Southwestern Business Administration Journal

Volume 16 | Issue 1 Article 1

2017

Leveraging Decision Making in Cyber Security Analysis through Data Cleaning

Chen Zhong
Indiana University Kokomo, Indiana

Hong Liu
Indiana University Kokomo, Indiana

Awny Alnusair
Indiana University Kokomo, Indiana

Keywords*: Cyber Situational Awareness, Data Cleaning, Decision Making, Data Analysis, Cyber Security Analysis*

ABSTRACT

Security Operations Centers (SOCs) have been built in many institutions to support intrusion detection and incident response. A SOC employs various cyber defense technologies to continually monitor and control network traffic. Given the voluminous monitoring data, cyber security analysts need to identify suspicious network activities to detect potential attacks. As the network monitoring data are generated at a rapid speed and contain a lot of noise, analysts are so bounded by tedious and repetitive data triage tasks that they can hardly concentrate on in-depth analysis for further decision making. Therefore, it is critical to employ data cleaning methods in cyber situational awareness. In this paper, we investigate the main characteristics and categories of cyber security data with a special emphasis on its heterogeneous features. We also discuss how cyber analysts attempt to understand the incoming data through the data analytical process. Based on this understanding, this paper discusses five categories of data cleaning methods for heterogeneous data and addresses the main challenges for applying data cleaning in cyber situational awareness. The goal is to create a dataset that contains accurate information for cyber analysts to work with and thus achieving higher levels of data-driven decision making in cyber defense.

INTRODUCTION

Equifax, a major credit reporting agency in the US, announced the largest data breach in history on September 7. According to Equifax, the data breach lasts from May to July, 2017 and has impacted about 143 million customers. The customers' personal identifiable information has been leaked, including names, social security numbers, birth dates, addresses, and driver's license numbers. This incident alerted us again that the risk of cyber security incidents should never be underestimated. Many prominent companies, government organizations and military departments have invested a lot of money to construct Security Operations Centers (SOCs) to combat the increasingly sophisticated cyber attacks (Nathans, 2014)

Technology, people and process are three main components in SOCs. In order to properly monitor network activities, various multiple cyber defense technologies have been employed in SOCs. These include Intrusion Detection/Prevention Systems (IDS/IPS), firewalls, network and system status monitors, traffic monitors, and vulnerability scanners. These defense technologies continually monitor and generate network data. Given such network data from multiple sources, cyber security analysts are playing a critical role in achieving Cyber Situational Awareness (Cyber SA) by performing a series of data analysis. More specifically, their goal is to find answers for the following questions: Whether a network is under attack? How does attacks happen? What will attackers do next? To achieve these goals, analysts usually perform a series of analyses, including data triage, escalation analysis, correlation analysis, threat analysis, incident response and forensic analysis (D'Amico & Whitley, 2008).

However, analysts are faced with multiple challenges in today's SOCs. The network monitoring data includes IDS/IPS alerts, firewall logs, vulnerability scanning reports, reports generated by security information and event management (SIEM) products, etc. Such data sources, which are being continuously generated in large volumes, are very overwhelming for analysts to

process in a timely manner. Compared to a computing system, the information processing capability of human brains is quite limited. Therefore, cyber security analysts are overwhelmed by the influx of network data. Furthermore, the data sources contain many false positives (e.g., false IDS/IPS alerts) which require analysts to apply their expertise and experience knowledge to identify the critical information for further investigation in a timely manner. Besides, the data which is usually arriving from various sources is collected separately so that further aggregation and correlation is necessary through analysts' dedicated manual efforts. At the end of the analysis, analysts need to report the suspicious incidents with the evidence found in the data. Considering the difficulties in data analysis in Cyber SA, it is critical to employ data cleaning methods to support cyber security analysts while performing their daily activities.

The remaining part of the paper is structured as follows. We first identify the main characteristics of the cyber security data and demonstrate some of the data samples in Section 2. After that, we investigate the existing data cleaning methods and point out what data cleaning methods can be adopted in cyber situational awareness.

