Improved Cyber Threat Indicator Sharing by Scoring Privacy Risk

Daniel M. Best[1], Jaspreet Bhatia[2], Elena S. Peterson[1], Travis D. Breaux[2]
[1]Pacific Northwest National Laboratory, Richland, Washington, United States
[2]Carnegie Mellon University, Pittsburgh, Pennsylvania, United States
daniel.best@pnnl.gov, jbhatia@cmu.edu, elena@pnnl.gov, breaux@cmu.edu

***Abstract*—Information security can benefit from real-time cyber threat indicator sharing, in which companies and government agencies share their knowledge of emerging cyber-attacks to benefit their sector and society at large. As attacks become increasingly sophisticated by exploiting behavioral dimensions of human computer operators, there is an increased risk to systems that store personal information. In addition, risk increases as individuals blur the boundaries between workplace and home computing (e.g., using workplace computers for personal reasons). This paper describes an architecture to leverage individual perceptions of privacy risk to compute privacy risk scores over cyber threat indicator data. Unlike security risk, which is a risk to a particular system, privacy risk concerns an individual's personal information being accessed and exploited. The architecture integrates tools to extract information entities from textual threat reports expressed in the STIX format and privacy risk estimates computed using factorial vignettes to survey individual risk perceptions. The architecture aims to optimize for scalability and adaptability to achieve real-time risk scoring.**

Keywords—cyber security, threat indicators, information sharing, privacy, risk

I. Introduction

The proposed Cyber Intelligence Sharing and Protection Act "Directs the federal government to provide for the real-time sharing of actionable, situational cyber threat information between all designated federal cyber operations centers to enable integrated actions to protect, prevent, mitigate, respond to, and recover from cyber incidents." While many technical challenges to providing real-time sharing of data exist, one critical aspect is protecting individuals' and entities' privacy in that process. Several efforts to protect personally identifiable information (PII) and other privacy-related data have resulted in standards adopted by many. One example is the *Guide to Protecting the Confidentiality of Personally Identifiable Information* [1]. However this standard does not take into account cyber-related information that may also be identifiable (e.g., IP address). Research into which pieces of computer-related data are considered "private" is needed. Once risky data is identified, one approach is to obfuscate the data in some way before sharing it. One drawback to this is approach is that some of the obfuscated data may be critical to providing a meaningful response. Although we need to share data at all levels, we also need to understand the risk of doing so. Defining that risk is not straightforward and is based on a number of factors. In this paper, we will discuss our research on defining the risk of sharing cyber-related data and identifying what that data is. We will also present an algorithm for creating a unified score of privacy risk based on our research.

II. Related Work

A. Definitions of Privacy

Privacy has many definitions that characterize how technology influences individual self-preservation: Warren and Brandeis's the "right to be let alone" [2], Westin's four states (solitude, intimacy, anonymity, and reserve) [3], and Nissenbaum's contextual integrity, or data confidentiality [4]. Solove summarized a taxonomy of privacy intrusions from legal proceedings, including data breach, decisional interference, and surveillance, among others [5]. Calo distinguishes between objective and subjective harm: objective harm is external to the individual, such as data breach and decisional interference wherein the individual knows that the harm has occurred, whereas subjective harm is the "perception of unwanted observation" that can arise due to the mere suggestion of surveillance [6]. In cyber threat indicator sharing, personal information may be shared to investigate a data breach or other objective harm; meanwhile, the potential act of sharing such information can introduce the subjective harm of being observed.

B. Privacy Risk Perception

Risk has been studied in marketing, psychology, and economics. In marketing, risk is a choice among multiple options, which are values based on the likelihood and desirability of the consequences of the choice [7]. Starr first proposed that risk preferences could be *revealed* from economic data, in which both effect-likelihood and magnitude were previously measured (e.g., the acceptable risk of death due to skiing, smoking, or motor vehicle accidents) [8]. In psychology, Fischhoff et al. note that revealed preferences assume that past behavior is a predictor of present-day preferences, which cannot be applied to situations where technological risk or personal attitudes are changing [9]. The psychometric paradigm of *perceived* risk emerged to address these limitations by measuring personal attitudes about risks and benefit [10]. In Kahneman and Tversky's prospect theory, loss aversion explains that people are more sensitive to

978-1-5090-6356-7/17/$31.00 ©2017 IEEE

potential harm than potential benefit when they weigh the risk and benefit of a decision [11]. In contrast, Knightian economists argue that subjective estimates based on partial knowledge, which includes perceived risk, are measures of uncertainty and not measures of risk [12]. In this paper, *perceived privacy risk* is measured by an individual's willingness to share personal data in the presence of potential benefits and privacy loss. In other words, when people share information, they are accepting the risk.

