CYBER THREAT RECON

Bernard Low, Dimple Modi,
Emmanuel Asong, Rachelle Azulay

01

Defining the Problem

Problem Statement and Goals

02

Dataset Description

Introduction to main dataset, cleaning and processing

03

Presentation of Findings

Executive Summary and Dashboard Walkthrough

04

Primary Objectives

Topic Modelling and Clustering Insights

05

Secondary Objectives

Security Benchmarking and Trust Analysis

06

Wrapping Up

Conclusion, Limitations and Future

01 - Problem Statement

The Department of Defense (DoD) has requested Deloitte provide them with analytics to expand its detection of various Cyber threats such as data breaches, and recommend strategic action.

In turn, Deloitte has enlisted the help of the GWU team to build a series of views to predict and quantify key patterns, and classify incidents into distinct categories for agency assessment and response planning.

01 - Project Goals

Primary Goals

1. Find out data breach patterns
   - ○ Who is being breached?
   - ○ Where are they located?
   - ○ What industries do they belongs to ?
   - ○ How are the attacks conducted?

2. Create dashboard to showcase findings.

Secondary Goals

1. Find security measure benchmark for breached companies to follow.

2. Study effect of data breaches on organizational reputation and customer trust.

02 - Main Dataset Overview

Data Source

PrivacyRights.org

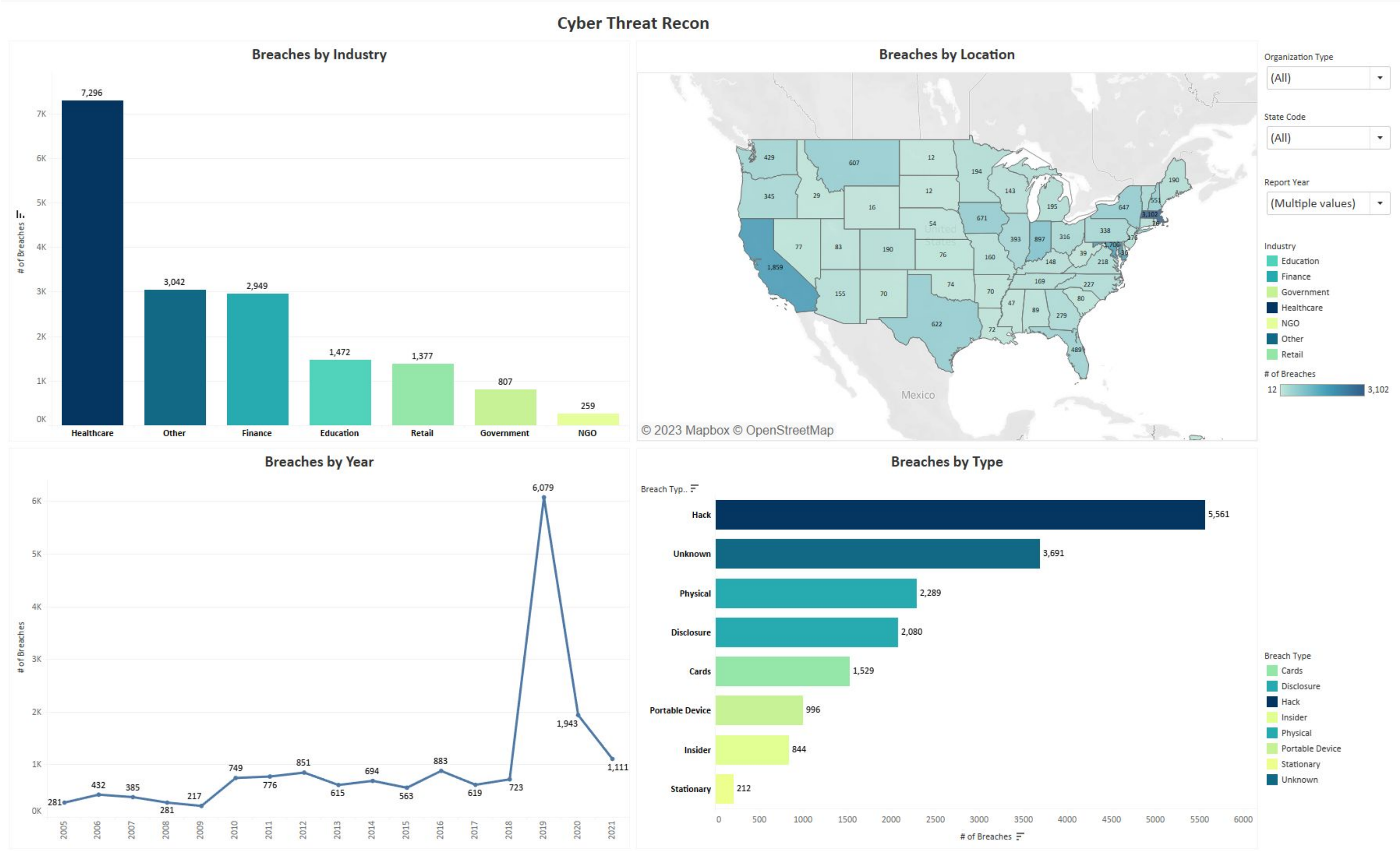
Nonprofit organization focused on increasing access to information and policy discussions.

