The Impact of Aggregated Leaks on Privacy

Caspar Martens Simon Kindhauser Pascal Lehmann Mitra Purandare

June 16, 2023

Abstract

Privacy is in a constant battle with economic, scientific and security interests. In our modern world enormous amounts of personal and identifying information about us is collected and analyzed. While significant resources are allocated to regulate this industry, criminals are not bound by such limitations.

The exploitation of illegally obtained information is already estimated to be a billion dollar industry and its growth has not plateaued yet. In order to be ready for the ever increasing capabilities of cyber criminals it is crucial to examine the challenges that await.

This paper studies network effects emerging from the aggregation and linkage of illegally obtainable personally identifiable information. We achieve this by analyzing a synthetic dataset which represents over 600 distinct data leaks that resulted in the exposure of over 12 billion records.

We show the effectiveness of mitigation strategies such as using different email addresses and introduce User-Initiated Differential Privacy as an effective defense, reducing the risk of link discovery by up to 50%.

1 Introduction

In recent years privacy has come under attack from various directions. Some companies are able to capitalize behavioral patterns in a practice sometimes referred to as surveillance capitalism, and the so called "Crypto Wars" continue to be fought as governments – including the European Union – try to limit the use of strong encryption, a proposal which undermines efforts by many security researchers advocating for privacy by design.

To share data, especially in research, the law often obligates to anonymize sensitive information. One particularly popular technique is k-anonymity. Various problems with this technique have been identified in the past but in 2022, Cohan demonstrated a practical attack on a professionally anonymized dataset and was able to re-identify multiple individuals[1]. He argues that k-anonymity does in fact not properly anonymize data as required by GDPR.

Whilst attacking k-anonymized datasets at scale requires considerable effort, these are not the only publicly available datasets containing potentially sensitive information. Leaked, stolen or scraped data is far more dangerous and arguably even more available. Additionally, it is being used by cyber criminals to whom regulations mean relatively little.

It is widely accepted that absolute security is unachievable and in the context of privacy we experience this through the leak of personally identifiable information (PII). In the year 2020 alone at least 4.6 billion assets of PII have been leaked in no less than 850 data breaches[2]. Related work is predicting the occurrence and size of data breaches[3][4]. Meanwhile others have quantified the impact of data breaches on companies in monetary terms[5] and on their reputation[6] as well as on the change of their hiring practices[7].

While the negative impact of leaked PII on an individual is hard to quantify, data leaks containing PII have been positively correlated with the amount of identity thefts[8]. However, the impact on privacy itself is established as soon as the rights of an individual are violated. In the context of the GDPR we argue that every asset of PII leaked and associated to a person is a violation of – an thus a negative impact on – their privacy[9][10].

To establish this impact we analyze a generated dataset (2.2) synthesized from *Have I Been Pwned*[11] metadata describing more than 12 billion records. By empirically studying the effects of aggregation in the context of leaked data we quantify the association of records to individuals. Additionally we evaluate different mitigation strategies experimentally using advanced big data processing tech-

niques to understand their impact on the privacy of individuals. This enables us to make recommendations on how individuals can protect their privacy optimally without depending on companies or regulations.

2 Methodology

Our research is based on an empirical approach to the problem of privacy. We will use a synthetic dataset to simulate the aggregation of leaked information and empirically validate mitigation strategies against it.

2.1 Leaked Data

Initially, the dataset was to be compiled from real data leaks that are freely available on the internet and darknet. While legal inquiries have been made before the start of the project, a more detailed analysis during the project has concluded that the risks for our institution and the researchers involved overwhelm the benefits of using actual data. We are fully convinced that using real data would be more beneficial and would be feasible but has legal and ethical caveats[12] that could only be overcome with a significant amount of resources that are out of scope for this project. Because of this a synthetic dataset has been generated for the analysis as described in Section 2.2.

2.1.1 Limitations

The decision not to use actual data prevents us from analysing the distribution of the actual dataset. Hence all distributions used in the synthetic dataset are estimations that we were unable to verify. The assumptions made are based on the available data from other sources[13][14][15][16] as very little literature in this context exists. These estimations are described in the Appendix A.

In general working with a synthetic dataset means that we are working on an idealized version of the problem. Errors that could be introduced for example by incomplete or invalid data standardization or missing values are not present in the synthetic dataset. As the results of our paper are based on the accuracy of the distribution estimations our results should be interpreted with caution.

2.1.2 Benefits

While considerable limitations of using a synthetic dataset have been identified, there are also benefits. By publishing our data generation pipeline we can ensure the reproducibility of our research which would be much harder with actual data.

As mentioned we are working on an idealized version of the problem which greatly reduces the complexity of the analysis. Tasks such as data normalization were not necessary. This allows us to focus on the core problems such as the data linkage and the impact of aggregation. The synthesized dataset also allows for an additional approach we call simulation that is significantly more efficient and described in Section 2.2.4.

The most important benefit of a generated dataset is that we know whether a linkage is correct or not. This allows us to verify our algorithms and methods against the ground truth of the dataset. As the algorithms developed are specifically crafted to work both on synthetic and real data this means that future research can be made with the same algorithms that were validated in this paper.

2.2 Synthetic Dataset

The synthetic pica dataset has been generated by the authors for the purpose of this paper. It uses only publicly available meta information regarding the occurrence, size and contained attribute classes of real data breaches, scrapes and collections. In combination with estimated distributions of attributes we configured our data generator and simulator to reproduce statistically similar synthetic versions of the real data.

Two versions of this dataset exist. The first is a simulated dataset which deterministically generates the data for one person at a time on the fly. The second is a generated subset of the entire dataset. Both datasets use the same code base to generate the data.

The simulated dataset consists of 12 billion records and would when generated consist of over 2^{35} unique entities which would require over 3 TB of disk space. The generated dataset is a subset one tenth the size of the simulated dataset representing the population of the United States.

2.2.1 Data Model

The pica data model is a simple yet highly extensible model consisting of only three major

types as shown in Figure 1.

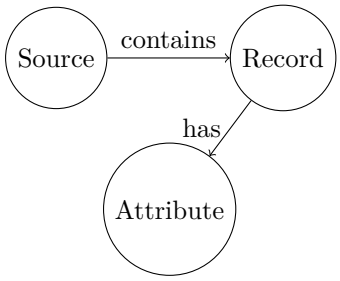

Figure 1: The pica data model

Attributes are everything from names to phone numbers. The node includes only the attribute class (e.g. email), a probability describing the uniqueness characteristics and a unique, irreversible identifier based on the occluded value of the attribute. The available attribute classes are described in Section 2.2.2.

Records are collections of attributes. Their id is based on the ids of the attributes they contain. A special property in this dataset is the inclusion of an artificial person identifier. This information is used to verify the linkage algorithms.

Sources represent the origin of the information. It contains meta information such as the date of the leak or publication. This can be used to simulate the aggregation of data over time.

2.2.2 Classes of Attributes

The used repository of data leaks defines a wide variety of attribute classes which can be contained in a source. The synthetic pica dataset limits itself to Names, Usernames, Genders, Dates of birth, Geographic locations, Email addresses, Phone numbers, Social media profiles and Passwords. These attribute classes were chosen because they belong to the most common or most identifying classes. To ensure the statistical similarity with reality each attribute class requires a parameterized distribution. The chosen parameters can be found in Appendix A.

Although our base algorithms do not distinguish between different attribute classes, we do add appropriate labels identifying their specific type. In practice they are needed to simulate applications of User-Initiated Differential Privacy and other mitigation strategies as described in Section 3.2.

2.2.3 Data Sources

The number of data leaks that have occurred in the past is unknown since many of them are not reported. A lower bound is set by the privacy rights clearing house which tracks obligatory data breach notifications in the USA[17]. They report that for the time frame from 2005 until 2022, there have been 20'030 data breaches with close to 2 billion records. This resource has been used in recent literature but it does not include complete information about the number of records and contained attribute classes for each incidence, which we require for our analysis. A renowned source of accurate and validated information regarding the contents of numerous leaks is *Have I Been Pwned*[11] by Troy Hunt. We use a version of this dataset which we have persisted as the basis for our data generation process. It includes information about 661 leaks and 12'482'354'793 records. A table with all contained sources can be found in Table 3.

2.2.4 Dataset Simulation

Due to a technical limitation it was impossible to import the full dataset into the database. To circumvent this limitation we developed an alternative approach which does not rely on the physical existence of the dataset. Instead it only generates the subgraph required for the current analysis step. This allows us to analyze the entire dataset without the need to import it first. The simulation starts with the creation of a person based on a given person identifier. The person identifier is associated with an inclusion probability based on the distribution of people in the sources. This probability is used to determine which sources this person is included in and in turn to produce records based on the persons information and the attribute types contained in the source. Finally, all this information is added to an in-memory graph for analysis. The drawback of this technique stems from the absence of other persons in the subgraph and the resulting impossibility of false positives. Interestingly, it is this limitation which enables us to efficiently implement a third-order algorithm.

2.2.5 Data Generation Pipeline

For the generated dataset we implemented a pipeline that produces records containing the attributes belonging to a specific virtual person. Our pipeline deploys a just-in-time approach, regenerating virtual persons from a

unique identifier as needed, since keeping 8 billion people in memory is unfeasible. Which is why we derive all attributes from the person identifier in a deterministic manner by seeding a random generator which guarantees the same attributes for persons with the same identifier. When synthesizing data, the first step is to sample person identifiers from a predefined distribution. Next, a virtual person is instantiated and queried for the attributes as defined in the configuration for the source. From that a record is produced which is then serialized into a CSV file, one entity per line. Once the records have been generated the pipeline ensures that duplicates exist neither for sources, records nor attributes. It then imports the data into our graph database for the analysis.

2.3 Record Linkage

Record linkage is becoming a researching tool that surges in popularity in the context of statistics on PHI-data[18] but also comes with caveats to be aware of[19]. There is a smorgasbord of available software that implements probabilistic data linkage on the basis of Newcombe et al[20][21] and Fellegi and Sunter[22] proposed formalized methods[18][23].

