Local Differential Privacy for Physical Sensor Data

Audra McMillan and Anna C. Gilbert

Department of Mathematics
University of Michigan

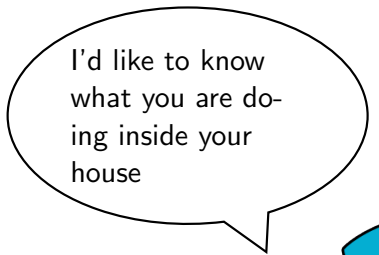
February 13, 2018

The right to privacy in your own home

Before the information age:

The right to privacy in your own home

Before the information age:



The right to privacy in your own home

Before the information age:

The right to privacy in your own home

Before the information age:

The right to privacy in your own home

Before the information age:

The right to privacy in your own home

Before the information age:

Privacy in the Information Age

Things are a lot murkier...

Privacy in the Information Age

Things are a lot murkier...

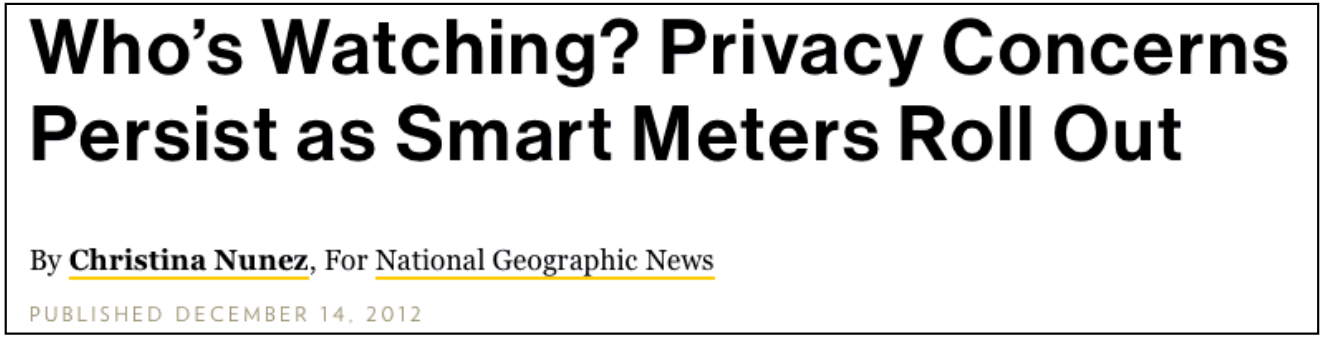
Privacy in the Information Age

Things are a lot murkier...

Guardian sustainable business Smart cities

Smart buildings monitor energy efficiency, but what are they really tracking?

Space utilisation technology aims to make offices more efficient and people friendly but there are concerns around privacy



Who's Watching? Privacy Concerns Persist as Smart Meters Roll Out

By **Christina Nunez**, For National Geographic News

PUBLISHED DECEMBER 14, 2012

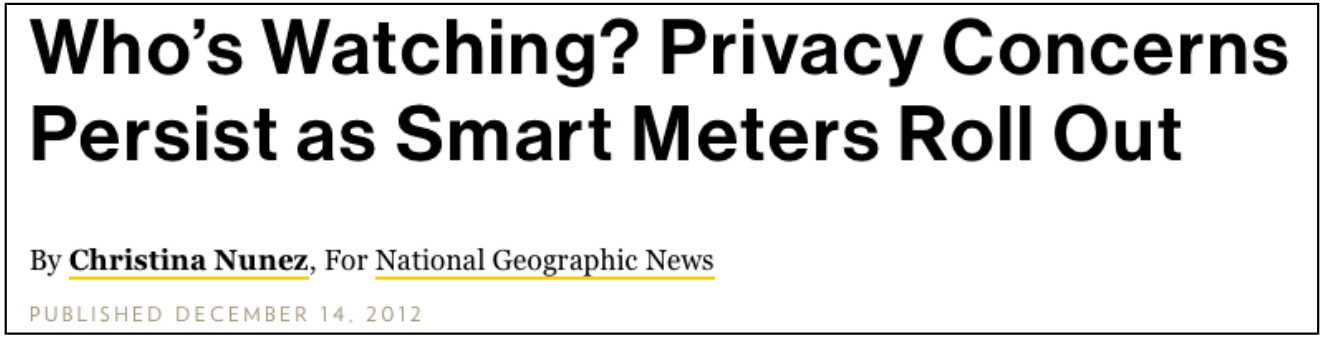
Privacy in the Information Age

Things are a lot murkier...

Guardian sustainable business Smart cities

Smart buildings monitor energy efficiency, but what are they really tracking?

Space utilisation technology aims to make offices more efficient and people friendly but there are concerns around privacy



Who's Watching? Privacy Concerns Persist as Smart Meters Roll Out

By **Christina Nunez**, For National Geographic News

PUBLISHED DECEMBER 14, 2012

Your Roomba May Be Mapping Your Home, Collecting Data That Could Be Shared

By MAGGIE ASTOR JULY 25, 2017

Privacy in the Information Age

Things are a lot murkier...


Guardian sustainable business Smart cities

Smart buildings monitor energy efficiency, but what are they really tracking?

Space utilisation technology aims to make offices more efficien
friendly but there are concerns around privacy


PRIVACY & THE INTERNET OF THINGS

In the privacy of your own home

That smart TV, your connected thermostat, even your washing machine—they're all tracking your daily habits. Why you need to know who's watching.

Published: April 30, 2015 06:00 AM


Who's Watching? Privacy Concerns Persist as Smart Meters Roll Out

By **Christina Nunez**, For National Geographic News

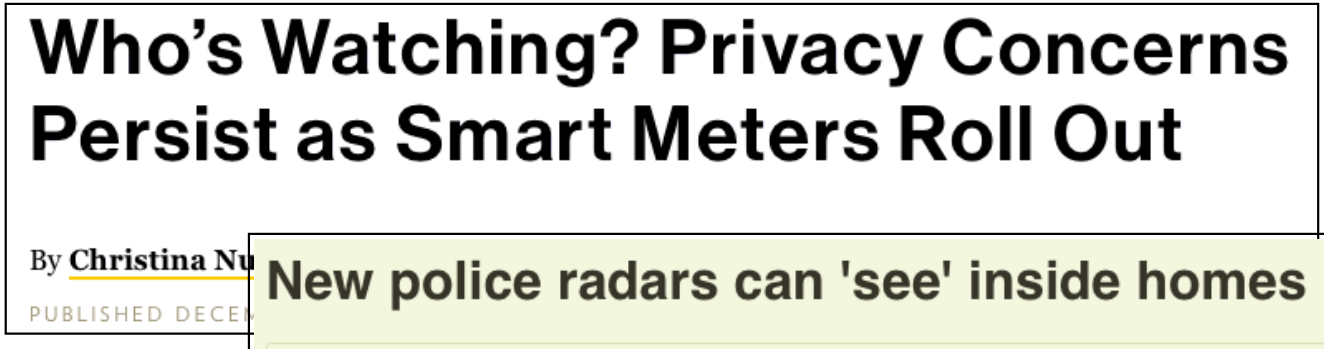
PUBLISHED DECEMBER 14, 2012


Your Roomba May Be Mapping Your Home, Collecting Data That Could Be Shared

By MAGGIE ASTOR JULY 25, 2017

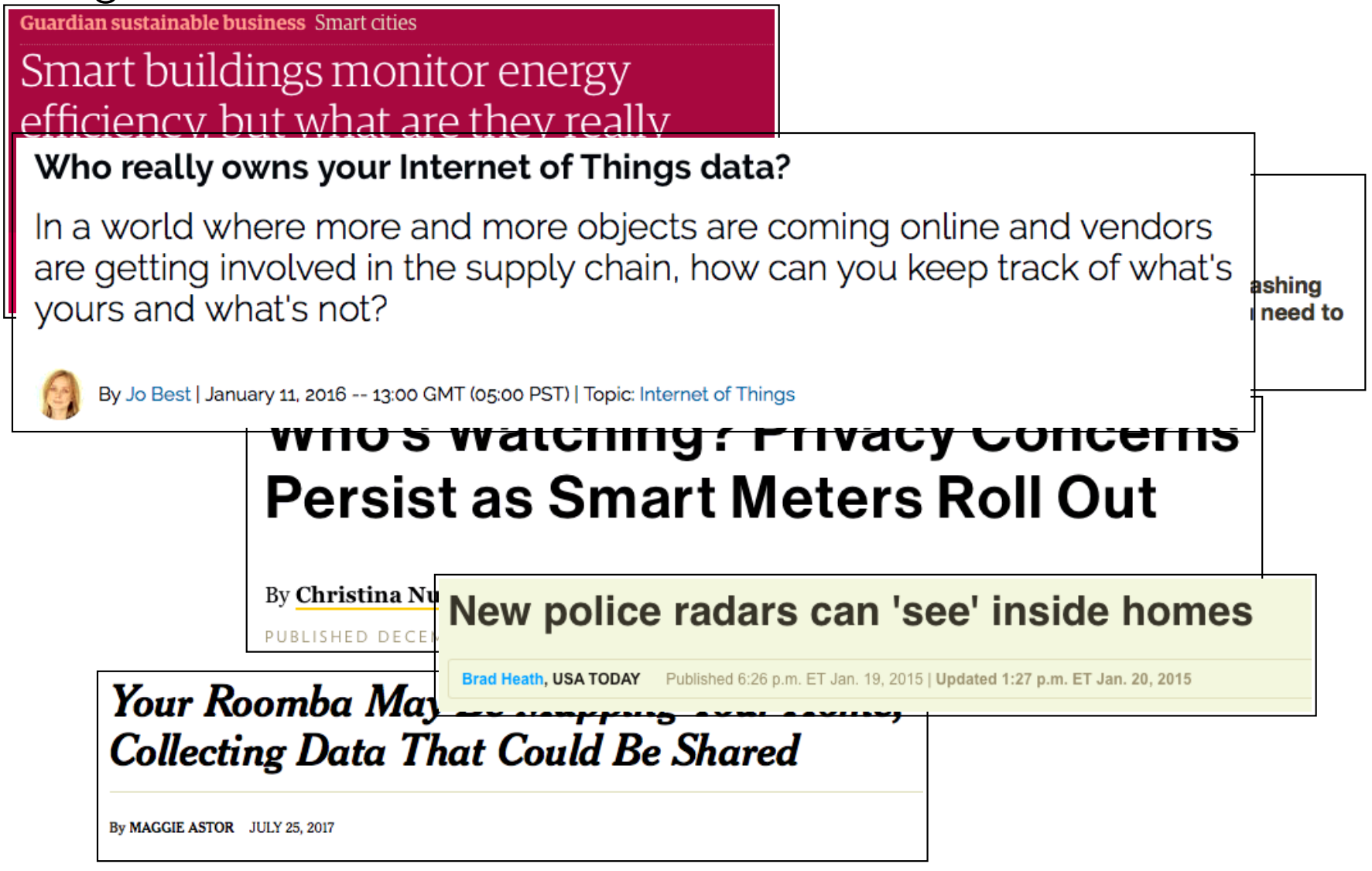
Privacy in the Information Age

Things are a lot murkier...



Privacy in the Information Age

Things are a lot murkier...



Privacy in the Information Age

Things are a lot murkier...

Guardian sustainable business Smart cities

Smart buildings monitor energy efficiency but what are they really

Who really owns your Internet of Things data?