- 20,146 records ranging from 2002 to 2022.
- Dataset logs data breaches and provides incident descriptions.

Cleaning and Processing

- Formatted dates and other columns.
- Dropped irrelevant columns.
- Cleaned entries, homogenizing unknown values.
- Filtered for US only data.
- Prepared relevant columns for Non-negative matrix factorization (NMF).

03 - Dashboard Development and Walkthrough



Cyber Threat Recon

Organization Type: (All)

State Code: (All)

Report Year: (Multiple values)

04 - Topic Modelling - NMF

Topic Modeling:

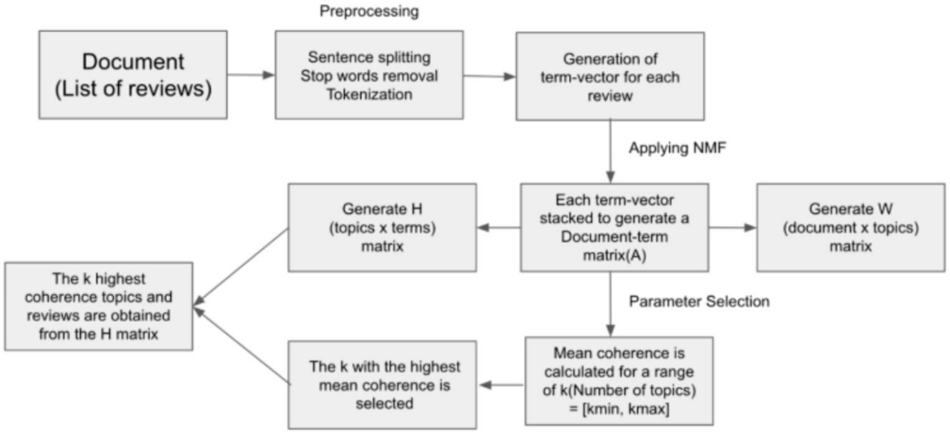
Topic modeling is a frequently used approach to discover hidden semantic patterns portrayed by a text corpus and automatically identify topics that exist inside it. Namely, it's a type of statistical modeling that leverages unsupervised machine learning to analyze and identify clusters or groups of similar words within a body of text.

Key Question: Are there any trends or patterns in breach occurrences?

NMF (Non-negative matrix factorization)

- NMF topic modeling is memory efficient and fast .
- NMF works best with sparse corpora .

Programming: Python.



Preprocessing

Document (List of reviews)

Sentence splitting Stop words removal Tokenization

Generation of term-vector for each review

Applying NMF

Generate H (topics x terms) matrix

Each term-vector stacked to generate a Document-term matrix(A)

Generate W (document x topics) matrix

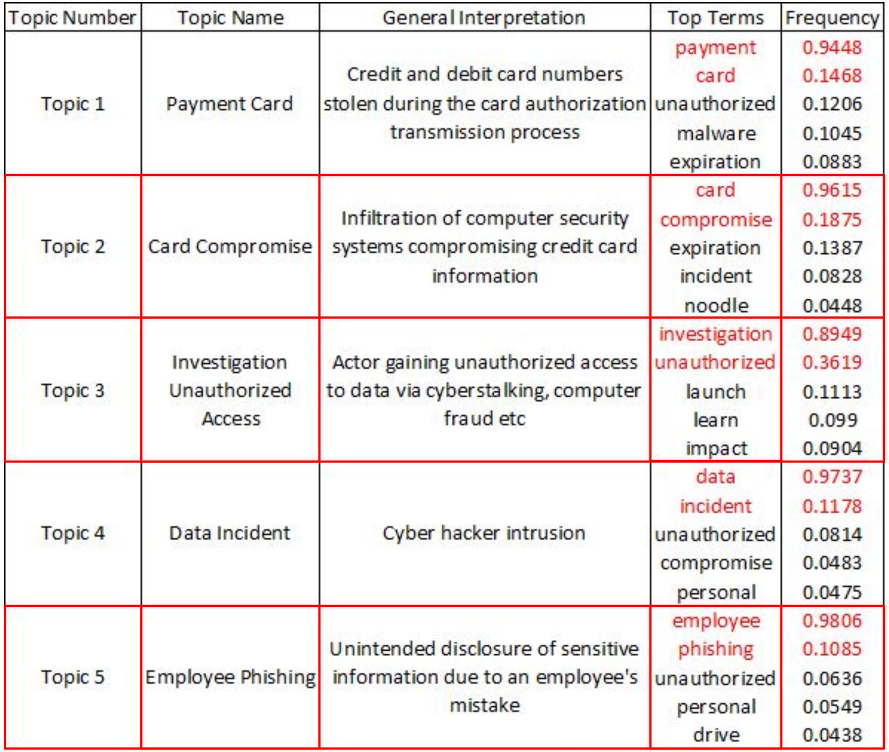
The k highest coherence topics and reviews are obtained from the H matrix

