Deploying Differential Privacy in Industry: Progress and Learnings

Ryan Rogers

Staff Software Engineer, LinkedIn

November 29, 2021, UCSB

| DATA SCIENCE |

Collaborators: Subbu Subramaniam, Sean Peng, Seunghyun Lee, Sajjad Moradi, Akash Kaura, Nikhil Gahlawat, Adrian Rivera Cardoso, Mark Cesar, Jinshuo Dong, David Durfee, Koray Mancuhan, Paul Ko, Santosh Kumar Kancha, Neha Jain, Shraddha Sahay, Parvez Ahammad, Ya Xu

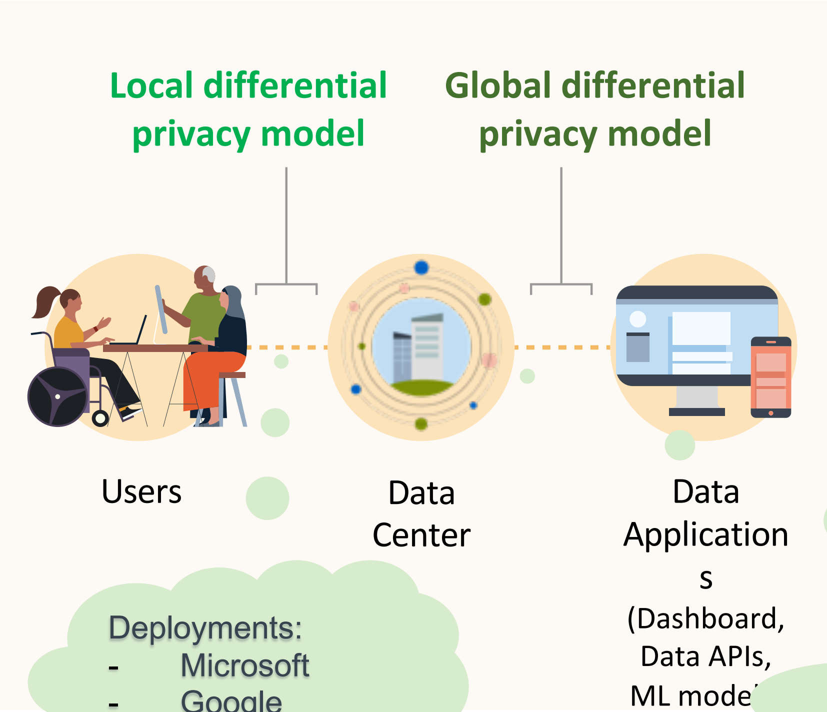
Agenda

1. Overview of Differential Privacy
2. Overview of DP approach at LinkedIn
3. DP Top-*k* Algorithms
4. Privacy Budget Management
5. Deployments and Conclusion

Mission

Utilizing data while protecting the privacy of members.

Models and Deployments of Differential Privacy



Deployments:
- Microsoft
- Google
- Apple

- Traditional data protection techniques are not sufficient to defend data privacy
- Differential Privacy ensures data learnings are similar with/without a single member's data

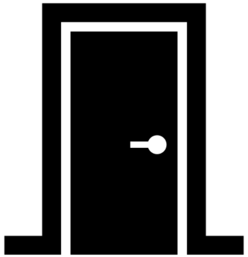
Deployments:
- 2020 Census
- Microsoft Open Data DP Project
- Google's Mobility Reports
- FB's release of publicly shared URLs

Agenda

1. Overview of Differential Privacy
2. Overview of DP approach at LinkedIn
3. DP Top-*k* Algorithms
4. Privacy Budget Management
5. Deployments and Conclusion

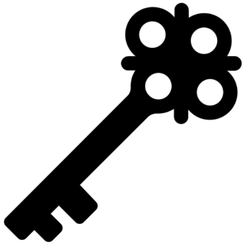
Initial Discussions

- What teams are interested in releasing aggregates?
- What are the general problems and what solutions would be the most applicable?
- What additional constraints are there?

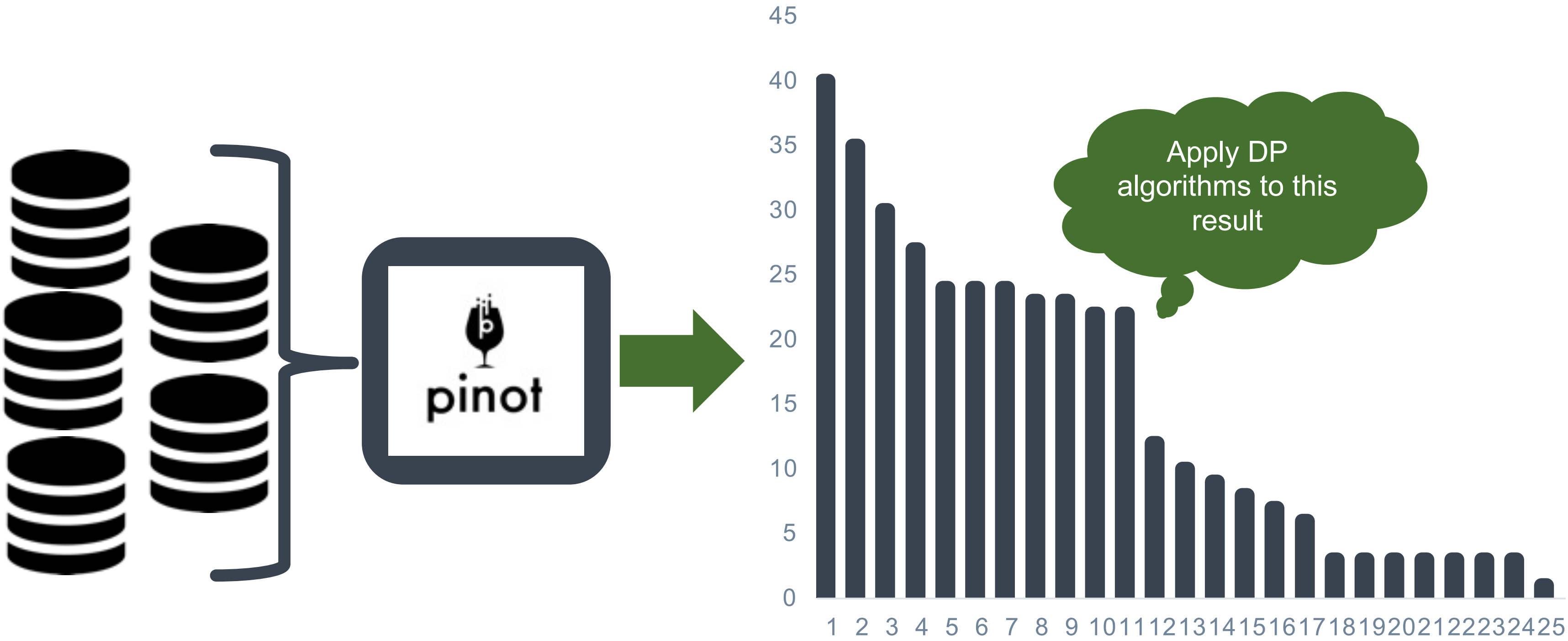


Key Takeaways

- Existing infrastructure for computing aggregates quickly.
- Want tunable privacy as well as tunable run time.
- Lots of data analytics can be reduced to histograms.
- Labels of the histograms are not always known.
- Typically, only want top-*k* results
- Want consistent results, see PriPeARL [Kenthapadi,Tran'18].