Much of the statistical literature and its implementations regarding record linkage are limited single-file record linkage and assume that each entity is contained only once in each file[23]. Our problem violates both limitations and thus another approach had to be used. Closest to our problem requirements came an approach by Aleshin-Guendel and Sadinle on multi-file partitioning for record linkage and duplicate detection[24]. But their approach was not applicable to our problem as our comparatively large dataset has to be processed with very high efficiency and a sophisticated statistical analysis is to be performed on the output. We have therefore decided to implement a custom algorithm that disregards the means of text similarity matching and relies on the uniqueness of the attributes as the predictor.

The synthetic pica dataset when loaded into a graph database allows for efficient traversal[25] of relationships between records via attributes. It also provides estimates for the probability of occurrence of the attribute values. As such, the system enables the analysis of the linkage process and the impact of mitigation strategies.

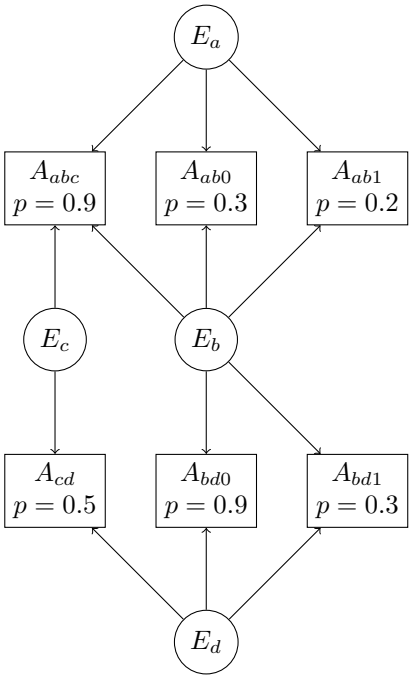

Figure 2: Record Linkage Graph

2.3.1 First-Order Record Linkage

First-order record linkage is limited to exploiting preexisting high probability connections. The linkage probability propagates with a simple multiplication of the individual attribute probabilities along the path. The border of a cluster is thus reached when the probability of the linkage falls below an arbitrary threshold.

$$\mathrm{p}(c,t) = \prod_{s \in \mathrm{spath}(c,t)} \mathrm{p'}(s)$$

where $\mathrm{spath}(c,t)$ is the shortest path in terms of probabilistic linkage between the two records c and t and $\mathrm{p'}(s)$ is the probability of the linkage in a segment:

$$\mathrm{p'}(s) = \max_{a \in s}$$

As such it is simple to implement and is relatively fast even on large datasets.

In the example in Figure 2 the first-order probability of the linkage between E_a and E_c is $0.9 \times 0.9 = 0.81$.

2.3.2 Second-Order Record Linkage

Second-order linkage is a generalization of first-order linkage. Similar to first-order linkage it also operates on the basis of segments and multiplies them along the path to calculate the link probability.

$$\mathrm{p}(c,t) = \prod_{s \in \mathrm{spath}(c,t)} \mathrm{p'}(s)$$

Where $\mathrm{spath}(c,t)$ is the shortest path in terms of probabilistic linkage between the two records c and t and p'(s) is the probability of the linkage in a segment:

$$\mathrm{p'}(s) = 1 - \prod_{a \in s}(1-a)$$

It is able to increase the linkage probability by considering all attributes in a segment. As such it is able to find links with no high probability attribute if enough attributes are present.

In the example in Figure 2 the second-order probability of the linkage between E_a and E_d is $[1-(1-0.9)(1-0.3)(1-0.2)] \times [1-(1-0.9)(1-0.3)] \approx 0.878$

2.3.3 Third-Order Record Linkage

Third-order linkage works differently than the other linkage algorithms. Dissimilar to the first- and second-order algorithms, it does not operate on segments but rather on sets of attributes. It maintains a mapping between attributes and the estimated probability of them being associated with the originating record. As such it is able to find transitive links, meaning that if the intersection of attributes between a cluster of records and a candidate record would suffice to establish a link, the algorithm is able to do so even if no segment in itself contains all the attributes.

The probability of the correctness of a cluster is calculated as follows:

$$p(c) = \prod_{t \in c} p'(c,t) \tag{1}$$

where $p'(c,t)$ is the probability a record t is part of the cluster c:

$$p'(c,t) = \sum_{s \in c \setminus t} p''(c,s) * o2(s \wedge q) \tag{2}$$

where $p''(c,s)$ is the weight of the subcluster s within the cluster c and $o2(A)$ is the second-order probability for a set of attributes A:

$$p''(c,s) = p(s) - \sum_{s \in q \in c} p''(q) \tag{3}$$

$$o2(A) = 1 - \prod_{a \in A}(1-p(a)) \tag{4}$$

Note that in these equations the decorrelation is part of the selection criteria in the sum and product formulas. This is required to ensure that the probability of a cluster is not overestimated by the inclusion of subclusters that are already part of the cluster.

2.3.4 Cutoff Prameter

Our algorithms include a cutoff probability that is used as a classification threshold with which we can tune the precision of the algorithm. Through experimenting with that value we have concluded that a cutoff probability calculated as follows yields the best results in the context of our research:

$$\text{cutoff probability} = 1 - \frac{1}{N - \mathrm{u}} \tag{5}$$

Where:

$$N = \text{population size}$$
$$\mathrm{u} = \text{attribute uniqueness}$$

The population depends on our approach and is either the number of people in the US population, or an estimate of the world population count. The attribute uniqueness is the number of people that are expected to share an attribute. A value of 3 would mean that an expected 3 people may share this attribute value. An attribute uniqueness of 1 means that the attribute is expected to be unique to a single person and hence the natural choice.

The cutoff probability can be adjusted depending on the use case. In an attack scenario the attribute uniqueness can be increased to achieve a higher number of true positives in exchange for a respective increase in false positives.

2.3.5 Analysis

The metric we use to quantify the effectiveness of a mitigation strategy is the discovery ratio. It is the ratio of the number of records in a cluster to the total number of records in all sources associated to a specific person. It is a measure of the effectiveness of the linkage algorithm. The higher the discovery ratio, the more records were successfully linked to the same person. The discovery ratio is calculated as follows:

$$\text{discovery ratio} = \frac{\text{discovered records}}{\text{total records in all sources}}$$

To compare the different mitigation strategies we calculate the average discovery rate over N samples. The samples are generated by randomly sampling the person identifier from a uniform distribution.

3 Results

First we establish the accuracy of our algorithms by comparing their results to the ground truth which we have access to since the synthetic dataset includes the necessary information. Based on this we show that our algorithms are able to link records effectively, supporting the hypothesis that aggregation and the abundance of personally identifiable information is a threat to privacy. To counter this threat we evaluate different mitigation strategies analyze their effectiveness to protection the privacy of individuals.

3.1 Algorithm Verification

The verification of our algorithms is based on the data generated as described in Section 2.2.5. We used a sample size of 100 thousand starting records. It is important to note, that although this approach is different from the technique described in Section 2.2.4 — which is used to analyze the mitigation strategies — both use the exact same algorithms for generating records and finding links.

Due to practical limitations in our implementation of the third-order algorithm we focus our analysis primarily on the second-order algorithm. Nevertheless the results shown in Figure 4 give an overview of the accuracy and performance of the first-order algorithm as well.

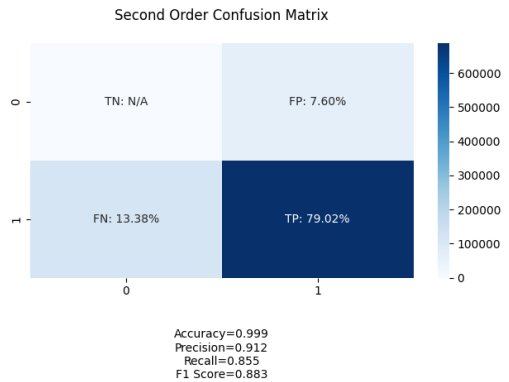

Figure 3: Confusion matrix of the second-order algorithm.

Our analysis suggest a true positive rate of 85% suggesting that our algorithms are able to link a majority of records together. Furthermore, the high precision of 91% satisfy our requirements for analyzing the effectiveness of our mitigation strategies.

As can be seen in Figure 3 we have refrained from including the true negative matches. This is because true negatives represent possible links from a single record to all other records not belonging to the same person and is therefore a huge number that would skew the percentages in the plot. However, we included the number in the calculation of the accuracy which is the reason for the very high result of 99.99%.

Figure 4 compares the true positive cluster sizes to the predicted cluster sizes. The difference between the two visualizes a major contributor to false positives. An in-depth analysis of the observed effect concluded that it is caused by the incorrect linkage of two clusters. Thus the graphic includes a notable overestimation of the cluster size at the higher end of the spectrum. This overestimation is visualized by the area that spans from the solid line representing the true positive cluster sizes to the actual cluster sizes recorded by our algorithms.

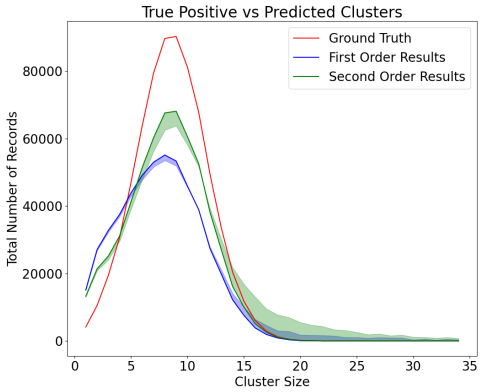

Figure 4: Difference between query results and true positive clusters.

3.1.1 Network Effect

We found that the discovery ratio improves with the number of sources. With more records the probability to find a link increases linearly for all tested algorithms. This can be seen in Figure 5. This illustration was created with the recommended mitigation strategy applied as described in Section 3.2.4. As the email address is the most common attribute

type in our dataset a graphic without the mitigation strategy applied would show less of a difference between the algorithms.

