Cybersecurity as an Industry: A Cyber Threat Intelligence Perspective

Sagar Samtani, Maggie Abate, Victor Benjamin, and Weifeng Li

Contents

S. Samtani
Department of Information Systems and Decision Sciences, Muma College of Business, University of South Florida, Tampa, FL, USA
e-mail: ssamtani@usf.edu

M. Abate
WellCare Health Plans Inc., Tampa, FL, USA
e-mail: megebar.abate@gmail.com

V. Benjamin (✉)
Department of Management Information Systems, W.P. Carey School of Business, Arizona State University, Phoenix, AZ, USA
e-mail: victor.benjamin@asu.edu

W. Li
Department of Management Information Systems, Terry College of Business, University of Georgia, Athens, GA, USA
e-mail: weifeng.li@uga.edu

T. Holt, A. M. Bossler (eds.), *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, https://doi.org/10.1007/978-3-319-90307-1_8-1

Abstract

The rapid integration of information technology has been met with an alarming rate of cyber-attacks conducted by malicious hackers using sophisticated exploits. Many organizations have aimed to develop timely, relevant, and actionable cyber threat intelligence (CTI) about emerging threats and key threat actors to enable effective cybersecurity decisions. To streamline and create efficient and effective CTI capabilities, many major cybersecurity companies such as FireEye, Anomali, ThreatConnect, McAfee, CyLance, ZeroFox, and numerous others have aimed to develop CTI platforms, enabling an unprecedented ability to prioritize threats, pinpoint key threat actors, understand their tools, techniques, and procedures (TTP), deploy appropriate security controls, and ultimately, improve overall cybersecurity hygiene. Given the significant benefits of such platforms, our objective for this chapter is to provide a systematic review of existing CTI platforms within industry today. Such a review can offer significant value to academics across multiple disciplines (e.g., sociology, computational linguistics, computer science, information systems, information science, etc.) and industry professionals across public and private sectors. Systematically reviewing existing CTI platforms identified five future possible directions CTI start-ups can explore: (1) shift from reactive to proactive OSINT-based CTI platforms, (2) enhancement of natural language processing (NLP) and text mining capabilities, (3) enhancement of data mining capabilities, (4) further integration of big data and cloud computing technologies, and (5) opportunities and strategies for academia to address identified gaps.

Keywords

Cyber threat intelligence · Platforms · Data sources · Data mining

Introduction: Cybersecurity Significance and Cyber Threat Intelligence (CTI) Overview

Computing technology has provided modern society with innumerable and unprecedented benefits. Many organizations across private and public sectors employ complex information systems (IS) to maintain critical infrastructure, execute financial transactions, maintain health records, and conduct many other day-to-day activities. Unfortunately, the rapid integration of IS has been met with an alarming rate of cyberattacks conducted by malicious hackers using sophisticated exploits. Cybersecurity experts have appraised the total cost of hacking activities against major entities such as Target, Home Depot, Marriott, Equifax, Uber, and Yahoo! at $450 billion annually (Graham 2017). To combat this major societal and global

issue, many organizations have aimed to develop timely, relevant, and actionable intelligence about emerging threats and key threat actors to enable effective cybersecurity decisions. This process, also referred to as cyber threat intelligence (CTI), has quickly emerged as a key aspect of cybersecurity.

CTI is fundamentally a data-driven process. Similar to other data analysis procedures, organizations will first define their intelligence needs by examining the existing threat landscape, monitoring their cyber assets, and modelling possible attack vectors. This information guides data collection from Intrusion Detection and Prevention Systems (IDS/IPS) and log files from databases, firewalls, and servers. Well-refined analytics such as malware analysis, event correlation, and forensics are utilized to derive the intelligence needed for CTI professionals to deploy appropriate security controls (e.g., two-factor authentication, malware signatures for antiviruses, form sanitation, etc.) and develop more robust cyber defenses (Friedman 2015; Kime 2016; Shackleford 2016). Figure 1 illustrates the four phases of the general CTI lifecycle.

To streamline and create efficient and effective CTI capabilities, many major cybersecurity companies such as FireEye, Anomali, ThreatConnect, McAfee, CyLance, ZeroFox, and numerous others have aimed to develop CTI platforms. Such platforms are purchasable by organizations looking to improve their overall cybersecurity posture by utilizing CTI platforms to enable an unprecedented ability to prioritize threats; pinpoint key threat actors; understand their tactics, techniques, and procedures (TTPs); deploy appropriate security controls; and, ultimately, improve overall cybersecurity hygiene. Such benefits have led Anomali, the vendor of the internationally renowned CTI platform ThreatStream, to remark that: "An effective Threat Intelligence Platform can enable analysts to determine patterns of malicious behavior learned from previous events to better address future attacks." (Anomali 2017).

Given the significant benefits of such platforms, our objective for this chapter is to provide a systematic review of existing CTI platforms within industry today. Such a review can offer significant value to academics across multiple disciplines (e.g., sociology, computational linguistics, computer science, information systems,

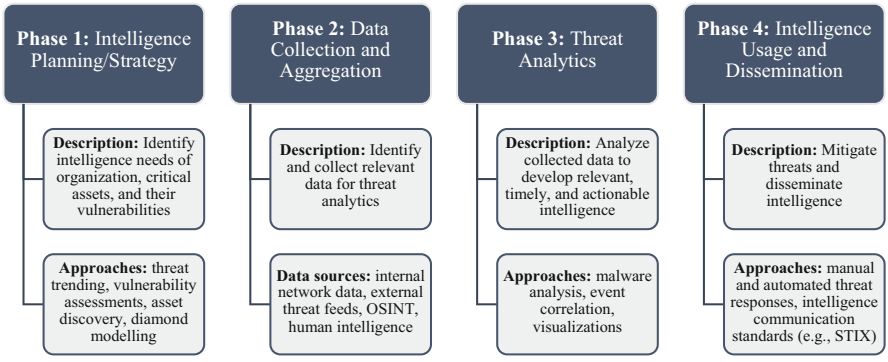

Fig. 1 General cyber threat intelligence (CTI) lifecycle

information science, etc.) and industry professionals across public and private sectors. For example, academics can use this review to serve as a basis to pursue novel, high-impact research inquiries that can advance future CTI platforms. Similarly, it can help provide grounding and justification for pursuing several federal funding opportunities to help support novel CTI initiatives that are of significant interest to academia, government, and industry. Further, security practitioners can form a better understanding of how CTI platforms are useful aggregators of intelligence sourced from ISACs (Information Sharing and Analysis Center), various data feeds (both paid and unpaid), network and system traffic, and more.

