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

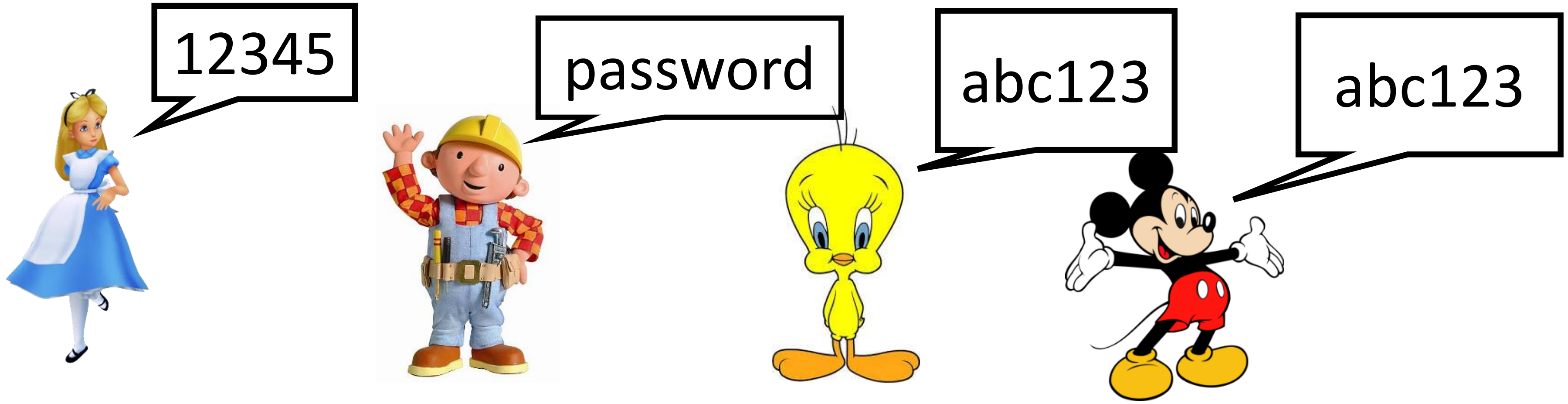
Jeremiah Blocki

Purdue University

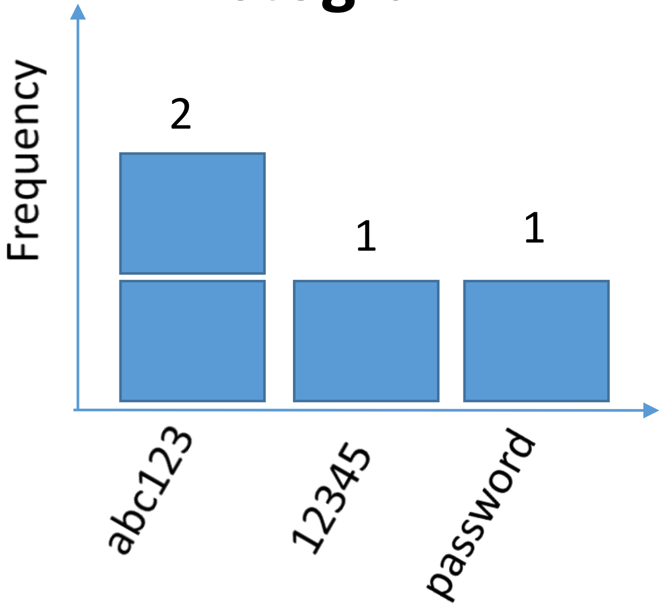
DIMACS/Northeast Big Data Hub Workshop on Overcoming Barriers to Data Sharing including Privacy and Fairness

What is a Password Frequency List?

Password Dataset:
(N users)

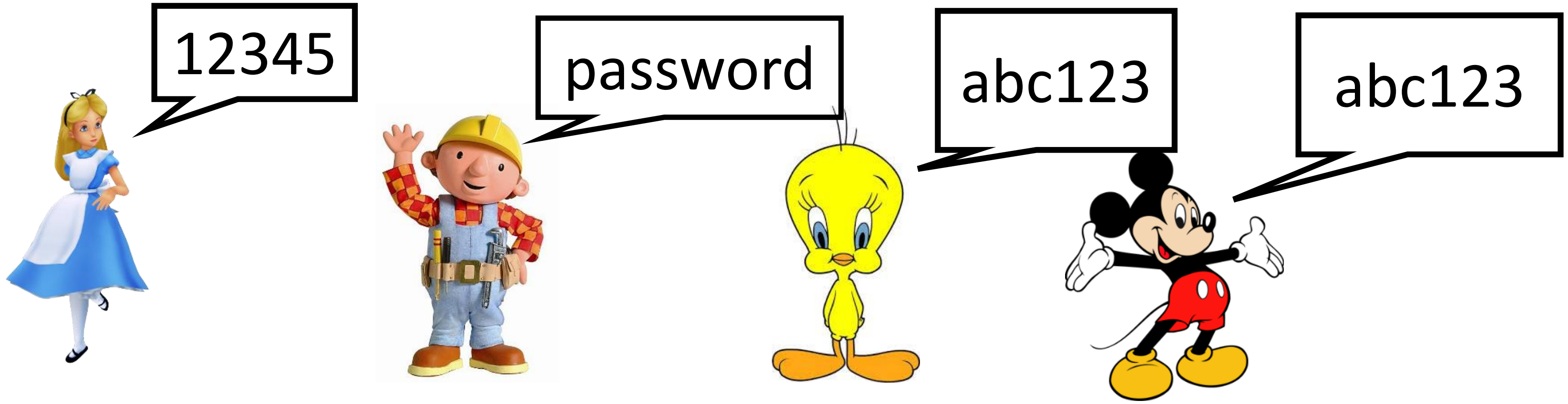


Histogram

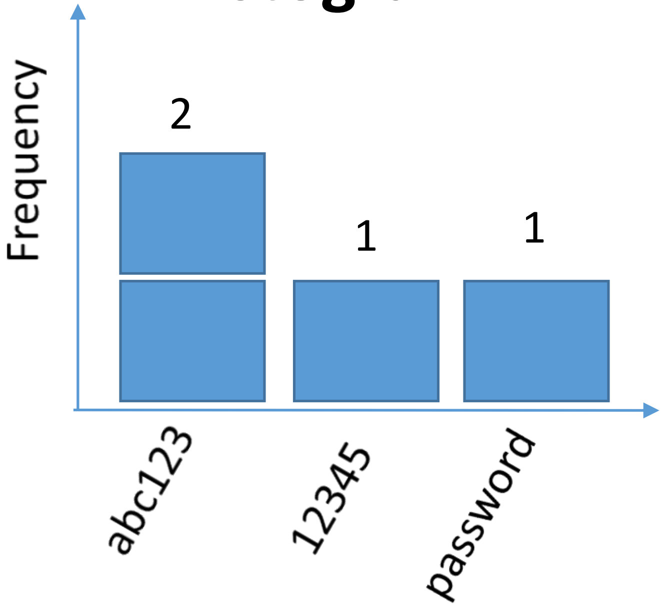


What is a Password Frequency List?

