Audience Engagement API: A Privacy Preserving Data Analytics System at Scale

Presenter: Ryan Rogers
Collaborators: Subbu Subramaniam, Sean Peng, David Durfee, Seunghyun Lee, Santosh Kumar Kancha, Shraddha Sahay, Parvez Ahammad
Agenda

1. Overview of Differential Privacy
2. Application
3. Overall Privacy System
Mission

Utilize data while protecting the privacy of users.
Reasons for Data Privacy

• We want to be "Members first"
• “Anonymized data isn’t” – Cynthia Dwork
  • 87% of U.S. is uniquely identified by (DOB, Gender, Zip)
• Potential attacks:
  • Reconstruction attacks
  • Differencing attacks
  • Membership inference attacks
Differential Privacy [Dwork, McSherry, Nissim, Smith ‘06]

\[ x: \text{Alice} \quad \text{Bob} \quad \text{Chuck} \quad \text{Doug} \quad \cdots \quad \text{Zulu} \rightarrow A \]

\[ \Pr[A(x) = y] \]

Outcome: \( y \)
Differential Privacy [Dwork, McSherry, Nissim, Smith ‘06]

Algorithm

\[ \Pr[A(x) = y] \]

Outcome: \( y \)
Differential Privacy [Dwork, McSherry, Nissim, Smith ‘06]

A randomized algorithm $A: \mathcal{D} \rightarrow \mathcal{Y}$ is $(\varepsilon, \delta) -$DP if for any neighboring data sets $x, x' \in \mathcal{D}$ and any outcome $S \subseteq \mathcal{Y}$ we have:

$$P(A(x) \in S) \leq e^\varepsilon P(A(x') \in S) + \delta$$

Privacy loss
Models and Deployments of Differential Privacy

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

Deployments:
- Microsoft
- Google
- Apple
- 2020 Census
- Microsoft Open Data DP Project with Harvard
- Google’s Open Source Library
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
• Questions that need to be addressed:
  • How much can a single user affect the outcome of these queries?
  • How many queries can the advertiser ask?
Existing Systems for Data Analytics

Apply DP algorithms to this result.
Overall Privacy System

Data

pinot

DP Algos

Application

Top-k

Marketing Partner
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?

Laplace Mechanism [DMNS06]: Add Noise to each count for DP

User can impact only one count

Aggregate Data
Query: Top-10 countries with certain skill set?

Releasing this histogram ensures $\epsilon$-DP.
Sensitivity of the Query

Query: Top-10 skills in the Bay Area?

Aggregate Data
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.

User can impact many counts!
Sensitivity of the Query

Query: Top-10 skills in the Bay Area?

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

Releasing only elements in top-$k$ (not their counts) ensures $k\varepsilon$-DP
### Known Algorithms for User Level DP

<table>
<thead>
<tr>
<th>$\Delta$-Restricted Sensitivity</th>
<th>Unrestricted Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algorithm: Laplace Mechanism</strong> [DMNS‘06]</td>
<td><strong>Algorithm: Exponential Mechanism</strong> [McSherry, Talwar ‘07]</td>
</tr>
</tbody>
</table>
Unknown Domain Setting

• Previous algorithms require knowing the full data domain
• They require adding noise to counts even when the true count is zero
• Typically, the domain is unknown or very large (e.g. all possible articles)
## Algorithms for User Level Privacy

<table>
<thead>
<tr>
<th>User Level DP Algorithms</th>
<th>Restricted Sensitivity</th>
<th>Unrestricted Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Domain</td>
<td>Laplace Mechanism [DMNS'06]</td>
<td>Exponential Mechanism [MT’07]</td>
</tr>
</tbody>
</table>
## Algorithms for User Level Privacy

<table>
<thead>
<tr>
<th>Known Domain</th>
<th>Known Domain</th>
<th>Restricted Sensitivity</th>
<th>Unrestricted Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace Mechanism [DMNS'06]</td>
<td>Exponential Mechanism [MT'07]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown Domain</td>
<td>Unknown Domain</td>
<td>UnkLap Mechanism [Durfee, R’19]</td>
<td>UnkExp Mechanism [Durfee, R’19]</td>
</tr>
</tbody>
</table>

Overall Privacy System

Alter the query: Top-2k

Return DP Top-k

Marketing Partner
Overall Privacy System

Data

pinot

DP Algos

Application

Marketing Partner

Budget Accesses
Overall Privacy System + Budget Manager
Overall Privacy System + Budget Manager

Privacy Budget Management

DP Algos

Application


Marketing Partner
Thank you!