Modified VAST Challenge with Applications to Data Breaches

Rosalyn Carr*

University of British Columbia

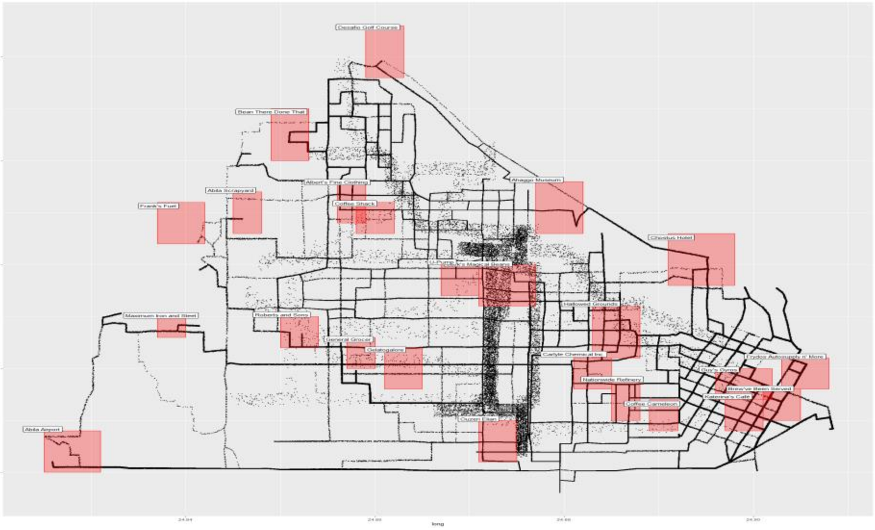


Fig 1: Aggregate view of global positioning system (GPS) data of artificial individuals. Locations of interest are outlined in red.

ABSTRACT – Data breaches are becoming an increasingly common and costly threat to personal privacy. Even then, most information available to lay people is dense, repetitive and focuses on showing preventative measures. This works in contrast to the target audience of many of these pieces, which is users who have already experienced a data breach (and therefore past the period of prevention). Data breaches are also difficult topics to approach directly as using any leaked data as an example for further learning raises ethical concerns. In this context, the 2021 VAST Challenge Mini Challenge 2 dataset was recontextualized as a potential surrogate for potential tools for spreading awareness among lay users of the severeness of data breaches.

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 personal 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 as multiple breaches can be linked together (for example, a repeated email or password in one breach could give access to information hosted on other platforms) [3]. A key starting point in educating individuals is understanding how simple the deanonymization process can be for a large portion of data, especially when multiple pieces of information can be linked. This data linking and deanonymization process, although computationally complicated, is often not conceptually difficult. This conceptual process is a key area for education of lay users.

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. Although full implementation of this tool was not possible within the time frame of this project due to the complexities and issues specific to the VAST provided data set, preliminary interactive tools showed potential for a future positive public implementation.

* e-mail: rosalyncarr@ieee.org

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

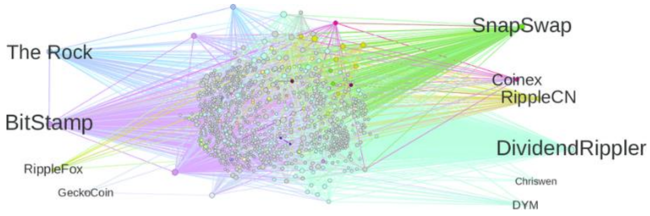
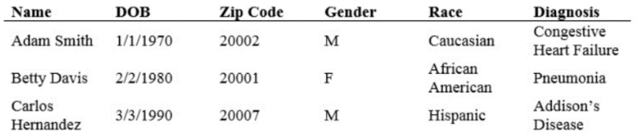


Fig 2: A visualization of the deanonymization process over a clustered graph [11].



Name	DOB	Zip Code	Gender	Race	Diagnosis
Adam Smith	1/1/1970	20002	M	Caucasian	Congestive Heart Failure
Betty Davis	2/2/1980	20001	F	African American	Pneumonia
Carlos Hernandez	3/3/1990	20007	M	Hispanic	Addison's Disease

Fig 3: An example of how data is shown in publication related to deanonymization processes [12].

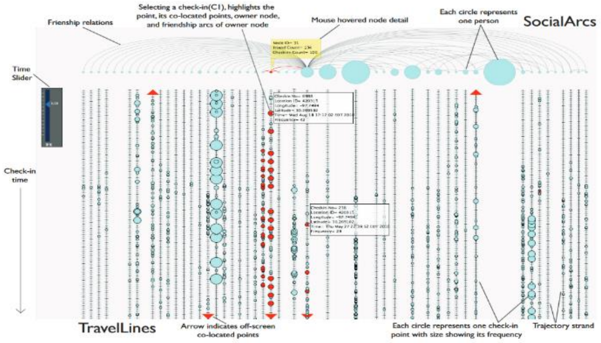

Friendship relations
Selecting a check-in(C1), highlights the point, its co-located points, owner node, and friendship arcs of owner node
Mouse hovered node detail
Each circle represents one person
SocialArcs
Time Slider
Check-in time
TravelLines
Arrow indicates off-screen co-located points
Each circle represents one check-in point with size showing its frequency
Trajectory strand

Fig 4: GSUVis tool visualization of deanonymized data showing the connections between users in a location-based social network when a user hovers over a specific node [13].

