Predicting Data Breach Trends Based on Historical Attack Patterns

Leonard Korreshi
Rishabh Sareen
Wen Bin

February 03rd, 2023

Abstract Data breaches are a significant concern for companies like ManpowerGroup looking to keep their client's information secure, but it is difficult to predict what methods attackers are using to illegitimately obtain data. A database published by Verizon which contains a collection of data breaches and incidents dating back to 2014, permitting the construction of a prediction model. Each incident is classified according to common patterns and a trend is established for each pattern. Using time series methods, these trends are predicted into the future with confidence.

Updates

- Fixed the TRSE errors. [LK]

Action Plan

- Modify it according to the predictive models built in due course of time [LK, RS]

Work done in partial fulfillment of the requirements of Michigan State University MTH 844; advised by Mr. Michael Friedel, ManpowerGroup; and Dr. Gee Lee, Dr. Peiru Wu, Michigan State University.

Table of Contents

Nomenclature (*In Order of Appearance*)

Introduction

ManpowerGroup (formerly known as Manpower Inc.) is a Fortune 500 American multinational corporation headquartered in Milwaukee, Wisconsin. Founded in 1948 by Elmer Winter and Aaron Scheinfeld, ManpowerGroup is the third-largest staffing firm in the world. One of the most important domains in which ManpowerGroup has actively been involved is cybersecurity. Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks. These cyberattacks are usually aimed at accessing, changing, or destroying sensitive data, and using that data to extract money from businesses via ransomware, information brokering, and selling the data on the dark web and other black markets.

The goal of this project is to predict the root causes of future 'Data Breaches' by extrapolating historical breach trends as detailed in Data Breach Investigation Report 2022 (DBIR) given by Verizon. A data breach refers to the security violation that involves releasing secure or private/confidential information into an untrusted environment (such as the public Internet or the dark web) where unauthorized individuals can access it (download, copy, view, or otherwise transmit it). Common types of data breaches can include basic application attacks and abuse, phishing users, Denial of Service (DoS) attacks, direct system intrusion, privilege misuse, and social engineering of customer service representatives. By being able to anticipate what sorts of data breaches will be common in the future, firms like the ManpowerGroup will be able to more efficiently allocate resources to protect against them.

The first phase of this project is to create a system that can summarize the raw data in a format useful for performing data analysis and then subsequently building predictive models. The data is available in the form of nested information which gives little insight into the nature of an incident or type of data breach. The reconstruction of data involves using various Python and R libraries that can simplify the data into usable data frames and provide additional information about the victim affected, timeline information and more. This can then give valuable insights into the types of models which should be built for accurate prediction of the future data breaches.

This report begins with the data section which outlines the available incident data contained in the VERIS Community Database (VCDB). Then, the analysis section illustrates the statistical underpinnings of classifying and reconstructing the data while the methods section demonstrates how the reconstruction was done programmatically. The results section presents the new reconstruction and its future predictions, followed the discussion section adding some commentary about these results and the future of data breaches. Finally, the conclusions section summarizes the report with brief, assertive points.

Updates

- Edited the paragraphs [RS, LK]

Action Plan

- Include a summary of each section as those sections get completed [RS, LK]

Data

The data used for the purpose of this project was retrieved from the VCDB GitHub page. It is a comprehensive open-source database with data breaches and incidents coded in two main formats: JSON and CSV. The JSON format uses a dictionary of dictionaries to contain the data while the CSV format codes each incident as a series of TRUE and FALSE fields for easy filtering between the numerous different kinds of events that can occur.

An example of what this data looks like can be seen in Table A1, in Appendix A. It is difficult for a human to read this kind of spreadsheet, but it is quite optimal for computer software as it can quickly filter through to find when certain fields are TRUE or FALSE. Additionally, at the end of both the JSON and CSV formats, there is noncoded information pertaining to the cases, such as timeline information of when the breach happened, and a textual summary of what occurred, but this information is limited to what the reporter was able to determine about the incident.

