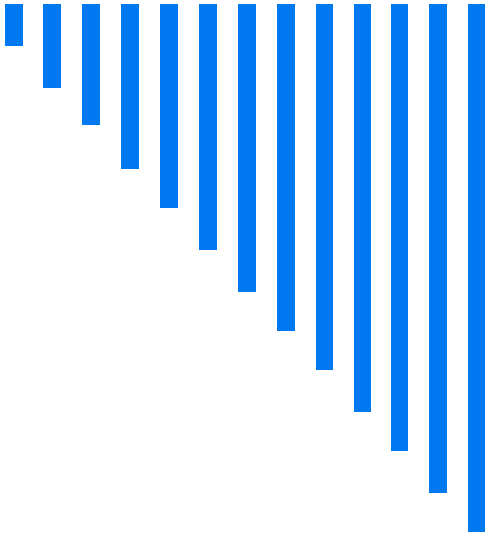
S. Raj Rajogopalan



Using Rate Distortion Theory for Privacy-Utility Tradeoffs in Databases

Lalitha Sankar†, S. Raj Rajagopalan*, H. Vincent Poor†

†Princeton University,
***HP Labs, Princeton (raj.raj@hp.com)**

Motivation: How to maximize privacy relative to a utility requirement?

Shannon's Rate Distortion Problem

- For a source X with a distribution p_X, alphabet $\mathcal{X}$, and reconstructed alphabet $\mathcal{X}'$, determine the minimum information rate R that guarantees a given average distortion (fidelity) between X and X'.
- For n samples of the source, average (per observation) distortion defined as

$$\Delta_d \equiv \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g\left(f(X_i), f(X_i')\right)\right] \leq D + \varepsilon$$

  - g : distance-based function (e.g.: Hamming, Euclidean, K-L divergence)
- Rate R is the number of information bits per observation revealed about X. (X can be a vector).

Adding a Privacy Constraint

- Privacy as equivocation per observation

$$\Delta_p \equiv \frac{1}{n} H\left(X^n \mid X'^n\right) > E - \varepsilon$$

  - D : distortion upper bound; E : equivocation lower bound
  - $\varepsilon \to 0$ as $n \to \infty$
  - Generalizes to equivocation of sets of private variables conditioned on sets of public variables
  - Can also introduce auxiliary (side) information to model external knowledge
- The RDE problem [Yamamoto 1982]:
  - Find an IT-code for a source (X,Y) with minimum rate that satisfies given distortion (on X) and equivocation (on Y) constraints

Model Database as a Source

- Database d with n rows, each row with k numeric attributes, is a sequence of n i.i.d. samples of $\boldsymbol{X}=[X_1\ X_2\ \ldots\ X_k]$ with the distribution (X_k: k^{th} attribute)

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1 X_2 \ldots X_K}(x_1, x_2, \ldots, x_k)$$

- $\mathcal{K}_r$: public (revealed) and $\mathcal{K}_h$: private (hidden) subsets of attributes and $\boldsymbol{X}$ has a revealed precision c (per sample)

- The Utility-Privacy (U-P) Problem:
  - Construct d' from d so that a user can reconstruct $\mathcal{K}_r$ but not $\mathcal{K}_h$ to the desired level of accuracy

Mapping U-P to R-D-E

Application Requirements	Principle/Abstraction
Utility	Distortion/Fidelity Functions
Privacy	Equivocation
Precision	Rate
Perturbation Technique	Information-theoretic Source Code

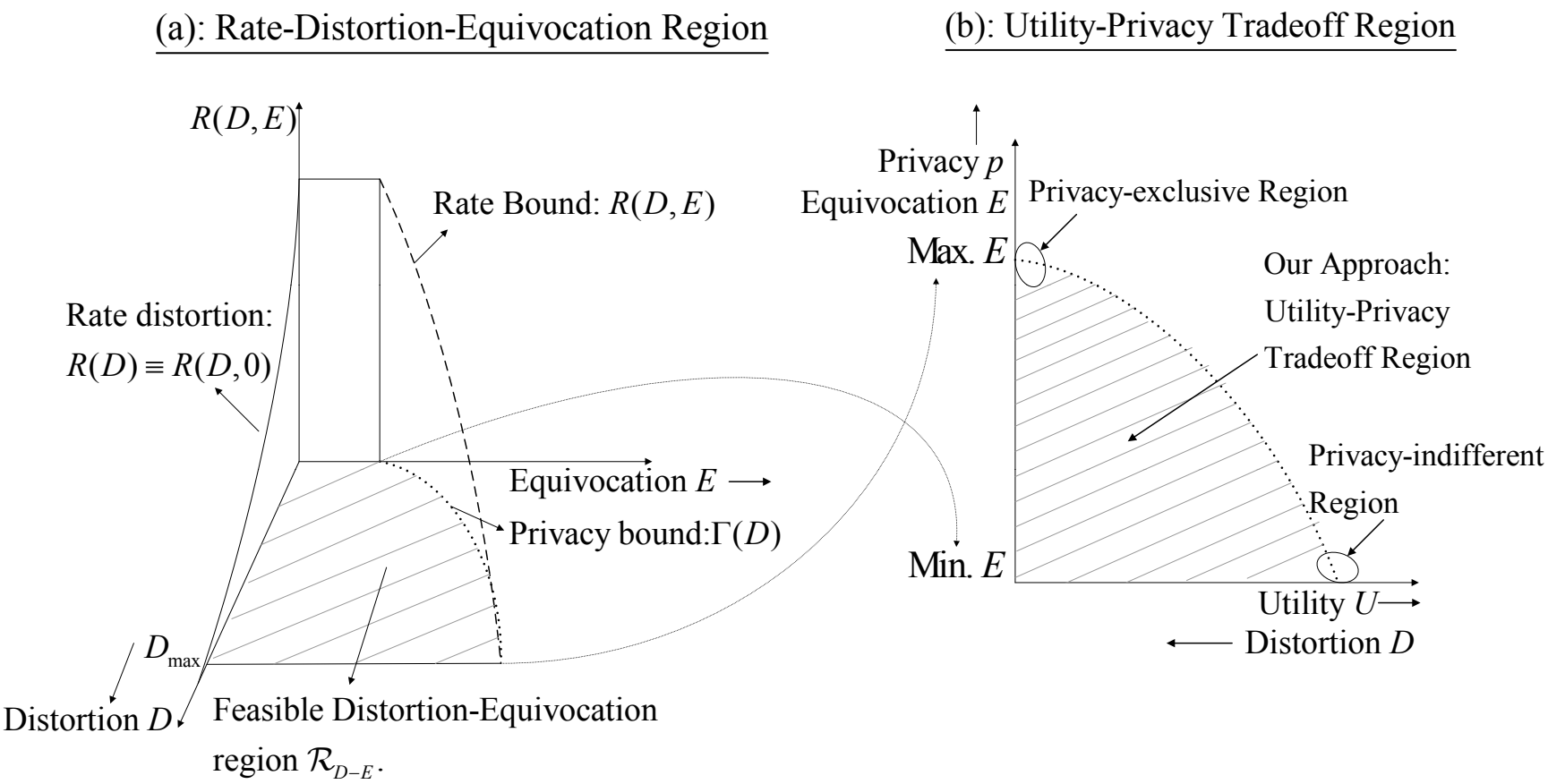
- Fidelity and Utility are measures of ‘closeness’ of original and published/revealed sources
- Equivocation and Privacy are measures of ‘uncertainty’ about private data given revealed data

Database Sanitization via Source Coding

- Map d ($\mathbf{X}^n$) to a sanitized database (SDB) d' by encoding public variables
- RDE achievable region $\mathcal{R}_{D\text{-}E}$ consists of feasible triples (R,D,E), equivocation $\Gamma(D)$ as a function of E, and rate R as function of D and E.

- Theorem: For a database with a given distribution, the set of feasible utility, privacy, precision triples is given by the corresponding $\mathcal{R}_{D\text{-}E}$.

RDE and Utility-Privacy Regions



- Corollary: A given utility constraint tightly constraints the maximum achievable privacy.

Strengths and Weaknesses

- Strengths
  - Analytic abstract measures of utility and privacy
  - Tradeoff region between utility and privacy for some standard source distributions
  - Model extensible to handle third party data and some multiple queries
- Challenges
  - Source distribution needs to be known, rows are iid
  - Utility and Privacy metrics are “on average” metrics – weaker results may be possible for stronger metrics
  - Concrete map from application-specific utility requirement needs to be built

Gerome Miklau

Optimizing Linear Queries Under Differential Privacy

Database is represented by counts: x_1, x_2, ... x_n

Given: a workload W of linear queries: each $w = c_1x_1 + c_2x_2 + ... + c_nx_n$

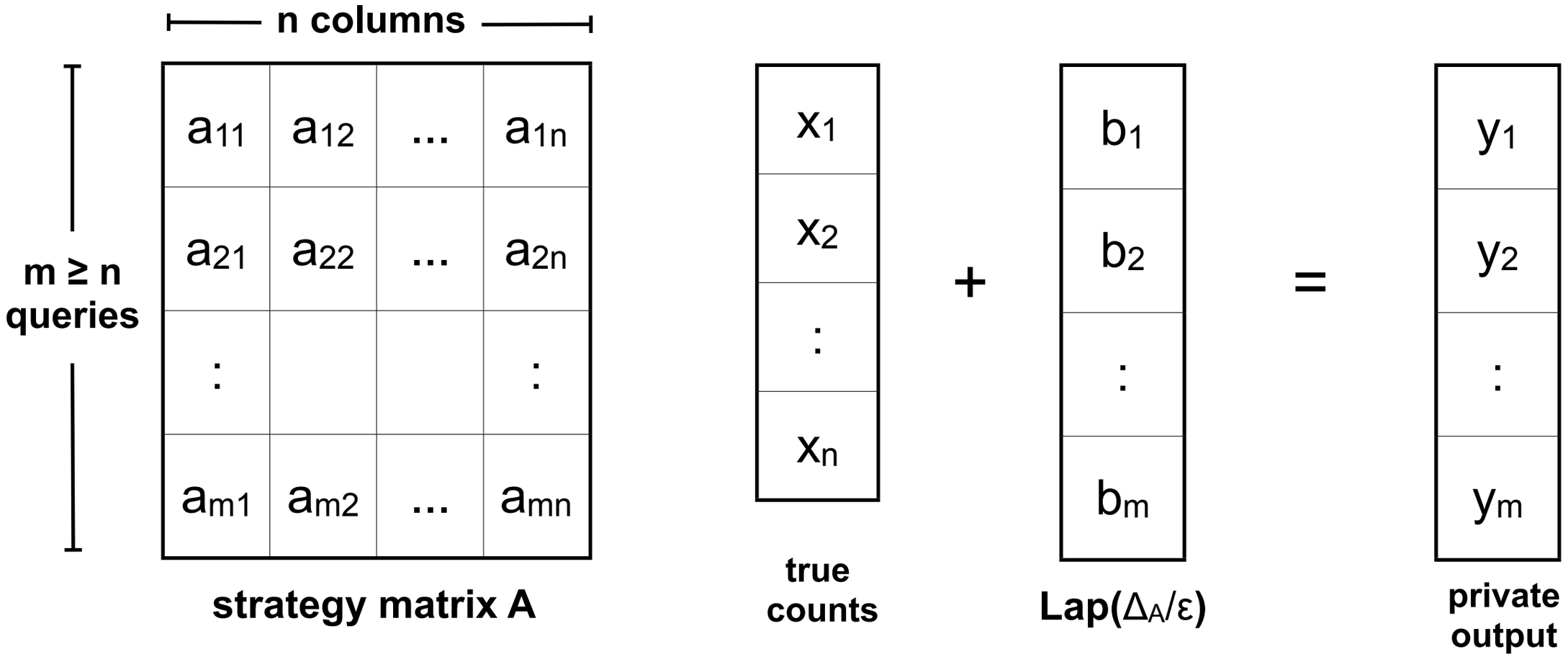
Goal: a set of linear queries A to answer W with least error.

Optimizing Linear Queries Under Differential Privacy

Database is represented by counts: $x_1, x_2, \ldots x_n$

Given: a workload W of linear queries: each $w = c_1x_1 + c_2x_2 + \ldots + c_nx_n$

Goal: a set of linear queries A to answer W with least error.

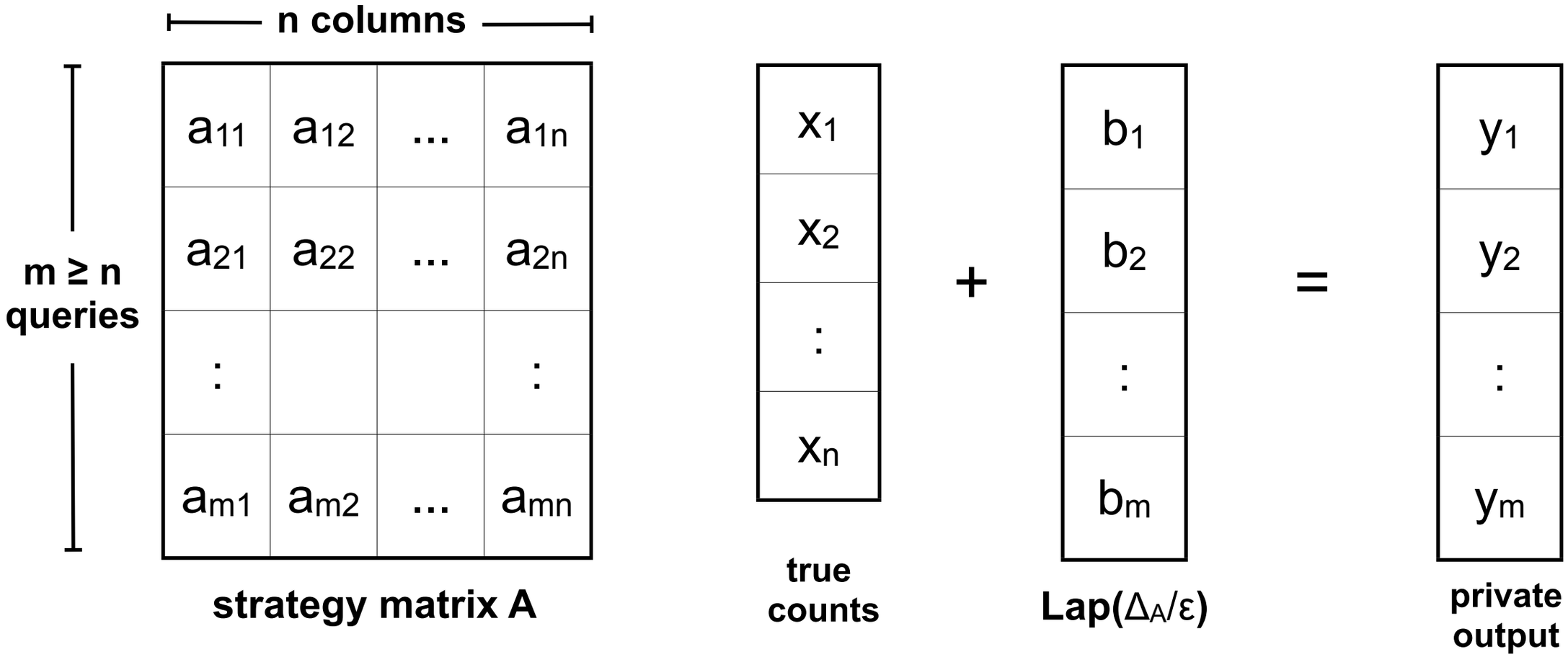


Optimizing Linear Queries Under Differential Privacy

Database is represented by counts: $x_1, x_2, \ldots x_n$

Given: a workload W of linear queries: each $w = c_1x_1 + c_2x_2 + \ldots + c_nx_n$

Goal: a set of linear queries A to answer W with least error.



