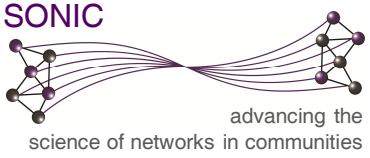
Lecture 5:
Network Data Collection Strategies

Noshir Contractor

Jane S. & William J. White Professor of Behavioral Sciences

Professor of Ind. Eng. & Mgmt. Sciences, McCormick School of Engineering
Professor of Communication Studies, School of Communication &
Professor of Management & Organizations, Kellogg School of Management,
Director, Science of Networks in Communities (SONIC) Research Laboratory
nosh@northwestern.edu



Agenda

- Strategies for the collection of network data:
- Traditional methods
- Digital harvesting of metadata for multidimensional networks

Setting Network Boundaries

- Minimal network database:
  one set of objects/nodes linked by one set of relationships/ties observed at one occasion.

- Three generic boundary specification strategies:
  1. Positional approach: characteristics of nodes or formal membership criteria
  2. Event-based approach: participation in some class of activities
  3. Relational approach: social connectedness

Data collection - Surveys

- Collect perceptions of interactions
- List of names or free recall
- Free vs. fixed choices
- Ratings vs. complete rankings

- **Pros**: Established methods, multidimensional relations, strong internal validity
- **Cons**: Expensive, not scalable, boundary conditions, bounded rationality

Respondent Accuracy

- A methodological challenge:
  - *"Informants are inaccurate; memory does decay exponentially with time… . And on top of all this there appears to be systematic distortion in how informants recall just about everything."*

— Bernard, Killworth, Kronenfeld and Sailer, 1984.

- "Accuracy" reconsidered:
  - Three realms of investigation:
    1. Behavioral patterns ("Behavioral" data)
    2. Cognitive patterns ("Cognitive" data)
    3. Relationship between the two

Data collection - Observations

- Face-to-Face interactions: Who talks to whom at a party?
- Who answers to what kinds of requests on a list server?

- **Pros**: Inexpensive, capture latent/hidden relationships, strong external validity
- **Cons**: Temporal censoring, entrée, very unscalable (only one set of eyes), limited multidimensionality

Data collection - Interviews

- Face-to-face, or telephone
  - Snowball principle: Who else is important in this network?

- **Pros**: Established methods, multidimensional data, strong internal validity
- **Cons**: Bounded rationality, expensive, entrée, boundary conditions

Questionnaire formats

- Question formats that can be used in a questionnaire include:
  - Roster vs. Free Recall
  - Free vs. Fixed Choice
  - Ratings vs. Complete Rankings

Data collection – Indirect data

- Archival records: past political interactions, co-authorship, court records, …
- Digital trace data: Log files of communication tools, online activities.

- **Pros**: Inexpensive, exhaustive, multidimensional, strong validity
- **Cons**: Need specialized skills, very large data, entrée, construct validity

Its all about "Relational Metadata"

- Technologies that "***capture***" communities' relational meta-data (Pingback and trackback in interblog networks, blogrolls, data provenance)

- Technologies to "***tag***" communities' relational metadata (from Dublin Core taxonomies to folksonomies ('wisdom of crowds') like
  - Tagging photos (Flickr), images (ESP), blogs (Technorati), news stories (digg)
  - Social bookmarking (del.icio.us)
  - Social citations (CiteULike.org)
  - Social libraries (discogs.com, LibraryThing.com)
  - Social shopping (SwagRoll, Kaboodle, thethingsiwant.com)
  - Social networks (FOAF, XFN, MySpace, Facebook)

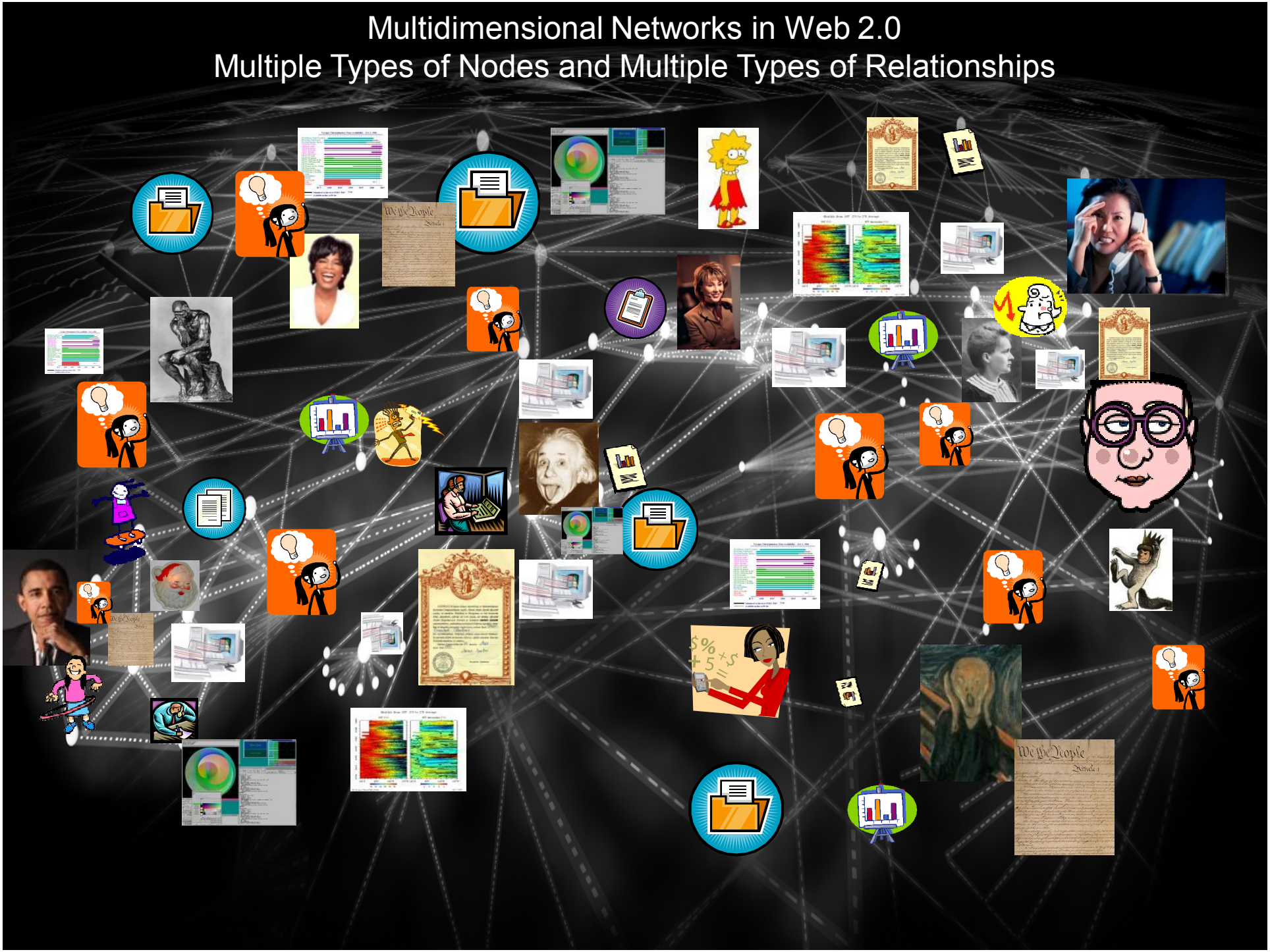
- Technologies to "***manifest***" communities' relational metadata (Tagclouds, Recommender systems, Rating/Reputation systems, ISI's HistCite, Network Visualization systems)

Data Collection

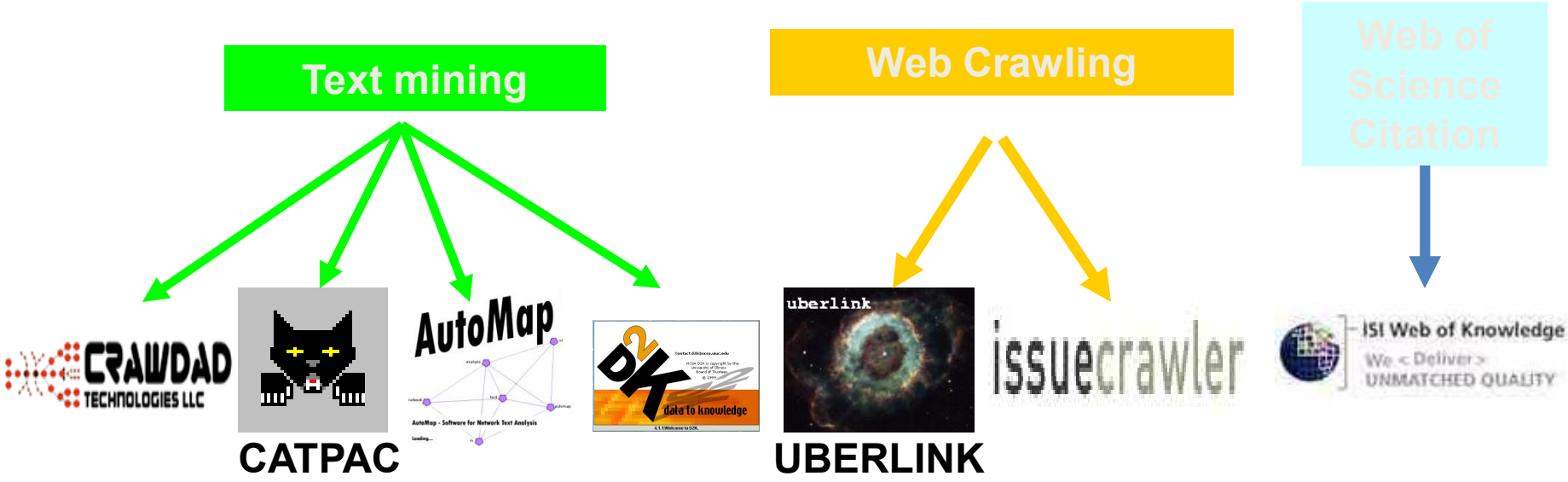
There are additional ways in which social network data can be gathered. These techniques include:

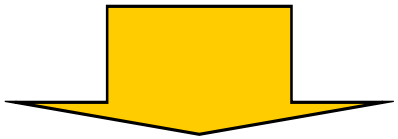
- Experiments
- Ego-centered
- Small World
- Diaries

Multidimensional Networks in Web 2.0
Multiple Types of Nodes and Multiple Types of Relationships