III. Scoring Architecture Overview

Architectural design for the privacy implementation focused on two key elements: scalability and adaptability. Scalability was a major consideration due to the likelihood that the implementation would be leveraged at enterprise scale for varying sizes of organizations. Another factor in the design is the ability to change components as continued research yields new insites on how to improve the models. Given these two elements, we designed a solution that would allow for a small initial implementation that could expand or decrease with components that can be adapted, replaced, or augmented as the solution matures.

To achieve scalability in the solution, we chose technologies that a) work well together and b) are designed to scale out with the addition of hardware or systems. The primary technologies chosen were Apache NiFi[1] and Apache Kafka[2], both of which scale and work well together. NiFi provides the processing workflow, while Kafka provides a message queue that allows producers or consumers of documents to interact with the solution.

The input and output of the privacy capability is designed to share cyber-security incident records. Incident records include details about a cyber-security event that an organization experienced or wishes to notify others about. Several incident record formats have been developed over the years, such as Indicators of Compromise and Structured Threat Information eXpression (STIX), which aim to provide structure to information sharing. Our initial data set leverages the STIX format due to its increasing use in governement organizations for data sharing.

The workflow for the process can change, new techniques can be introduced, or additional measures may be implemented in the future. Apache NiFi is a perfect solution in this instance not only because it scales (as mentioned previously) but also because the workflow is straightforward to implement and new processing units can be introduced with minimal effort. For our solution, we implement major components (recognizers and scoring) as NiFi processors. The workflow is then built around moving data from each processor and sending the data to a final area for consumption (as seen in Figure 1).

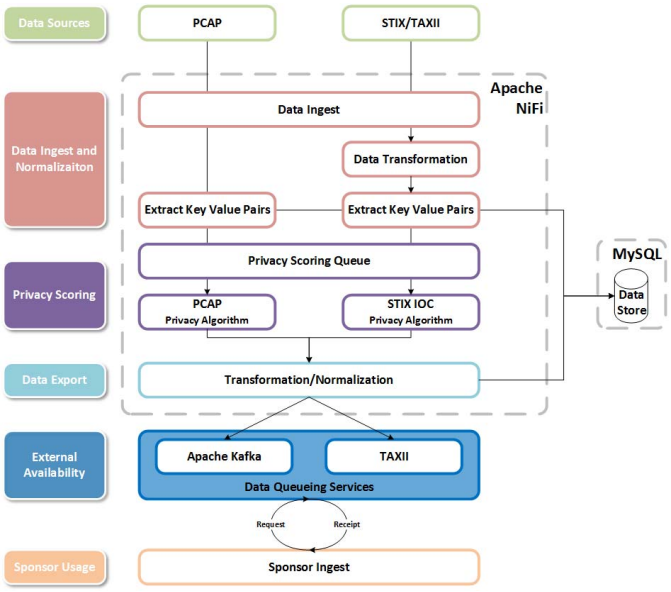

Fig. 1. Architecture design concept

Entry into the privacy capability starts with a NiFi processor to retrieve data or place data on the Kafka queue. The retrieval process can be adapted to interact with various data producers. In this case, due to the focus on STIX data, the retrieval is designed to pull from a TAXII server, querying for up-to-date incident records. The Kafka queue is provided as an additional means that could provide a streaming (or near-streaming) solution if a data producer publishes to the queue.

