A COMPRESSED SENSING APPROACH TO TAMING THE INTERNET MEASUREMENT DATA DELUGE

Matt Hall*, Joel Sommers^, Ramakrishnan Durairajan*
(*) University of Oregon, (^) Colgate University

PROBLEM: THE DATA DELUGE

Internet Measurement Campaigns generate tremendous amounts of data

Content Distribution Networks
- Performance Data

Security Firms
- Internet routing anomalies

Compounding Factors
- Longitudinal Analysis
- Multiple vantage points

EXISTING SOLUTIONS

Data-agnostic approaches use off-the-shelf compression to store Internet measurement data
- tar, gz, bz2, xz, zip

These compression algorithms are used to store multiple classes of Internet measurement data
- Round Trip Time (RTT), timestamp, IP addresses, etc.

OUR APPROACH

Compressed Sensing (CS) can be used to compress Internet measurement data

CS is used in other areas of networking research
- Internet Tomography, Sensor Networks

CS works best for sparse data with a high degree of repetition and occasional outliers

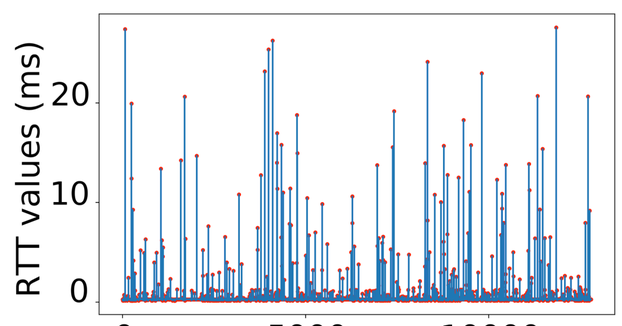
There are two key operations to implement CS
- Sampling input data
- Reconstructing the original trace

SAMPLING

Given a trace, x, of RTT values, select and store a subset, x' (where $x' \ll x$) of the values

How to choose x'?
- Peak finding algorithm finds the 'highs' and 'lows'

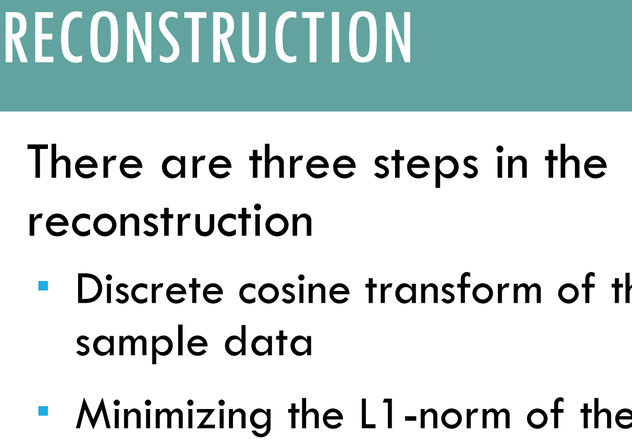
A trace of 12.8k RTT values from CAIDA's Ark dataset [2] are shown in blue. Red dots indicate sampled values

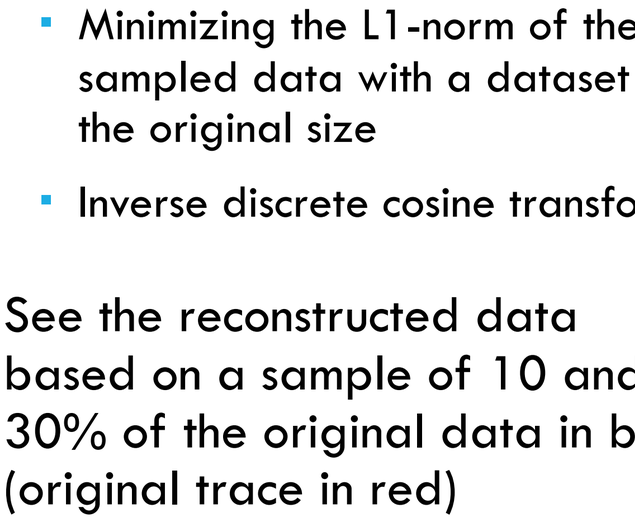


RECONSTRUCTION

There are three steps in the reconstruction
- Discrete cosine transform of the sample data
- Minimizing the L1-norm of the sampled data with a dataset of the original size
- Inverse discrete cosine transform

See the reconstructed data based on a sample of 10 and 30% of the original data in blue (original trace in red)





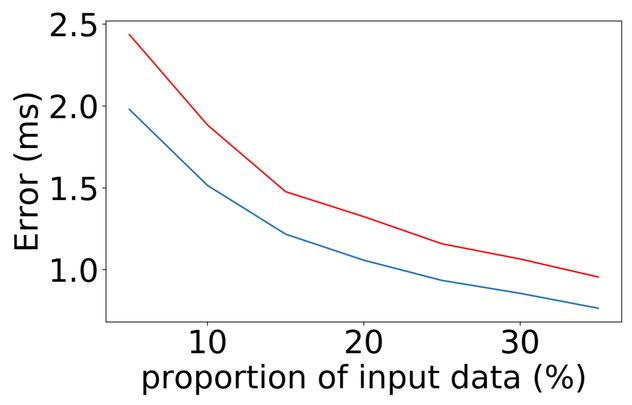
RESULTS

Results with the Ark dataset show that CS outperforms all off-the-shelf compression algorithms except xz with only 20% of the data

At 15% sampling, CS beats xz while keeping the mean absolute error (MAE) below 1.3 ms

Method	Size (kB)	MAE (ms)	RMSE (ms)
original	153	0	0
tar	164	0	0
gz	46	0	0
zip	46	0	0
bz2	38	0	0
CS (20%)	29	1.04	1.30
xz	27	0	0
CS (15%)	22	1.22	1.48

The figure below shows the change in the Root Mean Square Error (red) and Mean Absolute Error (blue) at different sampling percentages



These results show promise for the opportunity to apply CS to RTT data with varying degrees of accuracy based on domain applications

A larger scale study will be necessary to make guarantees about its usefulness

REFERENCES

[1] Donoho, David L. "Compressed sensing." *IEEE Transactions on information theory* 52.4 (2006): 1289-1306.
[2] http://www.caida.org/projects/ark