# Leveraging Similarity Joins for Signal Reconstruction

Abolfazl Asudeh | Azade Nazi | **Jees Augustine** | Saravanan Thirumuruganathan | Nan Zhang | Gautam Das | Divesh Srivastava







Motivation

Problem Formulation

Contribution

Algorithms

Experiments & Evaluation

Motivation

Problem Formulation

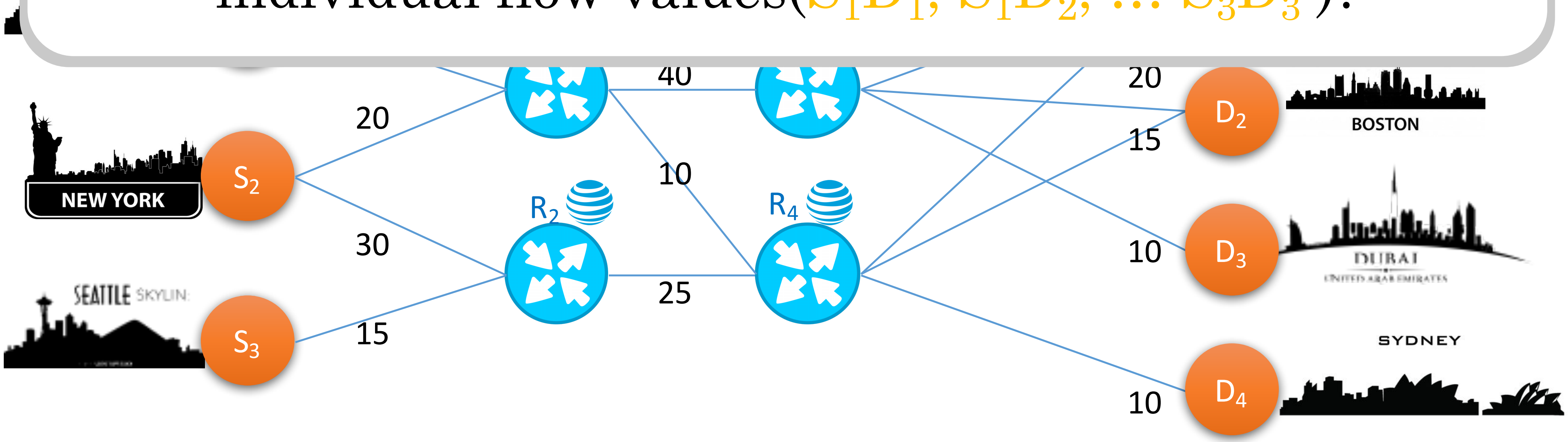
Contribution

Algorithms

Experiments & Evaluation

# Motivation

Given any traffic routing matrix and aggregated link level flow information, can we effectively infer the individual flow values($S_1D_1$, $S_1D_2$, … $S_3D_3$ )?



# Scope of Problem High Dimensional Signal

1. 3D image reconstruction from 2D images

2. Accurate temperature estimate from limited temperature sensors

Motivation

Problem Formulation

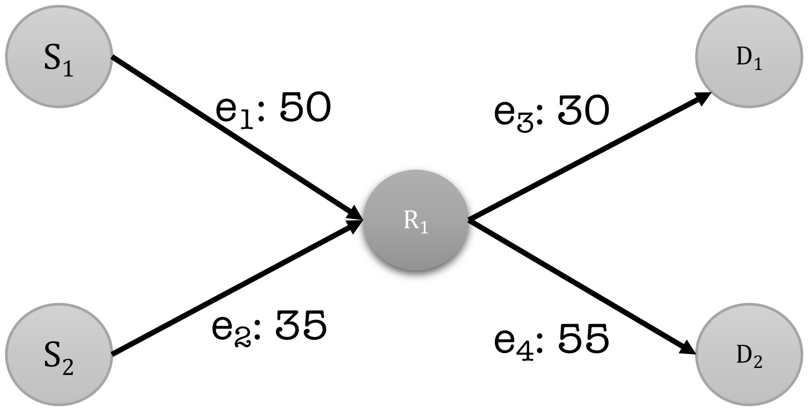
Contribution

Algorithms

Experiments & Evaluation

DBXLab
UT ARLINGTON

# Problem Representation




$$\Rightarrow \mathcal{A} = \begin{array}{c|cccc} & S_1D_1 & S_1D_2 & S_2D_1 & S_2D_2 \\ \hline e_1 & 1 & 1 & & \\ e_2 & & & 1 & 1 \\ e_3 & 1 & & 1 & \\ e_4 & & 1 & & 1 \end{array}$$

$$\Rightarrow b = \begin{array}{c|c} e_1 & 50 \\ e_2 & 35 \\ e_3 & 30 \\ e_4 & 55 \end{array}$$

