*Accepted to the 3rd ACM Workshop on Information Sharing and Collaborative Security. Updated 4 November 2016.*

# Privacy Risk in Cybersecurity Data Sharing

Jaspreet Bhatia[1], Travis D. Breaux[1], Liora Friedberg[2], Hanan Hibshi[1, 3], Daniel Smullen[1]
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States[1]
University of Pennsylvania, Philadelphia, Pennsylvania, United States[2]
College of Computing, King Abdul-Aziz University, Jeddah, Saudi Arabia[3]
{jbhatia, breaux}@cs.cmu.edu, lioraf@sas.upenn.edu, {hhibshi, dsmullen}@cs.cmu.edu

## ABSTRACT
As information systems become increasingly interdependent, there is an increased need to share cybersecurity data across government agencies and companies, and within and across industrial sectors. This sharing includes threat, vulnerability and incident reporting data, among other data. For cyberattacks that include socio-technical vectors, such as phishing or watering hole attacks, this increased sharing could expose customer and employee personal data to increased privacy risk. In the US, privacy risk arises when the government voluntarily receives data from companies without meaningful consent from individuals, or without a lawful procedure that protects an individual's right to due process. In this paper, we describe a study to examine the trade-off between the need for potentially sensitive data, which we call incident data usage, and the perceived privacy risk of sharing that data with the government. The study is comprised of two parts: a data usage estimate built from a survey of 76 security professionals with mean eight years' experience; and a privacy risk estimate that measures privacy risk using an ordinal likelihood scale and nominal data types in factorial vignettes. The privacy risk estimate also factors in data purposes with different levels of societal benefit, including terrorism, imminent threat of death, economic harm, and loss of intellectual property. The results show which data types are high-usage, low-risk versus those that are low-usage, high-risk. We discuss the implications of these results and recommend future work to improve privacy when data must be shared despite the increased risk to privacy.

## CCS Concepts
**• Software and it's engineering→Software design trade-offs • Security and privacy→Privacy protections • Social and professional topics→Government surveillance.**

## Keywords
Cybersecurity data sharing; risk perception; data usage; personal privacy

## 1. INTRODUCTION
Economic development, growth and stability over the last several decades has been driven by information technology (IT): while only 30% of US GDP between 1995 and 2005 is attributed to IT companies, 50% of economic growth during that time can be attributed to IT [1]. Modern IT systems are used to operate critical infrastructure in banking, energy, and transportation, and to create proprietary designs for these sectors in support of this growth. For these reasons, IT systems have become targets for cyberattacks by criminal organizations, nation states and individual hackers. The US Federal Bureau of Investigation reports over $1 billion in reported losses due to criminal cyberattacks [2] and, during the last year alone, attackers have used over 431 million variants of malware to conduct attacks [3]. IT and critical infrastructure are more frequently built from distributed services and multiple vendors, further obfuscating the ability to respond to attacks. In this distributed ecosystem, service providers and vendors can be unaware of the need to share sensitive vulnerability data with partners and competitors [4]. To address this challenge, the US White House introduced Presidential Decision Directive 63 (PDD-63) in 1998 that asks each critical infrastructure sector to establish sector-specific information sharing and analysis centers (ISACs). Today, there are ISACs for a wide range of sectors, including automotive, aviation, financial, health, retail, and water, among others.

In 2015, the White House introduced Executive Order (EO) 13691, which requires the US National Cybersecurity and Communications Integration Center (NCCIC) to coordinate data sharing with ISACs in a public-private partnership. The NCCIC allows companies to automatically receive cybersecurity threat indicators, as well as to voluntarily share their own indicators with other agencies and companies. As of July 2015, the NCCIC has established 125 partnerships with an additional 156 partnerships in negotiation, and they have shared over 28,000 indicators [5]. In 2016, the US legislature passed the Cybersecurity Sharing Act (CISA), which codifies portions of the EO 13691 into law, further supporting the sharing of private-sector incident data with the US government. Despite this support, however, companies report concerns about violating customer and employee privacy by sharing cybersecurity intelligence with others, including the government [6].

In this paper, we report results from a study to measure the trade-off between incident data usage estimates and perceived privacy risk while sharing incident data with the government. The study design consists of two parts: (1) an incident data usage estimate based on an incident reporting survey (see Appendix A) that was conducted with 76 security professionals with a mean of 8 years' experience; and (2) a privacy risk estimate that measures privacy risk based on an ordinal likelihood scale and nominal data type, which is based on a factorial vignette survey design that was conducted with 80 Internet users. The results include evidence as to which data types are high-usage, low-risk versus those that are low-usage, high-risk.

The paper is organized as follows: in Section 2, we present background and related work; in Section 3, we present the method, including the survey designs and formula for computing estimates; in Section 4, we present the survey results and trade-off analysis; in Section 5, we present threats to validity; and in Section 6, we discuss the results and future work.


*WISCS'16*, October 24, 2016, Vienna, Austria.

DOI: http://dx.doi.org/10.1145/2994539.2994541

# 2. BACKGROUND AND RELATED WORK

We now discuss background and related work as follows: Section 2.1 discusses work in risk perception across disciplines; 2.2 discusses data sharing techniques. Lastly, Section 2.3 discusses work in data utility in contrast to our work in data usage.

