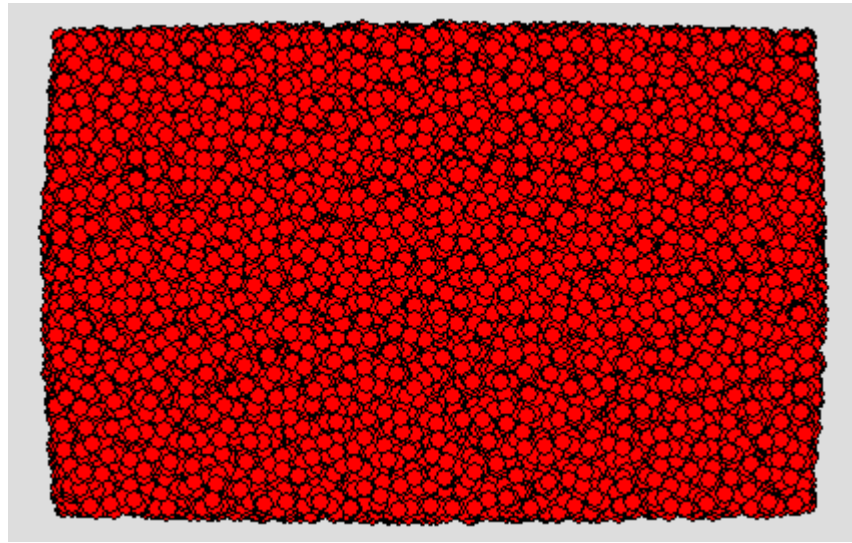


Automated statistical content analysis for identifying semantic and actor networks

James A. Danowski
Communication
University of Illinois at Chicago
jimd@uic.edu



Background

Who-to-Whom Networks

- Chase Manhattan Bank
- Office of Defense-Civil Preparedness
- Korean
- Retirement Communities





CBS concept cooccurrence across
message pairs in discussion list

Approach to Semantic Network Analysis

- Identify word cooccurrences based on proximity model.
- Not “bag of words” but word pairs within sliding window.
- Word bigrams +/- 3 words on either side of each word in text
- Linkage model not distance model, i.e. network not space

Assumptions

- Useful to have same analytical model for text, individual social actors, and other social units of analysis such as departments, organizations, etc.
- Consistent with cognitive psychology models.
- Message content is related to social network structures over time.

Examples of Questions

- Do changes in message content usually precede or follow changes in who-to-whom networks?
- What features of the message network predict changes in the social network?
- Are media representations of social networks predictive of audience perceptions?
- Does the medium of communication foster differences in message content networks?
- Does text file t1 significantly more or less relative occurrence of predicted message content than text file t2?

Sources of Data

- Open-ended survey responses
- News stories
- Blogs
- Discussion lists
- Email
- Combinations (survey X news stories)

Open-ended Survey Responses

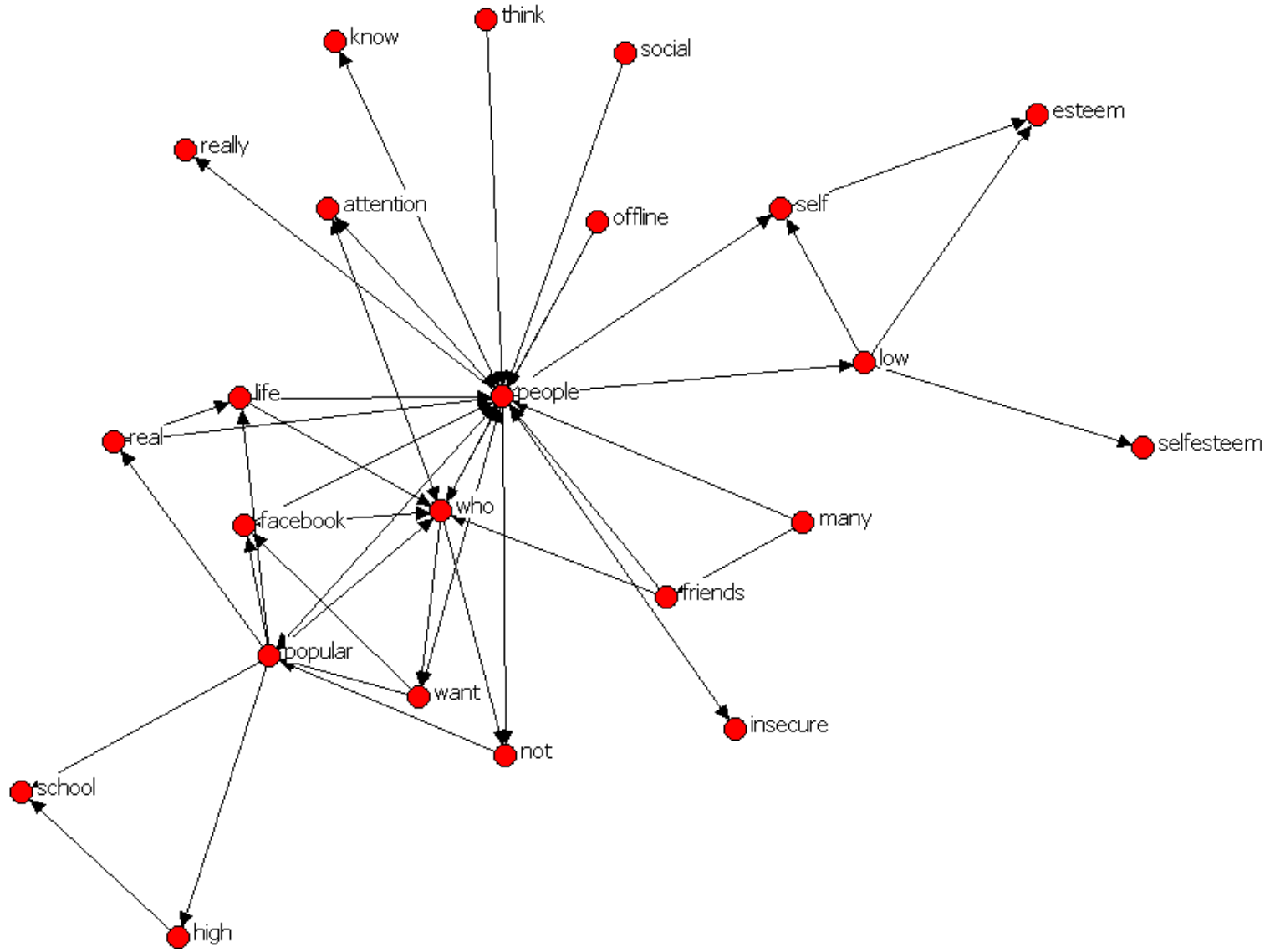
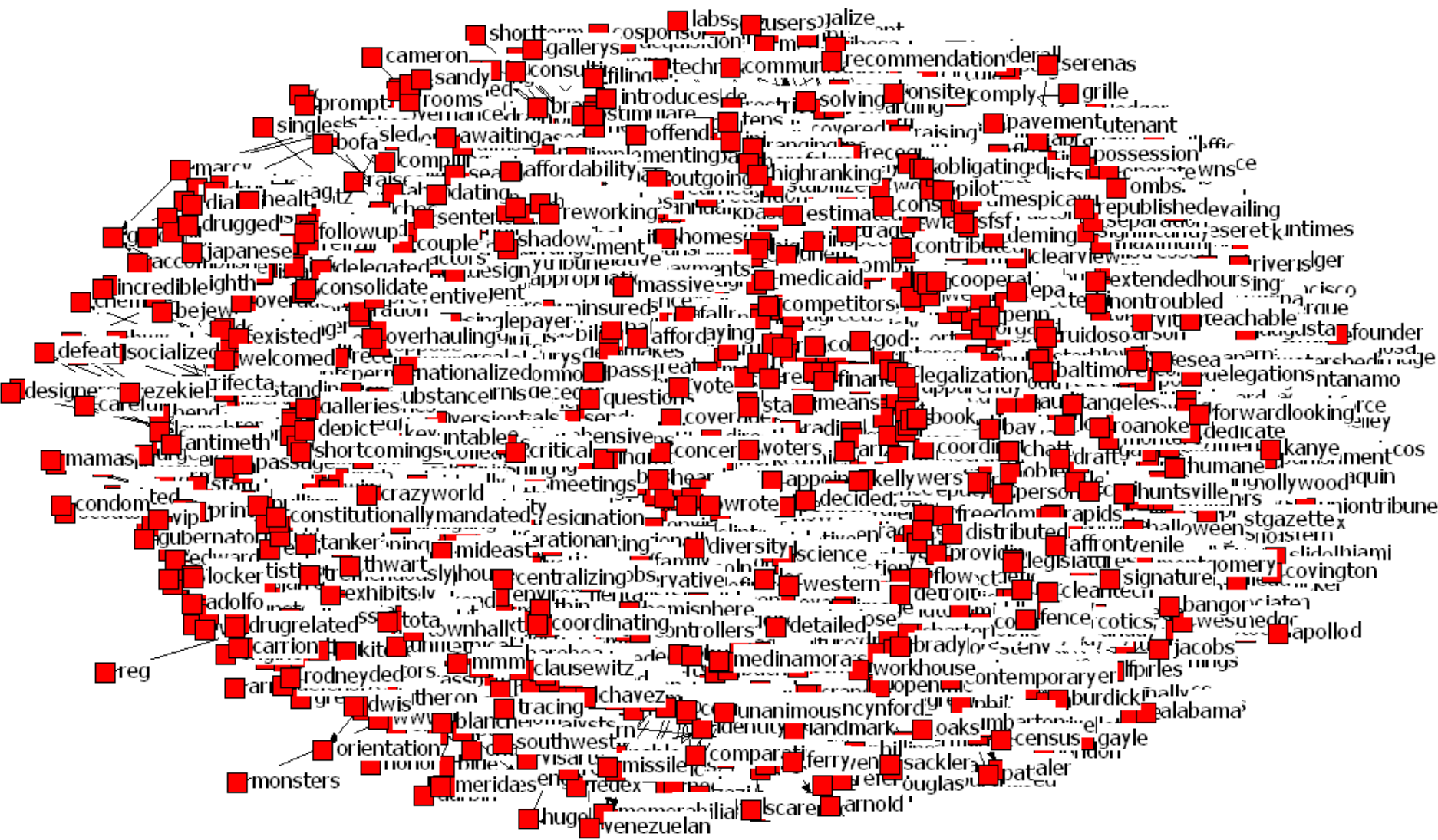