# Signal Reconstruction Problem(SRP)

$$\mathcal{A} \cdot \mathcal{X} \to b$$

X: SD Traffic Vector $\Rightarrow \begin{bmatrix} SD_1 \\ SD2 \\ \ldots\ldots \\ SD_m \end{bmatrix}$

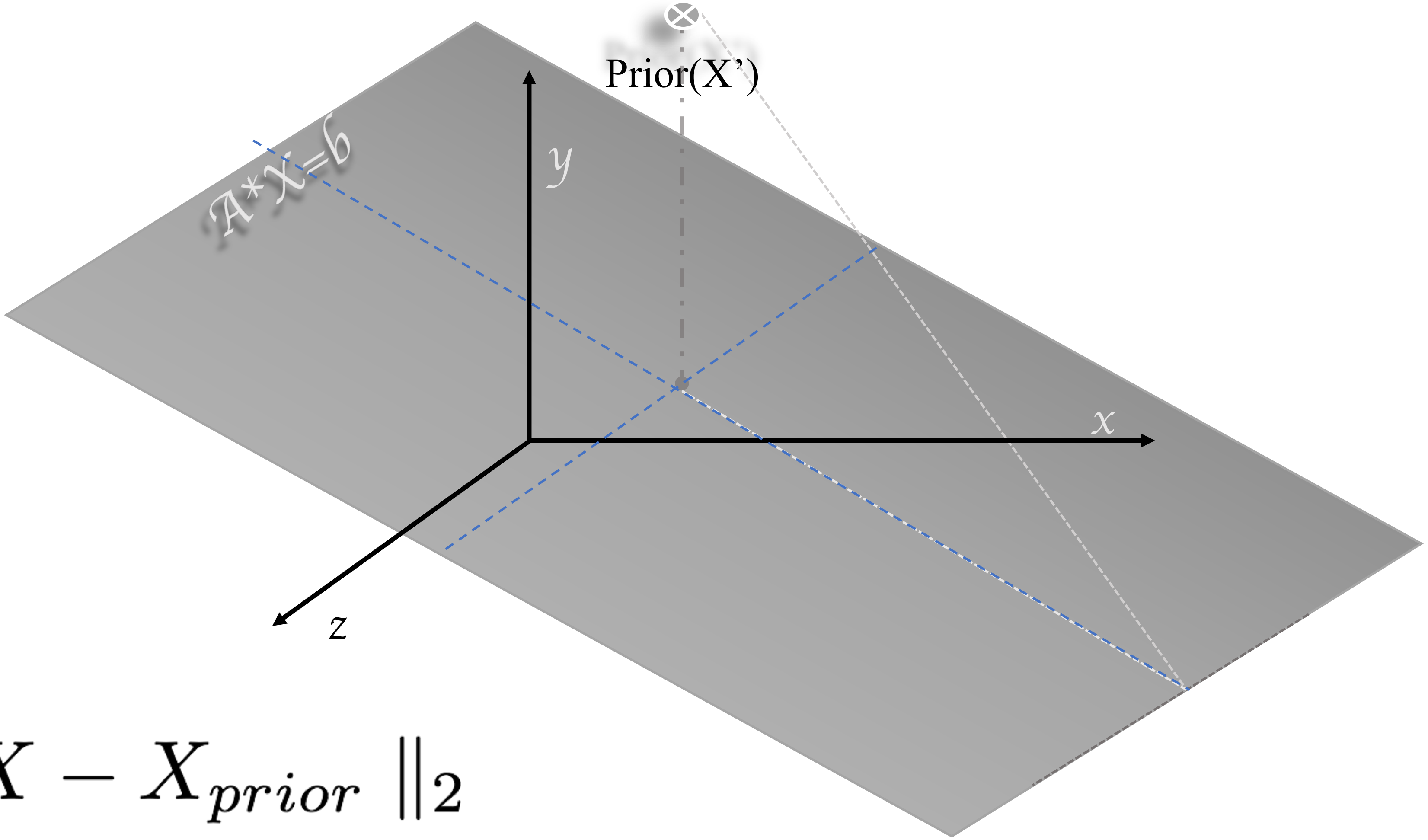
# Existing Solutions

- Compressive Sensing
  - Assume that most of signal elements are zeros(0), this sparsity could lead to reconstruction with fewer samples
  - Large Time requirement
  - Large error in answers

$$A \cdot x \to b$$

Can we do better with some prior information about the signal !