## 2.1 Risk Perception

Risk is a multidisciplinary topic which spans marketing, psychology, and economics. In marketing, risk is defined as 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 was previously measured (e.g., the acceptable risk of death in motor vehicle accidents) [8]. In psychology, Fischhoff et al. note that so-called 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]. To address these limitations, the psychometric paradigm of *perceived* risk emerged in which surveys are designed to measure personal attitudes about risks and benefits [10]. We define *privacy risk perception* as the act of identifying a choice or action that may have an impact on privacy. However, Knightian economists argue that subjective estimates based on partial knowledge, which includes perceived risk, are measures of uncertainty, and not measures of risk [11]. Two insights that emerged from this paradigm and inform our approach are: (a) people better accept technological risks when presented with enumerable benefits, and: (b) perceived risk can account for benefits that are not measurable in dollars, such as lifestyle improvements [10]. In other words, people who see technological benefits are more inclined to see lower risks than those who do not see benefits. Moreover, privacy is closely associated with a kind of lifestyle improvement, e.g., private communications with friends and family, or the ability to avoid stigmatization.

## 2.2 Data Sharing Techniques

Research in cybersecurity data sharing includes techniques for privacy-preserving protocols [12], [13]. Freudiger et al. proposed data quality metrics and a privacy-preserving protocol based on additive homomorphic encryption to enable sharing privacy-sensitive data [12]. The metrics are based on integrity constraints, which can measure whether data conforms to various tests of equality, comparison and interval membership, and dependency constraints that determine whether data pairs are consistent with each other. For example, whether a state and ZIP code are logically consistent. The goal is to test whether private, encrypted data conforms to these metrics without revealing the content of the data as it moves between client and server. Similarly, Khader proposed an approach to search encrypted data based on attributes [13].

## 2.3 Data Utility and Usage

Xu et al. define utility as the quantity and quality of the data [14]. This work illustrates the trade-off between anonymizing data to protect the seller, and the utility that the collector gains by aggregating and using the data. Their definition of utility has relevance to any interests who may need to use the data, because it introduces the notion that utility may vary when certain data or data quality are unavailable. Insufficient utility may prove inadequate to allow a business arrangement to be met. Utility, by this definition, is like an economic transaction; data is a good, and is traded for some form of other compensation. Under different contract specifications, compensation and utility differ. In Chen et al.'s study, privacy is modeled as a variable based on individual preferences, but no method is proposed to quantify it. With no means to concretely measure privacy in any regard, Chen et al. further provide no underlying theory to justify their representation of privacy as a variable. In our work, we evaluate privacy as an estimate of willingness to share data. Our estimate is derived numerically from survey answers, and we calculate estimates for how willingness to share changes for different data types, among other factors (see Section 3). We contrast this estimate with data usage, which is the frequency with which analysts use the data.

# 3. METHOD AND APPROACH

In this section, we present our study design for computing the trade-off between incident data usage and perceived privacy risk. This includes the survey designs used to collect data, and the method for computing the estimates in the trade-off analysis.

## 3.1 Estimating Incident Data Usage

Incident reports are prepared for evidentiary and training purposes, and include threat indicators derived from incidents [15]. Example indicators include the hostname or IP address from which an attack originates. For incidents where the attack vector is not easily generalized into a non-private indicator, organizations may wish to share more sensitive, detailed technical data with third-party analysts. To do so, they must balance the privacy and security risks of sharing too much data [15]. Thus, we designed a method for estimating incident data usage from the data types that are available for sharing. We assume that, in extreme situations, an organization may wish to share any type of forensic data compiled as evidence of the incident, although, in most cases less data will be shared as a matter of routine.

The incident data usage estimates are computed from the incident reporting survey results. In the survey (see Appendix A), participants describe the data type usage as a frequency interval of incident cases for which each data type is used: 100-50% of cases, 50-25%, <25%, and Never. We propose to estimate incident data usage using two methods: the *simulation* method, and the *relative, ranked usage* method, which we now discuss.

### *3.1.1 Estimating Usage by Simulation*

The simulation method aims to simulate the incident cases that a security analyst estimates what percent of cases each data type is used in. The method assumes that, when an analyst estimates the percent of cases in which the data type is used, they are using the same number of cases to describe each estimate. The simulation universe consists of the set of data types D, and the set of incident reports $r \in R$, where $r \subseteq D$ and R can be partitioned into disjoint subsets, one subset for each analyst. We first generate 100 reports per analyst. For each analyst's data type estimate, we randomly select a percentage within the reported interval wherein all percentages in the interval are equally likely (e.g., 64% is in 100-50%, and 64% is equally likely as 72% to be the analyst's true estimate). Next, we randomly select a corresponding subset of the 100 reports to match this percentage, and assign that data type to those reports. With this dataset, we can estimate the number of reports affected by removing a set of data types DR from all reports by computing the size of the set of affected reports $\{r \mid \forall r \in R, \exists d \in DR \text{ and } d \in r\}$. Estimate results appear in Section 4.2.