MASSIVE AND RAPIDLY CHANGING DATA IN CYBER SA

Aimed at gaining Cyber SA, multiple sensors are deployed in a network to monitor various network activities. Bass (Bass, 2000) first pointed out that the data collected from multiple sensors, which are used as input into intrusion detection systems, are **heterogeneous** in nature.
The data consist of "numerous distributed packet sniffers, system log-files, SNMP traps and queries, signature-based ID systems, user profile databases, system messages, threat databases and operator commands" (Bass, 2000). Apart from the data collected by computer/network sensors, there are other important sources of data generated by human intelligence, including the data of SIEM systems (e.g., threat databases), data from external sources (e.g., external attack or threat reports) and data collected from social media (e.g., Facebook and Twitter) (Mahmood & Afzal, 2013). The heterogeneous data vary significantly in types and formats, including quantitative and qualitative data (types), structured, semi-structured, and non-structured data (formats).

We first investigate the heterogeneous features of the data sources in Cyber SA and study the data analytical process of how security analysts make sense of these data sources. Figure 1 illustrates the massive and rapidly changing data sources in Cyber SA. The data is categorized into six different dimensions (which can be further extended). These dimensions are briefly explained below:

- **Sensor**: The network monitoring data can be categorized based on the sensors from which they are collected. The common data sources include alerts of Intrusion Detection Systems (IDS), firewall logs, traffic packages, vulnerability reports, network configurations, server logs, system security reports and anti-virus reports.
- **Data Format**: The format of data can be categorized into structured, semi-structured, and non-structured data.

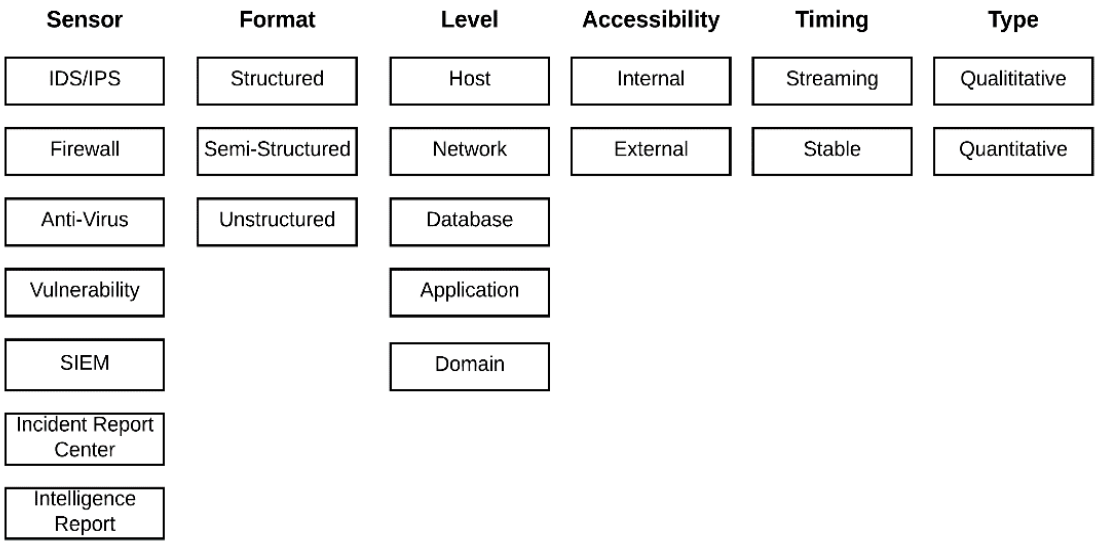

Figure 1: The heterogeneous network monitoring data collected in a SOC.

- **Level of Monitoring Scale**: The data can be divided into the activities of network, host, database, application, and directory.
- **Accessibility**: There are two types of data in this dimension: the internal and external data. The internal data refer to the data which can be directly accessed by the analysts within a SOC, while external data refer to the data outside the SOC which will be available by request only.
- **General Type**: This type of data include both qualitative and quantitative data.
- **Timing**: According to whether the data are time-sensitive or not (i.e., timing), the data can be divided into stable and streaming data. Stable data are relatively fixed and not necessarily changing over time, e.g., network configurations and vulnerability reports. According to whether the data are time-sensitive or not (i.e., timing), the data can be divided into stable and streaming data. Stable data are relatively fixed and not necessarily changing over time, e.g., network configurations and vulnerability reports.

In the following subsection, we describe an example case study that shows the variety of data sources generated by monitoring network activities.

A Case of Multiple Data Sources in Cyber SA

The VAST 2013 Challenge provides a typical case of cyber analysis in an international marketing company (Community, 2013). The organizational network topology of the VAST Challenge is shown in Figure 2.

