Webinar

NIST Differential Privacy Synthetic Data Challenge

Christine Task

Terese Manley

November 13, 2018

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

Why the Challenge?

- The Public Safety Communications Research Division (PSCR) of the National Institute of Standards and Technology (NIST) is sponsoring this exciting data science competition to help advance research for public safety communications technologies for America's First Responders
- As first responders utilize more advanced communications technology, there are opportunities to use data analytics to gain insights from public safety data, inform decision-making and increase safety.
- **But… we must assure data privacy.**

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow…today

2

Past & Potential Solutions for De-identification

- Current de-identification approaches, such as field suppression, anonymization and k-anonymity, can still be vulnerable to re-identification attacks, and may have a poorly understood trade-off between utility and privacy.

- Algorithms that satisfy the Differential Privacy guarantee provide privacy protection that's robust against re-identification attacks, independent of an attacker's background knowledge. They use randomized mechanisms and provide a tunable trade-off between utility and privacy.

- **Differentially Private Synthetic Data Generation** is a mathematical theory, and set of computational techniques, that provide a method of de-identifying data sets—under the restriction of a quantifiable level of privacy loss. It is a rapidly growing field in computer science.

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

What NIST wants to achieve for the Future

- The purpose of this empirical coding competition is to discover new tools to generate differentially private synthetic data while keeping the integrity of the original data distribution for analysis

- Developments coming out of this competition will drive major advances in the practical applications of differential privacy for applications such as public safety.

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

What do we mean by Privacy?

Privacy-preserving data-mining algorithms allow trusted data-owners to release useful, aggregate information about their data-sets (such as common user behavior patterns) while at the same time protecting individual-level information.

Intuitively, the concept of making large patterns visible while protecting small details makes sense. You just 'blur' things a bit:

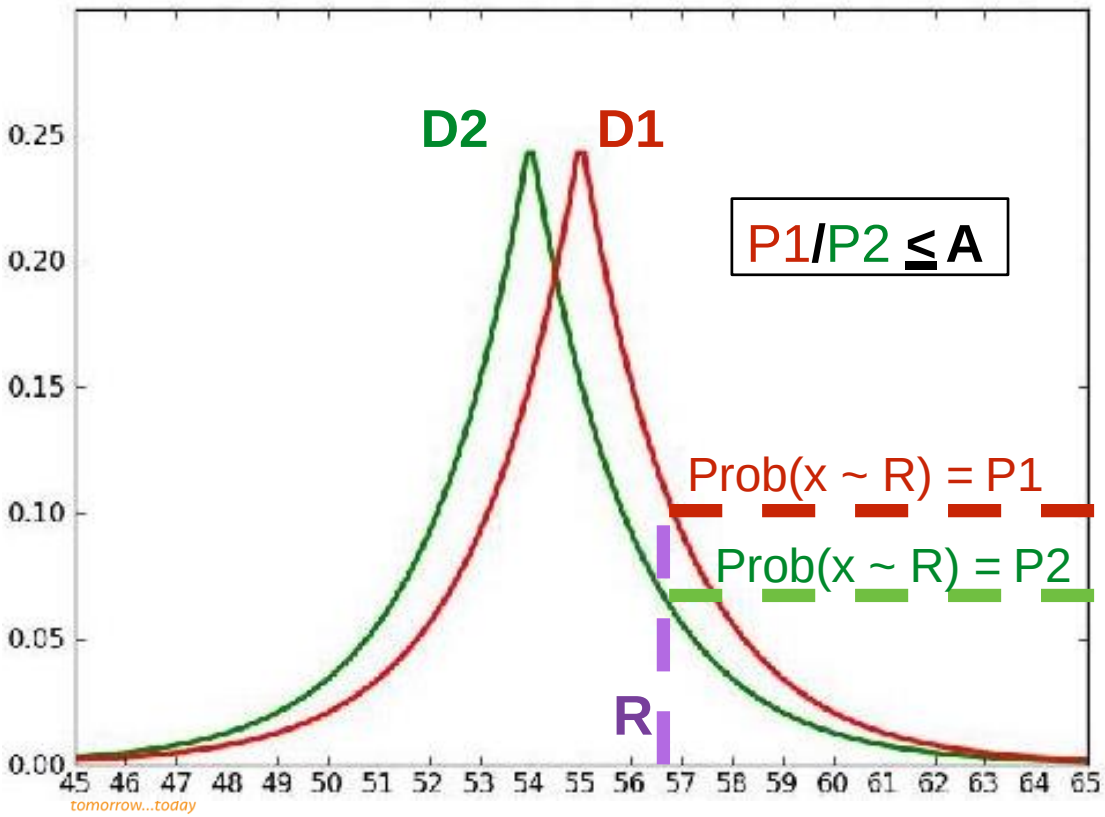
http://fryeart1.weebly.com/journals.html

If we refine this idea into a mathematically formal definition, we can create a standard for individual privacy.