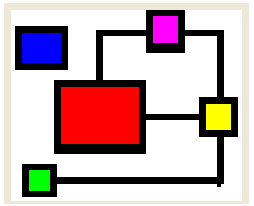


Digital Harvesting





IKNOW
Inquiring Knowledge Networks on Web

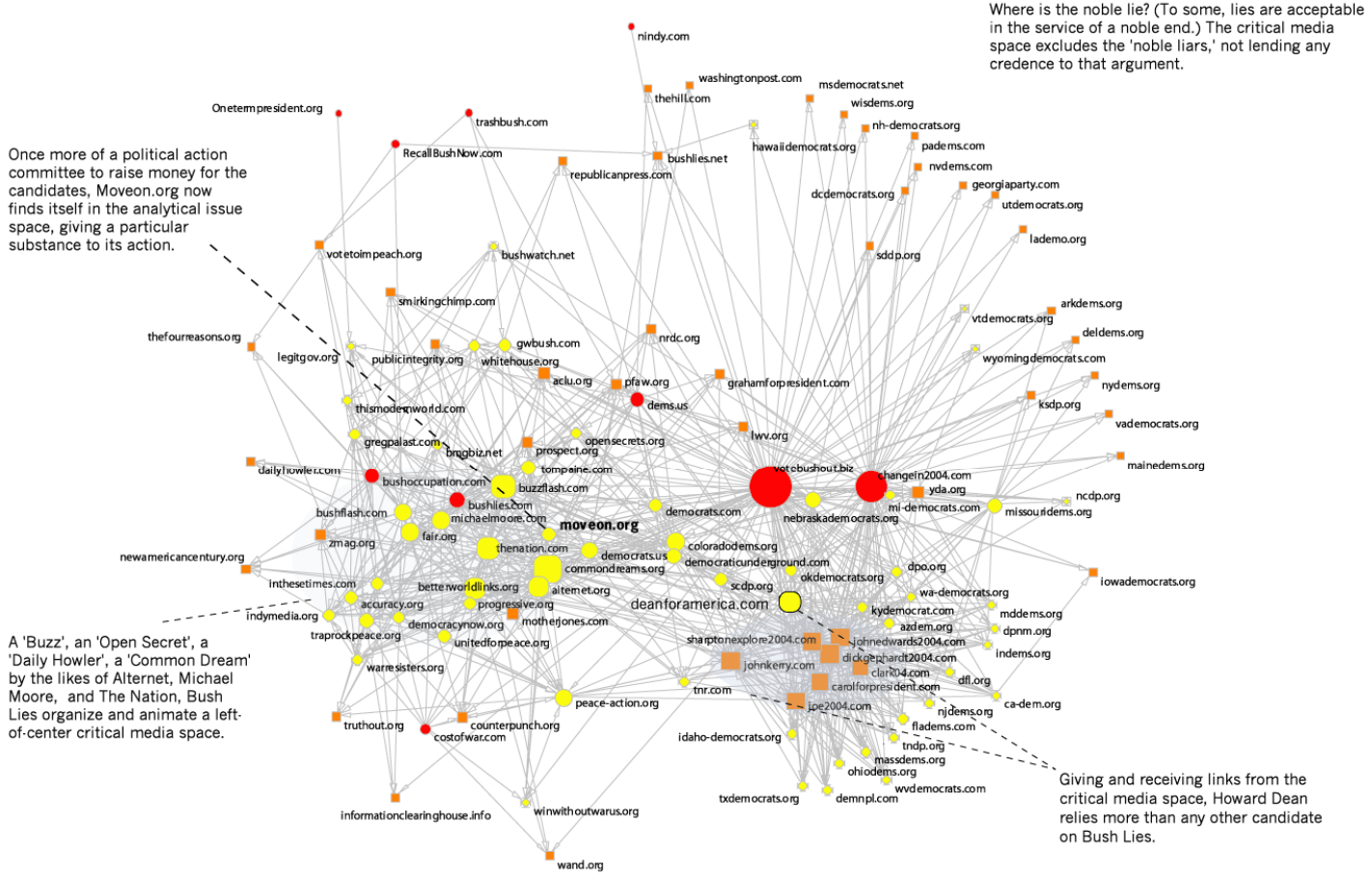


NETDRAW

Analyses and Visualizations

http://iknowinc.com/iknow/sb_digital_forum/www/iknow.cgi

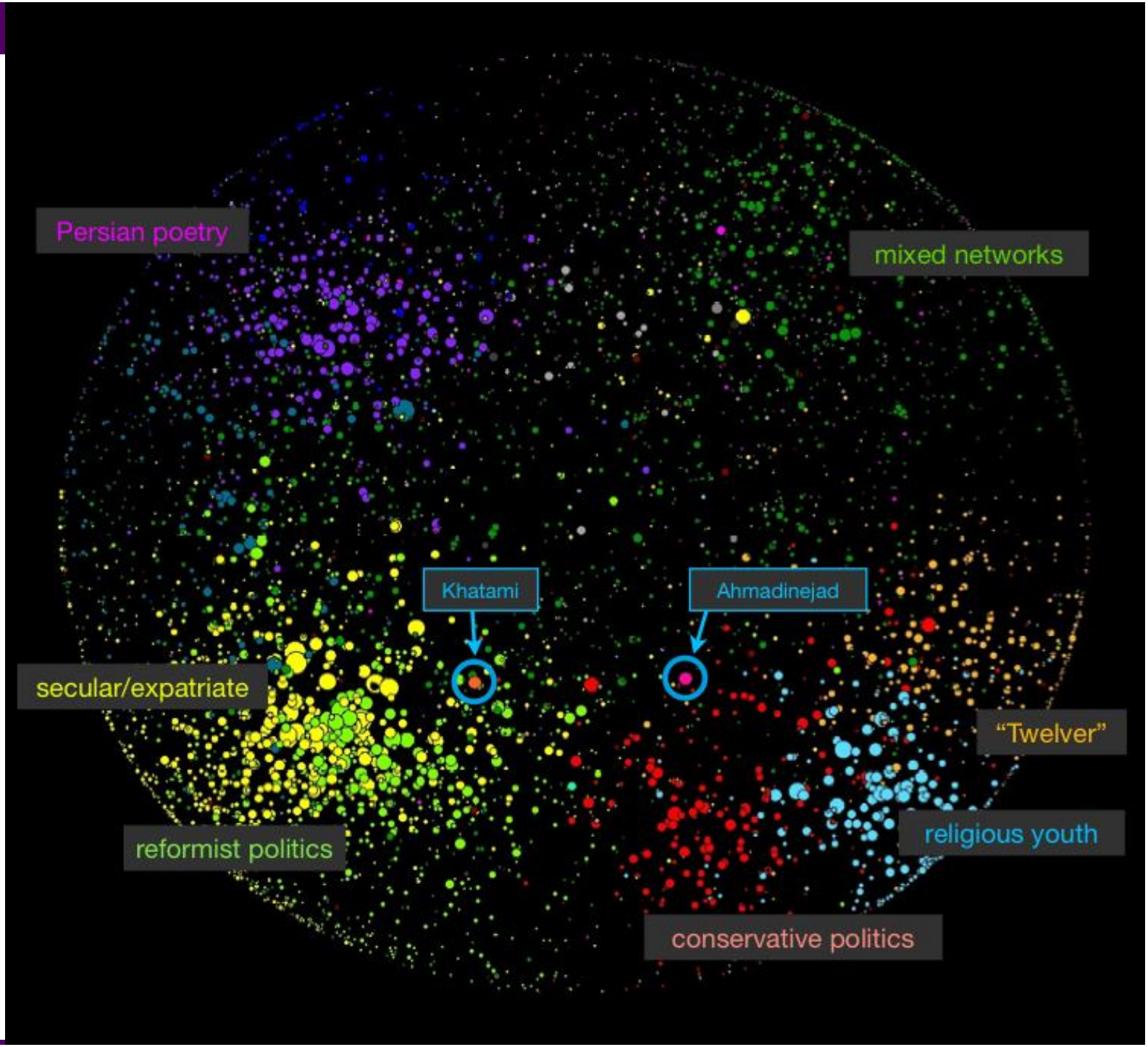
Issue Crawler (govcom.org)



NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Date Taken:
6 Apr 2008

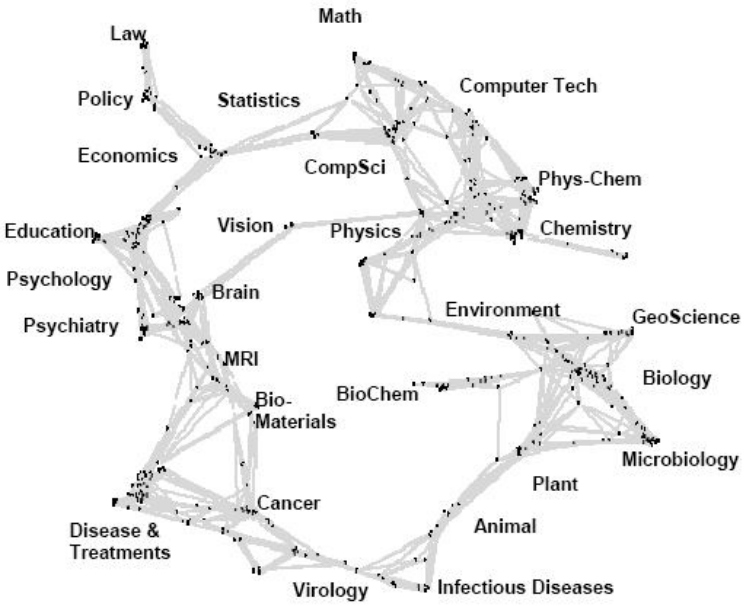


John Kelly & Bruce Etling (2008) Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere

NORTHWESTERN UNIVERSITY

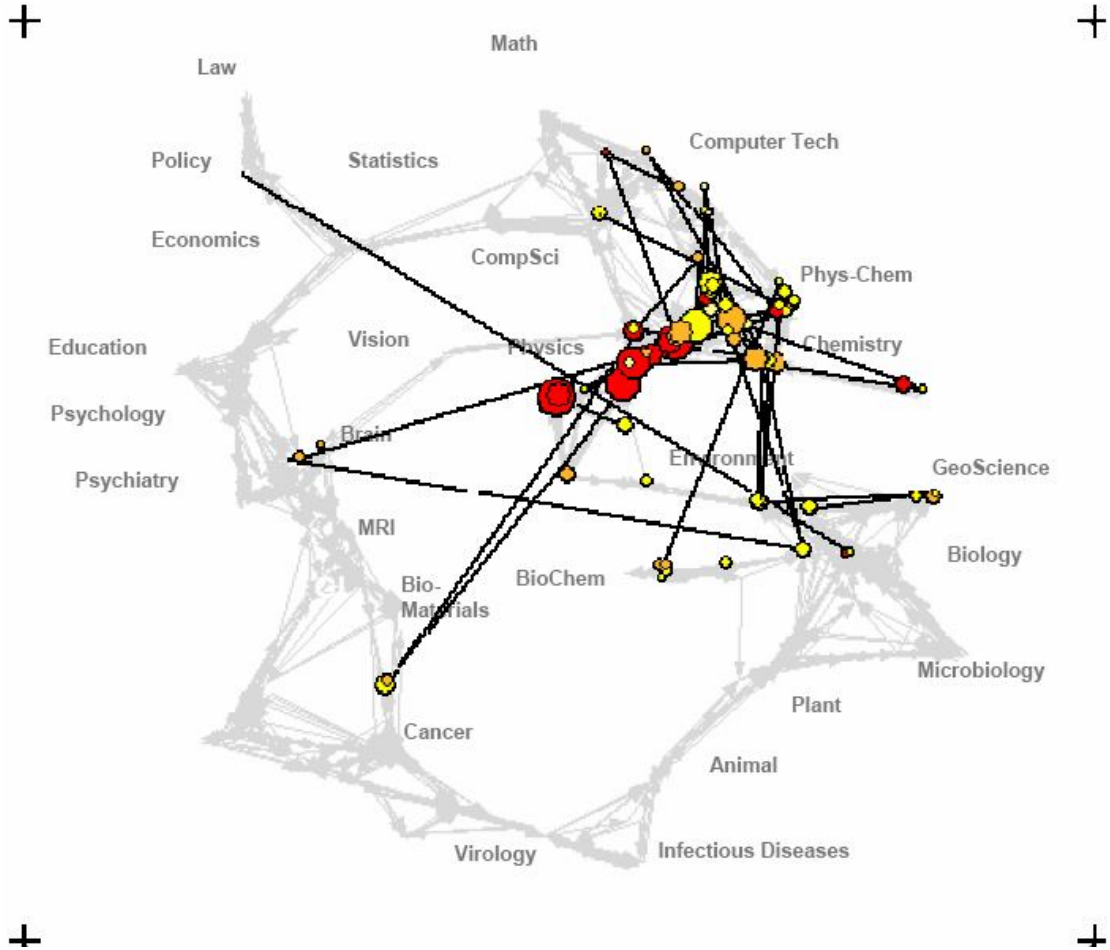
(1) Disciplinary Science Model - Details

- Uses combined SCIE/SSCI from 2002
  - 1.07M papers, 24.5M references, 7300 journals
  - Bibliographic coupling of papers, aggregated to journals
- Initial ordination and clustering of journals gave 671 clusters
- Coupling counts reaggregated at the journal cluster level; ordination of journal clusters
  - (x,y) positions for each journal cluster
  - by association, (x,y) positions for each journal



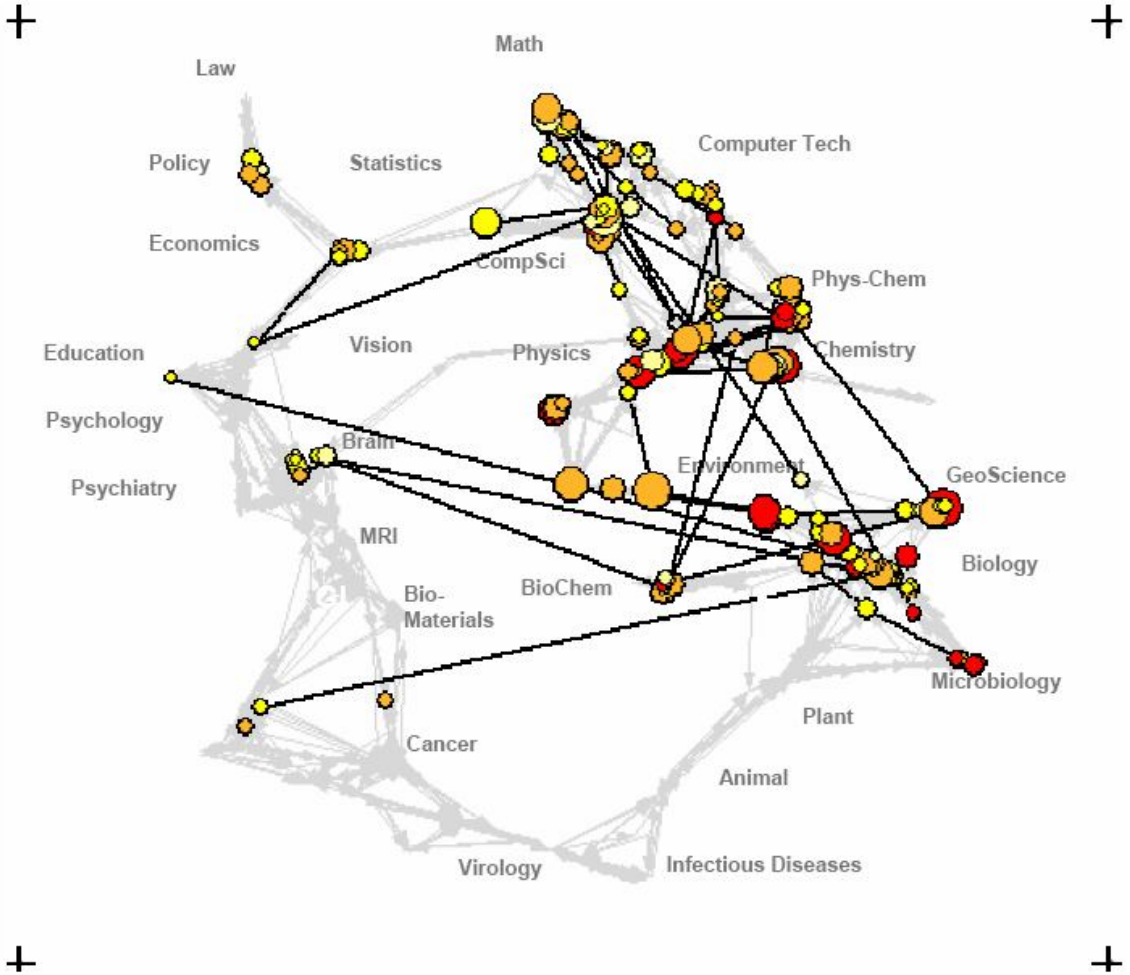
Klavans, R., & Boyack, K. W. (2005). Mapping world-wide science at the paper level. *ISSI05, Stockholm, Sweden, July 24-28, 2005.*

Funding patterns of the US Department of Energy (DOE)