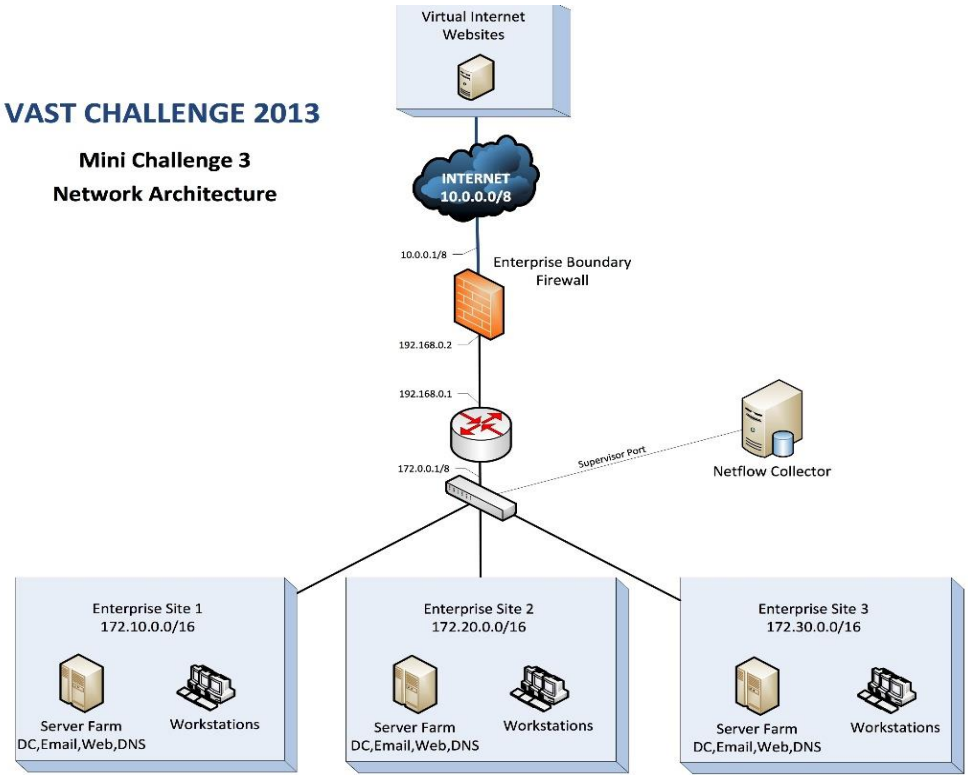

Figure 2: The network topology of the VAST Challenge 2013

The network under investigation has three subnetworks, each of which consists of about 400 employees and its own server farm (including web servers, email servers, Domain Name Servers (DNS) and domain controller servers). All the critical servers are located in the inner network protected by Firewalls. Several sources of data have been collected by special sensors deployed on the network to monitor the network activities, including network flow data, network health and status data, Intrusion Protection System (IPS) data, and some intelligence data.

The network flow data is shown in Figure 3. This flow is captured by the firewall which records all the incoming and outgoing network traffic. This data source is well-structured, including the fields of time/date, IP-layer protocol, source Internet Protocol (IP) and destination IP address, source port, destination port, connection duration, source payload, destination payload, number of packets at source and destination.

TimeSeconds	parsedDate	ipLayerProtocolCode	firstSeenSrcIp	firstSeenDestIp	firstSeenSrcPort	firstSeenDestPort	moreFragments	contFragments	durationSeconds	firstSeenSrcPayloadBytes	firstSeenDestPayloadBytes	firstSeenSrcTotalBytes	firstSeenDestTotalBytes	firstSeenSrcPacketCount	firstSeenDestPacketCount	recordForceOut
1364802616	4/1/13 7:50	UDP	172.20.0.3	172.255.255.255	137	137	0	0	29	600	0	1104	0	12	0	0
1364802621	4/1/13 7:50	UDP	172.10.0.40	172.255.255.255	137	137	0	0	0	100	0	184	0	2	0	0
1364803500	4/1/13 8:05	TCP	172.10.1.17	10.0.0.13	5040	80	0	0	0	176	409	454	633	5	4	0
1364803502	4/1/13 8:05	TCP	172.10.1.9	10.1.0.77	5067	80	0	0	0	176	409	454	633	5	4	0

Figure 3: The network flow data of the VAST Challenge 2013

In the VAST Challenge scenario, a status monitoring technology, called Big Brother, is installed to monitor the status of the network. It captures the latest status of workstations and servers every five minutes. The data is well-structured and has the fields of disk usage, page file usage, number of processes, working load, and physical memory usage, and number of connection made. A portion of such data is demonstrated in Figure 4.

