Releasing a Differentially Private Password Frequency Corpus from 70 Million Yahoo! Passwords

Jeremiah Blocki

Purdue University

DIMACS/Northeast Big Data Hub Workshop on Overcoming Barriers to Data Sharing including Privacy and Fairness
What is a Password Frequency List?

Password Dataset: (N users)

12345
password
abc123
abc123

Histogram
What is a Password Frequency List?

Password Dataset: (N users)

12345
password
abc123
abc123

Histogram

Frequency List
What is a Password Frequency List?

Password Dataset: (N users)

Formally:
\[ f \in \mathcal{P}(N) \]

Password Frequency List is just an integer partition.
Password Frequency List (Example Use)

Estimate #accounts compromised by attacker with $\beta$ guesses per user

- Online Attacker ($\beta$ small)
- Offline Attacker ($\beta$ large)

$$\lambda_\beta = \sum_{i=1}^{\beta} f_i$$

Password Frequency Lists allow us to estimate

- Marginal Guessing Cost (MGC)
- Marginal Benefit (MB)
- Rational Adversary: MGC = MB
Available Password Frequency Lists (2015)

<table>
<thead>
<tr>
<th>Site</th>
<th>#User Accounts (N)</th>
<th>How Released</th>
</tr>
</thead>
<tbody>
<tr>
<td>RockYou</td>
<td>32.6 Million</td>
<td>Data Breach*</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>6</td>
<td>Data Breach*</td>
</tr>
<tr>
<td>....</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

* entire frequency list available due to improper password storage
Yahoo! Password Frequency List

• Collected by Joseph Bonneau in 2011 (with permission from Yahoo!)
  • Store $H(s|\text{pwd})$
  • Secret salt value $s$ (same for all users)
  • Discarded after data-collection
• $\approx 70$ million Yahoo! Users

• Yahoo! Legal gave permission to publish analysis of the frequency list
Would it be possible to access the Yahoo! data? I am working on a cool new research project and the password frequency data would be very useful.
I would love to make the data public, but Yahoo! Legal has concerns about security and privacy. They won’t let me release it.
I would love to make the data public, but Yahoo! Legal has concerns about security and privacy. They won’t let me release it.
Available Password Frequency Lists

<table>
<thead>
<tr>
<th>Site</th>
<th>#User Accounts (N)</th>
<th>How Released</th>
</tr>
</thead>
<tbody>
<tr>
<td>RockYou</td>
<td>32.6 Million</td>
<td>Data Breach*</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>6</td>
<td>Data Breach*</td>
</tr>
<tr>
<td>....</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Yahoo! [B12]</td>
<td>70 Million</td>
<td>With Permission**</td>
</tr>
</tbody>
</table>

* entire frequency list available due to improper password storage
** frequency list perturbed slightly to preserve differential privacy.

Yahoo! Frequency data is now available online at: [https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937](https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937)
Yahoo! Frequency Corpus

Largest publicly available frequency corpus

The New York Times | https://nyti.ms/2xEvrP

Technology

All 3 Billion Yahoo Accounts Were Affected by 2013 Attack

By NICOLE PERLOTH OCT. 3, 2017

It was the biggest known breach of a company's computer network. And now, it is even bigger.

Verizon Communications, which acquired Yahoo this year, said on Tuesday that a previously disclosed attack that had occurred in 2013 affected all three billion of Yahoo’s user accounts.
Why not just publish the original frequency lists?

• Heuristic Approaches to Data Privacy often break down when the adversary has background knowledge
  • Netflix Prize Dataset [NS08]
    • Background Knowledge: IMDB
  • Massachusetts Group Insurance Medical Encounter Database [SS98]
    • Background Knowledge: Voter Registration Record
  • Many other attacks [BDK07,...]