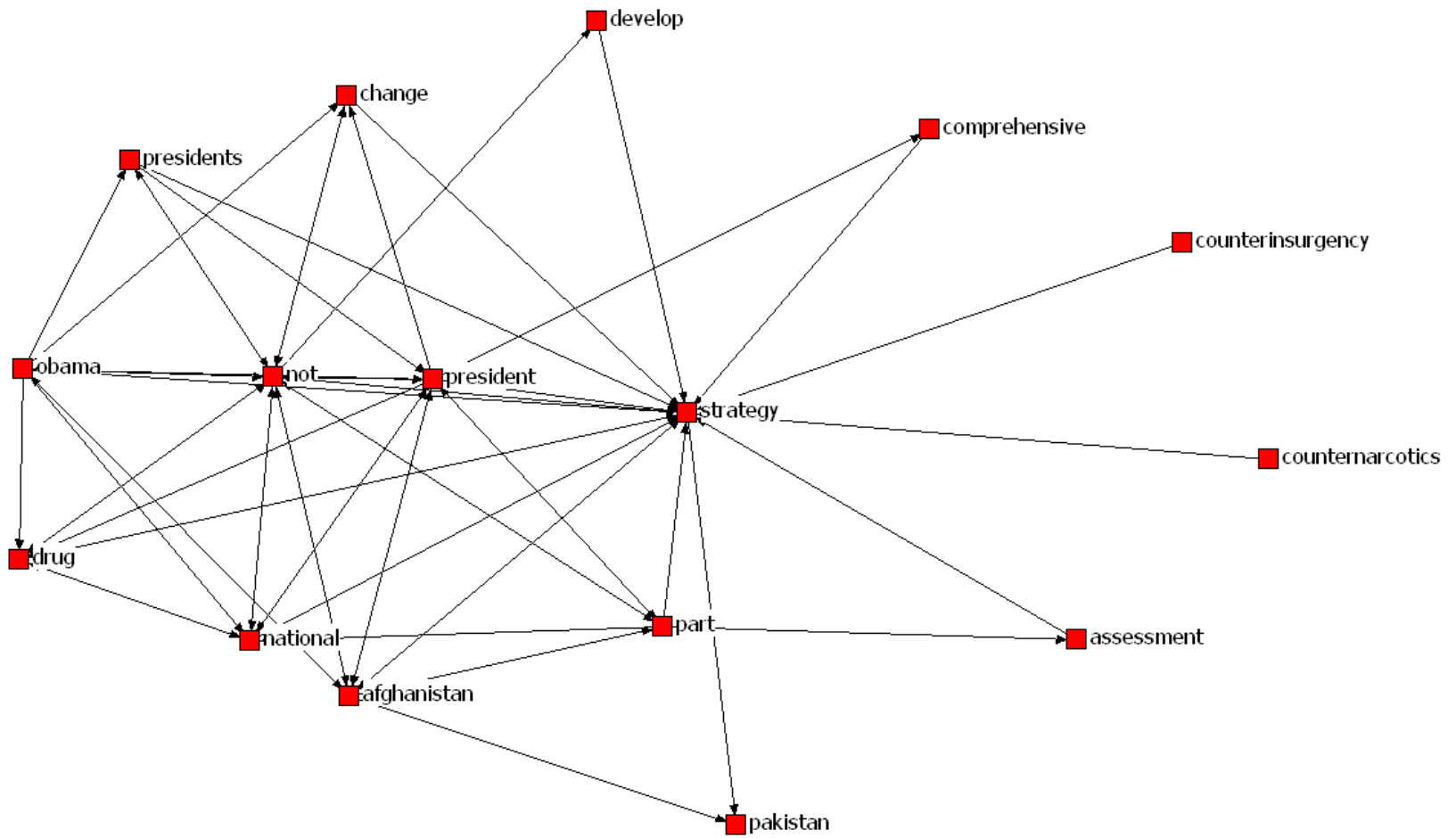
Figure 3. What types of people want to be popular on Facebook™ ? Word network of responses from unpopular participants.

Online reactions to political event

Node-centric (NodeTric) zoom ins



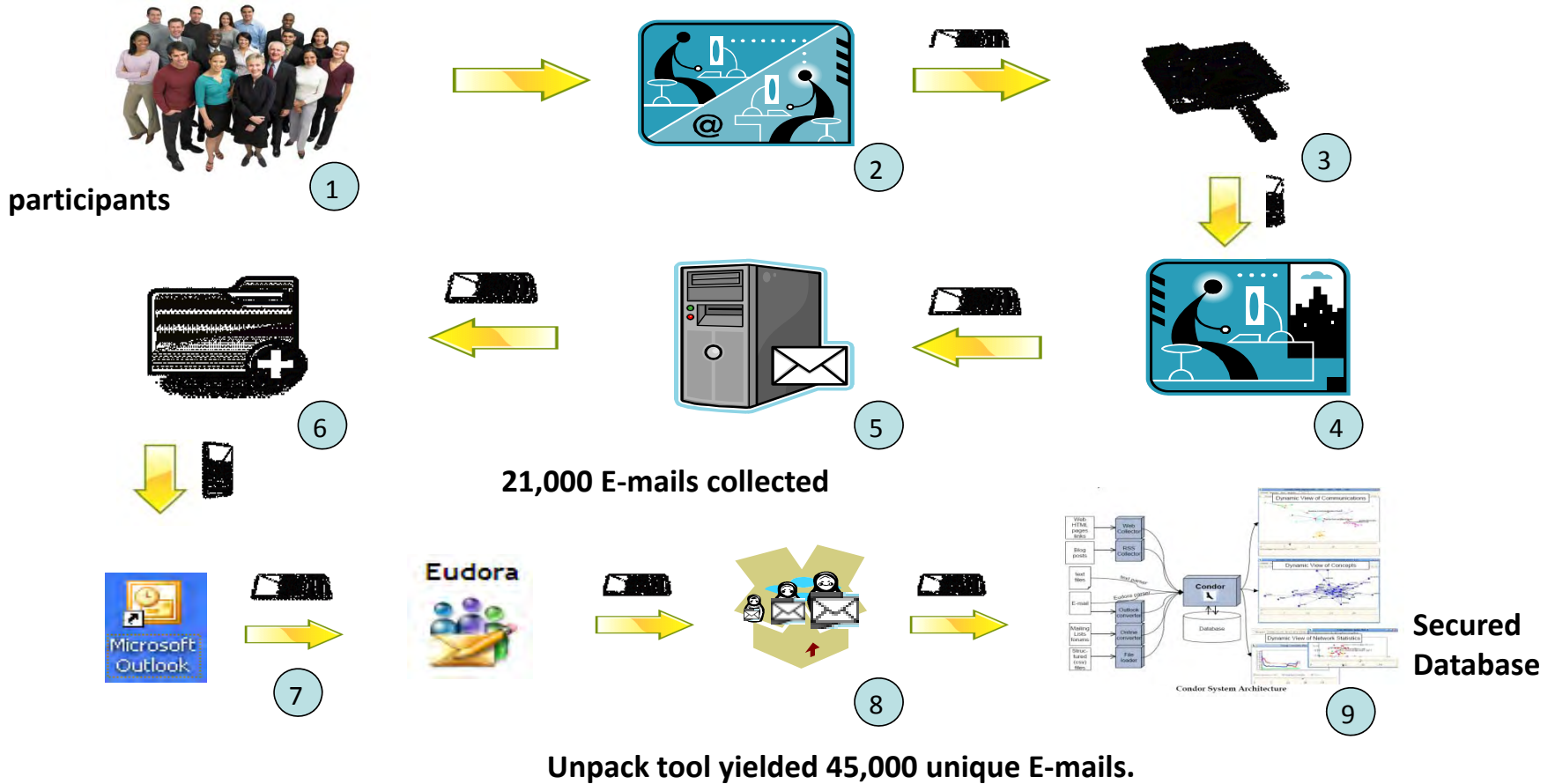
Node-Centric “Czar” Network (frequency GE 10) within “Czar” 2,403 News Stories



Node-Centric “Counterinsurgency” selection within “Czar” Network

Email content and who-to-whom
networks over time in organization

IT-based E-mail data collection Process Flow





Email Extraction

- Target 298 people for diffusion study
- Participants opt in & install Outlook rules
- Process all inbound /outbound emails with Outlook rules
- Proxy account receives 21,000 filtered emails
- Store on secure server
- Remove encrypted or secret email
- Convert Outlook email to Eudora .mbx
- Unpack email: remove forwarded headers & segment nested emails yielding 45,000 emails
- Network analyze who2whom links and word pairs

WORDij Software

- Input large volumes of text
- Index word pair frequencies
- Map networks of words
- Main Options:
 - Stop (drop) lists
 - Stemming
 - String replacement
 - Select text slices



Source Text File

Browse...

Drop List File

Browse...

Drop words appearing less often than

Drop pairs appearing less often than

Use linkage strength method

Window size for extracting word pairs

Source File:
text file in
UTF-8 format.