In a world where more and more objects are coming online and vendors are getting involved in the supply chain, how can you keep track of what's yours and what's not?

By Jo Best | Janu

ashing
need to

Key Issues in Perspective:

SMART METERS and DATA PRIVACY

The electric power industry is modernizing the nation's electric grid. Using advanced technologies, electric companies are building a smart grid that will deliver more reliable power to customers across the country and allow two-way communication between customers and their electric companies.

Installing smart meters is an important step in building the smart grid. These advanced meters enable customers to track their power usage and learn more about the way they use electricity. This will help customers better manage their electricity usage in the future.

Your Ro
Collecti

By MAGGIE ASTOR JULY 25, 2017

Privacy in the Information Age

Things are a lot m

Guardian sustainable business Smart

Smart buildings m
efficiency but wha

Who really owns you

In a world where mo
are getting involved
yours and what's not

By Jo Best | Janu

Illinois grapples with question of who owns energy data

WRITTEN BY

Kari Lydersen
August 28, 2014

Don't miss the latest headlines. Sign up for our Daily Digests.

Your email | Zip code | Subscribe

Key Issues in Perspective:

SMART METERS and DATA PRIVACY

The electric power industry is modernizing the nation's electric grid. Using advanced technologies, electric companies are building a smart grid that will deliver more reliable power to customers across the country and allow two-way communication between customers and their electric companies.

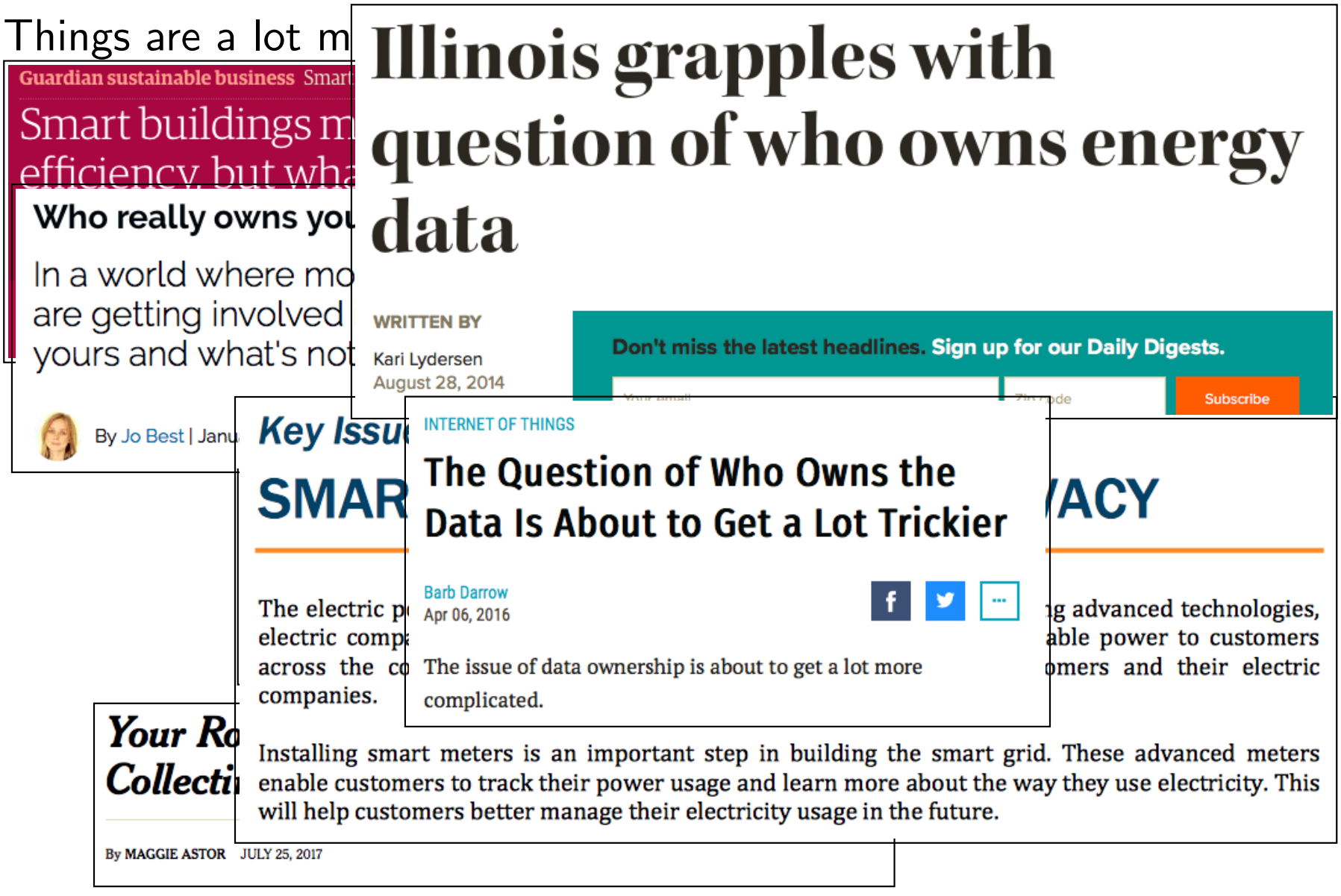
Installing smart meters is an important step in building the smart grid. These advanced meters enable customers to track their power usage and learn more about the way they use electricity. This will help customers better manage their electricity usage in the future.

Your Ro
Collecti

By MAGGIE ASTOR JULY 25, 2017

Privacy in the Information Age

Things are a lot m



What do these sensors do?

Lots of things! They help make our environment more comfortable, convenient and energy efficient. They help stabilise the grid, alert appropriate people about issues (power outage, fire, etc.)

What do these sensors do?

Lots of things! They help make our environment more comfortable, convenient and energy efficient. They help stabilise the grid, alert appropriate people about issues (power outage, fire, etc.)

We're going to focus today on sensors that measure the physical process of diffusion. Two motivating examples:

- Thermal sensors.
  - Building managers want to locate heat sources in a building, to help control HVAC systems.
  - Sensitive because people are heat sources.
- Information diffusion in social networks.
  - We want to be able to locate the sources of misinformation.
  - Sensitive because we want people to be able to spread information without the fear of retribution.

Our Goal: Produce "private" measurements of diffusion processes that

- permit recovery of the general vicinity of the source but
- do not permit recovery of the exact locations of sources.

Our Goal: Produce "private" measurements of diffusion processes that

- permit recovery of the general vicinity of the source but
- do not permit recovery of the exact locations of sources.

Why is diffusion interesting?

It is an example of ill-conditioning:

- The process is *technically* reversible.
- In the presence of even a small amount of noise, it is impossible to determine the original source location.

Our Goal: Produce "private" measurements of diffusion processes that

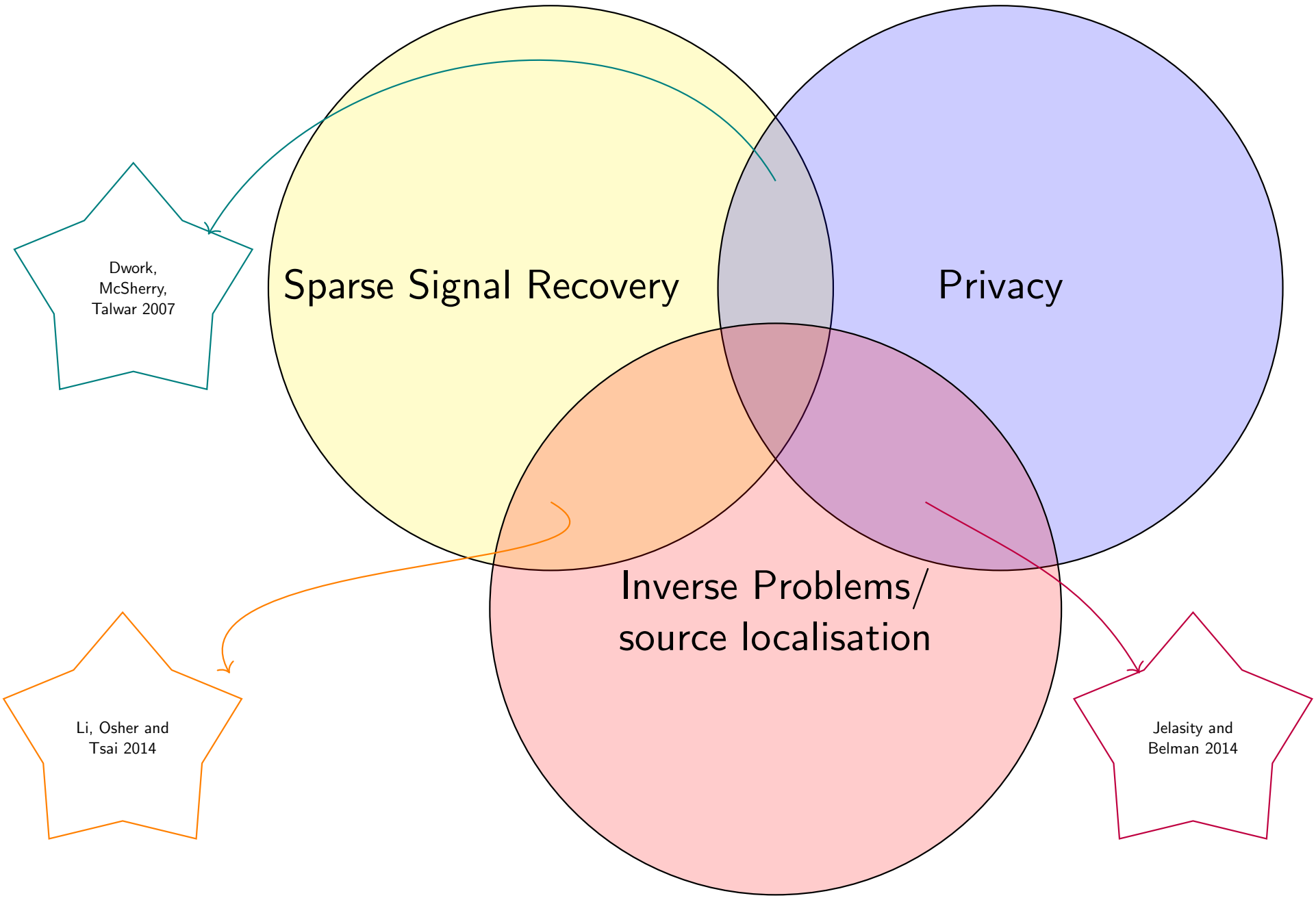
- permit recovery of the general vicinity of the source but
- do not permit recovery of the exact locations of sources.

