Northeastern University
Network Science Institute

LEARNING, MINING AND GRAPHS

Tina Eliassi-Rad

tina@eliassi.org *@tinaeliassi*

Supported by NSF, DTRA, DARPA, IARPA, DOE/LLNL & WaPo Labs

Roadmap

1. The reasonable effectiveness of roles in networks
2. A theoretical guide to tie-strength measures

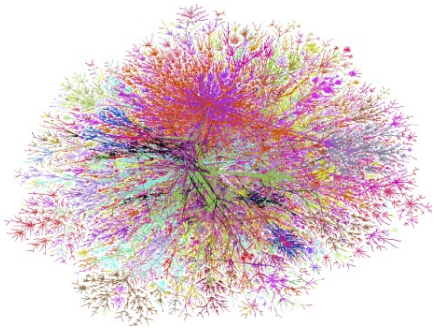
Roadmap

1. The reasonable effectiveness of roles in networks
2. A theoretical guide to tie-strength measures

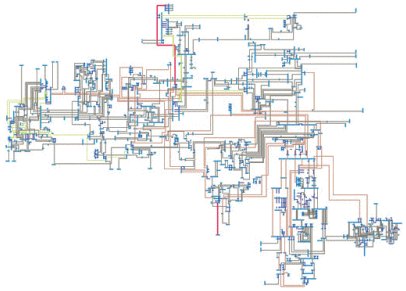
Complex Networks are Ubiquitous

Technological Networks

Internet

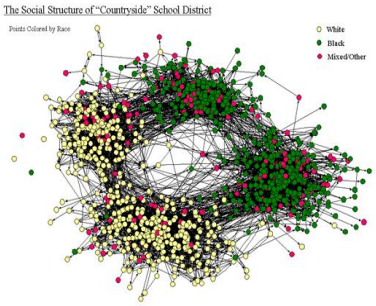


NY State Power Grid

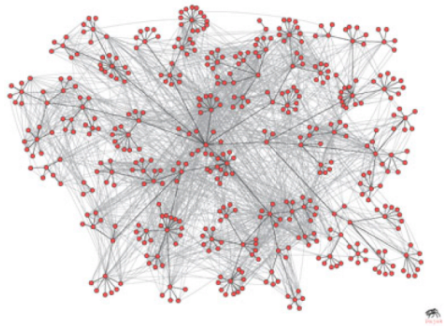


Social Networks

Friendship

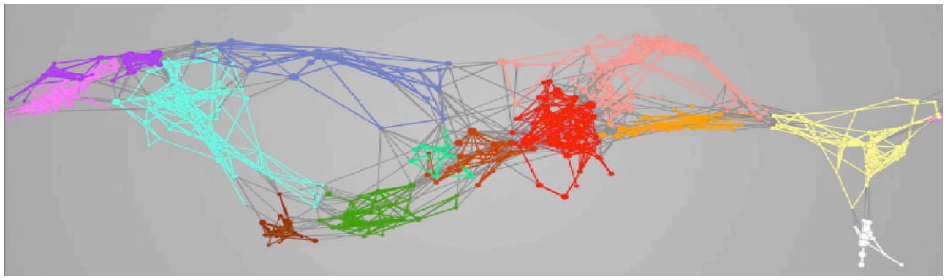


HP Emails



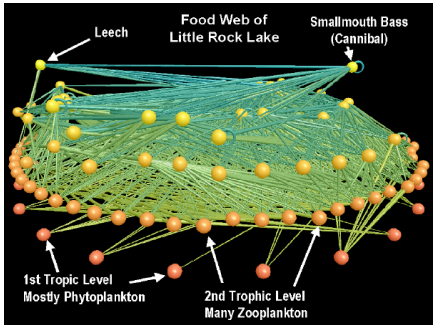
Information Networks

Map of Science

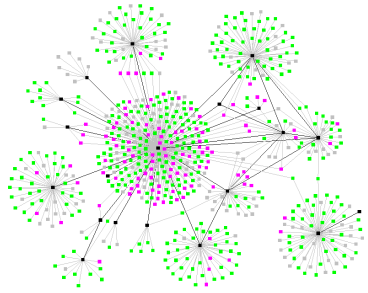


Biological networks

Food Web

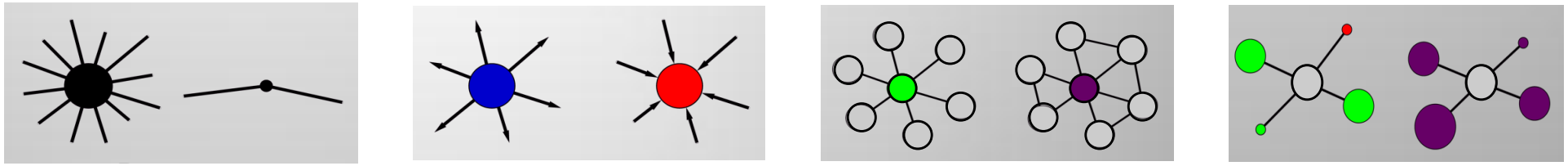


Contagion of TB

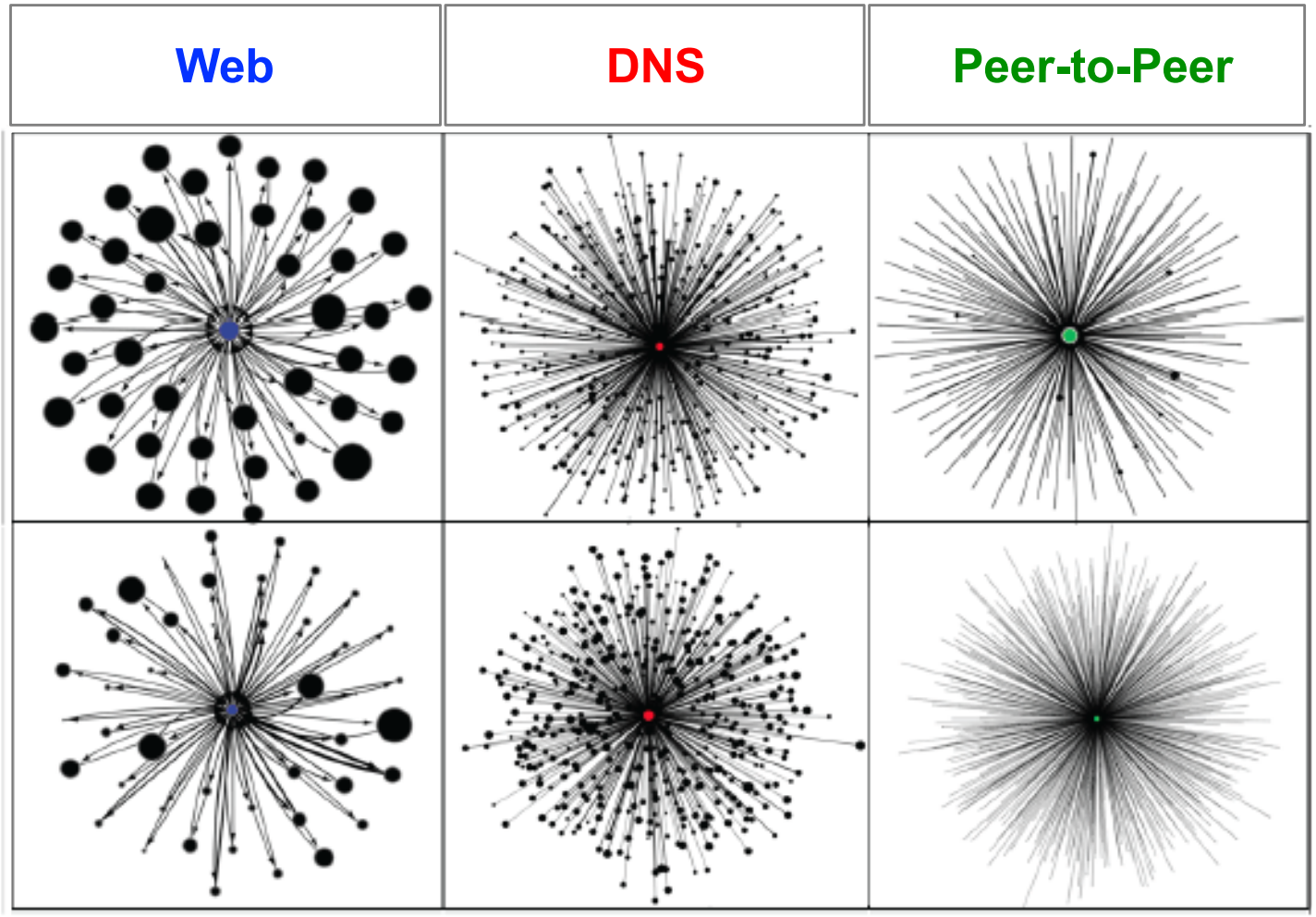


What are Roles?

- Functions of nodes in the network
  - Similar to functional roles of species in ecosystems
- Roles are defined in terms of structural behaviors
  - What is your connectivity pattern?
  - To what kinds of individuals are you connected?



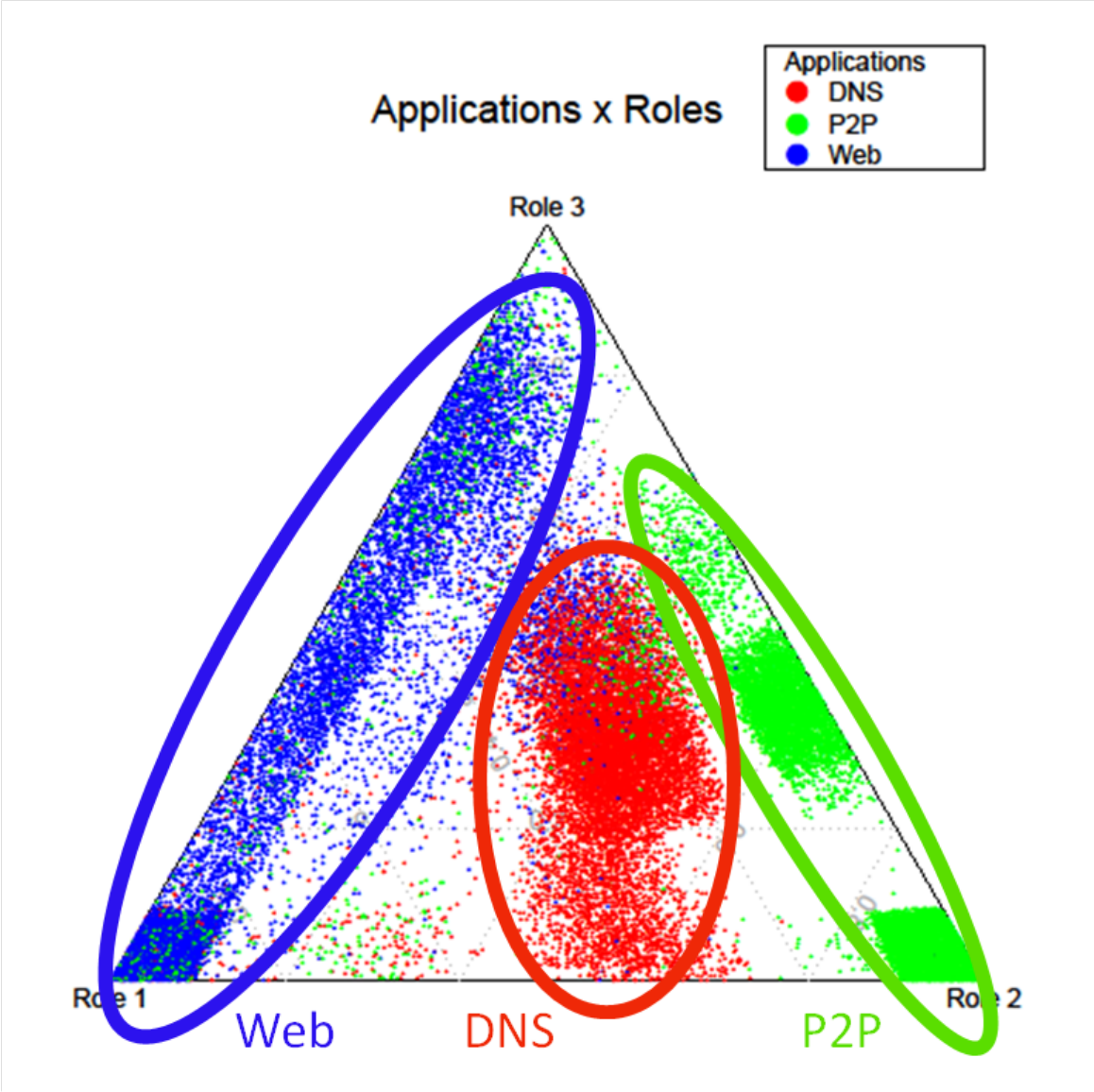
Example of Roles in an IP × IP Network



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

The types of neighbors that are connected to a given host are indicators of the host's role.

Each Node has a Mixture of Roles



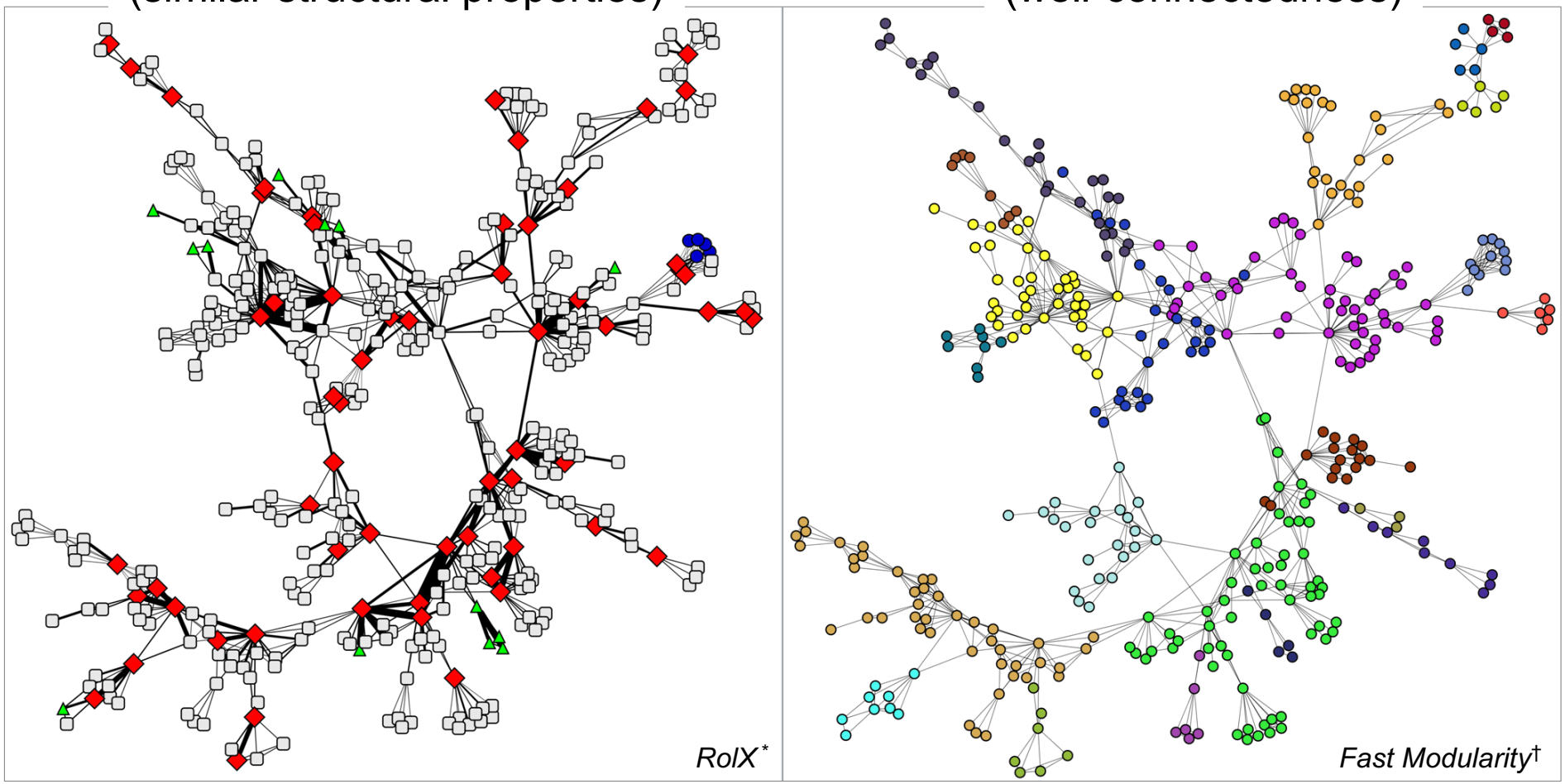
Research Questions

1. How are roles different from communities and from positions/equivalences (from sociology)?
2. Given a network, how can we automatically discover roles of nodes?
3. How can we make sense of these roles?
4. Are there good features that we can extract for nodes that indicate role-membership?
5. What are the applications in which these discovered roles can be effectively used?

Roles & Communities are Complementary

Roles
(similar structural properties)

Communities
(well-connectedness)

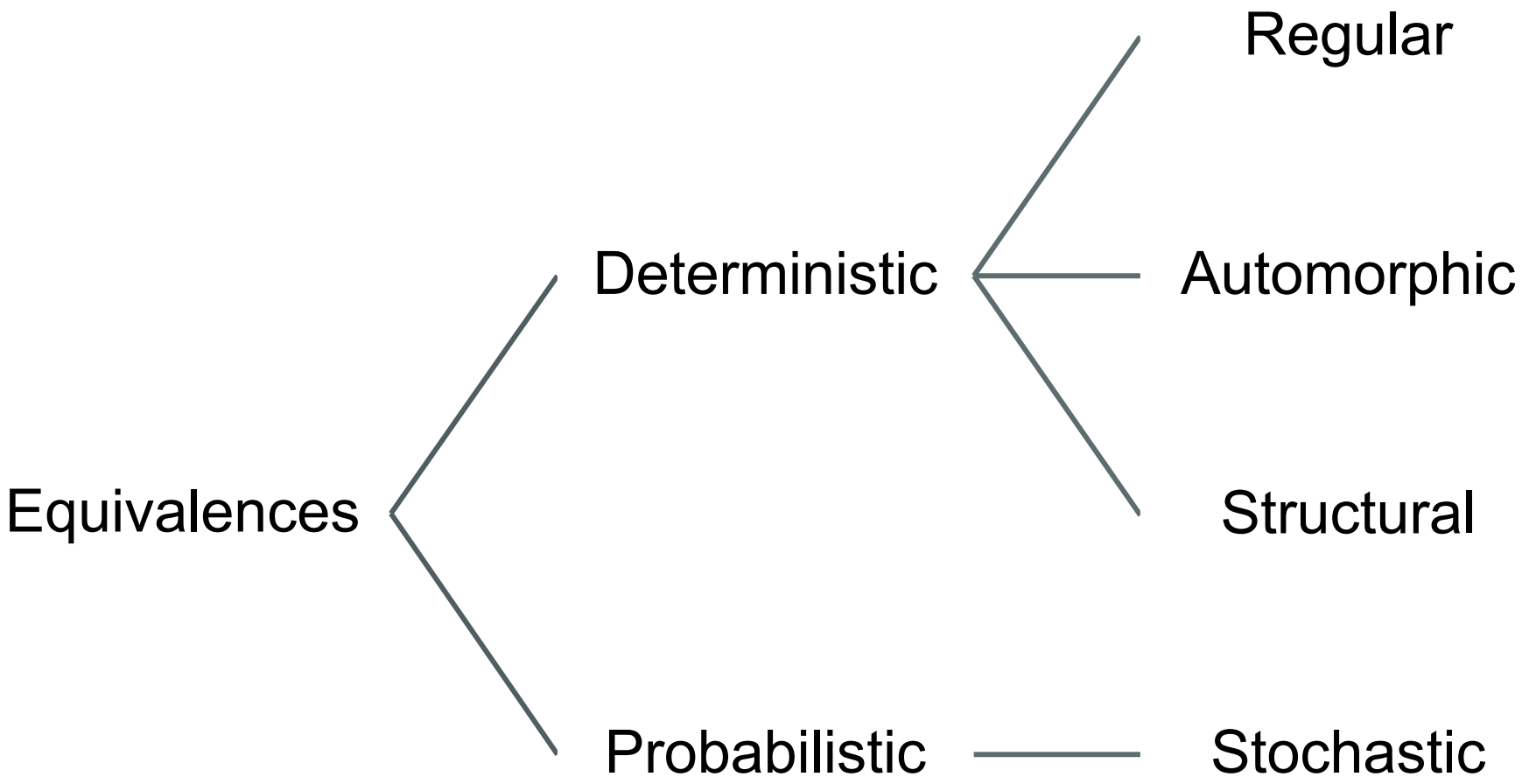


* Henderson, et al. 2012; † Clauset, et al. 2004

Roles are Similar to Positions from Sociology

- Two nodes with the same position are in an equivalence relation
- Equivalence, Q, is any relation that satisfies these three conditions:
  - Transitivity: (a,b), (b,c) $\in Q \Rightarrow$ (a,c) $\in Q$
  - Symmetry: (a, b) $\in Q$ if and only if (b, a) $\in Q$
  - Reflexivity: (a, a) $\in Q$

Taxonomy of Equivalences from Sociology

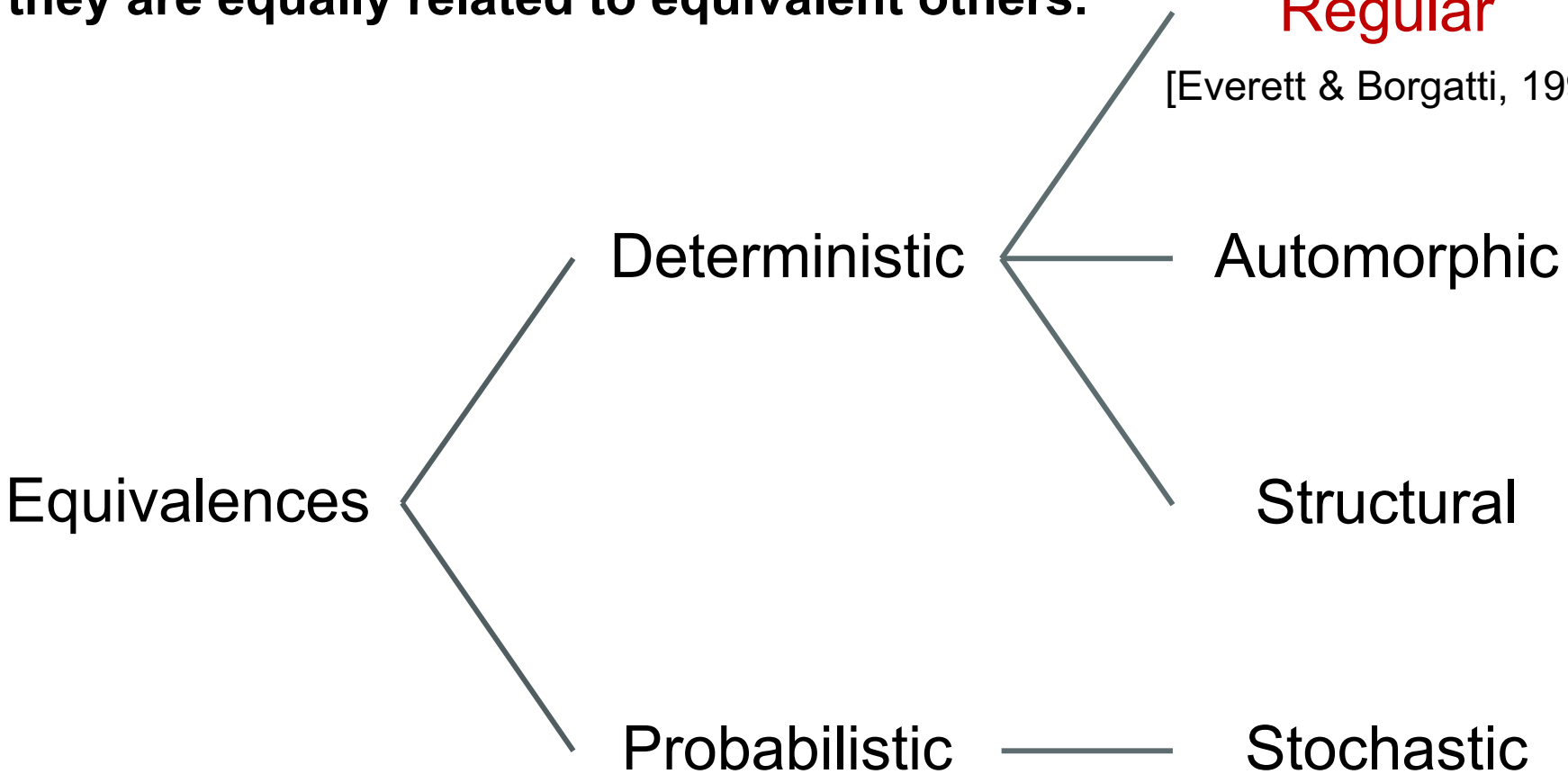


Roles find Regular Equivalences

Two nodes *u* and *v* are regularly equivalent if they are equally related to equivalent others.

