RESEARCH \ ENGAGEMENT \ INNOVATION

The DataBridge: A Social Network for Long Tail Science Data

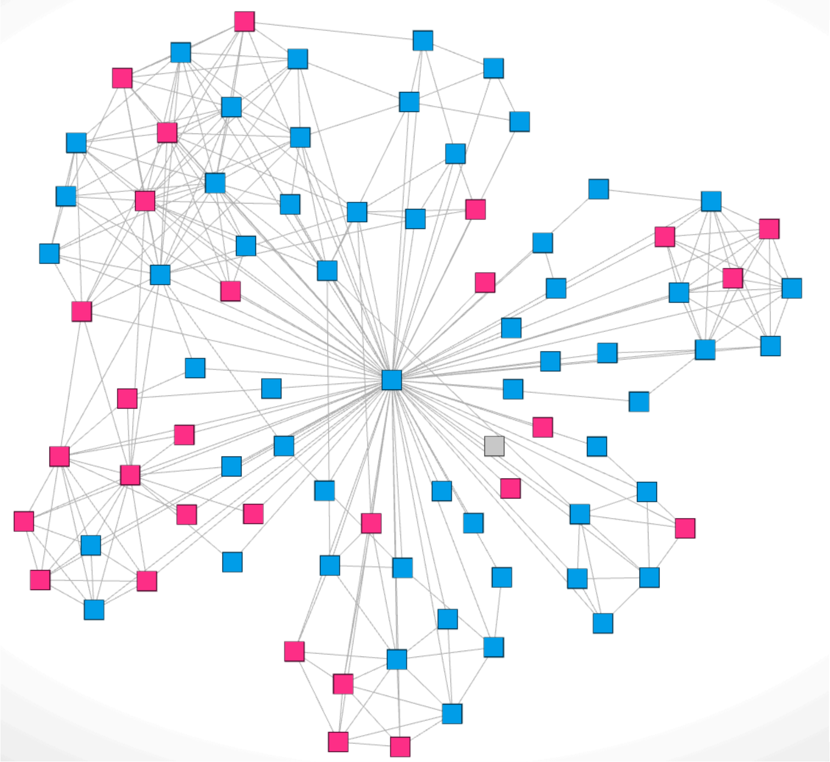
Howard Lander
howard@renci.org
Renaissance Computing Institute
The University of North Carolina at Chapel Hill

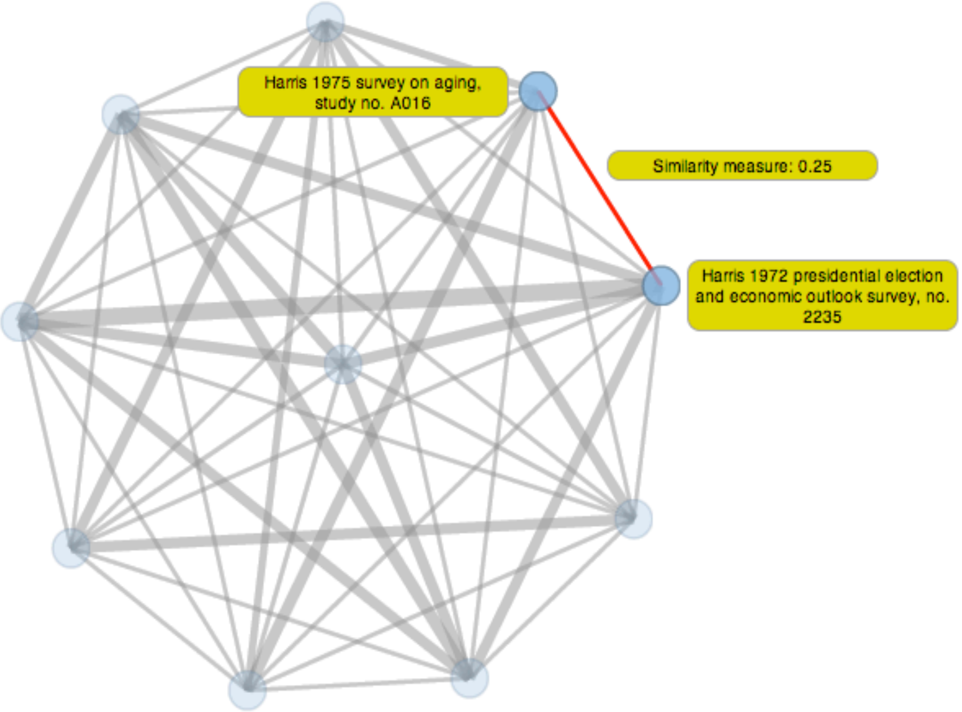
DATABRIDGE

Outline of This Talk

- The DataBridge
- DataBridge Vision
- Long Tail Problem
- DataBridge Strategy
- Progress to Date

The DataBridge: A Social Network for Data




Harris 1975 survey on aging, study no. A016
Similarity measure: 0.25
Harris 1972 presidential election and economic outlook survey, no. 2235

Dark Data from The Long Tail of Science

- Long tail data is the small data sets produced by numerous investigators
- From Brahe to Mendel discovery has come from relatively small data sets
- Much long tail data is dark data, data "not easily found by potential users" (Heidorn)

The DataBridge Vision: Shining a light on dark data

- Maximize the usefulness of long tail data for scientific research
- Facilitate searching for collaborators
- Enable data set publication as a means of communication
- Assist scientists in discovering “interesting” data sets by automatically forming communities of data

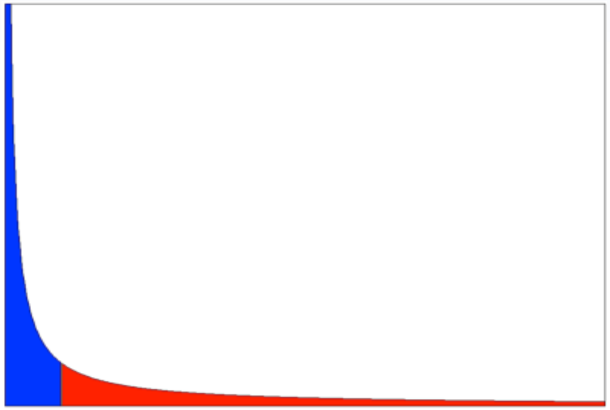
The Big Data: CERN, NASA, NOAA

These peta-scale data sets have some advantages

- Well defined automatically generated metadata
- Homogeneous formats
- Professional staff
- Known location
- Limited number of producers
- A large research agenda

