WSDM 2016 Tutorial

Understanding Offline Political Systems by Mining Online Political Data

David Lazer

Northeastern & Harvard

d.lazer@neu.edu

Oren Tsur

Northeastern & Harvard

orentsur@seas.harvard.edu

Tina Eliassi-Rad

Northeastern & Rutgers

tina@eliassi.org

Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

USING BIG DATA TO UNDERSTAND POLITICS

David Lazer

Northeastern/Harvard

Outline

- Some big themes in politics
- The opportunity to create a new science of politics
- Exemplar data
- Cautionary tales

Big themes in politics

- Collective action
- Political communication
- Power

The paradox of collective action

The paradox of collective action

- Social movements
- Voting
- Contributing to campaigns
- Vaccination

Why?

- “Selective benefits/penalties”
- Solidarity
- Norms

Political communication

THE GETTYSBURG ADDRESS

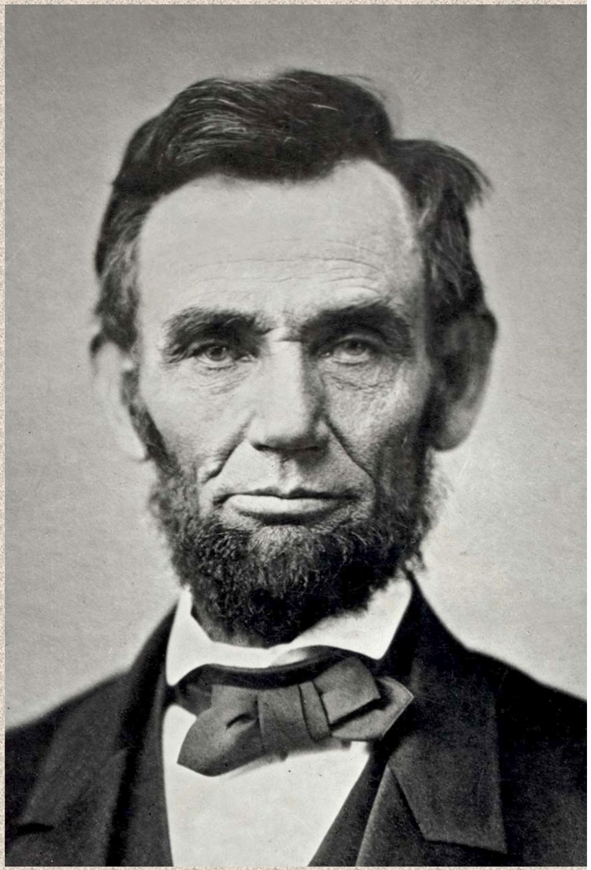
November 19, 1863
At the Dedication of the Soldiers' National Cemetery in Gettysburg, Pennsylvania:

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate – we can not consecrate -- we can not hallow – this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion – that we here highly resolve that these dead shall not have died in vain – that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Abraham Lincoln



The construction of language in politics

- Testing of different linguistic constructions (“estate taxes” vs “death taxes”)
  - Surveys, focus groups, etc
- Process of dissemination to elites
- Re-dissemination via media

Three dimensions of power (Lukes)

- Decision making: When A gets B to do something B otherwise would not do.
- Agenda setting: what is and is not discussed.
- Normative influence: what you think is in your interest

Time for a new science of politics

- Most social science is:
  - Static
  - Spatially and socially decontextualized
  - And small scale (hundreds or thousands of individuals)

Time for a new science of politics

- The new science of politics
  - dynamic
  - Spatially and socially embedded
  - And societal or even globally spanning

Exemplar big data

- $ in politics
- Political language

Federal elections commission data

Langevin for Congress

A.

Full Name (Last, First, Middle Initial)
Robert Popeo

Mailing Address One Financial Ctr., 40th Floor

City	State	Zip Code
Boston	MA	021 1

FEC ID number of contributing federal political committee. C

Name of Employer: Mintz Levin Cohn Ferris Glovsky & Pope
Occupation: Chairman

Receipt For: 2008
X Primary — General — Other (specify)

Election Cycle-to-Date: 500.00

Date of Receipt: M M / D D / Y Y Y Y — 03 / 10 / 2008

Transaction ID: C2017416

Amount of Each Receipt this Period: 500.00

Limit Increased Due to Opponent's Spending (2 U.S.C. 441a(i)/441a-1)

B.

Full Name (Last, First, Middle Initial)
Larry Rasky

Mailing Address 20 Bridle Path

City	State	Zip Code
Westwood	MA	02090

FEC ID number of contributing federal political committee. C

Name of Employer: Rasky Baerlein, Inc.
Occupation: Partner

Receipt For: 2008
X Primary — General — Other (specify)

Election Cycle-to-Date: 1000.00

Date of Receipt: M M / D D / Y Y Y Y — 03 / 10 / 2008

Transaction ID: C2017417

Amount of Each Receipt this Period: 1000.00

Limit Increased Due to Opponent's Spending (2 U.S.C. 441a(i)/441a-1)

C.

Full Name (Last, First, Middle Initial)
John Reger

Mailing Address 89 Farnham St
Mintz Lezin Cohn

City	State	Zip Code
Belmont	MA	02478-3172

FEC ID number of contributing federal political committee. C

Name of Employer
Occupation

Date of Receipt: M M / D D / Y Y Y Y — 03 / 10 / 2008

Transaction ID: C2017410

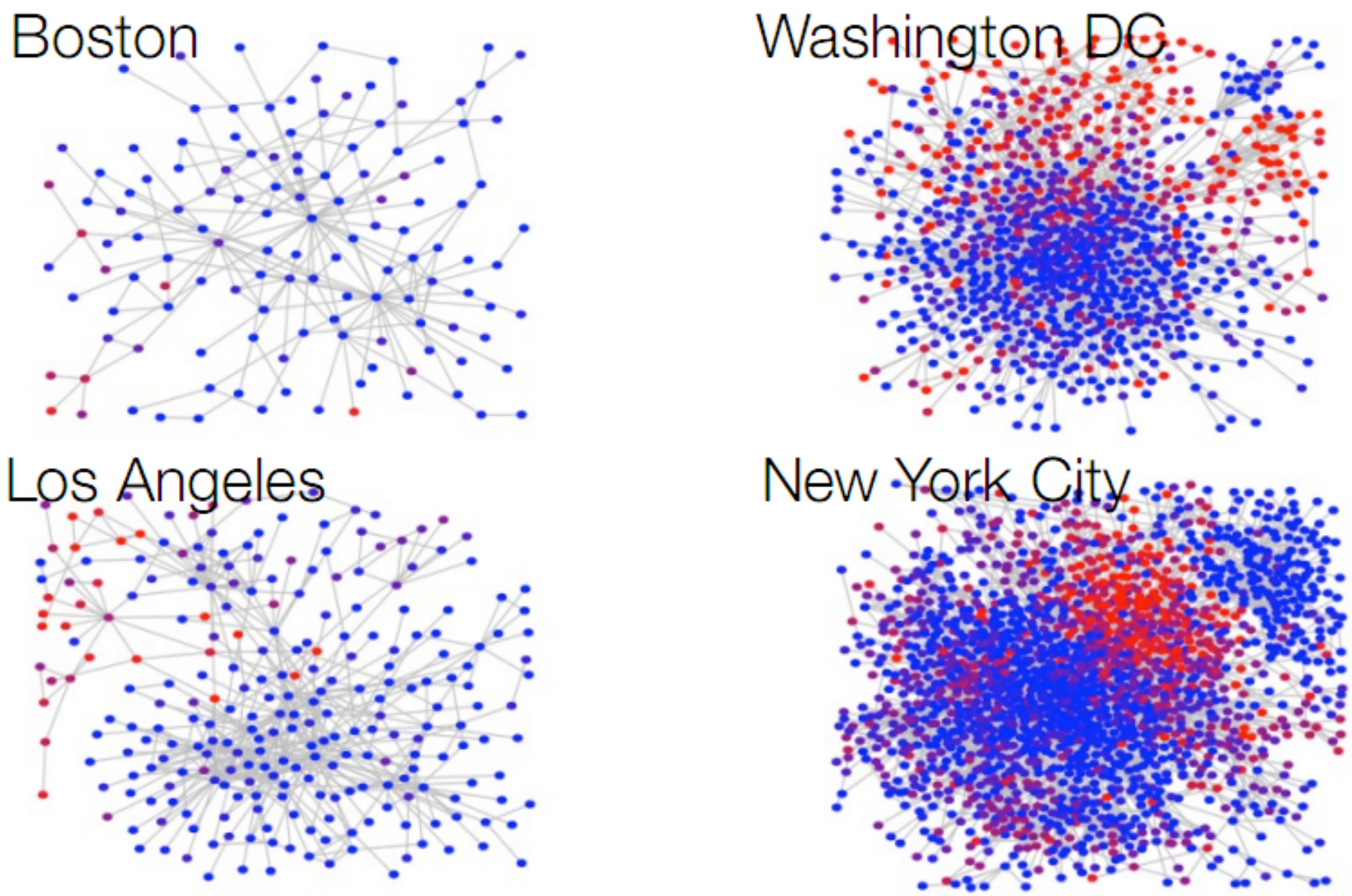
Amount of Each Receipt this Period: 500.00

FEC data

- Contributor name
- Occupation
- Employer
- Address/zip of contributor
- Receiving committee (unique id)
- Donation amount
- Date
- *http://www.fec.gov/finance/disclosure/ftpdet.shtml*


Boston
Organizations
MoneyBombs
DONATIONS
in log scale
100,000 $
50,000 $
40,000 $
30,000 $
20,000 $
10,000 $
Democratic
Republican
Bain Capital
Fidelity Investments
Harvard University
MIT
TA Associates
The Baupost Group
Thornton & Naumes
Sunday, 1/1/12
Medford
Arlington
Everett
Revere
North Waltham
Belmont
Somerville
Harvard University
Winthrop
Waltham
Watertown
Cambridge
North End
Logan International Airport
MIT
Boston Common
Brighton
South Boston
Newton
Northeastern University
Brookline
Thompson Island
Island Shelter
Jamaica Plain
Oak Hill
Dorchester
Marina Bay
Needham
Mattapan

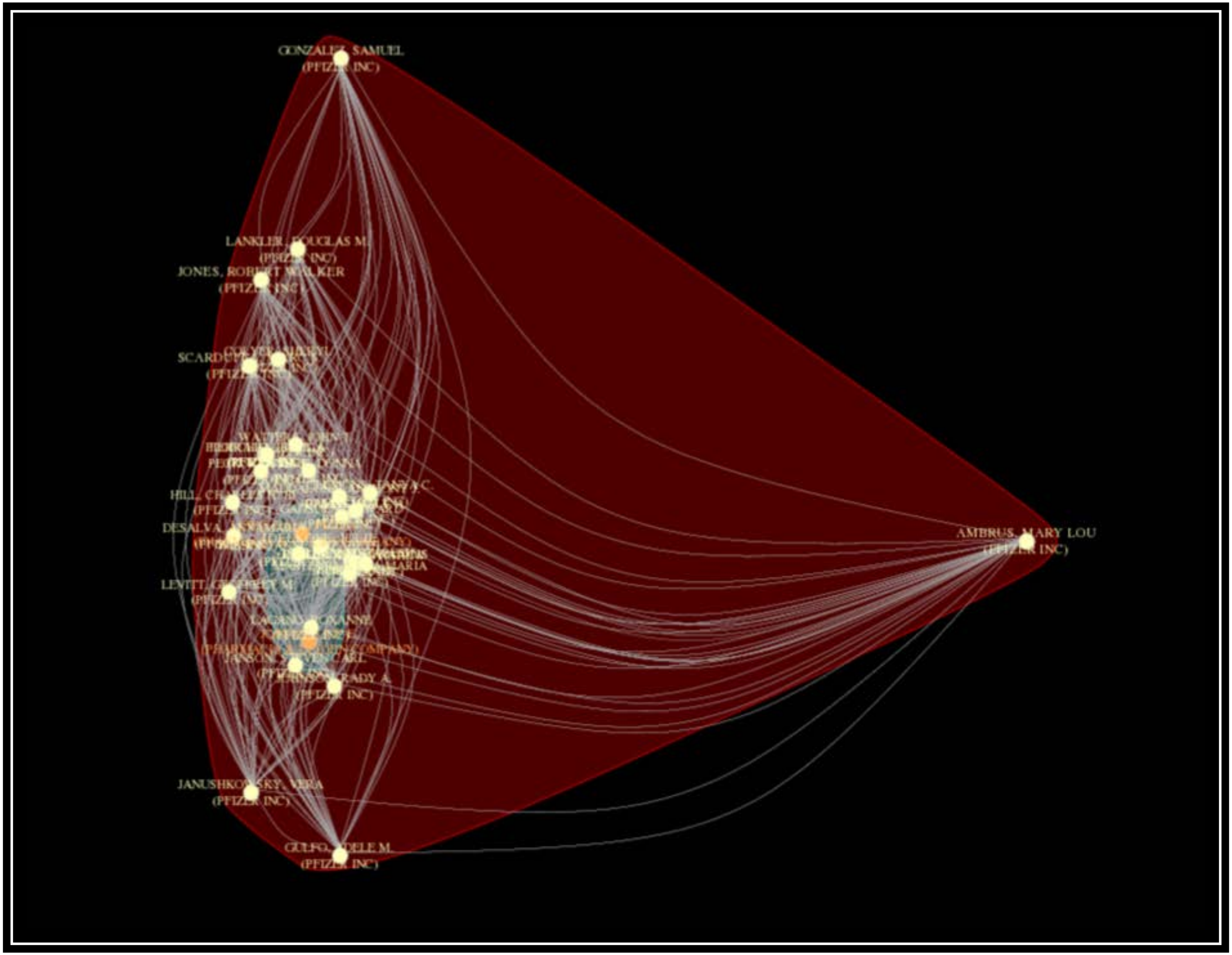
And inferred network among contributors



From Ruths and Lazer (2009)

Inferred relationships…

- Boston
  - Ryan Vincent & Carla Meyer - board members
- Washington DC
  - Ed Rogers & Lanny Griffith - partners in lobbying firm
- Los Angeles
  - Spielberg & Katzenberg - film producers
- NYC
  - Debra Black & Judith Hannan - board members

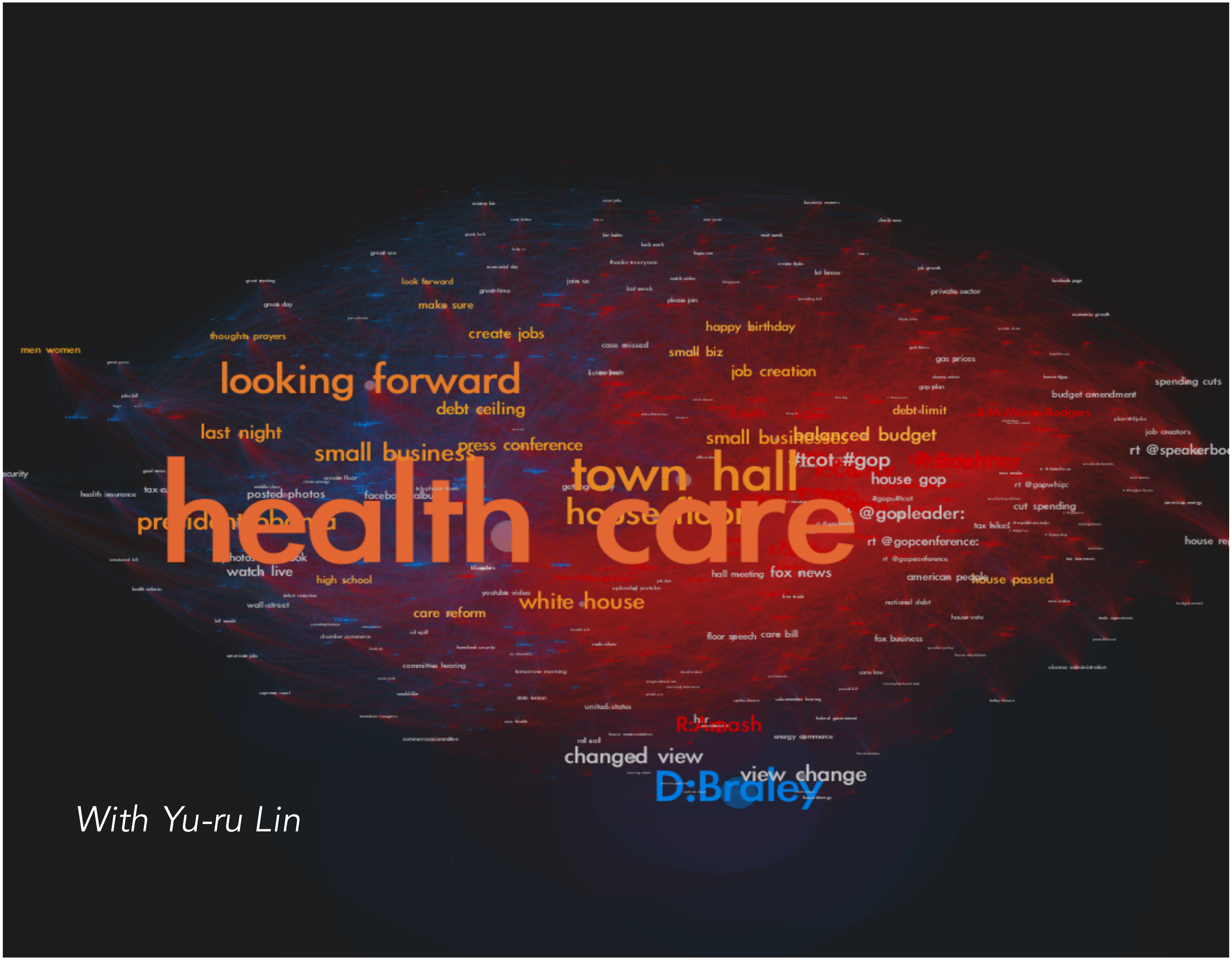

GONZALEZ, SAMUEL
(PFIZER INC)
LANKLER, DOUGLAS M.
(PFIZER INC)
JONES, ROBERT WALKER
(PFIZER INC)
AMBRUS, MARY LOU
(PFIZER INC)
JANUSHKOWSKY, VERA
(PFIZER INC)
GULFO, ADELE M.
(PFIZER INC)
JOHNSON, BRADY A.
(PFIZER INC)

FYI, new $ data..

- Federal Election Commission data do not have unique identifiers…
- And disambiguation is a big barrier to doing anything with the data
- So we synthesized unique identifiers
- http://politicalcents.cs.mcgill.ca/

Political language

- Myriad of sources…
- Public statements data from Votesmart



With Yu-ru Lin

osama bin laden

on the web press releases

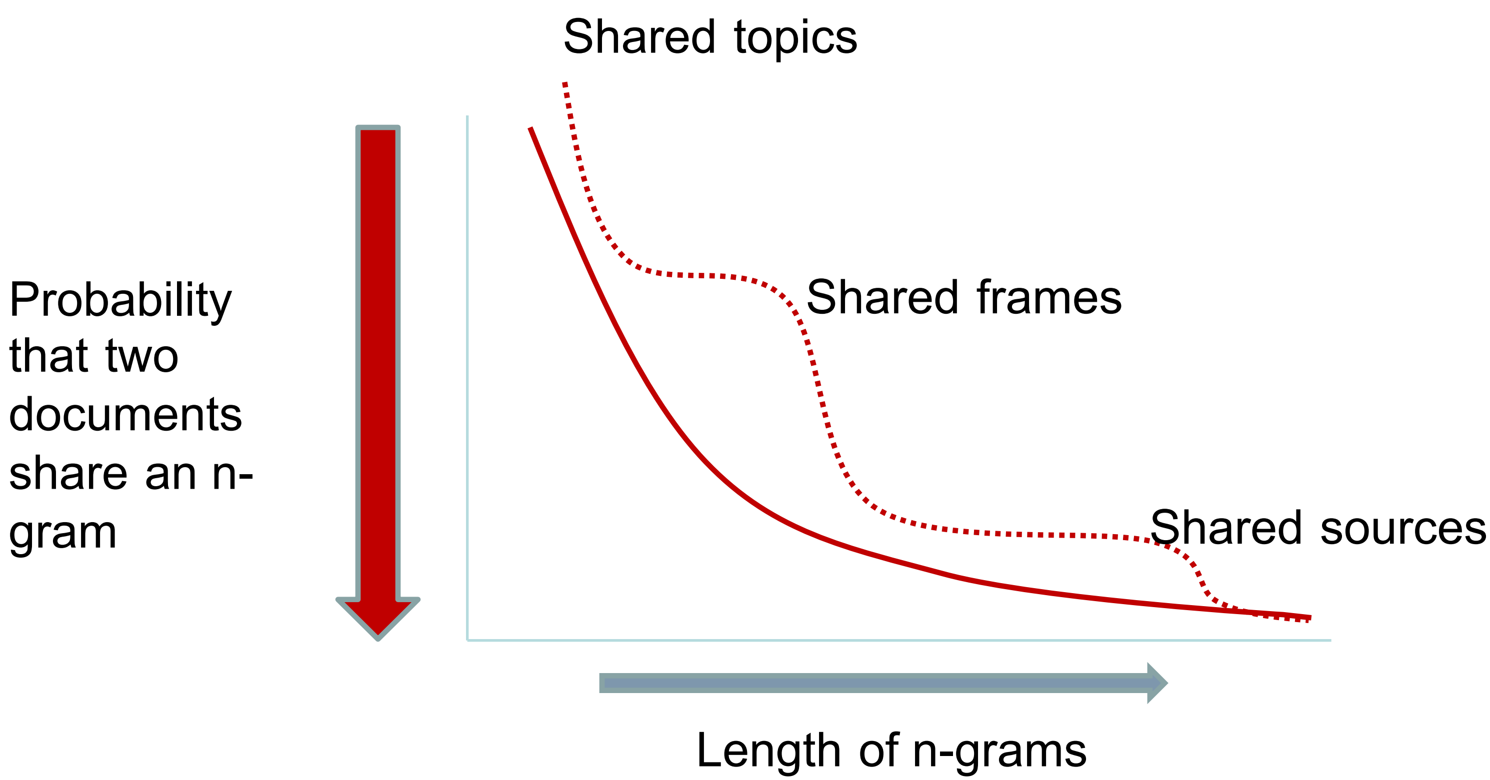
Example

- Data
  - 0.5 million documents from public statements of Members of US Congress from Votesmart
- Computational methods
  - Tracking semantic convergence
  - Randomized n-gram extraction

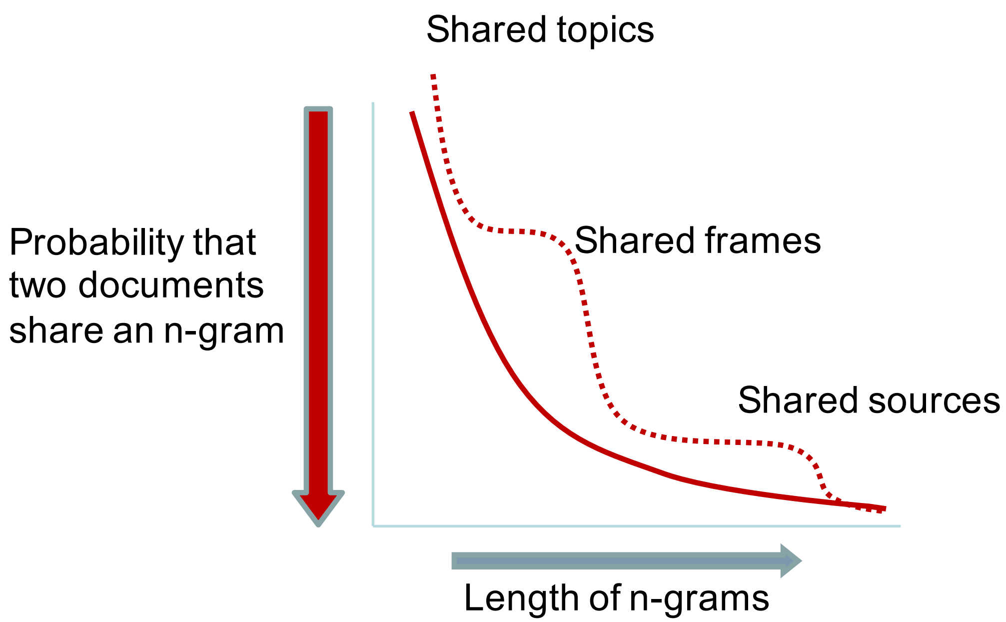
Lin et al 2015

In a large corpus, multiple types of convergence

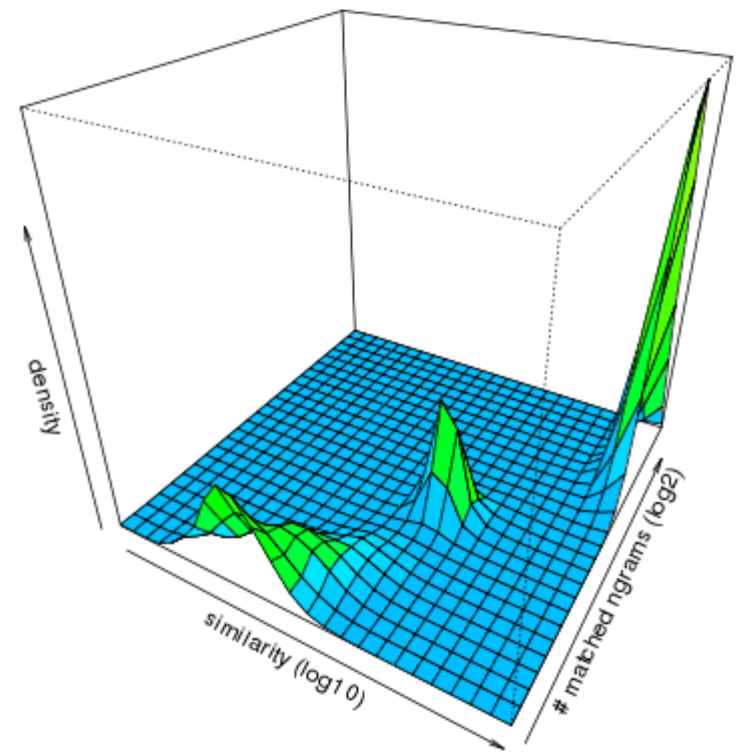
Together producing a “bumpy” distribution



