DBXLab

UT ARLINGTON

Leveraging Similarity Joins for Signal Reconstruction

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

Microsoft Research

PENNSTATE

UNIVERSITY OF MICHIGAN

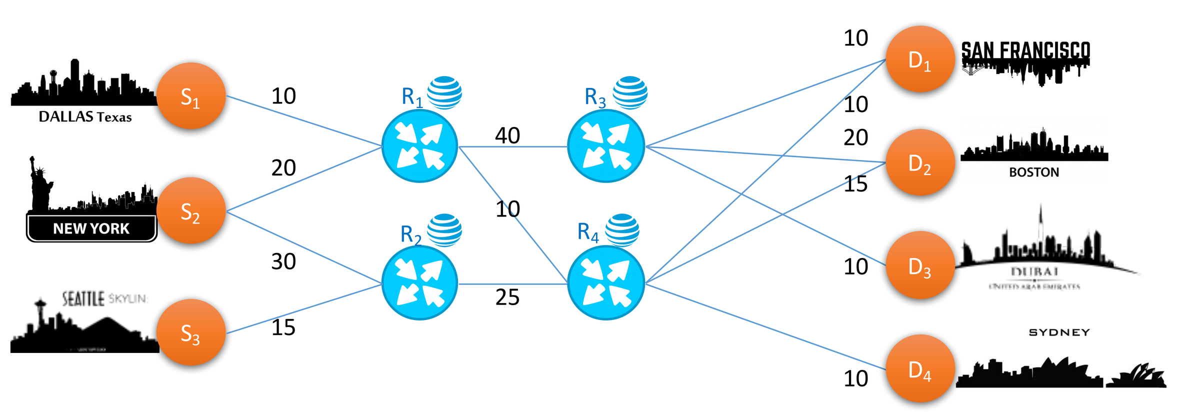
QCRI Qatar Computing Research Institute

HAMAD BIN KHALIFA UNIVERSITY

INTRODUCTION

Traffic Flow : Consider an IP network where sources and destinations are sending traffic to each other, while network admin is aware of link loads, and the routing table. Our objective is to find amounts of flow between source-destination pairs. As the number of source-destination pairs are larger than the number of edges, the system of equations is under determined and does not have a unique solution.

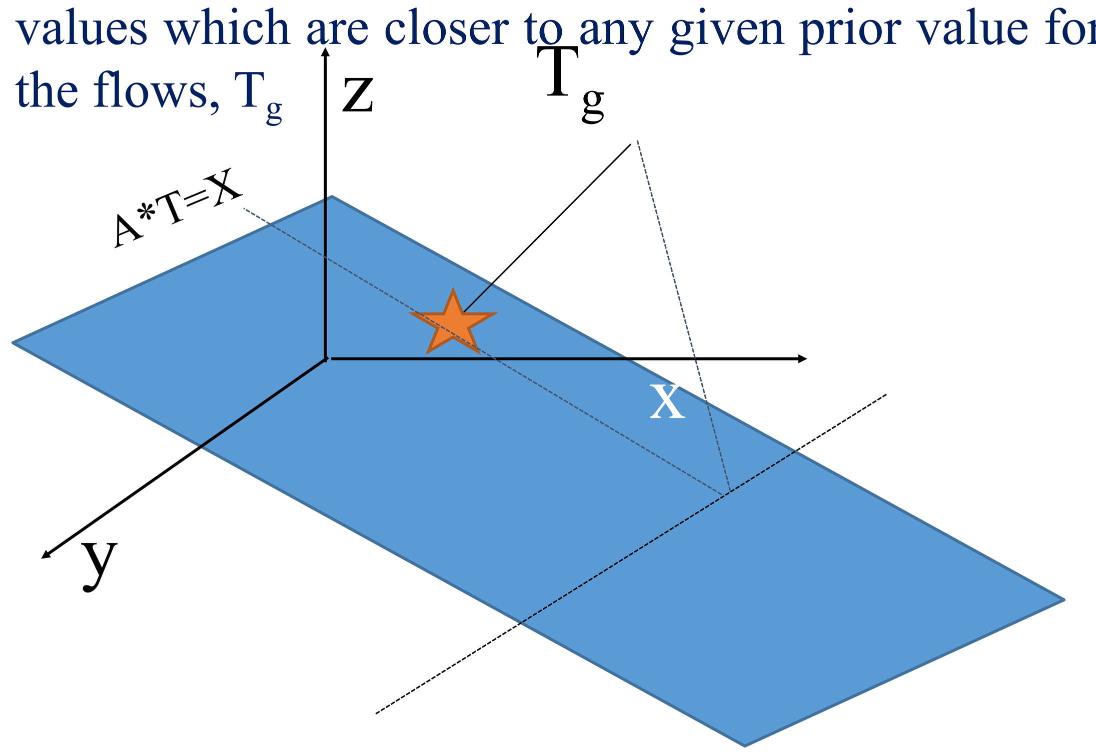
For the traffic matrix computation problem, Roughan et.al. proposed a prior that flows follow in practice. Among all possible outcomes, our objective is to find the one that minimizes the distance to it prior.



