APT), believed to have been initiated by government-coordinated cells in China targeting US defense networks and contractors such as Lockheed Martin, Sandia National Laboratories, and NASA, among others. The SolarWinds hack led to a massive impact for federal systems believed to be directed by the Russian government (Stubbs et al. 2020). In some cases, this becomes a global supply chain attack, where the hack from a trusted vendor leads to a major attack, such as in the case of FireEye, which was impacted by the SolarWinds attack (Geenans 2020)
- In addition to the aforementioned motivations, cyberattacks may be aimed at damaging reputation (Townsend et al. 2014), which can lead to substantial losses; making a splash for fun (Gesenhues 2014), such as a traditional search example using Google dorking; and impeding access to data and applications (Risk Based Security 2014). Most of these scenarios are carried out to gain unauthorized access to information.

1.1.2.1 Why Do We Have Security Risks?

Security risks arise due to various factors, which include the following: applications with several dependencies, logical errors in software code (such as Heartbleed), organizational risks (multiple partners, such as in cyber-attacks at Target and the Pacific Northwest National Laboratories [PNNL]), lack of user awareness of cybersecurity risks (such as in social engineering and phishing), personality traits of individuals using the systems (phishing), and inherent issues in the Internet protocol being used.

1.1.2.2 What Is the Level of Damage That Can Occur?

According to a McAfee report, the monetary loss resulting from cybercrime costs about $600 billion, which is about 0.8% of the world Gross Domestic Product (GDP) (McAfee–Cybercrime Impact 2018), with malicious actors

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

becoming more and more sophisticated. Such large numbers are difficult to estimate, and the report outlines a sound strategy for reaching the estimated loss.

The loss due to cyberattacks is not simply based on direct financial loss but also based on several indirect factors that may lead to a major financial impact. As an example, let us consider the Target cyberattack. According to a Reuters news article (Skariachan and Finkle 2014), Target reported $61 million in expenses related to the cyberattack out of which $44 million was covered by insurance. Thus, the direct financial impact to Target was $17 million. Now let us consider some other factors that led to the indirect loss for Target: there was a 46% drop in net profit in the holiday quarter and a 5.5% drop in transactions during the quarter, share price fluctuations led to further losses, cards had to be reissued to several customers, and Target had to offer identity protection to affected customers. All these losses amount to much more than the total $61 million loss. In addition, the trust of the customers was lost, which is not a quantifiable loss and has long-term impacts.

1.2 Handling Cyberattacks

Now the key question is how one secures the resources against unauthorized accesses. Understanding and preventing such risks can be done in several ways, including protecting resources, hardening defenses, capturing data logs, monitoring systems, tracing the attacks, predicting risks, predicting attacks, and identifying vulnerabilities.

Several subdisciplines of cybersecurity deal with these aspects.

1.2.1 Subareas of Cybersecurity

Various areas of cybersecurity have developed techniques to prevent and respond to cyberattacks. There are several overlaps in these areas of cybersecurity. These include the following (shown in Figure 1.3):

Application security: This area deals with incorporating security in the software development process. This includes following best practices for code development, designing strong and comprehensive test cases, and following rigorous maintenance practices.

Data and information security: Increasingly with the advent of cloud data storage and remote access to data warehouses, data and information are at a risk of unauthorized access and misuse. Data and information security deals with securing data from such threats.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

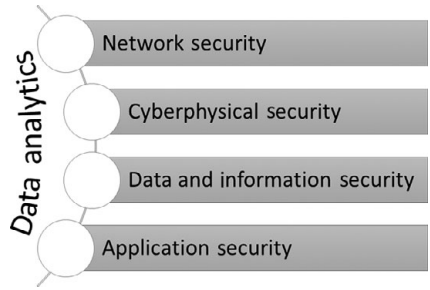

Figure 1.3 Areas of cybersecurity.

Network security: This deals with the challenges faced in securing the traditional computer networks and security measures adopted to secure, prevent unauthorized access and misuse of either the public or the private network.

Cyberphysical security: This focuses on the emerging challenges due to the coupling of the cyber systems with the physical systems. For example, the power plants being controlled by a cyber system present new security challenges that can arise due to the risk of disruption of the cyber component or risk of unauthorized control of the cyber system, thus gaining control of the physical systems.