Parameter Selection

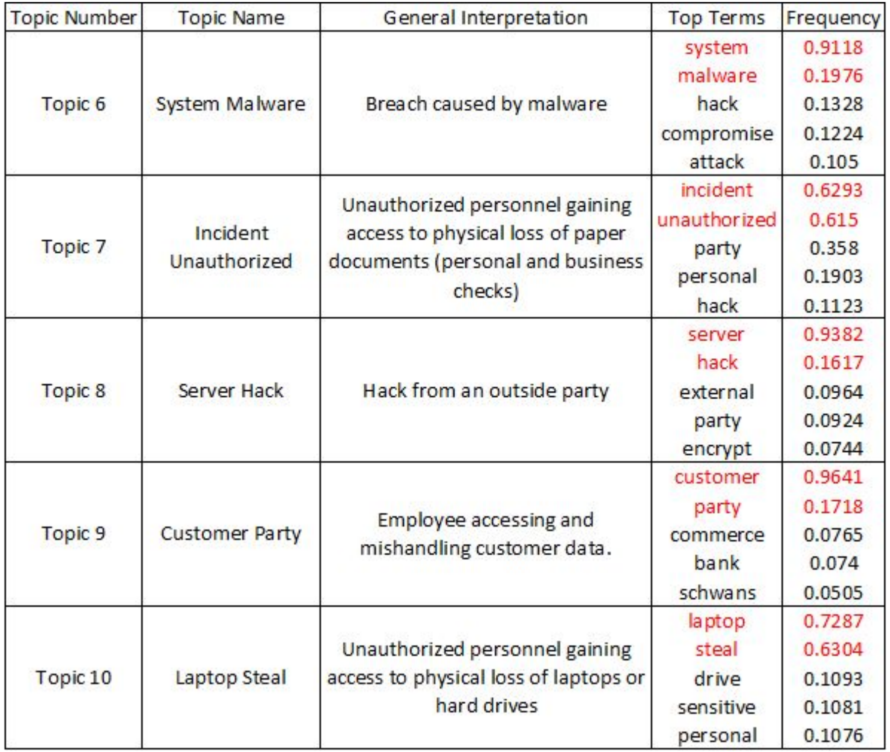
The k with the highest mean coherence is selected

Mean coherence is calculated for a range of k(Number of topics) = [kmin, kmax]

04 - Topic Modelling - Terms Generated



Topic Number	Topic Name	General Interpretation	Top Terms	Frequency
Topic 1	Payment Card	Credit and debit card numbers stolen during the card authorization transmission process	payment card unauthorized malware expiration	0.9448 0.1468 0.1206 0.1045 0.0883
Topic 2	Card Compromise	Infiltration of computer security systems compromising credit card information	card compromise expiration incident noodle	0.9615 0.1875 0.1387 0.0828 0.0448
Topic 3	Investigation Unauthorized Access	Actor gaining unauthorized access to data via cyberstalking, computer fraud etc	investigation unauthorized launch learn impact	0.8949 0.3619 0.1113 0.099 0.0904
Topic 4	Data Incident	Cyber hacker intrusion	data incident unauthorized compromise personal	0.9737 0.1178 0.0814 0.0483 0.0475
Topic 5	Employee Phishing	Unintended disclosure of sensitive information due to an employee's mistake	employee phishing unauthorized personal drive	0.9806 0.1085 0.0636 0.0549 0.0438



Topic Number	Topic Name	General Interpretation	Top Terms	Frequency
Topic 6	System Malware	Breach caused by malware	system malware hack compromise attack	0.9118 0.1976 0.1328 0.1224 0.105
Topic 7	Incident Unauthorized	Unauthorized personnel gaining access to physical loss of paper documents (personal and business checks)	incident unauthorized party personal hack	0.6293 0.615 0.358 0.1903 0.1123
Topic 8	Server Hack	Hack from an outside party	server hack external party encrypt	0.9382 0.1617 0.0964 0.0924 0.0744
Topic 9	Customer Party	Employee accessing and mishandling customer data.	customer party commerce bank schwans	0.9641 0.1718 0.0765 0.074 0.0505
Topic 10	Laptop Steal	Unauthorized personnel gaining access to physical loss of laptops or hard drives	laptop steal drive sensitive personal	0.7287 0.6304 0.1093 0.1081 0.1076