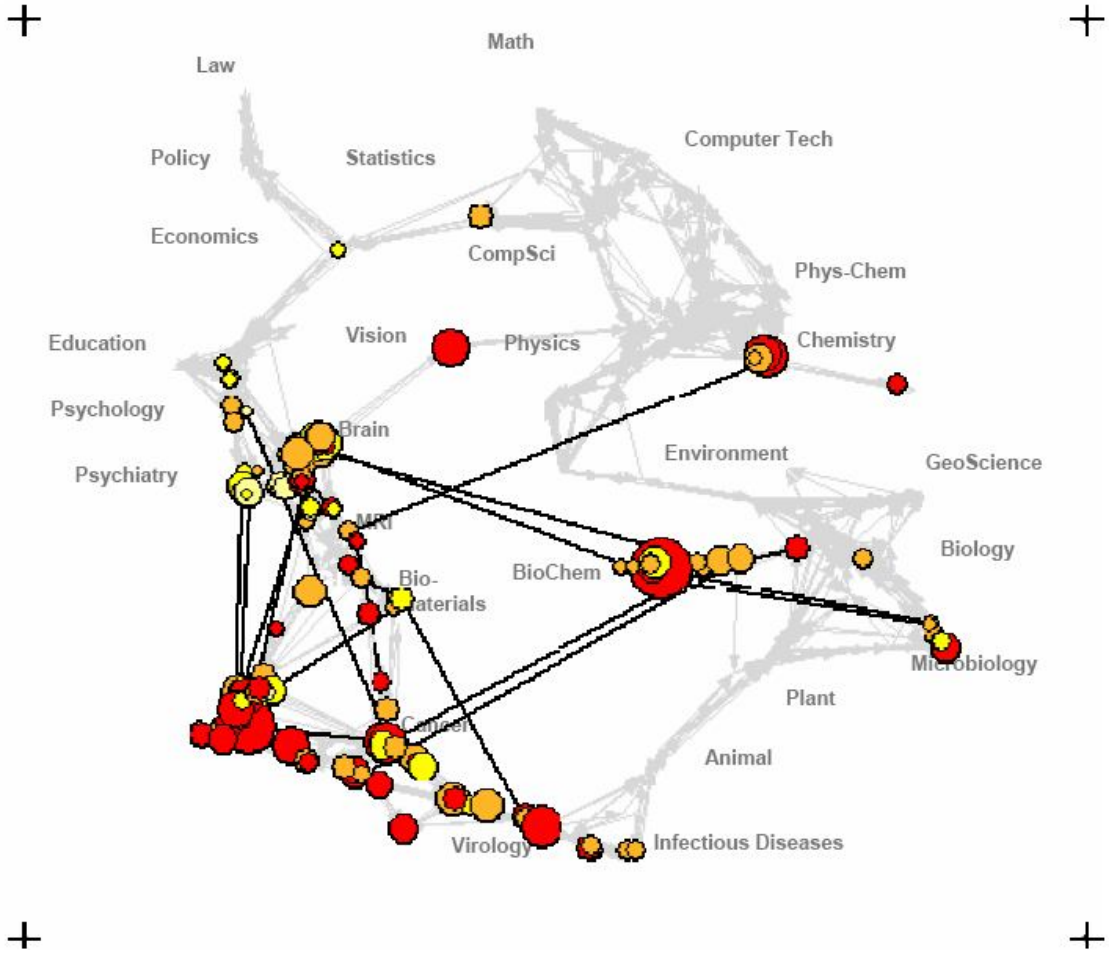
Klavans, R., & Boyack, K. W. (2005). Mapping world-wide science at the paper level. *ISSI05, Stockholm, Sweden, July 24-28, 2005.*

Funding patterns of the National Science Foundation (NSF)



Klavans, R., & Boyack, K. W. (2005). Mapping world-wide science at the paper level. *ISSI05, Stockholm, Sweden, July 24-28, 2005.*

Funding patterns of the National Institutes of Health (NIH)



Klavans, R., & Boyack, K. W. (2005). Mapping world-wide science at the paper level. *ISSI05, Stockholm, Sweden, July 24-28, 2005.*

Hurricane Katrina 2005

Formed:	Aug 23, 2005
Dissipated:	Aug 31, 2005
Highest wind:	175 mph
Lowest press:	902 mbar
Damages:	$81.2 Billion
Fatalities:	>1,836

Map source: http://hurricane.csc.noaa.gov/

SITREP Content

- Basic Format / Information
  1. Situation (What, Where, and When)
  2. Action in Progress
  3. Action Planned
  4. Probable Support Requirements and/or Support Available
  5. Other items

Typical SITREP

*Colorado Division of Emergency Management
SITUATION REPORT 2005-6
(Hurricane Katrina)
August 30, 2005*

Event Type: Hurricane Response

Situation: On August 29, Hurricane Katrina hit the gulf coast east of New Orleans. It was considered a Category 5 Hurricane, which brings winds of over 155mph and storm surge of 18 feet above normal. Massive property damage has occurred and undetermined number of deaths and injuries.

Colorado response to date include two deployments:
- Two members from the Division of Emergency Management to the Louisiana EOC, departed on August 29.

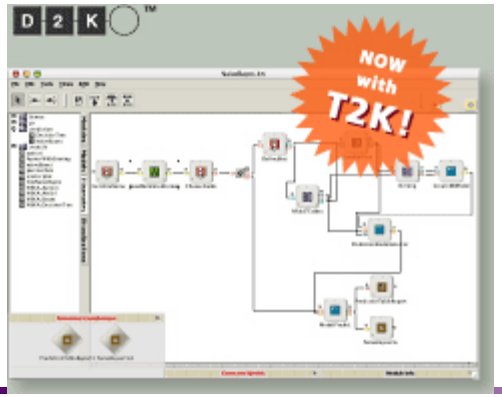
Weather Report: Katrina is moving toward the north-northeast near 18 mph. A turn toward the northeast and a faster forward speed is expected during the next 24 hours. This motion should bring the cent

Agencies Involved: Colorado Department of Military and Veteran Affairs, Department of Local Affairs, Division of Emergency Management, Governor's Office.* *

Additional Assistance Requested: Type III teams, consisting of Operations, Plans, and Logistics personnel (two individuals for each area). These teams could deploy to Alabama, Louisiana, and/or Mississippi. Teams will be at either working the State or Parish/County EOCs.

Automatic Coding

- T2K – The Text to Knowledge application environment is a rapid, flexible data mining and machine learning system
- Automated processing is done through creating itineraries that combine processing modules into a workflow
- Developed by the Automated Learning Group at NCSA

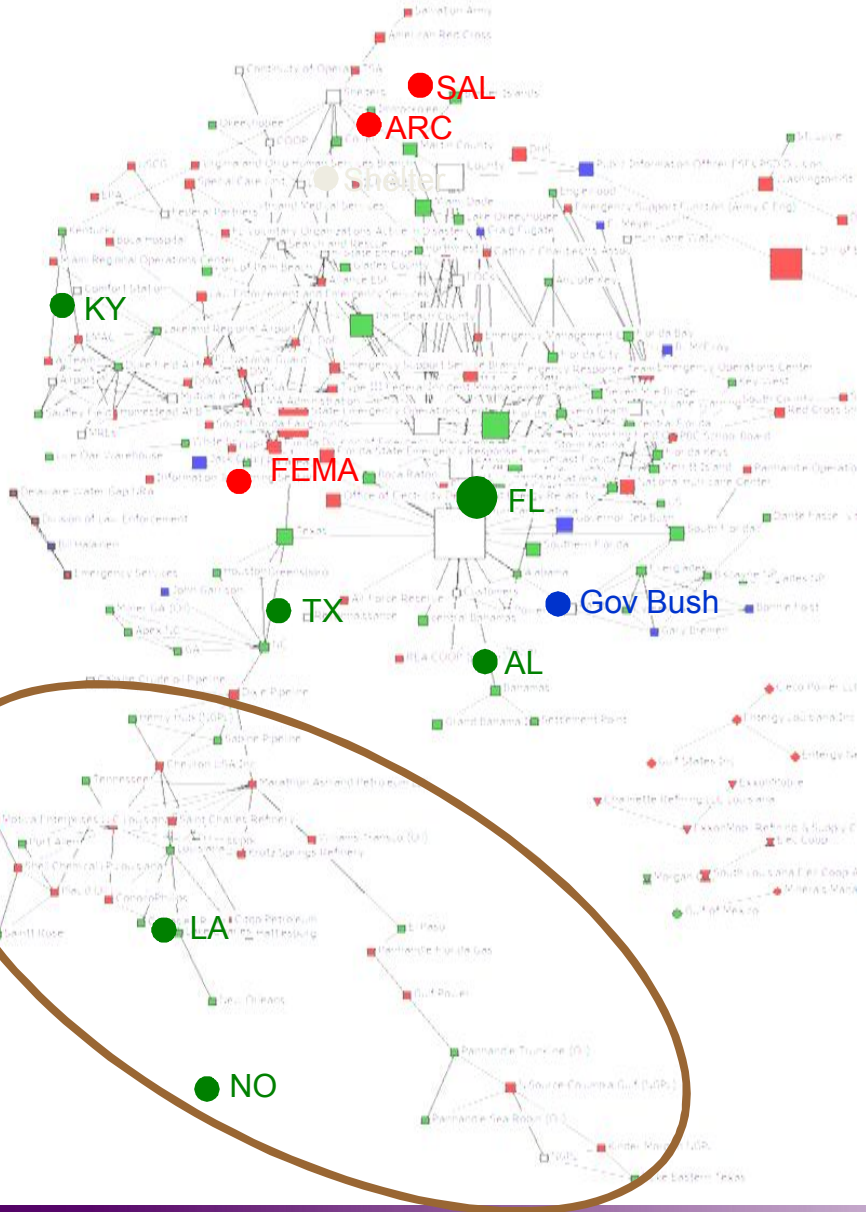


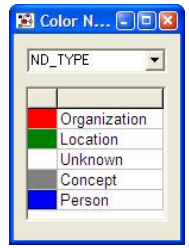
NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Time Slice 1: 8/23 to 8/25/2005