2.3 Deanonymization

2.3.1 Visualizing Deanonymization

Transaction data is core to the provided VAST challenge dataset, involving what is often geospatial data alongside identifiers including cards numbers and purchase prices. The deanonymization process often involving linking these data to other available identifiers. Although there have been attempts to involve visualizations in deanonymization as seen in Figure 2 [11], these visualizations focus heavily on visualizing the results of the process and not the methods that were used. Many articles tend to focus heavily on explaining a novel technique and use various static flowcharts to explain the process [12], [14], but these are often specific to the novel process and have a certain level of technical jargon (as is appropriate for academic but not for lay people). What is a notable standard across many papers is the use of tables as seen in Figure 3 to explicitly show the raw attributes of the data rather than building more traditional visualizations such as charts or plots that show relationships between attributes.

One exception is when deanonymization is targeted at a social network scale rather than an individual scale. The GSUVis tool by Tarameshloo et al. introduced an interactive method to show the benefits of anonymizing data [13]. Users can switch between anonymized and deanonymized views of the system to show how collocated data can be used to show possible relationships and how

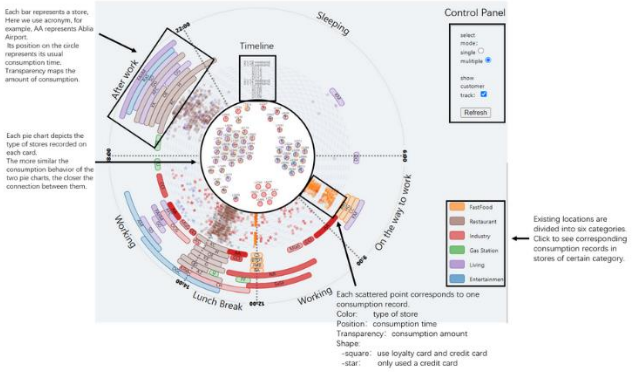

Each bar represents a store, Here we use acronym, for example, AA represents Ablia Airport. Its position on the circle represents its usual consumption time. Transparency maps the amount of consumption.
Each pie chart depicts the type of stores recorded on each card. The more similar the consumption behavior of the two pie charts, the closer the connection between them.
Sleeping
After work
Timeline
Working
Lunch Break
Working
On the way to work
22:00
18:00
14:00
12:00
9:00
6:00
Control Panel
select mode:
single ○
multiple ●
show customer track: ☑
Refresh
FastFood
Restaurant
Industry
Gas Station
Living
Entertainmen
Existing locations are divided into six categories. Click to see corresponding consumption records in stores of certain category.
Each scattered point corresponds to one consumption record.
Color: type of store
Position: consumption time
Transparency: consumption amount
Shape:
-square: use loyalty card and credit card
-star: only used a credit card

Fig 5: An example of a solution to the first task of 2021 VAST Challenge Mini Challenge 2, which is to locate data discrepancies and anomalies [15].

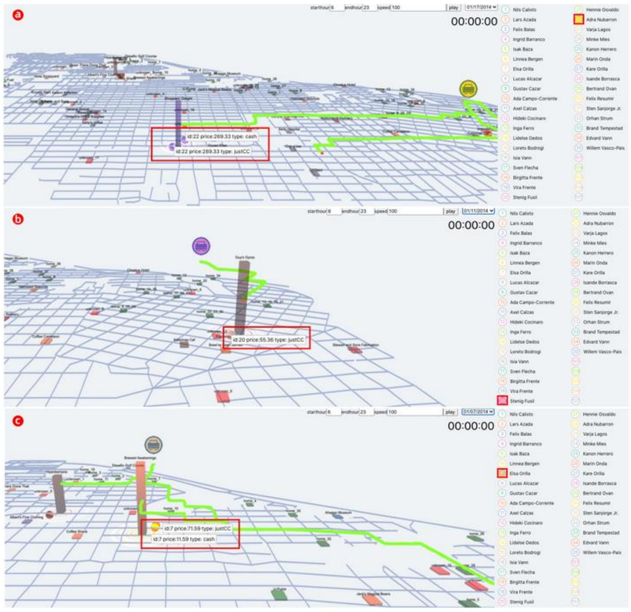
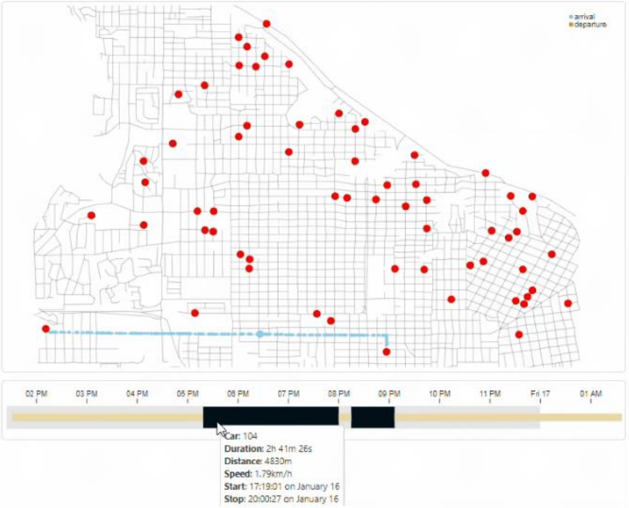