04 - Topic Clustering - Main Process

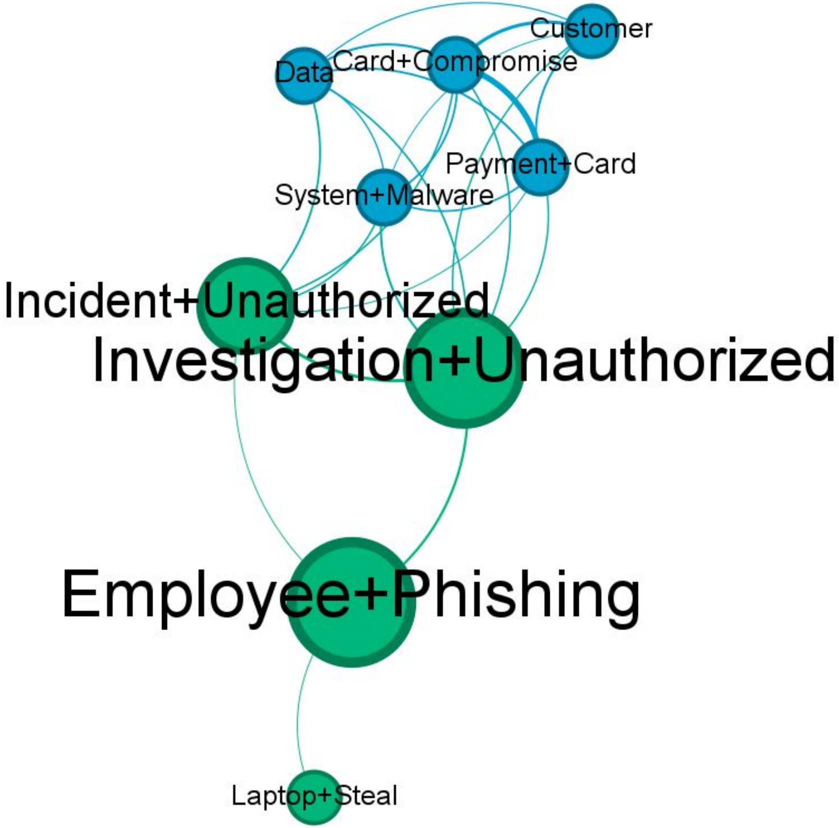
Cluster:

Derived from topic modeling using NMF, which utilizes **'H' matrices to identify topics.**

- **Sample Data Size:** 20,146 rows.
- **Software:** Gephi 0.10.1

Plot:

- **Investigation+Unauthorized** & **Employee+Phishing** shows high degree compare to other topics .
- **Laptop+Steal** is connected to **Employee+Phishing.**
- **Customer** connected to **Card+Compromise** and **Payment+Card**.



04 - Topic Clustering - Specific Terms View

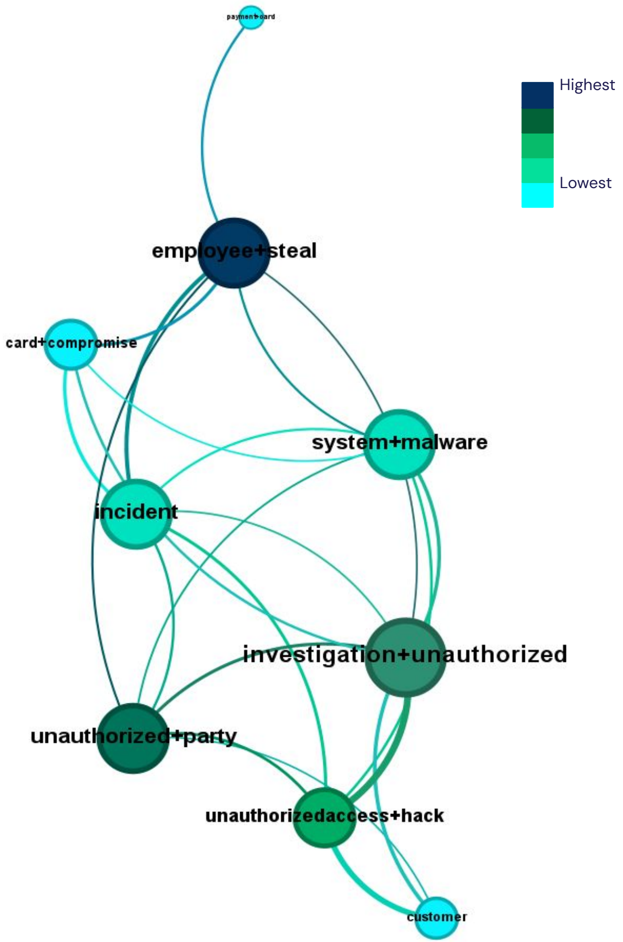
Cluster: Created with data rows where breach description contained specific terms.

- **Terms:** Card, Unauthorized, Steal, Stolen, Employee
- **Sample Data Size:** 10,096

Betweenness Centrality: Finds which nodes are most important for holding the network together.

- The **Employee+Steal** node shows the highest Betweenness Centrality score – **7.25**

As it connects to the **Unauthorized+Party** node, it can be inferred that **company employees** have created data breach incidents by **stealing customer information** via **unauthorized access** or **abuse of privileges.**