Deriving answers

If m=n: estimate $\bar{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{y}$

If m>n: estimate $\bar{\mathbf{x}} = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\,\mathbf{y}$

Derived answer for **w** is $\mathbf{w}\bar{\mathbf{x}}$

Computing error

$$\text{Error}(\mathbf{w}\bar{\mathbf{x}}) = (2/\varepsilon^2)\,\Delta_A^2\,\mathbf{w}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{w}^t$$

Example: answering interval queries

Workload

1	1	1	1
1	1	1	0
0	1	1	1
1	1	0	0
0	1	1	0
0	0	1	1
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Alternative strategies

Identity

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

$\Delta=1$

Hierarchical

1	1	1	1
1	1	0	0
0	0	1	1
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

$\Delta=\log_2 n$

Wavelet

1	1	1	1
1	1	-1	-1
1	-1	0	0
0	0	1	-1

$\Delta=\log_2 n$

Example: answering interval queries

Workload

1	1	1	1
1	1	1	0
0	1	1	1
1	1	0	0
0	1	1	0
0	0	1	1
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Alternative strategies

Identity

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

$\Delta=1$

Hierarchical

1	1	1	1
1	1	0	0
0	0	1	1
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

$\Delta=\log_2 n$

Wavelet

1	1	1	1
1	1	-1	-1
1	-1	0	0
0	0	1	-1

$\Delta=\log_2 n$

Max error on workload:

$O(n)$

$O(\log^3 n)$

$O(\log^3 n)$

$$\text{Error}(\mathbf{w}\bar{\mathbf{x}}) = (2/\varepsilon^2)\,\Delta_A^2\,\mathbf{w}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{w}^t$$

Singular value decomposition of A

$$\mathbf{A} = \mathbf{QDP}$$

Joint work with:

Michael Hay
Chao Li
Andrew McGregor
Gerome Miklau

University of Massachusetts Amherst

Vibhor Rastogi
Dan Suciu

University of Washington

Details here:

Boosting the accuracy of differentially-private queries through consistency.
Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu
To appear, Proceedings of the VLDB Endowment (PVLDB), 2010.

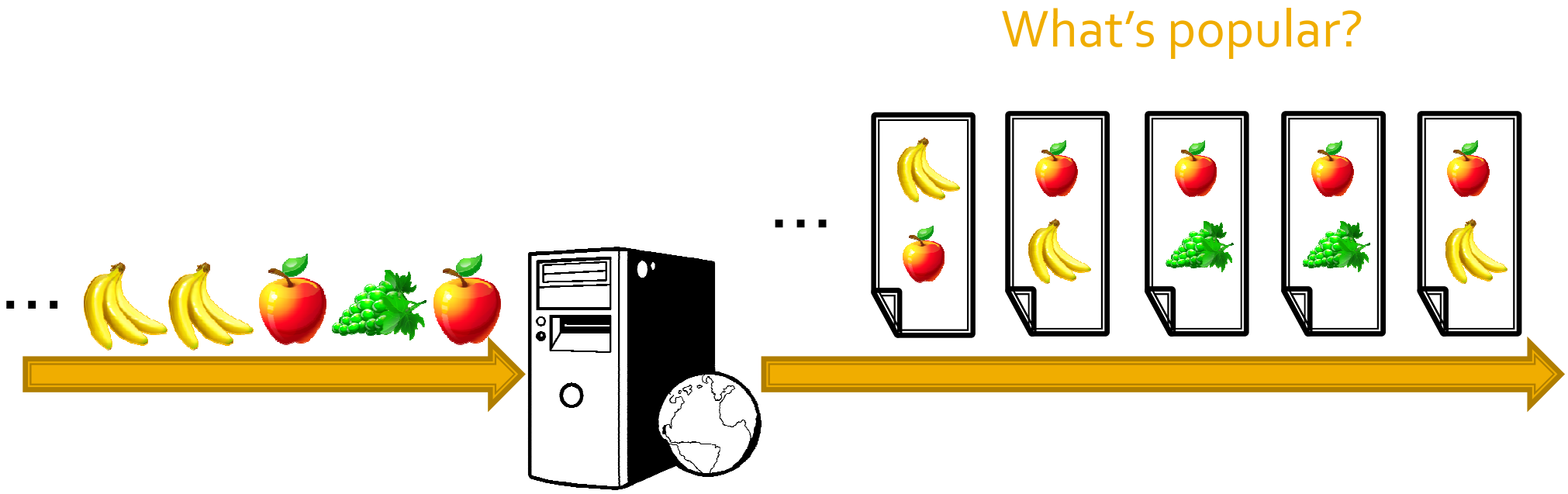
Optimizing Histogram Queries Under Differential Privacy
Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor
ArXiv Preprint, abs/0912.4742 2009

Elaine Shi

Hubert Chan (HKU)
Elaine Shi (PARC)
Dawn Song (Berkeley)

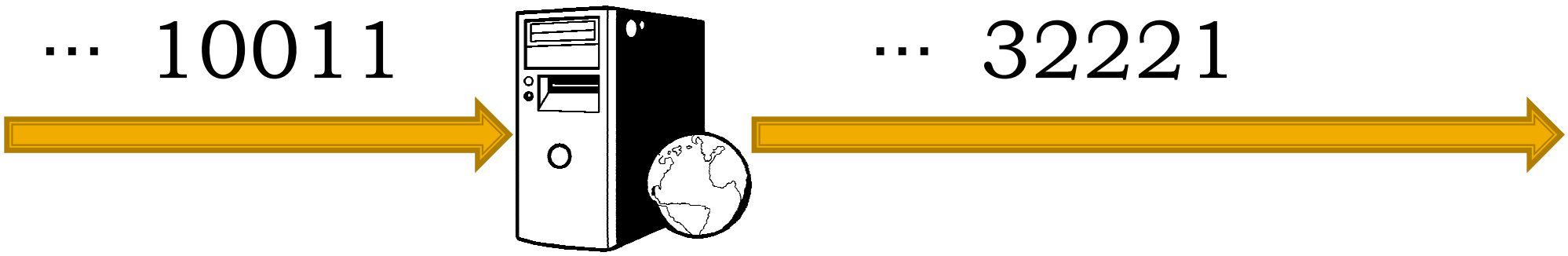
Private and Continual Release of Statistics

Continual Setting



Continual Counter

Number of 1's seen thus far



Main Result

- Privacy: ε-differentially private
- Utility:

$$O(\frac{1}{\varepsilon}) \cdot (\log t)^{1.5} \cdot \log(\frac{1}{\delta}) \text{ error}$$

with prob $1-\delta$

- Naïve scheme: $O(\sqrt{t})$ error

[Dwork et al.] -- STOC10

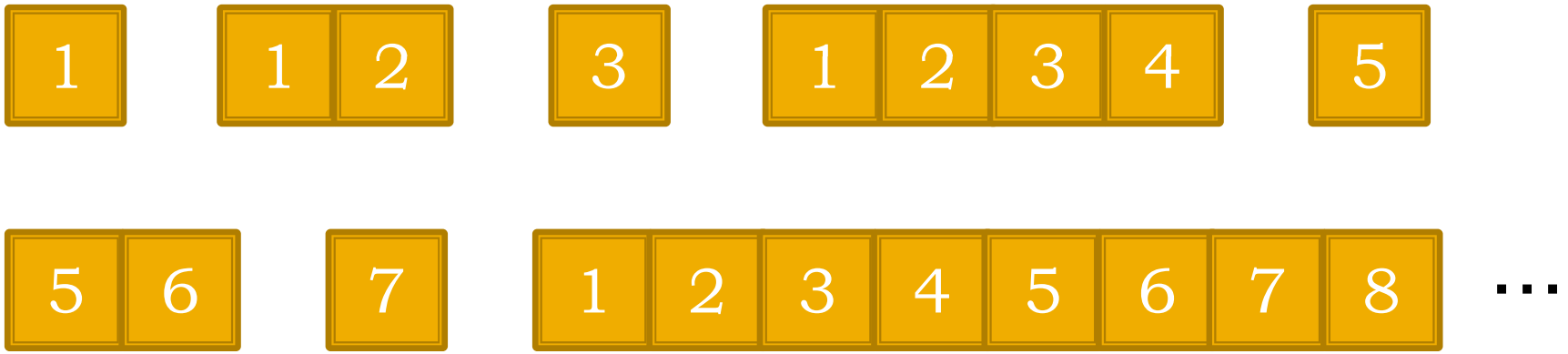
Also in the paper

- http://eprint.iacr.org/2010/076
- Pan privacy
- Applications

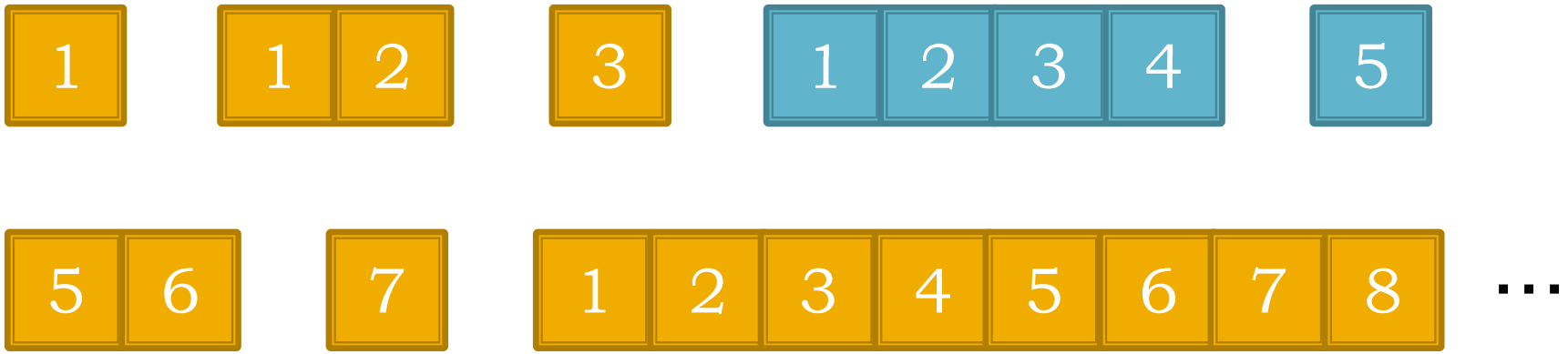
eshi@parc.com

Thank you!

Technique

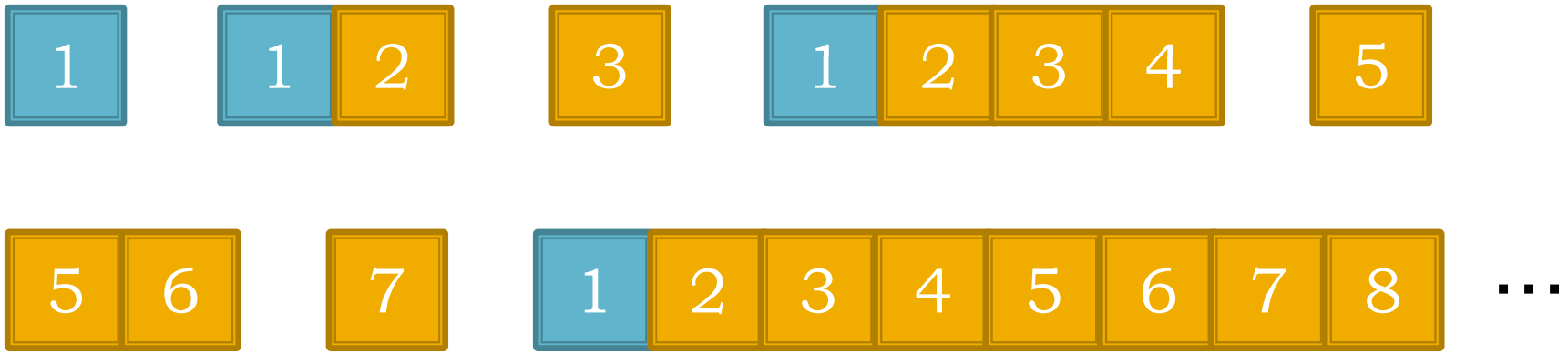


Technique



- Utility: Each count is the sum of O(log t) blocks

Technique



- Utility: Each count is the sum of O(log t) blocks
- Privacy: Each bit appears in O(log t) blocks

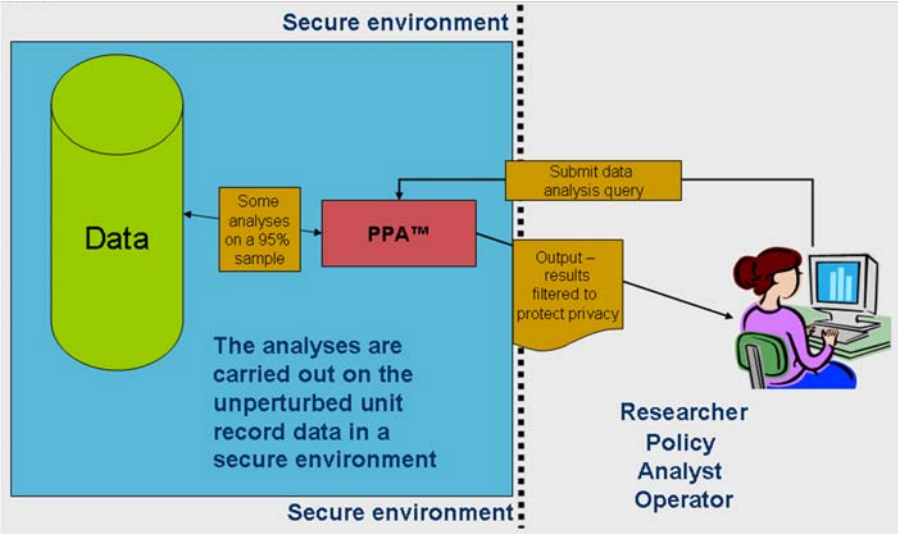
Christine O'Keefe

Confidentialising the output of a remote analysis server: Privacy-Preserving Analytics

Christine M O'Keefe, CSIRO Mathematics, Informatics and Statistics

Introduction

CSIRO's Privacy-Preserving Analytics is methods and software for analysing confidential data without compromising confidentiality.