Once data is in the privacy solution, the first step is to transform the data into a format that can be easily handled by NiFi. While this step could be skipped, doing it allows for data to pass to processors with ease. After transformation, the recognizer processor evaluates the data. The recognizer processor identifies key value pairs of information that could have privacy concerns. Privacy concerns are influenced based on the implementation. For the initial implementation, the focus was on data likely to be found in incident records that cyber subject matter experts reported to be concerning.

Next, the output from the recognizer processor is sent to the privacy scoring processor where the values are evaluated for their potential impact to privacy. Finally, the privacy scoring processor produces a privacy score. Currently, this score is simply reported; however, the intent is to have another processor that adds the privacy score to an appropriate location (for STIX, this location may be the markings section).

Once the data and score are ready to be made available for consumers, they are put onto the Kafka queue or made available in a data-type appropriate means (such as a TAXII server for STIX).

While the data focus for this first iteration has been on the STIX document format, the design and flow of data have been developed to allow for additional data types. In the future, packet capture or other cyber-security-related data may be added to the capability. Having data transformed on ingress

[1] Apache NiFi: An Easy to Use, Powerful, and Reliable System to Process and Distribute Data. https://nifi.apache.org/

[2] Apache Kafka: A distributed steaming platform. https://kafka.apache.org/

978-1-5090-6356-7/17/$31.00 ©2017 IEEE

and egress allows the core components to remain the same, with new processors for transformation to be implemented.

IV. Case Study Scenarios

The U.S. Department of Health and Human Services is required by the HITECH Act[3] to report the number and scope of data breaches affecting health information governed by the Health Insurance Portability and Accountability Act in the United States. In 2016, 329 data breaches were reported that affected an estimated 16.7 million individuals. In reporting these breaches, the details of specific records, including personal health information, may be exposed. Therefore, we prototyped and evaluated the approach described in Section III, based on information types used in a healthcare database breach scenario. The information types are:

- Credit card number
- Date of birth
- Diseases
- Driver's license information
- Email address
- Full name
- Home address
- Home state
- Medical procedure
- Phone number
- Social security number

The information types presented above present different degrees of technical challenge to automatically recognize the information type. For example, credit card numbers and birth dates have well-defined formats, which vary within largely predictable ways. Alternatively, disease names do not follow such formats and would require a lexicon to detect.

V. Information Type Recognizer

To allow the privacy risk scoring component to operate, the potential privacy information must first be extracted from the STIX documents using entity extraction. As discussed in Section IV, the entities of interest are common PII and other identifiable information that could have sharing sensitives.

To choose which entities to focus on for extraction, we used three criteria: entities that had a) known PII concerns, b) had possible sharing concerns as identified in user studies, or c) had sharing concerns as identified by the National Institute of Standards & Technology or other governing bodies.

For the proof-of-concept pipeline, we built a simple regular expression entity extraction method as part of the NiFi processing pipeline. While other techniques for finding entities in a document corpus exist, the regular expression implementation allowed us to focus on the privacy scoring implementation and overall pipeline solution.

Regular expressions developed for the initial implementation include credit card number, date of birth, email address, home state, IP address, media access control address, phone number, social security number, and websites visited. To accurately match patterns of the chosen features, we add as much context as possible to reduce false positives. For

[3] https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

Please rate your willingness to share your information below with the Federal government for the purpose of **$DP**, given the following risk.

Risk: In the last 6 months, while using this website, only **$RL** experienced a privacy violation due to **$PH**.

When choosing your rating for the information types below, consider the **$CT**, purpose and the risk, above.

	Extremely Willing	Very Willing	Willing	Somewhat Willing	Somewhat Unwilling	...
$DT	○	○	○	○	○	

Fig. 2. Template used for vignette generation: fields with dollar sign ($) are replaced with values selected from Table I

VI. Privacy Risk Scoring

Our approach to privacy risk scoring is based on data collected using factorial vingette surveys and multi-level modeling [13]. A factorial vignette is a scenario description with factors and corresponding levels. The survey design extends the design introduced by Bhatia et al. [14] to measure the interactions between five independent variables—the *computer type* ($CT) where the cyber incident occurs, the *data type* ($DT) shared with the U.S. federal government, the *data purpose* ($DP) for which data is shared, the *risk likelihood* ($RL) of a privacy violation, and the *privacy harm* ($PH)—and their combined effect on a dependent variable, the employee's *willingness to share* ($WtS) his or her data with the U.S. federal government [15]. The factorial vignettes are presented using the template shown in Fig. 2: the independent variables $CT and $RL are between-subject factors, so participants only see one level of these two factors, and the variables $DT, $DP, and $PH are within-subject factors, so participants see all combinations of these factors. The $DT factor levels, with the exception of age range, match the data types in the incident reporting survey design.