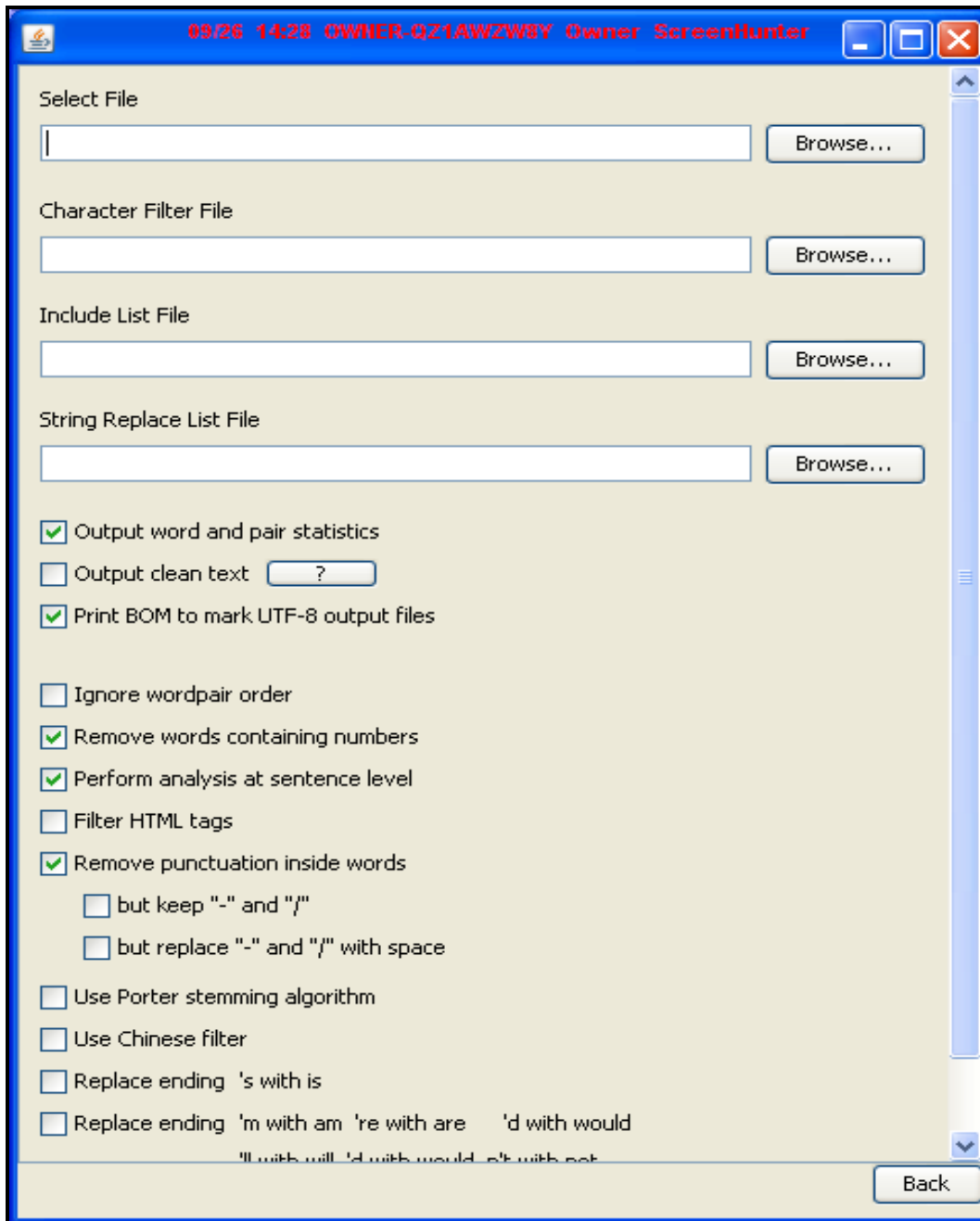
Drop List File:
file with a list
of words that
will be dropped.

Drop words /
pairs appearing
less often than:
words / pairs
appearing less
often will be not
included in the
output files.