The key information in each incident is tied to four fields: the actor, which describes who caused the incident, the action, which describes what the actor did to cause an incident, the assets, which describe what was lost in the incident, and the attributes, which note how people were affected by the incident. Collectively, these are referred to as the 4As. In Figure 1, an example of the action field is shown – every 'hacking' action taken since 2010 is displayed.

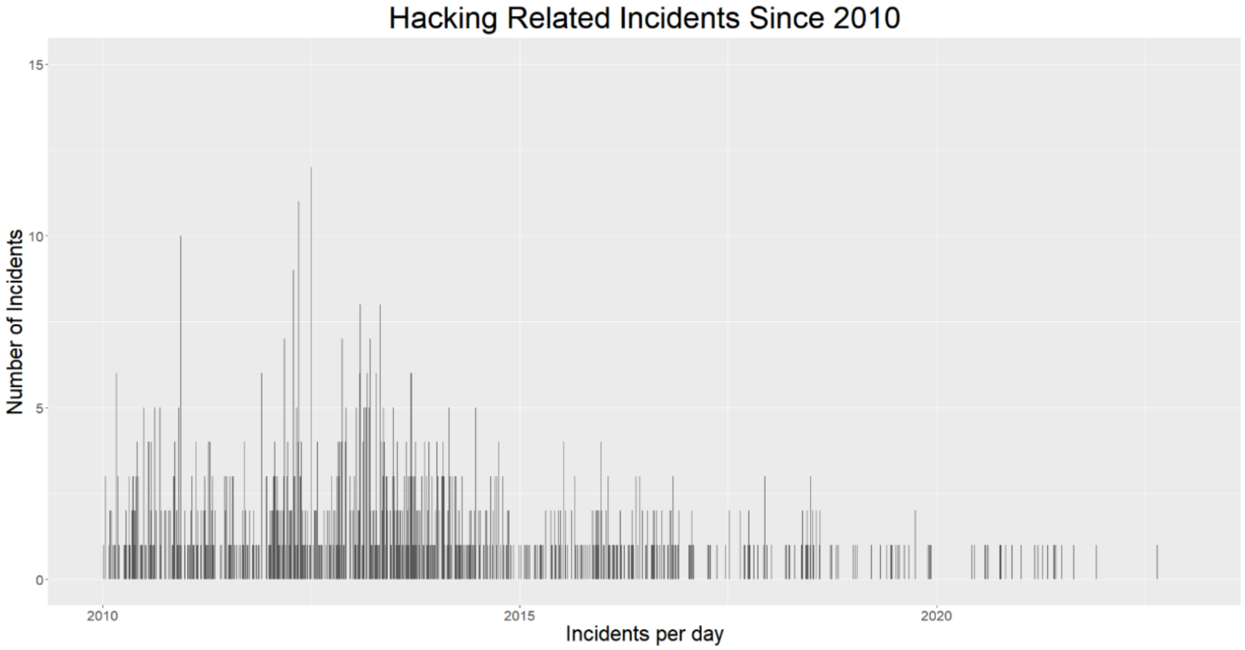


Figure 1. Hacking incidents per day since 2010.