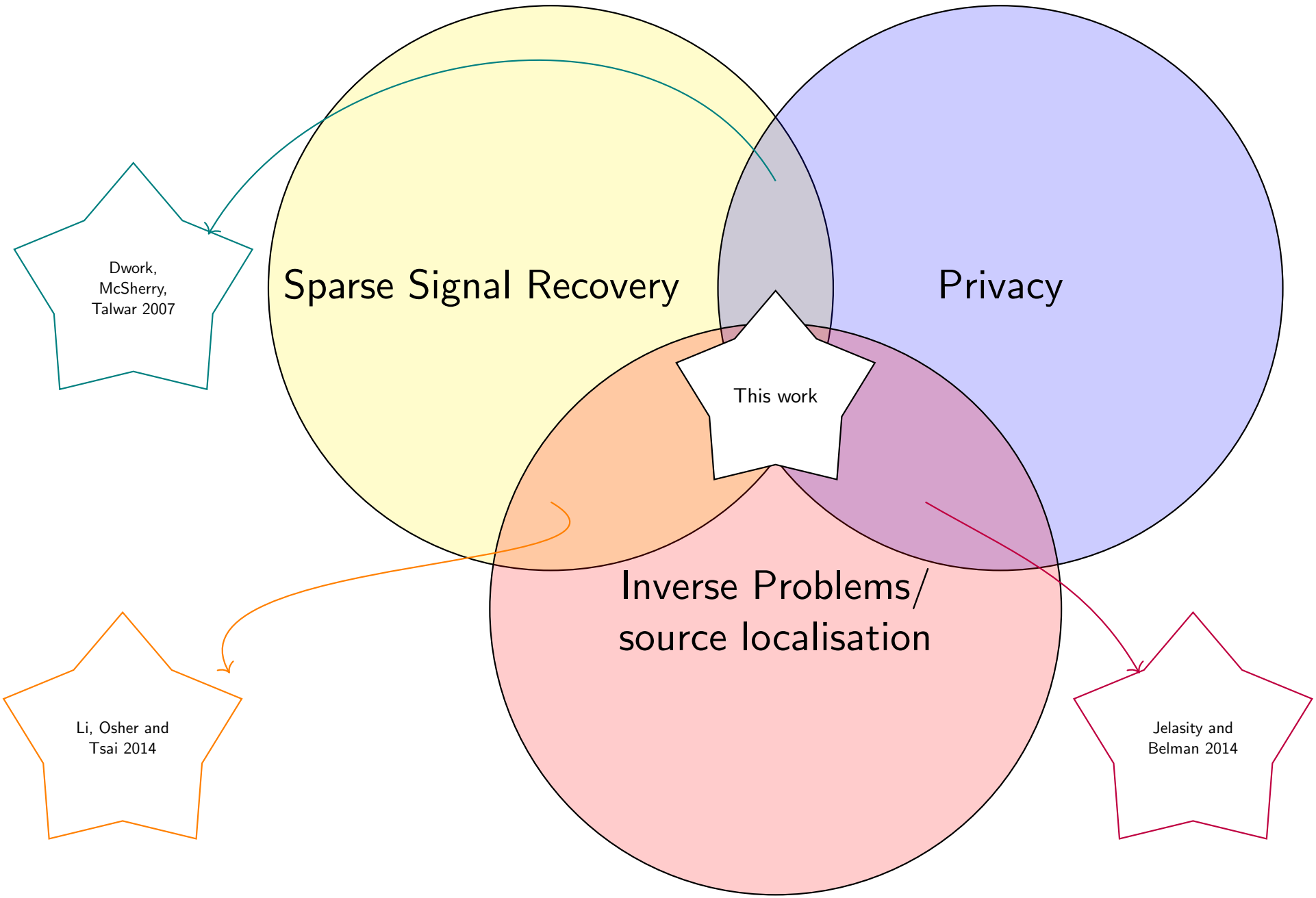
Why is diffusion interesting?

It is an example of ill-conditioning:

- The process is *technically* reversible.
- In the presence of even a small amount of noise, it is impossible to determine the original source location.

This is good for privacy!

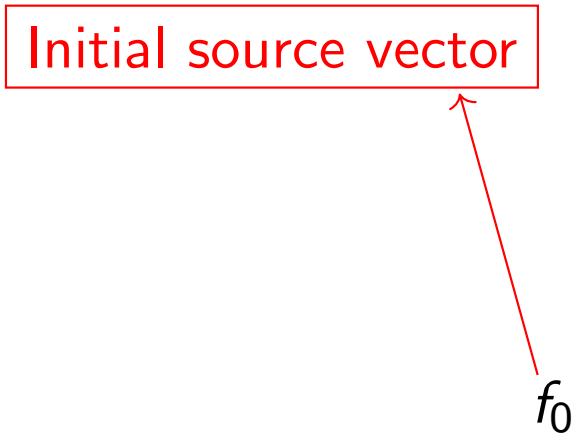

Dwork,
McSherry,
Talwar 2007
Sparse Signal Recovery
Privacy
Inverse Problems/
source localisation
Li, Osher and
Tsai 2014
Jelasity and
Belman 2014


Dwork,
McSherry,
Talwar 2007
Sparse Signal Recovery
Privacy
This work
Inverse Problems/
source localisation
Li, Osher and
Tsai 2014
Jelasity and
Belman 2014

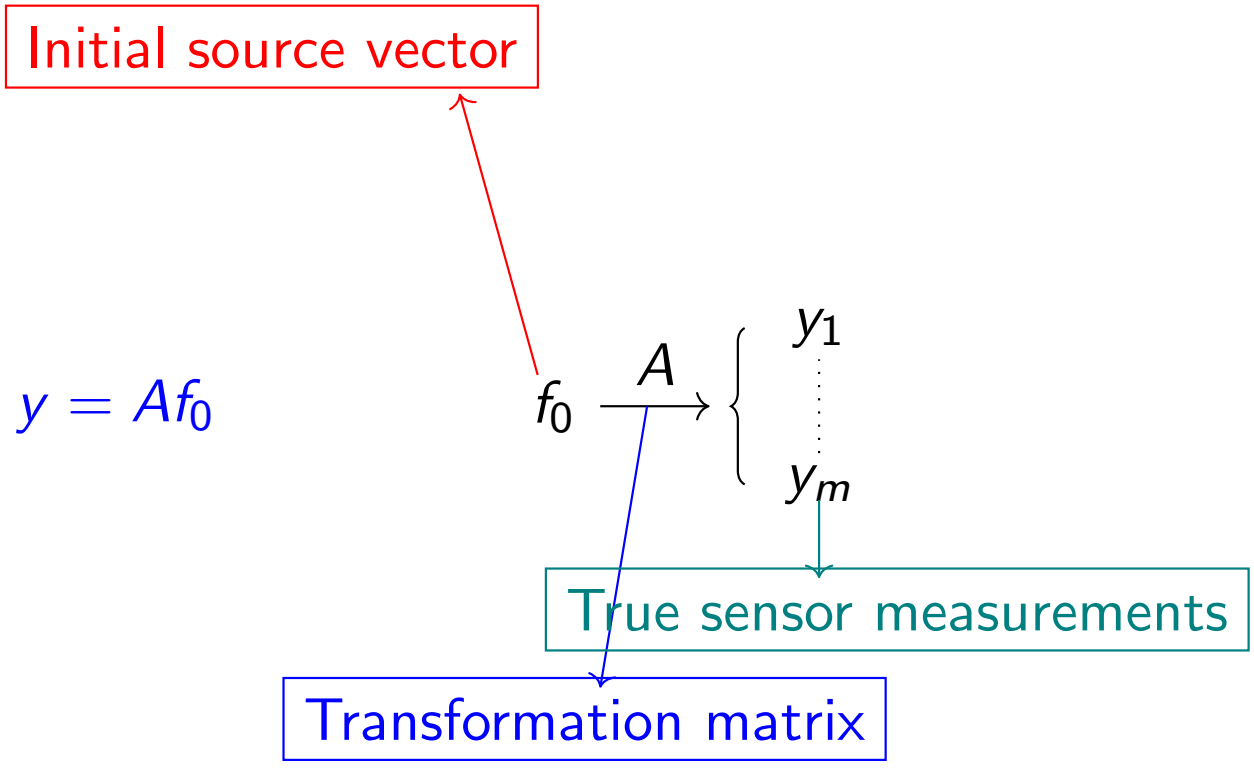
Structure:

1. Architecture of the problem
2. Definition of privacy
3. Relationship to Ill-conditioned inverse problems
4. Recovery Algorithm
5. Examples

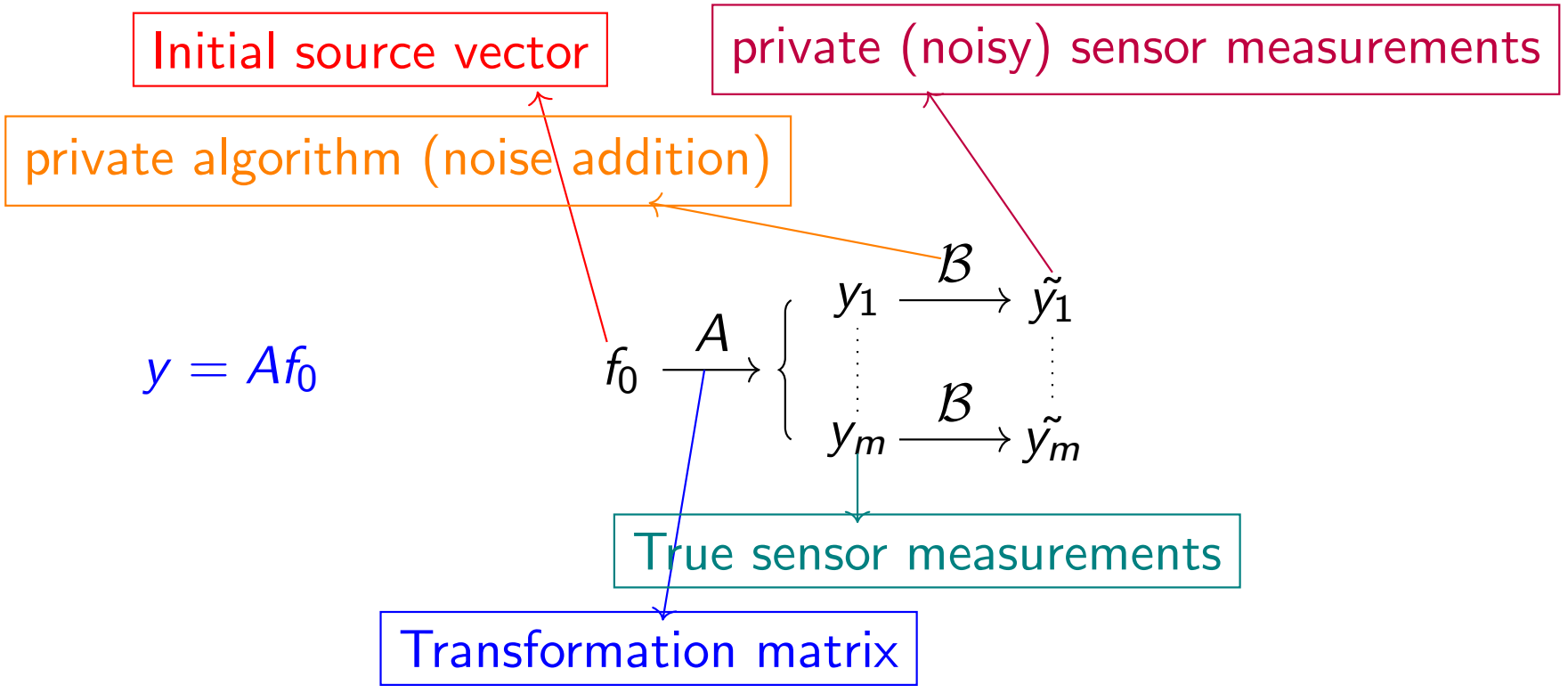
Architecture of (Locally) Private Linear Inverse Problems



