What all we've been up to at Knexus Research

Christine Task, PhD
Senior Computer Scientist
Knexus Research Corporation

Knexus Research is a small R&D company located in the DC area at National Harbor, MD.

We have over 13 years of experience moving research in AI and Data Science through the math, science and engineering steps necessary to make the jump from theory to practice.

We have three active projects in data privacy:

- With the DARPA Brandeis Program, Knexus PRESNA project is tackling the problem of developing noise-resistant decision metrics that can operate reliably over privatized data in critical contexts such as crisis detection.
- For the US Census Bureau, the Knexus CenSyn team is providing evaluation, research, engineering and production software development support for Census privacy efforts.
- As technical lead for the NIST Differentially Private Synthetic Data Challenge, Knexus is providing technical guidance for the first national challenge in Differential Privacy.

At Knexus, our priority is to develop the **tools and technologies necessary to facilitate safe, successful public adoption** of new research concepts. In this talk we'll describe our ongoing work and share lessons we've learned managing the challenges of data privacy applications.

Privacy Enhanced Social Network Analysis (PRESNA)

Project Details

Team: Dr. Christine Task (PI), Karan Bhagat (Software Engineer), Kevin Raoofi (Software Engineer)

Research Topic: Differentially private social network analysis

Start Date: July 2017

Application: Use privatized analysis of the call/text graph (***meta-data***) to alert emergency responders to crisis onset and track crisis features.

With **PRESNA** project on the DARPA Brandeis Program, Knexus is tackling the problem of developing noise-resistant decision metrics that can operate reliably over epsilon differentially privatized data in critical contexts such as crisis detection.

This material is based upon work supported by the **Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC Pacific)** under Contract No. **N66001-17-C-4012**. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of **DARPA or SSC Pacific**.

PRESNA System Overview:

City Block Call Generator and Privatized Crisis Localization

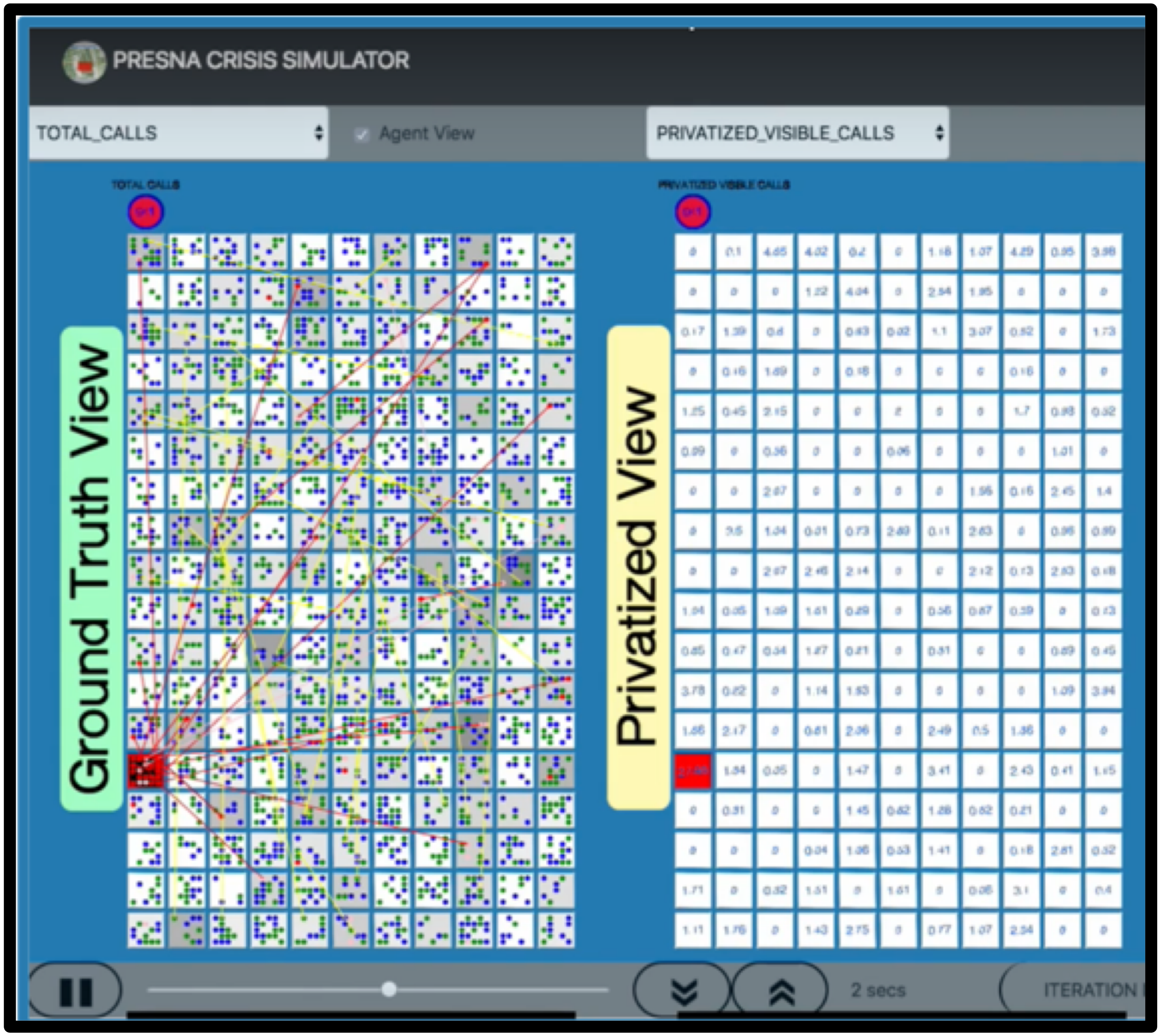
- Crisis localization using typed privatized cityblock call volume data
- Cityblock visualization, with statistics menu, agent view, and controllable playback
- Additional crisis detection metrics, improved crisis simulation logic and controls
- Can run over real data, simulated data, and **simulate synthetic crises over real data**.

PRESNA Evaluator

- Enables efficient parallelized execution, data collection, and chart generation for large experiments.
- Interface supports easy simultaneous comparison of multiple algorithms and parameter settings.

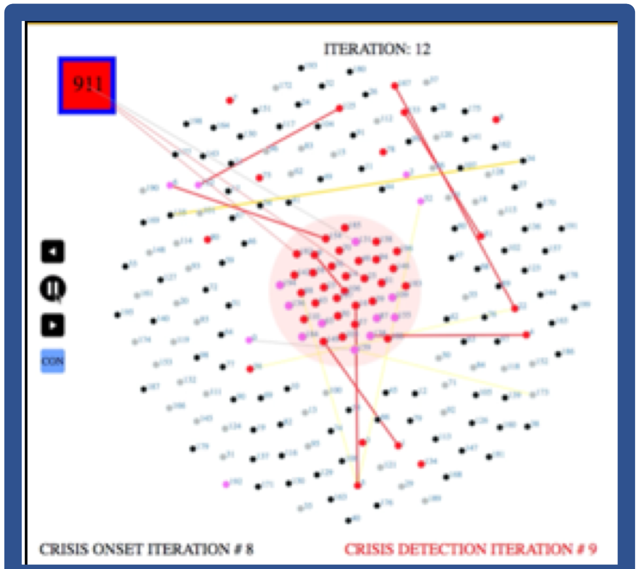
PRESNA BBN, Stealth Integration

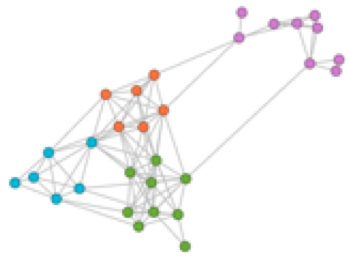
- Larger scale live crisis simulation with 2,000 simulated phones (AWS)
- Double spike crisis detection, using labeled edges: 911, incoming and outgoing
- Integration with Stealth FSS, securely collects and stores large time/space call count structure, returns query responses with Laplacian noise addition.



0	0.1	4.65	4.02	0.2	0	1.18	1.07	4.29	0.05	3.98
0	0	0	1.22	4.04	0	2.54	1.85	0	0	0
0.17	1.39	0.6	0	0.43	0.02	1.1	3.07	0.52	0	1.73
0	0.16	1.69	0	0.18	0	0	0	0.16	0	0
1.25	0.45	2.15	0	0	2	0	0	1.7	0.98	0.32
0.09	0	0.36	0	0	0.06	0	0	0	1.01	0
0	0	2.67	0	0	0	0	1.56	0.16	2.45	1.4
0	3.5	1.04	0.01	0.73	2.49	0.11	2.63	0	0.06	0.09
0	0	2.67	2.46	2.14	0	0	2.12	0.13	2.63	0.18
1.04	0.05	1.09	1.61	0.29	0	0.56	0.67	0.39	0	0.13
0.65	0.47	0.34	1.27	0.21	0	0.31	0	0	0.69	0.45
3.78	0.22	0	1.14	1.53	0	0	0	0	1.09	3.94
1.66	2.17	0	0.61	2.06	0	2.49	0.5	1.36	0	0
[illegible]	1.84	0.05	0	1.47	0	3.41	0	2.43	0.41	1.15
0	0.31	0	0	1.45	0.62	1.28	0.62	0.21	0	0
0	0	0	0.04	1.06	0.53	1.41	0	0.18	2.81	0.32
1.71	0	0.32	1.51	0	1.61	0	0.06	3.1	0	0.4
1.11	1.76	0	1.43	2.75	0	0.77	1.07	2.34	0	0

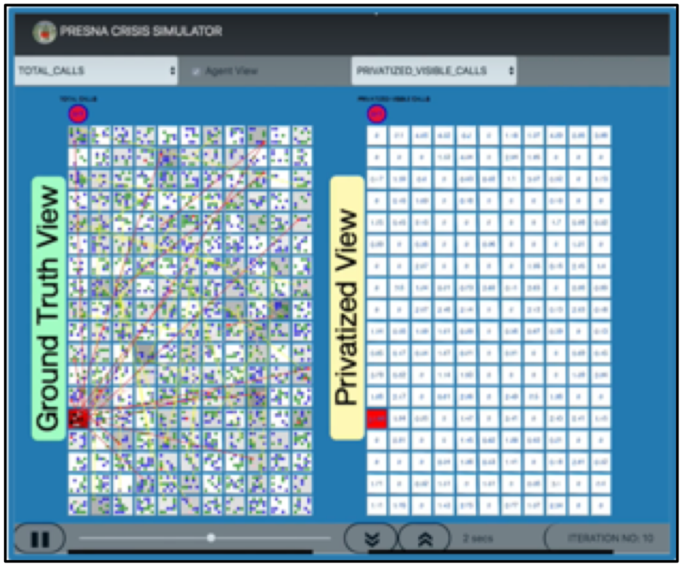
PRESNA Timeline

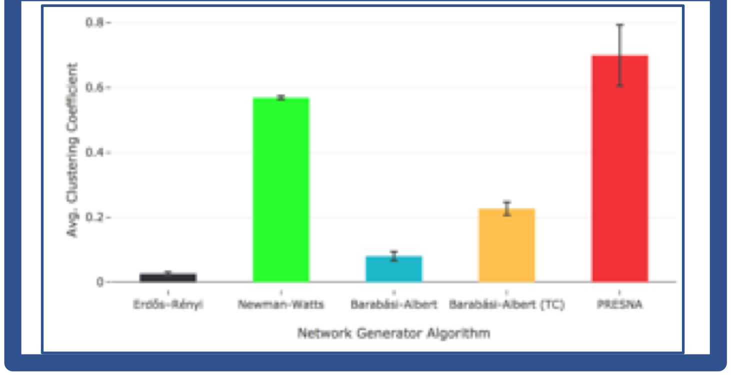




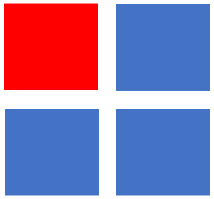
PRESNA

October 2017





May 2018



Query set after first spike:

🟦 $Total\_Incoming\_From(B_{crisis}, B\_i)$

$Total\_Incoming\_From(B_{j \neq i, crisis}, B\_i)$

$Total\_Outgoing(B\_i)$

🟥 $\emptyset$

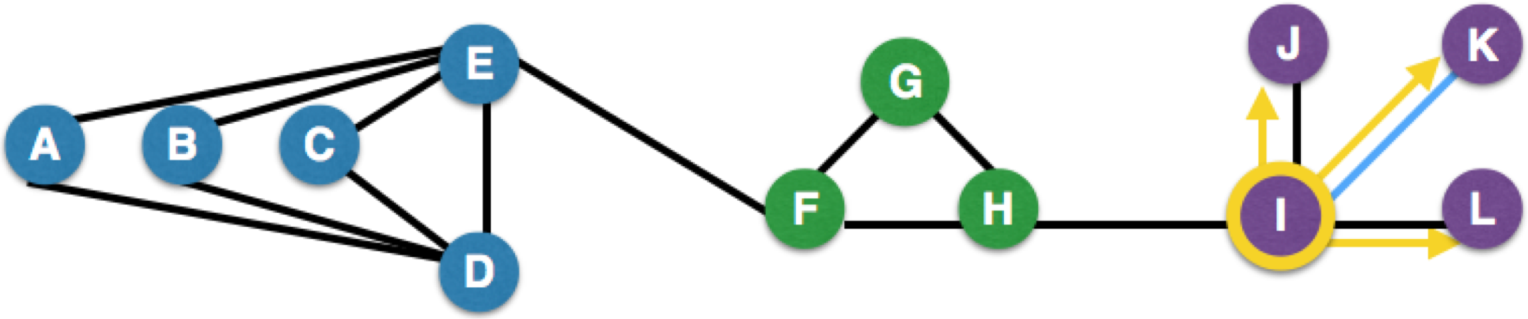
$\Delta\mathbf{F} = \mathbf{1}$

October 2018

KNEXUS RESEARCH CORPORATION

Single Slide Overview: Contributor Privacy for Social Networks

By changing our approach to network analysis, we can introduce new definitions of differential privacy that significantly reduce difficulties.