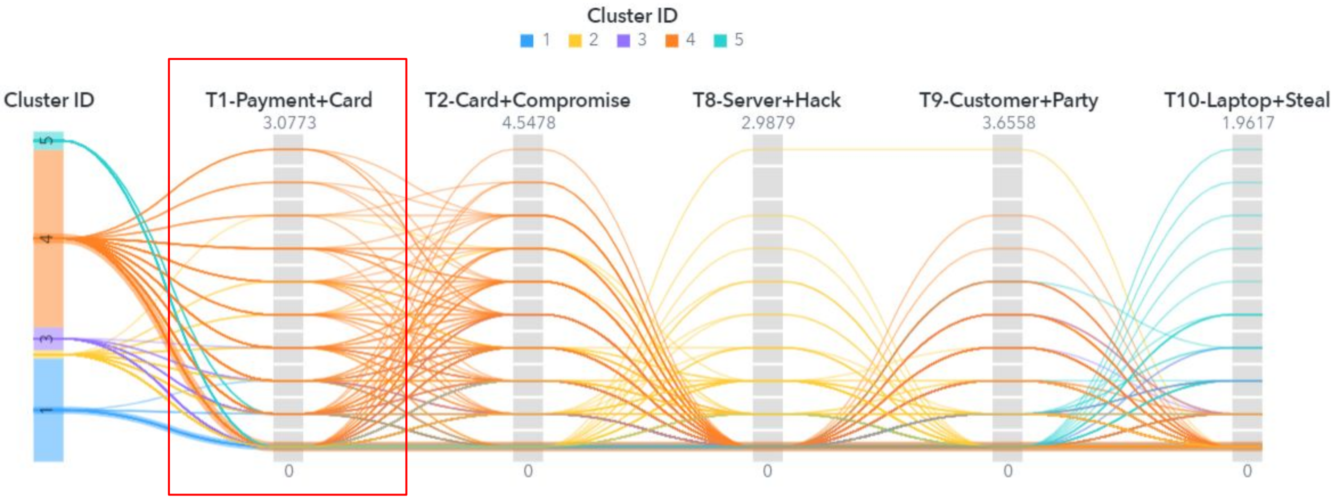


04 - Document Clustering - Process

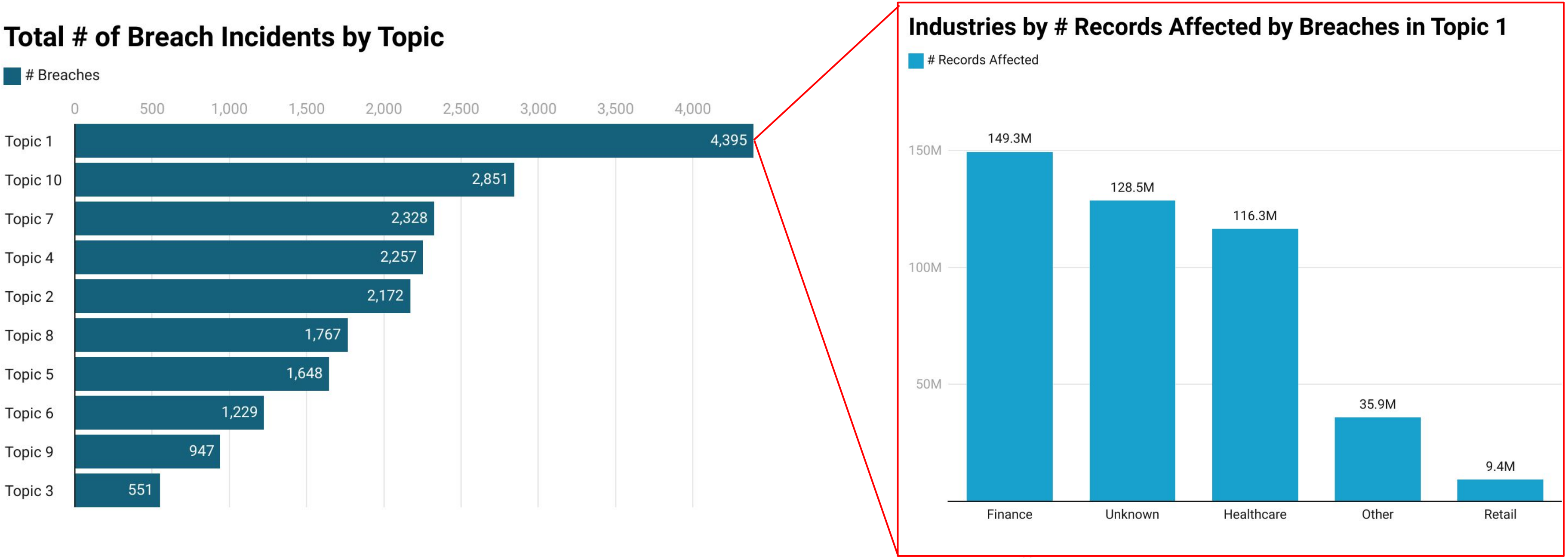
This cluster is derived from topic modeling using Non-negative Matrix Factorization (NMF), which utilizes 'H' and 'W' matrices to identify topics. We integrate the **'H' matrix** from the NMF with the **original dataset** to enhance the data structure and reveal topic associations.

Software: SAS Viya

The parallel coordinates plot suggests that Cluster 4 has the **highest number of observations** and thus **high association** with **Topic 1 – 'Payment+Card'** and **Topic 2 – 'Card+Compromise'**.



05 - Insights Into Breach Types - Benchmarking



- **Topic 1**, concerning breaches via theft of customer credit/debit card details during transactions, was found to have the highest frequency of attacks.

- These attacks were directed towards the financial industry more than other Industries.

- Companies in this industry will be focused on in our benchmarking exercise.

05 - Benchmarking - Introduction

- Benchmarking, in its essence, is the practice of measuring the performance of organization's policies, practices, and metrics against industry best practices.