a
b
c
00:00:00
00:00:00
00:00:00
id:22 price:269.33 type: cash
id:22 price:269.33 type: justCC
id:20 price:55.38 type: justCC
id:7 price:71.59 type: justCC
id:7 price:11.59 type: cash

Fig 6: An example of a solution to the first task of 2021 VAST Challenge Mini Challenge 2, showing a map view of the GPS data and vehicle id's [15].

this ability is lost when the data is anonymized. This tool was similarly developed on a synthetic dataset and was tested with domain experts on privacy. Although this tool was very well received by the domain experts through its ability to not only show specific trends (such as which locations a user might visit the most frequently, as well as when and who they go with), its scope was not aligned with lay users. The domain experts specifically saw value in educating their clients, specifically people who were already invested in their own online privacy. These experts had not encountered a similar tool for visualizing location-based social networks and could see multiple applications for the tool when an expert is present to explain the process and severity of trends.

2.4 VAST 2021 Solutions

Although solutions to the specific challenge are not contextualized similarly to this project, the original 2021 VAST Challenge Mini Challenge 2 does require participants to solve (or deanonymize) the data through visualizations. There was a total of seven solutions available alongside the official solution (which does not include visualizations). Two of the solutions were notable,



Car: 104
Duration: 2h 41m 26s
Distance: 4830m
Speed: 1.79km/h
Start: 17:19:01 on January 16
Stop: 20:00:27 on January 16

Fig 7: An example of a solution to the first task of 2021 VAST Challenge Mini Challenge 2, which shows an interactive tool that allows participants to scroll through GPS data [16].

however, it was interesting that generally the solutions to the challenge followed a pattern of using heavy computational practices to deanonymize the data and a visualization was built to show the results (which aligns with most academic practices). Similarly, visual idioms were only developed to the point of success, and many lacked polishes that would be needed before given to any user unfamiliar with the dataset (these visualizations were deliberate tools to solve the tasks and assumed that the human-in-the-loop would be a domain expert).

2.4.1 Data Star Observatory

The solution proposed by Data Star Observatory stood out as a solution provided with primarily visual encodings of the data [15]. Although very thorough and successful in completing the tasks, the actual visual encodings are quite noisy and are not meant to be legible by a user unfamiliar to the data, as seen in Figure 5 and Figure 6. However, this may be due to the interactivity lacking in this presentation of the tool (screenshots rather than the actual tool) so it is difficult to speculate how effective these tools would be out of this specific context.

Another interesting feature of this solution is the reliance on data annotations and labelling. It is not clear how these labels were generated, but much of the visual encoding (such as colour) was reliant on these labels.

2.4.2 UHasselt

The solution provided by UHasselt included a large variety of simple, static displays alongside demonstrations of an interactive tool for exploring the spatialtemporal data [16]. Although many of the idioms were very basic and repeated the same idiom, there was a strength in how thorough the visualizations were in explaining the computational methods done between visual encodings.

The solution however was not as successful in revealing some trends. The repeated bar charts, although good revealing outliers and anomalies, did lead to some complexities in the deanonymization process. This is not particularly notable to the challenge as much of the original goals were aligned with identifying specifically the abnormal behaviour among employees, however for the purpose of this project, this is important to recognize when trying to communicate the process to non-domain experts.

Dataset	Attribute Name	Attribute Description	Attribute Type
tourist map	.jpg image	Image of notable locations in the city.	Image
abila	abila	Street coordinates and labels for the city.	ESRI project files; corrupted.
kronos	kronos	Geospatial data with island shape.	ESRI project files
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 1: Data Attributes. Note that the geospatial files for Abila were not functional.

3 Data and Task Abstraction

Data and tasks were provided by the original 2021 VAST Challenge Mini Challenge 2, however these were adapted to fit the intents of this project regarding visualizing for lay users.

3.1 Domain

Data breaches affected hundreds of millions of individuals each year, however the data is still sensitive [17]. 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 deanonymization. The original context is a company, GASTech, is concerned about the actions of their employees. The company has significant control over the city, being the primary place of employment for many citizens and owning the majority of businesses. Knowing this, the company has attached geospatial trackers to company cars to track employee movements both during and outside of work hours. Additionally, they track both company-used credit card, tracking the purchase price, time, number and business name, along with company issued loyalty cards which track similar information but the time is aggregated by day [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 as well as identify other personal relationships not publicly declared, and to report the suspicious behaviour to law enforcement [4]. Because of this, there are both usual life patterns of "good" employees as well as outliers who will act differently than employees of a similar job description.