### *3.1.2 Estimating Usage by Relative, Ranked Usage*

In the relative, ranked usage method, we identify which data types are used more frequently than other data types in a relative, ranked order. To do this, we first compute a confusion matrix [16]. For each survey respondent, we perform pairwise

comparisons on their reported data type frequency intervals; this yields $n$ 26x26 tables for the 26 data types and $n$ survey respondents. When comparing two frequencies for a given respondent, we proceeded as follows. Given a pair of data types $i$ and $j$, if frequency($i$) < frequency($j$), we enter 0 in cell$_{i,j}$, else if frequency($i$) > frequency($j$) we enter 1 in this cell, else if frequency($i$) = frequency($j$), we enter 0.5 in this cell. Next, we compute the average per data type across the table rows, yielding a number between 0 and 1 per respondent per data type. This number is higher for data types that are more frequently used than other data types. Finally, for each data type, we compute the average across all $n$ respondents' individual averages for the type, yielding a final estimate for each data type between 0 and 1.

For both estimation methods, we treat all data types as semantically independent, despite the fact that some data type interpretations interact through subsumption (e.g. network information includes IP address). Without independence, there is a risk of overestimation, if a data subset is counted twice, (e.g. if the IP address estimate is included in the network information estimate). In addition, there is a risk of underestimation by excluding the IP address estimate from the network information estimate. This issue is addressed further in Section 5, Threats to Validity**.**

The estimation results are presented in Section 4.2.

## 3.2 Measuring Privacy Risk Perception

Factorial vignettes provide a method to measure the extent to which discrete factors contribute to human judgment [17]. The factorial vignette method employs a detailed scenario with multiple factors and their corresponding levels. Our factorial vignette survey extends the survey design of Bhatia et al. [18] to measure the interactions between five independent variables, the *computer type* ($CT) where the cyber incident occurs, the *data type* ($DT) shared with the US Federal government, the *data purpose* ($DP) for which data is shared, the *risk likelihood* ($RL) of a privacy violation, the *privacy harm* ($PH), and their combined effect on a dependent variable, the employee's *willingness to share* ($WtS) their data with the US Federal government.

For this study, we chose to control several factors that affect willingness to share. For example, Nissenbaum argues that privacy and data sharing are contextual, meaning that the factors, data type, data recipient, and data purpose affect willingness to share [19]. We control these factors in a single context—sharing cybersecurity incident data with the government—while varying the computer type affected, the data type and the data purpose.

**Vignette Survey Design**. The factorial vignettes are presented using a template in which factors correspond to independent variables and each factor takes on a level of interest. Figure 1 shows the vignette template: for each participant, each factor is replaced by one level (see Table 1 for the levels for each factor variable, which begin with $). 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. In the vignette survey design, the $DT levels were evenly divided into three groups 1 to 3, thus, each participant sees and responds to 3x4x1=12 vignette combinations. The allocation of $DT levels to groups was made to ensure that types that were technically related are shown together. These Age range was included in each group as a non-sensitive data type aimed at balancing the $WtS scale utilization.

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 | ○ | ○ | ○ | ○ | ○ | |

**Figure 1. Template used for vignette generation (fields with $ sign are replaced with values selected from Table 1)**

A key component in risk estimation is the likelihood of an adverse consequence. Guidance suggests that lay people can map ratios to physical people affected much better than they can map probabilities to people affected [9]. In prior work, Bhatia et al. found that lay people cannot distinguish among ratios to represent the probability of a privacy harm [18]. Alternatively, construal-level theory in psychology claims that people correlate larger spatial, temporal, social and hypothetical distances with decreased likelihood than they do with shorter psychological distances along these four dimensions [20]. We used Bhatia et al.'s empirically validated risk likelihood scale [18] that combines spatial and social distance as a correlate measure of likelihood (see $RL in Table 1): a privacy harm affecting *only one person in your family* is deemed a psychologically closer and more likely factor level than *one person in your city* or *one person in your country*, which are more distal and perceived less likely.

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*. This scale omits a midpoint, such as indifferent or unsure, which produces scale attenuation when responses cluster, and these midpoints are often more indicative of vague or ambiguous contexts than they are of respondents' attitudes [21]. Thus, we chose to force respondents to be either willing, or unwilling to share.

The 12 vignette combinations are presented in group-order: first participants see four vignettes for each group 1-3 in succession, where only the $DP level changes across each group. Prior to responding to each group of four vignettes, participants watch an approximately 60 second video that illustrates the meaning of each data type, because some data types are technical terms that lay people may not be familiar with, such as running processes or registry information. The $DT levels were assigned to each group to fit these narratives, thus the grouped data types had to be related in a technical manner. In addition, the videos offer a break between each group of four vignettes, so respondents can rest.

**Pre-Test Design**. Before the vignettes, we present 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 risk [9]. We ask participants to rank the risk likelihood levels from nearest to farthest proximity as an attention test. Next, we asked participants whether they store personal data on their workplace computer. These three questions aim to sensitize participants to the factorial

vignette levels in Table 1, especially the between-subject factors, prior to asking participants to report their willingness to share.

**Post-Test Design**. After the vignettes, participants are presented a post-test to elicit their demographic characteristics (gender, age range, race, education level, income range).

**Analysis Method**. Multilevel modeling 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 [22]. Multilevel modeling has been used to study security [23] and privacy requirements [18]. In our study, the main dependent variable of interest is willingness to share, labeled $WtS. As can be seen in Table 1, the fixed independent variables, which are within-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.