Great potential for science

The Motivating Problem: Long Tail Data



NSF in 2011

- 11,150 awards
- Median grant: $126,500
- Power law distribution
- Thousands of researchers
  - Primarily single investigator projects
- Data individually not petascale but **large in aggregate**

The Motivating Problem

Long tail data does not have

- Well defined automatically generated metadata
- Homogeneous formats
- Professional staff
- Known location
- Limited number of producers

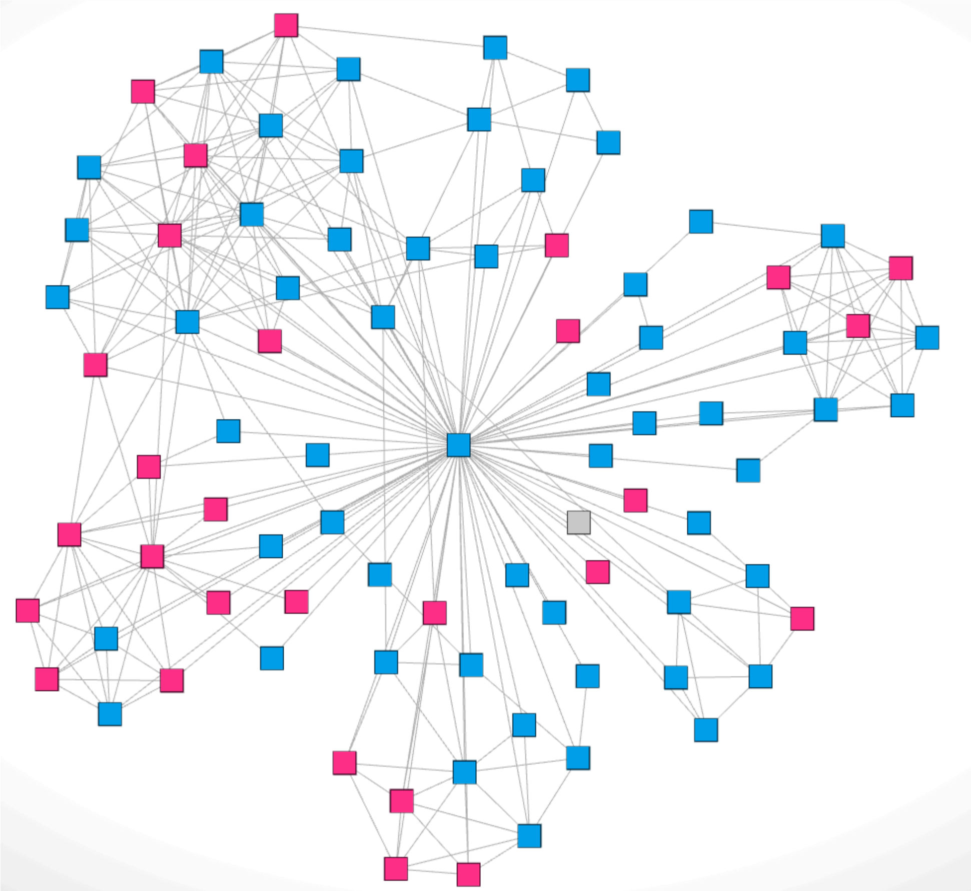
Enormous amounts of **atomistic** and **unconnected** dark data!

The DataBridge Strategy: Building a Social Network for Scientific Data

- Construct a multi-dimensional sociometric network for data. Three challenges:
  - Evaluate the similarity/relevancy of data sets
  - Perform community detection on the resulting set of similarities
  - Provide query interfaces on resulting multi-dimensional network

The DataBridge Strategy: Simple Social Network Example

- Basic similarity: people who are Facebook friends with me
- Not a lot of additional information



renci NSF DATABRIDGE

The DataBridge Strategy: Community Detection

- Doc Watson at Amazon.com
- Clusters represent related items
- Clusters are connected somehow

