How Well Does Privacy Compose?

Thomas Steinke

IBM Almaden

IPAM/UCLA, Los Angeles CA, 10 Jan. 2018
This Talk

• Composition!!
  • What is composition?
  • Why is it important?
  • Composition & high-dimensional (e.g. genetic) data

• Concentrated differential privacy
  • Reformulation of DP with tight composition
  • Understand & compare to \((\varepsilon, \delta)\)-DP
  • Useful analytical tool & valuable theoretical perspective
What is composition?

1:00 PM:

How many patients have DProsy?

631

2:00 PM:

How many patients have DProsy?

3:00 PM:

How many patients have DProsy?

632

Conclusion: Alice was diagnosed with DProsy!

Here composition led to a privacy compromise.
Fortunately, DP protects against attacks like this.
Why is composition important?

• Your data is held by many entities who do not coordinate on privacy. Information released by these entities can be combined to violate privacy.

• Allows complex algorithms to be built -- crucial for handling high-dimensional data (e.g. genetic data).
High-dimensional data & one-way marginals

<table>
<thead>
<tr>
<th></th>
<th>Dimension $d$</th>
<th>#individuals $n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>0 0 1 0 0 1 1 1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>1 0 1 1 0 1 1 0 0 1 1</td>
<td></td>
</tr>
<tr>
<td>Charles</td>
<td>1 0 1 0 1 1 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>0 1 0 1 1 1 0 1 0 0 1</td>
<td></td>
</tr>
</tbody>
</table>

- E.g. GWAS data. $d \approx 10^6$, $n \approx 1000$
- Key Question: For a given $n$ and $d$, how accurately can we release the one-way marginals of this dataset without imperiling privacy?
- I.e. how does privacy risk compose over the attributes?
Privacy risks of one-way marginals

- [Homer+08, Sankararaman+09, Bun+14, Dwork+15, etc.] showed that one-way marginals are susceptible to tracing.
- That is, given someone’s data and the one-way marginals of a case group, we can determine whether that person is in the group.
  - Surprising!
  - Led to privacy policy changes by NIH.
  - Works as long as \( d \gg n \).
  - Works even with approximate one-way marginals (but requires larger \( d \)).

### Table

<table>
<thead>
<tr>
<th></th>
<th>Dimension ( d )</th>
<th>#individuals ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>0 0 1 0 0 1 1 1 0 0 1</td>
<td>5 4 1 3 1 1 1 1  .5 .25 .75 .5 .5 1 .5 .5 0 .25 .75</td>
</tr>
<tr>
<td>Bob</td>
<td>1 0 1 1 0 1 1 0 0 1 1</td>
<td></td>
</tr>
<tr>
<td>Charles</td>
<td>1 0 1 0 1 1 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>0 1 0 1 1 1 0 1 0 0 1</td>
<td></td>
</tr>
</tbody>
</table>
This Talk

• Composition!!
  • What is composition?
  • Why is it important?
  • Composition & high-dimensional (e.g. genetic) data

• Concentrated differential privacy
  • Reformulation of DP with tight composition
  • Understand & compare to \((\varepsilon, \delta)\)-DP
  • Useful analytical tool & valuable theoretical perspective
Differential Privacy [DMNS06...]

**Definition**: A randomized algorithm $M$ is **differentially private** if, for all datasets $x$ and $x'$ differing only on one individual’s data,

$$\text{distribution}(M(x)) \approx \text{distribution}(M(x')).$$
Adding normally-distributed noise to all the values satisfies DP.

Does this give good privacy-utility tradeoff?
Quantifying Differential Privacy

Rényi divergence [R61]:

\[ D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left( \int_\Omega P(x)\alpha Q(x)^{1-\alpha} \, dx \right) \]

Interpolates between KL divergence (\(\alpha \to 1\)) & max divergence (\(\alpha \to \infty\)).

Exactly characterizes adding Normal noise.