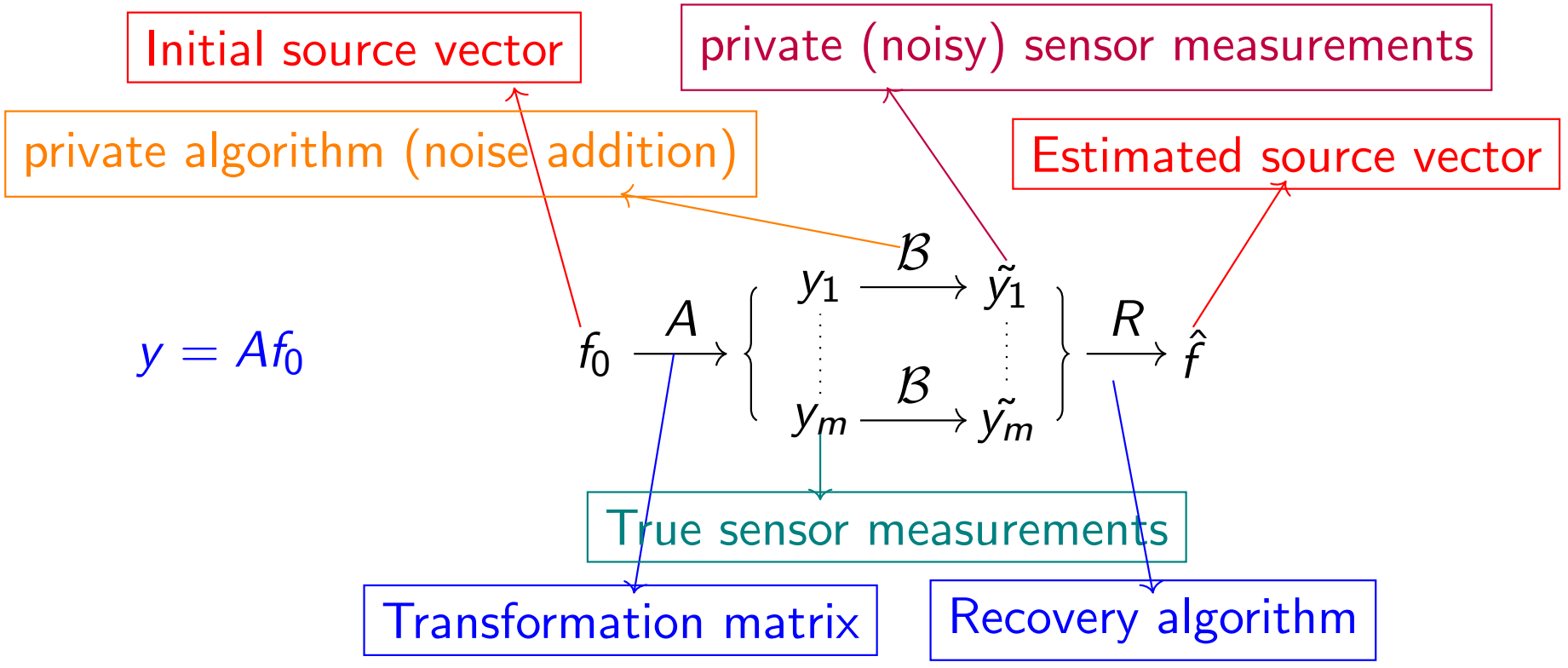
Architecture of (Locally) Private Linear Inverse Problems



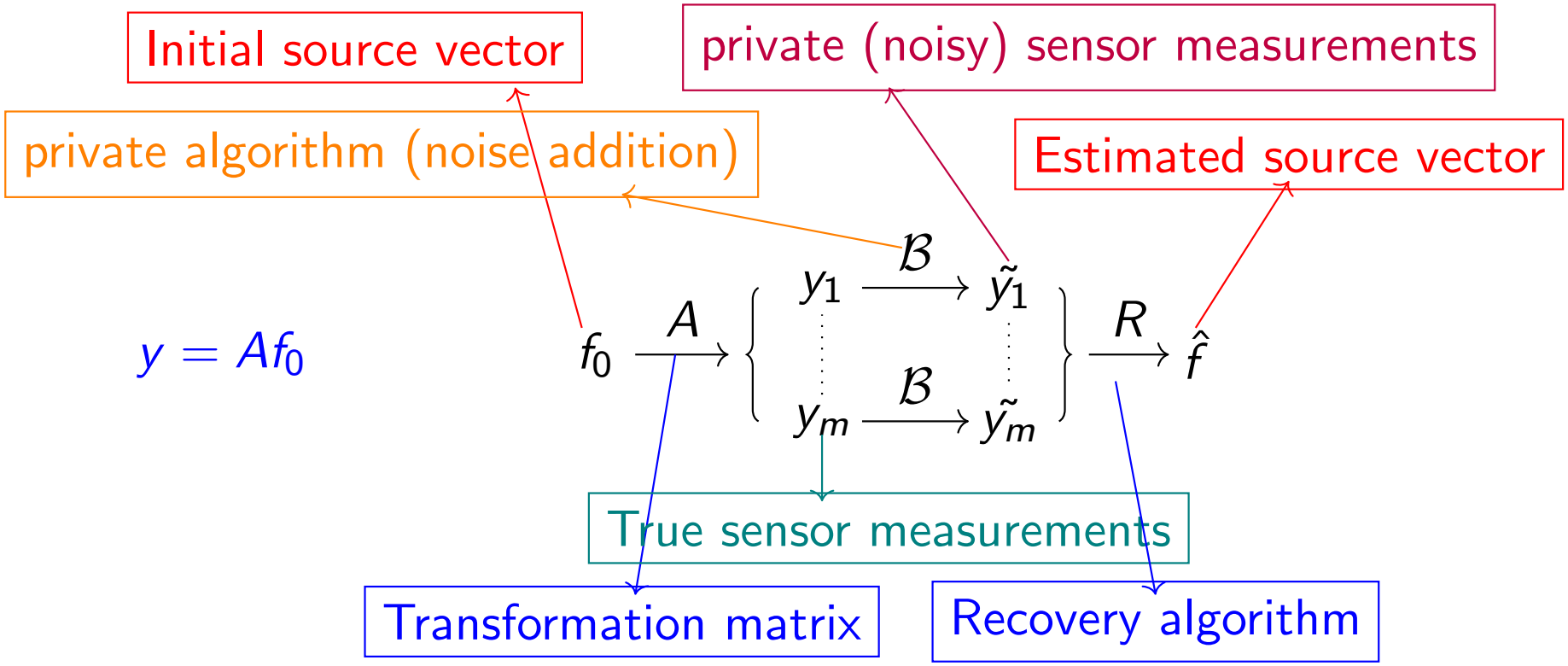
Architecture of (Locally) Private Linear Inverse Problems



Architecture of (Locally) Private Linear Inverse Problems



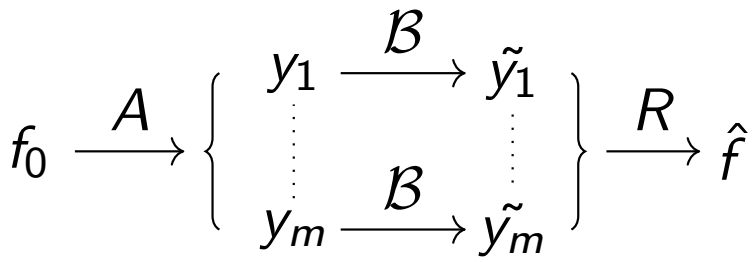
Architecture of (Locally) Private Linear Inverse Problems



We are applying $\mathcal{B}$ to individual sensor measurements:

- The data aggregator may be the person we don't trust.
- Physical sensor measurements are usually noisy already.
- There is less need to protect the data during transmission.
- Worse accuracy than if we collated the data first.

Main Questions



1. What does it mean for $\mathcal{B}$ to be private?
2. How should we design $\mathcal{B}$?
3. What algorithm should we use to recover?
4. In what way are $\hat{f}$ and f_0 close?

We're going to focus on Questions 1 & 2.

Earth Mover Distance (EMD)

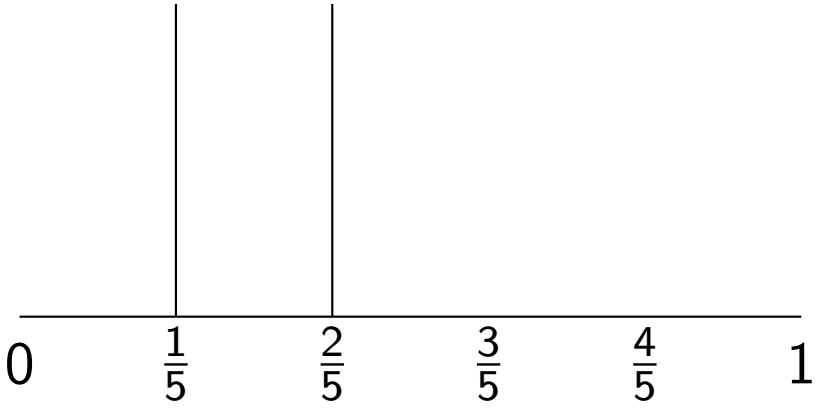
The EMD is a measure of how “geographically similar“ two probability distributions on a metric space are.

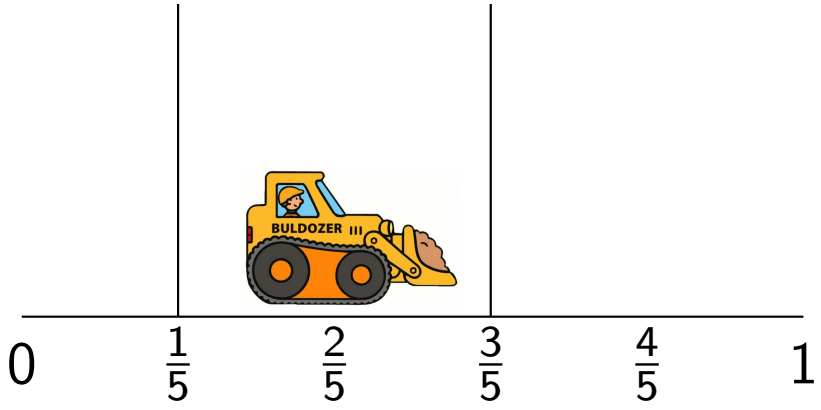
Earth Mover Distance (EMD)

The EMD is a measure of how "geographically similar" two probability distributions on a metric space are.

> The EMD between two probability distributions is the amount of **work** (mass $\times$ distance) required to transition between the two distributions.

For example, the following two distributions have EMD equal to 1/10.





Neighbouring source vectors

We assume that there exists a metric on the set of possible source locations, which induces the EMD on the set of source vectors. The coordinate $(f_0)_i$ corresponds to the intensity of the source at location i.

Two source vectors f_0 and f_0' are α-neighbours if $\mathrm{EMD}(f_0, f_0') \leq \alpha$.