Data analytics: This crosscutting theme can apply to each of these areas to learn from existing threats and develop solutions for novel and unknown threats toward networks, infrastructure, data, and information. Threat hunting (Sadhwani 2020), which proactively looks for malicious players across the myriad data sources in an organization, is a direct application of using data analytics on the various types of security data produced from the various types of security streams. However, this does not necessarily have to be a completely machine-driven process and should account for user behaviors as well (Shashanka et al. 2016), looking at the operational context. Data analytics can provide security analysts a much focused field of vision to zero in on solutions for potential threats.

Cybersecurity affects many different types of devices, networks, and organizations. Each poses different types of challenges to secure. It is important to understand and differentiate between these challenges due to the changing hardware and software landscape across each type of cybersecurity domain. While hardware configurations and types of connectivity are out of scope for this book, it is important to understand some of the fundamental challenges to study their impact and how they can be addressed using techniques such as

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

data analytics, which is crosscutting across the many different types of connectivity since data are ubiquitous across all these domains. In the current connected environment, multiple types of networks and devices are used, such as computer networks, cyberphysical systems (CPS), Internet of Things (IoT), sensor networks, smart grids, and wired or wireless networks. To some extent, IoT and sensor networks can be seen as specialized cases of CPS systems. We can consider these systems to study how data analytics can contribute to cybersecurity since any of these types of systems will generate data, which can be evaluated to understand their functioning and any potential benign or nonbenign malfunctions.

Computer networks are the most traditional type of networks where groups of computers are connected in prespecified configurations. These configurations can be designed using security policy deciding who has access to what areas of networks. Another way networks form is by determining patterns of use over a period of time. In both cases, zones can be created for access and connectivity where each computer in the network and subnetworks can be monitored.

Cyberphysical systems are an amalgamation of two interacting subsystems, cyber and physical, that are used to monitor a function. cyberphysical systems are used to monitor and perform the day-to-day functions of the many automated systems that we rely on, including power stations, chemical factories, and nuclear power plants, to name a few.

With the ubiquitous advent of connected technology, many "smart" things are being introduced into our connected environment. A connected network of such smart things has led to the evolution of Internet of Things. IoT has become excessively pervasive and prevalent from our smart homes to hospitals. These new types of connected systems bring about new challenges in securing them from attacks with malicious intent to disrupt their day-to-day functioning.

Throughout this book, we will use examples from various types of such connected systems to illustrate how data analytics can facilitate cybersecurity.

1.3 Data Analytics

Data analytics deals with analyzing large amounts of data from disparate sources to discover actionable information leading to gains for an organization. Data analytics includes techniques from data mining, statistics, and business management, among other fields.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

The large amount of data collected has led to the "big data" revolution. This is also the case in the domain of cybersecurity. Big data (Manyika et al. 2011, Chen et al. 2014) refers to not only massive datasets (volume) but also data that are generated at a rapid rate (velocity) and have a heterogeneous nature (variety), and that can provide valid findings or patterns in this complex environment (veracity). These data can also change by location (venue). Thus, big data encompasses the truly complex nature of data particularly in the domain of cybersecurity.

Every device, action, transaction, and event generates data. Cyber threats leave a series of such data pieces in different environments and domains. Sifting through these data can lead to novel insight into why a certain event occurred and potentially allow the identification of the responsible parties and lead to knowledge for preventing such attacks in the future.

Let us next understand how data analytics plays a key role in understanding cyberattacks.

1.3.1 Why Is Data Analytics Important for Cybersecurity? A Case Study of Understanding the Anatomy of an Attack

Cyberattacks evolve very fast. Let us consider a vulnerability in software code. A software patch can be developed to address the vulnerability; however, a slight modification may lead to a new vulnerability. There is a constant back and forth between the conspirators and the legitimate users. In several cases, the more serious attacks are not just a one-stop exploit that breaches a system through a single vulnerability, but a multipronged attack that uses several different channels to gain access to the secured systems.

To understand the intricacies of such a complex attack, let us consider the anatomy of an attack based on a scenario motivated from a recent attack on a federal lab (Dark Reading 2011). This example demonstrates the challenges posed by multiple aspects of cyberattacks coming from multiple sources and spread over time. Similar patterns of threat propagation have been seen in other attacks (such as Skariachan and Finkle 2014, FireEye 2020).

