Datasheet for Dataset of Data Breaches and Ransomware Attacks Over 15 Years*

Shivank Goel

April 10, 2024

This datasheet documents the datasets used in a studying data breaches and ransomware attacks over 15 years, from 2004 to 2019. The dataset in the study is compiled by researchers Tsen, Elinor, Ko, Ryan, and Slapnicar, Sergeja, and is based on public data.It serves as a great resource for analyzing trends and patterns in organizational cyber resilience. The focus is on the organizational size, industry sector, and impact of cybersecurity strategies, helping us to better understand how businesses deal with digital threats.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

The dataset was created to investigate the cybersecurity incidents impacting various organizations, considering their size and sector. It helps to understand the importance of cybersecurity and the need for a strong defense against such threats.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

The dataset was compiled by Tsen, Elinor, Ko, Ryan, and Slapnicar, Sergeja, affiliated with The University of Queensland.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

[*] Code and data are available at: https://github.com/shivankgoel003/DataBreach_Ransomware_Stat.

The dataset compilation was part of a PhD project, presumably supported by funding allocated for academic research by The University of Queensland. No specific grant number is associated with the dataset creation, as per the available information.

4. *Any other comments?*

The dataset is a resource for researchers, policymakers, and cybersecurity professionals to study cyber resilience in today's digital world.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The instances within the dataset represent cybersecurity incidents, especially data breaches and ransomware attacks on various organizations. Each instance shows details such as the nature of the attack, the sector and size of the organization affected, the geographical location, and the impact and response measures.

2. *How many instances are there in total (of each type, if appropriate)?*

The exact number of cyberattack instances are 1146 of each type withing 15 years of time frame from 2004 to 2020.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

The dataset does not cover all possible instances but represents a broad selection from multiple resources. It is a sample of large number of cyber incidents, focusing on publicly reported and confirmed attacks. It includes a diverse range of incidents from various sectors and countries to provide a representative overview of cyber threats.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

Each instance includes processed data with features such as the year of the attack, the name of the organization, critical industry classification, organization size, level of digital intensity, the sector, country, details of the cyber security role and frameworks, impact level, and actions taken. This provides a structured understanding of each incident.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

Each instance is labeled with the type of attack (e.g., ransomware, data breach) and classified by its impact level (e.g., low, medium, high) based on the response taken by the organization.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

There is many missing information in the dataset.This might be due to provacy concerns, or some orgnizations did not report the exact information. For example, instances such as number of users affected has lot of gaps in the dataset.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

The datasets do not explicitly mention relationships between individual instances, however they can be drawn while doing statistical analysis.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

There are no specific recommended data splits mentioned. Data splits would depend on the specific research or analysis goals.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

Specific errors or sources of noise are not mentioned. However, as with any databreach report data, there may be variabilities and inconsistencies due to natural fluctuations and measurement limitations.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The datasets are self-contained and available through the website of University of Queensland.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

No confidential or personally identifiable information is included in the dataset.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

There is no content in these datasets that would be considered offensive or insulting. The data is factual, focusing on data breaches and cyber incidents.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

The datasets do not specify sub-populations like age or gender.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

Individual identification is not possible with these datasets as they deal with aggregated data covering various countries.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

The datasets do not contain sensitive personal data.

16. *Any other comments?* N/A

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The data were directly observed and collected through various resources such as : Carnegie Mellon's List of Banking Cyber Incidents (2020), Repository of Industrial Security Incidents (2015), Privacy Rights ClearingHouse (2020) and Information is Beautiful (2020).

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

Not applicable since the dataset was compiled from various resources.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

The datasets seem to be collections rather than samples.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

Not applicable, since the dataset was compliation of data from other resouces related to cyber atatcks and cyber security.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

The data collection covers several years, as indicated by the datasets (e.g., 2004-2020 for the data breach incidents).

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

There is no specific mention of ethical review processes.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

Not applicable

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

Individual notification is not applicable, as the data do not pertain to individual persons.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

Individual consent is not relevant for this type of data, as it does not involve personal data collection from individuals.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

As individual consent was not required for this data collection.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

There is no information provided regarding an impact analysis on data subjects, likely because the datasets deal with environmental and health data and not individual-specific data.

12. *Any other comments?*

The data collection process is extensive and critical for understanding the environmental and public health in Alberta. The absence of personal data reduces privacy concerns.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

There is likely some level of preprocessing and cleaning involved, ensuring consistency, and possibly dealing with missing values. However, specific details are not provided in the information given.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

There is no mention of whether the raw data is saved alongside the processed data.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

Details about the specific software or tools used for preprocessing, cleaning, or labeling are not provided.

4. *Any other comments?*

Proper data preprocessing and cleaning are essential for the accuracy and reliability of analysis results. The datasets, as released, should be ready for use in various analyses and research.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   The dataset has been used in academic studies to analyze and understand cybersecurity incident trends. The author themselves have published an academic paper based on the dataset.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

The University of Queensland's eSpace repository contains links to publications and related datasets. The dataset and related works can be accessed at UQ eSpace.

3. *What (other) tasks could the dataset be used for?*

The data could be used for further statistical analysis in the field of cyber risks and cybersecurity.It may also be used for educational purposes, to train machine learning models, or for benchmarking organizational cybersecurity practices.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

The users should consider potential biases and reporting in data collection.Also they should be aware of geographic coverage to avoid misinterpretations.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset should not be used for tasks that could compromise the privacy or security of the affected organizations.

6. *Any other comments?*

The datasets are valuable for understanding trends and correlations in cyber incidents. They should be used appropriately, considering their scope and limitations.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

The dataset is already in the public domain and is available to third parties for research and analysis. It can be accessed globally, subject to the terms of use set by the University of Queensland.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

The dataset is distributed online through the University of Queensland's data collection portal. It does have a DOI, which can be used to cite the dataset in academic work.

3. *When will the dataset be distributed?*

The dataset is already available and has been since its publication year in 2020.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

Yes, the dataset is distributed under a license that permits reuse with acknowledgment, and the terms are outlined in the dataset's access conditions on the University of Queensland's website.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

There is no mention of any third-party IP restrictions on these datasets.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

There are no export controls or regulatory restrictions mentioned for this dataset.

7. *Any other comments?*

The open access nature of these datasets makes them a valuable resource for a wide range of users, from researchers to policymakers.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

The dataset will continue to be hosted and maintained by the University of Queensland, ensuring its accessibility for ongoing and future research.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

Enquiries about the dataset can be directed to the corresponding author of the research or the data management team at the University of Queensland. Contact can typically be made through the contact details provided on the dataset's webpage.

3. *Is there an erratum? If so, please provide a link or other access point.*

There is no mention of an erratum.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

Updates to the dataset, if any, will be managed by the research team at the University of Queensland. They would be responsible for correcting any errors, adding new data, or removing outdated instances.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

There is no specific information about the maintenance of older versions of the dataset.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

This is not applicable as the datasets do not contain individual-level data.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

There is no mentioned mechanism for external contributions to the dataset. The data are sourced from government monitoring and records, so external contributions might not be applicable.

8. *Any other comments?*

Regular maintenance and updates are crucial for the continued relevance and accuracy of these datasets.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.