Florida is the Topic of the Conversation



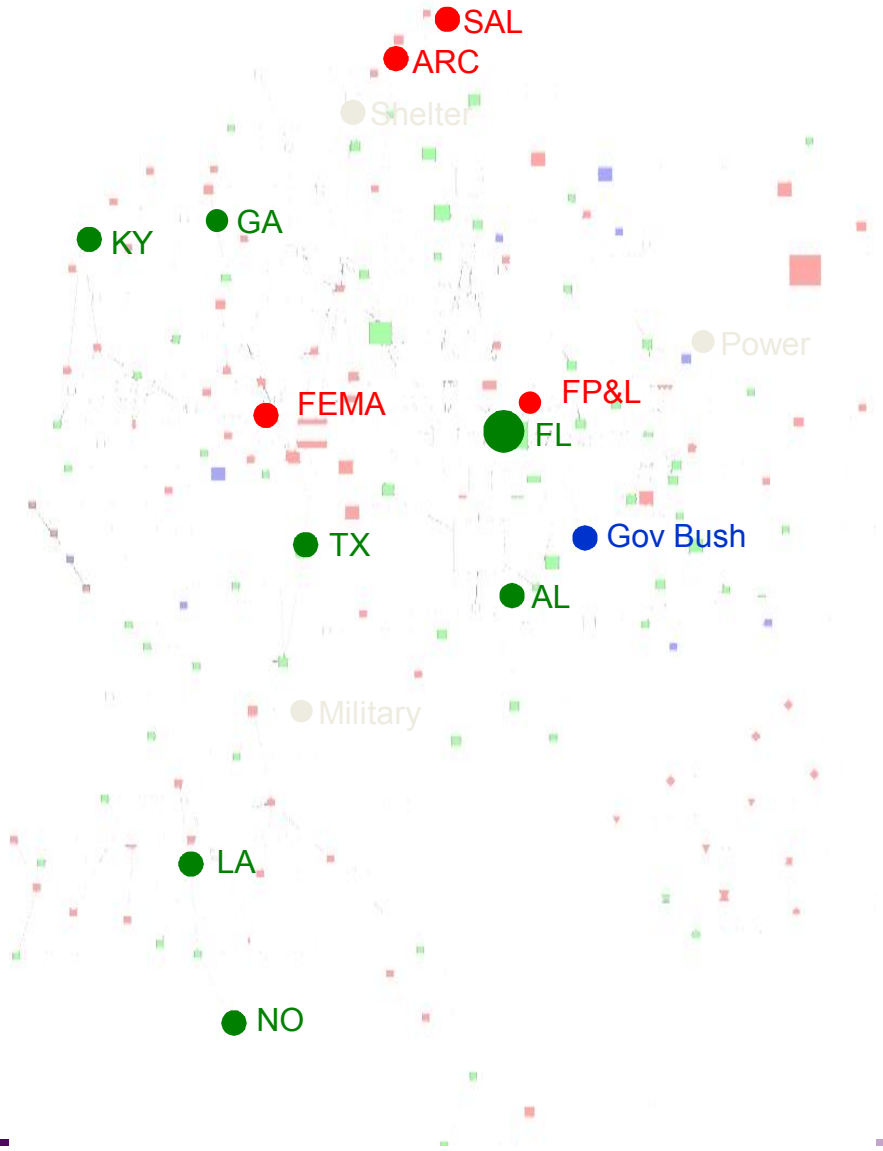


Petroleum Network formed Early

NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

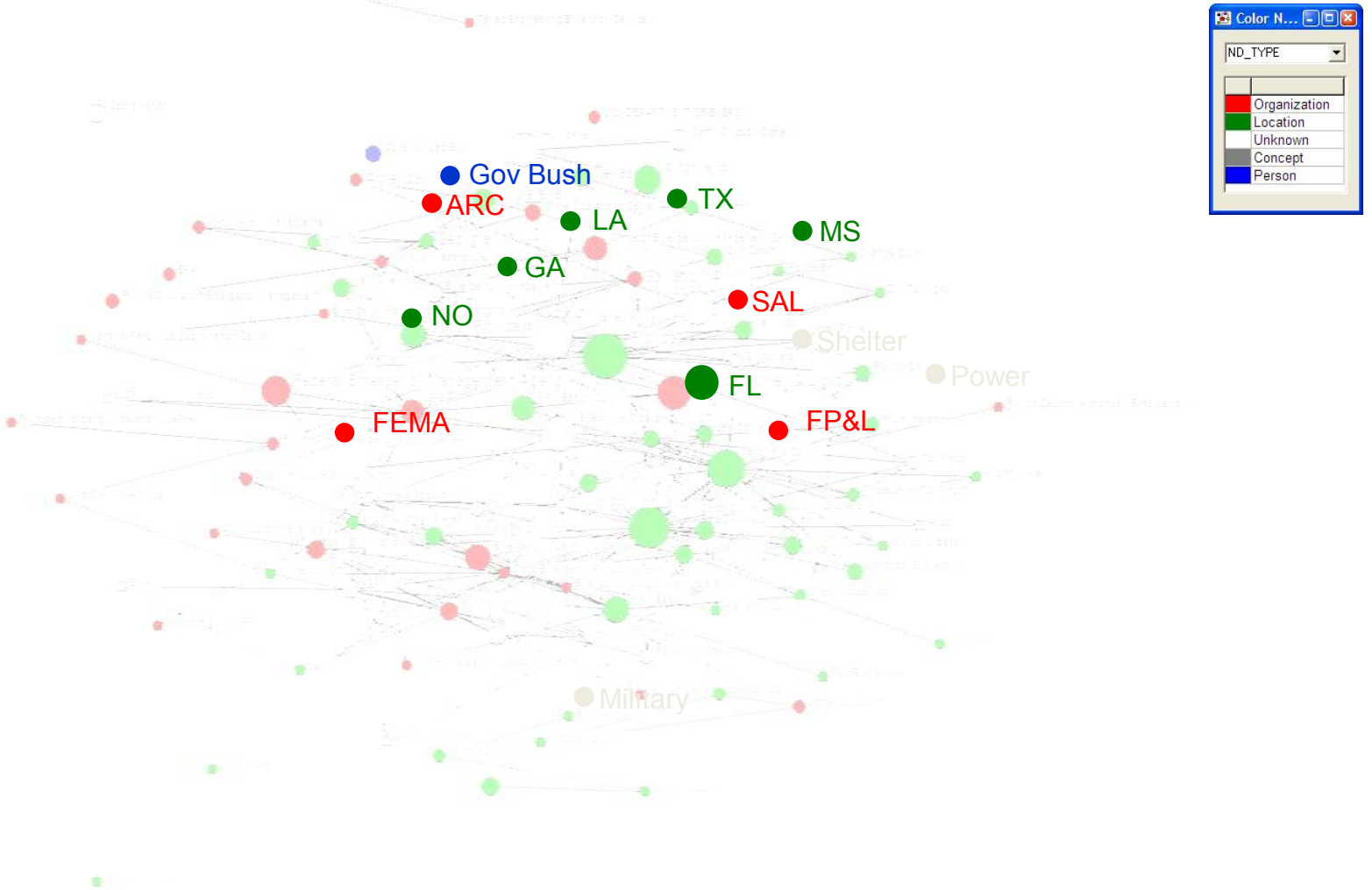
Time Slice 1 to 2



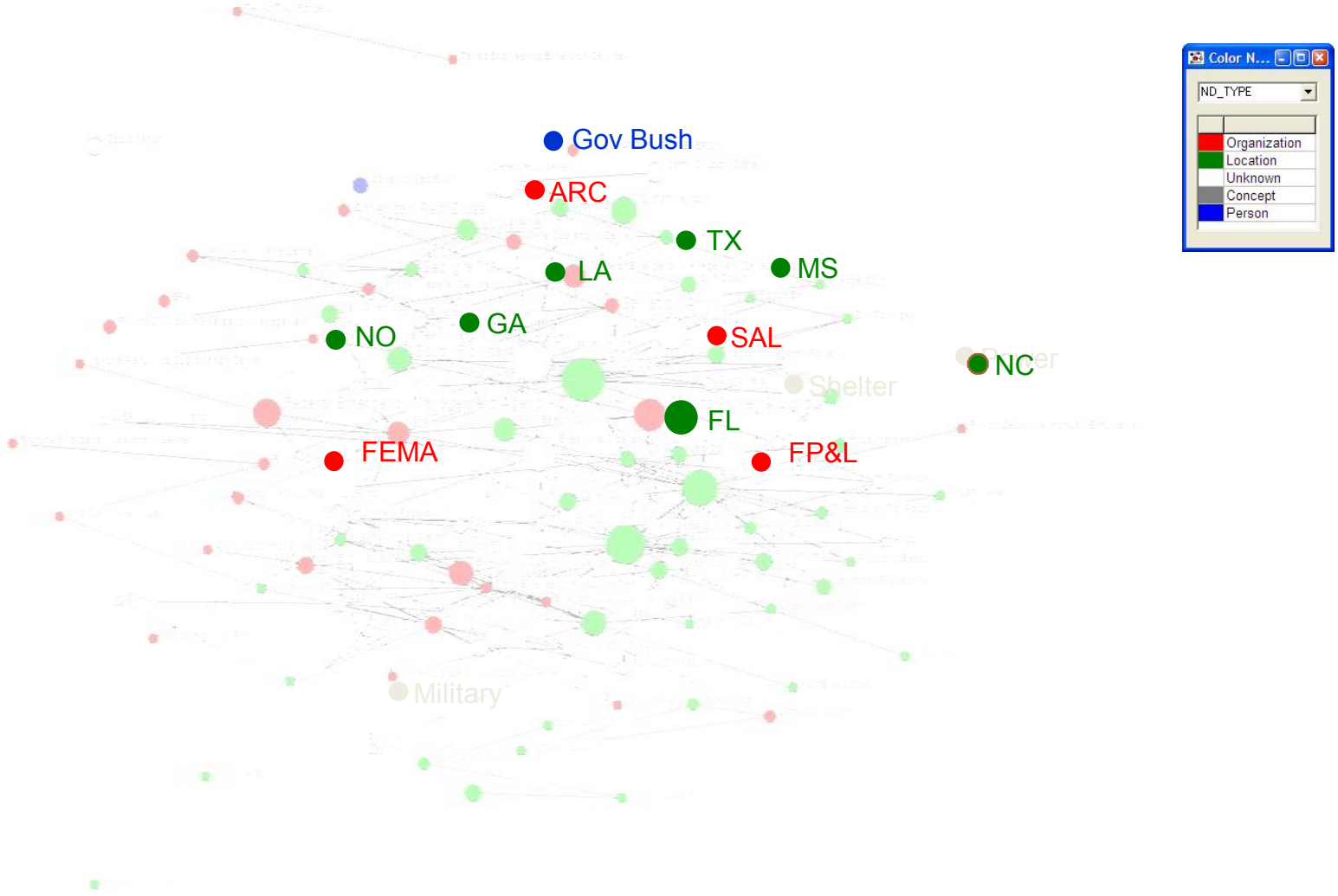
NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

