SDNist: Benchmark Data and Evaluation Tools for Data Synthesizers

Grégoire Lothe,[2] Christine Task,[3] Isaac Slavitt,[4] Nicolas Grislain,[2] Karan Bhagat,[3] Gary S Howarth (gary.howarth@nist.gov) ,[1]

1 National Institute of Standards and Technology, Public Safety Communications Research Division, 2 Sarus Technologies, 3 Knexus Research Corporation, 4 DrivenData, Inc.

National Institute of Standards and Technology
U.S. Department of Commerce

The Challenge

NIST, PSCR hosted the Differential Privacy Temporal Map and Synthetic Data Challenges, inviting contestants to train DP algorithms on public data and scoring them with novel metrics on private data. Many solutions were open sourced. We introduce generic benchmarks using those data and scoring functions.

Temporal Map Challenge Results

(a) ACS dataset, $k-$marginal score

Dataset	$\epsilon = 0.1$ (1)	$\epsilon = 0.1$ (2)	$\epsilon = 1$ (1)	$\epsilon = 1$ (2)	$\epsilon = 10$ (1)	$\epsilon = 10$ (2)
N - CRiPT	781±2	807±2	**851**	**865**±1	**893**	**901**
Duke Privacy	796±1	816±3	832	852	881	890
Minutemen	**822**±1	788±1	825±1	834	873	881
DPSyn	805±3	**822**±1	818±1	844±1	822	848±1
Jim King	782±2	803±2	790	814	840	846

(b) Taxi dataset, $k-$marginal score

Dataset	$\epsilon = 1$ 2016	$\epsilon = 1$ 2020	$\epsilon = 10$ 2016	$\epsilon = 10$ 2020
Minutemen	**464**±3	**455**±15	**556**±5	491±3
N - CRiPT	340±7	437±18	456±2	**700**±2
DPSyn	344±1	433±3	416±1	464±2
GooseDP-PSA	251±2	382±1	251±1	382±1

(c) Taxi dataset, HOC score

Dataset	$\epsilon = 1$ 2016	$\epsilon = 1$ 2020	$\epsilon = 10$ 2016	$\epsilon = 10$ 2020
DPSyn	922	**942**	917	**945**±1
N-CRiPT	924	872±1	924	880
DP Duke	857±22	982±7	900±27	898±15
Minutemen	**931**	918	**929**	817
Jim King	828	845±1	839	885±2
GooseDP-PSA	865	827	864	827

On each table and for each value of ϵ, the left and right column indicate the score on the public and the private leaderboard respectively. (1)=NY-PA, (2)=GA-NC-SC

Table 3: Subsampling baseline, $k-$marginal score

Fraction	Census (1)	Census (2)	Taxi 2016	Taxi 2020
1%	572±1	590±1	547±1	472±1
10%	831	839	721	703
50%	940	944	889	887

The Benchmark Problems

American Community Survey (ACS)

Motivation: This data set captures a realistic, challenging use case de-identifying diverse, complex, heterogenous survey/tabular data

Size: 35 features, 7 years, 185 map segments (PUMA)

Query Sensitivity: 7 records/individual

Challenges: Edit constraints/variable dependencies, large categorical variables

Evaluation: 3-marginal density distribution comparison, averaged across map/time segments to ensure fair modeling accuracy for diverse population subgroups

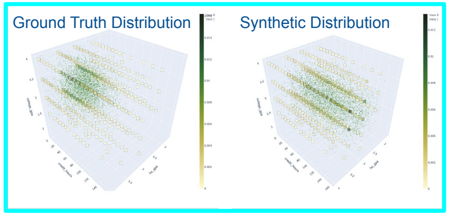

Fig. 1: Scoring synthetic data with 3-marginal density distribution comparison.

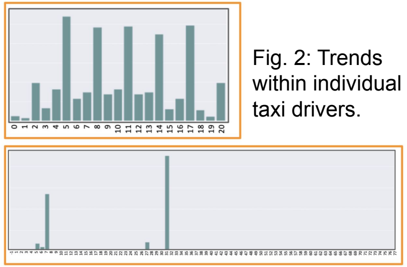

Fig. 2: Trends within individual taxi drivers.

Chicago Taxi Problem

Motivation: This data set requires de-identifying location sequences of real individuals while preserving both global and individual properties

Size: 13 features, 21 shifts, 78 map segments (neighborhoods)

Query Sensitivity: *200 records/individual*

Challenges: Mixed sparse/dense areas

Evaluation: Higher Order Conjunction comparison of individual sequences, network edges, and 3-marginal density averaged across map/time segments

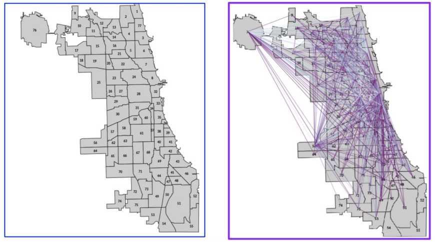

Fig. 3: Taxi trends between map segments.

The SDNist Python Library

Usage: Install from pypi or Github

```
$ pip install sdnist
  >>> import sdnist
# fetch public data
  >>> dataset, schema = sdnist.census()
# synthesize data or similar
  >>> sub = dataset.sample(n=20000)
# test on private data
  >>> score = sdnist.score(dataset,
sub, schema, challenge="census")
  >>> score
  CensusKMarginalScore(847)
# visualize on map
  >>> score.html()
# launches visualization in browser:
```

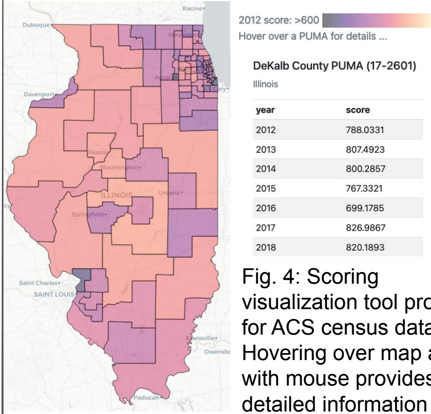

Fig. 4: Scoring visualization tool provided for ACS census data. Hovering over map areas with mouse provides detailed information per PUMA.