Analyses Implemented to Date

Exploratory data analysis

Statistical modelling

- Generalised additive modelling
- Generalised linear modelling
- Mixed linear effects modelling
- Robust linear modelling
- Time series modelling

Survival Analysis

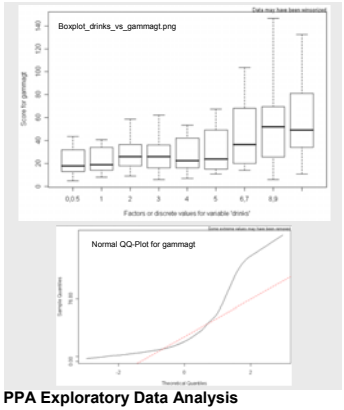
- Cox proportional hazards modelling
- Kaplan-Meier fitting
- Parametric survival modelling

Clustering

- k-means

Confidentialisation Measures – Regression

- Restricted access
  - Some analyses not permitted
- Restricted data
  - Sometimes a 95% sample is used
- Restricted queries
  - Control range of analyses permitted
  - Control transformations and interactions levels
- Restricted output
  - Confidentialisation of output of analyses



PPA Exploratory Data Analysis

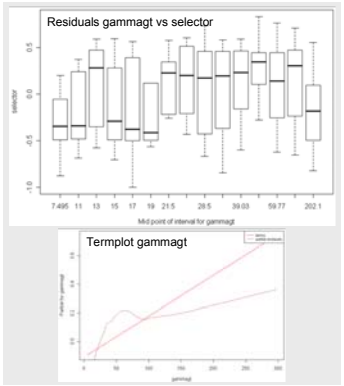
Analysis of Deviance Table
Model: gaussian, link: identity, terms: sequential
Response: selector

	Df	Deviance	Resid. Df	Resid. Dev
NULL			341	83.368
mcv	1	0.699	340	82.669
alkphos	1	0.998	339	81.671
sgpt	1	0.019	338	81.652
sgot	1	7.317	337	74.335
gammagt	1	2.003	336	72.332
factor(drinks)	15	7.811	321	64.5

Coefficients:

	Estimate	Pr(>|t|)
(Intercept)	2.912	p<0.005 ***
mcv	-0.013	0.01<p<0.05 *
alkphos	-0.004	p<0.005 **
sgpt	-0.009	p<0.005 ***
sgot	0.017	p<0.005 ***
gammagt	0.003	p<0.005 ***
factor(drinks)0.5	-0.185	0.2<p<0.5
…		
factor(drinks)16	-0.853	0.01<p<0.05 *
factor(drinks)20	-0.752	0.01<p<0.05 *

PPA Regression



PPA Regression Diagnostics

Conclusion

- Privacy-Preserving Analytics is an example of a remote analysis server where the *output* has been confidentialised, not the *input*
- Remote servers are likely to play an important role in the future of data dissemination
  - There are still important technical challenges to be addressed

References

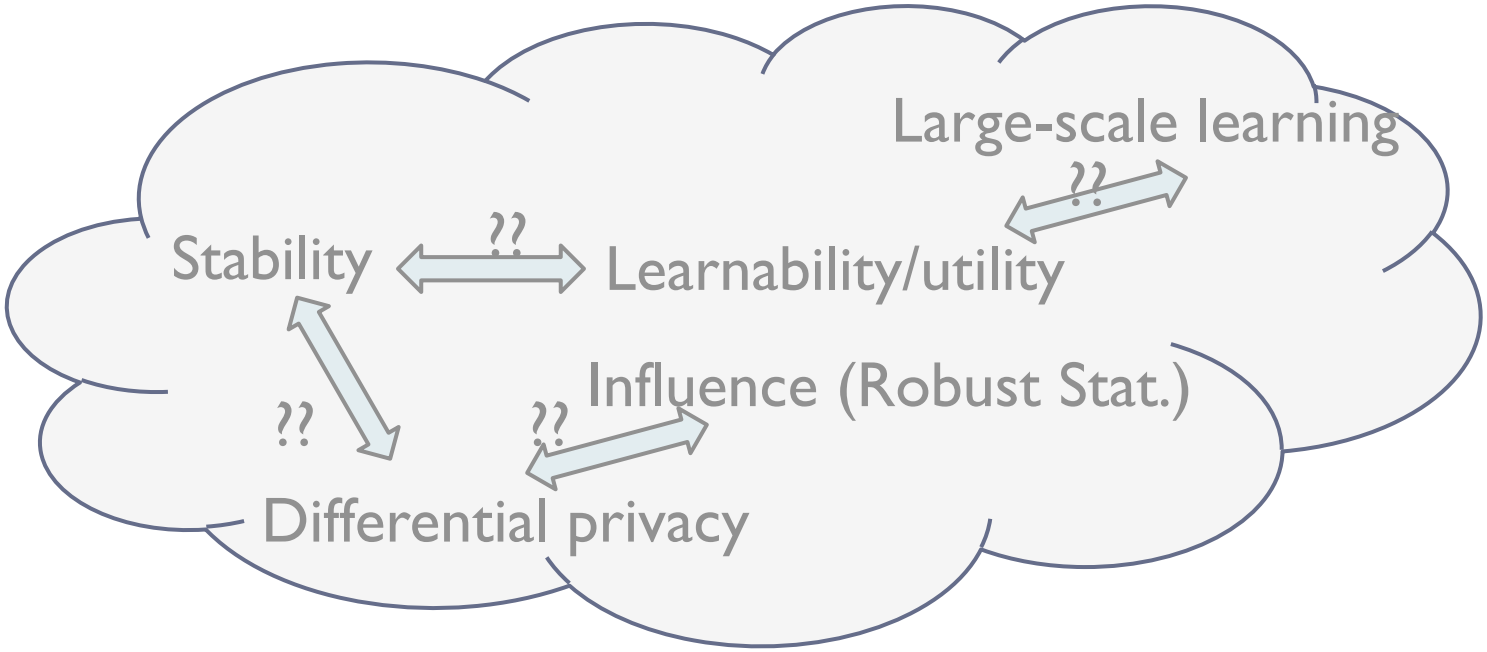
1. R. Sparks, C. Carter, J. Donnelly, C.M. O'Keefe, J. Duncan, T. Keighley and D. McAullay, Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics™, Comput Methods Programs Biomed., 91 (2008) 208-222.
2. C.M. O'Keefe, Privacy and the Use of Health Data - Reducing Disclosure Risk, electronic Journal of Health Informatics 2008; 3(1): e5.
3. C.M. O'Keefe and N.M. Good, Regression Output from a Remote Analysis Server, Data and Knowledge Engineering, 68 (2009), 1175-1186.
4. C.M. O'Keefe and N. Good, Risk and Utility of Alternative Regression Diagnostics in Remote Analysis Servers, Proceedings of the 56th Session of the ISI International Statistical Institute, 22-29 Aug 2007, Lisbon, Portugal.

National Research FLAGSHIPS CSIRO

Further information
contact: Professor Christine M O'Keefe
phone: +61 2 6216 7021
email: Christine.O'Keefe@csiro.au
web: www.csiro.au/org/CMIS.html
www.csiro.au

Eleanor Rieffel

Ben Rubinstein



Differentially Private SVMs:
Algorithmic Stability & Large-Scale Learning

Benjamin Rubinstein, CS Division, UC Berkeley

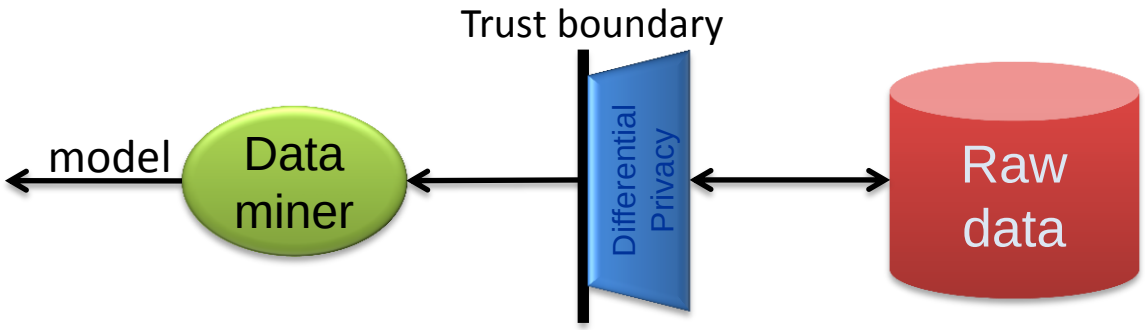
Poster Outline

- Algorithmic stability
  - Used in COLT to derive risk bounds
  - Property of learning map, not hypothesis class (like VC-dim)
- Large-scale learning
  - Techniques for dealing with large *n*
  - Often improving comp. complexity achieves regularization
- COLT'2010 submission – Rubinstein, Bartlett, Huang, Taft
  - Goal: release useful classifier while preserving data privacy
  - Mechanisms for SVM learning
    - Differential Privacy: via stability
    - Utility (L_∞-closeness of response & SVM whp): via large-scale learning
  - Lower bounds on diff. privacy achievable for useful mechanisms

Arik Friedman

Decision Tree Induction with Differential Privacy

Arik Friedman and Assaf Schuster
Technion, Israel Institute of Technology



Choosing an Attribute for splitting a node

1.Use noisy counts to approximate information gain (SuLQ [BDMN'05]):

$$V(A) = -\sum_{j \in A} \sum_{c \in C} -N^A_{j,c} \cdot \log \frac{N^A_{j,c}}{N^A_j}$$

2.Use the exponential mechanism with a query function based on a splitting criterion:

Splitting Criterion	Query function	Sensitivity
Information gain [Q'86]	$q_{IG}(T,A) = -\sum_{j \in A} \sum_{j \in C} \tau^A_{j,c} \cdot \log \frac{\tau^A_{j,c}}{\tau^A_j}$	$S(q_{IG}) = \log(\lvert T \rvert + 1) + 1/\ln 2$
Gini Index [BFOS'84]	$q_{GINI}(T,A) = \sum_{j \in A} \tau^A_j \left(1 - \sum_{c \in C} \left(\frac{\tau^A_{j,c}}{\tau^A_j} \right)^2 \right)$	$S(q_{GINI}) = 2$
Max (based on resubstitution estimate [BFOS'84])	$q_{Max}(T,A) = \sum_{j \in A} \left(\max_c (\tau^A_{j,c}) \right)$	$S(q_{MAX}) = 1$

Notation: T – a set of records, r_A and r_C refer to the values that record $r \in T$ takes on the attributes A and C respectively, $\tau^A_j = |\{r \in T : r_A = j\}|$, $\tau^A_{j,c} = |\{r \in T : r_A = j \wedge r_C = c\}|$. For noisy counts substitute N for τ.

Example – a single split

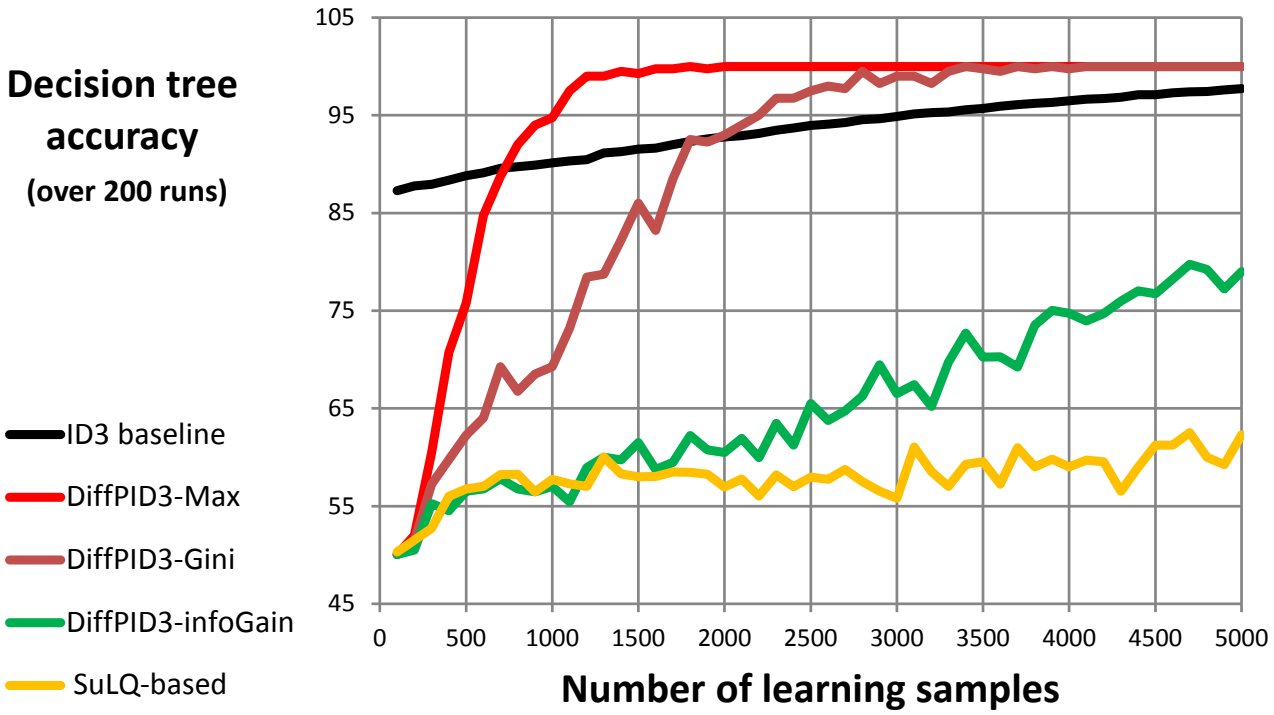


Figure 1. A single split: synthetic dataset with 10 binary attributes and a binary class, tree depth 1, ε=0.1, noise rate in learning data 0.1.

Sébastien Gambs

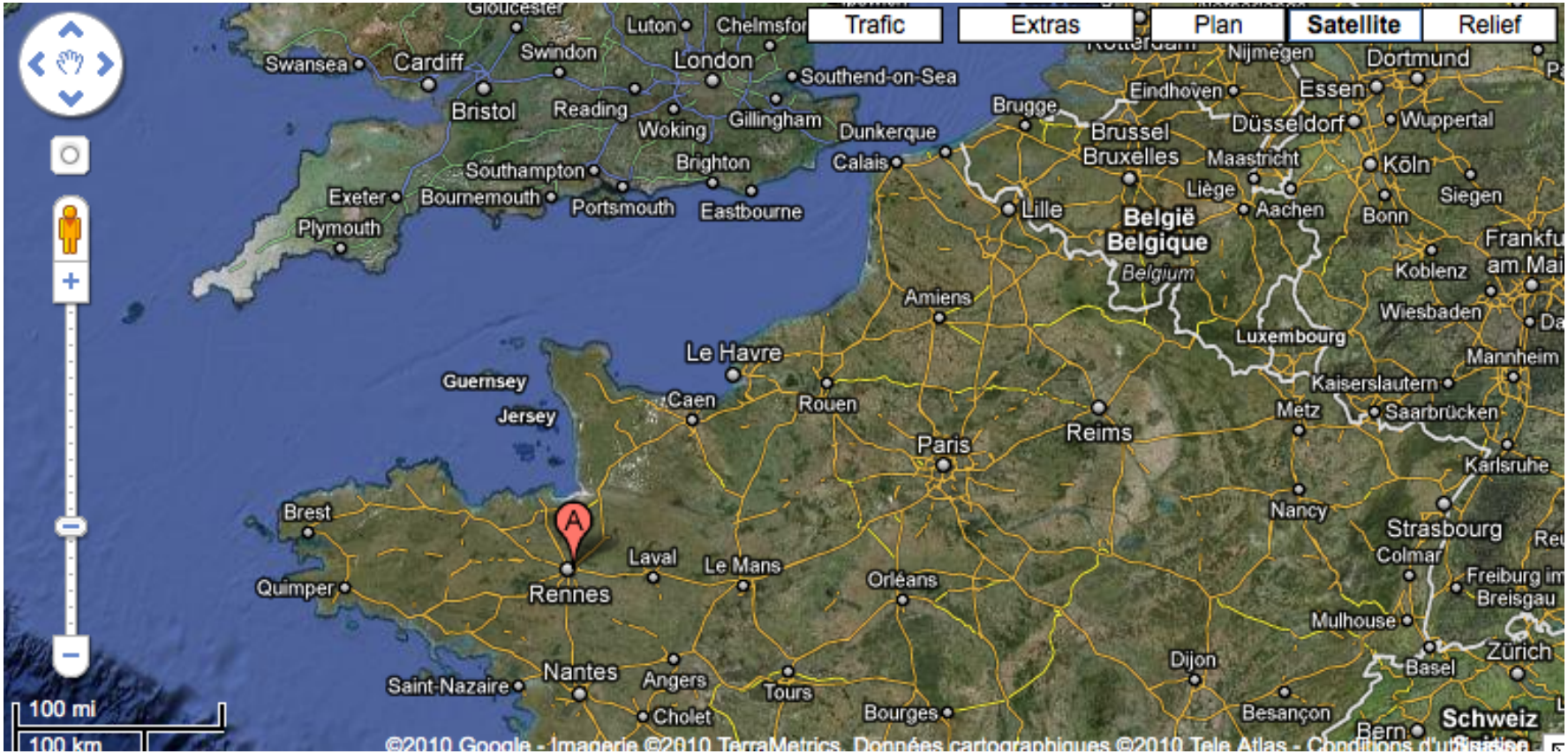
Defining and Quantifying Privacy for Geolocated Data

Sébastien Gambs

Université de Rennes 1 - INRIA / IRISA
sgambs@irisa.fr

February 2010

Where I am generally geolocated (when I am not away)



Geo-privacy

The main goal of geo-privacy is *to prevent an unauthorized entity from learning the past, current and future geographical location of an individual* (Beresford et Stajano 03).