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

This is the Laplace Mechanism

Adding **Laplacian noise** to the true answer means that the distribution of possible results from any data set overlaps heavily with the distribution of results from its neighbors.



$$Prob(R = x \mid D \text{ is the true world}) = \frac{\varepsilon}{2\Delta F} e^{\frac{|x - F(D)|\varepsilon}{\Delta F}}$$

Rcan be any publishable data- structure (not just a simple count), usually formally represented as a k-dimensional vector. It encapsulates all privatized published data (including repeated queries).

A provides a formal measure of individual privacy. The larger A is, the farther apart P1 and P2 can be, the less overlap between realities is required.

Challenge Approach

Stage 1

- Awarded $40k to competitors who wrote concept papers identifying new approaches to de-identification using differential privacy

Stage 2

- Competitors will design and implement synthetic data generation algorithms that satisfy differential privacy
- References and tutorials are available for data scientists who wish to learn about differential privacy
- Contest will use datasets based on San Francisco's Fire Dept Incident Data and Census Bureau American Community Survey Data
- Contest entails a sequence of 3 marathon matches
  - 5 weeks in length starting every 2 months
- Prize winners will have their solutions carefully reviewed and validated to ensure they satisfy differential privacy before prizes are awarded
- Total prize purse up to $150k

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

7

Marathon Matches

A **Marathon Match** is a long-running crowdsourcing challenge where participants can join any time during the match and solve complex algorithms on large datasets

The match proceeds in **two phases**: Testing (Provisional Scoring) and Sequestered (Final Scoring).

In the **Testing Phase** contestants will be supplied with Ground Truth data and will apply their synthetic data generation algorithm to create three synthetic data sets at three values of epsilon (using a specified value of delta). Then they'll submit their synthetic data for scoring, and their score will be updated on the leaderboard. Contestants can submit every 4 hours.

During the Testing Phase contestants can submit a clear, complete algorithm write-up and privacy proof to pass the "**Differential Privacy Prescreen**". This checks that they are making a good faith effort to satisfy differential privacy and are not committing any obvious errors. Prescreened entries will receive a significant score boost on the leaderboard.

After the Testing Phase concludes, the top scorers on the leaderboard will be invited to participate in the **Sequestered Phase**, to have a chance at winning a prize. They will submit their code, along with their written algorithm specification and privacy proof. The code will be run on the sequestered data (same format as the Ground Truth data used in the Testing phase, but a different subsample of the data), and the privacy proof and source code will be subjected to careful SME review. The top 5 scorers with valid differentially private solutions will be eligible for prizes.

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

8

*Requirement for Certification Process

Important Dates

Pre-registration	October 5, 2018
Challenge launch	October 31, 2018
Match #1	Oct 31 – Nov 29
Winners announced the week of January 4th	
Match #2	Jan 11 – Feb 9
Winners announced the week of March 3rd	
Match #3	March 10 – April 15
Final Winners announced the week of April 30th	

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Prize Awards

Match 1	Match 2	Match 3
1st Place: $10,000 2nd Place: $7000 3rd Place $5000 4th place $2000 5th place $1000	1st Place: $15,000 2nd Place: $10,000 3rd Place $5000 4th place $3000 5th place $2000	1st Place: $25,000 2nd Place: $15,000 3rd Place $10,000 4th place $5,000 5th place $3000
progressive prize: 4 @$1000 Total: $29,000	progressive prize: 4 @$1000 Total: $39,000	progressive prize: 4 @$1000 Total: $62,000
** An additional prize of $4000 may be awarded to each of the top 5 award winning teams at the end of the final marathon match who agree to provide and do provide their full code solution in an open source repository for use by all interested parties.* Total Prize Awards $150,000		

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Competition Details and Official Rules

Challenge.gov
https://challenge.gov/a/buzz/nist-pscr/differential-privacy-synthetic-data-challenge

Topcoder
https://www.topcoder.com/community/data-science/Differential-Privacy-Synthetic-Data-Challenge

Challenge Questions
PSPrizes@nist.gov

PSCR OPEN INNOVATION
Solving the Public Safety needs of tomorrow...today

Thank you!