**Table 1. 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 | |

**Recruitment**. We recruited English-speaking participants from Amazon Mechanical Turk, located in the US, with a ≥97% approval rating, and ≥5000 HITs completed. The mean time to complete in a pilot was ~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.

# 4. RESEARCH RESULTS

We now report the results of our incident response survey. What follows is the results of computing the incident data usage and privacy risk estimates, as well as the trade-off analysis.

For the incident response survey, we recruited a total 76 participants from the SANS Threat Hunting and Incident Response Summit in New Orleans during April 2016. The sample population consists of: 3.9% academic, 75% industry, and 22.4% government; 5.3% are female, and 94.7% are male; 39.5% were aged 25-34, and 31.6% were aged 35-44, and 22.4% were aged 45-60; 98.7% reported having at least some college level education; and they had a mean 8.29 years of experience as a security analyst in some capacity from entry-level to director. Table 2 shows standards and tools used by security analysts. Most respondents reported using Indicators of Compromise (71.4%), followed by Host Based Security System (60.0%).

**Table 2. Frequencies of Standards and Tools Used in Incident Investigations**

| Standards and Tools | Reporting |
|---|---|
| Host Based Security System | 60.0% |
| Structured Threat Information Expression | 35.7% |
| Assured Compliance Accreditation Solution | 11.4% |
| Indicators of Compromise | 71.4% |
| Cyber Federated Model | 2.9% |

Participants more often reported encountering desktop (81.6%) and laptop computers (88.2%), than USB drives (43.4%), smart phones (35.5%), and least commonly tablet computers (19.7%) during their work as incident responders.

Table 3 presents the total proportions of responses for the reported percent of cases in which data types were used in incident analysis. Network information (84.2%), including IP addresses and domain names (86.9%), and OS information (73.7%) were the most frequently encountered data types in 100-50% of cases. Personal data types, such as e-mails (40.8%), browser history (40.8%), passwords (25.0%), chat history (18.4%), and keylogging data (7.9%), were less commonly encountered, yet still present in 100-50% of cases.

**Table 3. Intervals of Reported Use per Data Type (n=76)**

| Data Type | 100-50% of cases | 50-25% of cases | <25% of cases | Never |
|---|---|---|---|---|
| Network information | 63 | 8 | 0 | 0 |
| IP addresses and domain names | 66 | 9 | 1 | 0 |
| Packet data | 33 | 16 | 24 | 2 |
| OS information | 56 | 12 | 7 | 1 |
| OS type and version | 56 | 12 | 6 | 1 |
| Usernames | 50 | 20 | 5 | 1 |
| Passwords | 19 | 10 | 33 | 12 |
| Running processing information | 41 | 26 | 8 | 1 |
| Registry information | 34 | 23 | 15 | 4 |
| Temporary files | 30 | 26 | 15 | 5 |
| Device information | 44 | 21 | 10 | 1 |
| Device identifiers | 36 | 22 | 11 | 6 |
| MAC address | 35 | 17 | 20 | 3 |
| UDID / IMEI | 9 | 16 | 25 | 23 |
| Memory data | 17 | 28 | 23 | 7 |
| Sensor data | 29 | 18 | 20 | 8 |
| Application information | 32 | 28 | 13 | 2 |
| Browser history | 31 | 24 | 16 | 4 |
| Keyword searches | 20 | 25 | 19 | 10 |
| Websites visited | 34 | 24 | 13 | 4 |
| Chat history | 14 | 11 | 31 | 19 |
| Application session data | 16 | 21 | 28 | 10 |
| E-mails | 31 | 23 | 18 | 3 |
| Contact information | 26 | 17 | 25 | 7 |
| Keylogging data | 6 | 17 | 24 | 28 |
| Video or image files | 15 | 15 | 28 | 16 |

Table 4 presents the frequencies of work tasks that applied to participants' current job positions (e.g., security analyst, cyber warfare operation, lead intrusion analyst, etc.) As expected with security analysts engaged in incident reporting, few participants performed preventative tasks, such as vulnerability assessment (41.3%) and patching software and firmware (20.0%). Most participants performed network monitoring (73.3%), forensic investigation (72.0%) and incident report preparation (69.3%).

**Table 4. Frequencies of Work Tasks that Apply to Participants' Current Position**

| Work Tasks | Reporting |
|---|---|
| Vulnerability assessment | 41.3% |
| Patch software and firmware | 20.0% |
| Network monitoring | 73.3% |
| Forensic investigation | 72.0% |
| Threat indicator development | 65.3% |
| Malware analysis | 60.0% |
| Prepare incident reports | 69.3% |
| Security policy compliance | 34.7% |

## 4.1 Privacy Risk Survey Results

We now discuss our results from the privacy risk surveys.

### *4.1.1 Descriptive Statistics*

We received 80 responses to our risk perception survey: 48.8% reported as female, 51.3% male; 80.0% reported white as their ethnicity; 85.0% reported having at least some college level education; and 82.4% reported having annual household income less than $75,000. Less than 5% report their age as 18-24 years, with 43.8% aged 25-34 and 23.5% aged 35-44, and 28.8% report being over 45 years old. With 80 responses, we achieved 97% actual power, calculated using G*Power [24].