- In the domain of Card Payment System, this means comparing your security posture, practices, and incident records against recognized standards such as the Common Vulnerability Scoring System (CVSS) and the Payment Card Industry Data Security Standard (PCI DSS).

- The goal here is to benchmark how 9 selected organizations within the financial industry handle payment card data with respect to the industry standards set out by CVSS AND PCI.

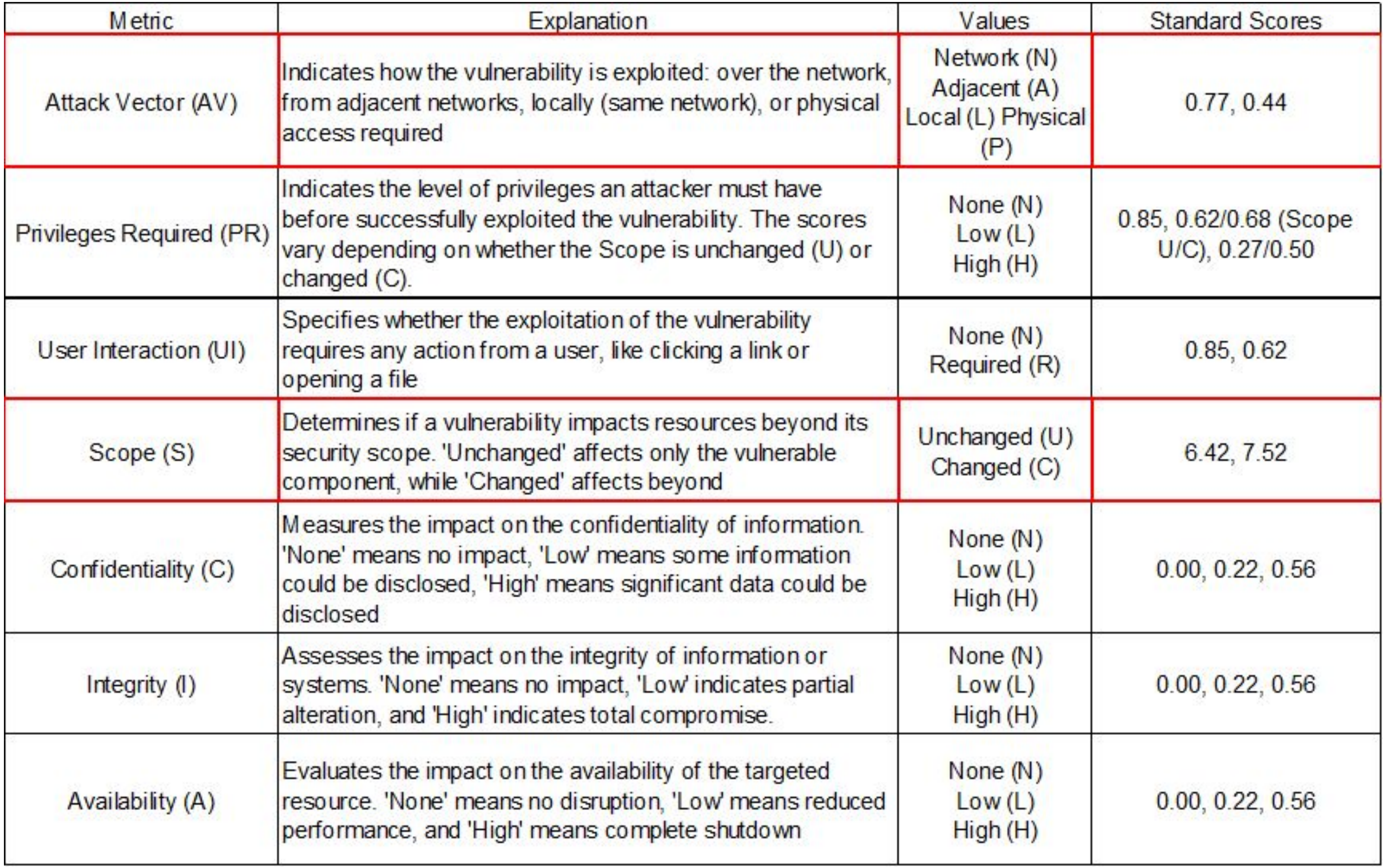
05 - Benchmarking - CVSS Framework & PCI

- The Common Vulnerability Scoring System (CVSS) is a framework for assessing the severity of security vulnerabilities in computer systems and networks.

- The Payment Card Industry Data Security Standard (PCI DSS) represents a common set of industry tools and measurements for ensuring the safe handling of sensitive customer card information. It details technical requirements for the secure storage, processing and transmission of cardholder data.

