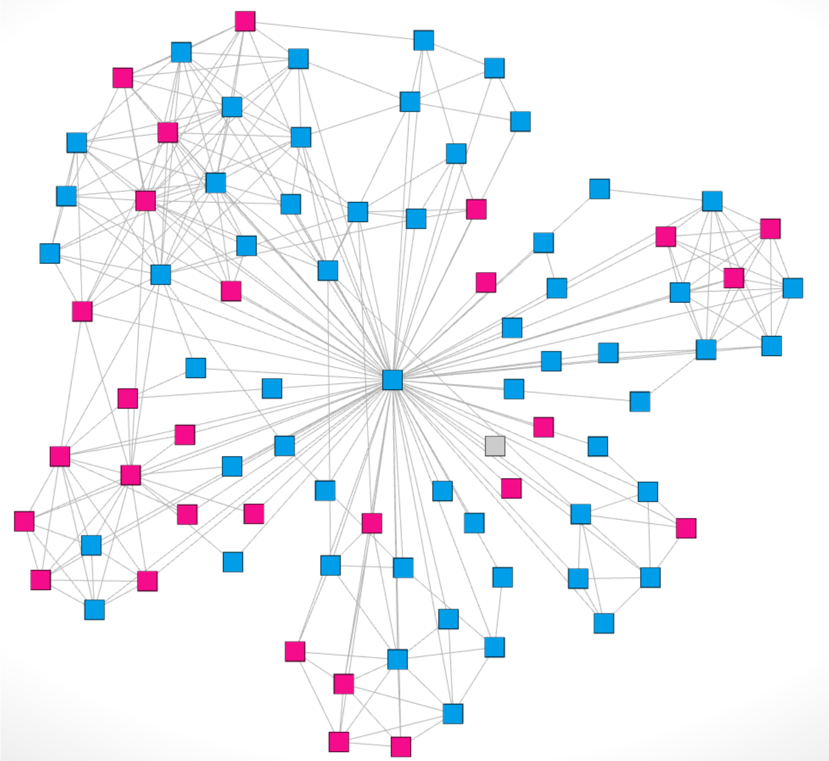
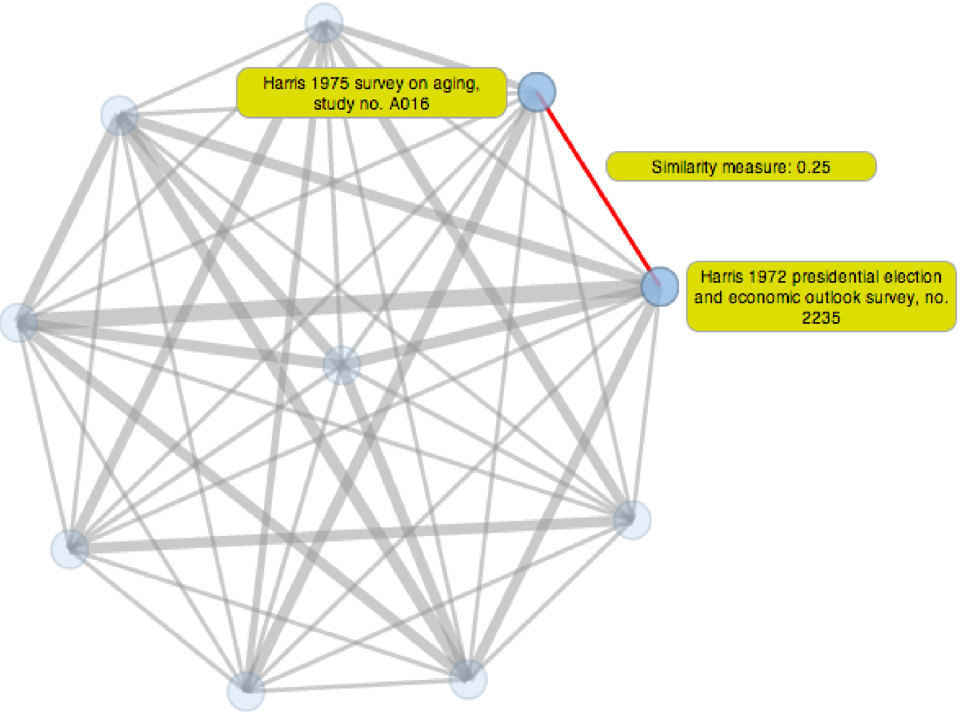
The DataBridge: A Social Network for Long Tail Science Data