Participants were asked to rank order the data purposes by their benefit to society. Overall, the majority ranked the data purposes as follows: investigating imminent threat of death (68.8%) was most beneficial, followed by terrorism (60.0%), followed by economic harm (63.8%), and ending with intellectual property (68.8%) as least beneficial.

Asked whether participants stored personal data on their workplace computer, 42.3% reported Yes, and 58.7% reported No.

### *4.1.2 Multilevel Model for Privacy Risk*

Equation 1 below is our main additive regression model with a random intercept grouped by participant's unique ID, the independent between-subject measures $CT, which is the computer type, and $RL, which is the likelihood of a privacy violation, and the independent within-subject measure $DP, which is the data purpose from one of the four categories and $DT, which is the data type (see Table 1 in Section 3.2). The additive model is a formula that defines the dependent variable $WtS, willingness to share, in terms of the intercept α and a series of components, which are the independent variables. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Eq. 1 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience.

$$\$WtS = \alpha + \beta_C\$CT + \beta_R\$RL + \beta_P\$DP + \beta_D\$DT + \epsilon \qquad (1)$$

To compare dependent variable $WtS across vignettes, we establish the baseline level for the factor $CT to be workplace computer, $RL to be "*only one person in your family*" who experiences the privacy violation, $DP to be investigating intellectual property and we set the factor $DT to "age range". The intercept (α) is the value of the dependent variable, $WtS, when the independent variables, $CT, $RL, $DP and $DT take their baseline values.

We found a significant contribution of the four independent factors for predicting the $WtS ($\chi^2(32)=2415.1$, $p<0.000$) over the null model, which did not have any of the independent variables. In our model, we did not observe any effect of the independent variable $CT, ($\chi^2(1)=2.2319$, $p=0.1352$), which means Computer Type did not affect the willingness to share. We also did not observe a statistically significant effect of the independent variable $RL, ($\chi^2(4)=1.5181$, $p=0.8234$), which means the Risk Level may not affect the willingness to share. In Table 5, we present the model *Term*, the corresponding model-estimated *Coefficient* (along with the p-value, which tells us the statistical significance of the term over the corresponding baseline level), and the coefficient's *Standard Error*. In our survey, the semantic scale option *Extremely Unwilling* has a value of 1, and *Extremely Willing* has a value of 8. A positive coefficient in the model signifies an increase in willingness to share and a negative coefficient signifies a decrease in willingness to share.

The results in Table 5 show that $WtS is significantly different and increasing for increasing levels of $DP, as compared to the baseline level, *investigating intellectual property and trade secrets*, except for *investigating economic harm, fraud, or identity theft*. For the $DP level *investigating imminent threat of death or harm to an individual, including children*, the $WtS increases by 1.153 over the baseline level, which denotes a statistically significant increase in willingness to share. However, for the $RL level *only 1 person in your country*, the $WtS decreases by 0.461 over the baseline level, *only 1 person in your family*, which denotes a very slight decrease in willingness to share, but the change is not statistically significant.

**Table 5. Multilevel Modeling Results**

| Term | Coefficient | Std. Error |
|---|---|---|
| Intercept (family + workplace PC + intellectual) | 6.340*** | 0.421 |
| Risk Level – 1 person in your workplace | -0.611 | 0.533 |
| Risk Level – 1 person in your city | -0.519 | 0.533 |
| Risk Level – 1 person in your state | -0.355 | 0.533 |
| Risk Level – 1 person in your country | -0.461 | 0.533 |
| Data Purpose – economic harm | 0.136** | 0.044 |
| Data Purpose – terrorism | 0.795*** | 0.044 |
| Data Purpose – imminent death | 1.153*** | 0.044 |
| Computer Type – personal smart phone | -0.512 | 0.337 |

*p ≤ 0.05 **p ≤ 0.01 ***p ≤ 0.001

The estimated dependent variable $WtS is presented for each data type in Table 6.

## 4.2 Trade-off Analysis Results

The trade-off analysis compares the incident data usage estimate to the perceived privacy risk for each $DT. The usage is estimated using the simulation method described in Section 3.1.1, and the relative, ranked method described in Section 3.1.2 and both estimates appear in Table 6 under the columns "Simulated Usage" and "Ranked Usage," respectively. The Pearson's $r$ correlation co-efficient for the two estimates is 0.995, which is a high correlation. The perceived privacy risk is measured by the estimated willingness to share $WtS (intercept=intellectual

property+1 person in your family+ workplace computer) on a scale of 1 to 8, wherein 1=Extremely Unwilling and 8=Extremely Willing, and which estimates an average Internet user's acceptance of the risk.

In the relative, ranked usage estimate function, each variable is binary, taking on a 1, if the data type is present in the incident report, and a 0 if the data type has been removed from the report. The final estimated value for a report's data types has no particular meaning, rather its meaning comes from its comparison to other estimates, as follows: the values are meaningful in units of distance, but not multiplicative distance. We can say that contact information has 0.202 more units of usage than keylogging data (see Table 6). However, we do not say that contact information has 1.84 times as much usage as keylogging data.

