XDATA SUMMER CAMP

Embedding Methodology and Statistics for Inference

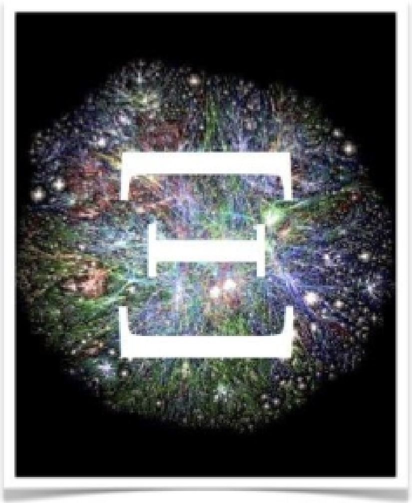
Sancar Adali

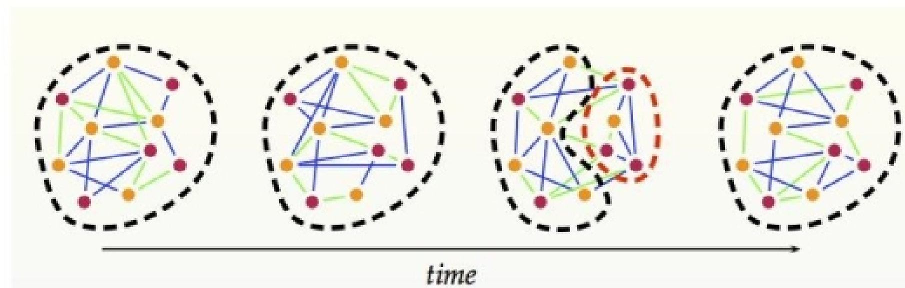
Johns Hopkins University

July 29, 2013

FROM DATA TO STATISTICS

h:





CAPABILITIES AND IMPLEMENTATIONS

- Compute statistics from time series of graphs (TSG)
- Out-of-sample extension for adjacency spectral embedding
- Faster Embedding by the use of OOS-embedding
- Dissimilarity computation for multivariate time series
- Tensor Decomposition for time series data (adj. matrices, multivariate data)
- Fast computation of local statistics in very large graphs