• In the absence of provable privacy guarantees Yahoo! was understandably reluctant to release these password frequency lists.
Security Risks (Example)

Adversary Background Knowledge
Security Risks (Example)
Differential Privacy (Dwork et al)

**Definition:** An (randomized) algorithm $A$ preserves $(\varepsilon, \delta)$-differential privacy if for any subset $S \subseteq \text{Range}(A)$ of possible outcomes and any we have

$$\Pr[A(f) \in S] \leq e^\varepsilon \Pr[A(f') \in S] + \delta$$

for any pair of adjacent password frequency lists $f$ and $f'$,

$$\|f - f'\|_1 = 1.$$

\[ \|f - f'\|_1 \stackrel{\text{def}}{=} \sum_i |f_i - f'_i| \]
Differential Privacy (Dwork et al)

Definition: An (randomized) algorithm $A$ preserves $(\varepsilon, \delta)$-differential privacy if for any subset $S \subseteq Range(A)$ of possible outcomes and any we have

$$\Pr[A(f) \in S] \leq e^\varepsilon \Pr[A(f') \in S] + \delta$$

for any pair of adjacent password frequency lists $f$ and $f'$,

$$\|f - f'\|_1 = 1.$$ 

$f$ – original password frequency list

$f'$ – remove Alice’s password from dataset
Differential Privacy (Dwork et al)

**Definition:** An (randomized) algorithm $A$ preserves $(\varepsilon, \delta)$-differential privacy if for any subset $S \subseteq \text{Range}(A)$ of possible outcomes and any we have

$$\Pr[A(f) \in S] \leq e^\varepsilon \Pr[A(f') \in S] + \delta$$

for any pair of adjacent password frequency lists $f$ and $f'$,

$$\|f - f'\|_1 = 1.$$
Differential Privacy (Dwork et al)

**Definition:** An (randomized) algorithm $A$ preserves $(\varepsilon, \delta)$-differential privacy if for any subset $S \subseteq \text{Range}(A)$ of possible outcomes and any $f$ and $f'$, we have

$$\Pr[A(f) \in S] \leq e^{\varepsilon}\Pr[A(f') \in S] + \delta$$

for any pair of adjacent password frequency lists $f$ and $f'$,

$$\|f - f'\|_1 = 1.$$

Small Constant (e.g., $\varepsilon = 0.5$)

Negligibly Small Value (e.g., $\delta = 2^{-100}$)

$f$ – original password frequency list

$f'$ – remove Alice’s password from dataset
Differential Privacy (Example)

Subset $S$ of all potentially harmful outcomes to Alice

Outcomes
Differential Privacy (Example)

\[
\Pr \left[ A(f) \in \text{HACKED} \right] \leq e^\epsilon \Pr \left[ A(f') \in \text{HACKED} \right] + \delta
\]
Differential Privacy (Example)

**Intuition:** Alice won’t be harmed because her password was included in the dataset.

\[
\Pr \left[ A(f) \in \text{HACKED} \right] \leq e^\varepsilon \Pr \left[ A(f') \in \text{HACKED} \right] + \delta
\]
Main Technical Result

**Theorem:** There is a computationally efficient algorithm $A: \mathcal{F} \times \mathcal{F} \to \mathcal{F}$ such that $A$ preserves $(\varepsilon, \delta)$-differential privacy and, except with probability $\delta$, $A(f)$ outputs $\tilde{f}$ s.t.

$$\|f - \tilde{f}\|_1 \leq O\left(\frac{1}{\varepsilon \sqrt{N}} + \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon N}\right).$$

Time($A$) = $O\left(\frac{N \sqrt{N} + N \ln\left(\frac{1}{\delta}\right)}{\varepsilon}\right)$ = Space($A$)
Main Tool: Exponential Mechanism [MT07]

Input: $f$

Output: $\Pr[\mathcal{E}^{\varepsilon}(f) = \tilde{f}] \propto e^{-\frac{\|f - \tilde{f}\|_1}{2\varepsilon}}$ Assigns very small probability to inaccurate outcomes.
Main Tool: Exponential Mechanism [MT07]

Input: f

Output: \( \Pr[\mathcal{E}^\varepsilon(f) = \tilde{f}] \propto e^{-\frac{\|f - \tilde{f}\|_1}{2\varepsilon}} \)

**Theorem [MT07]:** The exponential mechanism preserves \((\varepsilon, 0)\)-differential privacy.
Analysis: Exponential Mechanism

Input: $f$

Output: $\Pr[\mathcal{E}^\varepsilon(f) = \tilde{f}] \propto e^{-\frac{\|f - \tilde{f}\|_1}{2\varepsilon}}$

Assigns very small probability to inaccurate outcomes.