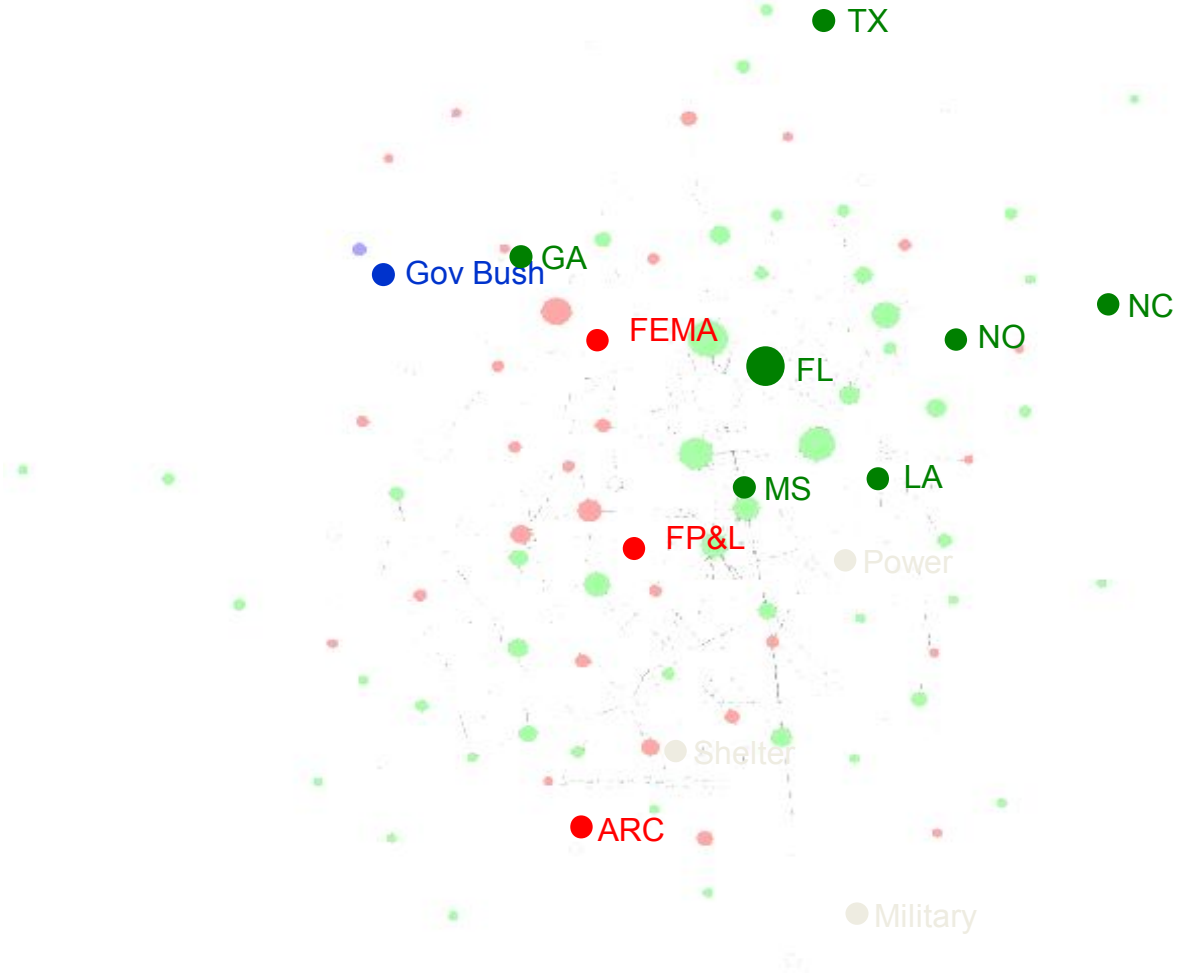
Time Slice 2: 8/26 to 8/27/2005

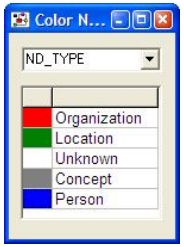


Time Slice 2 to 3

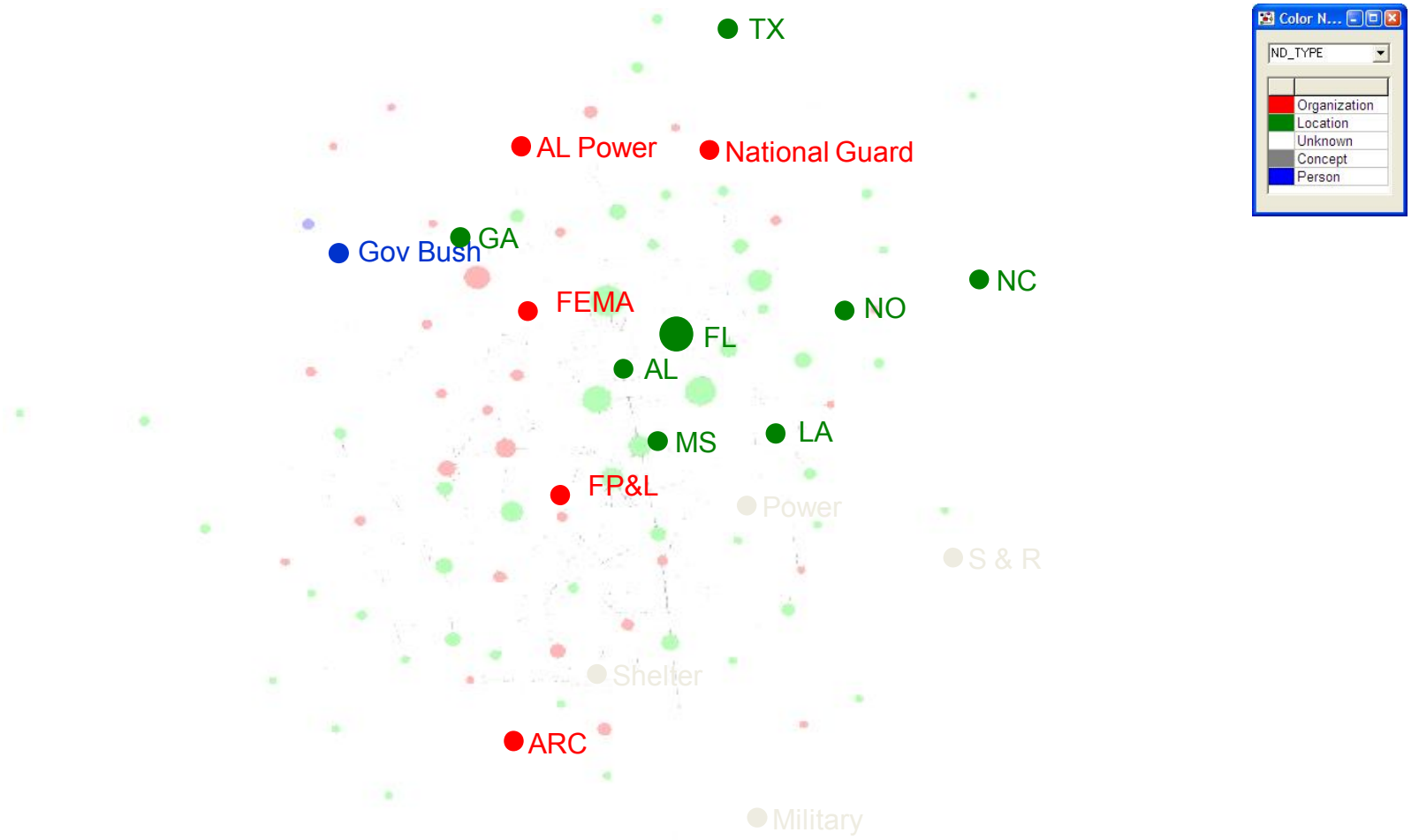


Time Slice 3: 8/28 to 8/29/2005





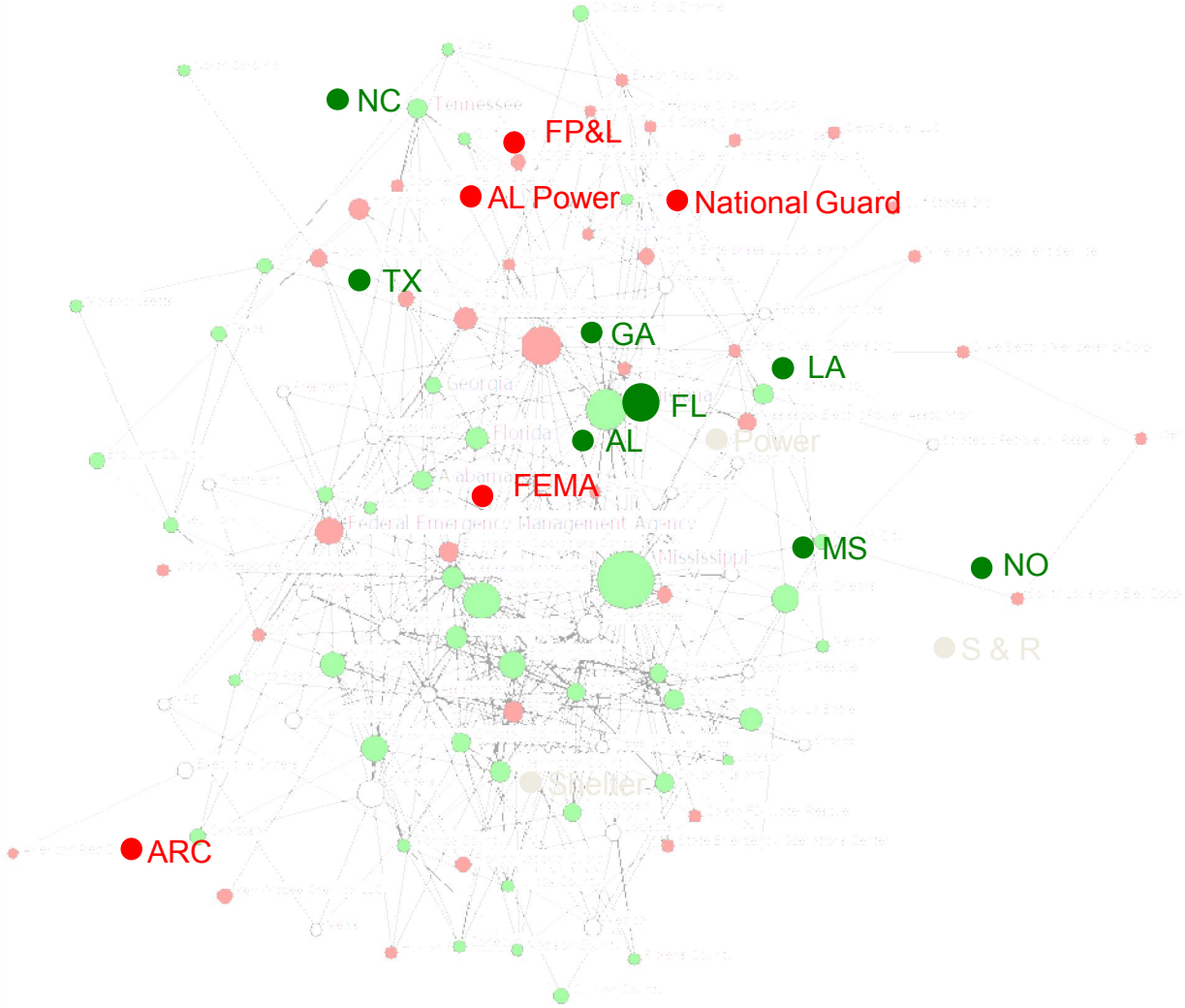
Time Slice 3 to 4



NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

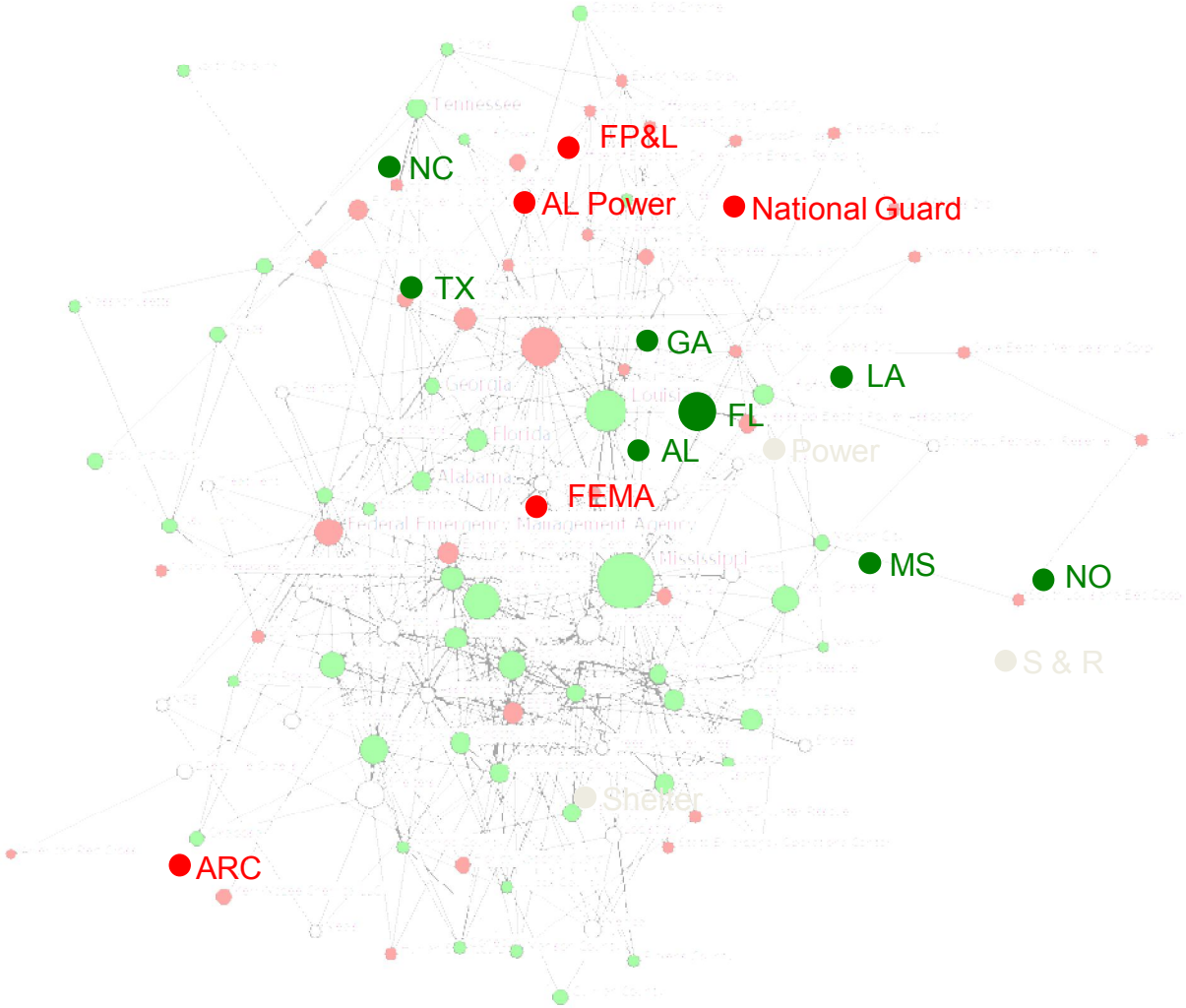
Time Slice 4: 8/30 to 8/31/2005



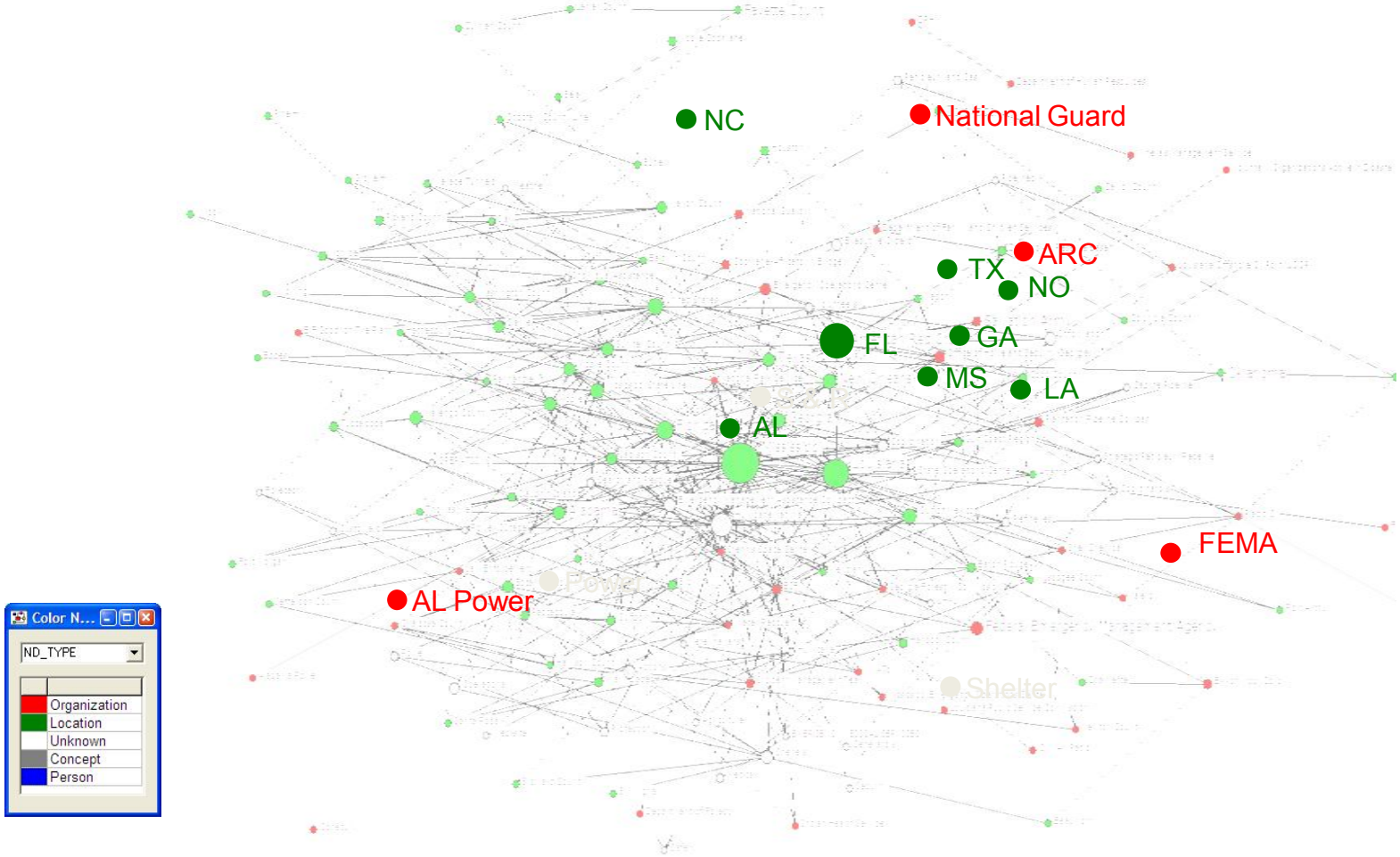
NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Time Slice 4 to 5