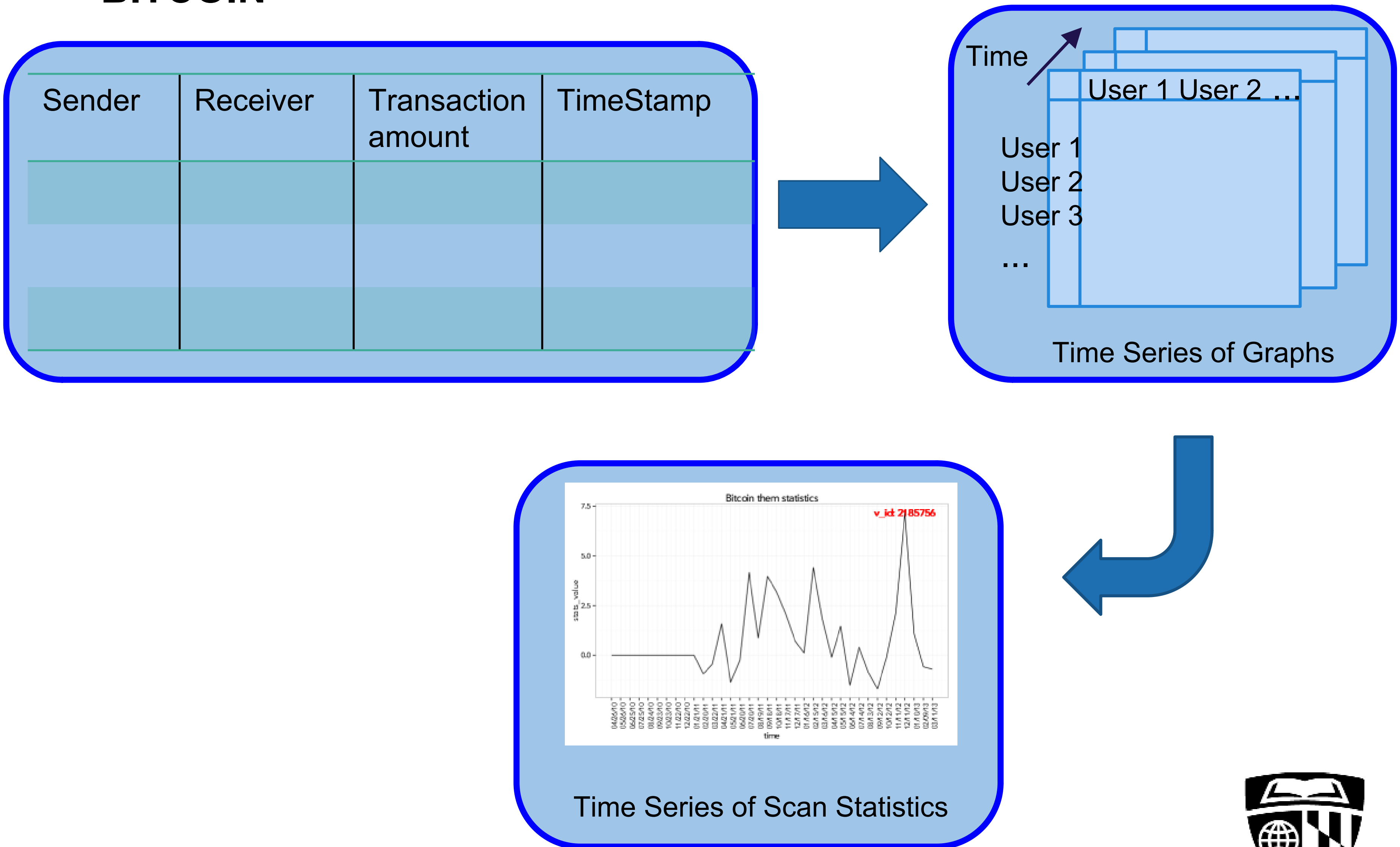
CAPABILITIES AND IMPLEMENTATIONS

Software

- R packages: ScanStats, AdjMatEmbed, DissTimeSeries
- Python: Large-graph invariants , MySQL-igraph for TSG
- igraph C/C++ library (devel branch)