Theorem [HR18]: There are $e^{O(\sqrt{N})}$ partitions of the integer $N$. 
Analysis: Exponential Mechanism

Input: $f$

Output: $\Pr[\mathcal{E}^\varepsilon(f) = \tilde{f}] \propto e^{-\frac{\|f - \tilde{f}\|_1}{2\varepsilon}}$

Assigns very small probability to inaccurate outcomes.

Theorem [HR18]: There are $e^{O(\sqrt{N})}$ partitions of the integer $N$.

Union Bound $\Rightarrow \|f - \tilde{f}\|_1 \leq O\left(\frac{\sqrt{N}}{\varepsilon}\right)$ with high probability when $\frac{1}{\varepsilon} = O(\sqrt{N})$. 
Analysis: Exponential Mechanism

Input: $f$

Output: $\Pr[\mathcal{E}^\varepsilon(f) = \tilde{f}] \propto e^{-\frac{\|f - \tilde{f}\|_1}{2\varepsilon}}$

Assigns very small probability to inaccurate outcomes.

Theorem: $\frac{\|f - \tilde{f}\|_1}{N} \leq O\left(\frac{1}{\varepsilon \sqrt{N}}\right)$ with high probability.

Theorem [MT07]: The exponential mechanism preserves $(\varepsilon, 0)$-differential privacy.
ONE DOES NOT SIMPLY (e.g., [U13])

"RUN" THE EXPONENTIAL MECHANISM
But, we did run the exponential mechanism

**Theorem:** There is an efficient algorithm $A$ to sample from a distribution that is $\delta$–close to the exponential mechanism $\mathcal{E}$ over integer partitions. The algorithm uses time and space

$$O\left(\frac{N\sqrt{N} + N \ln \left(\frac{1}{\delta}\right)}{\varepsilon}\right)$$

**Key Intuition:**

$$e^{-\varepsilon \sum_i |f_i - \tilde{f}_i|} = e^{-\varepsilon \sum_{i=t} |f_i - \tilde{f}_i|} \times e^{-\varepsilon \sum_{i>t} |f_i - \tilde{f}_i|}$$

Suggests Potential Recurrence Relationships
But, we did run the exponential mechanism

**Theorem:** There is an efficient algorithm A to sample from a distribution that is $\delta$–close to the exponential mechanism $\mathcal{E}$ over integer partitions. The algorithm uses time and space

$$O\left(\frac{N\sqrt{N} + N \ln\left(\frac{1}{\delta}\right)}{\varepsilon}\right)$$

**Key Idea 1:** Novel dynamic programming algorithm to compute weights $W_{i,k}$ such that

$$\Pr\left[\tilde{f}_i = k \mid \tilde{f}_{i-1}\right] = \frac{W_{i,k}}{\sum_{t=0}^{\tilde{f}_{i-1}} W_{i,t}}.$$
But, we did run the exponential mechanism

**Theorem:** There is an efficient algorithm $A$ to sample from a distribution that is $\delta$–close to the exponential mechanism $\mathcal{E}$ over integer partitions. The algorithm uses time and space

$$O\left(\frac{N \sqrt{N} + N \ln \left(\frac{1}{\delta}\right)}{\varepsilon}\right)$$

**Key Idea 1:** Novel dynamic programming algorithm to compute weights $W_{i,t}$

**Key Idea 2:** Allow $A$ to ignore a partition $\tilde{f}$ if $\|f - \tilde{f}\|_1$ very large.
Practical Challenge #1

• **Space is Limiting Factor:** $N=70$ million, $\varepsilon = 0.02$

\[
\frac{N\sqrt{N} + N \ln \left( \frac{1}{\delta} \right)}{\varepsilon} \text{ (8 bytes)} \approx 200 \ TB
\]

• **Workaround:** Initial pruning phase to identify relevant subset of DP table for sampling.