Figure 1.4 shows the flow of a multipronged attack. Despite the lab's well-protected information technology (IT) security perimeter, the attacks made it through in a very coordinated and prolonged process. Similar to a global supply chain attack (FireEye 2020), in the PNNL case, first there is an attack on the organization, and second, there is an attack on a partner that shares key resources. In the first part of the attack, intruders take advantage of vulnerabilities in public-facing web servers. In addition, hackers secretively scout the

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

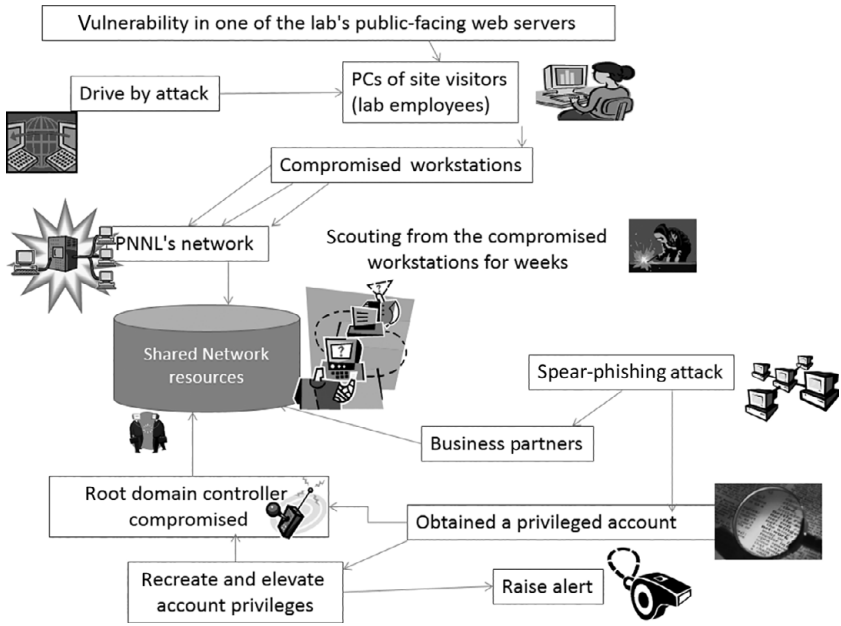

Figure 1.4 Anatomy of a multilevel attack.

network from compromised workstations that have already been targeted beforehand as part of a coordinated prolonged attack. The second part of the attack starts with spear phishing. Instead of casting out thousands of emails randomly, spear phishers target select groups of people with something in common, such as a common employer, a similar banking or financial institution, the same college, etc. So potentially a second group of hackers institutes a spear-phishing attack on the organization's major business partners with which it shares network resources. The hackers are able to obtain a privileged account and compromise a root domain controller that is shared by the organization and its partner. When the intruders try to recreate and assign privileges, it triggers an alarm, alerting the organization's cybersecurity team.

This scenario clearly demonstrates that simply looking at one dimension of the attack is not enough in such prolonged attacks. For such multipronged attacks, a multifaceted approach is required. Events of interest can be identified using a combination of factors such as proximity of events in time, in terms of series of communications, and even in terms of the geographic origin or destination of the communication, as shown in Figure 1.4.

For this specific example, we can evaluate three important aspects of analysis of the cyber traffic data, which can potentially generate new insights

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

in detecting unusual events of interest. The three aspects are temporal, spatial, and data-driven understanding of human behavioral aspects (particularly of attackers):

- Firstly, computer networks evolve over time, and communication patterns change over time. Can we identify these key changes, which deviate from the normal changes in a communication pattern, and associate them with anomalies in the network traffic?
- Secondly, attacks may have a spatial pattern. Sources and destinations in certain key geolocations are more important for monitoring and preventing an attack. Can key geolocations, which are sources or destinations of attacks, be identified?
- Thirdly, any type of an attack has common underpinnings of how it is carried out; this has not changed from physical security breaches to computer security breaches. Can this knowledge be leveraged to identify anomalies in the data where we can see certain patterns of misuse?

Utilizing the temporal, spatial, and human behavioral aspects of learning new knowledge from the vast amount of cyber data can lead to new insights of understanding the challenges faced in this important domain of cybersecurity.