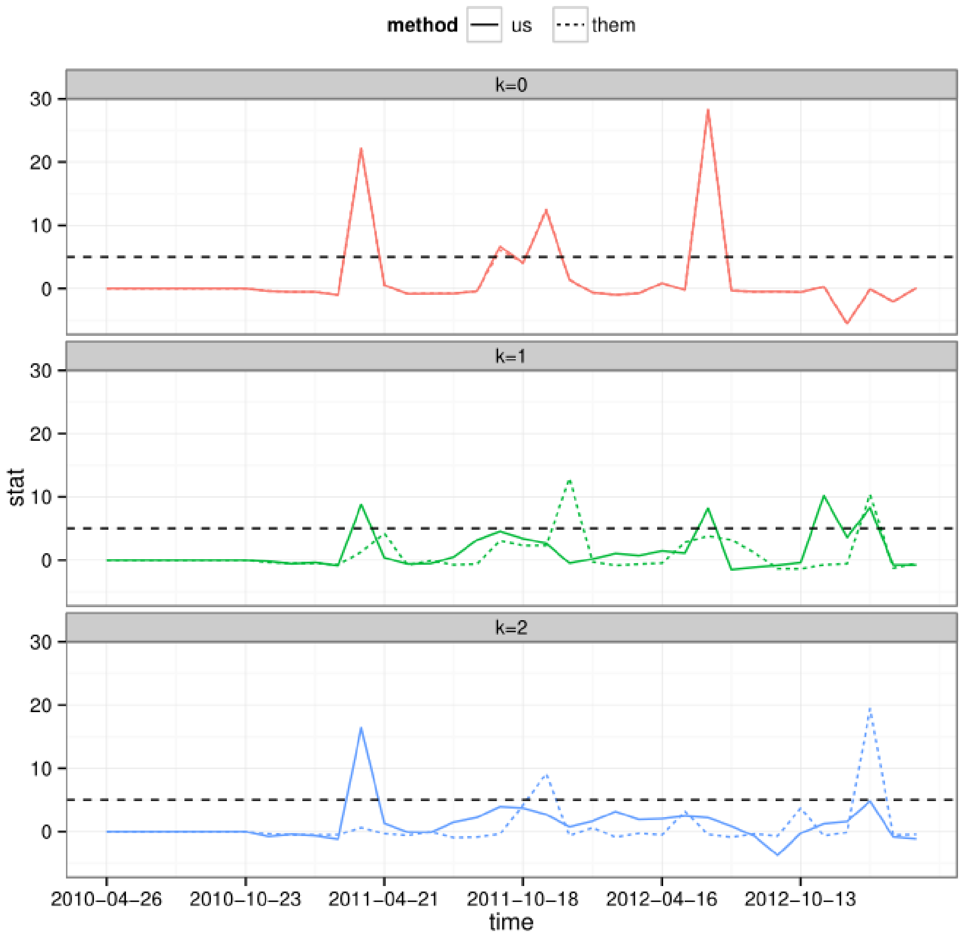
BITCOIN DATA ANALYSIS

BITCOIN



Sancar Adali

BITCOIN



- Various anomaly detections using the normalized statistics.
- The vertices which are the sources of anomalous activity should be investigated further.

Kiva

- Joint embedding of all entities (lender, loan, partner, borrower)
- Relationship between entities of different kinds

-> Adjacency matrix of graph (entities -> vertices)