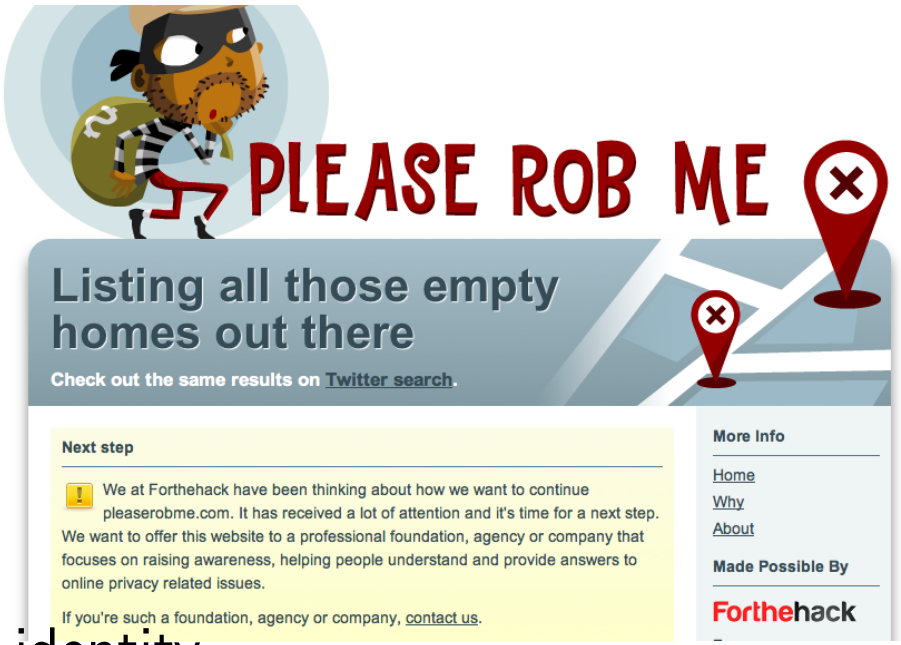
Among all the Personal Identifiable Information (PII), learning the location of an individual is one of the greatest threat against his privacy.

For instance, the locational data of an individual can be used to infer:

- his home and place of work,
- his identity,
- his center of interests,
- his habits or
- a deviation from his usual behaviour.

$\Rightarrow$ Privacy breach

From robbing your house . . .



to stealing your identity

Usurpation d'identité : plus de 210.000 victimes par an

AFP
06/10/2009 | Mise à jour : 11:59 Réactions(74)

Selon un rapport du Credoc, un Français a plus de chances de subir une usurpation d'identité qu'un cambriolage ou un vol de voiture. Le coût pour la société d'un tel phénomène frôle les 4 milliards d'euros.

C'est une enquête du Credoc, le Centre de recherche pour l'étude et l'observation des conditions de vie, qui l'affirme. Chaque année en France, plus de 210.000 personnes sont victimes d'une usurpation d'identité. Selon cette étude menée auprès d'un échantillon représentatif de la population française et présentée mardi, 4,2 % des personnes interrogées déclarent avoir été victimes d'une usurpation d'identité pendant les dix dernières années. «Cela représente plus de 210.00 cas avérés chaque année, un chiffre plus important que les cambriolages à domicile (150.000) et que les vols d'automobiles (130.000)»,

Some protection mechanisms for geolocated data

- Sanitization algorithms : pseudonymization, downsampling, geographical masks, removing records or adding fake ones, swapping, spatial cloacking (Gruteser and Grunwald 03), mix-zone (Beresford and Stajano 03), ...
- Remark: leads to a trade-off between the resulting level of privacy and the remaining utility of the sanitized data.
- Access-control methods,
- Secure multiparty computation,
- ...

How to define privacy for geolocated data?

- Fundamental interrogation : what does it mean to have a "good" preservation of privacy in a geolocated context?
- To be hidden inside a crowd gathered in a small area?

- To be alone in a desert?

- To have a behaviour indistinguishable from those of a non-negligible number of other individuals?

How to quantify privacy for geolocated data?

- Possible metric: measure how well an adversary can perform a particular on the sanitized vs unprotected version of the data?
- Example: identify with good confidence the home of individuals within the dataset.
- Does not take into account the auxiliary knowledge that the adversary might have.
- Avenue of research: derive the equivalent in geo-privacy of metrics coming from other domains (for instance privacy-preserving data mining).
- Crude global measure: mutual information between sanitized and original data.
- Question for the audience: natural extension of *differential privacy* to geolocated data?
- Other metrics?
- Interrogation: how to include the level of (un)linkability in the privacy measure?

Brief announcements

- I have a Postdoc position on this subject (under the INRIA recruitment campaign) and I am looking for talented candidates.
- I am also looking for a PhD candidate in the area of privacy-preserving data mining.
- There is a link to the descriptions of the two subjects on my website (http://www.irisa.fr/prive/sgambs/).
- Please contact me if you want more details or if you are interested.

Abhradeep Guha Thakurta

Differentially Private Ranking

Abhradeep Guha Thakurta
azg161@cse.psu.edu

Department of Computer Science
Pennsylvania State University

Joint work with Raghav Bhaskar and Srivatsan Laxman, Microsoft Research India and Adam Smith, Pennsylvannia State University

Problem Formulation

- **Ranking (generalizing [MT07])**
  - Consider a collection of elements $U = \{1, \cdots, u\}$.
  - Each element i has a real valued score $q_T(i)$ based on a data set T.
  - **Goal:** Output k elements with highest scores.
- **Privacy**
  - Data set T consists of n entries in domain $\mathcal{D}$.
  - **Differential privacy**: Protects privacy of entries in T.
- **Condition: Insensitive Scores**
  - for any element i, for any data sets T, T' that differ in one entry:

$$|q_T(i) - q_{T'}(i)| \leq 1 .$$

Problem Formulation

- **Ranking (generalizing [MT07])**
  - Consider a collection of elements $U = \{1, \cdots, u\}$.
  - Each element i has a real valued score $q_T(i)$ based on a data set T.
  - **Goal:** Output k elements with highest scores.
- **Privacy**
  - Data set T consists of n entries in domain $\mathcal{D}$.
  - **Differential privacy**: Protects privacy of entries in T.
- **Condition: Insensitive Scores**
  - for any element i, for any data sets T, T' that differ in one entry:

$$|q_T(i) - q_{T'}(i)| \leq 1 .$$

Problem Formulation

- **Ranking (generalizing [MT07])**
  - Consider a collection of elements $U = \{1, \cdots, u\}$.
  - Each element i has a real valued score $q_T(i)$ based on a data set T.
  - **Goal:** Output k elements with highest scores.
- **Privacy**
  - Data set T consists of n entries in domain $\mathcal{D}$.
  - **Differential privacy**: Protects privacy of entries in T.
- **Condition: Insensitive Scores**
  - for any element i, for any data sets T, T' that differ in one entry:

$$|q_T(i) - q_{T'}(i)| \leq 1 \,.$$

Definition (Differential Privacy)

A randomized algorithm $\mathcal{A}$ is ϵ-differentially private if for all data sets $T, T' \in \mathcal{D}^n$ differing in at most one entry and all events $\mathcal{O} \subseteq \mathit{Range}(\mathcal{A})$:

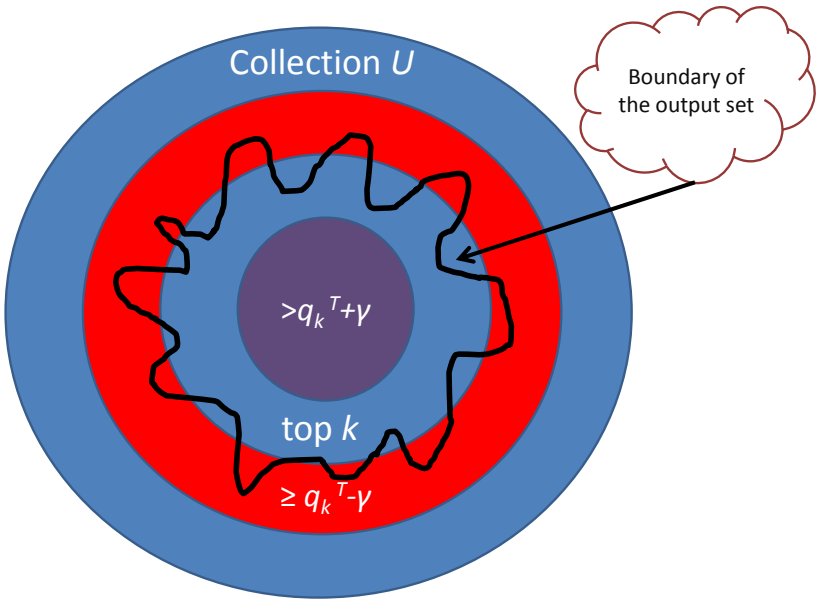
$$\Pr[\mathcal{A}(T) \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathcal{A}(T') \in \mathcal{O}] \,.$$

Approximate Ranking

Let q_k^T be the k^{th} highest score based on data set T.

An output list is γ-useful if:

- *(Soundness)* No element in the output has score less than $(q_k^T - \gamma)$.
- *(Completeness)* Every element with score greater than $(q_k^T + \gamma)$ is in the output.



Our Contributions: Two Algorithms

- **Score perturbation**
  - Perturb the scores of the elements with noise and then pick the top k elements in terms of noisy scores.
  - Faster and simpler implementation but slightly worse utility guarantee.
- **Exponential sampling**
  - Run the exponential mechanism [MT07] k times.
  - Slightly better utility guarantee but more complicated and slower implementation.
- In this talk we present the Score perturbation-based algorithm.

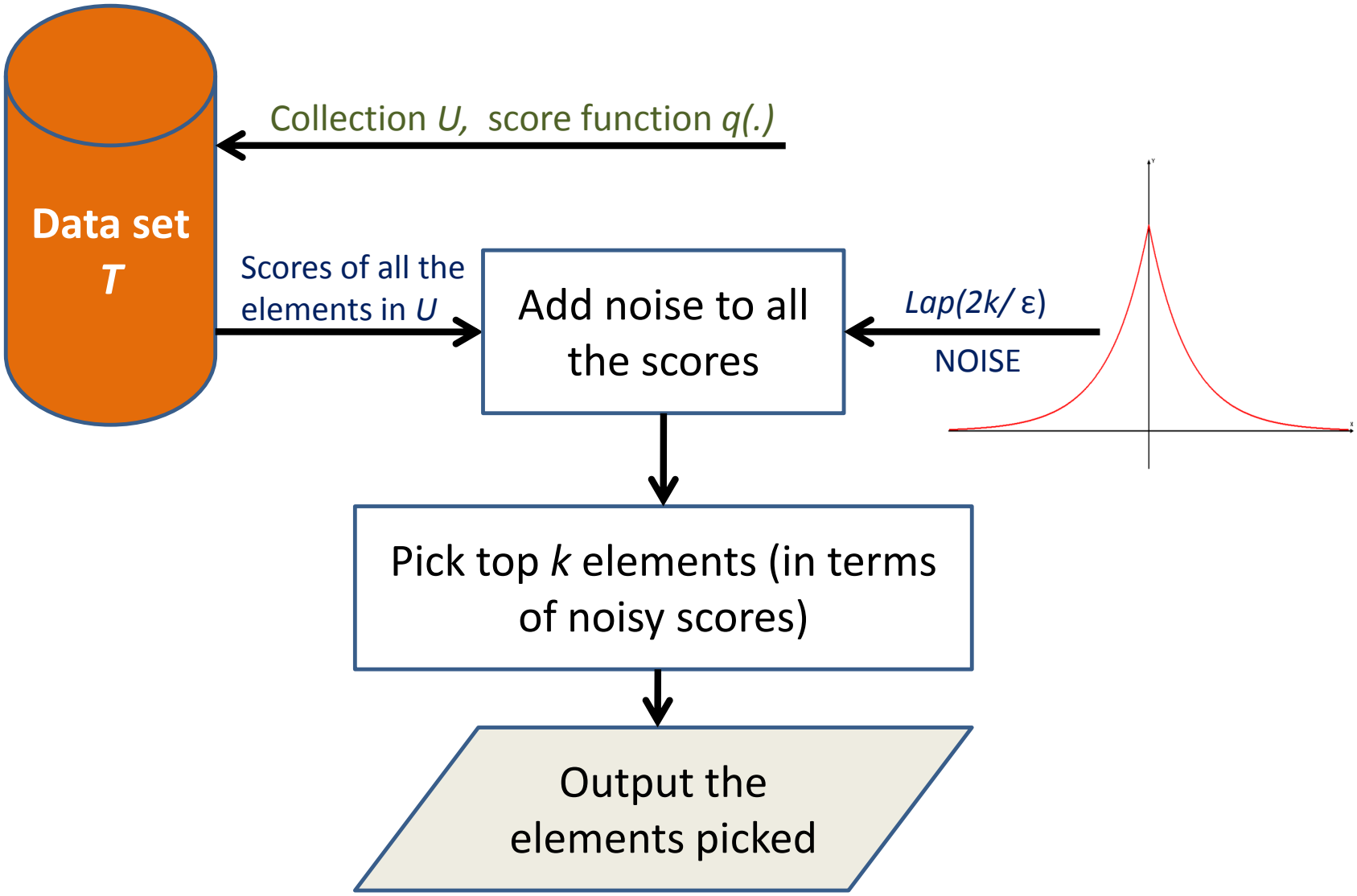
Our Contributions: Two Algorithms

- **Score perturbation**
  - Perturb the scores of the elements with noise and then pick the top k elements in terms of noisy scores.
  - Faster and simpler implementation but slightly worse utility guarantee.
- **Exponential sampling**
  - Run the exponential mechanism [MT07] k times.
  - Slightly better utility guarantee but more complicated and slower implementation.
- In this talk we present the Score perturbation-based algorithm.

Our Contributions: Two Algorithms

- **Score perturbation**
  - Perturb the scores of the elements with noise and then pick the top k elements in terms of noisy scores.
  - Faster and simpler implementation but slightly worse utility guarantee.
- **Exponential sampling**
  - Run the exponential mechanism [MT07] k times.
  - Slightly better utility guarantee but more complicated and slower implementation.
- In this talk we present the Score perturbation-based algorithm.

Score perturbation-based algorithm



Analysis (Privacy)

Theorem: *The algorithm is ϵ-differentially private.*

- **Naive analysis:** Requires $\Theta\left(\frac{u}{\epsilon}\right)$ noise for ϵ-differential privacy.
- **Our analysis:** $\Theta\left(\frac{k}{\epsilon}\right)$ noise suffices.

Analysis (Privacy)

Theorem: *The algorithm is ϵ-differentially private.*

- **Naive analysis:** Requires $\Theta\left(\frac{u}{\epsilon}\right)$ noise for ϵ-differential privacy.
- **Our analysis:** $\Theta\left(\frac{k}{\epsilon}\right)$ noise suffices.

Analysis (Performance)

Theorem (Utility): *For all* $\rho > 0$*: with probability at least* $1 - \rho$*, the output is* γ*-useful, where* $\gamma = \frac{4k}{\epsilon}\left(\ln u + \ln(\frac{1}{\rho})\right)$.

Theorem (Running Time): *The algorithm runs in time* $O(u)$.

Related Work

[KKMN09, GMW+09]

- Algorithms for differentially private ranking in search logs.
- Satisfy a weaker definition: (ϵ, δ)-differential privacy.
- Utility guarantee depends crucially on sensitivity
  - = the number of elements whose score can change if one entry is altered.
- Not useful for problems with high sensitivity.
  - Our utility guarantees are incomparable to those of [KKMN09, GMW+09].