• \(\epsilon\)-DP [DMNS06]:
  \[ \forall y \quad \mathbb{P}[M(x) = y] \leq e^\epsilon \mathbb{P}[M(x') = y] \]

• (\(\epsilon, \delta\))-DP [DKMMN06]:
  \[ \forall S \quad \mathbb{P}[M(x) \in S] \leq e^\epsilon \mathbb{P}[M(x') \in S] + \delta \]

\(\rho\)-CDP [DR16, BS16, M17, BDRS17]:
\[ \forall \alpha \in (1, \infty) \quad D_\alpha(M(x)||M(x')) \leq \rho \alpha \]
Why do we need a new definition?

• “Pure” $\varepsilon$-DP gives poor composition bounds
  • Gets “hung up on” very low probability events.
  • Composition is quadratically worse than it “should” be.

• “Approximate” ($\varepsilon, \delta$)-DP gives messy composition bounds
  • Can ignore events with probability $\leq \delta$.
  • Doesn’t sharply capture what’s going on.
  • Superfluous $\log(1/\delta)$ factors in composition analysis.

• Concentrated DP gives sharp composition bounds!

Composition & privacy loss are natural phenomena

$\delta = \mathbb{P}[\text{bad event}]$ needs to be cryptographically small.
Composition for CDP

**Theorem** (CDP composition [DR16,BS16]):
Let $M_1, \ldots, M_k$ be randomized algorithms. Suppose each $M_i$ is $\rho_i$-CDP. Then combining the outputs of $M_1, \ldots, M_k$ satisfies $(\rho' = \sum_i \rho_i)$-CDP.

- Simple and optimal (in contrast to $\varepsilon$-DP and $(\varepsilon, \delta)$-DP).

- Cf. Optimal $(\varepsilon, \delta)$-DP composition [KOV15,MV16]:
\[
\sum_{s \subseteq [k]} \max \left\{ 0, e^{\sum_{i \in s} \varepsilon_i - e^{\varepsilon'} + \sum_{i \in [k] \setminus s} \varepsilon_i} \right\} \leq 1
\]

- Computing optimal composition exactly is #P-hard [MV16]!!
Noisy one-way marginals

More sophisticated algorithms possible!

E.g.,

- Only identify the \( k \) most significant attributes.
- Attributes are sparse/structured.
- Exploit data distribution.

\[
\frac{1}{k} \sum_{j=1}^{k} \left| a_j - \frac{1}{n} \sum_{i=1}^{n} q_j(x_i) \right| \leq \frac{1}{100}
\]

<table>
<thead>
<tr>
<th>Individuals</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way</td>
<td>.5 .25 .75 .5 .5 1 .5 .5 0 .25 .75</td>
</tr>
<tr>
<td>Noisy marginals</td>
<td>.6 .1 .8 .6 .4 1 .4 .5 .1 .4 .9</td>
</tr>
</tbody>
</table>

Adding \( N(0, \sigma^2) \) to each marginal achieves \( \left( \rho = \frac{d}{2\sigma^2 n^2} \right) \)-CDP.

Sharp tradeoff between privacy \( \rho \), dimension \( d \), accuracy \( \sigma \), and number of individuals \( n \).

E.g. \( \rho = 0.5, \sigma = 0.1, d = 10^4 \) requires \( n = \frac{\sqrt{d}}{\sigma \sqrt{2\rho}} = 1000 \).
Composition Comparison

- **Pure $\epsilon$-DP:** $\epsilon' = \sum_i \epsilon_i$.  
  - Can approximate $d = \Theta(\epsilon n)$ one-way marginals to constant accuracy with $\epsilon$-DP.

- **Approx. $(\epsilon, \delta)$-DP:** $\epsilon' = O\left(\sqrt{\log(1/\delta)} \sum_{i=1}^k \epsilon_i^2\right)$, $\delta' = O\left(\sum_{i=1}^k \delta_i\right)$.  
  - #P-hard to compute optimal composition exactly.
  - Can approximate $d = \Theta(\epsilon^2 n^2 / \log(1/\delta))$ marginals to constant accuracy with $(\epsilon, \delta)$-DP.