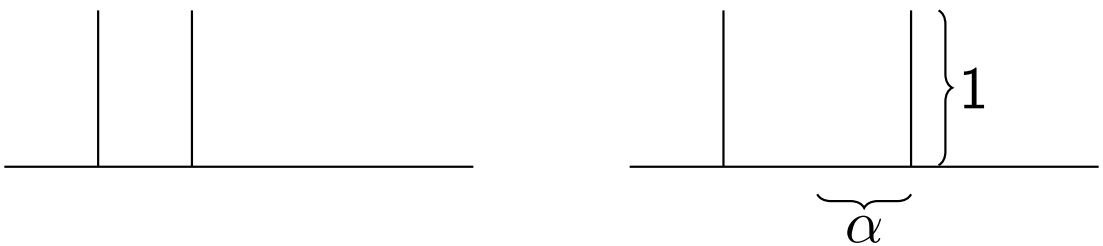
Neighbouring source vectors

We assume that there exists a metric on the set of possible source locations, which induces the EMD on the set of source vectors. The coordinate $(f_0)_i$ corresponds to the intensity of the source at location i.

Two source vectors f_0 and f_0' are α-neighbours if $\mathrm{EMD}(f_0, f_0') \leq \alpha$.

There are two important ways that vectors can be α-neighbours:

A large source can be moved a small distance.



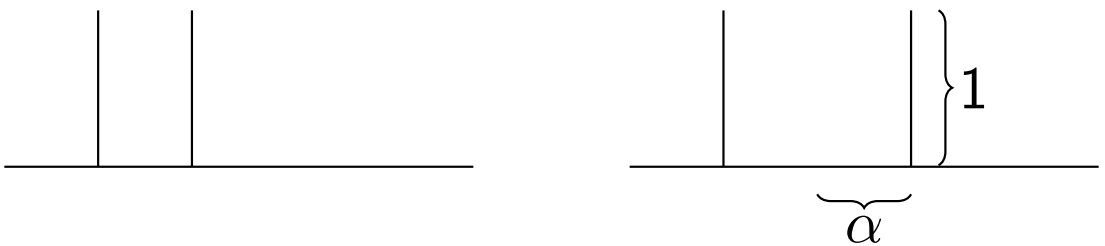
Neighbouring source vectors

We assume that there exists a metric on the set of possible source locations, which induces the EMD on the set of source vectors. The coordinate $(f_0)_i$ corresponds to the intensity of the source at location i.

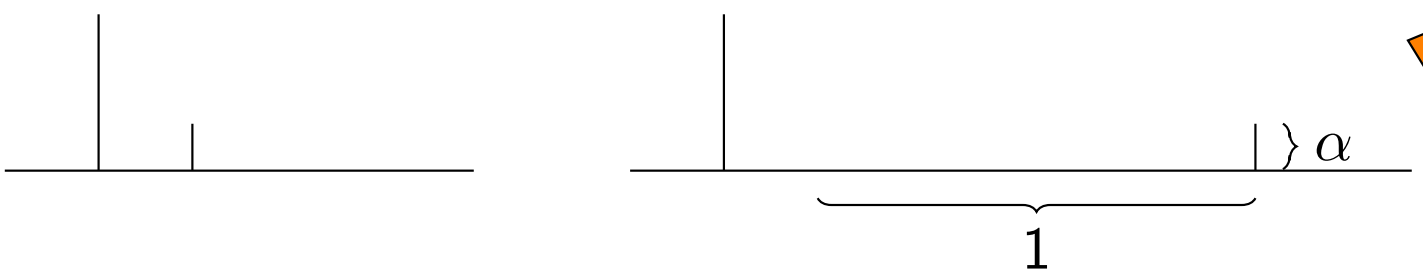
Two source vectors f_0 and f_0' are α-neighbours if $\mathrm{EMD}(f_0, f_0') \leq \alpha$.

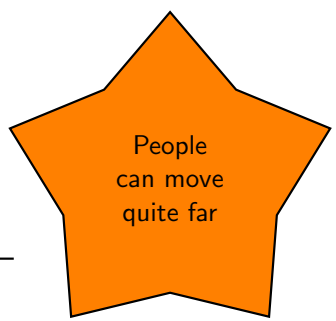
There are two important ways that vectors can be α-neighbours:

A large source can be moved a small distance.



A small source can be moved a large distance.





Privacy Guarantee

Our goal is to release a "version" of the measurement $y_i = (Af_0)_i$ that does not "reveal" the true source vector. We still want to be able to figure out the "approximate" source vector.

Privacy Guarantee

Our goal is to release a "version" of the measurement $y_i = (Af_0)_i$ that does not "reveal" the true source vector. We still want to be able to figure out the "approximate" source vector.

Returning to our heat source example: Suppose Mary would like to hid the fact that she was napping rather than studying. That is, she would like the sensor measurements to not reveal that she was on her bed rather than at her desk. Suppose also that the temperature distributions of these two scenarios are α-neighbouring.

Privacy Guarantee

Our goal is to release a "version" of the measurement $y_i = (Af_0)_i$ that does not "reveal" the true source vector. We still want to be able to figure out the "approximate" source vector.

Returning to our heat source example: Suppose Mary would like to hid the fact that she was napping rather than studying. That is, she would like the sensor measurements to not reveal that she was on her bed rather than at her desk. Suppose also that the temperature distributions of these two scenarios are α-neighbouring.

Our privacy guarantee to Mary:

- Someone looking at the private measurements will gain *almost* no new information about whether you were on you napping or studying.

Privacy Guarantee

Our goal is to release a "version" of the measurement $y_i = (Af_0)_i$ that does not "reveal" the true source vector. We still want to be able to figure out the "approximate" source vector.

Returning to our heat source example: Suppose Mary would like to hid the fact that she was napping rather than studying. That is, she would like the sensor measurements to not reveal that she was on her bed rather than at her desk. Suppose also that the temperature distributions of these two scenarios are α-neighbouring.

Our privacy guarantee to Mary:

- Someone looking at the private measurements will gain *almost* no new information about whether you were on you napping or studying.
- Any consequences you suffer as a result of the private measurement data being released were *almost* as likely to occur whether you were napping or studying.

Differential Privacy

A randomised algorithm $\mathcal{B}$ is $(\epsilon, \delta, \alpha)$-differentially private if for all α - neighbouring source vectors f_0, f_0' and events E,

$$e^{-\epsilon}\mathbb{P}(\mathcal{B}(y_i^{f_0'}) \in E) - \delta \leq \mathbb{P}(\mathcal{B}(y_i^{f_0}) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{B}(y_i^{f_0'}) \in E) + \delta.$$

The smaller ϵ and δ are, the "more private" the algorithm is.

Differential Privacy

A randomised algorithm $\mathcal{B}$ is $(\epsilon, \delta, \alpha)$-differentially private if for all α - neighbouring source vectors f_0, f_0' and events E,

$$e^{-\epsilon}\mathbb{P}(\mathcal{B}(y_i^{f_0'}) \in E) - \delta \leq \mathbb{P}(\mathcal{B}(y_i^{f_0}) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{B}(y_i^{f_0'}) \in E) + \delta.$$

The smaller ϵ and δ are, the "more private" the algorithm is.

Interpretation: Any outcome that occurs when the source vector is f_0 was almost as likely to have occurred if the source vector was instead f_0', α-neighbour of f_0. Which means that the outcome doesn't help us distinguish between f_0 and f_0'.

Differential Privacy

A randomised algorithm $\mathcal{B}$ is $(\epsilon, \delta, \alpha)$-differentially private if for all α - neighbouring source vectors f_0, f_0' and events E,

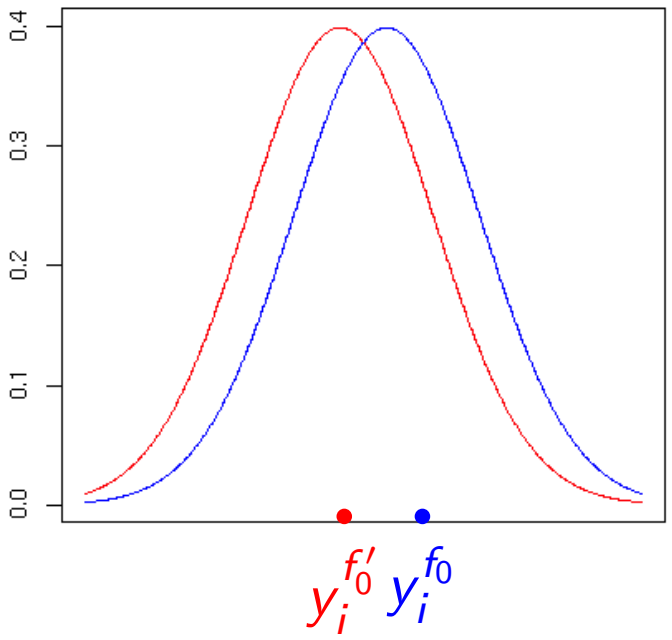
$$e^{-\epsilon}\mathbb{P}(\mathcal{B}(y_i^{f_0'}) \in E) - \delta \leq \mathbb{P}(\mathcal{B}(y_i^{f_0}) \in E) \leq e^{\epsilon}\mathbb{P}(\mathcal{B}(y_i^{f_0'}) \in E) + \delta.$$

The smaller ϵ and δ are, the "more private" the algorithm is.

Interpretation: Any outcome that occurs when the source vector is f_0 was almost as likely to have occurred if the source vector was instead f_0', α-neighbour of f_0. Which means that the outcome doesn't help us distinguish between f_0 and f_0'.

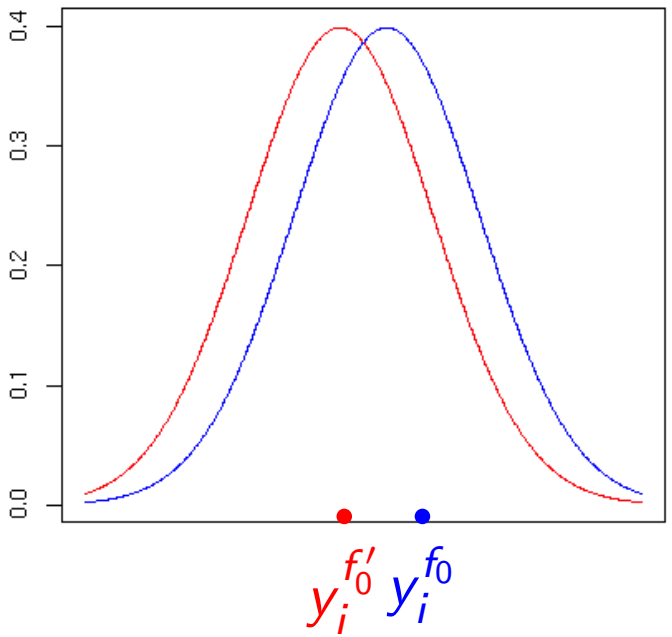
I can hide Mary's location within a much larger radius than I can hide the location of a fire.

The Gaussian mechanism (Dwork, Smith, McSherry, Nissim 2006)



- f_0 and f_0' are α-neighbouring source vectors.
- Choose the variance so that they look similar.
- If the true source vector is f_0 then y^{f_0} is sampled from the Gaussian centred at y^{f_0}.

The Gaussian mechanism (Dwork, Smith, McSherry, Nissim 2006)



- f_0 and f_0' are α-neighbouring source vectors.
- Choose the variance so that they look similar.
- If the true source vector is f_0 then y^{f_0} is sampled from the Gaussian centred at y^{f_0}.

Privacy: We can't tell which distribution it was sampled from because they are very similar.
Utility: Gaussians are light tailed so w.h.p. the sample will be close to the true value (the mean).

What should be variance be?