To achieve our objective, we organize this book chapter as follows: First, we provide a detailed background of CTI by summarizing the structure of CTI platforms, including common data sources, popular analytics, and operational intelligence (e.g., visualization, reporting capabilities). Second, we provide a systematic review of existing CTI industry platforms within the marketplace today. Third, we summarize existing gaps and offer suggestions on how these gaps can be addressed by academics and industry professionals. Finally, we summarize our overall findings and conclude this chapter.

Background of Cyber Threat Intelligence (CTI) Platforms

As indicated in the introduction, CTI is fundamentally a data analytics-oriented process. While numerous CTI platform vendors exist (detailed in section "Review of Cyber Threat Intelligence (CTI) Platforms"), CTI platforms generally have three major components: data collection, analytics, and operational intelligence. In the following subsections, we describe common CTI data sources, popular CTI analytics, and the common strategies for creating, organizing, and disseminating operational intelligence created from CTI analytics procedures.

Common Cyber Threat Intelligence (CTI) Data Sources

In general, five major categories of CTI data sources exist: (1) Open Source Intelligence (OSINT), (2) Internal Intelligence, (3) Human Intelligence (HUMINT), (4) Counter Intelligence, and (5) Finished Intelligence (FINTEL). Other data sources also include Signal Intelligence (SIGINT), Imagery (IMAGINT), Geospatial Intelligence (GEOINT), and Signature Intelligence. However, they are more common for military applications rather than CTI; thus, they are omitted from our review. Table 1 briefly summarizes each common data source, whether it is internal or external to an organization, several examples of each data source, and the value each data source provides.

Internal intelligence (most common and traditional CTI data source) is data collected from network logs generated from Intrusion Detection Systems, Intrusion Prevention Systems, databases, servers, routers, switches, and other network devices on an organization's networks. Such data is the most common and traditional data

Table 1 Common cyber threat intelligence (CTI) data sources used for analytics

CTI data source	Brief description	Internal or external?	Examples	Value
Open Source Intelligence (OSINT)	Data that can be collected from the Internet or from other CTI companies	External	Vulnerability/exploit feeds, social media, Darknet data, public statements, threat feeds	Provides a comprehensive view of an organization's external threat landscape
Internal Intelligence	Data collected from an organization's internal cyber assets	Internal	Network logs, database access events, Intrusion Detection Systems (IDS) logs, Intrusion Prevention Systems (IPS) logs	Provides information about activities internal to an organization
Human Intelligence (HUMINT)	Manual research and data collection	Both	Direct hacker interactions	Provides very precise and deep knowledge
Counter Intelligence	Providing false information to deceive hackers	Both	Honeypots, antihuman intelligence	Safely identifies tools and methods used by attackers
Finished Intelligence (FINTEL)	Finished Intelligence ready for dissemination	Both	Commercial data feeds	Refines and analyzes intelligence

source, as it is relatively straightforward to collect (e.g., dedicated packet sniffers, packet captures, vulnerability assessments, port scans, log aggregators, etc.), providing significant information regarding an organization's operations. OSINT offers organizations an opportunity to look at the "outside world" to identify relevant threats. Common data sources include traditional social media sites (e.g., Facebook, Twitter, PasteBin, Shodan, etc.) and Darknet (i.e., online hacker community) sources such as hacker forums, Darknet Marketplaces (DNMs), Internet-Relay-Chat (IRC), carding shops, and others. Such data sources can provide an excellent overview of potential cyber threats within cyberspace that are present within relevant industries. HUMINT relies on manual research and data collection (e.g., direct hacker interactions, ethnographies, etc.) to gain very precise and deep knowledge about particular threats (e.g., advanced persistent threats). Counter Intelligence is based on providing false information through automated or manual approaches (e.g., honeypots, antihuman intelligence) to deceive hackers. Such methods can offer a relatively safe approach to identify tools and methods used by attackers without directly engaging with attackers. Finally, FINTEL is finished intelligence ready for dissemination.

Table 2 Common cyber threat intelligence (CTI) analytics procedures

Analytical approach	Description	Examples	Value
Summary statistics	High-level summary of collected data	Number of blocked IPs, locations of attacks, counts over time	Good overview for cybersecurity executives for decision-making purposes
Event correlation	Analyzes relationships between events	Identifying machine sending malicious traffic by checking firewall log	Integrates multiple sources of data together (usually internal network)
IP reputation services	Identifying the quality of an IP	IP "X" has a poor reputation	Identify which IP addresses to block
Malware analysis	Analyzing (statically and/or dynamically) malicious files on a network or a system	Decompiling ransomware	Bolsters technical cyber defenses against malicious files and binaries
Anomaly detection	Detecting abnormal activities and behaviors	Unusual user logins, spurious network activities	Detect malicious activities
Forensics	Identifying and preserving digital evidence	Examining RAM from a malicious or hacked system	Identifying how an attack occurred
Machine learning	Algorithms that can learn from and make predictions/describe data	Classifying malware	Automating analysis and identifying patterns within data that are not possible by other methods

Popular Cyber Threat Intelligence (CTI) Analytics

Each of the aforementioned data sources summarized in the previous subsection can offer significant CTI value. However, each source requires significant analysis so that CTI value can be effectively drawn from them and insights can be developed. Broadly speaking, seven major types of CTI analytics exist: (1) summary statistics, (2) event correlation, (3) reputation services, (4) malware analysis, (5) anomaly detection, (6) forensics, and (7) machine learning. Table 2 briefly describes each analytics procedure, provides salient examples, and summarizes the value of technique. We note that the analytics summarized in Table 2 are neither exhaustive nor mutually exclusive. In practice, each of the listed analytical approaches can be used in silos or in conjunction with each other to maximize the potential CTI value. Moreover, companies may have developed proprietary approaches to effectively analyze the CTI data source(s) of interest and relevance to them.