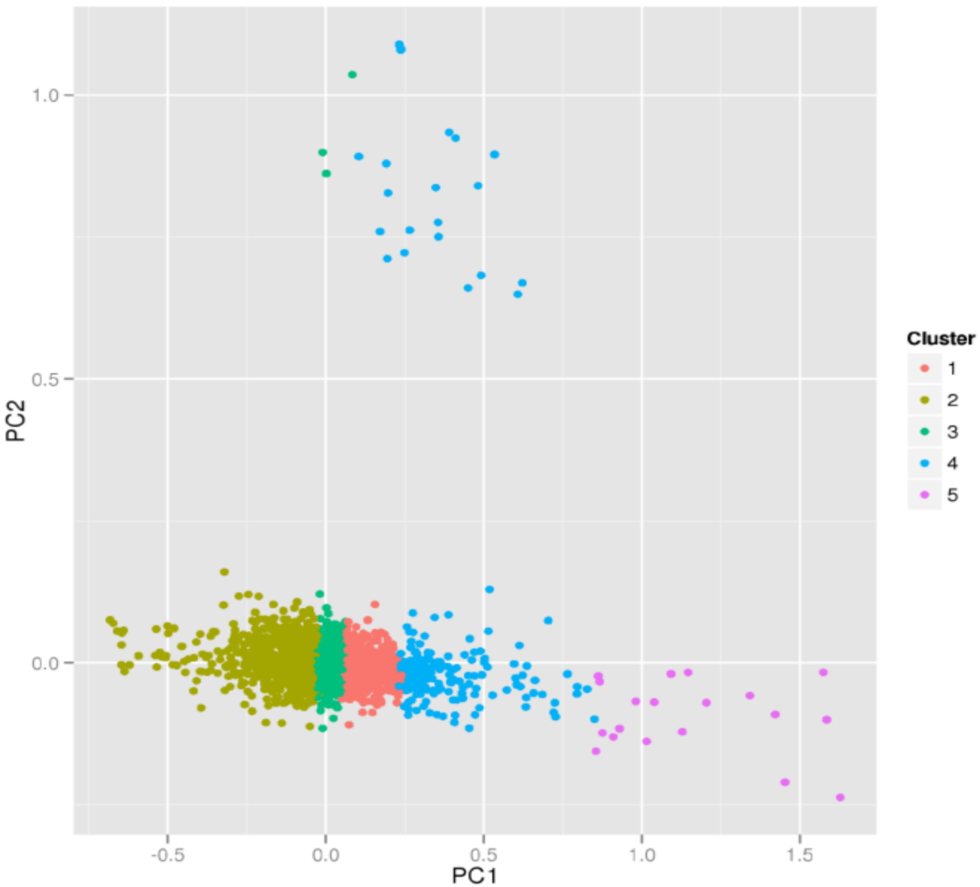
- Lender-lender graph: edges

EMBEDDING APPROACH

Fast Embedding via out-of-sample extension

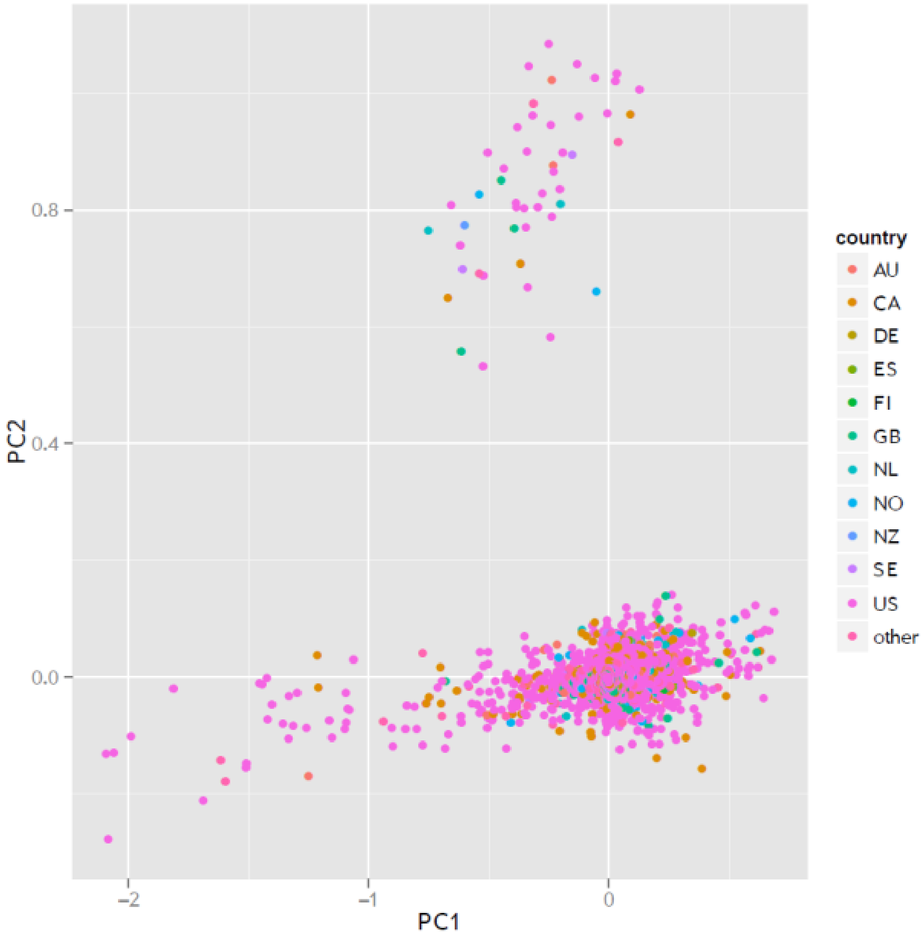
- Embed the most active lenders in-sample
- Repeat until all entities have been embedded:
  - Embed a batch from the remaining entities via out-of-sample extension
- Cluster the embedded entities

