Meet The Sims: Generating Private Synthetic Data

Zexi Song

Mentor: Dr. Anand Sarwate

DIMACS-REU

June 5, 2015

Overview

Data and Privacy

We are gathering more and more digital information about people. As the data becomes more and more detailed, it is easier to "de-anonymize" or re-identify individuals from their data alone. However, for most applications all we want is population-level statistics. Allowing others to re-identify individuals from data availabe is potentially harmful.

Data and Privacy

- Medical Records
- Genetic Data
- Financial Information
- Search Logs

Two Major Approaches

- Limit disclosures through limiting the questions asked and providing only approximate data
- Publish "anonymized" data

Simply Anonymizing Data Is Unsafe

Name	**Has Diabetes (X)**
Ross	1
Monica	1
Joey	0
Phoebe	0
Chandler	1

Table: Medical records of "Friends"

Simply Anonymizing Data Is Unsafe

Netflix anonymized movie records

Simply Anonymizing Data Is Unsafe

Netflix anonymized movie records

IMDb side information

Simply Anonymizing Data Is Unsafe

Netflix anonymized movie records

IMDb side information

Privacy Leakage!

Our Approach

Our goal is to publish synthetic data: we learn patterns and correlations in the population in a privacy-preserving way and then generate fake individuals/records so that the population statistics are similar.

Formally Define Privacy

Intuition:

- A condition on the data release mechanism not on the data set.
- The presence or absence of an individual would not effect the final output of a statistical query significantly.

Formally Define Privacy

Formula:
Let M be a randomized data release algorithm. We say that M is $\epsilon-$differentially private if for all D_1, D_2 that differ in only one element and any event S,

$$Pr(M(D_1) \in S) \leq e^{\epsilon} \times Pr(M(D_2) \in S)$$

Formally Define Privacy

Formula:
Let M be a randomized data release algorithm. We say that M is $\epsilon-$differentially private if for all D_1, D_2 that differ in only one element and any event S,

$$Pr(M(D_1) \in S) \leq e^{\epsilon} \times Pr(M(D_2) \in S)$$

If $M(D_1)$ and $M(D_2)$ have probability mass functions P_1, P_2 respectively, then

$$\left| log \frac{P_1(M(D_1) = x)}{P_2(M(D_2) = x)} \right| \leq \epsilon$$

An Example of Differential Private Algorithm

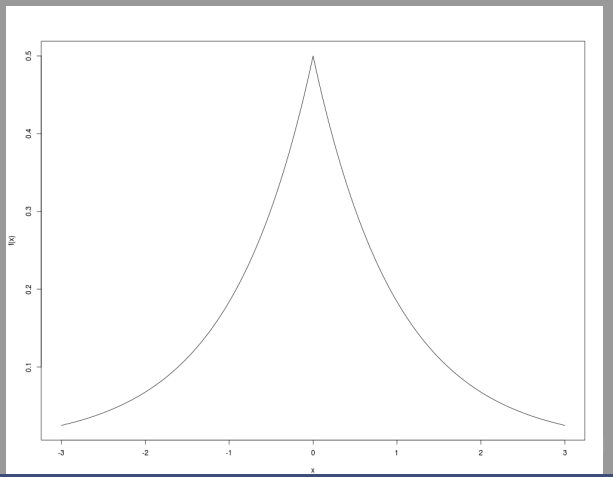
Data points $x_1, x_2, ..., x_n$ where $x_i \in [0, 1], \forall i$
Query for the mean.

An Example of Differential Private Algorithm

Data points $x_1, x_2, ..., x_n$ where $x_i \in [0, 1], \forall i$
Query for the mean.

Mechanism:

1. Calculate the true mean: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
2. Output:
$\overline{x} + \frac{1}{n\epsilon} Z$, where $Z \sim$ Lap $(0, 1)$



Recall the Process

- Given a dataset, study the patterns and correlations in the population.
- Build a probabilistic graphical model that yields similar stats.
- Generate synthetic data.

Preserve Privacy!

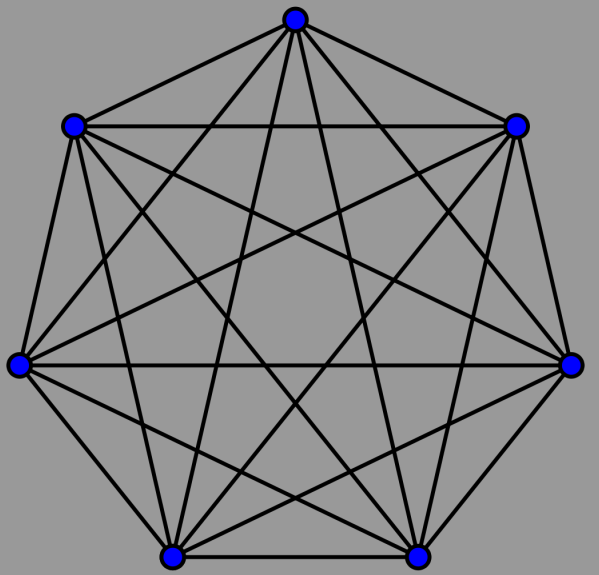
Graphical Model

	F_1	F_2	...	F_m
x_1	$x_1(1)$	$x_1(2)$	...	$x_1(m)$
x_2	$x_2(1)$	$x_2(2)$	...	$x_2(m)$
...	...	...	...	...
x_n	$x_n(1)$	$x_n(2)$	...	$x_n(m)$