Summary statistics provide CTI analysts with the ability to supply overviews and high-level summaries of vast quantities of data and/or carefully refined analytic results. Event correlation aims to fuse and integrate multiple data sources (usually internal network data sources) together to analyze relationships between events. IP reputation services aim to identify the quality of an IP address based on the type of data and payloads it is generating. Such analysis is valuable when identifying which

IPs to blacklist and block. Malware analysis is the process of systematically analyzing (through static and/or dynamic procedures) malicious files on a network or a system. Knowledge gleaned from these procedures can bolster cybersecurity controls against malicious files and binaries. Anomaly detection aims to detect abnormal activities and behaviors which deviate from a predefined set of normal activities (i.e., baseline). Forensics aims to carefully examine the factors which led to a cyberbreach by identifying and preserving digital evidence. Finally, machine learning offers a suite of algorithms that can learn from and make predictions and/or describe data. Such analysis is valuable for automation and identifying trends and patterns within data that are not possible by other methods. Taken together, each analytics procedure offers a valuable mechanism to systematically sift through the terabytes of generated data to glean valuable cybersecurity insights.

Operational Intelligence: Visualization, Report Generation, and Intelligence Dissemination

A key aspect of any CTI program and platform is the proper implementation of operation intelligence capabilities. This often takes form via three methods: visualization, report generation, and intelligence dissemination. Visualizations are visual representations of analyzed data. CTI analysts may have to deal with hundreds of thousands of log files, hacker data, etc. Generally speaking, visual cues can be more intuitive than raw data and pure textual information. Thus, visualizations can make threat data analytics and dissemination significantly easier. More importantly, they can enable better decision making, provide value to strategic level employees, and serve as an excellent reporting mechanism. Oftentimes, visualizations are the fundamental building blocks for biweekly, quarterly, semiannual, and annual CTI reports offered by major organizations. Common visualizations within the CTI landscape include, but are not limited to, bar charts, line charts for trend analysis and temporal evolution, network science-based representations, radar charts, box and whisker plots, geo-spatial maps, heat maps, and many others.

Following the generation of these visualizations, reports, and analyzed data is the sharing of information with key stakeholders. Cyber threat information sharing is critical in the fight against today's sophisticated cyber adversaries and emerging threats. Cyber threat information sharing relies on asking and answering a series of key questions, including (but not limited to) who to tell (e.g., incident response team, chief information security officer, staff, developer, clients, etc.), when to tell, what to tell, and how to tell. Ultimately, the sharing of this information can help organizations deploy automated and manual defenses (i.e., security controls) to help prevent and mitigate cyberattacks. Automated defenses are those in which security controls are automatically deployed after an event has been identified. Examples of automated defenses include deploying a firewall rule based on abnormal activities on a port, flagging specific emails based on its features, and automatically blocking a user account based on an abnormal activity. Such actions can significantly reduce an incident response team's overall workload. In contrast, manual defenses require forms of

human intervention. These include manual deployment of third-party controls, high-risk mitigation strategies, human interfaces, interacting with hackers in the Darknet, addressing insider threats, combating social engineering threats, and many others.

Review of Cyber Threat Intelligence (CTI) Platforms

In total, we reviewed 91 CTI companies focused on developing CTI capabilities. Most of the reviewed CTI companies are US-based, though a small number of international firms were included. The CTI companies were identified through a Gartner report and also through domain expertise regarding reputable vendors with advanced threat intelligence capabilities. Further, the review is focused on companies which offer paid platforms, rather than those found through open source feeds and resources (such as those found at https://github.com/theragnarpatel/awesome-threat-intelligence). This helps focus the review on the commercial space. Additionally, we focused on reviewing companies which offered CTI platforms with the three major components described in the previous subsections: data collection, analytics, and operational intelligence. For each company, we aimed to identify what data sources a company's platform uses, data sources when the company was founded, its headquarters, and other information. To conduct the review, we carefully examined each company's websites, platform data sheets, fact sheets, and other publicly accessible resources. This helps ensure that we gathered a comprehensive set of information for each company's CTI platform without purchasing the actual platform itself (out of the scope of this project).

A summary of our review is provided in this section. Specifically, we first provide an overview of the entire industry (e.g., age of the industry, locations of many companies, revenue streams, their public/private status, etc.). We then systematically review platforms on the three aspects found in CTI platforms: data collection efforts, CTI analytics efforts, and operational intelligence efforts. Throughout the summary, we highlight specific company names as examples and illustrations. Interested readers who wish to obtain the full review in table format can email the authors.

General Observations: Industry Age, Locations, Revenue Streams, and Public Standing

In contrast to other industries such as retail, banking, and healthcare that have existed for over half a century, 59 of 91 of the reviewed CTI companies were founded in the 2000s. Similar to many traditional and well-established technology companies (e.g., Facebook, Twitter, LinkedIn, Google, Apple, Adobe, etc.), many (34 of 91) CTI companies were founded in and currently headquartered in the greater Silicon Valley area (e.g., San Francisco, San Jose, Palo Alto, Mountain View, etc.) and in other areas of California (e.g., Los Angeles, San Diego). However, contrast to the traditional technology companies, the majority of reviewed CTI companies (56 of 91) were founded and are currently headquartered outside of California. These include Impulse

Table 3 Summary of CTI companies based on revenue brackets

Revenue range	Number of companies	Number of companies that are publicly traded	Number of private companies	Example companies
$1–$10 million	22	0	22	Ziften, ZeroFox, Lifars
$10–$100 million	31	1	30	LookingGlass, ForeScout, AlienVault
$101–$999 million	20	4	16	CarbonBlack, FireEye, NSFocus
Over $1 billion	18	14	4	Checkpoint, Verint, Juniper Networks

in Tampa, Florida, Threatq in Washington, DC, Forcepoint in Austin, Texas, and Insights in Israel. Such observations bode well for those researchers aiming to produce CTI platforms using federal seed funding and other mechanisms (refer to subsection e of the next section) outside of the numerous venture capitalist (VC) funding opportunities commonly found within Californian borders. At the time of this writing, 24 of the 91 companies are publicly traded, while the remainder (67 of 91) remain private. Table 3 summarizes the reviewed CTI companies based on their revenue bracket.

