Data Visualization
Project - Data Breaches in Top Tech Companies
Process Book

Suraj Khamkar	**skhamka@clemson.edu**	**C16741431**
Sravani Pati	**spati@clemson.edu**	**C79908900**

Project Link - https://github.com/khamkarsuraj/data-breaches

Website link - https://khamkarsuraj.github.io/data-breaches/

Contents

Overview:

The primary goal of our project is to construct a bubble chart that accounts for organizations and information lost due to hacking techniques such as email, SSN, credit card, Personal Details, and Full details. Following that, we want to use a bar chart to show the company name on the x-axis and the amount of data lost in millions on the y-axis. Finally, we created a donut graphic to ascertain how many businesses were compromised using a specific hacking technique.

Motivation:

Global business is evolving due to the fast advancement of communication, networking, and information technology, which is expected to last for some time. This evolution offers many benefits and drawbacks for the stakeholders in all organizations. Because they could pose serious problems for all business operations, information security and privacy are topics that information systems management is considering more and more.

Because multinational corporations rely so much on technology and are prone to technical flaws, data breaches and losses are unavoidable. Data is one of a company's most valuable assets, and the risk of losing data control is something that everyone must deal with. No matter what policies and rules businesses implement to lessen the risk of data breaches, hacking and phishing threats remain. Information security and privacy are determinants of a company's continuity and viability. Companies employ risk-reduction strategies, including user and employee orientation to the organization's information security policies and protocols, system authentication, data encryption, user access control, and firewalls. Despite these precautions, criminals are getting more skilled and coordinated, increasing the risk.

Numerous recent examples of businesses that experienced significant data breaches include Equifax, Anthem, eBay, JPMorgan Chase, Home Depot, Yahoo, and Target. Both accounting and information security management face difficulties when evaluating the financial impacts of data breaches (Schatz and Bashroush, 2001).

Although data breaches are common, little is known about people's awareness, perceptions, and reactions to breaches that affect them. Through an online study in which we exposed victims to up to two data breaches that have revealed their email addresses and other personal information, we offer new insights into this subject. Overall, 73% of users were affected by at least one breach, 5.36 breaches on average. Only 14% of users correctly identified external reasons like breached organizations and hackers as the source of being affected by a breach; most victims blamed their email and security habits. 74% of the displayed breaches were unknown to the victim and organization, who reacted differently upon learning about them.

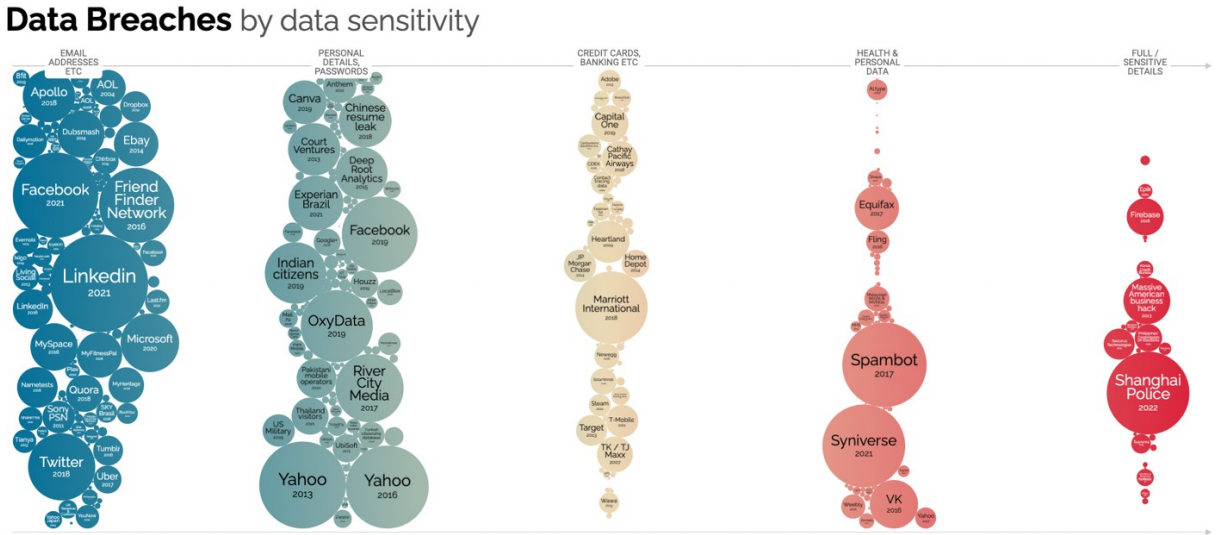
Most users thought they wouldn't be affected by the incident, despite others saying they intended to act. Our findings highlight the amount of data lost per data sensitivity and organizations vulnerable to data breaches. This will help us understand and learn more about how we can secure our data and what action we should take.

All of this information about data breaches at firms and other incidents and accounts of how sensitive data has affected organizations drew our attention to this particular issue for visualization. And we concluded that it is always better to be visually represented because it would immediately grab consumers' attention than to read all preventive measures in words.

Related work:

The research paper we took as a reference is “World’s biggest Data breaches and hacks.” In this Contemporary era, data plays a significant role. To perform some tasks, the primary thing needed is data. Besides advantages, there are vulnerabilities present in data, and this information increases security threats to all individuals. All these points dragged our attention to choose this particular topic. They have considered various organizations that are attacked by data breaches, and the year of the incident and the exciting story is shown visually, which fascinates us. The size indicates the record lost in the particular company.

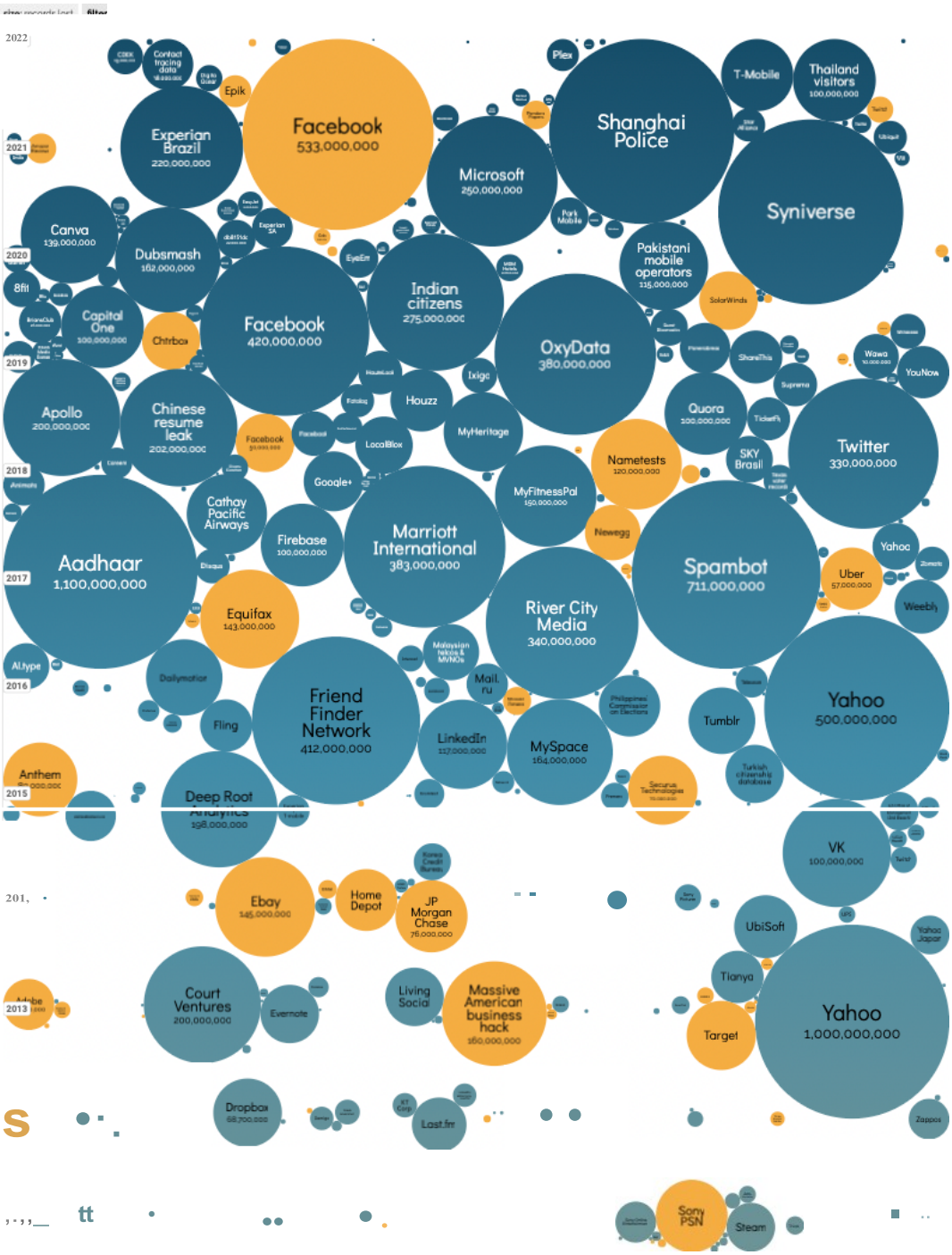
Moreover, it is linked to types of data sensitivity that caused these data breaches in the subsequent visualization they have created. They have used a high-end tool for creating beautiful and interactive data visualizations and stories using VIZsweet. There is only a little data present in this related work, so Dr. Federico suggested we focus on issues we had while doing bubble charts and bar charts. At first, we couldn’t make it more interactive as we had chosen the design choices ideal for our data. Moreover, he suggested that part of our visualization should be modified to address our problem. Finally, considering all his suggestions, we tried to improve our project and make it more interactive.