	Alice	Bob	Carla	Dan	Eun	Fran	Gigi	Hans	Isaac	Jyoti	Kyle	Lori
Alice	0	0	0	1	1	0	0	0	0	0	0	0
Bob	0	0	0	1	1	0	0	0	0	0	0	0
Carla	0	0	0	1	1	0	0	0	0	0	0	0
Dan	1	1	1	0	1	0	0	0	0	0	0	0
Eun	1	1	1	1	0	1	0	0	0	0	0	0
Fran	0	0	0	0	1	0	1	1	0	0	0	0
Gigi	0	0	0	0	0	1	0	1	0	0	0	0
Hans	0	0	0	0	0	1	1	0	1	0	0	0
Isaac	0	0	0	0	0	0	0	1	0	1	1	1
Jyoti	0	0	0	0	0	0	0	0	1	0	0	0
Kyle	0	0	0	0	0	0	0	0	1	0	0	0
Lori	0	0	0	0	0	0	0	0	1	0	0	0

Edge privacy protects one edge, but leaves node centered information exposed.

Contributor privacy protects information contributed by one node, while considering ego networks independently, allowing for very low sensitivity analyses.

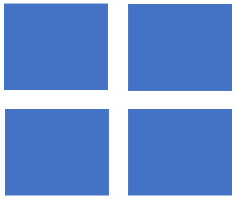
KNEXUS RESEARCH CORPORATION

PRESNA

A starting point: Simple Single Spike Detection

Motivation:

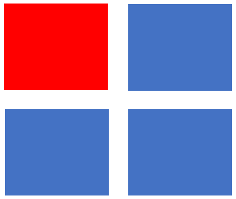
A simple first approach--A crisis will cause a spike in communication behavior. Which block has a spike in call volume?



Query set before first spike:

🟦 $Total\_Calls(B_i)$

$\mathbf{\Delta F = 1}$



Query set after first spike:

🟦 $Total\_Calls(B_i)$

🟥 $\emptyset$

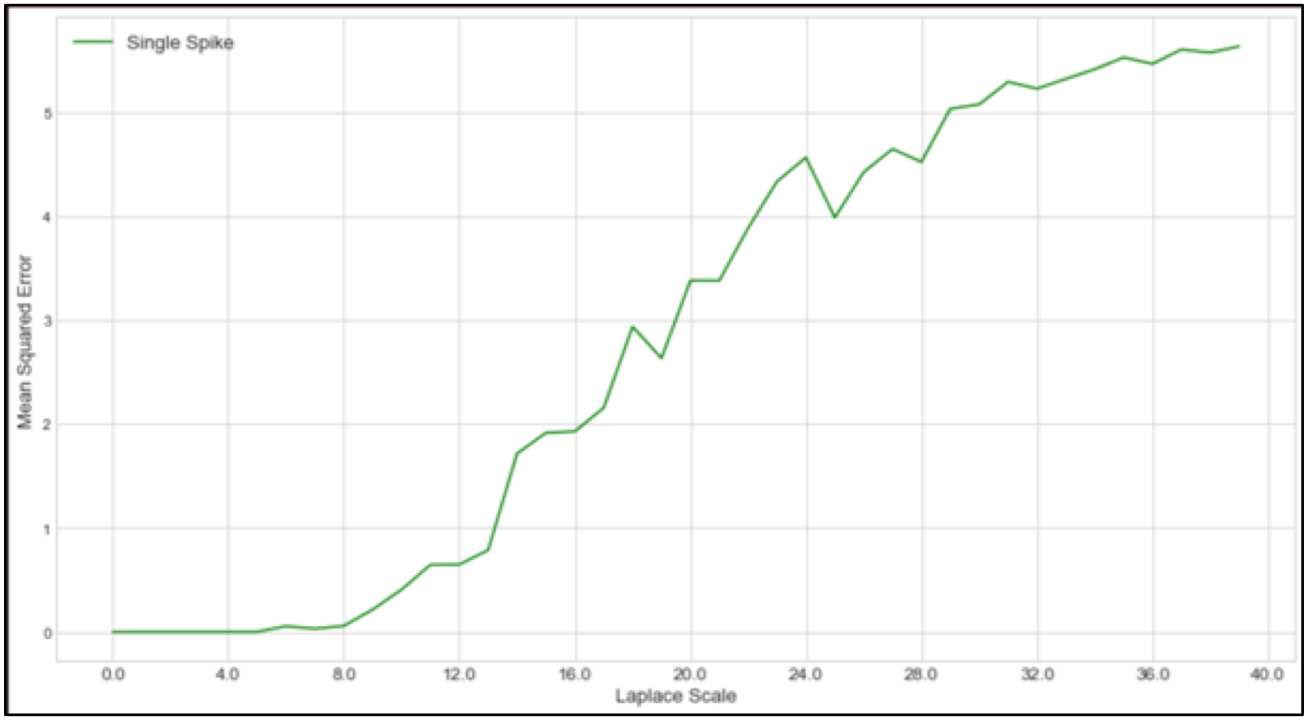
$\mathbf{\Delta F = 1}$

Crisis Detection Metric:

$$Total\_Calls(B_i)[iteration\ i-1] - Total\_Calls(B_i)[iteration\ i] > \mathrm{T}$$

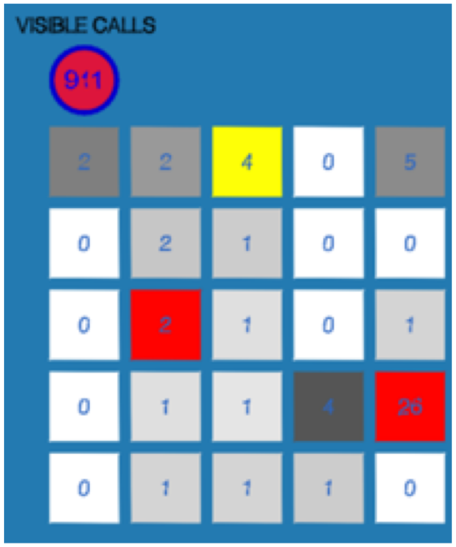
Spike due to crisis… or noise?

A starting point: Simple Single Spike Detection



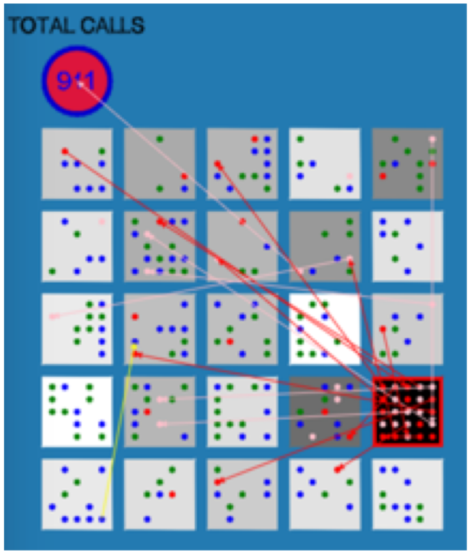
Mean Squared Error is computed for each simulation by comparing the prediction (crisis, warning, normal) made at each iteration on each city block, against the true value (crisis, normal). False Detections are given an additional penalty.

Predicted

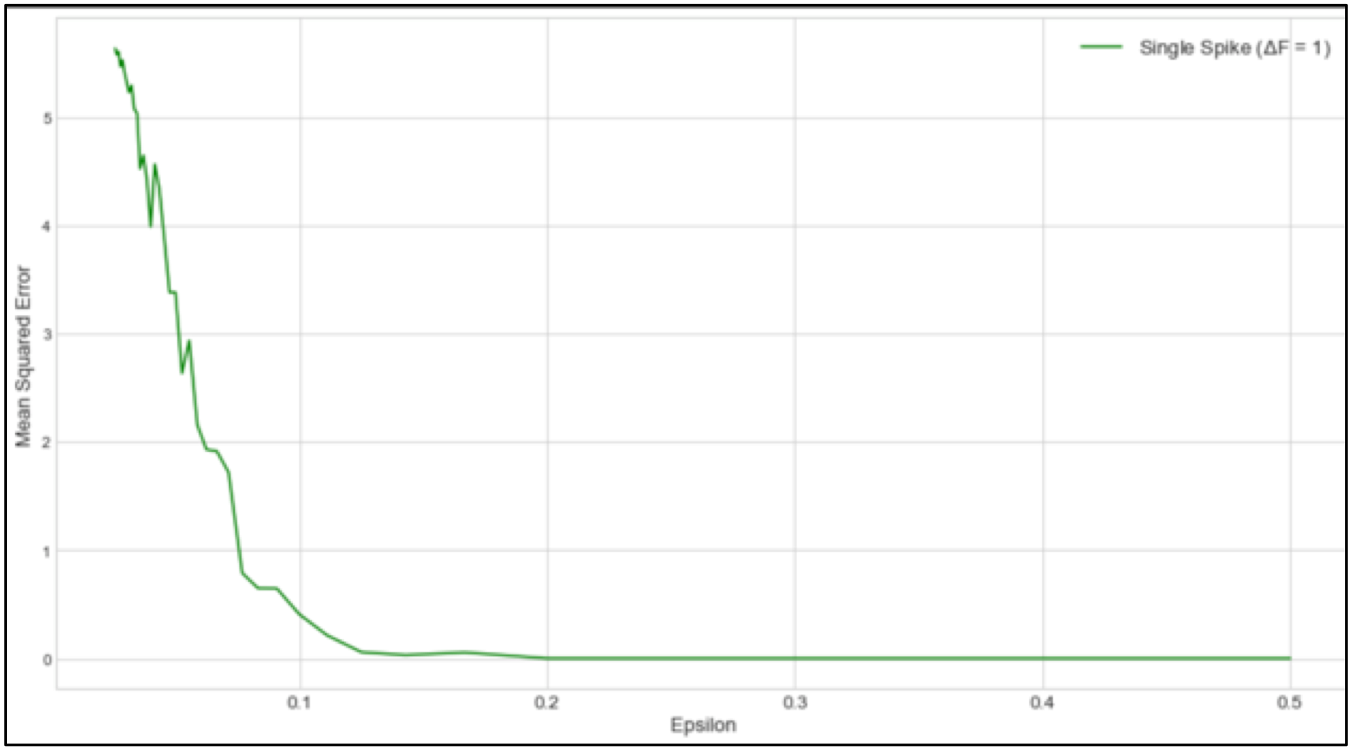


2	2	4	0	5
0	2	1	0	0
0	2	1	0	1
0	1	1	4	26
0	1	1	1	0

Ground Truth

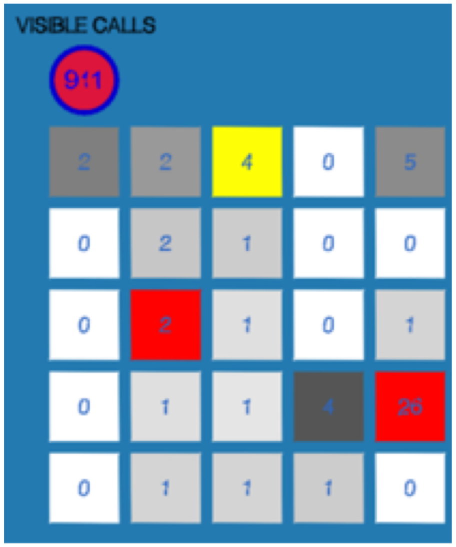


A starting point: Simple Single Spike Detection



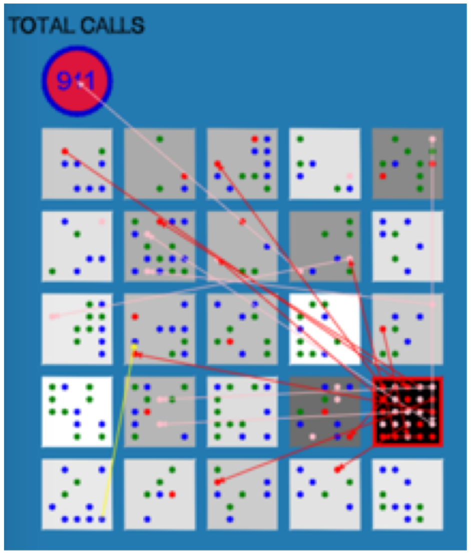
Mean Squared Error is computed for each simulation by comparing the prediction (crisis, warning, normal) made at each iteration on each city block, against the true value (crisis, normal). False Detections are given an additional penalty.