The relative youth of the CTI industry has resulted in the majority of companies (53 of 91) falling within the $1–$10 million (22 of 91) and $10–$100 million (31 of 91) revenue brackets. Companies in the former bracket include Ziften, ZeroFox, and Lifars, while the latter comprise of LookingGlass, ForeScout, and AlienVault. Only one of these companies, NSFocus, is publicly traded; the remainder are private. In contrast, 14 of the 18 companies in the $1 billion+ revenue bracket (e.g., Checkpoint, Verint, Juniper Networks) are public. These 14 make up 73.68% of all publicly traded companies within the CTI industry. Overall, these results indicate that the CTI industry remains one that is rapidly emerging and growing (boding well for budding entrepreneurs in this space), and not one which has reached saturation.

Data Collection Efforts

Carefully examining the data sources used by the reviewed CTI companies revealed that 90%+ companies rely either solely or primarily on internal network data (e.g., log files generated from servers, databases, IDS/IPS, security information and event management (SIEM), and other networked devices). Oftentimes, the log files are collected with one or a combination of two strategies. The first is deploying sensors and log aggregators on client networks to gather data. This allows the CTI platforms provided by the CTI vendors to gather and analyze data closest to the client. One such platform that employs this strategy is ThreatConnect, who offers services to put monitoring technologies on selected endpoints within an enterprise network to collect data. The second strategy relies on sensor networks deployed worldwide.

Such a deployment enables the collection of vast amounts of data and provides a global perspective on potential cyberattack events worldwide. One such company deploying worldwide sensor networks is FireEye.

While network data remains the prevailing data source, the Darknet is slowly emerging as a viable data source for selected CTI companies. The Darknet consists of a multitude of underground online communities inhabited by cybercriminals. Darknet communities span across web forums, Internet-Relay-Chat (IRC), black

Table 4 Summary of selected CTI companies using Darknet as a data source

CTI company using Darknet	Date founded	Location	Public or private?	Revenue	Target audience	Only data source Darknet?
Lifars (consulting firm)	2013	New York, NY	Private	$6.2 million	Small to medium organizations	No
LookingGlass	2006	Washington, DC	Private	$12 million	Small to medium organizations	No
Recorded Future	2007	San Francisco, CA	Private (Series E)	$5 million	Medium to large organizations	No
Digital Shadows	2011	San Francisco, CA	Private (Series C)	$5 million	Small to medium organizations	No
Skybox Security	2002	Silicon Valley	Private	$15 million	Medium to large organizations	No
Blueliv	2009	European Union (EU)	Private (Series A)	$11.3 million	Small to medium organizations	No
Verint	1994	New York, NY	Public	$1.135 billion	Large organizations	No
Cyber4Sight Booz Allen Hamilton	1914	McLean, VA	Public	$5.48 billion	Large organizations and government	No
SurfWatch Labs	2013	Washington, DC	Private (Series B)	$3 million	Small to medium organizations	No
Flashpoint	2010	New York, NY	Private (Series C)	$5.6 million	Medium to large organizations	No
Insights	2015	Israel	Private (Series D)	$2.5 million	Small to medium organizations	No
ZeroFox	2013	Baltimore, MD	Private (Series C)	$8 million	Small to large organizations	No
DarkOwl	2015	Denver, CO	Private	$6.1 million	Small to large organizations	Yes

markets, stolen data shops, and more. Participants will often share or trade cybercriminal assets such as hacking tools, tutorials, and services. Discussion topics are often related to new attack techniques, emerging threats, potential targets, victims, and so on. The availability of these resources and information has also enabled many lesser-skilled Internet miscreants to conduct advanced cybercriminal operations that may cause disruption and financial loss. The Darknet exacerbates existing cybersecurity issues and is a critical data source for academics and security practitioners.

To the best of our knowledge and to the extent of our review, we identified 13 of 91 CTI companies who explicitly mentioned their use of the Darknet as a CTI data source. These include Lifars, LookingGlass, Recorded Future, Digital Shadows, Skybox Security, Blueliv, Verint, Cyber4Sight Booz Allen Hamilton, SurfWatch Labs, Flashpoint, Insights, ZeroFox, and DarkOwl. While the specific aspect of the Darknet collected and used for CTI analytics is not detailed in many company's data and fact sheets, Table 4 summarizes each of these companies based on the date they were founded, their location, whether they are private or public, their revenue, their target audience, and whether the Darknet is the only CTI data source these companies use.

Several key observations are drawn from our review of CTI companies using Darknet data sources. First, 11 of the 13 reviewed companies are private; only Verint and Cyber4Sight Booz Allen Hamilton are publicly traded. Among the 11 that are private, 8 were founded in the past decade (i.e., since 2009). Second, revenues for each of these companies range between 2.5 million and 5.48 billion. These observations suggest that many established companies have not yet forayed extensively into Darknet data sources. However, this appears to be the focus of 50% (11 of 22) of the companies within the $1–$10 million range. Third, about half of the companies using the Darknet aim to provide services to small- to medium-sized organizations. This indicates that medium- to large-sized organizations are often not the target audience of these CTI companies. Finally, carefully examining each company's data sheets indicates that the Darknet is not the only data source employed for analytics purposes for 12 of 13 companies. This indicates that internal network data is used to correlate and analyze events occurring on network devices and Darknet platforms. The only exception is DarkOwl, who focuses their value proposition on collecting a comprehensive set of Darknet data for selected clients to develop intelligence from (whereas other companies are more targeted in nature).