Existing Systems for Data Analytics

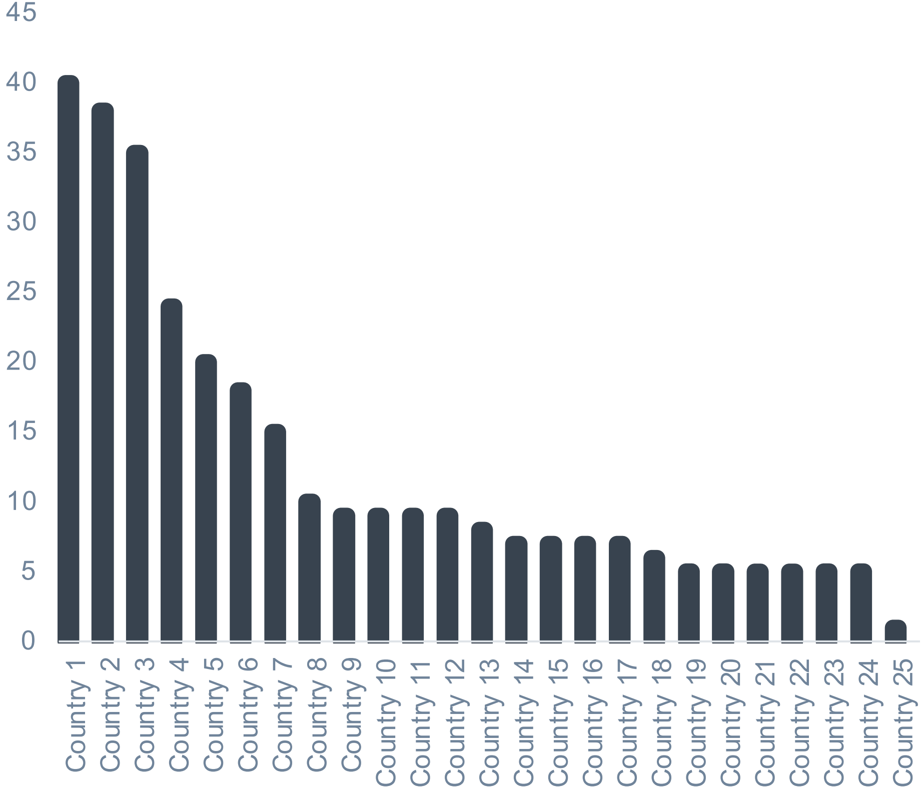


Agenda

1. Overview of Differential Privacy
2. Overview of DP approach at LinkedIn
3. DP Top-*k* Algorithms
4. Privacy Budget Management
5. Deployments and Conclusion

Sensitivity of the Query

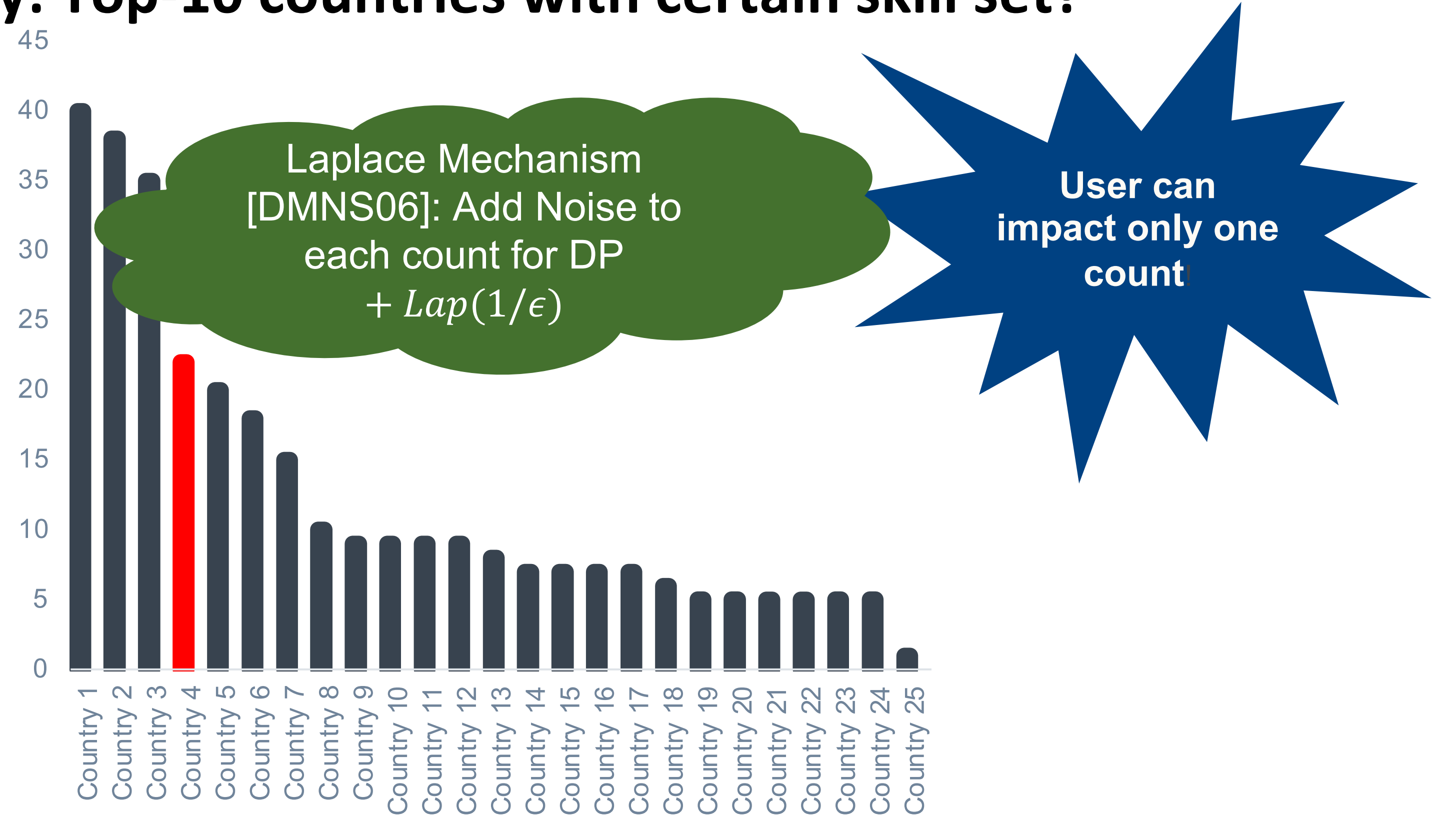
Query: Top-10 countries with certain skill set?



Aggregate Data

Sensitivity of the Query

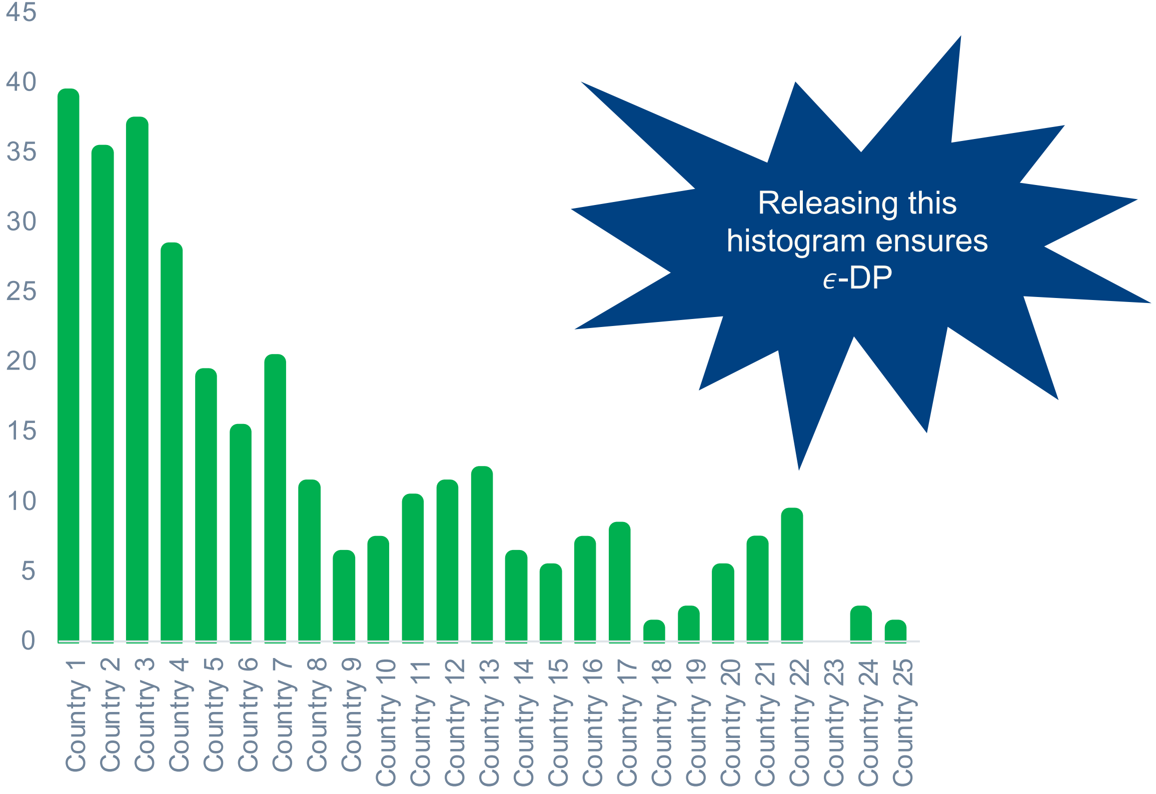
Query: Top-10 countries with certain skill set?