Lemma

$$\mathcal{B}(y_i^{f_0}) \sim N\left(y_i^{f_0}, \left(\frac{2\log(1.25/\delta)\triangle_2(A)}{\epsilon}\right)^2\right)$$

is a (ϵ, δ)-differentially private algorithm where

$$\triangle_2(A) = \max_{(f_0, f_0') \ \alpha\text{-neighbours}} \|y^{f_0} - y^{f_0'}\|_2$$

$$= \alpha \max_{e_i, e_j \text{ nearby sources}} \|A_i - A_j\|_2$$

Notice that this depends on ALL the sensor measurements! The more measurements you want to have, the larger your variance

What should be variance be?

Lemma

$$\mathcal{B}(y_i^{f_0}) \sim N\left(y_i^{f_0}, \left(\frac{2\log(1.25/\delta)\triangle_2(A)}{\epsilon}\right)^2\right)$$

is a (ϵ, δ)-differentially private algorithm where

$$\triangle_2(A) = \max_{(f_0, f_0') \ \alpha\text{-neighbours}} \|y^{f_0} - y^{f_0'}\|_2$$

$$= \alpha \max_{e_i, e_j \text{ nearby sources}} \|A_i - A_j\|_2$$

Notice that this depends on ALL the sensor measurements! The more measurements you want to have, the larger your variance

If the amount of noise we need to add to achieve privacy, $\triangle_2(A)$, is small then A must be nearly rank 1. In particular, it's spectrum is almost 1-sparse.

What should be variance be?

Lemma

$$\mathcal{B}(y_i^{f_0}) \sim N\left(y_i^{f_0}, \left(\frac{2\log(1.25/\delta)\triangle_2(A)}{\epsilon}\right)^2\right)$$

is a (ϵ, δ)-differentially private algorithm where

$$\triangle_2(A) = \max_{(f_0, f_0') \ \alpha\text{-neighbours}} \|y^{f_0} - y^{f_0'}\|_2$$

$$= \alpha \max_{e_i, e_j \text{ nearby sources}} \|A_i - A_j\|_2$$

Notice that this depends on ALL the sensor measurements! The more measurements you want to have, the larger your variance

If the amount of noise we need to add to achieve privacy, $\triangle_2(A)$, is small then A must be nearly rank 1. In particular, it's spectrum is almost 1-sparse. This is almost (but not quite) an equivalence.

What should be variance be?

Lemma

$$\mathcal{B}(y_i^{f_0}) \sim N\left(y_i^{f_0}, \left(\frac{2\log(1.25/\delta)\triangle_2(A)}{\epsilon}\right)^2\right)$$

is a (ϵ, δ)-differentially private algorithm where

$$\triangle_2(A) = \max_{(f_0, f_0') \ \alpha\text{-neighbours}} \|y^{f_0} - y^{f_0'}\|_2$$

$$= \alpha \max_{e_i, e_j \text{ nearby sources}} \|A_i - A_j\|_2$$

Notice that this depends on ALL the sensor measurements! The more measurements you want to have, the larger your variance

If the amount of noise we need to add to achieve privacy, $\triangle_2(A)$, is small then A must be nearly rank 1. In particular, it's spectrum is almost 1-sparse. This is almost (but not quite) an equivalence.

Side note: Gaussian noise is not the only type of noise we could have used. It is desirable because sensors usually already have Gaussian noise present. It's also computationally efficient to sample, which is nice for lightweight sensors.

Preserving Privacy for Ill-Conditioned Inverse Problems

A problem is ill-conditioned if the relative error of recovery is large. That is,

$$\kappa(A) = \left(\frac{\|A^{-1}(y + \triangle y)\|_2}{\|A^{-1}y\|_2} \Big/ \frac{\|(y + \triangle y)\|_2}{\|y\|_2} \right) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

Preserving Privacy for Ill-Conditioned Inverse Problems

A problem is ill-conditioned if the relative error of recovery is large. That is,

$$\kappa(A) = \left(\frac{\|A^{-1}(y + \triangle y)\|_2}{\|A^{-1}y\|_2} / \frac{\|(y + \triangle y)\|_2}{\|y\|_2} \right) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

Intuition: A matrix being ill-conditioned means that we only need a small amount of noise to mask the original data.

Preserving Privacy for Ill-Conditioned Inverse Problems

A problem is ill-conditioned if the relative error of recovery is large. That is,

$$\kappa(A) = \left(\frac{\|A^{-1}(y + \triangle y)\|_2}{\|A^{-1}y\|_2} / \frac{\|(y + \triangle y)\|_2}{\|y\|_2} \right) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

Intuition: A matrix being ill-conditioned means that we only need a small amount of noise to mask the original data.

If $\triangle_2(A)$ is small then the spectrum is dominated by the maximum singular value. The condition number $\kappa(A)$ only requires the minimum singular value to be much smaller than the maximum singular value.

Preserving Privacy for Ill-Conditioned Inverse Problems

A problem is ill-conditioned if the relative error of recovery is large. That is,

$$\kappa(A) = \left(\frac{\|A^{-1}(y + \triangle y)\|_2}{\|A^{-1}y\|_2} / \frac{\|(y + \triangle y)\|_2}{\|y\|_2} \right) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

Intuition: A matrix being ill-conditioned means that we only need a small amount of noise to mask the original data.

If $\triangle_2(A)$ is small then the spectrum is dominated by the maximum singular value. The condition number $\kappa(A)$ only requires the minimum singular value to be much smaller than the maximum singular value.

If the amount of noise, $\triangle_2(A)$, we need to add to maintain privacy is small then the problem is necessarily ill-conditioned.

Lemma Let A be a matrix such that $\|A\|_2 = 1$, then

$$\kappa(A) \geq \frac{\alpha}{\triangle_2(A)}$$

The relationship between privacy and ill-conditioning

Well-conditioned problems need to have a lot of noise added to maintain privacy

The relationship between privacy and ill-conditioning

Well-conditioned problems need to have a lot of noise added to maintain privacy

However, the converse is not necessarily true;

there is a fundamental difference between the notion of a problem being ill-conditioned and being easily kept private.

The relationship between privacy and ill-conditioning

Well-conditioned problems need to have a lot of noise added to maintain privacy

However, the converse is not necessarily true;

there is a fundamental difference between the notion of a problem being ill-conditioned and being easily kept private.

A matrix may be ill-conditioned but still require a large amount of noise to maintain privacy:
Example: Assume $\rho << 1$ and consider the general inverse problem $y = Ax$ where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \rho \end{pmatrix}, \qquad A \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ \rho x_1 \end{pmatrix}$$

Then $\kappa(A) = 1/\rho$ is large so recovery is "difficult" but we still need to add a significant amount of noise to mask the first coordinate.

Are the noisy measurements useful? Basis Pursuit Denoising

If $\tilde{y}$ is the noisy version of our measurement vector y then

$$\hat{f} = \arg \min_{f \in [0,1]^n} \|f\|_1 \quad \text{s.t.} \quad \|Af - \tilde{y}\|_2 \leq \sigma\sqrt{m}$$

the ℓ_1-norm promotes sparsity in the solution

This ensures that f_0 is a feasible point w.h.p.

Are the noisy measurements useful? Basis Pursuit Denoising

If $\tilde{y}$ is the noisy version of our measurement vector y then

$$\hat{f} = \arg \min_{f \in [0,1]^n} \|f\|_1 \quad \text{s.t.} \quad \|Af - \tilde{y}\|_2 \leq \sigma\sqrt{m}$$

the ℓ_1-norm promotes sparsity in the solution

This ensures that f_0 is a feasible point w.h.p.

Where does this algorithm come from?

- This algorithm is commonly used for sparse signal recovery when the matrix A is well-behaved.

Are the noisy measurements useful? Basis Pursuit Denoising

If $\tilde{y}$ is the noisy version of our measurement vector y then

$$\hat{f} = \arg \min_{f \in [0,1]^n} \|f\|_1 \quad \text{s.t.} \quad \|Af - \tilde{y}\|_2 \leq \sigma\sqrt{m}$$

the ℓ_1-norm promotes sparsity in the solution

This ensures that f_0 is a feasible point w.h.p.

Where does this algorithm come from?

- This algorithm is commonly used for sparse signal recovery when the matrix A is well-behaved.
- The use of this algorithm for the poorly behaved heat matrix was proposed in [Li, Osher and Tsai 2014].

Are the noisy measurements useful? Basis Pursuit Denoising

If $\tilde{y}$ is the noisy version of our measurement vector y then

$$\hat{f} = \arg \min_{f \in [0,1]^n} \|f\|_1 \quad \text{s.t.} \quad \|Af - \tilde{y}\|_2 \leq \sigma\sqrt{m}$$

the ℓ_1-norm promotes sparsity in the solution

This ensures that f_0 is a feasible point w.h.p.

Where does this algorithm come from?

- This algorithm is commonly used for sparse signal recovery when the matrix A is well-behaved.
- The use of this algorithm for the poorly behaved heat matrix was proposed in [Li, Osher and Tsai 2014].
- Matrices that describe diffusion are very far from well-behaved so it's surprising that this algorithm would be effective.

Are the noisy measurements useful? Basis Pursuit Denoising

If $\tilde{y}$ is the noisy version of our measurement vector y then

$$\hat{f} = \arg \min_{f \in [0,1]^n} \|f\|_1 \quad \text{s.t.} \quad \|Af - \tilde{y}\|_2 \leq \sigma\sqrt{m}$$

the ℓ_1-norm promotes sparsity in the solution

This ensures that f_0 is a feasible point w.h.p.

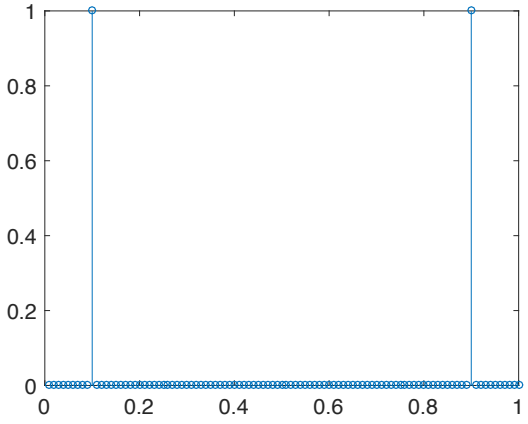
Where does this algorithm come from?

- This algorithm is commonly used for sparse signal recovery when the matrix A is well-behaved.
- The use of this algorithm for the poorly behaved heat matrix was proposed in [Li, Osher and Tsai 2014].
- Matrices that describe diffusion are very far from well-behaved so it's surprising that this algorithm would be effective.