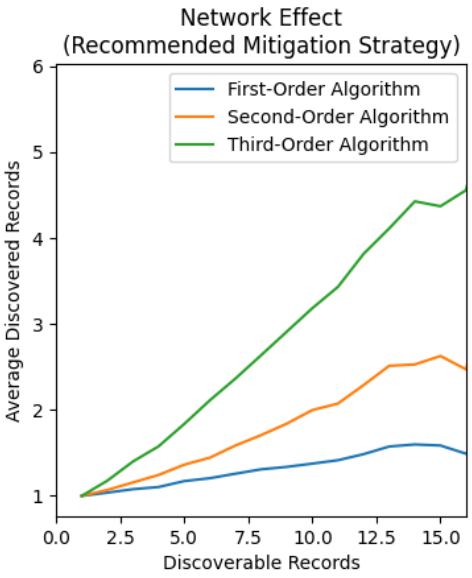

Figure 5: Impact of the number of sources on the average discovery ratio.

3.2 Mitigations

We looked at different mitigation strategies and measured their impact on our dataset. All are based on the occlusion of attributes which will make it harder to establish links between records.

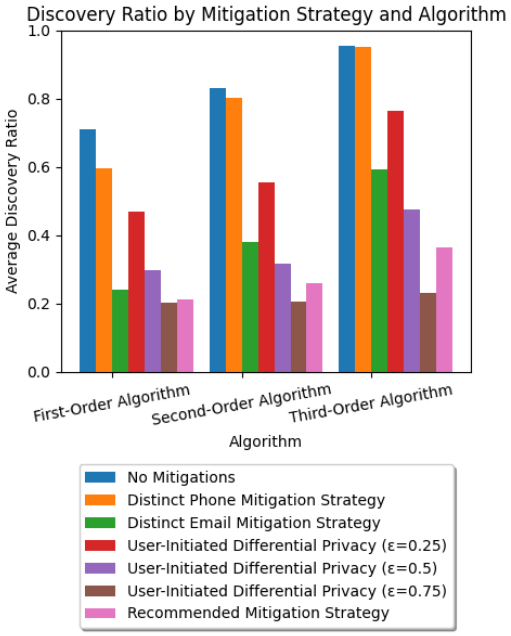

Figure 6: Impact of mitigation strategies on the discovery ratio.

3.2.1 Using Unique Phone Numbers

The occlusion of phone numbers has a comparatively small impact on the discovery ratio. One factor contributing to this is the relatively low number of sources including the phone numbers. Additionally, services which include phone numbers often also include other QIs such as email addresses or location thus we see the benefits of this mitigation strategy getting smaller with more sophisticated algorithms.

3.2.2 Using Unique Email Addresses

On our dataset, the occlusion of email addresses has a significant impact on the discovery ratio. This is due to the fact that email addresses are the most common QI in our dataset. But it is important to note the performance of higher order algorithms, which are still able to uncover up to 60% of records when this mitigation strategy is used exclusively.

3.2.3 User-Initiated Differential Privacy

Figure 6 shows that using unique email addresses and the occlusion of phone numbers is not enough to meaningfully mitigate advanced

linkage attacks. Thus we introduce a new mitigation strategy we call User-Initiated Differential Privacy. With this mitigation the user decides to occlude an attribute with a probability ϵ. We show that this mitigation strategy is able to significantly reduce the discovery ratio for all algorithms. However, with relatively low values of ϵ this mitigation alone preforms worse than the unique email mitigation alone.

3.2.4 Recommended Mitigation Strategy

Our results suggest that the best possible protection can be achieved with high ϵ values in User-Initiated Differential Privacy. However, this approach might not be practical for various reasons including that the information is truly relevant for the service requesting it. Thus we recommend the combination of unique email, unique password and User-Initiated Differential Privacy as a mitigation strategy which offers a 50% reduction in Discovery Rate as compared to No Mitigation even with a relatively low ϵ value of 0.25. Furthermore, tooling to assist with this mitigation strategy is already available[26][27][28].

4 Conclusion

We have shown that linkage attacks on datasets statistically similar to what is publicly available to cyber criminals are effective in identifying information belonging to the same individual. Through our analysis we also demonstrated the feasibility of such attacks by sufficiently motivated actors. As such we argue for the urgency of mitigating countermeasures to protect the privacy and security of individuals and organizations alike.

The analysis of the proposed mitigation strategies shows that recently popularized counter measurements such as using unique email addresses are not sufficient on their own to effectively prevent linkage attacks. Therefore we recommend using unique email addresses and passwords in combination with User-Initiated Differential Privacy. This method presents a good tradeoff between practicality and security offerings for most people. This strategy reduces the average risk of intra-record link discovery by half when tested against our third-order linkage algorithm, the most powerful attack discussed in this paper.

We recognize that this reduction seems underwhelming at first glance. However, considering the sophistication of the recommended mitigation we believe it to reflect the inherit risk posed by linkage attack in general. As such we believe that hard privacy as a guidance during data collection deserves more attention as it reduces the attack surface of all costumers or citizens of a company or government respectively.

Lastly, we want to emphasize that the risk posed by linkage attack is increasing with every new data breach. The documented network effect suggests that the risk grows linearly with the number of leaked records. This was a surprising finding to us but is not at all a sign of relieve.

4.1 Future Work

The area of research surrounding linkage attacks still has a lot of unanswered questions. The insights we gained from our analysis are but a fraction of what could be uncovered from our dataset in future work.

The documentation of new and improved linkage algorithms could further publicize the dangers of linkage attacks. Such work needs to respect the ethical implications of the findings they present and carefully consider their impact. We believe that access to high quality analysis of linkage algorithms will benefit the public discourse on privacy and security.

As the total number of leaked records gets larger and algorithms get better, the need for more resilient mitigation methods increases too. In our work we limited our analysis to a few promising mitigation strategies. Future work could improve on our mitigations or propose new ones. A possible mitigation strategy not discussed in this paper is the overall reduction in accounts and services used by a person or the use of single sign-on solutions.

A meta analysis of the long term feasibility of user-initiated mitigation strategies in the face of ever increasing amounts of available data could yield valuable insights into the benefits of stricter privacy policies and regulations. Research in this area could benefit the larger privacy and security community and potentially produce new grounds for arguments for hard privacy.

Finally, research conducted with real data would be invaluable. The increase in accuracy and overall realism of the results further increases the legitimacy of the core arguments of this paper. Research in this area is difficult and local laws and regulation as well as ethical consideration need to be taken seriously.

However, producing a properly anonymized dataset containing real data could give the community easier access to conduct independent research on the topic.

References

[1] A. Cohen and K. Nissim, "Towards formalizing the GDPR's notion of singling out," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 8344–8352, Apr. 2020. Publisher: Proceedings of the National Academy of Sciences.

[2] SpyCloud, "Credential Exposure Report," 2021.

[3] M. Barati and B. Yankson, "Predicting the Occurrence of a Data Breach," *International Journal of Information Management Data Insights*, vol. 2, p. 100128, Nov. 2022.

[4] S. Wheatley, T. Maillart, and D. Sornette, "The extreme risk of personal data breaches and the erosion of privacy," *The European physical journal. B, Condensed matter physics*, vol. 89, no. 1, pp. 1–12, 2016. Place: Berlin/Heidelberg Publisher: Springer Berlin Heidelberg.

[5] IBM, "Cost of a data breach 2022," Apr. 2023.

[6] K. Whitler and P. Farris, "The impact of cyber attacks on brand image: Why proactive marketing expertise is needed for managing data breaches," *Journal of Advertising Research*, vol. 57, pp. 3–9, 2017.

[7] S. Bana, E. Brynjolfsson, W. Jin, S. Steffen, and X. Wang, "Human capital acquisition in response to data breaches."

[8] F. Bisogni and H. Asghari, "More Than a Suspect: An Investigation into the Connection Between Data Breaches, Identity Theft, and Data Breach Notification Laws," *Journal of information policy (University Park, Pa.)*, vol. 10, pp. 45–82, 2020. Publisher: Pennsylvania State University Press.

[9] GDPR, "Art. 6 GDPR - Lawfulness of processing," Nov. 2018. Section: Uncategorized.

[10] GDPR, "Art. 7 GDPR - Conditions for consent," Nov. 2018. Section: Uncategorized.

[11] "Have i been pwned: API v3."

[12] M. Ienca and E. Vayena, "Ethical requirements for responsible research with hacked data," *Nature Machine Intelligence*, vol. 3, pp. 744–748, Sept. 2021.

[13] "Forebears: Names & Genealogy Resources," 2023. https://forebears.io/.

[14] "FiveThirtyEight Most Common Name Dataset," 2019. https://www.kaggle.com/datasets/fivethirtyeight/fivethirty most-common-name-dataset.

[15] "World Cities Database | Simplemaps.com," 2023. https://simplemaps.com/data/world-cities.

[16] "US Cities Database | Simplemaps.com," 2023. https://simplemaps.com/data/us-cities.

[17] P. Rights, "Data Breach Chronology | Privacy Rights Clearinghouse."

[18] S. March, M. Antoni, J. Kieschke, B. Kollhorst, B. Maier, G. Müller, M. Sariyar, M. Schulz, S. Enno, J. Zeidler, and F. Hoffmann, "Quo Vadis Data Linkage in Germany? An Initial Inventory," *Gesundheitswesen*, vol. 80, no. 3, pp. 20–31, 2018. Num Pages: 12 Number: 3 Publisher: Georg Thieme Verlag.

[19] M. A. Bohensky, D. Jolley, V. Sundararajan, S. Evans, D. V. Pilcher, I. Scott, and C. A. Brand, "Data Linkage: A powerful research tool with potential problems," *BMC Health Services Research*, vol. 10, p. 346, Dec. 2010.

[20] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, "Automatic Linkage of Vital Records," *Science*, vol. 130, pp. 954–959, Oct. 1959. Publisher: American Association for the Advancement of Science.

[21] H. B. Newcombe and J. M. Kennedy, "Record linkage: making maximum use of the discriminating power of identifying information," *Communications of the ACM*, vol. 5, pp. 563–566, Nov. 1962.

[22] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, Dec. 1969. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1969.10501049.

[23] M. Sadinle, "Bayesian Estimation of Bipartite Matchings for Record Linkage," *Journal of the American Statistical Association*, vol. 112, pp. 600–612, Apr. 2017. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1148612.

