Modified VAST Challenge with Applications to Data Breaches

Update

Rosalyn Carr*
University of British Columbia

1 Introduction

Although a formal definition is not widely agreed upon, data breaches are often defined by the illegal use or disclosure of confidential information and are categorized into internal and external breaches. Internal breaches involve the assistance of individuals within the affected organization, whether voluntarily or not, to distribute personal or confidential information. External data breaches are caused by external entities such as hackers or other parties. Hacking and IT incidents comprise the majority of these breaches. [1]

Data breaches pose a threat to both the individual client and organization. Potential harm includes financial setbacks, lost clientele, tarnished reputation, and compromised personal information leading to identify theft. [1] A recent survey found that 76% of those affected by a data breach felt serious stress afterwards, however, surprisingly less than half took any steps to protect themselves from future identity theft or other data breaches. Common themes for lack of action included the overwhelming amount of data security information to process before taking preventative action, or lack of education prior to a breach. [2]

Successful data breach prevention often funnels down to education and modification of basic security best practices. Many users do not always understand where a data breach can happen, and often dismiss a single data breach, unaware of the compounding issues that could be taking place. [3] A key starting point in educating individuals is understanding which data is most crucial to protect. The primary goal of this work was to develop a tool understandable by lay people to put in perspective the risk of identity theft when a data breach occurs through an interactive visualization tool of data often found in breaches. This tool is based on the dataset provided for the 2021 Visual Analytics Science and Technology (VAST) Challenge Mini Challenge 2 [4], as although real data breach datasets exist, the individuals to which the information belongs to have likely not consented to the further distribution of their personal details, or are even unaware that it is publicly available.

1.1 Personal Experience

My specific research topic is related to bi-directional data sharing between patients and clinicians, of which breaches are always a significant risk not only due to the sharing of information over multiple platforms but the type of data associated. My work aims to develop an appropriate and ethical procedure for systematic dissemination of individual (non-aggregate) research results for the purpose of expanding informed consent and engagement with clinical research. This process must be incredibly secure, as the consequences of a data breach of clinical data are almost always catastrophic for the patients and families involved.

2 Related Work

Data breaches are becoming more discussed as they become more and more frequent; however, many tools are simply web articles with extensive lists of recommendations that are visually overwhelming. These articles often only phrased as a response to a data breach (targeting users who have discovered they are a victim of a data breach and are looking for solutions) or a general "protect everything" argument that lacks targeted information for users. [5] Not every data breach is the same, and it can be difficult for a lay person to navigate these sources. In contrast, some academic sources have aimed to develop risk factors and other tools to help communicate the consequences and risks associated with carrying data breaches. [6], [7]

2.1 Criminological Contextual Risk of Breaches

The academic paper by Sen and Borle aims to develop a risk factor model to estimate and classify data breaches. The risk of data breach was measured in the context of an organization's physical location, its primary industry, and the type of data breach that it may have suffered in the past. Multiple theories were applied to create a measurement system, including institutional theory and the opportunity theory of crime. These measurements were then built into a statistical model to identify key indicators for future data breaches. [6]

Although this paper follows key criminological theories and follows a strict empirical framework for identifying risk factors, the results are not easily interpreted by a lay person and the application of the system seems quite limited by the availability of information (such as industrial classification and internal spending of a company). [6]

2.2 Visualization of Data Breaches

2.2.1 Breach Reidentification

The academic paper by Liu et al. uses a real-life data breach as well as publicly available income and transport statistics to create a series of visuals to demonstrate the risk of identity theft among Americans. Using a neural network, it was found the individual income could be predicted using the breached data and the publicly available income statistics. This cross referencing between public and private (now breached) data combined with the visuals aimed to show how risky even existing data breaches can be to the public, however there are some pitfalls. [7]

Many laypeople unfamiliar to artificial intelligence are unlikely to understand how these methods work. [8] The visuals are limited to frequency of breaches within certain categories (such as which professions more frequently experience data breaches), however it does not contextualize (certain professions may have more data storage inherently in their work) these conclusions nor control for population size (instead just shows the raw number of records breached). [7]

2.2.2 Visualizations of Breach Statistics for Lay Users

Multiple infographic approaches have been taken to visualize statistics regarding breached data. Some idioms are focused heavily on the sector the data was leaked from, but this often diminishes what the individual impact is. [9] Other implementations will focus on a specific sector, but the visualizations are often just traditional static bar graphs or choropleths. [10] Both of these, although

* e-mail: rosalyncarr@ieee.org

suitable for a lay audience, are quite limited to aggregated data that can remove the individual connection.