parsedDate	receivedfrom	statusVal	diskUsage Percent	pageFileUsage Percent	numProcs	loadAverage Percent	physical Memory UsagePer	connMade
4/1/13 9:05	172.10.0.40	1	0	0	23	0	9	0
4/1/13 9:05	172.30.0.3	2	0	0	53	3	20	0
4/1/13 9:05	172.30.0.6	1	0	0	40	0	12	0

Figure 4: The network health and status data of the VAST Challenge 2013

In addition, an intrusion protection system (IPS) is also employed to monitor and log network activities. The IPS detects suspicious activities based on a set of rules predefined by domain experts and it blocks or prevents the malicious ones. The log/alerts generated by the IPS is well-structured, containing the field of the date/time, priority, operation, alert message, protocol, source IP address, destination IP address, source port, destination port, destination service, direction, flags and command. While Figure 3 shows a portion of the network flow data. Figure 5 shows a sample of the IPS alerts.

dateTime	priority	operation	messageCod	protocol	SrcIp	destIp	srcPort	destPort	destService	direction	flags	command
4/10/13 7:02	Info	Built	ASA-6-30201	TCP	172.10.2.35	10.1.0.75	2507	80	http	outbound	(empty)	(empty)
4/10/13 7:02	Info	Teardown	ASA-6-30201	TCP	172.30.1.104	10.0.0.14	2651	80	http	outbound	TCP FINs	(empty)
4/10/13 7:02	Info	Teardown	ASA-6-30201	TCP	172.10.1.246	10.1.0.77	2504	80	http	outbound	TCP FINs	(empty)
4/10/13 7:02	Info	Built	ASA-6-30201	TCP	172.10.1.138	10.1.0.100	1893	80	http	outbound	(empty)	(empty)
4/10/13 7:02	Info	Teardown	ASA-6-30201	TCP	172.10.1.203	10.1.0.77	2506	80	http	outbound	TCP FINs	(empty)
4/10/13 7:02	Info	Teardown	ASA-6-30201	TCP	172.10.1.64	10.0.0.7	2260	80	http	outbound	TCP FINs	(empty)
4/10/13 7:02	Info	Built	ASA-6-30201	TCP	172.30.2.109	10.0.0.8	2673	80	http	outbound	(empty)	(empty)
4/10/13 7:02	Info	Built	ASA-6-30201	TCP	172.10.2.39	10.0.0.7	2509	80	http	outbound	(empty)	(empty)
4/10/13 7:02	Info	Built	ASA-6-30201	TCP	172.10.1.39	10.1.0.75	2261	80	http	outbound	(empty)	(empty)

Figure 5: The IPS alert of the VAST Challenge 2013

Most of the streaming data are structured but there may be different formats across various sources. The common ground of streaming data sources is that they can be viewed as a sequence of data entries in temporal order considering that they are collected over time. A data entry can be an alert, a report or a log item. The Cyber SA raw data report the network events perceived by the monitoring sensors (including human intelligence). From the view point of Cyber SA data analysis, we define a network event as a unit of analysis. A network event can be identified as one or more data entries from different data sources.

Data Analysis in Cyber SA

Given the multiple data sources collected in a SOC, cyber analysts need to perform a series of data analysis to make sense of the data to achieve Cyber SA. Data triage is the first and fundamental step of the analysis (Zhong, Yen, Liu, & Erbacher, 2016). During data triage, analysts rule out the false alerts and identify the suspicious data that worth further investigation from the multiple data sources in a timely manner. To make a quick decision, an analyst need to go through a very

complicated cognitive process as demonstrated in Figure 6. This process involves actions, observations and hypotheses (Yen, Erbacher, Zhong, & Liu, 2014). As such, the analyst conducts data filtering or searching operations to narrow down his/her search scope; these operations may result in the observations of suspicious evidence; through comprehending the observations, the analyst may generate some hypotheses about the potential attack incidents, which in turn may trigger following actions for further investigation. Through these iterative actions, observations and hypotheses, the analyst finally make his/her decision regarding what suspicious incidents exist in the current situation.