Thus, simply looking at one dimension of the data is not enough in such prolonged attack scenarios. For such a multipronged attacks, we need a multilevel framework that brings together data from several different databases. Events of interest can be identified using a combination of factors such as proximity of events in time, in terms of series of communications and even in terms of the geographic origin or destination of the communication, as shown in Figure 1.5. Some example tasks that can be performed to glean actionable information are the following:

(a) **Clustering based on feature combinations**: One important piece of data collected in most organizations is Intruder Detection System (IDS) logs such as SNORT. These can be leveraged, and a keyword matrix and a word frequency matrix can be extracted to use for various analytical tasks. For example, the keyword matrix can be used to perform alarm clustering and alarm data fusion to identify critical alerts that may have been missed. Instead of clustering the entire set of features seen in a snort alarm, we can perform clustering based on a combination of features.

(b) **Collusions and associations**: Using the keyword matrix, we also extract associations to identify potentially repeated or targeted communications. This information in conjunction with network mapping can also be used to determine which attacks are consistently targeted to specific types of machines.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

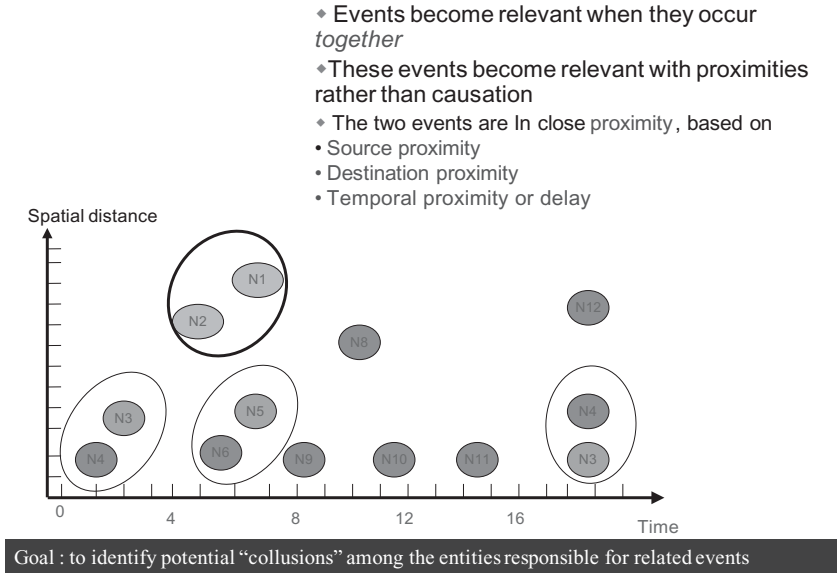

Figure 1.5 Multidimensional view of threats.

(c) **Time proximity and network evolution**: Clustering can be performed after creating time intervals that accounts for time proximity. This not only allows mining the data in proximity of time but also evaluating how the networks evolve over time and which time interval may be critical. For instance, we might want to identify if there are repeated events of interest in certain time periods. Lastly, looking at the clustering in different segments of time, we can look at mining for possible attack paths based on variations in cluster content and cluster cohesion.

1.3.2 How Can Data Analytics Help?

The attack case study shows the potential of data analytics and how data from multiple sources can be used to glean novel information. Data analytics can help in several ways to support the defense of cyber systems: mining system logs, monitoring systems, tracing the attacks, predicting risks by identifying critical systems in a network flow, predicting attacks based on prior or similar attacks, identifying vulnerabilities by mining software code, understanding user behavior by mining network logs, and creating robust access control rules by evaluating prior usage and security policies.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press

*What this book **is not** about:* This book does not address the traditional views of security configurations and shoring up defenses, including setting up computer networks, setting up firewalls, web server management, and patching of vulnerabilities.

*What this book **is** about*: This book addresses the challenges in cybersecurity that data analytics can help address, including analytics for threat hunting or threat detection, discovering knowledge for attack prevention or mitigation, discovering knowledge about vulnerabilities, and performing retrospective and prospective analysis for understanding the mechanics of attacks to help prevent them in the future.

This book will provide multiple examples across different types of networks and connected systems where data can be captured and analyzed to derive actionable knowledge fulfilling one of the several aims of cybersecurity, namely prevention, detection, response, and recovery.

https://doi.org/10.1017/9781108231954.002 Published online by Cambridge University Press