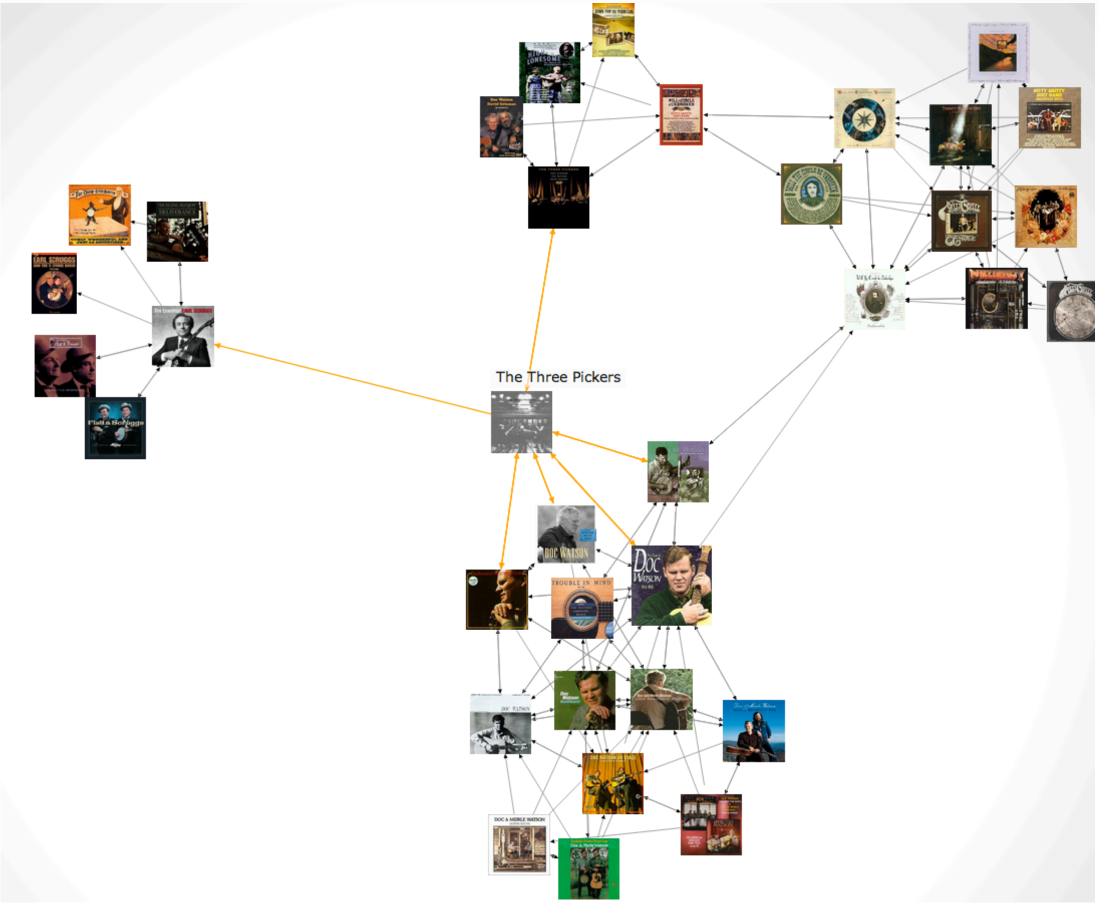
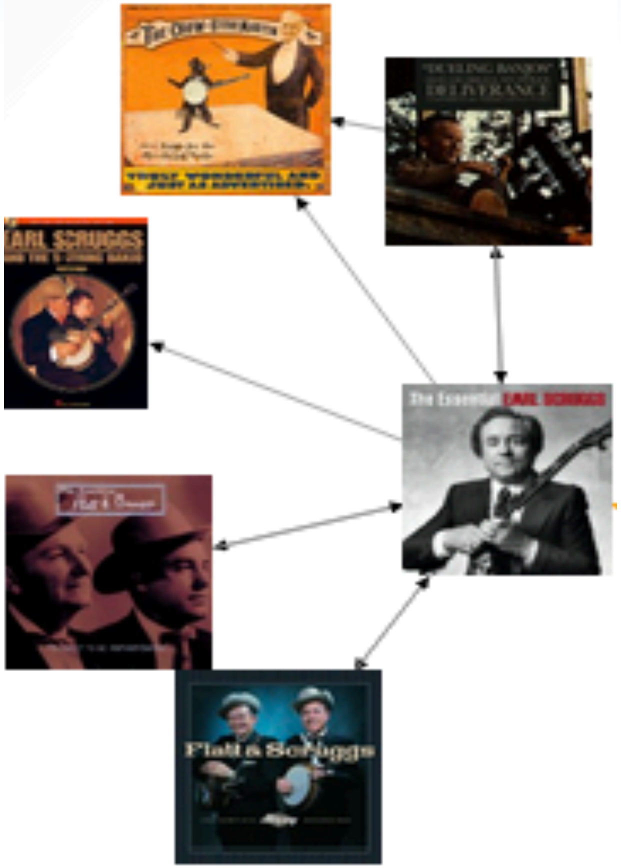
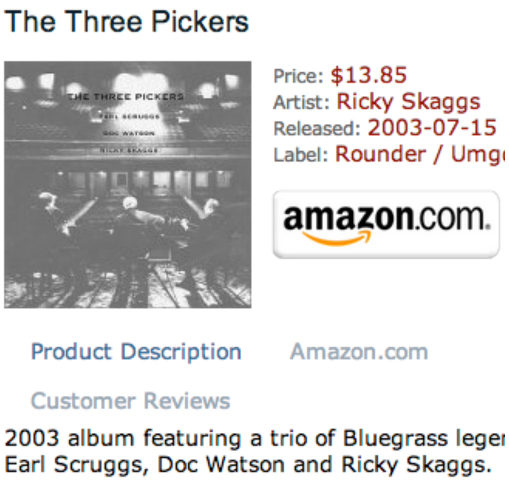


Image from yasiv.com

The DataBridge Strategy: Community Detection



With Earl Scruggs



The Three Pickers

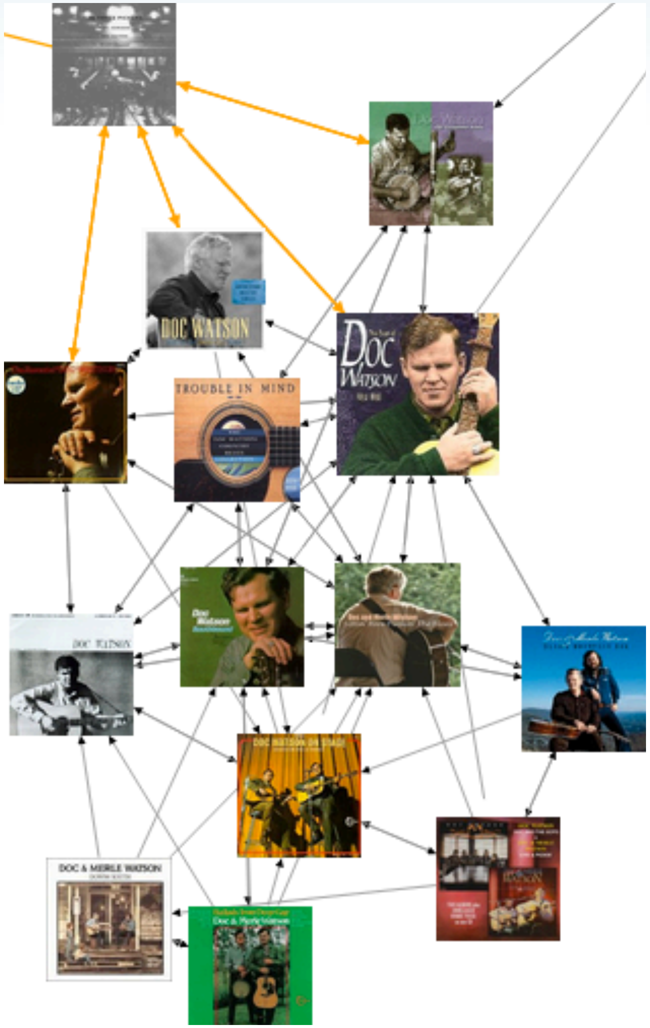
Price: $13.85
Artist: Ricky Skaggs
Released: 2003-07-15
Label: Rounder / Umg