When participants see the vignette, they rate their willingness to share their data with the government on an eight-point, bipolar semantic scale, labeled *Extremely Willing, Very Willing, Willing, Somewhat Willing, Somewhat Unwilling, Unwilling, Very Unwilling* and *Extremely Unwilling*.

Before the vignettes, we presented a pre-test that asks participants to rank order and score $DP based on their benefit to society. Fischhoff et al. argue that individuals should be presented with enumerable benefits before judging the risk of a specific event [9]. We asked participants to rank order the risk likelihood levels from nearest to farthest proximity as an attention test. Finally, we asked participants whether they store their personal data on their workplace computer. Each of these three questions aims to sensitize participants to the factorial vignette levels in Table I, especially the between-subject factors, before asking participants to report their willingness to share.

We analyzed the survey results using multilevel modeling, which is a statistical regression model with parameters that account for multiple levels in datasets and limits the biased covariance estimates by assigning a random intercept for each subject [16]. In our study, the main dependent variable of interest is willingness to share, labeled $WtS. As can be seen in Table I, the fixed independent variables, which are within-

978-1-5090-6356-7/17/$31.00 ©2017 IEEE

TABLE I. VIGNETTE FACTORS AND THEIR LEVELS

Factors	Factor Levels	
Computer Type ($CT)	personal smart phone	
	workplace computer	
Data Purpose ($DP)	investigating intellectual property and trade secrets	
	investigating economic harm, fraud or identity theft	
	investigating imminent threat of death or harm to an individual, including children	
	investigating terrorism	
Risk Likelihood ($RL)	only one person in your family	
	only one person in your workplace	
	only one person in your city	
	only one person in your state	
	only one person in your country	
Privacy Harm ($PH)	a privacy violation due to government surveillance	
Data Type ($DT)	Group 1	
	age range	sensor data
	usernames & passwords	network information
	device information	IP address & domain names
	device ID	packet data
	UDID / IMEI	MAC address
	Group 2	
	age range	registry information
	OS information	running processes
	OS type & version	application information
	memory data	application session data
	temporary files	
	Group 3	
	age range	contact information
	emails	keyword searches
	chat history	keylogging data
	browser history	video & image files
	websites visited	

subject factors, are $DT (with 28 levels), $DP (with 4 levels), and $PH (with one level, which is called a blank dimension). For the within-subject design, subject-to-subject variability is accounted for by using a random effect variable $PID, which is unique to each participant.

Each participant sees all 4x3x1=12 combinations of our three within-subject factors (the 28 data type levels are divided into three groups). The analysis accounts for dependencies in the repeated measures, calculates the coefficients (weights) for each explanatory independent variable, and tests for interactions. We test the multilevel models' significance using standard likelihood ratio test: we fit the regression model of interest; we fit a null model that excludes the independent variables used in the first model; we compute the likelihood ratio; and then, we report the chi-square, p-value, and degrees of freedom [16]. We performed *a priori* power analysis using G*Power [17] to test for the required sample size for repeated measures analysis of variance. The power analysis estimate is at least five participants per combination of the within-subject factors to achieve 95% power and a medium effect size.

We recruited participants from Amazon Mechanical Turk who are located in the United States, who have a 97% approval rating or higher and 5000 or more Human Intelligence Tasks (HIT) completed. We piloted the survey and found the mean time to complete was approximately 20 minutes; thus, we allowed 45 minutes for recruited participants to complete the survey. We paid $6 per participant, and we ran the survey using SurveyGizmo.