Regular

[Everett & Borgatti, 1992]



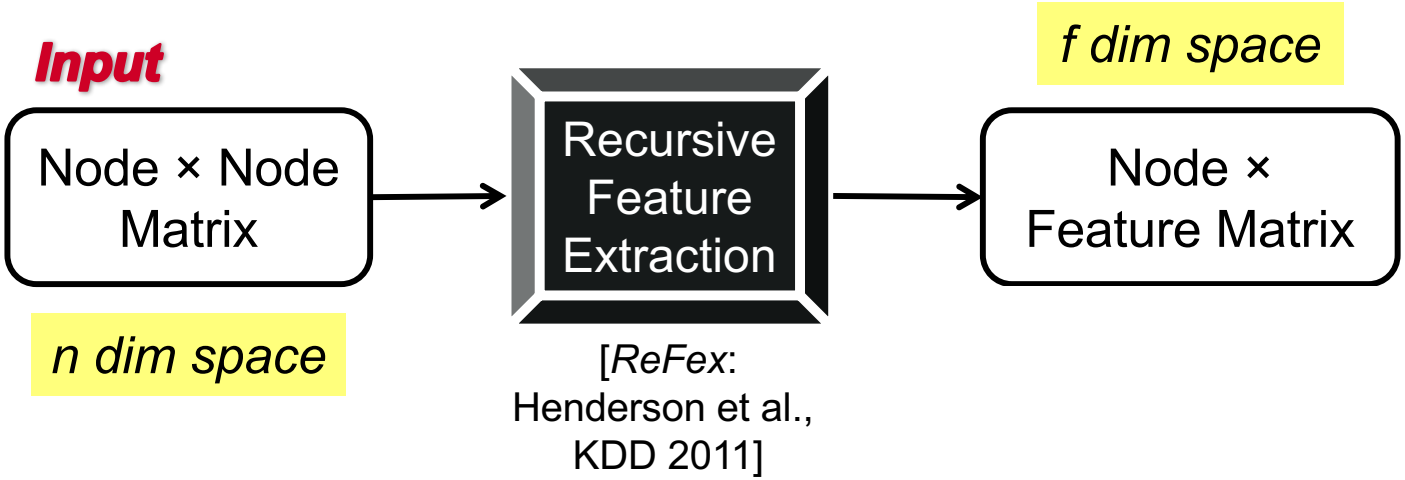
Finding Roles in a Network

Input

Node × Node Matrix

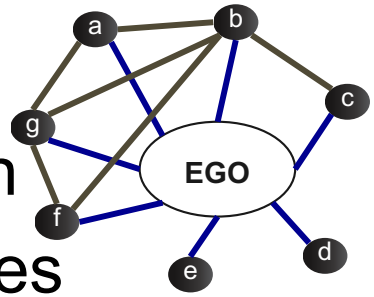
n dim space

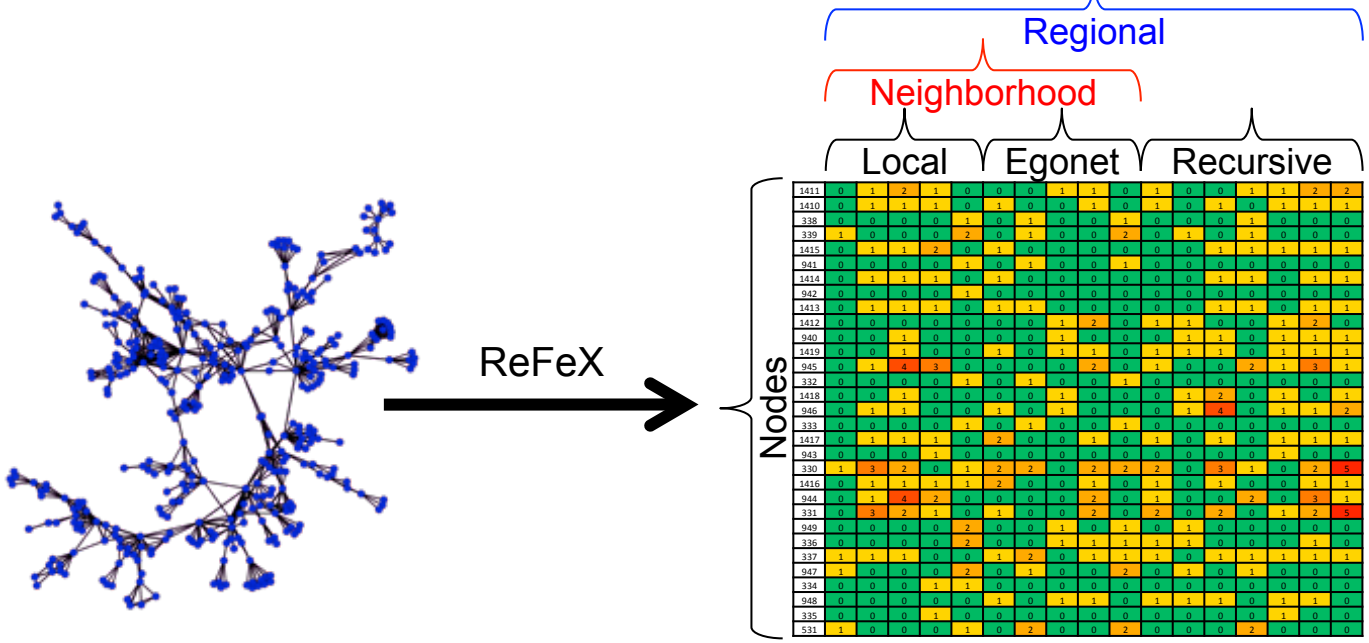
Finding Roles in a Network



ReFeX: Recursive Feature Extraction

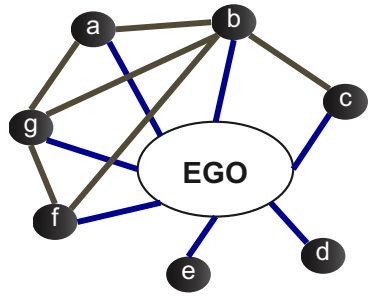
- [Henderson *et al*., KDD 2011]
- Recursively combines node-based features with egonet-based features to output regional features

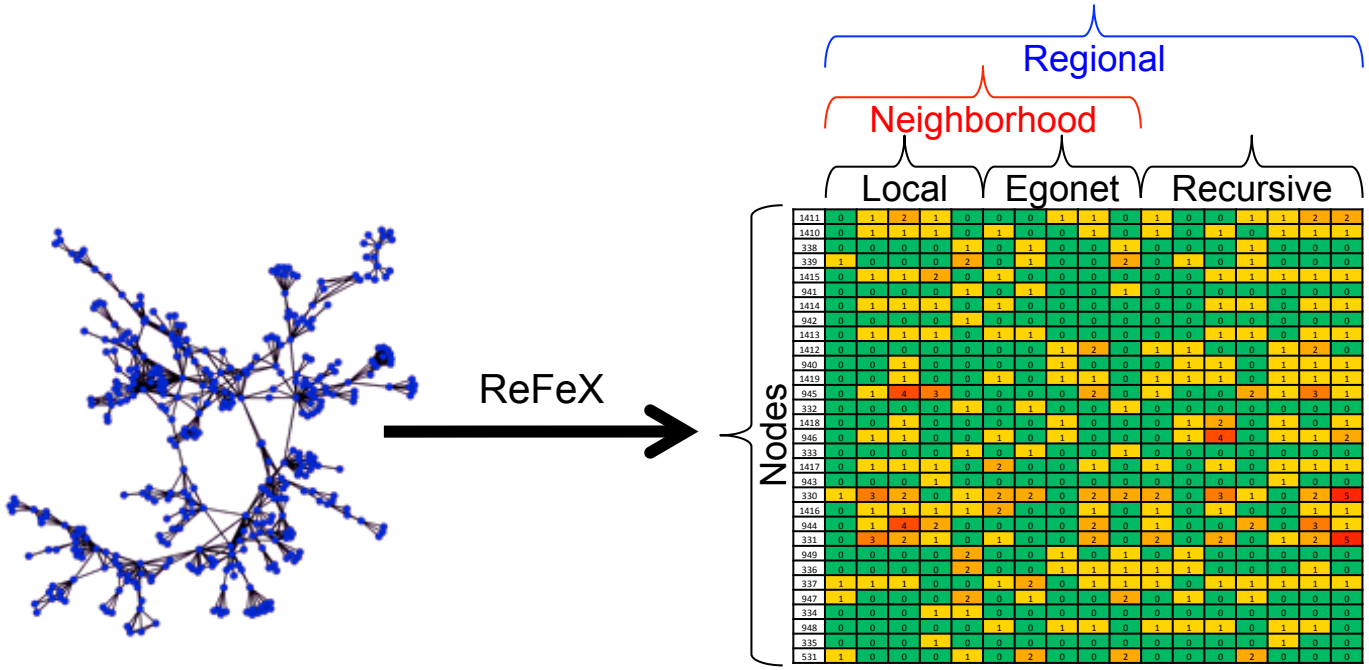




ReFeX: Recursive Feature Extraction

- [Henderson *et al*., KDD 2011]
- Recursively combines node-based features with egonet-based features to output regional features

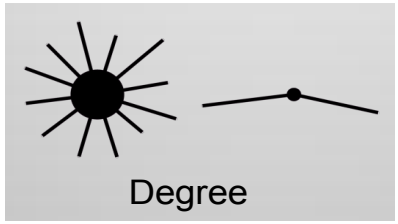


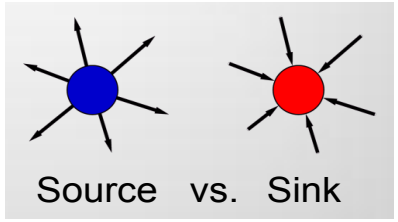


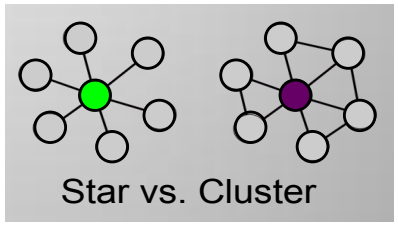
- Neighborhood features: What is your connectivity pattern?
- Recursive Features: To what kinds of nodes are you connected?

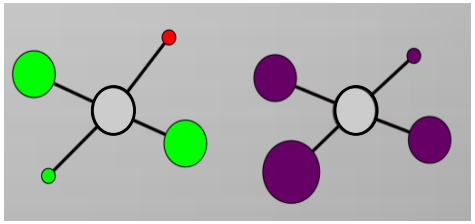
ReFeX: Structural Features

- Local
  - Essentially measures of the node degree
- Egonet
  - Computed based on each node's ego network
  - Examples
    - # of within-egonet edges
    - # of edges entering & leaving the egonet
- Recursive
  - Some aggregate (mean, sum, max, min, …) of another feature over a node's neighbors
  - Aggregation can be computed over any real-valued feature, including other recursive features





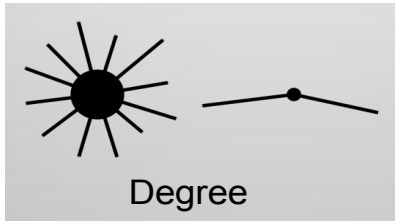


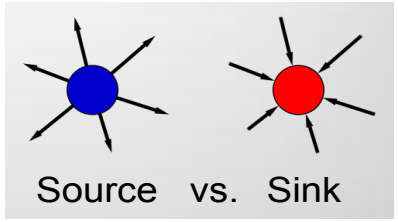


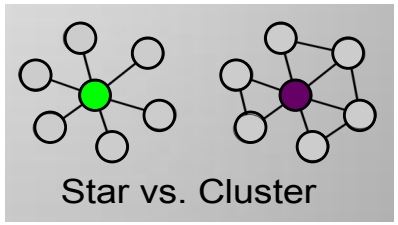
ReFeX: Structural Features

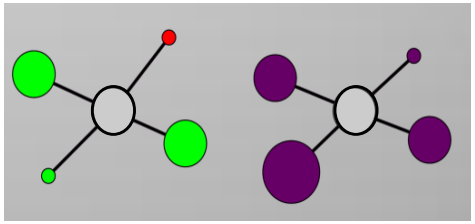
Neighborhood

- Local
  - Essentially measures of the node degree
- Egonet
  - Computed based on each node's ego network
  - Examples
    - # of within-egonet edges
    - # of edges entering & leaving the egonet
- Recursive
  - Some aggregate (mean, sum, max, min, …) of another feature over a node's neighbors
  - Aggregation can be computed over any real-valued feature, including other recursive features







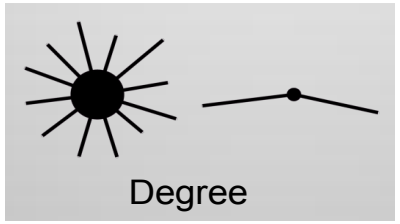


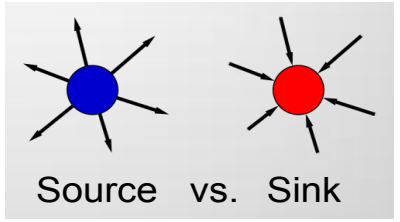
ReFeX: Structural Features

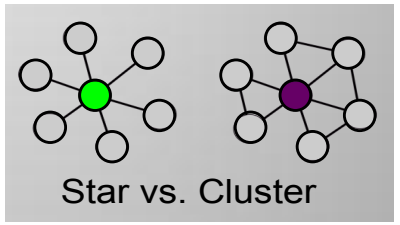
Regional

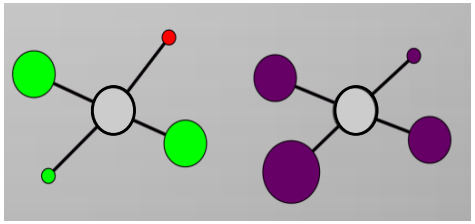
Neighborhood

- Local
  - Essentially measures of the node degree
- Egonet
  - Computed based on each node's ego network
  - Examples
    - # of within-egonet edges
    - # of edges entering & leaving the egonet
- Recursive
  - Some aggregate (mean, sum, max, min, …) of another feature over a node's neighbors
  - Aggregation can be computed over any real-valued feature, including other recursive features









ReFeX (continued)

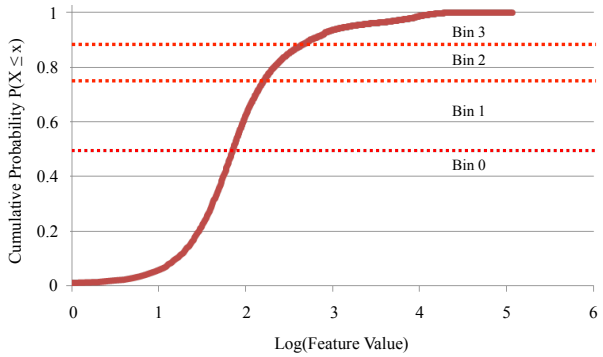
- Number of possible recursive features is infinite

ReFeX (continued)

- Number of possible recursive features is infinite
- ReFeX pruning

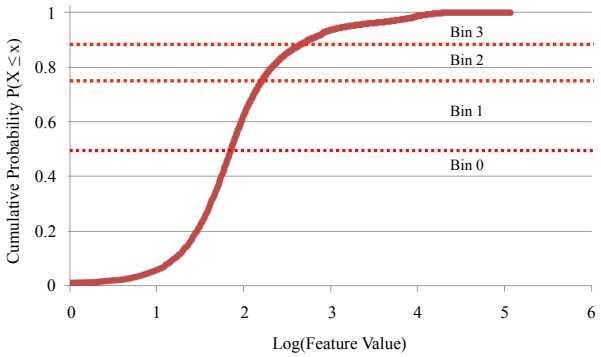
ReFeX (continued)

- Number of possible recursive features is infinite
- ReFeX pruning
  - Feature values are mapped to small integers via vertical logarithmic binning
    - Log binning places most of the discriminatory power among sets of nodes with large feature values



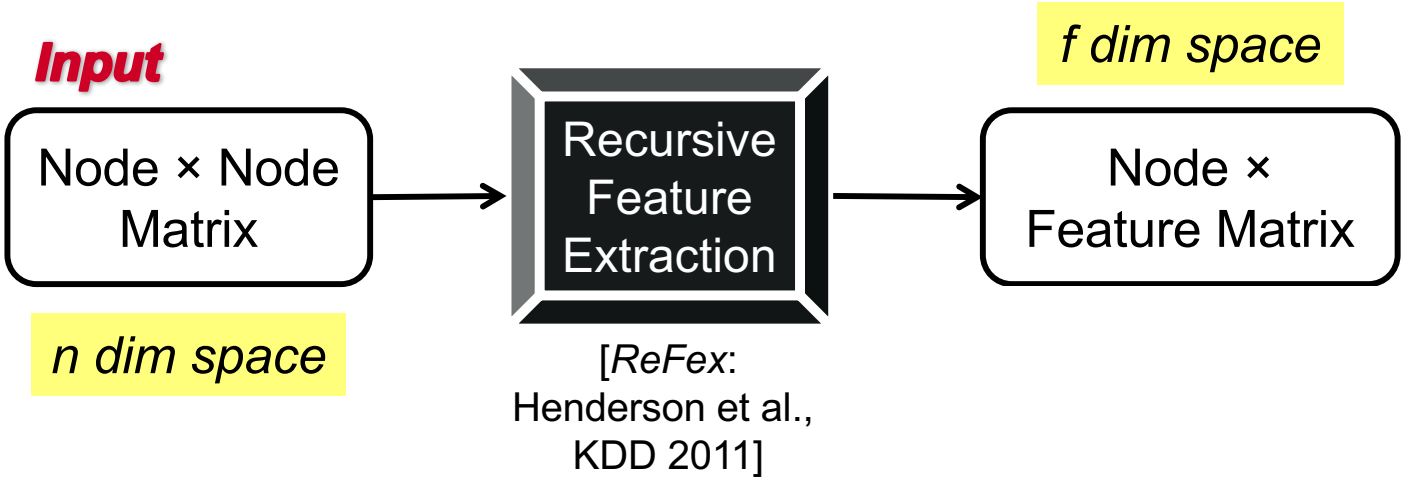
ReFeX (continued)

- Number of possible recursive features is infinite
- ReFeX pruning
  - Feature values are mapped to small integers via vertical logarithmic binning
    - Log binning places most of the discriminatory power among sets of nodes with large feature values

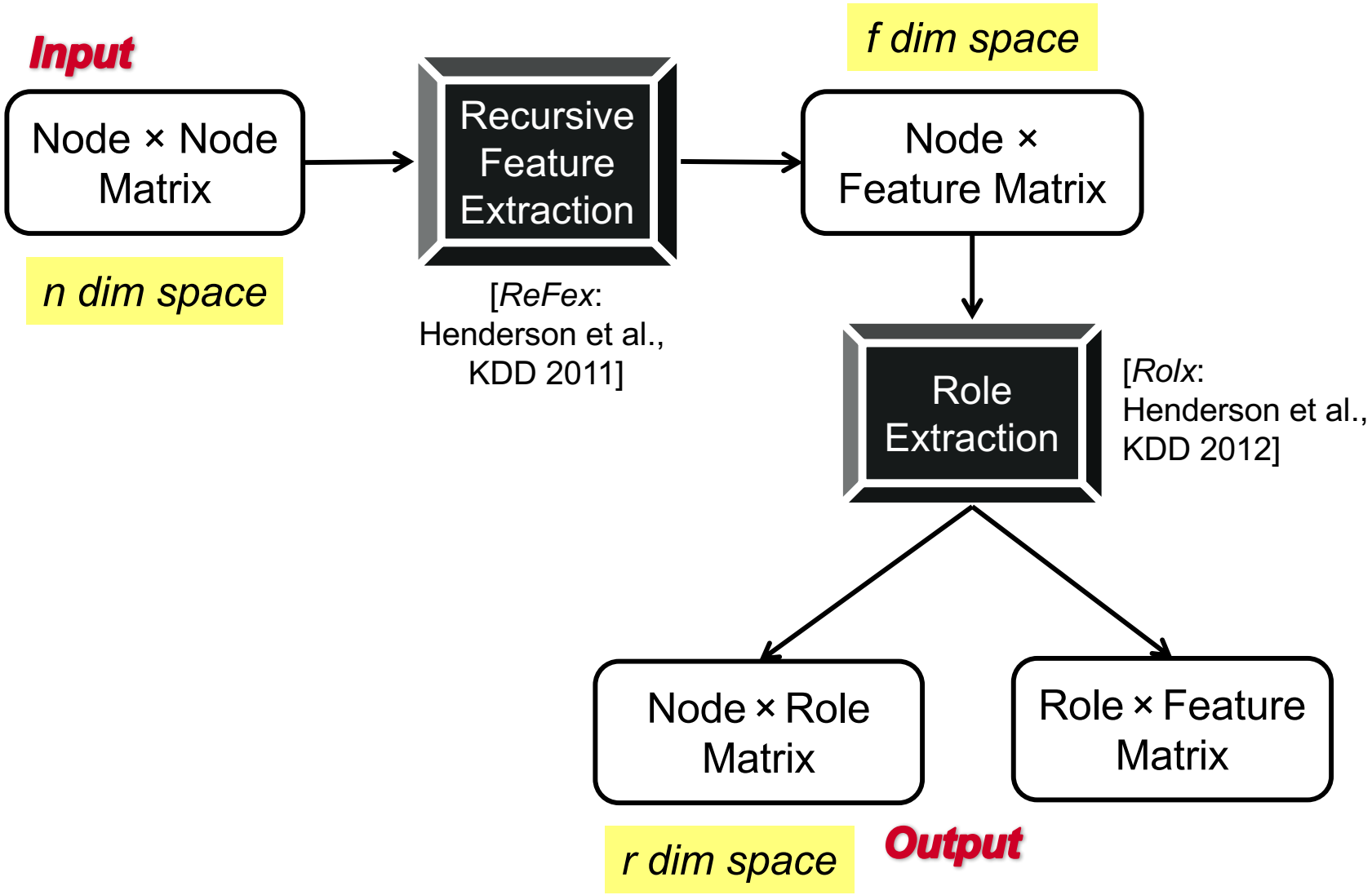


  - Look for pairs of features whose values never disagree by more than a threshold
    - A graph-based approach
    - Threshold automatically set
    - Details in the KDD'11 paper

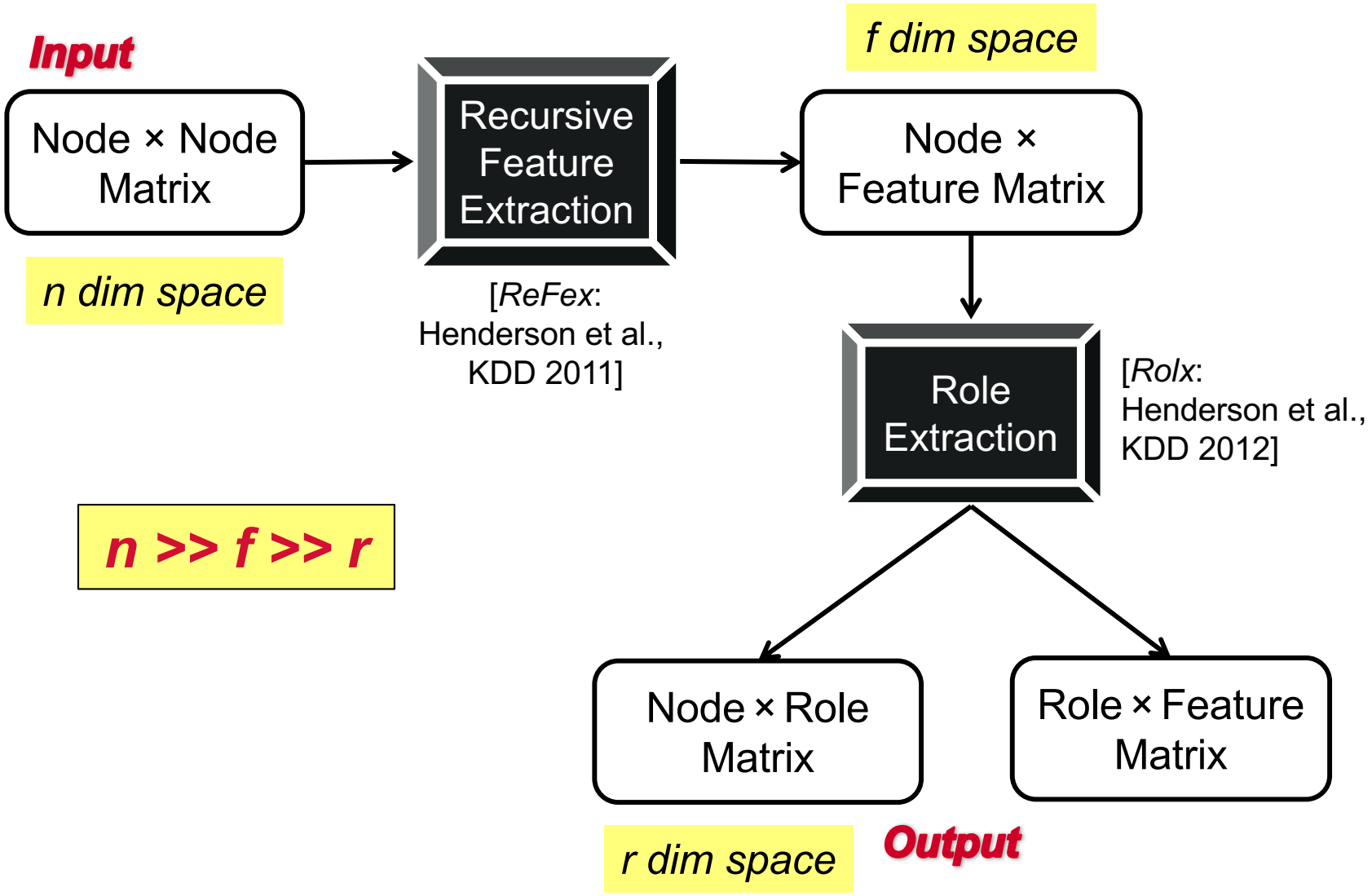
Finding Roles in a Network



Finding Roles in a Network

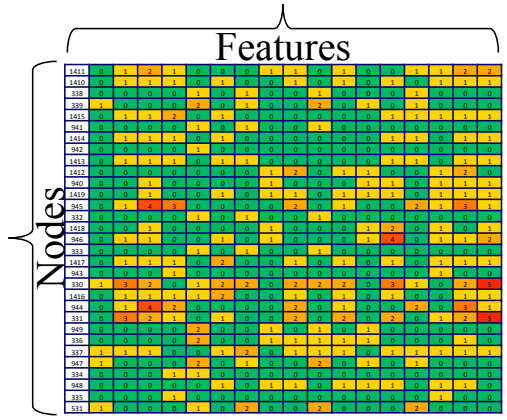


Finding Roles in a Network



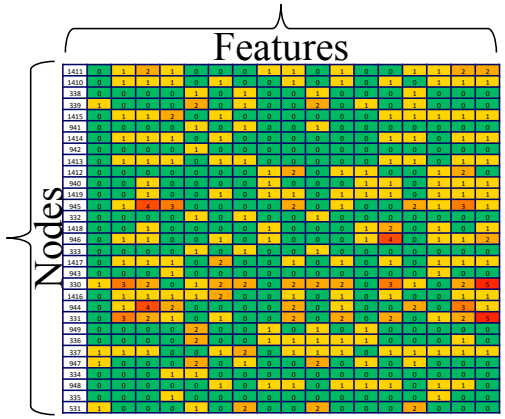
Role Extraction: Feature Grouping

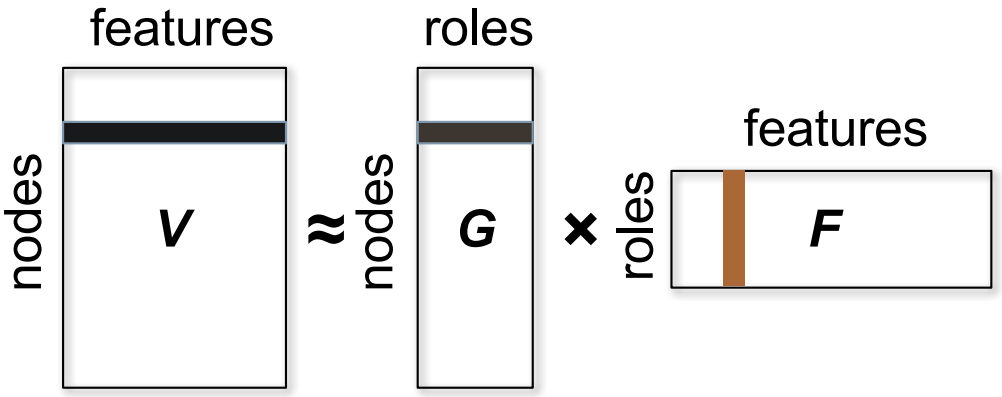
- Soft clustering in the structural feature space
  - Each node has a mixed-membership across roles