in theory

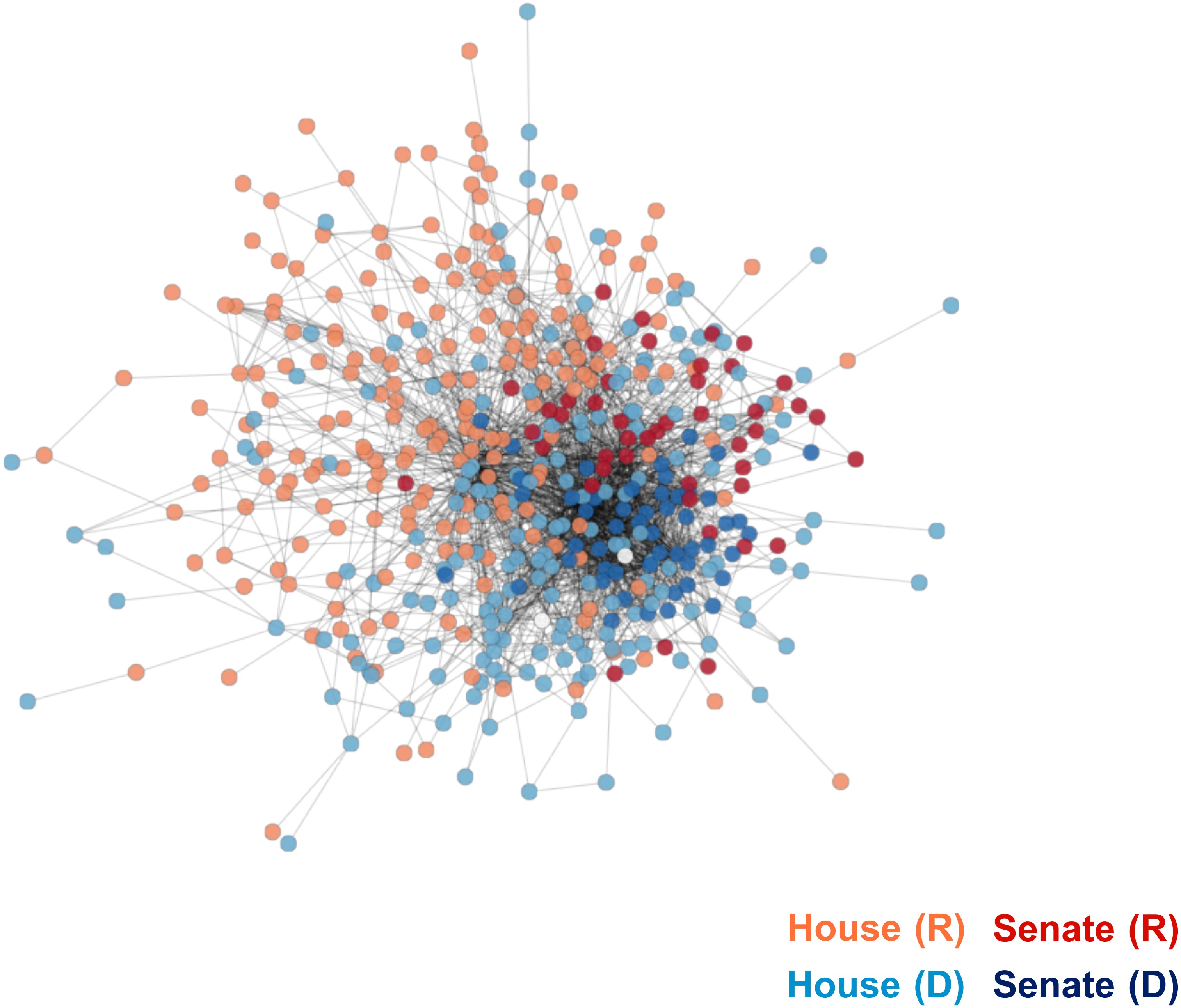


in data

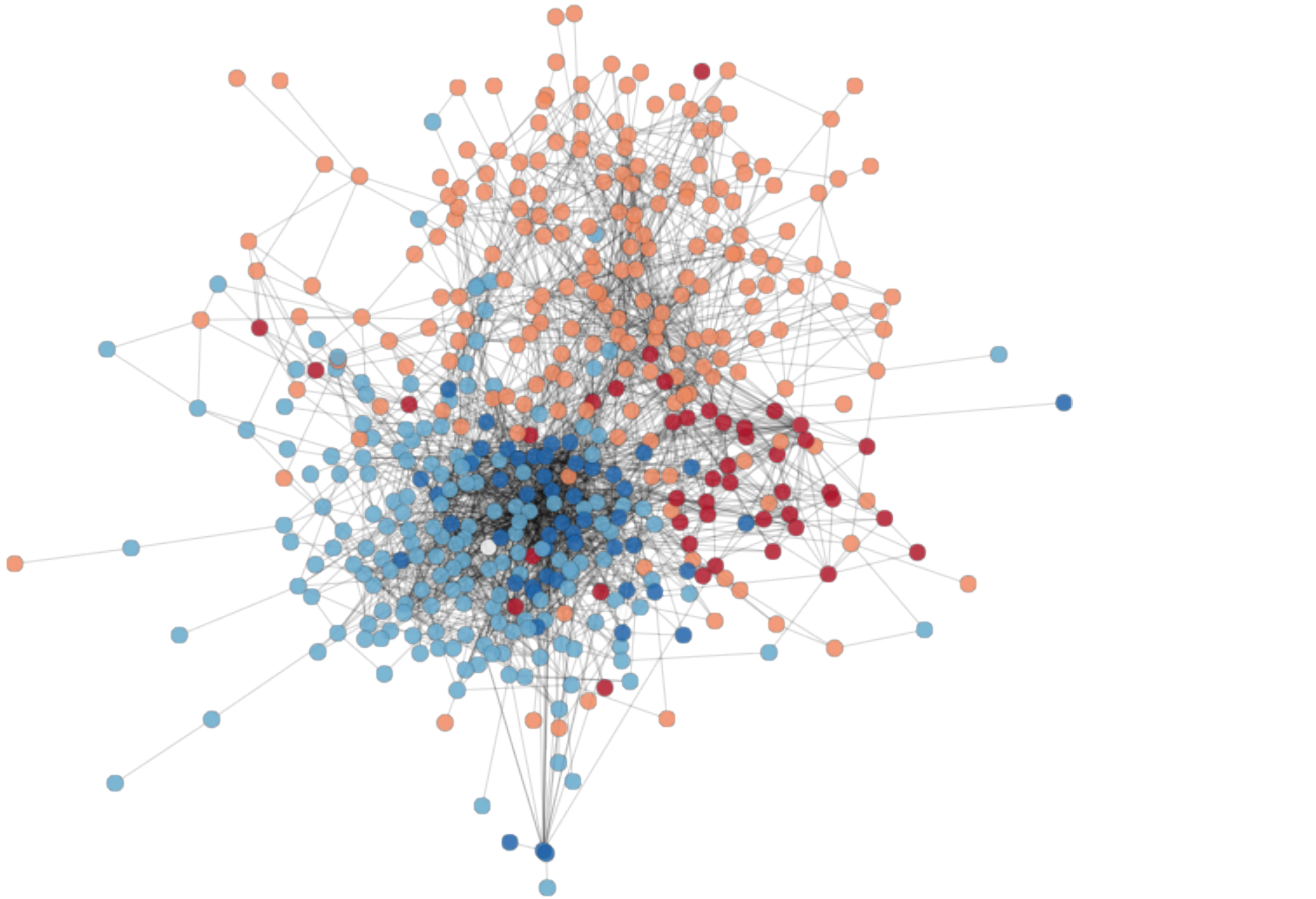


8-gram (jaccard) similarity between document pairs

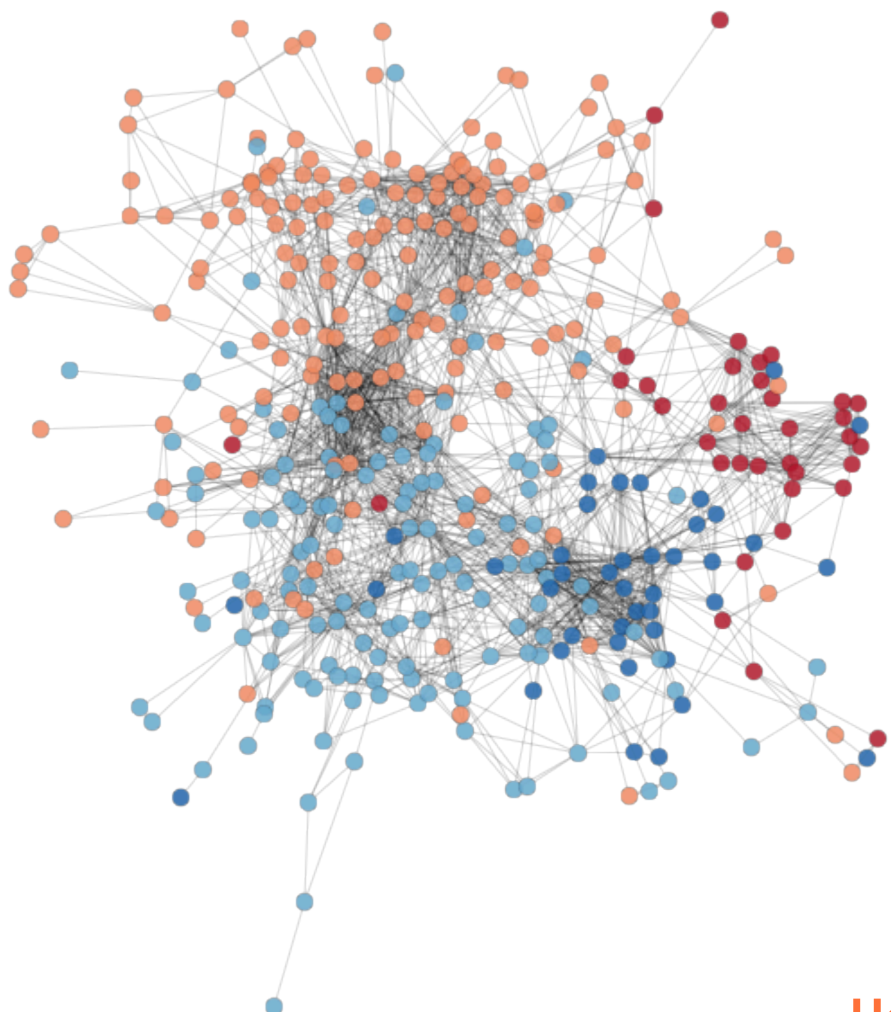
2-gram



4-gram



32-gram



House (R) Senate (R)
House (D) Senate (D)

Social media

- Twitter
- Facebook (?)
- Tumblr
- Anything you can scrape from the Web.
- Etc etc etc

But handle (big) data with care, a few quick lessons from the failure of Google Flu Trends

- Nobody can tell you're a dog on the Internet– and that's not a good thing if you are trying to understand *humans*.
  - Value of curated data– sometimes < 1% of the data is way better than 100% of the data
- Algorithmic changes– e.g., algorithmic sorting in Facebook and Twitter
- Evolving norms– example of hashtags in Twitter

Lazer et al 2015

Example…

- Can we classify people as liberal or conservative based on the language they use?
- **Answer**: yes, lit suggests 90+% accuracy is possible based on snapshots of language use.
- **But: these findings turn out to be ephemeral** *(Cohen and Ruths ICWSM 2013)*

Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

Outline

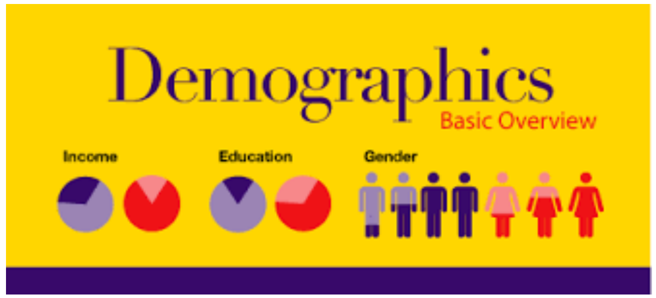
9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

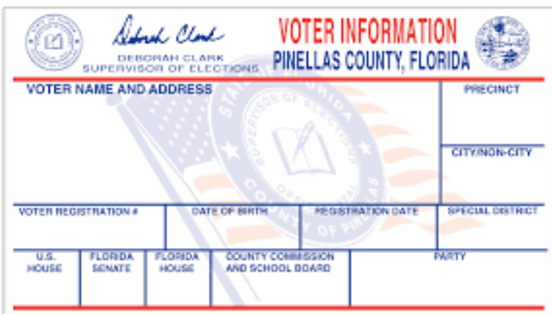
Exponential Random Graph Models

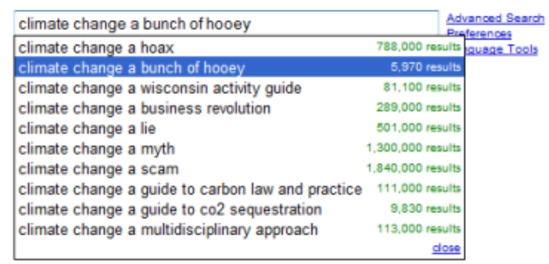
Oren Tsur

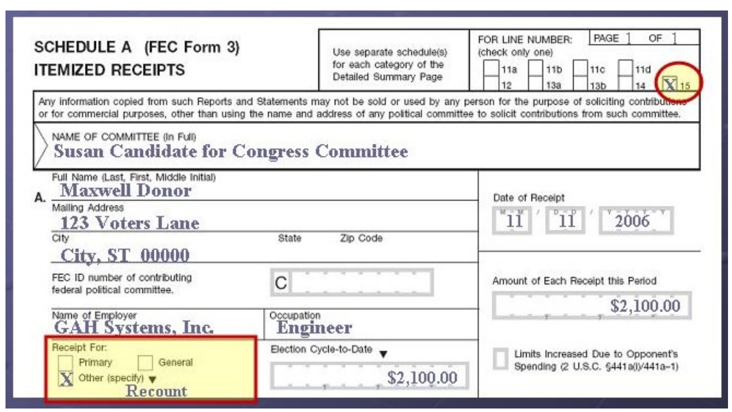
Northeastern/Harvard

Available datasets









SCHEDULE A (FEC Form 3)
ITEMIZED RECEIPTS

Use separate schedule(s) for each category of the Detailed Summary Page

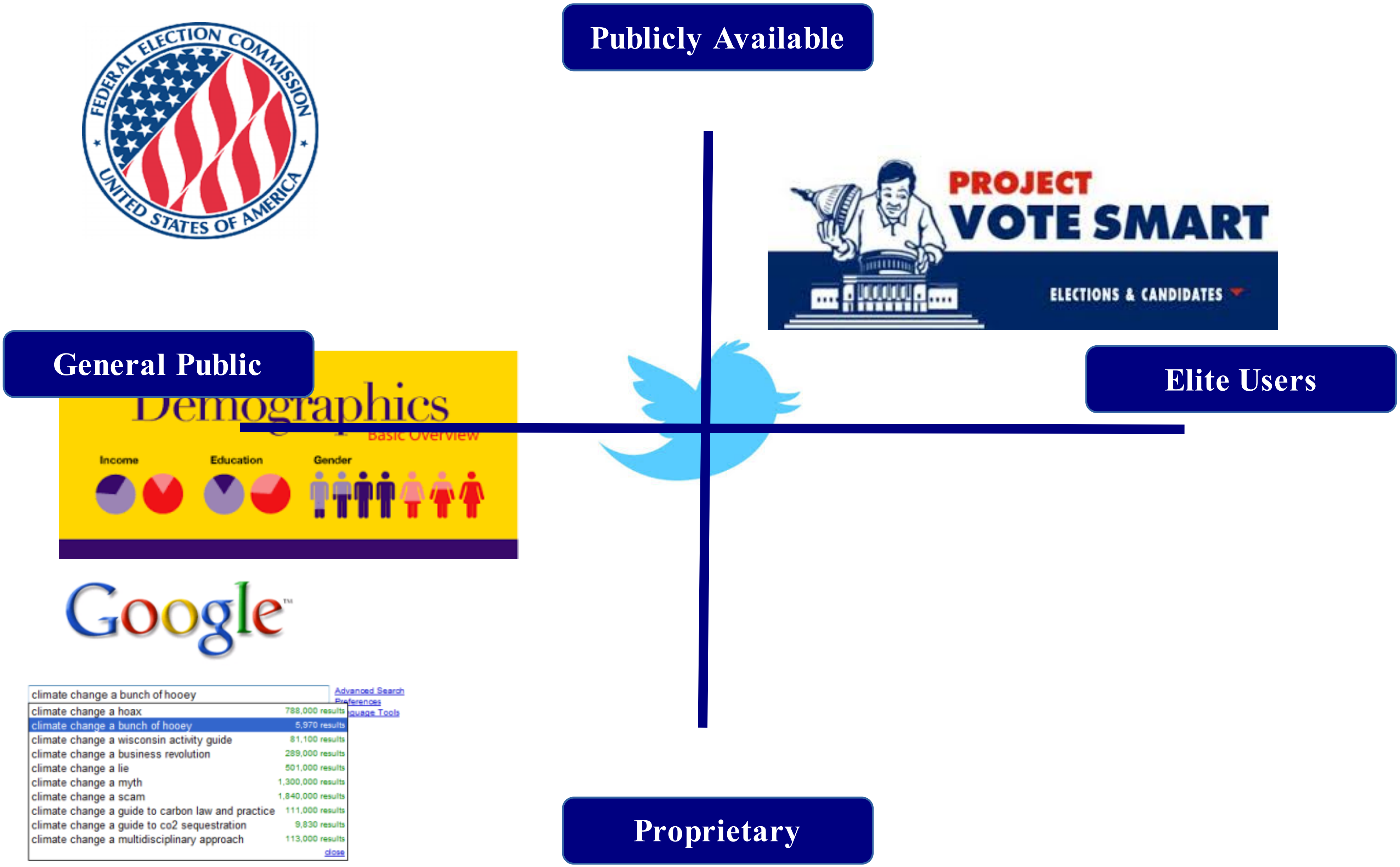
FOR LINE NUMBER: (check only one) PAGE OF
11a 11b 11c 11d 12 13a 13b 14 X 15

Any information copied from such Reports and Statements may not be sold or used by any person for the purpose of soliciting contributions or for commercial purposes, other than using the name and address of any political committee to solicit contributions from such committee.

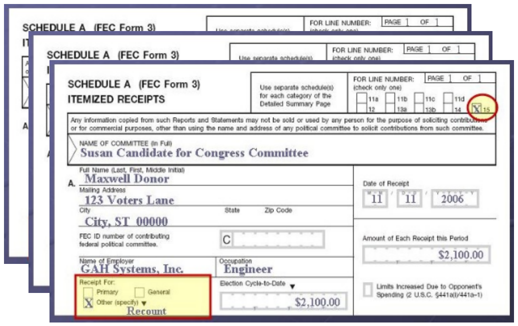
NAME OF COMMITTEE (In Full)
Susan Candidate for Congress Committee

A. Full Name (Last, First, Middle Initial)
Maxwell Donor
Mailing Address
123 Voters Lane
City State Zip Code
City, ST 00000
FEC ID number of contributing federal political committee. C
Name of Employer
GAH Systems, Inc.
Occupation
Engineer
Receipt For:
Primary General
X Other (specify)
Recount
Election Cycle-to-Date
$2,100.00
Date of Receipt
11 / 11 / 2006
Amount of Each Receipt this Period
$2,100.00
Limits Increased Due to Opponent's Spending (2 U.S.C. §441a(i)/441a-1)

Types of “political” datasets



Example: “Money Talks”

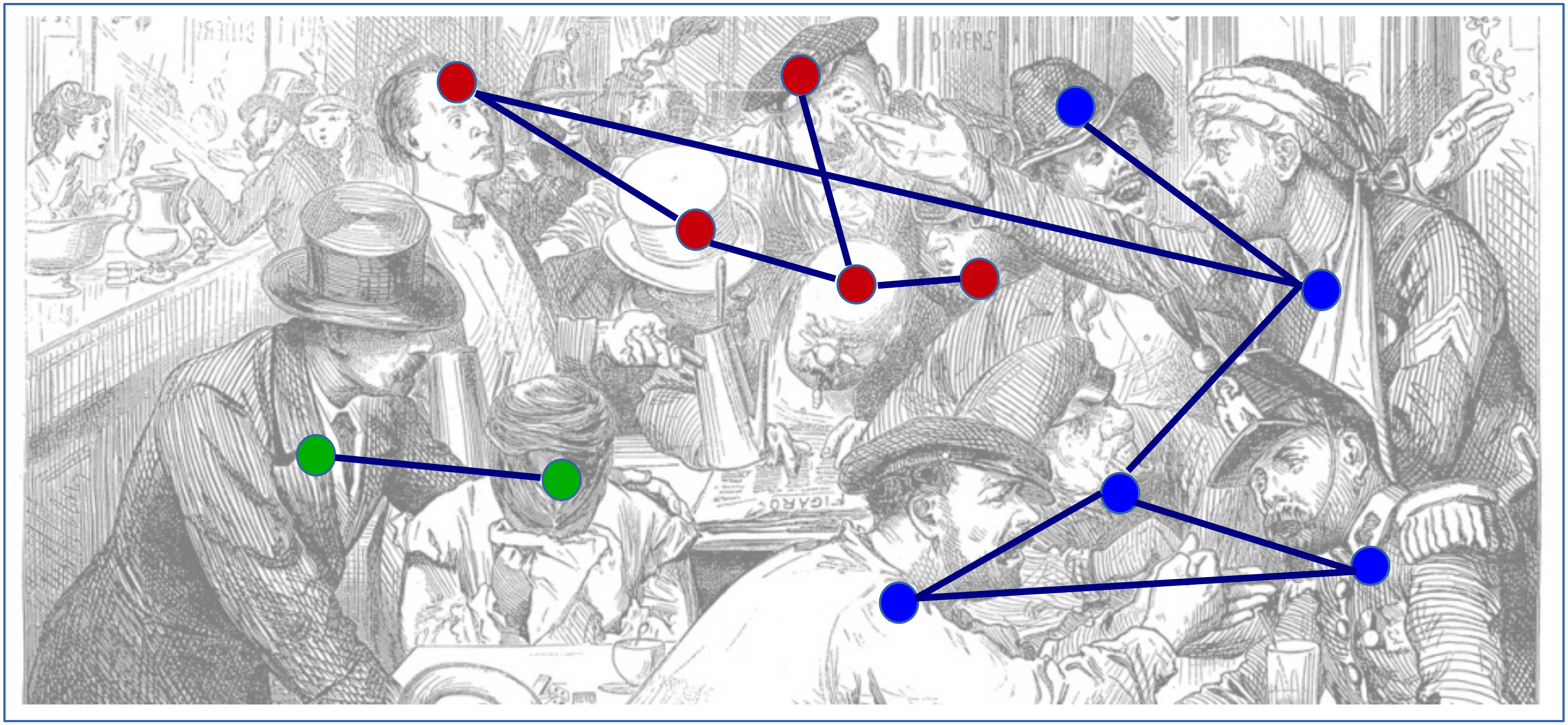

SCHEDULE A (FEC Form 3)
ITEMIZED RECEIPTS
Use separate schedule(s) for each category of the Detailed Summary Page
FOR LINE NUMBER: (check only one)
PAGE OF
11a 11b 11c 11d 12 13a 13b 14 15
NAME OF COMMITTEE (in Full)
Susan Candidate for Congress Committee
A. Full Name (Last, First, Middle Initial)
Maxwell Donor
Mailing Address
123 Voters Lane
City State Zip Code
City, ST 00000
FEC ID number of contributing federal political committee.
C
Name of Employer
GAH Systems, Inc.
Occupation
Engineer
Receipt For: Primary General
X Other (specify)
Recount
Election Cycle-to-Date
$2,100.00
Date of Receipt
11 11 2006
Amount of Each Receipt this Period
$2,100.00

- Finding network/cluster/community similarities btw. contributions networks and speech.

Networks – Informal Introduction

A political setting

Is this a random network?

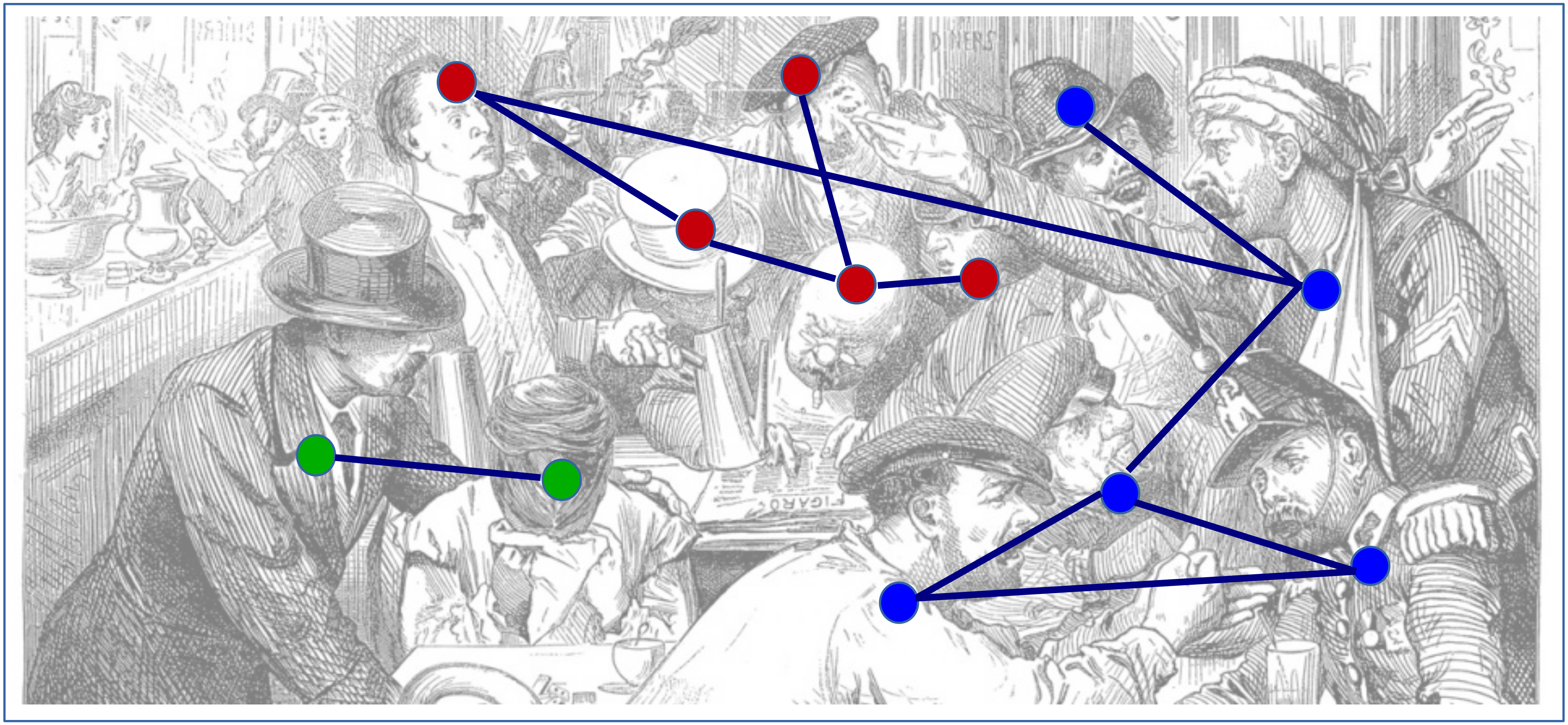


Network analysis

- Given a network
  - Link prediction
  - Community detection
  - Role discovery
- Network dynamics
  - Evolution
  - Contagion, diffusion, cascades
- Network formation
  - Social factors for the above

Networks Basics

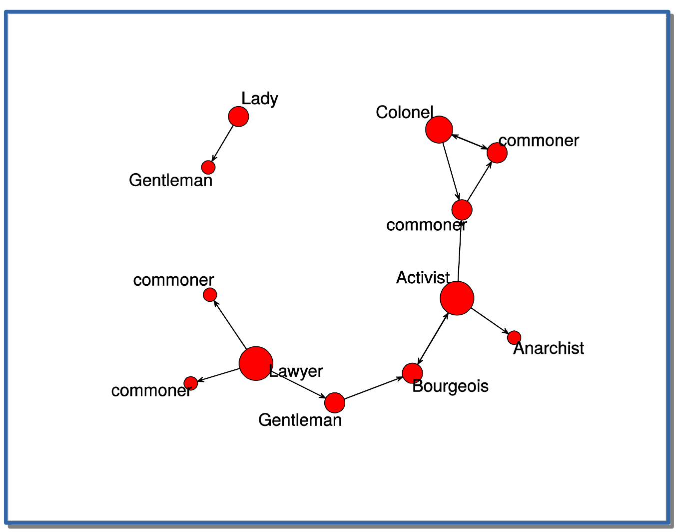
Is this a random network?

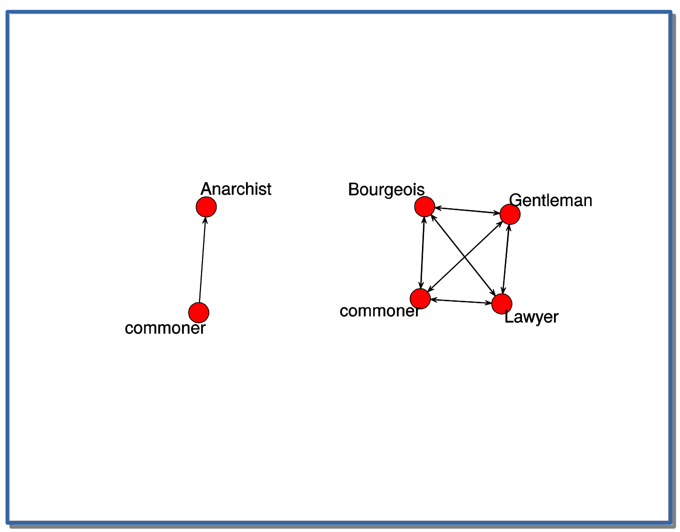


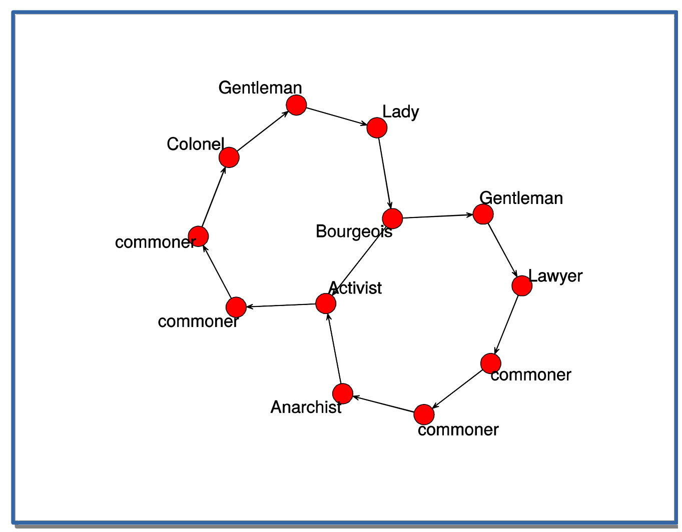
Likelihood of an observed graph

Given |N|=12 and |E|=13 (directed)

- There are $2 \cdot \binom{12}{2} = 132$ options for edge placement
- Edges are distributed independently
- So the number of possible graphs is $\binom{132}{13} = 3.22 \times 10^{17}$
- All of these graphs are equally likely...







Erdos-Renyi networks: A generative-probabilistic approach

- We assume **edge independence.**
- Edges are generated by a **Bernoulli process** with a parameter *p.*
- We generate a graph $G(N, p)$ by:
  - For each ordered pair (u, v) of nodes from *N:*
    - E += (u, v) with a probability p
- Each graph with n nodes and m edges has the following likelihood:

$$p^m \cdot (1-p)^{2\cdot\binom{n}{2}-m}$$

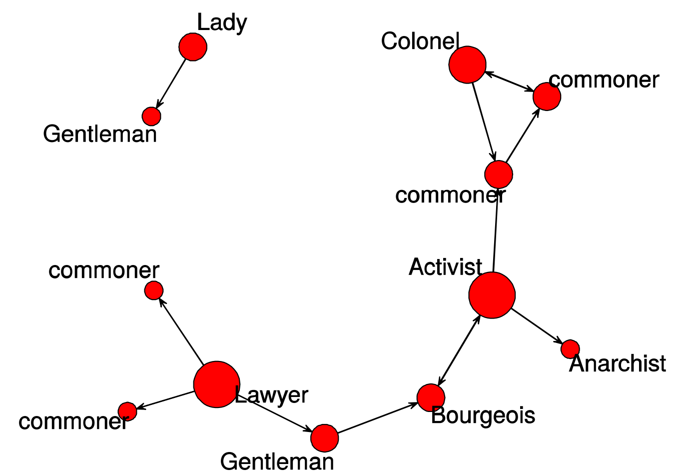
Terminology: “Graphs” vs. “Networks”

[informal:]

- Graphs are mathematical (topological) concepts defined by nodes and edges.
- [Social] Networks represent the outcome of some social ***process*** (can be dynamic)
- In networks we care about ***dependency*** between nodes and edges.

Networks are not "random" graphs

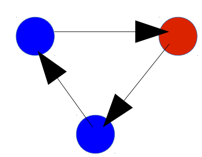
- **Goal**: Find a plausible (and interesting) model explaining the creation of an observed network.
- Assumptions (for simplicity):
  - Observed network is ***fixed***.
  - Edge formation is ***not*** random
  - Network was generated based on latent factors
  - We can speculate about the factors:
    - Common sense
    - Social science theory
    - Guess in the wild



Exponential Random Graph Models (ERGM)

- General form: $$Pr(Y=y)=\left(\frac{1}{k}\right) \exp \sum_A \theta_A g_A(y)$$
- Where:
  - A is a specific "configuration" (e.g. reciprocity)
  - θ_A is a parameter corresponding to configuration A.
  - $g_A(y)=\prod_{y_{ij}\in A} y_{ij}$ the network statistic corresponding to A.
    - $y_{ij}\in\{0,1\}$ 1 iff the ij edge is observed in y.
    - For simplicity we generalize g and A (sum on edges inst. mult)
  - k is a normalization factor, making the general form a proper probability distribution.