World's Biggest Data Breaches & Hacks

Selected events over 30.000 records

UPDATE% Sep10:;



Questions:

1) **The number of accounts or data thefts taking data sensitivity or hacking type into account in different organizations?**
   Here, we attempted a bubble chart to display sensitive data, such as email addresses, Social Security numbers, credit cards, and other personal information, distinguished by distinct colors. Additionally, the sizes in the chart below imply data loss in various organizations. The data loss increases with size. Additionally, a slider on the bottom of the bubble chart allows users to view data losses broken down by years.

2) **Variations of data lost as per year in particular organizations?**
   Using a bar chart, we calculated records lost in various companies, displaying records lost in millions on the y-axis and organization names on the x-axis. We then used color to distinguish company names. For all graphs, we used a slider representing years starting from 2017 to 2022. This slider bar chart varies yearly because the number of records lost in that particular year varies.

3) **Total count of particular sensitive data with comparison of organizations?**
   This donut chart represents various hacking methods such as email, SSN, credit card, personal details, and full details. This shows the total number of companies affected by one hacking method. The count of companies that are damaged by sensitive data has been mentioned. This donut chart also varies accordingly with the year. So, the total count of hacking in various companies keeps changing.
   At first, answering all of these questions would be simple. However, when we started creating the visualizations, we realized that there were many things we needed to do with the data. Moreover, we also came across various hurdles. We used numerous JavaScript frameworks and performed a significant amount of data preprocessing. We ultimately achieved the desired result.

Data:

For raw data collection, we first reviewed the news and articles about data breaches in different countries in the last couple of years. After that, we got some pointers to head toward actual data. The actual data we collected was taken from this link, but we worked on it for cleaning data and some rectification of data. Here are some other links and references we used to collect the data.

- https://docs.google.com/spreadsheets/d/1wPgM8ye1AUTVxlZOFsyiKEPWp6iFt34xpp2XA5iM6P0/edit#gid=25233212

- https://docs.google.com/spreadsheets/d/1i0oIJJMRG-7t1GT-mr4smaTTU7988yXVz8nPlwaJ8Xk/edit#gid=2

- https://www.ibm.com/reports/data-breach
- https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/

Data Scraping Method:
We haven't used any automated data scraping or software that automatically scraps data. Instead, we did it manually. We first chose the data and files from various websites and files that we have mentioned, and then, using an excel sheet, we kept the data in the form of tables. We also added and removed unnecessary data by filtering out helpful information.

Data Cleanup:
We did some data cleanup for the links provided above. We extracted the number of data lost for each company per year. So far, the quantities we have derived are the number of accounts lost, year, revenue, country, data recovered, passwords, etc.

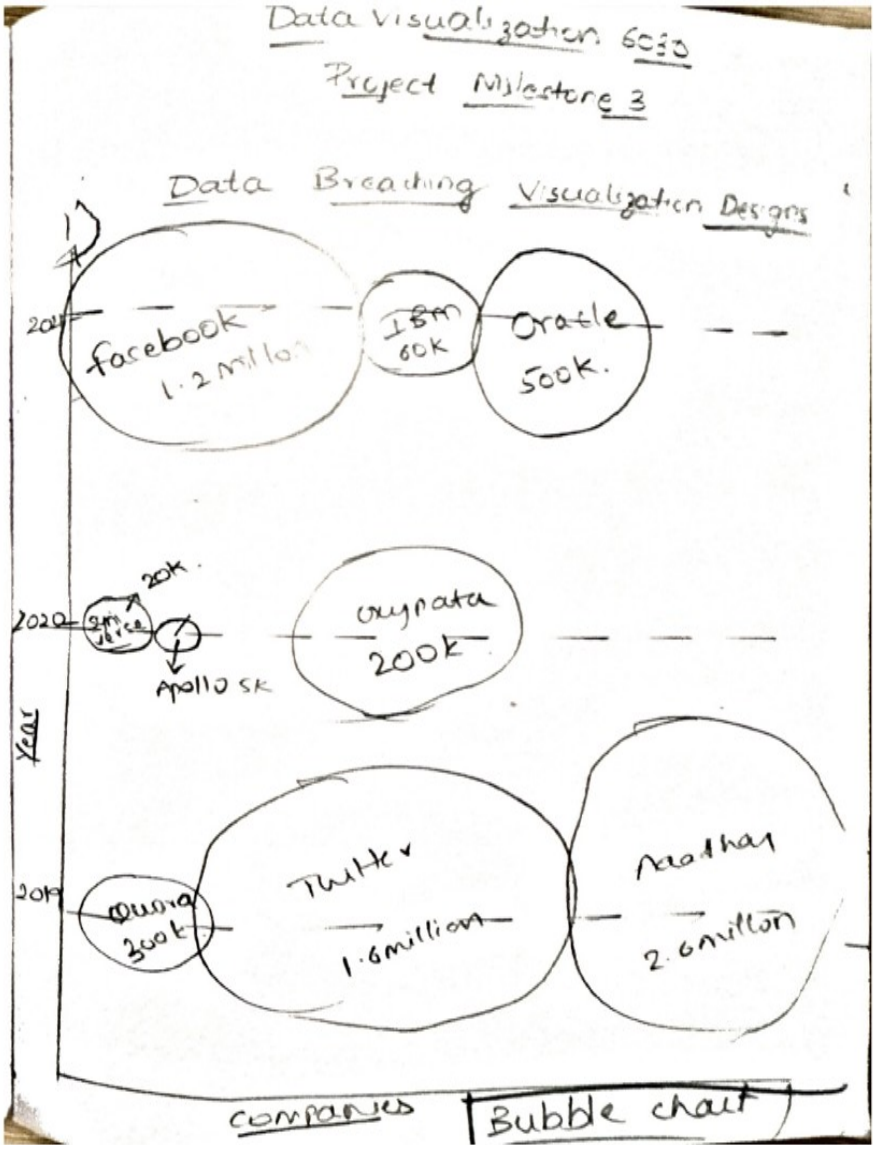
Exploratory Data Analysis:
Initially, we could not design grouped bar chart and bubble chart interactively, and we thought of adding a donut chart, but at first, we didn't add it as it needed to be simpler. But after taking suggestions from Dr. Federico and the help of assignments he gave helped us a lot. He made us learn javascript and any libraries inside javascript to create the visualization. He made us play with the data in all possible ways. At first, he allowed us to use the data we took as a reference and make prototypes for our visualizations using tableau To get the required visualization. Then the fundamental part started. Firstly, we have begun using CSV files. We created a duplicate CSV file to make a bar chart, bubble chart, and donut chart. It has helped to get desired output for our visualizations. Secondly, we created a JSON file and created x-axis and y-axis coordinates and grid cells to make our data look more clearly in our bar chart. We used labels to show hacking methods and what organizations got hacked, and the count of the record lost is also demonstrated in the bubble chart. The donut chart shows the total count of hacking done in various organizations.