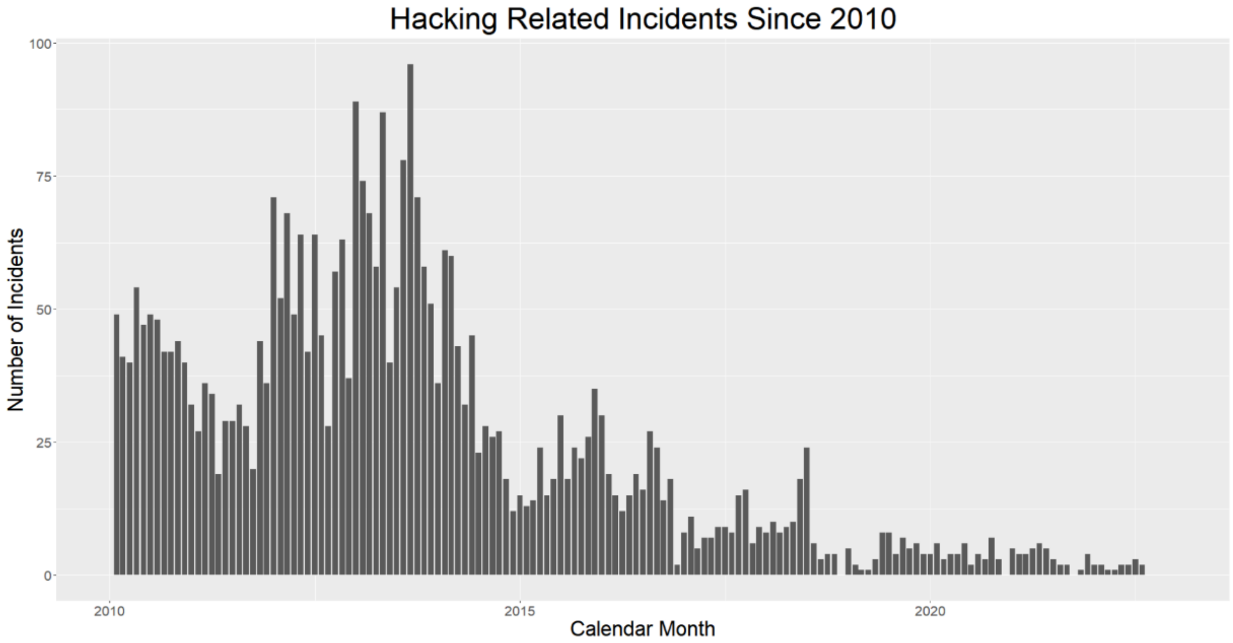


Figure 2. Hacking incidents per month since 2010.

There can also be more than one actor involved in any incident, and their actions can be malicious or non-malicious, intentional or unintentional, causal or contributory.

Updates

- Description of Data [LK]
- Hacking Plots in R [WB]

Action Plan

- Understand and perform classification of various attributes. [RS, LK]