Example of some model features

- General form: $Pr(Y=y)=\left(\frac{1}{k}\right) \exp\sum_A \theta_A g_A(y)$
- Baseline model (Erdos Renyi): $Pr(Y=y)=\left(\frac{1}{k}\right) \exp\sum_{ij} \theta\, y_{ij}$
- Examples for other terms:
  - Formal leadership (nodal): $\sum_{ij,\, j\in Leaders} \theta_{leadership}\, y_{ij}$ 
  - Reciprocity (dyad): $\sum_{ij} \theta_{reciprocity}\, y_{ij}\, y_{ji}$ 
  - Cyclic triad (dyad): $\sum_{ijk} \theta_{cTriad}\, y_{ij}\, y_{jk}\, y_{ki}$ 
- So "simple" toy model to estimate:

$$Pr(Y=y)=\left(\frac{1}{k}\right) \exp\sum_A \theta_A g_A(y)=\left(\frac{1}{k}\right) \exp\left(\sum_{ij} \theta\, y_{ij}+\sum_{ij,\, j\in Leaders} \theta_{leadership}\, y_{ij}+\sum_{ij} \theta_{reciprocity}\, y_{ij}\, y_{ji}+\sum_{ijk} \theta_{cTriad}\, y_{ij}\, y_{jk}\, y_{ki}\right)$$

Parameter estimation

- Markov Chain Monte Carlo
  - Metropolis Hastings
  - [There are other algs + new developments]
- Issues:
  - Degeneracy
  - Stability (over subsampling, incomplete networks, thresholding)
  - No direct temporal modeling
  - Not suitable for large network (estimation is problematic)
  - Interpretation needed

ERGM resources

- MCMC estimation of ERGMs
  - http://www.cmu.edu/joss/content/articles/volume3/Snijders.pdf
- R packages: statnet, network, ergm
- ERGM introduction, package documentation and examples
  - https://cran.r-project.org/web/packages/ergm/vignettes/ergm.pdf
- [New] Generalized-ERGM (+beta implementation)
  - http://arxiv.org/pdf/1505.04015.pdf
- Many other tutorials, variations and examples (online)

Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

Data

Political Twitter

- 6 month full stream
- 158817 tweets
- Graph is directed
- Edge threshold (@>3)
- 439 Members of 114th Congress (Current)
- |E| = 9167 (after thresholding)

Joint Statements (JS)

- Full term (112th congress)
- 8979 statements
- Graph is undirected
- Edge threshold: normalized-weighted
- 435 members of 112th Congress (2011-13)
- |E| = 3188 (after thresholding)

Model features (factors/terms)

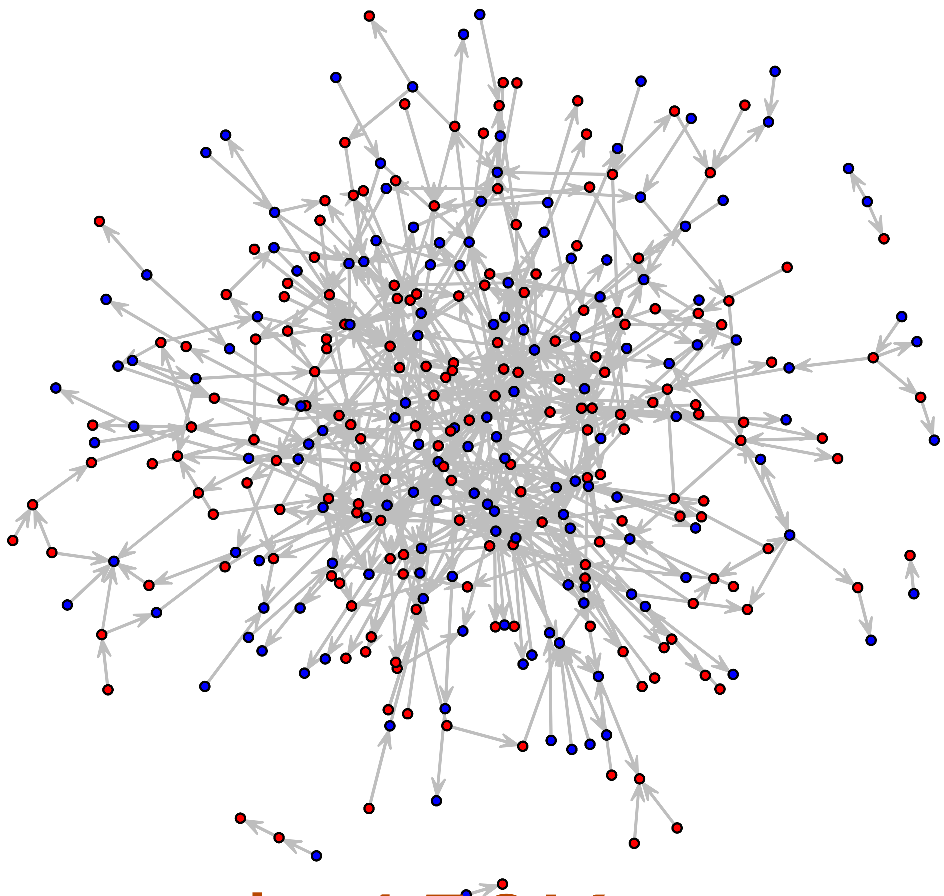
Nodal Factors

- Party
- Age
- Gender
- Seniority (terms in congress)
- State, region, district
- Formal leadership position
- Committee membership

Dyad Factors

- Reciprocity
- Cyclic triads
- Transitive triads
- Shared committee membership
- In/out-star

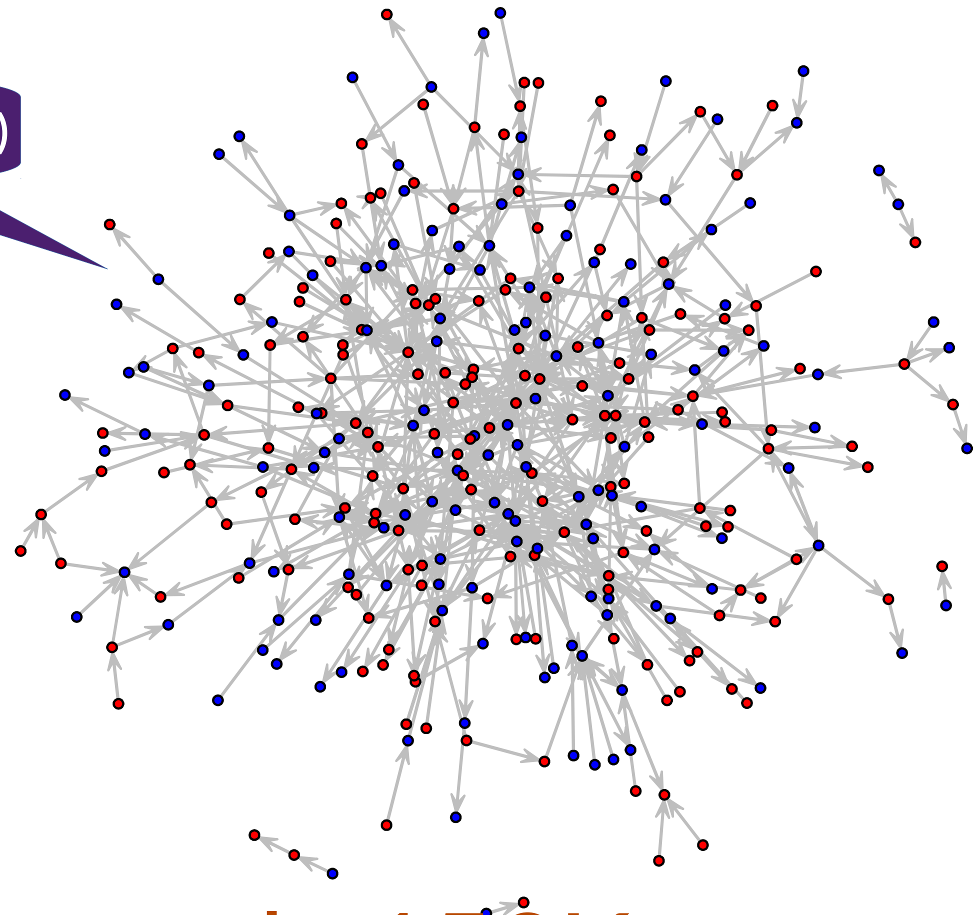
Members of U.S. Congress (Twitter)



6 month, 159K tweets
directed, by @, threshold=3

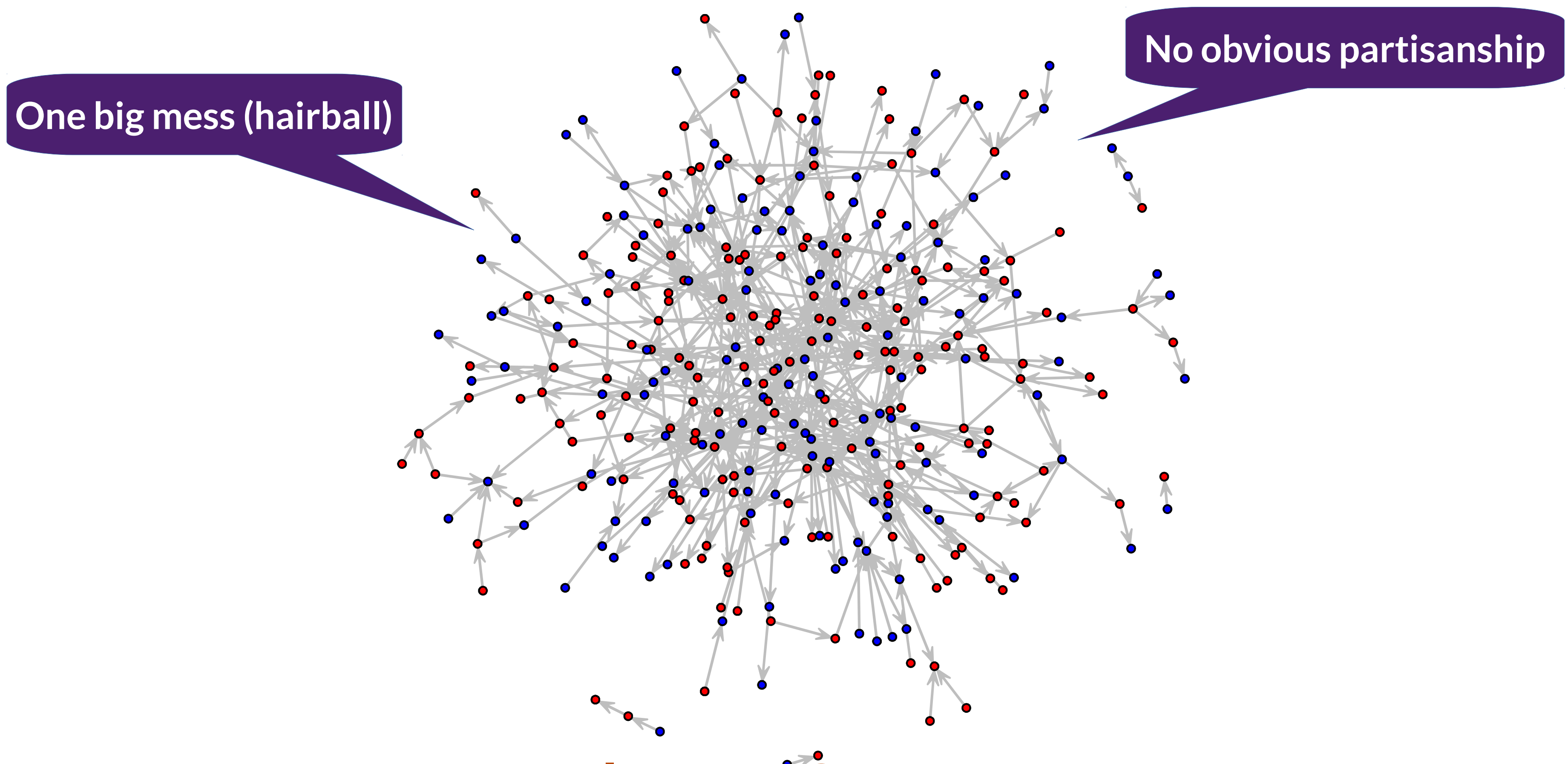
Members of U.S. Congress (Twitter)

One big mess (hairball)



6 month, 159K tweets
directed, by @, threshold=3

Members of U.S. Congress (Twitter)

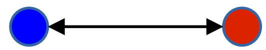
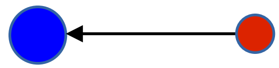
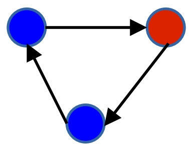


6 month, 159K tweets
directed, by @, threshold=3

Research questions

- What latent factors dominate link formation?
- Does network analysis fit with what we know (Political Science theory, other quant. works)?

ERGM results

- Significant:
  - **Reciprocity** matters 
  - **Seniority** matters 
  - **Cyclic triads** 
- Not found significant (surprising):
  - **Partisan homophily** 
  - **Formal leadership role** 
  - **We checked for other terms, e.g.:**
    - **Gender**
    - **State**
    - **Region**
    - **Committee membership network**

ERGM results - Twitter

Significant factors in **nodal** (independent) model:

- Number of edges (Bernoulli)
- Seniority (senior members attract incoming nodes)
- Surprising: state, party, shared committees and formal leadership were not found significant.

Significant **dyadic** (dependent) factors:

- Reciprocity (could this be a bias of the @ mechanism?)
- 2-in-star
- Cyclic-triads
- Transitive-triads term yielded degenerate models
- Seniority significance disappeared after introducing dyadic factors
  - Probably covered by the 2-in-star

Interpretation (1)

- Seniority matters
- Reciprocity rules (in politics; in conversing; in life?)
- Leadership is not a factor (masked by seniority?!)

But:

- This is not aligned with the JS network
  - Lack of seniority is a significant factor (new members are connected)
  - Leadership is a significant factor
- **Why?** (Is there a political scientist in the room?)
  - Technical: different networks (directed, vs. undirected)
  - Social 1: different networks (“wild” vs. collaborative by definition)
  - Social 2: different social processes shape different network dynamics

Interpretation **(cont.)**

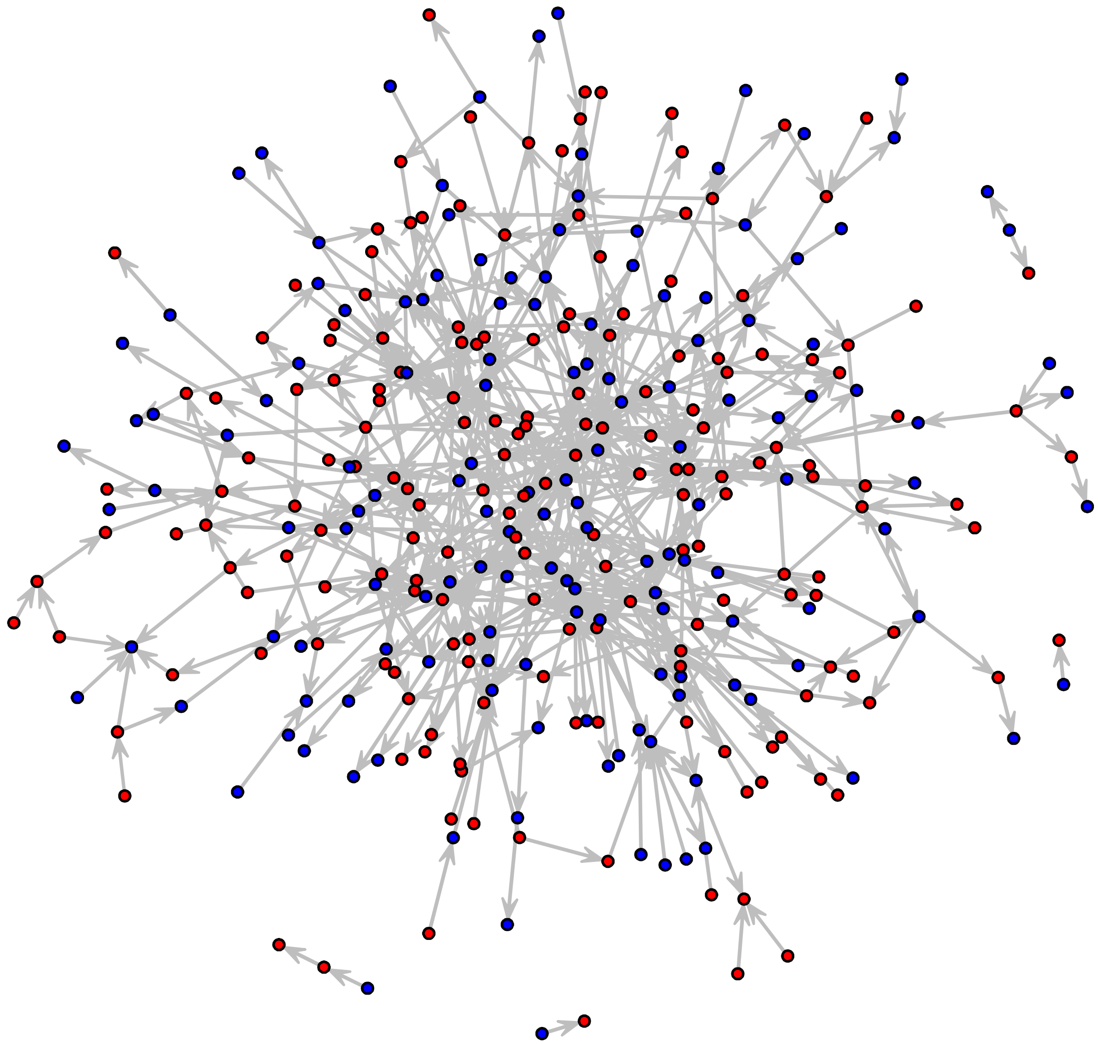
But

- Leadership (speakers, whips, majority/minority leader) has high/top centrality in relevant centrality measures (in/out/deg, betweenness)

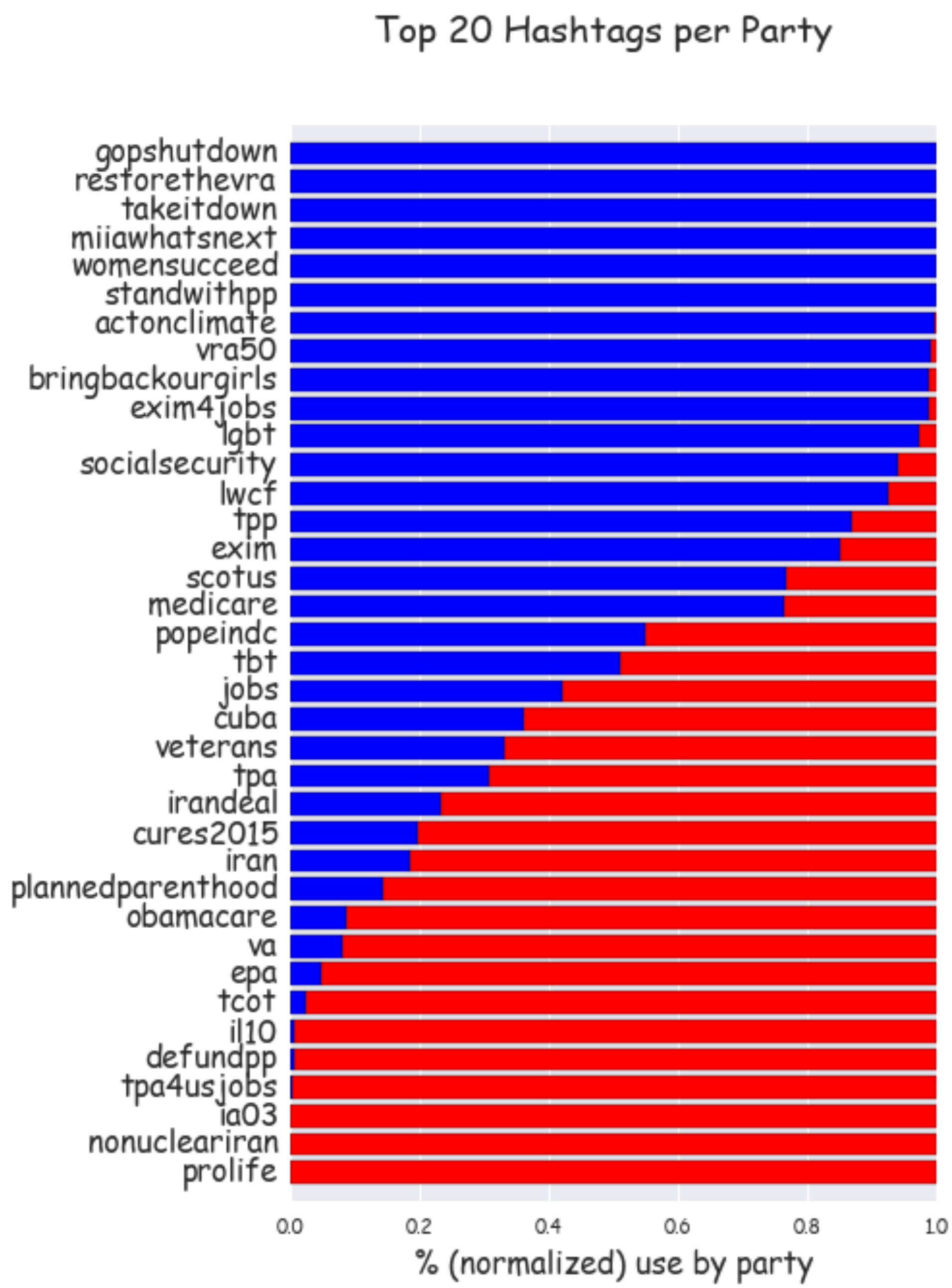
And

- In a frame of mind (**Tsur et al. ACL 2015**), we find:
  - strong partisanship even in subtle topics (=framing campaigns)
  - Strong party discipline (stronger for Republicans)

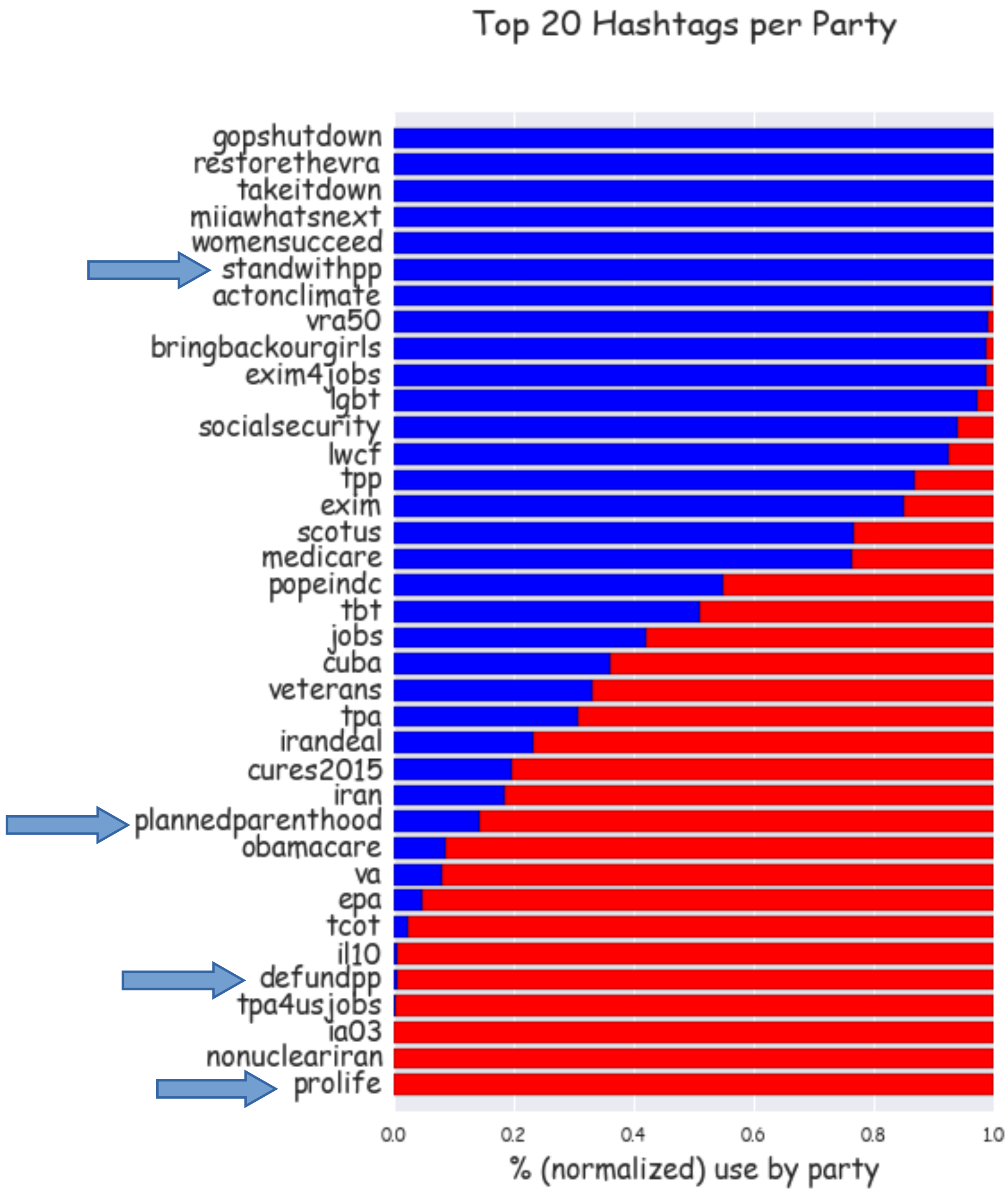
Living happily ever after?!