[24] S. Aleshin-Guendel and M. Sadinle, "Multifile Partitioning for Record Linkage and Duplicate Detection," *Journal of the American Statistical Association*, vol. 0, pp. 1–10, Dec. 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.2013242.

[25] M. Besta, P. Emanuel, R. Gerstenberger, M. Fischer, M. Podstawski, C. Barthels, G. Alonso, and T. Hoefler, "Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries," *arXiv.org*, 2022. Place: Ithaca Publisher: Cornell University Library, arXiv.org.

[26] "Firefox relay." https://relay.firefox.com/.

[27] "SimpleLogin." https://simplelogin.io/.

[28] "AnonAddy." https://anonaddy.com/.

A Distribution Estimation

The distributions for the attributes in our dataset were estimated on a best effort basis given the time constraint of the project. The following tables show the exact parameters used.

Table 1: Distributions for Generated Dataset

Attribute	Count	Distribution	US Parameters
Names	1	Linear Spline Distribution	points=(0, 0), (1, 31250), (10, 19850), (50, 11350), (100, 8000), (200, 5150), (300, 3250), (350, 2150), (375, 1200), (400, 250), (1000000, 10), (2000000, 2), (25000000, 1), (50000000, 0)
Usernames	1	Gauss Distribution	average=10'000'000, standard deviation=100'000
Genders	1	Bernoulli Distribution	probability=0.5
Dates of birth	1	Gauss Distribution	average=365*50, standard deviation=365*10
Geographic locations	1	Linear Spline Distribution	(0, 18 * 10**6), (1, 12 * 10**6), (2, 8.5 * 10**6), (10, 5 * 10**6), (20, 3 * 10**6), (50, 1 * 10**6), (100, 500 * 10**3), (200, 250 * 10**3), (500, 100 * 10**3), (750, 65 * 10**3), (1000, 50 * 10**3), (2000, 25 * 10**3), (3000, 15 * 10**3), (4000, 10 * 10**3), (5000, 7.5 * 10**3), (10000, 2.5 * 10**3), (20000, 500), (30000, 200), (100000, 0)
Email addresses	3	Uniform Distribution	from=0 to 10**12
Phone numbers	2	Uniform Distribution	from=0 to 10**12
Social media profiles	1	Uniform Distribution	from=0 to 10**12
Passwords	10	Gauss Distribution	average=336'500'000, standard deviation=10'000'000

Table 2: Distributions for Simulated Dataset

Attribute	Count	Distribution	World Parameters
Names	1	Linear Spline Distribution	points=(0, 0), (1, 750000), (10, 470000), (50, 245000), (100, 185000), (200, 125000), (300, 80000), (350, 50000), (375, 30000), (400, 6000), (100000000, 10), (200000000, 2), (250000000, 1), (250000000, 0)
Usernames	1	Gauss Distribution	average=10'000'000, standard deviation=100'000
Genders	1	Bernoulli Distribution	probability=0.5
Dates of birth	1	Gauss Distribution	average=365*50, standard deviation=365*10
Geographic locations	1	Linear Spline Distribution	points=(0, 30 * 10**6), (200, 5 * 10**6), (10**3, 850 * 10**3), (10 * 10**3, 12 * 10**3), (4 * 10**6, 0)
Email addresses	3	Uniform Distribution	from=0 to 10**12
Phone numbers	2	Uniform Distribution	from=0 to 10**12
Social media profiles	1	Uniform Distribution	from=0 to 10**12
Passwords	10	Gauss Distribution	average=8'000'000'000, standard deviation=100'000'000

B Data Sources

In Table 3 we list the data sources we used to generate the synthetic pica dataset (2.2). The information in this table is sourced from *have i been pwned*[11].

Table 3: Data Sources

Source	Records Count	Attributes
000webhost	14936670	Email addresses, Names, Passwords
123RF	8661578	Email addresses, Names, Passwords, Phone numbers, Usernames
126	6414191	Email addresses, Passwords
17173	7485802	Email addresses, Passwords, Usernames
17Media	4009640	Email addresses, Passwords, Usernames
2844Breaches	80115532	Email addresses, Passwords
2fast4u	17706	Email addresses, Passwords, Usernames
500px	14867999	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Usernames
7k7k	9121434	Email addresses, Passwords, Usernames
8fit	15025407	Email addresses, Genders, Geographic locations, Names, Passwords
8tracks	17979961	Email addresses, Passwords
Abandonia2022	919790	Email addresses, Passwords, Usernames
Abandonia	776125	Email addresses, Passwords, Usernames
ABFRL	5470063	Email addresses, Genders, Names, Passwords, Phone numbers
AbuseWithUs	1372550	Email addresses, Passwords, Usernames
AcneOrg	432943	Dates of birth, Email addresses, Passwords, Usernames
ActMobile	1583193	Email addresses

Table 3, continued

Source	Records Count	Attributes
Adapt	9363740	Email addresses, Names, Phone numbers, Social media profiles
Adecco	4284538	Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers
Adobe	152445165	Email addresses, Passwords, Usernames
AdultFanFiction	186082	Dates of birth, Email addresses, Names, Passwords
AdultFriendFinder2016	169746810	Email addresses, Passwords, Usernames
AdultFriendFinder	3867997	Dates of birth, Email addresses, Genders, Geographic locations, Usernames
AerServ	66308	Email addresses, Names, Passwords, Phone numbers
AgusiQTorrents	90478	Email addresses, Passwords, Usernames
AhaShare	180468	Email addresses, Genders, Geographic locations, Passwords, Usernames
Aimware	305470	Email addresses, Passwords, Usernames
Aipai	6496778	Email addresses, Passwords
AIType	20580060	Dates of birth, Email addresses, Genders, Geographic locations, Names, Phone numbers, Social media profiles
Ajarn	266399	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers
AKP	917461	Email addresses
AmartFurniture	108940	Email addresses, Names, Passwords, Phone numbers
Ancestry	297806	Email addresses, Passwords
AndroidForums	745355	Dates of birth, Email addresses, Passwords
AnimalJam	7104998	Dates of birth, Email addresses, Genders, Names, Passwords, Usernames
AnimeGame	1431378	Email addresses, Passwords, Usernames
AnimePlanet	368507	Dates of birth, Email addresses, Passwords, Usernames
Animoto	22437749	Dates of birth, Email addresses, Geographic locations, Names, Passwords
AntiPublic	457962538	Email addresses, Passwords
Apollo	125929660	Email addresses, Geographic locations, Names, Phone numbers, Social media profiles
Appartoo	49681	Email addresses, Genders, Names, Passwords, Social media profiles
Appen	5888405	Email addresses, Names, Passwords, Phone numbers
Aptoide	20012235	Email addresses, Names, Passwords
ArmorGames	10604307	Dates of birth, Email addresses, Genders, Geographic locations, Passwords, Usernames
ArmyForceOnline	1531235	Email addresses, Geographic locations, Names, Passwords, Usernames
Artsy	1079970	Email addresses, Names, Passwords
Artvalue	157692	Email addresses, Names, Passwords, Usernames
AshleyMadison	30811934	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
Astoria	11498146	Dates of birth, Email addresses, Names, Phone numbers
AstroPID	5788	Email addresses, Names, Passwords, Usernames
Aternos	1436486	Email addresses, Passwords, Usernames
AtlasQuantum	261463	Email addresses, Names, Phone numbers
Audi	2743539	Dates of birth, Email addresses, Names, Phone numbers
Autocentrum	143717	Email addresses, Passwords
Autotrader	20032	Email addresses, Phone numbers
Avast	422959	Email addresses, Passwords, Usernames
Avvo	4101101	Email addresses, Passwords
B2BUSABusinesses	105059554	Email addresses, Names, Phone numbers

Table 3, continued

Source	Records Count	Attributes
BabyNames	846742	Email addresses, Passwords
Badoo	112005531	Dates of birth, Email addresses, Genders, Names, Passwords, Usernames
BannerBit	213415	Email addresses, Passwords
Banorte	2107000	Email addresses, Genders, Names, Phone numbers
BattlefieldHeroes	530270	Passwords, Usernames
Battlefy	83610	Email addresses, Passwords, Usernames
BeautifulPeople	1100089	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords
Bell2017	2231256	Email addresses, Geographic locations, Names, Passwords, Phone numbers, Usernames
Bell	20902	Genders, Passwords, Usernames
Benchmark	93343	Email addresses, Passwords, Usernames
Bestialitysextaboo	3204	Dates of birth, Email addresses, Genders, Geographic locations, Passwords, Usernames
Bhinneka	1274340	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
bigbasket	24500011	Dates of birth, Email addresses, Names, Passwords, Phone numbers
BigMoneyJobs	36789	Email addresses, Names, Passwords, Phone numbers
BinWeevils	1287073	Email addresses, Genders, Passwords, Usernames
BiohackMe	3402	Email addresses, Passwords, Usernames
BitcoinTalk	501407	Dates of birth, Email addresses, Genders, Passwords, Usernames
Bitly	9313136	Email addresses, Passwords, Usernames
BitTorrent	34235	Email addresses, Passwords, Usernames
BlackBerryFans	174168	Email addresses, Passwords, Usernames
BlackHatWorld	777387	Dates of birth, Email addresses, Passwords, Usernames
BlackSpigotMC	140029	Email addresses, Genders, Geographic locations, Passwords, Usernames
BlankMediaGames	7633234	Email addresses, Passwords, Usernames
BlueSnapRegpack	104977	Email addresses, Names, Phone numbers
Bolt	995274	Email addresses, Passwords, Usernames
BombujEu	575437	Email addresses, Passwords
Bonobos	2811929	Email addresses, Names, Passwords, Phone numbers
Bookchor	498297	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles
Bookmate	3830916	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Usernames
BotOfLegends	238373	Email addresses, Passwords, Usernames
BourseDesVols	1460130	Dates of birth, Email addresses, Names, Phone numbers
Boxee	158093	Dates of birth, Email addresses, Geographic locations, Passwords, Usernames
BrandNewTube	349627	Email addresses, Genders, Passwords, Usernames
Brazzers	790724	Email addresses, Passwords, Usernames
BTCAlpha	362426	Email addresses, Passwords, Usernames
BTCE	568340	Email addresses, Passwords, Usernames
BtoBet	444241	Dates of birth, Email addresses, Geographic locations, Names, Usernames
BTSec	4789599	Email addresses, Passwords
Bukalapak	13369666	Email addresses, Names, Passwords, Usernames
BulgarianNationalRevenueAgency	471167	Email addresses, Names, Phone numbers
BusinessAcumen	26596	Email addresses, Names, Passwords, Usernames
CafeMom	2628148	Email addresses, Passwords