Design Evolution:
During every milestone, we can see changes in our project. He constantly gave suggestions and ideas to implement the project. After developing interactions and making them more interactive, we still needed help with missing names for axes and legends. We then cross- checked by asking our professor, and also we discussed more concepts regarding the downside of our project and collected more information to make it perfect. Finally, All our results came the right way, and we gained all the critical insights. These insights are used until our project's end and help us get desired outputs.

Initial Proposal work:

Visualization 1: Bubble Chart

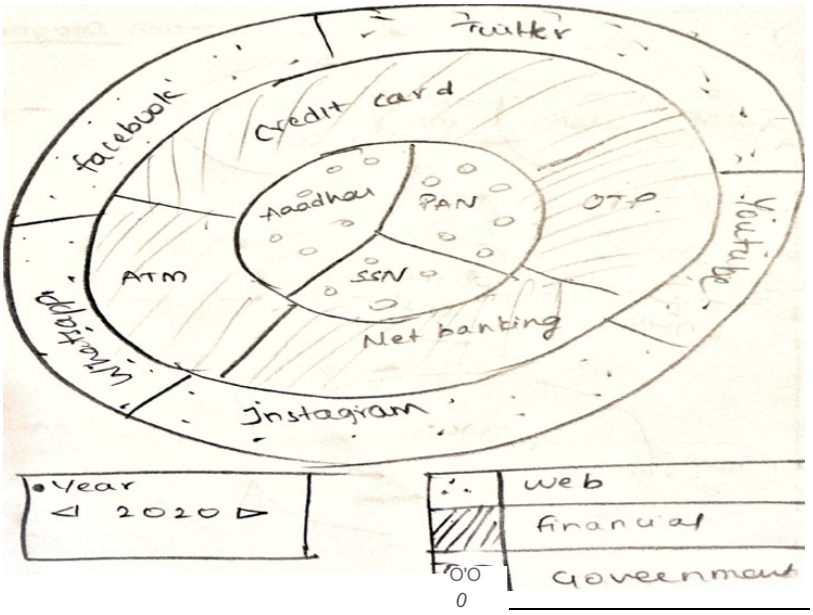
A bubble chart will display relationships and distribution for the number of accounts or data thefts, taking data sensitivity or hacking type into account. However, in this variation, we'll use bubbles in place of the data points. To represent a third kind of data, we will also alter the size of the bubble. A category axis is not used in a bubble chart. Instead, it displays the data sets as X-, Y-, and now Z-values (bubble size).



Visualization 2: Multi-layer pie chart

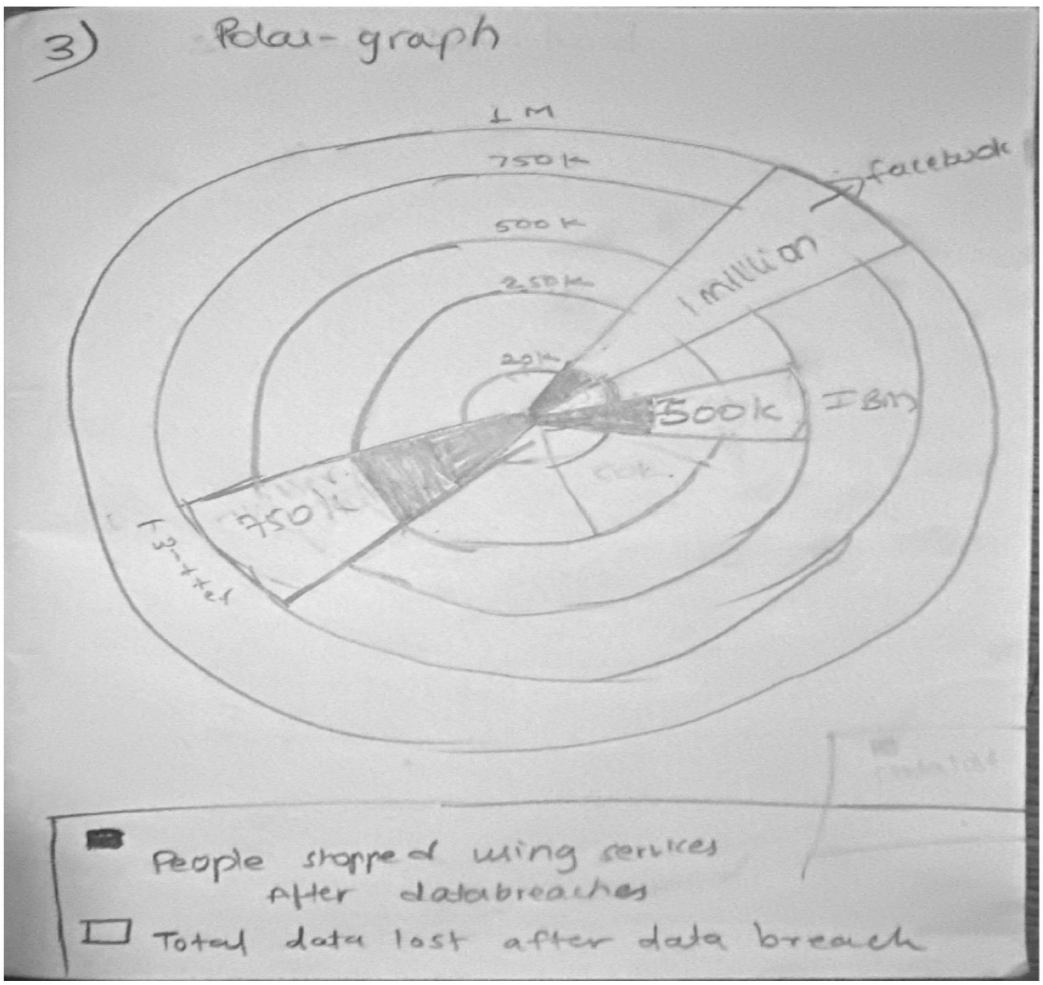
The infographic below shows a multi-layered pie chart that shows variances in data loss in several industries. We can see the web activity sector on the top layer, and the financial industries that fall under each primary category are on the bottom. A thin layer in the middle divides all government sectors into three categories. It takes some planning to have all the classes fit together and be simple to grasp in this sort of data visualization, which makes it more challenging to produce than other types. Technically speaking, this representation consists of three pie charts stacked on top of one another.

2) Multi-layer Pie chart



Visualization 3: Polar Chart

Polar graphs have a circular foundation, but the data is plotted differently. Wedge shapes extend from the center rather than join points together. The main visual distinction is present. Because the data values are so dissimilar, we choose a polar graph. Otherwise, it could not be easy to read quickly. It is ideal for measuring the number of users who ceased using services after data breaches.



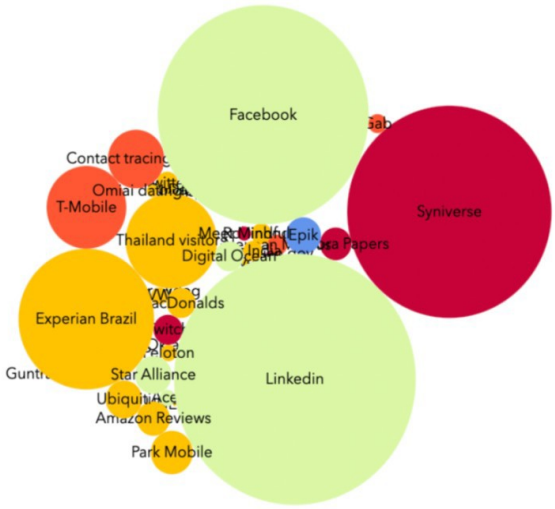
Prototype:

Dr. Federico asked us to write the code in a JavaScript file. Then we started using d3 to do all visualizations. While working on our visualizations, initially, we encountered some issues, but later, with the help of Dr. Federico‘s suggestions and assignments, we tackled the problems that we had before. Halfway, we somehow managed to create the same graphs that we thought back, but we felt a simple bar chart for comparing companies and records lost from 2017-2022 would be appropriate to visualize our dataset. Another thought was that the donut chart would suit the given dataset more instead of a multi-layered pie chart. But the bubble chart remained as it was, leading to our prototype's design.