3 Data and Task Abstraction

3.1 Domain

Data breaches affected hundreds of millions of individuals each year, however the data is still sensitive. [11] While datasets of real-life breaches exist, [7] in an attempt to be respectful to those personally affected, these were not chosen. Instead, this project follows the 2021 VAST Challenge Mini Challenge 2. [4]

The 2021 VAST Challenge is a reprise of the 2014 challenge, with similar associated tasks related to personal information collection and individual identification. The context is a company is concerned about the actions of their employees and has attached geospatial trackers to company cars. [4] The "data" that was fabricated to the challenge was originally intended to be used to identify individual employees, monitor their behaviour, identify patterns consistent with crimes reported, and to report the suspicious behaviour to law enforcement. [4]

Instead of following the usual trajectory of the challenge and presenting a list of suspicious individuals to the "law enforcement", the data was instead recontextualized as "breached" data. This tool, contrary to others built previously, [7] involves an interactive component to allow users to see which combinations of data are crucial in identifying an individual compared to others, and how their own choices in protecting certain pieces of data over others can lead to better or worse results in the eyes of someone acting as an identity thief.

3.2 Task

Some of the original tasks from the 2021 VAST Challenge Mini-Challenge 2 are preserved as they are required for later synthesis. [4] This includes:

- Identify Locations of Interest and Find Data Discrepancies (Q1 and Q2 from the original challenge
- Infer the owners of each credit card and loyalty card (Q3 from the original challenge)
- Identify most crucial information for identification

The original challenge required synthesis of multiple datasets to identify individual employees and track their actions. There were deliberate anomalies in the fabricated data (such as the owner of a credit card being in a different location at the time the credit card was used), as one of the original prompts was to identify "suspicious individuals". Rather than follow the original prompts to locate these individual employees as describe their suspicious behavior, these anomalies will be disregarded.

3.3 Data

Data will be used from the 2021 VAST Challenge Mini Challenge 2. [4] A table outlining all attributes across the four datasets can be seen in the appendix.

4 Proposed Solution

The proposed solution will involve replicating the 2021 VAST Challenge Mini Challenge 2 as an interactive tool where a user can "build" the visualization themselves through personally selecting attributes of the dataset to see whether identification of employees is possible and using which specific attributes.

A map of the fabricated town created for the 2021 VAST Challenge Mini Challenge 2 is provided in the dataset for the final visualization, with intent that patterns in geographically associated data be marked directly on top. [4] This map will not be utilized as it, instead a less visually distracting replicate will be used.

Attributes will be provided as a list for the user, with the ability to select and deselect attributes to add or remove them from the map. If an attribute requires another to be visualized (for example, the user selects only categorical data without any geographic data to associate it to the map), an error message will be presented. Attributes will also have associated descriptions that can be toggled on or off by users, including information about the context of the attribute (which dataset it came from, what it means) as well as associated information on where a similar piece of data could be collected in a data breach (such as latitude and longitude data being available from many Bluetooth tracking applications). The goal is to allow the users to form a mental model on how the information arrived in front of them; how it could have gotten there, who could have acquired it, and what they could do with it.

4.1 Implementation

Exploration of the datasets and possible idioms was performed in R. All visualizations are built with D3 to incorporate the necessary interactive features, with the later goal of the tool being publicly available on a web platform.

4.2 Scenario of Use

The scenario of use for a potential user would be exploratory. Unlike existing static models, the goal of this visualization would be to involve users directly in seeing a more cause and effect approach to data breaches. Somewhat of a education tool, it would provide information about the risks and consequences of data breaches in a more hands on fashion that would allow users to understand the complexities and risks associated with putting their data online.

More expanded versions of the tool, time permitting, will aim to incorporate the usual information presented on the blog-style websites covering data breaches, but in a less reactionary and more educational manner aimed towards guiding users to take more preventative measures. Asking users to protect every last piece of data they put online is difficult and out of touch, as many have already been victims of data breaches they can't avoid. Instead, users can contextualize their own experiences and hopefully be guided to more safe data practices.