• **Running Time:** $\approx 12$ hours on this laptop
Practical Challenge #2

• $W_{i,k}$ can get very large (too big for native floating point types in C#)

• **Workaround**: Store $\log(W_{i,k})$ instead of $W_{i,k}$.

• **Important Implementation Question**: Where do your random bits come from?
  • Default random number generator is much easier for developer to use.
  • **Example**: `Rand.NextDouble()` vs `CryptoRand.NextBytes()`
Practical Challenge #3

Does Yahoo! have any preference about the privacy parameter $\varepsilon$?
Practical Challenge #3

Are there standardized guidelines to select $\varepsilon$?
Practical Challenge #3

No, I was thinking $\varepsilon = \frac{1}{2}$ would be reasonable....
Practical Challenge #3

Yahoo! is fine with $\epsilon = \frac{1}{2}$

**Risk:** Industry deployments become *de facto* standard for selecting $\epsilon$?

**Suggested Dinner Discussion Topic:** What role should academia play in influencing these standards?
### Yahoo! Results

<table>
<thead>
<tr>
<th></th>
<th>Original Data</th>
<th></th>
<th></th>
<th>Sanitized Data</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N)</td>
<td>(\log_2\left(\frac{N}{\lambda_1}\right))</td>
<td>(\log_2\left(\frac{N}{\lambda_{100}}\right))</td>
<td>(\log_2(G_{0.5}))</td>
<td>(\bar{N})</td>
<td>(\log_2\left(\frac{\bar{N}}{\bar{\lambda}_1}\right))</td>
</tr>
<tr>
<td>All</td>
<td>69,301,337</td>
<td>6.5</td>
<td>11.4</td>
<td>21.6</td>
<td>69,299,074</td>
<td>6.5</td>
</tr>
<tr>
<td>gender (self-reported)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>30,545,765</td>
<td>6.9</td>
<td>11.5</td>
<td>21.1</td>
<td>30,545,765</td>
<td>6.9</td>
</tr>
<tr>
<td>Male</td>
<td>38,624,554</td>
<td>6.3</td>
<td>11.3</td>
<td>21.8</td>
<td>38,624,554</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>language preference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>1,564,364</td>
<td>6.5</td>
<td>11.1</td>
<td>22.0</td>
<td>1,571,348</td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Yahoo! Frequency data is now available online at: [https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937](https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937)
## Yahoo! Results

<table>
<thead>
<tr>
<th></th>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N$</td>
<td>$\log_2 \left( \frac{N}{\lambda_1} \right)$</td>
</tr>
<tr>
<td>All</td>
<td>69,301,337</td>
<td>6.5</td>
</tr>
<tr>
<td>gender (self-reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>30,545,765</td>
<td>6.9</td>
</tr>
<tr>
<td>Male</td>
<td>38,624,554</td>
<td>6.3</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,564,364</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Yahoo! Frequency data is now available online at: [https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937](https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937)
Yahoo! Results (Selecting Epsilon)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
<th>( \log_2(G_{0.5}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>69,301,337</td>
<td>69,299,074</td>
<td>21.6</td>
</tr>
<tr>
<td>Female</td>
<td>30,545,765</td>
<td>30,545,765</td>
<td>21.1</td>
</tr>
<tr>
<td>Male</td>
<td>38,624,554</td>
<td>38,624,554</td>
<td>21.8</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,564,364</td>
<td>1,571,348</td>
<td>21.8</td>
</tr>
</tbody>
</table>