Aggregate Data

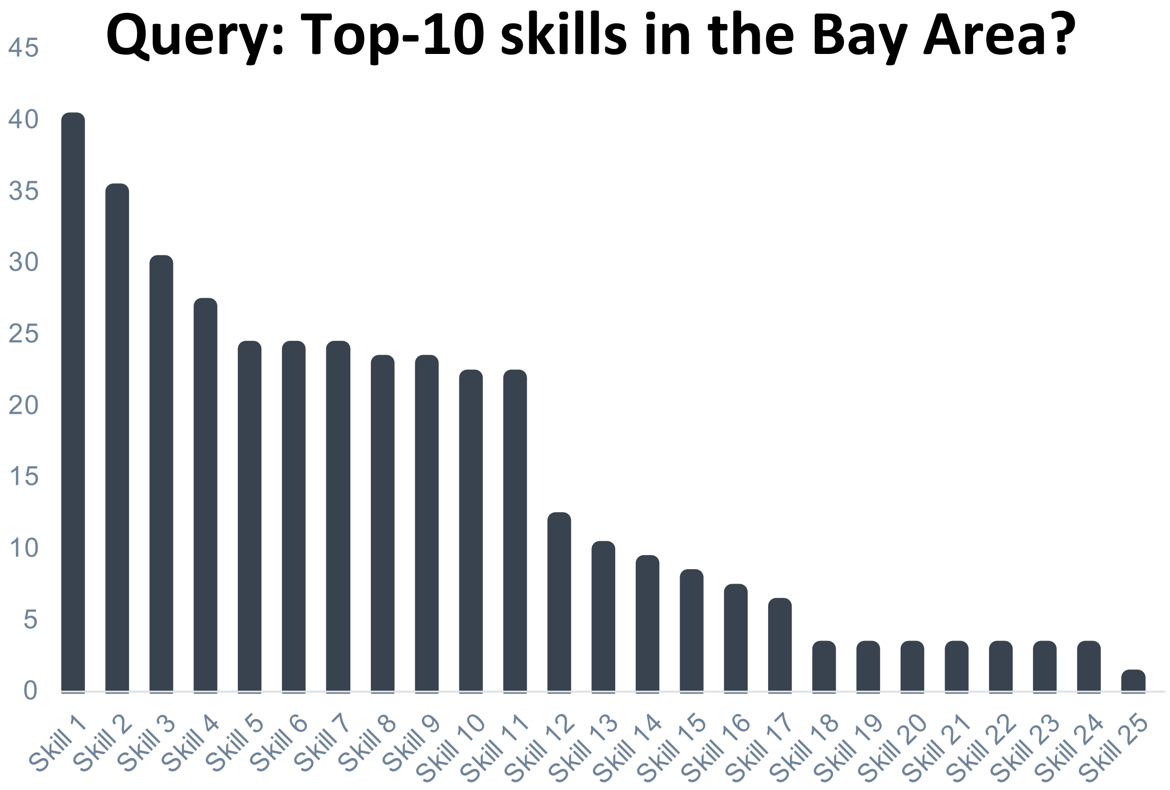
Sensitivity of the Query

Query: Top-10 countries with certain skill set?



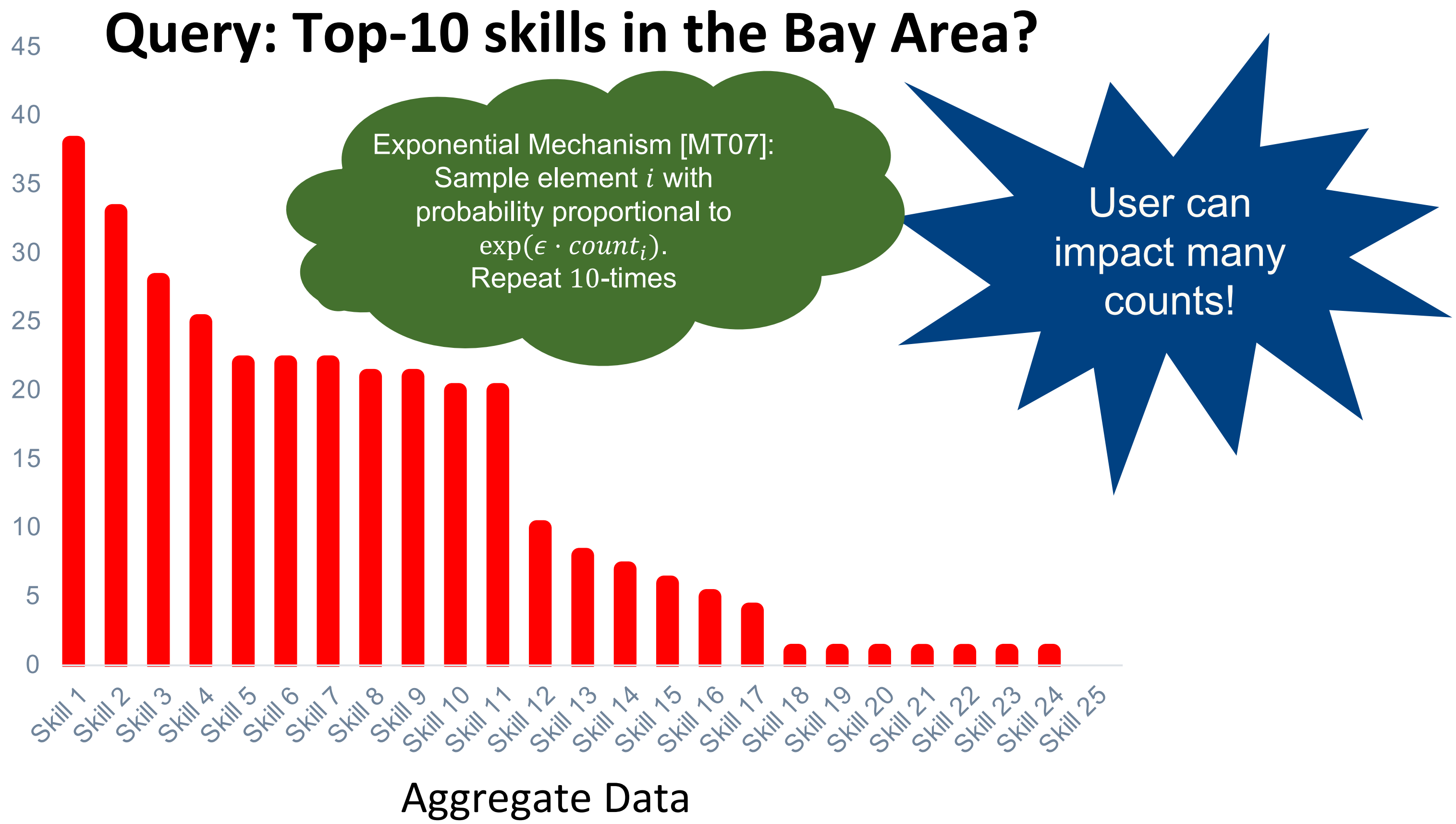
Aggregate Data

Sensitivity of the Query



Aggregate Data

Sensitivity of the Query



Sensitivity of the Query

Query: Top-10 skills in the Bay Area?