Need to approximate
$P(F_1, F_2, ..., F_m)$

Graphical Model

	F_1	F_2	...	F_m
x_1	$x_1(1)$	$x_1(2)$	...	$x_1(m)$
x_2	$x_2(1)$	$x_2(2)$	...	$x_2(m)$
...	...	...	...	...
x_n	$x_n(1)$	$x_n(2)$	...	$x_n(m)$



Simplify the Graph

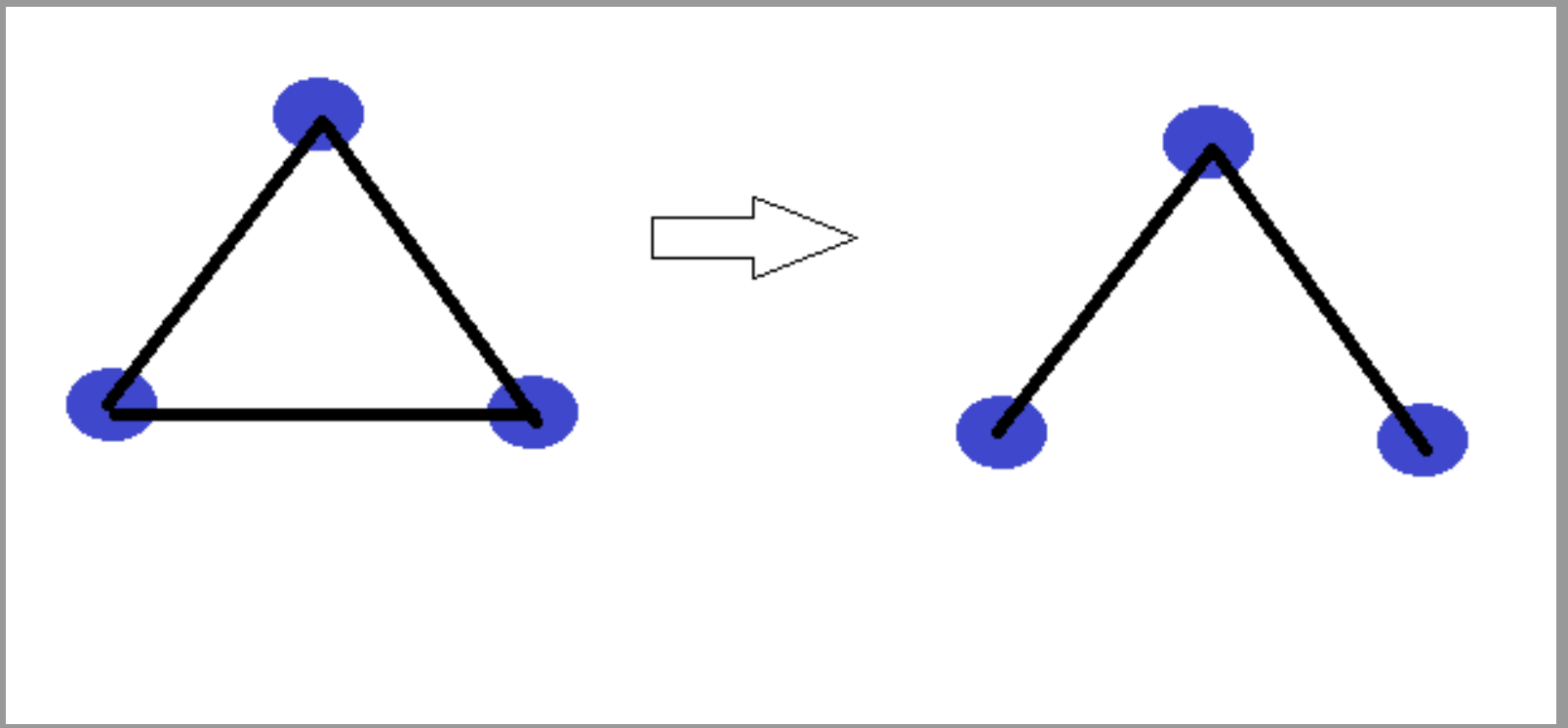
Independence:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

Conditional Independence:

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z)$$

Simplify the Graph



We want a tree structure!