CTI Analytics Efforts

Our review revealed that all companies use one or a combination of the CTI analytics procedures detailed earlier (e.g., summary statistics, event correlation, malware analysis, IP reputation services, anomaly detection, forensics, etc.). Employing this breadth of analytics provides two key benefits. First, it enables companies to have a rich toolbox of approaches that can effectively extract intelligence from the diverse data often collected. Second, providing a suite of analytics options allows these companies to offer multiple selections and pricing plans to organizations interested

in purchasing such offerings. Such strategies enable CTI companies to maximize their revenues and marketability to organizations interested in adopting a CTI platform as part of their overall cybersecurity strategy. CTI companies providing multiple offerings include FireEye, McAfee (multiple malware analysis offerings), and Symantec.

Our review also revealed that artificial intelligence-based methods (e.g., data mining, machine learning, natural language processing) have also rapidly permeated and emerged in the cyber threat intelligence industry. While not as widespread as the traditional CTI analytics, these methodologies offer a promising mechanism to automate CTI analytics and discover patterns, relationships, and associations within large amounts of cybersecurity data which would otherwise be undetectable. Examples of traditional CTI analytics which have seen significant improvements and benefits include malware analysis and anomaly detection. Oftentimes, these analytics are conducted on a "Big Data" scale (e.g., terabytes of data moving in real time). Companies leading the forefront of AI-based CTI analytics include McAfee, CyLance, Cybersift, and Insights. However, carefully examining the data sheets from these companies suggests that many of the algorithms used for analytics are "out-of-the-box" (i.e., provided by standard data mining software packages such as scikit-learn, Rapidminer, or WEKA). One such example is using the standard k-means clustering algorithm to automatically group similar malware samples. Ultimately, such algorithms can enable more effective and efficient threat detection, mitigation, and incident response.

Operational Intelligence Efforts

Reviewing the operational intelligence aspect revealed two key discoveries. First, to our surprise, less than half of the companies (43 of 91) offered visualization services as part of their CTI platform services. Those companies providing such services relied heavily on dashboards that contained multiple types of visualizations (e.g., bar charts, line charts, network diagrams, etc.) together. Oftentimes, these dashboards are real time, dynamic, interactive, and carefully laid out such that key stakeholders receive actionable (i.e., relevant and timely) intelligence. Examples of companies providing visualization services include Splunk, Insights, Rapid7, Checkpoint, and others. Interestingly, nearly all of the companies collecting and analyzing Darknet data offer some visualization capabilities.

The second key insight drawn from our review is the provision of threat intelligence feeds. Oftentimes, companies have developed numerous technical mechanisms for storing threat analysis data. These include Structured Threat Information eXpression (STIX), Cyber Observable Expression (CybOX), and/or Malware Attribution Enumeration and Characterization (MAEC). Each uses data formats such as eXtensible Markup Language (XML) and/or JavaScript Object Notation (JSON). Moreover, each follows a prespecified schema and data dictionary definitions. Automatic data sharing technologies such as Trusted Automated eXchange of Indicator Information (TAXII) and other Application Programming Interfaces

(APIs) serve as technical mechanisms for interested clients or consumers to automatically collect selected data from companies. Such feeds can provide actionable intelligence for other organizations within and across industries, as well as input for other organization's threat analytics. Organizations can serve as providers of threat feeds, consumers, or a combination of both. Ultimately, this enables organizations to share data within and across industries at near real-time speeds. Selected considerations when choosing threat feeds can include the following:

- Data sources used for analytics and operational intelligence
- Analytics employed to analyze selected data
- Cost of feed (if appropriate and relevant)
- Functionalities of feed (e.g., dynamic updating, etc.)
- Formatting of the data feed for ingestion into existing databases and warehouses
- Visualization capabilities to present collected and analyzed data

Existing Gaps Within CTI Platform Landscape and Potential Opportunities

To date, staggering advances have been made in progressing CTI platforms. However, the development of any nascent industry often has its gaps. Identifying and working progressively to addressing those gaps can help develop novel insights and capabilities beyond those currently seen within the marketplace. Throughout our review, we identified four major gaps and areas of further expansion: (1) lack of proactive OSINT-based CTI platforms, (2) enhancement of natural language processing (NLP) and text mining capabilities, (3) enhancement of data mining capabilities, and (4) further integration of big data and cloud computing technologies. We note that issues such as customization, cost, vendor selection, vendor support, and others are not unique to the CTI industry but are common issues seen among various technological industries. As such, we omit such general drawbacks and considerations common across the technological industry. Rather, we focus on the four abovementioned issues that are unique to the CTI industry. The final subsection summarizes some of the possible opportunities available to academia (with special focus on current federal funding opportunities) to solve some of these gaps and progressively improve the CTI platform landscape.

Shift from Reactive to Proactive OSINT-Based CTI Platforms

Despite its value, existing CTI practices have been criticized as reactive to known exploits, rather than proactive to new and emerging threats from the hackers themselves. To combat these concerns, CTI experts have suggested proactively examining emerging exploits in the vast, international, and rapidly evolving online hacker community platforms (i.e., "Darknet"). As indicated earlier, Darknet platforms include hacker forums, Darknet Marketplaces, Internet-Relay-Chat (IRC), and

carding shops. Hackers have obtained exploits in forums to execute well-known breaches (e.g., Target in 2013). Innovative solutions for salient cybersecurity issues require interdisciplinary efforts cutting across private and public sectors. Recognizing these challenges, there is a need for developing advanced, proactive CTI capabilities by (1) identifying and automatically collecting a multimillion record testbed of hacker community posts and (2) analyzing the rich textual nature of these posts to identify emerging threats, specifically malicious hacker exploits (malware). While this is a growing body of literature in these areas (Benjamin and Chen 2013; Benjamin et al. 2015, 2016, 2019; Benjamin 2016; Li 2017; Li and Chen 2014; Li et al. 2016a,b; Samtani et al. 2015, 2016; 2017; Samtani and Chen 2016; Grisham et al. 2017), significant work is required to effectively transition proof-of-concept and proof-of-value methodologies demonstrated in these studies to practice.