Exponential Mechanism [MT07]:
Sample element i with
probability proportional to
$\exp(\epsilon \cdot count_i)$.
Repeat 10-times

Releasing only elements in top-k
(not their counts) ensures
$k\epsilon$-DP

Known Algorithms for Private Data Analytics

ℓ_0-Restricted Sensitivity	ℓ_0-Unrestricted Sensitivity
Algorithm: Laplace Mechanism [DMNS‘06]	**Algorithm: Exponential Mechanism** [MT’07]

Implementing Exp Mech

- Folklore result: Exp Mech = Adding $\mathit{Gumbel}\left(\frac{1}{\epsilon}\right)$ to each count and reporting the arg noisy max.
- [DR'19] Can simulate repeated Exponential Mechanisms in one-shot this way to get $\left(\approx \epsilon\sqrt{k \log \frac{1}{\delta}}, \delta\right)$-DP.
- Improves on work from [Dwork, Su, Zhang '15] and [Garg, Su, Zhang '21] that adds $\mathit{Laplace}\left(\frac{1}{\epsilon}\right)$ to each count … reports the k largest noisy count elemen… order.

See blog post:
https://differentialprivacy.org/one-shot-top-k/

Known Algorithms for Private Data Analytics

ℓ_0-Restricted Sensitivity	ℓ_0-Unrestricted Sensitivity
Algorithm: Laplace Mechanism [DMNS'06]	**Algorithm: Exponential Mechanism** [MT'07]

Known Algorithms for Private Data Analytics

ℓ_0-Restricted Sensitivity	ℓ_0-Unrestricted Sensitivity
Algorithm: Known Laplace [DMNS‘06]	**Algorithm: Known Gumbel** [MT‘07]

Solving Top-k subject to DP

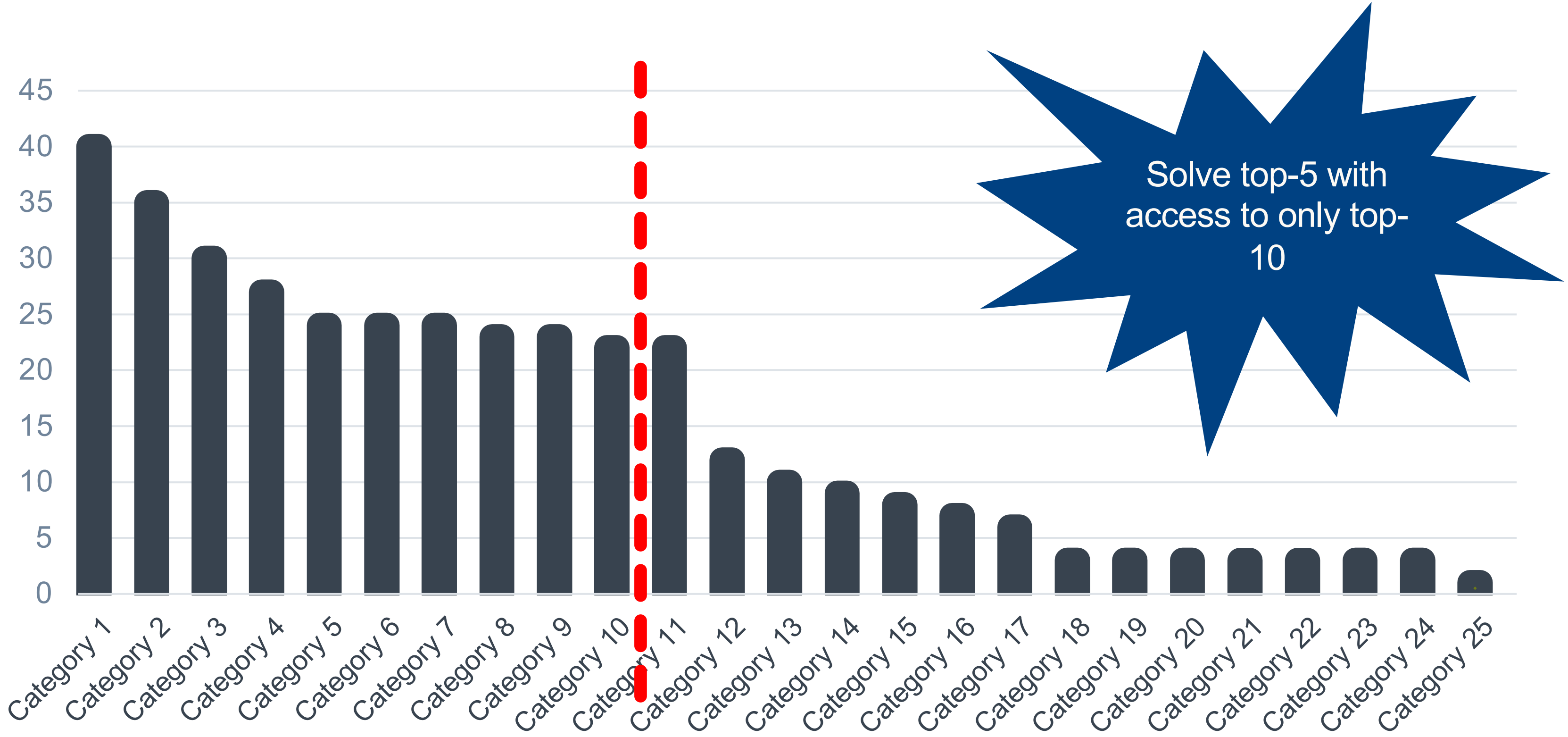
- One of the most fundamental problems in exploratory data analytics
- Lots of work in DP on solving top-k
  - Local Model of DP (Heavy Hitters)
    - Bassily and Smith STOC'15
    - Fanti, Pihur, Erlingsson PoPETS'16
    - Bassily, Nissim, Stemmer, Thakurta NIPS'17.
  - Global Model of DP
    - Bhaskar, Laxman, Smith, Thakurta KDD'10
    - Li, Qardaji, Su, Cao VLDB'12
    - Zeng, Naughton, Cai VLDB'12
    - Lee and Clifton, KDD'14
    - Chaudhuri, Hsu, Song NIPS'14
    - Zhu, Kairouz, Sun, McMahan, Li '19

All require knowing structure about the data domain or need the count of every element