Table 3, continued

Source	Records Count	Attributes
CafePress	23205290	Email addresses, Names, Passwords, Phone numbers
CannabisForum	227746	Dates of birth, Email addresses, Geographic locations, Passwords, Usernames
Canva	137272116	Email addresses, Geographic locations, Names, Passwords, Usernames
CapialEconomics	263829	Email addresses, Names, Phone numbers
CardingMafia	297744	Email addresses, Passwords, Usernames
CardingMafiaDec2021	303877	Email addresses, Passwords, Usernames
CashCrate	6844490	Email addresses, Names, Passwords
Catho	1173012	Email addresses, Names, Passwords, Usernames
CDEK	19218203	Email addresses, Names, Phone numbers
CDProjektRed	1871373	Email addresses, Passwords, Usernames
Chatbooks	2520441	Email addresses, Names, Passwords, Phone numbers, Social media profiles
CheapAssGamer	444767	Email addresses, Passwords, Usernames
Chegg	39721127	Email addresses, Names, Passwords, Usernames
Chowbus	444224	Email addresses, Names, Phone numbers
Cit0day	226883414	Email addresses, Passwords
CityBee	110156	Email addresses, Names, Passwords
CivilOnline	7830195	Email addresses, Passwords, Usernames
ClashOfKings	1604957	Email addresses, Passwords, Usernames
ClearVoiceSurveys	15074786	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
ClixSense	2424784	Dates of birth, Email addresses, Genders, Names, Passwords, Usernames
CloudPets	583503	Email addresses, Passwords
ClubPenguinRewritten	1688176	Email addresses, Passwords, Usernames
ClubPenguinRewrittenJul2017	2007909	Email addresses, Passwords, Usernames
Coachella	599802	Email addresses, Passwords, Usernames
Coinmama	478824	Email addresses, Passwords, Usernames
CoinMarketCap	3117548	Email addresses
CoinTracker	1557153	Email addresses
Collection1	772904991	Email addresses, Passwords
Comcast	616882	Email addresses, Passwords
COMELEC	228605	Dates of birth, Email addresses, Genders, Names, Phone numbers
Convex	150129	Email addresses, Names, Phone numbers
CouponMomAndArmorGames	1019525	Email addresses, Passwords
CrackCommunity	19210	Email addresses, Passwords, Usernames
CrackedTO	749161	Email addresses, Passwords, Usernames
CrackingForum	660305	Email addresses, Passwords, Usernames
Creative	483015	Email addresses, Passwords, Usernames
CrimeAgencyVBulletin	942044	Email addresses, Passwords, Usernames
CrossFire	12865609	Email addresses, Passwords, Usernames
CTARS	12314	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
CyberServe	1107034	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers, Usernames
D3scene	568827	Email addresses, Passwords, Usernames
DaFont	637340	Email addresses, Passwords, Usernames
Dailymotion	85176234	Email addresses, Passwords, Usernames
DailyObjects	464260	Email addresses, Names, Passwords, Phone numbers
DailyQuiz	8032404	Email addresses, Passwords, Usernames
Dangdang	4848734	Email addresses

Table 3, continued

Source	Records Count	Attributes
DaniWeb	1131636	Email addresses, Passwords
DataAndLeads	44320330	Email addresses, Names, Phone numbers
DataCamp	760561	Email addresses, Geographic locations, Names, Passwords
DataEnrichment	8176132	Dates of birth, Email addresses, Names, Phone numbers
DatPiff	7476940	Email addresses, Passwords, Usernames
Dave	2964182	Dates of birth, Email addresses, Names, Passwords, Phone numbers
db8151dd	22802117	Email addresses, Names, Phone numbers, Social media profiles
DDO	1580933	Dates of birth, Email addresses, Passwords, Usernames
DecoratingTheHouse	1298651	Email addresses, Names, Phone numbers, Usernames
Deezer	229037936	Dates of birth, Email addresses, Genders, Geographic locations, Names, Usernames
DemonForums	52623	Email addresses, Passwords, Usernames
Descomplica	4845378	Email addresses, Names, Passwords
DevilTorrents	63451	Email addresses, Passwords
devkitPro	1508	Email addresses, Passwords
DietCom	1383759	Dates of birth, Email addresses, Names, Passwords, Usernames
Digimon	7687679	Email addresses, Names
Disqus	17551044	Email addresses, Passwords, Usernames
DivXSubTitles	783058	Email addresses, Passwords, Usernames
DLH	3264710	Dates of birth, Email addresses, Names, Passwords, Usernames
Dodonew	8718404	Email addresses, Usernames
Dominos	648231	Email addresses, Names, Passwords, Phone numbers
DominosIndia	22527655	Email addresses, Names, Phone numbers
Doomworld	34478	Email addresses, Passwords, Usernames
DoorDash	367476	Email addresses, Geographic locations, Names
Doxbin	370794	Email addresses, Passwords, Usernames
DriveSure	3675099	Email addresses, Names, Passwords, Phone numbers
Drizly	2479044	Dates of birth, Email addresses, Names, Passwords, Phone numbers
Dropbox	68648009	Email addresses, Passwords
Dubsmash	161749950	Email addresses, Geographic locations, Names, Passwords, Phone numbers, Usernames
DucksUnlimited	1324364	Dates of birth, Email addresses, Names, Passwords, Phone numbers
DuelingNetwork	6486626	Email addresses, Passwords, Usernames
Dunzo	3465259	Email addresses, Geographic locations, Names, Phone numbers
Duowan	2639894	Email addresses, Passwords, Usernames
DVDShopCH	67973	Email addresses, Passwords
Eatigo	2789609	Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles
EatStreet	6353564	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles
Edmodo	43423561	Email addresses, Passwords, Usernames
Elance	1291178	Email addresses, Geographic locations, Passwords, Phone numbers, Usernames
Elanic	2325283	Email addresses, Geographic locations, Usernames
ElasticsearchSalesLeads	5788169	Email addresses, Names
Emotet	4324770	Email addresses, Passwords
Emuparadise	1131229	Email addresses, Passwords, Usernames

Table 3, continued

Source	Records Count	Attributes
EPal	108887	Email addresses, Usernames
EpicBot	816662	Email addresses, Passwords, Usernames
EpicGames	251661	Email addresses, Passwords, Usernames
EpicNPC	408795	Email addresses, Passwords, Usernames
Epik	15003961	Email addresses, Names, Phone numbers
Eroticy	1370175	Email addresses, Names, Passwords, Phone numbers, Usernames
Eskimi	1197620	Dates of birth, Email addresses, Genders, Geographic locations, Passwords, Usernames
Estonia	655161	Email addresses, Passwords
eThekwiniMunicipality	81830	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
Ethereum	16431	Email addresses, Passwords, Usernames
EuropaJobs	226095	Dates of birth, Email addresses, Geographic locations, Names, Passwords, Phone numbers
Evermotion	435510	Dates of birth, Email addresses, Passwords, Usernames
EverybodyEdits	871190	Email addresses, Usernames
Evite	100985047	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
Evony	29396116	Email addresses, Passwords, Usernames
Exactis	131577763	Dates of birth, Email addresses, Genders, Names, Phone numbers
Experian2020	1284637	Email addresses, Names, Phone numbers
Experian	7196890	Dates of birth, Email addresses, Genders, Names, Phone numbers
ExploitIn	593427119	Email addresses, Passwords
Eye4Fraud	16000591	Email addresses, Names, Passwords, Phone numbers
EyeEm	19611022	Email addresses, Names, Passwords, Usernames
Facebook	509458528	Dates of birth, Email addresses, Genders, Geographic locations, Names, Phone numbers
Facepunch	342913	Dates of birth, Email addresses, Passwords, Usernames
FaceUP	87633	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
Factual	2461696	Email addresses, Phone numbers
Famm	535240	Dates of birth, Email addresses, Genders, Names, Passwords
Fanpass	112251	Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles
FantasyFootballHub	66479	Email addresses, Names, Passwords, Usernames
FashionFantasyGame	2357872	Email addresses, Passwords
FFShrine	620677	Email addresses, Passwords, Usernames
FilmaiIn	645786	Email addresses, Passwords, Usernames
Flashback	40256	Email addresses
FlashFlashRevolution	1771845	Email addresses, Passwords, Usernames
FlashFlashRevolution2019	1858124	Dates of birth, Email addresses, Passwords, Usernames
FlexBooker	3756794	Email addresses, Names, Passwords, Phone numbers
Fling	40767652	Dates of birth, Email addresses, Genders, Geographic locations, Passwords, Phone numbers, Usernames
FLVS	542902	Dates of birth, Email addresses, Names, Passwords, Usernames
Foodora	582578	Email addresses, Names, Passwords, Phone numbers
Forbes	1057819	Email addresses, Passwords, Usernames
ForumCommunity	776648	Email addresses, Passwords, Usernames
Fotolog	16717854	Email addresses, Passwords, Usernames