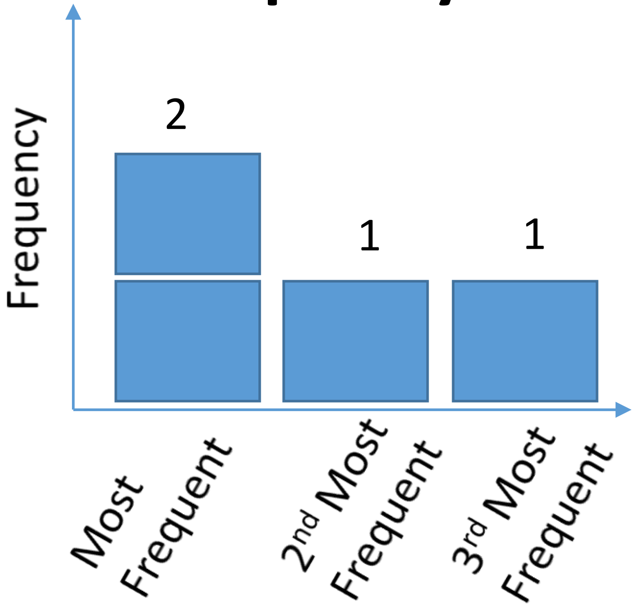
Password Dataset:
(N users)



Histogram

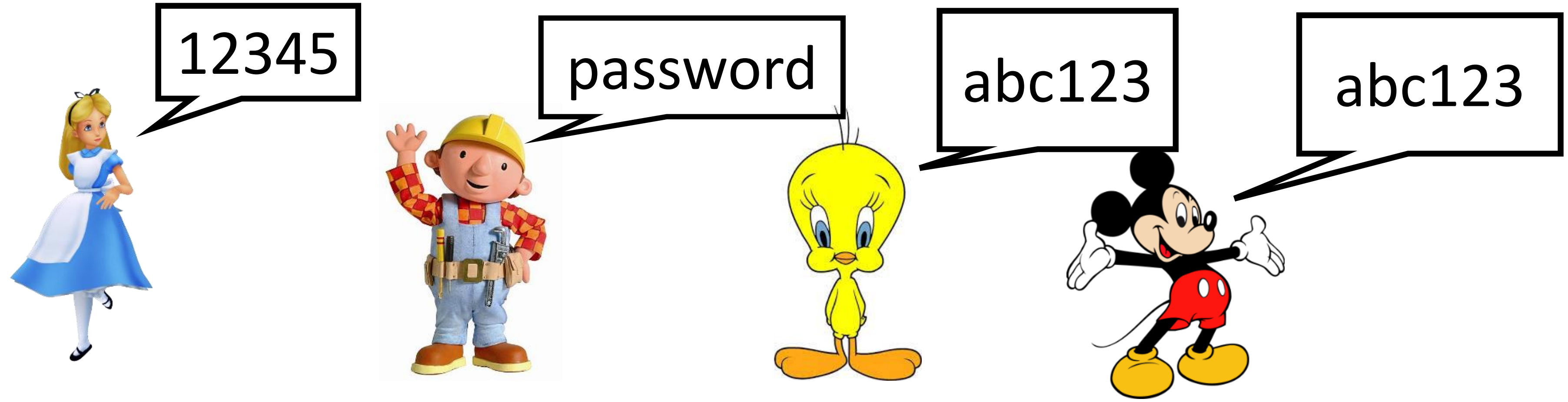


Frequency List

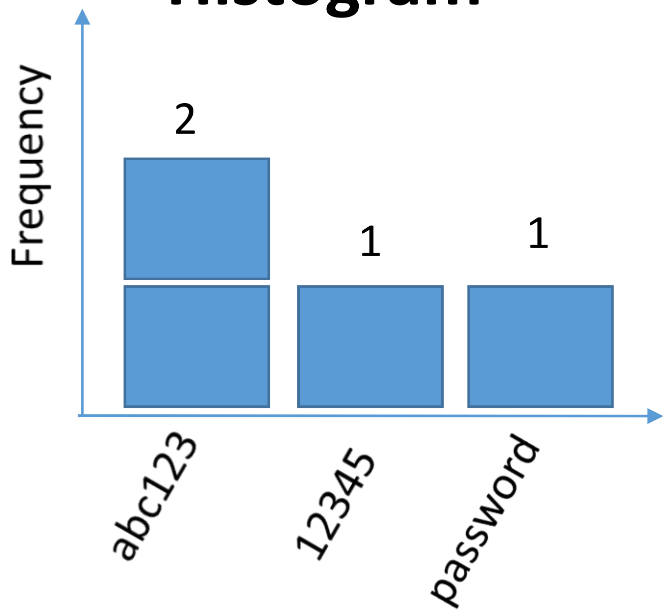


What is a Password Frequency List?

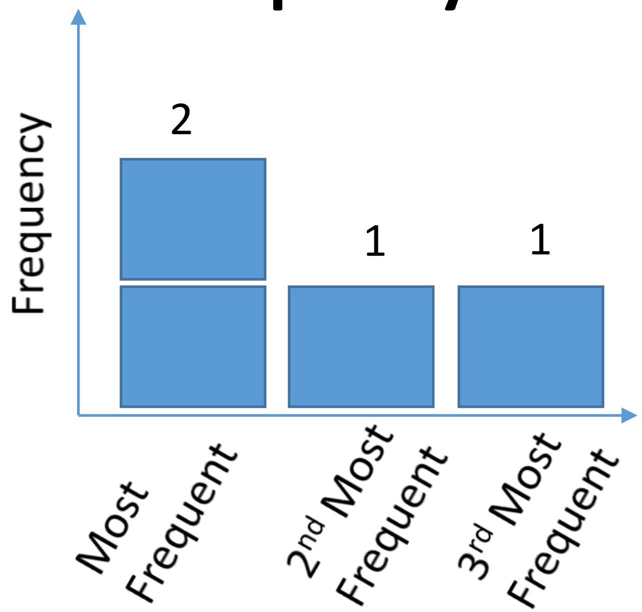
Password Dataset:
(N users)



Histogram



Frequency List



Formally:

$$\mathbf{f} \in \wp(\boldsymbol{N})$$

Password Frequency List is just an integer partition.