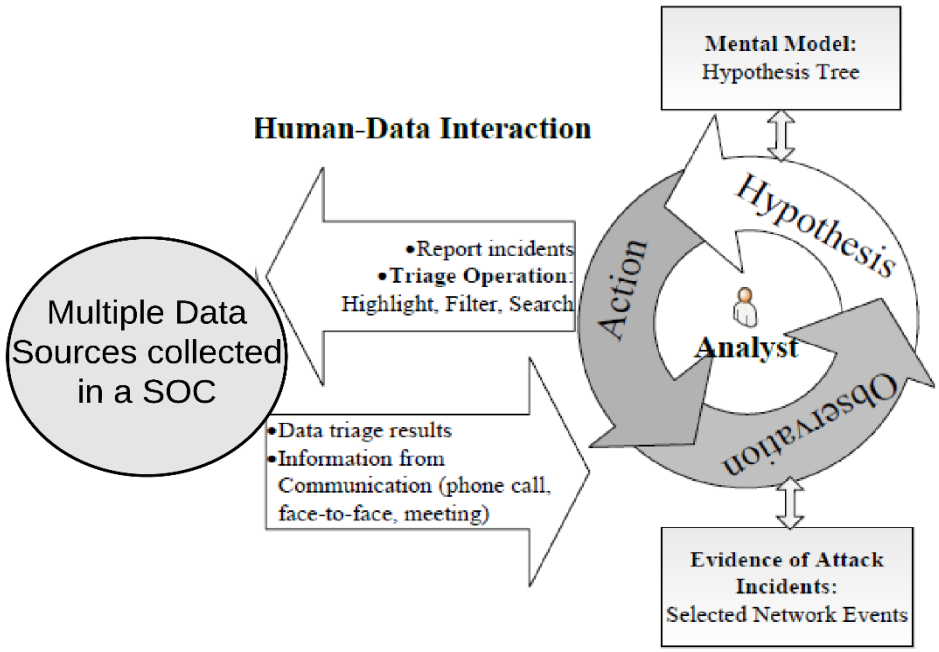

Figure 6: The data analytical process of cyber security analysts in a SOC

Due to the large volumes of Cyber SA data, the need for data cleaning methods that can be used to smooth out the noise in network monitoring data is evident. Additionally, since Cyber SA data is coming from different sources, an effective data cleaning approach would consolidate multiple sources of cyber data sources to a single stable and usable source. In the following section, we provide a brief overview of such cleaning methods that can help in reducing the time and efforts needed by cyber analysts in identifying cyber threats in a timely fashion.

DATA CLEANING METHODS IN CYBER SA

With the development of Cyber SA techniques, the management of Cyber SA data become far more complex. A number of problems attribute to the difficulty of managing such kind of data. These problems include:

- Incorrect data: examples of such incorrect data can be errors in data entry or typos
- Scattered data: data are distributed into different systems. Many processes involve interactions among multiple systems.

- Heterogeneous data: data is captured using varied data models including object-oriented models, hierarchical data models, XML data models, and relational models.
- Quality of data: quality refers to how well the data is fit for intended uses, and it is at varying levels, with some data being of high quality and some data containing lots of errors. In general, data quality is defined as a set of quality criteria (completeness, accuracy, timeliness, consistency).
- Autonomous data: data are derived from nonrelated systems and defined by vary schemas.
- Sensitive data: data contain personal information of individuals or data are classified information related to an organization.

With those problems, the data quality problem, obtaining high quality data, is not an easy task. Nowadays, most data cleaning methods could not find out a satisfied solution other than focusing on specific formulation of the problem. In real-world, the complex nature of data makes data quality problem being more evident in any data-related system since useless or even harmful results could be given due to errors propagation and exemplification. Therefore, data cleaning becomes more and more important for data mining and data analysis.

Classifications of Data Error

Data cleaning is a term with no common clear definition because data cleaning targets errors in data. Therefore, data cleaning is based on how to define the error. Whether it is an error or not is highly application specific. Therefore, before cleaning errors in data, it is important to clarify the variety of errors that may be present within the data along with how to deal with them. In the subsections below, we provide a review of the current classifications of data errors.

Rahm and Do's Classification of Data Error

A classification of data quality problems with different levels in data sources is provided in (Rahm & Do, 2000). As shown in Figure 7, data errors could be divided into single data source and multiple data sources problems, and under each of them, it is divided into schema level problems and instance-level problems.