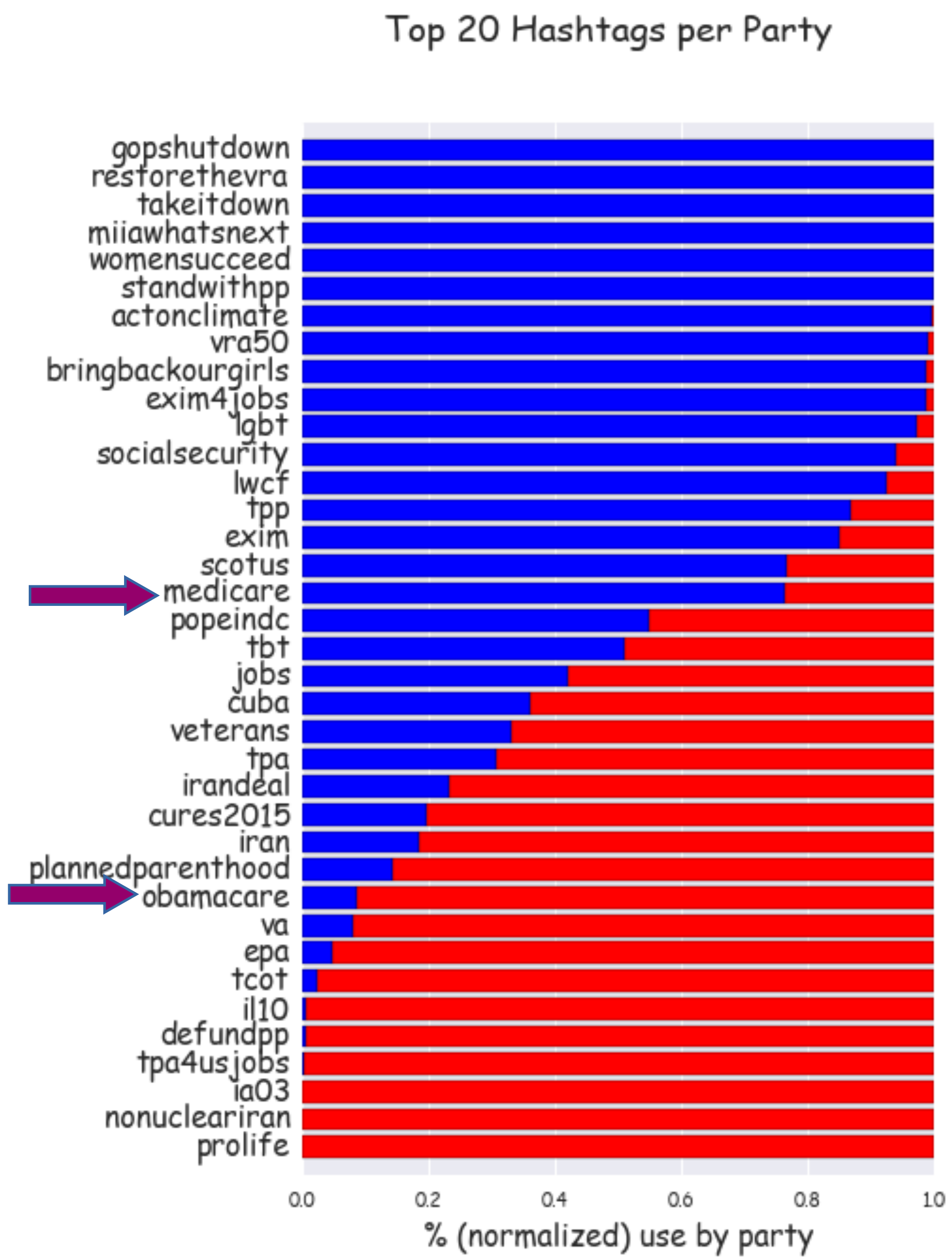
Partisan divergence and discipline



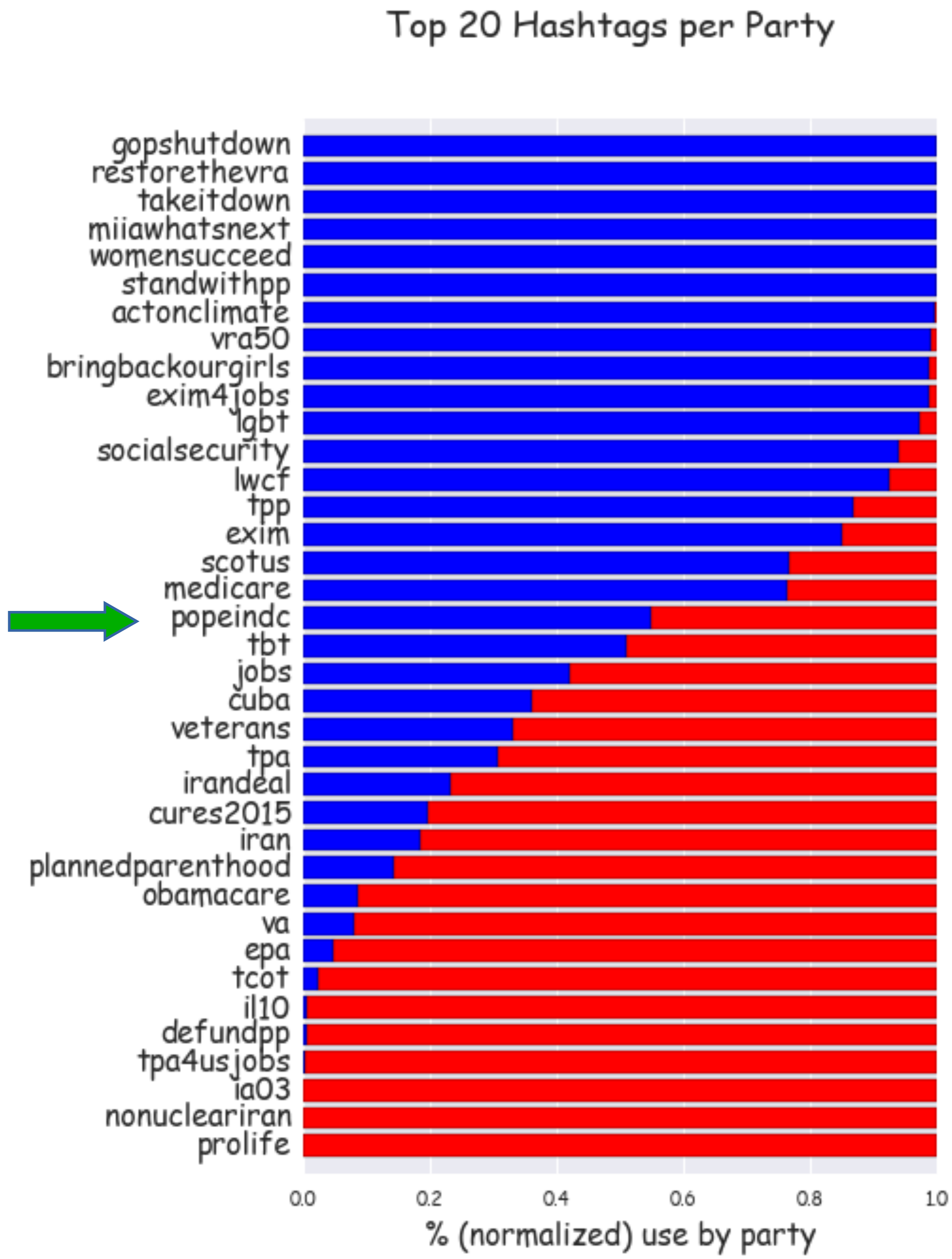
Partisan divergence and discipline



Partisan divergence and discipline

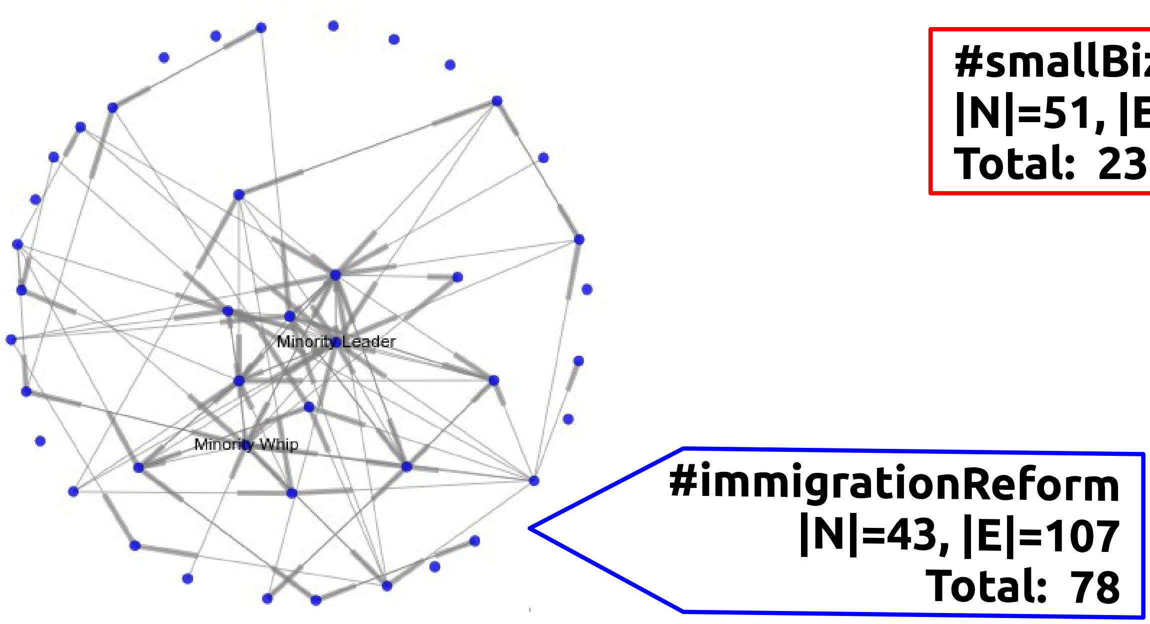


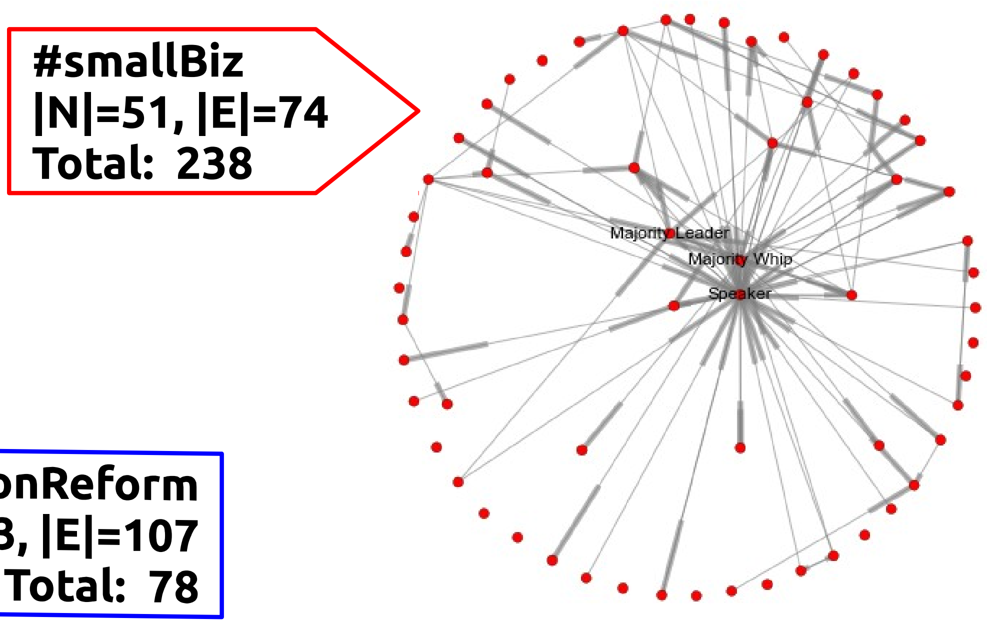
Partisan divergence and discipline



Leadership and sub-communities

- Leadership is central in the Twitter network.
- Even in marginal “campaigns”
  - Involving ~20% of the party members





Collaborative partisan hashtagging

%party users	ht_D	ht_R	sum_D	sum_R	avg_D	avg_R
[0.02, 0.05]	724	419	8154	6359	11	15
[0.05, 0.10]	179	110	5010	4875	27	44
[0.10, 0.20]	91	54	4897	5442	53	100
[0.20, 0.30]	48	15	5575	3843	116	256
[0.30, 0.40]	18	9	2718	3132	151	348
[0.40, 0.50]	14	3	3229	1837	230	612
[0.50, 0.60]	9	0	3141	0	349	0
[0.60, 0.70]	2	1	1706	1514	853	1514
[0.70, 0.80]	1	0	918	0	918	0
[0.80, 0.90]	0	0	0	0	0	0

- Democrat use more hashtags, less effectively
- Republicans have higher average uses per member in ALL ranges
- In line with findings at *Tsur et al. ACL 2015*

Workshop Announcements

2 WS on NLP and Computational Social Science (**NLP+CCS**):

- **WebSci** – Hannover, Germany, May 2016

  (deadline: March 25)
- **EMNLP** – Austin, Texas, November 2016

  (deadline: TBA)

Politics and networks

- **Political Networks** (PolNets) – St. Louis, Missouri, June, 2016

  Abstract based. (deadline: April 15)

Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

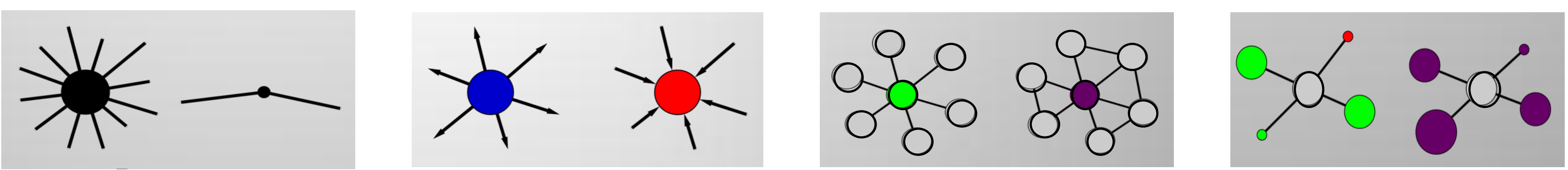
Roles in Socio-political Data

Tina Eliassi-Rad

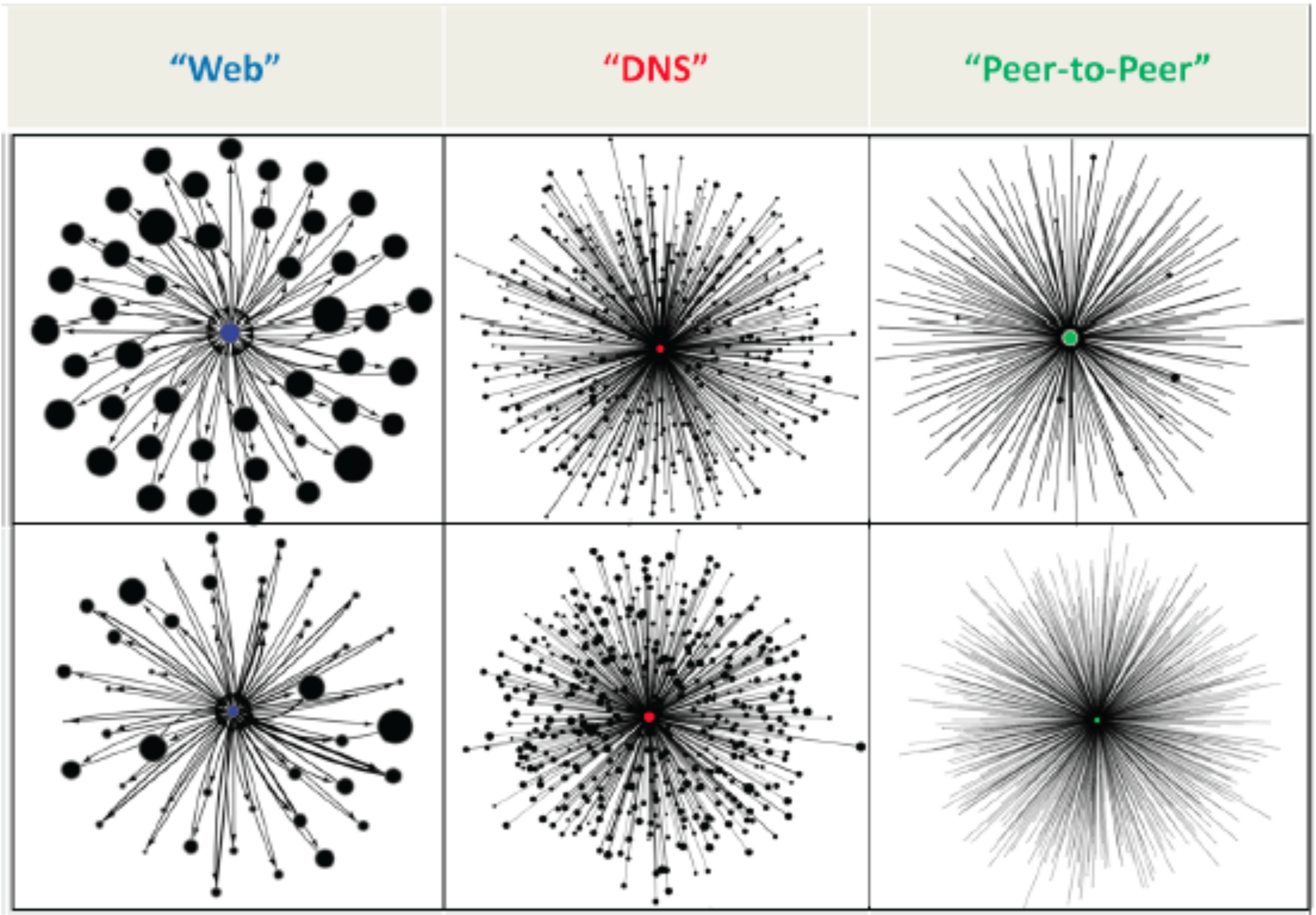
Northeastern/Rutgers

A network is an eco-system

- Individuals have a mixture of *roles* in this eco-system
  - Roles = functions = positions
- Roles are defined in terms of structural behaviors
  - What is your connectivity pattern?
  - To what kinds of individuals are you connected?



Intuition: Types of neighbors matter



Node sizes indicate communication volume relative to the central node in each frame.

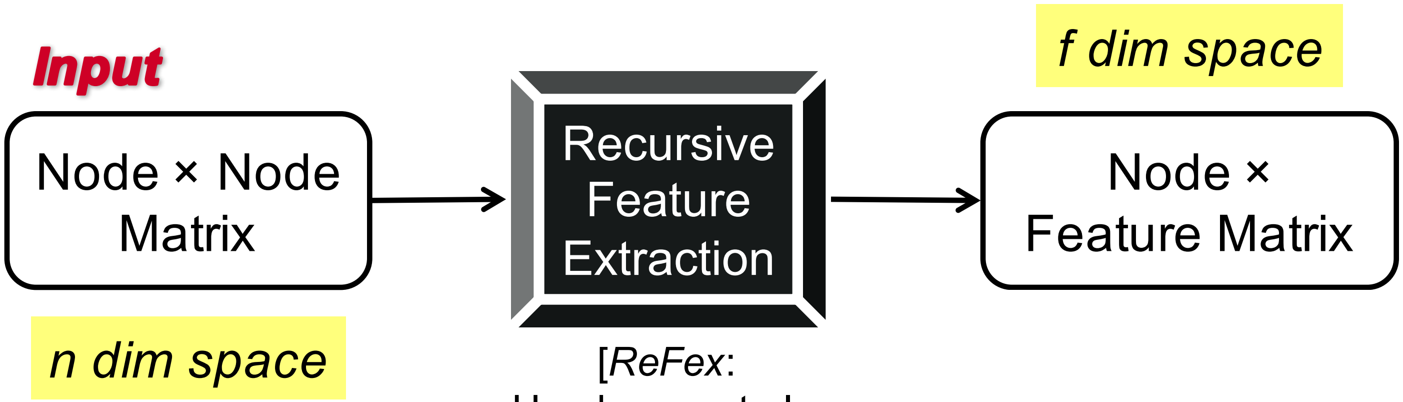
Finding roles in a network

Input

Node × Node
Matrix

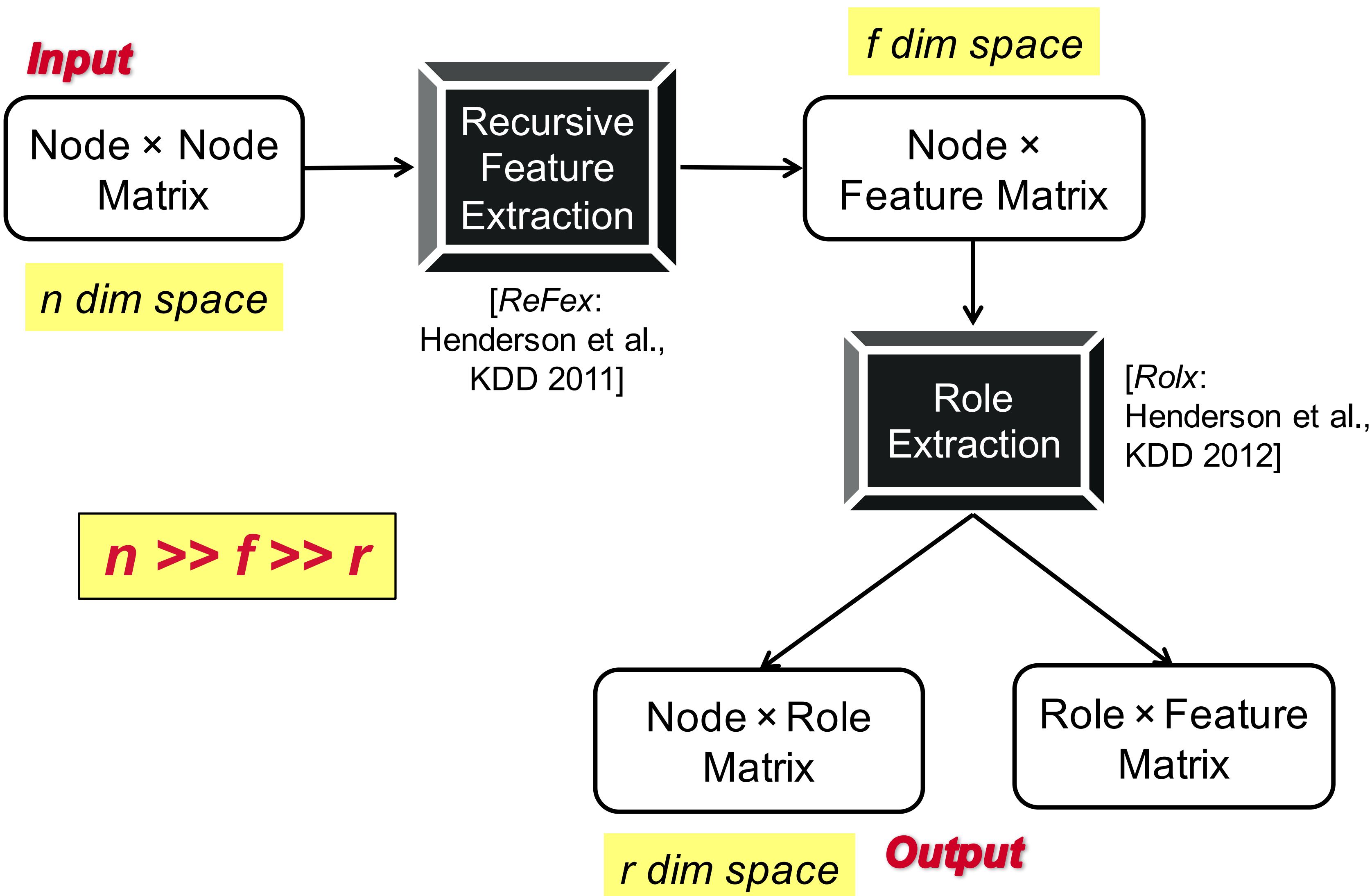
n dim space

Finding roles in a network

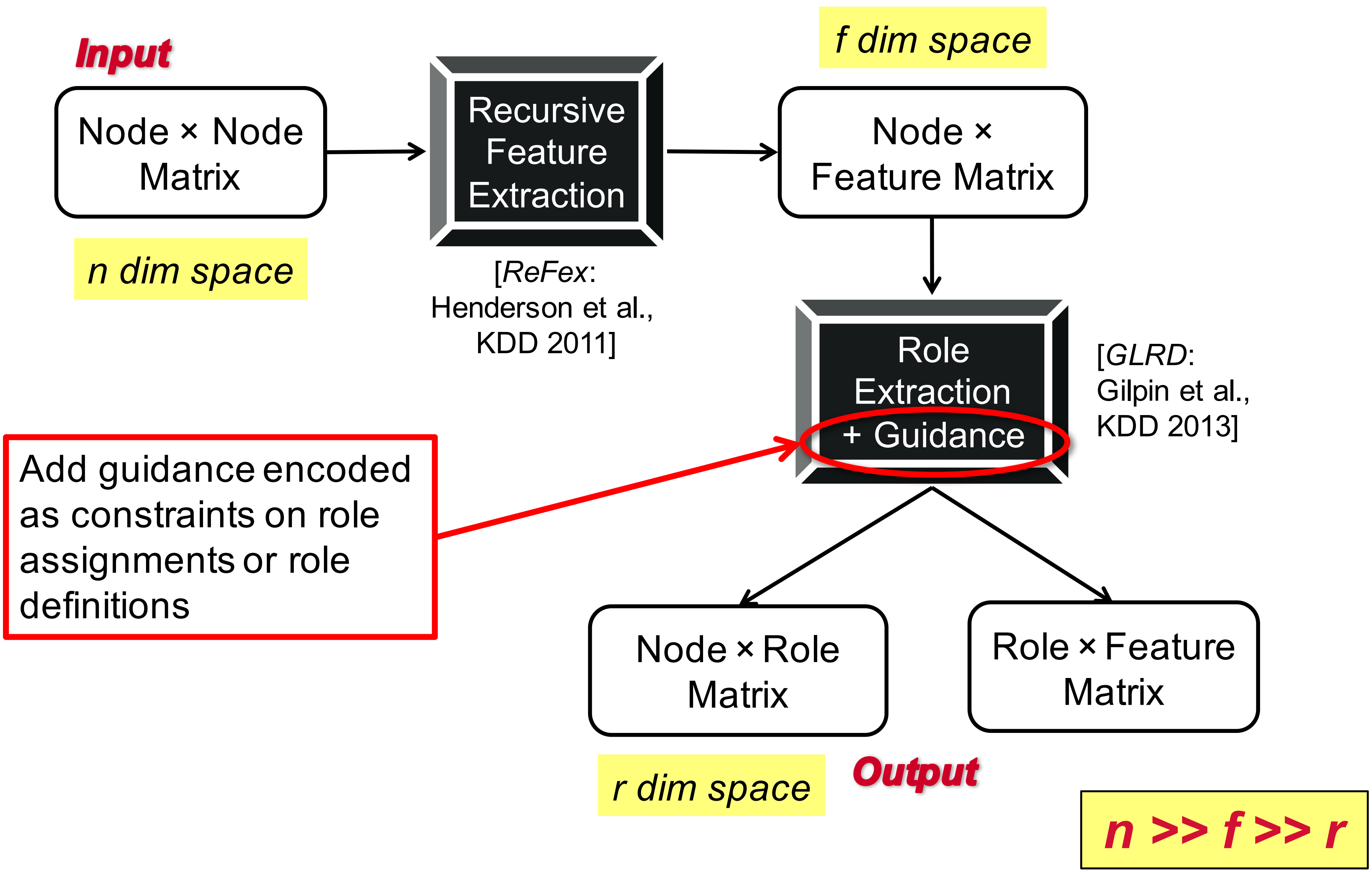


[*ReFex*: Henderson et al., KDD 2011]

Finding roles in a network

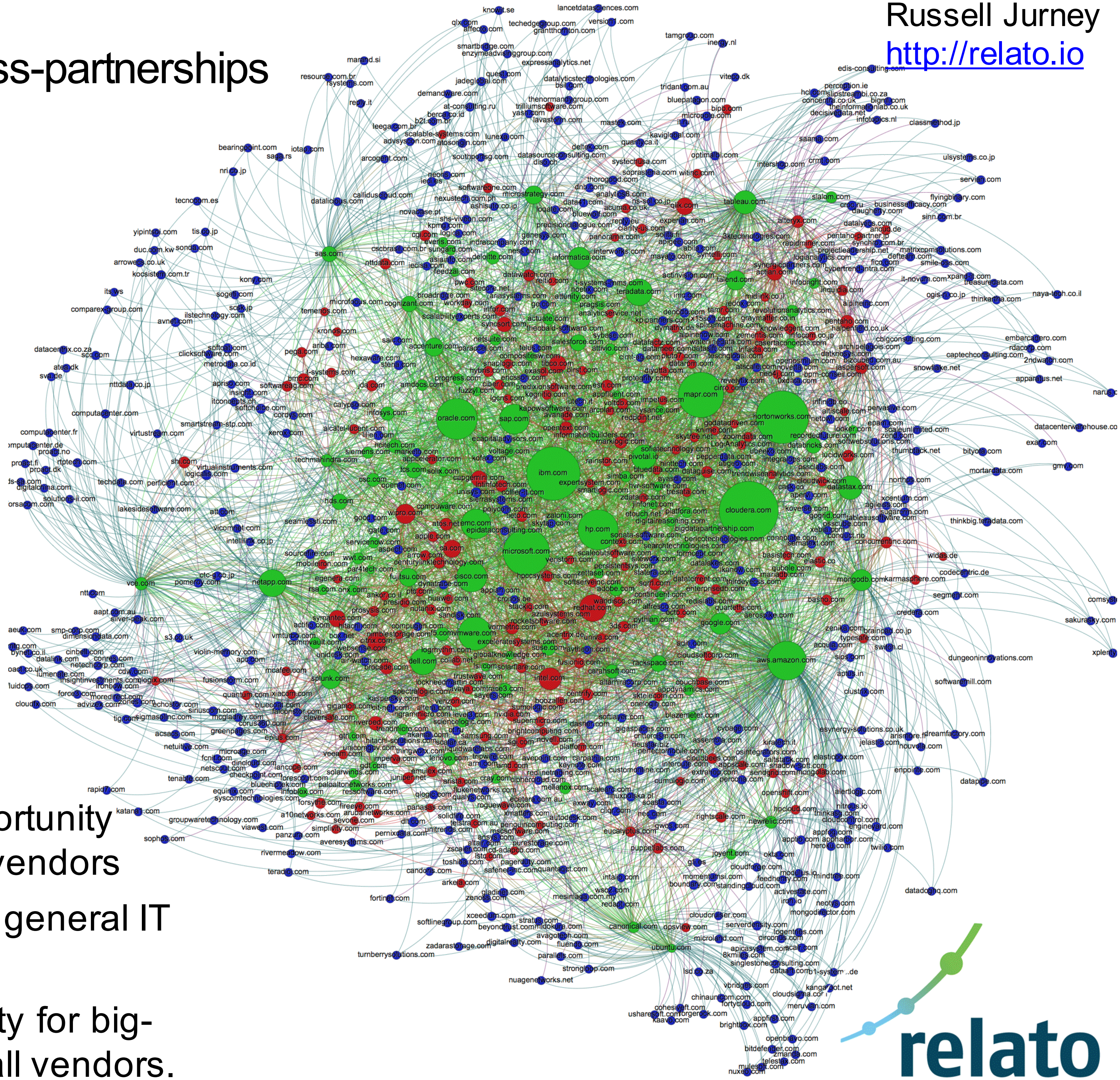


Finding roles in a network



Big-data business-partnerships

Russell Jurney
http://relato.io

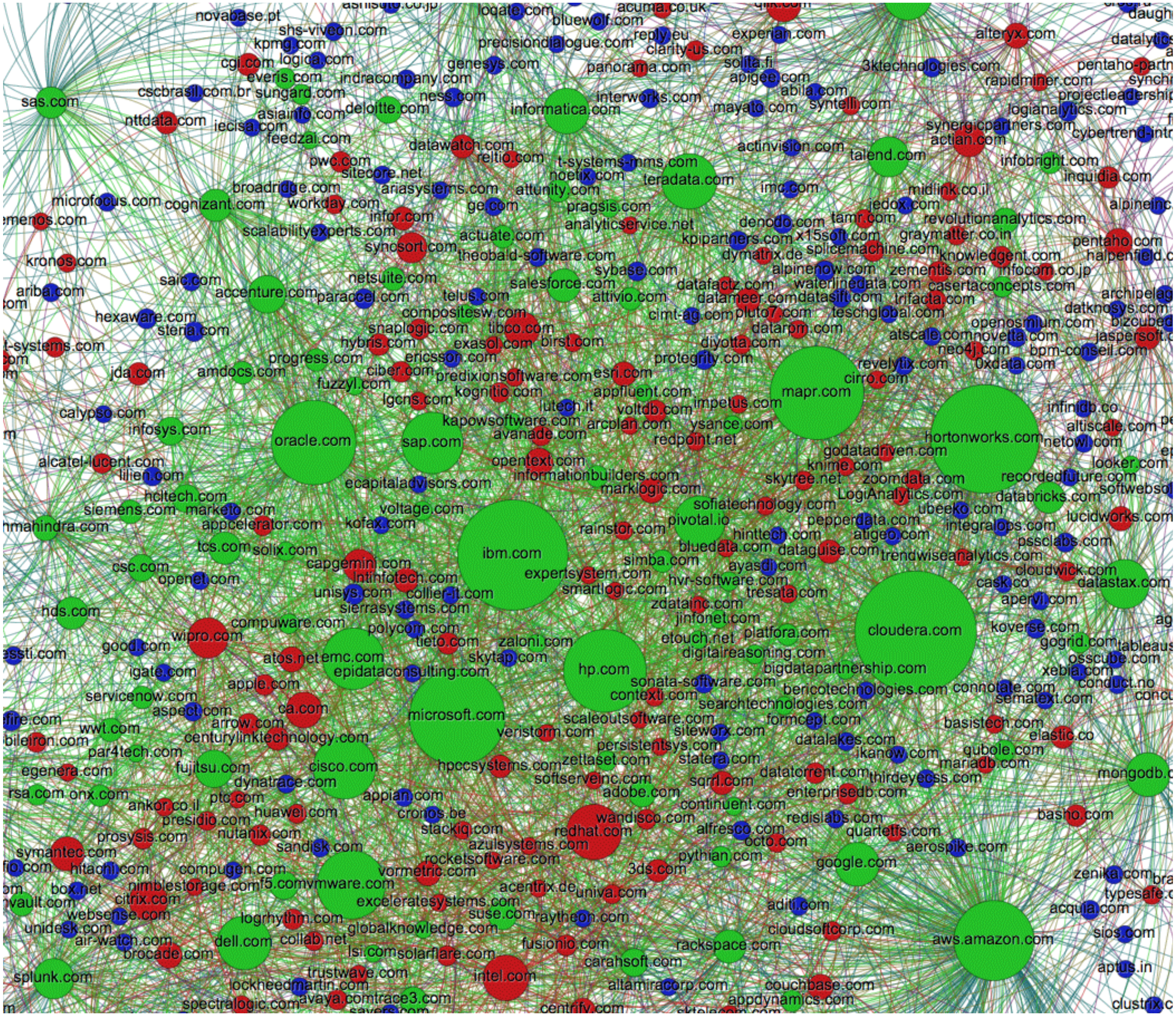


- **Green**: equal opportunity bridges; big-data vendors
- **Red**: middle-men; general IT vendors.
- **Blue**: Strong affinity for big-data vendors; small vendors.