Password Frequency List (Example Use)

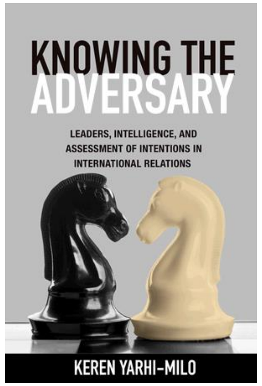
Estimate #accounts compromised by attacker with β guesses per user

- Online Attacker (β small)
- Offline Attacker (β 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)

Site	#User Accounts (N)	How Released
RockYou	32.6 Million	Data Breach*
LinkedIn	6	Data Breach*
….	…	…

* 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|pwd)
  - Secret salt value s (same for all users)
  - Discarded after data-collection
- ≈ 70 million Yahoo! Users

- Yahoo! Legal gave permission to publish analysis of the frequency list

Project Origin

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.

Project Origin

I would love to make the data public, but Yahoo! Legal has concerns about security and privacy. They won't let me release it.

Project Origin

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

Site	#User Accounts (N)	How Released
RockYou	32.6 Million	Data Breach*
LinkedIn	6	Data Breach*
….	…	…
Yahoo! [B12]	70 Million	With Permission**

* 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

Yahoo! Frequency Corpus

Largest publicly available frequency corpus

FORTUNE

Sc)

ANDY GREENB

HACI
BILL

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

TECHNOLOGY

All 3 Billion Yahoo Accounts Were Affected by 2013 Attack

By NICOLE PERLROTH 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.

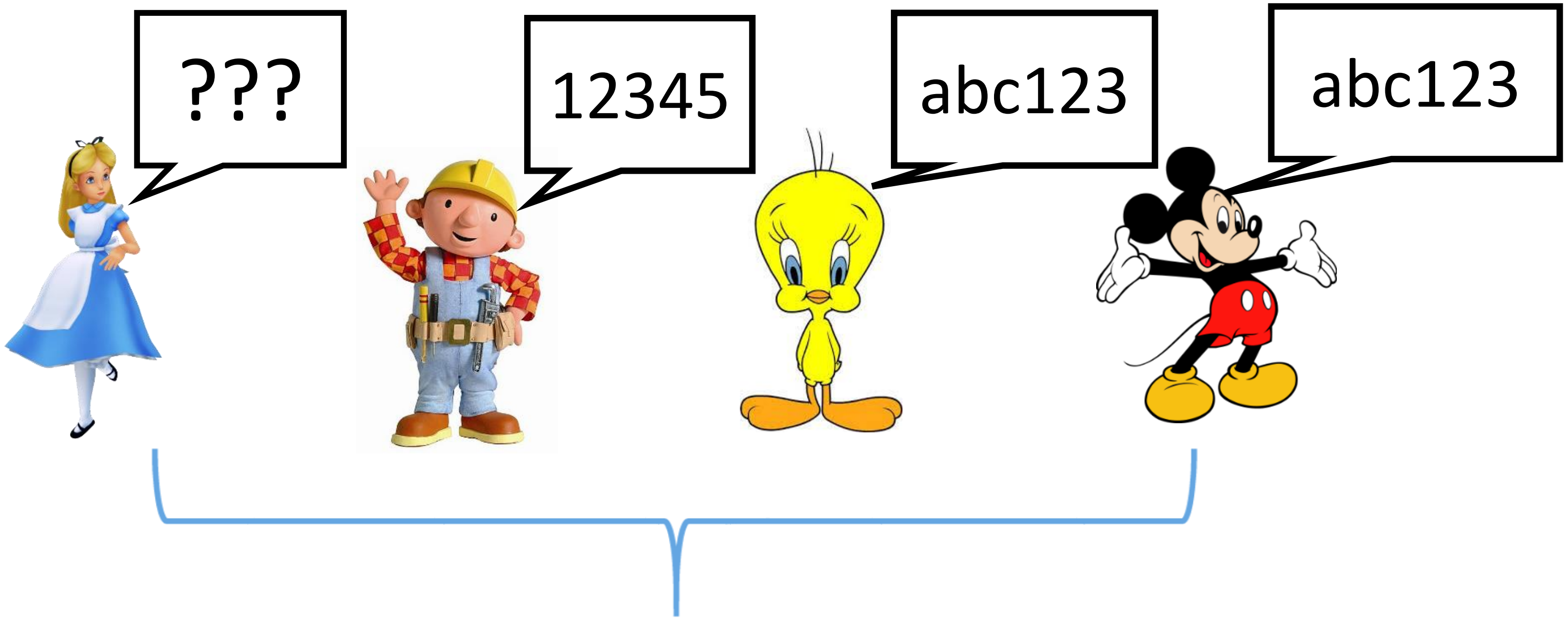
A laptop computer bearing the LinkedIn logo. Photograph by David Paul Morris—Bloomberg via Getty Images

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,…]

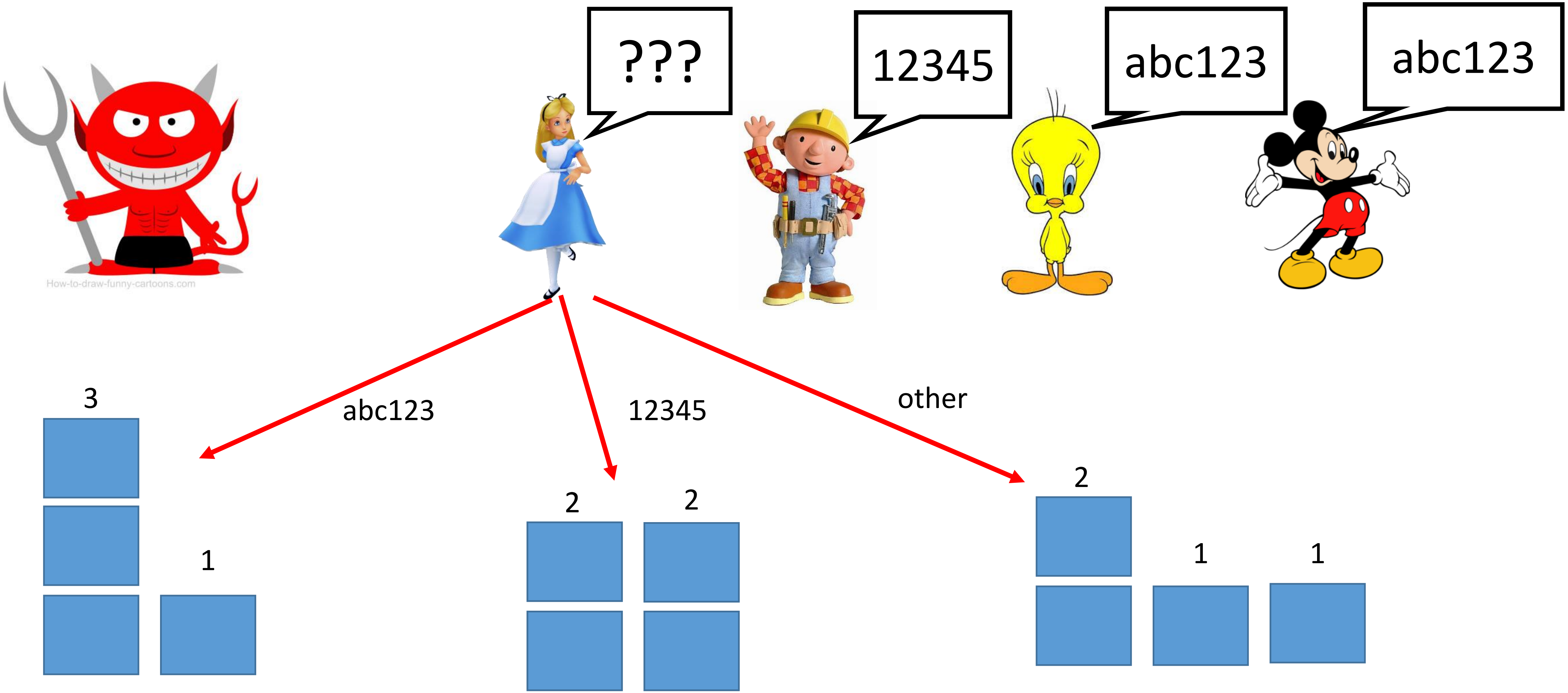
- In the absence of provable privacy guarantees Yahoo! was understandably reluctant to release these password frequency lists.