amazon.com

Product Description Amazon.com

Customer Reviews

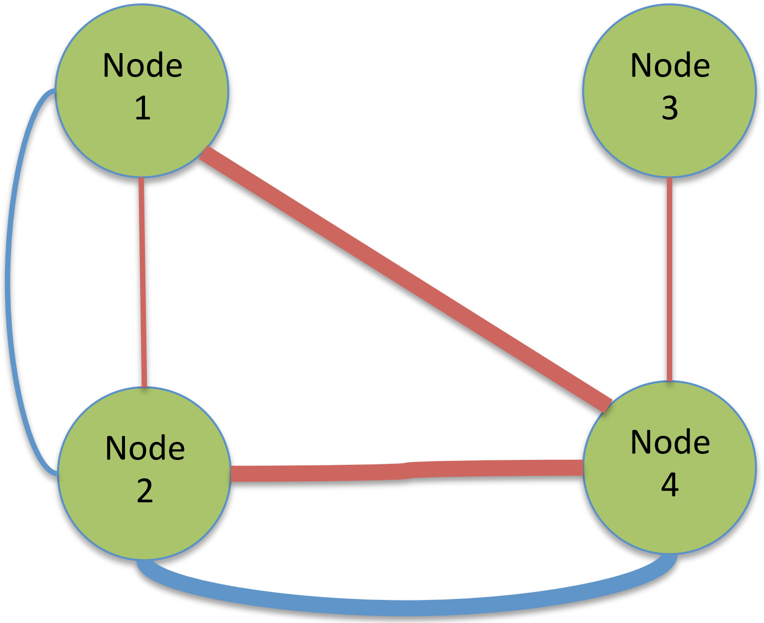
2003 album featuring a trio of Bluegrass lege
Earl Scruggs, Doc Watson and Ricky Skaggs.



With Doc Watson

The DataBridge Strategy: Multidimensional Networks

- Nodes represents a single data set
- Edges represents the similarity of the two data sets
- Line thickness denotes strength of similarity



Similarity Measure 1
Similarity Measure 2

The DataBridge Strategy: Build a Social Network for Scientific Data

- Instrument known data
  - Use DataVerse Network and iRODS
  - DataVerse contains social science and political data
  - iRODS used by many academic and government agencies around the world

The DataBridge Strategy: Building a Social Network for Scientific Data

- Investigate similarity measures:
  - Data to Data Connections: metadata and derived data about the data set
  - User to Data Connections: metadata about the usage and users of the data set
  - Method to Data Connections: metadata about the analyses of the data set

The DataBridge Strategy: Data to Data Similarity Measures

- Use native and “derived” metadata.
  - Native metadata provided with the dataset
  - Derived metadata e.g.: from the Hive ontology engine
- Use “categorical” similarity measures such as occurrence frequency to produce a similarity matrix for non-numeric data.

The DataBridge Strategy: User to Data Similarity Measures

- Create audit trails tracking
  - Use of data sets in published papers
  - Views and downloads of data sets
  - owners of data sets
- Calculate similarity of data sets from audit trails
- Use frequency and recency of access as a measure of data value

The DataBridge Strategy: Methods to Data Similarity Measures

- Create an ontology of analytic methods and applications
- Gather information about the usage of methods on data sets
- Calculate similarity from ontology and usage information

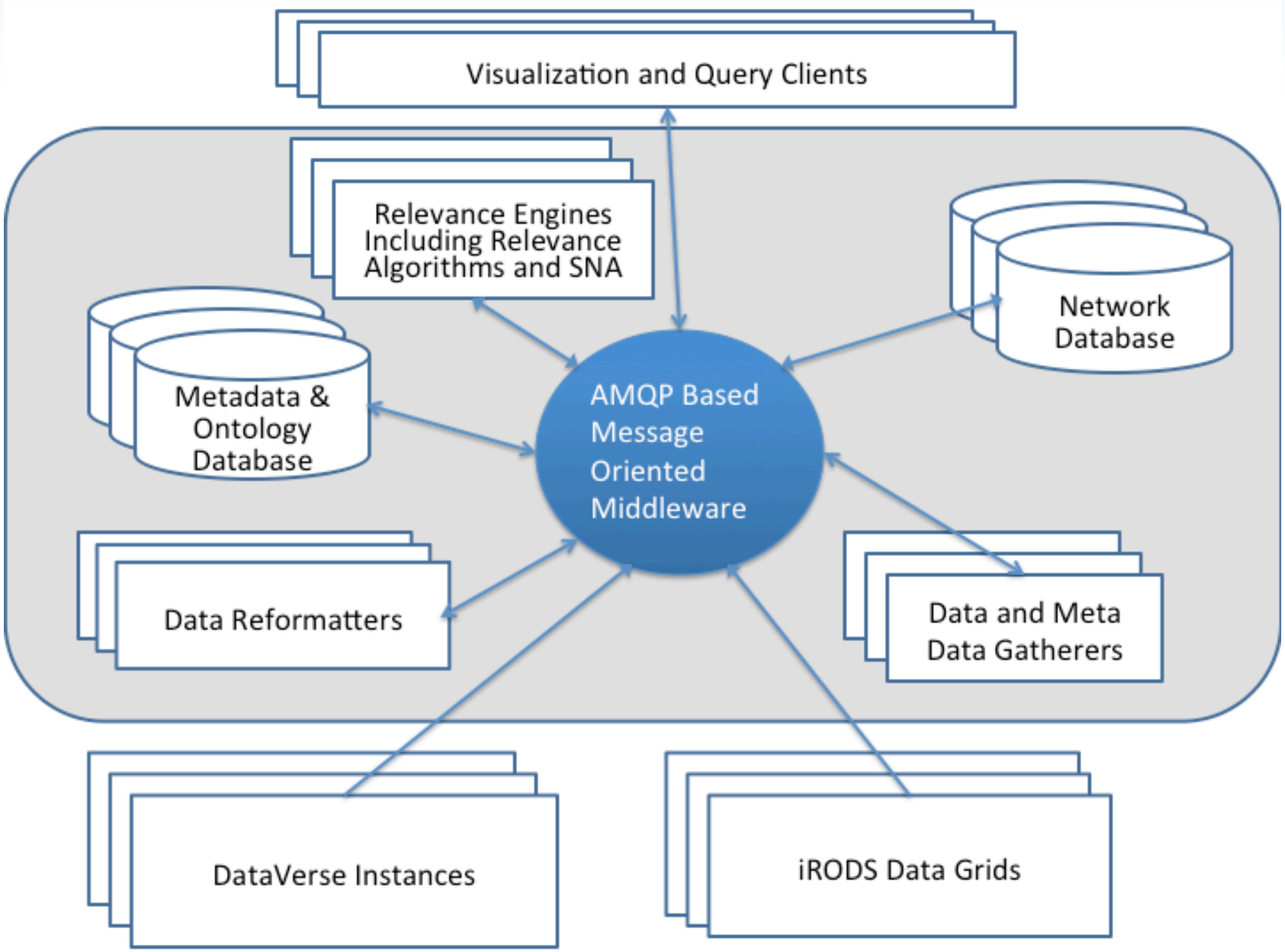
The DataBridge Strategy: Community Detection