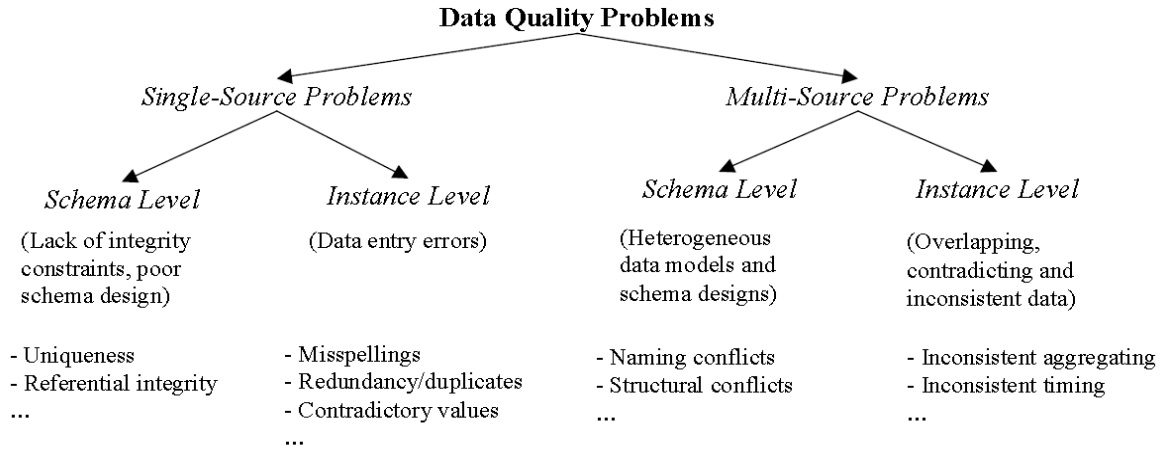

Figure 7: Classification of data errors in data sources

At the schema level, the data quality problem can be improved by designing schema, translating schema and integrating schema. At instance level, they include data entry errors and inconsistent data which are not accessible at the schema level. For single-source problems, the problems cause by lack of appropriate model-specific or application-specific integrity constraints which are defined as schema-level data quality problems. For multiple data sources problems, data sources are designed and maintained independently, and later when these data sources are integrated, data quality problems become more complicated. The main problems at the schema-level are naming conflicts and structural conflicts. Table 1 shows the data error types introduced at the schema level.

Table 1: Data error types from Rahm and Do

Data Error Type	Annotation
Uniqueness violation	Uniqueness for a field
Referential integrity violation	Referenced value not defined
Misspellings	Typo
Duplicated records	Same records represent multiple times
Contradicting records	Same unit described by different values
Naming conflicts	Two fields has same name
Structural conflicts	Two relations has same structure
Inconsistent aggregating	Different aggregating
Inconsistent time	Different times
Illegal values	values outside of domain range
Violated attribute dependences	Ex: age = current date – birth data
Missing values	Unavailable values
Cryptic values, Abbreviations	Vague information
Embedded values	Multiple values entered in on field
Misfielded values	Value in the other field
Word transpositions	Usually in a free-form field
Wrong references	Referenced value is defined but wrong

Muller and Freyta's Data Anomalies

Müller and Freyta classified data anomalies into syntactical, semantic, and coverage anomalies (Müller & Freytag, 2005). Syntactical anomalies describe characteristics focusing on the format and values which are used to represent entities. Semantic anomalies hinder the data collection from being a comprehensive and non-redundant representation. Coverage anomalies decrease the amount of entities and properties of entity from datasets which are represented in the data collection. Each category has its own data error types. All errors types are shown in Table 2.

Table 2: Data anomalies from Muller and Freytag

Syntactical Anomalies	Lexical error
	Domain format error
	Irregularities error
Semantic Anomalies	Integrity constraint violations error
	Contradictions
	Duplicate records
	Invalid tuple
Coverage Anomalies	Missing values
	Missing tuple

- Lexical errors identify discrepancies between the structure of the data items and the specified format.
- Domain format errors specify errors where the given value for an attribute *A* does not conform with the specified domain format $G(dom(A))$.
- Irregularities are concerned with the non-uniform use of values, units and abbreviations.
- Integrity constraint violations describe tuples (or sets of tuples) that do not satisfy one or more of the integrity constraints in Γ.
- Duplicates are two or more tuples representing the same entity.
- Invalid tuples represent the tuples that do not display anomalies, but they do not represent the true entities.
- Missing values are the result of omissions while collecting data.
- Missing tuples result from omissions of complete entities existent in the dataset that are not represented by tuples in the data collection.