Enhancement of Natural Language Processing (NLP) and Text Mining Capabilities

Significant quantities of the data found within the realms of cybersecurity are text. Traditional internal network devices (e.g., Intrusion Detection Systems, Intrusion Prevention Systems, databases, workstations, routers, switches, gateways, etc.), emerging significant hacker community data sources (e.g., hacker forums, Internet-Relay-Chat, carding shops, and Darknet Marketplaces), and traditional Open Source Intelligence (OSINT) sources offered by major commercial entities (e.g., Facebook, Twitter, PasteBin, Shodan, etc.) are ripe with rich information that can significantly aide organizations in developing comprehensive and holistic cyber defenses. To date, many cybersecurity companies such as FireEye, Splunk, IBM, Webroot, and many others are looking beyond traditional structured data to mine novel insights out of these rich textual data sources. However, such practices have not yet been widely adopted across the entire CTI industry.

A common paradigm which many companies and researchers can adopt is natural language processing (NLP) and text mining. Such methodologies can offer significant value in traditional CTI analytics, including but not limited to malware analysis, phishing (e.g., fake email and/or website detection), anomaly detection, and many others. These include semantic matching, coreference resolution, distant supervision, tagging, parsing, named entity recognition (NER), entity resolution, feature selection/reduction, ontology development, topic modelling (e.g., latent Dirichlet allocation, latent semantic analysis), and others. In recent years, there has been a shift to emerging deep learning based NLP approaches. These include neural information retrieval (neural IR), hacker language modelling, diachronic linguistics (i.e., mapping how language evolves over time to detect emerging threats), deep structured semantic modelling for short text matching (offers value for data fusion tasks), text-augmented social network analysis, and numerous others. Despite remarkable advances in the fundamental principles for each aforementioned methodology, the unique characteristics of cybersecurity data (e.g., version names, rapidly evolving hacker terminology, significant multilingual content, computer-

generated content, etc.) necessitate the development of novel computationally oriented NLP and text processing approaches inspired by these domain-specific features. Consequently, this can be the feature and value offering proposition of novel CTI platforms offered by nascent CTI companies entering the existing marketplace.

Enhancement of Data Mining Capabilities

Data mining holds significant promise in advancing numerous traditional analytics commonplace within CTI, including malware analysis, IP reputation services, phishing email detection, event correlation, anomaly detection, and others. Data mining can assist CTI efforts from two perspectives. First, it can help organizations and researchers identify patterns within datasets which are not readily apparent by other analytics approaches (e.g., summary statistics, manual inspection, basic malware analysis, etc.). Second, it can help CTI researchers and practitioners process large amounts of data in an effective and efficient manner. In a domain where the amount of data is being generated at staggering rates from a variety of data sources (e.g., internal network devices, OSINT, etc.), these benefits are critical to ensuring that an organization is able to effectively extract key insights from all collected data. Beyond enhancing the aforementioned traditional CTI analytics, data mining can provide an array of new inquiries for cybersecurity. These include, but are not limited to, grouping similar types of network events together; clustering similar threat actors in hacker community platforms (e.g., hacker forums); classification of log files into predefined bins; detecting an adversary's tactics, techniques, and procedures (TTPs); stream data mining to deliver real-time cyber threat intelligence data feeds; and many other analytics possibilities.

Further Integration of Big Data and Cloud Computing Technologies

Cybersecurity analytics, such as cyber threat intelligence (CTI) processes, is fundamentally a Big Data analytics problem. Promising CTI data sources include Open Source Intelligence (OSINT) such as Facebook, Twitter, PasteBin, hacker forums, IRC, Darknet Marketplaces, and carding shops. They may also include data from internal network devices such as IDS/IPS, routers, databases, firewalls, switches, servers, SIEM. These various data sources are aggregated to generate terabytes of heterogeneous (structured and unstructured, of varying quality) data, often at real-time speeds. Consequently, CTI shares the same five Vs commonly associated with traditional Big Data contexts (e.g., e-commerce, business intelligence, health informatics, etc.): volume, variety, velocity, veracity, and value. The similarities of CTI characteristics vis-à-vis these traditional, high-impact domains suggest that technologies such as Hadoop (MapReduce + Hadoop Distributed File System (HDFS)), Apache Spark (GraphX, SparkSQL, Spark Streaming, and Machine Learning Library (MLlib)), Hive, Sqoop, Mahout, and others can assist scholars and practitioners to efficiently collect, process, and present threat data, analytics, and key

insights. For example, Hadoop can provide a highly scalable solution to systematically extract features. Moreover, the rich set of analytical programs offered by technologies built upon Hadoop provide access to common feature reduction algorithms (e.g., principal components analysis (PCA), autoencoders, etc.) to extract the most critical features from provided inputs for subsequent malware classification and clustering tasks. Storing feature lists using HDFS and then using MapReduce-based programs and technologies to distribute the feature extraction and selection process to a cluster of machines can achieve significant savings time and finances. Apache Spark's Spark Streaming functionality can allow researchers to analyze large amounts of real-time streaming data (commonplace in the CTI domain). Taken together, leveraging Big Data technologies in conjunction with fundamental CTI principles and approaches (e.g., honeypots and malware analysis) can push the frontier and boundary of novel CTI Big Data analytics.