Security Risks (Example)



Adversary Background Knowledge

Security Risks (Example)



Differential Privacy (Dwork et al)

Definition: An (randomized) algorithm A preserves (ε, δ)-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',

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

$$\|f - f'\|_1 \stackrel{\text{def}}{=} \sum_i |f_i - f_i'|$$

Differential Privacy (Dwork et al)

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

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

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 (ε, δ)-differential privacy if for *any* subset S⊆ $Range(A)$ of possible outcomes and *any* we have

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

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

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

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

f – original password frequency list

f' – remove Alice's password from dataset

Differential Privacy (Dwork et al)

Definition: An (randomized) algorithm A preserves (ε, δ)-differential privacy if for *any* subset S⊆ $Range(A)$ of possible outcomes and *any* we have

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

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

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

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

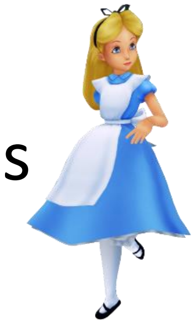
Negligibly Small Value (e.g., $\delta = 2^{-100}$)

f – original password frequency list

f' – remove Alice's password from dataset

Differential Privacy (Example)

2 2

minus  =

2

1

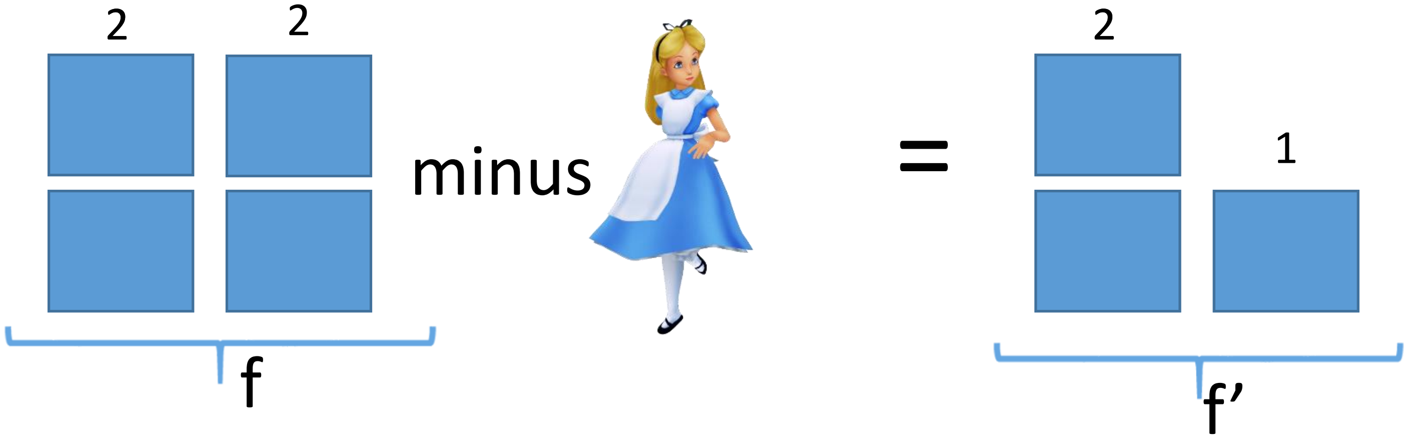
f

f’

Outcomes

Subset S of all potentially harmful outcomes to Alice

Differential Privacy (Example)



Subset S of all potentially harmful outcomes to Alice

$\mathbf{Pr}[A(f) \in$ 

$] \leq e^{\varepsilon}\,\mathbf{Pr}[A(f') \in$ 

$] + \delta$

Differential Privacy (Example)

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

$\mathbf{Pr}\left[A(f) \in$ 

$\right] \leq e^{\varepsilon}\, \mathbf{Pr}\left[A(f') \in$ 

$\right] + \delta$

Main Technical Result

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

$$\wp = \bigcup_{n=1}^{\infty} \wp(\boldsymbol{n})$$

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

$$\mathbf{Time}(\mathbf{A}) = O\left(\frac{N\sqrt{N} + N\ln\left({}^{1}/_{\delta}\right)}{\varepsilon}\right) = \mathbf{Space}(A)$$

Main Tool: Exponential Mechanism [MT07]

Input: f

Output: $\Pr\left[\mathcal{E}^{\varepsilon}(f) = \tilde{f}\right] \propto e^{-\frac{\|f-\tilde{f}\|_1}{2\varepsilon}}$ → **Assigns very small probability to inaccurate outcomes.**

Main Tool: Exponential Mechanism [MT07]

Input: f

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

Theorem [MT07]: The exponential mechanism preserves $(\varepsilon, 0)$-differential privacy.