Data Cleaning Techniques

Most data mining, decision making, and data analysis technologies focus on the analysis process to get the results, and usually assume that the data has already been gathered, cleaned, explored, and understood before it is used. However, the true challenge is to create a dataset that contains

relevant and accurate information for cyber analysts to work with. This challenge makes the data cleaning process significantly important. Data cleaning refers to the process of identifying and repairing damaged or inaccurate data. Nowadays, several commercial tools such as Informatics, Oracle Warehouse Builder are able to process erroneous and abnormal data. These tools are often referred to as ETL (Extraction, Transformation and Loading) tools (Rahm & Do, 2000). In addition, for the past years, several of data cleaning techniques have been proposed such as noise removal, outlier detection, entity resolution and imputation, with recent efforts examining integrity constraints, such as functional/inclusion dependencies and their extensions, and Bayes Wipe (De, Hu, Chen, & Kambhampati, 2014). Table 3 shows the five categories of data cleaning techniques.

Table 3. Categories of Data Cleaning Techniques

Categories	**Description**	**Related Work**
Error detection	Error detection refers to the problem of finding patterns in data that do not conform to the expected value.	(Chambers, Cleveland, Kleiner, & Tukey, 1983) (Tukey, 1977) (Dixon & Massey, 1950) (Barnett & Lewis, 1994) (Rosner, 1975)
Functional dependency	Functional dependency is a set of fields which can determine another field.	(Chomicki, 2007) (Bertossi, Hunter, & Schaub, 2005) (Bohannon, Fan, Flaster, & Rastogi, 2005) (Lopatenko & Bravo, 2007) (Cong, Fan, Geerts, Jia, & Ma, 2007)
Missing/Incomplete data	Missing data is a constant feature of massive data, where individual cells, columns, rows or entire sections of the data may be missing.	(Little & Rubin, 2014) (Soley-Bori, 2013) (Allison, 2012) (McKnight, McKnight, Sidani, & Figueredo, 2007)
String/field/record matching	Matching is a method to determine that two things are similar, and compute how similar they are.	(Brizan & Tansel, 2006) (Cohen & Richman, 2002) (Gravano, Ipeirotis, Koudas, & Srivastava, 2003) (Bernstein, Madhavan, & Rahm, 2011)

Duplicate elimination	The process of combining multiple entries into a single entry is called duplicate elimination.	(Bilenko & Mooney, 2003) (Sarawagi & Bhamidipaty, 2002) (Cong, Fan, Geerts, Jia, & Ma, 2007)

A variety of string matching techniques have been proposed for performing approximate string matching, data duplication, entity linking, and fuzzy string searching. Data imputation uses relational learning to get the characteristics of the attributes relationships. Then, the learnt model is used to impute the missing values. This approach requires a priori setting of the relationships among the attributes to build the learning model. The main challenges for these techniques are the scalability for large dataset to be modeled with all correlations among the dataset and most data imputation techniques are limited to numerical and categorical attributes. Apparently, techniques that are not scalable enough are the least effective in the context of Cyber SA unless a prior filtering and processing of the data using other techniques has been applied.

There are two main steps in constraint-based data repairing process. Firstly, identify a set of constraints by domain expert which should be followed by all related data. Secondly, construct a new consistent database that minimally differs from the original database by using the constraints. Most constrain based data repairing works use functional dependencies and their variants conditional functional dependency. One drawback of these techniques is that all constraints should be specified by domain experts, which may be an expensive and time consuming manual process.

CONCLUSION

Due to the increasing levels of cyber-attacks and data breaches, cyber security analysts in Security Operations Centers (SOCs) are faced with massive and rapidly changing data that they have to analyze to combat potential attacks. In order to promote and facilitate cyber analysis, it is critical to employ data cleaning and data analysis methods that can remove noise from the large amounts of cyber SA data. This paper addresses the main characteristics of cyber SA data and how data cleaning methods can be applied to support cyber analysis and facilitate higher levels of decision making. The paper is meant to bring awareness to the need of such data cleaning and analysis methods in Cyber SA and provides concrete directions to the importance of conducting further research on data-driven decision making in cyber defense.

To this end, we have identified the main problems that attribute to increasing the difficulty of managing the complexity of data in Cyber SA. We focused the discussion on how to clean errors in such data sources in order to create a cleaned dataset with accurate and relevant information for cyber analysts to work with. Due to the massive amounts of Cyber SA data, we have also shown that a multitude of data cleaning techniques are needed to be utilized effectively to achieve higher levels of data-driven decision making about potential security threats.