Table 3, continued

Source	Records Count	Attributes
FoxyBingo	252216	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
FreedomHostingII	380830	Email addresses, Passwords, Usernames
FreshMenu	110355	Email addresses, Names, Phone numbers
Fridae	35368	Email addresses, Passwords, Usernames
Funimation	2491103	Dates of birth, Email addresses, Passwords, Usernames
FunnyGames	764357	Email addresses, Passwords, Usernames
FurAffinity	1270564	Email addresses, Passwords, Usernames
Gaadi	4261179	Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers, Usernames
Gab	66521	Email addresses, Names, Passwords, Usernames
GamerzPlanet	1217166	Email addresses, Passwords, Usernames
GameSalad	1506242	Email addresses, Passwords, Usernames
GameTuts	2064274	Email addresses, Passwords, Usernames
Gamigo	8243604	Email addresses, Passwords
GateHub	1408078	Email addresses, Passwords
Gawker	1247574	Email addresses, Passwords, Usernames
GeekedIn	1073164	Email addresses, Geographic locations, Names, Usernames
Gemini	5274214	Email addresses
GeniusU	1301460	Email addresses, Genders, Names, Passwords, Social media profiles
GetRevengeOnYourEx	79195	Email addresses, Names, Passwords, Phone numbers
Gett	2481121	Email addresses, Names, Passwords, Social media profiles
GFAN	22526334	Email addresses, Passwords, Usernames
GGCorp	2376330	Email addresses, Passwords, Usernames
GiveSendGo	89966	Email addresses, Geographic locations, Names
Glofox	2330735	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
GoGames	3430083	Email addresses, Passwords, Usernames
GoldSilver	242715	Email addresses, Names, Phone numbers
gPotato	2136520	Dates of birth, Email addresses, Genders, Names, Passwords, Usernames
GPSUnderground	669584	Dates of birth, Email addresses, Passwords, Usernames
Gravatar	113990759	Email addresses, Names, Usernames
GTAGaming	197184	Dates of birth, Email addresses, Passwords, Usernames
GunAuction	565470	Email addresses, Genders, Passwords, Phone numbers, Usernames
GunsDotCom	375928	Dates of birth, Email addresses, Names, Passwords, Phone numbers
Guntrader	112031	Email addresses, Geographic locations, Names, Passwords, Phone numbers
HackForums	191540	Dates of birth, Email addresses, Passwords, Usernames
HackingTeam	32310	Email addresses
HalloweenSpot	10653	Email addresses, Names, Phone numbers
HauteLook	28510459	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords
Havenly	1369180	Email addresses, Geographic locations, Names, Passwords, Phone numbers
HealthNowNetworks	321920	Dates of birth, Email addresses, Genders, Names, Phone numbers
Hemmakvall	47297	Email addresses, Names, Passwords, Phone numbers
Hemmelig	28641	Email addresses, Genders, Passwords, Usernames
HeroesOfGaia	179967	Email addresses, Usernames
HeroesOfNewerth	8089103	Email addresses, Passwords, Usernames

Table 3, continued

Source	Records Count	Attributes
HIAPK	13873674	Email addresses, Passwords, Usernames
HLTV	611070	Email addresses, Names, Passwords, Usernames
HomeChef	8815692	Email addresses, Geographic locations, Names, Passwords, Phone numbers
HongFire	999991	Dates of birth, Email addresses, Passwords, Usernames
HookersNL	290955	Email addresses, Passwords, Usernames
HoundDawgs	45701	Email addresses, Passwords
Houzz	48881308	Email addresses, Geographic locations, Names, Passwords, Social media profiles, Usernames
HTCMania	1488089	Dates of birth, Email addresses, Passwords, Usernames
HTHStudios	411755	Dates of birth, Email addresses, Names, Phone numbers, Usernames
Hub4Tech	36916	Email addresses, Passwords
Hurb	20727771	Dates of birth, Email addresses, Names, Passwords, Phone numbers, Social media profiles
IDCGames	3966871	Email addresses, Passwords, Usernames
iDressup	2191565	Email addresses, Passwords
iDTech	415121	Dates of birth, Email addresses, Names, Passwords
IGF	3200	Email addresses, Names, Passwords, Usernames
IIMJobs	4216063	Dates of birth, Email addresses, Geographic locations, Names, Passwords, Phone numbers
ILikeCheats	188847	Email addresses, Passwords, Usernames
Imavex	878209	Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
iMesh	49467477	Email addresses, Passwords, Usernames
imgur	1749806	Email addresses, Passwords
IndiaMART	20154583	Email addresses, Names, Phone numbers
IndianRailways	583377	Email addresses, Passwords, Usernames
Insanelyi	104097	Email addresses, Passwords, Usernames
InstantCheckmate	11943887	Email addresses, Names, Passwords, Phone numbers
Intelimost	3073409	Email addresses, Passwords
Interpals	3439414	Dates of birth, Email addresses, Geographic locations, Names, Passwords, Usernames
iPmart	2460787	Dates of birth, Email addresses, Passwords, Usernames
ixigo	17204697	Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles, Usernames
James	1541284	Email addresses, Geographic locations, Passwords
JD	77449341	Email addresses, Passwords, Phone numbers, Usernames
Jefit	9052457	Email addresses, Passwords, Usernames
JobAndTalent	10981207	Email addresses, Names, Passwords
JobStreet	3883455	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers, Usernames
JoomlArt	22477	Email addresses, Names, Passwords, Usernames
JukinMedia	314290	Email addresses, Names, Passwords, Phone numbers
JustDate	24451312	Dates of birth, Email addresses, Geographic locations, Names
KayoMoe	41826763	Email addresses, Passwords
Kickstarter	5176463	Email addresses, Passwords
Kimsufi	504565	Email addresses, Passwords, Usernames
KiwiFarms	4606	Dates of birth, Email addresses
KMRU	1476783	Dates of birth, Email addresses, Genders, Geographic locations, Usernames
KnownCircle	1957600	Email addresses, Genders, Names, Passwords, Phone numbers

Table 3, continued

Source	Records Count	Attributes
Knuddels	808330	Email addresses, Geographic locations, Names, Passwords, Usernames
KomplettFritid	139401	Email addresses, Names, Passwords, Phone numbers
Kreditplus	768890	Dates of birth, Email addresses, Genders, Names, Phone numbers
Lanwar	45120	Email addresses, Names, Passwords, Usernames
LaPosteMobile	533886	Dates of birth, Email addresses, Genders, Names, Phone numbers
Lastfm	37217682	Email addresses, Passwords, Usernames
Lazada	1107789	Email addresses, Names, Passwords, Phone numbers
LBB	39288	Email addresses, Names
LeadHunter	68693853	Email addresses, Genders, Names, Phone numbers
LeagueOfLegends	339487	Email addresses, Passwords, Usernames
Ledger	1075241	Email addresses, Names, Phone numbers
Leet	5081689	Email addresses, Passwords, Usernames
Lifebear	3670561	Email addresses, Passwords, Usernames
Lifeboat	7089395	Email addresses, Passwords, Usernames
LightsHope	30484	Dates of birth, Email addresses, Geographic locations, Passwords, Usernames
Liker	465141	Dates of birth, Email addresses, Geographic locations, Names, Passwords, Phone numbers, Social media profiles, Usernames
LimeVPN	23348	Email addresses, Names, Passwords, Phone numbers
LinkedIn	164611595	Email addresses, Passwords
LinkedInScrape	125698496	Email addresses, Genders, Geographic locations, Names, Social media profiles
LinuxForums	275785	Email addresses, Passwords, Usernames
LinuxMint	144989	Dates of birth, Email addresses, Geographic locations, Passwords
LittleMonsters	995698	Dates of birth, Email addresses, Passwords, Usernames
LiveAuctioneers	3385862	Email addresses, Names, Passwords, Phone numbers, Usernames
LiveJournal	26372781	Email addresses, Passwords, Usernames
Livpure	269552	Email addresses, Names, Phone numbers
LizardSquad	13451	Email addresses, Passwords, Usernames
Lolzteam	398011	Email addresses, Usernames
Lookbook	1074948	Dates of birth, Email addresses, Names, Passwords, Usernames
LOTR	1141278	Dates of birth, Email addresses, Passwords, Usernames
LoungeBoard	45018	Email addresses, Names, Passwords, Usernames
LuminPDF	15453048	Email addresses, Genders, Names, Passwords, Usernames
LyricsMania	109202	Email addresses, Passwords, Usernames
MacForums	326714	Dates of birth, Email addresses, Passwords, Usernames
MacGeneration	101004	Email addresses, Passwords, Usernames
Mac-Torrents	93992	Email addresses, Passwords, Usernames
MailRu	16630988	Email addresses, Passwords
MajorGeeks	269548	Email addresses, Passwords, Usernames
MallCZ	735405	Email addresses, Names, Passwords, Phone numbers
Malwarebytes	111623	Dates of birth, Email addresses, Passwords, Usernames
MangaDex	2987329	Email addresses, Passwords, Usernames
MangaFox	1311610	Dates of birth, Email addresses, Passwords, Usernames
Mangatoon	23040238	Email addresses, Genders, Names, Passwords, Social media profiles, Usernames
MangaTraders	855249	Email addresses, Passwords