Predicted



2	2	4	0	5
0	2	1	0	0
0	2	1	0	1
0	1	1	4	26
0	1	1	1	0

Ground Truth

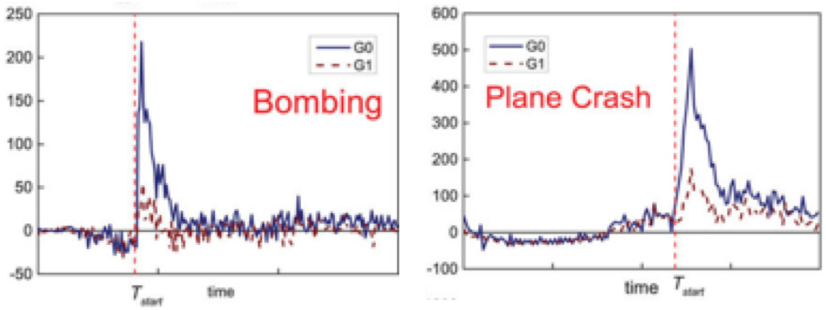


Privatized Crisis Detection Evaluation Results:

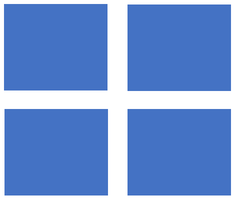
Double Spike Detection

Motivation:

Research in crisis communication has shown that crisis locations experience a spike in incoming calls that rapidly follows the initial spike in outgoing calls.



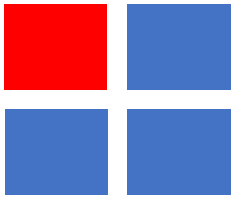
From Quantifying Information Flow During Emergencies [Gao 2014]



Query set before first spike:

🟦 $Total\_Calls(B_i)$

$\mathbf{\Delta F = 1}$



Query set after first spike:

🟦 $Total\_Calls(B_i)$

🟥 $Total\_Incoming\_Calls(B_{crisis})$

$\mathbf{\Delta F = 1}$

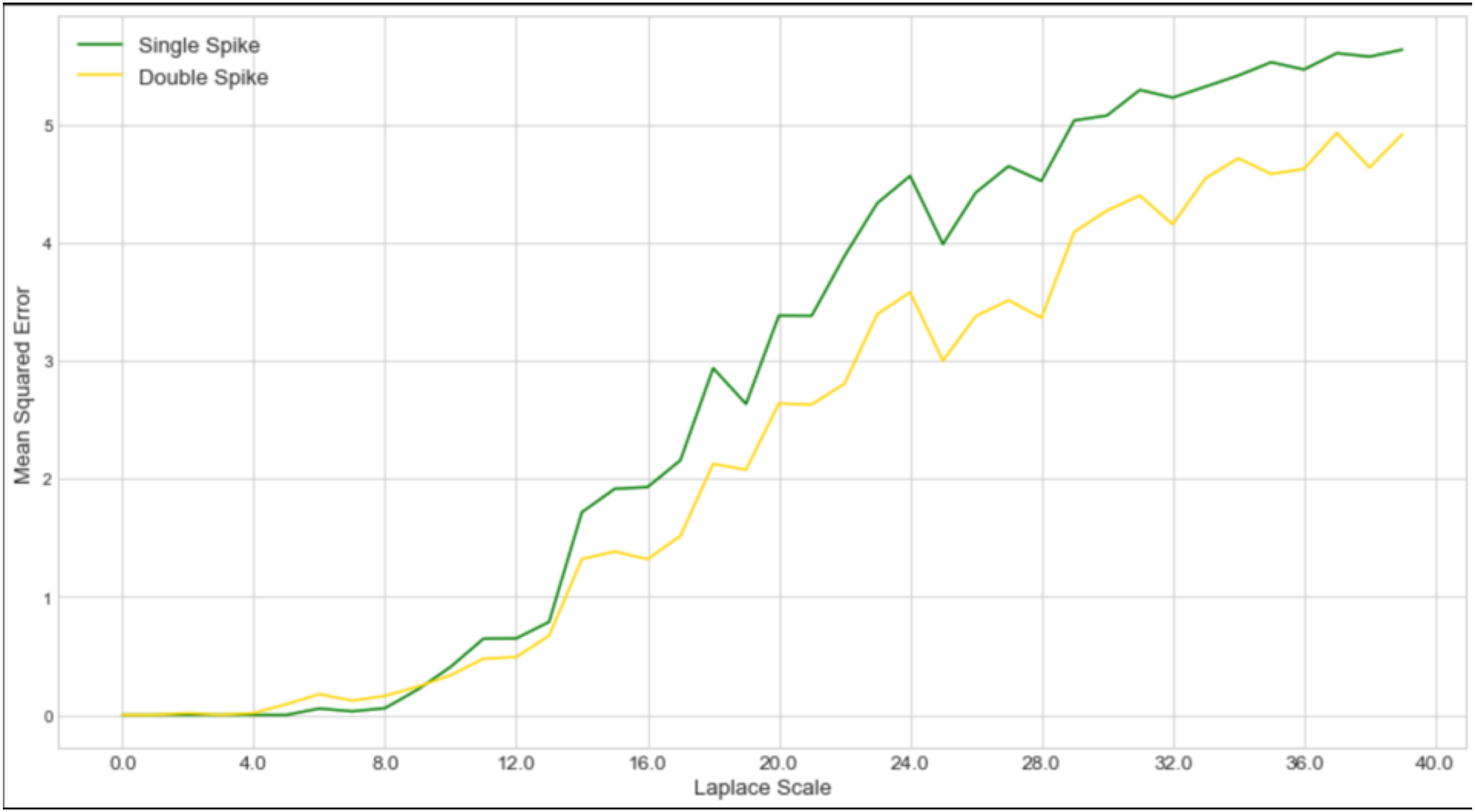
Crisis Detection Metric:

First Spike: $Total\_Calls(B_i)[iteration\ i-1] - Total\_Calls(B_i)[iteration\ i] > T_1$

Second Spike: $Total\_Incoming\_Calls(B_i)[iteration\ i_{crisis}-1] - Total\_Incoming\_Calls(B_i)[iteration\ i] > T_2$

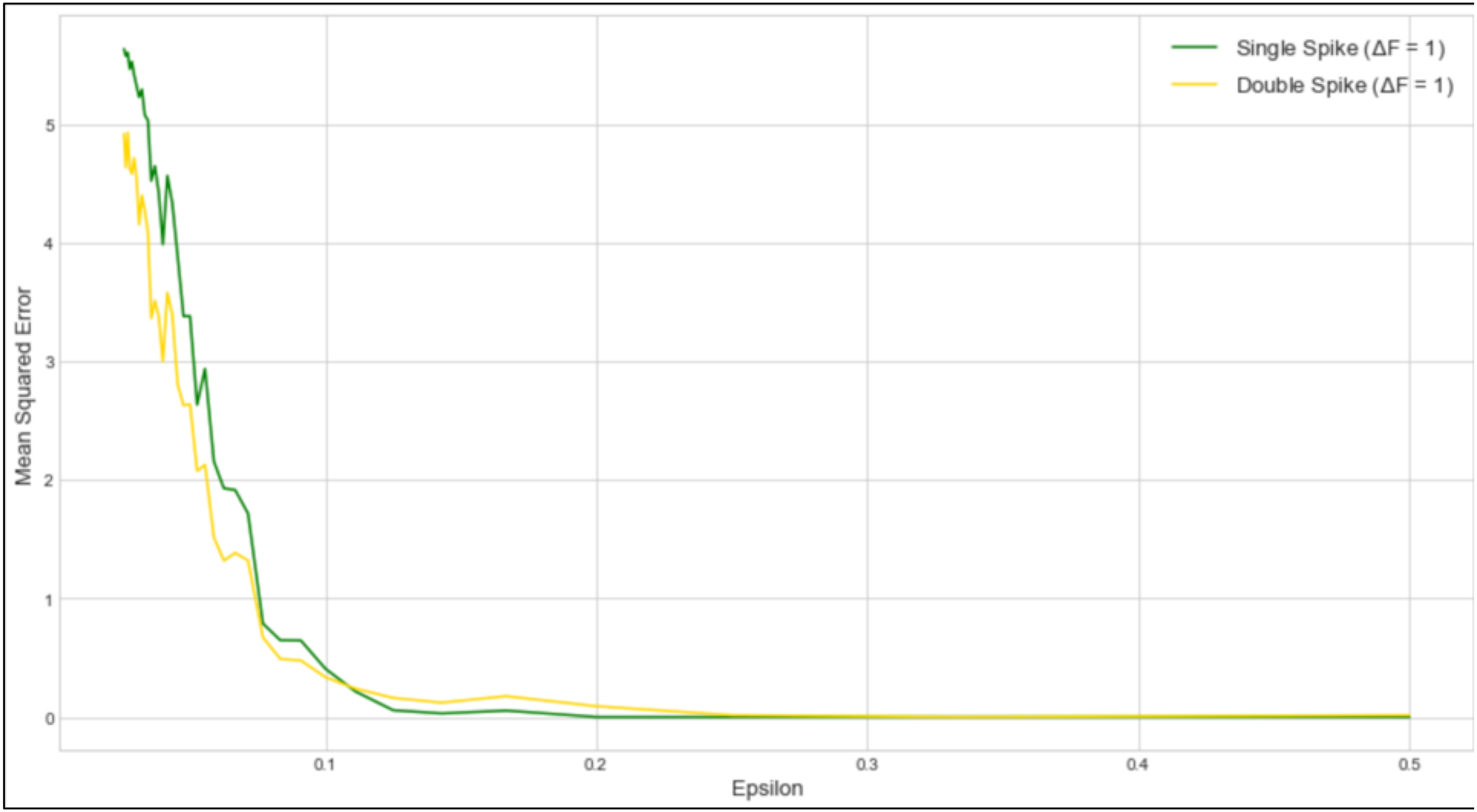
Privatized Crisis Detection Evaluation Results:

Double Spike Detection



Privatized Crisis Detection Evaluation Results:

Double Spike Detection

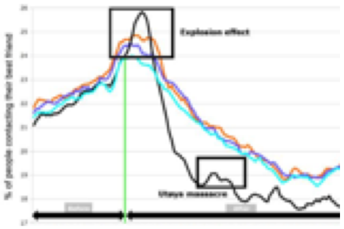


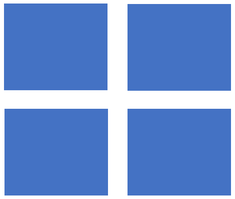
Distribution A. Approved for public release: distribution unlimited

Olso Adaption: Favorite Spike

Motivation:

[Sunsoy 2012] Found that during a crisis, calls to ‘favorite’ contacts first spiked and then dropped below normal. We adapted their measures for this effect to privatized queries, applying PE Android’s privacy streams concept to collect typed call information.