Related Work

[KKMN09, GMW+09]

- Algorithms for differentially private ranking in search logs.
- Satisfy a weaker definition: (ϵ, δ)-differential privacy.
- Utility guarantee depends crucially on sensitivity
  - = the number of elements whose score can change if one entry is altered.
- Not useful for problems with high sensitivity.
  - Our utility guarantees are incomparable to those of [KKMN09, GMW+09].

Related Work

[KKMN09, GMW+09]

- Algorithms for differentially private ranking in search logs.
- Satisfy a weaker definition: (ϵ, δ)-differential privacy.
- Utility guarantee depends crucially on sensitivity
  - = the number of elements whose score can change if one entry is altered.
- Not useful for problems with high sensitivity.
  - Our utility guarantees are incomparable to those of [KKMN09, GMW+09].

Related Work

[KKMN09, GMW+09]

- Algorithms for differentially private ranking in search logs.
- Satisfy a weaker definition: (ϵ, δ)-differential privacy.
- Utility guarantee depends crucially on sensitivity
  - = the number of elements whose score can change if one entry is altered.
- Not useful for problems with high sensitivity.
  - Our utility guarantees are incomparable to those of [KKMN09, GMW+09].

References


Michaela Götz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke.
Privacy in search logs.
CoRR, abs/0904.0682, 2009.


Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas.
Releasing search queries and clicks privately.
In *WWW*, pages 171–180, 2009.


Frank McSherry and Kunal Talwar.
Mechanism design via differential privacy.
In *FOCS*, pages 94–103, 2007.

Aaron Roth

How to Privately Answer More Count Queries Interactively

Aaron Roth

Tim Roughgarden

Count Queries:

"What Fraction of the people in the dataset have blue eyes and brown hair?"

…

"What Fraction of the people in the dataset satisfy complicated condition C?"

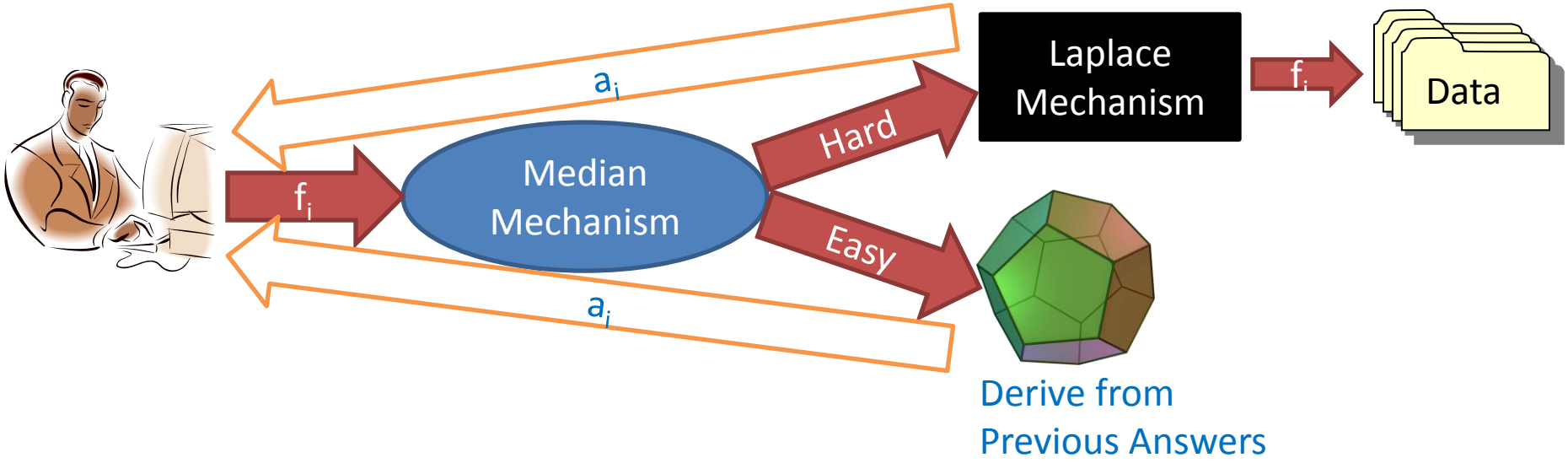
Setting

Setting

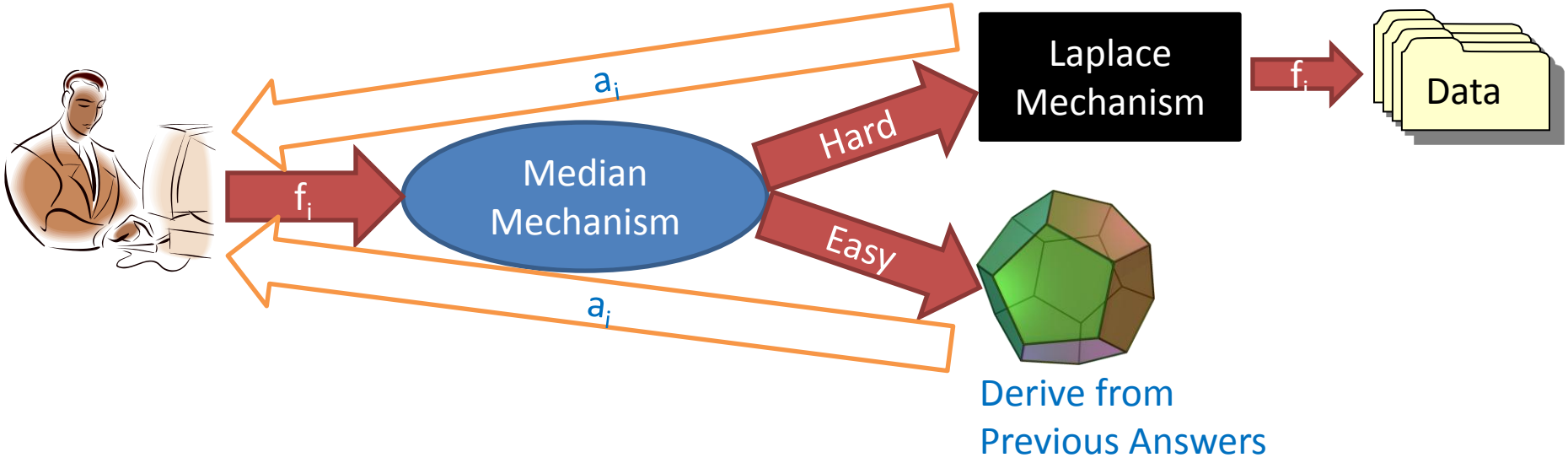
Want ε-differential privacy, and all queries to be accurate up to α.

Laplace Mechanism can answer ≈ nεα queries

Don’t always have to add independent noise

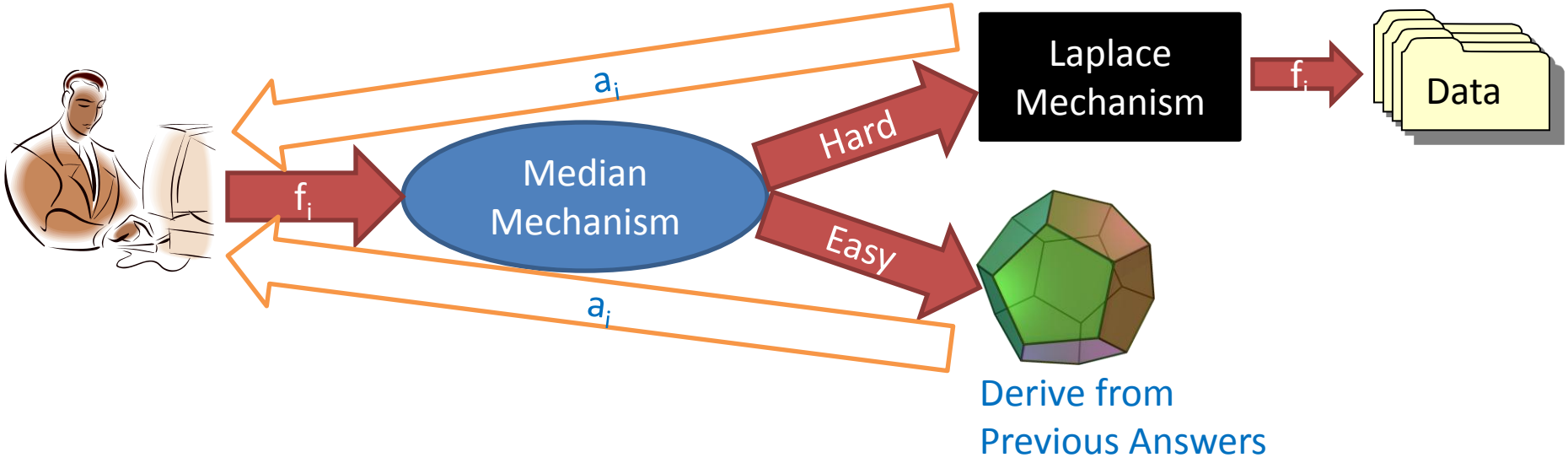


Don’t always have to add independent noise



Theorem: For any set of k queries, there exist ≈ log k queries, the approximate answers to which imply the approximate answers to *all* other queries.

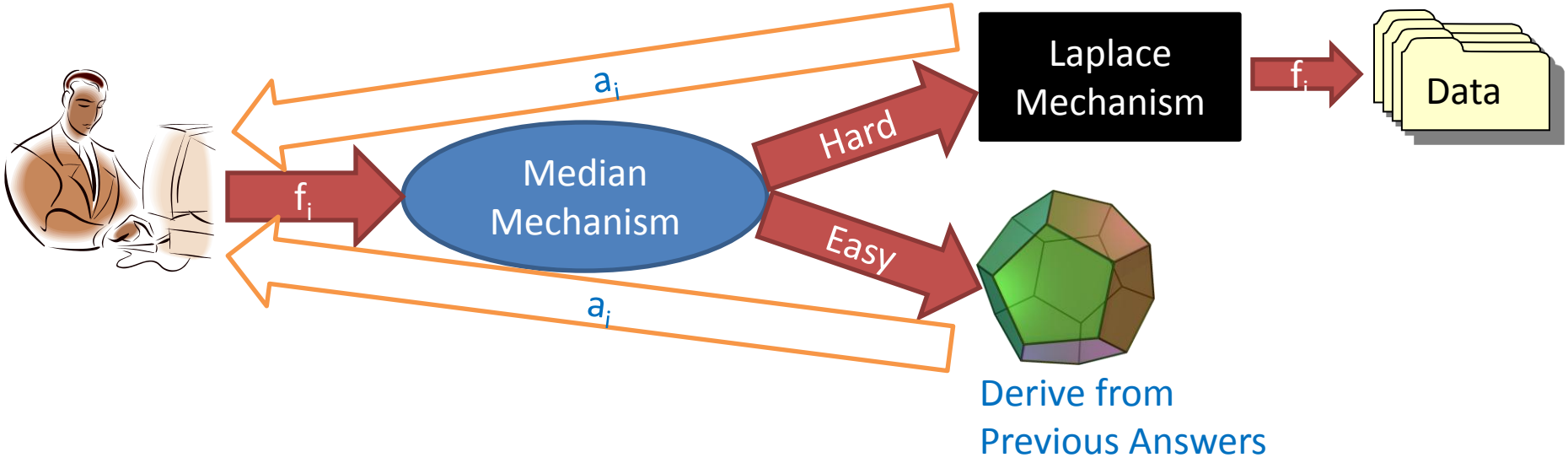
Don’t always have to add independent noise



We can privately identify hard queries adaptively as they arrive.

Result: We only have to spend our privacy budget on hard queries.

Don’t always have to add independent noise



Want ε-differential privacy, and all queries to be accurate up to α.

End Result: We can answer $2^{(n\alpha\varepsilon)^{1/3}\alpha}$ queries.

Shiva Kasiviswanathan

Price of Private Data Analysis and Spectra of Random Matrices

Shiva Kasiviswanathan
Los Alamos National Lab

Joint work with:
Mark Rudelson (Missouri/Michigan)
Adam Smith (Penn State)
Jonathan Ullman (Harvard)

Contingency (Marginal) Tables

Database: Table of observations of size n × d

	Brown Eye	Black Eye	Red Hair	Black Hair
Alice	0	1	1	0
Bob	1	0	0	1
Charlie	1	0	1	0
Dave	1	0	1	0

Marginal table for subset $S \subseteq [d]$ of size k:

- frequency of all 2^k possible combinations of attributes in S

2-way marginal table (conjunction table)

Black Hair and Brown Eye	0	1
0	1	0
1	2	1

Lower Bounds under (ε,δ)-Diff. Privacy

Let D be a database with n rows and d columns

Treat k, ε, and δ as constants

Suppose we want to release all k-way marginal tables

- $O(d^k)$ real numbers

D.P. Mechanism	Upper bound – Noise	Lower bound - Noise
Instance-Independent	$O(d^{k/2})$ [BDMN05]	$\Omega(d^{k/2})$
General	$O(\min\{n,(n^2d)^{1/3},d^{k/2}\})$ [BDMN05, BLR08]	$\Omega(\min\{n^{1/2},d^{k/2}\})$

Idea: Project the mean squared matrix of A(D) in various directions

Lower Bounds under Minimal Privacy

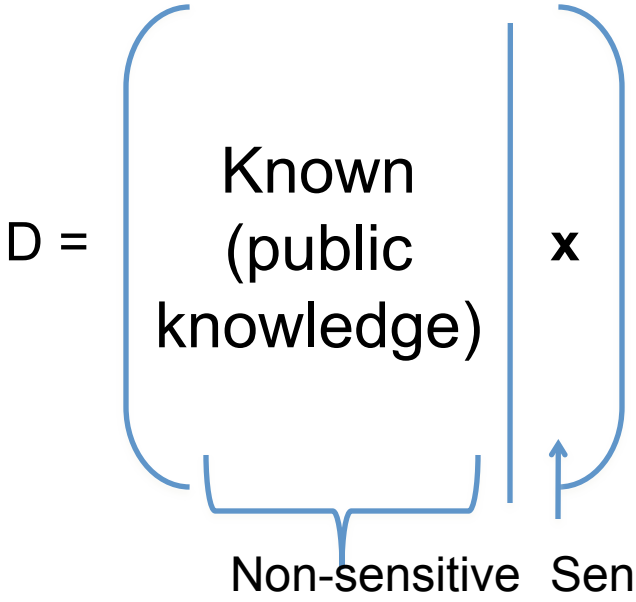
We consider two other "simpler notions" of privacy

Privacy Guarantee	Upper Bound - Noise	Lower Bound - Noise
Attribute Non-Privacy	$O(\min\{n^{1/2}, d^{k/2}\})$	$\Omega(\min\{n^{1/2}, d^{(k-1)/2}\})$
Row Non-Privacy	$O(\min\{n^{1/2}, d^{k/2}\})$	$\Omega(\min\{n^{1/2}, d^{k/2}\})$

Idea: Lower bound is based on new techniques for analyzing spectra of random correlated matrices

Lower Bounds under Minimal Privacy

We consider two other "simpler notions" of privacy

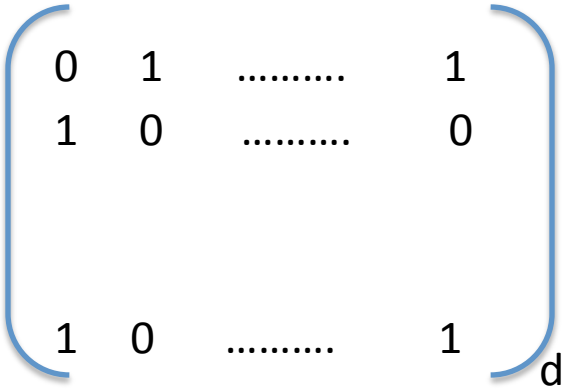


Privacy Guarantee	Upper Bound - Noise	Lower Bound - Noise
Attribute Non-Privacy	$O(\min\{n^{1/2}, d^{k/2}\})$	$\Omega(\min\{n^{1/2}, d^{(k-1)/2}\})$
Row Non-Privacy	$O(\min\{n^{1/2}, d^{k/2}\})$	$\Omega(\min\{n^{1/2}, d^{k/2}\})$