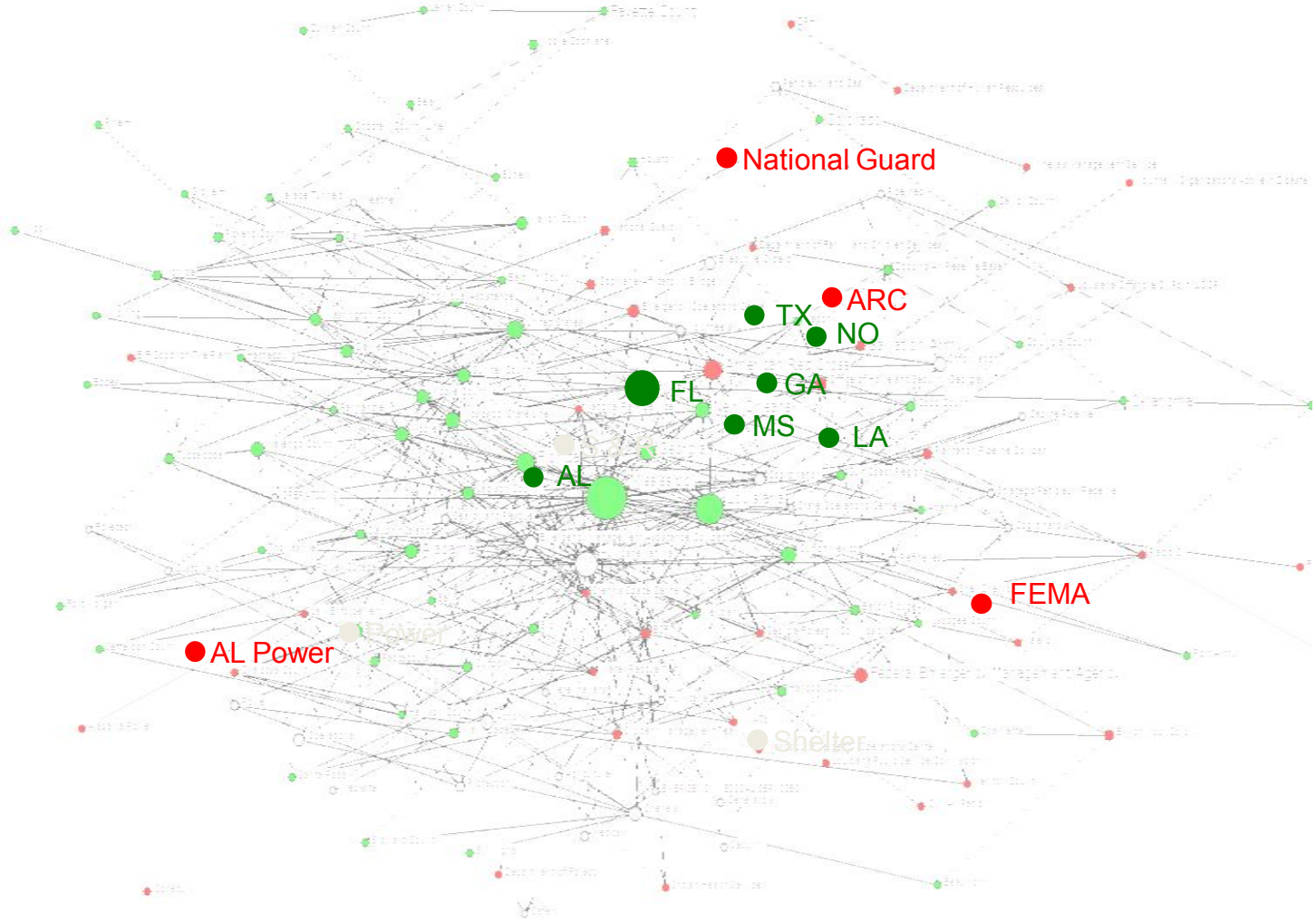
Time Slice 5: 9/1 to 9/2/2005

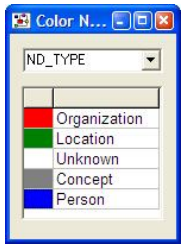


NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Time Slice 5 to 6

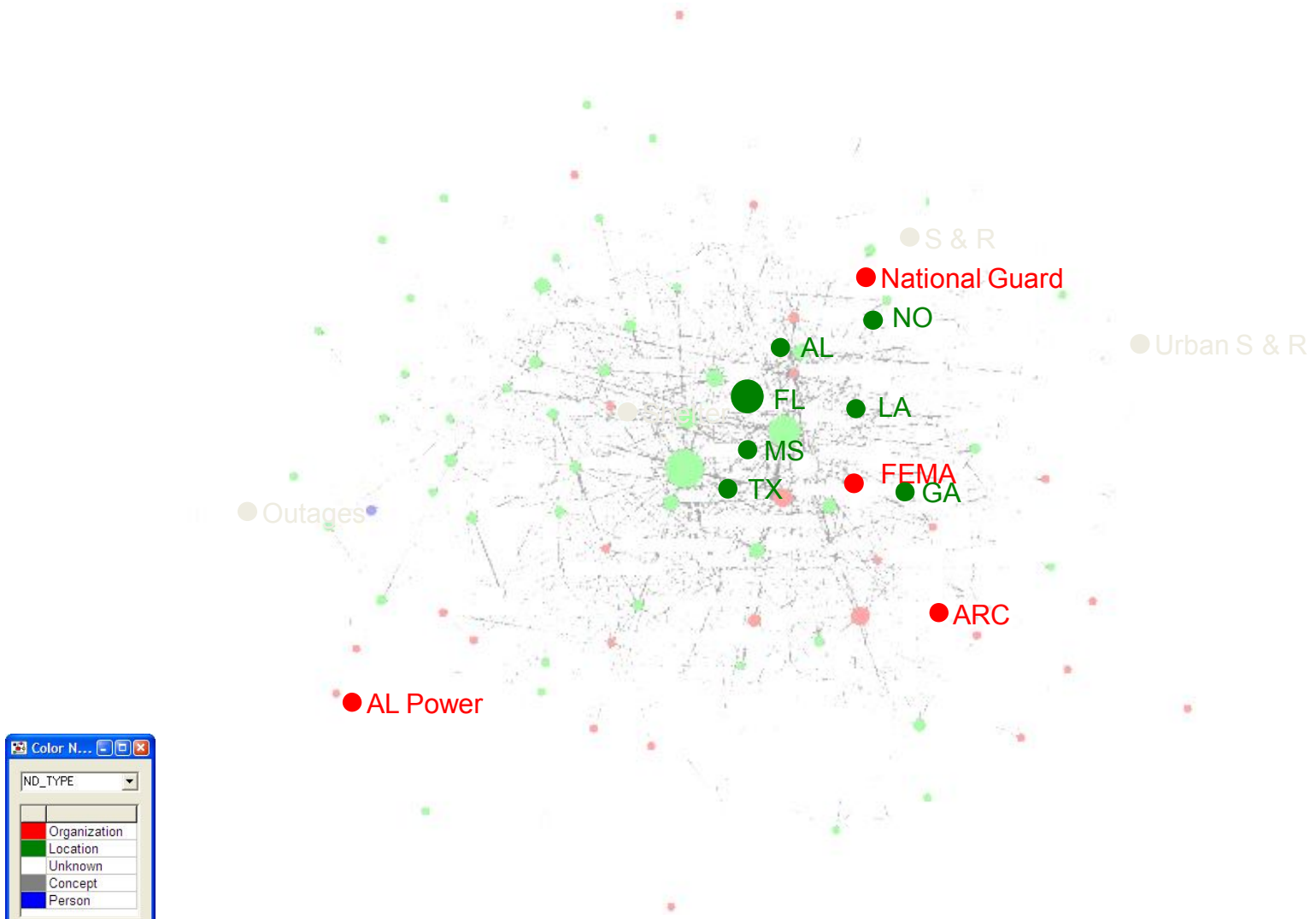




NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Time Slice 6: 9/3 to 9/4/2005

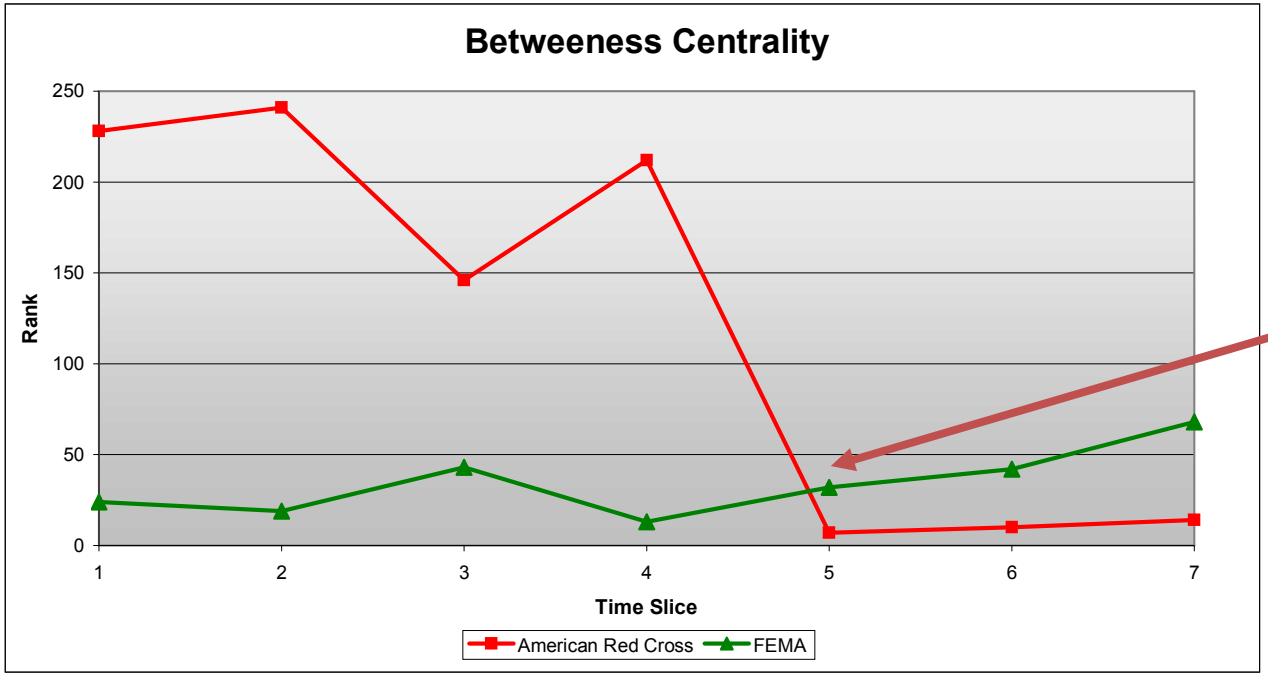


NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Change in Network Centrality Rankings

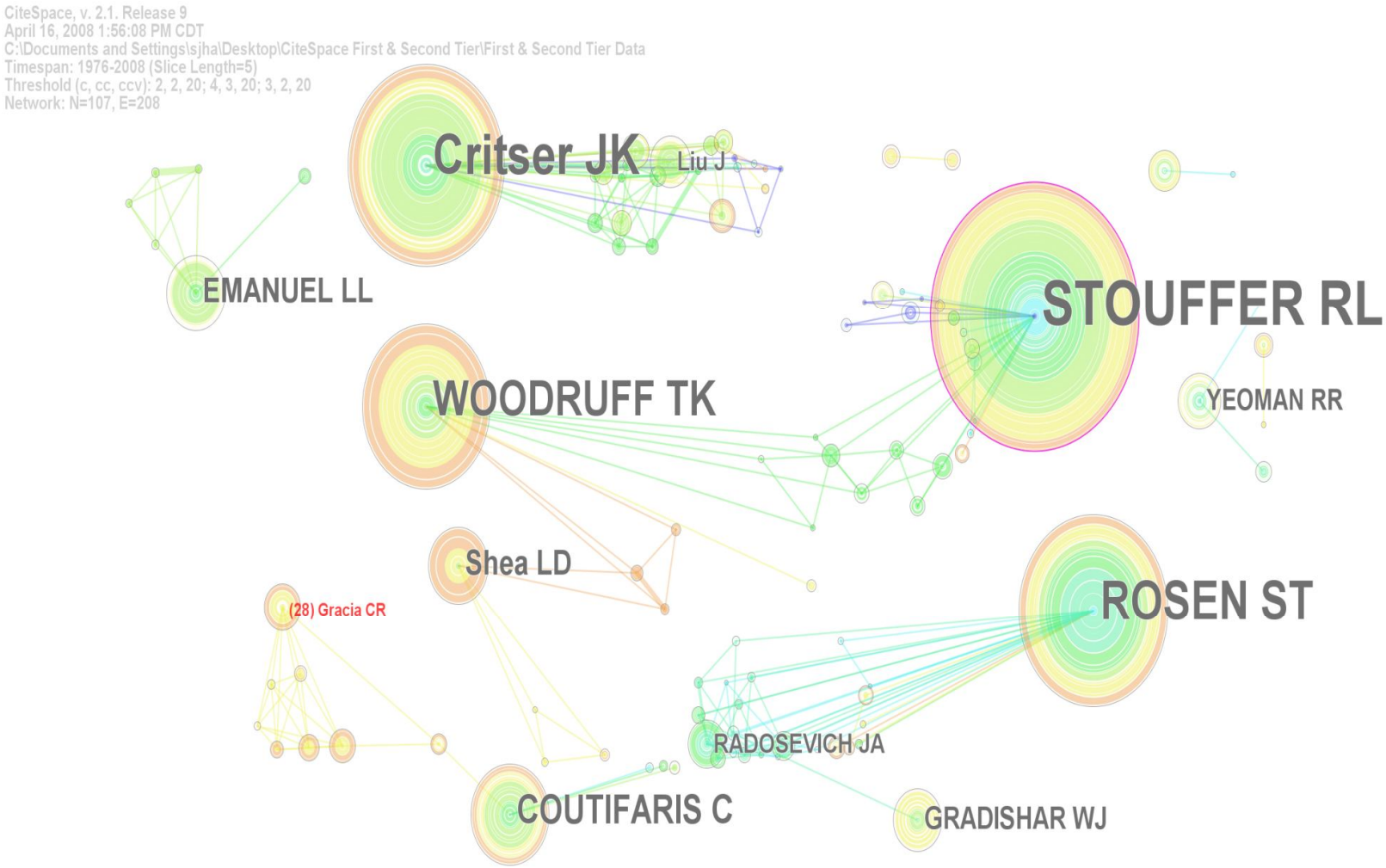
- **“American Red Cross” starts in the 200s and moves to the teens**
- **“FEMA” starts in the 20s, moves to the teens, and ends in the 60s**