Analysis

Methods

Results

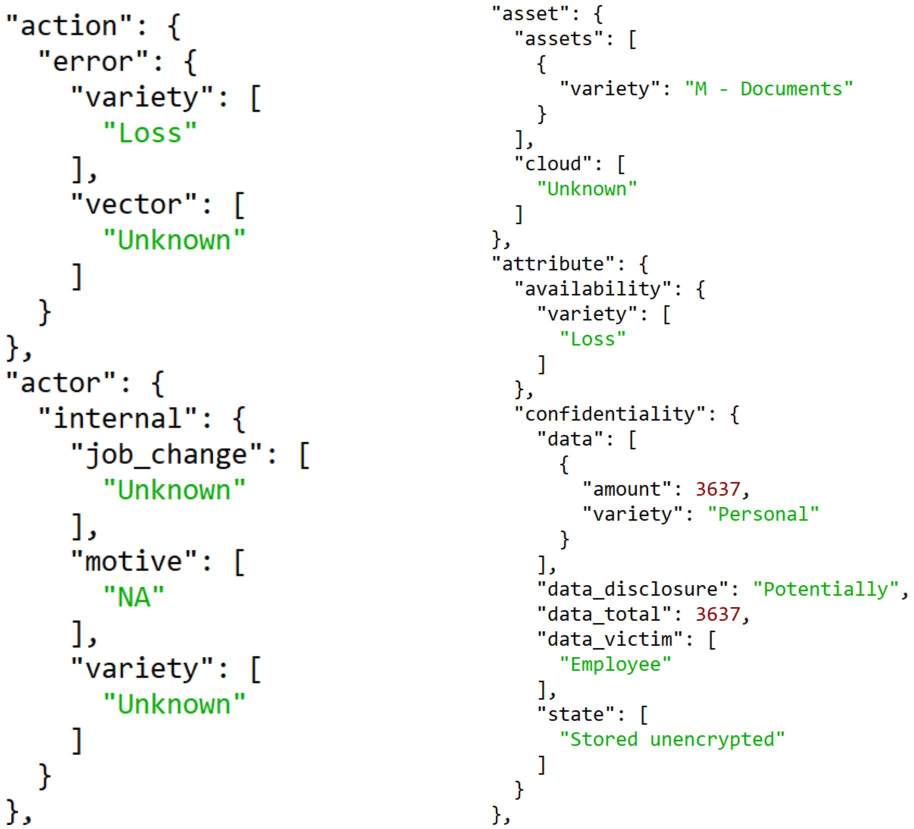
Discussion

Conclusions

References

Appendix A

The summarized raw data as obtained from the VCDB's Github page is encoded in the following VERIS format in a .json file and is shown in Figure A1. It's like a dictionary class with nested attributes.



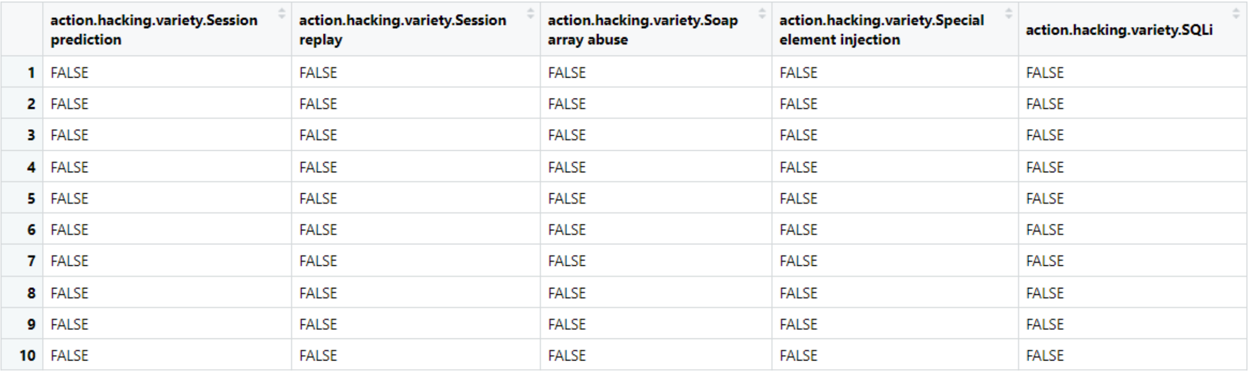
```
"action": {
  "error": {
    "variety": [
      "Loss"
    ],
    "vector": [
      "Unknown"
    ]
  }
},
"actor": {
  "internal": {
    "job_change": [
      "Unknown"
    ],
    "motive": [
      "NA"
    ],
    "variety": [
      "Unknown"
    ]
  }
},
```

```
"asset": {
  "assets": [
    {
      "variety": "M - Documents"
    }
  ],
  "cloud": [
    "Unknown"
  ]
},
"attribute": {
  "availability": {
    "variety": [
      "Loss"
    ]
  },
  "confidentiality": {
    "data": [
      {
        "amount": 3637,
        "variety": "Personal"
      }
    ],
    "data_disclosure": "Potentially",
    "data_total": 3637,
    "data_victim": [
      "Employee"
    ],
    "state": [
      "Stored unencrypted"
    ]
  }
},
```

Figure A1. Format of VERIS schema in a .json file.

The data as described in Figure A1 was simplified and transformed into a functional data frame by using the VERISR package in R and is shown in Table A1.

Table A1. Representation of data in .csv format.



	action.hacking.variety.Session prediction	action.hacking.variety.Session replay	action.hacking.variety.Soap array abuse	action.hacking.variety.Special element injection	action.hacking.variety.SQLi
1	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE

It has all Boolean True/False entries.

Updates

- Description of data. [RS, LK]

Action Plan

- Simplify the data frame even further. [RS, LK]

Appendix B

Appendix C