Idea: Lower bound is based on new techniques for analyzing spectra of random correlated matrices

Least Singular Value of Random Matrix with Correlated Rows

First we need to define a "Conjunction Matrix"



$d \times n$

M

k-way conjunction matrix obtained by taking entry-wise product of every set of k rows

$d^k \times n$

Entry-wise product of

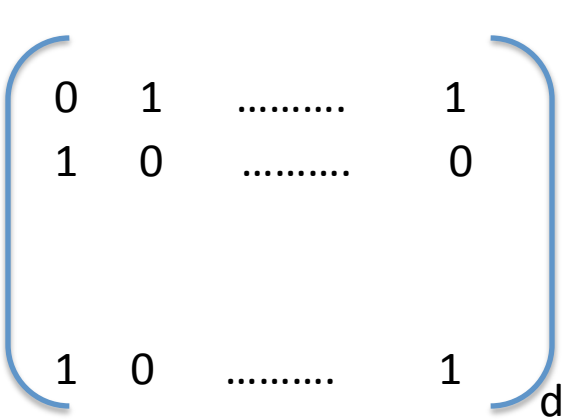
0 0 1
1 0 0

0 0 0

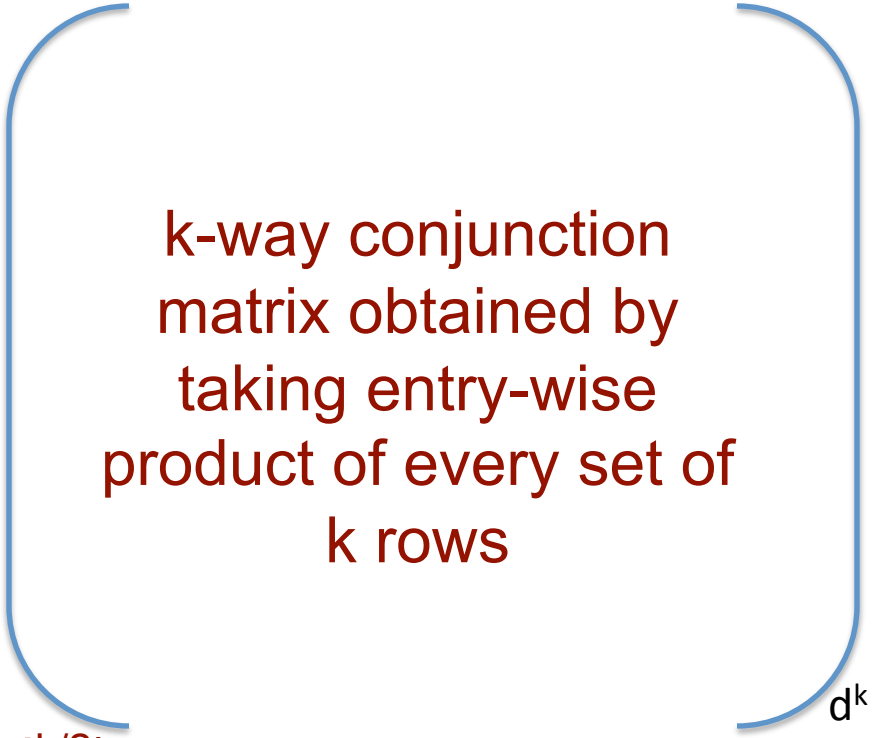
$M^{(k)}$ = k-way conjunction matrix

Least Singular Value of Random Matrix with Correlated Rows

Least singular value of a random d^k x n matrix = $O(d^{k/2})$



Random Matrix M



Least singular value of $M^{(k)} = O(d^{k/2})$

$M^{(k)}$ = k-way conjunction matrix

More Details: Ask us
Proceedings version appearing soon

Preliminary version:
http://www.cse.psu.edu/~kasivisw/public.pdf

Vibhor Rastogi

Differential Privacy for Distributed Time-Series Data

Vibhor Rastogi @ University of Washington
Suman Nath @ Microsoft Research

Motivating Example

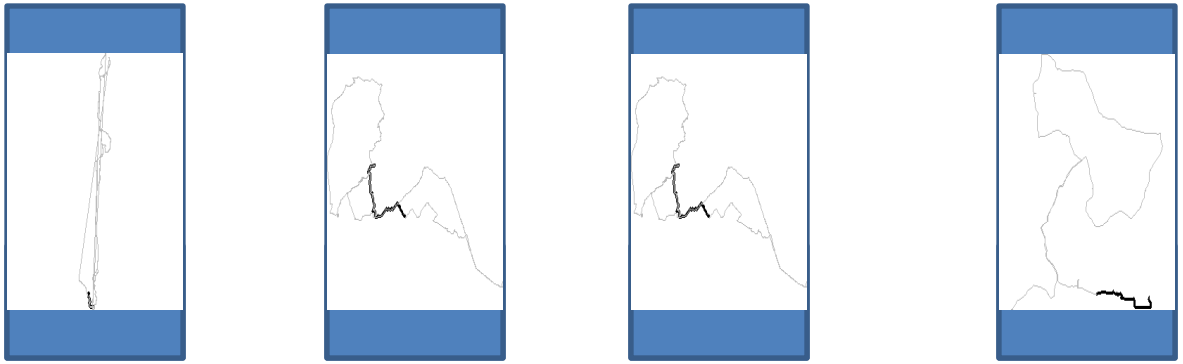
- Distributed location-tracking system

Aggregate Queries

Traffic Analyzer

How many people drive through the 520 bridge and take the 148th avenue exit before 9AM?

Privacy Concerns



Alice Bob Charlie Delta

Each user collects location data using GPS

Two main challenges

Challenge #1: Accuracy Problem

q_1 = # of people in ***148th & Sr 520*** at 5:00 PM
q_2 = # of people in ***148th & Sr 520*** at 5:15 PM
…
…
q_N = # of people in ***148th & Sr 520*** at 1:25 AM

Name	Age	Location	Time
Alice	25	Building 92	5 PM
Alice	25	Building 99	5:02 PM
Alice	25	148th & 36th St	5:04 PM
Bob	32	148th & Sr 520	5:35 PM

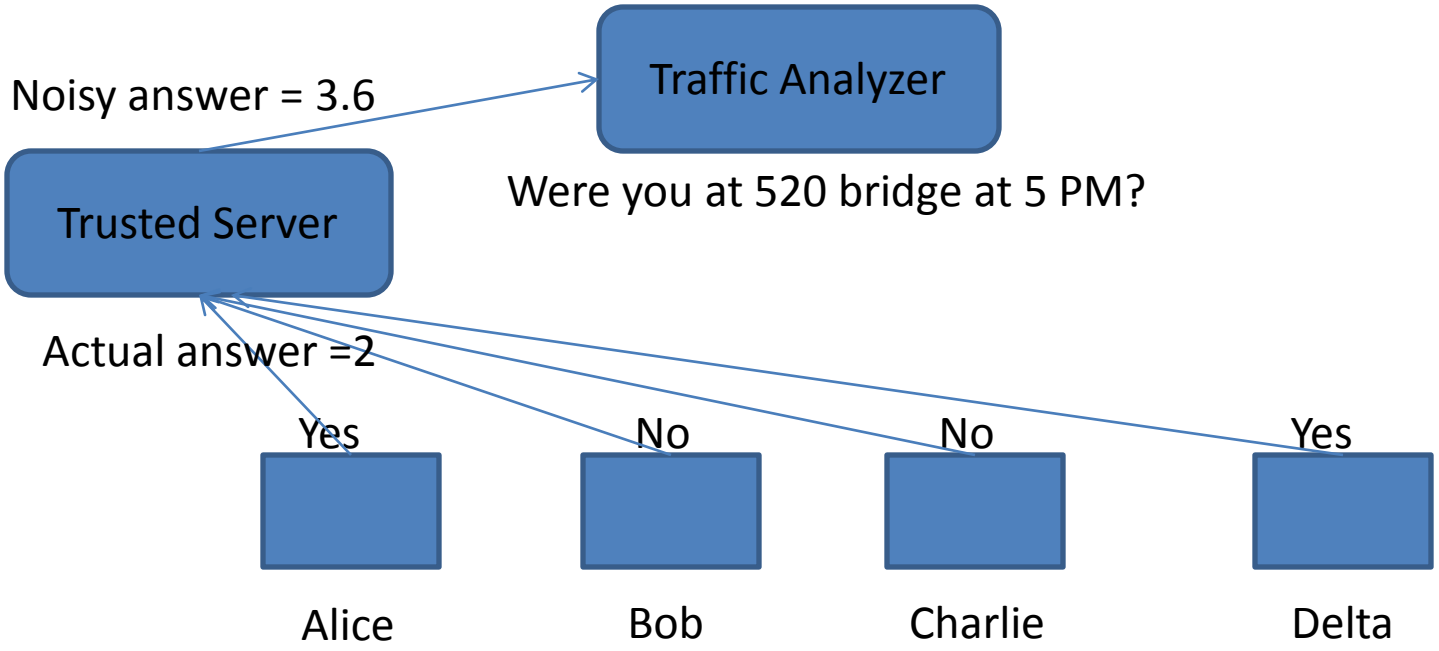
Time-Series Data

Answer of each query can change by 1
L1 sensitivity is N
Θ(N) noise required in each answer

Noise too large for long sequences!

Two main challenges

Challenge #2: No trusted server



A Third Challenge

- I am graduating – find a job!
- Still an open problem
- Apparently, solved by several researchers
  - Will be happy to talk

Ilya Mironov

Collaborative Filtering and Differential Privacy: Building Privacy into the Netflix Contenders

Frank McSherry and Ilya Mironov
Microsoft Research, Silicon Valley Campus

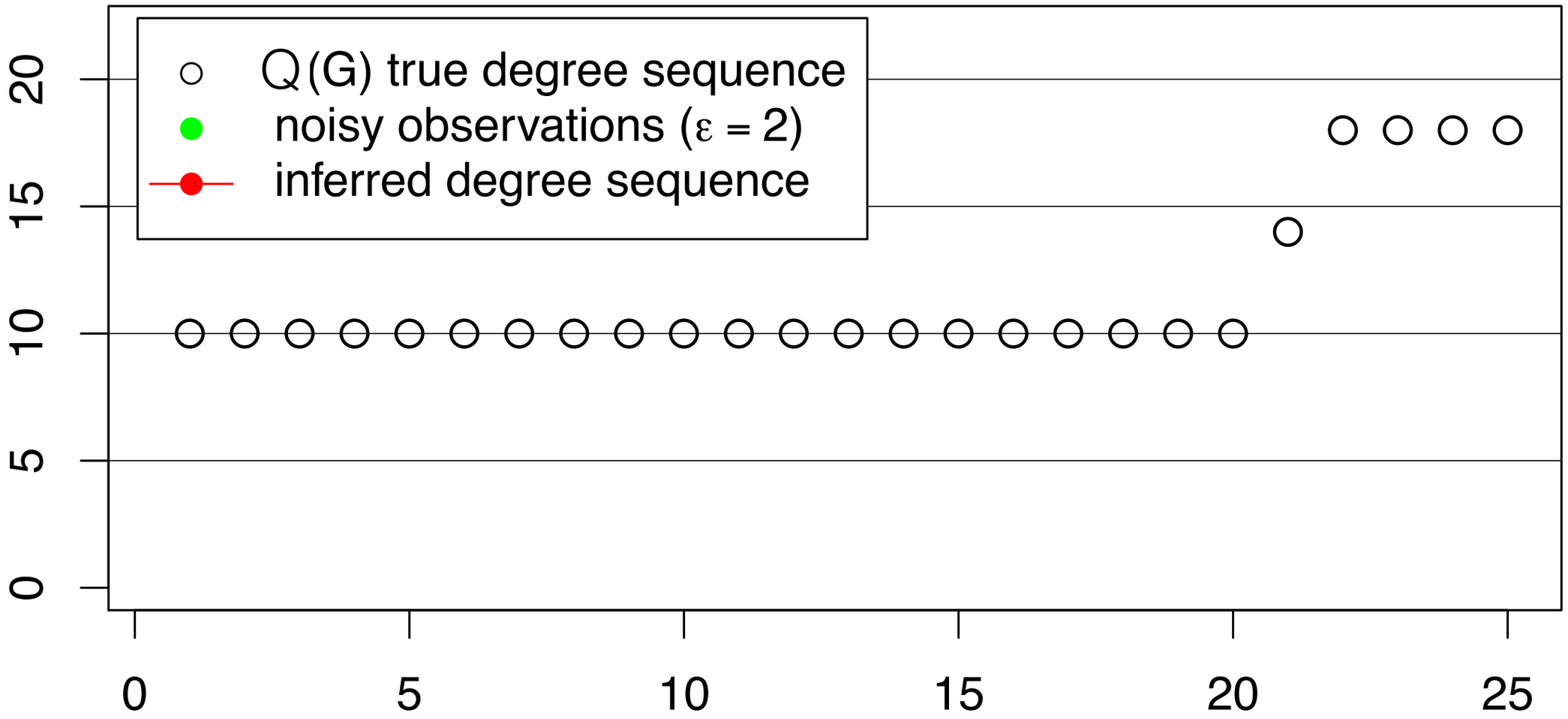
Appeared in KDD'09

Michael Hay

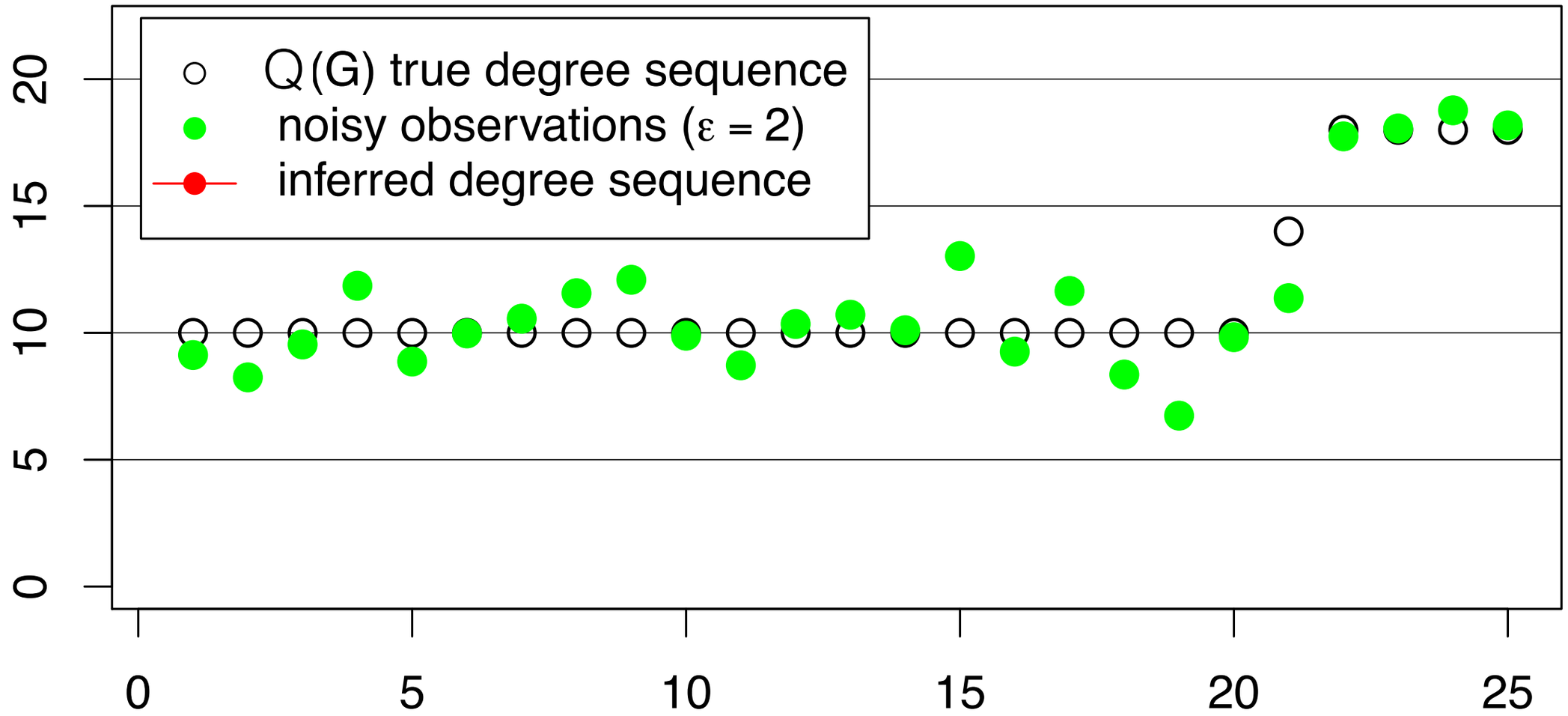
Unattributed histograms under differential privacy