Error Metric: $\text{EMD}(f_0, \hat{f})$

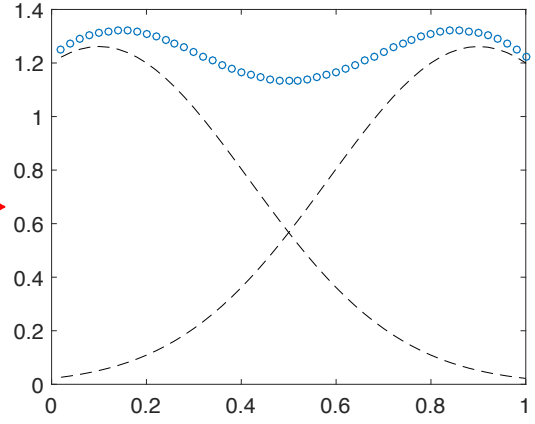
Example 1: Heat diffusion on the discrete, 1D unit interval

Source Vector

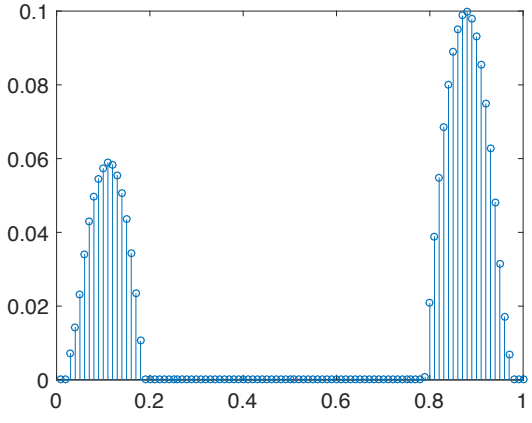


time elapses

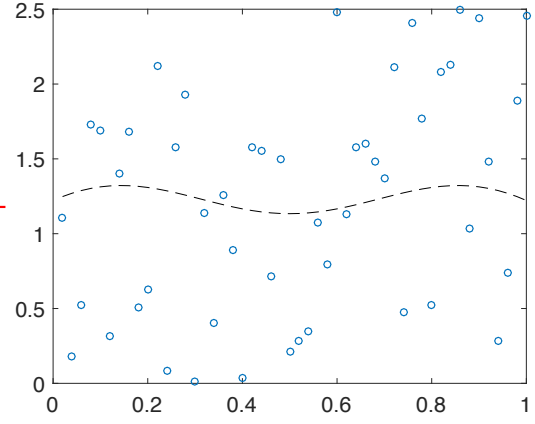
Sensor measurements



Each sensor adds noise $\Theta\left(\frac{\alpha\sqrt{m}\ln(1/\delta)}{T^{1.5}\epsilon}\right)$



Recovery



Estimated source vector

Noisy sensor measurements

The Error Guarantee

We added enough noise to ensure that there is uncertainty in the exact location of heat source. Thus, by design, our recovery algorithm can not perform well in norms like the ℓ_1 and ℓ_2 norms.

The Error Guarantee

We added enough noise to ensure that there is uncertainty in the exact location of heat source. Thus, by design, our recovery algorithm can not perform well in norms like the ℓ_1 and ℓ_2 norms.

If $\mathrm{EMD}(f_0, \hat{f})$ is small means then even though we may not be able to pinpoint exactly where the heat sources are, we can say *approximately* where they are. The previous experimental results suggest that $\hat{f}$ may indeed be close in EMD.

The Error Guarantee

We added enough noise to ensure that there is uncertainty in the exact location of heat source. Thus, by design, our recovery algorithm can not perform well in norms like the ℓ_1 and ℓ_2 norms.

If $\mathrm{EMD}(f_0, \hat{f})$ is small means then even though we may not be able to pinpoint exactly where the heat sources are, we can say *approximately* where they are. The previous experimental results suggest that $\hat{f}$ may indeed be close in EMD.

We have upper bounds on $\mathrm{EMD}(f_0, \hat{f})$ in the paper, but the most interesting take-aways are:

The Error Guarantee

We added enough noise to ensure that there is uncertainty in the exact location of heat source. Thus, by design, our recovery algorithm can not perform well in norms like the ℓ_1 and ℓ_2 norms.

If $\mathrm{EMD}(f_0, \hat{f})$ is small means then even though we may not be able to pinpoint exactly where the heat sources are, we can say *approximately* where they are. The previous experimental results suggest that $\hat{f}$ may indeed be close in EMD.

We have upper bounds on $\mathrm{EMD}(f_0, \hat{f})$ in the paper, but the most interesting take-aways are:

- The error increases both as $t \to \infty$ and $t \to 0$.
  - If t is large then it's hard to recover.
  - If t is small then it's hard to maintain privacy

The Error Guarantee

We added enough noise to ensure that there is uncertainty in the exact location of heat source. Thus, by design, our recovery algorithm can not perform well in norms like the ℓ_1 and ℓ_2 norms.

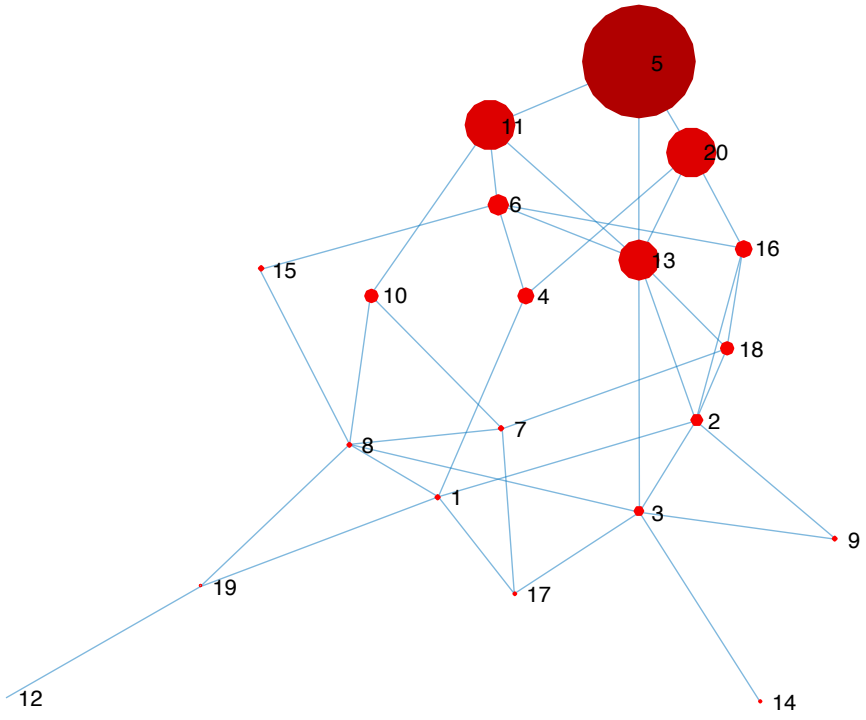
If $\text{EMD}(f_0, \hat{f})$ is small means then even though we may not be able to pinpoint exactly where the heat sources are, we can say *approximately* where they are. The previous experimental results suggest that $\hat{f}$ may indeed be close in EMD.

We have upper bounds on $\text{EMD}(f_0, \hat{f})$ in the paper, but the most interesting take-aways are:

- The error increases both as $t \to \infty$ and $t \to 0$.
  - If t is large then it's hard to recover.
  - If t is small then it's hard to maintain privacy
- The error is asymptotically constant in the number of sensors.
  - More (noisy) measurements typically means more accurate data.
  - If we have more sensors then we need to increase the variance of the noise.

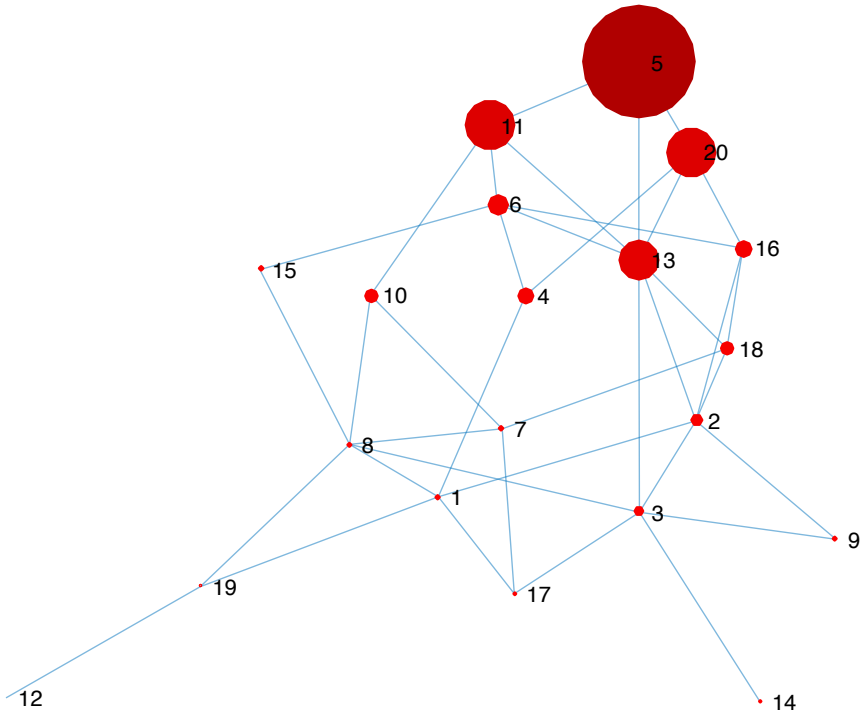
Example 2: Diffusion on graphs (spread of information in social networks)

Let G be a connected, undirected graph on n nodes.



Example 2: Diffusion on graphs (spread of information in social networks)

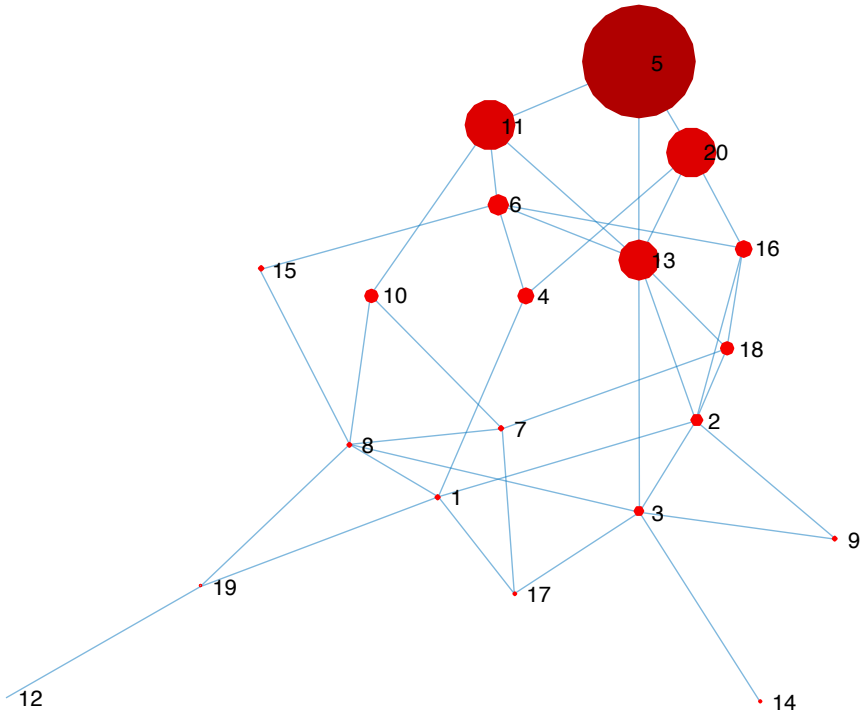
Let G be a connected, undirected graph on n nodes.



- D be the diagonal matrix such that D_{ii} is the degree of the ith vertex of G
- W be the weight matrix, W_{ij} is the weight of the edge between i and j.
- $L = D - W$ is called the Laplacian of G.