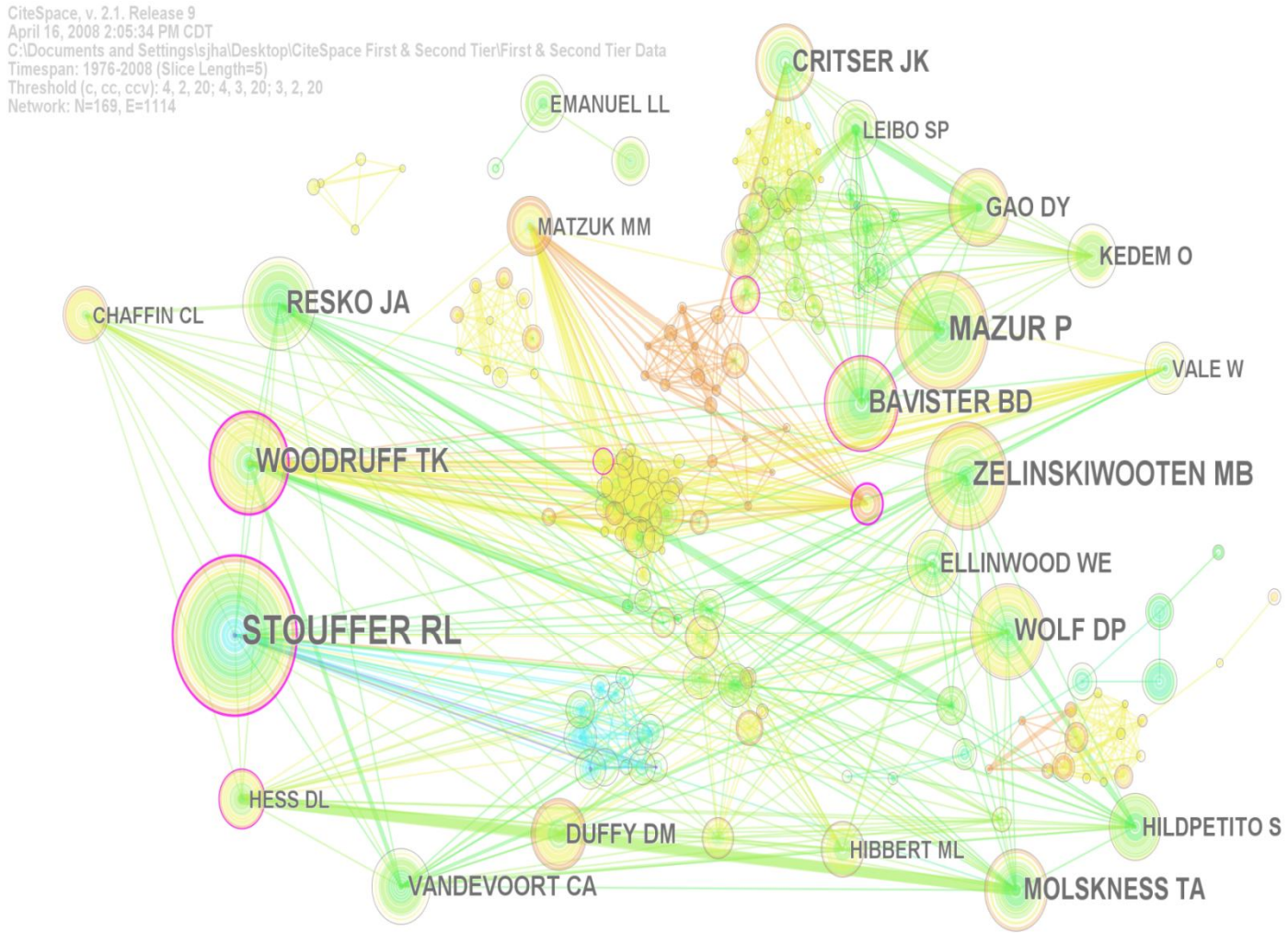
Crossover where American Red Cross becomes relatively more central than FEMA (Sep 1, 2005)

FEMA drops rank and American Red Cross moves up

Oncofertility Consortium Co-authorship Network



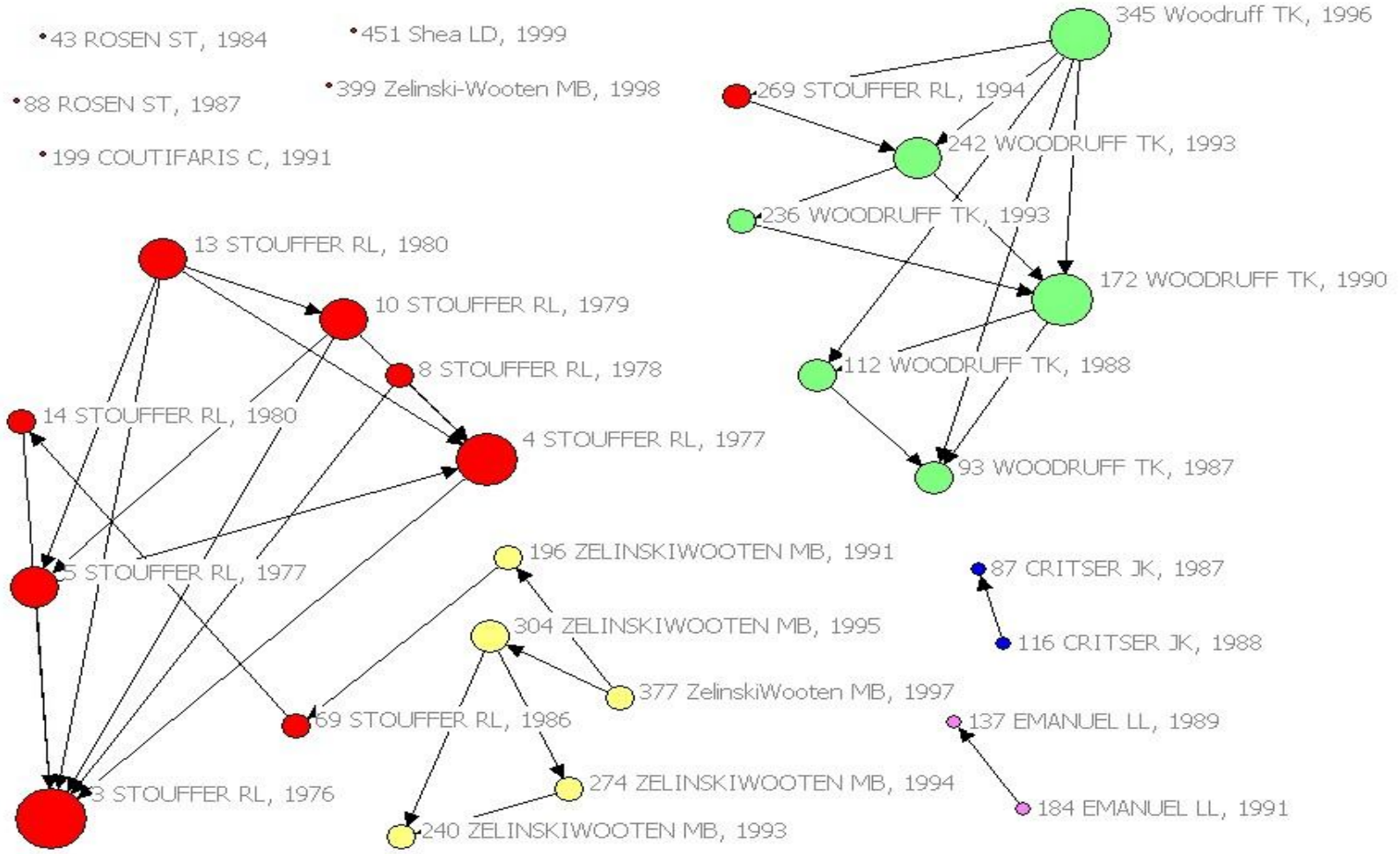
Oncofertility Consortium Author's Co-citation Network



NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

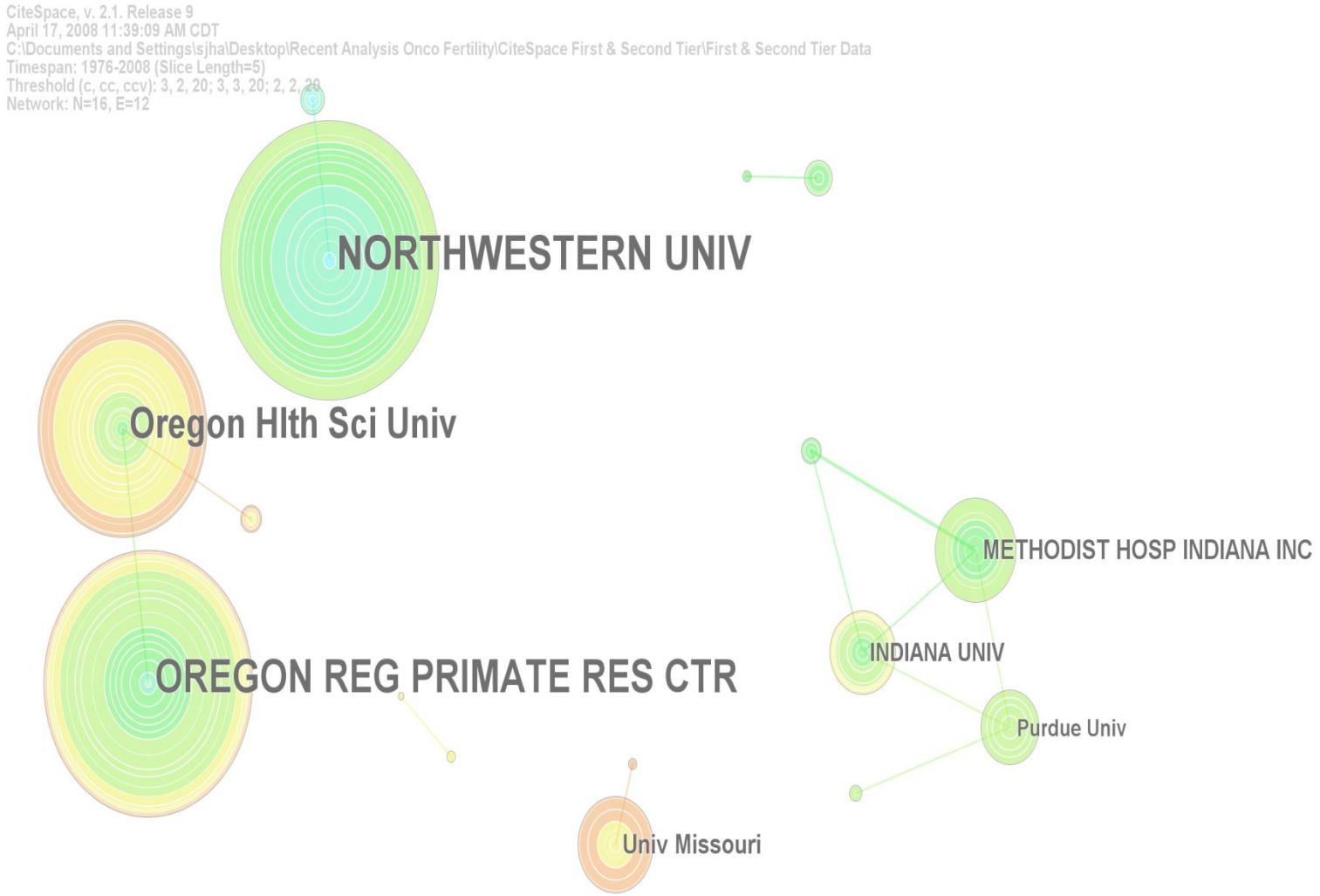
Oncofertility Consortium Citation Network



NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

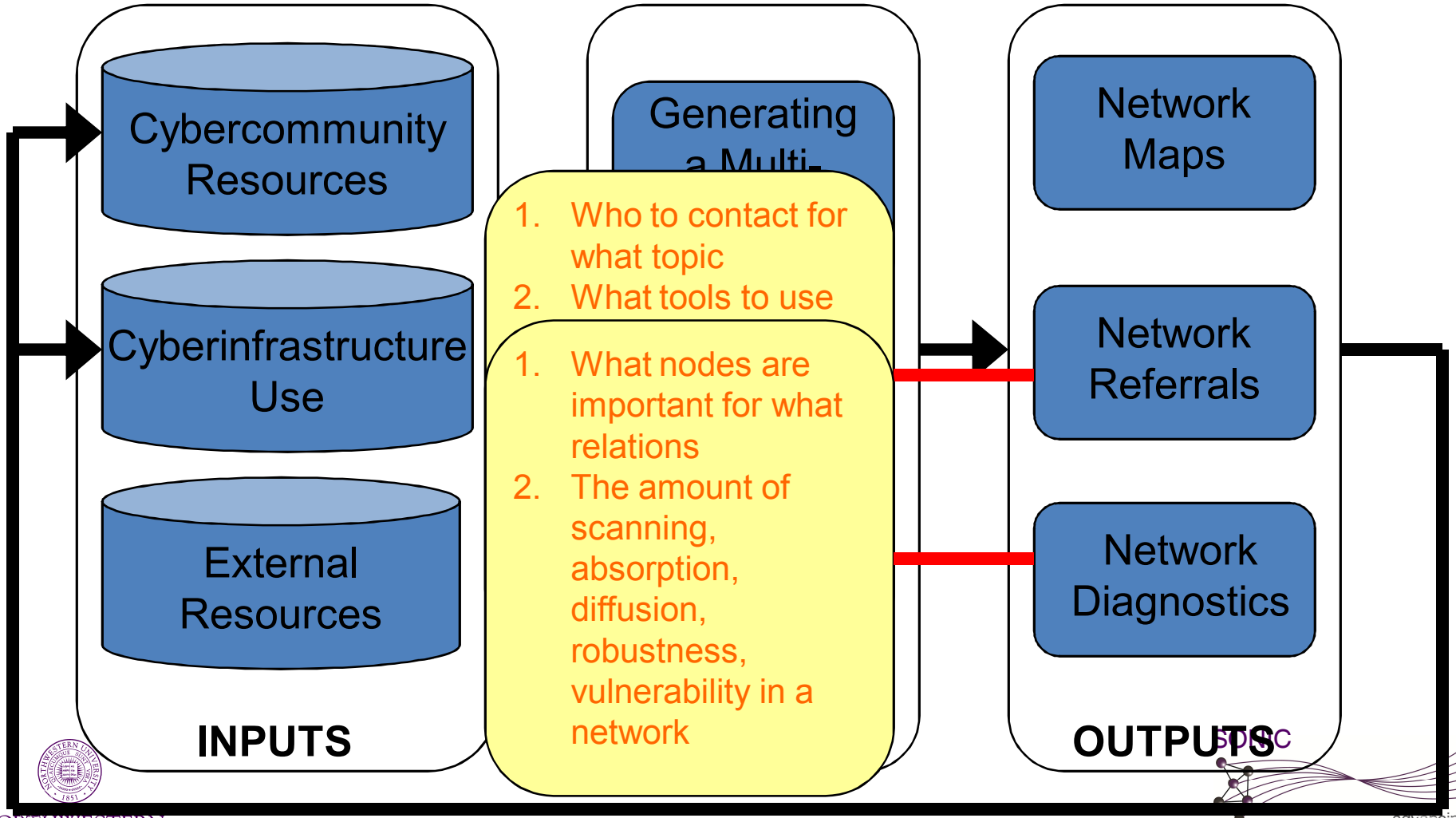
Oncofertility Consortium Co-author's Institutions Network



NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

CI-KNOW: Harvesting the online community's relational meta-data



Design Examples: Mapping & Enabling Networks in …

Tobacco Research: TobIG Demo

Computational Nanotechnology: nanoHUB Demo

Cyberinfrastructure: CI-Scope Demo

Oncofertility: Onco-IKNOW

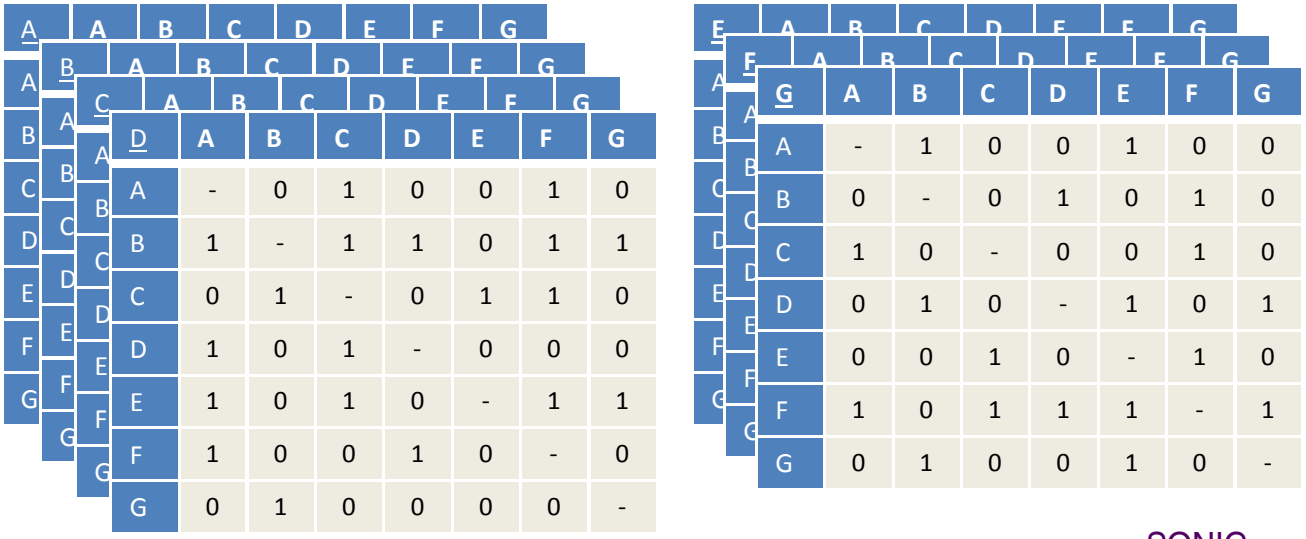
Social Structure Data

- The structure of any system is defined as a set of relational statements between all pairs of actors in the system
  - $R_{i,j}$ (*R*: structure-defining relation; i: "sender"; j: "receiver").
  - N-actor social structure: *NxN* matrix for each R relation.

	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	1	0
D	1	0	1	-	0	0	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

Cognitive Social Structure (CSS)

- Variation in the social perception of networks
  - $R_{i,j,k}$ (i: "sender"; j: "receiver"; k: "perceiver")
  - Cognitive Social Structure: *NxNxN* matrices for each R relation



D	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	1	0
D	1	0	1	-	0	0	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

G	A	B	C	D	E	F	G
A	-	1	0	0	1	0	0
B	0	-	0	1	0	1	0
C	1	0	-	0	0	1	0
D	0	1	0	-	1	0	1
E	0	0	1	0	-	1	0
F	1	0	1	1	1	-	1
G	0	1	0	0	1	0	-

Two kinds of reductions

- Locally-aggregated structures
  - **Row**– self-reports of which *i* actors go to which *j* actors $(R'_{i,j} = R_{i,j,i})$
  - **Column** – self-reports of which *j* actors come to which *i* actors $(R'_{i,j} = R_{i,j,j})$
  - **Intersection** – *i* and *j* both agree a tie exists $(R'_{i,j} = \{R_{i,j,i} \cap R_{i,j,j}\})$
  - **Union** – how many people think a tie exists $(R'_{i,j} = \{R_{i,j,i} \cup R_{i,j,k}\})$
- Consensus structures
  - Tie exists if threshold of everyone else agrees it should exist
  - Tie exists EVEN IF actors report it does not exist
  - $R'\ i,j = f\ (R_{i,j,k1}, R_{i,j,k2}, \ldots, R_{i,j,kn})$

Row LAS

- Take each self-reported *i* row out of its matrix, and stitch together into a new matrix

A	A
A	-
B	1
C	0
D	1
E	1
F	1
G	0

B	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	1						
E	1						
F	1						
G	0						

C	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	1						
E	1						
F	1						
G	0						

D	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	1	0
D	1	0	1	-	0	0	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

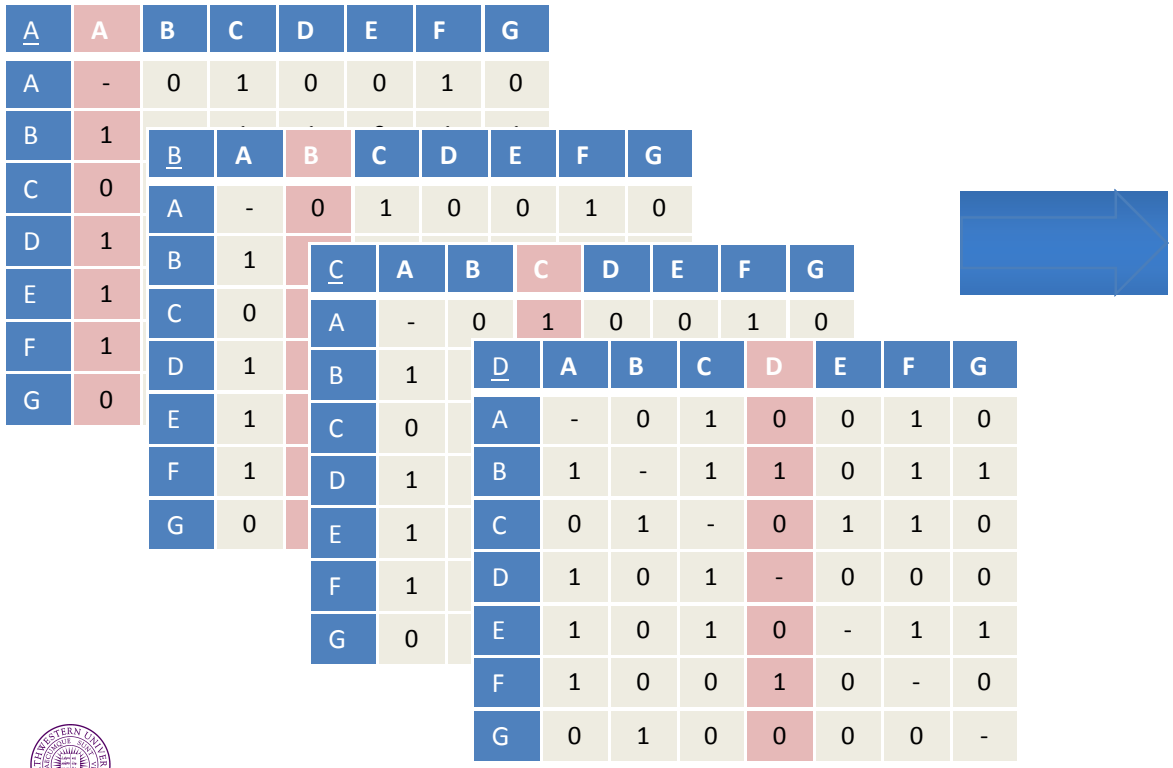
	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	1	0
D	1	0	1	-	0	0	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Column LAS

- Take each self-reported *j* column out of its matrix and stitch together into a new matrix



A	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	1						
E	1						
F	1						
G	0						

B	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	1						
E	1						
F	1						
G	0						

C	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	1						
E	1						
F	1						
G	0						

D	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	1	0
D	1	0	1	-	0	0	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	1	0
D	1	0	1	-	0	0	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Intersection LAS

- *i* and *j* both agree a tie exists, doesn't matter what others say

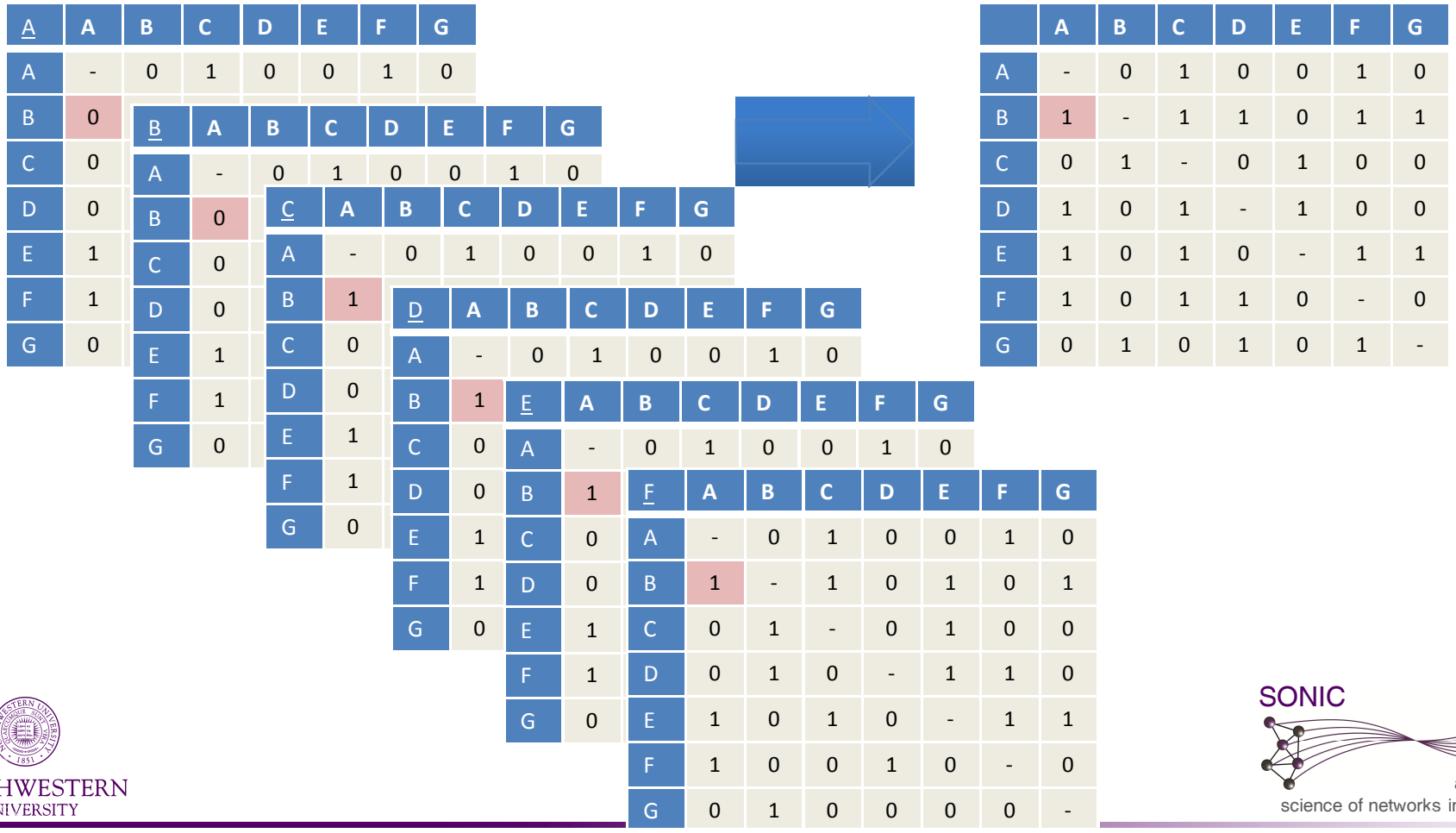
A	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1					
D	1	0					
E	1	0					
F	1	0					
G	0	1					

B	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	0	1	0	1
C	0	1	-	0	1	0	0
D	0	1	0	-	1	1	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	0	0
D	1	0	1	-	1	0	0
E	1	0	1	0	-	1	1
F	1	0	1	1	0	-	0
G	0	1	0	1	0	1	-

Consensus

- A threshold of other people agree a tie exists



A	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	0						
C	0						
D	0						
E	1						
F	1						
G	0						

B	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	0						
C	0						
D	0						
E	1						
F	1						
G	0						

C	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	0						
E	1						
F	1						
G	0						

D	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	0						
E	1						
F	1						
G	0						

E	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1						
C	0						
D	0						
E	1						
F	1						
G	0						

F	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	0	1	0	1
C	0	1	-	0	1	0	0
D	0	1	0	-	1	1	0
E	1	0	1	0	-	1	1
F	1	0	0	1	0	-	0
G	0	1	0	0	0	0	-

	A	B	C	D	E	F	G
A	-	0	1	0	0	1	0
B	1	-	1	1	0	1	1
C	0	1	-	0	1	0	0
D	1	0	1	-	1	0	0
E	1	0	1	0	-	1	1
F	1	0	1	1	0	-	0
G	0	1	0	1	0	1	-

CSS in Research Design

- Which is the “most” accurate network?
- Do central people have more accurate perceptions of the network?
- Do people in similar positions (equivalence) have similar perceptions?
- Are there ties believed to exist which don’t exist and vise versa?