relato

Russell Jurney
http://relato.io

Big-data business-partnerships



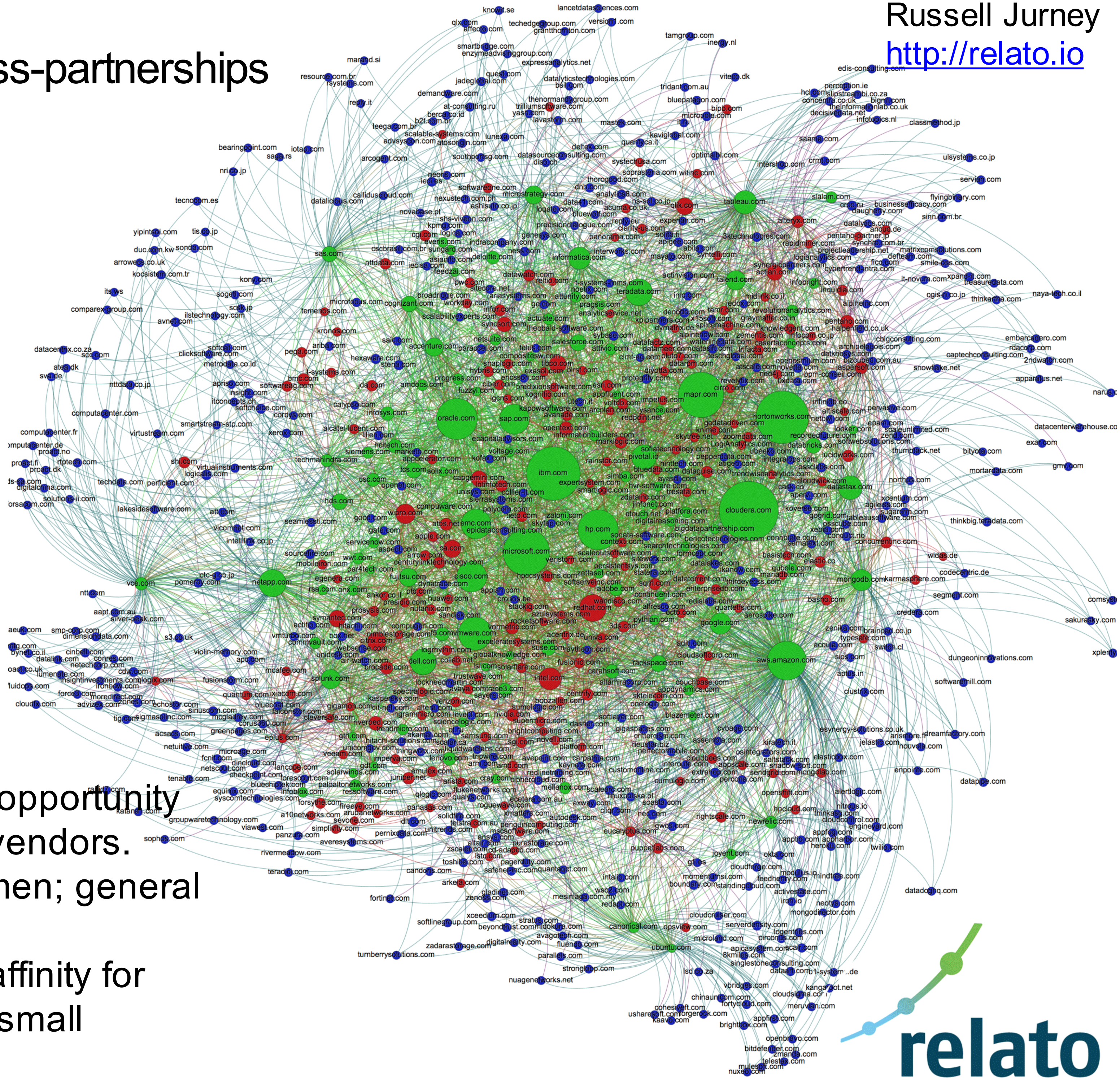
Roles & communities are complementary

- Roles group nodes with similar structural properties

- Communities group nodes that are well-connected to each other

Big-data business-partnerships

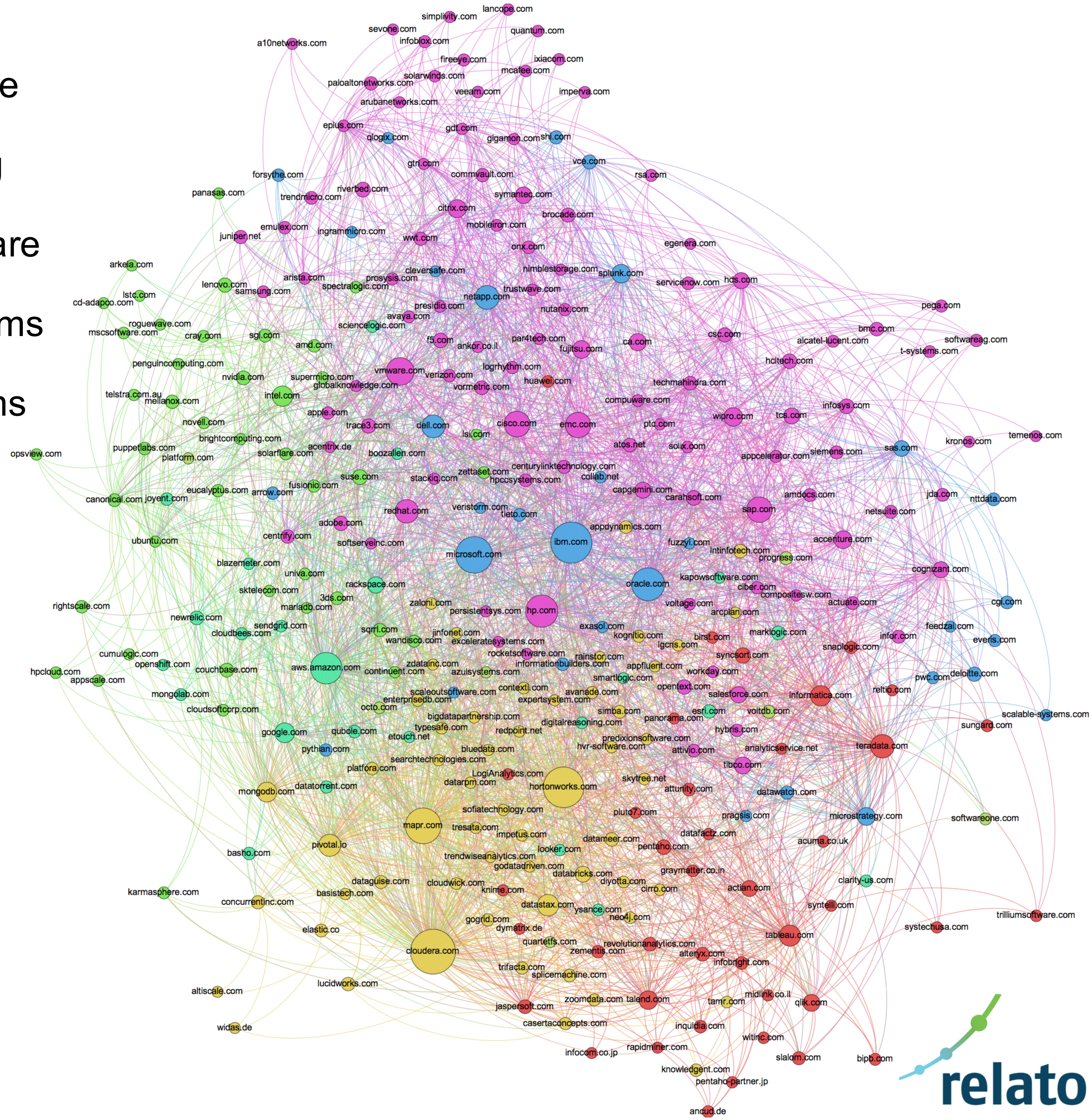
Russell Jurney
http://relato.io



- Green role: equal opportunity bridges; big-data vendors.
- Red role: middle-men; general IT vendors.
- Blue role: Strong affinity for big-data vendors; small vendors.

relato

- Analytics Software
- Cloud Computing
- Enterprise Software
- New Data Platforms
- Old Data Platforms
- Servers

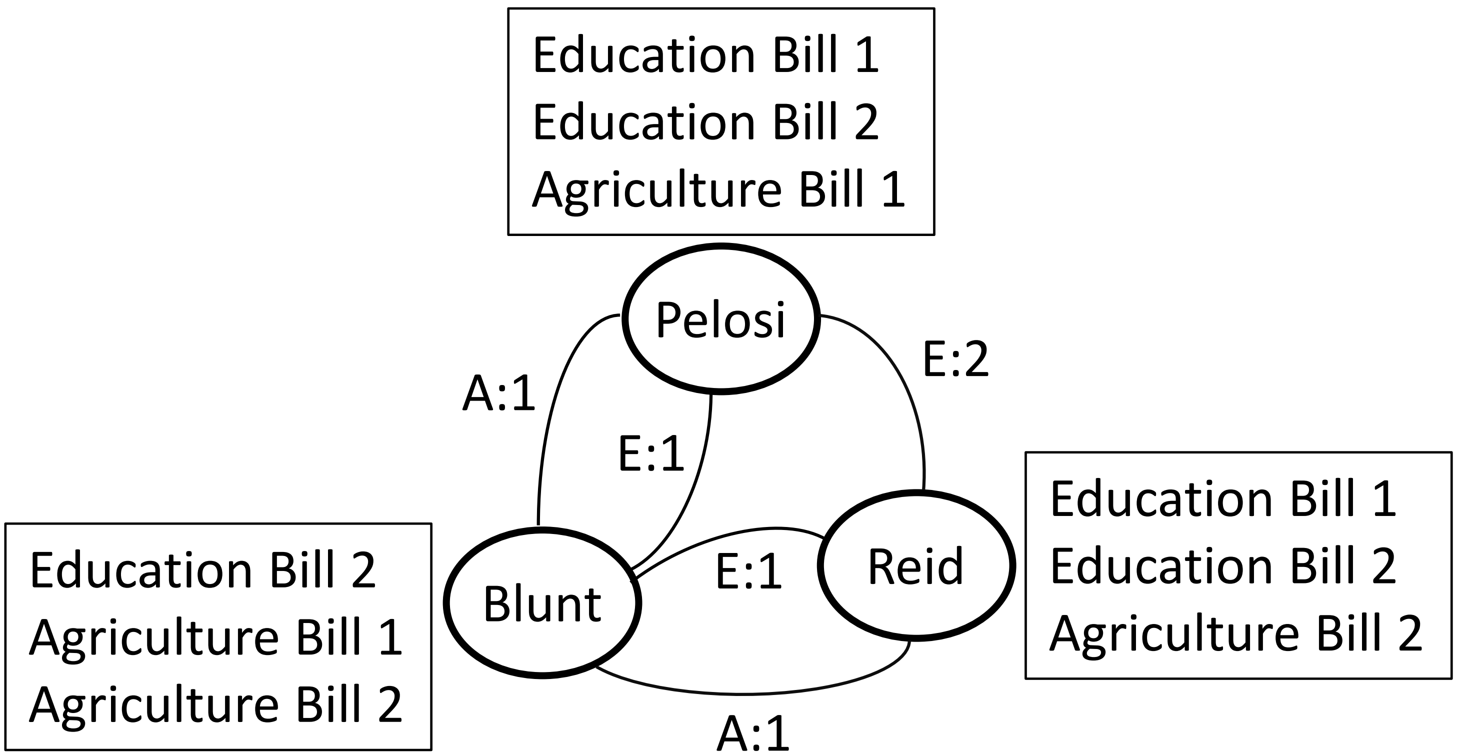


Russell Jurney

http://relato.io

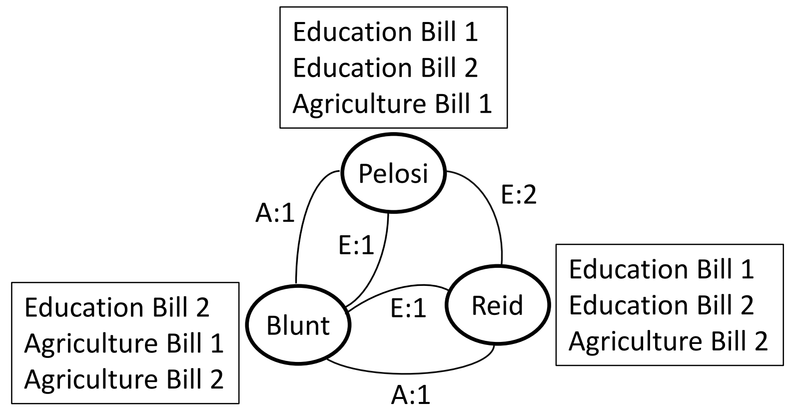
Moving beyond simple networks

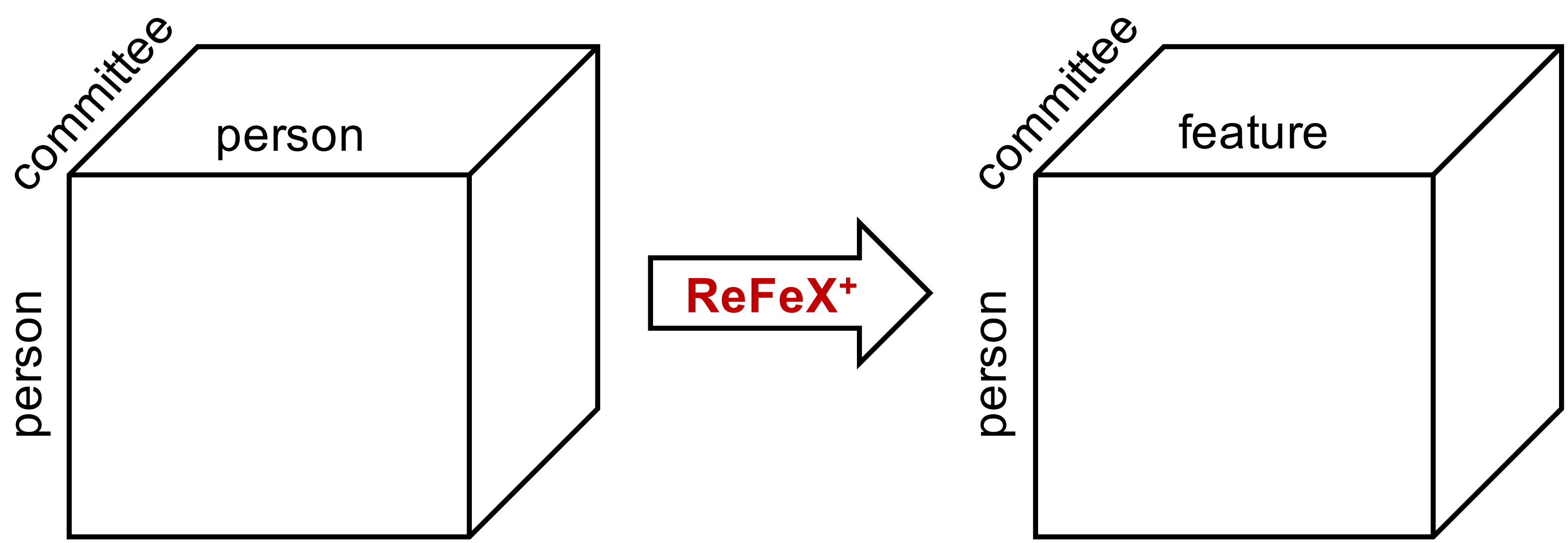
- Suppose you have a multi-relational networks
- Example: Congressional co-sponsorship data



No longer have an adjacency matrix

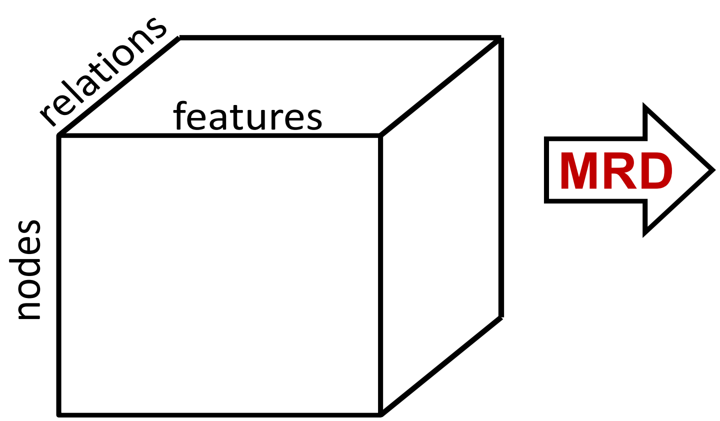
- We have a `person` × `person` × `committee` tensor
- Entry at (i , j, k) indicates how often congress-person i and j co-sponsored a bill that was sent to committee k for a particular congressional committee





Finding roles in a multi-relational network

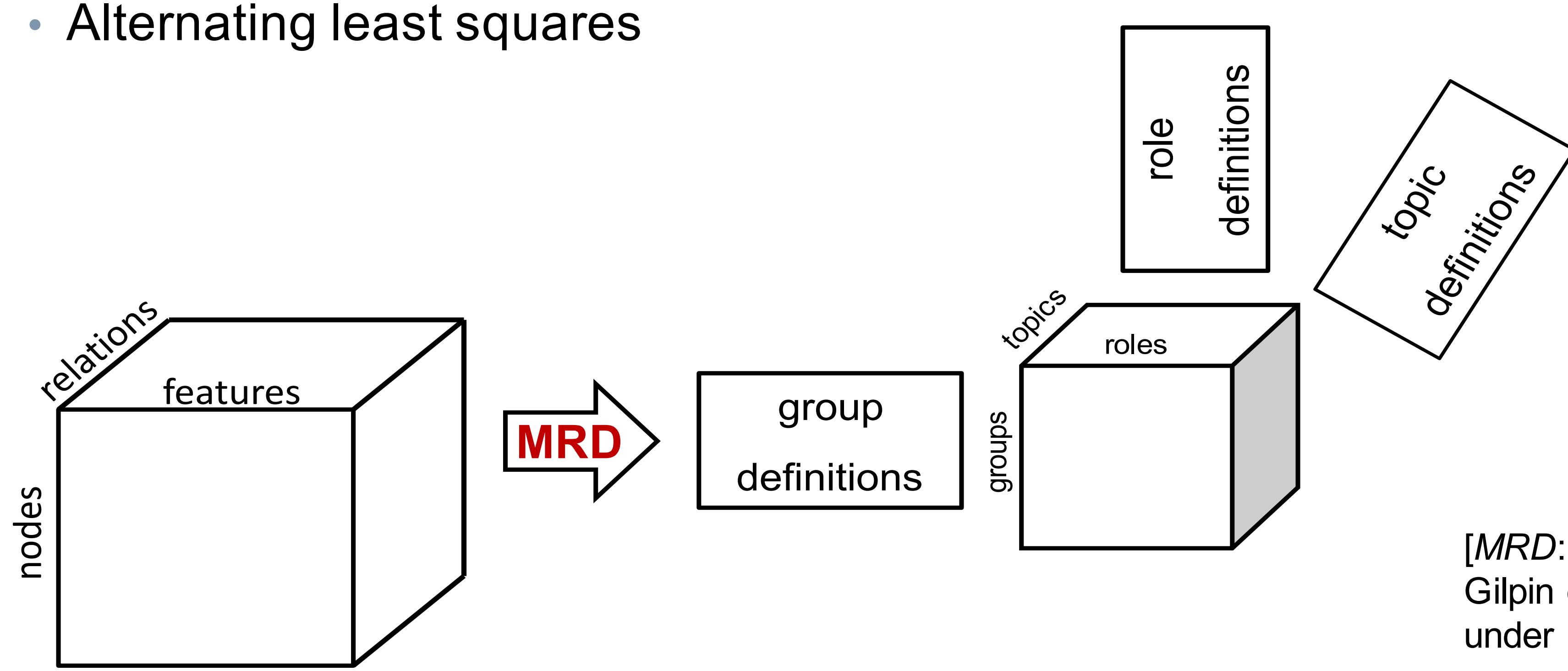
- *Multi-relational Role Discovery* (MRD)
  - No orthogonality constraint on factors
  - Nonnegative Tucker decomposition
  - Alternating least squares



[*MRD*: Gilpin et al., under review]

Finding roles in a multi-relational network

- *Multi-relational Role Discovery* (MRD)
  - No orthogonality constraint on factors
  - Nonnegative Tucker decomposition
  - Alternating least squares
- The factor matrices are:
  - groups of features (*role* definitions)
  - groups of entities (*groups*)
  - groups of relations (*topics*)
- Tucker core



[*MRD*: Gilpin et al., under review]

MRD Algorithm

Algorithm 1 Multi-relational Role Discovery (MRD) using Alternating Least Squares Nonnegative Tucker decomposition.

Initialize $\mathbf{G}, \mathbf{F}, \mathbf{R}$ and $\mathcal{H}$ to any non-negative values
while Stop condition not met **do**
    $\mathbf{G} \leftarrow \underset{\mathbf{G} \geq \mathbf{0}}{\operatorname{argmin}} \; \|\mathcal{V}_G - \mathbf{G}\mathcal{H}_G(\mathbf{R} \otimes \mathbf{F})^T\|_{Fro}$
    Normalize the columns of $\mathbf{G}$
    $\mathbf{F} \leftarrow \underset{\mathbf{F} \geq \mathbf{0}}{\operatorname{argmin}} \; \|\mathcal{V}_F - \mathbf{F}\mathcal{H}_F(\mathbf{R} \otimes \mathbf{G})^T\|_{Fro}$
    Normalize the columns of $\mathbf{F}$
    $\mathbf{R} \leftarrow \underset{\mathbf{R} \geq \mathbf{0}}{\operatorname{argmin}} \; \|\mathcal{V}_R - \mathbf{R}\mathcal{H}_R(\mathbf{F} \otimes \mathbf{G})^T\|_{Fro}$
    Normalize the columns of $\mathbf{R}$
    $\mathcal{H} \leftarrow \underset{\mathcal{H} \geq \mathbf{0}}{\operatorname{argmin}} \; \|\mathtt{vec}(\mathcal{V}) - (\mathbf{R} \otimes \mathbf{F} \otimes \mathbf{G})\mathtt{vec}(\mathcal{H})\|_{Fro}$
end while
return $\mathbf{G}, \mathbf{F}, \mathbf{R}, \mathcal{H}$

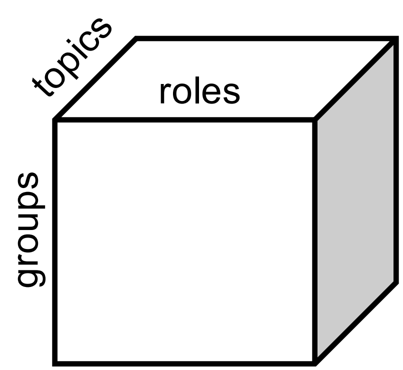
Experiments

- Data from U.S. House of Representatives
- Bill co-sponsorship data from 1979 (the start of the 96^{th} Congress) to 2009 (the end of the 110^{th} Congress)
- 15 committees, for which there were legislation in each congress from 96^{th} to 110^{th}
- 110^{th} Congress (from 2007-09)
  - 453 representatives & 10,613 bills
  - Average degree in aggregated graph = 8.37
  - Median value of average degree across committee co-sponsorship graphs = 0.48

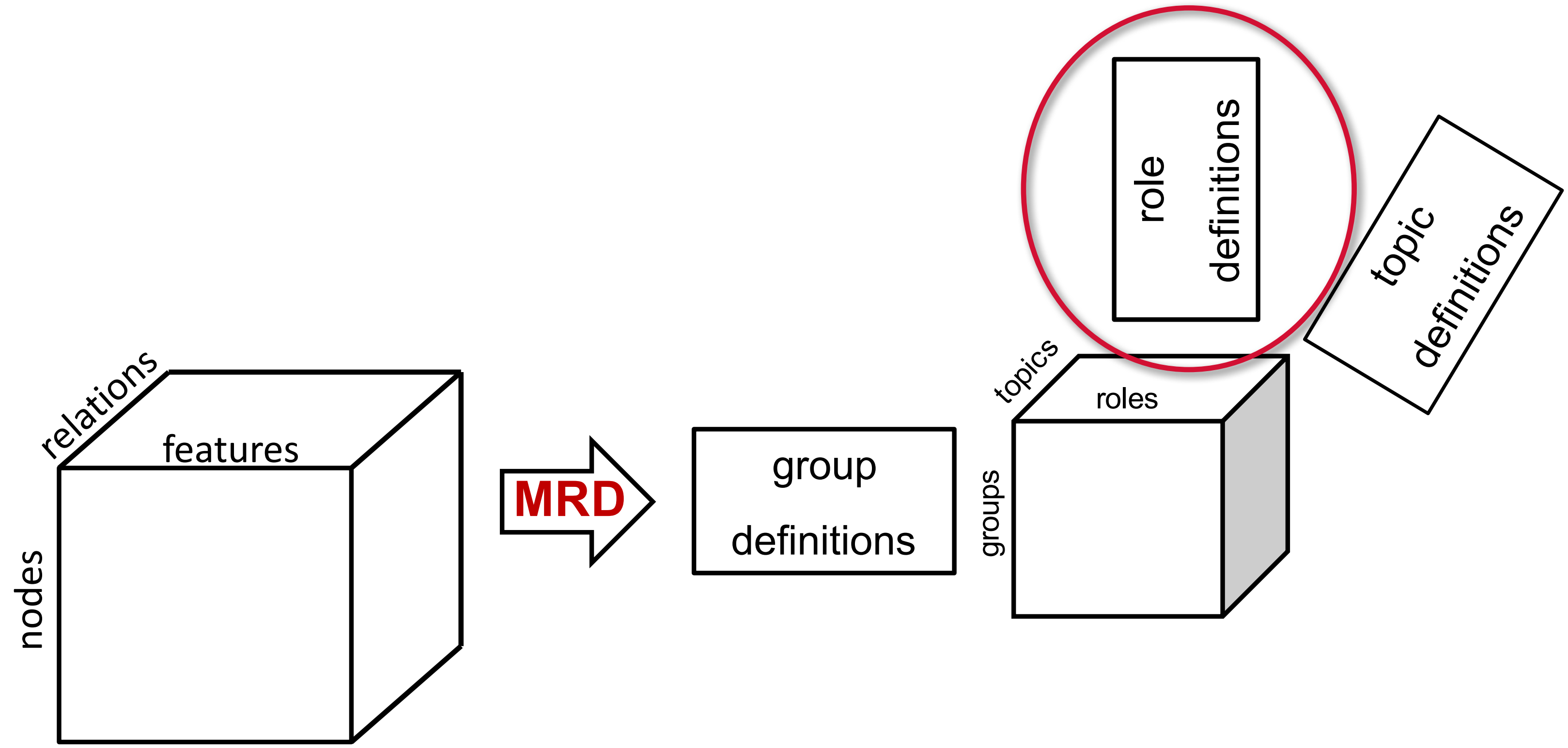
Sci & Tech
Judiciary
Ways & Means
VA
Small Business
Budget
Oversight & Gov't Reform
Agriculture
Appropriations
Rules
Natural Resources
Financial Services
Education & Labor
Transportation & Infrastructure
Energy & Commerce

Model order selection

- Can do model order selection with Tucker
  - Morten Morup and Lars Kai Hansen. 2009. Automatic relevance determination for multi-way models. *Journal of Chemometrics*, 23: 352–363.
    - Automatic relevance determination (ARD)
    - A Bayesian approach that estimates the adequate degree of regularization
- In these experiments, we set the model order to a 5 × 5 × 5 core



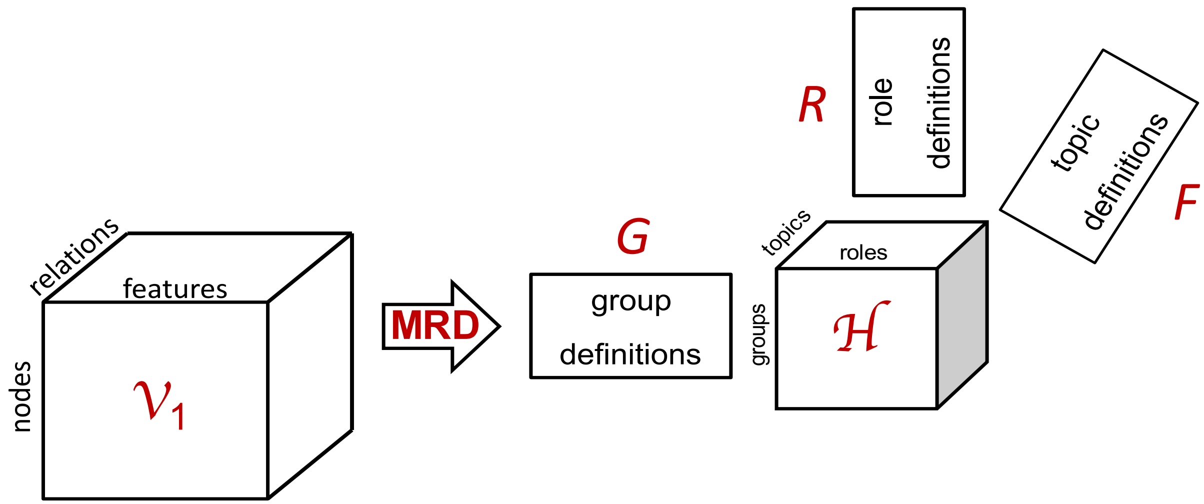
Role definitions



[*MRD*:
Gilpin et al.,
under review]

Role sense-making procedure

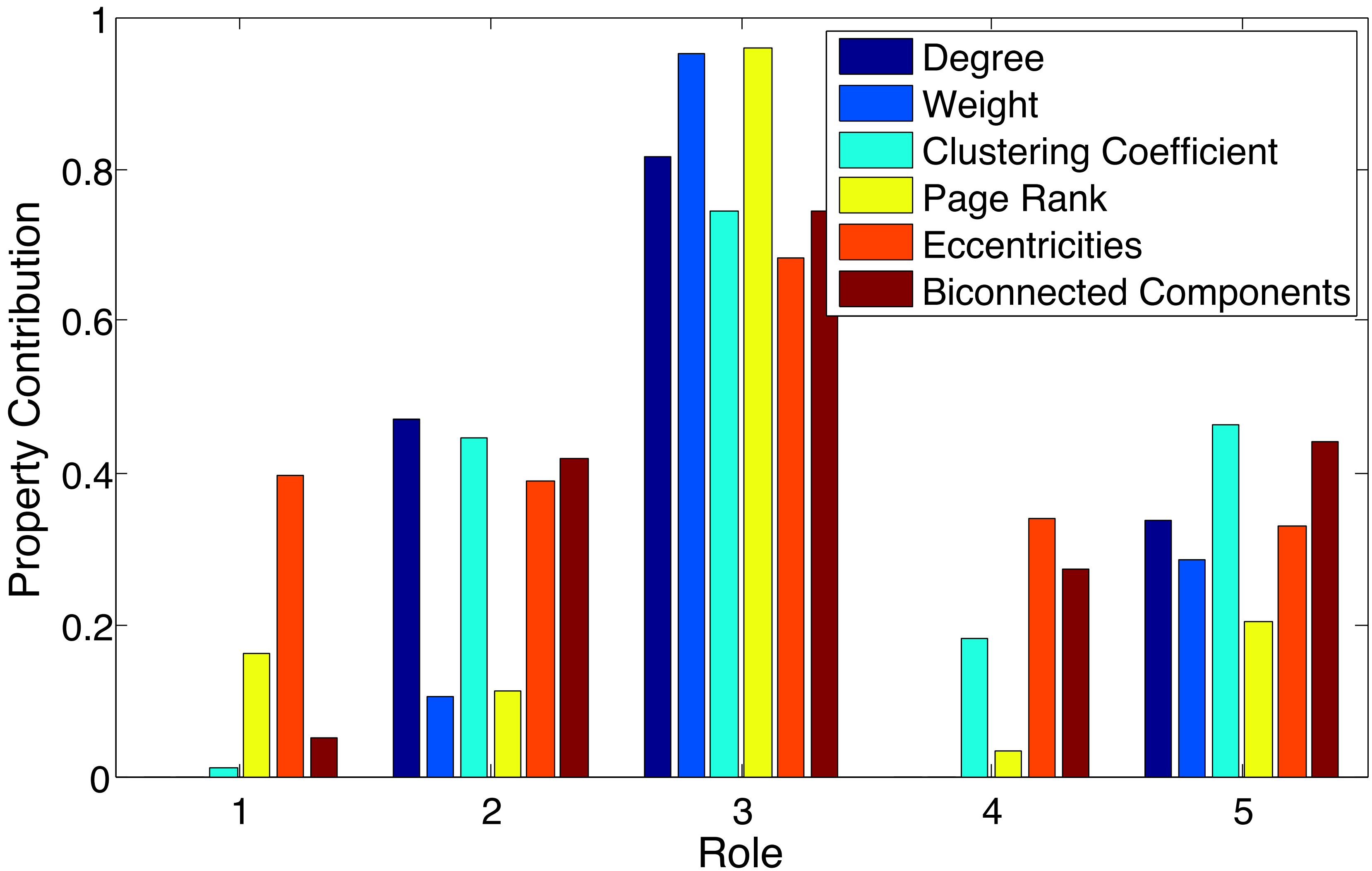
1. Run MRD to get the core and factor matrices: $\mathcal{V}_1$, $\mathcal{H}_1$, G_1, R_1, F_1.
2. Generate a new input tensor (nodes × relations × features), where the features are from a reference set of widely used and known features: $\mathcal{V}_2$.
3. Use $\mathcal{V}_2$, $\mathcal{H}_1$, G_1, and F_1 to compute a new R_2 role definitions that make “sense” to a human.