The number of accounts or data thefts taking data sensitivity or hackingtype into account in different organizations?

Here, we attempted bubble charts to display data lost in data sensitivity, such as email addresses, Social Security numbers, credit cards, and other personal information, distinguished by distinct colors.

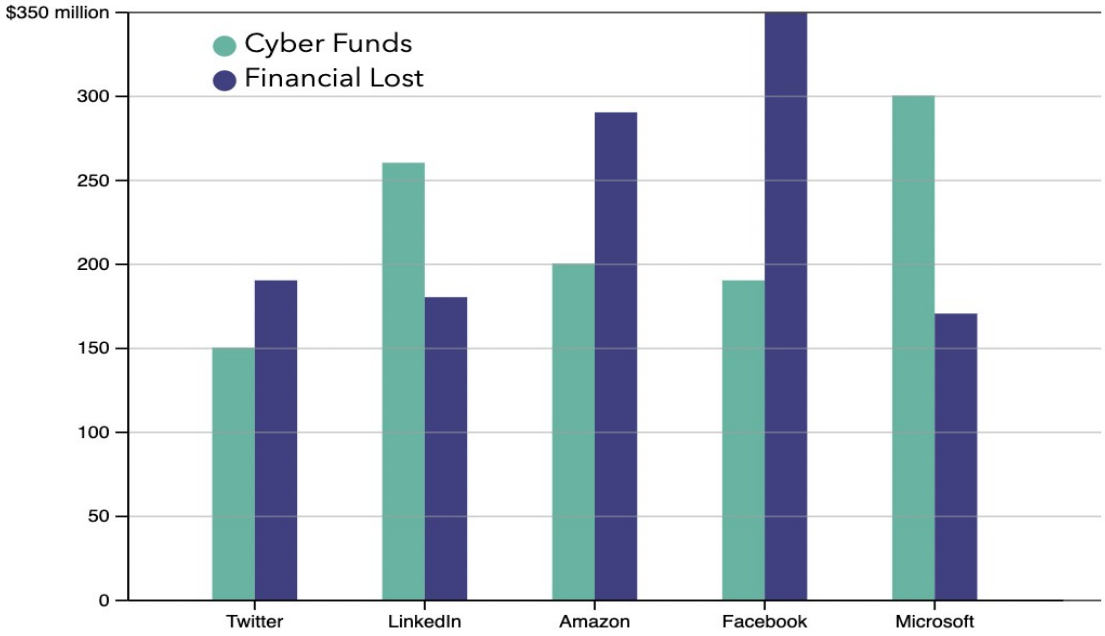
Additionally, the sizes listed in the chart below imply the amount of data loss. The data loss increases with size. There is a drop-down box on the lower left side of the screen where we may view data losses broken down by years. Also, we added a legend for the given visualization to show various data sensitivities in color.



Financial lost caused for companies categorized by sectors as well as fundsallocated for data lost

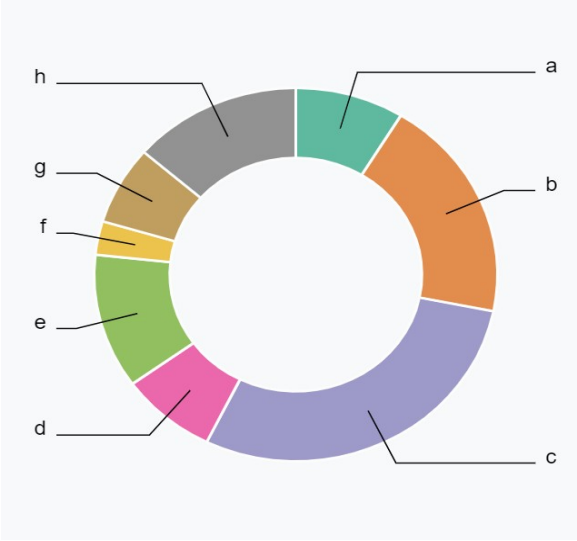
Using a grouped chart, we aimed to show financial losses and the funds allotted for them. We then used color to distinguish between the two. The color light green represents cyber funds, or

money allocated by cyber security to enterprises, while the other hue represents a financial loss for businesses. We tried to display cost in millions on the y-axis, showing numerous tech gains on the x-axis, including Amazon, Facebook, Twitter, LinkedIn, Microsoft, etc. Also, we have horizontal grid lines, which will help to learn the exact numbers for both bars.

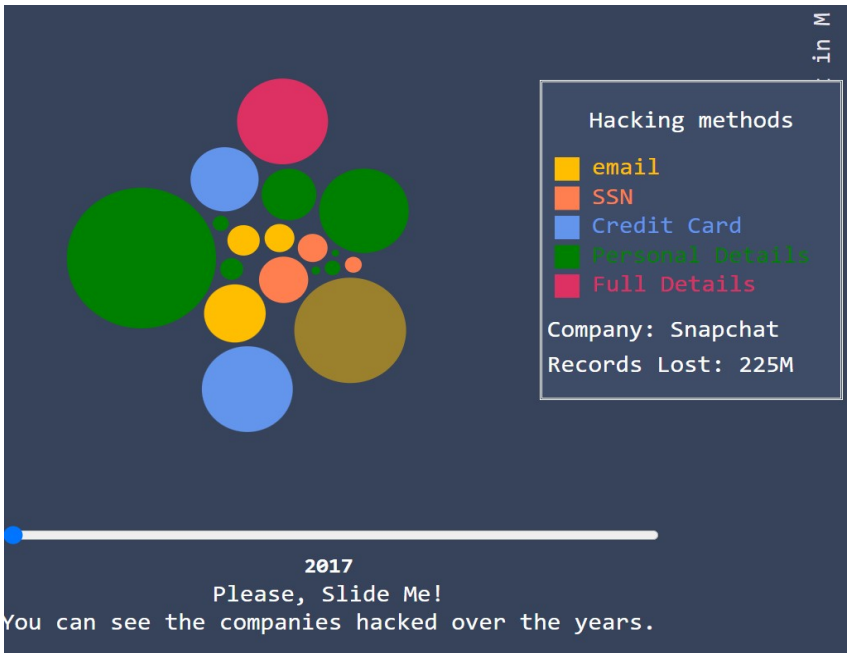


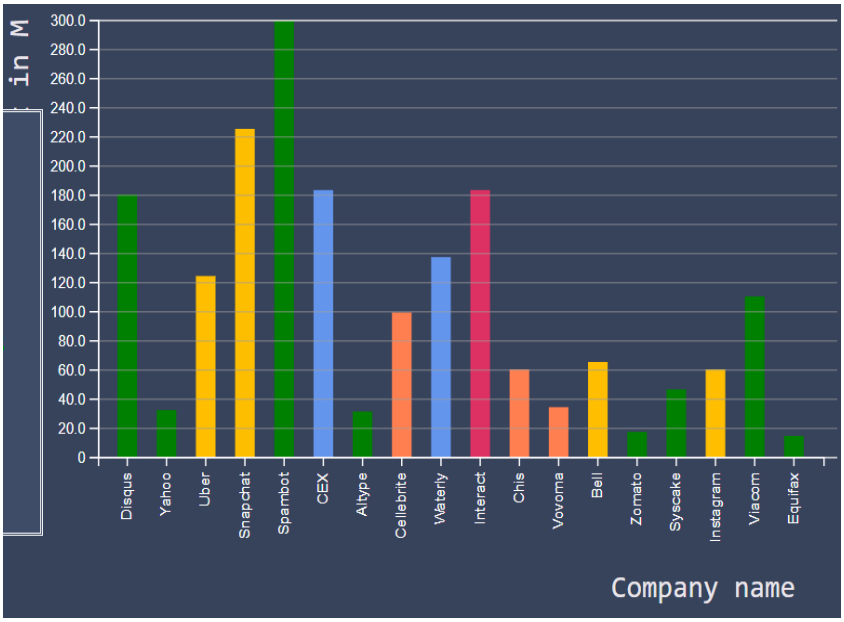
Variations in data lost as per years in particular sectors