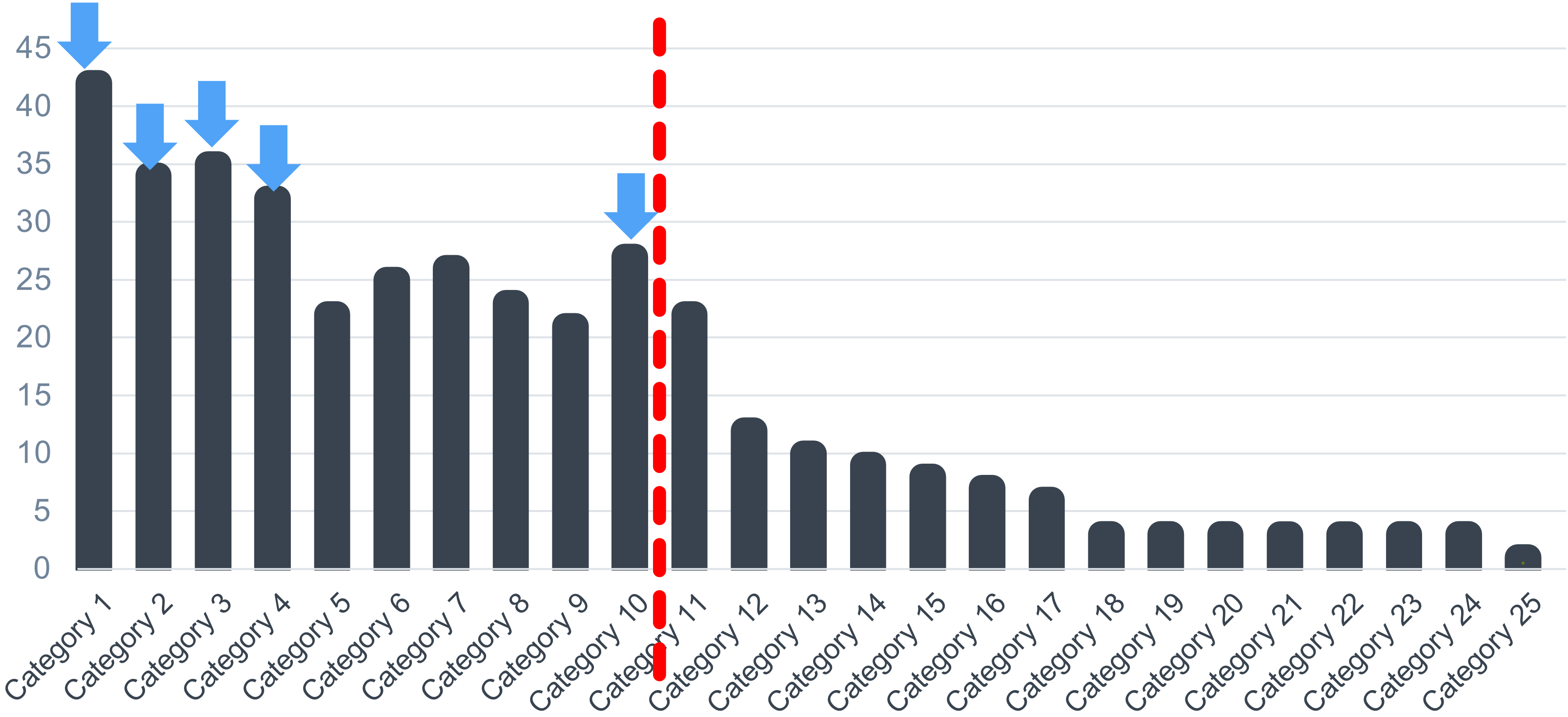
Unknown Domain Setting

- Typically, the domain is unknown or very large and we restrict how many elements to consider
- Lots of prior work for Frequent Itemsets, but requires knowing structure of the data domain universe.
  - Can prune the number of things we need to query.
- Related work requiring full histogram:
  - [Korolova, Kenthapadi, Mishra, Ntoulas '09]
  - [Wilson, Zhang, Lam, Desfontaines, Simmons-Marengo, Gipson'20]
  - DP Set Union [Gopi, Gulhane, Kulkarni, Shen, Shokouhi, Yekhanin '20]