Kiva Entity Embedding



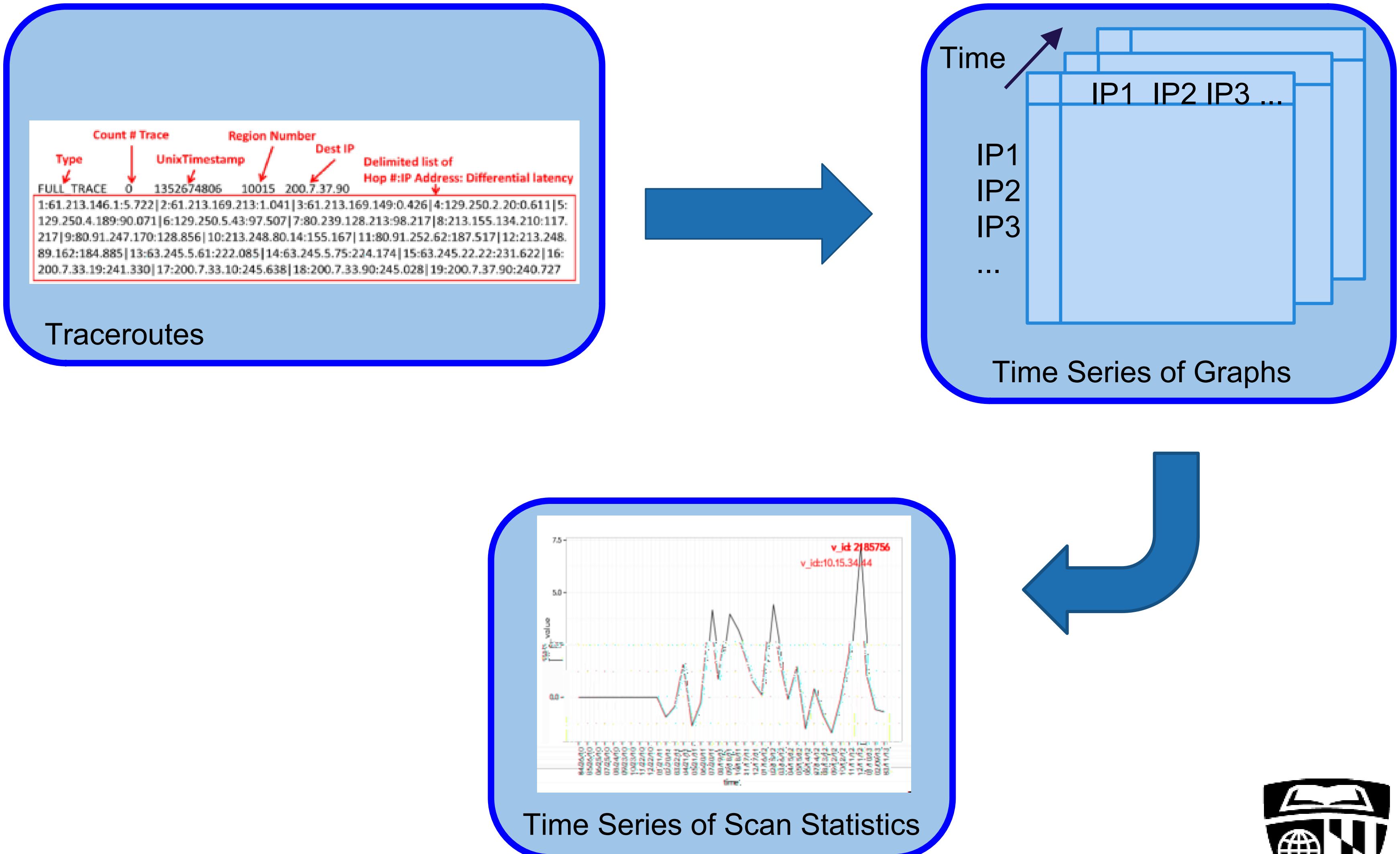
- Embedding of 720K Kiva lenders
- Other entity types will be OOS embedded