Any individual participates in at most 23 groups (including All)

\[ \varepsilon = \varepsilon_{all} + 22\varepsilon' \]
Yahoo! Results (Selecting Epsilon)

<table>
<thead>
<tr>
<th></th>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( N )</td>
<td>( \log_2 (\frac{N}{\varepsilon}) )</td>
</tr>
<tr>
<td>All</td>
<td>69,301,337</td>
<td>6.5</td>
</tr>
<tr>
<td>Female</td>
<td>30,545,765</td>
<td>6.9</td>
</tr>
<tr>
<td>Male</td>
<td>38,624,554</td>
<td>6.3</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,564,364</td>
<td>6.5</td>
</tr>
</tbody>
</table>

\( \varepsilon = \varepsilon_{all} + 22\varepsilon' \)

\( \varepsilon_{all} = 0.25 \)

\( \varepsilon' = \frac{\varepsilon_{all}}{22} \)
Yahoo! Results (Selecting Epsilon)

<table>
<thead>
<tr>
<th></th>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>$\log_2 \left( \frac{N}{\lambda_1} \right)$</td>
</tr>
<tr>
<td>All</td>
<td>69,301,337</td>
<td>6.5</td>
</tr>
<tr>
<td>gender (self-reported)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>30,545,765</td>
<td>6.9</td>
</tr>
<tr>
<td>Male</td>
<td>38,624,554</td>
<td>6.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>language preference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>1,564,364</td>
<td>6.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

$\varepsilon = 0.5$
## Yahoo! Results (Selecting Epsilon)

<table>
<thead>
<tr>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( \log_2 \left( \frac{N}{\lambda_1} \right) )</td>
</tr>
<tr>
<td>All</td>
<td>69,301,337</td>
</tr>
</tbody>
</table>

### gender (self-reported)

<table>
<thead>
<tr>
<th></th>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>38,624,554</td>
<td>6.3</td>
</tr>
</tbody>
</table>

### language preference

<table>
<thead>
<tr>
<th></th>
<th>Original Data [B12]</th>
<th>Sanitized Data [BDB16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>1,564,364</td>
<td>6.5</td>
</tr>
</tbody>
</table>

\[ \varepsilon = 0.5, \quad \delta = 2^{-100} \]
An Open Problem

**Conjecture:** For $\frac{1}{\varepsilon} = O\left(\frac{3}{\sqrt{n}}\right)$

$$E[\|\mathcal{E}^\varepsilon(f) - f\|_1] \leq O\left(\frac{\sqrt{n}}{\sqrt{\varepsilon}}\right)$$

Application to Social Networks: Degree Distribution with Node Privacy
Lower Bounds on L1 Error

\[
E[\|A(f) - f\|_1] = \Omega \left( \sqrt{\frac{N}{\varepsilon}} \right) \quad [\text{AS16,B16}]
\]

\[
E[\|A(f) - f\|_1] = \Omega \left( \frac{1}{\varepsilon^2} \right) \quad \text{relevant when } \frac{1}{\varepsilon} = \Omega(\sqrt{N})
\]
Empirical Evidence

$n = 32.6$ million users

$L_1$ Error (100 Samples)

$(\sqrt{\varepsilon})^{-1}$
More Empirical Evidence

\[ \varepsilon \approx \frac{1}{\sqrt{N}} \]
More Empirical Evidence

\[ \varepsilon \approx \frac{1}{\sqrt{N}} \]

\[ \varepsilon \approx \frac{1}{3\sqrt{N}} \]
Comparison with Prior Techniques

![Comparison with Prior Techniques](image.png)

Bitcoin Trust Network

- Laplace (Post Process)
- Exponential Mechanism

Mean Squared Error (200 Samples) vs. $1/\varepsilon^2$
Conclusions

• Differential Privacy Enables Analysis of Sensitive Data

• The exponential mechanism is not always intractable
  • integer partitions
  • Other practical settings?

• Applications to Social Networks?
Thanks for Listening

Anupam Datta
CMU

Joseph Bonneau
NYU