Role Extraction: Feature Grouping

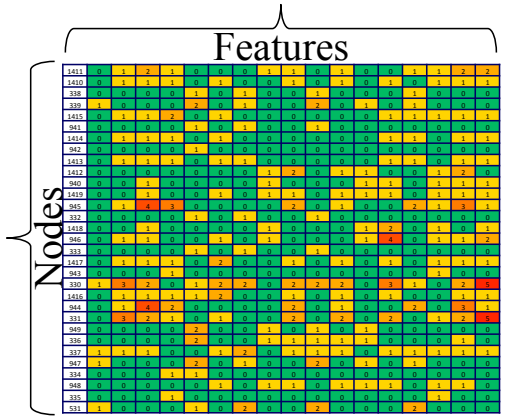
- Soft clustering in the structural feature space
  - Each node has a mixed-membership across roles
- Generate a rank *r* approximation of $V \approx GF$

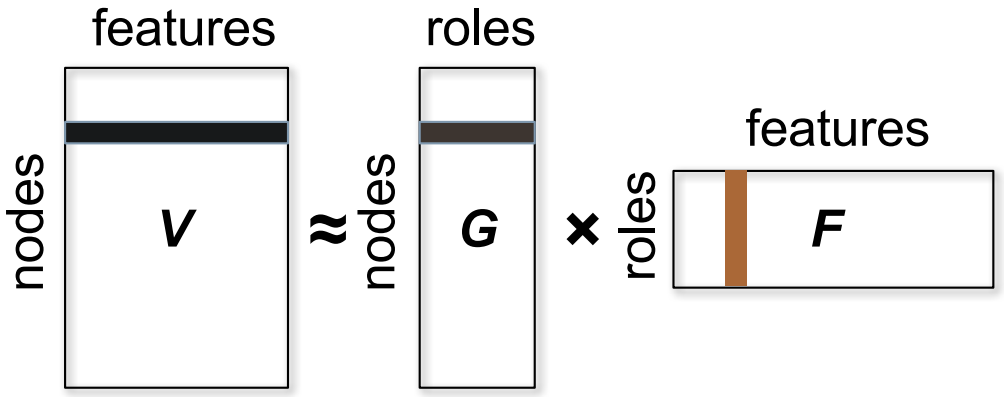




Role Extraction: Feature Grouping

- Soft clustering in the structural feature space
  - Each node has a mixed-membership across roles
- Generate a rank *r* approximation of $V \approx GF$





- RolX uses NMF for feature grouping
  - Computationally efficient
  - Non-negative factors simplify interpretation of roles and memberships

$$\text{argmin}_{G,F} \|V - GF\|_{fro}, \text{s.t. } G \geq 0,\ F \geq 0$$

Role Extraction: Model Selection

- Roles summarize behavior
  - Or, they compress the feature matrix, *V*

Role Extraction: Model Selection

- Roles summarize behavior
  - Or, they compress the feature matrix, V
- Use MDL to select the model size r that results in the best compression
  - L: description length
  - M: # of bits required to describe the model
  - E: cost of describing the reconstruction errors in $V - GF$
  - Minimize $L = M + E$

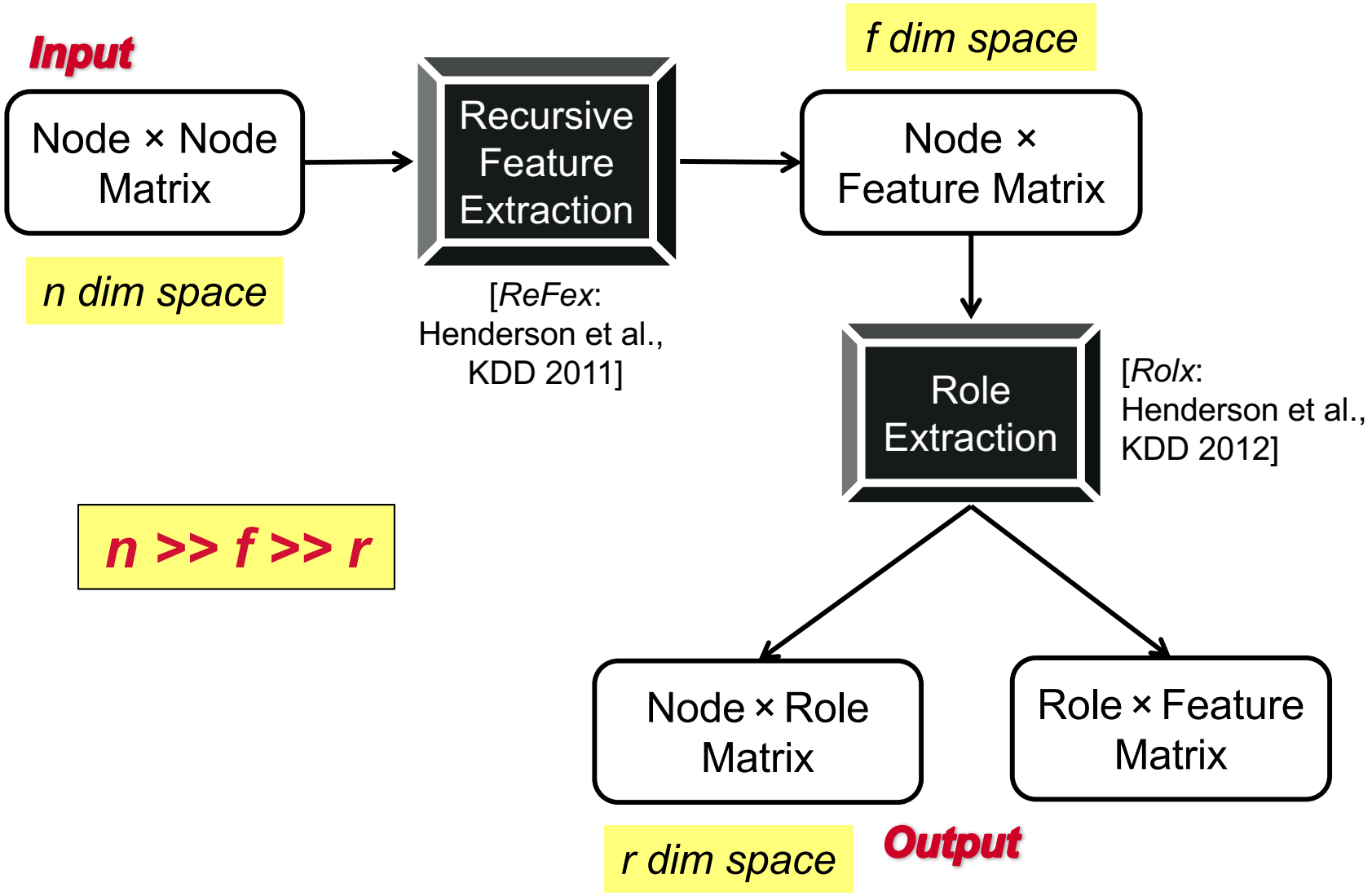
Role Extraction: Model Selection

- Roles summarize behavior
  - Or, they compress the feature matrix, *V*
- Use MDL to select the model size *r* that results in the best compression
  - *L*: description length
  - *M*: # of bits required to describe the model
  - *E*: cost of describing the reconstruction errors in *V* – *GF*
  - Minimize *L* = *M* + *E*
    - To compress high-precision floating point values, RolX combines Llyod-Max quantization with Huffman codes
    - Errors in *V*-*GF* are not distributed normally, RolX uses KL divergence to compute *E*

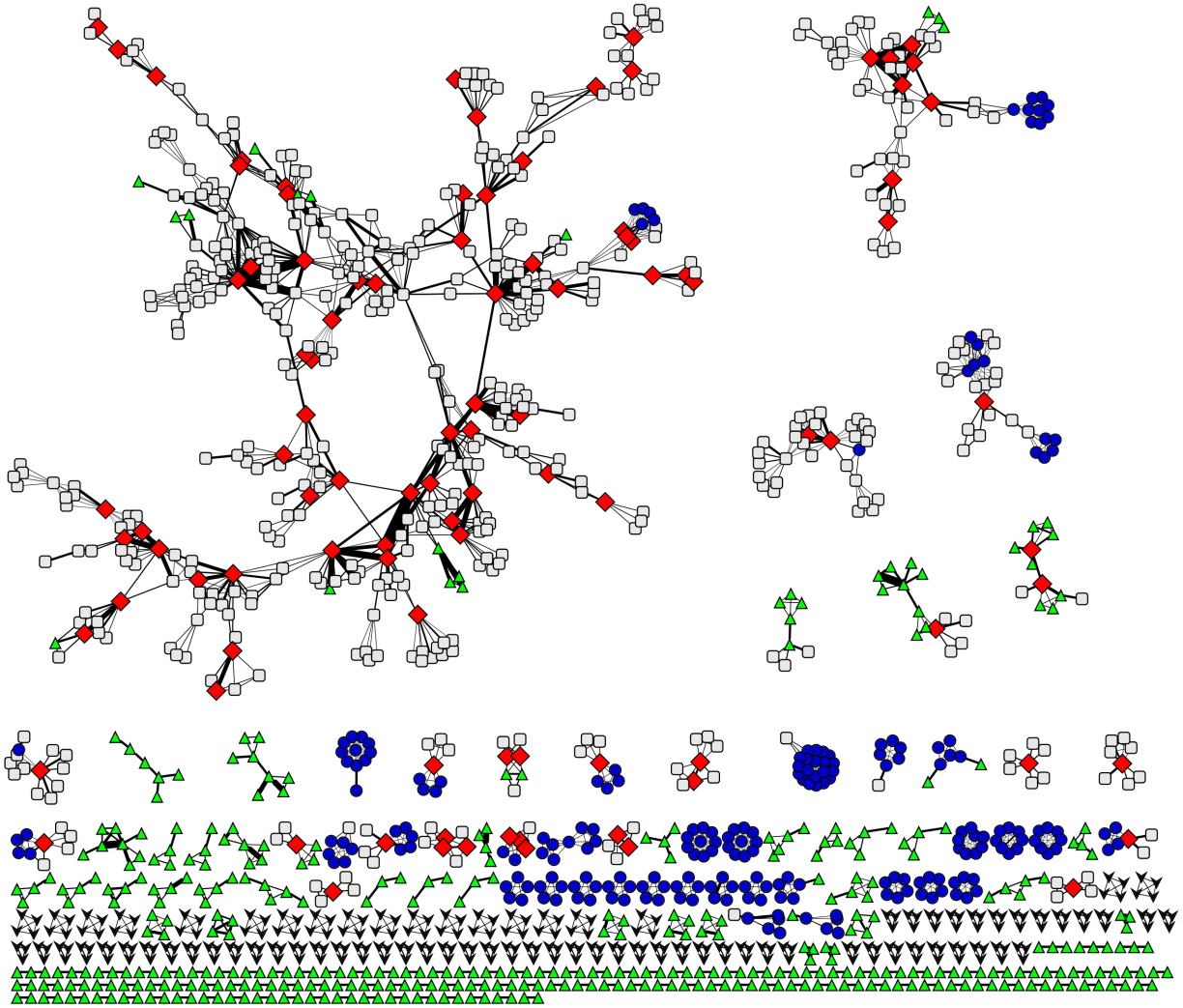
$$M = \bar{b}r(n+f)$$

$$E = \sum_{i,j}\left(V_{i,j}\log\frac{V_{i,j}}{(GF)_{i,j}} - V_{i,j} + (GF)_{i,j}\right)$$

Finding Roles in a Network

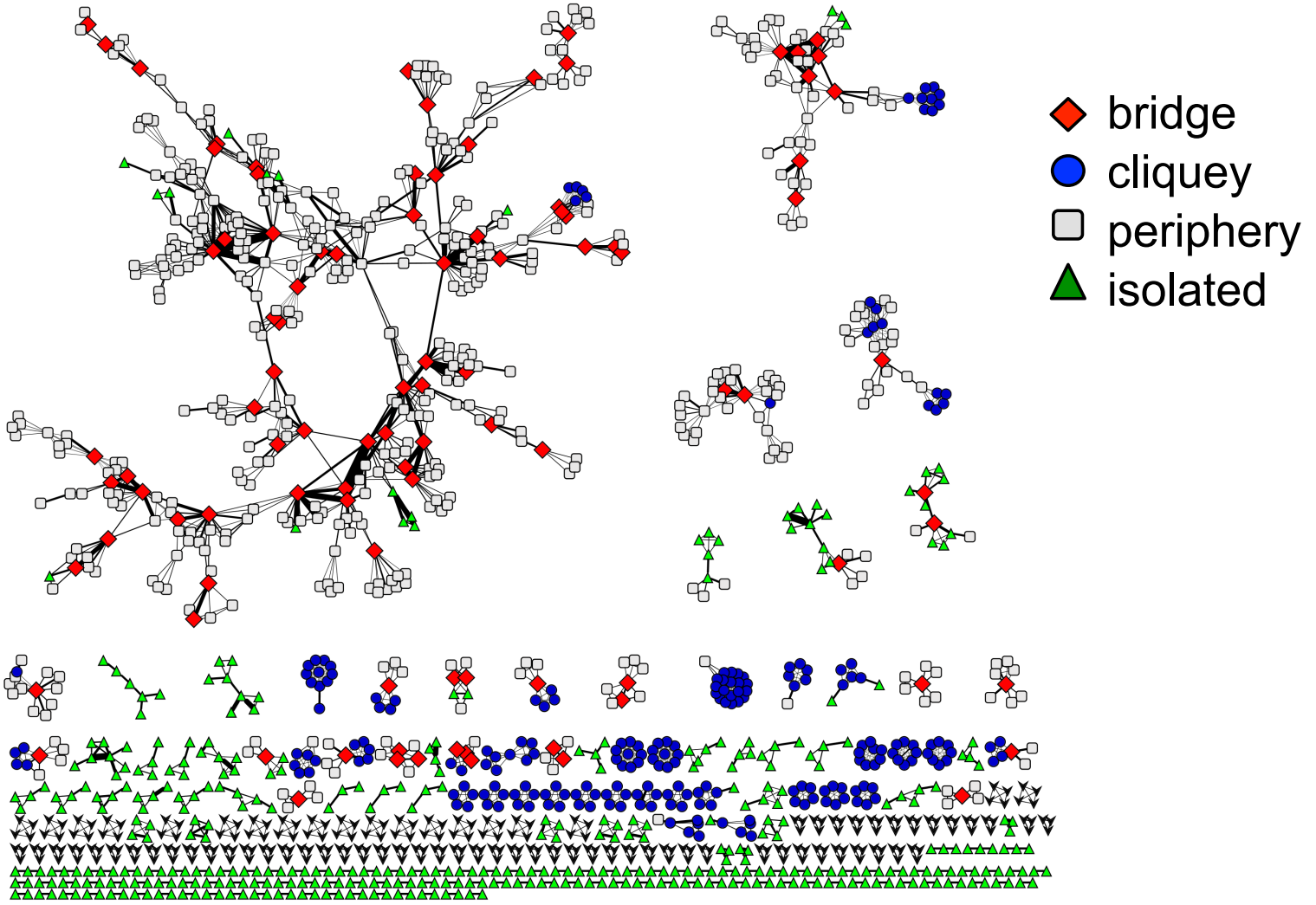


Automatically Discovered Roles



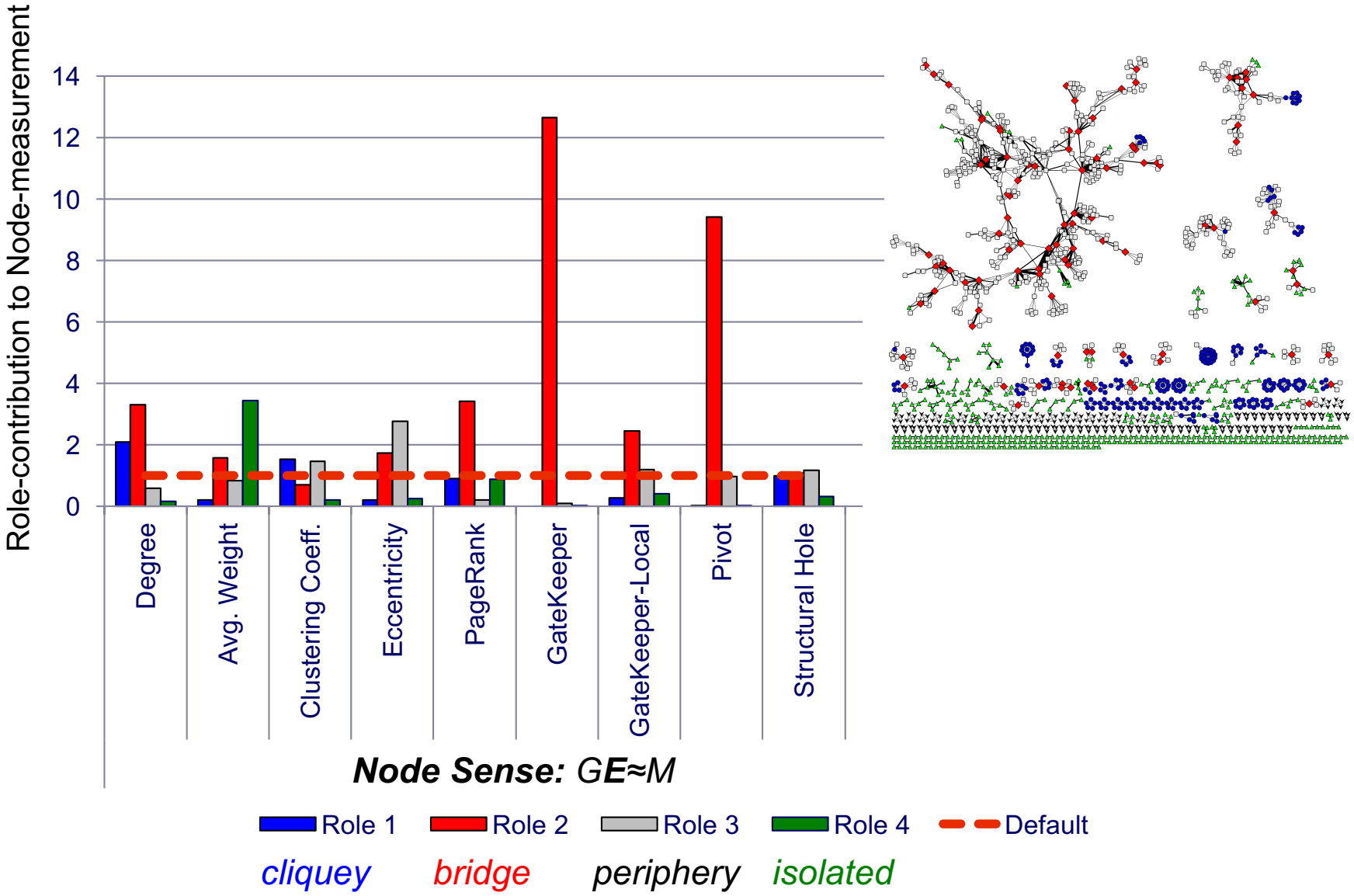
Network Science Co-authorship Graph
[Newman 2006]

Automatically Discovered Roles

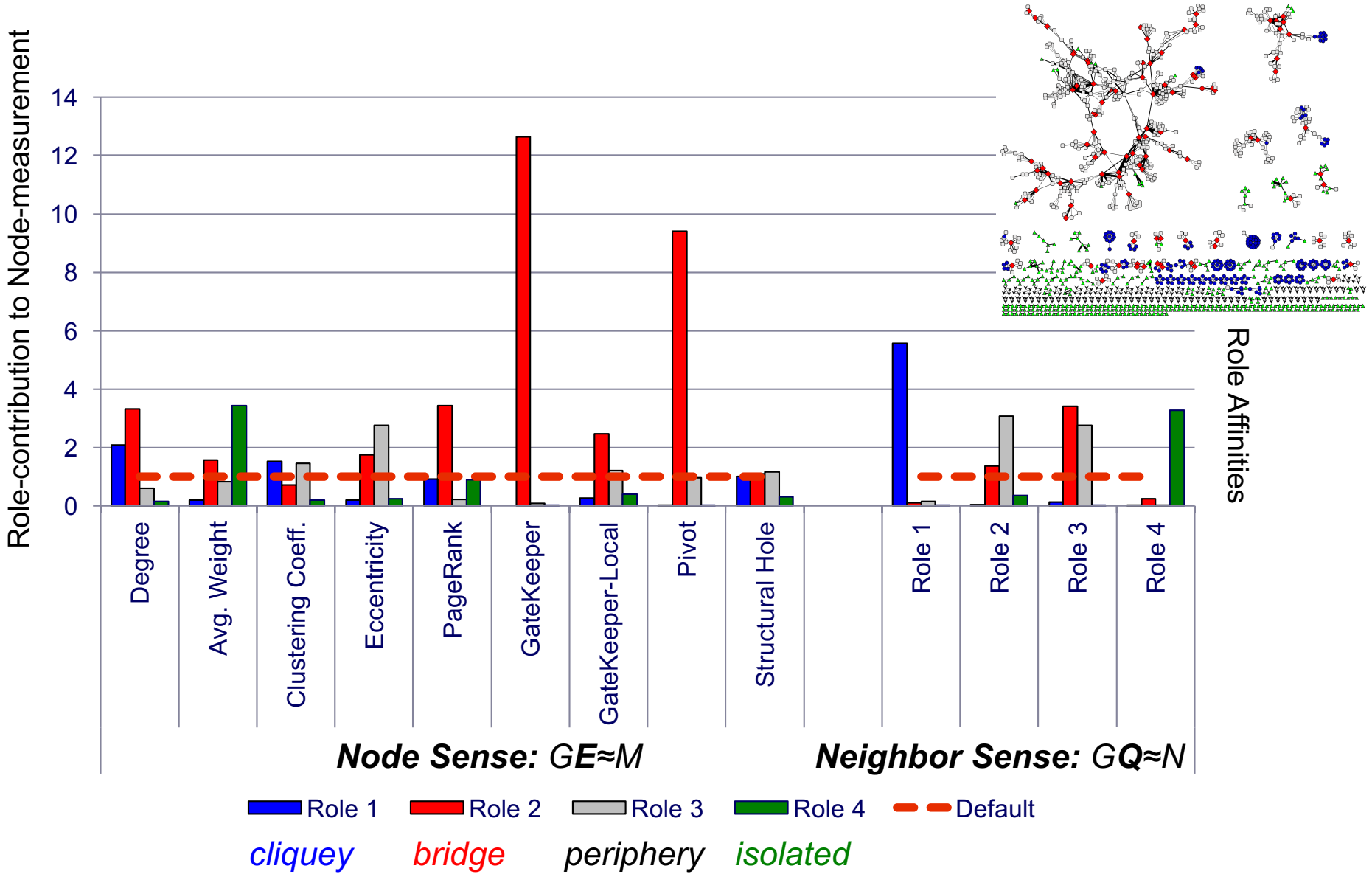


Network Science Co-authorship Graph
[Newman 2006]

Making Sense of Roles

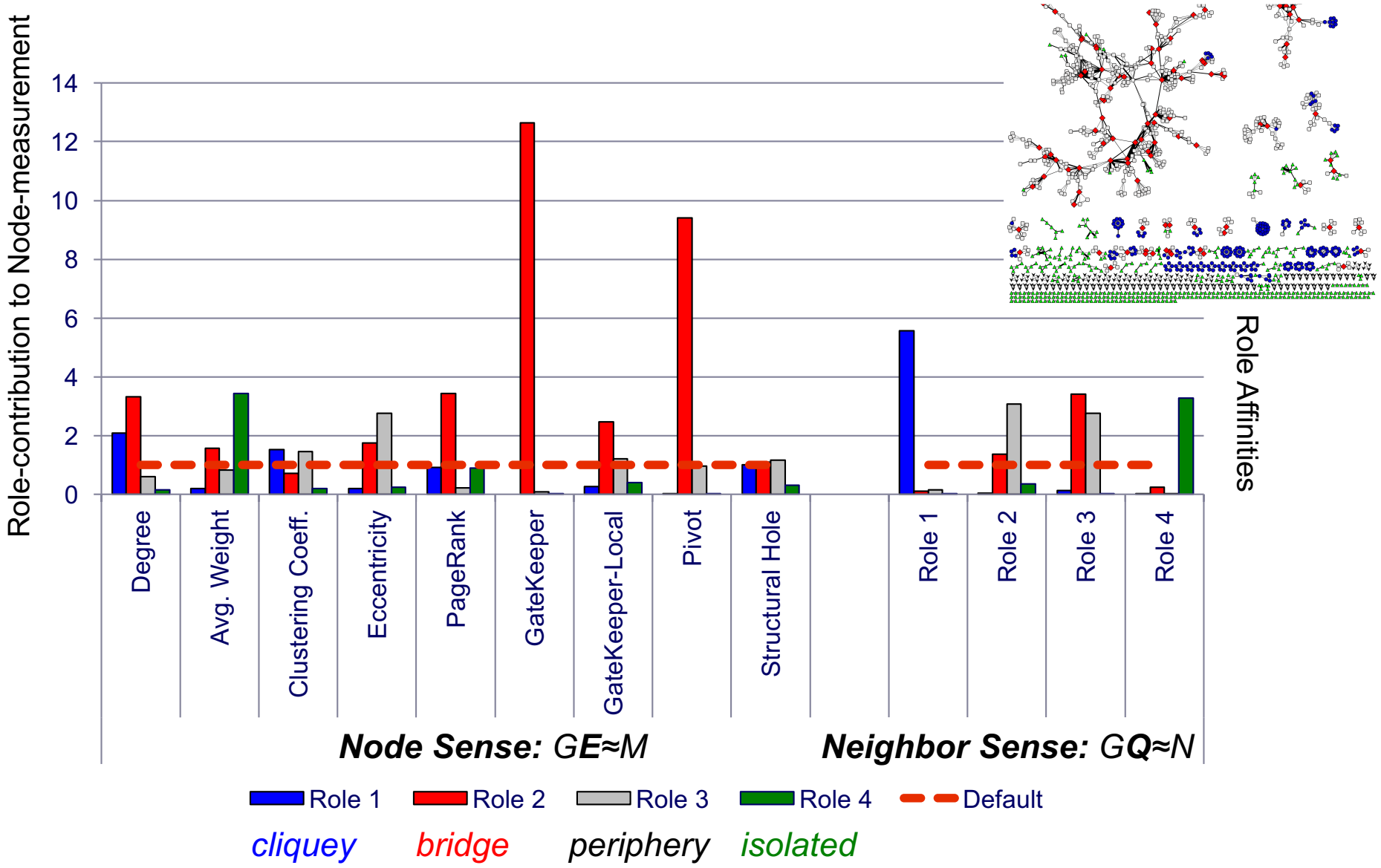


Making Sense of Roles



Making Sense of Roles

Topological measures & role homophily help interpret roles.