Table 3, continued

Source	Records Count	Attributes
Mappery	205242	Email addresses, Geographic locations, Passwords, Usernames
Mashable	1414677	Email addresses, Genders, Geographic locations, Names, Social media profiles
MastercardPricelessSpecials	89388	Email addresses, Names, Phone numbers
MasterDeeds	2257930	Dates of birth, Email addresses, Genders, Names, Phone numbers
Mate1	27393015	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Usernames
Mathway	25692862	Email addresses, Names, Passwords, Social media profiles
MCBans	119948	Email addresses, Passwords, Usernames
MDPI	845012	Email addresses, Names
MechoDownload	437928	Email addresses, Passwords, Usernames
MeetMindful	1422717	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Social media profiles, Usernames
MGM2022Update	24842001	Dates of birth, Email addresses, Names, Phone numbers
MGM	3081321	Dates of birth, Email addresses, Names, Phone numbers
MindJolt	28364826	Dates of birth, Email addresses, Names
MinecraftPocketEditionForum	16034	Email addresses, Passwords, Usernames
MinecraftWorldMap	71081	Email addresses, Passwords, Usernames
Minefield	188343	Dates of birth, Email addresses, Passwords, Usernames
Minehut	396533	Email addresses, Passwords
Minted	4418182	Email addresses, Names, Passwords, Phone numbers
MMGFusion	2660295	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
MobiFriends	3512952	Dates of birth, Email addresses, Genders, Passwords, Usernames
MoDaCo	879703	Email addresses, Passwords, Usernames
ModernBusinessSolutions	58843488	Dates of birth, Email addresses, Genders, Names, Phone numbers
MoneyBookers	4483605	Dates of birth, Email addresses, Names, Phone numbers
Moneycontrol	762874	Email addresses, Genders, Geographic locations, Passwords, Phone numbers
MoreleNet	2467304	Email addresses, Names, Passwords, Phone numbers
MortalOnline	606637	Email addresses, Names, Passwords, Usernames
MPGH	3122898	Email addresses, Passwords, Usernames
MrExcel	366140	Dates of birth, Email addresses, Passwords, Usernames
mSpy	699793	
MuslimDirectory	37784	Email addresses, Names, Passwords, Phone numbers
MuslimMatch	149830	Email addresses, Geographic locations, Passwords, Usernames
MyFHA	972629	Email addresses, Names, Passwords
MyFitnessPal	143606147	Email addresses, Passwords, Usernames
MyHeritage	91991358	Email addresses, Passwords
myRepoSpace	252751	Email addresses, Passwords, Usernames
MySpace	359420698	Email addresses, Passwords, Usernames
MyVidster	19863	Email addresses, Passwords, Usernames
NamelessMalware	1121484	Email addresses
NapsGear	287071	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
NaughtyAmerica	1398630	Dates of birth, Email addresses, Passwords, Usernames
NemoWeb	3472916	Email addresses, Names
Neopets	26892897	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Usernames

Table 3, continued

Source	Records Count	Attributes
NetEase	234842089	Email addresses, Passwords
Neteller	3619948	Dates of birth, Email addresses, Genders, Names, Phone numbers
NetGalley	1436435	Dates of birth, Email addresses, Names, Passwords, Phone numbers, Usernames
Netlog	49038354	Email addresses, Passwords
NetProspex	33698126	Email addresses, Names, Phone numbers
Netshoes	499836	Dates of birth, Email addresses, Names
NextGenUpdate	1194597	Email addresses, Passwords, Usernames
NexusMods	5915013	Email addresses, Passwords, Usernames
Nihonomaru	1697282	Email addresses, Passwords, Usernames
Nitro	77159696	Email addresses, Names, Passwords
Nival	1535473	Dates of birth, Email addresses, Genders, Names, Usernames
NonNudeGirls	75383	Email addresses, Names, Passwords, Usernames
NotAcxiom	51730831	Email addresses, Names, Phone numbers
Nulled	599080	Dates of birth, Email addresses, Passwords, Usernames
NulledCH	43491	Email addresses, Passwords, Usernames
NurseryCam	10585	Email addresses
NVIDIA	71335	Email addresses, Passwords
OGUsers	161143	Email addresses, Passwords, Usernames
OGUsers2020	263189	Email addresses, Passwords, Usernames
OGUsers2021	348302	Email addresses, Passwords, Usernames
OnlinerSpambot	711477622	Email addresses, Passwords
Onverse	800157	Email addresses, Passwords, Usernames
OpenCSGO	512311	Email addresses, Phone numbers, Social media profiles, Usernames
OpenSubtitles	6783158	Email addresses, Geographic locations, Passwords, Usernames
OrderSnapp	1304447	Dates of birth, Email addresses, Names, Passwords, Phone numbers
OrdineAvvocatiDiRoma	41960	Email addresses, Geographic locations, Passwords, Phone numbers
OVH	452899	Email addresses, Passwords, Usernames
OwnedCore	880331	Email addresses, Passwords, Usernames
Oxfam	1834006	Dates of birth, Email addresses, Genders, Names, Phone numbers
PaddyPower	590954	Dates of birth, Email addresses, Names, Phone numbers, Usernames
ParagonCheats	188089	Email addresses, Usernames
Parapa	4946850	Email addresses, Passwords, Usernames
ParkMobile	20949825	Email addresses, Names, Passwords, Phone numbers
Patreon	2330382	Email addresses, Passwords
PayAsUGym	400260	Email addresses, Names, Passwords, Phone numbers
PayHere	1580249	Email addresses, Names, Phone numbers
Paytm	3395101	Dates of birth, Email addresses, Genders, Geographic locations, Names, Phone numbers
PDL	622161052	Email addresses, Geographic locations, Names, Phone numbers, Social media profiles
Peatix	4227907	Email addresses, Names, Passwords
Pemiblanc	110964206	Email addresses, Passwords
PeoplesEnergy	358822	Dates of birth, Email addresses, Names, Passwords, Phone numbers
PetFlow	990919	Email addresses, Passwords

Table 3, continued

Source	Records Count	Attributes
PhoneHouse	5223350	Dates of birth, Email addresses, Genders, Names, Phone numbers
PHPFreaks	173891	Dates of birth, Email addresses, Passwords, Usernames
PixelFederation	38108	Email addresses, Passwords
Pixlr	1906808	Email addresses, Geographic locations, Names, Passwords, Social media profiles
piZap	41817893	Email addresses, Genders, Geographic locations, Names, Passwords, Social media profiles, Usernames
PlanetCalypso	62261	Email addresses, Passwords, Usernames
PlanetIce	240488	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
Playbook	50538	Email addresses, Names, Passwords, Phone numbers, Social media profiles
Playgar	143569	Email addresses, Passwords, Usernames
Plex	327314	Email addresses, Passwords, Usernames
PlutoTV	3225080	Dates of birth, Email addresses, Genders, Names, Passwords, Social media profiles, Usernames
Pokebip	657001	Email addresses, Passwords, Usernames
PokemonCreed	116465	Email addresses, Genders, Passwords, Usernames
PokemonNegro	830155	Email addresses, Passwords
PoliceOne	709926	Email addresses, Passwords, Usernames
Poshmark	36395491	Email addresses, Genders, Geographic locations, Names, Passwords, Usernames
Powerbot	503501	Email addresses, Passwords, Usernames
PPCGeeks	492518	Dates of birth, Email addresses, Passwords, Usernames
PreenMe	236105	Email addresses, Names, Social media profiles, Usernames
ProctorU	444453	Email addresses, Names, Passwords, Phone numbers, Usernames
ProgrammingForums	707432	Email addresses, Passwords, Usernames
Promo	14610585	Email addresses, Genders, Names, Passwords
Promofarma	1277761	Email addresses, Names
PropTiger	2156921	Dates of birth, Email addresses, Genders, Names, Passwords
Protemps	49591	Email addresses, Genders, Names, Passwords, Phone numbers
PS3Hax	447410	Email addresses, Passwords, Usernames
PSPISO	1274070	Email addresses, Passwords, Usernames
PSX-Scene	341118	Email addresses, Passwords, Usernames
QatarNationalBank	88678	Dates of birth, Genders, Geographic locations, Names, Passwords, Phone numbers
QIP	26183992	Email addresses, Passwords, Usernames
QuantumBooter	48592	Email addresses, Passwords, Usernames
QuestionPro	22229637	Email addresses
Quidd	3805863	Email addresses, Passwords, Usernames
QuinStreet	4907802	Dates of birth, Email addresses, Passwords, Usernames
R2-2017	1023466	Email addresses, Passwords, Usernames
R2Games	22281337	Email addresses, Passwords, Usernames
Rambler	91436280	Email addresses, Passwords, Usernames
Rankwatch	7445067	Email addresses, Names, Phone numbers
Raychat	938981	Email addresses, Names, Passwords
RbxRocks	149958	Email addresses, Names, Passwords
ReadNovel	22424472	Email addresses, Genders, Passwords, Phone numbers, Usernames
RealDudesInc	101543	Email addresses, Passwords, Usernames

Table 3, continued

Source	Records Count	Attributes
RealEstateMogul	307768	Email addresses, Names, Phone numbers
RedDoorz	5890277	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
RedLineStealer	441657	Email addresses, Passwords, Usernames
Reincubate	616146	Email addresses, Names, Passwords
RepublicanPartyOfTexas	72596	Email addresses, Geographic locations, Names
RetinaX	71153	Email addresses, Passwords
Reverb-Nation	7040725	Email addresses, Passwords
RiverCityMedia	393430309	Email addresses, Names
Robinhood	5003937	Email addresses
Roll20	3994436	Email addresses, Names, Passwords
Romwe	19531820	Geographic locations, Names, Passwords, Phone numbers
RosebuttBoard	107303	Email addresses, Passwords, Usernames
RoyalEnfield	420873	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles
RussianAmerica	182717	Email addresses, Names, Passwords, Phone numbers
SaverSpy	2457420	Email addresses, Genders, Names
SCDailyPhoneSpamList	32939105	Dates of birth, Email addresses, Genders, Names
Scentbird	5814988	Dates of birth, Email addresses, Genders, Names, Passwords
SchoolDistrict42	18850	Email addresses, Names
Seedpeer	281924	Email addresses, Passwords, Usernames
Sephora	780073	Dates of birth, Email addresses, Genders, Names
ServerPact	73587	Dates of birth, Email addresses, Passwords, Usernames
Shadi	2021984	Email addresses, Passwords
ShareThis	40960499	Dates of birth, Email addresses, Names, Passwords
SHEIN	39086762	Email addresses, Passwords
Shitexpress	23817	Email addresses, Names
ShockGore	73944	Email addresses, Genders, Passwords, Usernames
ShopBack	20529819	Email addresses, Geographic locations, Names, Passwords, Phone numbers
ShortEdition	505466	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Social media profiles, Usernames
Shotbow	1052753	Email addresses, Passwords, Usernames
SIAE	14609	Email addresses, Names, Passwords, Phone numbers
SirHurt	90655	Email addresses, Passwords, Usernames
SitePoint	1021790	Email addresses, Names, Passwords, Usernames
SkTorrent	117070	Email addresses, Passwords, Usernames
Slickwraps	857611	Email addresses, Names, Phone numbers
SlideTeam	1464271	Email addresses, Names, Passwords
Smogon	386489	Email addresses, Genders, Geographic locations, Passwords, Usernames
Snail	1410899	Email addresses, Passwords, Usernames
Snapchat	4609615	Geographic locations, Phone numbers, Usernames
SocialEngineered	89392	Email addresses, Passwords, Usernames
Solomid	442166	Email addresses, Passwords, Usernames
Sonicbids	751700	Email addresses, Names, Passwords, Usernames
Sony	37103	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
Soundwave	130705	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords
SpecialKSpamList	30741620	Dates of birth, Email addresses, Genders, Names
Spirol	55622	Email addresses, Names, Phone numbers