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

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

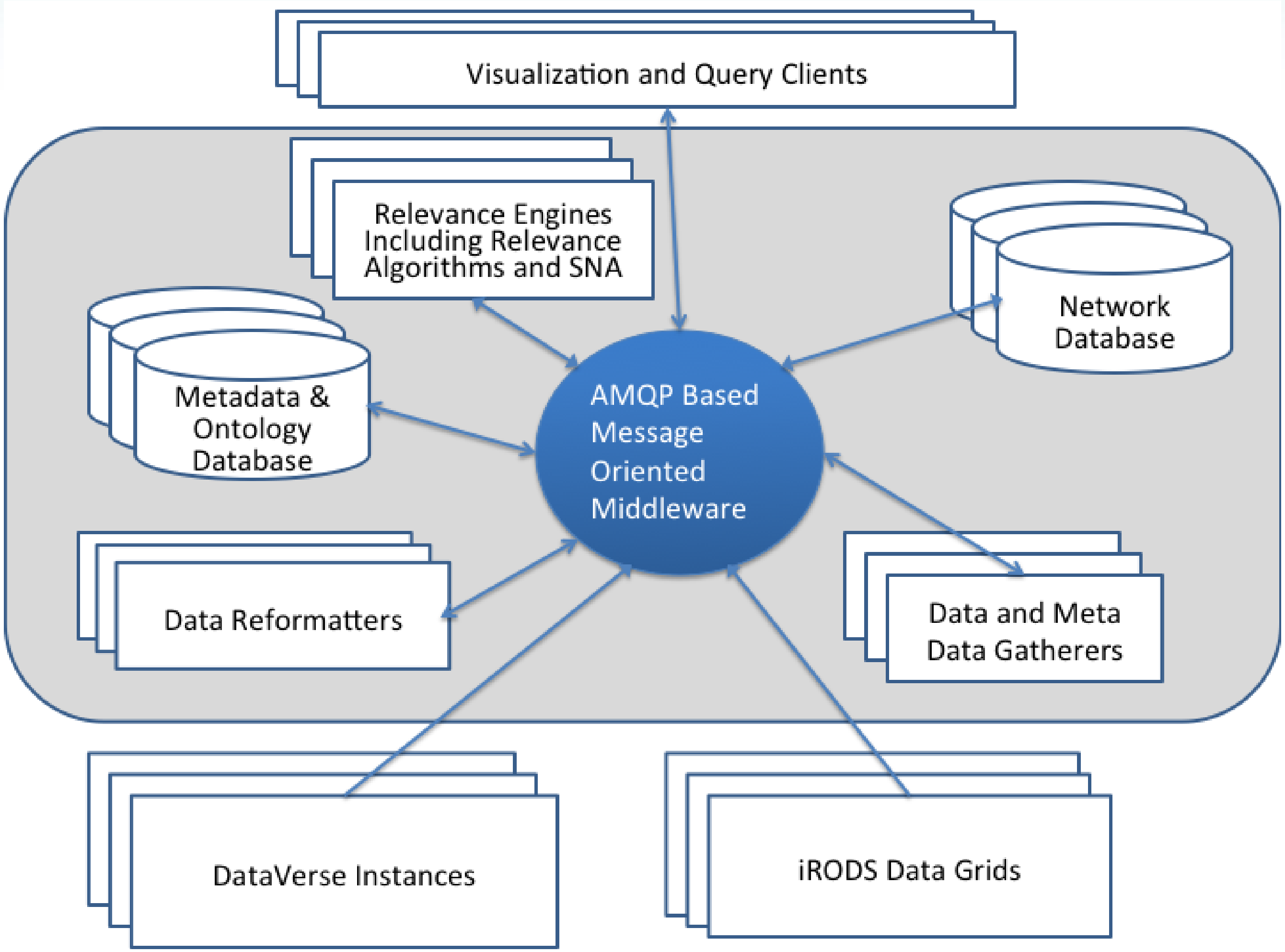
Dark Data from The Long Tail of Science

- 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)
- Long tail data sets lack structural advantages of “classic” Big 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