Similarly, fusing multiple sources of data together that are naturally disparate enables the access to all attributes (commonly referred to as features or dimensions in machine learning and data mining literature) in each fused data source. This aggregation of attributes can significantly improve the performance of traditional data mining task classification, clustering, dimensionality reduction, and recommendation algorithms. Each is employed for significant cybersecurity analytics applications such as Big Data malware analysis, phishing detection, OSINT analysis (both hacker community and traditional social media sources), event correlation, and others. Moreover, comprehensively collecting and aggregating a diverse set of features across multiple datasets increases the CTI researcher's capability to develop new features for enhanced data mining algorithm performances and/or Key Performance Indicators (KPIs) (e.g., those that can assist organizations and researchers in systematically quantifying and prioritizing risk). Cloud computing services provided by major providers such as Amazon Web Services (AWS), Microsoft Azure, Digital Ocean, and others can provide effective mechanisms for organizations and researchers to deploy environments (e.g., on-demand networks across multiple geographic regions) to significantly streamline Big Data CTI collection and analytics for selected CTI platform capabilities.

Opportunities and Strategies for Academia to Address Identified Gaps

Addressing the aforementioned gaps summarized in the previous subsections is a nontrivial task. Each requires well-constructed teams containing a diverse set of expertise, interests, and experiences. Such teams can include perspectives drawn from multiple disciplines. These include, but are not limited to, technical fields such as computational linguistics, computer science, information systems, electrical and computer engineering, and information science, as well as social science-based ones such as cognitive science, communications, criminology, psychiatry, and psychology. Each offers their unique perspectives to tackle separate yet related and

intertwined issues to help deliver unique CTI capabilities within novel CTI platforms.

One of the most efficient mechanisms to quickly make impact in the CTI landscape while simultaneously harnessing the collective knowledge across disciplines is acquiring grant funding from world-renowned federal agencies. Acquiring such funding has numerous benefits. These include the ability to develop long-term, sustainable research programs around major CTI issues (as opposed to forming ad hoc teams), the ability to generate a strong reputation, recruiting high-caliber research scientists, strong graduate students, and the ability to foster and facilitate industry/government and interdisciplinary academic collaborations. Among other options such as Defense Advanced Research Projects Agency (DARPA) or Intelligence Advanced Research Projects Activity, the National Science Foundation (NSF) regularly promotes solicitations relevant to CTI and, in a more broad sense, cybersecurity. Common recent solicitations include Cybersecurity Innovation for

Table 5 Summary of selected National Science Foundation (NSF) cybersecurity relevant funding opportunities with descriptions and possible areas of CTI investigation

NSF funding opportunity	Brief description of program	Possible areas of CTI investigation (selected)
Secure and Trustworthy Cyberspace (SaTC)	Supports fundamental research related to cybersecurity and privacy	Transitioning CTI analytics platforms to practice Identifying emerging threats and key threat actors via emerging machine learning, text mining, and deep learning analytics
Scholarship-for-Service (SFS)	Designed to enhance cybersecurity workforce development	Developing a workforce of government agents with significant CTI expertise (e.g., collection, analytics)
Community CISE Research Infrastructure (CCRI)	Provides resources to launch a computational research infrastructure	Designing a highly customized environment supporting advanced CTI data collection and analytics for multiple CISE research communities
Data Infrastructure Building Blocks (DIBBs)	Designing a computational research testbed to support transformative research opportunities	Developing a long-term storage source for maintaining and curating CTI data collection
EArly-concept Grants for Exploratory Research (EAGER)	Offers funding to potentially transformative yet high-risk project	Data fusion of multiple CTI and traditional CTI platforms for advanced and holistic analytics
CISE Research Initiation Initiative (CRII)	Provides seed funding to early career junior faculty to initiate their research	Launching CTI research streams and teams to begin conducting selected CTI research
CAREER	Provides funding to junior faculty to promote a lifetime of research and teaching excellence	Integrative, long-term CTI research and education opportunities

Cyberinfrastructure (CICI), Secure and Trustworthy Cyberspace (SaTC), EArly-concept Grants for Exploratory Research (EAGER), Scholarship-for-Service (SFS), Community Computer and Information Science and Engineering (CISE) Research Infrastructure (CCRI), and Data Infrastructure Building Blocks (DIBBs). Table 5 presents a summary of selected NSF solicitation relevant to cybersecurity and also details possible areas of CTI investigation for each solicitation. For the purposes of this chapter, we focus our review of promising solicitations to NSF only, as they are often viewed as the "gold standard" and most prestigious and portable funding source for academics across a multitude of technical and nontechnical disciplines.

Over the past decade, each program has funded multimillion dollar, large-scale, multi-institutional projects. Each program encourages multidisciplinary, high-impact cybersecurity research with significant intellectual merit that can be publishable at premier and top-tier journals and conferences across a multitude of disciplines. Additionally, each program aims to fund projects which can offer significant broader impacts to practice. Examples include CCRI, DIBBs, and SFS. The first two can offer scholars with enough seed funding to pursue and develop innovative CTI infrastructure such that professionals from academia, government, and industry can pursue high-impact opportunities not otherwise possible. On the other hand, SFS can help address the significant shortage of trained cybersecurity professionals for eventual placement into federal, state, and local government positions by offering cutting-edge cybersecurity curricula and education opportunities.

While each program offers significant promise in helping academics achieve and execute advanced CTI research, infrastructure, and capabilities, the program which has emerged as the premier source for cybersecurity funding is SaTC. Since its inception in 2012, the NSF SaTC program supports research that addresses cybersecurity and privacy. Cutting across multiple CISE divisions and drawing upon numerous technical perspectives, SaTC project PIs have made remarkable advances in adversarial data mining, anomaly detection, real-time log analytics, and many other areas. SaTC offer several funding mechanisms within including CORE research, education (EDU), Transition to Practice (TTP), CISE Research Initiation Initiative (CRII) and CAREER. The latter two serve as prestigious early seed funding for junior faculty at the start of their academic career (SaTC, a program offered within CRII and CAREER). Among these, TTP presents a promising opportunity for teams of researchers with current SaTC funding to transition their research into the marketplace. Additional funding mechanisms to help transition the research into practice and ensure sustainability of the technologies can be attained via mechanisms such as Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) sources, NSF, or other agencies such as the Department of Defense (DoD) and Defense Advanced Research Projects Agency (DARPA).