- Investigate a number of community detection algorithms e.g.:
  - Spatial algorithms such as Euclidean or Manhattan distance
  - Algorithms based on adjacency relationships

The DataBridge Strategy: Query Interfaces

- Simple network visualization as shown already
  - Add thresholds and other filters
- Multiple dimensional queries
  - Simple relational calculus
  - Ontology of similarity and community detection methods combined with SPARQL

DataBridge Implementation


Visualization and Query Clients
Relevance Engines Including Relevance Algorithms and SNA
Network Database
Metadata & Ontology Database
AMQP Based Message Oriented Middleware
Data Reformatters
Data and Meta Data Gatherers
DataVerse Instances
iRODS Data Grids

renci

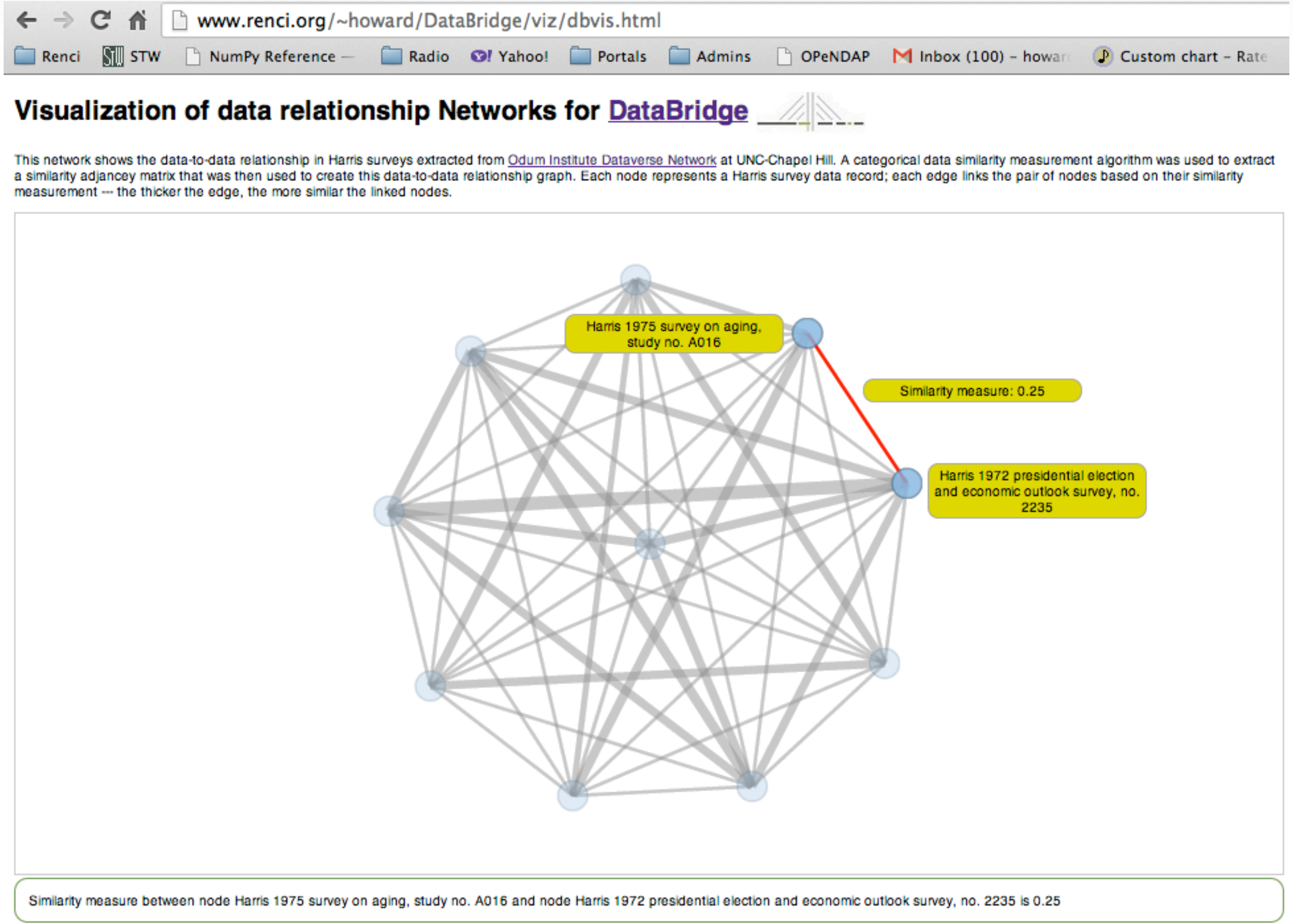
NSF

DATABRIDGE

DataBridge Progress to Date

- Initial gatherer that populates metadata store from our local DVN
- Metadata store is the file system
- Relevance engine with one algorithm
- Network database is currently both Neo4j and Titan
- JavaScript based visualization
- AMQP based middleware for connections