- **$\rho$-CDP:** $\rho' = \sum_i \rho_i$.  
  - Can approximate $d = \Theta(\rho n^2)$ marginals to constant accuracy with $\rho$-CDP.
This Talk

• Composition!!
  • What is composition?
  • Why is it important?
  • Composition & high-dimensional (e.g. genetic) data

• Concentrated differential privacy
  • Reformulation of DP with tight composition
  • Understand & compare to \((\varepsilon, \delta)\)-DP
  • Useful analytical tool & valuable theoretical perspective
Concentrated DP [DR16, BS16, BDRS17]

**Definition [BS16]:** A randomized algorithm $M$ is $\rho$-CDP if, for all datasets $x$ and $x'$ differing only on one individual’s data,

$$\forall \alpha \in (1, \infty) \quad D_\alpha(M(x)||M(x')) \leq \rho \alpha$$

Rényi divergence [R61]:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \left( \int_{\Omega} P(x)^\alpha Q(x)^{1-\alpha} \, dx \right)$$

Interpolates between KL divergence ($\alpha \to 1$) & max divergence ($\alpha \to \infty$). Exactly characterizes Gaussian mechanism.
CDP versus \((\varepsilon, \delta)\)-DP

**Theorem [BS16]:** \(\forall \rho, \delta > 0\)

\[
\sqrt{2\rho} \text{-DP} \implies \rho \text{-CDP} \implies (\rho + 2\sqrt{\rho} \cdot \log(1/\delta), \delta) \text{-DP}.
\]

- CDP is a relaxation of pure \(\varepsilon\)-DP.
  - Relaxation is strict. E.g. Gaussian mechanism satisfies CDP, but not pure DP.
- CDP is roughly equivalent to approx. \((\varepsilon, \delta)\)-DP with this \(\forall \delta\) quantification.
  - However, there are algorithms that satisfy approx. DP, but not CDP.
- Think of CDP as being intermediate between pure and approx. DP.
- Open-ended question: How to interpret \(\rho\)?
CDP versus $(\varepsilon, \delta)$-DP

$\varepsilon$-DP

$\rho$-CDP

Smooth control of bad events: For $\rho$-CDP
\[ \forall t > \rho \quad \mathbb{P}[\text{PrivLoss} > t] \leq e^{-(t-\rho)^2/4\rho} \]

CDP does not allow arbitrarily bad events!

PrivLoss$(y) = \log \left( \frac{\mathbb{P}[M(x) = y]}{\mathbb{P}[M(x') = y]} \right)$

PrivLoss = PrivLoss$(M(x))$

With probability $\delta$ arbitrarily bad things can happen!
Bounding Bad Events with CDP [M17]

**Proposition [M17]:** If $M$ is $\rho$-CDP and $x, x'$ are neighbouring inputs, then
\[ \forall S \forall \alpha \quad \mathbb{P}[M(x) \in S] \leq e^{(\alpha-1)\rho} \cdot (\mathbb{P}[M(x') \in S])^{1-1/\alpha} \]

E.g.:

- Suppose, when not in dataset, bad event happens with
  \[ \mathbb{P}[M(x') \in S] \leq 10^{-10} \]

- If $M$ is $\rho$-CDP, then, when in data, bad event happens with
  \[
  \alpha = 2: \quad \mathbb{P}[M(x) \in S] \leq e^\rho \sqrt{10^{-10}} \\
  \alpha = 10: \quad \mathbb{P}[M(x) \in S] \leq e^{9\rho} 10^{-9}
  \]
This Talk

• Composition!!
  • What is composition?
  • Why is it important?
  • Composition & high-dimensional (e.g. genetic) data

• Concentrated differential privacy
  • Reformulation of DP with tight composition
  • Understand & compare to $\left( \varepsilon, \delta \right)$-DP
  • Useful analytical tool & valuable theoretical perspective
What can we do with CDP?

\((\varepsilon, \delta)\)-DP

\(\rho\)-CDP

\(\varepsilon\)-DP

- Basic composition,
- Laplace mechanism,
- Exponential mechanism,
- Randomized response,
- Sparse vector,
- BLR mechanism

Advanced composition,
Gaussian mechanism,
Private multiplicative weights,
Projection mechanism

Propose-Test-Release framework,
Smooth sensitivity
Truncated CDP [BDRS17]