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; data that was made public in some way. Geospatial data is often collected through cellular devices or any device equipment with Bluetooth, and credit cards are a commonly targeted piece of information in breaches due to the possibility for quick financial gain.

3.2 Data Abstraction

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 Table 1 on the previous page. It is important to note that ESRI, the program used to develop the geospatial data, no longer supports a student license. The data was found to not be fully complete and therefore was generally not recoverable, meaning that although it is included in the data abstraction it was not usable in whole during this project.

3.3 Task

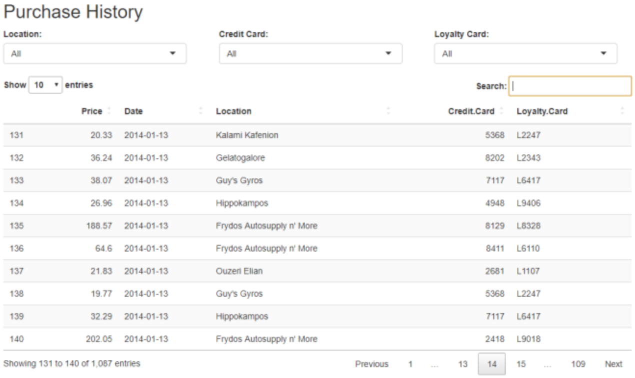
The original tasks presented by the 2021 VAST Challenge Mini Challenge 2 implicitly followed the deanonymization process, however the primary tasks for this project related to communication with lay users and not the domain experts seen in the original challenge. [4].

T1	Infer the owners of each credit card and loyalty card
T2	Understand the deanonymization process that was used to infer the owners

The first task was chosen as an amalgamation of the original tasks presented in the Mini Challenge, specifically focusing on the deanonymization of the data given. The second task is specific to this project, which is allowing users to form their own understand of the process such that they

3.3.1 Task Abstraction

As a general goal, the users would be able to search through the data in a way that does not require significant domain knowledge. Users should be able to complete most *search* actions, though *exploring* and *browsing* will be priority actions. *Exploring* will be important as not only is this an artificial dataset which users will be unfamiliar with but *browsing* will play a larger role across multiple idioms as users become familiar with the data and will look for specific patterns shown by specific artificial individuals.



Purchase History

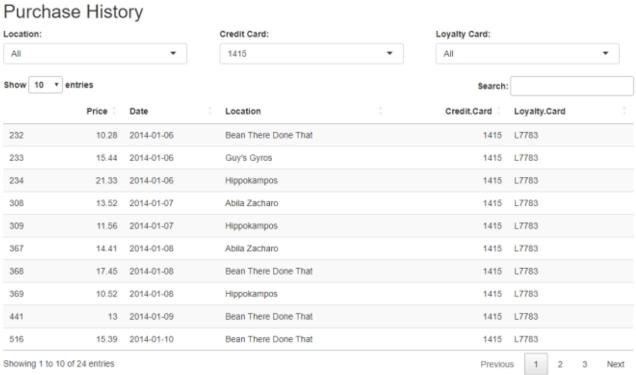
Location: All
Credit Card: All
Loyalty Card: All

Show 10 entries
Search:

	Price	Date	Location	Credit.Card	Loyalty.Card
131	20.33	2014-01-13	Kalami Kafenion	5368	L2247
132	36.24	2014-01-13	Gelatogalore	8202	L2343
133	38.07	2014-01-13	Guy's Gyros	7117	L6417
134	26.96	2014-01-13	Hippokampos	4948	L9406
135	188.57	2014-01-13	Frydos Autosupply n' More	8129	L8328
136	64.6	2014-01-13	Frydos Autosupply n' More	8411	L6110
137	21.83	2014-01-13	Ouzeri Elian	2681	L1107
138	19.77	2014-01-13	Guy's Gyros	5368	L2247
139	32.29	2014-01-13	Hippokampos	7117	L6417
140	202.05	2014-01-13	Frydos Autosupply n' More	2418	L9018

Showing 131 to 140 of 1,087 entries

Fig 8: A general view of the first interactive display, showing the results of the linked credit card and loyalty card purchases.



Purchase History

Location: All
Credit Card: 1415
Loyalty Card: All

Show 10 entries
Search:

	Price	Date	Location	Credit.Card	Loyalty.Card
232	10.28	2014-01-06	Bean There Done That	1415	L7783
233	15.44	2014-01-06	Guy's Gyros	1415	L7783
234	21.33	2014-01-06	Hippokampos	1415	L7783
308	13.52	2014-01-07	Abila Zacharo	1415	L7783
309	11.56	2014-01-07	Hippokampos	1415	L7783
367	14.41	2014-01-08	Abila Zacharo	1415	L7783
368	17.45	2014-01-08	Bean There Done That	1415	L7783
369	10.52	2014-01-08	Hippokampos	1415	L7783
441	13	2014-01-09	Bean There Done That	1415	L7783
516	15.39	2014-01-10	Bean There Done That	1415	L7783

Showing 1 to 10 of 24 entries

Fig 9: A view of the first interactive display, showing the results of the linked credit card and loyalty card purchases when filtering for a specific credit card.