3D Image Reconstruction : 3D image reconstruction has a wide range of applications in different fields such as medical imaging. Given a set of 2D projections of a 3D image, the objective is to reconstruct of the same 3D image. Many 3D images may produce the same projections. Prior knowledge can alleviate hardness of this problem. We may use a 3D image template and, among all possible 3D reconstructions from 2D images, find the one that is closest to the template.

OBJECTIVE

- Given any traffic routing matrix, and link level flow information, how to effectively infer the flow values which are closer to any given prior value for the flows, T_g



PROBLEM FORMULATION

The routing matrix A is represented as follows:

- A binary matrix where each column represents a source-destination pair

$$A = \begin{bmatrix} . & s_1d_1 & s_1d_2 & \dots & s_2d_1 & \dots & s_nd_n \\ e_1 & 1 & 0 & \dots & 1 & \dots & 1 \\ e_i & \dots & \dots & \dots & \dots & \dots & \dots \\ e_l & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Each column is an edge in the network or a 2D image in a 3D image reconstruction problem.
- '1' in a row indicates the usage of an edge in a particular source-destination routing
- 'b' matrix represents the traffic accumulated in each edge (represented as w_{ei})
- 'x' is the actual flow information about the source-destination pairs. Our aim is to find the 'x' as close as possible to a given prior 'T_g'

$$b = \begin{bmatrix} W_{e1} \\ W_{e2} \\ \vdots \\ W_{eL} \end{bmatrix}_{L \times 1} \quad x = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{nm} \end{bmatrix}_{nm \times 1} \quad T_g = \begin{bmatrix} T_{ge1} \\ T_{ge2} \\ \vdots \\ T_{geL} \end{bmatrix}_{L \times 1}$$

CHALLENGES

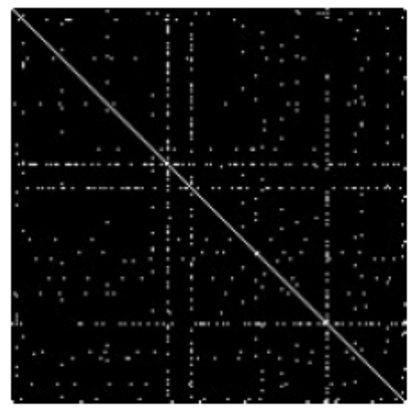
- Routing matrix is binary and sparse
- Routing matrix size is proportional to number of flows (which are in millions)
- Number of variables (flows) is much higher than number of equations (edges)
- We have infinite number of solutions
- Calculating A^{-1} is computationally not feasible

SOLUTION

- We formulate the problem as a Quadratic Programming problem
- Derive its Lagrangian dual form
- Propose an exact algorithm DIRECT to solve the dual problem
- Optimize DIRECT to propose a threshold based approximate algorithm
- Hybrid algorithm that combines database techniques to compute set similarity