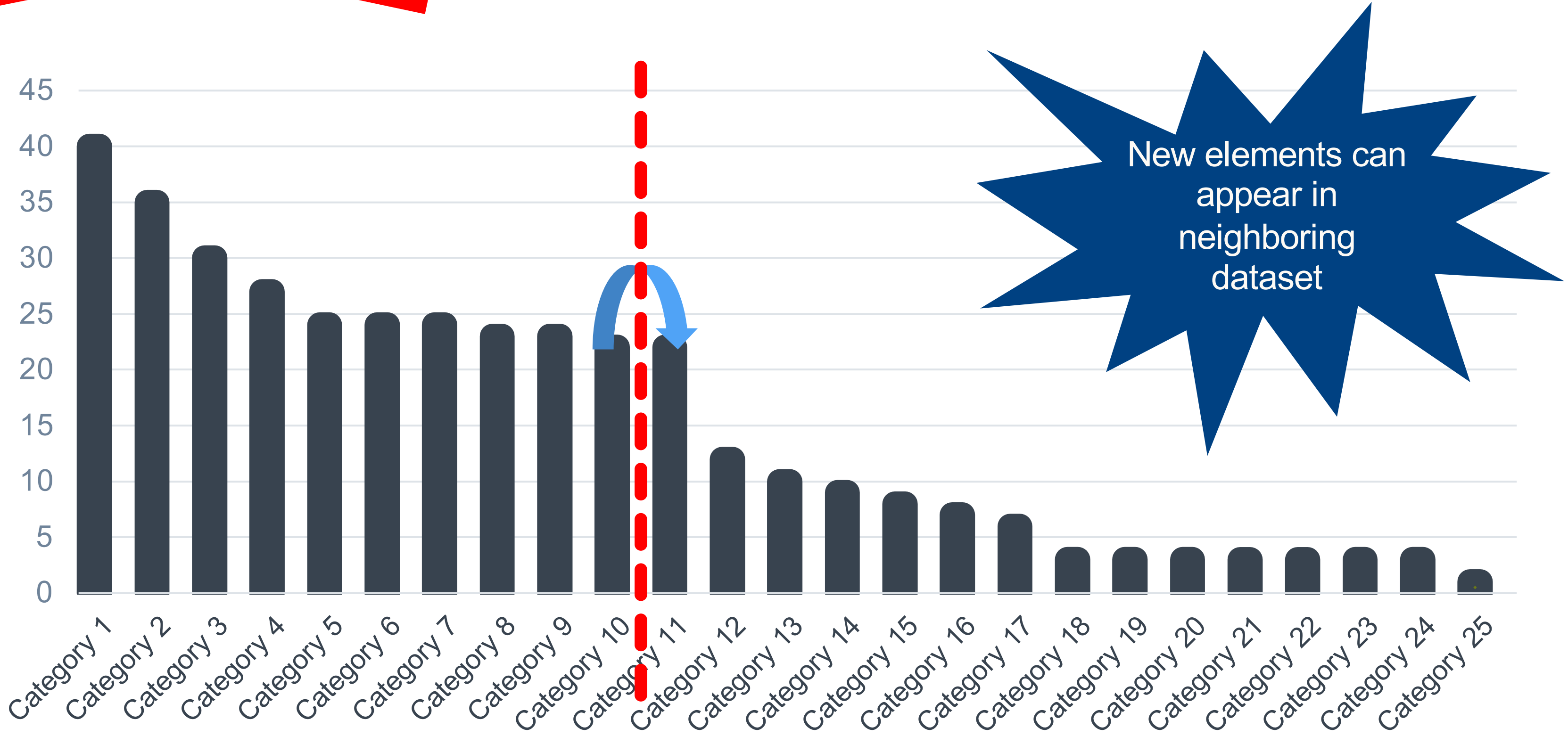
First Attempt



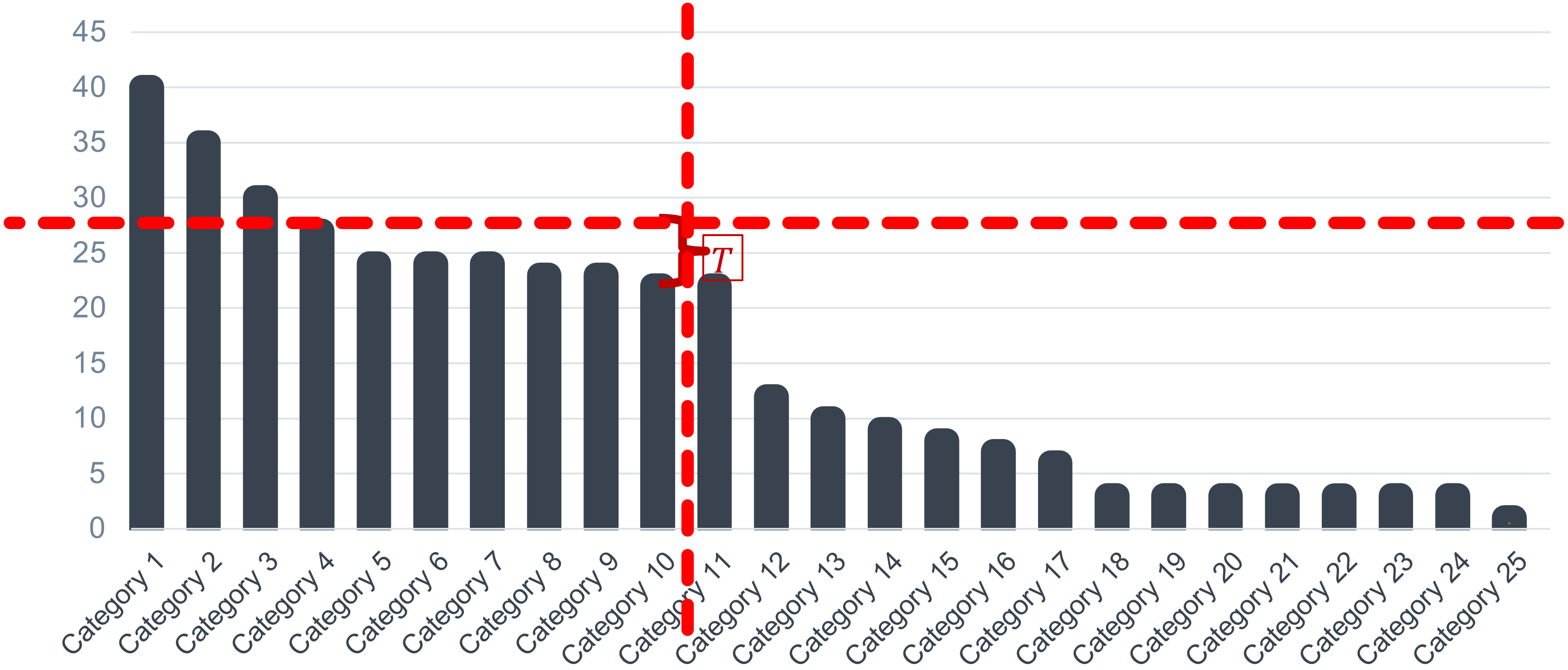
First Attempt



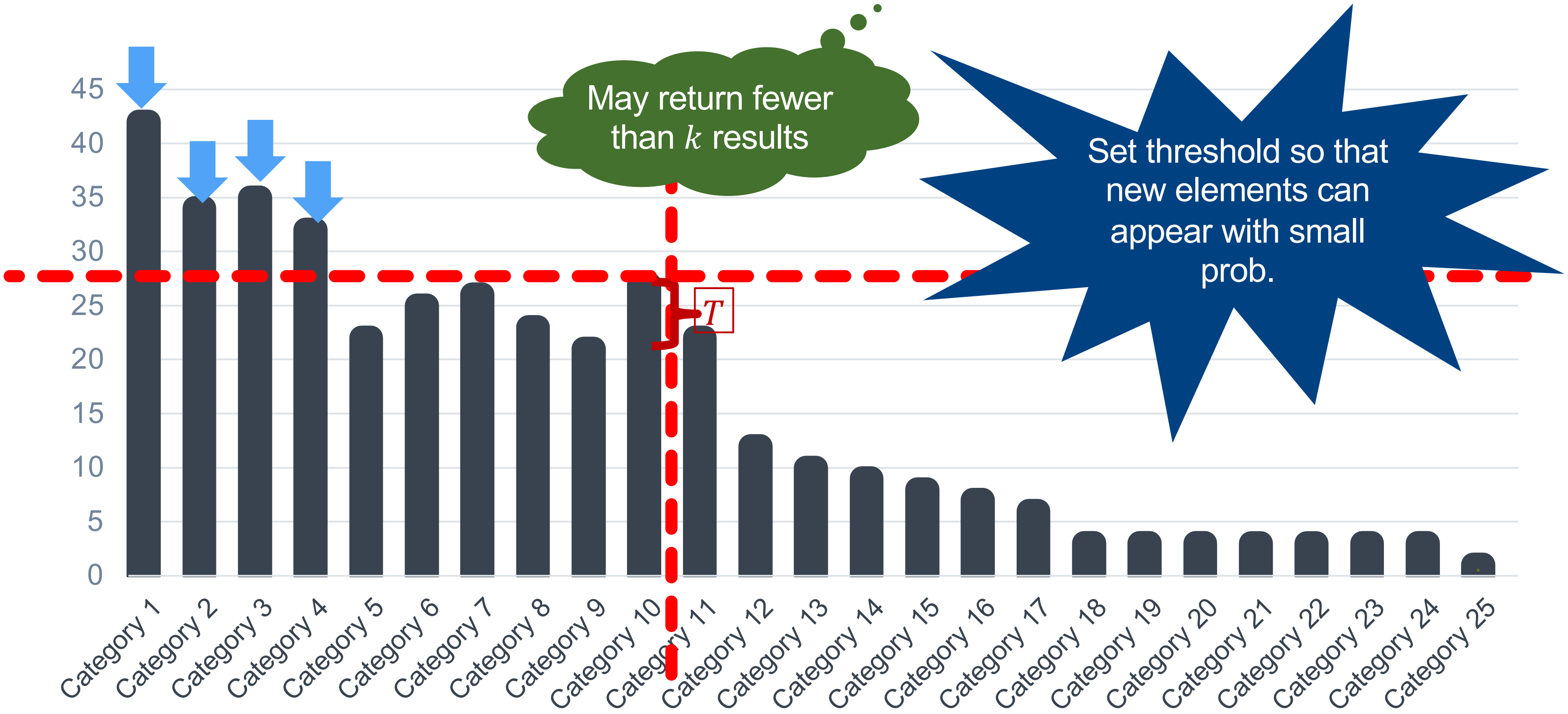
~~First Attempt~~



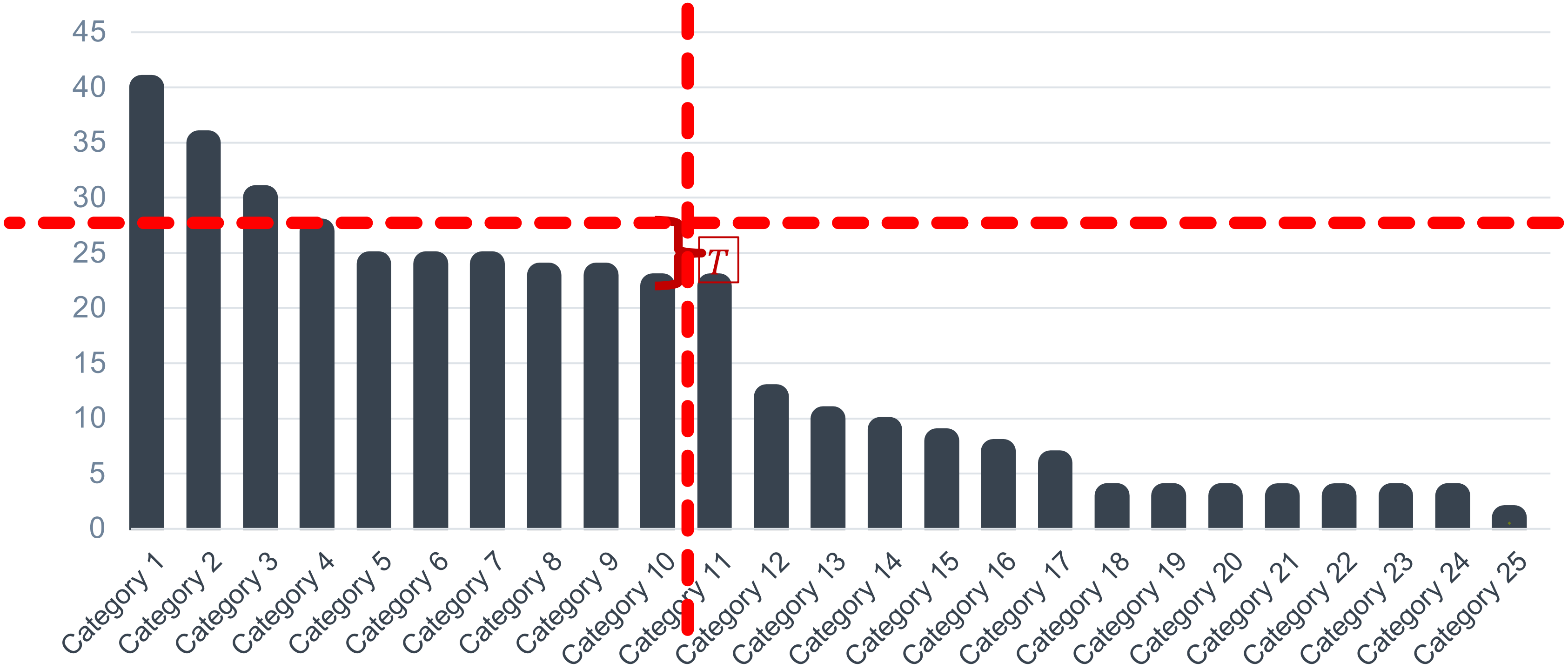
Second Attempt – Include a Threshold



Second Attempt – Include a Threshold



Second Attempt – Include a Threshold

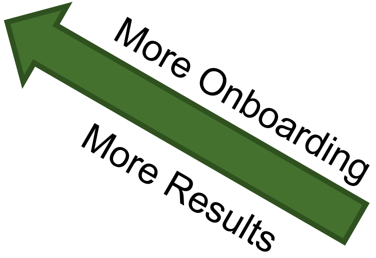


Algorithms for Private Data Analytics

DP Algorithms	**ℓ_0-Restricted Sensitivity**	**ℓ_0-Unrestricted Sensitivity**
Known Domain	Known Laplace [DMNS’06]	Known Gumbel [MT’07]

Algorithms for Private Data Analytics

DP Algorithms	ℓ_0-Restricted Sensitivity	ℓ_0-Unrestricted Sensitivity
Known Domain	Known Laplace [DMNS’06]	Known Gumbel [MT’07]
Unknown Domain	Unknown Laplace [Durfee, R’19]	Unknown Gumbel [Durfee, R’19]



Agenda

1. Overview of Differential Privacy
2. Overview of DP approach at LinkedIn
3. DP Top-*k* Algorithms
4. Privacy Budget Management
5. Deployments and Conclusion

What is the Overall Privacy Loss?