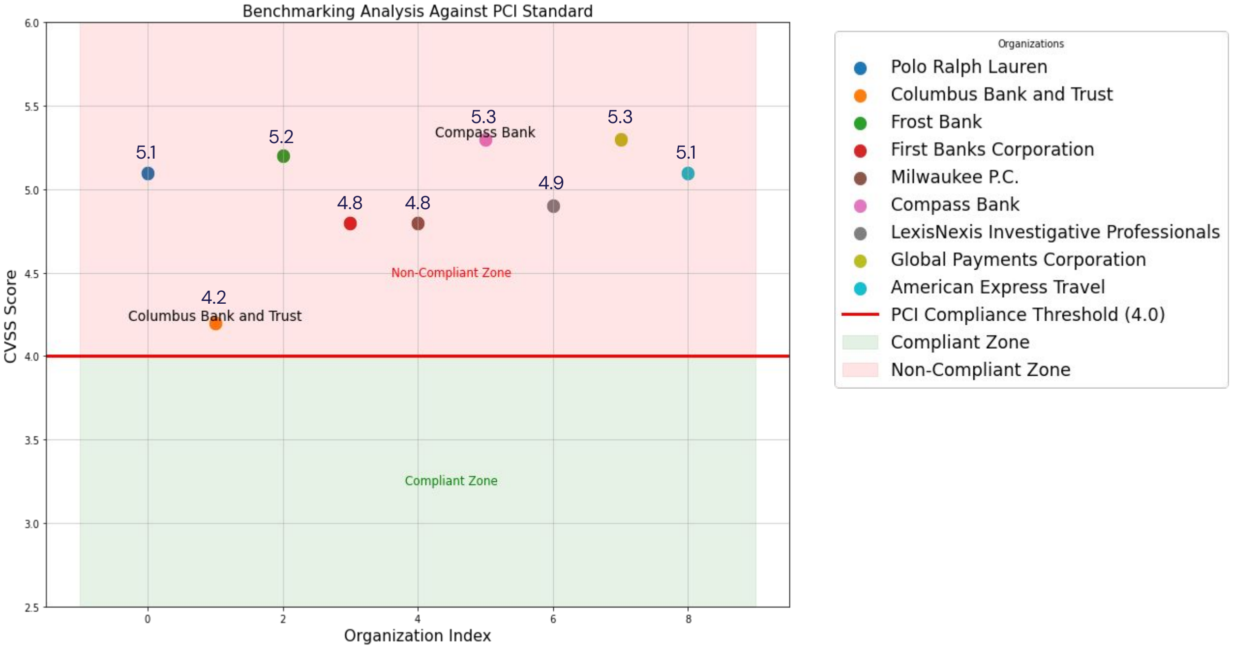
- Scores range from 0 to 10.0, with 4.0 or higher indicating failure to comply with PCI standards. Any asset that contains at least one vulnerability with CVSS score of 4.0 or higher is considered non-compliant. And, if at least one asset is non-compliant, the entire organization is considered to be non-compliant.

05 - Benchmarking - CVSS Framework & PCI



Metric	Explanation	Values	Standard Scores
Attack Vector (AV)	Indicates how the vulnerability is exploited: over the network, from adjacent networks, locally (same network), or physical access required	Network (N) Adjacent (A) Local (L) Physical (P)	0.77, 0.44
Privileges Required (PR)	Indicates the level of privileges an attacker must have before successfully exploited the vulnerability. The scores vary depending on whether the Scope is unchanged (U) or changed (C).	None (N) Low (L) High (H)	0.85, 0.62/0.68 (Scope U/C), 0.27/0.50
User Interaction (UI)	Specifies whether the exploitation of the vulnerability requires any action from a user, like clicking a link or opening a file	None (N) Required (R)	0.85, 0.62
Scope (S)	Determines if a vulnerability impacts resources beyond its security scope. 'Unchanged' affects only the vulnerable component, while 'Changed' affects beyond	Unchanged (U) Changed (C)	6.42, 7.52
Confidentiality (C)	Measures the impact on the confidentiality of information. 'None' means no impact, 'Low' means some information could be disclosed, 'High' means significant data could be disclosed	None (N) Low (L) High (H)	0.00, 0.22, 0.56
Integrity (I)	Assesses the impact on the integrity of information or systems. 'None' means no impact, 'Low' indicates partial alteration, and 'High' indicates total compromise.	None (N) Low (L) High (H)	0.00, 0.22, 0.56
Availability (A)	Evaluates the impact on the availability of the targeted resource. 'None' means no disruption, 'Low' means reduced performance, and 'High' means complete shutdown	None (N) Low (L) High (H)	0.00, 0.22, 0.56

05 - Benchmarking - PCI Score Comparison



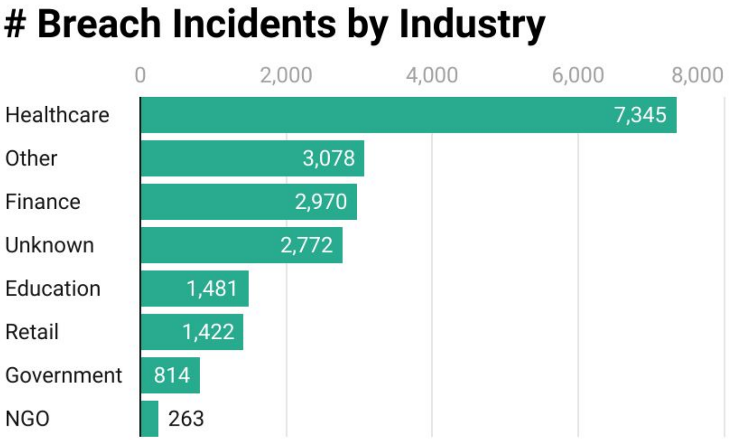
Methodology:

- Top 9 companies with the highest number of total records affected were taken.