Analysis: Exponential Mechanism

Input: f

Output: $\Pr\left[\mathcal{E}^{\varepsilon}(f) = \tilde{f}\right] \propto e^{-\frac{\|f-\tilde{f}\|_1}{2\varepsilon}}$ → **Assigns very small probability to inaccurate outcomes.**

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

Analysis: Exponential Mechanism

Input: f

Output: $\Pr\left[\mathcal{E}^{\varepsilon}(f) = \tilde{f}\right] \propto e^{-\frac{\|f-\tilde{f}\|_1}{2\varepsilon}}$ → **Assigns very small probability to inaccurate outcomes.**

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

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

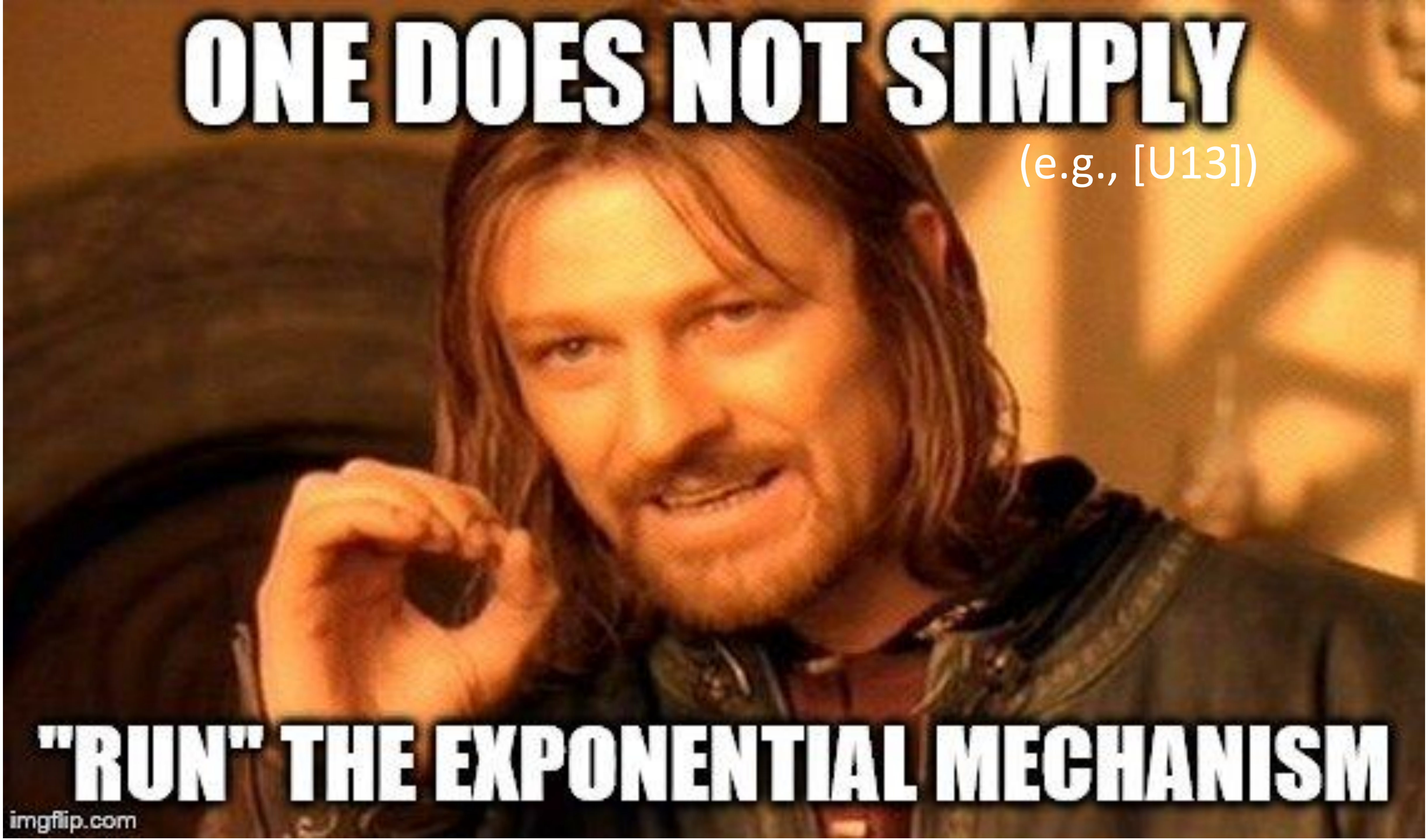
Analysis: Exponential Mechanism

Input: f

Output: $\Pr\left[\mathcal{E}^{\varepsilon}(f) = \tilde{f}\right] \propto e^{-\frac{\|f-\tilde{f}\|_1}{2\varepsilon}}$ → **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
imgflip.com

But, we did run the exponential mechanism

Theorem: There is an efficient algorithm A to sample from a distribution that is δ–close to the exponential mechanism Ɛ over integer partitions. The algorithm uses time and space

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

Key Intuition:

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

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 δ–close to the exponential mechanism Ɛ over integer partitions. The algorithm uses time and space

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

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

$$\mathbf{Pr}\left[\tilde{f}_i = k \,\middle|\, \tilde{f}_{i-1}\right] = \frac{\mathrm{W}_{\mathrm{i},k}}{\sum_{\mathrm{t}=0}^{\tilde{f}_{i-1}} \mathrm{W}_{\mathrm{i,t}}}.$$

But, we did run the exponential mechanism

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

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

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, $\varepsilon = 0.02$

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

- **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 ε?

Practical Challenge #3

Are there standardized guidelines to select ε?

Practical Challenge #3

No, I was thinking $\varepsilon = \frac{1}{2}$ would be reasonable….

Practical Challenge #3

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

Risk: Industry deployments become *de facto* standard for selecting ε?

Suggested Dinner Discussion Topic: What role should academia play in influencing these standards?

Yahoo! Results

	Original Data				Sanitized Data			
	N	$\log_2\left(\frac{N}{\lambda_1}\right)$	$\log_2\left(\frac{N}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$	$\tilde{N}$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_1}\right)$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_{100}}\right)$	$\log_2(G_{0.5})$
All	69,301,337	6.5	11.4	21.6	69,299,074	6.5	11.4	21.6
gender (self-reported)								
Female	30,545,765	6.9	11.5	21.1	30,545,765	6.9	11.5	21.1
Male	38,624,554	6.3	11.3	21.8	38,624,554	6.3	11.3	21.8
…	…	…	…	…	…	…	…	…
language preference								
Chinese	1,564,364	6.5	11.1	22.0	1,571,348	6.5	11.1	*21.8*
…	…	…	…	…	…	…	…	…

Yahoo! Frequency data is now available online at:
https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937

Yahoo! Results

	Original Data [B12]				Sanitized Data [BDB16]			
	N	$\log_2\left(\frac{N}{\lambda_1}\right)$	$\log_2\left(\frac{N}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$	$\tilde{N}$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_1}\right)$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_{100}}\right)$	$\log_2(G_{0.5})$
All	69,301,337	6.5	11.4	21.6	69,299,074	6.5	11.4	21.6
gender (self-reported)								
Female	30,545,765	6.9	11.5	21.1	30,545,765	6.9	11.5	21.1
Male	38,624,554	6.3	11.3	21.8	38,624,554	6.3	11.3	21.8
…	…	…	…	…	…	…	…	…
language preference								
Chinese	1,564,364	6.5	11.1	22.0	1,571,348	6.5	11.1	*21.8*
…	…	…	…	…	…	…	…	…