Advanced Options

Quit

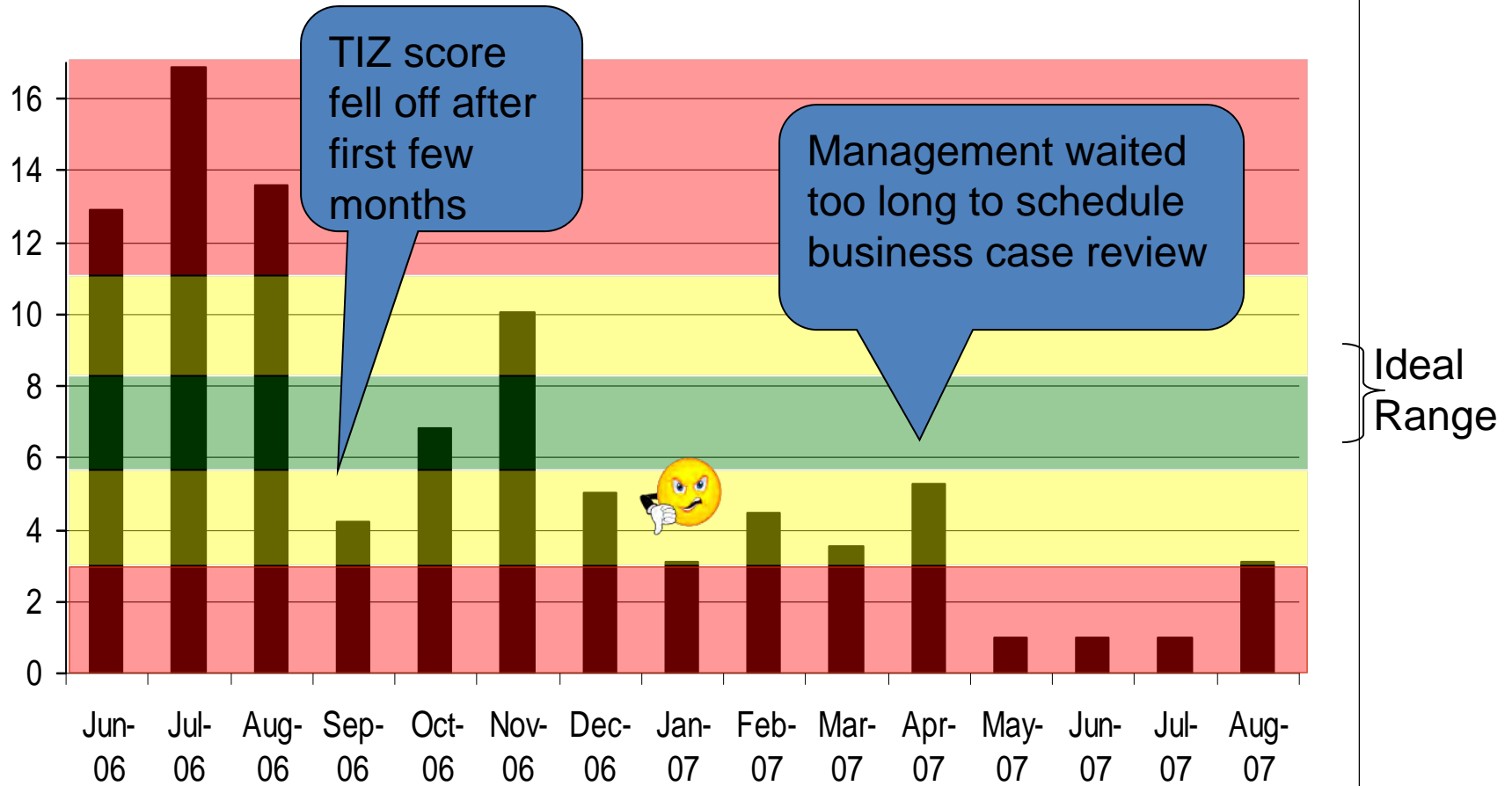
Analyze Now



Positivity Index: Losada Line

Ratio of Positive / Negative Emotion Score for TIZ

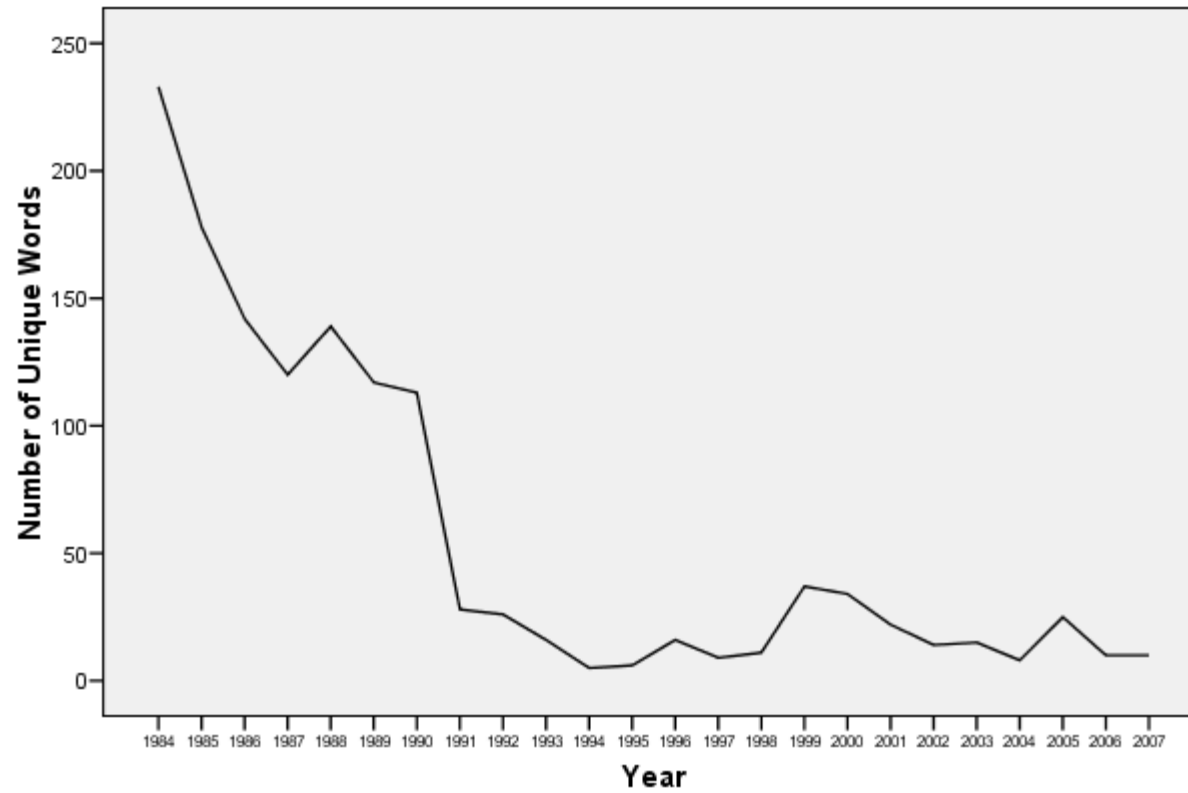
Positive/Negative Ratio



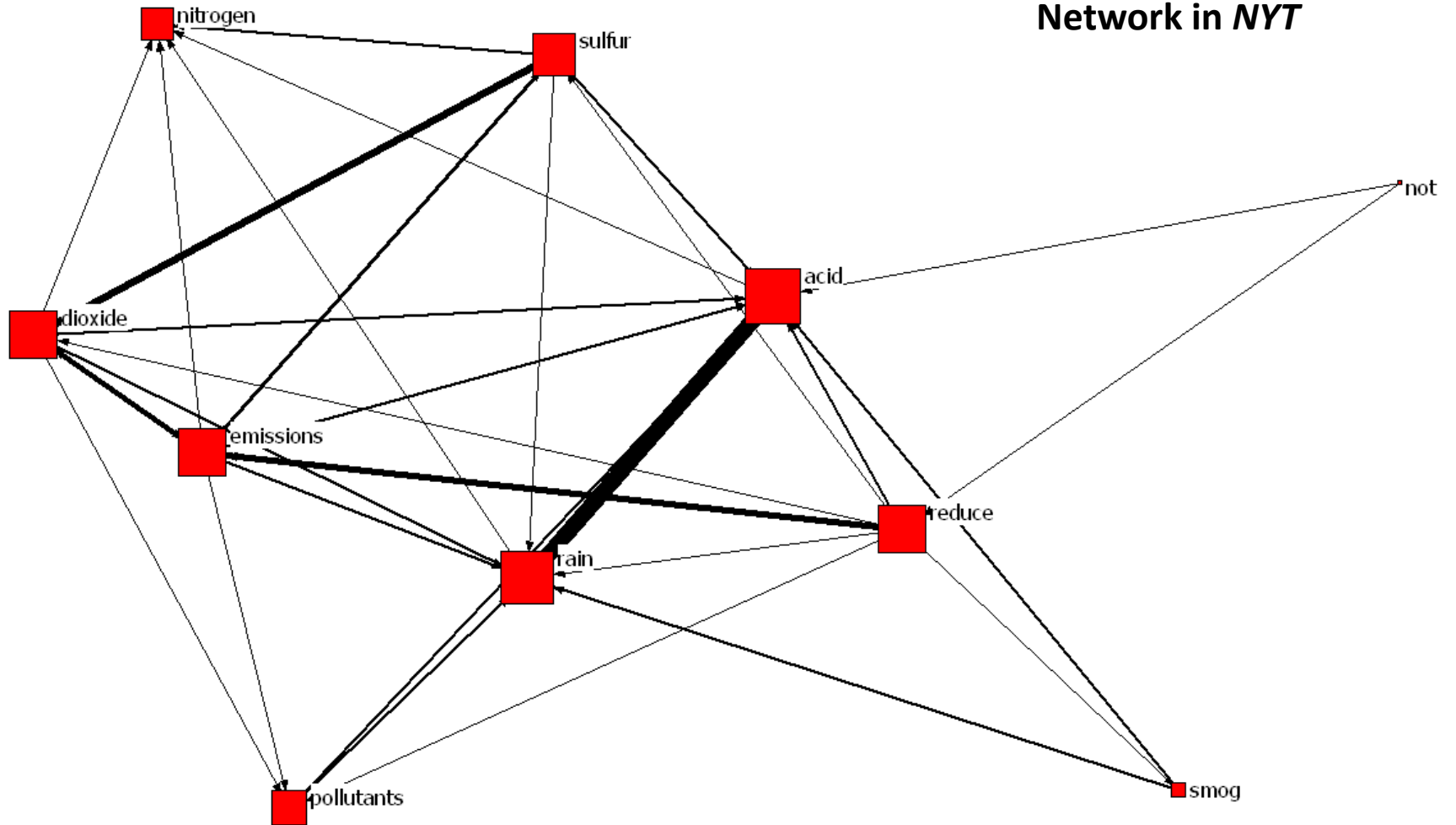
Media coverage over time

Similarity of networks

Unique Words Within 5 Links of "Acid Rain:" 1984-2007



2007 Acid Rain Network in NYT



Proper noun extraction, string
replacement

- Here are screenshots of a sample Proper Nouns run and sample output.

The screenshot shows a Windows desktop environment. In the background, the WORDij software interface is visible, with a 'Log' window overlaid on top. The 'Log' window displays the following text:

```

Input file: C:\Documents and Settings\Administrator\My Documents\Danowski\Wordij_Documentation\Twitter\twit02062009.TXT
Output list file: C:\Documents and Settings\Administrator\My Documents\Danowski\Wordij_Documentation\Twitter\2009 twitter proper nouns
Output replacement file: C:\Documents and Settings\Administrator\My Documents\Danowski\Wordij_Documentation\Twitter\2009twitter string file
  
```

In the foreground, two Notepad windows are open. The left window, titled '2009 twitter proper nouns.txt - Notepad', shows a list of proper nouns with a 'Name' column on the left. The right window, titled '2009twitter string file.txt - Notepad', shows a list of string replacements. To the right of the string file window, the words 'ent' and 'ent' are visible, likely representing the output of the string replacement process.

| Name | Proper Nouns | String File Output |
|------|-------------------|--------------------------------------|
| | A Bibliography | [Tuesday->Tuesday |
| | A Feel Good Guide | Industry->Industry |
| | A Team | Twitter->Twitter |
| | A***e | Rebecca Heslin->Rebecca_Heslin |
| 2009 | A-list | Pg->Pg |
| | A-listers | Krums->Krums |
| 2009 | A16 | Sarasota->Sarasota |
| | A38 | Fla->Fla |
| | ALZselftest | Hudson River->Hudson_River |
| | ALZselftest.com | River->Hudson_River |
| | ANC's | Jan->Jan |
| | AOL's | Airways Flight->Airways_Flight |
| | ASmallWorld | Flight->Airways_Flight |
| | Aaron C | Manhattan->Midtown_Manhattan |
| | Abbott | Midtown Manhattan->Midtown_Manhattan |
| | Abby Kohn | |

Automatic Time-Series Mapping of Social Networks of Political Actors From Large Collections of News Stories

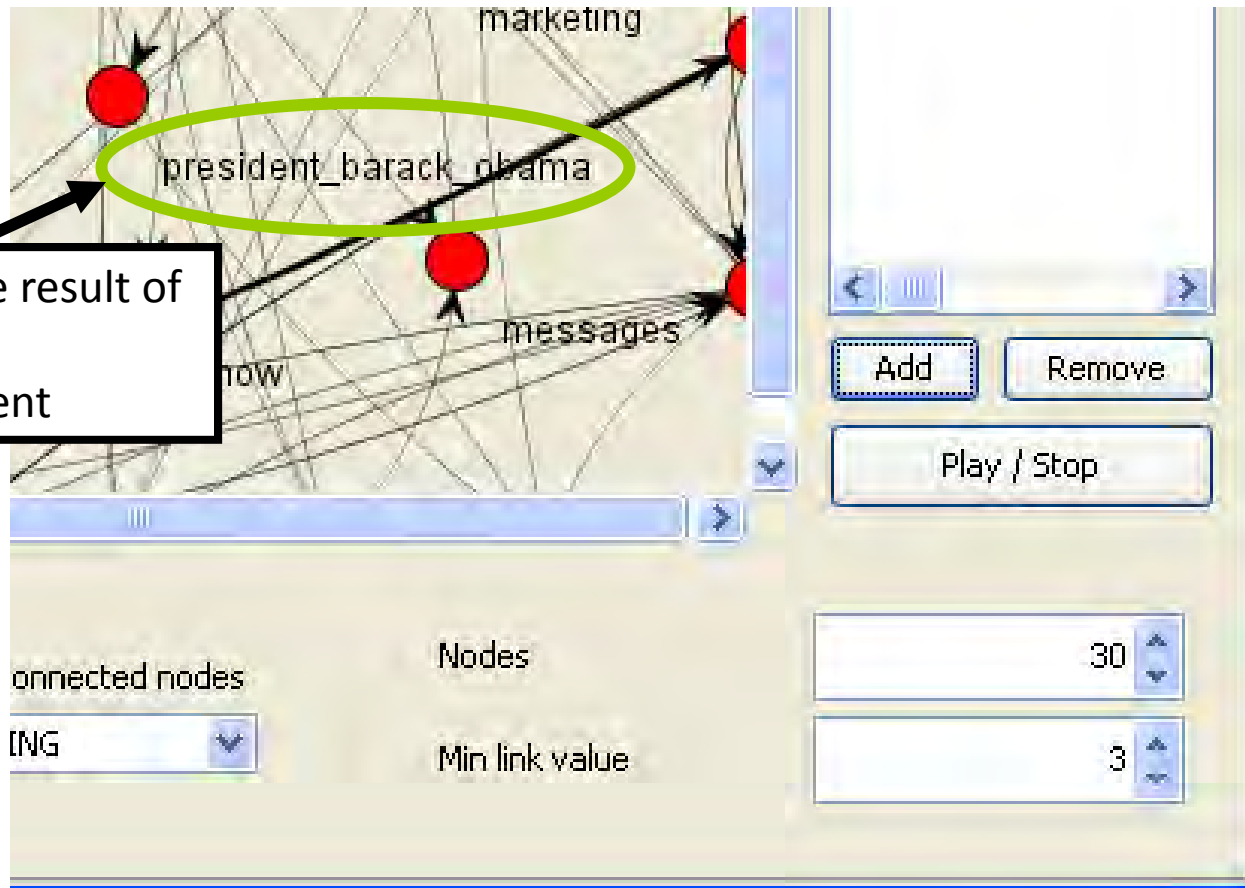
- Test hypotheses about presidential cabinet network centrality and presidential job approval over time.
- Illustrate automatic social network identification from large volumes of text.
- Mined the social networks among the cabinets of Presidents Reagan, G.H.W. Bush, Clinton, and G.W. Bush--members' co-occurrence in news stories.
- Each administration's data was sliced into time intervals corresponding to the Gallup presidential approval polls to synchronize the social networks with presidential job approval ratings.
- Centrality president in cabinet network computed.