- DIRECT: Solving for the Lagrangian Dual form yields $X = X' - A^T(AA^T)^{-1}(AX' - b)$
- Finding $(AA^T)^{-1}$ is expensive and is computed by Gaussian elimination of $(AA^T)\xi = AX' - b$

OPTIMIZATIONS



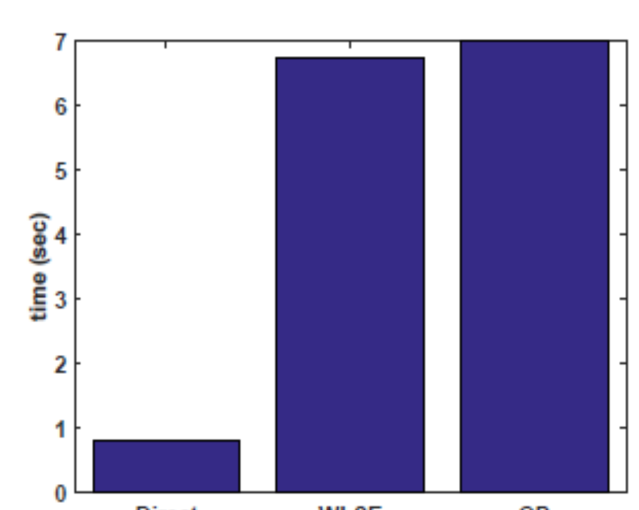
- (AA^T) is sparse as you see in Figure above
- Each cell is bounded by the value corresponding its diagonal
- Parametrized threshold and set similarity techniques yield faster but approximate results

DATA AND EXPERIMENTS

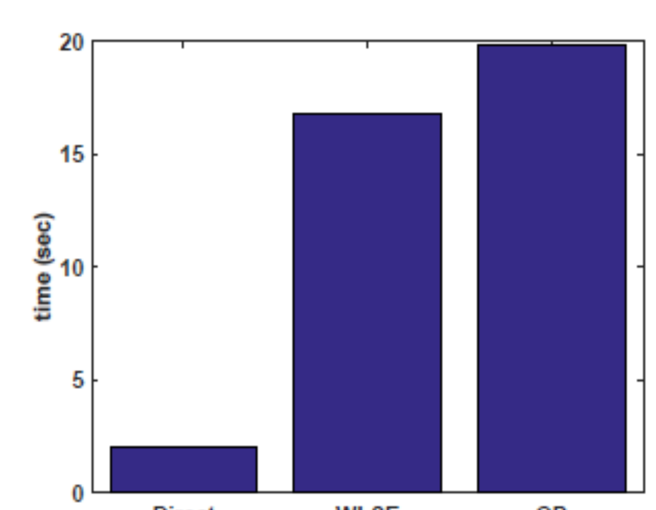
- P2P dataset from 'Gnutella' network from August 2002
- Traffic was generated using a Pareto distribution with alpha = 1 and minimum = 20
- '*networkx*' in python was used to form the network and the links
- '*gravity*' was assumed to be the prior
- '*Direct*', '*Threshold*', and '*Set Similarity techniques*' were implemented in MATLAB

RESULTS AND DISCUSSION

- Graph presents the speedup obtained using '*Direct*'
- Improvements in time is in orders of magnitude