- (Attributed) Histogram **x = x_1, ..., x_n**
  - **x_i** = # of records of type i
- **Unattributed** Histogram **y = y_1, ..., y_n**
  - **y_i** = # of records of the i^{th} most frequent type
- For each type of histogram, adding Laplace(1/ε) noise achieves ε-differential privacy.
- For unattributed histogram, we introduce post-processing step that reduces error (at no sacrifice to privacy).
- "Killer app": estimating degree sequence of a social network graph

Using the sort constraint



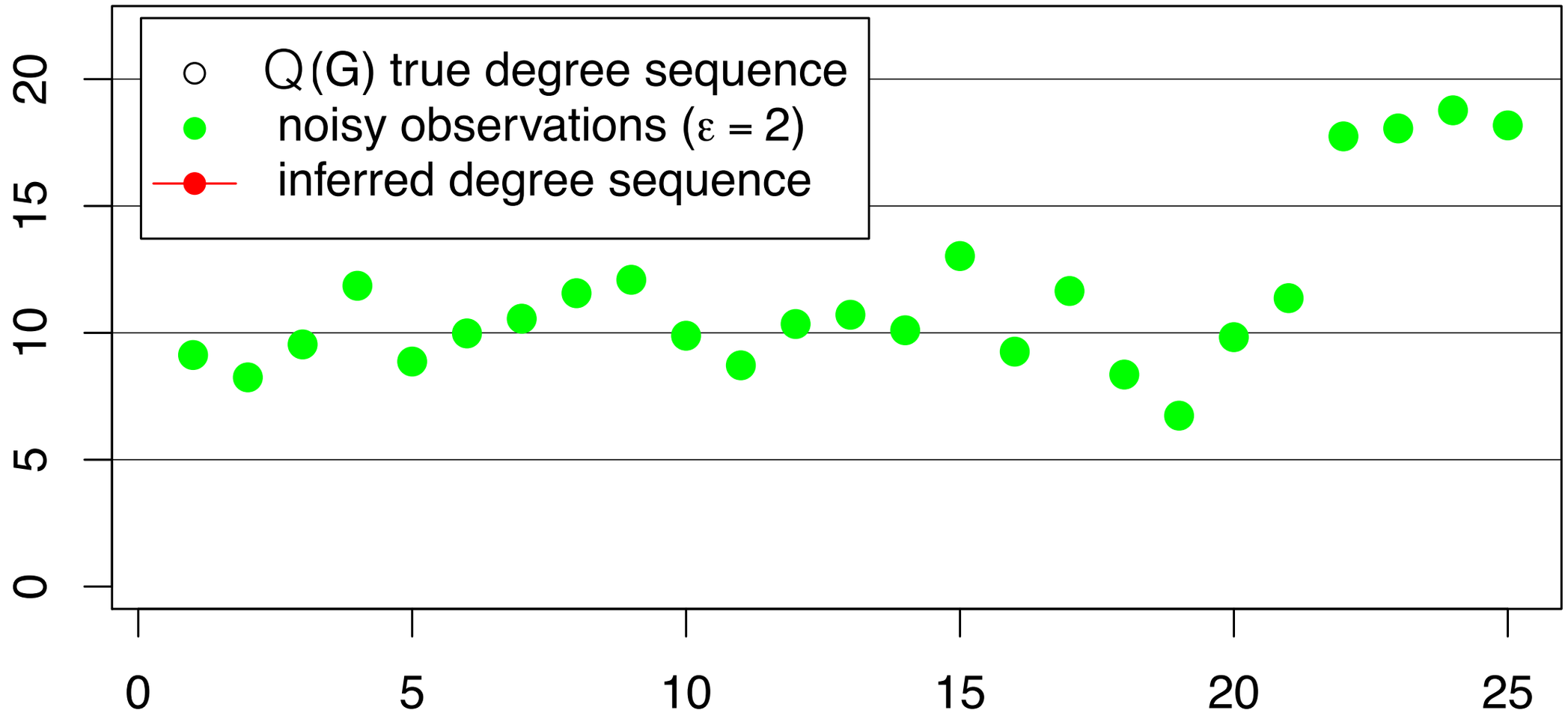
Using the sort constraint



- The output of the sorted degree query is not (in general) sorted.

- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

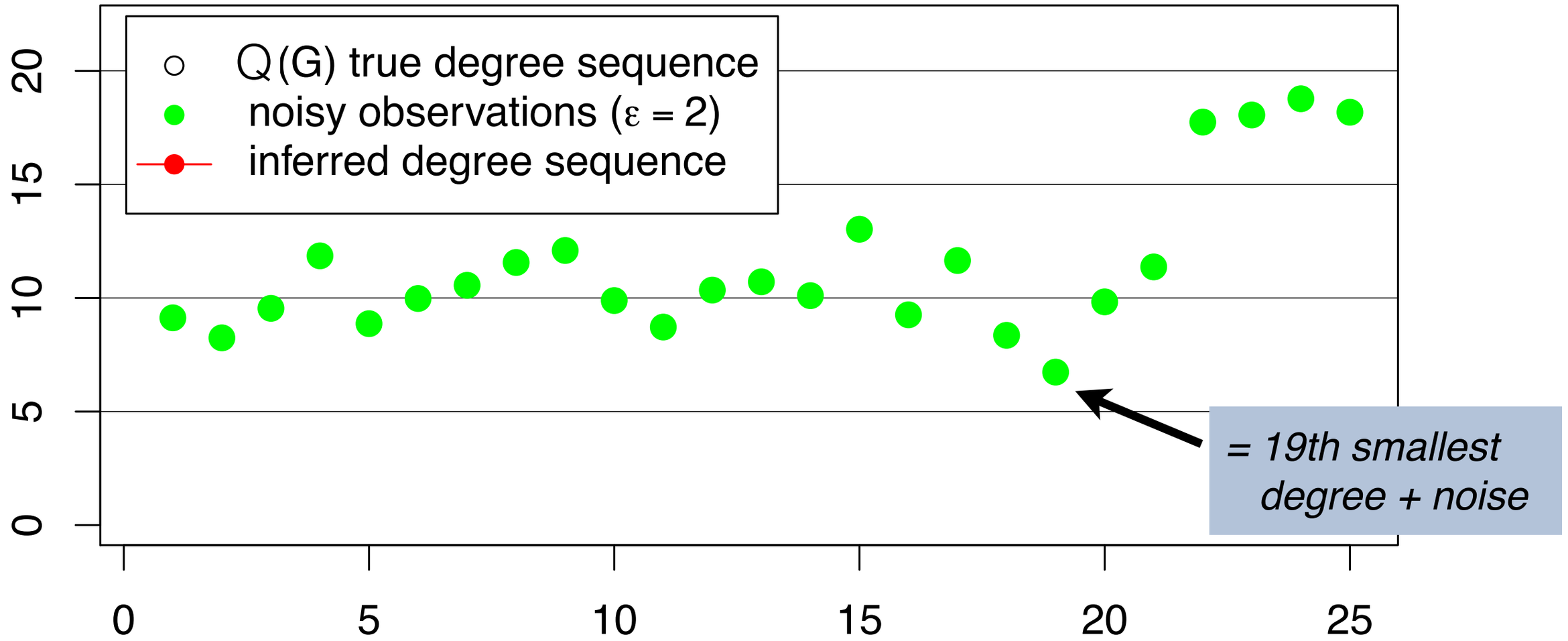
Using the sort constraint



- The output of the sorted degree query is not (in general) sorted.

- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

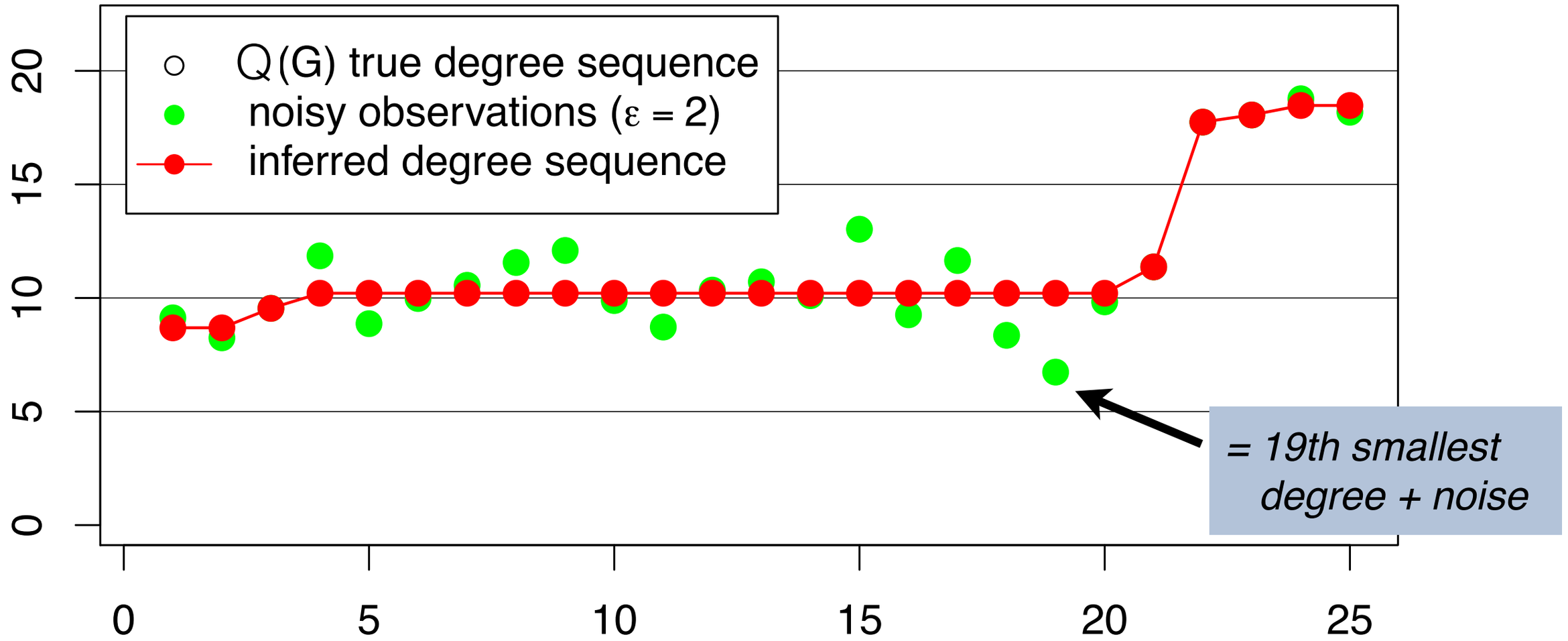
Using the sort constraint



- The output of the sorted degree query is not (in general) sorted.

- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

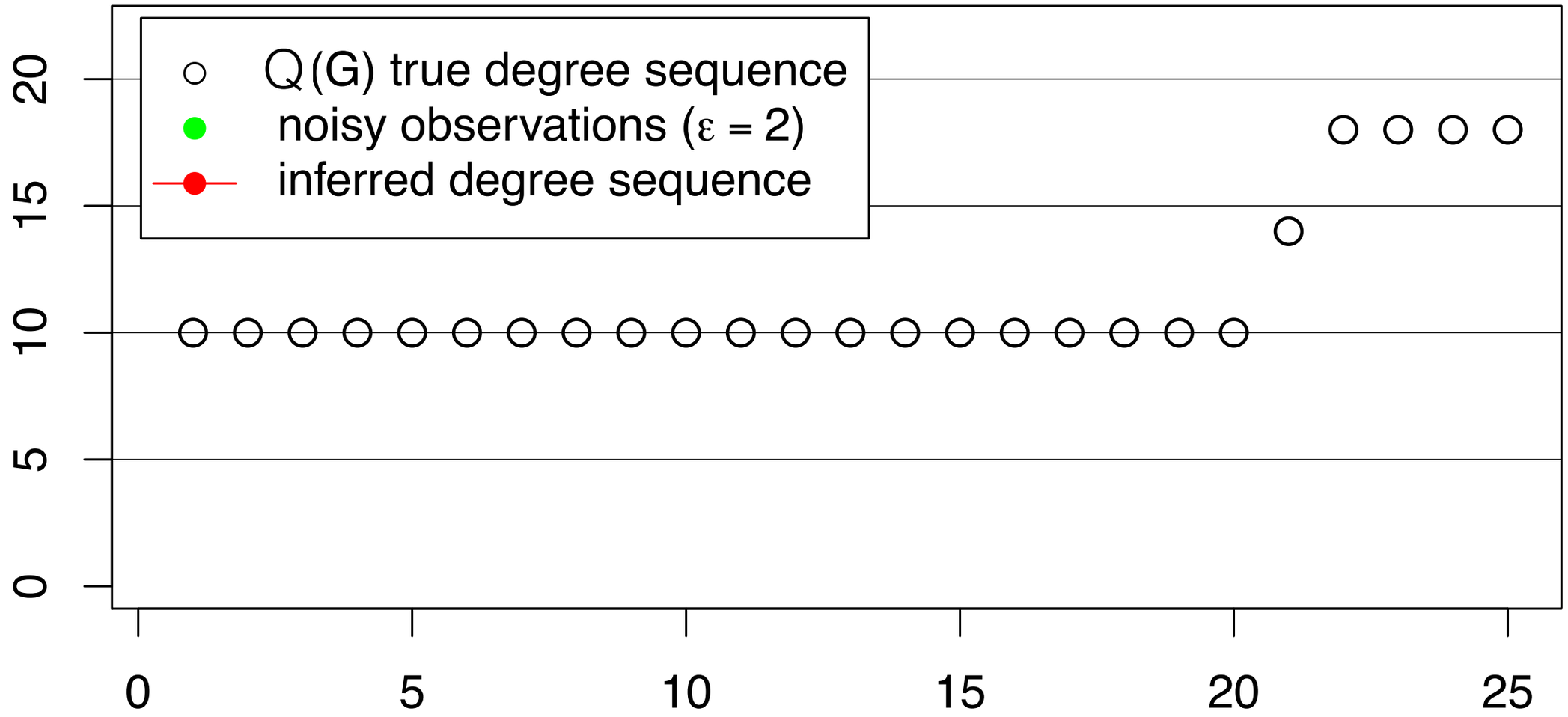
Using the sort constraint



- The output of the sorted degree query is not (in general) sorted.