Definition [BDRS17]: A randomized algorithm $M$ is $(\rho, \omega)$-tCDP if, for all datasets $x$ and $x'$ differing only on one individual’s data,

$$\forall \alpha \in (1, \omega) \quad D_\alpha(M(x)||M(x')) \leq \rho \alpha$$

- $\omega = \infty$ recovers $\rho$-CDP.
- Similar to Rényi DP [M17] – consider single $\alpha$, rather than interval.
- Extends CDP to permit analogs of key algorithmic techniques.
  - Analog of propose-test-release framework [DL09].
  - Smooth sensitivity [NRS07].
  - Privacy amplification by subsampling.
This Talk

• Composition!!
  • What is composition?
  • Why is it important?
  • Composition & high-dimensional (e.g. genetic) data

• Concentrated differential privacy
  • Reformulation of DP with tight composition
  • Understand & compare to $(\varepsilon, \delta)$-DP
  • Useful analytical tool & valuable theoretical perspective
Separation: $\varepsilon$-DP $\neq$ CDP $\neq$ $(\varepsilon, \delta)$-DP

Point Queries/Histograms:
Input: $x_1, \ldots, x_n \in \Omega$.
Output: For each $z \in \Omega$, return $\text{freq}(z) = |\{i : x_i = z\}| \pm \frac{n}{100}$.

- Possible with $\varepsilon$-DP iff $n = \Theta(\log |\Omega| / \varepsilon)$.
- Possible with $(\varepsilon, \delta)$-DP iff $n = \Theta(\log(1/\delta) / \varepsilon)$.
- Possible with $\rho$-CDP iff $n = \Theta(\sqrt{\log |\Omega|} / \rho)$.
  - Upper bound: Add noise from $\mathcal{N}(0, \frac{1}{\rho})$ to each frequency.

Quadratic separation

“Infinite” separation
CDP & Mutual Information

Theorem [BS16]: If $M$ is $\rho$-CDP and $X$ is a random input consisting of $n$ individuals, then

$$I(X; M(X)) \leq \rho \cdot n^2$$

• Follows from group privacy property of CDP.

• Idea: If $M$ is accurately answers many queries, then mutual information must be high.

• $\implies$ Lower bound on $n$. 
Separation: \( \varepsilon\text{-DP} \neq \text{CDP} \neq (\varepsilon, \delta)\text{-DP} \)

**Point Queries/Histograms:**
- **Input:** \( x_1, \ldots, x_n \in \Omega \).
- **Output:** For each \( z \in \Omega \), return \( \text{freq}(z) = |\{i : x_i = z\}| \pm \frac{n}{100} \).

- Possible with \( \varepsilon\text{-DP} \) iff \( n = \Theta(\log |\Omega| / \varepsilon) \).

- Possible with \((\varepsilon, \delta)\text{-DP} \) iff \( n = \Theta(\log(1/\delta) / \varepsilon) \).

- Possible with \( \rho\text{-CDP} \) iff \( n = \Theta(\sqrt{\log |\Omega| / \rho}) \).
- Upper bound: Add noise from \( \mathcal{N}(0, \frac{1}{\rho}) \) to each frequency.

 Quadratic separation

“Infinite” separation
Optimal CDP Algorithm [BBNS17] (see poster)

Linear Query Release problem:

\[
\begin{array}{c|c}
  x_1 & \vdots \\
  x_2 & \\
  \vdots & \\
  x_n & \\
\end{array}
\xrightarrow{M} \text{conv}(X) \subseteq \mathbb{R}^d
\]

Accuracy Goal:

\[
M(x) \approx _\alpha \frac{1}{n} \sum_{i=1}^{n} q(x_i)
\]
Optimal CDP Algorithm [BBNS17] (see poster)

Linear Query Release problem:

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{bmatrix} \rightarrow M \rightarrow \text{conv}(X) \subseteq \mathbb{R}^d
\]

Average Squared Accuracy Goal:

\[
\mathbb{E} \left[ \frac{1}{\text{diam}(X)^2} \left\| M(x) - \frac{1}{n} \sum_{i=1}^{n} q(x_i) \right\|_2^2 \right] \leq \alpha^2
\]

[BBNS17]: New CDP algorithm that is provably optimal for \( \alpha = \Omega(1) \).
Optimal CDP Algorithm [BBNS17] (see poster)

Let $X = \text{range}(q) \subseteq \mathbb{R}^d$ be the set of possible answers.