DataBridge Progress to Date: JavaScript based network visualization tool



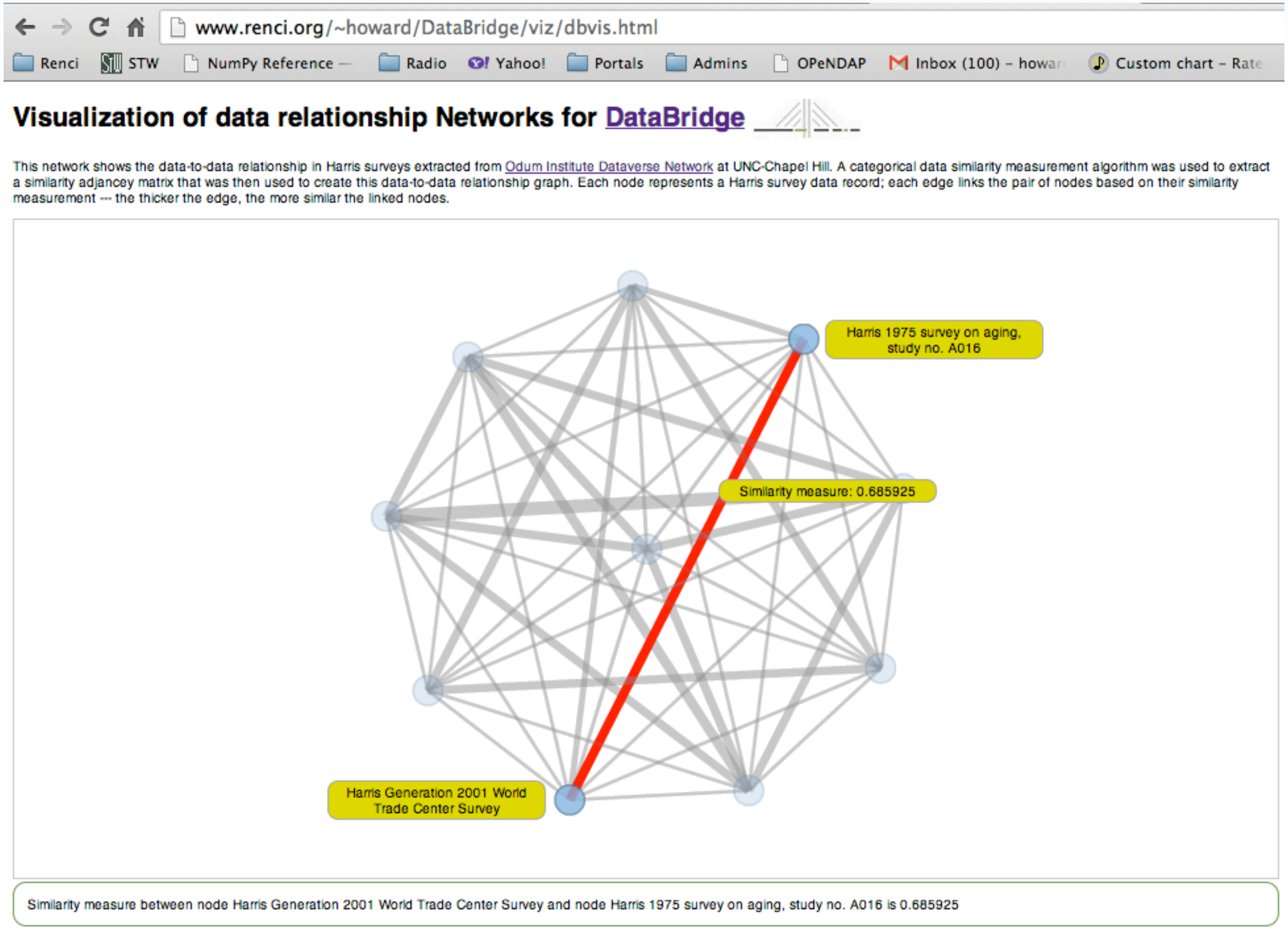
www.renci.org/~howard/DataBridge/viz/dbvis.html
Renci
STW
NumPy Reference —
Radio
Yahoo!
Portals
Admins
OPeNDAP
Inbox (100) – howar
Custom chart – Rate
Harris 1975 survey on aging, study no. A016
Similarity measure: 0.25
Harris 1972 presidential election and economic outlook survey, no. 2235

Visualization of data relationship Networks for DataBridge

This network shows the data-to-data relationship in Harris surveys extracted from Odum Institute Dataverse Network at UNC-Chapel Hill. A categorical data similarity measurement algorithm was used to extract a similarity adjancey matrix that was then used to create this data-to-data relationship graph. Each node represents a Harris survey data record; each edge links the pair of nodes based on their similarity measurement --- the thicker the edge, the more similar the linked nodes.

Similarity measure between node Harris 1975 survey on aging, study no. A016 and node Harris 1972 presidential election and economic outlook survey, no. 2235 is 0.25