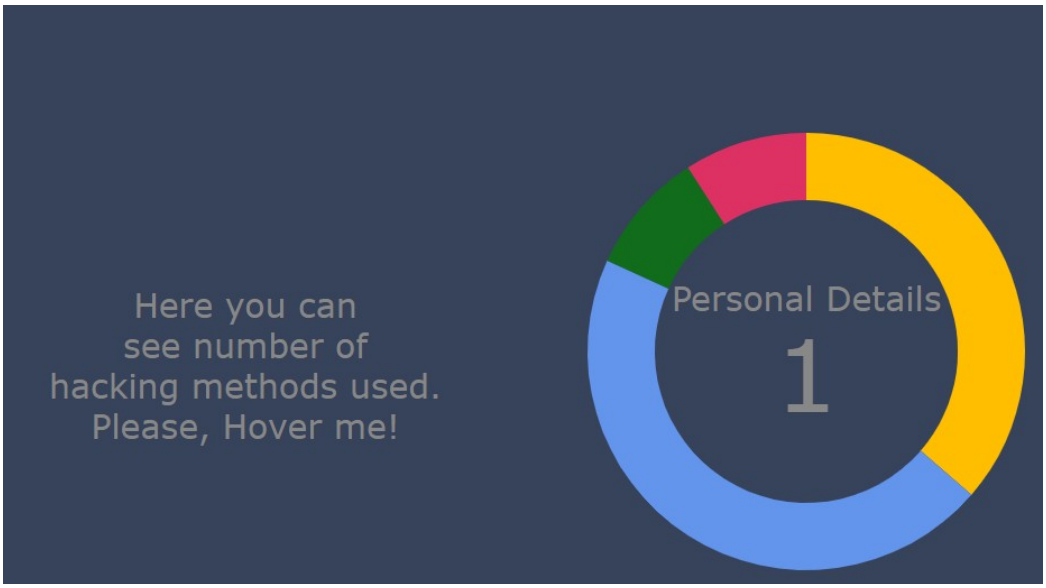
Here we visualize variations in different sectors per year. We have used a donut chart for this. We represent each sector's share when all sectors are involved in data theft. We are facing some errors in the d3 code, but we have added the expected visualization model at the time of the proposal.



After our prototype design, we almost did everything and had to work on the last visualization. Still, according to a review by Dr. Federico, after seeing our prototype, we realized that we missed the interactions such as mouse hover, and sliding, which makes it easy for the users to understand the data and interact with the visualization. In the next step, we also linked all the visualizations together. We designed it in such a way that all sensitive data that leads to data breaches in various organizations can be shown clearly. If we slide the slider linked to all other graphs indicating years, every piece of information change according to that particular year. What's more interesting is that when we mouse hover anywhere in one chart, similar data is highlighted in the other two graphs. The below visualizations show the changes that we have made.



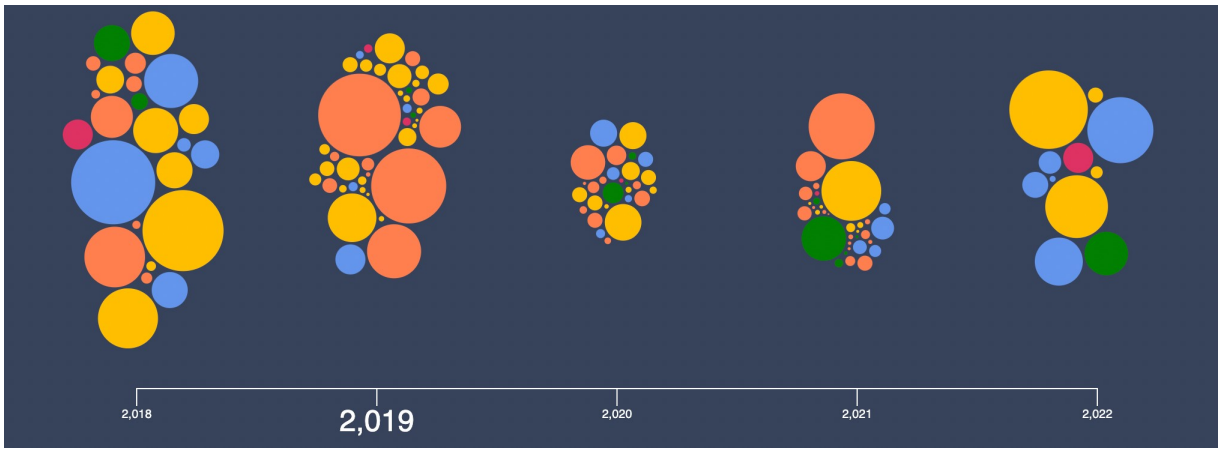




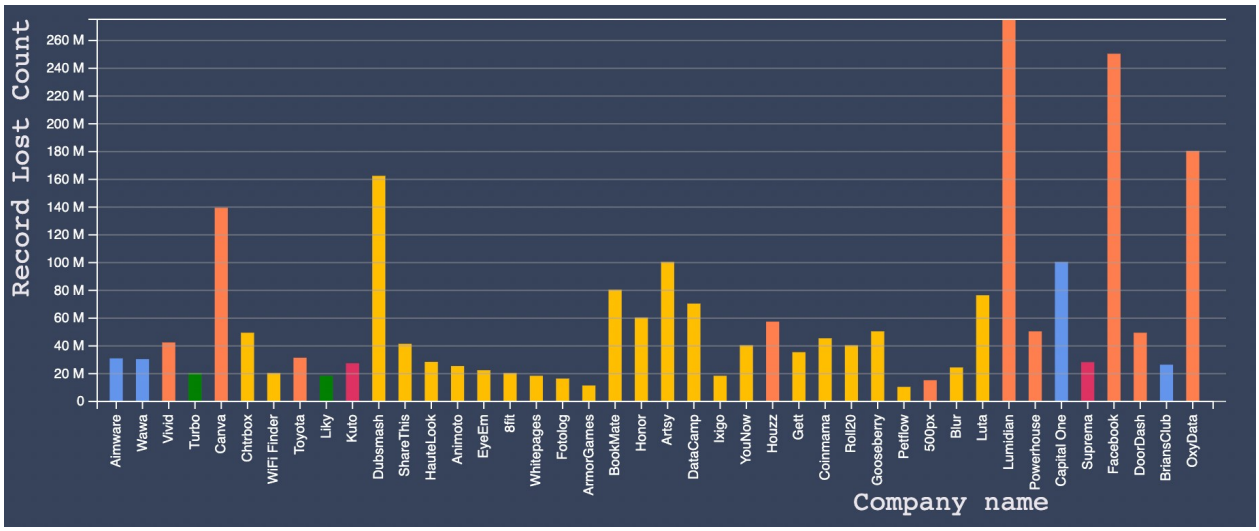
We thought this would be our final destination, but we want to ensure whether our visualizations are going in right direction or not. So, we again discussed with Dr.Federico to explore our visualizations and suggest if we should make any further changes. He then told me one weakness that is missing names for some of the axis/legends. Also, he mentioned that the design of our bubble chart could be more motivating. Another mistake we should have made is we forced the user to scroll through the slides to move from one year to another. After considering all these issues, we have worked on removing the possibility of making comparisons and reducing the information the users can grasp. While doing this, we even had some problems, and with the help of Dr. Federico, we worked on changes that we need to do still, such as the x-axis for the bubble chart, as we missed and used the data in the bar chart and pie chart similar. The second change we worked on is the Onclick event for the same x-axis to change data over the years. Last but not least, we removed text and made sure data should be visualized as graphically as possible.