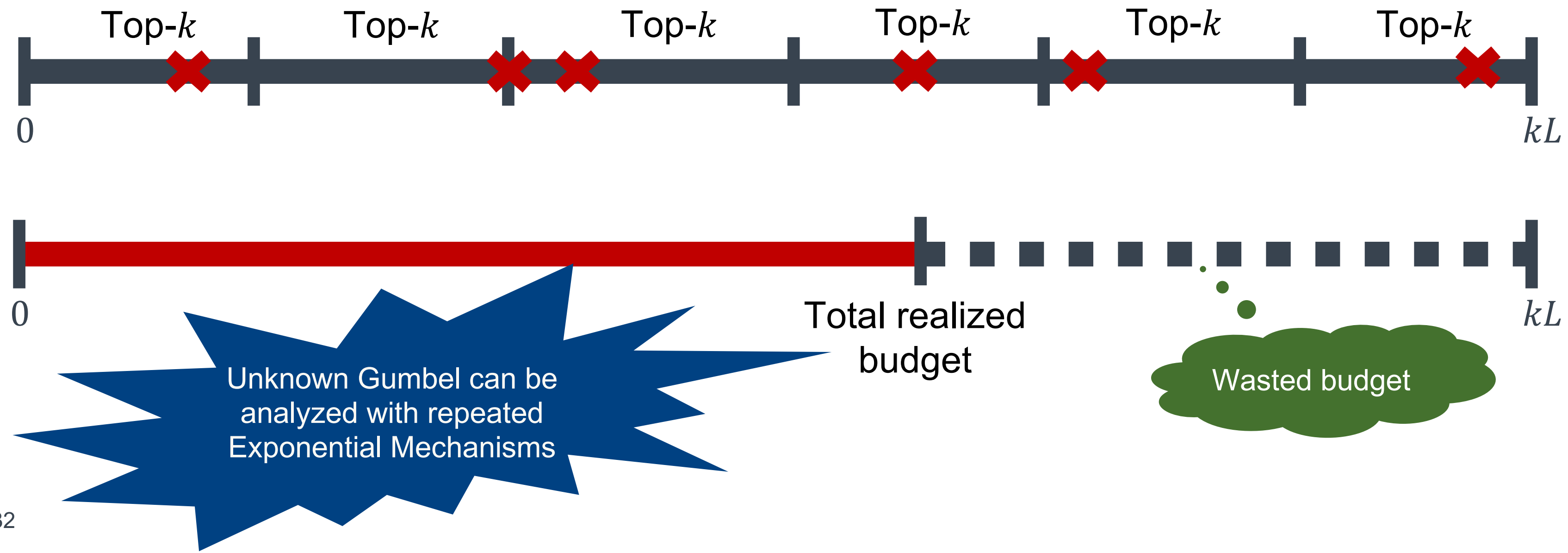
- Assume that the k in each top-k query is the same, at most L queries are allowed, and only using Unknown Gumbel.
  - Advanced Composition [Dwork, Rothblum, Vadhan '10]:

$$\left(\approx \epsilon\sqrt{Lk\log\frac{1}{\delta}}, (L+1)\delta\right)\text{-DP}$$

- Algorithm can give fewer results than what is asked.
  - Is it possible to only pay for what you get?

Pay-what-you-get Composition [DR'19]

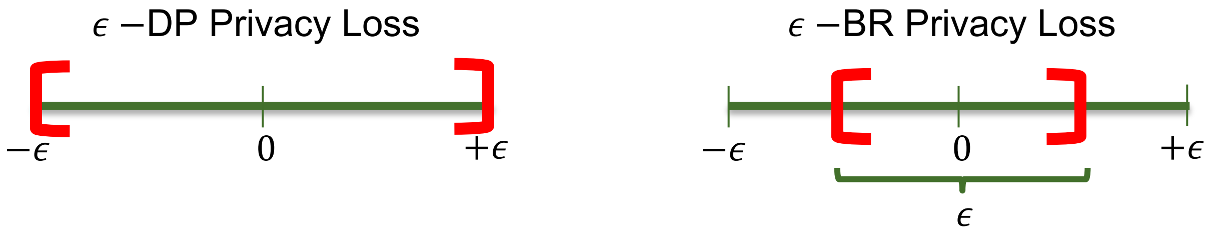
- Assume there is a global budget (ϵ_g, δ_g) with ϵ-parameter in each Unknown Gumbel



Bounded Range Mechanisms [DR'19]

- Can we improve general DP composition when we restrict to using Exp Mech only?
  - Don't rely on black box DP composition.
- **Defn**: A mechanism $M: X \to Y$ is ϵ-***Bounded Range*** (BR) if for any neighbors $x, x' \in X$ and outcomes $y, y' \in Y$, we have:

$$\frac{\Pr[M(x)=y]}{\Pr[M(x')=y]} \le e^{\epsilon} \frac{\Pr[M(x)=y']}{\Pr[M(x')=y']}$$



Bounded Range Mechanisms [DR'19]

- Can we improve general DP composition when we restrict to using Exp Mech only?
  - Don't rely on black box DP composition.
- **Defn**: A mechanism $M: X \to Y$ is ϵ-***Bounded Range*** (BR) if for any neighbors $x, x' \in X$ and outcomes $y, y' \in Y$, we have:

$$\frac{\Pr[M(x)=y]}{\Pr[M(x')=y]} \leq e^{\epsilon} \frac{\Pr[M(x)=y']}{\Pr[M(x')=y']}$$

- Note that ϵ-BR $\Longrightarrow$ ϵ-DP and ϵ-DP $\Longrightarrow$ 2ϵ-BR.
- **Lemma** [DR'19]: Exp Mech satisfies ϵ-BR and composing k^* of them gives:

$$\left(\approx \epsilon\sqrt{\frac{k^*}{2}\log\frac{1}{\delta}}, \delta\right)\text{-DP}$$

[Cesar, R '21]
ϵ-BR $\Longrightarrow$ $\epsilon^2/8$-zCDP

Optimal Comp of Exp Mech [Dong,Durfee,R ICML'20]

Optimal Differential Privacy Composition for Exponential Mechanisms and the Cost of Adaptivity

Jinshuo Dong*[1], David Durfee[2], and Ryan Rogers[2]

[1]Applied Mathematics and Computational Sciences, University of Pennsylvania
[2]Applied Research, LinkedIn

June 26, 2020

Abstract

Composition is one of the most important properties of differential privacy (DP), as it allows algorithm designers to build complex private algorithms from DP primitives. We consider precise composition bounds of the overall privacy loss for exponential mechanisms, one of the most fundamental class of mechanisms in DP. We give explicit formulations of the optimal privacy loss for both the adaptive and nonadaptive settings. For the nonadaptive setting in which each mechanism has the same privacy parameter, we give an efficiently computable formulation of the optimal privacy loss. Furthermore, we show that there is a difference in the privacy loss when the exponential mechanism is chosen adaptively versus nonadaptively. To our knowledge, it was previously unknown whether such a gap existed for any DP mechanisms with fixed privacy parameters, and we demonstrate the gap for a widely used class of mechanism in a natural setting. We then improve upon the best previously known upper bounds for adaptive composition of exponential mechanism with efficiently computable formulations and show the improvement.