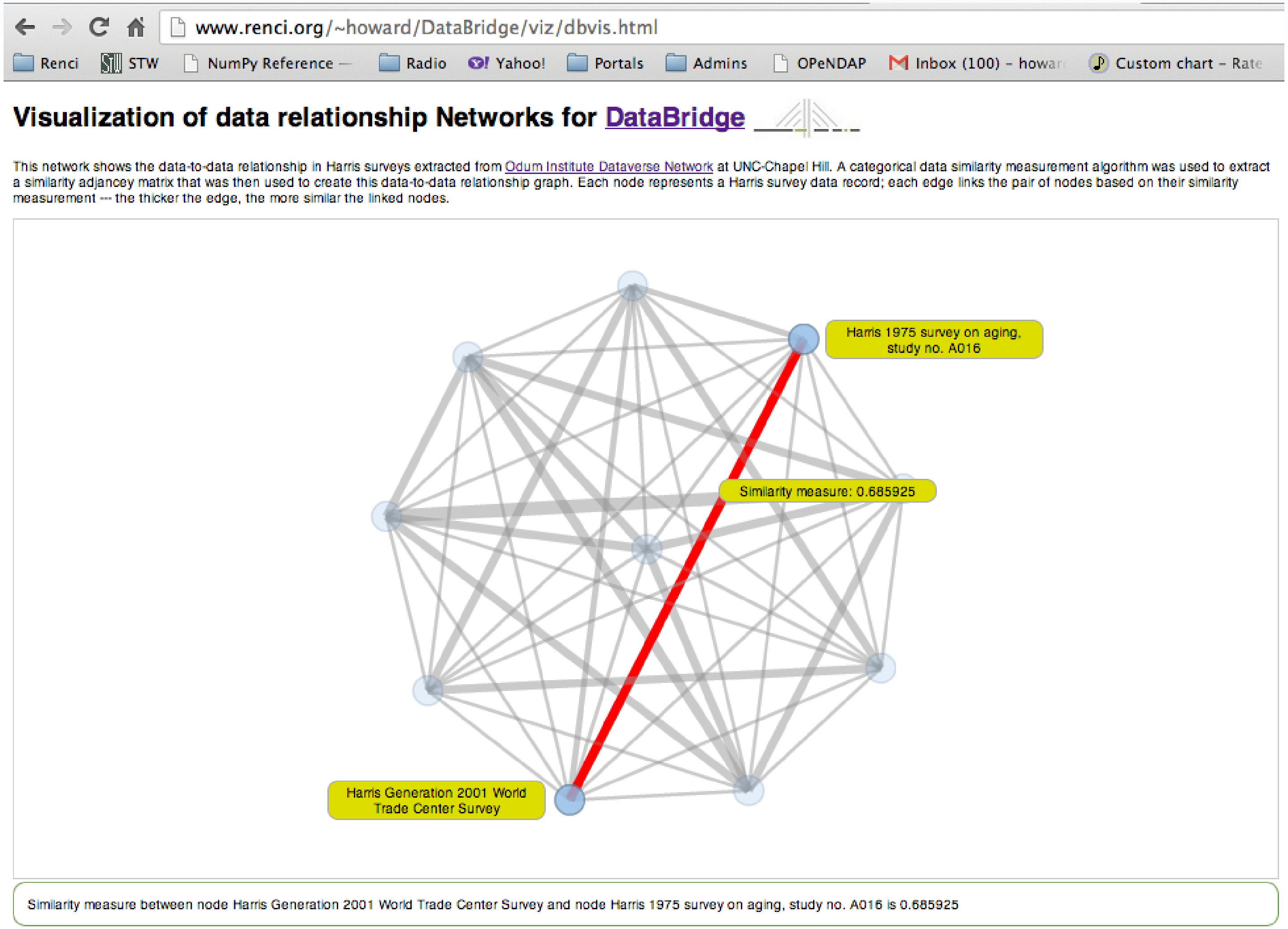
DataBridge Implementation


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

DATABRIDGE

DataBridge Progress to Date: JavaScript based network visualization tool

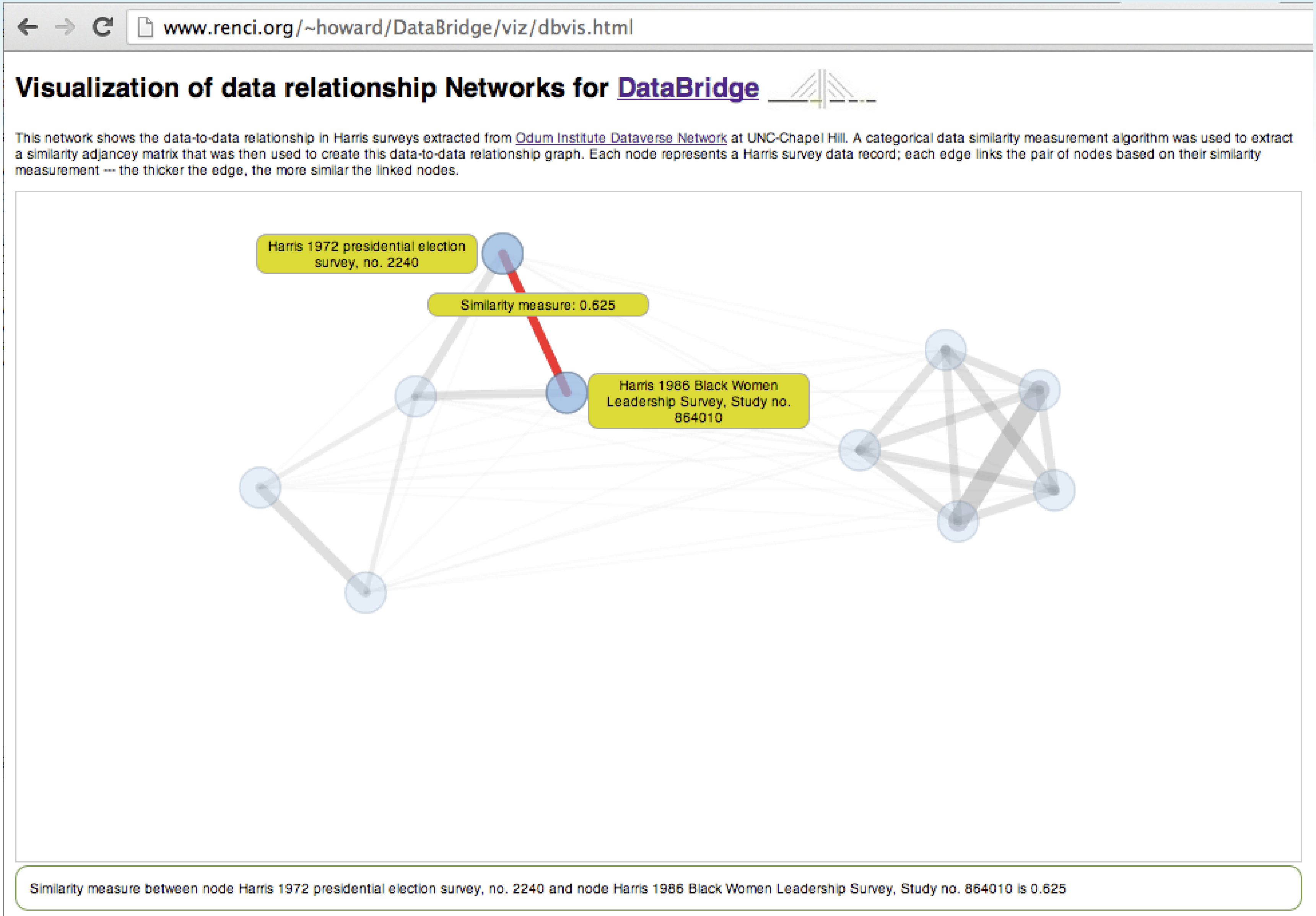

www.renci.org/~howard/DataBridge/viz/dbvis.html
Renci
STW
NumPy Reference —
Radio
Yahoo!
Portals
Admins
OPeNDAP
Inbox (100) - howar
Custom chart - Rate
Harris 1975 survey on aging, study no. A016
Similarity measure: 0.685925
Harris Generation 2001 World Trade Center Survey

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

renci NSF DATABRIDGE



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 1972 presidential election survey, no. 2240
Similarity measure: 0.625
Harris 1986 Black Women Leadership Survey, Study no. 864010

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 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