# Privacy Risk in Cybersecurity Information Sharing

Jaspreet Bhatia, Travis D. Breaux, Liora Friedberg, Hanan Hibshi, Daniel Smullen

## Introduction

- Information systems increasingly interdependent → increased **need to share cybersecurity data** → **potential exposure** of personal data
- We 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

## Methodology

**Building perceived privacy risk regression model:**

- The perceived privacy risk is measured by the estimated willingness to share $WtS, estimating acceptance of risk
- Uses an ordinal semantic scale in **factorial vignette surveys**:
  - 1 to 8, Very Unwilling to Very Willing
- Factors in data purposes with different nominal societal benefit levels

**Building data usage estimates:**

- Built from a survey in which security professionals describe data type usage as a frequency interval of incident cases

- **Simulation Method**
  - **Simulates the incident cases** that a security analyst has in mind
  - With this dataset, we can **estimate the number of reports affected by removing a set of data types**
- **Relative, Ranked Usage Method**
  - **Which data types are used more frequently** than other data types
  - Determined from **confusion matrices**

**Vignette Factors and 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 | |

**Data Usage Estimates and $WtS**

| # | 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 |

## Results

$$\$WtS = \alpha + \beta_C \$CT + \beta_R \$RL + \beta_P \$DP + \epsilon$$

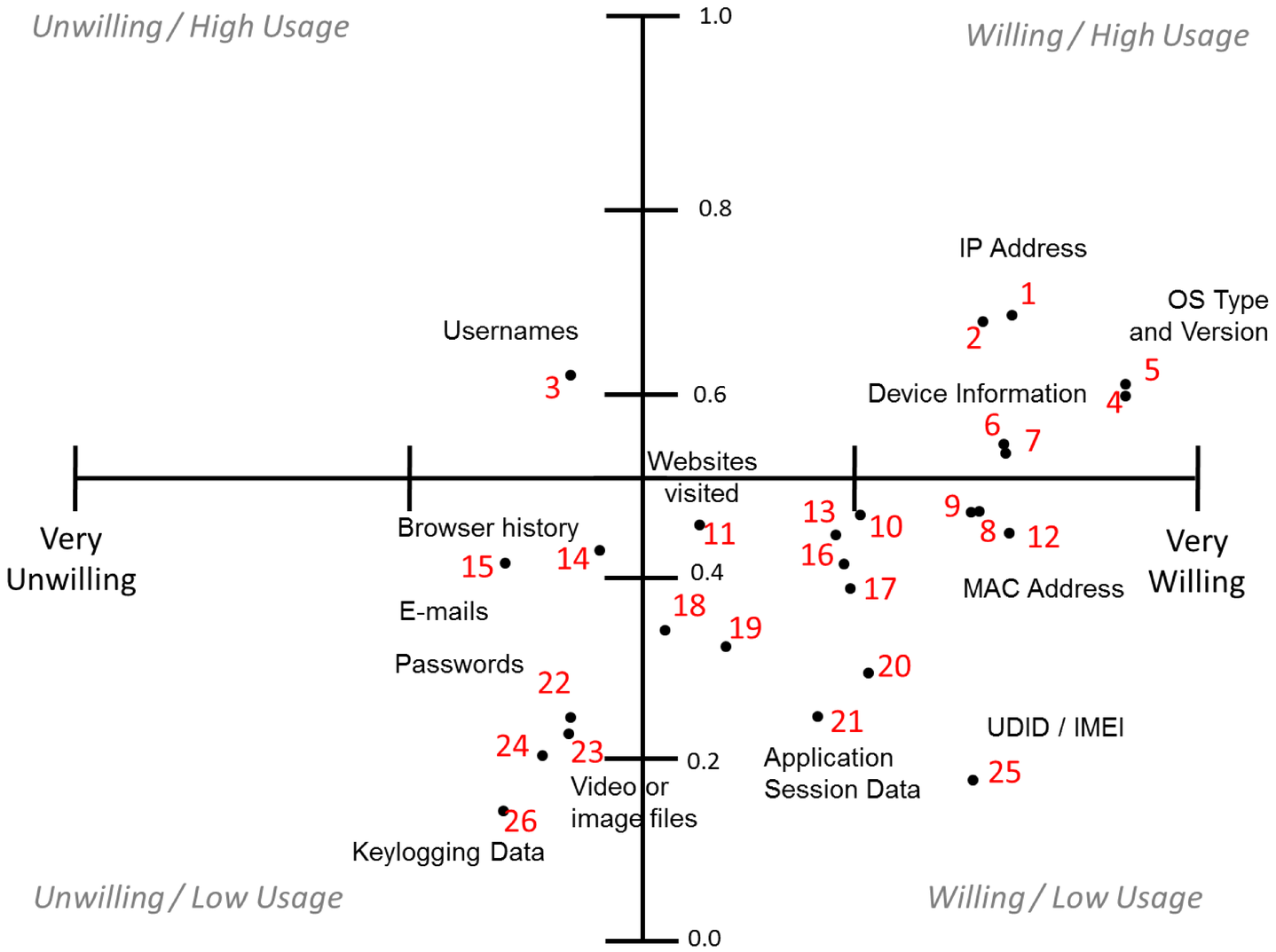
- No statistical significance for computer type ($CT) or risk level ($RL) effect on $WtS regression model
- $WtS significantly increases for data purposes of increasing societal benefit
- Trade-off revealed between usage and risk for data types

**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

**Trade-off Between Data Usage and $WtS**




## Future Work

- Investigate data sharing using the *Eddy privacy requirements language*
- Simulate data and data sharing with *dynamic microsimulation*

**Special thanks to Dr. Stephen Broomell and the CMU Requirements Engineering Lab. Supported by NSA Award #141333 and ONR Award #N00244-16-1-0006.**