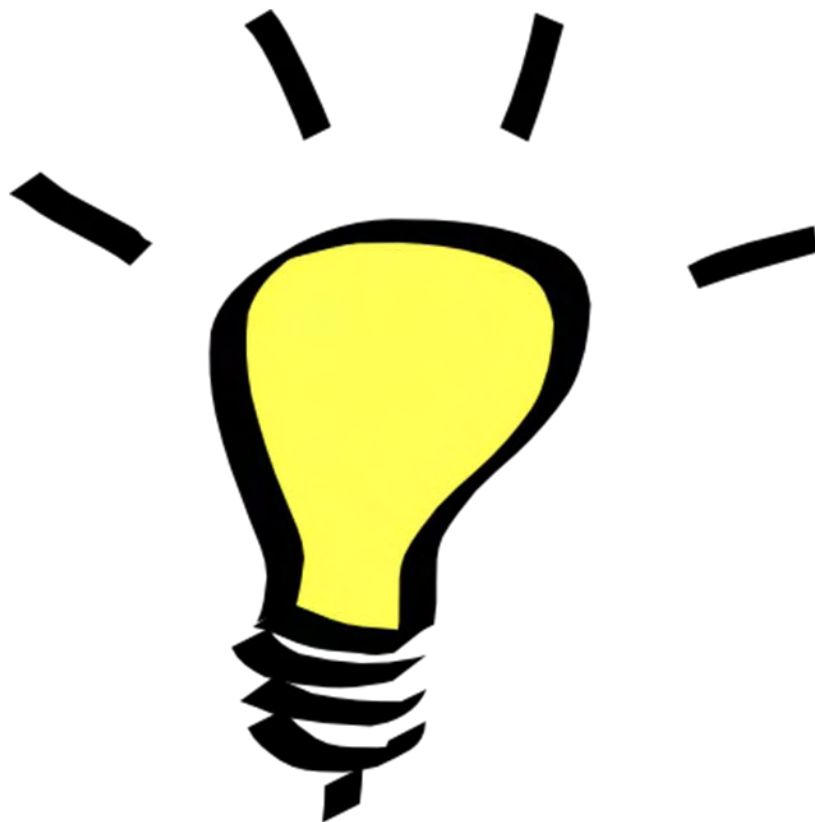
Table 1. Examples of String Replacement (Partial File for Nixon Cabinet)

Richard Nixon->richard_nixon
richard nixon->richard_nixon
nixon-richard_nixon
President->richard_nixon
president->richard_nixon
Vice President->spiro_agnew
vice president->spiro_agnew
Spiro Agnew->spiro_agnew
spiro agnew->spiro_agnew
agnew->spiro_agnew
Gerald Ford->gerald_ford
gerald ford->gerald_ford
ford->gerald_ford
William Rogers->william_rogers
william rogers->william_rogers
rogers->william_rogers
Henry Kissinger->henry_kissinger
henry kissinger->henry_kissinger
kissenger->kissinger
David Kennedy->david_kennedy
david kennedy->david_kennedy
kennedy->david_kennedy
John Connally->john_connally
john connally->john_connally
connallh->john_connally
George Shultz->george_shultz
george shultz->george_shultz
shultz->george_shultz

Table 2. Example Include List for Nixon Cabinet

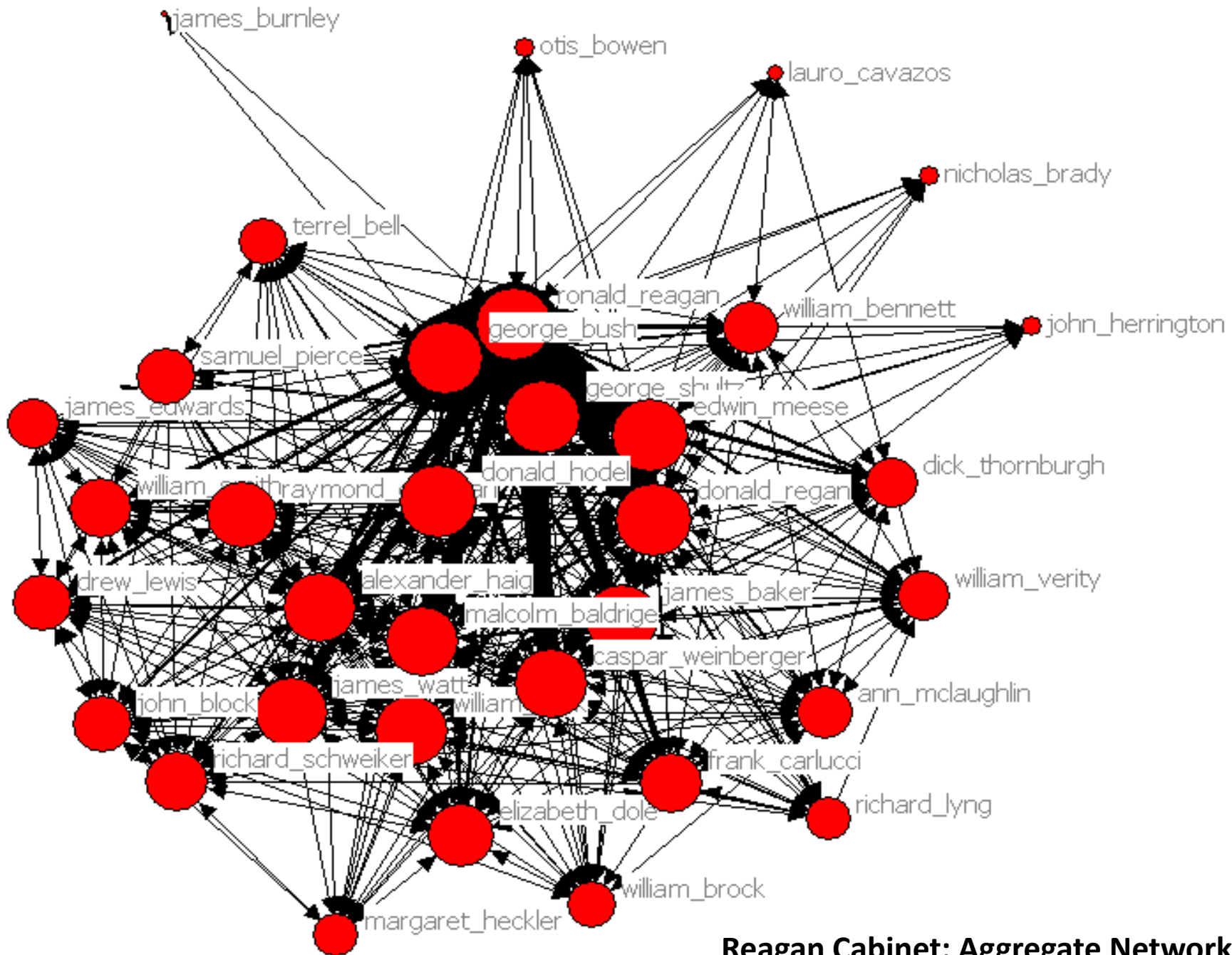
caspar_weinberger
claude_brinegar
clifford_hardin
clifford_hardin
david_kennedy
earl_butz
elliott_richardson
frederick_dent
george_romney
george_shultz
gerald_ford
henry_kissinger
james_hodgson
james_schlesinger
john_connally
john_mitchell
maurice_stans
melvin_laird
peter_brennan
peter_peterson
rich._kleindienst
richard_nixon
robert_finch
rogers_morton
spiro_agnew
walter_hickel
william_rogers
william_saxbe
william_simon
winton_blount