**Table 6. Estimates for Incident Data Usage using the Simulated and Relative, Ranked Usage methods**

| # | Data Type | Simulated Usage | Ranked Usage | $WtS |
|---|---|---|---|---|
| 1 | Passwords | 0.244 | 0.350 | 4.149 |
| 2 | Usernames | 0.610 | 0.661 | 4.149 |
| 3 | Keylogging data | 0.144 | 0.240 | 4.231 |
| 4 | E-mails | 0.408 | 0.524 | 4.340 |
| 5 | Chat history | 0.203 | 0.300 | 4.378 |
| 6 | Video or image files | 0.225 | 0.320 | 4.603 |
| 7 | Browser history | 0.422 | 0.526 | 4.649 |
| 8 | Web sites visited | 0.449 | 0.545 | 4.871 |
| 9 | Contact information | 0.336 | 0.442 | 4.874 |
| 10 | Keyword searches | 0.319 | 0.421 | 4.921 |
| 11 | Temporary files | 0.439 | 0.499 | 5.209 |
| 12 | Application session data | 0.244 | 0.545 | 5.268 |
| 13 | Memory data | 0.291 | 0.405 | 5.353 |
| 14 | Registry information | 0.459 | 0.534 | 5.371 |
| 15 | Packet data | 0.407 | 0.505 | 5.437 |
| 16 | Sensor data | 0.381 | 0.468 | 5.524 |
| 17 | Application information | 0.463 | 0.545 | 5.721 |
| 18 | Running process information | 0.526 | 0.610 | 5.790 |
| 19 | Network information | 0.667 | 0.715 | 5.862 |
| 20 | UDID / IMEI | 0.177 | 0.258 | 5.928 |
| 21 | Device identifiers | 0.464 | 0.543 | 6.984 |
| 22 | MAC address | 0.440 | 0.519 | 6.028 |
| 23 | Device information | 0.535 | 0.618 | 6.043 |
| 24 | IP addresses / Domain names | 0.673 | 0.741 | 6.093 |
| 25 | Operating system information | 0.600 | 0.670 | 6.603 |
| 26 | OS type and version | 0.588 | 0.673 | 6.603 |

In Figure 2, we present a scatterplot comparing the data usage and $WtS. The data types are arranged along the x-axis in the numbered order from Table 6, starting with password, usernames, keylogging data, and so on. Along the y-axis, we scaled the data usage and privacy estimate for each data type: the simulated usage (blue dots) appears alongside the $WtS-Normed (orange dots), which is $WtS values rescaled by normalizing the value by dividing by 8.0. For $WtS-Normed values from 0.5 to 0.0, the users are increasingly less willing to share their data; from 0.5 to 1.0, the users are increasingly more willing to share their data. Figure 2 shows that as the privacy risk estimate decreases from left to right, the data usage trend appears to be increasing, overall.

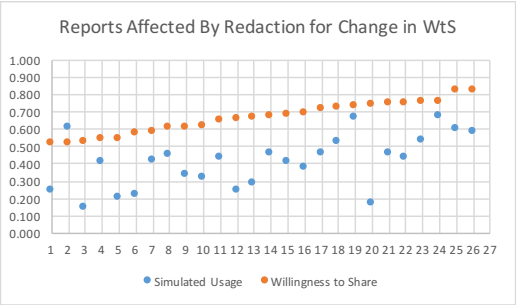


**Figure 2. The normative disparity between incident data usage and privacy risk (simulated usage and $WtS-Normed)**

In Figure 3, we present a histogram to show the distribution of the willingness to share ratings. The x-axis, shows the willingness to share scale options, and the y-axis shows the number of ratings for each scale option. In Figure 3, we observe that participant's scale use is skewed slightly toward "extremely willing," with 65% of all ratings lying between "willing" and "extremely willing." The histogram consists of all 1800 ratings from all 80 participants, regardless of whether participants chose similar ratings across all questions. To investigate responses by participants who utilize the full scale, we calculated the standard deviation (SD) for all ratings by participant. We found 19 participants, or 25% of the sample, with a SD ≥ 2. Figure 4 presents the trade-off of data use versus privacy risk for these 19 participants: as shown, why participants are more willing to share information about who they are (e.g., IP address, UDID, MAC address), they are less willing to share information about what they do (e.g., browser history, e-mails, websites visited, etc.)

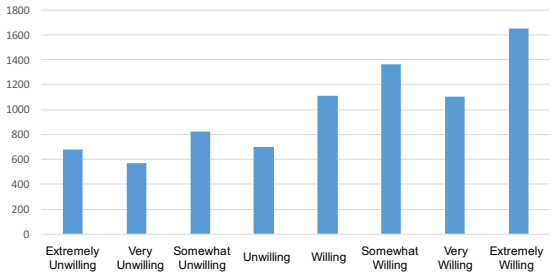


**Figure 3. Distribution of Willingness to Share Ratings**

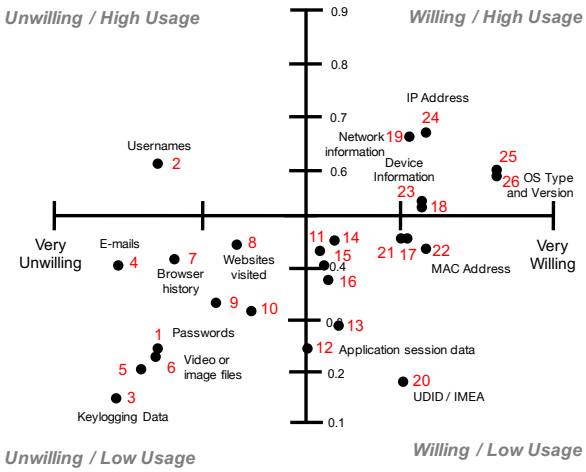


**Figure 4. Trade-off between incident data usage and user's willingness to share their data**

The results are further discussed in Section 6, after we present threats to validity.