Finding *dependencies* between attributes will also be a key target for this project – deanonymization requires a significant amount of data linking, and dependencies between attributes is absolutely core to this process.

4 Solution

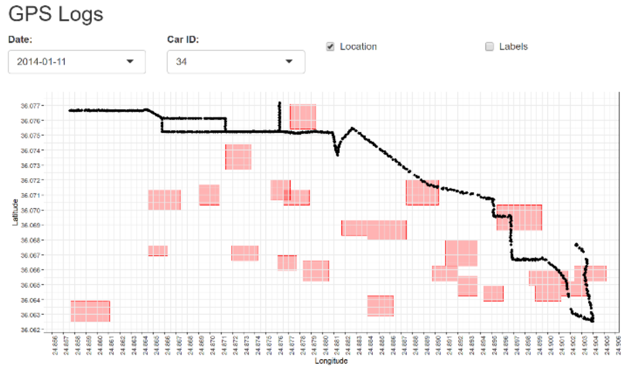
The final solution involved 3 interactives tools to allow users to search through the provided 2021 VAST Challenge Mini Challenge 2 datasets alongside generated annotations and identifiers.

4.1 Credit Card and Loyalty Card Linking

The first implemented idiom was an interactive data table that showed the results of the first step of data linking between the credit card purchases and the loyalty card purchases. The idiom has a number of filtering options for users to use to explore the data, specifically allowing users to sort by location, credit card number, and loyalty card number to support the *exploring* and *browsing* actions.

4.2 GPS Navigator and Location Mapping

The second interactive implementation was a similar interactive tool allowing users to sort through the data given, but



GPS Logs

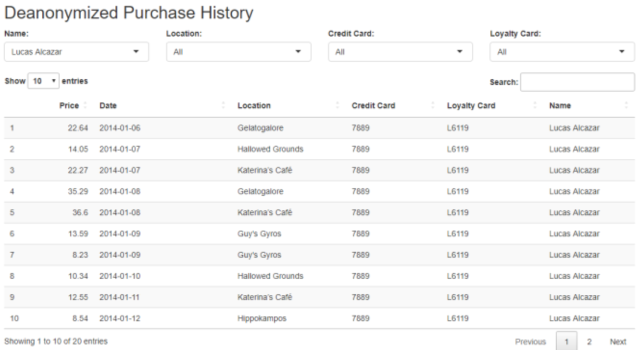
Date: 2014-01-11

Car ID: 34

☑ Location

□ Labels

Fig 10: A view of the second interactive display, showing the gps position for each car ID for each day of recorded data. Locations of interest from the Tourist Map are shown in red, with optional labels.



Deanonymized Purchase History

Name: Lucas Alcazar

Location: All

Credit Card: All

Loyalty Card: All

Show 10 entries

Search:

	Price	Date	Location	Credit Card	Loyalty Card	Name
1	22.64	2014-01-06	Gelatogalore	7889	L6119	Lucas Alcazar
2	14.05	2014-01-07	Hallowed Grounds	7889	L6119	Lucas Alcazar
3	22.27	2014-01-07	Katerina's Café	7889	L6119	Lucas Alcazar
4	35.29	2014-01-08	Gelatogalore	7889	L6119	Lucas Alcazar
5	36.6	2014-01-08	Katerina's Café	7889	L6119	Lucas Alcazar
6	13.59	2014-01-09	Guy's Gyros	7889	L6119	Lucas Alcazar
7	8.23	2014-01-09	Guy's Gyros	7889	L6119	Lucas Alcazar
8	10.34	2014-01-10	Hallowed Grounds	7889	L6119	Lucas Alcazar
9	12.55	2014-01-11	Katerina's Café	7889	L6119	Lucas Alcazar
10	8.54	2014-01-12	Hippokampos	7889	L6119	Lucas Alcazar

Showing 1 to 10 of 20 entries

Fig 10: A view of the third interactive display, showing the results of the linked credit card and loyalty card purchases when deanonymized from the GPS data and car assignments data.

this included the option of viewing and browsing the GPS data for different car IDs and dates. To reduce visual clutter as well as computational complexity, only one single day and one single car ID is viewable at a time, though users can switch between these through the attribute menus.

4.3 Deanonymized Dataset

The final implemented idiom was a similar interactive data table to the first, however this showed the results of the second step of data linking between the credit card purchases and the loyalty card purchases and the GPS data. The idiom similarly has a number of filtering options for users to use to explore the data, specifically allowing users to sort by employee name, location, credit card number, and loyalty card number to support the *exploring* and *browsing* actions.

5 Implementation

Exploration of the datasets and preliminary visualizations for tracking anomalies were performed in R using the ggplot2, lubridate, dplyr and sf packages [18]–[21]. All interactive visualizations were built with R Shiny and ggplot2 [18], [22], with the later goal of the tool being publicly available on a web platform, however this was not implemented due to various constraints.