**Definition** (Covering number): Let $N(X, \gamma)$ be the smallest number of $\gamma$-balls whose union covers $X$.

**Algorithm** [BBNS17]: $\rho$-CDP $\alpha$-accurate algorithm for $X$ as long as

$$n \geq O\left(\frac{1}{\alpha^2} \sqrt{\frac{\log N(X, \alpha \cdot \text{diam}(X)/2)}{\rho}}\right)$$

**Average Squared Accuracy Goal:**

$$\mathbb{E}\left[\frac{1}{\text{diam}(X)^2} \left\|M(x) - \frac{1}{n} \sum_{i=1}^{n} q(x_i)\right\|_2^2\right] \leq \alpha^2$$

**Lower Bound** [BBNS17]: Need

$$n \geq \Omega\left(\sqrt{\log N(X, 3\alpha \cdot \text{diam}(X))/\rho}\right).$$

Based on Projection Mechanism [NTZ13]
What can’t we do with CDP?

$$\mathcal{C}(\epsilon, \delta)$$-DP

$$\rho$$-CDP

$$\epsilon$$-DP

- Basic composition,
- Laplace mechanism,
- Exponential mechanism,
- Randomized response,
- Sparse vector,
- BLR mechanism,
- Subsampling*

Advanced composition,
Gaussian mechanism,
Private multiplicative weights,
Projection mechanism

Propose-Test-Release framework,
Smooth sensitivity,
Privacy amplification by subsampling
Truncated CDP [BDRS17]

Definition [BDRS17]: A randomized algorithm $M$ is $(\rho, \omega)$-tCDP if, for all datasets $x$ and $x'$ differing only on one individual’s data,

$$\forall \alpha \in (1, \omega) \quad D_\alpha (M(x)||M(x')) \leq \rho \alpha$$

- $\omega = \infty$ recovers $\rho$-CDP.
- Similar to Rényi DP [M17] – consider single $\alpha$, rather than interval.
- Extends CDP to permit analogs of key algorithmic techniques.
  - Propose-test-release framework.
  - Smooth sensitivity.
  - Privacy amplification by subsampling.
Separation: $\varepsilon$-DP $\neq$ CDP $\neq$ tCDP

Point Queries/Histograms:
Input: $x_1, \ldots, x_n \in \Omega$.
Output: For each $z \in \Omega$, return $freq(z) = |\{i : x_i = z\}| \pm \frac{n}{100}$.

- Possible with $\varepsilon$-DP iff $n = \Theta(\log |\Omega| / \varepsilon)$.
- Possible with $(\varepsilon, \delta)$-DP iff $n = \Theta(\log(1/\delta) / \varepsilon)$.
- Possible with $\rho$-CDP iff $n = \Theta(\sqrt{\log |\Omega| / \rho})$.
- Possible with $(\rho, \omega)$-tCDP iff $n = \Theta(\omega \cdot \log \log |\Omega|)$ (for $\omega \ll \sqrt{\log |\Omega| / \rho}$).

Related to propose-test-release.

How do these compare in practice?
Subsampling

\[
M: X^n \rightarrow Y \\
M': X^N \rightarrow Y \\
s = \frac{n}{N} \ll 1
\]

- If \( M \) is \((\varepsilon, \delta)\)-DP, then \( M' \) is \((\approx s \cdot \varepsilon, s \cdot \delta)\)-DP.

- If \( M \) is \( \rho \)-CDP, then \( M' \) is \( \rho \)-CDP.  
  No gain in parameters!

- If \( M \) is \((\rho, \omega)\)-tCDP, then \( M' \) is \((\approx s^2 \cdot \rho, \Omega(\min\{\omega, \log(1/s)/\rho\}))\)-tCDP.
What can’t we do with tCDP?

(\( \varepsilon, \delta \))-DP

(\( \rho, \omega \))-tCDP

\( \rho \)-CDP

\( \varepsilon \)-DP

- Basic composition,
- Laplace mechanism,
- Exponential mechanism,
- Randomized response,
- Sparse vector,
- BLR mechanism,
- Subsampling*

Advanced composition,
Gaussian mechanism,
Private multiplicative weights,
Projection mechanism

Propose-Test-Release framework,
Smooth sensitivity,
Privacy amplification by subsampling

???