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
…	…

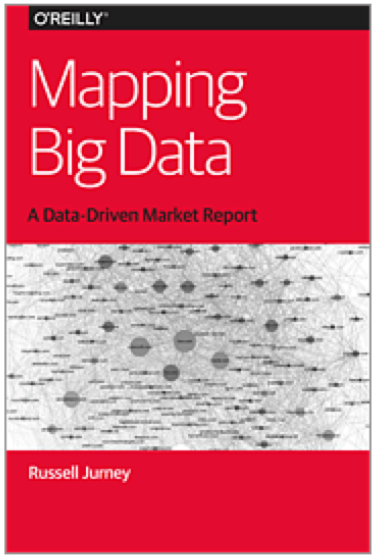
Applications of role discovery

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
…	…

O'REILLY®

Search

Home | Shop Video Training & Books | Radar | Safari Books Online | Conferences



Mapping Big Data

A Data-Driven Market Report

By Russell Jurney

Publisher: O'Reilly
Released: September 2015

Description

To discover the shape and structure of the big data market, the San Francisco-based startup Relato took a unique approach to market research and created the first fully data-driven market report. Company CEO Russell Jurney and his team collected and analyzed raw data from a variety of sources to reveal a boatload of business insights about the big data space. This exceptional report is now available for free download.

Using data analytic techniques such as social network analysis (SNA), Relato exposed the vast and complex partnership network that exists among tens of thousands of unique big data vendors. The dataset Relato collected is centered around Cloudera, Hortonworks, and MapR, the major platform vendors of Hadoop, the primary force behind this market.

From this snowball sample, a 2-hop network, the Relato team was able to answer several questions, including:

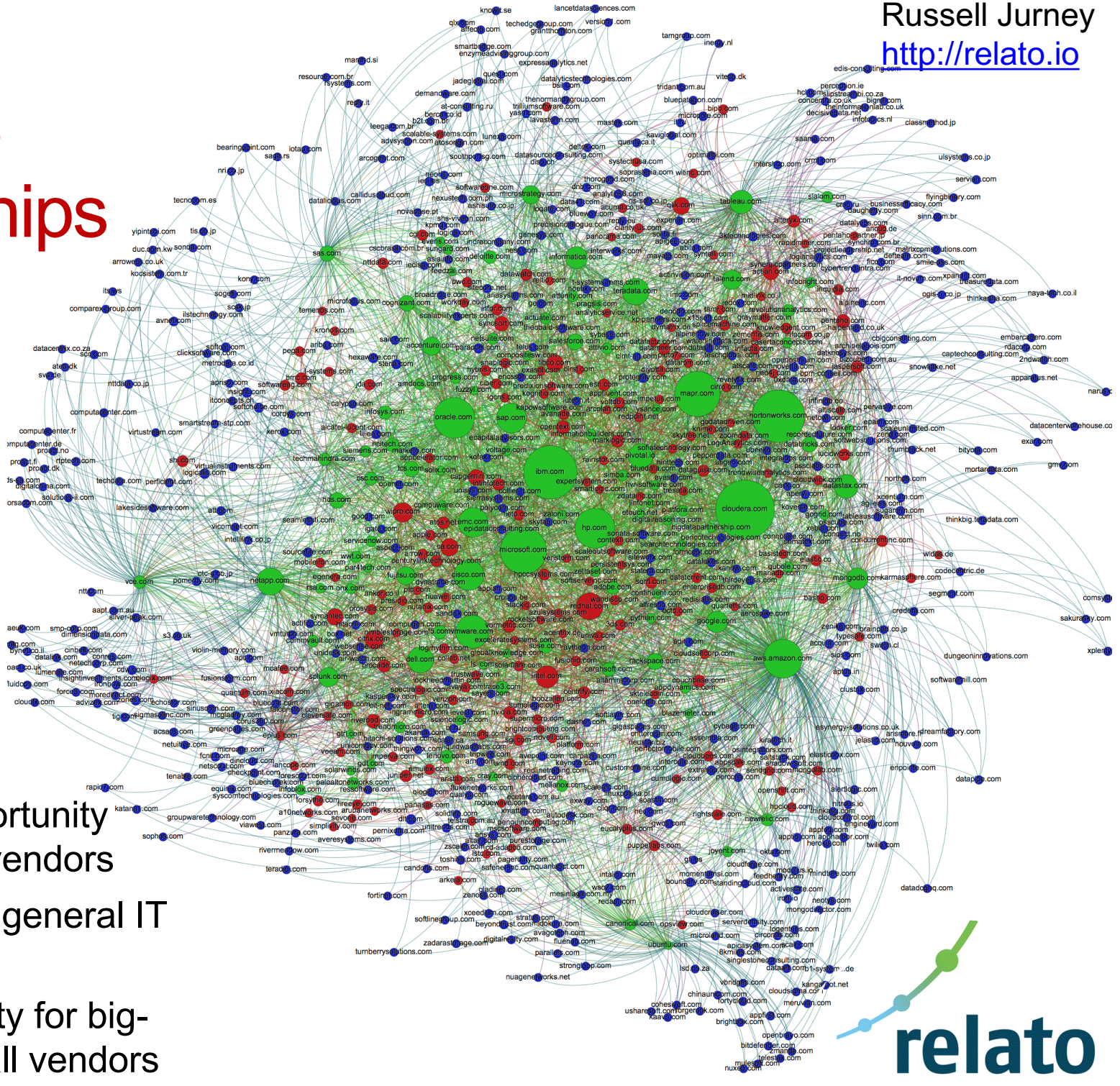
- Who are the major players in the big data market?
- Which is the leading Hadoop vendor?
- What sectors are included in this market and how do they relate?
- Which among the thousands of partnerships are most important?
- Who's doing business with whom?

Metrics used in this report are also visible in Relato's interactive web application, via a link in the report, which walks you through the insights step-by-step.

Russell Jurney is CEO of Relato, a San Francisco area startup that maps markets to drive sales and marketing. He is the author of Agile Data Science *and co-author of* Big Data for Chimps *(both O'Reilly). In addition, Russell is an Apache Committer on the Incubating DataFu project. Russell is a full stack engineer.*

Big data business partnerships

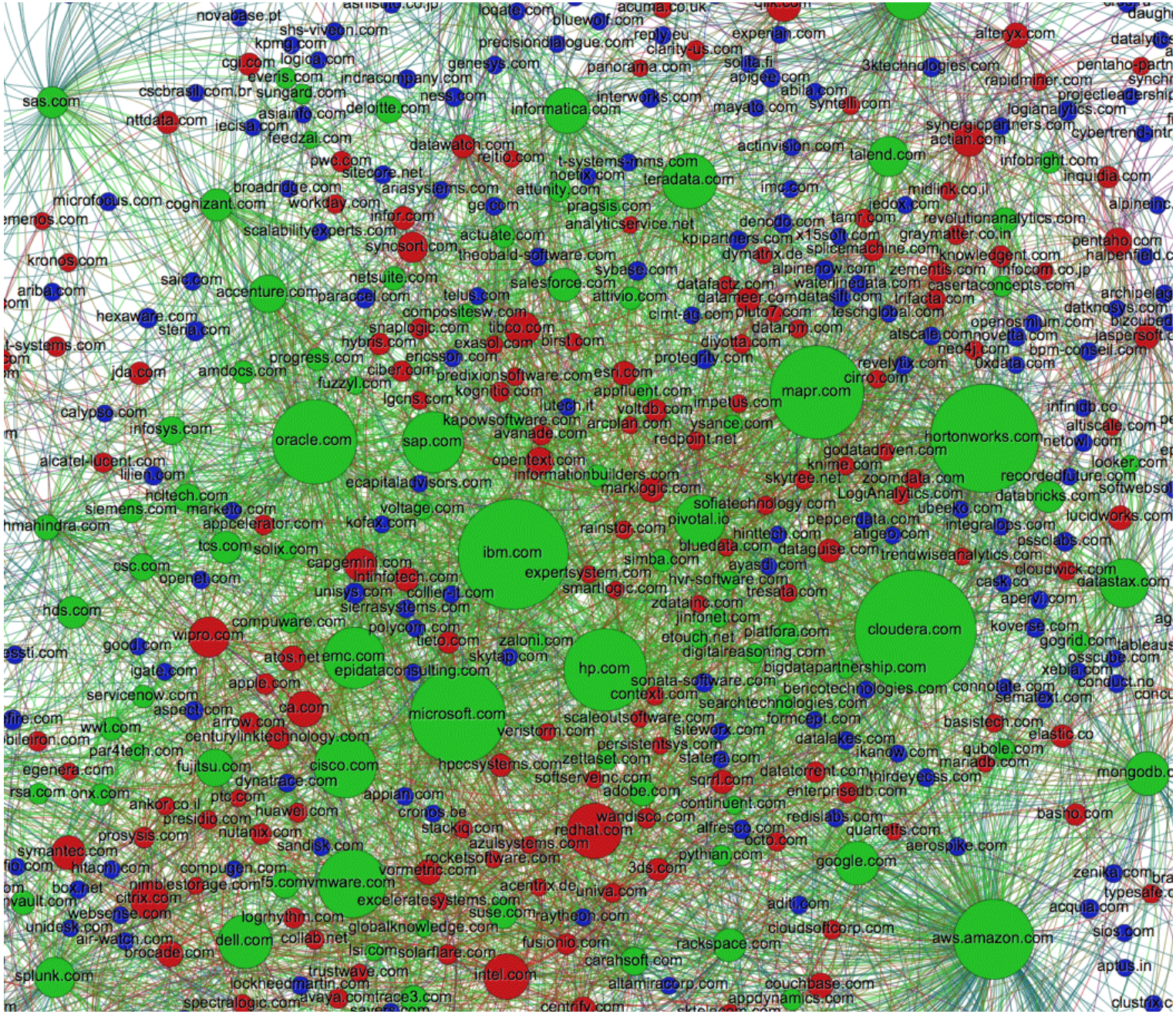
Russell Jurney
http://relato.io



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

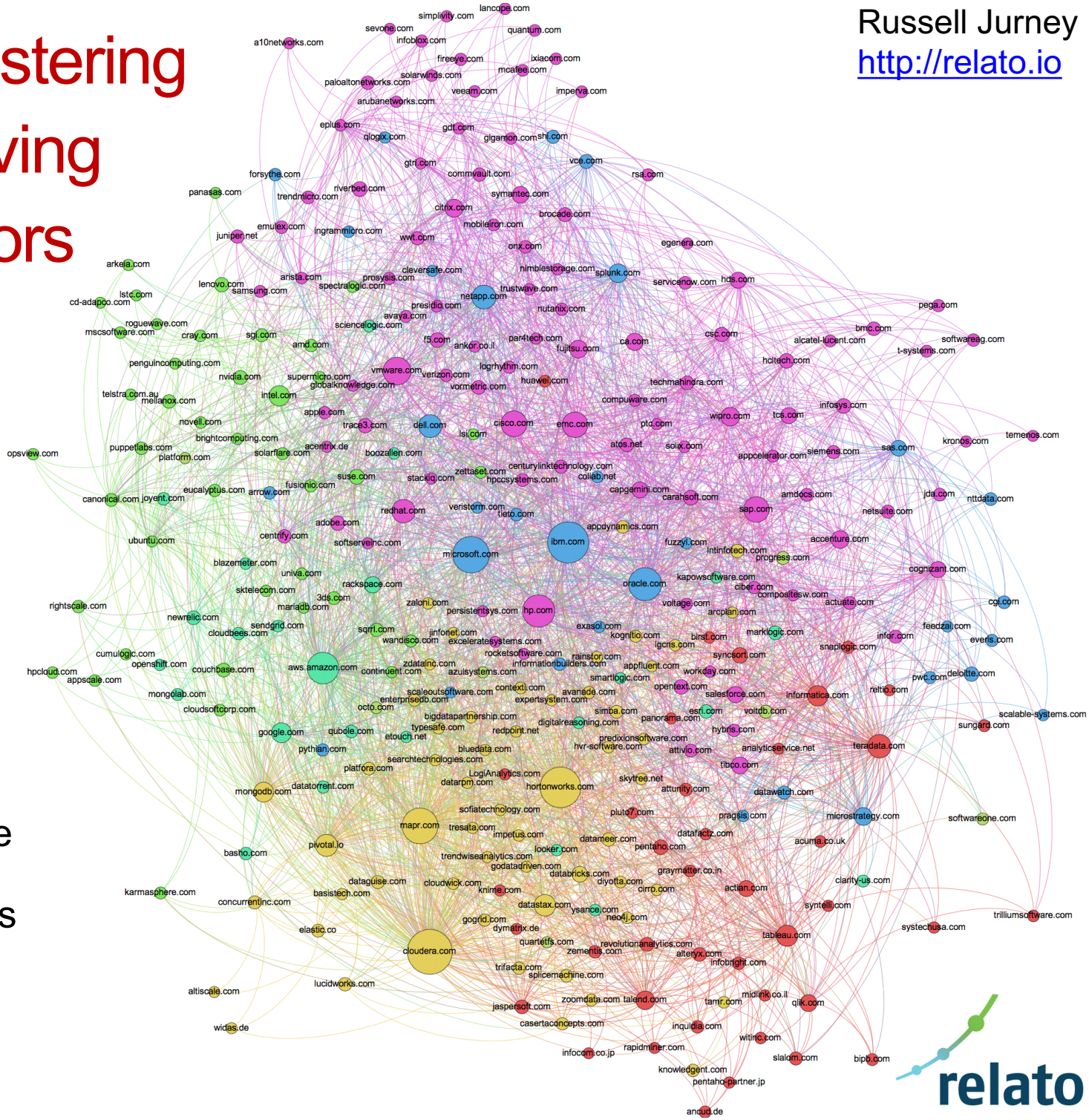
Big-data business-partnerships

Russell Jurney
http://relato.io



Louvain Clustering After Removing Small Vendors (Blue Role)

Russell Jurney
http://relato.io

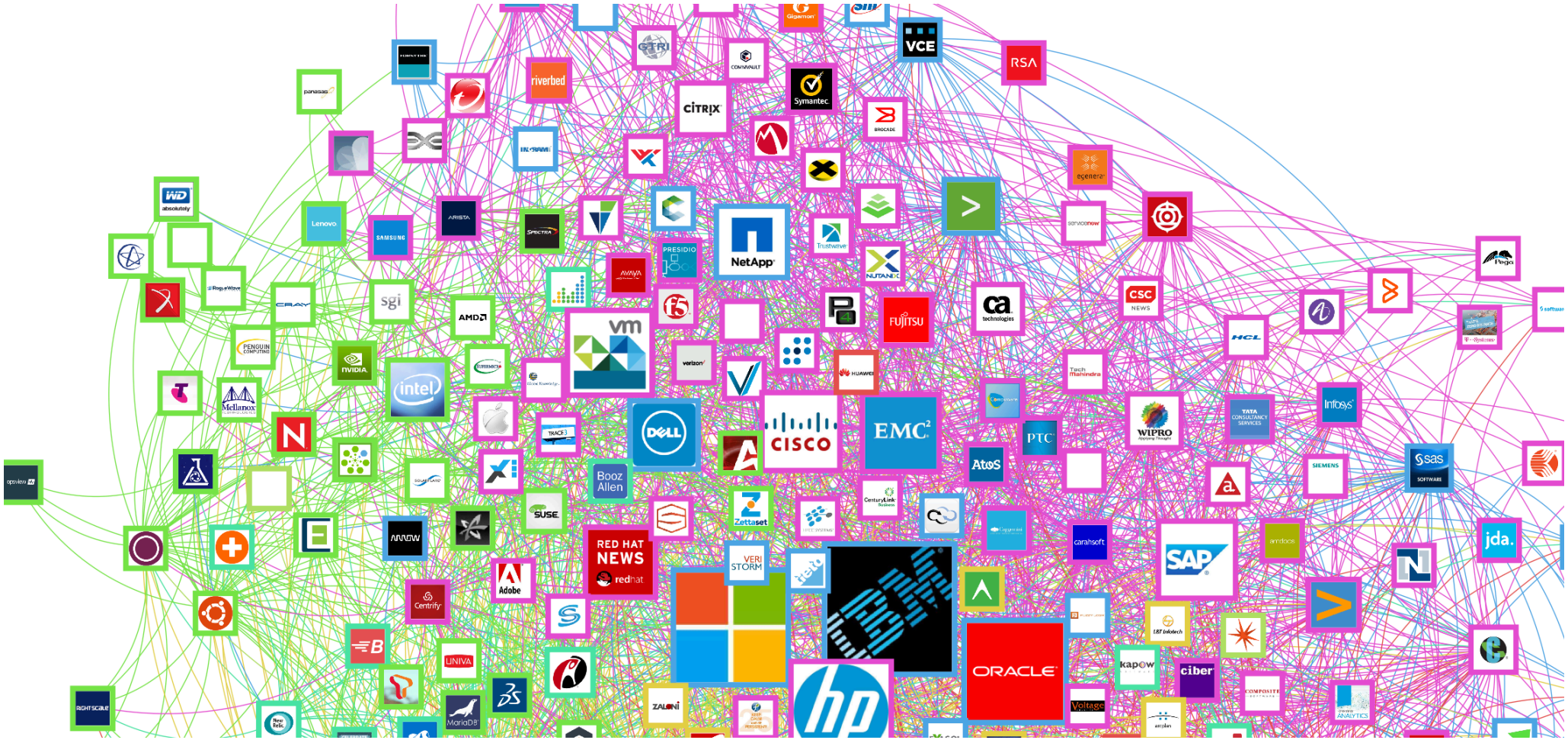


- Analytics Software
- Cloud Computing
- Enterprise Software
- New Data Platforms
- Old Data Platforms
- Servers

relato

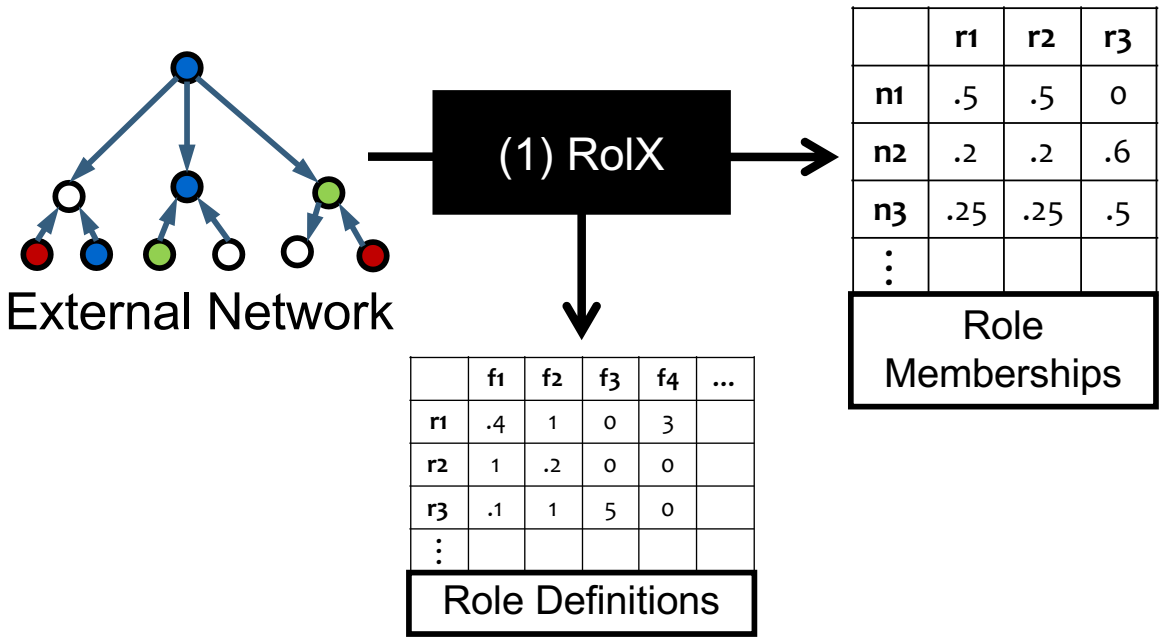
An Interactive Market Map of the Big Data Space

- http://demo.relato.io/oreilly and http://demo.relato.io/public



Role Transfer = RolX + SL

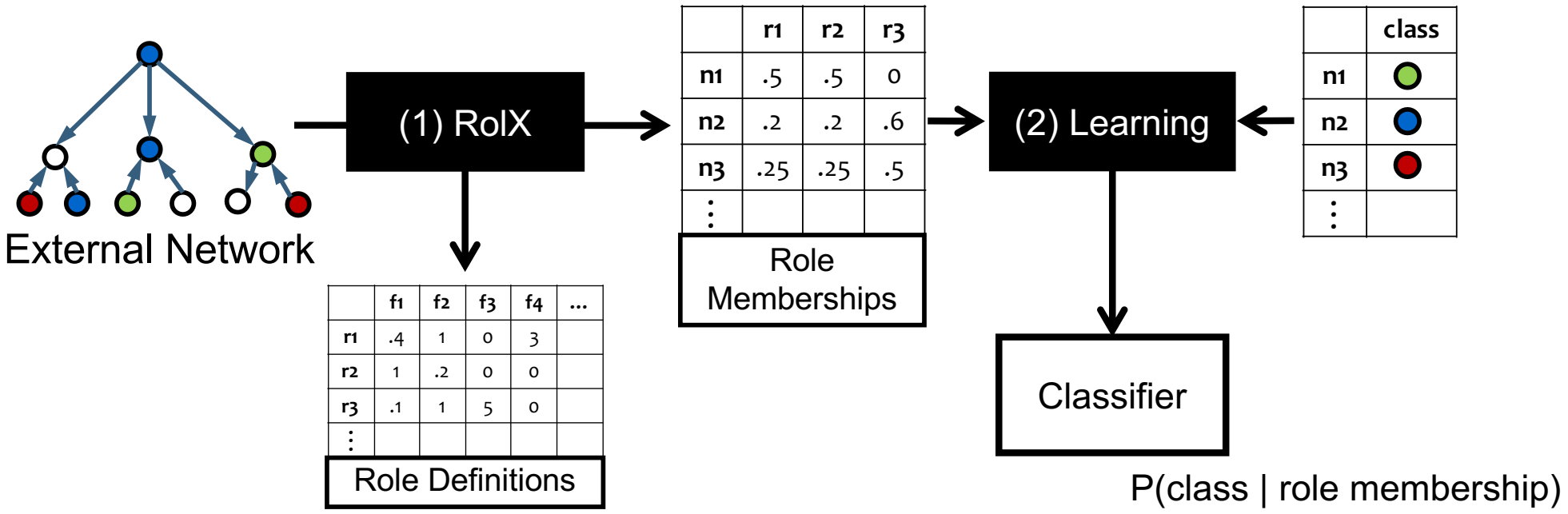
Role Transfer = RolX + SL



	r1	r2	r3
n1	.5	.5	0
n2	.2	.2	.6
n3	.25	.25	.5
⋮			

	f1	f2	f3	f4	...
r1	.4	1	0	3	
r2	1	.2	0	0	
r3	.1	1	5	0	
⋮					

Role Transfer = RolX + SL

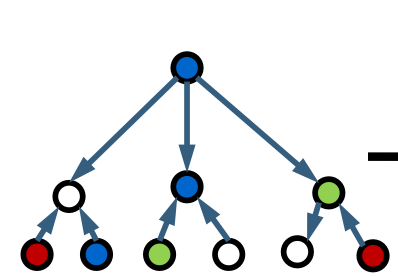


	f1	f2	f3	f4	...
r1	.4	1	0	3	
r2	1	.2	0	0	
r3	.1	1	5	0	
⋮					

	r1	r2	r3
n1	.5	.5	0
n2	.2	.2	.6
n3	.25	.25	.5
⋮			

	class
n1	
n2	
n3	
⋮	

Role Transfer = RolX + SL



External Network

(1) RolX

	f1	f2	f3	f4	...
r1	.4	1	0	3	
r2	1	.2	0	0	
r3	.1	1	5	0	
⋮					

Role Definitions

	r1	r2	r3
n1	.5	.5	0
n2	.2	.2	.6
n3	.25	.25	.5
⋮			

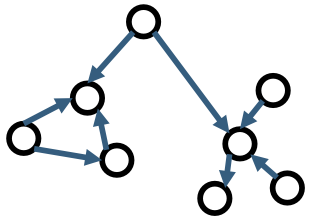
Role Memberships

(2) Learning

	class
n1	
n2	
n3	
⋮	

Classifier

P(class | role membership)