References

[1] A. H. Seh *et al.*, 'Healthcare Data Breaches: Insights and Implications', *Healthcare*, vol. 8, no. 2, Art. no. 2, Jun. 2020, doi: 10.3390/healthcare8020133.

[2] 'How Does It Feel To Be The Victim of A Breach? | Proofpoint US', *Proofpoint*, Sep. 22, 2020. https://www.proofpoint.com/us/blog/insider-threat-management/how-does-it-feel-be-victim-breach (accessed Nov. 02, 2022).

[3] 'Data Breach Detection 101'. https://www.echosec.net/blog/data-breach-detection (accessed Nov. 02, 2022).

[4] 'Mini-Challenge 2': https://vast-challenge.github.io/2021/MC2.html (accessed Nov. 02, 2022).

[5] 'Data Breach Response: A Guide for Business | Federal Trade Commission'. https://www.ftc.gov/business-guidance/resources/data-breach-response-guide-business (accessed Nov. 15, 2022).

[6] R. Sen and S. Borle, 'Estimating the Contextual Risk of Data Breach: An Empirical Approach', *J. Manag. Inf. Syst.*, vol. 32, pp. 314–341, Apr. 2015, doi: 10.1080/07421222.2015.1063315.

[7] L. Liu, M. Han, Y. Wang, and Y. Zhou, 'Understanding Data Breach: A Visualization Aspect', in *Wireless Algorithms, Systems, and Applications*, Cham, 2018, pp. 883–892. doi: 10.1007/978-3-319-94268-1_81.

[8] A. Schouten, 'AI Literacy 101 — What is it and why do you need it?', *Medium*, Aug. 25, 2020. https://towardsdatascience.com/ai-literacy-101-what-is-it-and-why-do-you-need-it-73238ec7c2db (accessed Nov. 02, 2022).

[9] C. Nwosu, 'Visualizing The 50 Biggest Data Breaches From 2004–2021', *Visual Capitalist*, Jun. 01, 2022. https://www.visualcapitalist.com/cp/visualizing-the-50-biggest-data-breaches-from-2004-2021/ (accessed Nov. 15, 2022).

[10] S. Schmeelk, 'Where is the Risk? Analysis of Government Reported Patient Medical Data Breaches', in *IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume*, New York, NY, USA, Oct. 2019, pp. 269–272. doi: 10.1145/3358695.3361754.

[11] B. Fowler, 'Data breaches break record in 2021', *CNET*. https://www.cnet.com/news/privacy/record-number-of-data-breaches-reported-in-2021-new-report-says/ (accessed Nov. 02, 2022).

APPENDIX

Dataset	Attribute Name	Attribute Description	Attribute Type
car-assignments	LastName	Last name of employee (text).	Categorical 45 non-unique labels
	FirstName	First name of employee (text), 45 unique labels.	Categorical 45 unique labels
	CarID	Numeric label.	Categorical (0-35) or blank (if employee title is “truck driver”)
	CurrentEmployeeType	Text label of employee classification.	Categorical 45 non-unique labels
	CurrentEmployeeTitle	Text label of title.	Categorical 45 non-unique labels
cc_data	timestamp	Time (date, hour and minute).	Interval 1490 non-unique values.
	location	Text label of a store, restaurant or establishment.	Categorical 1490 non-unique values.
	price	Numeric value for the cost charged to a specific card.	Interval 1490 non-unique values
	last4ccnum	Numeric label.	Categorical 4 digit label, 1490 non-unique labels.
gps	Timestamp	Time (date, hour and minute).	Interval 685169 non-unique values.
	id	Numeric label.	Categorical (0-107)
	lat	Latitude position at a given time.	Ratio 685169 non-unique values.
	long	Longitude position at a given time.	Ratio 685169 non-unique values.
loyalty_data	Timestamp	Time (date, hour and minute).	Interval 1392 non-unique values.
	location	Text label of a store, restaurant or establishment.	Categorical 1392 non-unique values.
	price	Numeric value for the cost charged to a specific card.	Interval 1392 non-unique values
	loyaltynum	Text label of employee classification.	Categorical 1392 non-unique labels

Table: Data Attributes