Example 2: Diffusion on graphs (spread of information in social networks)

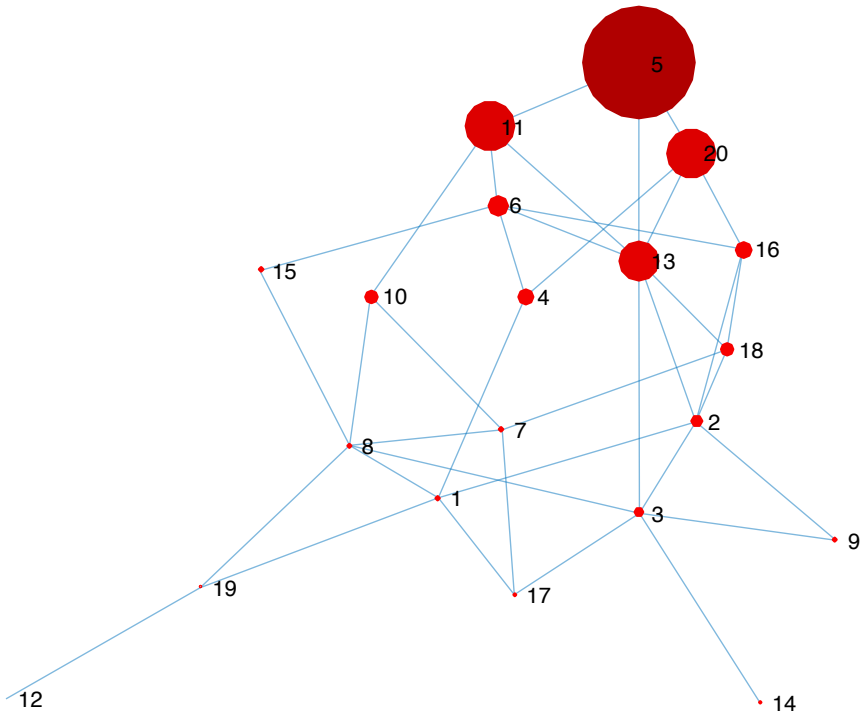
Let G be a connected, undirected graph on n nodes.



- D be the diagonal matrix such that D_{ii} is the degree of the ith vertex of G
- W be the weight matrix, W_{ij} is the weight of the edge between i and j.
- $L = D - W$ is called the Laplacian of G.
- $A_G = e^{-\tau L}$ is the diffusion matrix where τ is the rate of diffusion.

Example 2: Diffusion on graphs (spread of information in social networks)

Let G be a connected, undirected graph on n nodes.



- D be the diagonal matrix such that D_{ii} is the degree of the ith vertex of G
- W be the weight matrix, W_{ij} is the weight of the edge between i and j.
- $L = D - W$ is called the Laplacian of G.
- $A_G = e^{-\tau L}$ is the diffusion matrix where τ is the rate of diffusion.

The concentration (or probability) at the vertices after diffusion is given by

$$y = A_G f_0.$$

Spectral Properties of $\Delta_2(A_G)$

Let u_i be the ith row of the matrix whose columns are the left singular vectors of the Laplacian L. Let $s_1 \leq s_2 \leq \cdots$ be the singular values of L.

Lemma For any graph G, the singular values of A_G are $e^{-\tau s_1} \geq e^{-\tau s_2} \geq \cdots$

$$\Delta_2(A_G) \leq \max_{i,j \text{ nearby}} \sum_{k=2}^{\min\{n,m\}} e^{-\tau s_i} |(u_i)_k - (u_j)_k| \leq \max_{i,j \text{ nearby}} e^{-\tau s_2} \|u_i - u_j\|_1$$

Spectral Properties of $\Delta_2(A_G)$

Let u_i be the ith row of the matrix whose columns are the left singular vectors of the Laplacian L. Let $s_1 \leq s_2 \leq \cdots$ be the singular values of L.

Lemma For any graph G, the singular values of A_G are $e^{-\tau s_1} \geq e^{-\tau s_2} \geq \cdots$

$$\Delta_2(A_G) \leq \max_{i,j \text{ nearby}} \sum_{k=2}^{\min\{n,m\}} e^{-\tau s_i} |(u_i)_k - (u_j)_k| \leq \max_{i,j \text{ nearby}} e^{-\tau s_2} \|u_i - u_j\|_1$$

The second smallest eigenvalue, s_2, is called the algebraic connectivity of G. Aptly named for it's relationship to the connectivity of the graph (e.g. average distance, Cheeger constant etc.). It gets bigger the more connected the graph is.

The more connected the graph is, the less noise we need to add to maintain privacy.

Spectral Properties of $\Delta_2(A_G)$

Let u_i be the ith row of the matrix whose columns are the left singular vectors of the Laplacian L. Let $s_1 \leq s_2 \leq \cdots$ be the singular values of L.

Lemma For any graph G, the singular values of A_G are $e^{-\tau s_1} \geq e^{-\tau s_2} \geq \cdots$

$$\Delta_2(A_G) \leq \max_{i,j \text{ nearby}} \sum_{k=2}^{\min\{n,m\}} e^{-\tau s_i} |(u_i)_k - (u_j)_k| \leq \max_{i,j \text{ nearby}} e^{-\tau s_2} \|u_i - u_j\|_1$$

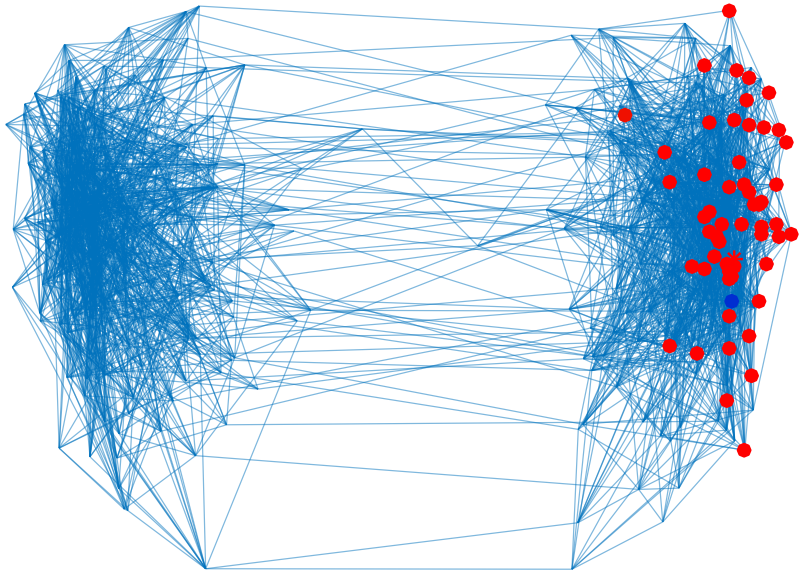
The second smallest eigenvalue, s_2, is called the algebraic connectivity of G. Aptly named for it's relationship to the connectivity of the graph (e.g. average distance, Cheeger constant etc.). It gets bigger the more connected the graph is.

The more connected the graph is, the less noise we need to add to maintain privacy.

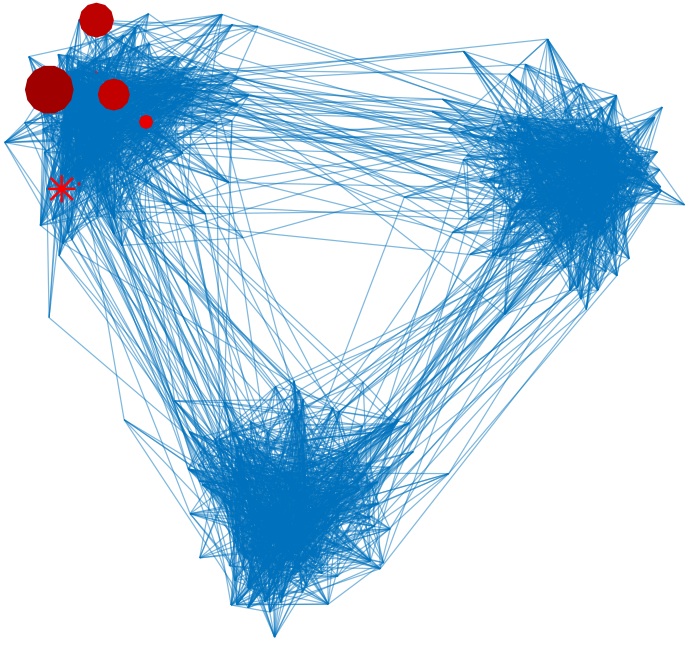
This makes sense because the more connected the graph is, the faster diffusion occurs.

Recovery on Graphs with Community Structure

The graph G was drawn from a stochastic block model with intracommunity probability 5% and intercommunity probability 0.1%.



$\tau = 2, n = 500, \delta = 0.1, \epsilon = 4$



$\tau = 3, n = 750, \delta = 0.1, \epsilon = 5$

Take-home messages

- It is possible to produce locally differentially private sensor measurements from which one **can** determine the general vicinity of the diffusion sources but **can not** infer the exact locations of the sources.
- Your error metric and specific notion of privacy matter a lot when deciding whether a problem is feasible.

Future Direction: Continual Monitoring

In practice sensor measurements are collated and analysed continually.

- Privacy degrades over time at a rate of about $\sqrt{t}$. If I want to release t statistics about my data and each is (ϵ, δ)-DP then the result is $(\sqrt{2t \ln 1/\delta'}\epsilon + t\epsilon(e^{\epsilon} - 1), t\delta + \delta')$-DP.

Future Direction: Continual Monitoring

In practice sensor measurements are collated and analysed continually.

- Privacy degrades over time at a rate of about $\sqrt{t}$. If I want to release t statistics about my data and each is (ϵ, δ)-DP then the result is $(\sqrt{2t \ln 1/\delta'}\epsilon + t\epsilon(e^{\epsilon} - 1), t\delta + \delta')$-DP.
- However, it is generally assumed that each "secret" has the same effect on every statistic.

Future Direction: Continual Monitoring

In practice sensor measurements are collated and analysed continually.

- Privacy degrades over time at a rate of about $\sqrt{t}$. If I want to release t statistics about my data and each is (ϵ, δ)-DP then the result is $(\sqrt{2t \ln 1/\delta'}\epsilon + t\epsilon(e^{\epsilon} - 1), t\delta + \delta')$-DP.
- However, it is generally assumed that each "secret" has the same effect on every statistic.
- This assumption is not true for time series data of physical measurements.

Imagine we are trying to keep secret that Mary had a party on Friday night. The party affects the thermal measurements on Friday night a lot, but hardly affects the thermal measurements on Saturday at all.

Perhaps rather than losing $\sqrt{t}$ privacy, we only lose privacy proportional to the length of time the secret *meaningfully* affects the statistics.

Individual vs. Event Level Privacy

Generally, we are not looking to keep individual events private, but rather patterns of behaviour

Imagine Mary has such a party every week. Even though each individual party has limited effect, the collection of parties has a long term effect on the private data.

Question: Can modeling patterns of behaviour be used to assist in minimising the privacy loss in continual monitoring problems?

Future Directions

- Are there other statistics that perform well when we consider the EMD as a privacy notion and/or recovery metric?
- Are there better algorithms than Basis Pursuit Denoising?
- Are the measurements of ill-conditioned matrices somehow "morally" private? or partially private?
- The rows u_i that appeared in the analysis of diffusion on graphs also arise elsewhere. In numerical analysis, their ℓ_2 norms are called *leverage scores*... They are used in clustering algorithms and low dimensional embeddings.. Is there a connection here?
- Can similar techniques be used to argue about other types of functional data?

Thank you!