Output: R_2, where roles are redefined in terms of a set of reference features each of which is normalized for comparison purposes

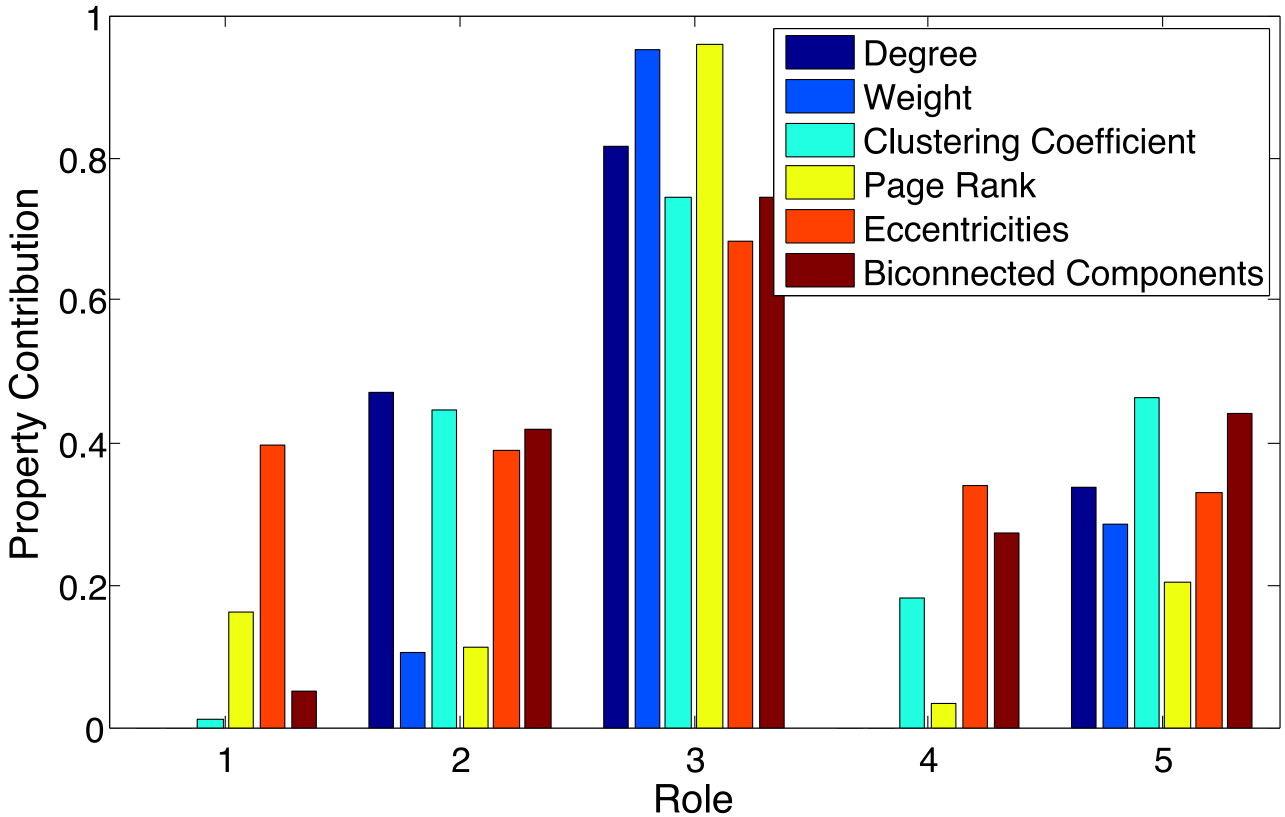
Role sense-making in the 110th Congress

- Role 3: Power brokers, high on every features



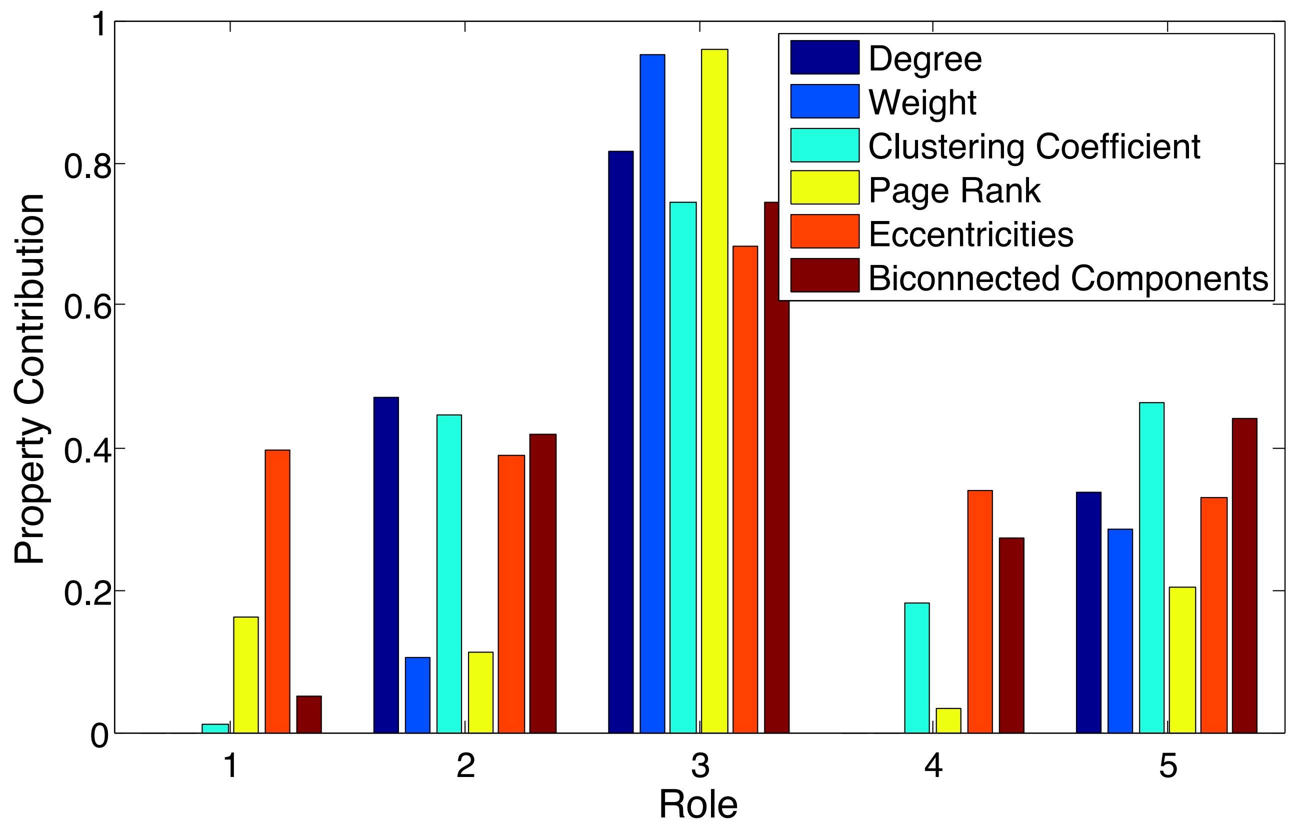
Role sense-making in the 110th Congress

- Role 1 & Role 4:
  - Both are path-y and on the periphery (high eccentricity values)
  - Both have very low degrees
  - But Role 4 nodes are more clique-y than Role 1 nodes (higher clust coeff) and less important (as measured by PageRank)

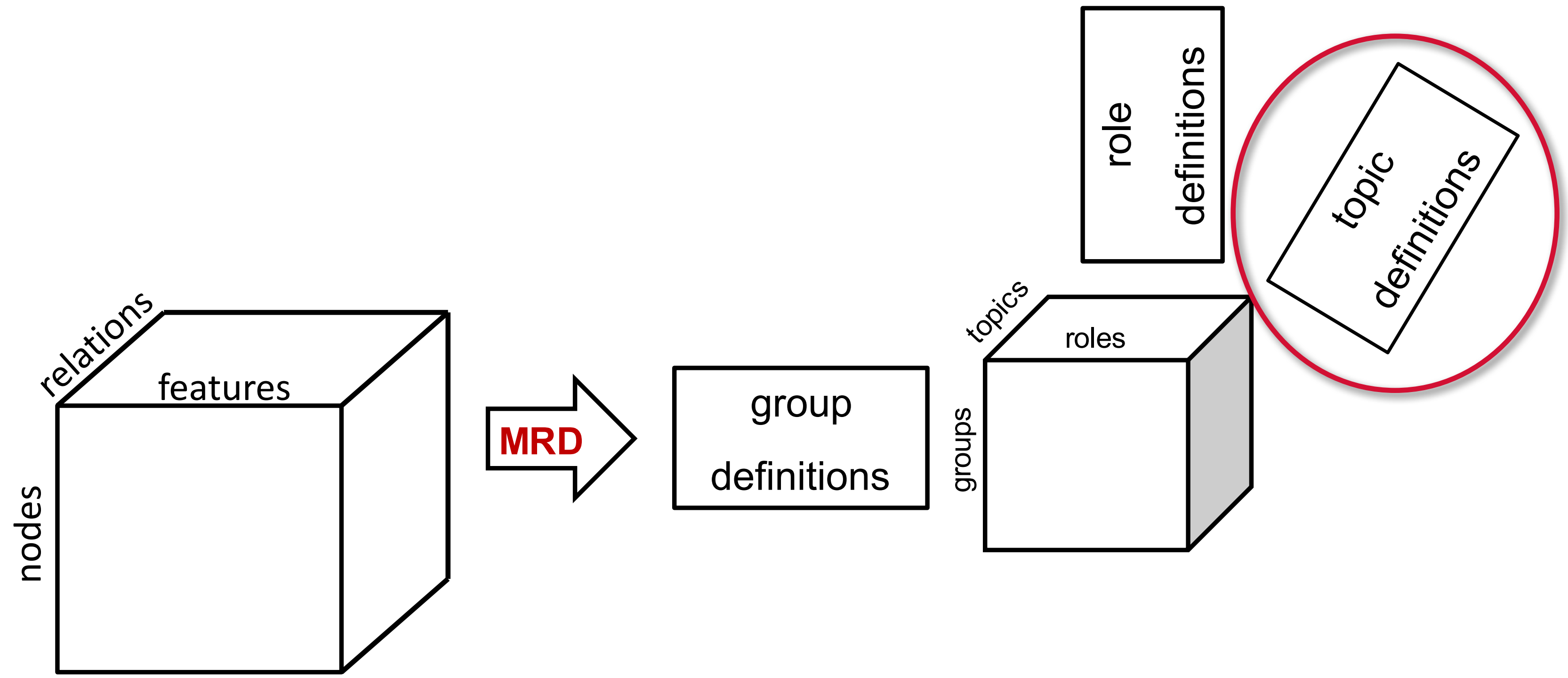


Role sense-making in the 110th Congress

- Role 2 & Role 5:
  - Both have high degrees and clust coeff
  - But Role 5 nodes have higher weight and higher PageRank
    ➔ Role 5 folks co-sponsor with the same people more often

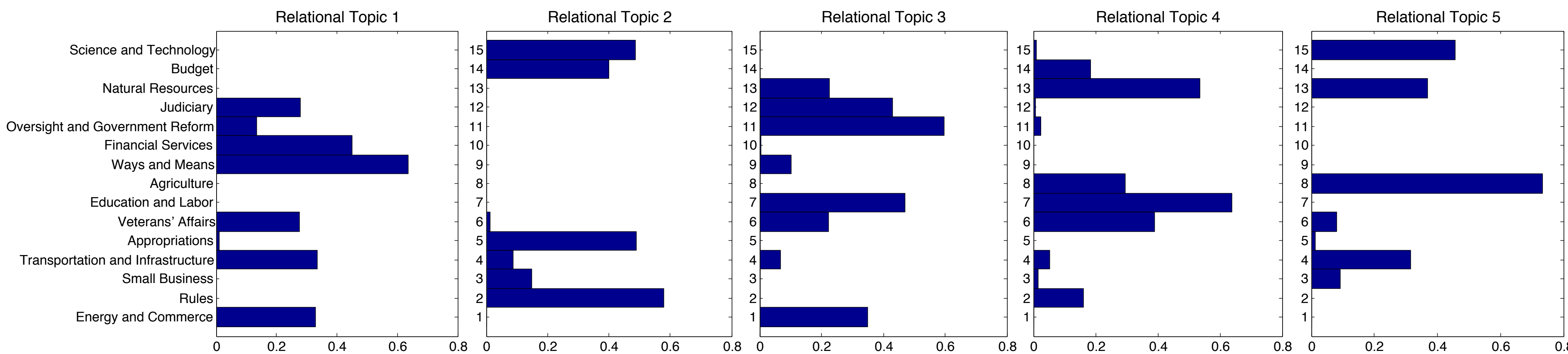


Relational topic definitions



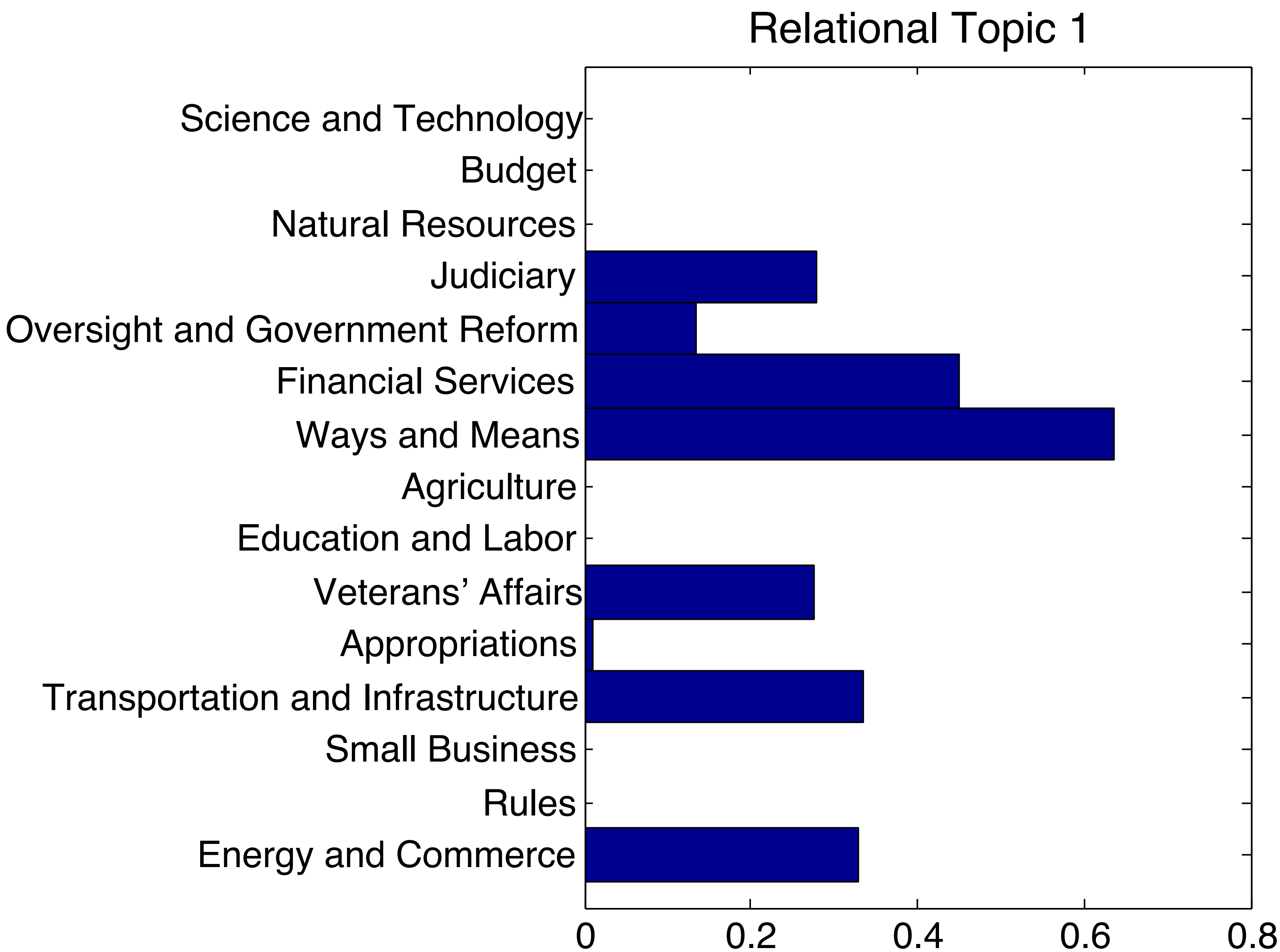
[*MRD*:
Gilpin et al.,
under review]

Relational topics found

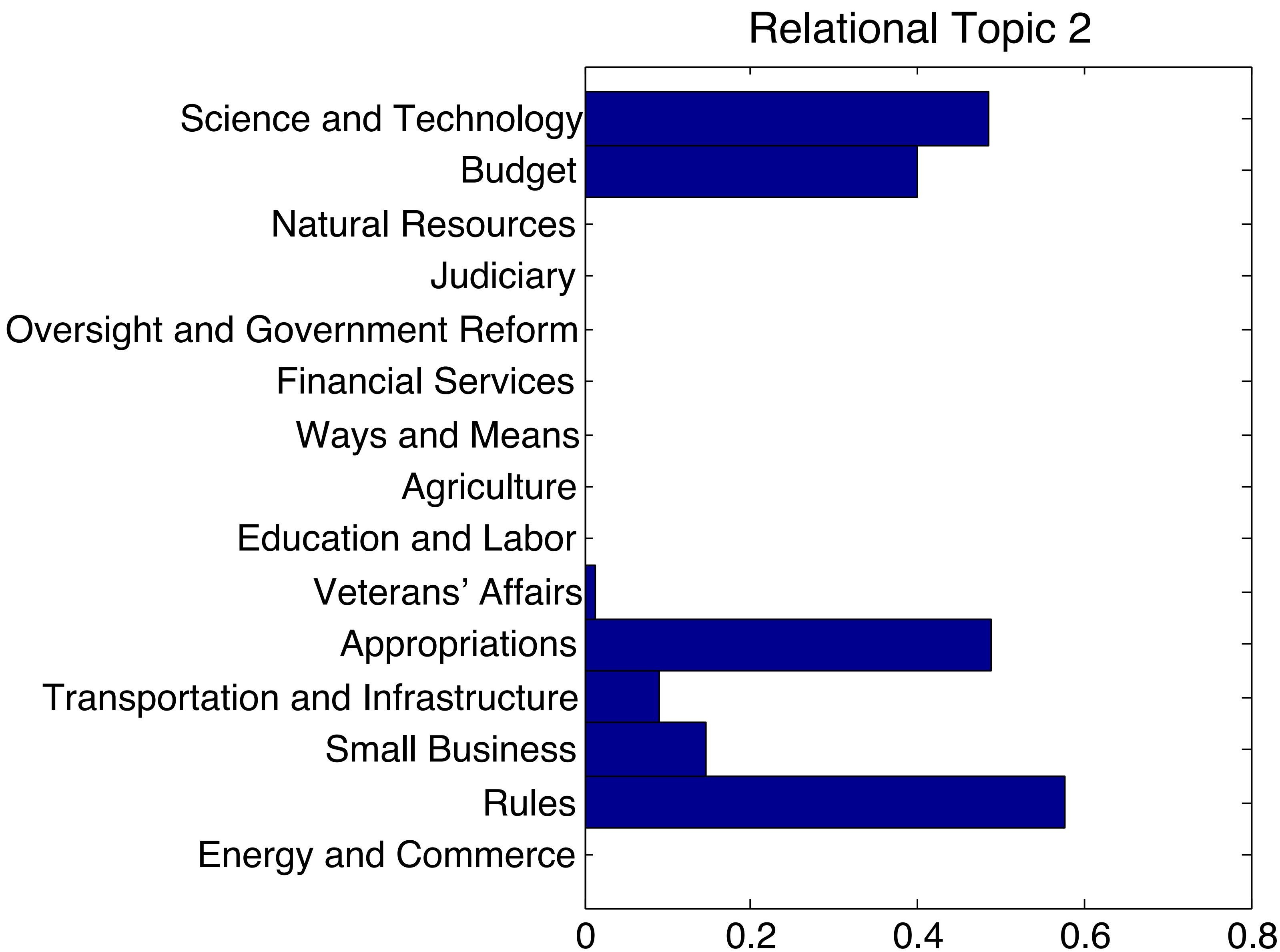


Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Ways & Means	Rules	Oversight & Gov't Reform	Education & Labor	Agriculture
Financial Services	Appropriations	Education & Labor	Natural Resources	Science & Technology
Transportation & Infrastructure	Science & Technology	Judiciary	VA	Natural Resources

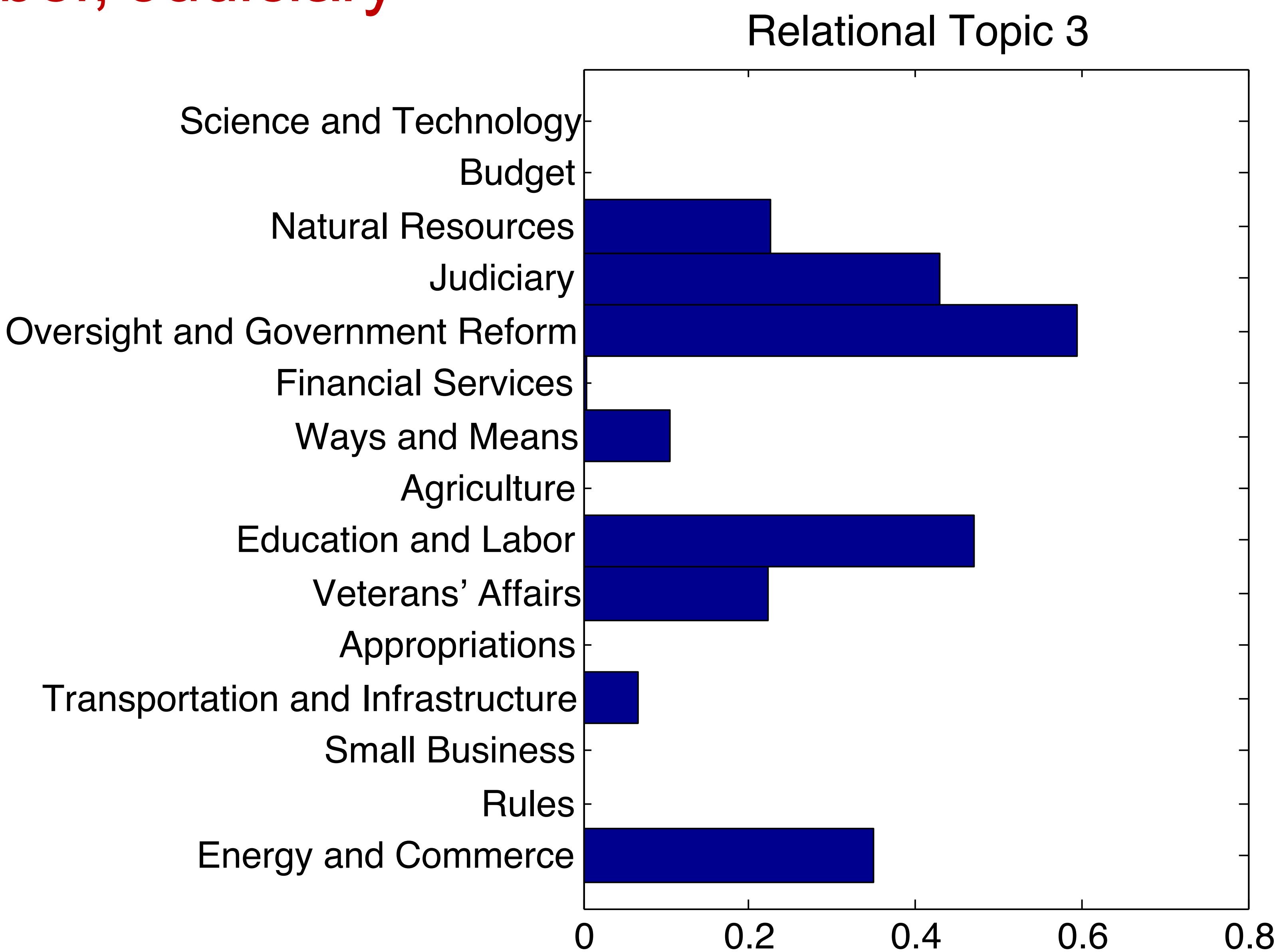
Topic 1: Ways & Means, Financial Services