# Visual Representation




$$\min \|X - X_{prior}\|_2$$
$$s.t.\ AX = b$$

Motivation

Problem Formulation

Contribution

Algorithms

Experiments & Evaluation

# Contributions

- Derived the Lagrangian Dual form of the problem and proposed DIRECT-Exact algorithm
- Identified computational bottleneck
- Leveraged Database techniques for Optimized DIRECT-Approximate as a scalable solution using set similarity join techniques
- Performed Extensive Experiments to confirm the efficiency and accuracy

Motivation

Problem Formulation

Contribution

Algorithms

Experiments & Evaluation

# Lagrangian Dual Expression

- Any general optimization problem in the form of

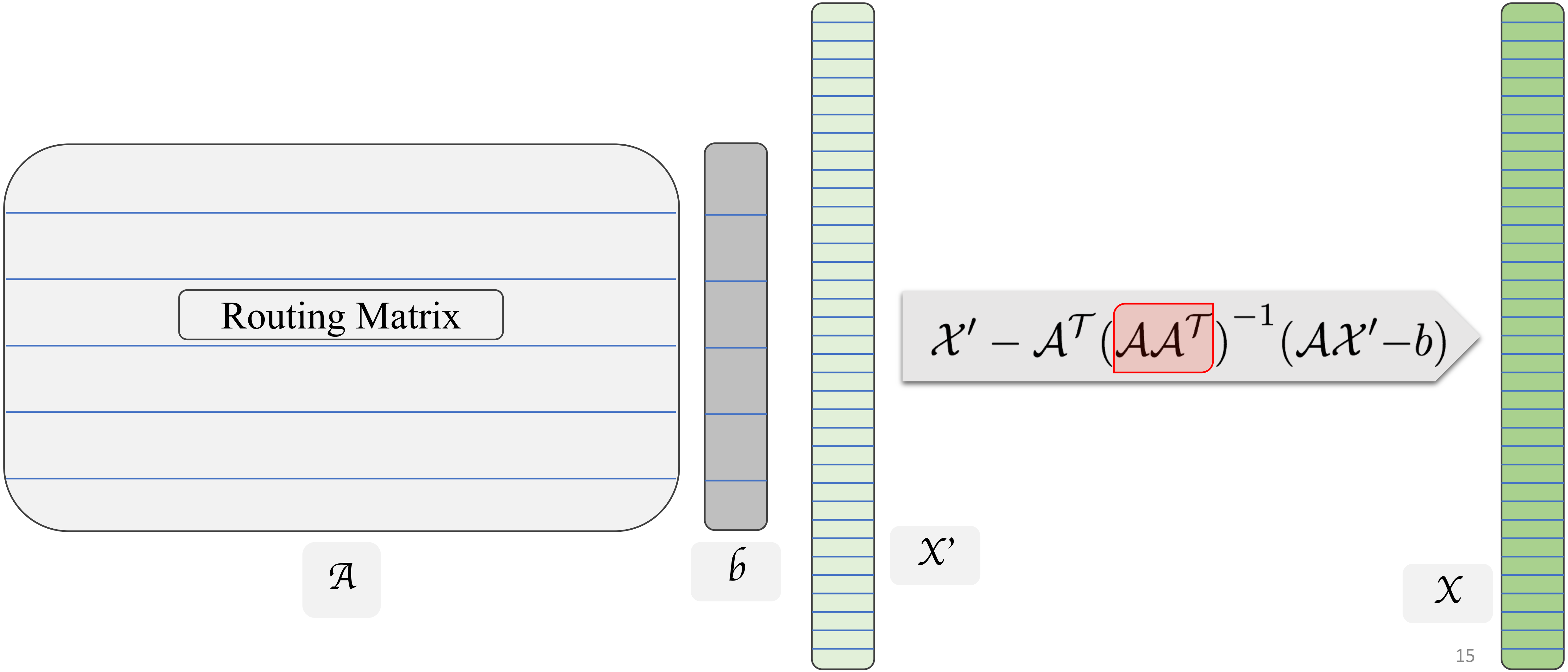
$$min\ f(X)$$
$$s.t.\ g(X) = b$$

- Can be rewritten as

$$L(X, \lambda) = f(X) + \lambda^T(g(X) - b)$$

$$\boxed{L(X, \lambda) = \frac{1}{2}X^T X - X'^T X + \lambda^T(AX - b)}$$

# Direct


Routing Matrix
$\mathcal{X}' - \mathcal{A}^T(\boxed{\mathcal{A}\mathcal{A}^T})^{-1}(\mathcal{A}\mathcal{X}' - b)$
$\mathcal{A}$
$b$
$\mathcal{X}'$
$\mathcal{X}$