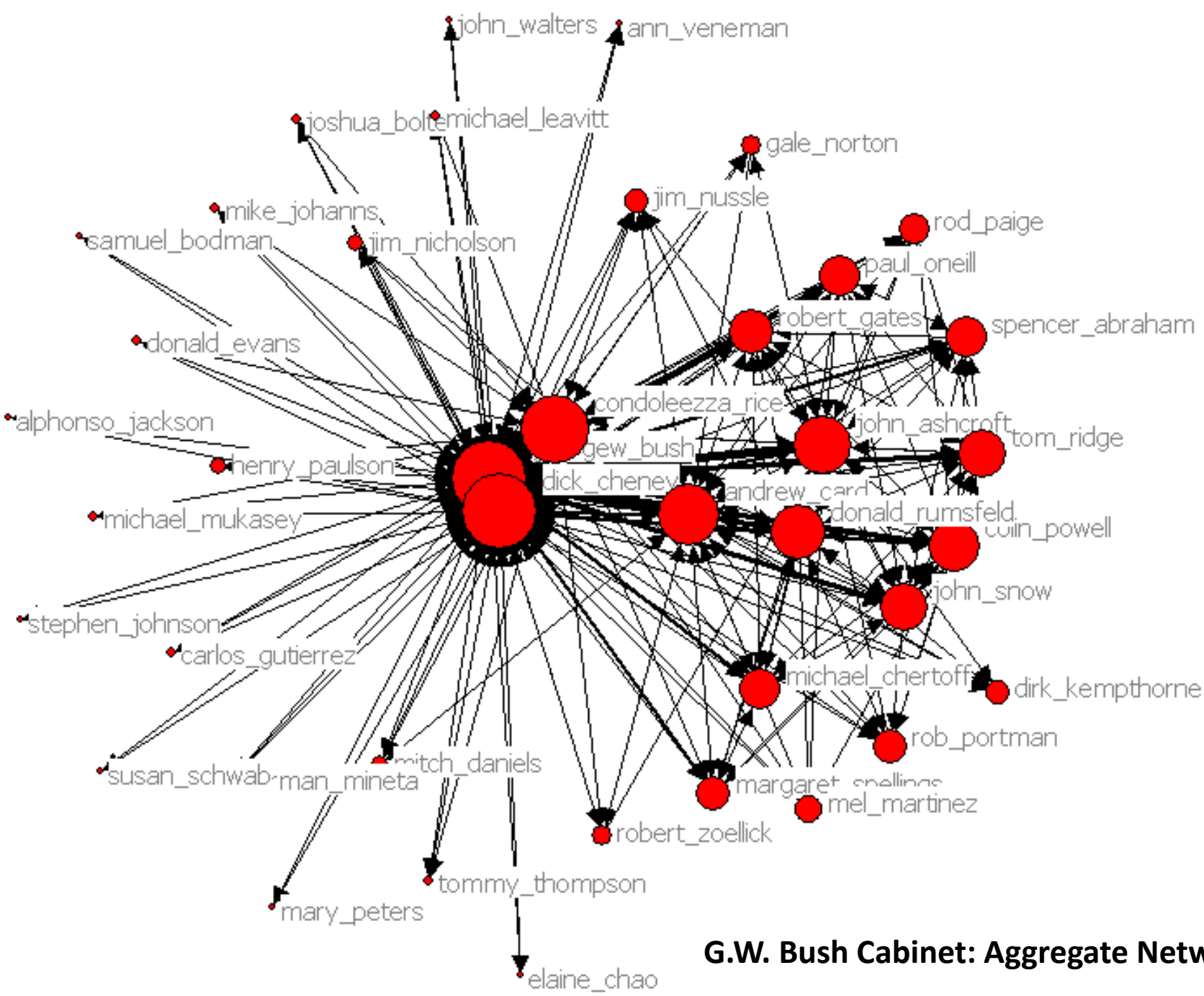


Automatic social network analysis from text mining

Organizational member networks from news stories

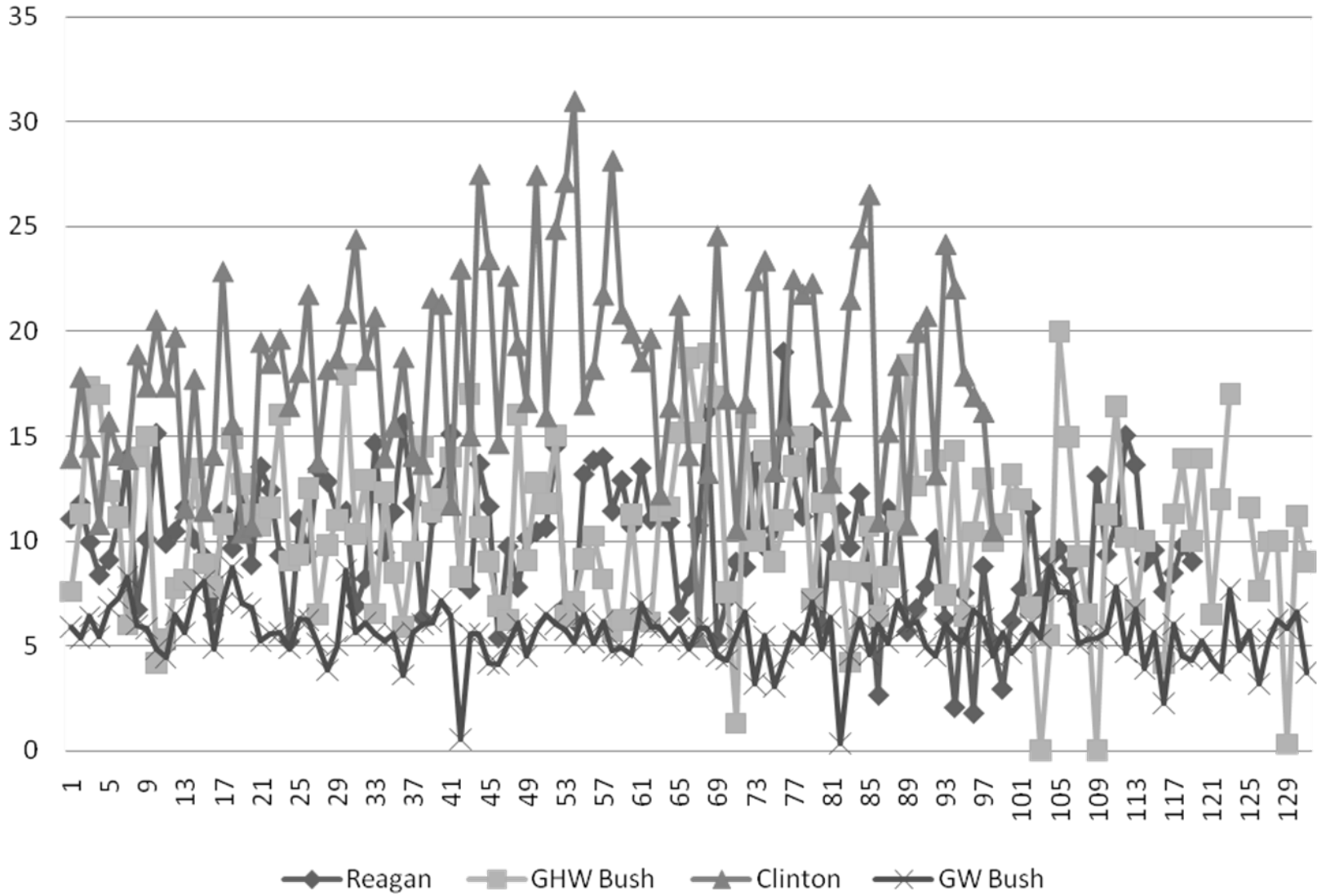


Reagan Cabinet: Aggregate Network



G.W. Bush Cabinet: Aggregate Network

Change over time in presidential
cabinet network centrality of
president from news stories