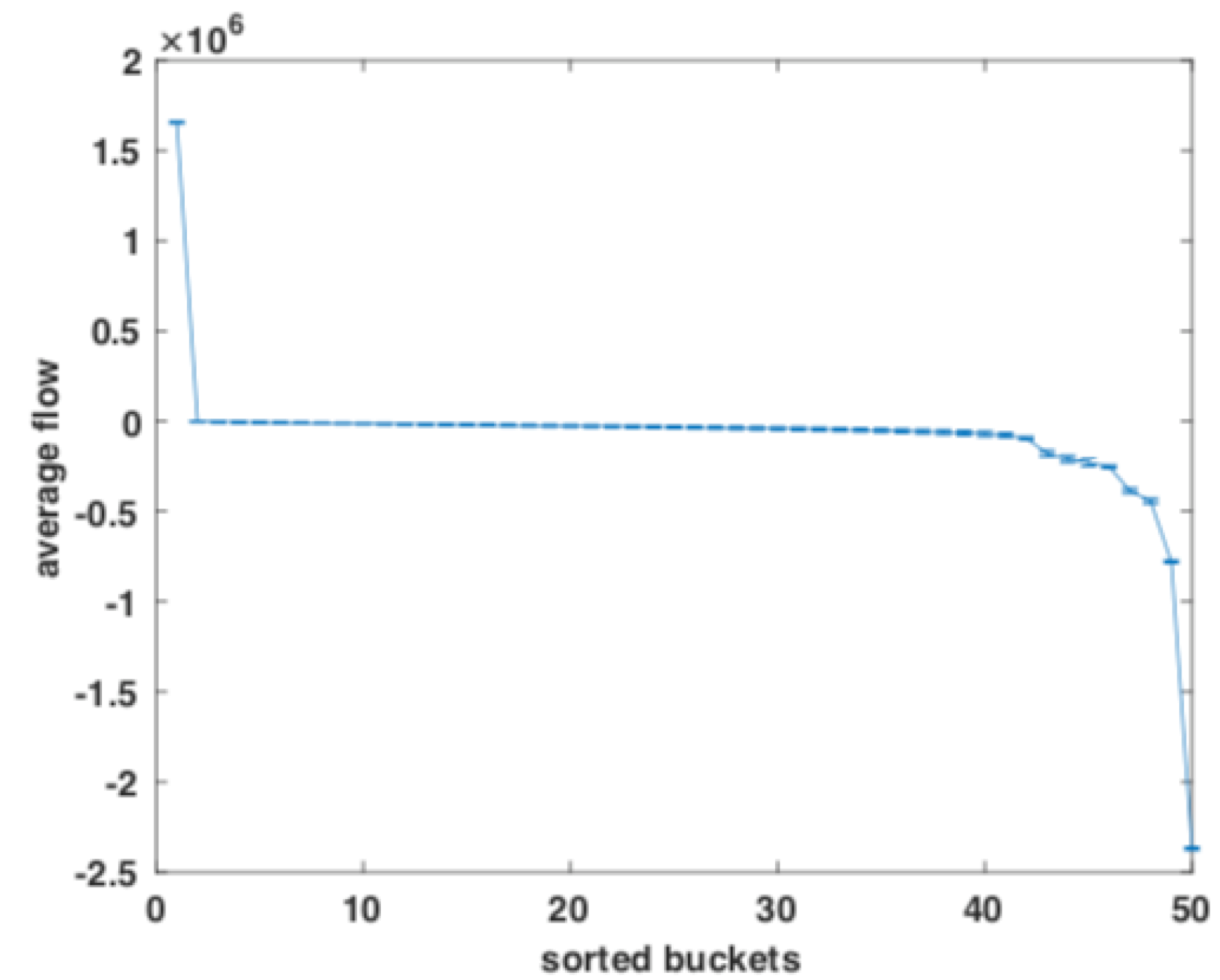
DIRECT v.s. baselines (1,278 x 9,330)



DIRECT v.s. baselines (1,421 x 21,058)

CONCLUSION

- Orders of magnitude improvement in time
- Newer quadratic programming solution for binary linear equations
- Approximate pseudo-inverse algorithm was obtained
- Faster signal reconstruction algorithm
- Faster algorithms in 3D image reconstruction
- Threshold based algorithms for faster approximate



Absolute Error of the DIRECT–A (τ = 3737) in p2p

ACKNOWLEDGEMENT

- We thank AT&T for the benevolent financial support and insightful discussions with Dr. Divesh Srivastava.