Yahoo! Frequency data is now available online at:
https://figshare.com/articles/Yahoo_Password_Frequency_Corpus/2057937

Yahoo! Results (Selecting Epsilon)

	Original Data [B12]				Sanitized Data [BDB16]			
	N	$\log_2\left(\frac{N}{\lambda_1}\right)$	$\log_2\left(\frac{N}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$	$\tilde{N}$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_1}\right)$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_{100}}\right)$	$\log_2(G_{0.5})$
All	69,301,337	6.5	11.4	21.6	69,299,074	6.5	11.4	21.6
[illegible]								
Female	30,545,765	[illegible]	11.5	[illegible]	[illegible]	6.9	11.5	21.1
Male	38,624,554	6.3	11.3	21.8	38,624,554	6.3	11.3	21.8
…	…	…	…	…	…	…	…	…
language preference								
Chinese	1,564,364	6.5	11.1	22.0	1,571,348	6.5	11.1	*21.8*
…	…	…	…	…	…	…	…	…

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

$$\varepsilon = \varepsilon_{all} + 22\varepsilon'$$

Yahoo! Results (Selecting Epsilon)

	Original Data [B12]				Sanitized Data [BDB16]			
	N	$\log_2\left(\frac{N}{\lambda_1}\right)$	$\log_2\left(\frac{N}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$	$\tilde{N}$	$\log_2\left(\frac{\tilde{N}}{\lambda_1}\right)$	$\log_2\left(\frac{\tilde{N}}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$
All	69,301,337							21.6
gender (self-reported)								
Female	30,545,765							21.1
Male	38,624,554							21.8
…	…							…
language preference								
Chinese	1,564,364	6.5	11.1	22.0	1,571,348	6.5	11.1	*21.8*
…	…	…	…	…	…	…	…	…

$$\varepsilon_{all} = 0.25$$

$$\varepsilon' = \frac{\varepsilon_{all}}{22}$$

$$\varepsilon = \varepsilon_{all} + 22\varepsilon'$$

Yahoo! Results (Selecting Epsilon)

	Original Data [B12]				Sanitized Data [BDB16]			
	N	$\log_2\left(\frac{N}{\lambda_1}\right)$	$\log_2\left(\frac{N}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$	$\tilde{N}$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_1}\right)$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_{100}}\right)$	$\log_2(G_{0.5})$
All	69,301,337	6.5	11.4	21.6	69,299,074	6.5	11.4	21.6
gender (self-reported)								
Female	30,545,765	6.9	11.5	21.1	30,545,765	6.9	11.5	21.1
Male	38,624,554	6.3	11.3	21.8	38,624,554	6.3	11.3	21.8
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
language preference								
Chinese	1,564,364	6.5	11.1	22.0	1,571,348	6.5	11.1	*21.8*
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**

$$\varepsilon = 0.5$$

Yahoo! Results (Selecting Epsilon)

	Original Data [B12]				Sanitized Data [BDB16]			
	N	$\log_2\left(\frac{N}{\lambda_1}\right)$	$\log_2\left(\frac{N}{\lambda_{100}}\right)$	$\log_2(G_{0.5})$	$\tilde{N}$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_1}\right)$	$\log_2\left(\frac{\tilde{N}}{\tilde{\lambda}_{100}}\right)$	$\log_2(G_{0.5})$
All	69,301,337	6.5	11.4	21.6	69,299,074	6.5	11.4	21.6
gender (self-reported)								
Female	30,545,765	6.9	11.5	21.1	30,545,765	6.9	11.5	21.1
Male	38,624,554	6.3	11.3	21.8	38,624,554	6.3	11.3	21.8
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
language preference								
Chinese	1,564,364	6.5	11.1	22.0	1,571,348	6.5	11.1	*21.8*
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**

$$\varepsilon = 0.5, \qquad \delta = 2^{-100}$$

An Open Problem

Conjecture: For $\frac{1}{\varepsilon} = O(\sqrt[3]{n})$

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

Application to Social Networks: Degree Distribution with Node Privacy

Lower Bounds on L1 Error

$$\mathrm{E}[\|A(f) - f\|_1] = \Omega\left(\sqrt{\frac{N}{\varepsilon}}\right)$$

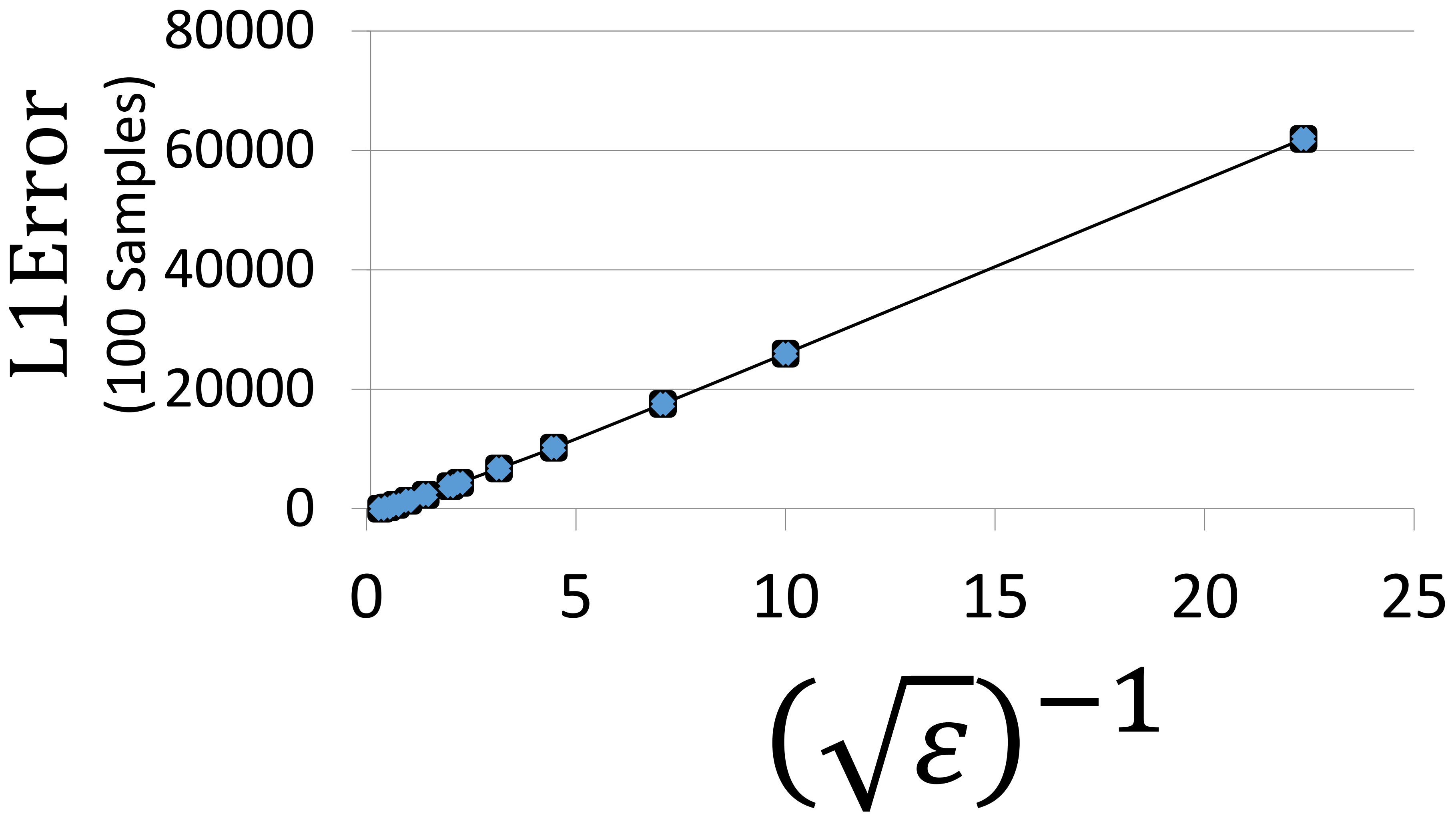
[AS16,B16]