ACKNOWLEDGEMENT

The authors of this manuscript were supported in part by the IUK Summer Faculty Fellowship Grant.

REFERENCES

Allison, P. D. (2012). *Handling missing data by maximum likelihood* (Vol. 23). Statistical Horizons Haverford, PA, USA.

Barnett, V., & Lewis, T. ,. (1994). *Outliers in statistical data* (Vol. 1). New York: Wiley.

Bass, T. (2000). Intrusion detection systems and multisensor data fusion. *Communications of the ACM, 43*(4), 99--105.

Bernstein, P. A., Madhavan, J., & Rahm, E. (2011). Generic schema matching, ten years later. *VLDB Endowment, 4*(11), 695--701.

Bertossi, L., Hunter, A., & Schaub, T. (2005). *Inconsistency Tolerance* (Vol. 3300). Springer. doi:https://doi.org/10.1007/b104925

Bilenko, M., & Mooney, R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 39--48). ACM.

Bohannon, P., Fan, W., Flaster, M., & Rastogi, R. (2005). A cost-based model and effective heuristic for repairing constraints by value modification. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 143--154). ACM.

Brizan, D. G., & Tansel, A. U. (2006). A. survey of entity resolution and record linkage methodologies. *Communications of the IIMA, 6*(3).

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis* (Vol. 1). Belmont, CA: Wadsworth.

Chomicki, J. (2007). Consistent query answering: Five easy pieces. *ICDT. 4353*, pp. 1--17. Springer.

Cohen, W. W., & Richman, J. (2002). Learning to match and cluster large high-dimensional data sets for data integration. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 475--480). ACM.

Community, V. A. (2013). *VAST Challenge 2013 Data Set*. Retrieved from http://vacommunity.org/VAST+Challenge+2013%3A+Mini-Challenge+3

Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007). Improving data quality: Consistency and accuracy. *Proceedings of the 33rd international conference on Very large data bases* (pp. 315--326). VLDB Endowment.

D'Amico, A., & Whitley, K. (2008). The real work of computer network defense analysts. *VizSEC*, pp. 19--37.

De, S., Hu, Y., Chen, Y., & Kambhampati, S. (2014). BayesWipe: A multimodal system for data cleaning and consistent query answering on structured bigdata. *2014 IEEE International Conference on Big Data* (pp. 15--24). IEEE Computer Society.

Dixon, W. J., & Massey, F. J. (1950). *Introduction To Statistical Analsis.* New York: McGraw-Hill Book Company, Inc.

Gravano, L., Ipeirotis, P. G., Koudas, N., & Srivastava, D. (2003). Text joins for data cleansing and integration in an rdbms. *Proceedings. 19th International Conference on Data Engineering* (pp. 729--731). IEEE Computer Society.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data.* John Wiley & Sons.

Lopatenko, A., & Bravo, L. (2007). Efficient approximation algorithms for repairing inconsistent databases. *IEEE 23rd International Conference on Data Engineering* (pp. 216--225). IEEE Computer Society.

Mahmood, T., & Afzal, U. (2013). Security Analytics: Big Data Analytics for cybersecurity: A review of trends, techniques and tools. *2nd national conference on Information assurance (ncia)* (pp. 129--134). IEEE Computer Society.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* Guilford Press.

Müller, H., & Freytag, J. C. (2005). Problems, methods, and challenges in comprehensive data cleansing. *Professoren des Inst. Für Informatik*.

Nathans, D. (2014). *Designing and Building Security Operations Center.* Waltham, MA, USA: Syngress, Elsevier.

Rahm, E. a. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3--13.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3--13.

Rosner, B. (1975). On the detection of many outliers. *Technometrics, 17*(2), 221--227.

Sarawagi, S., & Bhamidipaty, A. (2002). Interactive deduplication using active learning. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 269--278). ACM.

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University*.

Tukey, J. W. (1977). *Exploratory data analysis.* Pearson.

Yen, J., Erbacher, R. F., Zhong, C., & Liu, P. (2014). Cognitive process. *Cyber Defense and Situational Awareness*, 119-144.

Zhong, C., Yen, J., Liu, P., & Erbacher, R. F. (2016). Automate Cybersecurity Data Triage by Leveraging Human Analysts' Cognitive Process. *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)* (pp. 357-363). IEEE Computer Society.