Implementation:

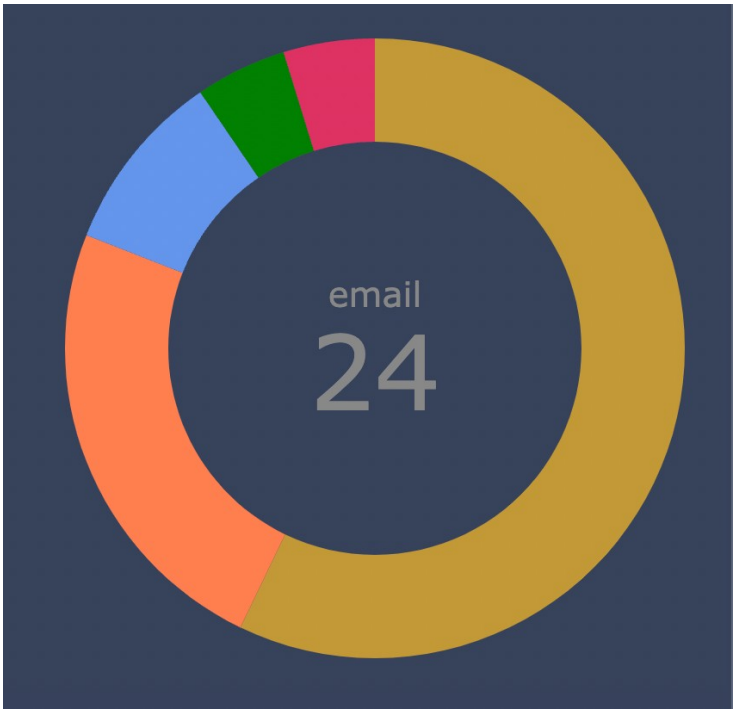
Here is our final implementation of bubble chart. After all, we tried to show how accounts got hacked by 5 different hacking methods in various organizations year wise from 2018 to 2022.



This visualization shows variations of record lost in various sectors . In X- axis we have given company names to show which organizations has lost the data by different type of hacking methods year wise .In y-axis we have used count of record lost in millions. We also used grid cells in order make the data count more clearly.



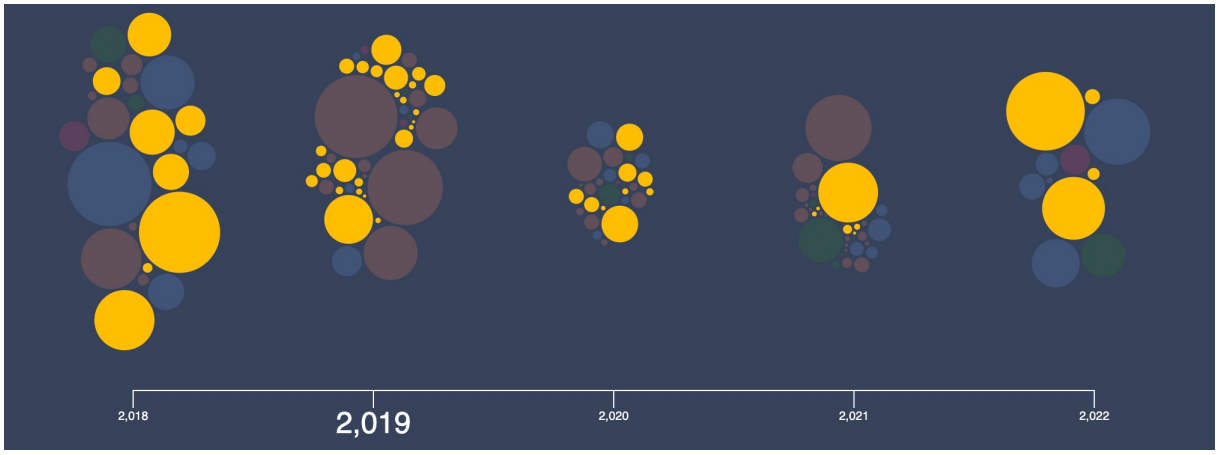
The below donut chart visualization shows total count of each individual hacking methods in given years. So , this pie chart varies accordingly as the data changes according to year.



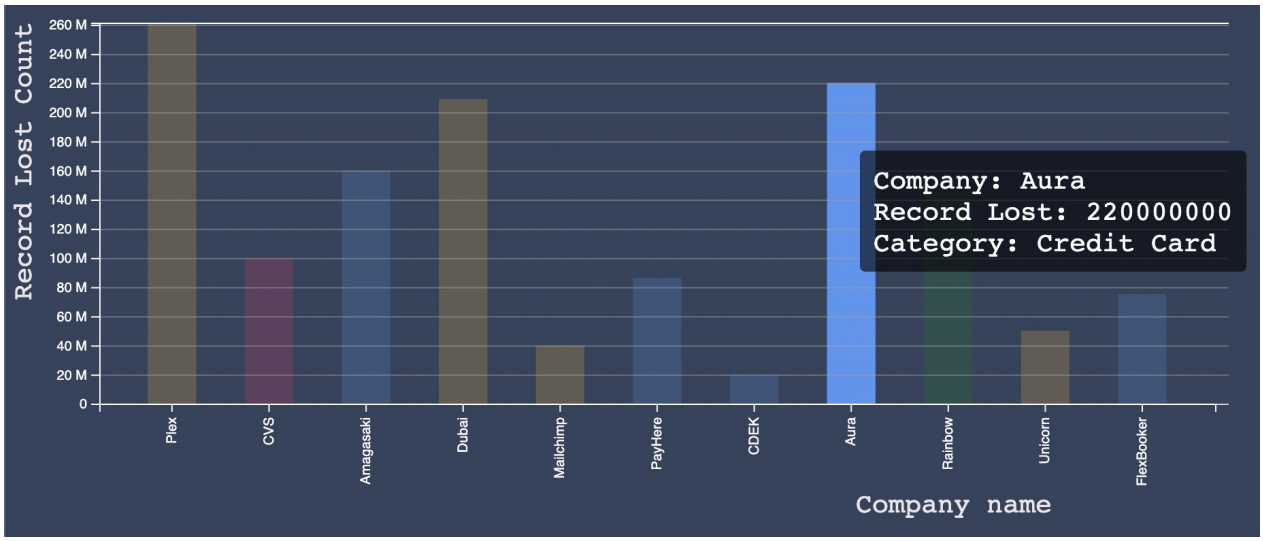
The below visualizations represent how they are linked, how they interact, and the relation between them. Here we focused on interlinking the graphs so that if we click something, every other graph should also react with that. Rather than adding legends, we wed tooltips in the bar chart, and we legends for the bubble chart.

How each individual graph is interactive?

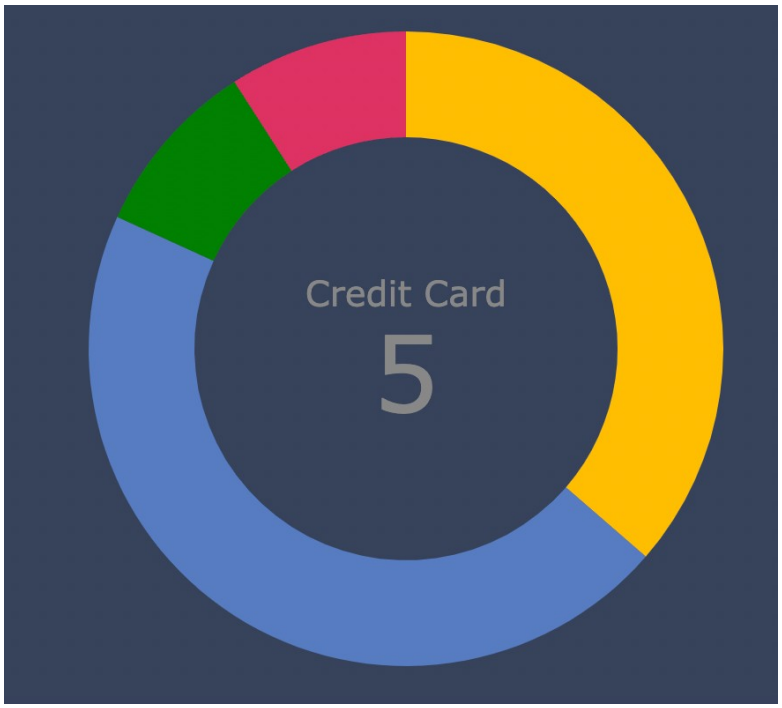
In the below diagram, when the pointer is just placed anywhere on the name of hacking methods, then the bubble chart reacts and highlights the email hacking methods in all given years as email is represented by Yellow color. It got highlighted, making all hacking techniques dim. Also, the tooltip pops out showing the total number of accounts that got hacked via email and counts in numbers such as 3.5 billion from the above diagram in the year 2019.