# 5. THREATS TO VALIDITY

Construct validity addresses whether what we measure is actually the construct of interest [25]. The trade-off analysis includes vague superordinate types, such as network information, which can have different meanings between the sampled security analysts and lay people. This ambiguity is an artifact of the problem, in which terminology to describe data types is poorly defined. With respect to lay people's perceptions, in a separate survey of 40 participants we observed that respondents were statistically more comfortable sharing "technical information" with law enforcement than they were sharing "IP address," which is commonly considered by companies to be included in the category of technical information.

Internal validity concerns whether the study procedures limit drawing correct inferences from the data [25]. One threat to internal validity is that our survey participants may not have understand the meaning of the different data types for which they needed to rate $WtS. In order to mitigate this, instructional videos illustrating the definition of each data type accompanied each group of survey questions, which included closed captions for the hearing impaired. Additionally, each data type presented with a semantic scale (as seen in Figure 1) had a definition that was displayed when the participant hovered their mouse over the data type.

External validity refers to the extent to which we can generalize the results to other situations [25]. In the privacy risk survey, 42.3% of participants reported storing personal data on their workplace computer. This frequency may be higher or lower depending on the sectorial culture and company policies influencing employee behavior, which can affect the level of perceived risk in the ideal population. People who store less personal data on their workplace computer may report lower perceived privacy risk, but we found no statistical significance to this effect ($p = 0.1352$, compared to the null model). In fact, it is possible that people view their workplace computer as storing as much personal data as their smart phone.

# 6. DISCUSSION AND FUTURE WORK

In Section 4, we presented results from estimated incident data usage and perceived privacy risk. The incident data usage estimates are based on a simulation and confusion matrix that were computed from surveys conducted with 76 professional security analysts. The privacy risk estimates were computed using factorial vignette surveys with 80 Internet users. We present results measuring the normative disparity between usage and risk (see Figure 2). In addition, we measured statistically significant differences between the privacy risk estimates based on the data purpose for which these types are shared with the US Federal government and found that users were more willing to share their data for purposes with higher societal benefit, e.g., terrorism and imminent threat of death (see Table 5).

In addition, results show a trade-off exists between data usage and privacy risk, in particular, that few types have high use and high risk (e.g., usernames) and most types have low or high use and low risk (see Figure 4). For low-risk data types, incident responders may feel comfortable sharing these data types using routine procedures for securing the data. These low-risk procedures likely include access control, and disk- and network-based encryption, for example, and security analysts who have access to the data may also be permitted to conduct their investigations by exploring the data, and with access to a broader set of data in the low-risk categories. For moderate- and high-risk data categories, however, security analysts may need to use data minimization techniques, such as redaction, to remove these data types before sharing. In addition, they may need to restrict access to those security analysts who are investigating specific incidents where the data is needed, and excluding such data from uncontrolled, exploratory practices. Ensuring such restrictions in cross-agency sharing environments is difficult due to the lack of transparency and lack of consistent data type terminology.

In future work, we propose to conduct additional surveys to improve our estimates and add new context. For example, we propose to investigate how Chief Security Officers and incident responders perceive privacy risk, and looking at how regulatory frameworks restrict data sharing (e.g., the IP address is considered identifiable data in Europe). In addition, we propose to study data sharing using the Eddy privacy requirements language. The Eddy language allows privacy policy authors to express their data collection, use and transfer requirements and to identify conflicts between permitted and prohibited data usage and sharing practices [26]. In addition, Eddy-based tools exist to trace data flows across agencies to ensure that requirements follow the data [27]. To support the challenge of restricting flows across agencies, the Eddy language should be extended to support data minimization techniques, such as redaction and perturbation, which includes the introduction of noise into a high-risk dataset.

# 7. ACKNOWLEDGMENTS

We would like to acknowledge that all authors on this paper contributed equally. We thank Dr. Stephen Broomell for his statistics guidance, and the CMU Requirements Engineering Lab, including João Caramujo, Morgan Evans, Mitra Bokaei Hosseini, and David Widder, for their helpful feedback. This work was supported by NSA Award #141333 and ONR Award #N00244-16-1-0006.

# 8. REFERENCES

[1] D. W. Jorgenson, M. S. Ho, and K. J. Stiroh, *Productivity, Volume 3: Information Technology and the American Growth Resurgence*, Postwar U.S. Economic Growth. MIT Press, 2005.

[2] FBI, "2015 Internet Crime Report," 2016.

[3] Symantec, "Internet Security Threat Report 2016," April, p. 81, 2016.

[4] D. Shackleford, "Combatting cyber risks in the supply chain," SANS.org, 2015.