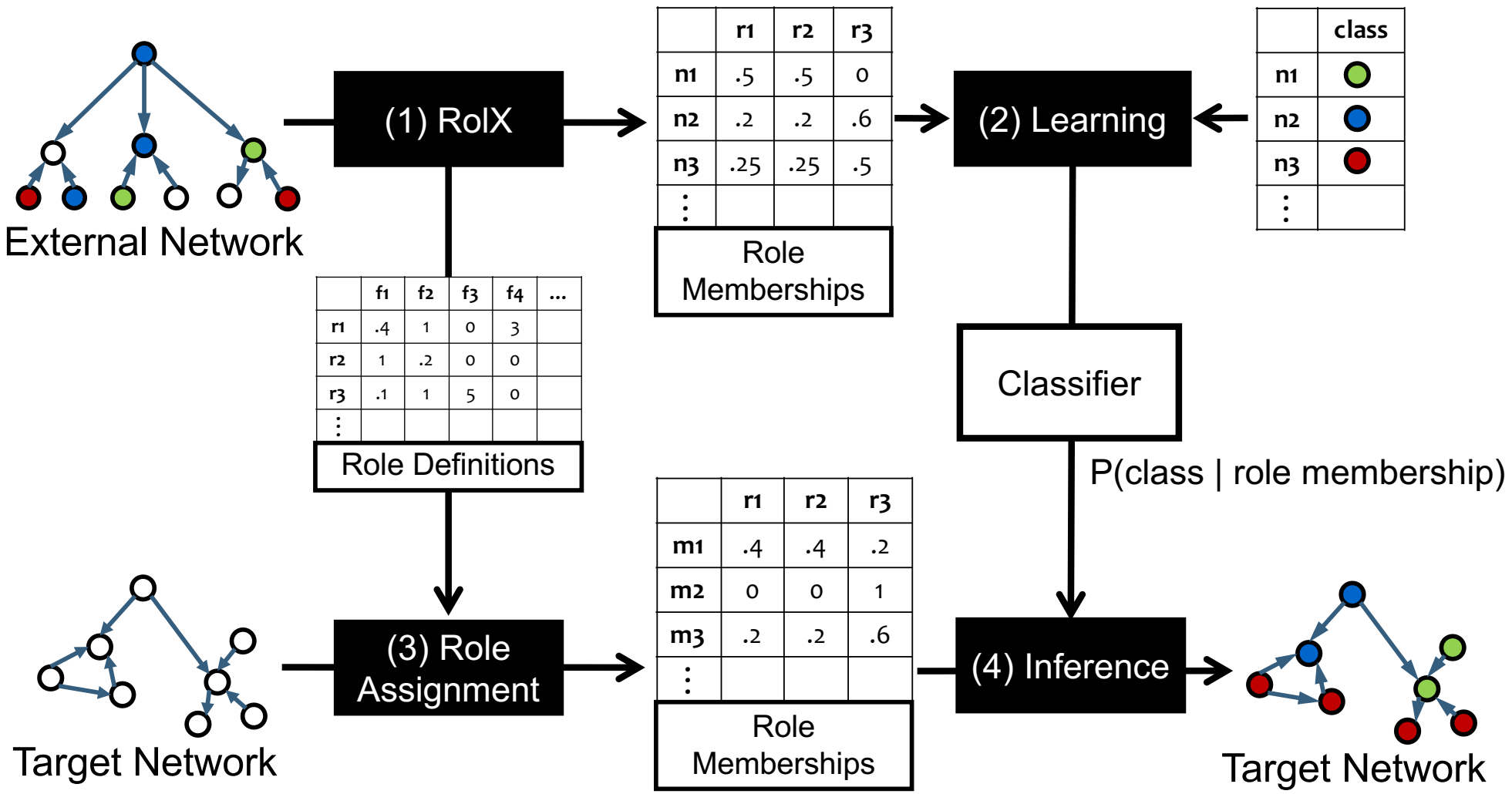
Target Network

(3) Role Assignment

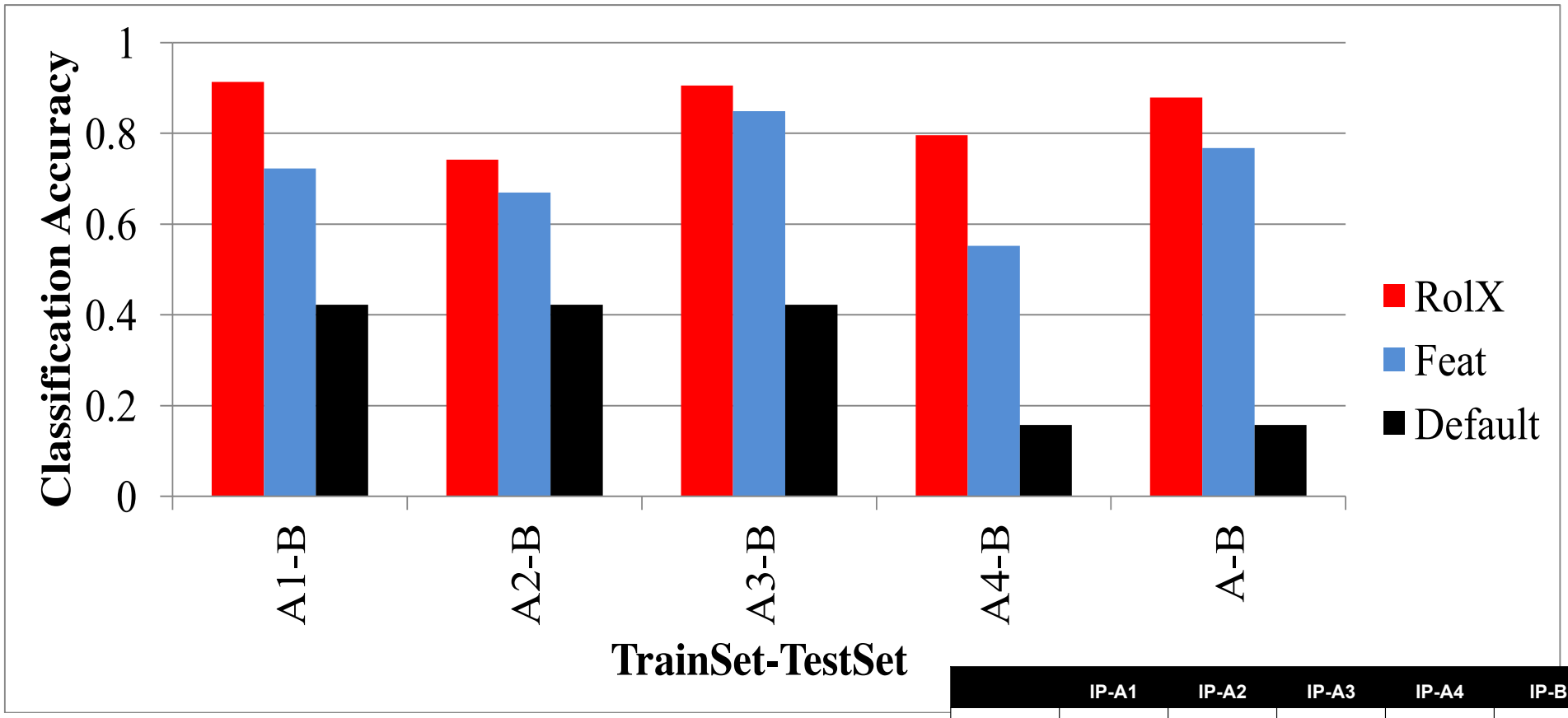
	r1	r2	r3
m1	.4	.4	.2
m2	0	0	1
m3	.2	.2	.6
⋮			

Role Memberships

Role Transfer = RolX + SL



Roles Generalize Across Disjoint Networks



	IP-A1	IP-A2	IP-A3	IP-A4	IP-B
# Nodes	81,450	57,415	154,103	206,704	181,267
% labeled	36.7%	28.1%	20.1%	32.9%	15.3%
# Links	968,138	432,797	1,266,341	1,756,082	1,945,215
(# unique)	206,112	137,822	358,851	465,869	397,925
Class Distribu-tion					

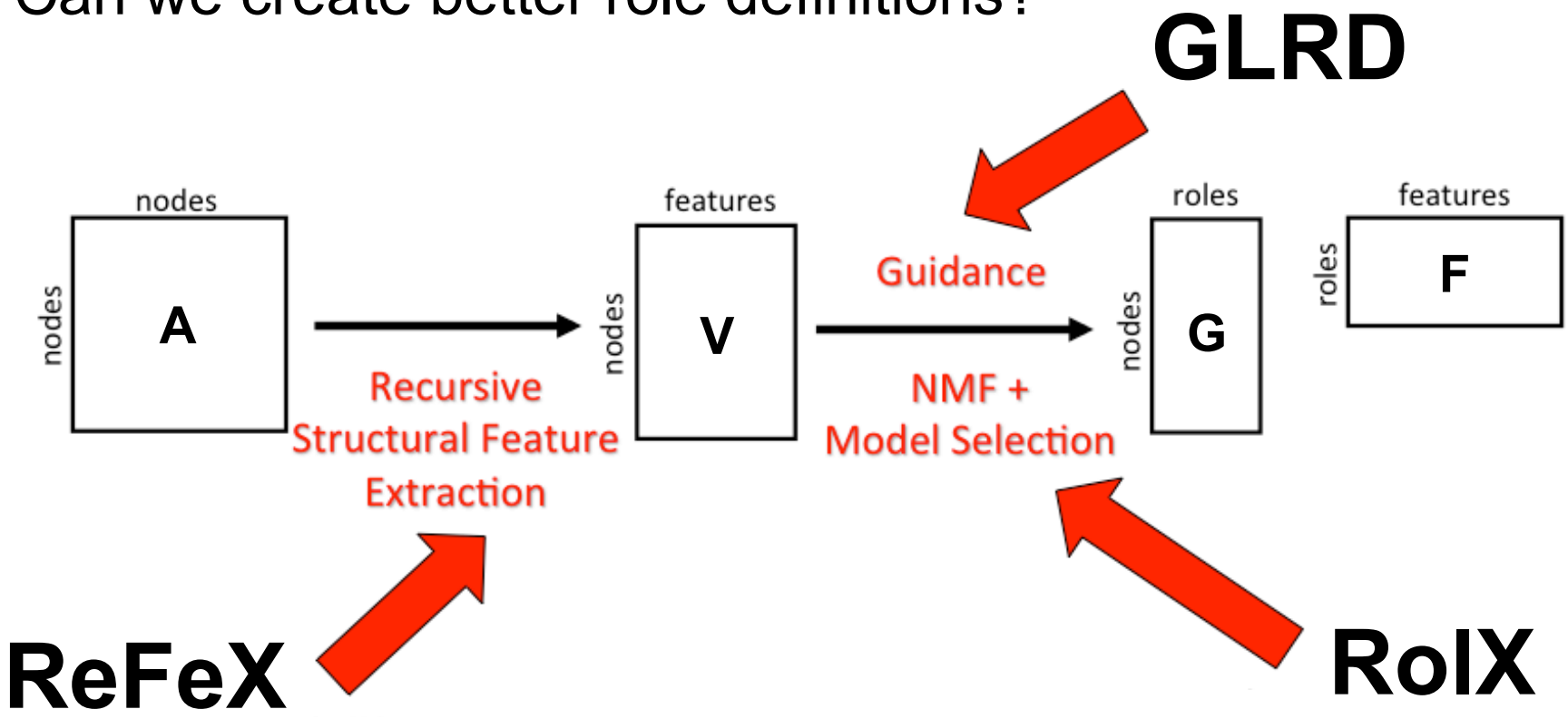
2nd Generation Algorithms for Role Discovery

- *GLRD*: guided learning for role discovery
  - [Gilpin et al., KDD 2013]
- *DBMM*: dynamic behavioral mixed-membership model
  - [Rossi et al., WSDM 2013]
- *RC-Joint*: simultaneous detection of communities and roles
  - [Ruan & Parthasarathy, COSN 2014]
- Motif-Role-Fingerprints
  - [McDonnell et al., PLoS ONE 9(12), 2014]
- Dynamic inference of social roles in information cascades
  - [Choobdar et al., DMKD 29(5), 2015]

- *MRD*: multi-relational role discovery
  - [Gilpin et al., ArXiv 2016]
- *DERM:* dynamic edge role mixed-membership model
  - [Ahmed et al., ArXiv 2016]
- node2vec: Scalable Feature Learning for Networks
  - [Grover and Leskovec, KDD 2016]
- A combinatorial approach to role discovery
  - [Arockiasamy et al., ICDM 2016]
- struc2vec: Learning Node Representations from Structural Identity
  - [Ribeiro et al., KDD 2017]
- …

Guided Learning for Role Discovery

- [Gilpin *et al.*, KDD 2013]
- What if we had guidance on roles?
  - Guidance as in weak supervision encoded as constraints
- Can we create better role definitions?



GLRD Framework

- Constraints on columns of G (i.e., role assignments) or rows of F (i.e. role definitions) are convex functions

$$\begin{aligned} \underset{\mathbf{G},\mathbf{F}}{\text{minimize}} \quad & ||\mathbf{V} - \mathbf{GF}||_2 \\ \text{subject to} \quad & g_i(\mathbf{G}) \leq d_{Gi},\ i = 1, \ldots, t_G \\ & f_i(\mathbf{F}) \leq d_{Fi},\ i = 1, \ldots, t_F \\ \text{where} \quad & g_i \text{ and } f_i \text{ are convex functions.} \end{aligned}$$

- Use an alternative least squares (ALS) formulation
  - Do not alternate between solving for the entire G and F
  - Solve for one column of G or one row of F at a time
    - This is okay since we have convex constraints

Sparsity Constraint

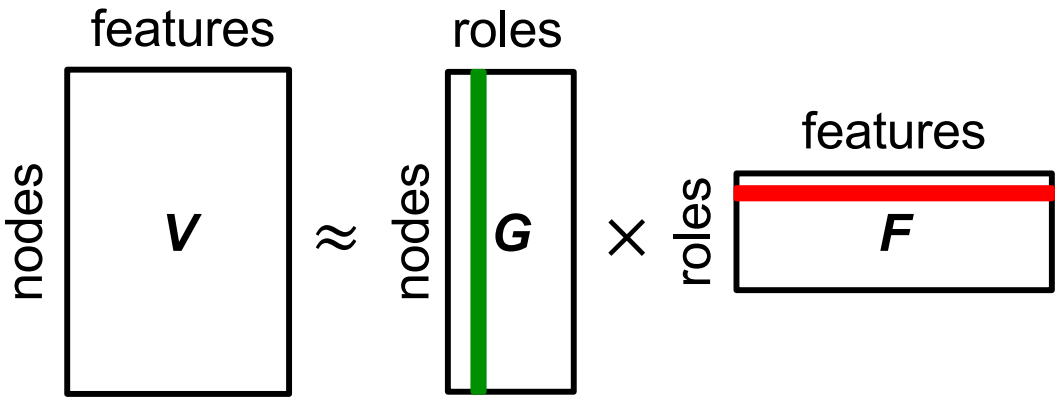
$$\underset{\mathbf{G},\mathbf{F}}{\text{argmin}} \quad ||\mathbf{V} - \mathbf{GF}||_2$$

subject to: $\mathbf{G} \geq 0, \mathbf{F} \geq 0$

$\forall i \quad ||\mathbf{G_{\bullet i}}||_1 \leq \epsilon_G$

$\forall i \quad ||\mathbf{F_{i\bullet}}||_1 \leq \epsilon_F$

where ϵ_G and ϵ_F define upperbounds for the sparsity constraints (amount of allowable density).



Diversity Constraint

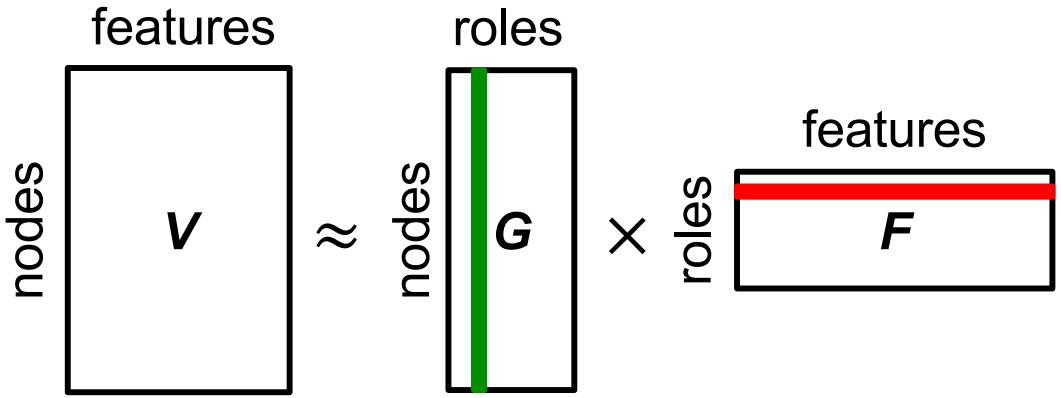
$$\underset{\mathbf{G},\mathbf{F}}{\text{argmin}} \quad ||\mathbf{V} - \mathbf{GF}||_2$$

subject to: $\mathbf{G} \geq 0, \mathbf{F} \geq 0$

$$\forall i,j \quad \mathbf{G}_{\bullet i}^T \mathbf{G}_{\bullet j} \leq \epsilon_G \quad i \neq j$$

$$\forall i,j \quad \mathbf{F}_{i\bullet} . \mathbf{F}_{j\bullet}^T \leq \epsilon_F \quad i \neq j$$

where ϵ_G and ϵ_F define upperbounds on how angularly similar role assignments and role definitions can be to each other.



Alternative Constraint

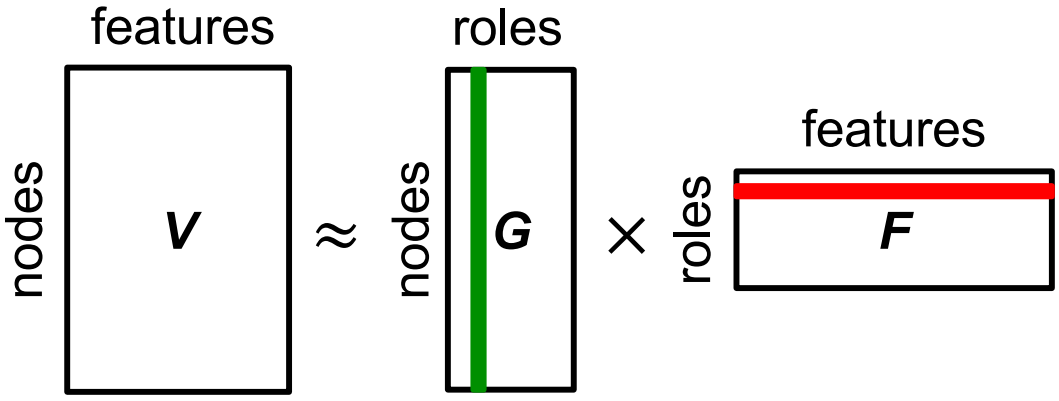
$$\underset{\mathbf{G},\mathbf{F}}{\text{argmin}} \quad ||\mathbf{V} - \mathbf{GF}||_2$$

subject to: $\mathbf{G} \geq 0, \mathbf{F} \geq 0$

$$\forall i, j \quad \mathbf{G}^{*T}_{\bullet i} \mathbf{G}_{\bullet j} \leq \epsilon_G$$

$$\forall i, j \quad \mathbf{F}^{*}_{i\bullet} \mathbf{F}^{T}_{j\bullet} \leq \epsilon_F$$

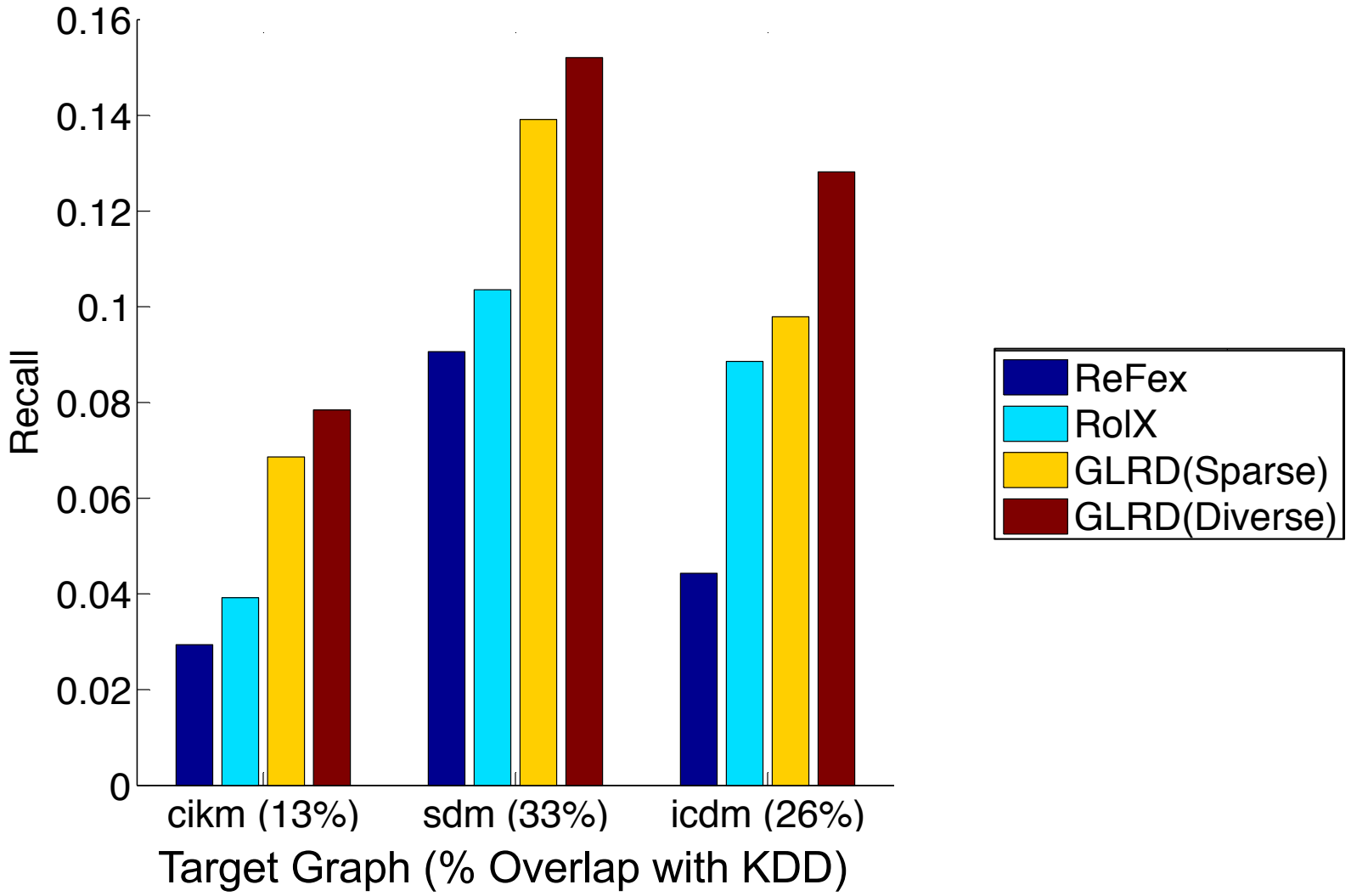
where ϵ_G and ϵ_F define upperbounds on how similar the results can be to $\mathbf{G}^*$ and $\mathbf{F}^*$.



Diverse Roles and Sparse Roles

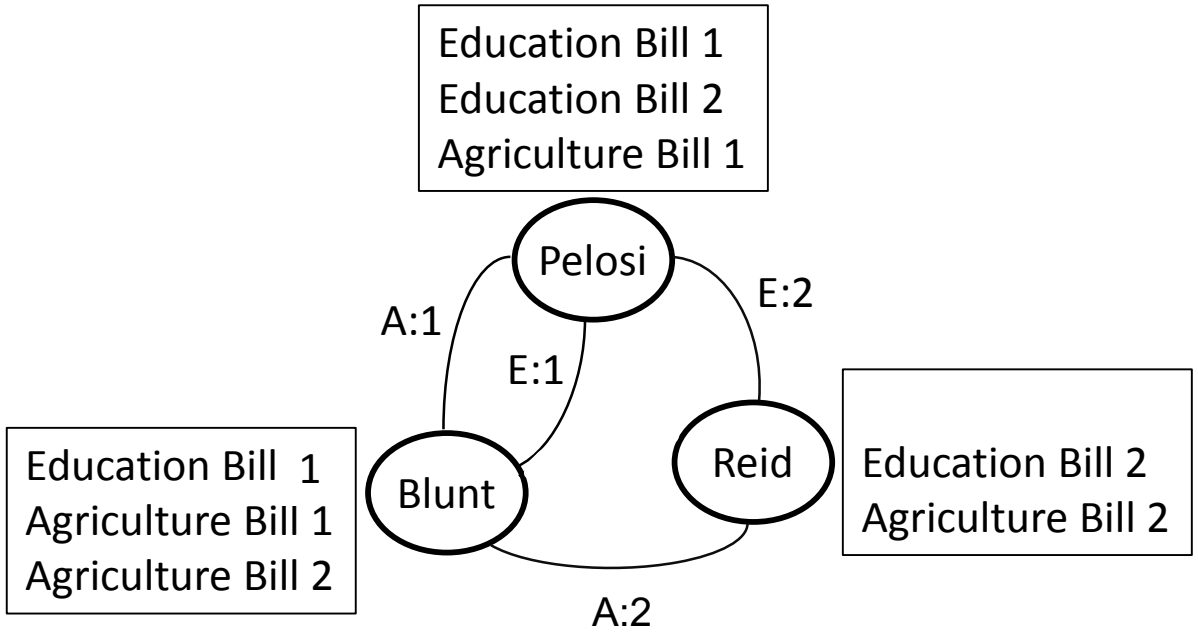
- Question: Can diversity and sparsity constraints create better role definitions?
- Conjecture: Better role definitions will better facilitate other problems such as node re-identification across graphs
- Experiment: Compare results from various methods for role discovery

GLRD on the DBLP Node Re-ID Task



Multi-relational Role Discovery (MRD)

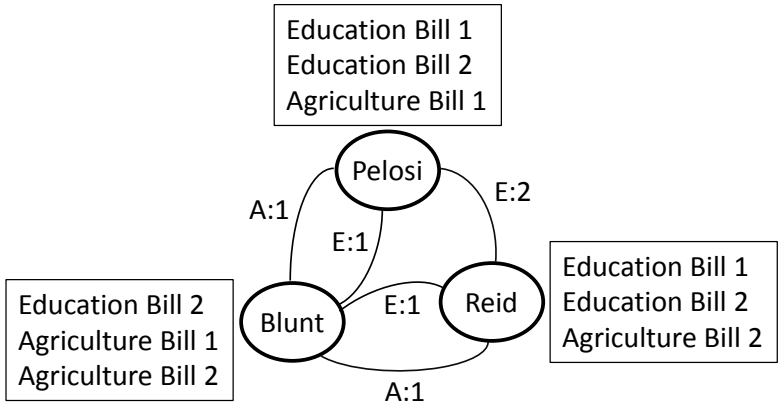
- Moving beyond simple networks
- Suppose you have a multi-relational networks
- Example: Congressional co-sponsorship data

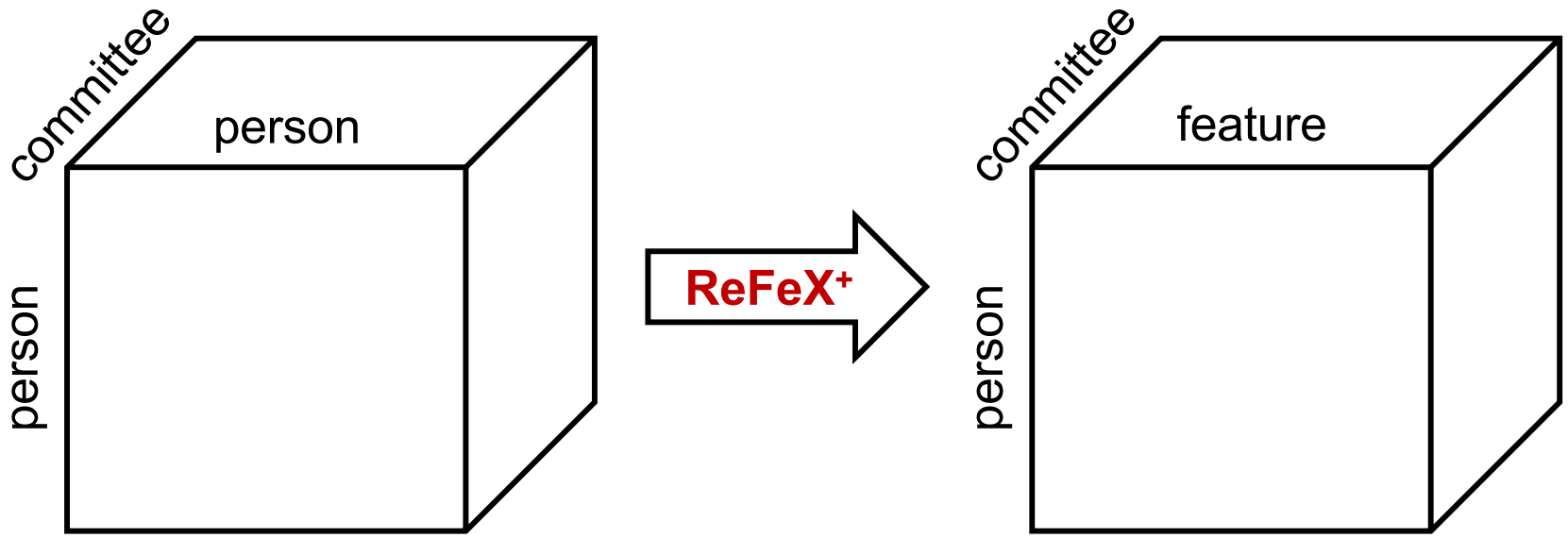


[Gilpin et al., ArXiv 2016]