DataBridge Progress to Date: JavaScript based network visualization tool

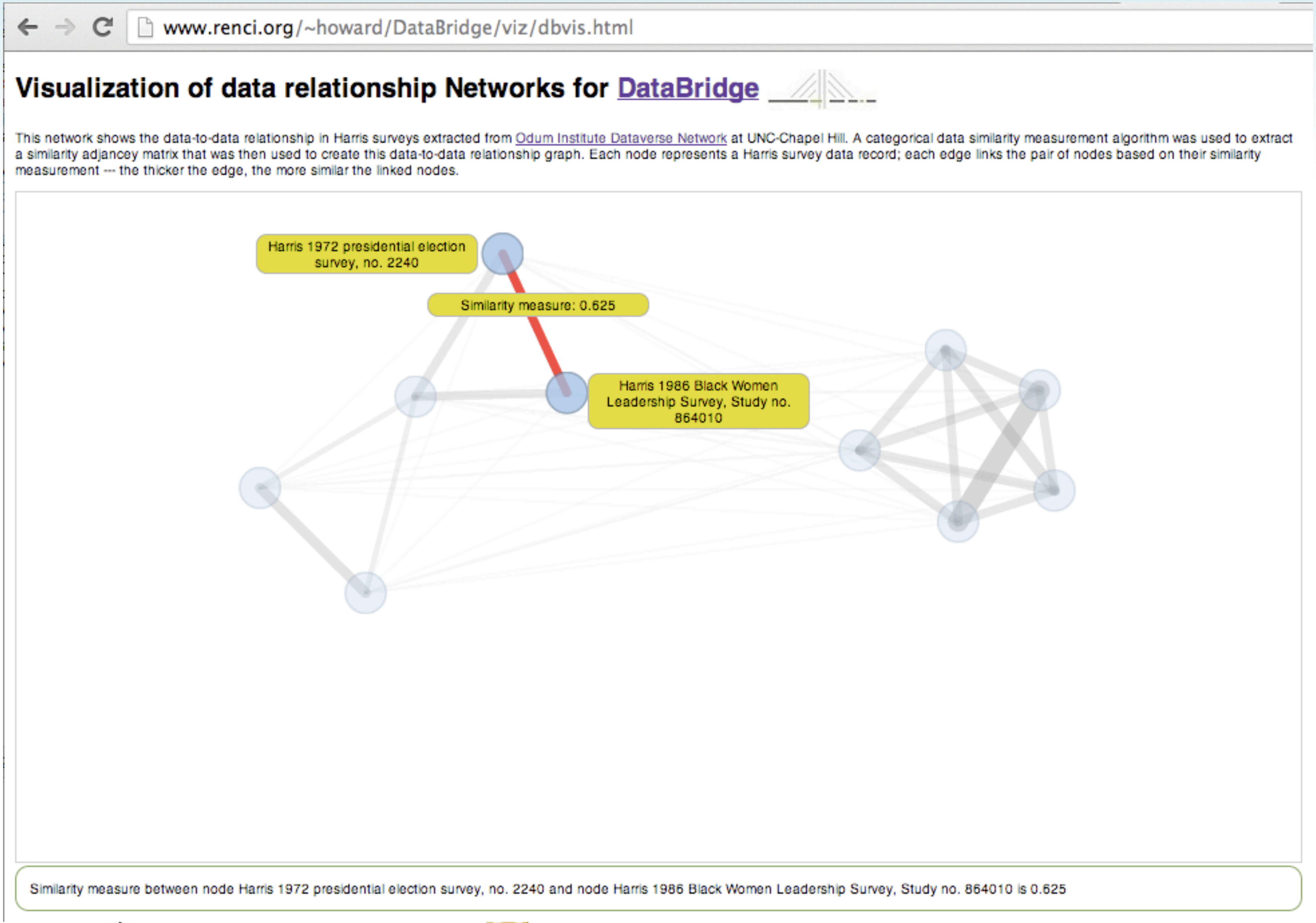


Visualization of data relationship Networks for DataBridge

This network shows the data-to-data relationship in Harris surveys extracted from Odum Institute Dataverse Network at UNC-Chapel Hill. A categorical data similarity measurement algorithm was used to extract a similarity adjancey matrix that was then used to create this data-to-data relationship graph. Each node represents a Harris survey data record; each edge links the pair of nodes based on their similarity measurement --- the thicker the edge, the more similar the linked nodes.

Harris 1975 survey on aging, study no. A016
Similarity measure: 0.685925
Harris Generation 2001 World Trade Center Survey

Similarity measure between node Harris Generation 2001 World Trade Center Survey and node Harris 1975 survey on aging, study no. A016 is 0.685925


www.renci.org/~howard/DataBridge/viz/dbvis.html
Harris 1972 presidential election survey, no. 2240
Similarity measure: 0.625
Harris 1986 Black Women Leadership Survey, Study no. 864010

Visualization of data relationship Networks for DataBridge

This network shows the data-to-data relationship in Harris surveys extracted from Odum Institute Dataverse Network at UNC-Chapel Hill. A categorical data similarity measurement algorithm was used to extract a similarity adjancey matrix that was then used to create this data-to-data relationship graph. Each node represents a Harris survey data record; each edge links the pair of nodes based on their similarity measurement --- the thicker the edge, the more similar the linked nodes.

Similarity measure between node Harris 1972 presidential election survey, no. 2240 and node Harris 1986 Black Women Leadership Survey, Study no. 864010 is 0.625

DataBridge Conclusion

- A promising start on a challenging research problem.
- Outstanding issues:
  - What metrics we will use to compare various similarity and community detection algorithms?
  - What happens when we scale up?
  - What sort of queries and query interfaces will be most effective?
  - How best to encourage publication?