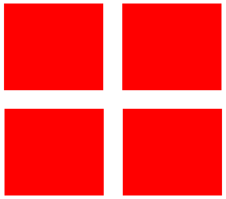




Query set before first spike:

🟦 $Total\_Calls(B_i)$

$$\Delta\mathbf{F} = \mathbf{1}$$



Query set after first spike:

🟥 $Total\_Calls(B_i)$

$Total\_Favorite\_Calls(B_i)$

$$\Delta\mathbf{F} = \mathbf{2}$$

Crisis Detection Metric:

First spike:

$$\frac{(\sum_i Total\_Favorite\_Calls(B_i))}{\sigma(\sum_i Normal(B_i))} > T_1$$

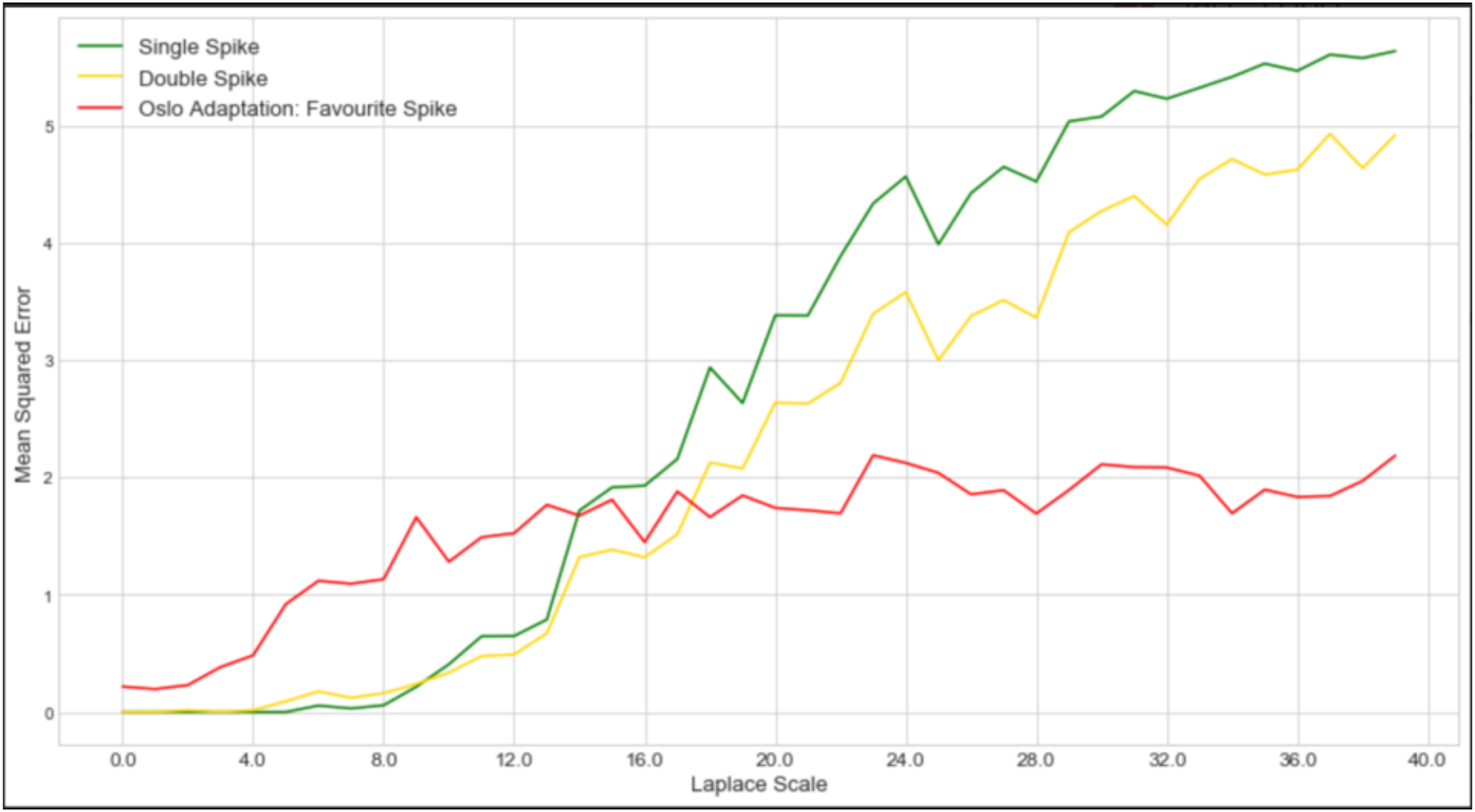
Favorite spike drop:

$$\frac{(\sum_i Total\_Favorite\_Calls(B_i))}{\sigma(\sum_i Normal(B_i))} < T_2$$

Localization: $\max_i \; Total\_Calls(B_i)$

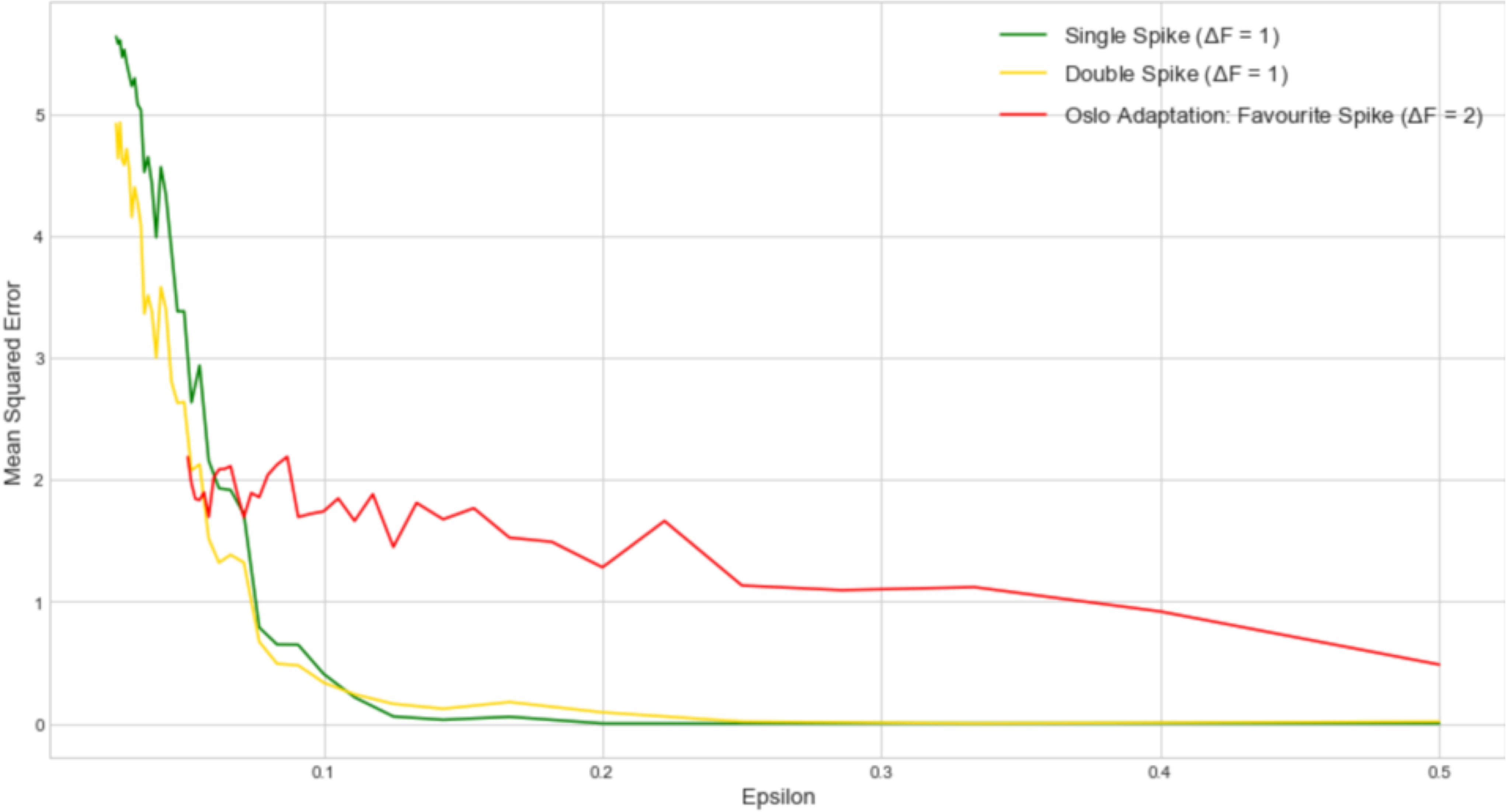
Privatized Crisis Detection Evaluation Results:

Oslo Adaptation: Favorite Spike



Privatized Crisis Detection Evaluation Results:

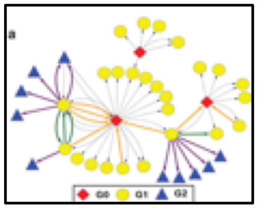
Oslo Adaptation: Favorite Spike

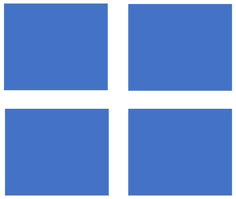


Privatized Crisis Detection Evaluation Results:

Gao Adaption: Info Flow Spike

Motivation: [Gao 2014] followed the way social networks naturally distribute crisis information. They examined the crisis affected population G0, and compared the behavior of recipients of G0's calls during crisis and non-crisis situations. Tracking individuals isn't feasible in a privacy-preserving real time crisis detection, but the phenomenon they identified can be captured in privacy-preserving queries.

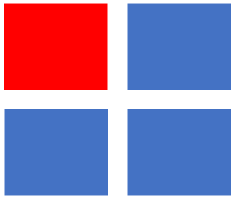




Query set before first spike:

🟦 $Total\_Outgoing(B_i)$

$\mathbf{\Delta F = 1}$



Query set after first spike:

🟦 $Total\_Incoming\_From(B_{crisis}, B\_i)$

$Total\_Incoming\_From(B_{j\neq i,crisis}, B\_i)$

$Total\_Outgoing(B\_i)$

🟥 $\emptyset$

$\mathbf{\Delta F = 1}$

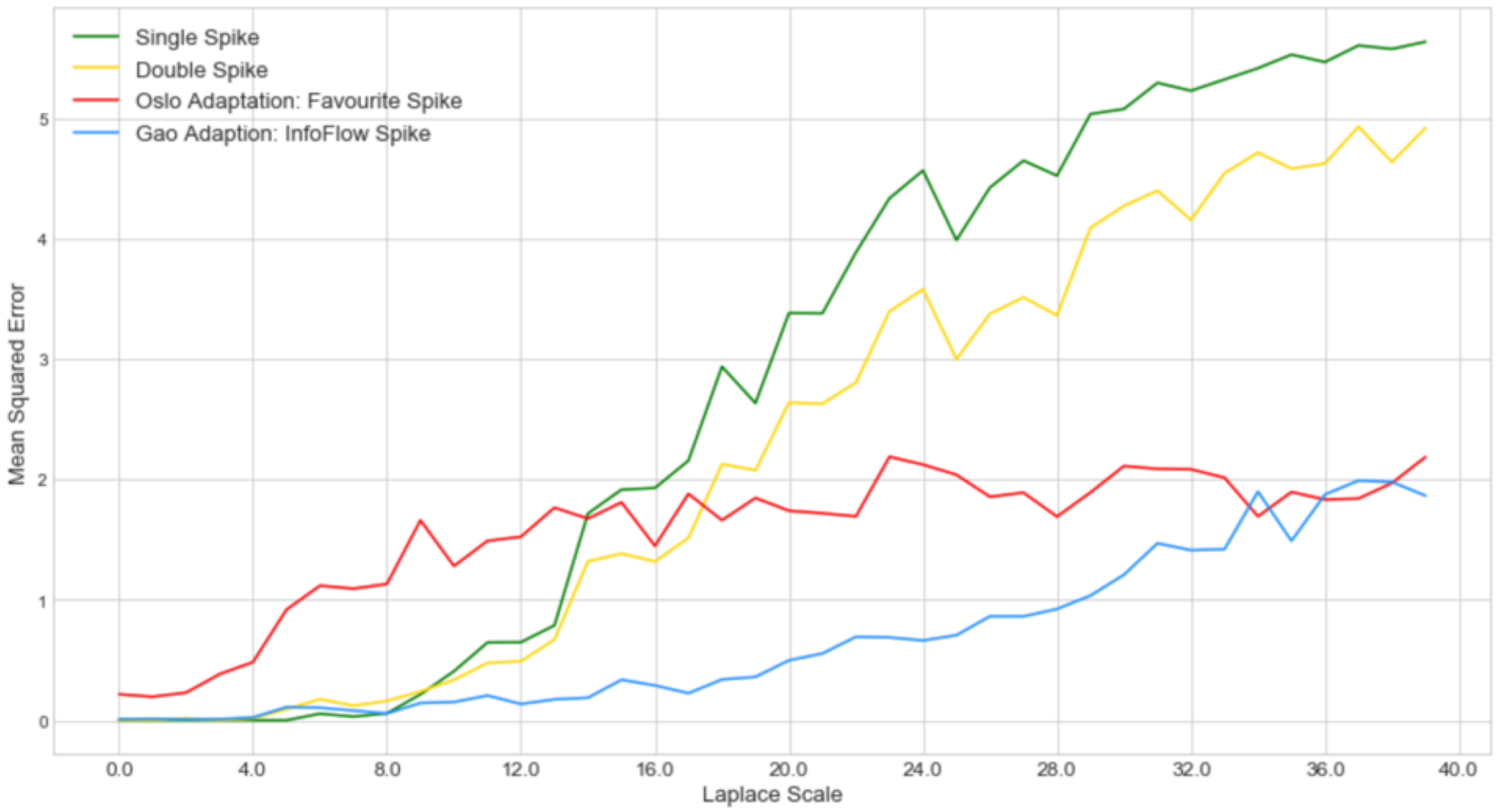
Crisis Detection Metric:

First spike: $\frac{Total\_Outgoing(B_i)}{\sigma(Normal(B_i))} > T_1$

Second spike: $\sum_{i\neq crisis} Total\_Outgoing(B_i)\times \frac{Total\_Incoming\_From\left(B_{crisis,}B_i\right)}{Total\_Incoming\_From\left(B_{j\neq i,crisis)}, B_i\right)} > T_2$

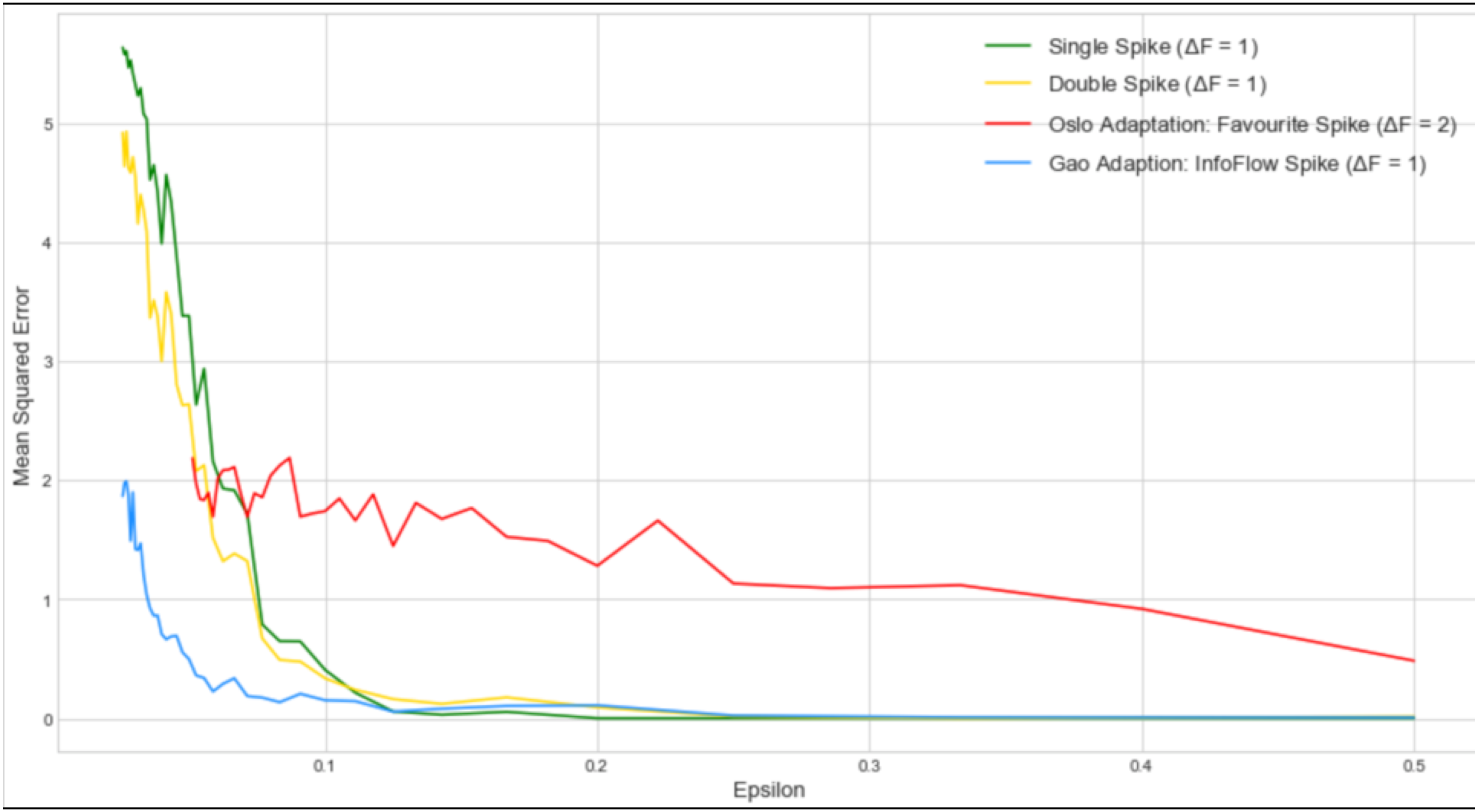
Estimates percentage of B_i's call increase due to receiving calls from B_{crisis}

Gao Adaptation: Info Flow Spike

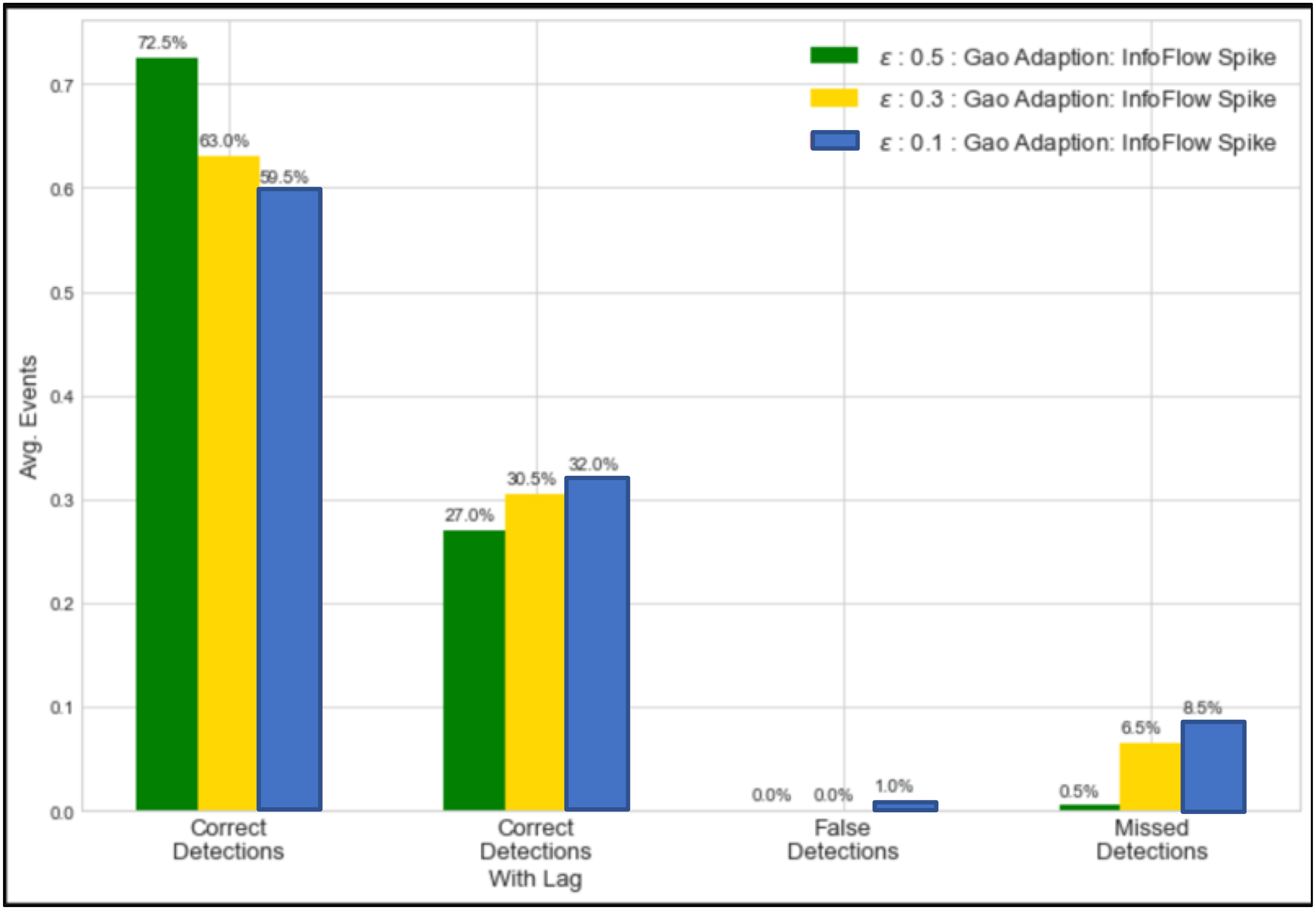


Privatized Crisis Detection Evaluation Results:

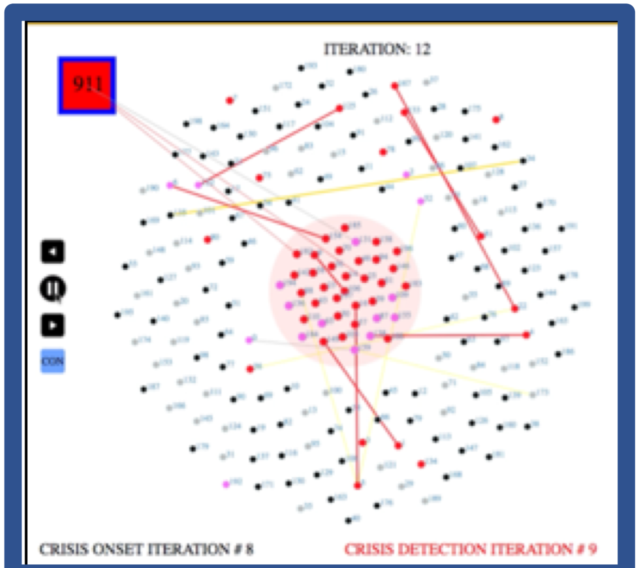
Gao Adaptation: Info Flow Spike

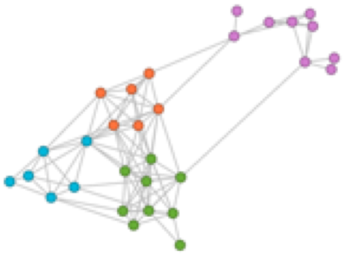


Gao Adaptation: Info Flow Spike



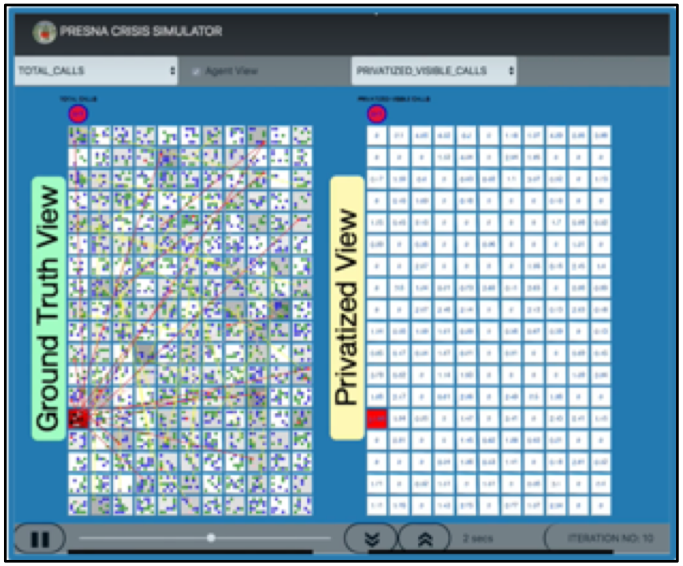
PRESNA Timeline

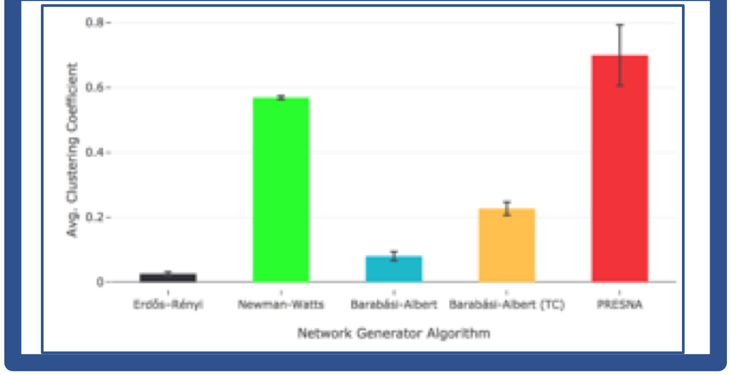




PRESNA

October 2017





May 2018

🟥 🟦

🟦 🟦

Query set after first spike:

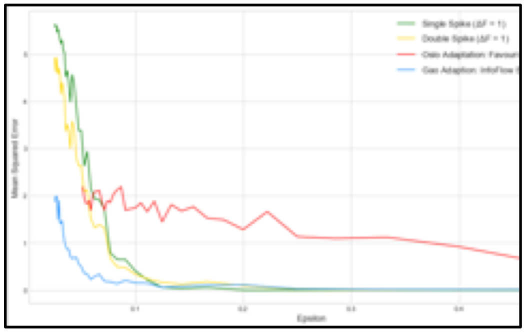
🟦 $Total\_Incoming\_From(B_{crisis}, B\_i)$

$Total\_Incoming\_From(B_{j \neq i, crisis}, B\_i)$

$Total\_Outgoing(B\_i)$

🟥 $\emptyset$

$\mathbf{\Delta F = 1}$



October 2018

The CenSyn Project

Project Details

Team: Christine Task (Research Lead), Micah Heineck (Software Engineering Lead)
Jason Suagee, Christine Heiss, Karan Bhagat, Joe Graus,
Konrad Rauscher, Jeffrey Hodges, Jonathan Woodell

Research Topic: Synthetic and Privatized Survey Data Generation

Start Date: October 2018

Application: ACS and other Census products

Overview:

The Knexus **CenSyn** team is providing research, engineering and software development support for ongoing Census privacy efforts. We're charged with developing effective, usable software systems that address the problem of data privacy.

There are several speakers at this workshop who are working with the Census. To avoid repeating topics covered in others' talks, in our CenSyn segment we will focus on one unique contribution: our **Evaluation Suite**. Originally developed for work on non-DP synthetic data for the 2019 ACS, the evaluation suite will be applicable to synthetic data generation research in general.

The views and opinions expressed in this talk are those of the authors and not the U.S Census Bureau.

CenSyn Evaluation Suite: Motivation

- Given a data-set **D** fitting a schema **S**, a *synthetic* replacement for **D** refers to an artificially generated data-set $\mathbf{D}_{\mathbf{synth}}$ that fits schema **S** and is distributionally similar to **D**. Any overlap on individual rows between the two data-sets is purely coincidental, and should occur with reasonably low probability.

- There are a myriad of techniques for producing synthetic data. At a high level, synthetic data generation is generally accomplished by capturing the distribution of the data (in some fashion), and then sampling another data-set from that same distribution

- High quality synthetic survey data should ideally be as similar to the original data as two uniform random partitions of the original data are to each other.

- Privacy-preserving data generators can be iteratively tested and improved safely using previously released public data-sets of the target schema, where available. All results shown here come from the publicly released 2016 ACS PUMS data.

- So… the obvious question is:

CenSyn Evaluation Suite: Motivation

- Given a data-set **D** fitting a schema **S**, a *synthetic* replacement for **D** refers to an artificially generated data-set $\mathbf{D}_{\mathbf{synth}}$ that fits schema **S** and is distributionally ***similar*** to **D**. Any overlap on individual rows between the two data-sets is purely coincidental, and should occur with reasonably low probability.

- There are a myriad of techniques for producing synthetic data. At a high level, synthetic data generation is generally accomplished by capturing the distribution of the data (in some fashion), and then sampling another data-set from that same distribution

- High quality synthetic survey data should ideally be as ***similar*** to the original data as two uniform random partitions of the original data are to each other.

- Privacy-preserving data generators can be iteratively tested and improved safely using previously released public data-sets of the target schema, where available. All results shown here come from the publicly released 2016 ACS PUMS data.

- So… the obvious question is:

***Similar how*?**

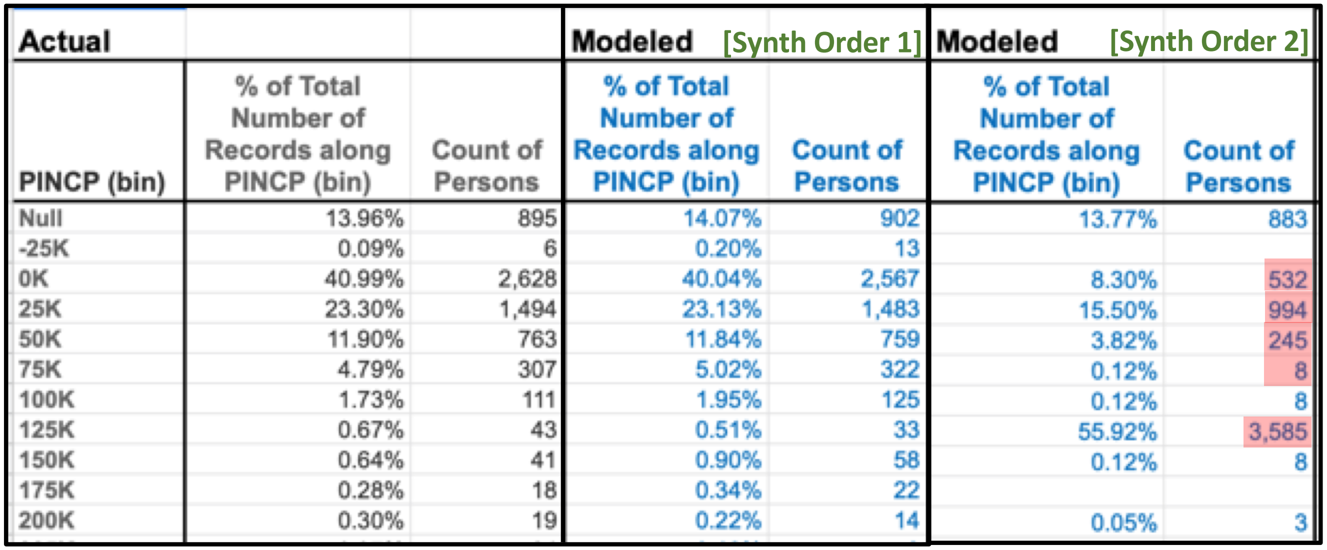
CenSyn Evaluation Suite: Tables

As a first take on the problem, it makes sense to look at individual variables of interest to verify that their values are distributed similarly between the two data-sets.

Tables give information on the distribution of populations across variables such as demographics, income, etc. in the form of counts/histograms/percentages. These tables combine data from a single or small number of related variables of interest. A wide variety of tables are released by the USCB and available on the **American Fact Finder** website.

AMERICAN FactFinder

The **Table Based Evaluator** will highlight table cells with significant deviation between the real and synthetic data.

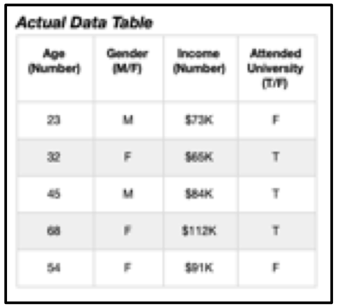


Actual			Modeled [Synth Order 1]		Modeled [Synth Order 2]	
PINCP (bin)	% of Total Number of Records along PINCP (bin)	Count of Persons	% of Total Number of Records along PINCP (bin)	Count of Persons	% of Total Number of Records along PINCP (bin)	Count of Persons
Null	13.96%	895	14.07%	902	13.77%	883
-25K	0.09%	6	0.20%	13		
0K	40.99%	2,628	40.04%	2,567	8.30%	532
25K	23.30%	1,494	23.13%	1,483	15.50%	994
50K	11.90%	763	11.84%	759	3.82%	245
75K	4.79%	307	5.02%	322	0.12%	8
100K	1.73%	111	1.95%	125	0.12%	8
125K	0.67%	43	0.51%	33	55.92%	3,585
150K	0.64%	41	0.90%	58	0.12%	8
175K	0.28%	18	0.34%	22		
200K	0.30%	19	0.22%	14	0.05%	3

CenSyn Evaluation Suite: Dr. Sergey Pogodin's K-Marginal

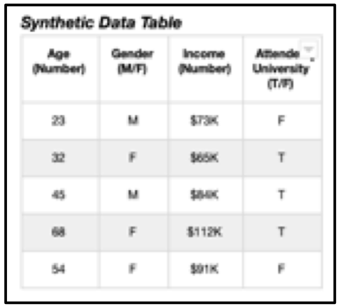
But what about preserving patterns in the data more generally? Correlations between variables in social survey data can form complex tangles, and these are difficult to reduce or summarize. Very few variables pairs show complete independence.

As part of the NIST Differential Privacy Synthetic Data Challenge (more on that later), we needed to address the problem of efficiently snapshotting complex correlation/distribution properties. To this end, TopCoder's Dr. Sergey Pogodin proposed a quickly computable metric based on randomly sampled 3-marginals.



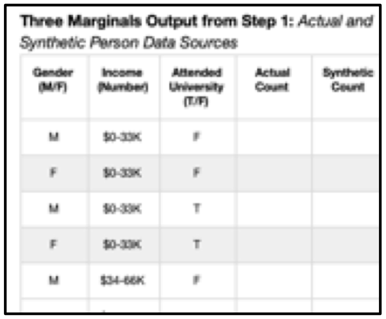
Actual Data Table

Age (Number)	Gender (M/F)	Income (Number)	Attended University (T/F)
23	M	$73K	F
32	F	$65K	T
45	M	$84K	T
68	F	$112K	T
54	F	$91K	F



Synthetic Data Table

Age (Number)	Gender (M/F)	Income (Number)	Attende University (T/F)
23	M	$73K	F
32	F	$65K	T
45	M	$84K	T
68	F	$112K	T
54	F	$91K	F



Three Marginals Output from Step 1: *Actual and Synthetic Person Data Sources*

Gender (M/F)	Income (Number)	Attended University (T/F)	Actual Count	Synthetic Count
M	$0-33K	F		
F	$0-33K	F		
M	$0-33K	T		
F	$0-33K	T		
M	$34-66K	F		

For our K-marginal Evaluator, we've expanded this approach to production level software, including configurable bucketing, data sub-selection, the ability to adjust the span and width of the marginals, and detailed reporting that assists in tracking down deviations in the data.

Still efficiently computable, this tool gives us an easy, effective, single score output that is invaluable for optimization and parameter fitting during synthesizer development.

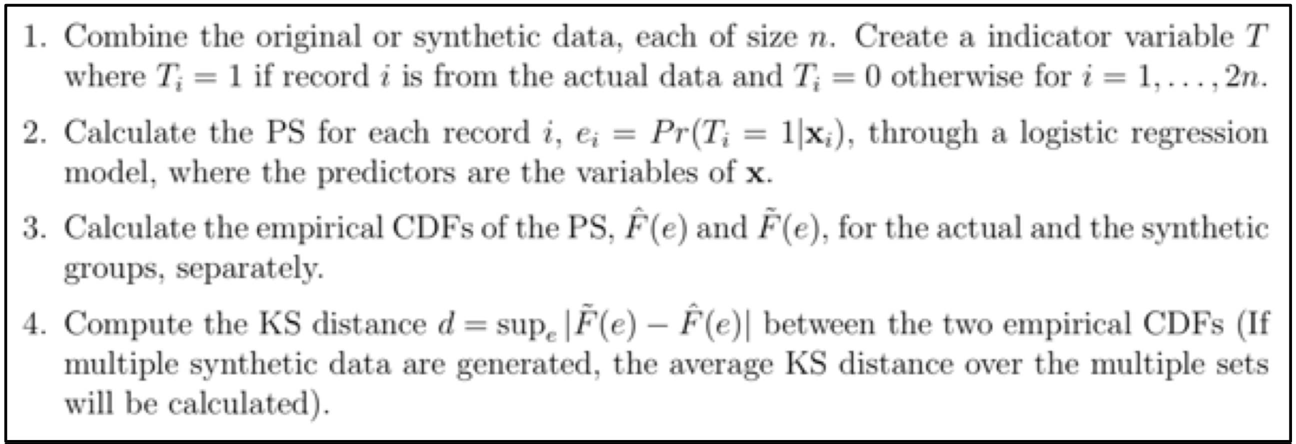
CenSyn Evaluation Suite: Dr. Claire Bowen’s SPECKS Method

Randomized heuristics are useful, but it’s important to have holistic evaluations as well. The **SPECKS method** for comparing high dimensional data sets was developed by Dr. Claire Bowen as part of her doctoral research in differentially private synthetic data at the University of Notre Dame.