No longer have an adjacency matrix

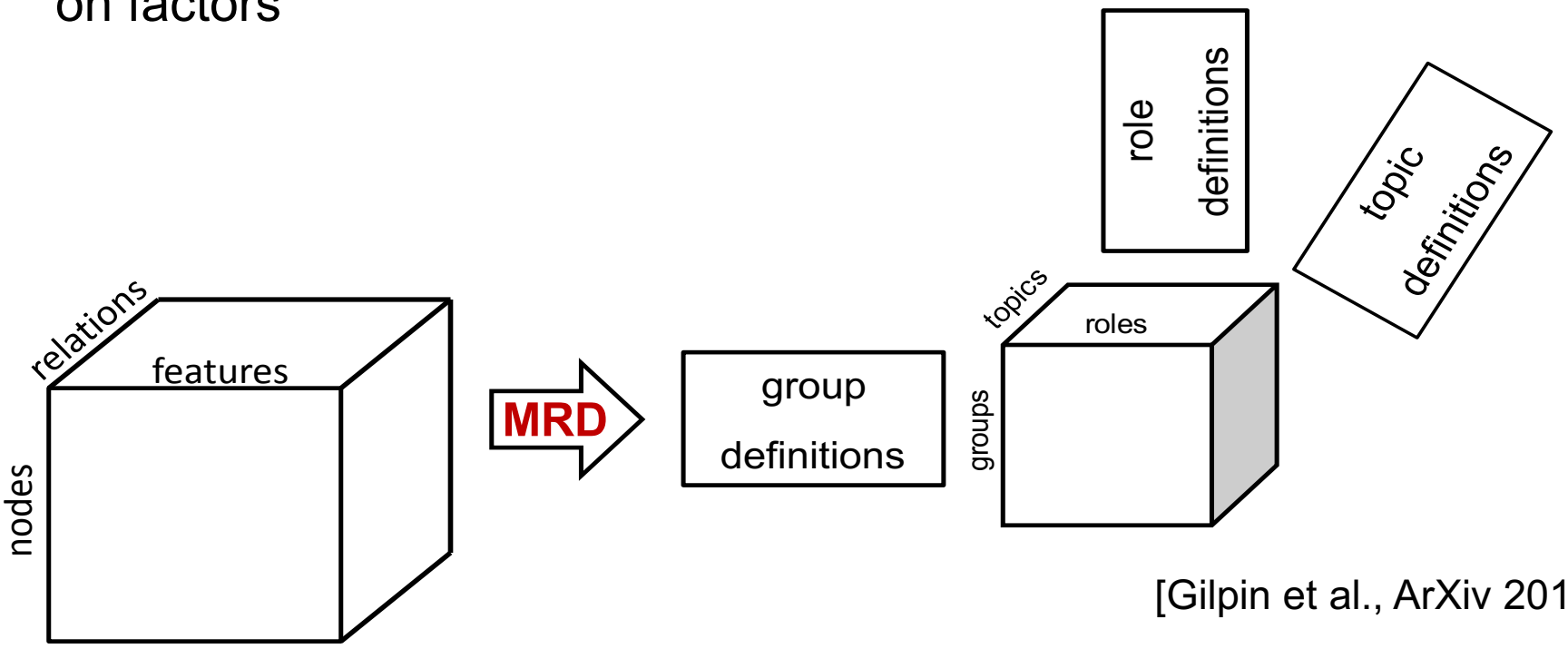
- We have a `person` × `person` × `committee` tensor
- Entry at (*i* , *j*, *k*) indicates how often congress-person *i* and *j* co-sponsored a bill that was sent to committee *k* for a particular congressional committee





Multi-relational Role Discovery (MRD)

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



[Gilpin et al., ArXiv 2016]

Multi-relational Role Discovery (MRD)

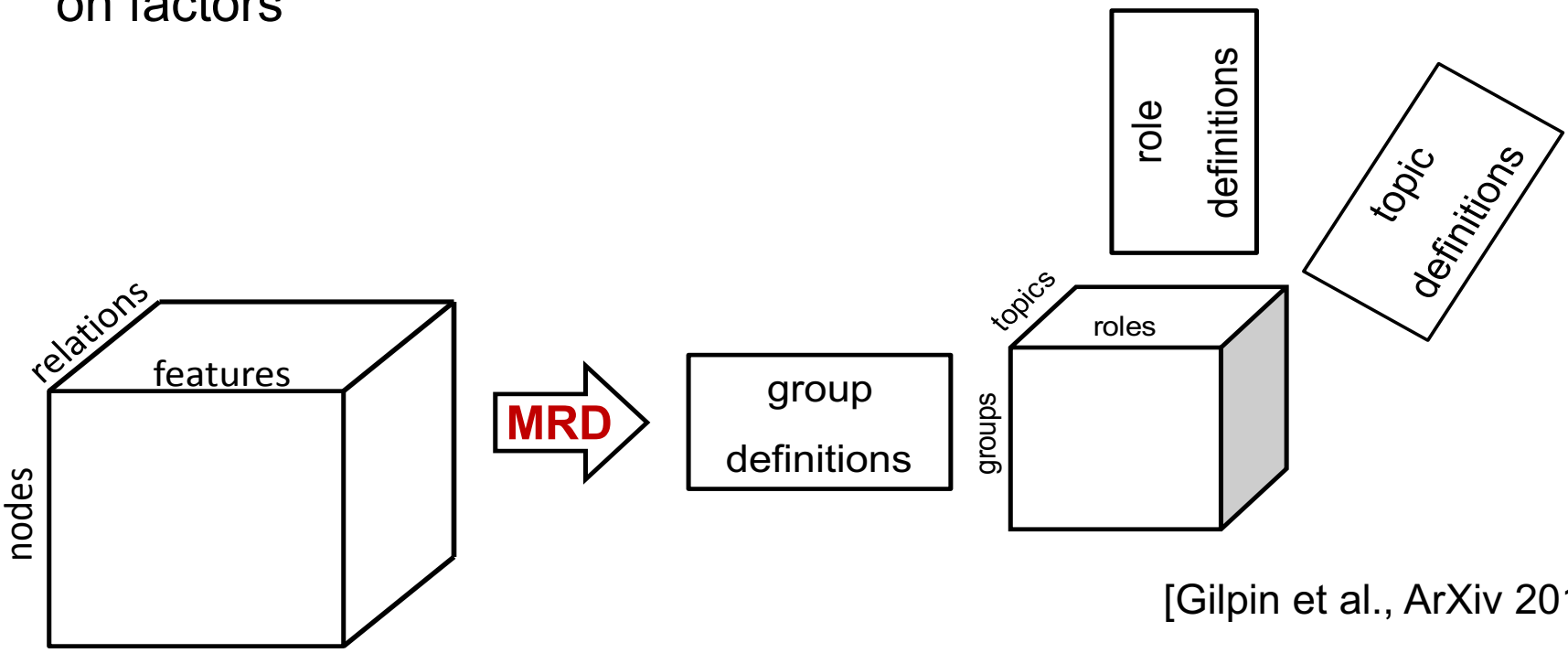
- *Multi-relational Role Discovery* (MRD)
  - Nonnegative Tucker decomposition
  - Alternating least squares
  - No orthogonality constraint on factors

$$\underset{\mathbf{G},\mathbf{F},\mathbf{R},\mathcal{H}}{\text{argmin}} \quad ||\mathcal{V} - \sum_i \sum_j \sum_k h_{ijk} * \mathbf{g_k} \circ \mathbf{f_k} \circ \mathbf{r_k}||_{Fro}$$

$$\text{subject to:} \quad \mathbf{G} \geq \mathbf{0}, \mathbf{F} \geq \mathbf{0}, \mathbf{R} \geq \mathbf{0}, \mathcal{H} \geq \mathbf{0}$$

$$g_i(\mathcal{H}) \leq d_{\mathcal{H}_i}, i = 1 \ldots t_{\mathcal{H}}$$

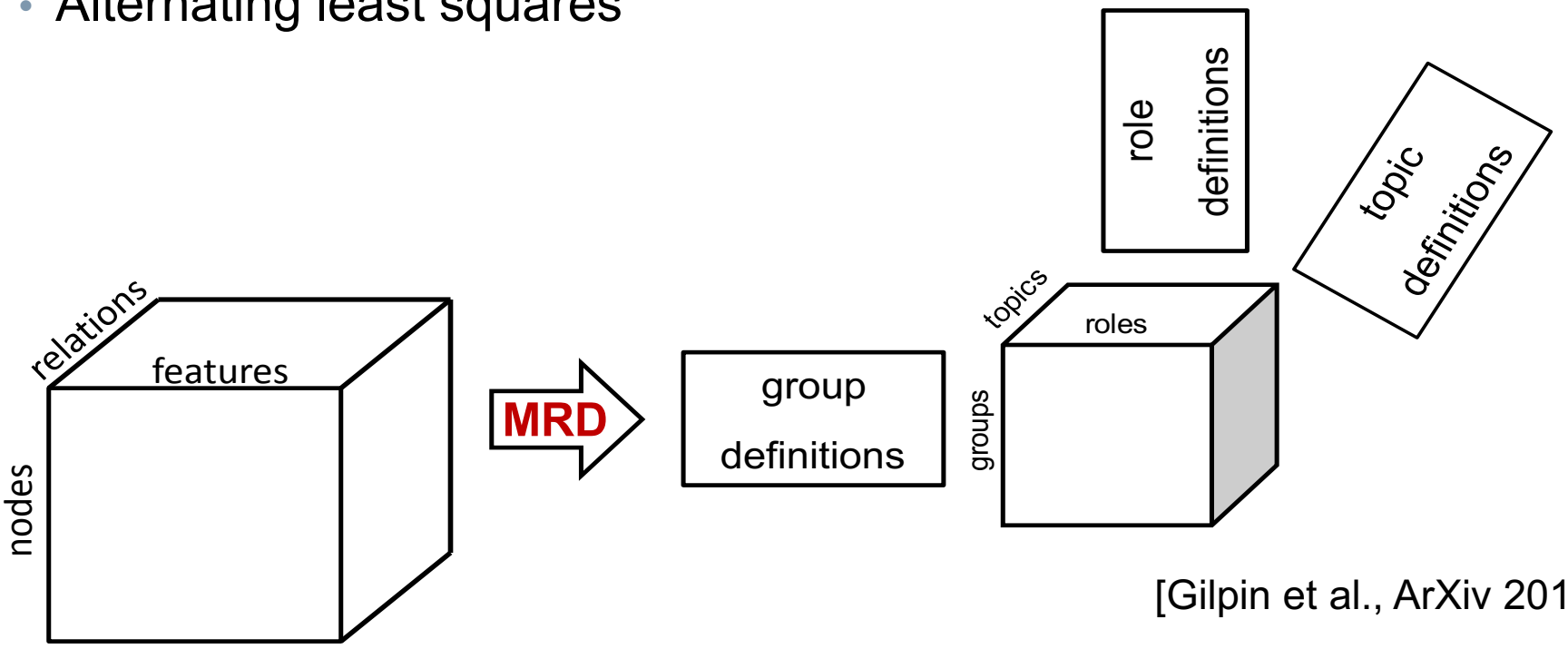
$$where\ g_i\ is\ a\ convex\ function$$



[Gilpin et al., ArXiv 2016]

Multi-relational Role Discovery (MRD)

- *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



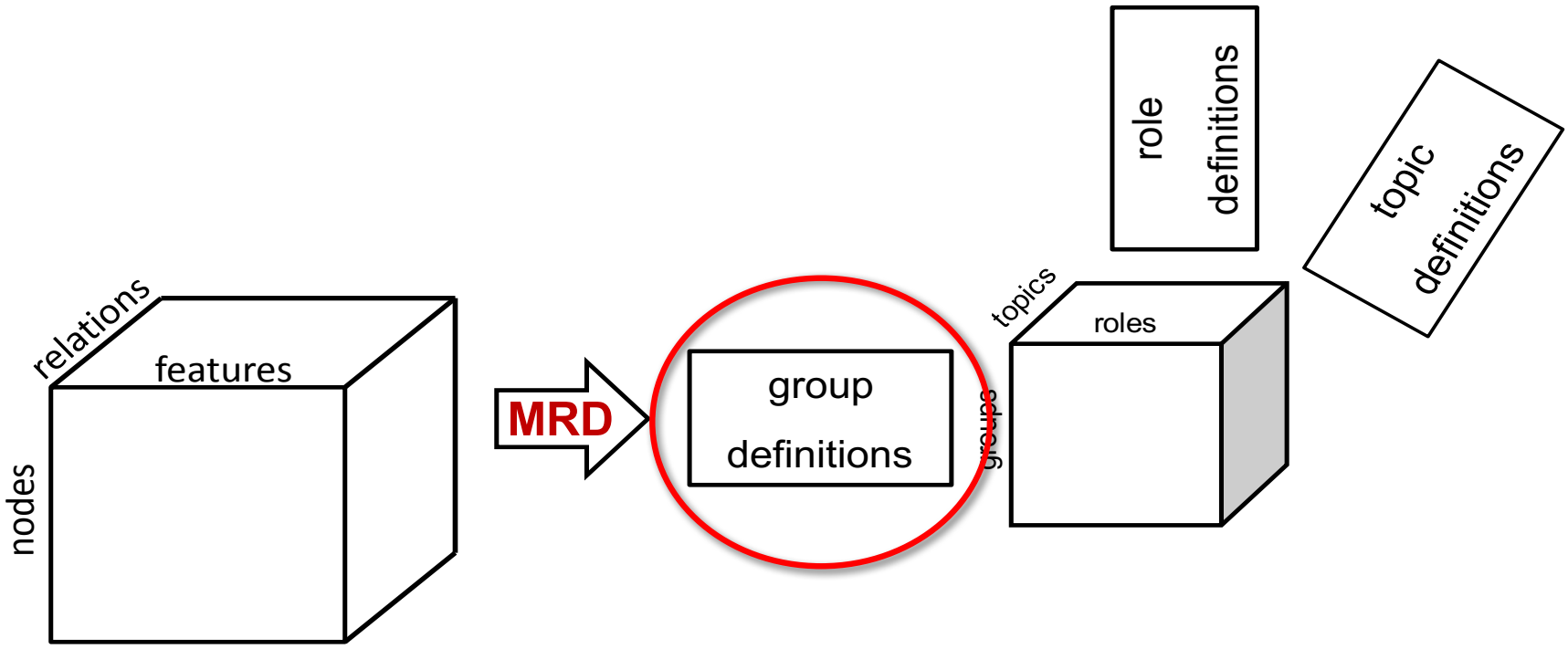
[Gilpin et al., ArXiv 2016]

Experiments

- Data from U.S. House of Representatives
- Bill co-sponsorship data from 1979 (the start of the 96th Congress) to 2009 (the end of the 110th Congress)
- 15 committees, for which there were legislation in each congress from 96th to 110th
- 110th Congress (from 2007-09)
  - 453 representatives & 10,613 bills

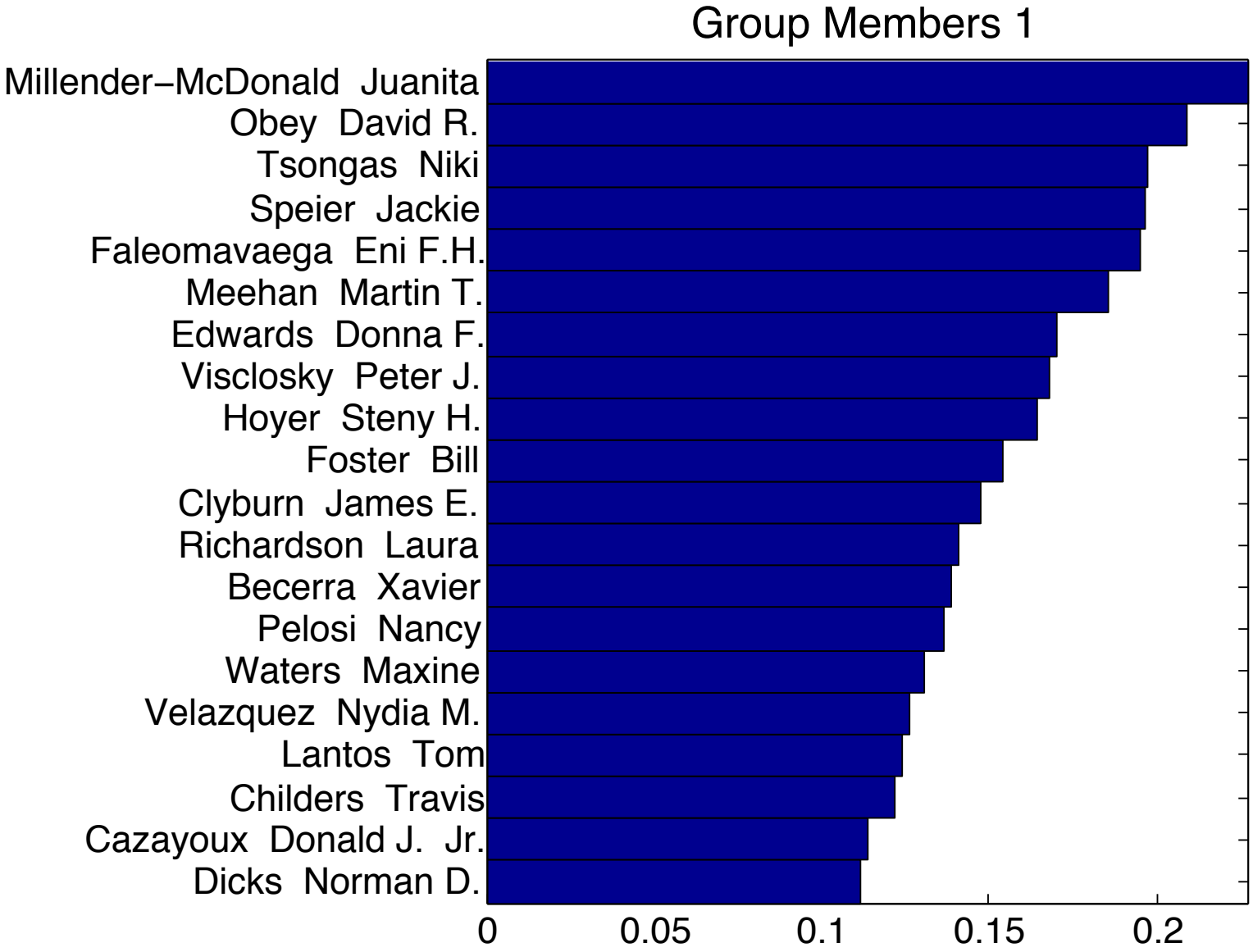
Sci & Tech
Judiciary
Ways & Means
VA
Small Business
Budget
Oversight & Gov't Reform
Agriculture
Appropriations
Rules
Natural Resources
Financial Services
Education & Labor
Transportation & Infrastructure
Energy & Commerce

Multi-relational Role Discovery (MRD)



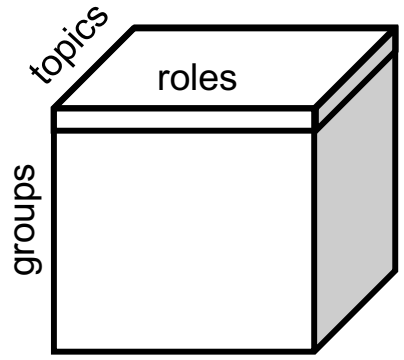
[Gilpin et al., ArXiv 2016]

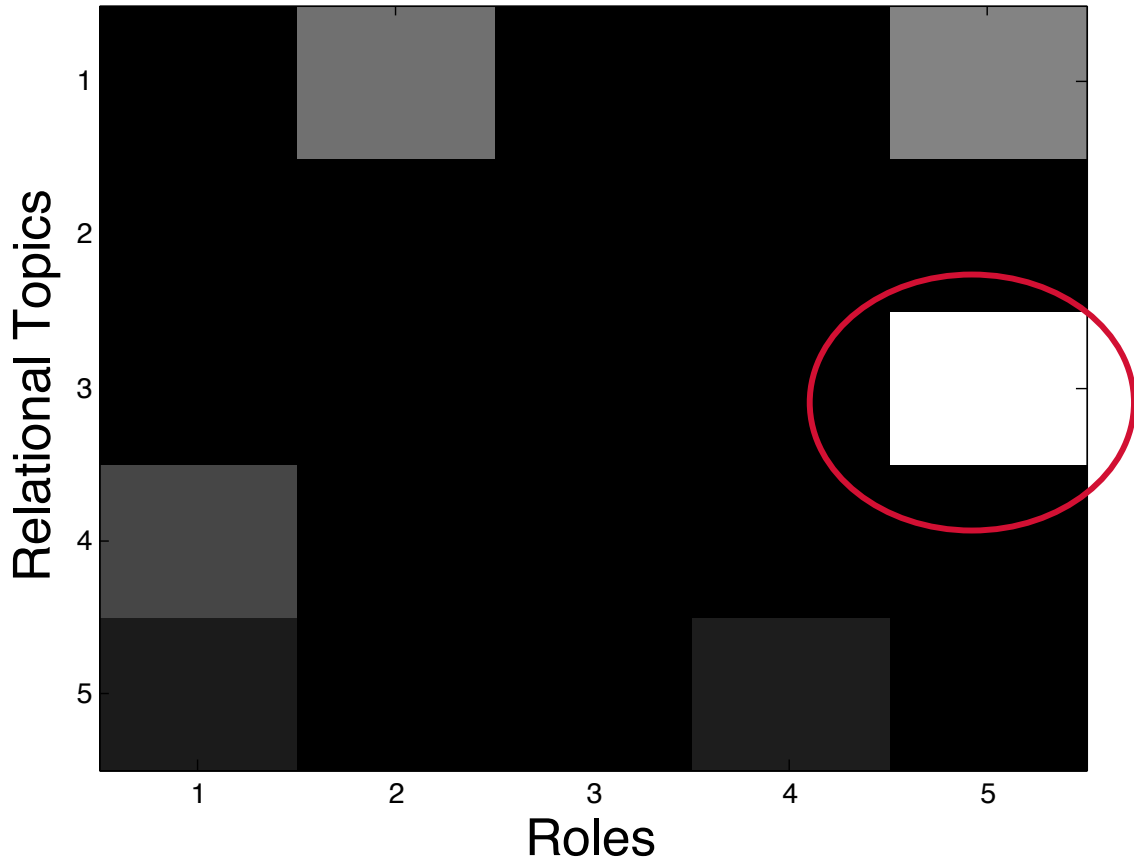
Groups of representatives



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

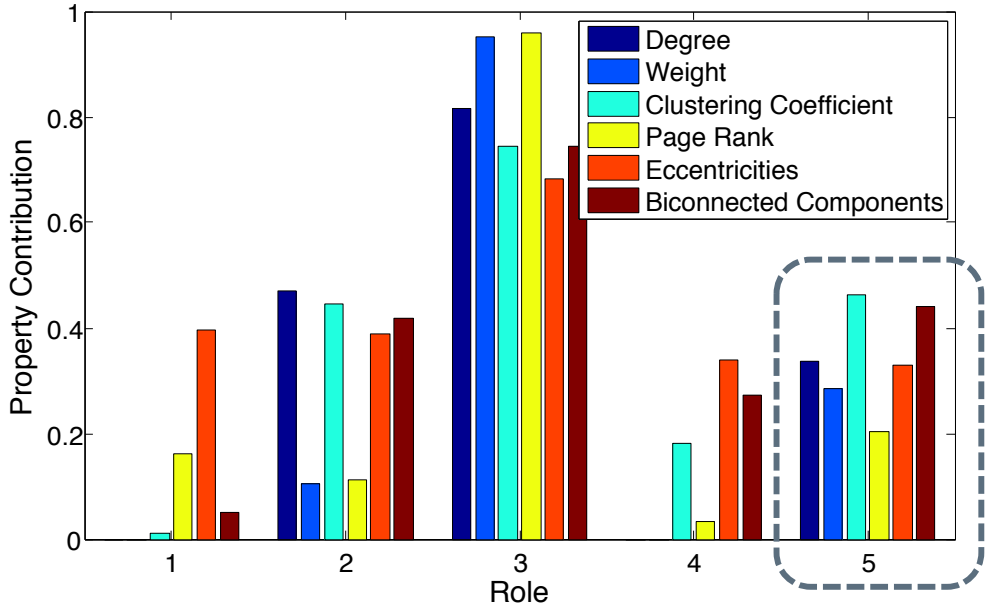


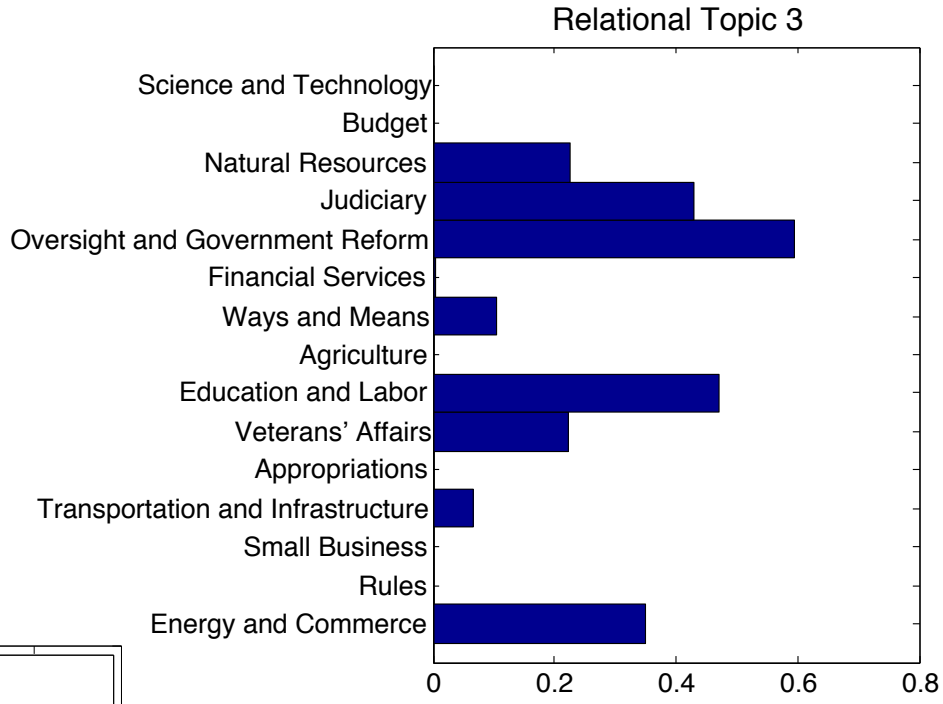


More insights into Group 1

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



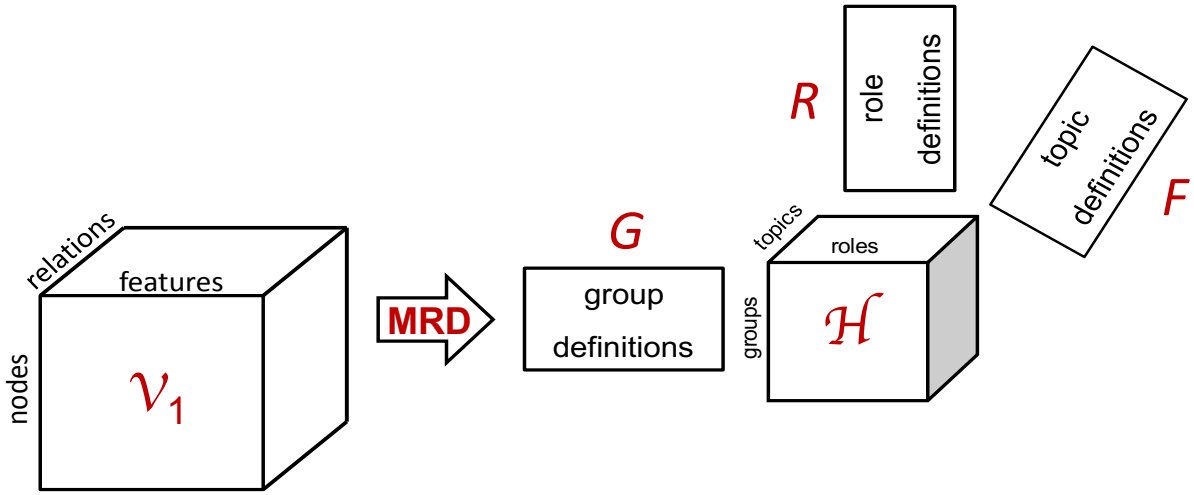


Group 1

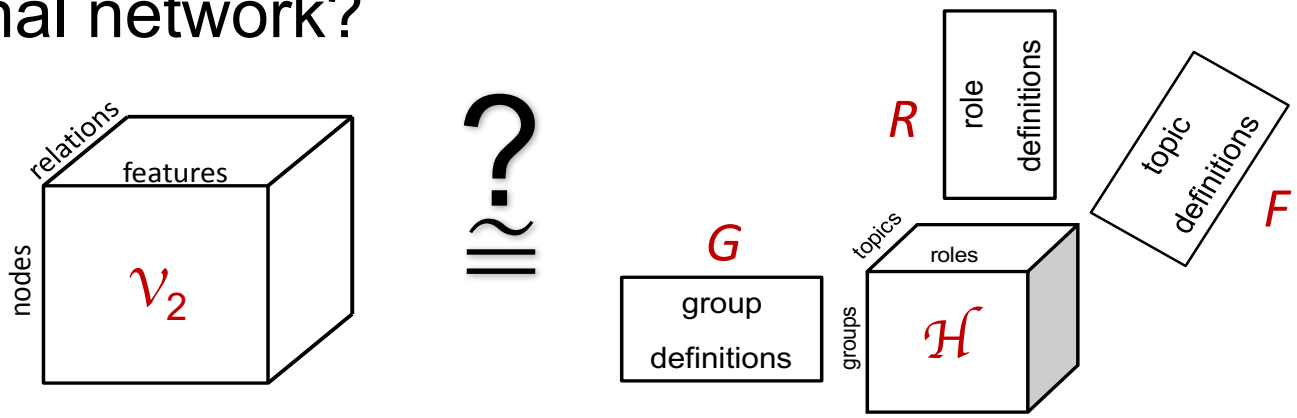
- Democrats; mostly not mid-career
- Active in oversight & gov't reform
- On the periphery, but lots of triangles