# Optimizing computation of $AA^T$

- Sparse representation of A & $A^T$

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 |

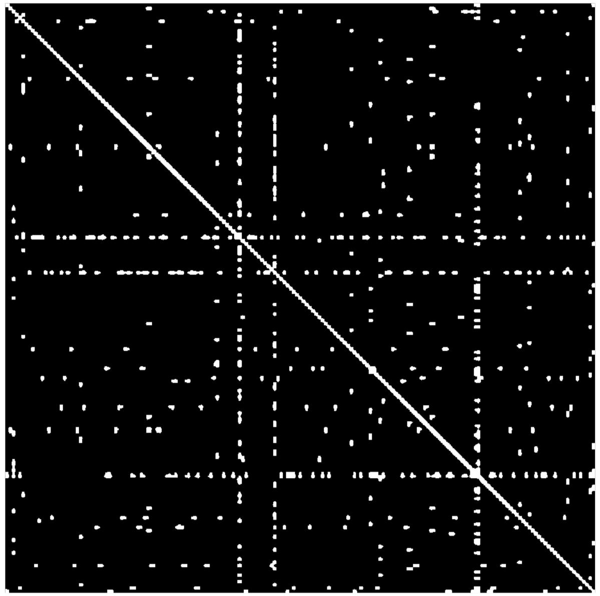
| |
|---|
| < 3, 6 > |
| < 2 > |
| < 4, 6, 7 > |
| < 1, 5 > |