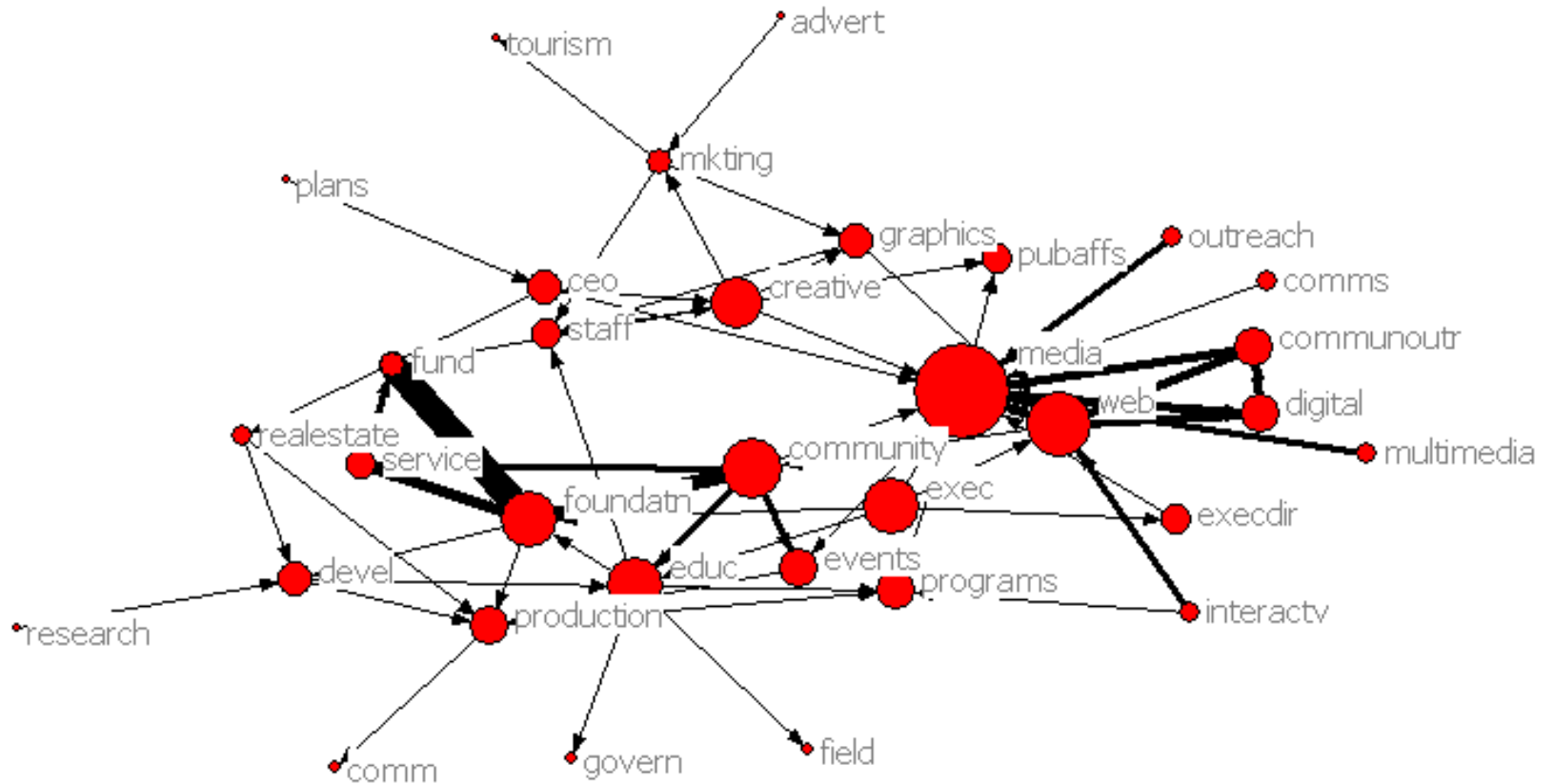


Ratio of President Centrality to Average Cabinet Member Centrality

Survey-derived departments
network from news stories

American Public Television

Automatic Interdepartmental Network



Interdepartmental collaboration networks from news stories

Table 1. Example of String Replacement File

Department of Advertising and Design->Advert_Design

Dept. of Advertising and Design->Advert_Design

Dept of Advertising and Design->Advert_Design

Advertising and Design Department->Advert_Design

Advertising and Design Dept.->Advert_Design

Advertising and Design Dept->Advert_Design

Advertising and Design->Advert_Design

Department of Accessory Design->Accessory

Dept. of Accessory Design->Accessory

Dept of Accessory Design->Accessory

Accessory Design->Accessory

Accessory Design Department->Accessory

Accessory Design Dept.->Accessory

Accessory Design Dept->Accessory

2005 Department Flow Betweenness

| | |
|------------|-------|
| fashion | 12.62 |
| animation | 12.54 |
| interior | 11.89 |
| filmtv | 5.05 |
| architec | 3.62 |
| performing | 3.62 |
| painting | 2.56 |
| print | 1.81 |
| sequential | 0.92 |
| photog | 0.58 |
| illus | 0.40 |

2006 Department Flow Betweenness

| | |
|------------|-------|
| painting | 11.22 |
| teaching | 9.34 |
| performing | 4.86 |
| photog | 2.28 |
| animation | 0.85 |
| writing | 0.78 |
| foundation | 0.13 |
| fashion | 0.06 |

2007 Department Flow Betweenness

| | |
|------------|-------|
| painting | 17.25 |
| fashion | 15.60 |
| animation | 14.42 |
| performing | 12.40 |
| interior | 11.72 |
| architec | 7.32 |
| print | 6.97 |
| filmtv | 3.35 |
| jewelry | 1.69 |
| urban | 1.38 |
| writing | 1.28 |
| accessory | 0.95 |

2008 Department Flow Betweenness

| | |
|------------|-------|
| photog | 27.82 |
| painting | 26.39 |
| interior | 17.50 |
| performing | 10.67 |
| architec | 10.25 |
| teaching | 9.83 |
| fashion | 4.16 |
| sequential | 3.56 |
| animation | 3.05 |
| filmtv | 2.89 |
| jewelry | 2.49 |
| foundation | 1.74 |
| urban | 1.25 |
| sculpture | 0.16 |