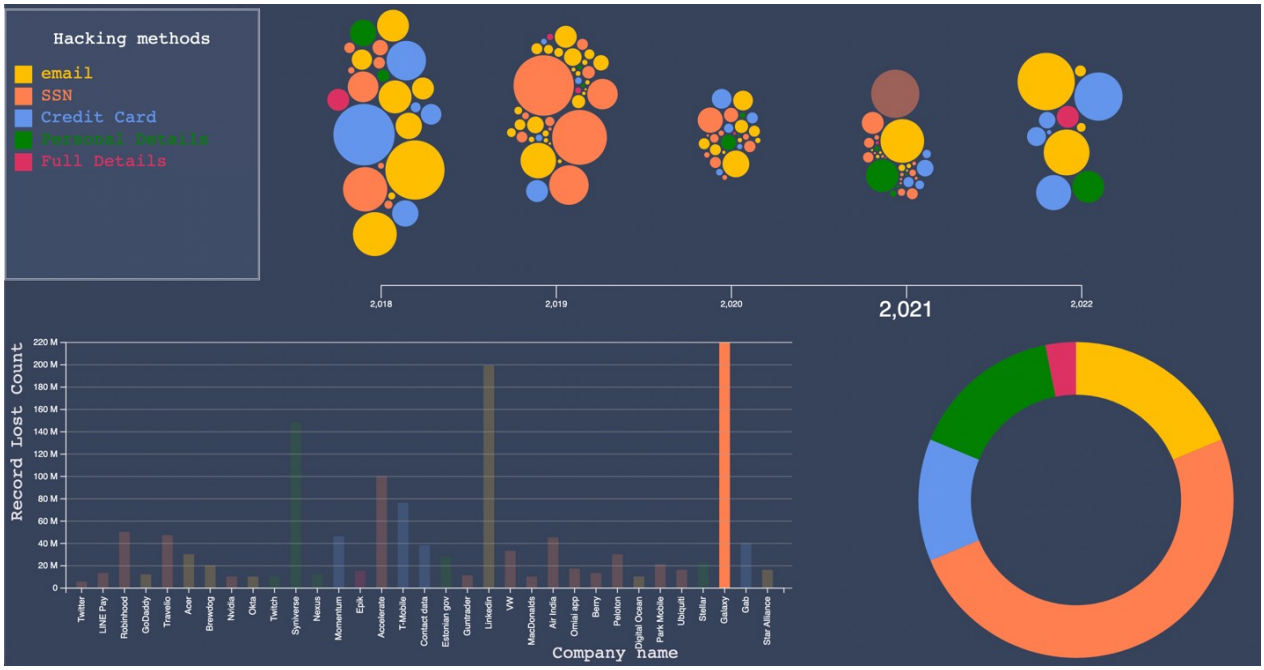
In the given bar graph below, if we mouse hover over any bar present in the bar chart, then it will be highlighted, making others un-emphasized. A tool tip also pops out describing the company name, record loss, and category of the hacking methods.



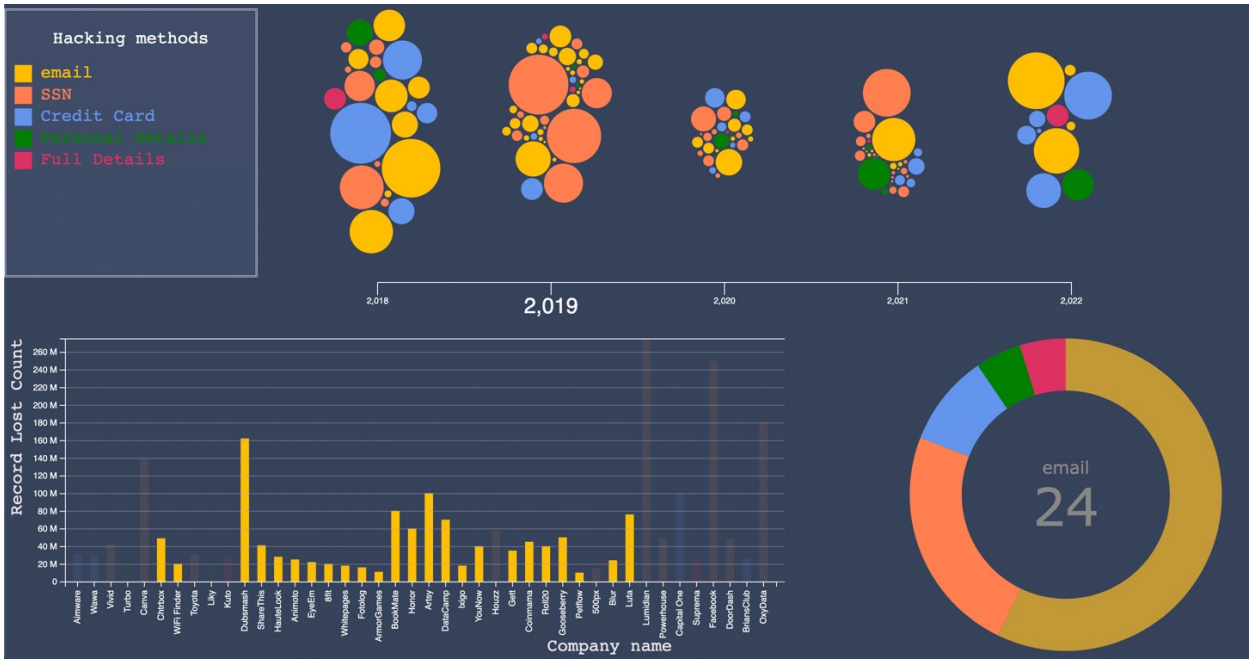
The interactive element in the given donut chart is mouse hover, where the data representing each area in the donut chart hovers. It displays the name of the hacking method and the individual responsible for the total count of data lost when compared to others in all organizations. In the below diagram, the mouse hover is blue, representing the credit card, and the total count is 5; in a sense, five companies in particular years lost their data with a hacking method called the credit card.



How two charts are connected with each other?