# Approximation: Trading off Accuracy with Efficiency

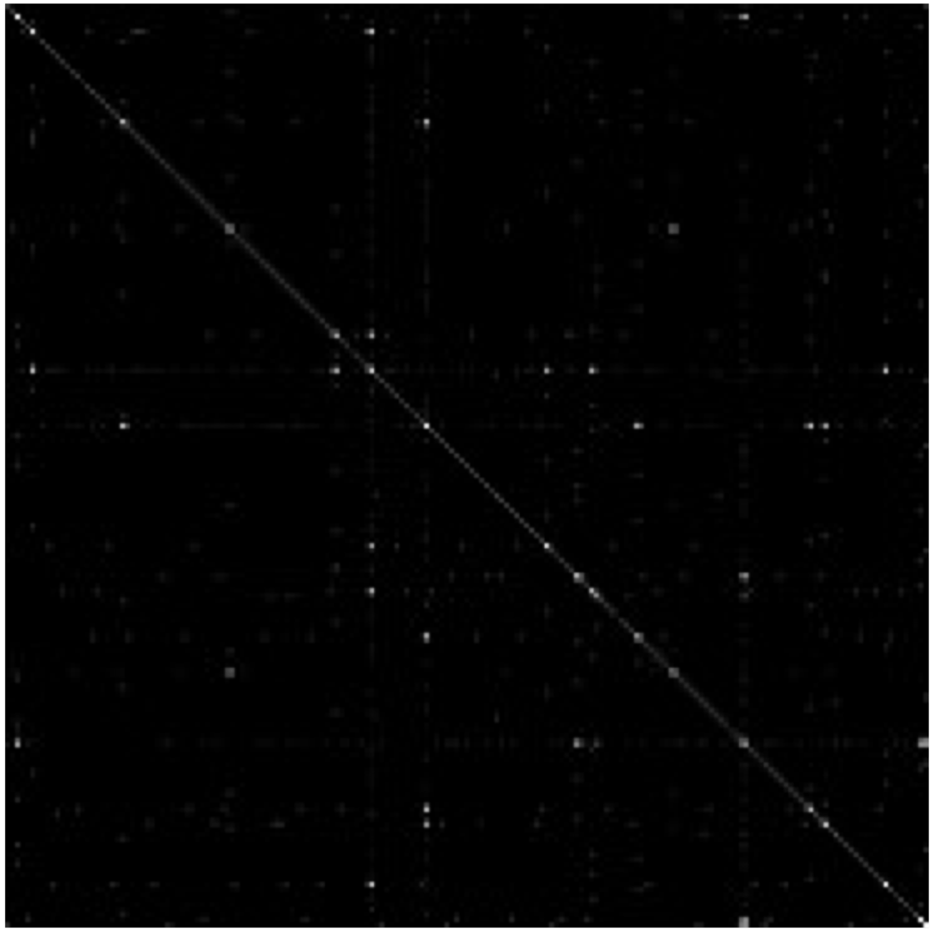
# Bounding Values in $AA^T$

$AA^T$ *Small number of entries take bulk of the values*

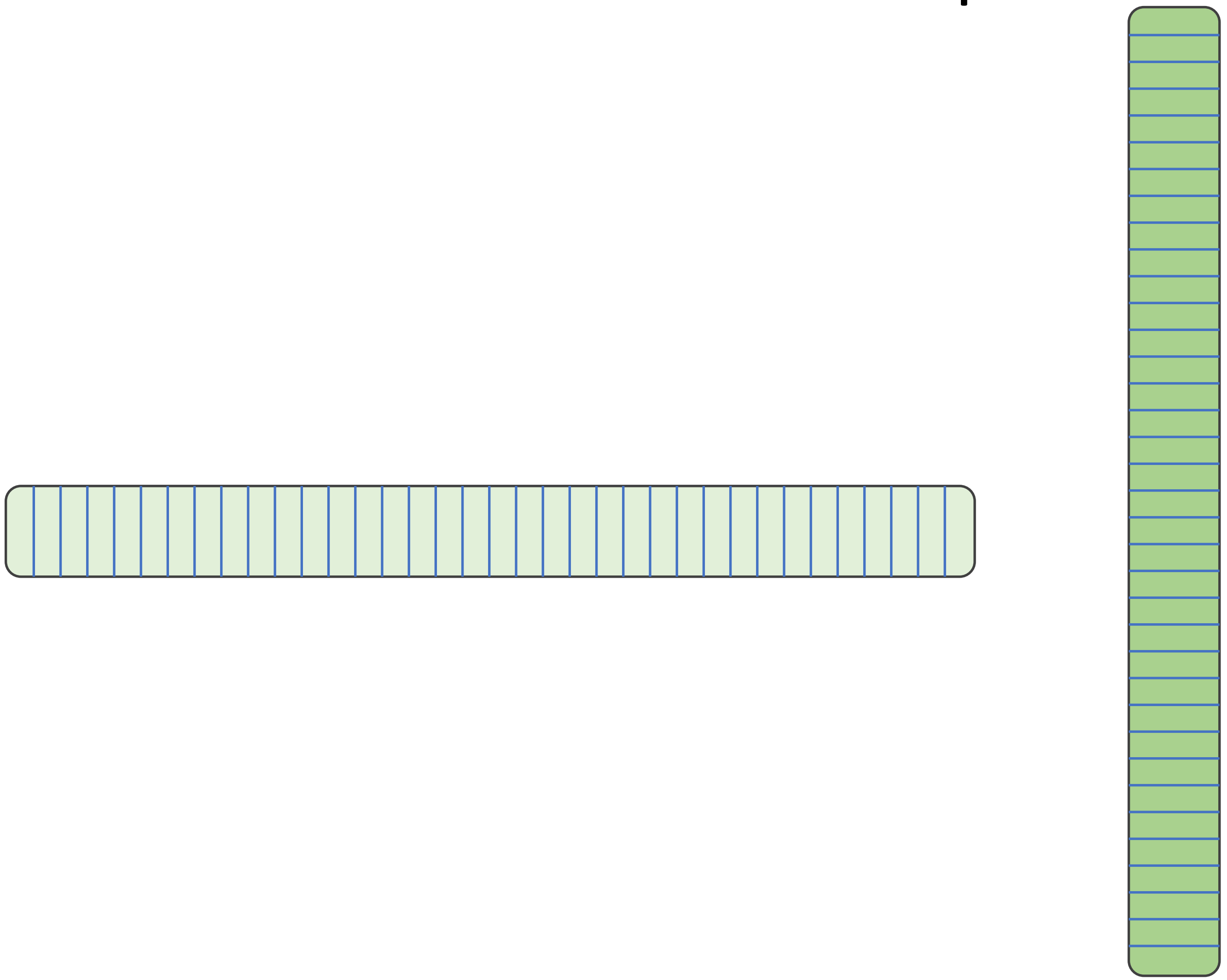
*Threshold based on the diagonal values*