[5] O. of the W. H. P. Secretary, "Fact Sheet: Administration Cybersecurity Efforts 2015," 2015.

[6] PWC, "The Global State of Information Security® Survey 2016: Turnaround and transformation in cybersecurity," 2016.

[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 (80-. ).*, vol. 165, no. 3899, p. 1232, Sep. 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.*, vol. 9, no. 2, pp. 127-152, Apr. 1978.

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

[11] F. Knight, "Risk, Uncertainty, and Profit," *Hart Schaffner Marx Prize essays*, vol. XXXI, pp. 1-173, 1921.

[12] J. Freudiger, S. Rane, A. E. Brito, and E. Uzun, "Privacy Preserving Data Quality Assessment for High-Fidelity Data Sharing," *WISCS '14 Proc. 2014 ACM Work. Inf. Shar. Collab. Secur.*, pp. 21-29, 2014.

[13] D. Khader, "Attribute Based Search in Encrypted Data," *Proc. 2014 ACM Work. Inf. Shar. Collab. Secur. - WISCS '14*, pp. 31-40, 2014.

[14] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. R. Liu, "Privacy or utility in data collection? A contract theoretic approach," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 7, pp. 1256-1269, Oct. 2015.

[15] P. Cichonski, T. Millar, T. Grance, and K. Scarfone, "Computer Security Incident Handling Guide: Recommendations of the National Institute of Standards and Technology, 800-61. Revision 2," *NIST Spec. Publ.*, vol. 800–61, p. 79, 2012.

[16] J. C. Baird and E. Noma, *Fundamentals of scaling and psychophysics*. John Wiley & Sons, Inc., 1978.

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

[18] J. Bhatia, T. D. Breaux, J. R. Reidenberg, T. B. Norton, "A Theory of Vagueness and Privacy Risk Perception," *IEEE 24th International Requirements Engineering Conference (RE'16)*, 2016.

[19] H. Nissenbaum, *Privacy in context: Technology, policy ,and the integrity of social life*. Stanford Law Books, 2009.

[20] C. Wakslak and Y. Trope, "The effect of construal level on subjective probability estimates," *Psychol. Sci.*, vol. 20, no. 1, pp. 52-58, Jan. 2009.

[21] J. T. Kulas and A. A. Stachowski, "Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators," *J. Res. Pers.*, vol. 47, no. 4, pp. 254-262, Aug. 2013.

[22] A. Gelman and J. Hill, "Data analysis using regression and multilevel/hierarchical models," *Policy Anal.*, pp. 1-651, 2007.

[23] H. Hibshi, T. D. Breaux, and S. B. Broomell, "Assessment of risk perception in security requirements composition," *2015 IEEE 23rd Int. Requir. Eng. Conf. (RE)*, pp. 146-155, 2015.

[24] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences.," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175-91, May 2007.

[25] J. Creswell, "Research design : qualitative, quantitative, and mixed methods approaches," 2014, p. 273.

[26] T. D. Breaux, H. Hibshi, and A. Rao, "Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements," *Requir. Eng.*, vol. 19, no. 3, pp. 281-307, Sep. 2014.

[27] T. D. Breaux, D. Smullen, and H. Hibshi, "Detecting repurposing and over-collection in multi-party privacy requirements specifications," *2015 IEEE 23rd Int'l Req'ts Engr. Conf. (RE)*, 2015, pp. 166-175.

**Appendix A. Incident Reporting Survey Questions**

1. Check the following standards and tools that you currently use:
   - ☐ Host Based Security System (HBSS)
   - ☐ Structured Threat Information Expression (STIX)
   - ☐ Assured Compliance Accreditation Solution (ACAS)
   - ☐ Indicators of Compromise (IOC)
   - ☐ Cyber Federated Model (CFM)

2. When investigating an incident, what device types do you typically encounter?
   - ☐ Desktops
   - ☐ Laptops
   - ☐ Smart phones
   - ☐ Tablets
   - ☐ USB drives
   - ☐ Other, please list: _____

3. When investigating an incident, which kind of information do you collect or use in your analysis?

| Information Type | 100-50% of cases | 50-25% of cases | <25% of cases | Never |
|---|---|---|---|---|
| Network information | | | | |
| IP addresses and domain names | | | | |
| Packet data | | | | |
| OS information | | | | |
| OS type and version | | | | |
| Usernames | | | | |
| Passwords | | | | |
| Running processing information | | | | |
| Registry information | | | | |
| Temporary files | | | | |
| Device information | | | | |
| Device identifiers | | | | |
| MAC address | | | | |
| UDID / IMEA | | | | |
| Memory data | | | | |
| Sensor data | | | | |
| Application information | | | | |
| Browser history | | | | |
| Keyword searches | | | | |
| Websites visited | | | | |
| Chat history | | | | |
| Application session data | | | | |
| E-mails | | | | |
| Contact information | | | | |
| Keylogging data | | | | |
| Video or image files | | | | |

4. Check all the work tasks that apply to your current position:
   - ☐ Vulnerability assessments
   - ☐ Patch software and firmware
   - ☐ Network monitoring
   - ☐ Forensic investigations
   - ☐ Threat indicator development
   - ☐ Malware analysis
   - ☐ Prepare incident reports
   - ☐ Security policy compliance