Leaked Databases and User Risk

Alex Bainbridge, Kenny Cohen, Frank Liao

Problem

As data breaches continue to be prevalent in todays society, it is important that the information contained in them be analyzed and studied. In order to understand the risks of these security issues, we measure the perceived harmfulness of data inferred from leaked data.

Goals

Construct User Profile

Develop Codebook for Risk Analysis and Inference Targets

Cross-Reference Leaked Databases

Data Sets

(millions of records)

117

32

15

000webhost

.007

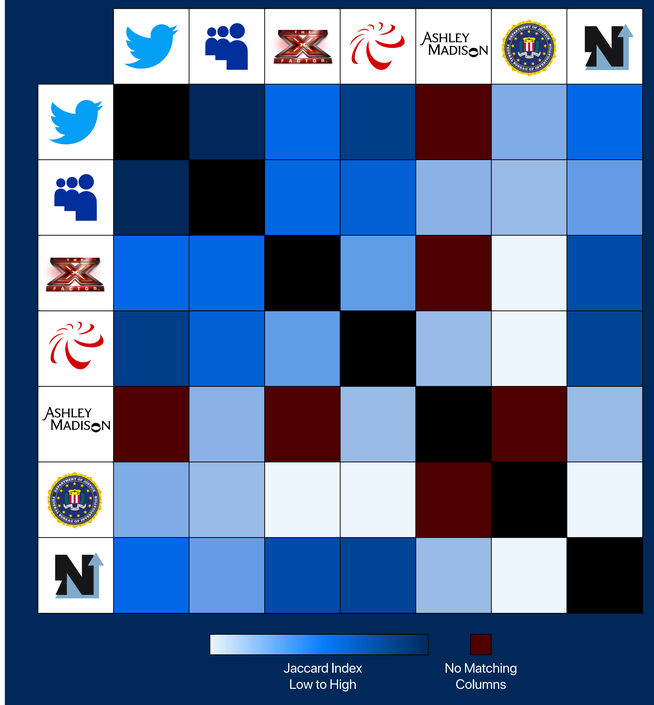
.15

11

Ashley Madison

.0001

Findings and Code Book



Variable	Risk Level	Value Label
Username	No Risk	Class 1
Email	Small Risk	Class 1
IP Address	Small Risk	Class 2
Password	No Risk	Class 1
SHA(Password Truncated)	No Risk	Class 1
SHA(Salted Password)	No Risk	Class 1
First Name + Last Name	Detrimental Risk	Class 1
First Name	Small Risk	Class 1
Last Name	Small Risk	Class 1
Location	Detrimental Risk	Class 2
Date of Birth	Small Risk	Class 1
Phone	Small Risk	Class 1
Gender	Small Risk	Class 1
Zip Code	Small Risk	Class 2
Timestamp	No Risk	Class 2

Label Definitions

No Risk – The field offers no risk to the user if leaked OR no effect on the anonymity set

Small Risk – The field offers minimal risk to the user, minimal in that something is learned about the user, but has no lasting impact OR slight changes to anonymity set

Large Risk – The field offers large risk to the user, the user may be severly impacted by this AND decreases the anonymity set signifcantly

Detrimental Risk – The field offers large risk to the user, the user may be wholey compromised with this information AND small anonymity set

Class 1 – The materialization of an entity

Class 2 – The materialization of an activity

Class 3 – The materialization of a sensitive relationship between materialized entities

Class 4 – The materialization of a sensitive relationship between materialized activities

Class 5 – The materialization of a sensitive relationship between one or more materialized activities or entities.

Class 6 – The materialization of a sensitive relationship between sensitive relationships

Class 7 – The materialization of a sensitive rule from existing classes

Inferences Coding

	Combined Fields	Resultant	Inference Target
Small	Username + IP Address	Persona Tagging	Class 5
Large	Username + Password	Persona Account Access	Class 3
	Email + Password	Persona Account Access	Class 3
	Zip Code + Birth Date + Gender	Persona Identifier	Class 5
	Phone + Email	Contact Information	Class 3
	IP Address + Password	Login Access	Class 5
Detrimental	Location + Timestamp	User Location at Time	Class 4
	User Location at Time + User Tagging	User Tied to Internet Activity	Class 6
	First Name + Last Name	Legal name	Class 3

Limitations

Computational Power

Hashed data made partial matches impossible

Small Number of Data points in comparison to internet population

Future Work

Including public data sets, web scraping and social media websites

Expert system integration

Matching of databases to vulnerable products to put users at risk for more attacks