Table 3, continued

Source	Records Count	Attributes
SprashivaiRu	3474763	Dates of birth, Email addresses, Genders, Geographic locations, Passwords
SpyFone	44109	Email addresses, Geographic locations, Names, Passwords
Staminus	26815	Email addresses, Passwords, Usernames
StarNet	139395	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
Start	7455386	Email addresses, Geographic locations, Names, Passwords
StarTribune	2192857	Dates of birth, Email addresses, Genders, Names, Passwords, Usernames
SterKinekor	1619544	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
StockX	6840339	Email addresses, Names, Passwords, Usernames
StoryBird	1047200	Email addresses, Names, Passwords, Usernames
Straffic	48580249	Email addresses, Genders, Names, Phone numbers
Stratfor	859777	Email addresses, Names, Passwords, Phone numbers, Usernames
StreetEasy	988230	Email addresses, Names, Passwords, Usernames
Stripchat	10001355	Email addresses, Usernames
StrongholdKingdoms	5187305	Email addresses, Passwords, Usernames
SubaGames	6137666	Email addresses, Passwords, Usernames
SumoTorrent	285191	Email addresses, Passwords, Usernames
SuperVPNGeckoVPN	20339937	Email addresses, Geographic locations
SvenskaMagic	30327	Email addresses, Passwords, Usernames
SweClockers	254867	Email addresses, Passwords, Usernames
Swvl	4195918	Email addresses, Names, Passwords, Phone numbers
TaiLieu	7327477	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers, Usernames
Tamodo	494945	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords
Taobao	21149008	Email addresses, Passwords
TAPAirPortugal	6083479	Dates of birth, Email addresses, Genders, Names, Phone numbers
Taringa	27971100	Email addresses, Passwords, Usernames
Technic	265410	Email addresses, Passwords
Teespring	8234193	Email addresses, Geographic locations, Names, Social media profiles
Teracod	97151	Email addresses, Passwords, Usernames
Tesco	2239	Email addresses, Passwords
TGBUS	10371766	Email addresses, Passwords, Usernames
TheCandidBoard	178201	Dates of birth, Email addresses, Geographic locations, Passwords, Usernames
TheFappening	179030	Email addresses, Passwords, Usernames
TheFlyOnTheWall	84011	Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
TheTVDB	181871	Email addresses, Passwords, Usernames
Thingiverse	228102	Dates of birth, Email addresses, Names, Passwords, Usernames
ThisHabboForum	612414	Email addresses, Passwords, Usernames
Tianya	29020808	Email addresses, Names, Usernames
Ticketcounter	1921722	Dates of birth, Email addresses, Genders, Names, Phone numbers
Ticketfly	26151608	Email addresses, Names, Phone numbers
Tokopedia	71443698	Dates of birth, Email addresses, Genders, Names, Passwords

Table 3, continued

Source	Records Count	Attributes
ToonDoo	6002694	Email addresses, Genders, Geographic locations, Passwords, Usernames
TorrentInvites	352120	Dates of birth, Email addresses, Passwords, Usernames
Tout	652683	Email addresses, Geographic locations, Names, Passwords, Usernames
TRAI	107776	Email addresses
Travelio	471376	Dates of birth, Email addresses, Names, Passwords, Phone numbers
TravelOK	637279	Dates of birth, Email addresses, Genders, Names
TrikSpamBotnet	43432346	Email addresses
Trillian	3827238	Dates of birth, Email addresses, Names, Passwords, Usernames
TruckersMP	83957	Email addresses, Passwords, Usernames
TrueFire	599667	Dates of birth, Email addresses, Names, Passwords, Phone numbers, Usernames
TruthFinder	8159573	Email addresses, Names, Passwords, Phone numbers
Tumblr	65469298	Email addresses, Passwords
TunedGlobal	985586	Email addresses, Names, Passwords, Phone numbers
Twitter200M	211524284	Email addresses, Names, Social media profiles, Usernames
Twitter	6682453	Email addresses, Geographic locations, Names, Phone numbers, Usernames
UC	547422	Dates of birth, Email addresses, Genders, Names, Phone numbers
Uiggy	2682650	Email addresses, Genders, Names
Ulmon	777769	Email addresses, Names, Passwords, Phone numbers, Social media profiles
UnderworldEmpire	428779	Email addresses, Passwords, Usernames
UnicoCampania	166031	Email addresses, Passwords
Universarium	564962	Email addresses, Passwords
UnrealEngine	530147	Email addresses, Passwords, Usernames
Upstox	111002	Dates of birth, Email addresses, Genders, Passwords, Phone numbers
UtahGunExchange	235233	Email addresses, Genders, Passwords, Usernames
uTorrent	395044	Email addresses, Passwords, Usernames
uuu9	7485802	Email addresses, Passwords, Usernames
Vakinha	4775203	Dates of birth, Email addresses, Names, Passwords, Phone numbers
Vastaamo	30433	Email addresses, Names
VBulletin	518966	Dates of birth, Email addresses, Passwords
Vedantu	686899	Email addresses, Genders, Names, Passwords, Phone numbers
VerificationsIO	763117241	Dates of birth, Email addresses, Genders, Geographic locations, Names, Phone numbers
Verified	16919	Email addresses, Passwords, Usernames
Vianet	94353	Email addresses, Names, Phone numbers
VictoryPhones	166046	Dates of birth, Email addresses, Names, Phone numbers
ViewFines	777649	Email addresses, Names, Passwords, Phone numbers
VINs	396650	Dates of birth, Email addresses, Genders, Names, Phone numbers
VK	93338602	Email addresses, Names, Passwords, Phone numbers
VNG	24853850	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers, Usernames
Vodafone	56021	Email addresses, Names, Passwords, Phone numbers, Usernames

Table 3, continued

Source	Records Count	Attributes
VoidTO	95431	Email addresses, Passwords, Usernames
VTech	4833678	Dates of birth, Email addresses, Genders, Names, Passwords, Usernames
VTightGel	2013164	Email addresses, Names, Phone numbers
Wakanim	6706951	Email addresses, Names, Usernames
Wanelo	23165793	Email addresses, Names, Passwords
Warframe	819478	Email addresses, Usernames
WarInc	1020136	Email addresses, Passwords, Usernames
Warmane	1116256	Dates of birth, Email addresses, Passwords, Usernames
Wattpad	268765495	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Social media profiles, Usernames
WedMeGood	1306723	Email addresses, Genders, Names, Passwords, Phone numbers
Weee	1117405	Email addresses, Names, Phone numbers
WeHeartIt	8600635	Email addresses, Passwords, Usernames
WeLeakInfo	11788	Email addresses, Names
Wendys	52485	Email addresses, Names, Passwords, Phone numbers
Whitepages	11657763	Email addresses, Names, Passwords
WhiteRoom	1279263	Dates of birth, Email addresses, Genders, Names, Passwords, Phone numbers
WHMCS	134047	Email addresses, Names, Passwords
WienerBuchereien	224119	Dates of birth, Email addresses, Names, Phone numbers
WifeLovers	1274051	Email addresses, Names, Passwords, Usernames
WIIUISO	458155	Email addresses, Passwords, Usernames
WildStar	738556	Dates of birth, Email addresses, Passwords, Usernames
Win7Vista	202683	Email addresses, Names, Passwords, Usernames
Wishbone2020	9705172	Dates of birth, Email addresses, Genders, Geographic locations, Names, Passwords, Phone numbers, Social media profiles, Usernames
Wishbone	2247314	Dates of birth, Email addresses, Genders, Names, Phone numbers, Usernames
WiziShop	2856769	Dates of birth, Email addresses, Names, Passwords, Phone numbers
Wongnai	3924454	Dates of birth, Email addresses, Geographic locations, Names, Passwords, Phone numbers, Social media profiles
WPSandbox	858	Email addresses, Passwords
WPT	148366	Email addresses, Passwords
xat	5968783	Email addresses, Passwords, Usernames
Xbox360ISO	1296959	Email addresses, Passwords, Usernames
Xbox-Scene	432552	Email addresses, Passwords, Usernames
xHamster	377377	Email addresses, Passwords, Usernames
Xiaomi	7088010	Email addresses, Passwords, Usernames
XKCD	561991	Email addresses, Passwords, Usernames
XPGameSaves	890341	Email addresses, Passwords, Usernames
XSplit	2983472	Email addresses, Names, Passwords, Usernames
Yahoo	453427	Email addresses, Passwords
Yam	13258797	Dates of birth, Email addresses, Names, Passwords, Phone numbers, Usernames
Yandex	1186564	Email addresses, Passwords
Yatra	5033997	Dates of birth, Email addresses, Names, Passwords, Phone numbers
YoteprestoCom	1444629	Email addresses, Passwords, Usernames
Youku	91890110	Email addresses, Passwords
YouNow	18241518	Email addresses, Names, Social media profiles, Usernames

Table 3, continued

Source	Records Count	Attributes
YouPorn	1327567	Email addresses, Passwords
YouveBeenScraped	66147869	Email addresses, Geographic locations, Names, Social media profiles
ZAPHosting	746682	Email addresses, Names, Phone numbers
Zhenai	5024908	Email addresses, Passwords
Zomato	16472873	Email addresses, Passwords, Usernames
Zoomcar	3589795	Email addresses, Names, Passwords, Phone numbers
Zoosk2020	23927853	Dates of birth, Email addresses, Genders, Geographic locations, Names
Zoosk	52578183	Email addresses, Passwords
Zooville	71407	Dates of birth, Email addresses, Passwords, Usernames
Zurich	756737	Dates of birth, Email addresses, Genders, Names
Zynga	172869660	Email addresses, Passwords, Phone numbers, Usernames

[**?**]