VII. CONCLUSION AND FUTURE WORK

The area of privacy and privacy risk for cyber security is a growing field of research and application. Although some basic guidelines for protecting privacy related data exist, they are not enough to truly support data sharing and integration at the level that is needed for true cyber security. The work presented here is just a start to the type of basic research needed in the area of identifying privacy and the risk of sharing private data. The key is to scientifically assess privacy concerns in the context of cyber security and not just in areas of medical records, etc. Different factors in identifying and sharing cyber related data have an impact on the public and private entities involved.

Once privacy risks have been assessed for a particular domain, in our case cyber security, a method is needed to apply the privacy risk understanding with minimal impact to sharing of data. The prototype architecture presented here attempts to do just that by automatically scoring and tagging documents to enable informed decisions about sharing quickly so that the flow of information is not impeded unduly. Solutions such as these will be a key factor in enabling real-time information sharing.

We intend to continue this research and development and provide this algorithm and tool to agencies who are tasked with sharing cyber data. We will expand our research as well. Specifically, we will look at estimating privacy risk for data composition with the idea of asking, "Are people willing to share 'what they do' if 'who they are' is protected?" We will continue to study the effects of different purposes on risk evaluation, estimating risk for under-represneted populations, estimating risk from data analysis, and determining how we can interact with other mitigation strategies to increase privacy (lower the risk of sharing).

Architecturally, we have two areas where we will continue to pursue improvements. The first is the value recognizer module. While the current solution is reasonable for the current use case, efficiency and accuracy can still be improved. Additionally, drawing out additional context around features will enable a richer input into the privacy scoring mechanism. The second area is updating the privacy module based on the research being conducted to understand privacy risks and features.

ACKNOWLEDGMENT

We thank Liora Friedberg and Daniel Smullen for their help in conducting the factorial vignette surveys. This research was funded by the National Science Foundation, NSF Award #1330596. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.

REFERENCES

[1] E. Mcallister,T. Grance, K.A. Scarfone. *Guide to Protecting the Confidentiality of Personally Identifiable Information*. NIST Special

Publication 800-122. National Institute of Standards & Technology. 2010.

[2] S. Warren, L. Brandeis. "The right to privacy," *Harvard Law Review,* 4(5): 1890.

[3] A. Westin. *Privacy and Freedom*, The Bodley Head Ltd., 1970.

[4] H. Nissenbaum. *Privacy in Context*, Stanford Law Books, 2009.

[5] D.J. Solove, "A taxonomy of privacy," *University of Pennsylvania Law Review*, no. 3, 2006.

[6] M.R. Calo. "The boundaries of privacy harm," *Indiana Law Journal,* 86(3): 1131-1162.

[7] R. A. Bauer, "Consumer behavior as risk taking," in Risk Taking and Information Handling in Consumer Behavior, 1960, pp. 389-398.

[8] C. Starr, "Social benefit versus technological risk," Science 165(3899): 1232, 1969.

[9] B. Fischhoff, P. Slovic, S. Lichtenstein, S. Read, and B. Combs, "How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits," Policy Sci., 9(2): 127-152, Apr. 1978.

[10] P. Slovic, The perception of risk. 2000.

[11] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decision under risk," Econometrica, 47(2): 263, 1979.

[12] F. Knight, "Risk, Uncertainty, and Profit," Hart Schaffner Marx Prize essays, vol. XXXI, 1921.

[13] K. Auspurg and T. Hinz, *Factorial Survey Experiments*. 2014.

[14] J. Bhatia, T.D. Breaux, J.R. Reidenberg, T.B. Norton. "A theory of vagueness and privacy risk perception." *In Press: 24th IEEE International Requirements Engineering. Conference, 2016.*

[15] J. Bhatia, T.D. Breaux, L. Friedberg, H. Hibshi, D. Smullen. "Privacy risk in cybersecurity data sharing." In Submission: *3rd ACM Workshop on Information Sharing & Collaborative Security*, 2016.

[16] A. Gelman, J. Hill, "Data analysis using regression and multilevel/hierarchical models," *Policy Anal.*, 2007.

[17] F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, "G*Power 3," *Behav. Res. Methods*, 39(2): 175-91, 2007.

978-1-5090-6356-7/17/$31.00 ©2017 IEEE