Chow-Liu Algorithm

First, for every pair of features F_i, F_j construct the empirical joint distribution function,

$$\hat{P}(F_i = a, F_j = b) = \frac{1}{n} \sum_{k=1}^{n} 1(x_k(i) = a, x_k(j) = b)$$

Chow-Liu Algorithm

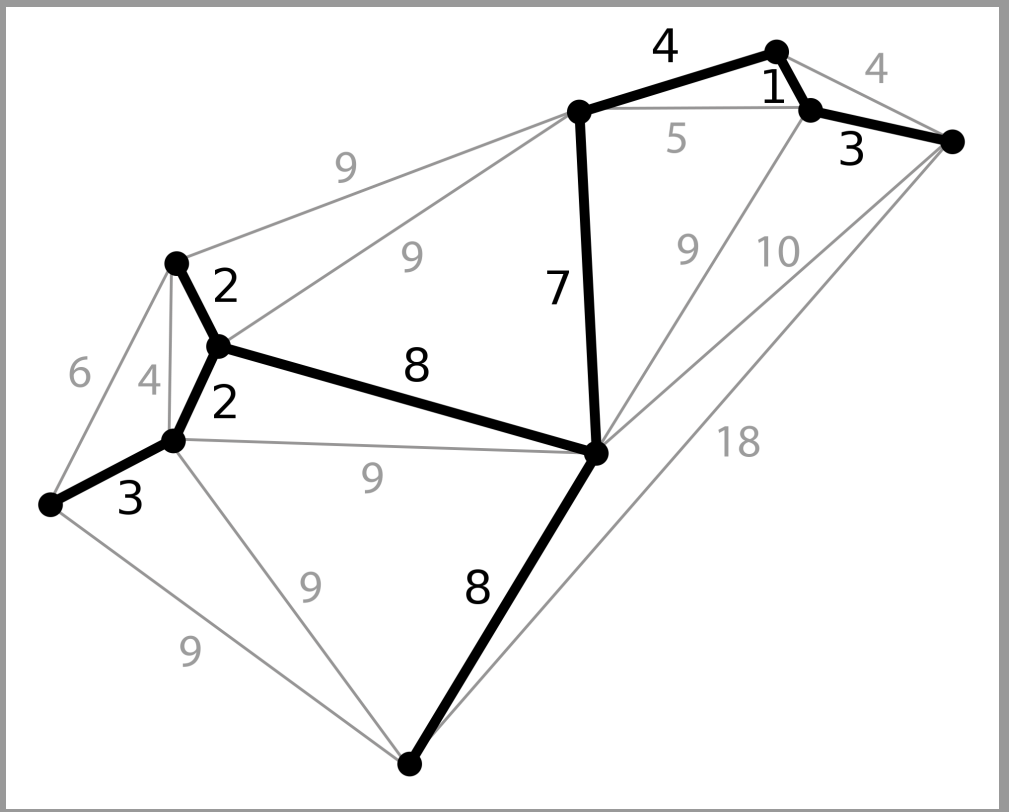
Then calculate the Mutual Information:

$$I_{\hat{P}}(F_i; F_j) = \sum_{a,b} \hat{P}(F_i = a, F_j = b) \cdot log \frac{\hat{P}(F_i = a, F_j = b)}{\hat{P}(F_i = a)\hat{P}(F_j = b)}$$

Chow-Liu Algorithm

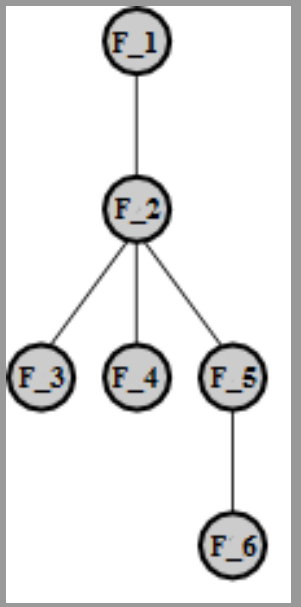
For each pair of F_i, F_j assign weight $w_{ij} = I_{\hat{P}}(F_i; F_j)$

Based on these weights, build a maximum spanning tree (MST).



Chow-Liu Algorithm

Now we can reconstruct the system based on the Chow-Liu Tree, Example,



Then the joint distribution can be approximated by,

$$P(F_1, F_2, F_3, F_4, F_5, F_6)$$

$$\approx P(F_6 \mid F_5)P(F_5 \mid F_2)P(F_4 \mid F_2)P(F_3 \mid F_2)P(F_2 \mid F_1)P(F_1)$$

Short-term Tasks

- Implement Chow-Liu algorithm
- Refine Chow-Liu algorithm:
  How to determine root node
  Forest model
  CLThres model
- Test different strategies of persevering privacy

References


Chaudhuri and Sarwate
Differential Privacy and Machine Learning


Dwork and Roth (2014)
The Algorithmic Foundations of Differential Privacy
Foundations and Trends in Theoretical Computer Science Vol. 9, 3–4.


Koller and Friedman(2009)
Probabilistic Graphical Models

Thank You