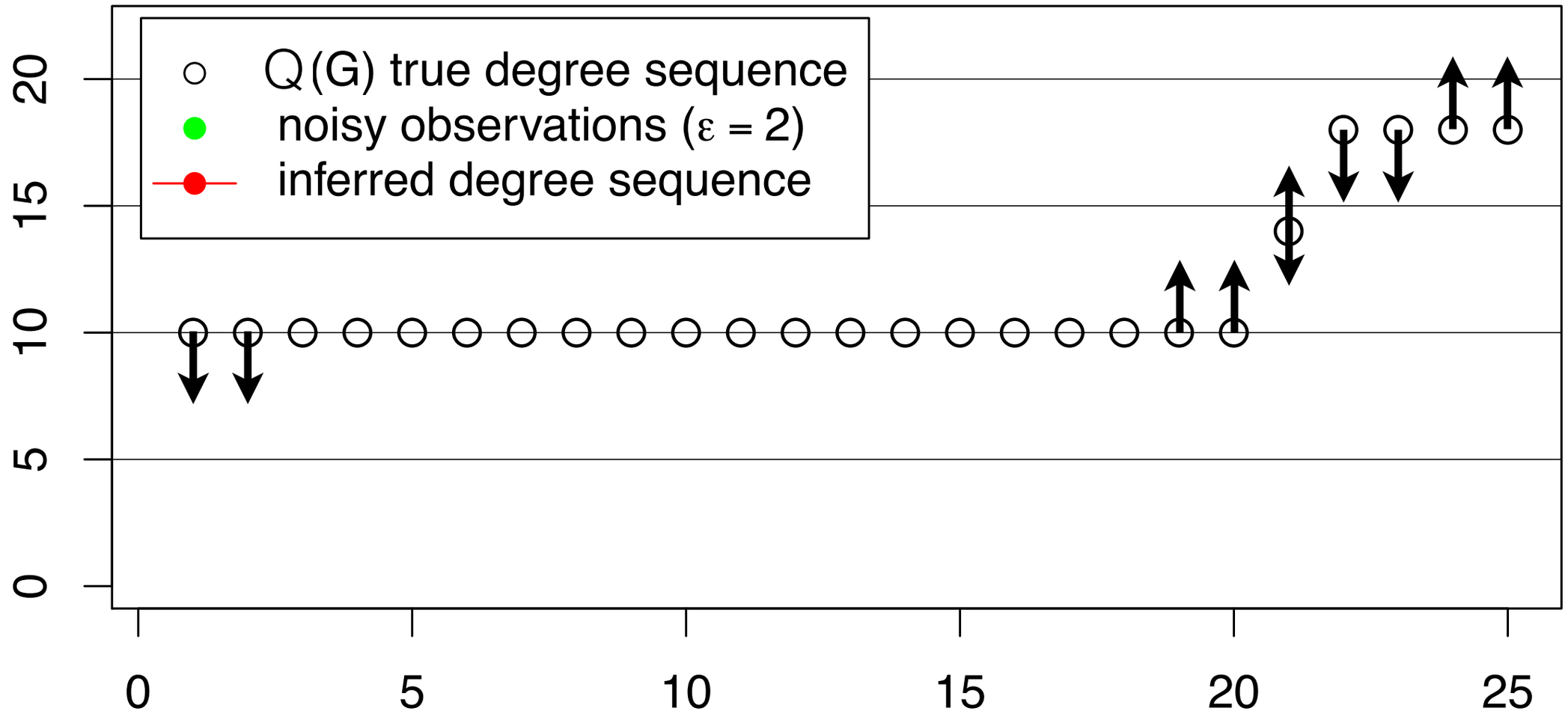
- We derive a new sequence by computing the **closest** non-decreasing sequence: i.e. minimizing L2 distance.

After inference, noise only where needed



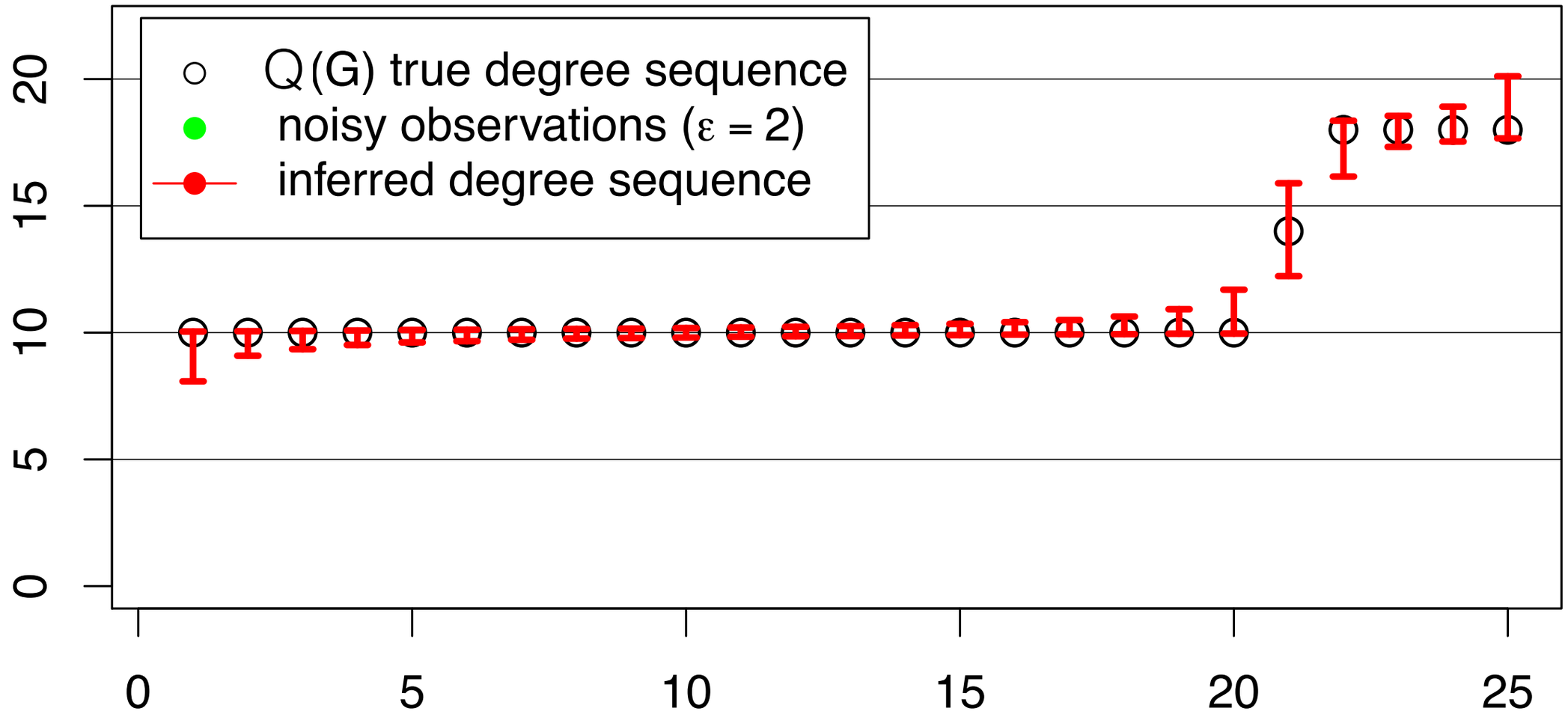
- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
  - Improvement in accuracy **depends on sequence**

After inference, noise only where needed



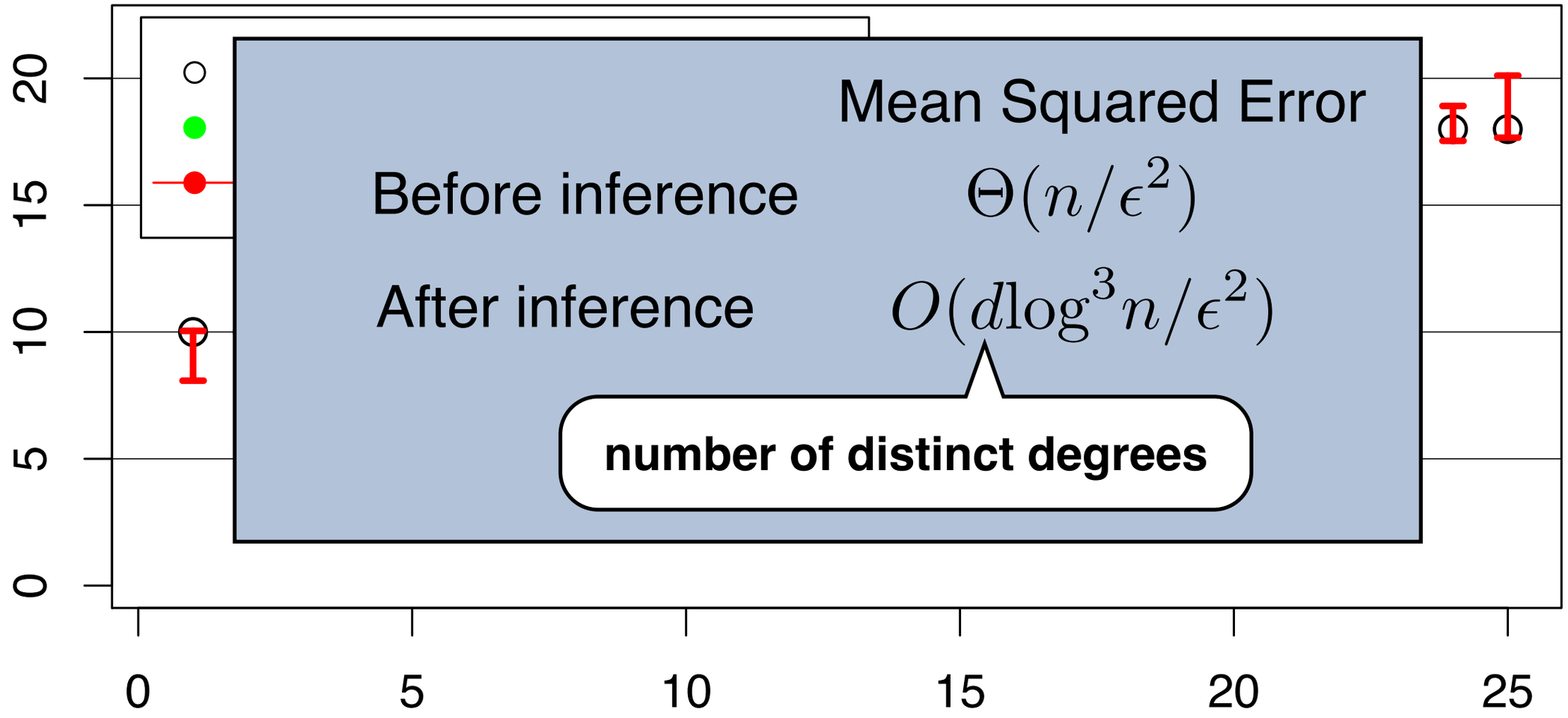
- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
  - Improvement in accuracy **depends on sequence**

After inference, noise only where needed



- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
  - Improvement in accuracy **depends on sequence**

After inference, noise only where needed



- Standard Laplace noise is sufficient *but not necessary* for differential privacy.
- By using inference, effectively apply a different noise distribution -- more noise where it is needed, less otherwise.
  - Improvement in accuracy **depends on sequence**

References

For additional details, references below or **come see the poster**!

- **[ICDM 09]** M. Hay, C. Li, G. Miklau, and D. Jensen. Accurate estimation of the degree distribution of private networks. In International Conference on Data Mining (ICDM), 2009.

- **[PVLDB 10]** M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially-private histograms through consistency. In Proceedings of the VLDB (PVLDB), to appear 2010.

Robert Hall

Scott Coull

Network Data Privacy

Scott Coull
University of North Carolina
Chapel Hill, NC
scoull@cs.unc.edu

Joint work with: Fabian Monrose and Michael Reiter

Problem Overview

- How do we publish generally **useful** network traces without revealing **private information**?
  - Many existing repositories:
    - DHS PREDICT
    - Dartmouth CRAWDAD
    - CAIDA DatCat
- Short-term solutions are necessary
- Long-term solutions grounded in definitions

Network Data

- **Network Data** is:
  - Non-interactive database of packets/flows

- Packet traces:
  - Records for every packet viewed by monitor

Timestamp	Source IP	Source Port	…	Payload
Feb. 2, 2010 05:10:02.43	10.0.0.1	80	…	GET /index.html HTTP/1.1 …

Network Data

- **Network Data** is:
  - Non-interactive database of packets/flows
- Flow logs:
  - Records summarize all packets in a connection

Start Time	End Time	Source IP	Source Port	…	Bytes Sent
Feb. 2, 2010 05:10:02.43	Feb. 2, 2010 05:45:26.17	10.0.0.1	80	…	1024 Bytes

Network Data Anonymization

- Typical anonymization policy:
  - Truncate payloads
    - Removes plaintext user names, passwords, etc.
  - Quantize timestamps
    - Prevents clock skew attacks [Kohno et al., 2005]
  - Replace IP addresses with linkable pseudonyms
    - Specifically, prefix-preserving pseudonyms

Network Data Anonymization

- Problem: Policies are defined based on intuition and expert knowledge
  - Bias toward utility
  - Fields are altered in *reaction* to new attacks

- Result: Unexpected areas of information leakage occur within the anonymized data
  - No privacy guarantees
  - No methods for verifying efficacy

Relation to Microdata Privacy

- Seems like microdata privacy notions apply…

Relation to Microdata Privacy

- Seems like microdata privacy notions apply…

	Microdata	Network Data
Privacy Goal:	Protect individual records	Protect objects made up of multiple records *(workstations, users, etc.)*
Data Types:	Categorical and numeric data	Complex, non-traditional data *(IP addresses, etc.)*
Semantics:	Weak semantics among records and attributes	Strong semantic relationships due to network protocols
Size:	Millions of records	Billions or trillions of records

Relation to Microdata Privacy

- Seems like microdata privacy notions apply…

	Microdata	Network Data
Privacy Goal:	Protect individual records	Protect objects made up of multiple records *(workstations, users, etc.)*
Data Types:	Categorical and numeric data	Complex, non-traditional data *(IP addresses, etc.)*
Semantics:	Weak semantics among records and attributes	Strong semantic relationships due to network protocols
Size:	Millions of records	Billions or trillions of records

Relation to Microdata Privacy

- Seems like microdata privacy notions apply…

	Microdata	Network Data
Privacy Goal:	Protect individual records	Protect objects made up of multiple records *(workstations, users, etc.)*
Data Types:	Categorical and numeric data	Complex, non-traditional data *(IP addresses, etc.)*
Semantics:	Weak semantics among records and attributes	Strong semantic relationships due to network protocols
Size:	Millions of records	Billions or trillions of records

Relation to Microdata Privacy

- Seems like microdata privacy notions apply…

	Microdata	Network Data
Privacy Goal:	Protect individual records	Protect objects made up of multiple records *(workstations, users, etc.)*
Data Types:	Categorical and numeric data	Complex, non-traditional data *(IP addresses, etc.)*
Semantics:	Weak semantics among records and attributes	Strong semantic relationships due to network protocols
Size:	Millions of records	Billions or trillions of records

Relation to Microdata Privacy

- Seems like microdata privacy notions apply…

	Microdata	Network Data
Privacy Goal:	Protect individual records	Protect objects made up of multiple records *(workstations, users, etc.)*
Data Types:	Categorical and numeric data	Complex, non-traditional data *(IP addresses, etc.)*
Semantics:	Weak semantics among records and attributes	Strong semantic relationships due to network protocols
Size:	Millions of records	Billions or trillions of records

Roadmap

- Repeat mistakes from microdata privacy
  - Helps to understand how to map to network data
  - Hopefully find good solutions in less than 30 years
- Inference attacks on “anonymized” network data
  - Re-identification of workstations, user behaviors
  - Revelation of network data
- Creation of risk analysis framework based partly on microdata privacy notions

Xiaolin Yang