$$\mathrm{E}[\|A(f) - f\|_1] = \Omega\left(\frac{1}{\varepsilon^2}\right)$$

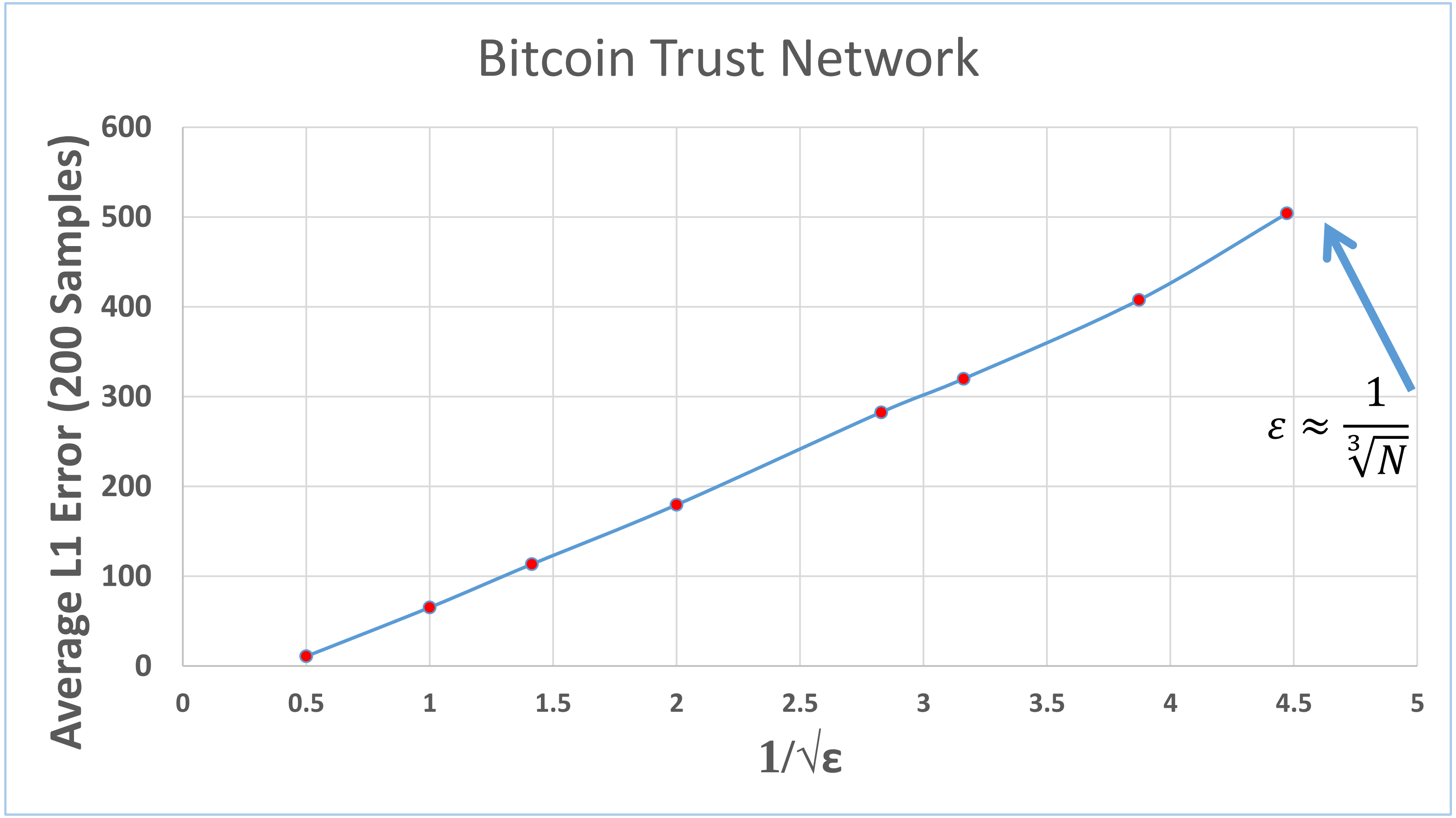
relevant when $\frac{1}{\varepsilon} = \Omega(\sqrt{N})$