Role Transfer in MRD

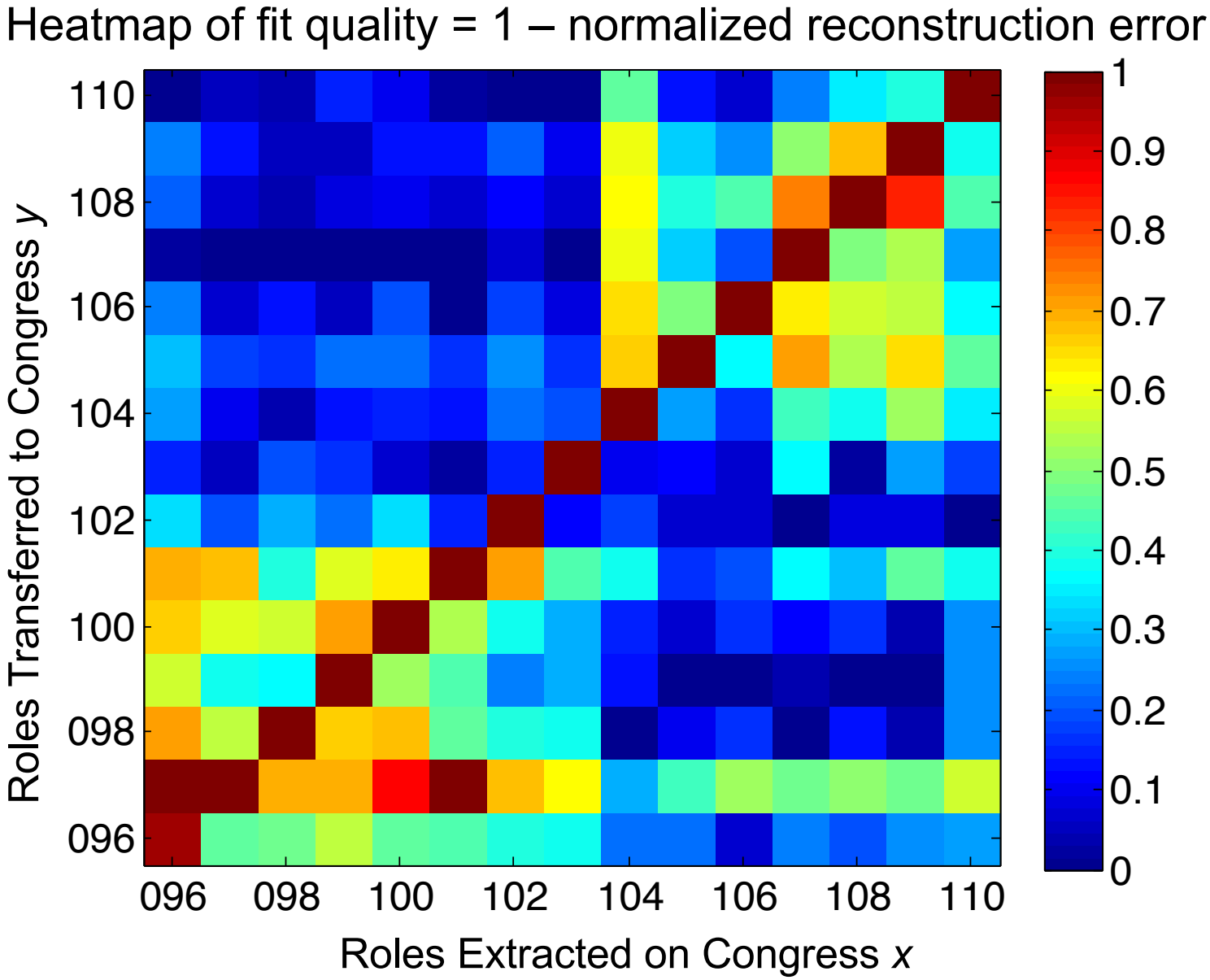
- Extract roles on one multi-relational network



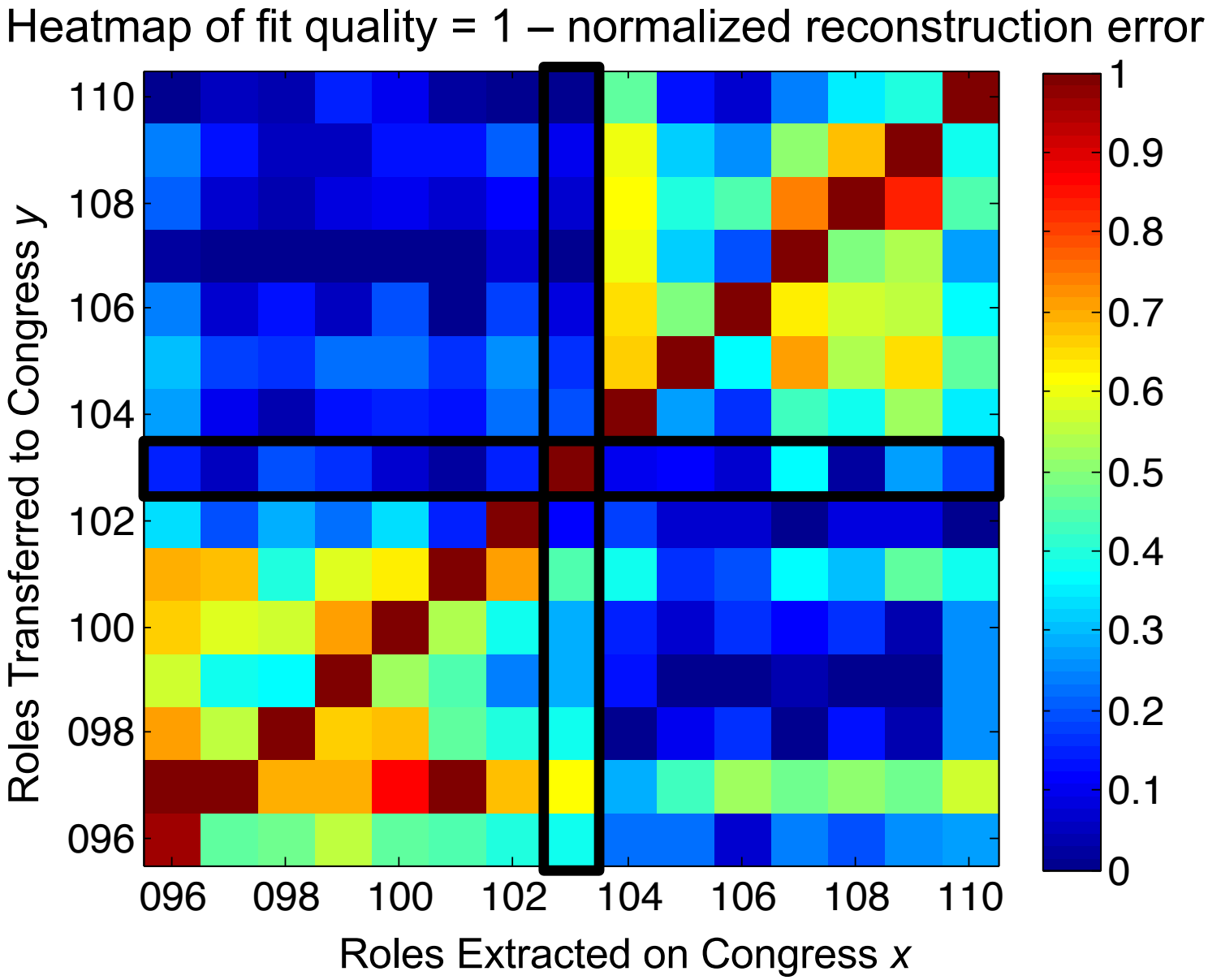
- How well do the extracted roles transfer to another multi-relational network?



Role Transfer on Multi-relational Networks

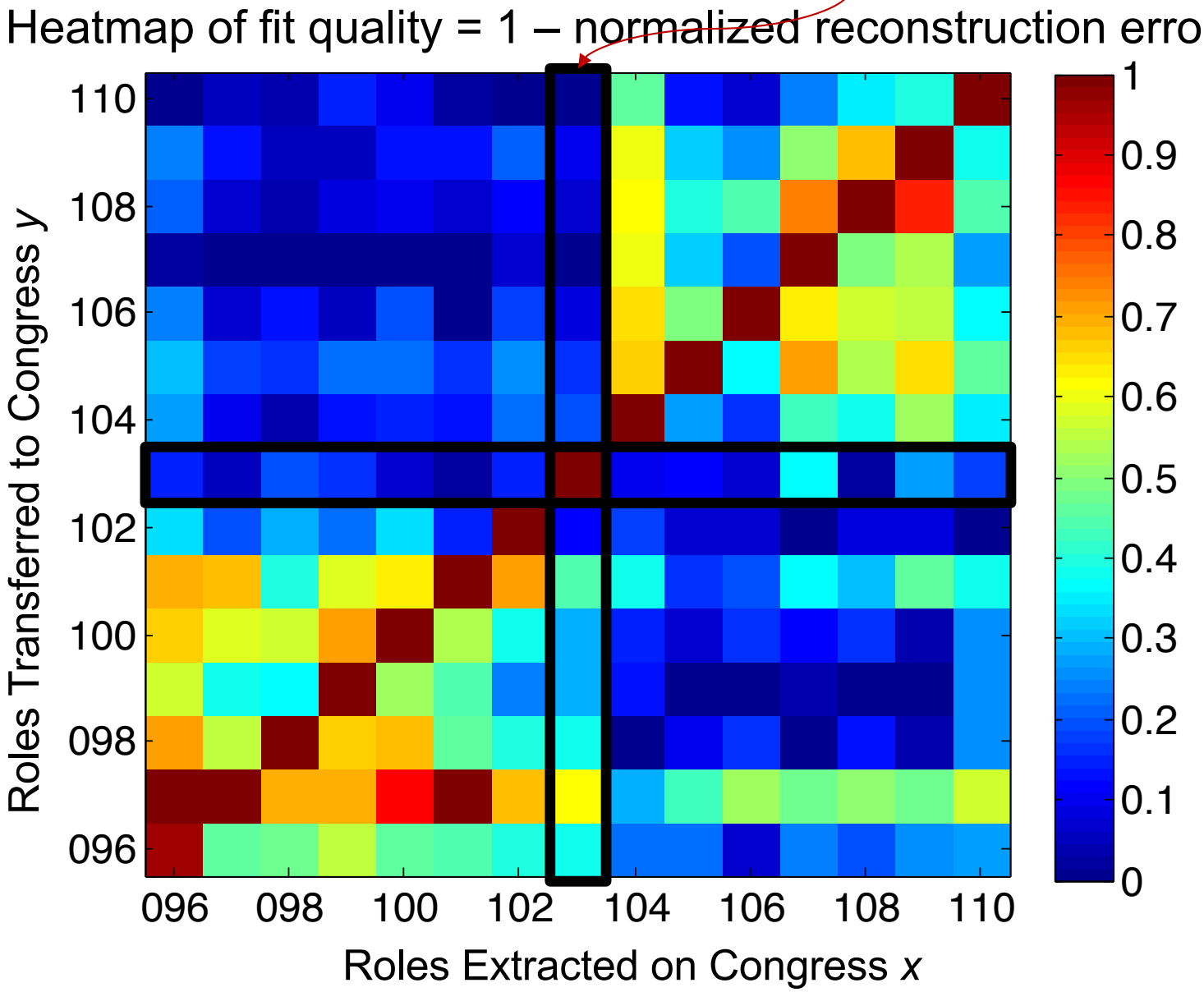


Role Transfer on Multi-relational Networks



Role Transfer

Hastert Rule: the Speaker will not allow a floor vote on a bill unless a majority of the majority party supports the bill.



Why are Roles Effective in Many Applications?

- Encode complex behavior
- Map nodes into a useful lower dimensional space
- Generalize across networks

Papers, Tutorials, Code

- Papers at http://eliassi.org/pubs.html
- Tutorials at http://eliassi.org
- Open-source code at https://snap.stanford.edu/snap-2.3/
- Role discovery is joint work with
  - LLNL (Keith Henderson & Brian Gallagher)
  - CMU (Christos Faloutsos, Leman Akoglu et al.)
  - Google (Sugato Basu)
  - UC Davis (Ian Davidson et al.)

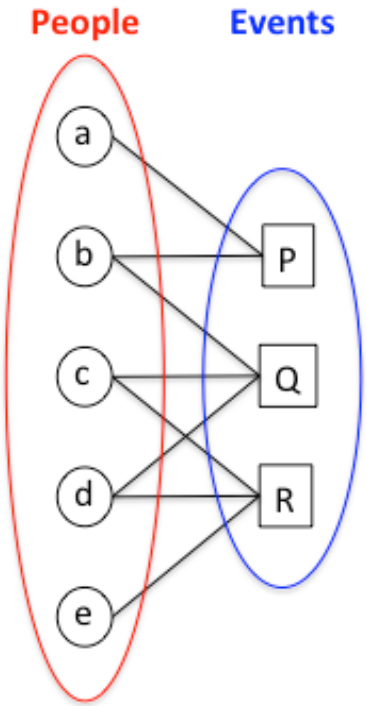
Roadmap

1. The reasonable effectiveness of roles in networks
2. A theoretical guide to tie-strength measures

Problem Definition

- Given a bipartite graph with people as one set of vertices and events as the other set, measure *tie strength* between each pair of individuals

- Assumption
  - Attendance at mutual events implies an implicit weighted social network between people



Motivation

- Most real-world networks are bipartite and are converted to unipartite (e.g., via AA^T)
- Explicitly declared friendship links can suffer from a low signal-to-noise ratio (e.g., Facebook friends)

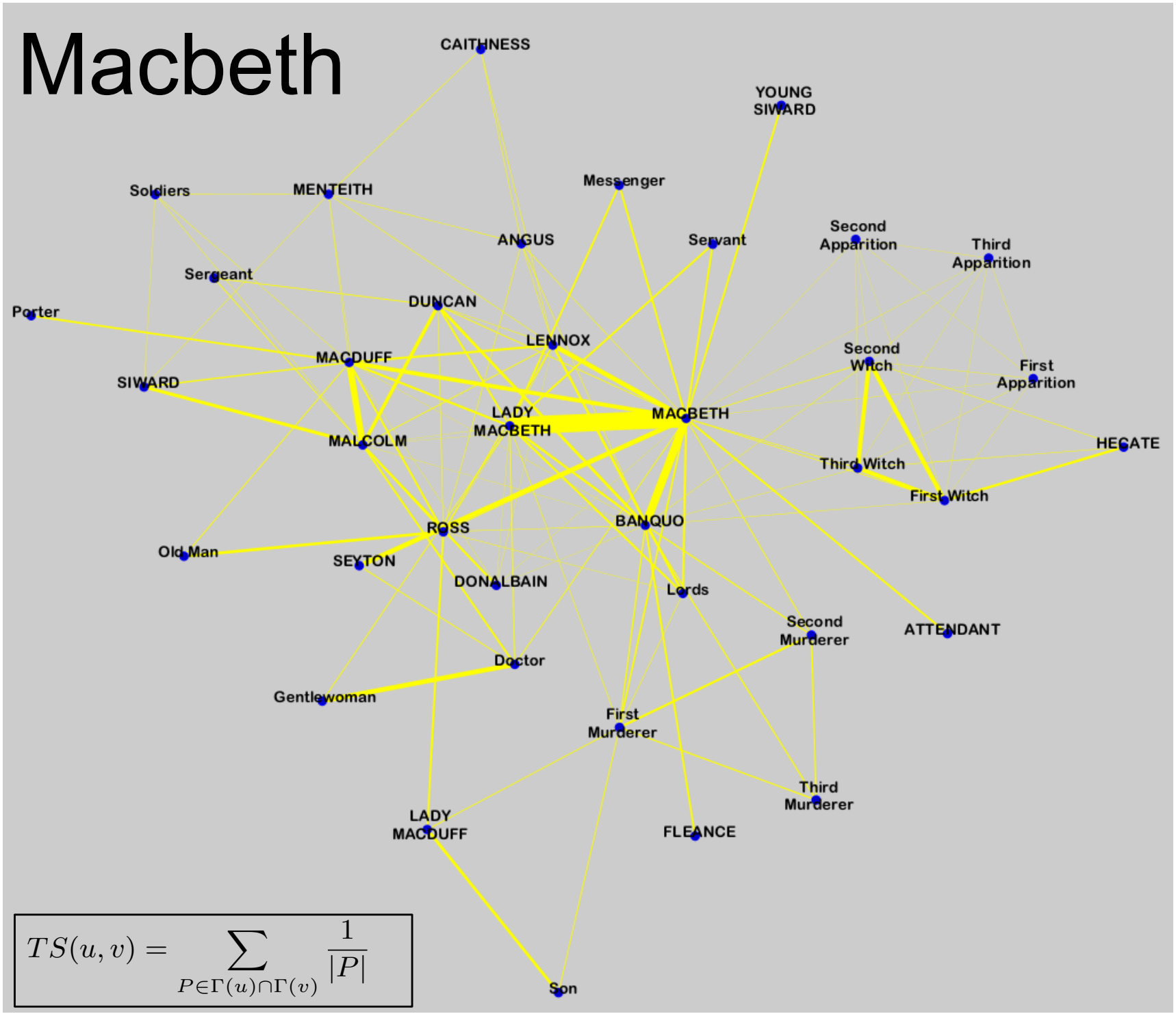
- **Challenge:** Detect which of links in the unipartite graph are important
- **Goal:** Infer the implicit weighted social network from people's participation in mutual events

Tie Strength

- A measure of tie strength induces
  - a ranking on all the edges, and
  - a ranking on the set of neighbors for every person

- Example of a simple tie-strength measure
  - Common neighbor measures the total number of common events to a pair of individuals

Macbeth



$$TS(u,v) = \sum_{P \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|P|}$$

Decisions, Decisions

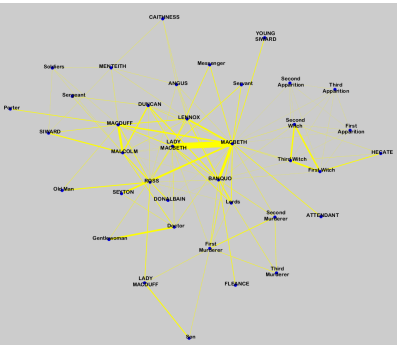
- There are many different measures of tie-strength
  1. Common neighbor
  2. Jaccard index
  3. Max
  4. Linear
  5. Delta
  6. Adamic and Adar
  7. Preferential attachment
  8. Katz measure
  9. Random walk with restarts
  10. Simrank
  11. Proportional
  12. …

Which one should you choose?

Outline for Tie-Strength

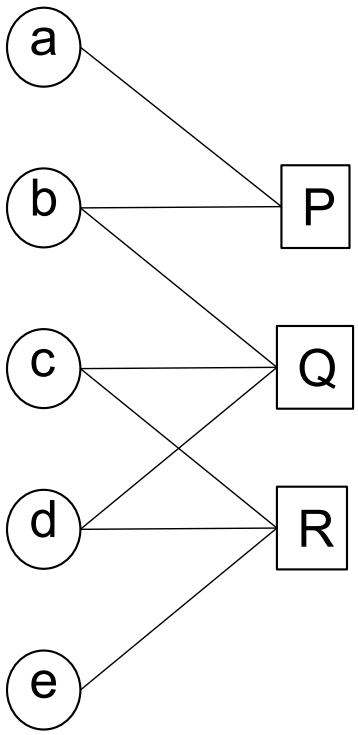
- An axiomatic approach to the problem of inferring implicit social networks by measuring tie strength
- A characterization of functions that satisfy all our axioms
- Classification of prior measures according to the axioms that they satisfy
- Experiments

Running Example



Input

People × Event Bipartite Graph



infer

Output

Partial Order of Tie Strength among People

(a,b) (c,d) *high*

(b,c), (b,d), (c,e), (d,e)

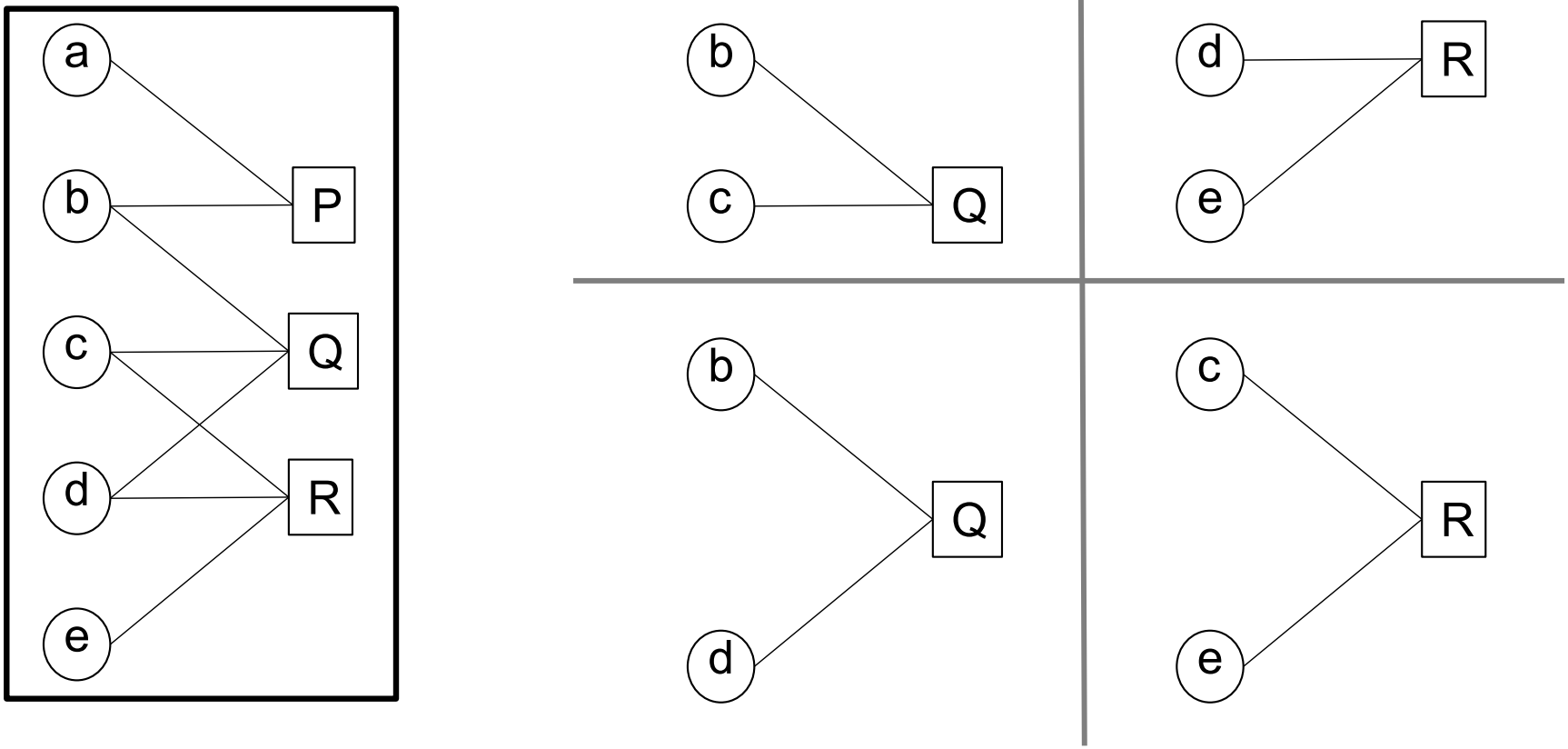
(a,c), (a,d), (a,e), (b,e) *low*

Axioms

- Axiom 1: Isomorphism
- Axiom 2: Baseline
- Axiom 3: Frequency
- Axiom 4: Intimacy
- Axiom 5: Popularity
- Axiom 6: Conditional Independence of People
- Axiom 7: Conditional Independence of Events
- Axiom 8: Submodularity