Conclusion

Hackers' regular exploitation of numerous information systems technologies costs the global economy nearly half a trillion dollars yearly. Cyber threat intelligence offers a promising mechanism for organizations to select appropriate cybersecurity controls (e.g., authentication, protocols, cryptography, etc.) to improve their overall cybersecurity posture. CTI platforms are developed to streamline the CTI process to prioritize threats; pinpoint key threat actors; understand their tactics, techniques, and procedures (TTPs); deploy appropriate security controls; and improve the overall cybersecurity hygiene and posture of organizations. These platforms draw upon (1) Open Source Intelligence (OSINT), (2) Internal Intelligence, (3) Human Intelligence (HUMINT), (4) Counter Intelligence, and (5) Finished Intelligence (FINTEL) coupled with (1) summary statistics, (2) event correlation, (3) reputation services, (4) malware analysis, (5) anomaly detection, (6) forensics, and (7) machine learning to develop threat intelligence. Through visualization, report generation, and intelligence dissemination, the resulting threat intelligence can help organizations effectively deploy automated and manual defenses and ultimately improve the cybersecurity posture of organizations.

Systematically reviewing dozens of CTI platforms revealed the CTI industry remains one that is rapidly emerging and growing, and not one which has reached saturation. This bodes well for budding entrepreneurs interested in exploring this space. Systematically reviewing existing CTI platforms identified five future possible directions CTI start-ups can explore: (1) shift from reactive to proactive OSINT-based CTI platforms, (2) enhancement of natural language processing (NLP) and text mining capabilities, (3) enhancement of data mining capabilities, (4) further integration of big data and cloud computing technologies, and (5) opportunities and strategies for academia to address identified gaps. Numerous funding opportunities from highly visible sources (e.g., NSF CCRI, SaTC, CRII, CAREER, DIBBs, SFS, EAGER, SBIR/STTR, TTP, etc.) can help scholars attain requisite funds to assemble teams, conduct high-impact and relevant CTI research within these identified gaps, and transition their findings into the CTI industry for adoption by broader society. Ultimately, addressing one or more of these gaps currently can help ensure a safer and more secure society.

Acknowledgments This work was supported in part by NSF CRII CNS-1850362.

References

Anomali. (2017). ThreatStream 6.0 Data Sheet. https://anomali.cdn.rackfoundry.net/files/ThreatStream_6.0.pdf.

Benjamin, V. A. (2016). *Securing cyberspace: Analyzing cybercriminal communities through web and text mining perspectives.* Doctoral dissertation, University of Arizona.

Benjamin, V. A., & Chen, H. (2013). Machine learning for attack vector identification in malicious source code. In *2013 IEEE international conference on intelligence and security informatics (ISI)* (pp. 21–23). IEEE.

Benjamin, V., Li, W., Holt, T., & Chen, H. (2015). Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops. In *2015 IEEE international conference on intelligence and security informatics (ISI)* (pp. 85–90). IEEE.

Benjamin, V., Zhang, B., Nunamaker, J. F., & Chen, H. (2016). Examining hacker participation length in cybercriminal internet-relay-chat communities. *Journal of Management Information Systems, 33*(2), 482–510.

Benjamin, V., Valacich, S. J., & Chen, H. (2019). DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics. *MIS Quarterly, 43*(1), 1–22.

Friedman, J. 2015. Definitive guide to cyber threat intelligence. CyberEdge Group, LLC. https://cryptome.org/2015/09/cti-guide.pdf.

Luke Graham. (2017). Cybercrime costs the global economy $450 billion: CEO. Retrieved June 5, 2017, from https://www.cnbc.com/2017/02/07/cybercrime-costs-the-global-economy-450-billion-ceo.html.

Grisham, J., Samtani, S., Patton, M., & Chen, H. (2017). Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *2017 IEEE international conference on intelligence and security informatics: Security and big data, ISI 2017* (pp. 13–18).

Kime, B. P. (2016). *Threat intelligence: Planning and direction*. SANS Institute. https://www.sans.org/reading-room/whitepapers/threats/threat-intelligence-planning-direction-36857. Accessed 5 June 2017.

Li, W. (2017). *Towards secure and trustworthy cyberspace: Social media analytics on hacker communities*. Doctoral dissertation, University of Arizona.

Li, W., & Chen, H. (2014). Identifying top sellers in underground economy using deep learning-based sentiment analysis. In *2014 IEEE joint intelligence and security informatics conference* (pp. 64–67). IEEE.

Li, W., Chen, H., & Nunamaker, J. F. (2016a). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *Journal of Management Information Systems, 33*(4), 1059–1086.

Li, W., Yin, J., & Chen, H. (2016b). Targeting key data breach services in underground supply chain. In *IEEE international conference on intelligence and security informatics: cybersecurity and big data, ISI 2016* (pp. 322–324).

Samtani, S., & Chen, H. (2016). Using social network analysis to identify key hackers for keylogging tools in hacker forums. In *2016 IEEE conference on intelligence and security informatics (ISI)* (pp. 319–321). IEEE.

Samtani, S., Chinn, R., & Chen, H. (2015). Exploring hacker assets in underground forums. In *2015 IEEE international conference on intelligence and security informatics (ISI)* (pp. 31–36). IEEE.

Samtani, S., Chinn, K., Larson, C., & Chen, H. (2016). AZSecure hacker assets portal: Cyber threat intelligence and malware analysis. In *2016 IEEE conference on intelligence and security informatics (ISI)* (pp. 19–24). IEEE.

Samtani, S., Chinn, R., Chen, H., & Nunamaker, J. F. (2017). Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *Journal of Management Information Systems, 34*(4), 1023–1053.

Shackleford, D. (2016). 2016 security analytics survey. SANS Institute. https://www.sans.org/reading-room/whitepapers/analyst/2016-security-analytics-survey-37467. Accessed 5 June 2017.