Scoring:

- Higher scores imply Severe system vulnerability to attacks.
- All the organizations failed to meet the PCI threshold of 4.0
  - Best: Columbus Bank And Trust (4.2)
  - Worst: Compass Bank/Global Payments Corp (5.3)
  - Payment cards data of customers in these organizations need to be protected from unauthorized access.

05 - Insights Into Breached Industries - Trust



Key Question:

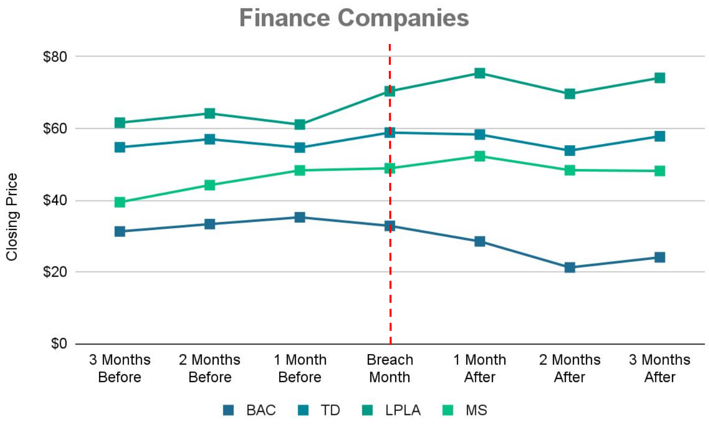
What are the effects on company reputation/consumer trust?

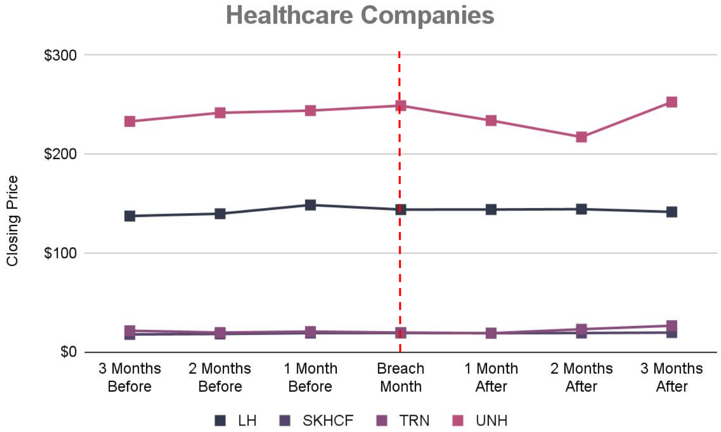
- **Healthcare** and **Finance** were shown to be the highest in frequency breached.
- Top 4 private companies in each industry with the highest number of total records affected were taken.

Methodology :

- Historical NASDAQ data of top breached companies was compiled from Yahoo! Finance.
- Stock closing price at end of month was used to gauge public impression of companies.
- Data 3 months before and after the month of a company's largest breach incident was used to create time series graphs.

05 - Trust Analysis - Consumer Effect





Finance Companies

- BAC (Bank of America)
- TD (TD Bank)
- LPLA (LPL Financial Holdings)
- MS (Morgan Stanley)

Healthcare Companies

- LH (Laboratory Corporation of America)
- SKHCF (Sonic Healthcare)
- TRN (Trinity Industries)
- UNH (UnitedHealth Group)

Impressions: Companies being breached doesn’t impact public perception.

Caveat: Stock prices are a multifaceted subject, and data breach incidents are just one of many factors that influence.

06 - Conclusion

Internal Factors: Employees are leaking data intentionally or through phishing attacks.

Strategic Action: Pressure companies to institute comprehensive training programs for their staff to prevent data breaches caused by careless mistakes.

External Factors: Credit/debit card payment attacks are commonplace.

Strategic Action: Hold companies to a strict security standard. Companies with high CVSS scores should emulate others without breach incidents or with low CVSS scores like Columbus Bank And Trust.

Consumers do not seem to be affected by news of breaches. Campaigns could be implemented to educate the public on the dangers of breaches, and to hold companies leaking their data accountable to incentivize security standard compliance.

06 - Limitations and Future

Limitations:

- Nature of dataset makes it difficult to properly classify some data points, limited description of incidents reports on attacks by various organizations.
- Additional datasets would have provided a more holistic view, more topics would have given the analysis more nuance, but might also have cause overfitting.

Next steps:

- Follow up on breached companies.
- Employ topic modelling on social media data or company reviews for more granular look at public perception for trust analysis.
- Research and implement processes that can identify and account for new threats born of generative AI.

06 - Appendices

References

- Dataset Source
  - https://privacyrights.org/
- Data Breaches in 2019 - 2022
  - https://selfkey.org
- PCI Severity Levels
  - https://pci.qualys.com

Clustering Performance

- Silhouette Score: 0.46
- Calinski- Harabasz Index: 3117.88
- Davies Bouldin Index: 1.08

Thank You!