Axiom 1: Isomorphism

- Tie strength between *u* and *v* is independent of the labels of *u* and *v*



Axiom 2: Baseline

- If there are no events, then tie strength between each pair *u* and *v* is 0

$$TS_{\varnothing}(u, v) = 0$$

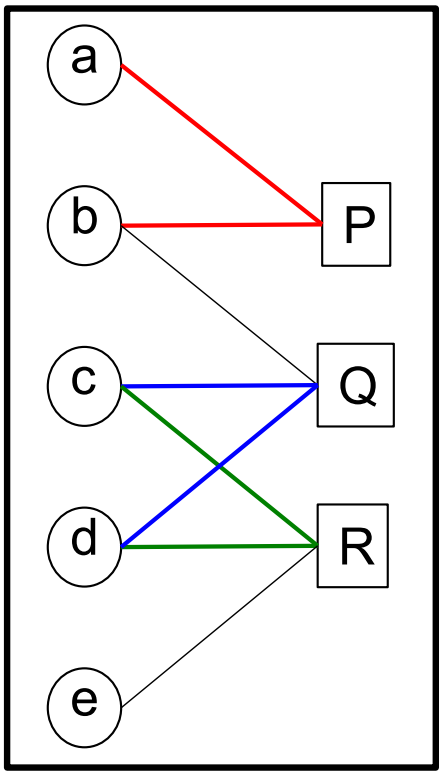
- If there are only two people *u* and *v* and a single event *P* that they attend, then their tie strength is at most 1

$$TS_{P}(u, v) \leq 1$$

  - Defines an upper-bound for how much tie strength can be generated from a single event between two people

Axiom 3: Frequency & Axiom 4: Intimacy

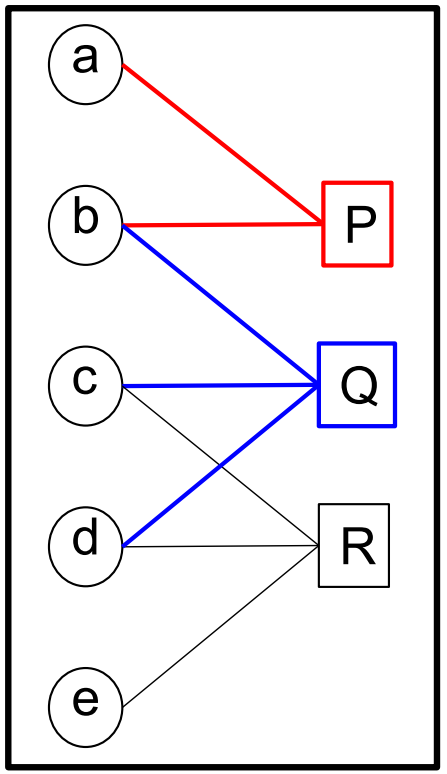
- Axiom 3 (**Frequency**)
  - More events create stronger ties
  - All other things being equal, the more events common to *u* and *v*, the stronger their tie-strength



- Axiom 4 (**Intimacy**)
  - Smaller events create stronger ties
  - All other things being equal, the fewer invitees there are to any particular event attended by *u* and *v*, the stronger their tie-strength

Axiom 5: Popularity

- Larger events create more ties
- Consider two events *P* and *Q*
- If |*Q*| > |*P*|, then the total tie strength created by *Q* is more than that created by *P*



Axioms 6 & 7: Conditional Independence of People and of Events

- Axiom 6: **Conditional Independence of People**
  - A node *u*'s tie strength to other people does **not** depend on events that *u* does **not** attend
- Axiom 7: **Conditional Independence of Events**
  - The increase in tie strength between *u* and *v* due to an event *P* does **not** depend on other events, just on the existing tie strength between *u* and *v*
  - $TS_{(G+P)}(u, v) = g(TS_G(u, v), TS_P(u, v))$
    - where *g* is some monotonically increasing function

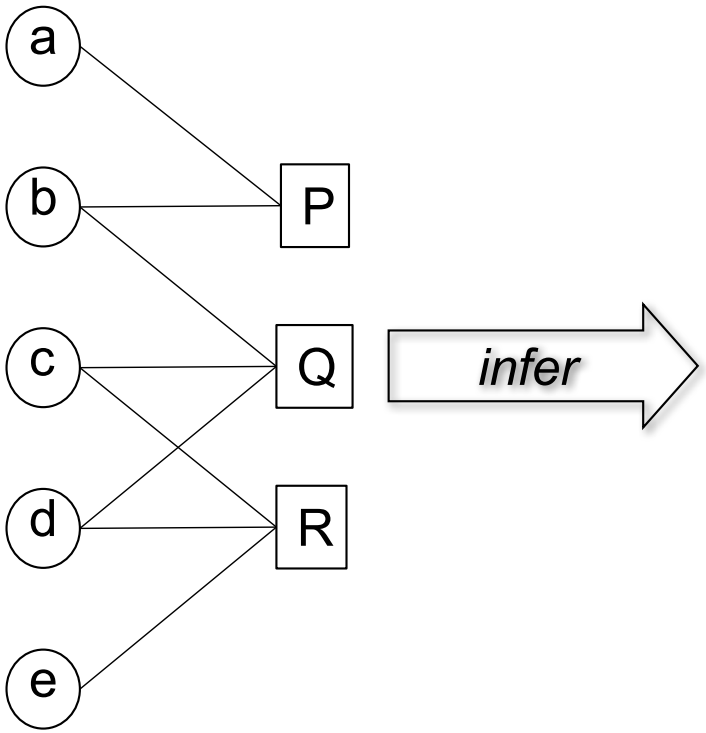
Axiom 8: Submodularity

- The marginal increase in tie strength of *u* and *v* due to an event *Q* is at most the tie strength between *u* and *v* if *Q* was their only event
- If *G* is a graph and *Q* is a single event, then $TS_{(G+Q)}(u, v) - TS_G(u, v) \leq TS_Q(u, v)$

Example – Mapping to Axioms

Input
People × Event Bipartite Graph

Output
Partial order of Tie Strength



infer

(a,b) (c,d) *high*

(b,c), (b,d), (c,e), (d,e)

(a,c), (a,d), (a,e), (b,e) *low*

Axiom 4 (Intimacy) & Axiom 3 (Freq)

Axiom 1 (Isomorphism)

Axiom 2 (Baseline) & Axiom 6 (Cond. Indep. of Vertices) & Axiom 7 (Cond. Indep. of Events)

Observations on the Axioms

- Our axioms are fairly intuitive

A1: Isomorphism	A2: Baseline	A3: Frequency	A4: Intimacy
A5: Popularity	A6: Cond. Indep. of people	A7: Cond. indep. of events	A8: Submodularity

- But, several previous measures in the literature break some of these axioms
- Satisfying all the axioms is not sufficient to uniquely identify a measure of tie strength
  - One reason: inherent tension between Axiom 3 (Frequency) and Axiom 4 (Intimacy)

Inherent Tension Between Frequency & Intimacy

- Scenario #1 (intimate)
  - Mary and Susan go to 2 parties, where they are the only people there.
- Scenario #2 (frequent)
  - Mary, Susan, and Jane go to 3 parties, where they are the only people there.
- In which scenario is Mary's tie to Susan stronger?

Observations on the Axioms (cont.)

A1: Isomorphism	A2: Baseline	A3: Frequency	A4: Intimacy
A5: Popularity	A6: Cond. Indep. of people	A7: Cond. indep. of events	A8: Submodularity

- Axioms are equivalent to a natural partial order on the strength of ties
  - Pertinent to ranking application
- Choosing a particular tie-strength function is equivalent to choosing a particular linear extension of this partial order
  - Non-obvious decision
  - Details in WebSci 2012 paper

Characterizing Tie Strength

A way to explore the space of valid functions for representing tie strength and find which work given particular applications

Theorem. Given a graph $G = (L \cup R, E)$ *and two vertices u and v*, *if* the tie-strength function *TS* follows Axioms (1-8), then the function has to be of the form

$$TS_G(u, v) = g(h(|P_1|), h(|P_2|), \ldots, h(|P_k|))$$

- $\{P_i\}_{1 \le i \le k}$ are the events common to both *u* and *v*
- h is a monotonically decreasing function bounded by $1 \geq h(n) \geq \frac{1}{\binom{n}{2}}, n \geq 2; h(1) = 1; h(0) = 0.$
- g is a monotonically increasing submodular function

Many Measures of Tie Strength

1. Common neighbor
2. Jaccard index
3. Max
4. Linear
5. Delta
6. Adamic and Adar
7. Preferential attachment
8. Katz measure
9. Random walk with restarts
10. Simrank
11. Proportional

$$TS(u,v) = |\Gamma(u) \cap \Gamma(v)|$$

$$TS(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

$$TS(u,v) = \max_{P \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|P|}$$

$$TS(u,v) = \sum_{P \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|P|}$$

$$TS(u,v) = \sum_{P \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\binom{|P|}{2}}$$

$$TS(u,v) = \sum_{P \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |P|}$$

$$TS(u,v) = |\Gamma(u)| \cdot |\Gamma(v)|$$

$$TS(u,v) = \sum_{q \in \text{ path between } u,v} \gamma^{-|q|}$$

$$TS(u,v) = \begin{cases} 1 & \text{if u = v} \\ \gamma \cdot \frac{\sum_{a \in \Gamma(u)} \sum_{b \in \Gamma(v)} TS(a,b)}{|\Gamma(u)| \cdot |\Gamma(v)|} & \text{otherwise} \end{cases}$$

$$TS(u,v) = \sum_{P \in \Gamma(u) \cap \Gamma(v)} \frac{\epsilon}{|P|} + (1-\epsilon) \frac{TS(u,v)}{\sum_{w \in \Gamma(u)} TS(u,w)}$$

Tie Strength Measures Fall into Two Groups

Non-Self Referential

- **Common neighbor**
  - The total # of common events that both u and v attended
- **Jaccard Index**
  - Similar to common neighbor
  - Normalizes for how "social" *u* and *v* are
- **Adamic and Adar [2003], Delta, and Linear**
  - Tie strength increases with the number of events
  - Tie strength is 1 over a simple function of event size
- **Max**
  - Tie strength does not increase with the number of events

Self Referential

- **Katz measure [1953]**
  - Tie strength is the number of paths between u and v, where each path is discounted exponentially by the length of the path
- **Random walk with restarts**
  - A non-symmetric measure of tie strength
  - Tie strength is the stationary probability of a Markov chain process
- **Simrank [Jeh & Widom, 2002]**
  - Tie strength is captured by recursively computing the tie strength of neighbors
- **Proportional**
  - Tie strength increases with # of events
  - People spend time proportional to their tie-strength at a party

Measures of Tie Strength that Satisfy All the Axioms

A1: Isomorphism	A2: Baseline	A3: Frequency	A4: Intimacy
A5: Popularity	A6: Cond. indep. of P	A7: Cond. indep. of E	A8: Submodularity

	A1	A2	A3	A4	A5	A6	A7	A8	$g(a_1, \ldots, a_k)$ $h(\lvert P_i \rvert) = a_i$
Common Neighbors	✓	✓	✓	✓	✓	✓	✓	✓	$g(a_1, \ldots, a_k) = \Sigma a_i$ $h(n) = 1$
Delta	✓	✓	✓	✓	✓	✓	✓	✓	$g(a_1, \ldots, a_k) = \Sigma a_i$ $h(n) = 2(n(n-1))^{-1}$
Adamic & Adar	✓	✓	✓	✓	✓	✓	✓	✓	$g(a_1, \ldots, a_k) = \Sigma a_i$ $h(n) = (\log(n))^{-1}$
Linear	✓	✓	✓	✓	✓	✓	✓	✓	$g(a_1, \ldots, a_k) = \Sigma a_i$ $h(n) = n^{-1}$
Max	✓	✓	✓	✓	✓	✓	✓	✓	$g(a_1, \ldots, a_k) = \max\{a_i\}$ $h(n) = n^{-1}$

Measures of Tie Strength that Do **Not** Satisfy All the Axioms

A1: Isomorphism	A2: Baseline	A3: Frequency	A4: Intimacy
A5: Popularity	A6: Cond. indep. of P	A7: Cond. indep. of E	A8: Submodularity

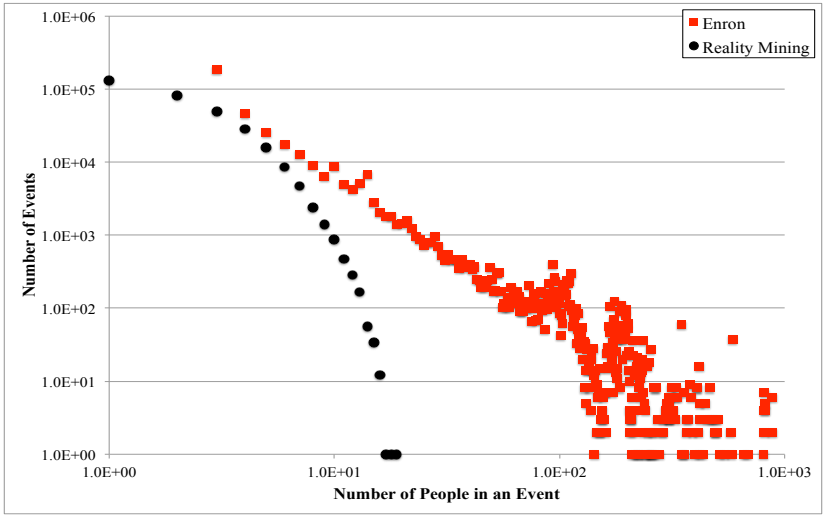
	A1	A2	A3	A4	A5	A6	A7	A8	$g(a_1, \ldots, a_k)$ $h(\lvert P_i \rvert) = a_i$
Jaccard Index	✓	✓	✓	✓	✓	✗	✗	✗	✗
Katz Measure	✓	✗	✓	✓	✓	✓	✗	✗	✗
Preferential Attachment	✓	✓	✗	✓	✓	✓	✗	✗	✗
RWR	✓	✗	✗	✗	✓	✓	✗	✗	✗
Simrank	✓	✗	✗	✗	✗	✗	✗	✗	✗
Proportional	✓	✗	✗	✓	✗	✓	✗	✗	✗

Data Sets

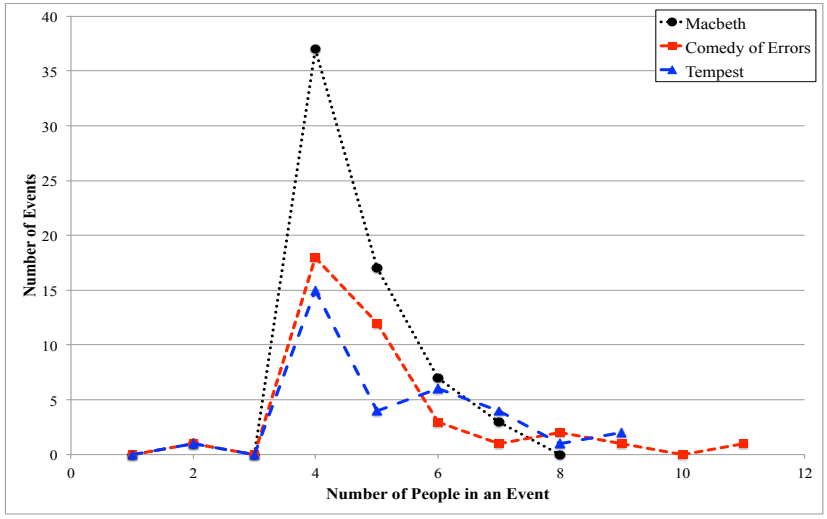
Graphs	# of People	# of Events
Southern Women	18	14
The Tempest	19	34
A Comedy of Errors	19	40
Macbeth	38	67
Reality Mining Bluetooth	104	326,248
Enron Emails	32,471	371,321

Degree Distributions

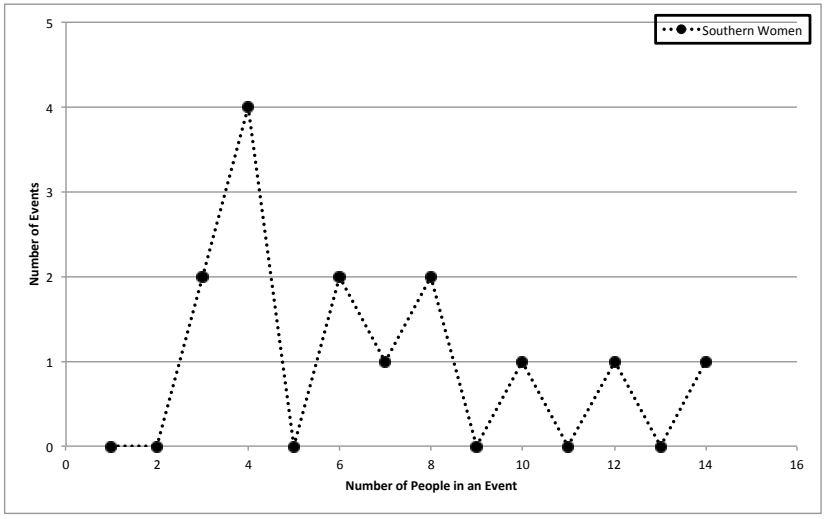
Enron & Reality Mining



Shakespeare's Plays



Southern Women



Completeness of Axioms 1-8

(Number of Ties **Not** Resolved by the Partial Order)

Dataset	Tie Pairs	Incomparable Pairs (%)
Southern Women	11,628	683 (5.87)
The Tempest	14,535	275 (1.89)
A Comedy of Errors	14,535	726 (4.99)
Macbeth	246,753	584 (0.23)
Reality Mining	13,794,378	1,764,546 (12.79)

- % of tie-pairs where different tie-strength functions can differ
  - Smaller is better
  - Generally, percentages are small
  - Large real-world networks have more unresolved ties

# of tie pairs = $\binom{\binom{n}{2}}{2}$

Take-away Point #1

% of tie pairs on which different tie-strength functions can differ is small.*

* Disclaimer: For ranking applications and tie-strength functions that satisfy the axioms

Soundness of Axioms 1-8

(Number of Conflicts Between the Partial Order and Tie-Strength Functions **Not** Satisfying the Axioms)

Dataset	Tie Pairs	Jaccard (%)	Temporal (%)
Southern Women	11,628	1,441 (12.39)	665 (5.72)
The Tempest	14,535	488 (3.35)	261 (1.79)
A Comedy of Errors	14,535	1,114 (7.76)	381 (2.62)
Macbeth	246,753	2,638 (1.06)	978 (0.39)
Reality Mining	13,794,378	290,934 (0.02)	112,546 (0.01)

- % of tie-pairs in conflict with the partial order
  - Smaller is better
  - Generally, percentages are small
  - They decrease as the dataset increases

More on Soundness

- **Question 1:**

  Are the number of conflicts, between the partial order and tie-strength functions not satisfying the axioms, small because most of the tie-strengths are zeros (sparsity of real graph)?

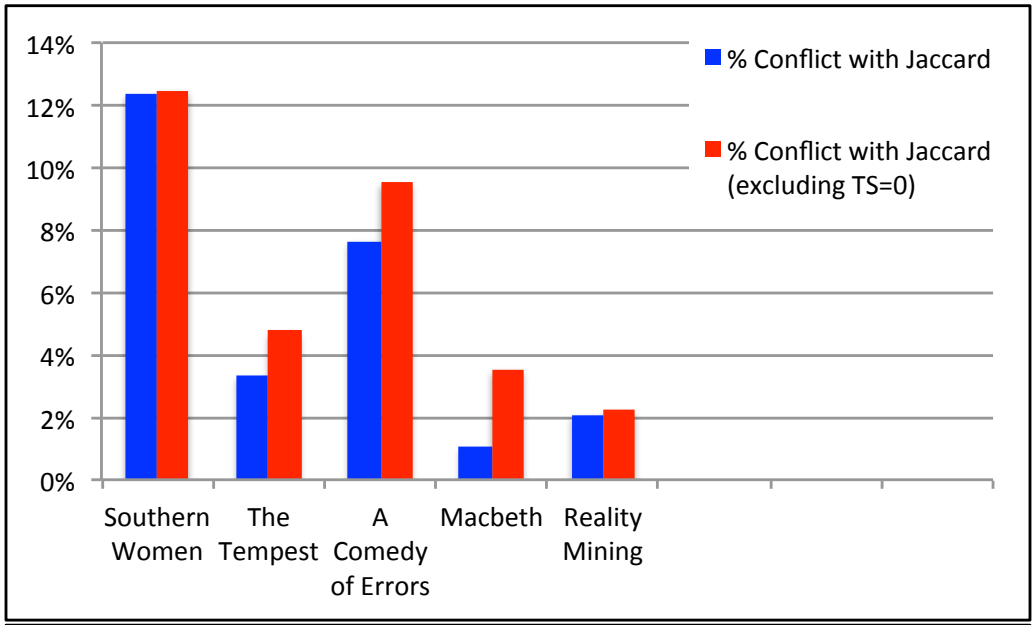
- **Answer:**
  - This is partially true.
  - For some pairs, the tie-strength being set to zero is caused by the axioms.
  - It may or may not be true that all these pairs have tie-strength zero in the actual function used.
    - For example, this won't be true for some self-referential functions like Simrank, Random Walk with Restart, etc.

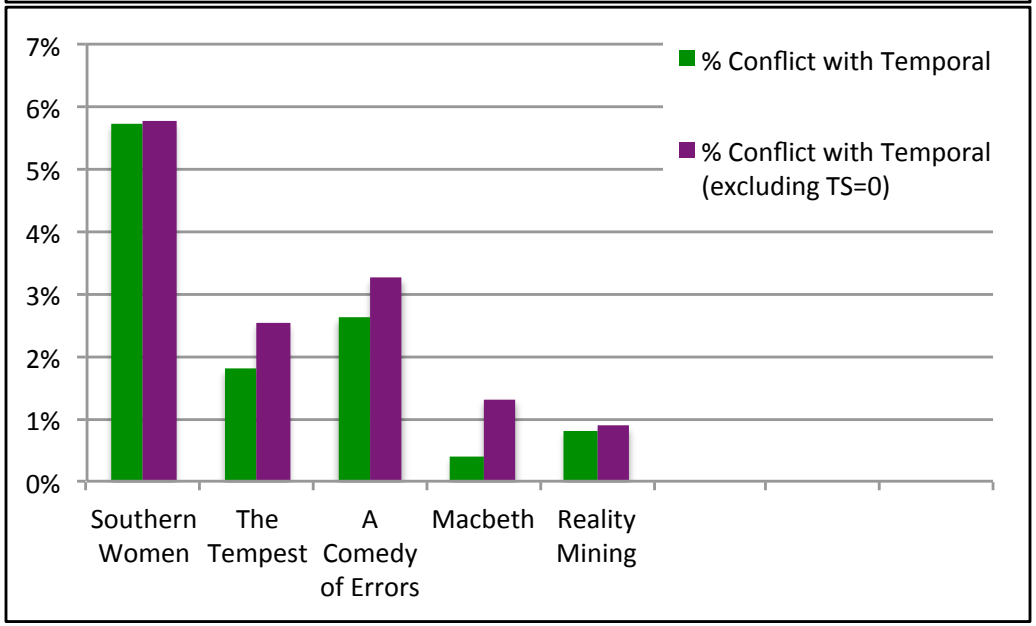
Even More on Soundness

- **Question 2:** How do the conflict numbers change if we only looked at tie pairs that have nonzero tie-strengths?
- **Answer:** The percentages go up but not by much.

Dataset	Tie Pairs	Tie Pairs (excluding TS=0)	Jaccard	Temporal
Southern Women	11,628	11,537	1,441	665
The Tempest	14,535	10,257	488	261
A Comedy of Errors	14,535	11,685	1,114	381
Macbeth	246,753	74,175	2,638	978
Reality Mining	13,794,378	12,819,272	290,934	112,546

Even More on Soundness





Take-away Point #2

% of conflicts between our axioms and tie-strength functions **not** satisfying our axioms is small.*

* Disclaimer: For ranking applications

Putting Take-away Points #1 & #2 Together

1. % of tie pairs on which different tie-strength functions can differ is small

2. % of conflicts between our axioms and tie-strength functions not satisfying our axioms is small

3. If your application is ranking, just pick the most computationally efficient tie-strength measure (e.g. common neighbor).

Disclaimer: For ranking applications

Scalability Issue

- # of tie pairs = $\binom{\binom{n}{2}}{2}$
- Enron has 32,471
- # of tie pairs in Enron ≈ 138 quadrillion

$$\binom{\binom{32471}{2}}{2} = 138,952,356,623,361,270$$

- Ignore zero tie-strengths

A Real-world Application

- **Input:** Data from an online friendship network and its social reader*
- **Q1:** How can we effectively capture the similarities between the reading behaviors of a user and her friends over time?
- **Q2:** How can we effectively summarize such similarities across users?

Details in our *NewsKDD* 2014 paper.
http://eliassi.org/papers/le-newskdd14.pdf

* A reading application deployed on a social network

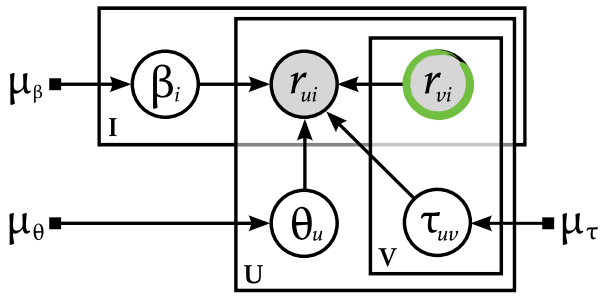
Related Work

- Strength of ties
  - Spread of information in social networks [Granovetter, 1973]
  - Use external information to learn strength of tie
    - [Gilbert & Karahalios, 2009], [Kahanda & Neville, 2009]
- Very few axiomatic work approaches to graph measures
  - PageRank axiomatization [Altman & Tennenholtz, 2005]
- Link prediction
  - [Adamic & Adar, 2003]
  - [Liben-Nowell & Kleinberg, 2003]
  - [Sarkar, Chakrabarti, Moore, 2010 & 2011]

Papers available at http://eliassi.org/pubs

- Measuring Tie Strength in Implicit Social Networks (with Mangesh Gupte), *ACM WebSci* 2012
- Measuring Coverage and Divergence of Reading Behaviors Among Friends (with Long T. Le), *ACM NewsKDD* 2014
- A Probabilistic Model for Using Social Networks in Personalized Item Recommendation (with Allison Chaney and David Blei), *ACM RecSys* 2015

$$r_{ui} \,|\, r_{-u,i} \sim \text{Poisson}\left(\theta_u^\top \beta_i + \sum_{v \in N(u)} \tau_{uv} r_{vi}\right)$$



Thank You!

- Contact info
  - *tina@eliassi.org*
  - *@tinaeliassi*

A postdoc position in my lab, starting Sep 2017.