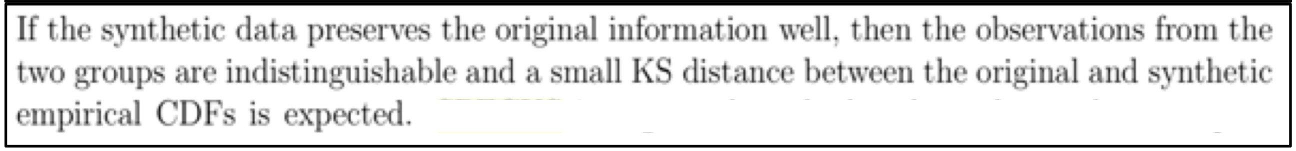
The acronym abbreviates the algorithm itself:

Synthetic Data Generation,
Propensity Score Matching,
Empirical CDF,
Comparison based on
Kolmogorov-
Smirnov

Clare Bowen, "Data Privacy Via Integration of Differential Privacy and Data Synthesis”, PhD diss, University of Notre Dame, Indiana, 2018

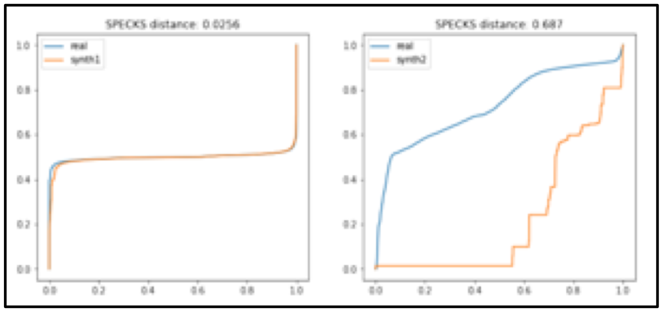


1. Combine the original or synthetic data, each of size n. Create a indicator variable T where $T_i = 1$ if record i is from the actual data and $T_i = 0$ otherwise for $i = 1, \dots, 2n$.
2. Calculate the PS for each record i, $e_i = Pr(T_i = 1|\mathbf{x}_i)$, through a logistic regression model, where the predictors are the variables of $\mathbf{x}$.
3. Calculate the empirical CDFs of the PS, $\hat{F}(e)$ and $\tilde{F}(e)$, for the actual and the synthetic groups, separately.
4. Compute the KS distance $d = \sup_e |\tilde{F}(e) - \hat{F}(e)|$ between the two empirical CDFs (If multiple synthetic data are generated, the average KS distance over the multiple sets will be calculated).



If the synthetic data preserves the original information well, then the observations from the two groups are indistinguishable and a small KS distance between the original and synthetic empirical CDFs is expected.

Bowen, Claire McKay, and Fang Liu. "STatistical Election to Partition Sequentially (STEPS) and Its Application in Differentially Private Release and Analysis of Youth Voter Registration Data." *arXiv preprint arXiv:1803.06763* (2018).

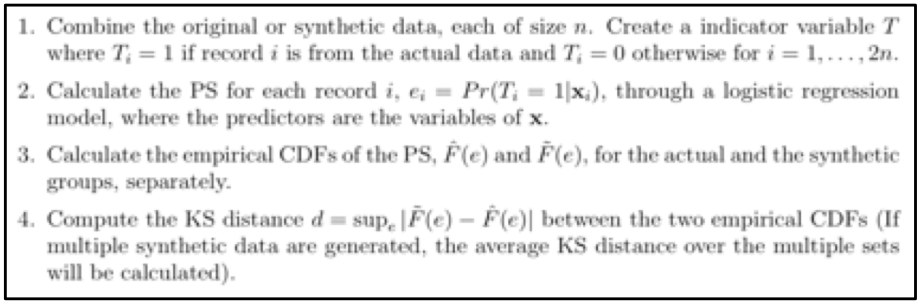


CenSyn Evaluation Suite: Dr. Claire Bowen’s SPECKS Method

Randomized heuristics are useful, but it’s important to have holistic evaluations as well. The **SPECKS method** for comparing high dimensional data sets was developed by Dr. Claire Bowen as part of her doctoral research in differentially private synthetic data at the University of Notre Dame.

The Kenxus **SPECKS Evaluator**, prototyped by Jason Suagee, supports easy comparison between different classification models for computing the Propensity Scores in Step 2 of the algorithm.

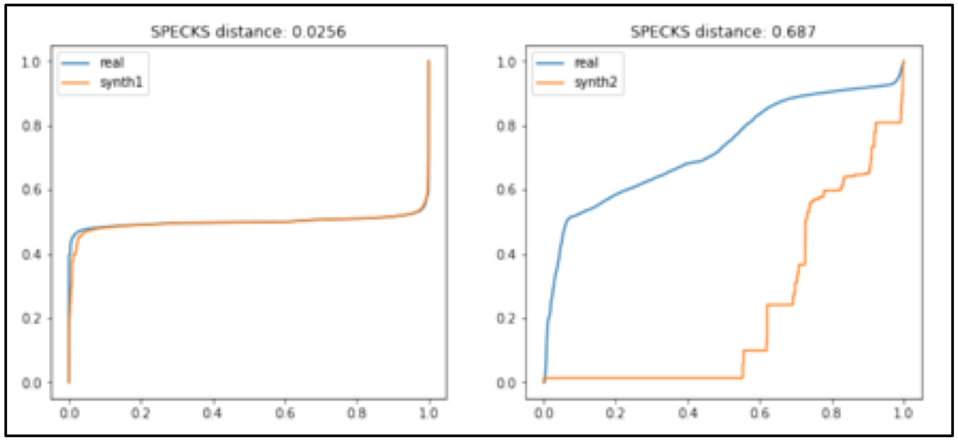
Shown here is an example comparing two different parameter conditions for our ACS synthesizer. Condition one (left) produces high quality synthetic data on the variables tested, condition two (right) produces very low quality data. The **neural net model** in this example has more discriminative power than the **logistic regression model**.



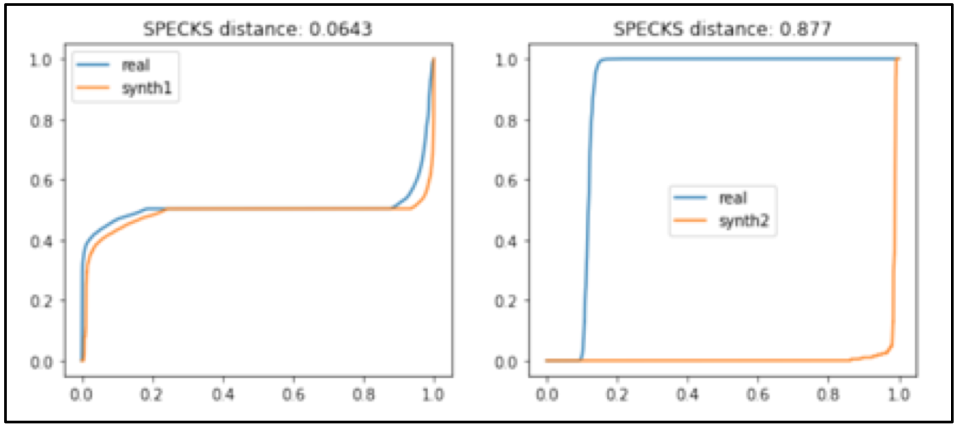
1. Combine the original or synthetic data, each of size n. Create a indicator variable T where $T_i = 1$ if record i is from the actual data and $T_i = 0$ otherwise for $i = 1, \ldots, 2n$.
2. Calculate the PS for each record i, $e_i = Pr(T_i = 1|\mathbf{x}_i)$, through a logistic regression model, where the predictors are the variables of $\mathbf{x}$.
3. Calculate the empirical CDFs of the PS, $\hat{F}(e)$ and $\tilde{F}(e)$, for the actual and the synthetic groups, separately.
4. Compute the KS distance $d = \sup_e |\bar{F}(e) - \hat{F}(e)|$ between the two empirical CDFs (If multiple synthetic data are generated, the average KS distance over the multiple sets will be calculated).

Bowen, Claire McKay, and Fang Liu. "STatistical Election to Partition Sequentially (STEPS) and Its Application in Differentially Private Release and Analysis of Youth Voter Registration Data."

Logistic Regression Model



Multilayer Perceptron Network Model



CenSyn Evaluation Suite: Additional Approaches

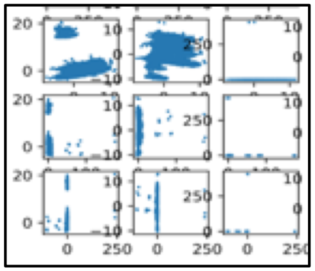
Evaluation for synthetic data is a topic of ***challenging***, ***active***, and ***vitally important*** research …..and this is a 15 minute presentation segment. We won't be providing any canonical answers today.

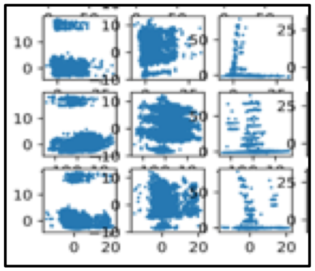
But the Knexus Evaluation Suite is designed to support the type of research, both in evaluating and in generating synthetic data, that we expect to be requisite for the engineering and development of practical, usable synthetic data generators.

Additional Evaluation Suite Tools that I did not cover in detail here:

- Analytics Package: This tool will evaluate data from the social scientist data user's perspective, focusing on ensuring that the performance of important analytics (such as pay gap analysis or measuring the lasting effects of red-lining) are not problematically impacted by the substitution of synthetic data.

- PCA Techniques: Dimension reduction can provide a holistic visual tool for exploring complex data sets. Our evaluation prototypes include several promising PCA-based approaches for understanding data.





- Other NIST Challenge Heuristics: The K-marginal technique was introduced for scoring in Match #1. Match #2 introduced (and *Match #3 will introduce)* new scoring metrics, capturing new features of data similarity. ***More on that…* about now, actually.**

KNEXUS
RESEARCH CORPORATION

The NIST Differential Privacy Synthetic Data Challenge

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

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

KNEXUS
RESEARCH CORPORATION

Simons Institute: Foundations to Applications Workshop 3/2019

NIST Challenge

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

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

Project Details

NIST-Challenge Oversight, PSCR Expertise, Metrology Expertise
Mary Theofanos (PI),
Terese Manley (Prize Manager)

Knexus Research -- Differential Privacy Expertise
Christine Task (NIST Technical Lead)

topcoder -- Phase II Challenge Platform
Ward Loving (Project Manager),
Sergey Pogodin (Technical Lead)

HeroX -- Phase I Challenge Platform
Kyla Jeffrey (Project Manager)

Research Topic: Differentially Private Synthetic Data
Start Date: May 2018
Application: The First National Challenge in Differential Privacy

As technical lead for the **NIST Differential Privacy Synthetic Data Challenge**, Knexus is providing technical guidance for the first national challenge in differential privacy. Developments coming out of this competition are expected to drive major advances in the practical applications of differential privacy for contexts such as public safety.

Winners from Match #2 will be announced this week

Match #3 begins Next Week, on March 10th 2019. Registration is open now!

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

NIST Challenge Motivation

PSCR Needs Data Analysis:

The **Public Safety Communications Research Division (PSCR)** of the **National Institute of Standards and Technology (NIST)** is sponsoring the Differential Privacy Synthetic Data Challenge to help advance research for public safety communications technologies for America’s First Responders

PSCR OPEN INNOVATION

Solving the Public Safety needs of tomorrow...today

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.

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. Differentially private synthetic data sets can be safely released to the public, **allowing state and local public safety departments to leverage the power of crowd-sourced analysis to understand and improve their systems**.

NIST Challenge Motivation

KNEXUS
RESEARCH CORPORATION

Tech Challenges Have Advantages:

Challenges provide researchers with a visible, open and accessible, **shared pathway from theory to practice**.