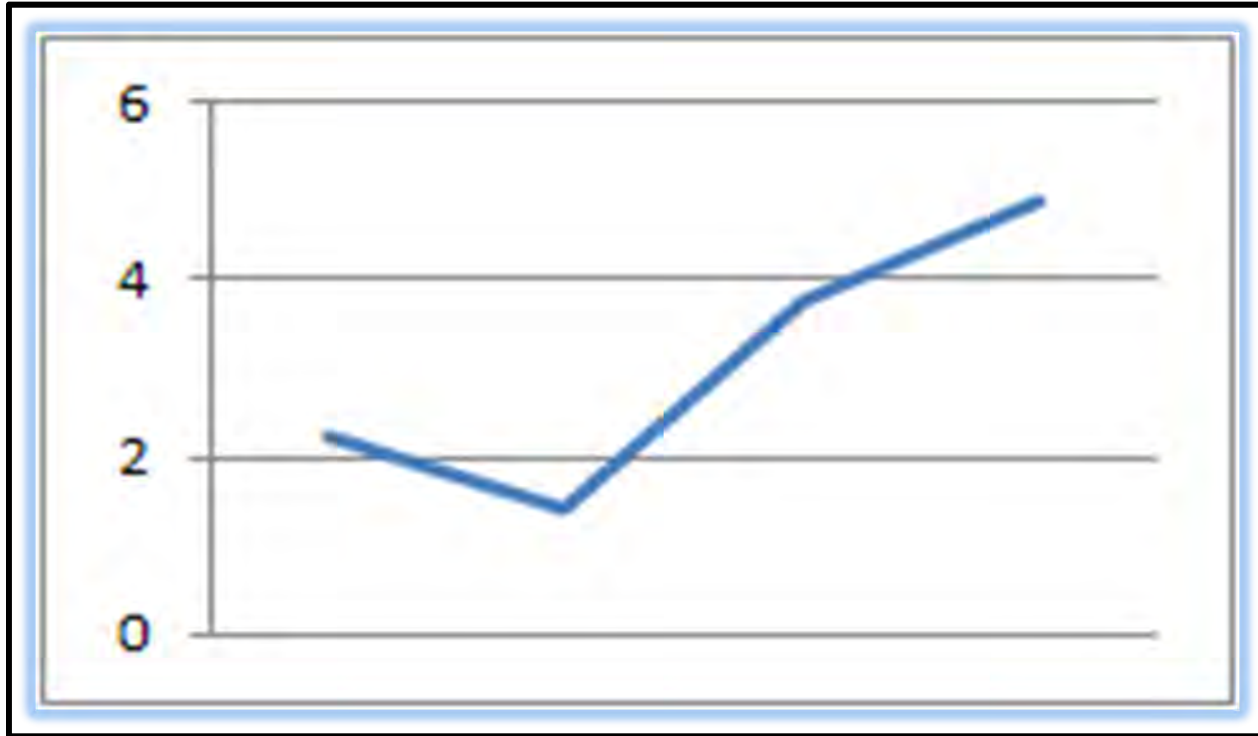


Figure 1. Average Flow Betweenness by Year

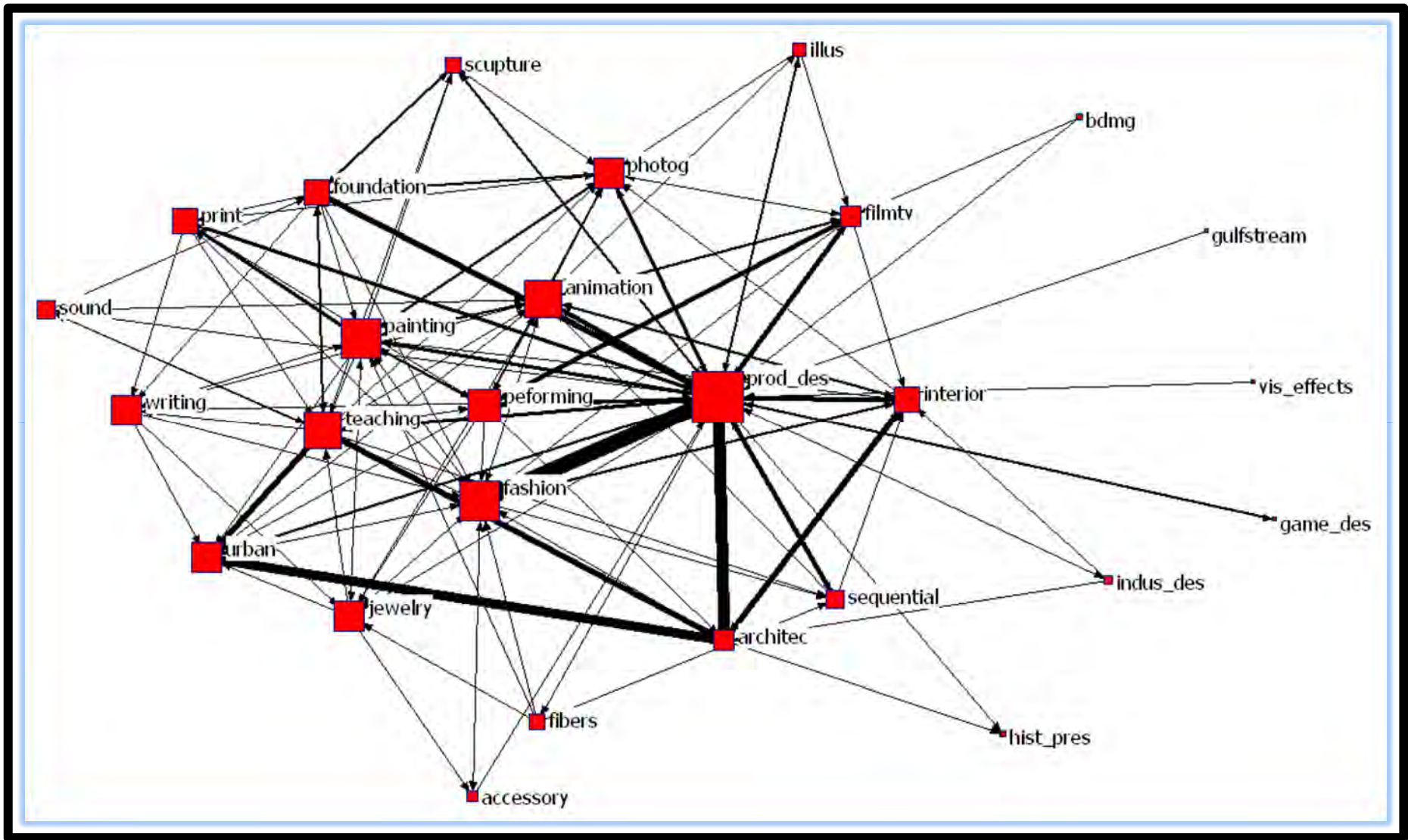
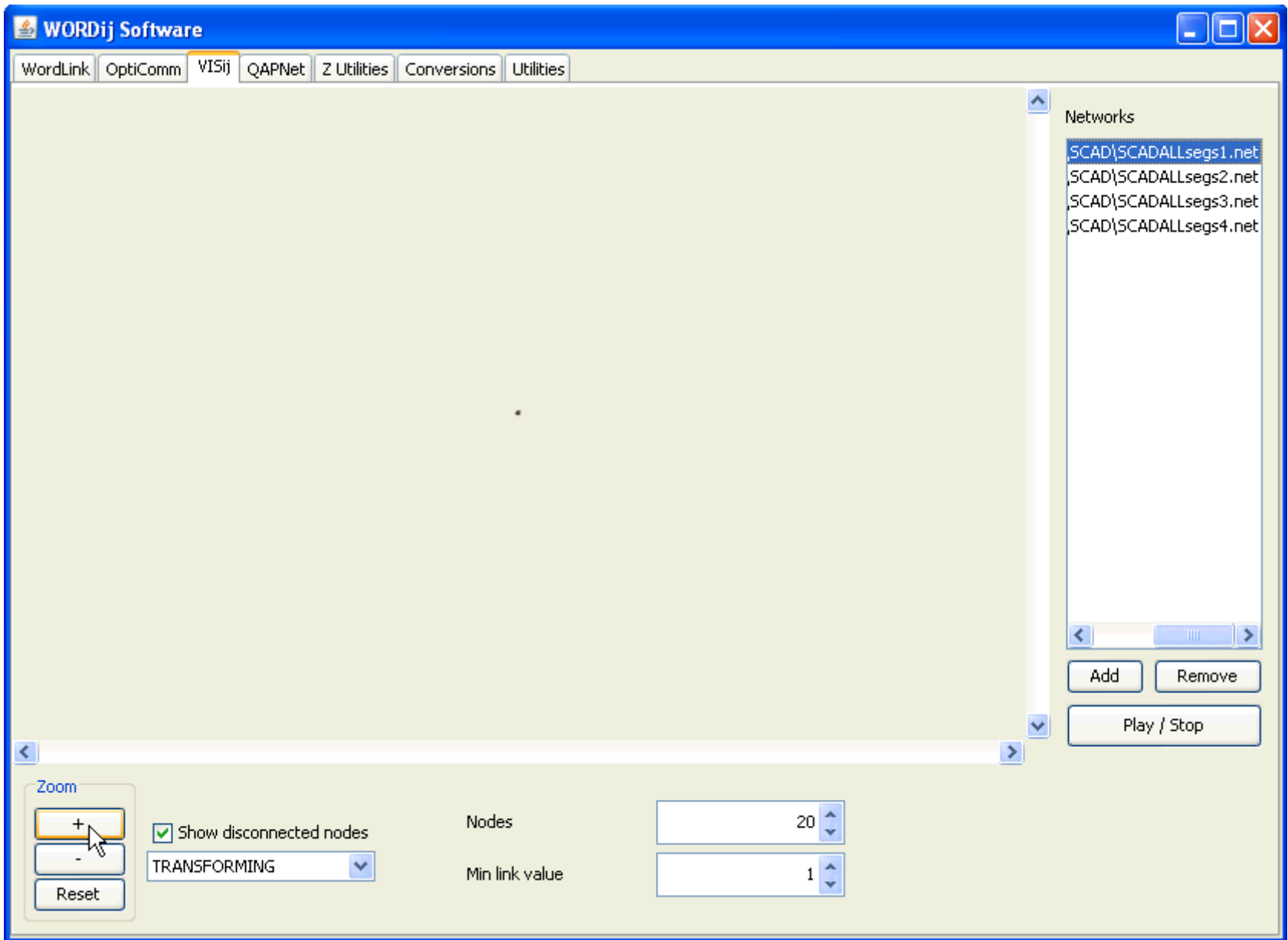
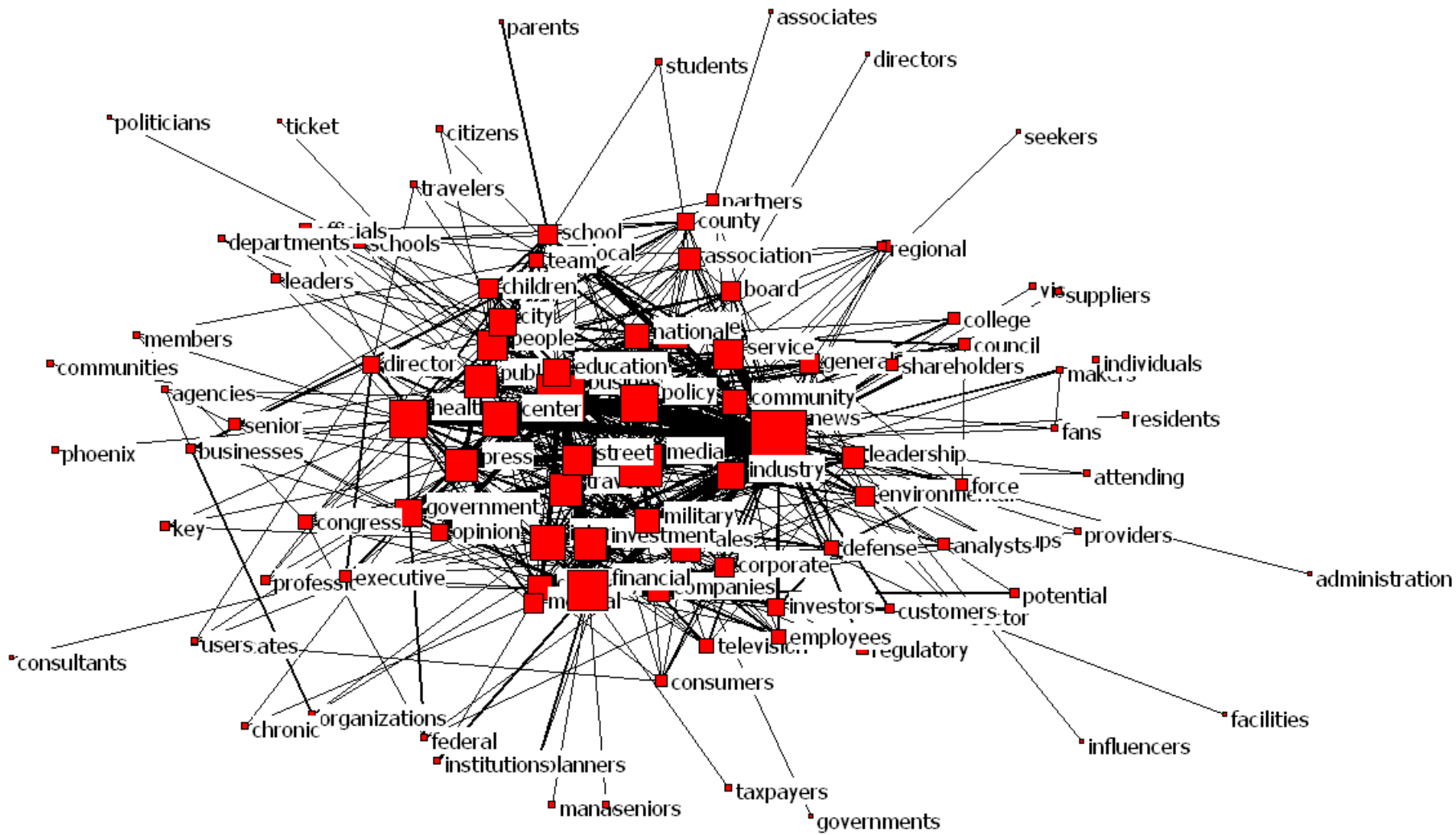


Figure 3. Aggregate Interdepartmental Network: 2005-2008

Interdepartmental collaboration
network change over time in
organization



Survey-derived list of stakeholders
by blog content for organizations



Brown-Forman Beverages Stakeholder Networks in Blogs

Survey-derived stakeholders in
news stories about organizations

Summary

- Large volumes of text can be mined for semantic and social networks.
- Node-centric analysis enables zoom in on local network structure.
- Time-slicing allows for more causal evidence.
- Include lists generate networks combining data from multiple sources.
- Networks can be statistically compared for similarity and differences.

Work in Progress

- Do automatic identification of social network members not on an include list, but whose semantic networks are highly correlated with members.
- Improve automation of string replacement and include list creation from proper noun extraction.
- Sentiment and message and social network structures over time behaviors.
- Build model linking message text content features to who-to-who network features over time: orgo-mechanical screw model.

Obtain WORDij from
<http://wordij.net>

Questions, please.

Thank you for your attention