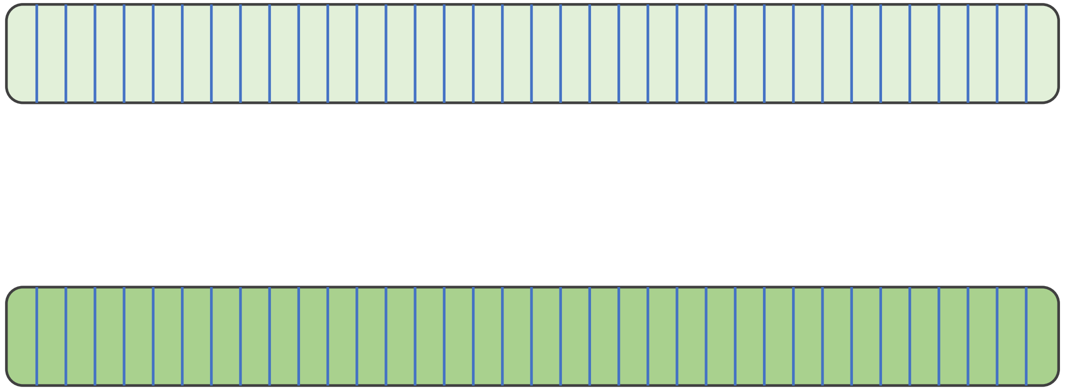
# Direct Approx – Threshold Based



# Matrix Multiplication



# Matrix Multiplication



# Set Similarity Joins

# Set Similarity

- Used - data cleaning, deduplication, product recommendation

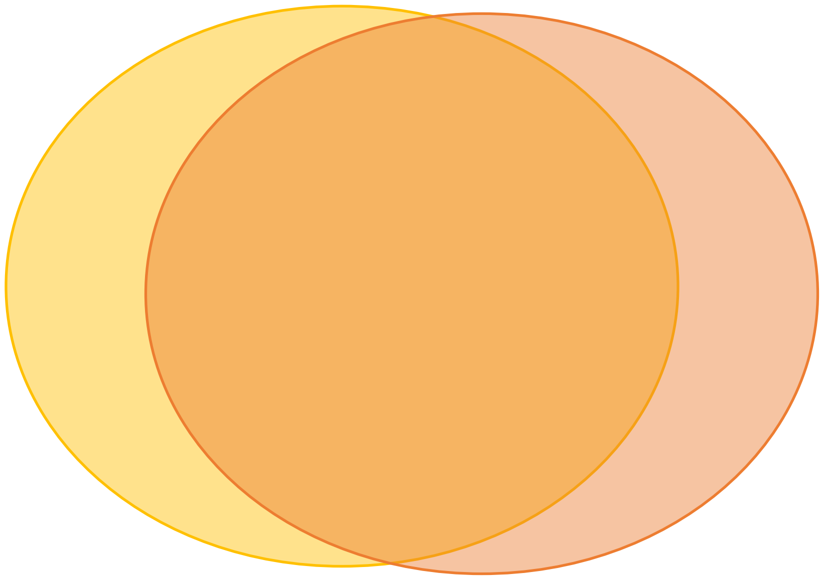
- Identify tuples, which are ‘close enough’, on multiple attributes

Designed Algorithm **SIM**

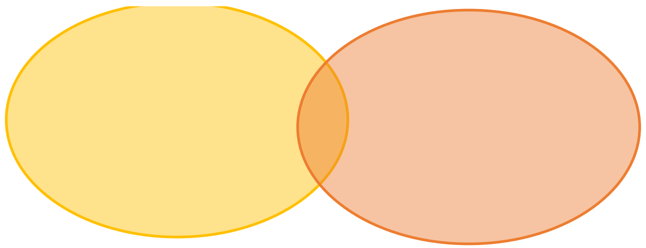
# Threshold Based – Set Similarity Join

- Surajit Chaudhuri et.al.