Challenges grab attention. They **educate the public and potential investors** about new technological possibilities, inviting the audience to follow along in the excitement as those possibilities are fulfilled.

Challenges often precede significant acceleration in the development of **commercial products** for new tech.

DARPA URBAN CHALLENGE

Google LUNAR XPRIZE

COLLEGIATE WIND COMPETITION
U.S. DEPARTMENT OF ENERGY

kaggle

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

Overview of the NIST Differential Privacy Synthetic Data Challenge:

- The Challenge began in the summer of 2018 with a concept-building phase where contestants submitted concept papers proposing a mechanism to enable the protection of personally identifiable information while maintaining a dataset's utility for analysis.
- The second phase consists of a sequence of empirical matches throughout fiscal year 2019, where participants with implemented systems compete to produce high quality synthetic data from real data sets.
- This is the **first national challenge** in differential privacy, but ***it's already not the last***.
- The "**data challenge**" format is a well-established archetype, and for public accessibility it makes sense to echo this format… but hosting a ***differentially private*** data challenge requires some **non-trivial adaptations**.
- For the rest of this talk, we'll discuss the complications we encountered and how we chose to address each of them.   These are certainly not the only possible solutions, and we invite your feedback and input. In general, we feel these **are important questions for the community to consider.**

NIST Challenge *Challenges*: Contest Procedure

Objective: Create a shared, *competitive* benchmarking process for Differentially Private Synthetic Data Generators.

Constraints: We… haven't really had one of those before. Contestants will be learning about the behavior of their solutions at the same time we do.

Conclusion: The contest needs to support teams iteratively evaluating, exploring and then refining and resubmitting their solutions.

Solution details: A Sequence of Three topcoder Marathon Matches, of increasing difficulty. Each Match has five weeks of **provisional leaderboard** scoring on submitted synthetic data sets, followed by three weeks rigorous **sequestered evaluation** of executable systems and source code, followed by a winners announcement and awarding of prizes.

To support iterative refinement without violating differential privacy, we assume provisional data is public, and keep sequestered data private (using data sets from different years/areas, with different individuals and different distributions).

NIST Challenge *Challenges*: Differential Privacy Validation

Objective: Prevent 'cheating' from (often unintentional) violations of Differential Privacy, which will generally result in high accuracy scores. DP validation must keep the leaderboard reasonably reliable during the provisional phase, and make every effort during the sequestered phase to ensure prizes are only awarded to valid solutions.

Constraints: Even though probabilistic black box DP verification systems exist, we didn't have the time or budget to implement and adapt them to our needs on this particular project.

Conclusion: SME Review Panel. Relying on human resources instead, we needed to be as efficient as possible, and considerate of volunteers' time.

Solution details:

Provisional Phase**:** To earn a **1000x score boost**, contestants must submit clear, complete privacy proofs to pass a **Differential Privacy Prescreen**, occurring as a weekly SME review telcon. The prescreen is a quick check to ensure the contestant is making a good faith effort to satisfy differential privacy and there are no obvious errors.

Sequestered Phase: Invited contestants submit source code, code guide/documentation, updated algorithm specification and privacy proof for a thorough **Final Differential Privacy Validation** by the SME review team. Solutions failing validation are eliminated from prize eligibility.

NIST Challenge *Challenges*: Data and Scoring

Objective: Select the **data sets** to use as the basis for the contest.

Constraints: Data should be relevant to PSCR's applications, of interest to the public audience generally, reflect current synthetic data needs, publicly available (to avoid access restrictions), and it should start out as an achievable objective (not too many variables, not too many values per variable, not too complex correlations)… and then get harder.

Conclusion: **Event Data and Survey Data** (not time series data, transaction data, or image/audio data this time)

Solution Details:

Match #1 and Match #2: San Francisco Fire Event Data. Over a decade of data with features such as priority, response time, location, unit type, etc.

Match #3: [redacted]

NIST Challenge *Challenges*: Data and Scoring

Objective: A functional definition of whether one synthetic data set is ’better’ than another.

Constraints: It has to run very efficiently, be fair, reasonably data independent, and we don’t need only one… we need at least three *non-redundant* metrics, in order to increase the rigor of the scoring across each of the three matches. It should also capture the needs and preferences of the data user community.

Conclusion: Randomized Heuristics! Each match **adds a new scoring metric to the existing set.**

Solution Details:

Provisional Leaderboard Scoring is done on three submitted synthetic data-sets, generated at three specified values of epsilon, with the resulting three scores averaged together.
Sequestered Final Scoring is done with repeated trials, additional values of epsilon as needed, and final score is computed a privacy/accuracy AUC (Area Under Curve)

Match #1: Randomized, normalized 3-Marginal based distance metric
Match #2: 3-Marginal, and randomized Row-pool similarity based metric
Match #3: 3-Marginal, Row-pool, and [redacted]

[Match #1 & #2 scoring metrics developed by Sergey Pogodin]

NIST Challenge *Challenges*: Algorithms That Exist As Software

Objective: NIST is part of the US Department of Commerce. NIST provides guidance that helps US corporations address technical needs. As a vital outcome of this contest, we would like to have stable, usable, well-engineered (ideally open sourced) software solutions that can be further evaluated by NIST experts, contributing towards NIST's efforts to issue official guidance on DP Synthetic Data.

Constraints: Teams may begin with academic prototypes that have only been used inside their research groups, to generate results for specific research papers.

Conclusion: Make initial participation in the match accessible for research prototypes, and increase code requirements over the course of the match.

Solution Details: Each Match begins with minimal software requirements (**simply submit correctly formatted synthetic data sets to earn a provisional leaderboard score**), and these are increased throughout the match: Invitees to the sequestered round must have standardized delta/epsilon input, no hardcoded data schemas (schema given as input), and thorough code documentation aligned with algorithm documentation. Their solutions then undergo source code review by multiple SME, and their docker containers are run by the TC tech lead—If either encounter problems, they are informed and may be able to fix and resubmit. Prize-winners leave the match with money, but also with an **externally evaluated code base** that will be more easily shared, tested, and used by other researchers, potentially forming a stable basis for future production-level solutions. ***Participate in the contest and we'll provide a free two-month bootcamp for your DP Synthetic Data solution.***

NIST Challenge Match #3

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

[redacted]

[redacted]

[redacted]

NIST Challenge Match #3

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

Data:

Provisional Phase Data—1940 Census Persons Level Data for Colorado

Sequestered Phase Data—1940 Census Persons Level Data for [redacted]

[redacted]

[redacted]

NIST Challenge Match #3

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

Data:

Provisional Phase Data—1940 Census Persons Level Data for Colorado

Sequestered Phase Data—1940 Census Persons Level Data for [Not Colorado]

[redacted]

[redacted]

NIST Challenge Match #3

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

Data:

Provisional Phase Data—1940 Census Persons Level Data for Colorado

Sequestered Phase Data—1940 Census Persons Level Data for [Not Colorado]

Scoring:

- To Catch **Long Tails**: Income Inequality based metrics
- To Catch Degradation of Accuracy over **Differences of Differences**: Pay-gap based metrics

[redacted]

NIST Challenge Match #3

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

Data:

Provisional Phase Data—1940 Census Persons Level Data for Colorado

Sequestered Phase Data—1940 Census Persons Level Data for [Not Colorado]

Scoring:

- To Catch **Long Tails**: Income Inequality based metrics
- To Catch Degradation of Accuracy over **Differences of Differences**: Pay-gap based metrics

Final Outcomes:

- $62K in prizes for Match #3!
- $4K bonus for top 5 challenge winners who provide their full solution in an open source repository for use by all interested parties , under a license such as BSD or Apache 2.0
- Further NIST research, as a metrology lab, to identify and establish metrics and methods for evaluating synthetic data
- Goal of disseminating the lessons learned and approaches taken, through journal special issue, conferences, workshops, talks. Challenge participants will be invited to contribute

NIST Challenge Match #3

KNEXUS RESEARCH CORPORATION

Match #3 Begins Next Week, 3/10/19, Registration Is Open Now:

Data:

Provisional Phase Data—1940 Census Persons Level Data for Colorado

Sequestered Phase Data—1940 Census Persons Level Data for [Not Colorado]

Scoring:

- To Catch **Long Tails**: Income Inequality based metrics
- To Catch Degradation of Accuracy over **Differences of Differences**: Pay-gap based metrics

Final Outcomes:

- $62K in prizes for Match #3!
- $4K bonus for top 5 challenge winners who provide their full solution in an open source repository for use by all interested parties, under a license such as BSD or Apache 2.0
- Further NIST research, as a metrology lab, to identify and establish metrics and methods for evaluating synthetic data
- Goal of disseminating the lessons learned and approaches taken, through journal special issue, conferences, workshops, talks. Challenge participants will be invited to contribute

Register Here: https://www.topcoder.com/community/data-science/Differential-Privacy-Synthetic-Data-Challenge
(…or just google "NIST Differential Privacy Challenge")

KNEXUS RESEARCH CORPORATION

NIST Challenge Motivation

KNEXUS
RESEARCH CORPORATION

PSCR Research Priorities:

PSCR
OPEN
INNOVATION

Solving the Public Safety needs of tomorrow...today

NIST's **Public Safety Communications Research** division has strong commitments to both **public safety research** and the preservation of security and privacy, including the use of de-identification.

It is well known that privacy in data release is an **important area for the Federal Government (which has an Open Data Policy), state governments, the public safety sector and many commercial non-governmental organizations**. Developments coming out of this competition would hopefully drive major advances in the practical applications of differential privacy for these organizations.

PSCR is sponsoring this exciting data science competition to help advance research for public safety communications technologies for **America's First Responders**

KNEXUS
RESEARCH CORPORATION

Questions?

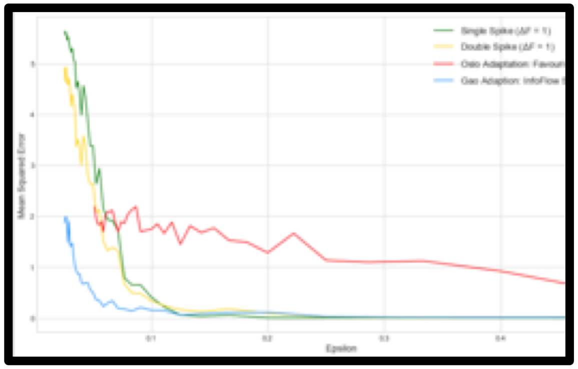
…and this is what all we've been up to at Knexus Research.

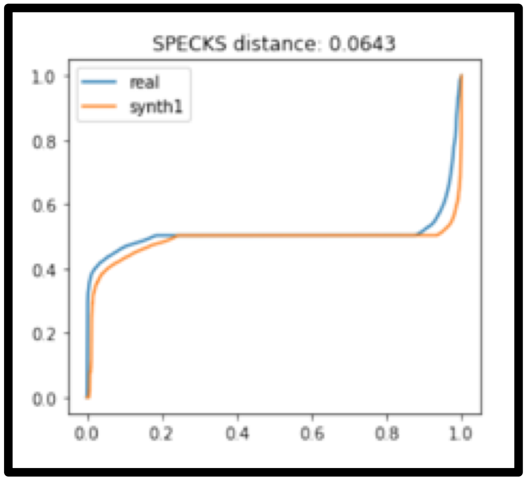
Knexus Research

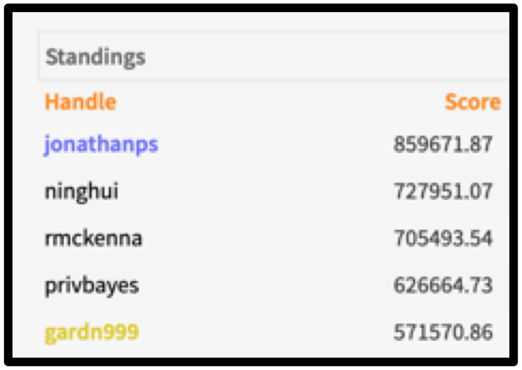
Privacy Team: Christine Task, Micah Heineck, Jason Suagee, Christine Heiss, Karan Bhagat, Joe Graus, Kevin Raoofi, Jonathan Woodell, Konrad Rauscher, Jeffrey Hodges, Destiny Ridguard, Kylie Berry

Research Topics: Synthetic Data, Differential Privacy, Noise-resistant Analytics

Privacy Contact Email: christine.task@knexusresearch.com







Standings

Handle	Score
jonathanps	859671.87
ninghui	727951.07
rmckenna	705493.54
privbayes	626664.73
gardn999	571570.86

Simons Institute: Foundations to Applications Workshop 3/2019