Kiva Entity Embedding



- Embedding of 720K Kiva lenders
- Other entity types will be OOS embedded

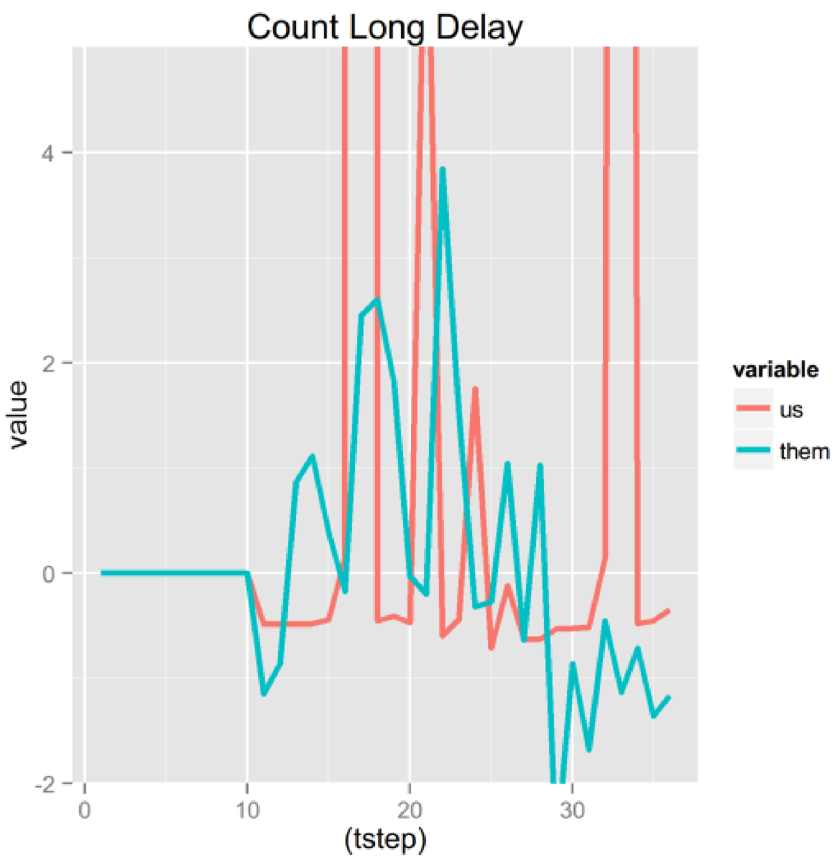
Scan Statistics for Anomaly Detection

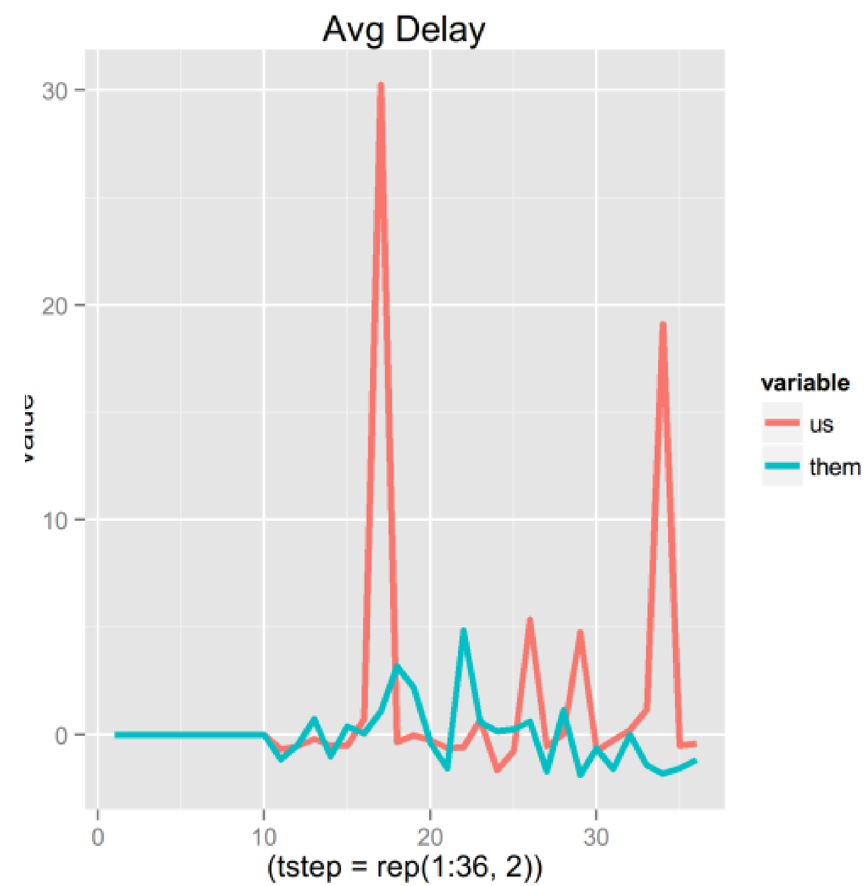


Sancar Adali

AKAMAI-TRACEROUTE

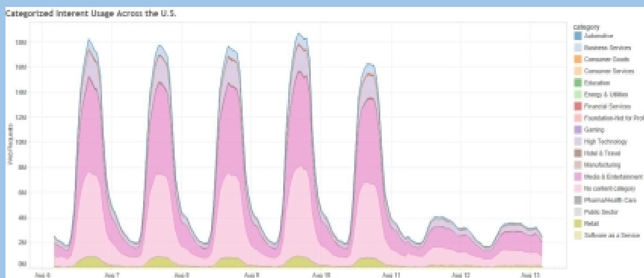
Scan Statistics for Anomaly Detection





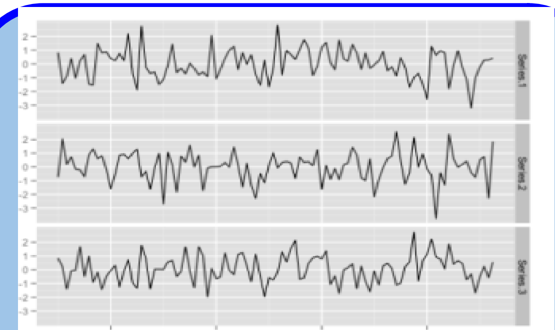
Sancar Adali

Embedding dissimilarities between CIDR Traffic

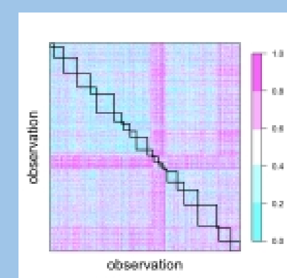