Empirical Evidence

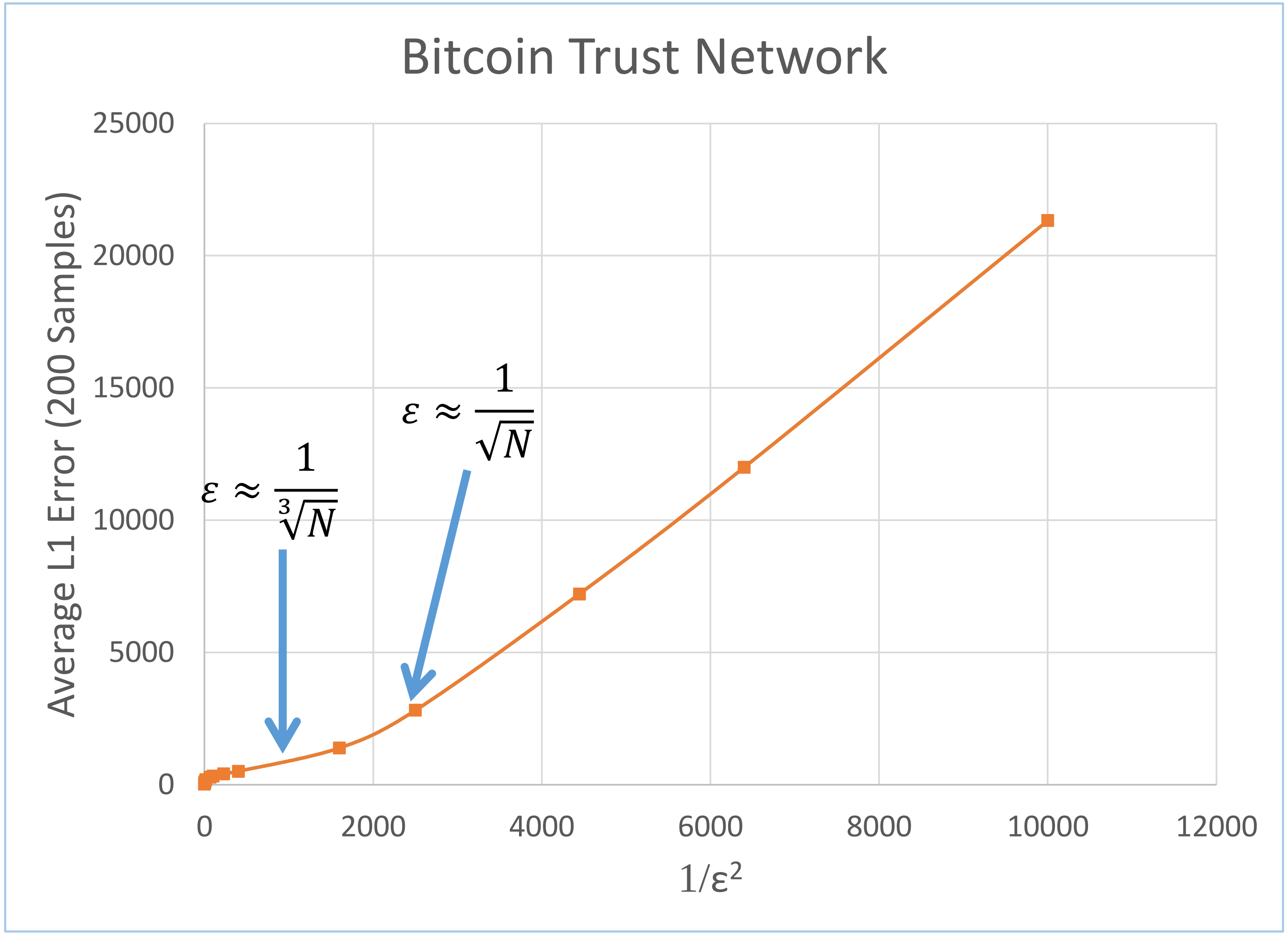
n=32.6 million users



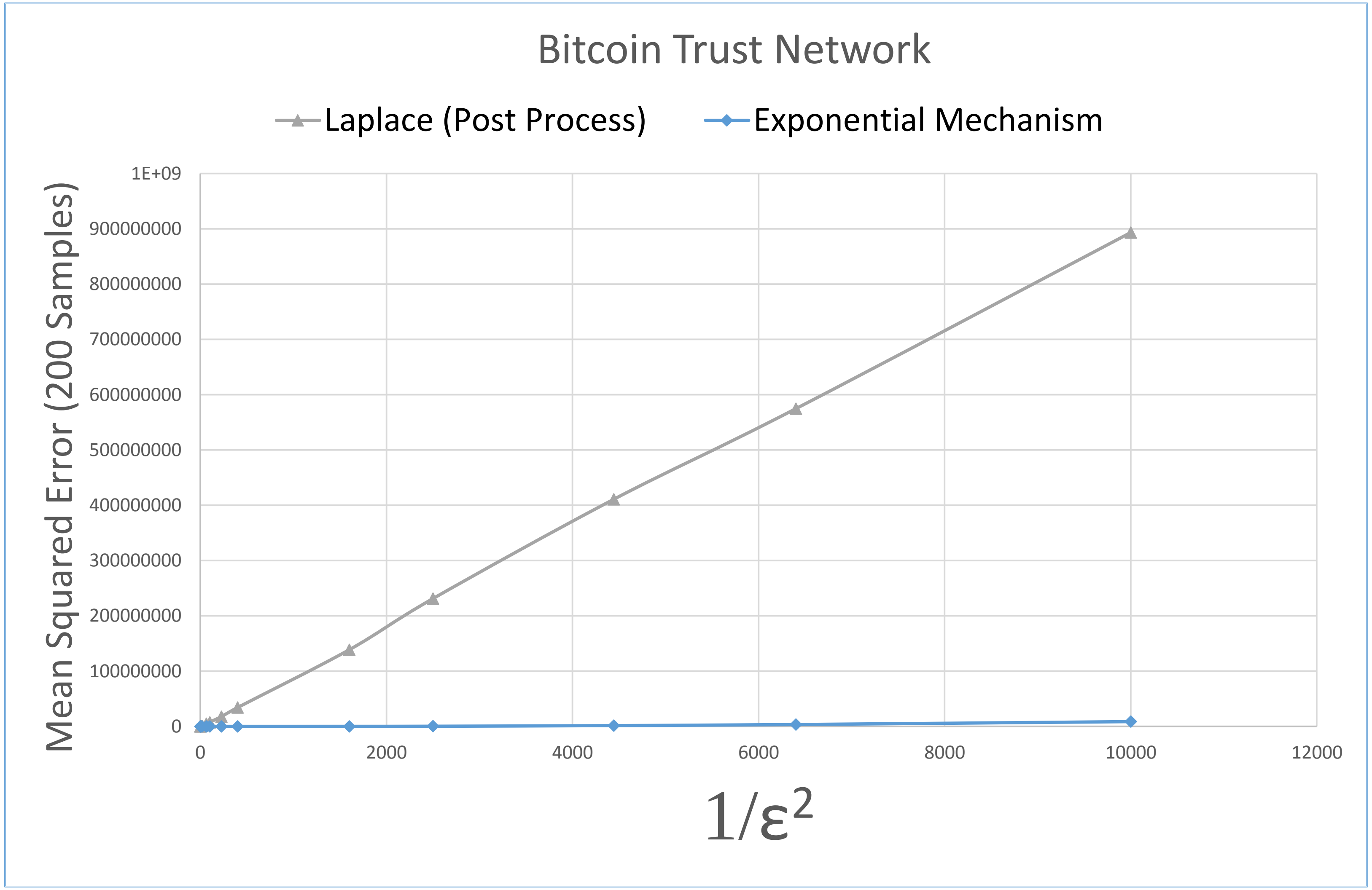
More Empirical Evidence



More Empirical Evidence



Comparison with Prior Techniques



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