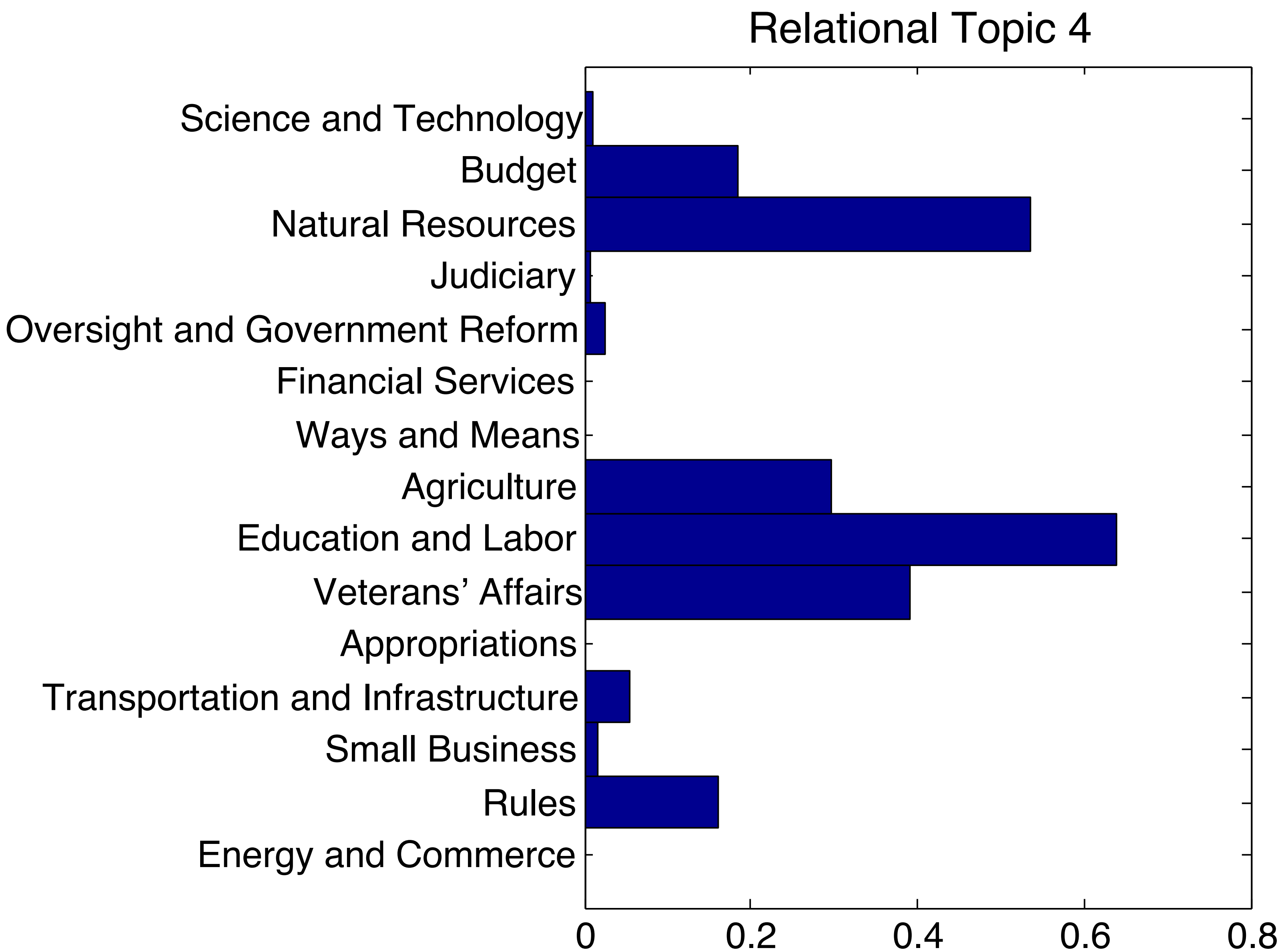
Topic 2: Rules, Appropriations, S&T



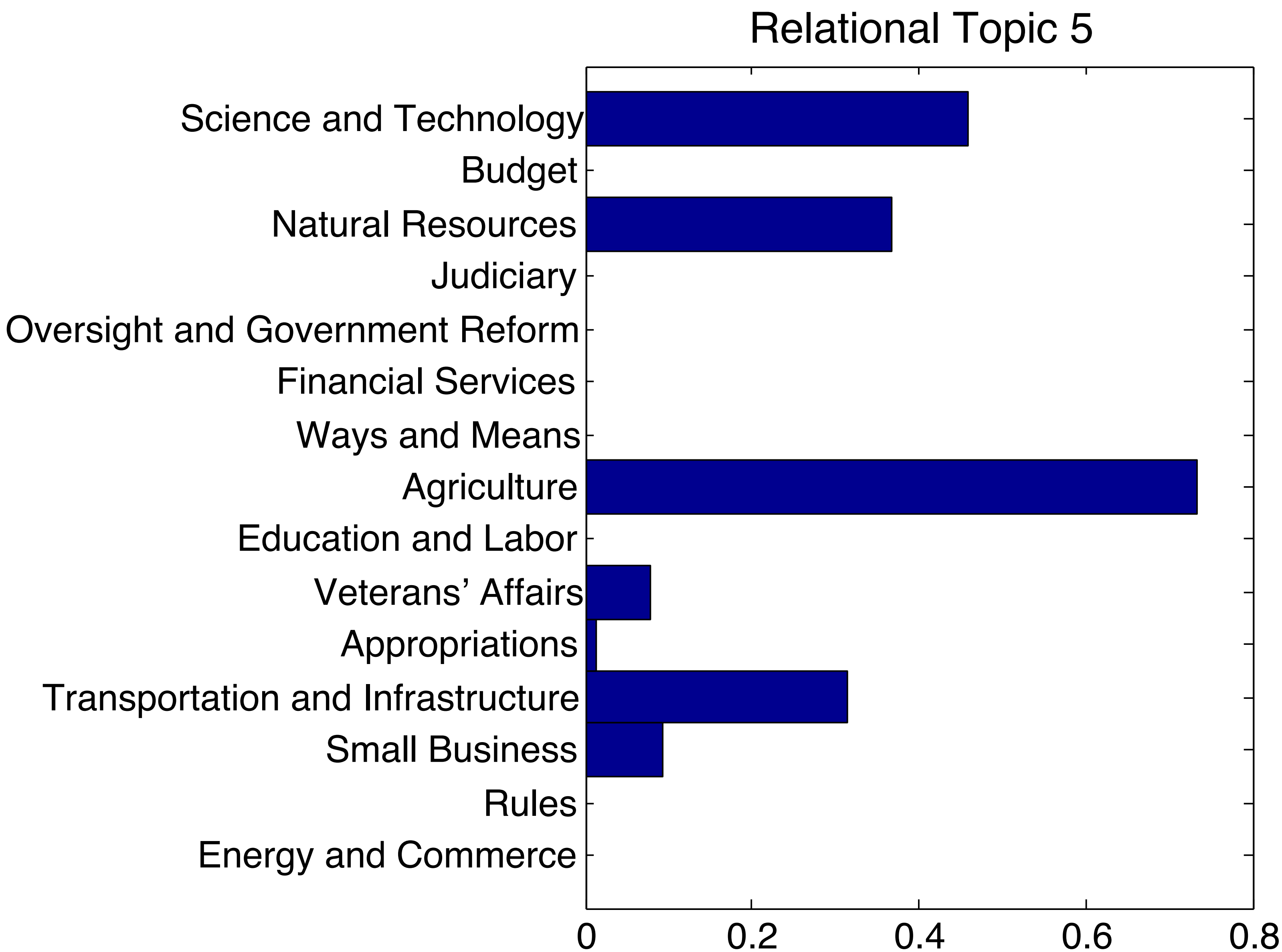
Topic 3: Oversight & Gov't Reform, Education & Labor, Judiciary



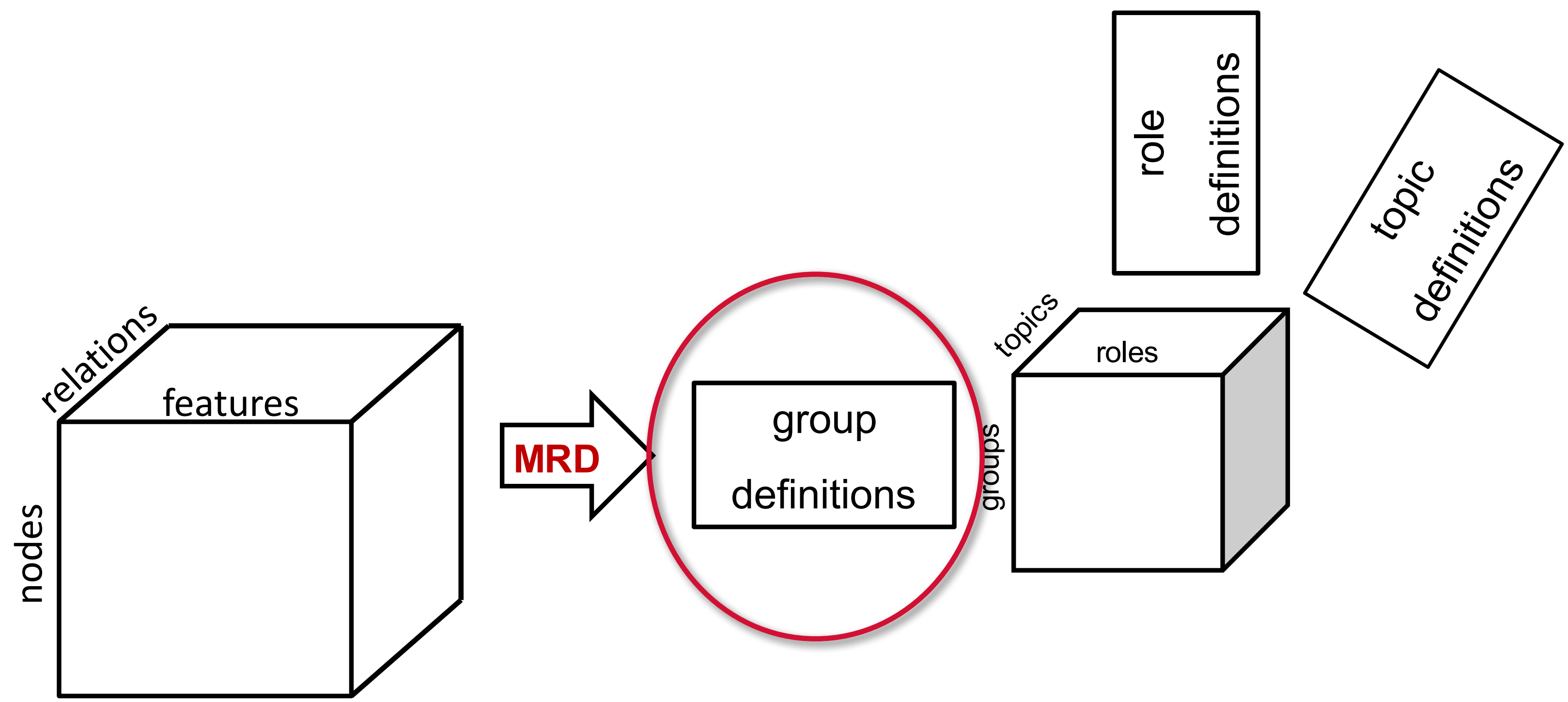
Topic 4: Education & Labor, Natural Resources, VA



Topic 5: Agriculture, S&T, Natural Resources

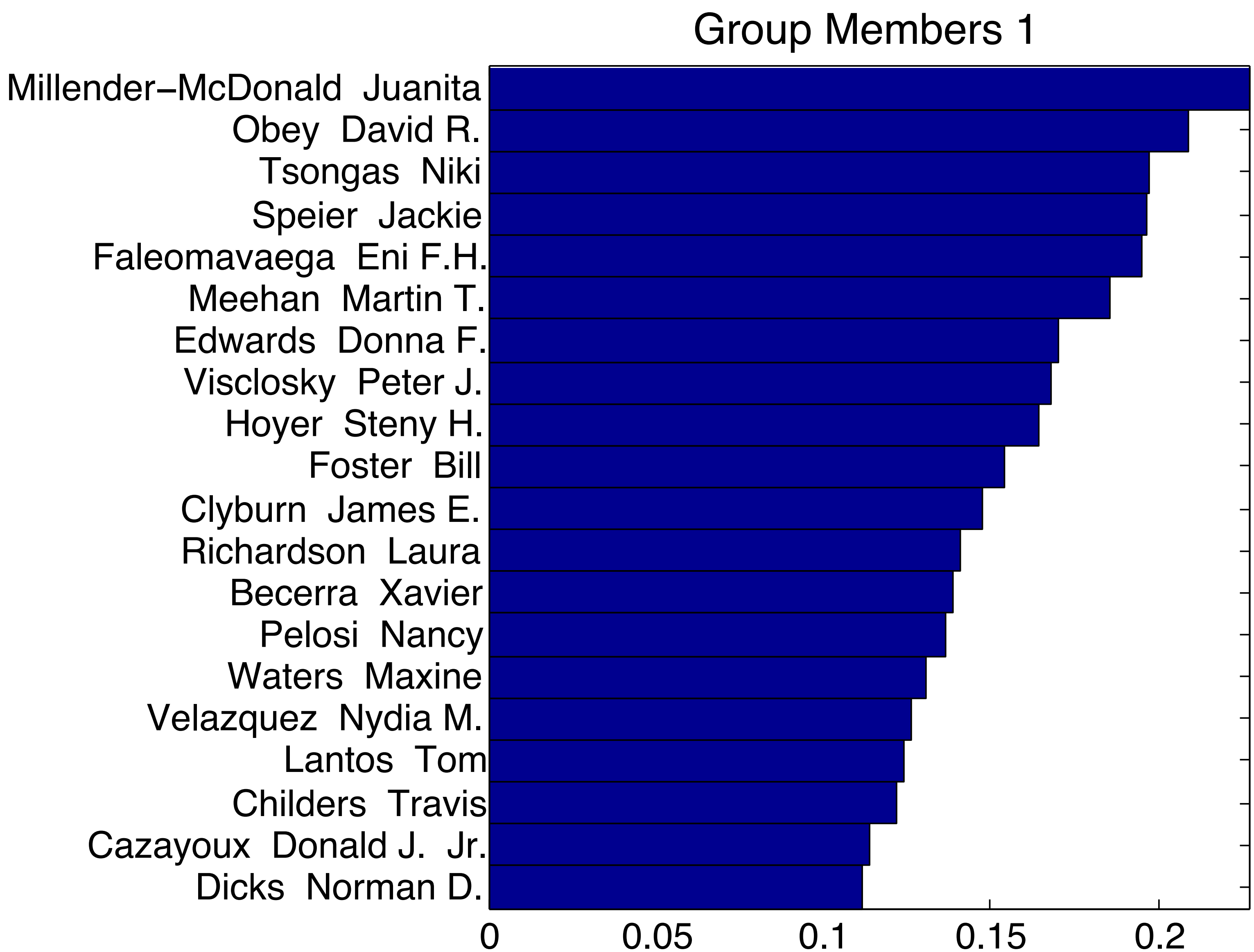


Group definitions



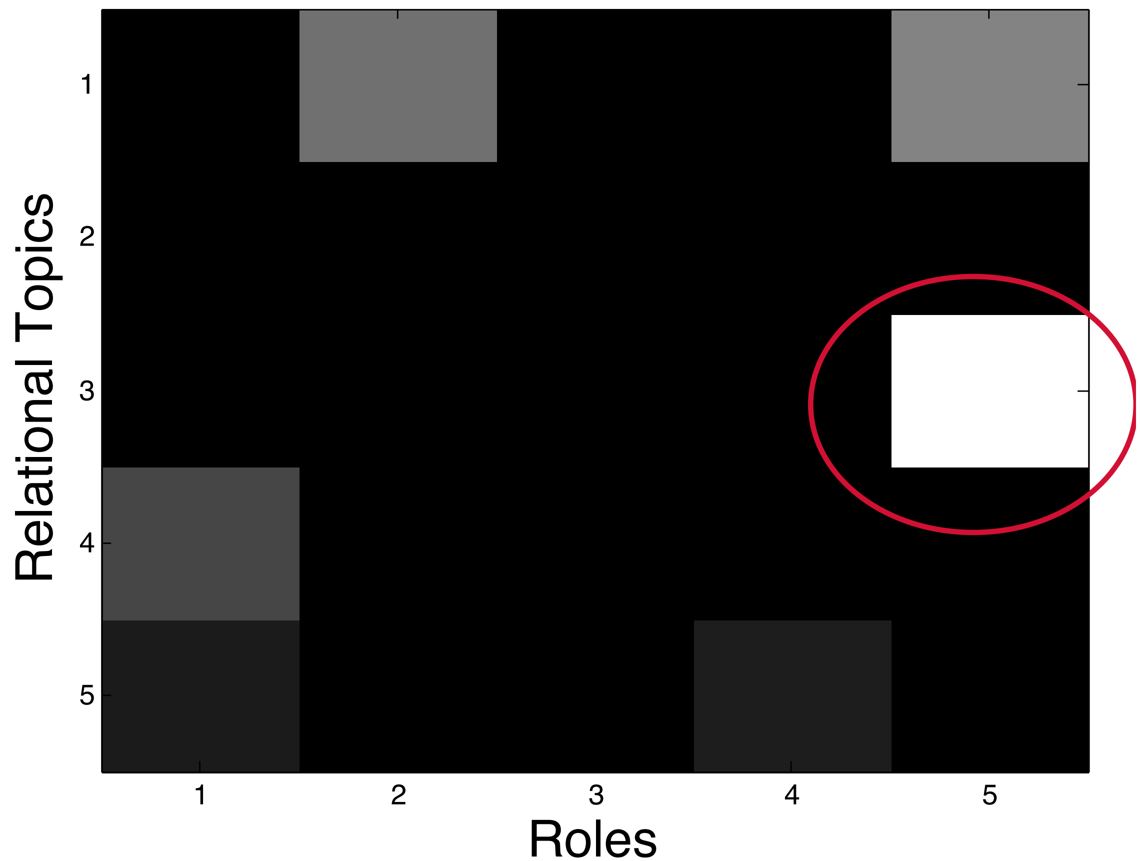
[*MRD*:
Gilpin et al.,
under review]

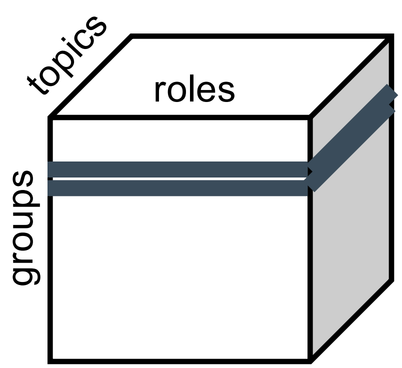
Groups of representatives



Group 1 of representatives

Name	Party	Exp
Millender-McDonald	D	11
Obey, David	D	38
Tsongas, Niki	D	0
Speier, Jackie	D	0
Faleomavaega, Eni	D	18
Meehan, Martin	D	14
Edwards, Donna	D	0
Visclosky, Peter	D	22
Hoyer, Steny	D	26
Foster, Bill	D	0

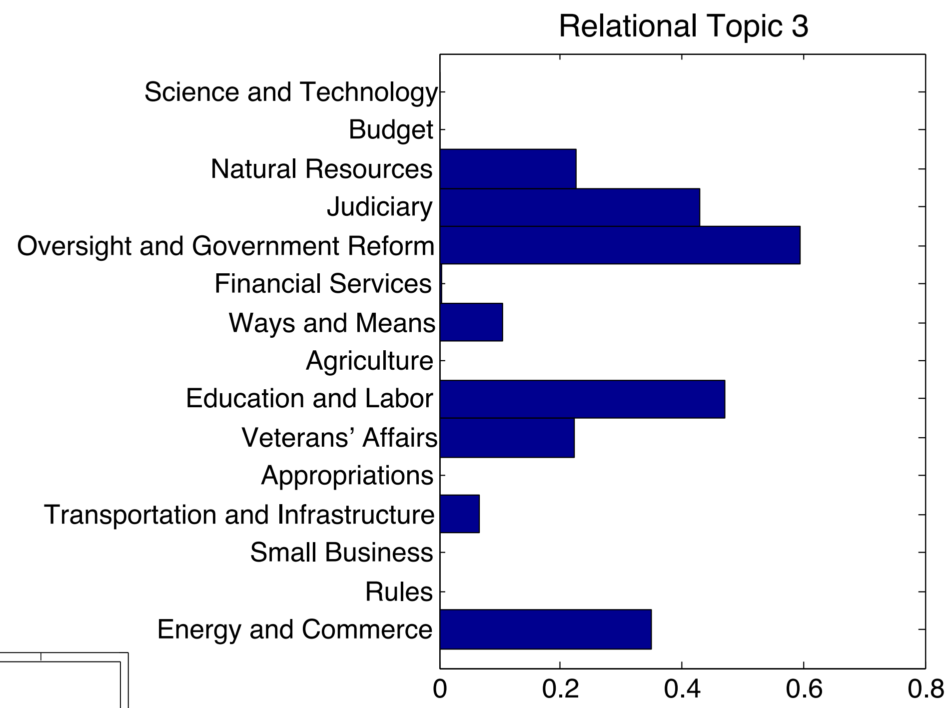


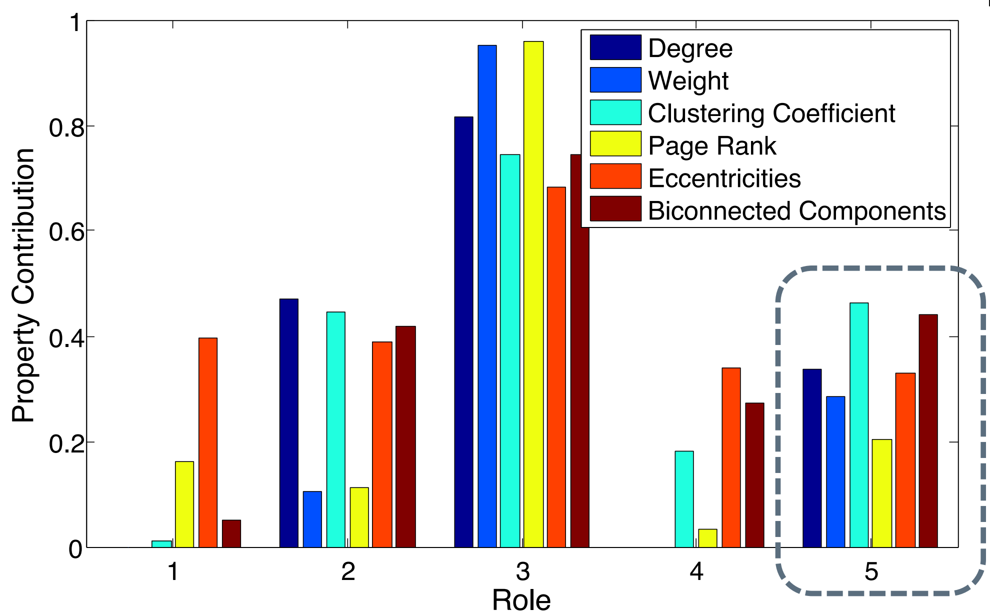


More insights into Group 1

Group 1

Name	Party	Exp
Millender-McDonald	D	11
Obey, David	D	38
Tsongas, Niki	D	0
Speier, Jackie	D	0
Faleomavaega, Eni	D	18
Meehan, Martin	D	14
Edwards, Donna	D	0
Visclosky, Peter	D	22
Hoyer, Steny	D	26
Foster, Bill	D	0



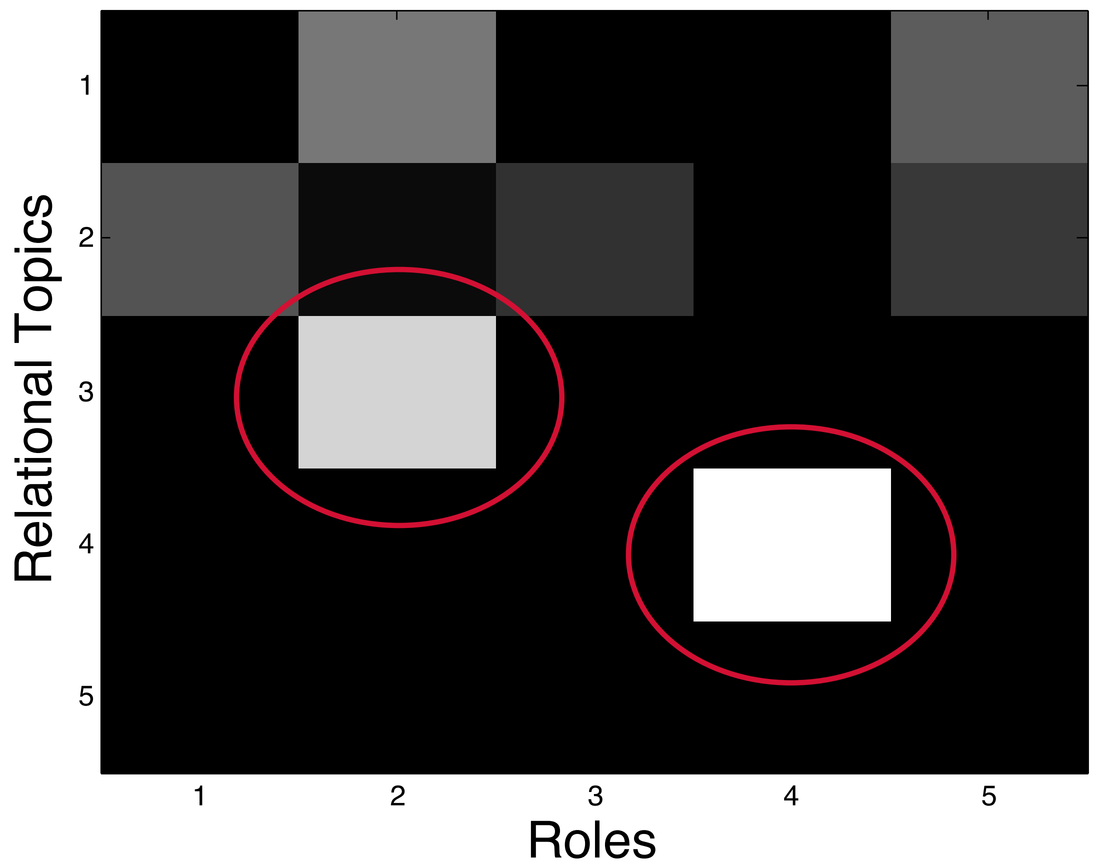


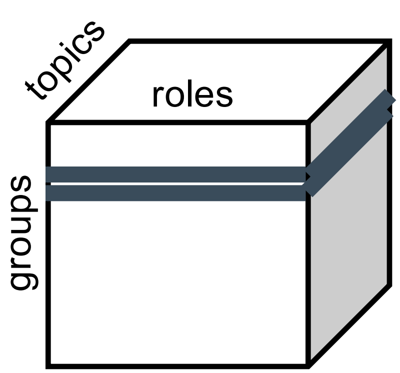
<u>Group 1</u>

- Democrats; mostly not mid-career
- Active in oversight & gov't reform
- On the periphery, but lots of triangles

Group 2

Name	Party	Exp
Hensarling, Jeb	R	4
Boehner, John	R	16
Thornberry, Mac	R	12
Broun, Paul	R	0
Shadegg, John	R	12
Hastert, Dennis	R	8
Scalise, Steve	R	11
Latta, Robert	R	6
Flake, Jeff	R	6
McCrery, Jim	R	14



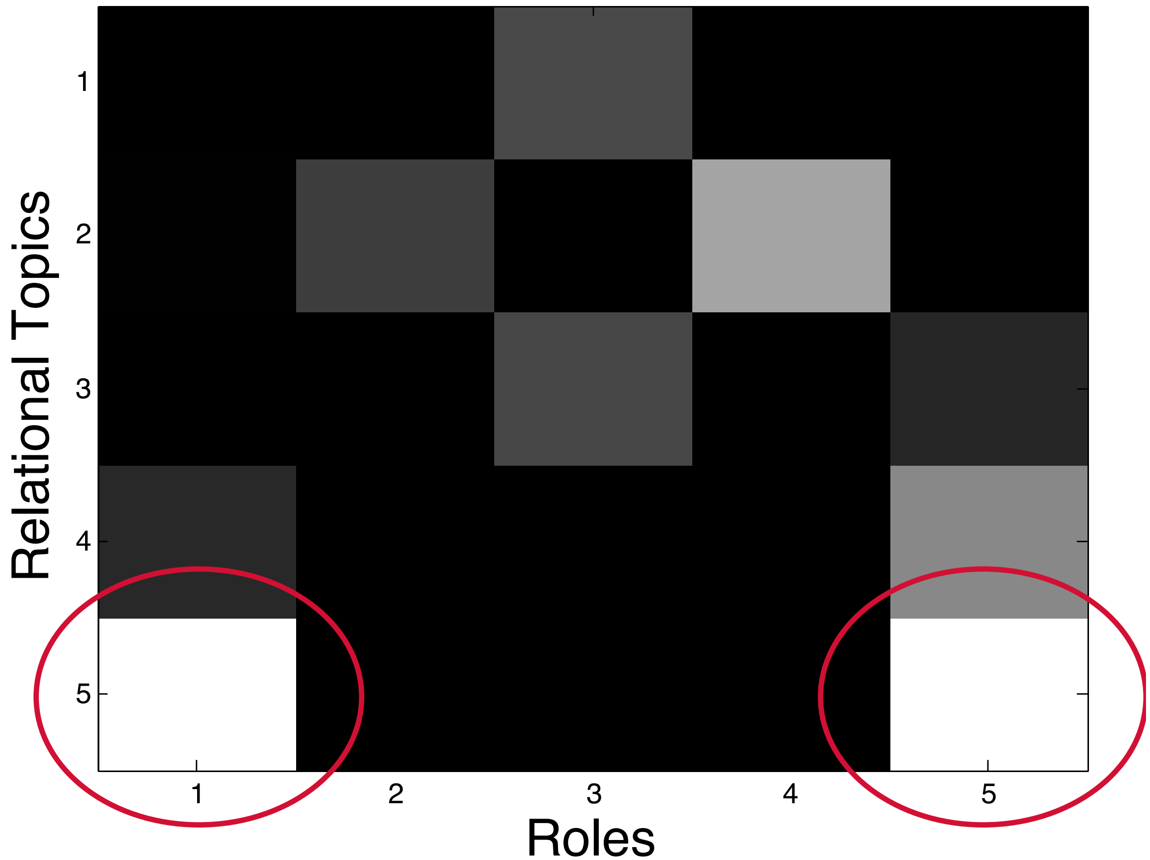


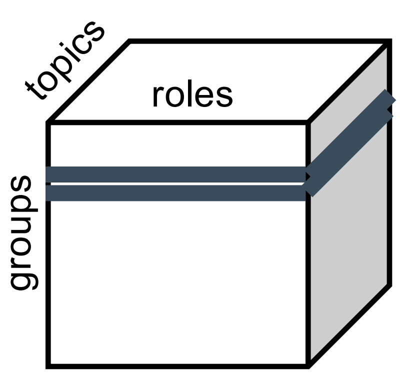
Group 2

- Republicans
- Different topics
- Different roles

Group 3

Name	Party	Exp
Cooper, Jim	D	16
Johnson, Henry	D	0
Ryan, Tim	D	4
DeGette, Diana	D	10
Engel, Eliot L.	D	14
Doggett, Lloyd	D	12
Pastor, Ed	D	16
Meek, Kendrick	D	4
Murphy, C.	D	0
Crowley, Joseph	D	8