5.1 Identifying Anomalies

The original VAST challenge included multiple data anomalies that would ideally have been removed prior to the final

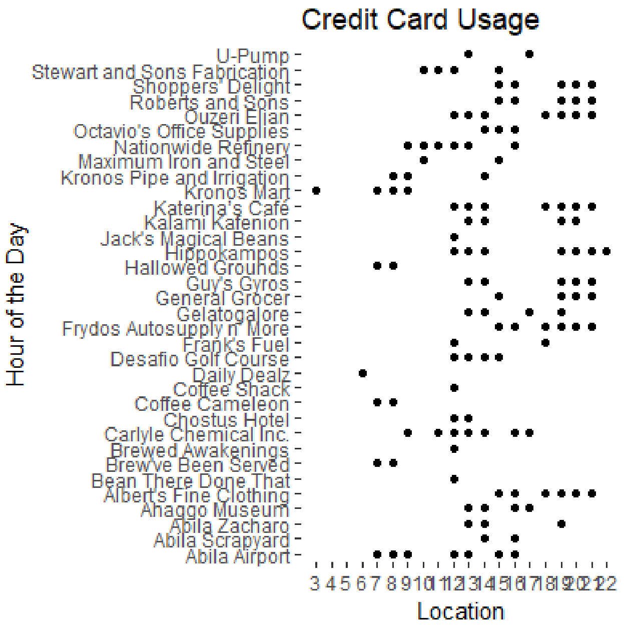


Fig 11: Aggregated Credit Card Usage

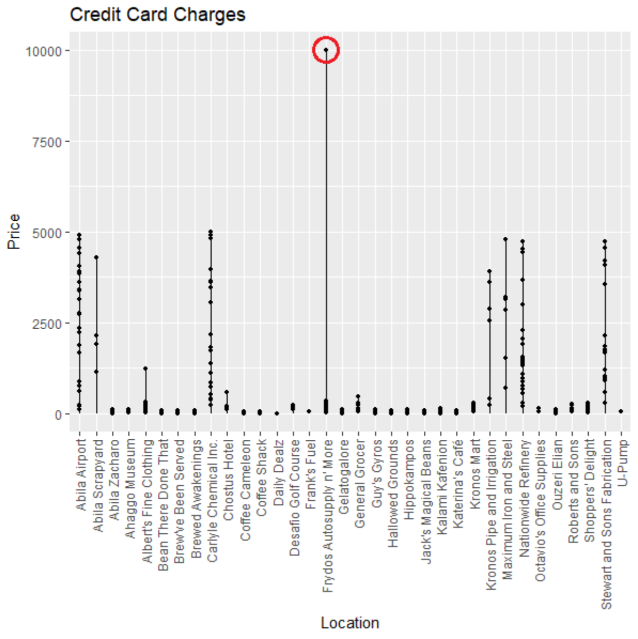


Fig 12: Aggregated credit card charges with an outlier charge highlighted in red.

Implementation to avoid user confusion. This was not feasible in the timeframe of the project but was informative in the process of becoming familiar with the data.

To determine which data would have been candidates for removal the original dataset was visualized in R using ggplot2 as seen in Figure 11. Starting with the credit card and loyalty card datasets, charges to the credit cards and charges made alongside the loyalty cards were visualized to identify any outliers. The following visualizations were produced, seen in Figure 12, and one outlier as a $10,000 charge to "Frydos Autosupply n' More" was identified as a planted anomaly.

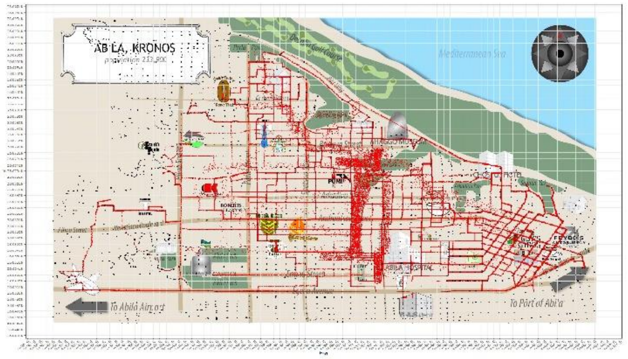


Fig 12: Overlaid GPS data and unlabeled geospatial points to recover location coordinates.

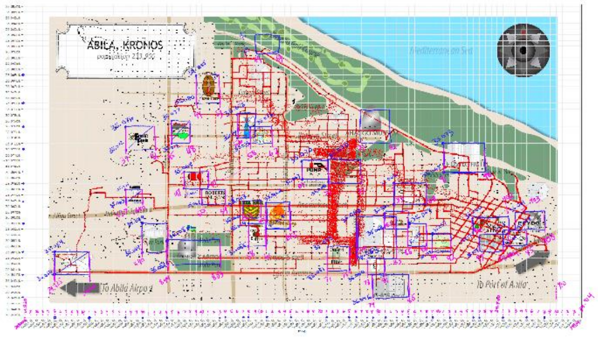


Fig 13: Overlaid GPS data and unlabeled geospatial points to recover location coordinates with annotated location coordinates.