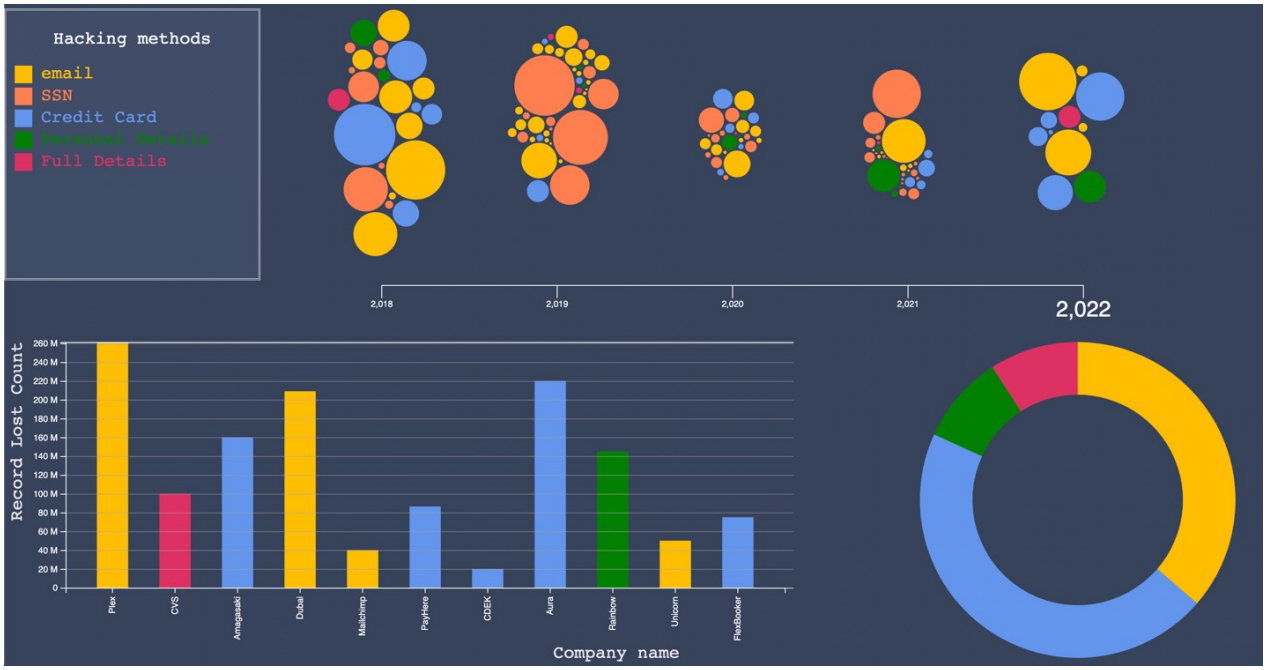


In above diagram bubble chart and bar chart are interlinked through category-named organizations. So, when we place the cursor at any color of our choice, the bar graph with the same company name is highlighted automatically as it is linked with each other.



Connection between donut chart and bar chart takes place in above diagram. when I place my cursor at donut chart the same data which is linked got highlighted .As , the donut chart shows total count of companies got hacked through email.In Bar chart all the organizations or companies which Is related or linked that is all 24 companies are shown which is in yellow color. Even year is common attribute between them, when year changes all the data also varies in both bar and donut chart.All 24 companies are different but the common thing between them is hacking method that is email. Email considered as sensitive data in all 24 companies got highlighted. So , we have interconnected them.

Overall Visualization:



Evaluation:

By performing all these visualizations, various insights were gathered.
The main highlight of our project is to show data breaches in various organizations per year. To understand the data, we used. We selected particular attributes which drag the attention of users.

- Firstly, we differentiated common or usual hacking methods and found five types of sensitive information: losing data in various organizations. We then decided to spot hacking methods in various companies per year and count the total loss of each sensitive data causing data loss.
- Secondly, we spotted variations of companies affected by data loss considering years from 2018 to 2022.
- Finally, we decided to show all organizations the count of specific sensitive data in a particular year.

Overall, our visualizations answered all the questions mentioned above.

- The number of accounts or data thefts taking data sensitivity or hacking type into an account in different organizations?
- Variations of data lost as per year in particular organizations?
- Total count of companies for particular data sensitivity with comparison of organizations?

For the first visualization:

The visualization displays how various organizations got hacked from data sensitivity from 2018 to 2022. We interactively visualized all of these elements using a Bubble chart.

Visual encodings:

The bubble chart we created to display the data breaches in various companies. We accurately translated the data and made the representation so that each component of the map was transparent, and any user could easily comprehend the facts they presented. We all are aware of it each time the data is altered in some way. We need to demonstrate it visually, so things like color, size, grouping, and so forth should be considered—where we tried to accomplish practically everything.

Interaction techniques:

The mouse hover, tooltip, legends, and tooltip to describing our attributes more interactive make it the most engaging component. We used the color scheme to distinguish the various hacking methods, and the backdrop hue's central theme was to highlight all the charts we used. It draws the focus of every user. We employed Filtering, which enables us to limit or customize the data displayed in the visualization.

Design Quality:

We followed all the guidelines and reasoned to provide their data interactively to the user. We used five different colors to view hacking types. We also used size as a criterion, as the bubble size is more significant than that particular company has lost more data through hacking. Additionally, we examined all the data per year to visualize the data more clearly; if we didn't mention the years, then the data would be more and look like a huge mess. The data lost in all organizations between 2018 and 2022 and how the data changes are evident in this comparison. And when we place a cursor on the legend, which displays five hacking methods, it shows the accounts that got hacked via a particular data sensitivity method. A unique pop-up dialog box says this. We aim to make the most crucial details stand out the most. Overall, everything went smoothly.

For the second visualization:

The visualization shows the company name on one side and numerous records lost on the y- axis, showing the data transition. We displayed the variation of data lost per year using a bar chart.

Visual encodings:

The representation we tried to show is comprehensive, and an attempt was made to translate the material. To demonstrate it graphically, the use of color and grouping is employed. To make it more user-interactive, a tooltip box shows the company name, category, and count of records lost in number. Also, the data keep changing according to the years.

Interaction techniques:

The colors were employed to effectively distinguish the hacking type we chose to display. Filtering allowed us to restrict or alter the data shown in the visualization, such as companies and records lost. To draw visitors' attention, we tried to be more interactive by displaying the data variation.

Design Quality:

We used all available techniques to increase user interaction. The various color's used to depict the various methods in hacking are well-designed to distinguish between them. Additionally, a new visualization is displayed for each portion of the year when a specific section is selected. They try to highlight the essential elements. Overall, everything went quite smoothly. Additionally, the numbers on the x and y axis are in millions is apparent as well.

For the third visualization:

The visualization shows the total count of particular sensitive data with comparison of organizations showing transition of data. We displayed this variation of this count as per year using a donut chart.

Visual encodings:

The representation we tried to show is comprehensive, and an attempt was made to translate the material. To demonstrate it graphically, the use of color and grouping is employed. To make it more user-interactive, there is a tool-tip box shows count and name of sensitivity data that causing vulnerabilities for hacking Is shown in middle of donut chart. Also, the data keep changing according to years.

Interaction techniques:

The colours were employed to effectively distinguish between the hacking type that we chose to display. Filtering allowed us to restrict or alter the data shown in the visualization, such as count of organizations that got hacked with similar data sensitivity. To draw visitors' attention, we tried it to be more interactive for showing variation of data.

Design Quality:
We used all available techniques to increase user interaction. The various colors used to depict the various methods in hacking are well-designed to distinguish between them. Additionally, a new visualization is displayed for each portion of the year when a specific section is selected. They try to highlight the essential elements. Overall, everything went quite smoothly. Additionally, the count that we tried to display is a tooltip which appears when we mouse hover at one color which represents the data sensitivity type.

Future Scope:

We will add more data regarding data breaches to make it comprehensive. We are also considering adding a Tool tip for bubbles. So, when we place the cursor in a bubble, we can see the popup of the tooltip, which displays a story of data lost news that happened in a particular year. We will also add more interactive components.

References:

[1] https://www.informationisbeautiful.net/visualizations/worlds-biggest-data- breaches-hacks/

[2] Juma'ah, A. and Alnsour, Y. "The Effect of Data Breaches on Company Performance" International Journal of Accounting and Information Management (IJAIM). Vol. 28, no. 2, 2020

[3] https://www.ftc.gov/system/files/documents/public_events/1582978/now_im_a_bit_angry_-_individuals_awareness_perception_and_responses_to_data.pdf

[4] https://www.trendmicro.com/vinfo/es/security/news/cyber- attacks/understanding-targeted-attacks-goals-and-motives

[5] https://www.statista.com/statistics/273550/data-breaches-recorded-in- the-united-states-by-number-of-breaches-and-records-exposed/

[6] https://www.ibm.com/reports/data-breach

[7] https://informationisbeautiful.net/visualizations/top-500-passwords- visualized/

[8] https://informationisbeautiful.net/visualizations/ransomware-attacks/

[9] https://www.hipaajournal.com/healthcare-data-breach-statistics/