CIDR Hourly Traffic for each category

Aggregate by summing the traffic for each week



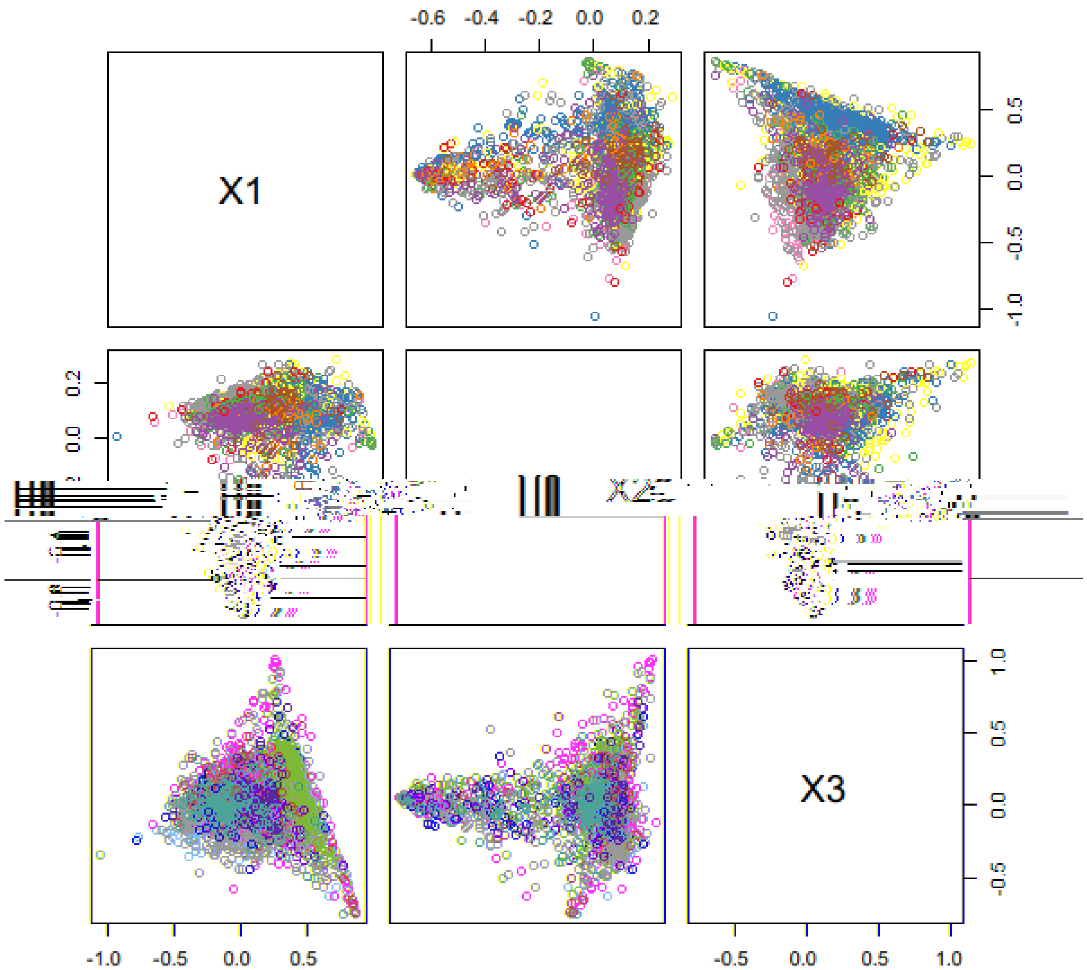
Multivariate Time Series



Dissimilarity Representation

Embedding via MDS

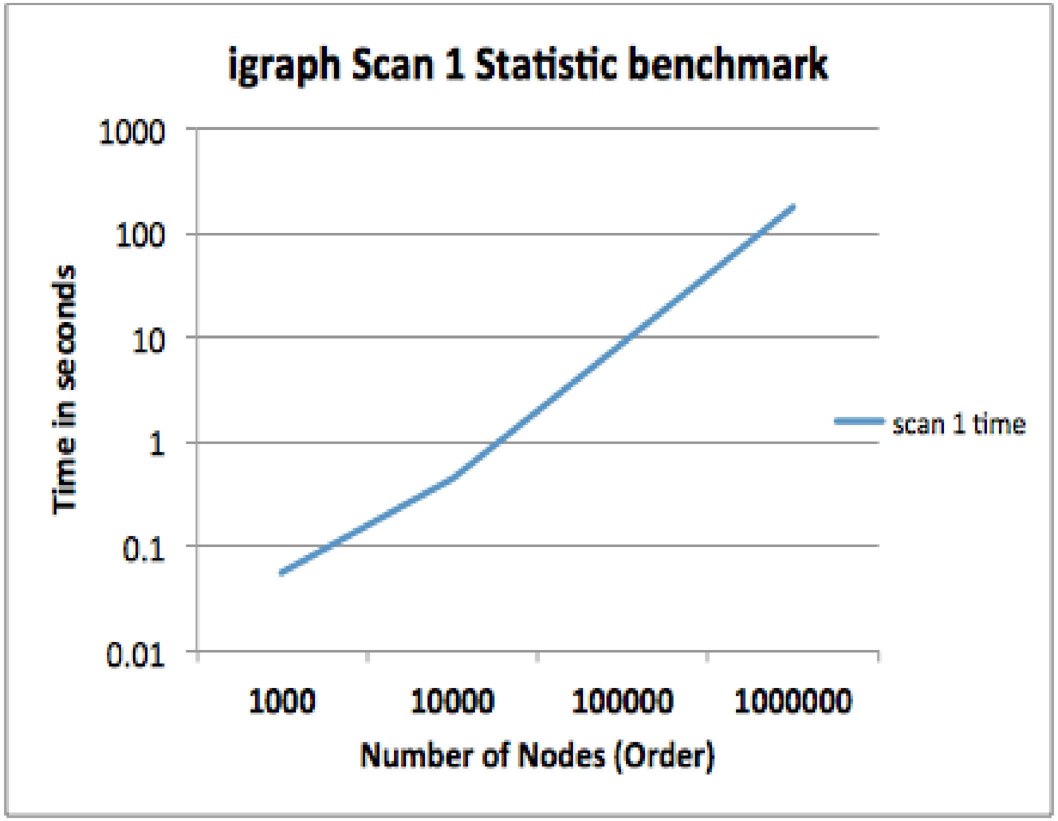
AKAMAI-CIDR



- CIDRs based on China (blue) show a clustering pattern

3D Plot of CIDR Embeding

Scan 1 Statistic & Spectral Embedding



igraph 0.7 introduces:

- Fast implementation of Scan 1 Statistic exact and approximate invariant
- Fast spectral embedding of adjacency matrices using ARPACK

COLLABORATORS

PNNL/Stanford/Purdue : Ryan Hafen (Akamai-Traceroute, Akamai-CIDR)

BBN/Raytheon: Walter Andrews (Kiva)

Oculus: Peter Schrettlen (Bitcoin)

Thanks to Peter Wang (Continuum) and Ryan Hafen (PNNL) for providing derived data

Thanks to everybody in DARPA XDATA program for supporting this work.

Thank you.

Contact information:

Sancar Adali

sadali1@jhu.edu
Johns Hopkins University
3400 N. Charles St.
100 Whitehead Hall
Baltimore, MD 21218