Agenda

1 Overview of Differential Privacy

2 Overview of DP approach at LinkedIn

3 DP Top-*k* Algorithms

4 Privacy Budget Management

5 Deployments and Conclusion

Audience Engagement API

- API Product to provide insights on LinkedIn engagement content and audience data
- Provides information about member data to external marketing partners
- Built on top of ***Pinot*** for fast, real-time data analytics

Understanding the Task

- Advertiser can interact adaptively with the API
- Differencing attacks are a concern
- Want to provide both real-time analytics and privacy
- Queries are general top-*k* queries

Audience Engagement API

- For more information, see https://arxiv.org/abs/2002.05839

LinkedIn's Audience Engagements API: A Privacy Preserving Data Analytics System at Scale

Ryan Rogers[1], Subbu Subramaniam[1], Sean Peng[1], David Durfee[1], Seunghyun Lee[1], Santosh Kumar Kancha[1], Shraddha Sahay[1], and Parvez Ahammad[1]

[1]LinkedIn Corporation

November 17, 2020

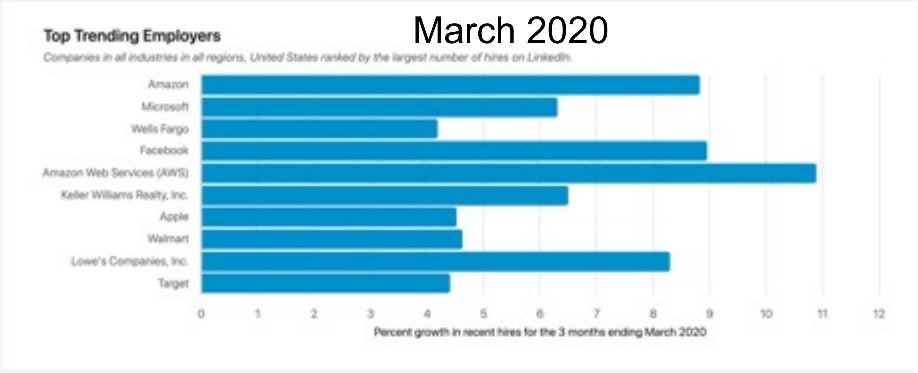
Abstract

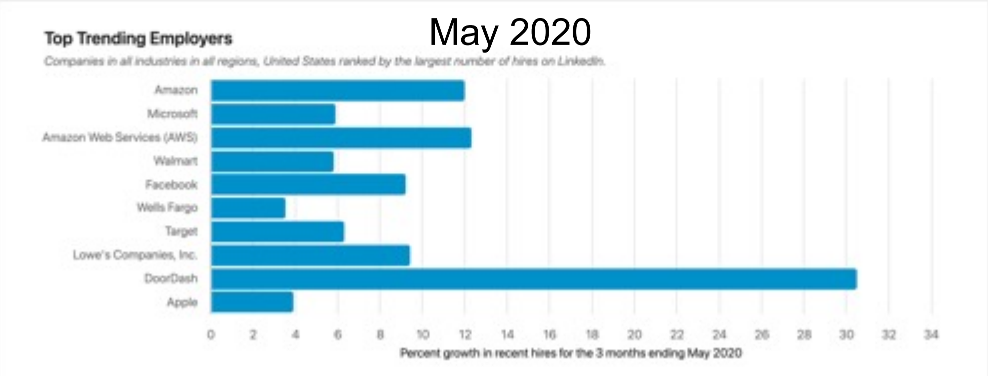
We present a privacy system that leverages differential privacy to protect LinkedIn members' data while also providing audience engagement insights to enable marketing analytics related applications. We detail the differentially private algorithms and other privacy safeguards used to provide results that can be used with existing real-time data analytics platforms, specifically with the open sourced Pinot system. Our privacy system provides user-level privacy guarantees. As part of our privacy system, we include a budget management service that enforces a strict differential privacy budget on the returned results to the analyst. This budget management service brings together the latest research in differential privacy into a product to maintain utility given a fixed differential privacy budget.

Labor Market Insights

- Tracking labor market trends is incredibly important especially during this pandemic.
- Leverage LinkedIn's Economic Graph to show these trends across different regions:
  - What employers are hiring the most?
  - What jobs are most in demand?
  - What are the top skills from these most in demand jobs?
- Global Skilling Event: https://news.microsoft.com/skills/

Labor Market Insights





Privacy considerations

In publishing these labor market insights from LinkedIn's Economic Graph, we wanted to provide accurate statistics while ensuring our members' privacy. We applied privacy techniques, such as differential privacy, to aggregate insights from our datasets without learning about specific individuals. Our differentially private algorithms have been deployed in other LinkedIn products used by our marketing partners, including the Audience Engagement API in LinkedIn Marketing Solutions.

graph.linkedin.com/insights/labor-market

Labor Market Insights

- For more information, see https://arxiv.org/abs/2010.13981

A Members First Approach to Enabling LinkedIn's Labor Market Insights at Scale

Ryan Rogers*, Adrian Rivera Cardoso*, Koray Mancuhan, Akash Kaura, Nikhil Gahlawat, Neha Jain, Paul Ko, Parvez Ahammad

LinkedIn Corporation

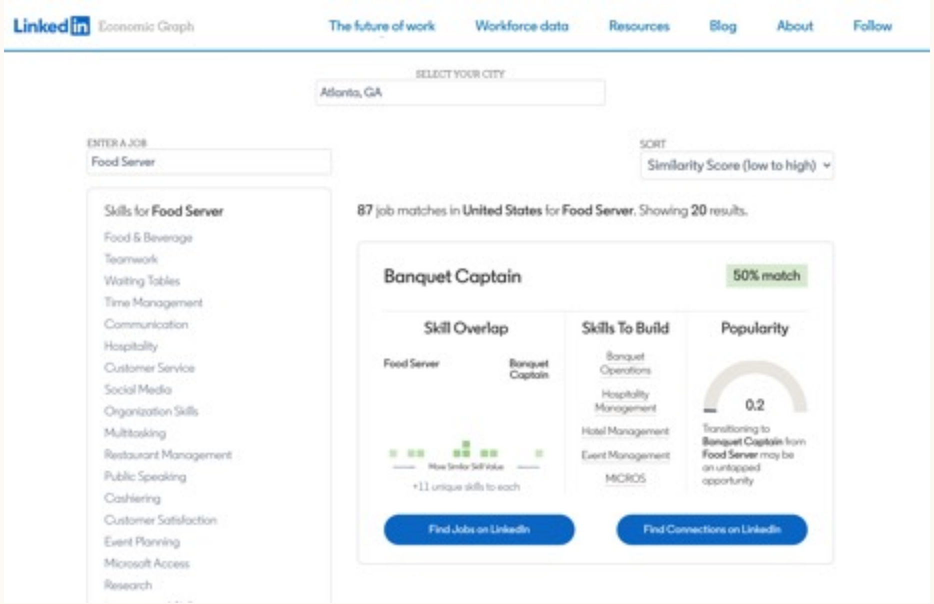
October 28, 2020

Abstract

We describe the privatization method used in reporting labor market insights from LinkedIn's Economic Graph, including the differentially private algorithms used to protect member's privacy. The reports in `https://graph.linkedin.com/insights/labor-market` show who are the top employers, as well as what are the top jobs and skills in a given country/region and industry. We hope this data will help governments and citizens track labor market trends during the COVID-19 pandemic while also protecting the privacy of our members.

Career Explorer

- https://linkedin.github.io/career-explorer/#explore
- Helps members discover new occupations based on the skills they have
- Helps members understand how the acquisition of new skills can lead to new opportunities.



Continual Observation

DP Algorithms	ℓ_0-Restricted Sensitivity	ℓ_0-Unrestricted Sensitivity
Known Domain	Binary Mechanism [Chan, Shi, Song ‘11 and Dwork, Naor, Pitassi, Rothblum, Yekhanin ’10]	Sparse Gumbel [Cardoso, R ’21]
Unknown Domain	Unknown Base [Cardoso, R ‘21]	Meta Algo

Concluding Remarks

- View privacy as a spectrum, not binary
- Can easily incorporate more privacy into systems that already are DP.
- How to rationalize large privacy loss (e.g. Census)?
  - There needs to be more open source attacks.
- Open Research Questions
  - How large is the gap between optimal adaptive vs non-adaptive composition for exponential mechanisms?
  - What about hardness results for some of these bounds?
  - How much can ordering impact the overall privacy loss?
    - See [Cesar, R'21]

Thank you!