- If intersection of two sets are large
  - Intersection of small subsets of them are non-zero



$$h - \tau + 1$$



DBXLab
UT ARLINGTON

# Sketch Based - Set Similarity Join

- Uses Min-hashing
  - Use a random ordering of all items in universe
  - Min-hash = element with the minimum hash value
  - Jaccard Similarity of two sets A and B, J(A, B) = $\frac{|A \cap B|}{|A \cup B|}$

$$P(h[A] = h[B]) = J(A, B)$$

# Sketch Based - Set Similarity Join

- Bottom-k sketch
  - Uses only first k elements of the hash
- Works well for large size sets

# Algorithm SIM

**if** $|U_i| \geq \log(m)$ **and** $|U_j| \geq \log(m)$ **then**

apply bottom-$k$ sketch based estimation

$$E[\cap_{i,j}] = \frac{k_{\cap}(i,j)}{k} \frac{m(k-1)}{h_{i,j}[k]}$$

**else**

apply threshold-based estimation

Motivation

Problem Formulation

Contribution

Algorithms

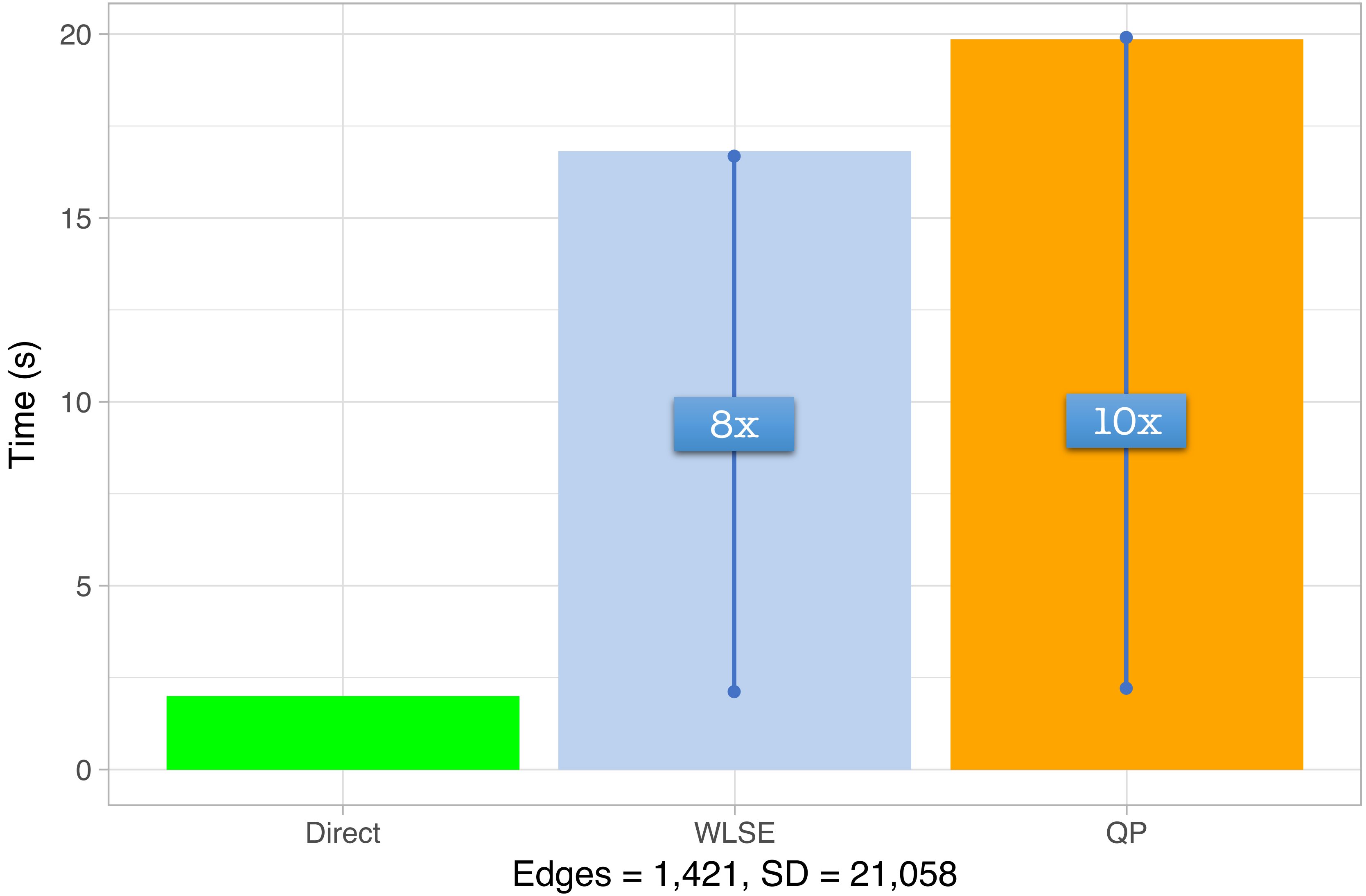
Experiments & Evaluation

# Experiments & Evaluation

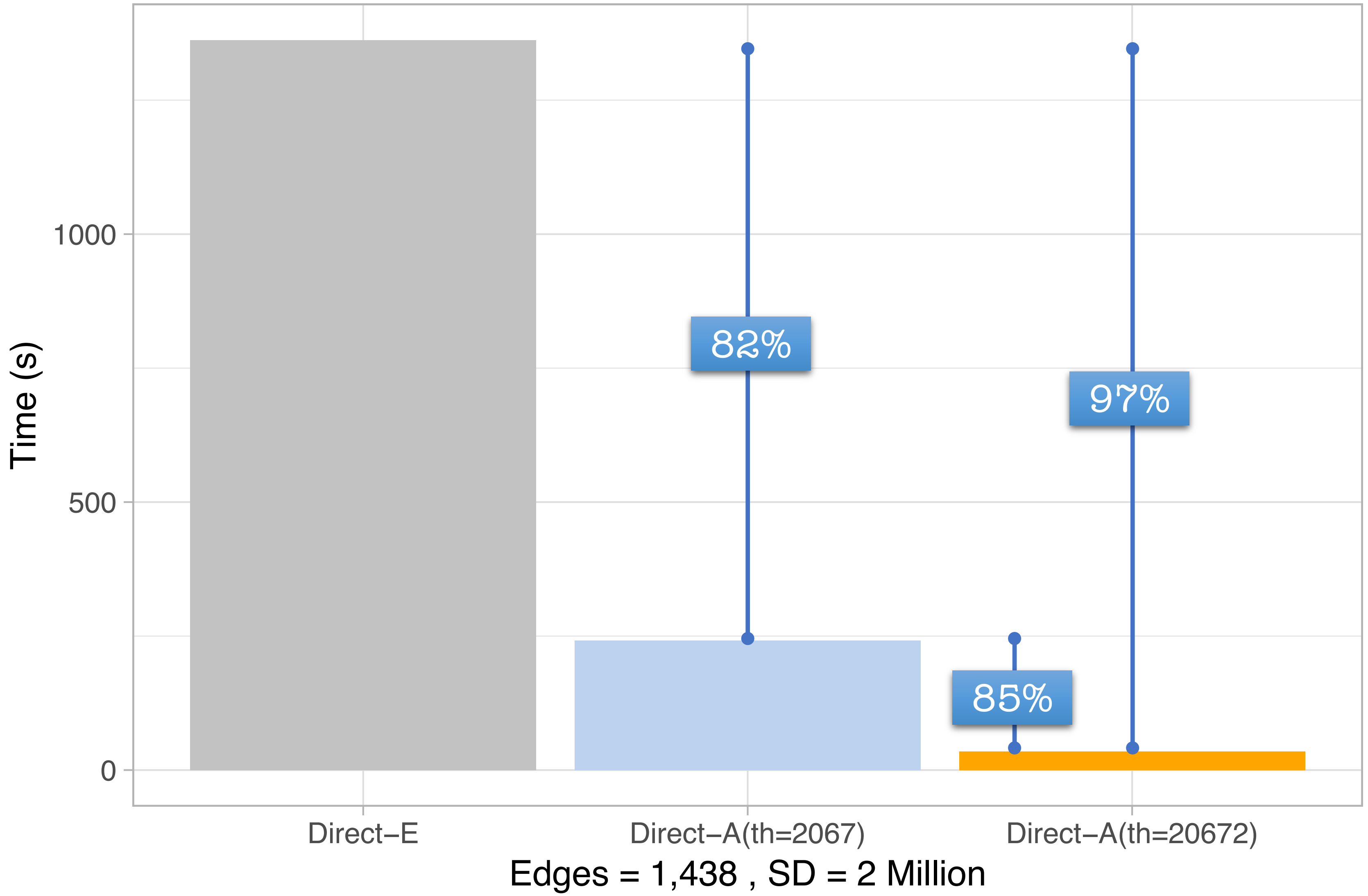
# Evaluation Setup

- Implementation: Matlab & Python2.7
- Synthetic Datasets: constructed as a random, Erdos-Renyi graph(Networkx)
- P2P dataset from SANP dataset of Stanford
  - 10786 Nodes & 39994 Edges

# Direct VS Baselines



# Direct-Exact VS Direct-Approximate



# Direct-Exact VS Direct-Approximate