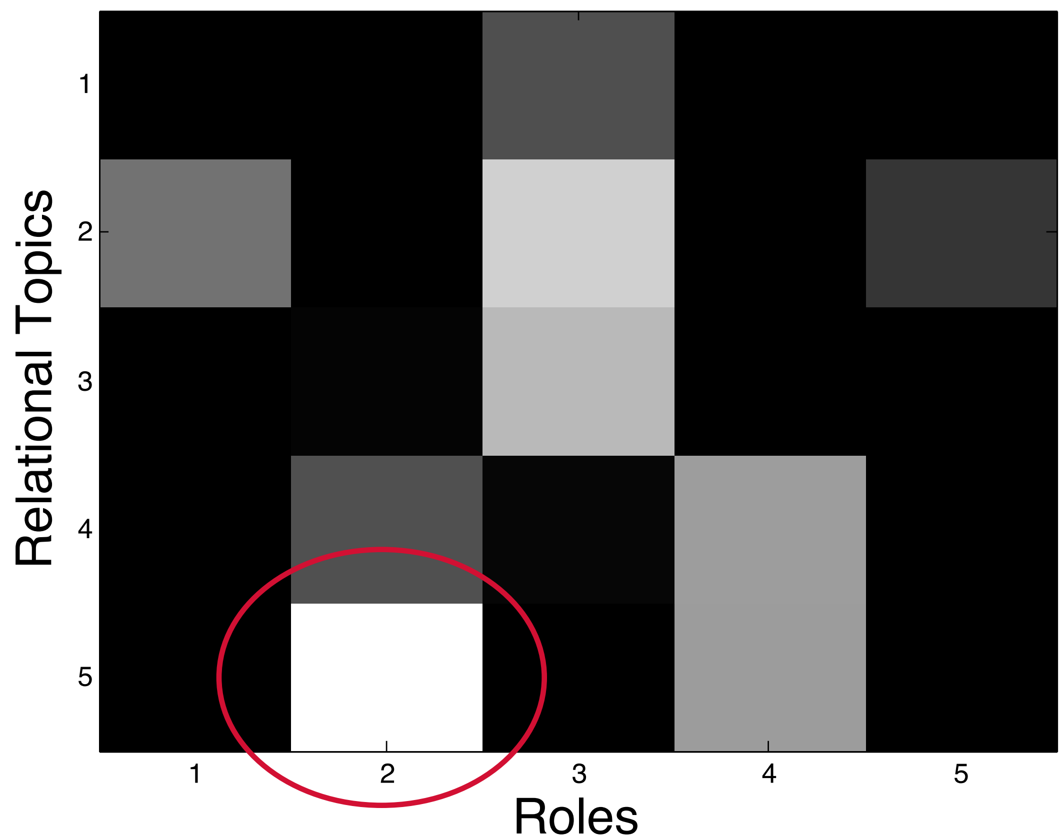


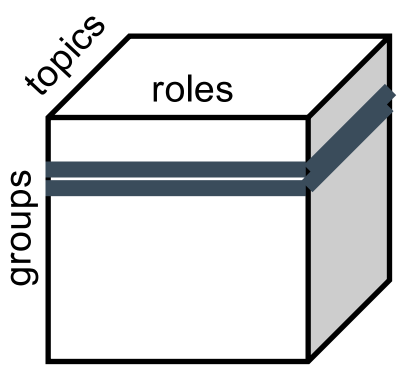
Group 3

- Democrats
- Same topic
- Different roles

Group 4

Name	Party	Exp
Hall, Ralph	R	16
Rodgers, Cathy	R	2
Myrick, Sue	R	12
Issa, Darrell	R	6
Drake, Thelma	R	2
Kuhl, Randy	R	2
Poe, Ted	R	2
Boozman, John	R	6
Conaway, Michael	R	2
Wamp, Zach	R	12



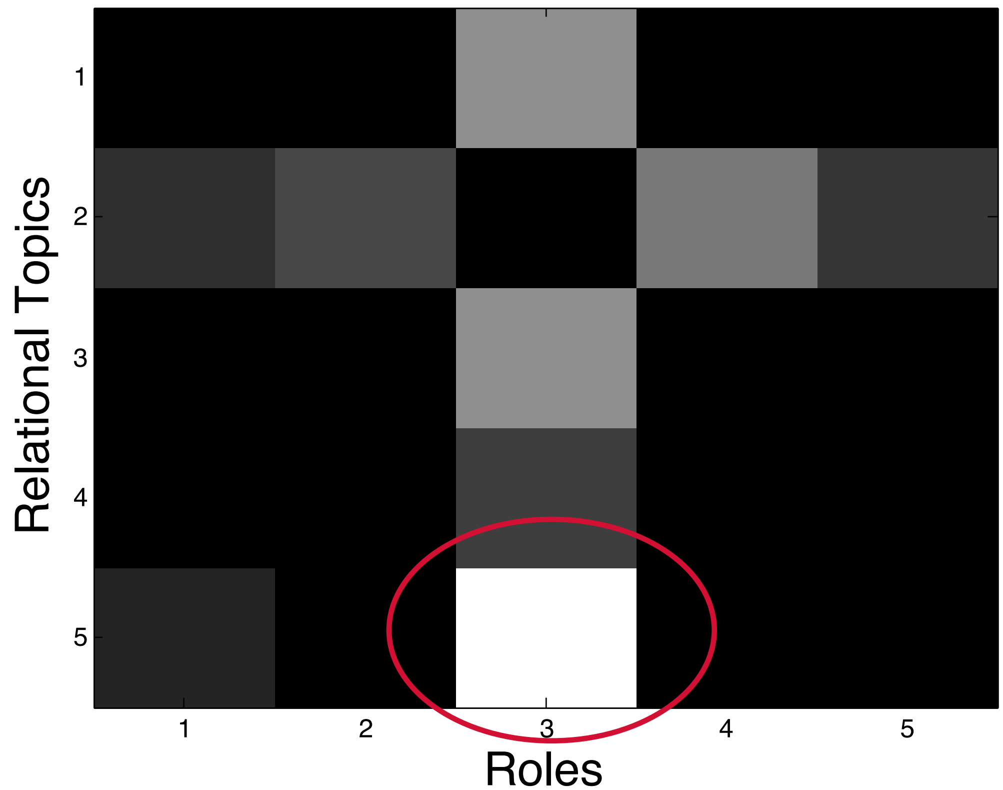


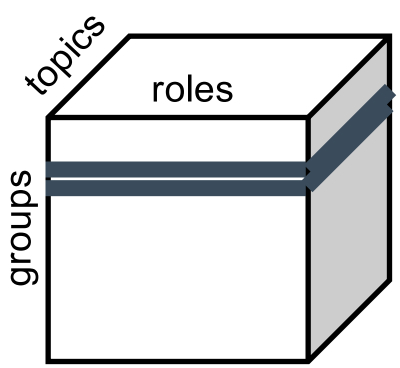
<u>Group 4</u>

- Republicans
- Active in Agriculture
- High degree & very clique-y

Group 5

Name	Party	Exp
Jackson-Lee, Sheila	D	12
Cohen, Steve	D	0
Hare, Phil	D	0
Grijalva, Raul	D	4
English, Phil	R	12
Honda, Michael	D	6
McCotter, Thaddeus	R	4
Filner, Bob	D	14
Hinchey, Maurice	D	14
Gonzalez, Charles	D	8

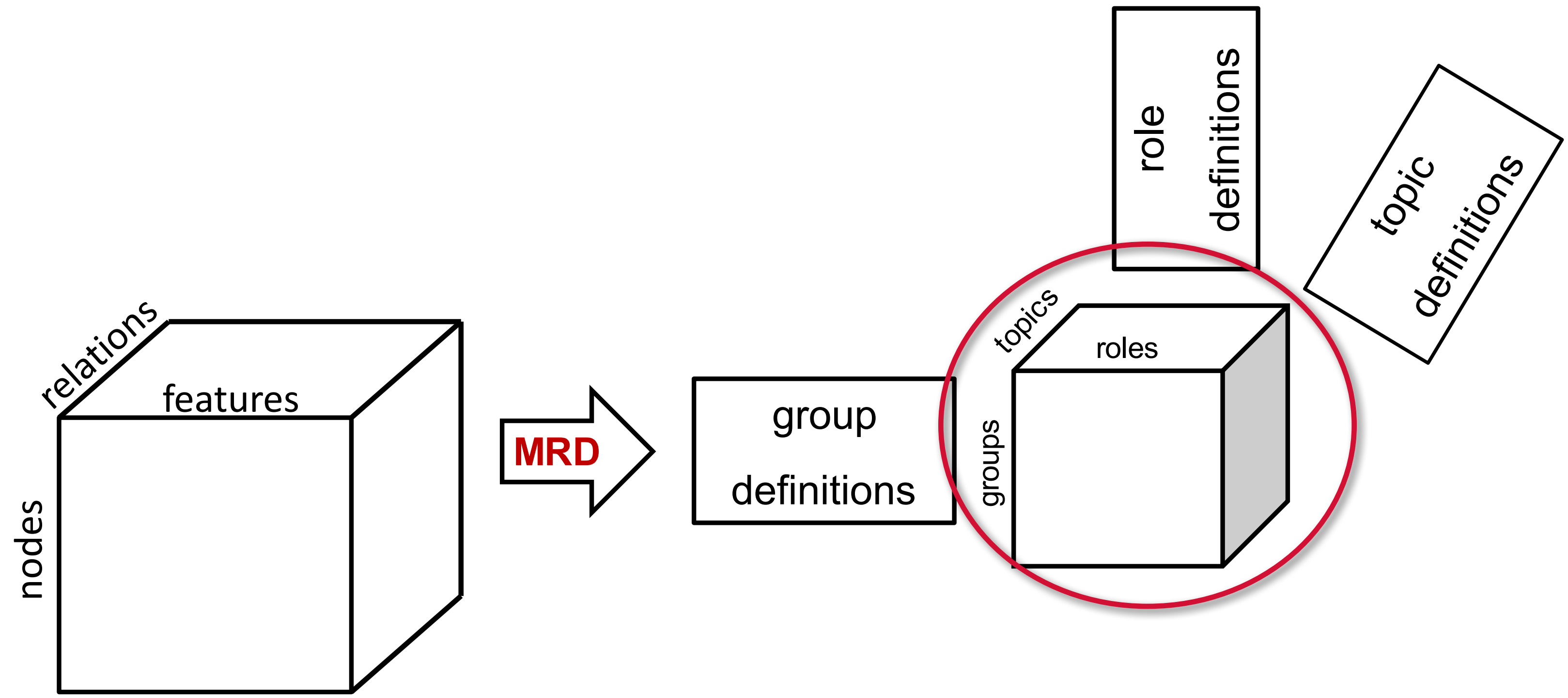




Group 5

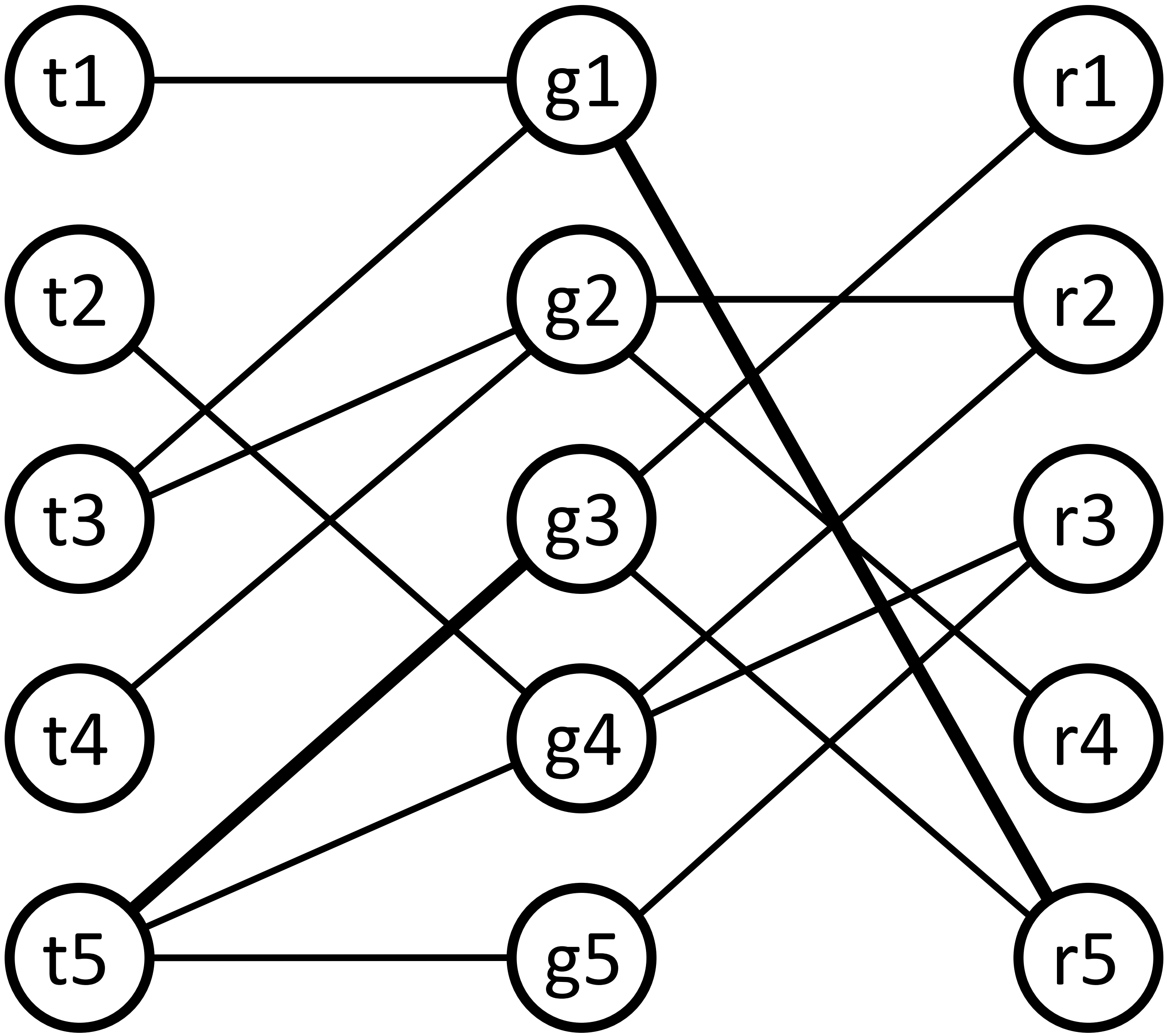
- Bipartisan
- Active in Agriculture
- Power brokers

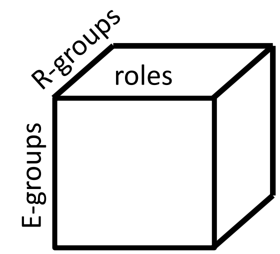
Tucker core



[*MRD*:
Gilpin et al.,
under review]

Interaction graph from the Tucker core



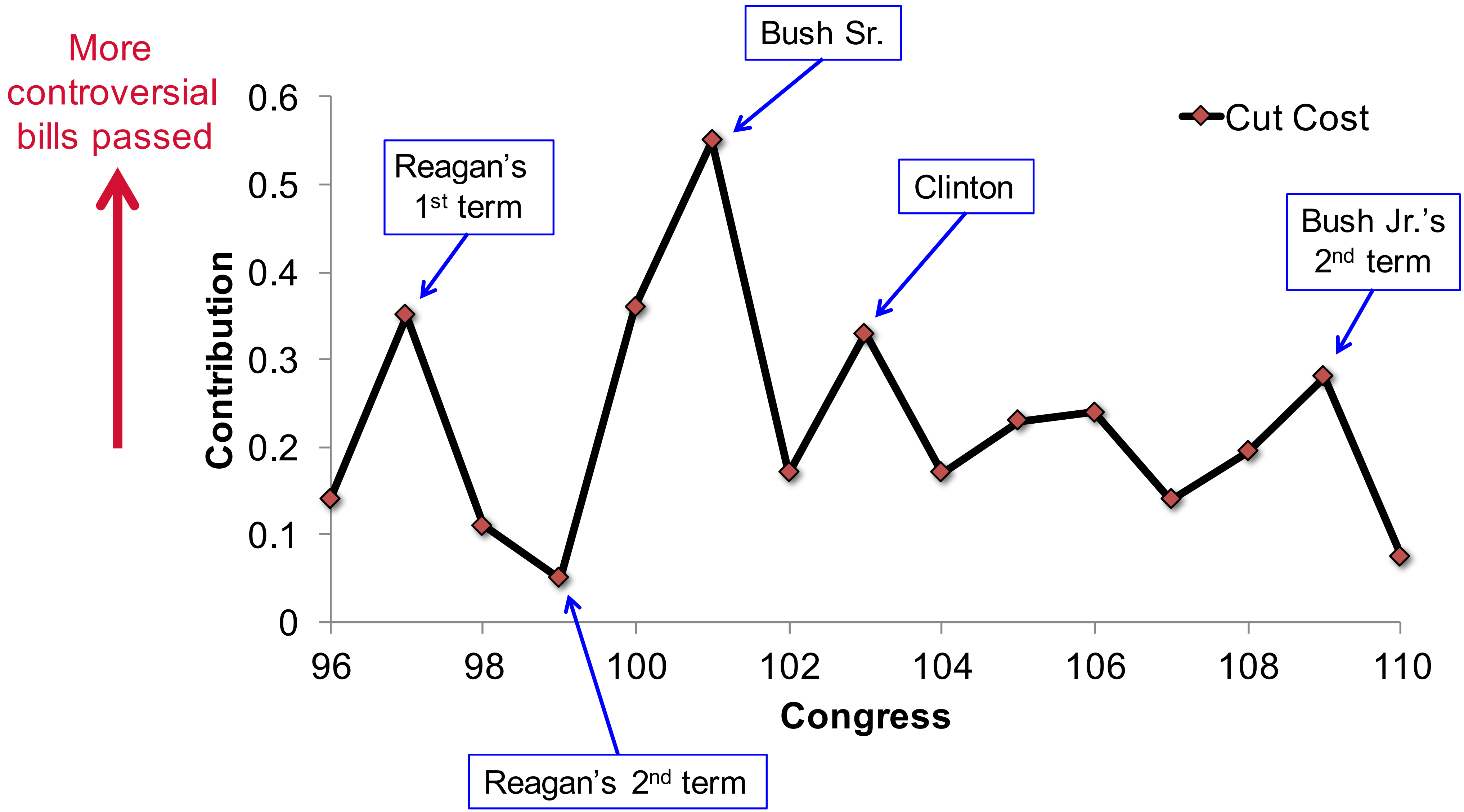


110th Congress
Co-sponsorship
Graph

Measure properties on the interaction graph

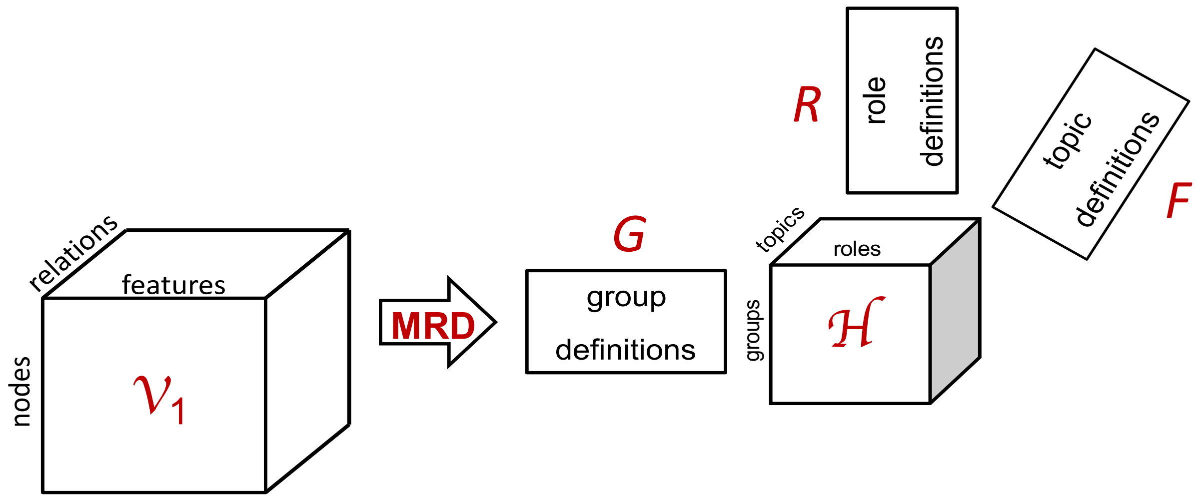
Property	Description	Computation
Simplicity	To what extent are nodes connected to (role) similar types of nodes?	Average Node Degree
Sharing	How much can a group be separated into independent parts?	Mincut cost
Variability	How does the simplicity of nodes vary across the interaction graph?	Variance of node degree
Stability	How stable are the interactions between roles, groups, and topics?	Spectral gap

Cut cost of the interaction graphs from Tucker cores

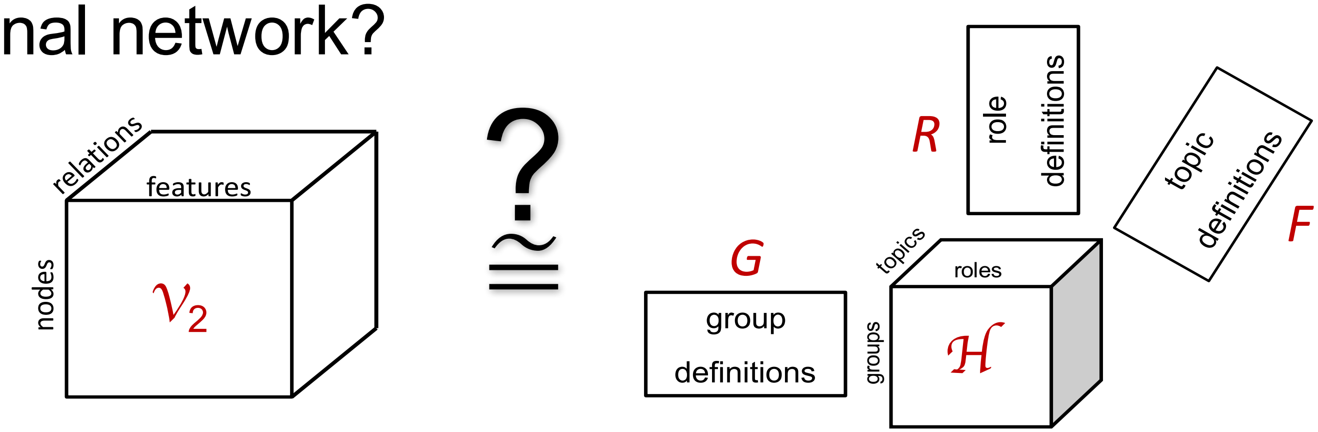


Role transfer (in this context)

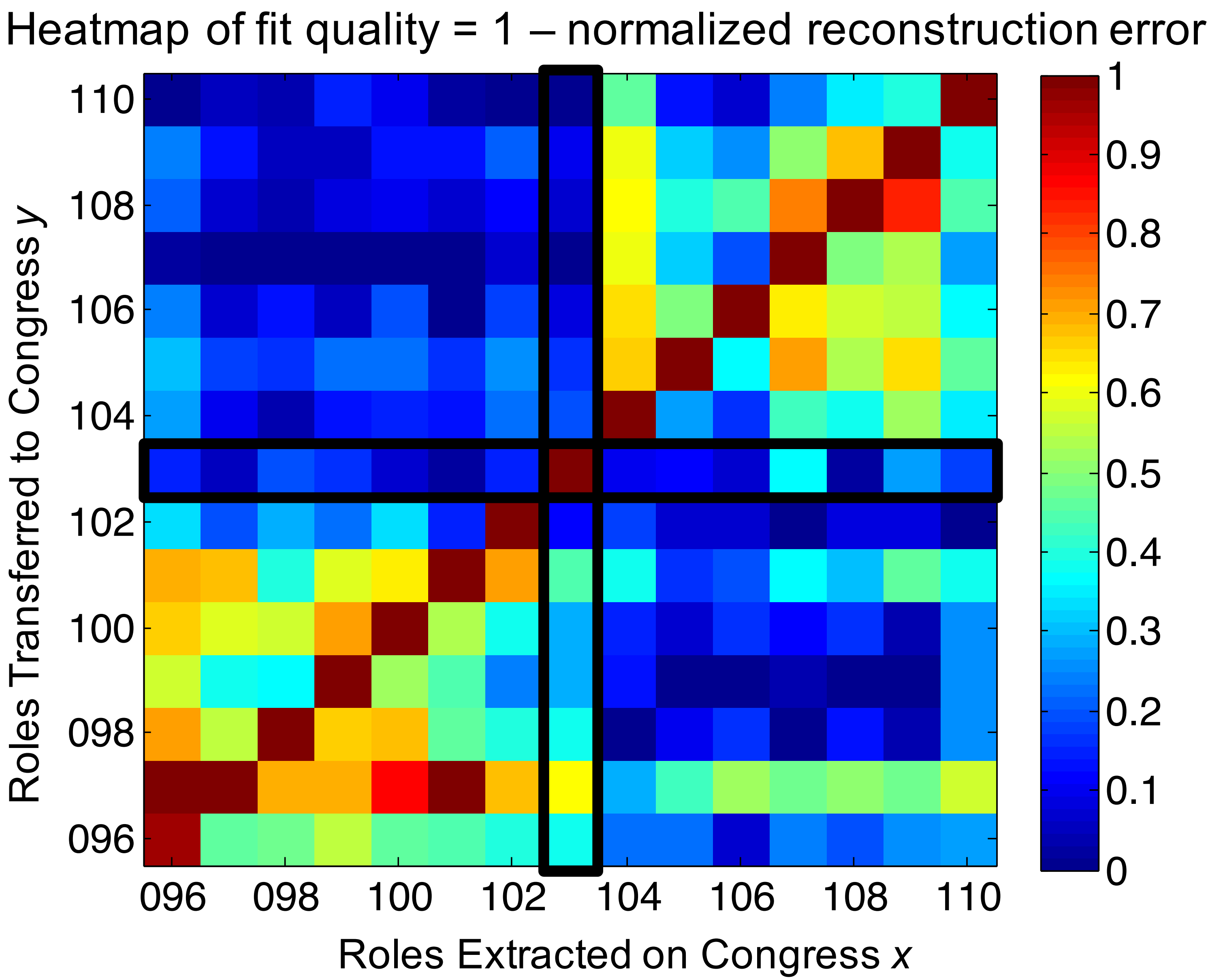
- Roles extracted on one multi-relational network



- How well do the extracted roles transfer to another multi-relational network?



Role transfer on multi-relational Networks



Applications of role discovery

Task	Use Case
Role query	Identify individuals with similar behavior to a known target
Role outliers	Identify individuals with unusual behavior
Role dynamics	Identify unusual changes in behavior
Re-identification	Identify individuals in an anonymized network
Role transfer	Use knowledge of one network to make predictions in another
Network comparison	Determine network compatibility for knowledge transfer
Exploration in role space	Exploratory analysis of network data in the role space
…	…

Why are roles effective?

- Encode complex behavior
- Map nodes into a useful lower dimensional space
- Generalize across networks
- Common language over a common alphabet

Funding from LLNL, NSF, IARPA, DARPA, and DTRA.

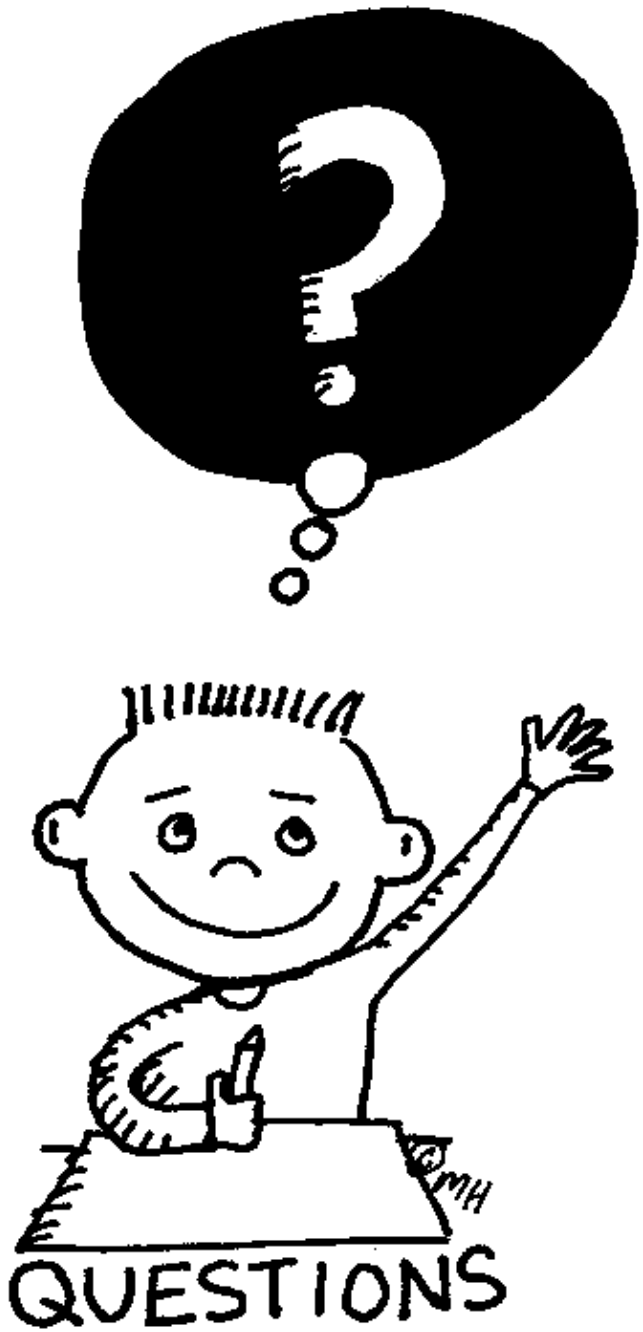
Outline

9:00 - 9:50	David	Political inquiry, new science of politics, exemplary data
9:50 - 10:30	Oren	Exponential Random Graph Models
10:30 - 11:00	Coffee Break	
11:00 - 11:20	Oren	Networks of political figures on Twitter
11:20 - 11:50	Tina	Roles in socio-political networks
11:50 - 12:00	David	Wrap-up & questions

Wrap-up

- Tutorial website includes slides, resources (data & code)
  - http://bit.ly/1Qs8bIA

- Seize the opportunity to create a new science of politics



THANKS!