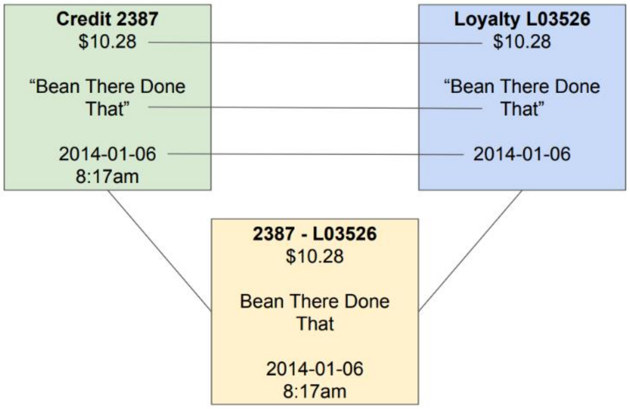


Fig 14: Diagram showing the condition for matching purchases across Credit and Loyalty cards.

After, the credit card data and loyalty card data are cross referenced against each other for similarities. In order to discover which loyalty card user matches which credit card user, the purchase histories must be compared. Although the loyalty card data does not include a time of purchase and only includes the day, it can be assumed that if a transaction is listed on both the credit card and loyalty card datasets with a matching date, price, and location as seen in Figure 14, there parsed and a list of 1087 matches were found. All 1087 matches were grouped into their own dataset and included in the first interactive tool available for users.

5.2 Geospatial Data

The geospatial data provided for the original challenge ended up being unworkable. ESRI [23], the program used to generate the artificial maps, no longer offers affordable licenses, and most other free versions can only handle portions of the data (which generally removes labels). Additionally, ESRI is a powerful program that is designed to handle incomplete data like what was generated for the challenge; specifically handling linestrings that are not fully defined. An unneeded amount of time was spent attempting to construct the incomplete linestrings (with help from multiple classmates in excel, python, and R) and was ultimately unsuccessful. Instead, the points that would have made up each incomplete linestring that should have denoted a labelled street were extracted using R sf and overlaid with the GPS data to create a grid. This gride was adjusted to fit the Tourist map image and each location was manual annotated and the coordinates recorded. This data was used to create the red "zones" on the interactive GPS tool, with the annotations also available for view.

5.3 Owner Identification

The only available dataset that lists employee's identity is the car assignments dataset, which outlines which company car is assigned to which employee. This data can be compared against the GPS data, which also lists the company vehicle number.

The first step in identifying the owners of each card is to correlate the GPS data with the map supplied by the VAST challenge. Using coordinate locations, the business locations shown on the credit card and loyalty card charges can be identified. From there, a similar process to the credit card and loyalty card matching can be performed. It can be assumed that if a car is at a location at the time of a purchase on either the loyalty or credit card, that car (and its owner) are a candidate for being an owner of said card. This is not a definitive approach however, as multiple charges are often made by different cards in very short time periods. Instead, candidacy was determined through a process of elimination, where once a card has a determinate owner (such as through an isolation purchase log), that owner is removed from all further candidacies. As an interactive process, this was completed manually due to continued issues with computational complexity and continued geospatial issues.

6 Discussion and Further Work

6.1 Unrealistic Project Scope

One key area the author identified over the course of the project was the incorrect and unfeasible scope. Although it had initially seemed feasible to recontextualize the original 2021 VAST Challenge Mini Challenge 2 and build an additional tool, there was oversight in how the reduction of scope for the challenge would go. More careful considerations would have revealed that all preceding tasks in the challenge prior to reidentification are necessary to achieve reidentification, so removing prior tasks did not remove the scope of work. Completely the challenge in full would have been a difficult but feasible challenge (especially given most solutions did not aim to have a polished view for non-domain expert users) but trying to broaden the applications of the data past the challenge was an irresponsible move that showed in the final scope of implementation.

6.1.1 Difficulties of Determining Level of Abstraction

One key question that came up throughout the project but was unsolvable was the extent that the process of deanonymization needed to be abstracted from users. The deanonymization process and data linking are incredibly complex, especially when dealing with time and space margins due to uncertainties and non-ideal conditions. Further work might either push for a simpler and more ideal dataset if it is determined lay users cannot understand more complexity beyond an ideal case, or an expansion of the current approach to include more steps of the process.

6.2 Unworkable Spatial Data

Although the manual labour put into this project seemed excessive, the presence of a seemingly similar amount of effort into manual labels in other existing VAST solutions implies this was a common path most participants of the challenge took. However, the extent that manual intervention was needed was not originally a problem the challenge had intended.

The unworkable geospatial data issues led to the less accurate method of manual annotation that led to some results (which later required manual reidentification as there was not adequate computation power to work with the large margins of error). These issues were not designed to be a obstacle in the original challenge as ESRI was more wildly available at the time of creation, however this did significantly impede progress and meant final design decisions and implementations were infeasible.