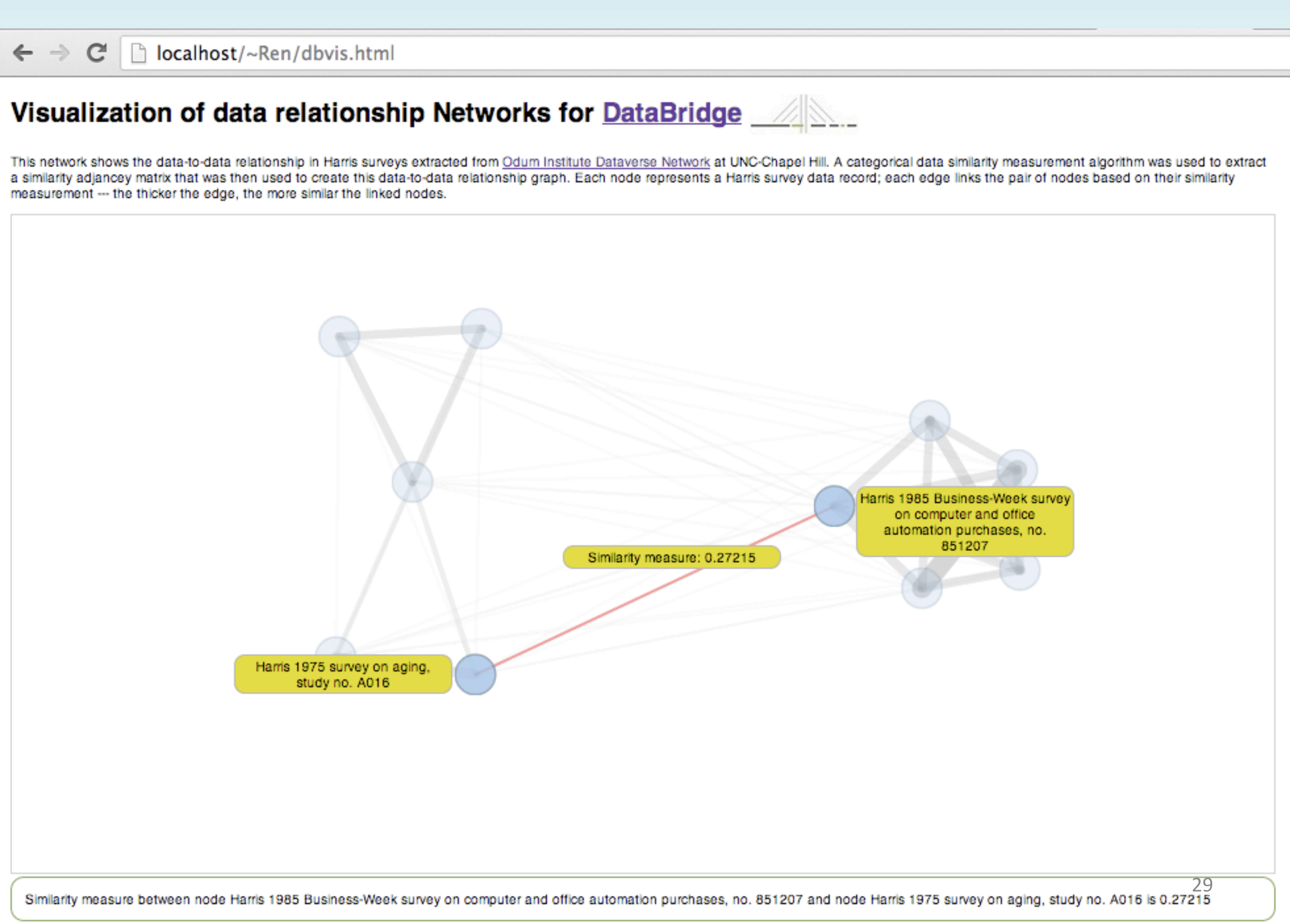
DataBridge Team

- PI: Arcot Rajasekar RENCI and SILS, UNC-Chapel Hill
- Collaborators:
  - Odum Institute, UNC-Chapel Hill
  - Population Informatics Research Group, UNC-Chapel Hill, Texas A & M University
  - iLab, North Carolina A&T University
  - The Institute for Quantitative Social Science, Harvard University
- Funded by: NSF Office of Cyberinfrastructure Awards OCI-1247562, OCI-1247602 and OCI-1247663

renci NSF DATABRIDGE

The Motivating Problem

- The "Classic" 5 V Problem
  - Examples: CERN/Twitter/Facebook
  - Volume: massive data quantity
  - Velocity: data flow
  - Variety: structured or unstructured
  - Veracity: data quality
  - Value: uncertain economic/ scientific utility



Visualization of data relationship Networks for DataBridge

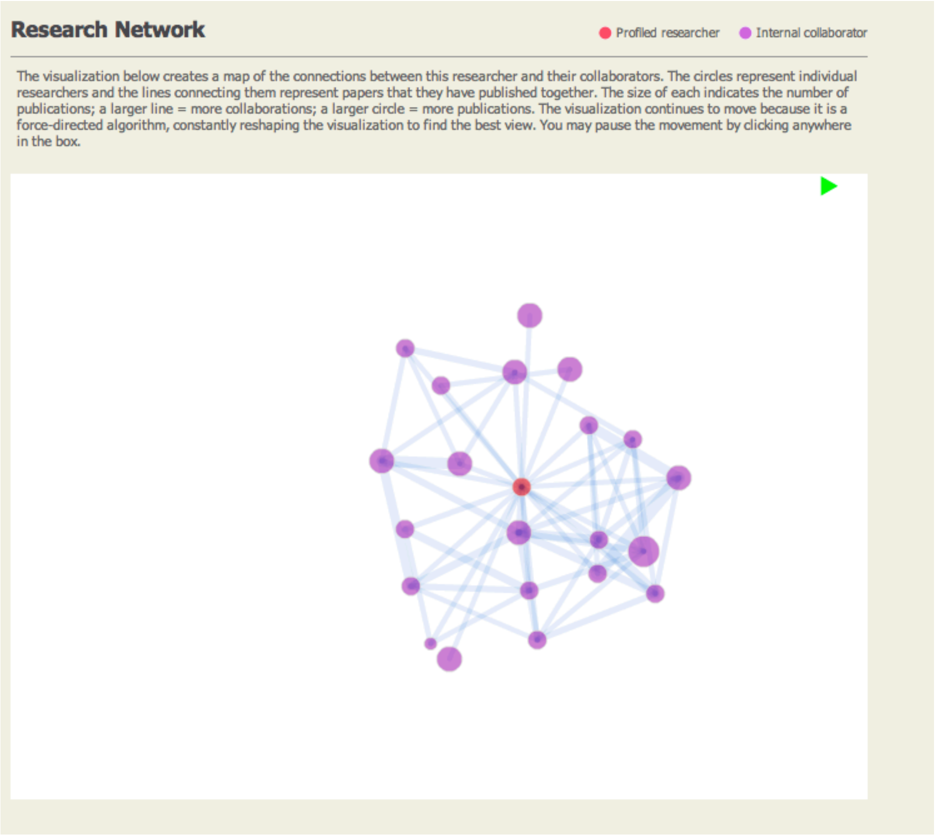
This network shows the data-to-data relationship in Harris surveys extracted from Odum Institute Dataverse Network at UNC-Chapel Hill. A categorical data similarity measurement algorithm was used to extract a similarity adjancey matrix that was then used to create this data-to-data relationship graph. Each node represents a Harris survey data record; each edge links the pair of nodes based on their similarity measurement --- the thicker the edge, the more similar the linked nodes.

Harris 1985 Business-Week survey on computer and office automation purchases, no. 851207
Similarity measure: 0.27215
Harris 1975 survey on aging, study no. A016

Similarity measure between node Harris 1985 Business-Week survey on computer and office automation purchases, no. 851207 and node Harris 1975 survey on aging, study no. A016 is 0.27215

DataBridge Implementation Goals

- Define protocols, methods and an implementation for exchanging data about data
- Provide data social network to existing systems
- Enable scientists to browse/query the data network



Research Network

● Profiled researcher ● Internal collaborator

The visualization below creates a map of the connections between this researcher and their collaborators. The circles represent individual researchers and the lines connecting them represent papers that they have published together. The size of each indicates the number of publications; a larger line = more collaborations; a larger circle = more publications. The visualization continues to move because it is a force-directed algorithm, constantly reshaping the visualization to find the best view. You may pause the movement by clicking anywhere in the box.

reachnc.org

- From Mendel to Brahe discovery has come from relatively small data sets
- Data enabled computational science co-equal with theory and experiment