7 Conclusion

Data breaches are a serious risk for anyone who uses online services, especially those who are not aware of privacy best practices. Although articles of preventative measures are available, they often try to reach users after their data has already been compromised, which limits its effectiveness. Discussing breaches and how the process of deidentification works directly is risky as the data is inherently sensitive, bringing up the possibility of the effectiveness of a surrogate dataset. The 2021 VAST Challenge Mini Challenge 2 dataset was presented as an option if recontextualized, which is possible if further time was spent to overcome technical difficulties that complicate the dataset outside what is realistic.

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] P. Moreno-Sanchez, M. Zafar, and A. Kate, ‘Listening to Whispers of Ripple: Linking Wallets and Deanonymizing Transactions in the Ripple Network’, *Proc. Priv. Enhancing Technol.*, vol. 2016, Feb. 2016, doi: 10.1515/popets-2016-0049.

[12] ‘Re-Identification of “Anonymized” Data’, *Georgetown Law Technology Review*, Apr. 12, 2017. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/ (accessed Nov. 02, 2022).

[13] E. Tarameshloo, M. H. Loorak, P. W. L. Fong, and S. Carpendale, ‘Using Visualization to Explore Original and Anonymized LBSN Data’, *Comput. Graph. Forum*, vol. 35, no. 3, pp. 291–300, Jun. 2016, doi: 10.1111/cgf.12905.

[14] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, ‘Estimating the success of re-identifications in incomplete datasets using generative models’, *Nat. Commun.*, vol. 10, no. 1, Art. no. 1, Jul. 2019, doi: 10.1038/s41467-019-10933-3.

[15] ‘Data Star Observatory-Gao-MC2’. http://visualdata.wustl.edu/varepository/VAST%20Challenge%202021/challenges/MC2/entries/Data%20Star%20Observatory/ (accessed Dec. 16, 2022).

[16] ‘UHasselt-Aerts-MC2’. http://visualdata.wustl.edu/varepository/VAST%20Challenge%202021/challenges/MC2/entries/UHasselt/ (accessed Dec. 16, 2022).

[17] 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).

[18] H. Wickham, *Ggplot2: Elegant graphics for data analysis*, 2nd ed. Cham, Switzerland: Springer International Publishing, 2016.

[19] ‘Dates and Times Made Easy with lubridate | Journal of Statistical Software’. https://www.jstatsoft.org/article/view/v040i03 (accessed Nov. 22, 2022).

[20] H. Wickham, R. François, L. Henry, and K. Müller, ‘dplyr: A Grammar of Data Manipulation’, 2022. https://dplyr.tidyverse.org/ (accessed Nov. 22, 2022).

[21] ‘Simple Features for R’. https://r-spatial.github.io/sf/ (accessed Dec. 16, 2022).

[22] ‘Shiny’. https://shiny.rstudio.com/ (accessed Dec. 16, 2022).

[23] ‘Buy GIS Software | ArcGIS Product Pricing - Esri Canada Store’. https://www.esri.ca/en-ca/store/overview (accessed Dec. 16, 2022).

APPENDIX – MILESTONES

Milestone	Sub Milestones	Estimated Time / Date	Actual Time / Date
Pitch	• Finding Dataset • Building Slides	3 hours / Sept 28	3 hours / Sept28
Proposal	• Data Abstraction • Related Work Research • Writing	6 hours / Oct 21	6 hours / Oct 28* *Extension due to change of topic
Data Management	• Static Views of Provided Data to Explore Possible Idioms • Data Linking Credit Card and Loyalty Card Purchase Logs • Loading Geospatial ESRI Files to Convert Tourist Map to GPS Coordinates • Data Linking Credit Card and GPS Data • Interactive Linking to Remove Missing and Outlier Values • Selection of Candidate Individuals for Final Dataset	30 hours / Dec 1st	100 hours / Incomplete* *ESRI removed student access, meaning geospatial projects could not be opened; significant time was spent exploring alternatives, but manual processing was eventually the only option to extract business coordinates. This manual processing eventually impacted the data linking between the GPS and credit card purchase logged, which also had to be done mostly manually.
Update	• Writing	5 hours / Nov 15th	5 hours / Nov 20th* *Extension due to change of topic.
Peer Review	• Review	2 hours / Nov 15th	2 hours / Nov 15th
Implementation	• Exploration of Possible Idioms on Reduced Dataset • Learning of R Shiny • Building and Testing Idioms • Integration of R Shiny Applications into Single View	35 hours / Dec 12th	20 hours / Dec 14th* *Time learning R Shiny was much longer than expected, including management of missing data expected in data linking. Time lost to data management issues meant final idioms were not built.
User Interaction	• Exposure of Integrated Views to Test Users	5 hours / Dec 13th	2 hours / Incomplete* *Loss of time to data management meant final implementation was not possible.
Final Presentation	• Gathering Materials • Building Slides	4 hours / Dec 14th	4 hours / Dec 14th
Final Report	• Expanding Related Work • Updating Task and Data Abstraction • Writing	10 hours / Dec 16th	10 hours / Dec 16th
TOTAL		**100 HOURS**	**152 HOURS**

Table 2: Milestone Breakdown