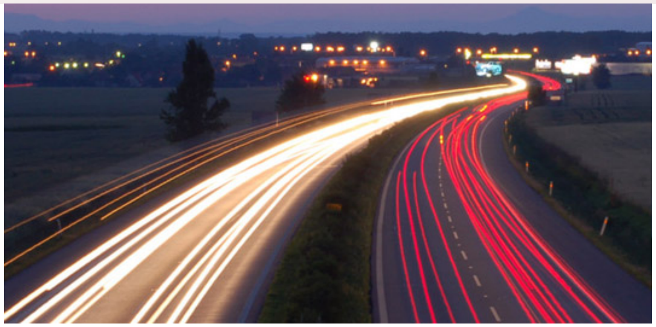
STRUCTURAL FEATURES THREATEN PRIVACY ACROSS SOCIAL GRAPHS: A GRAPH MINER'S VIEW

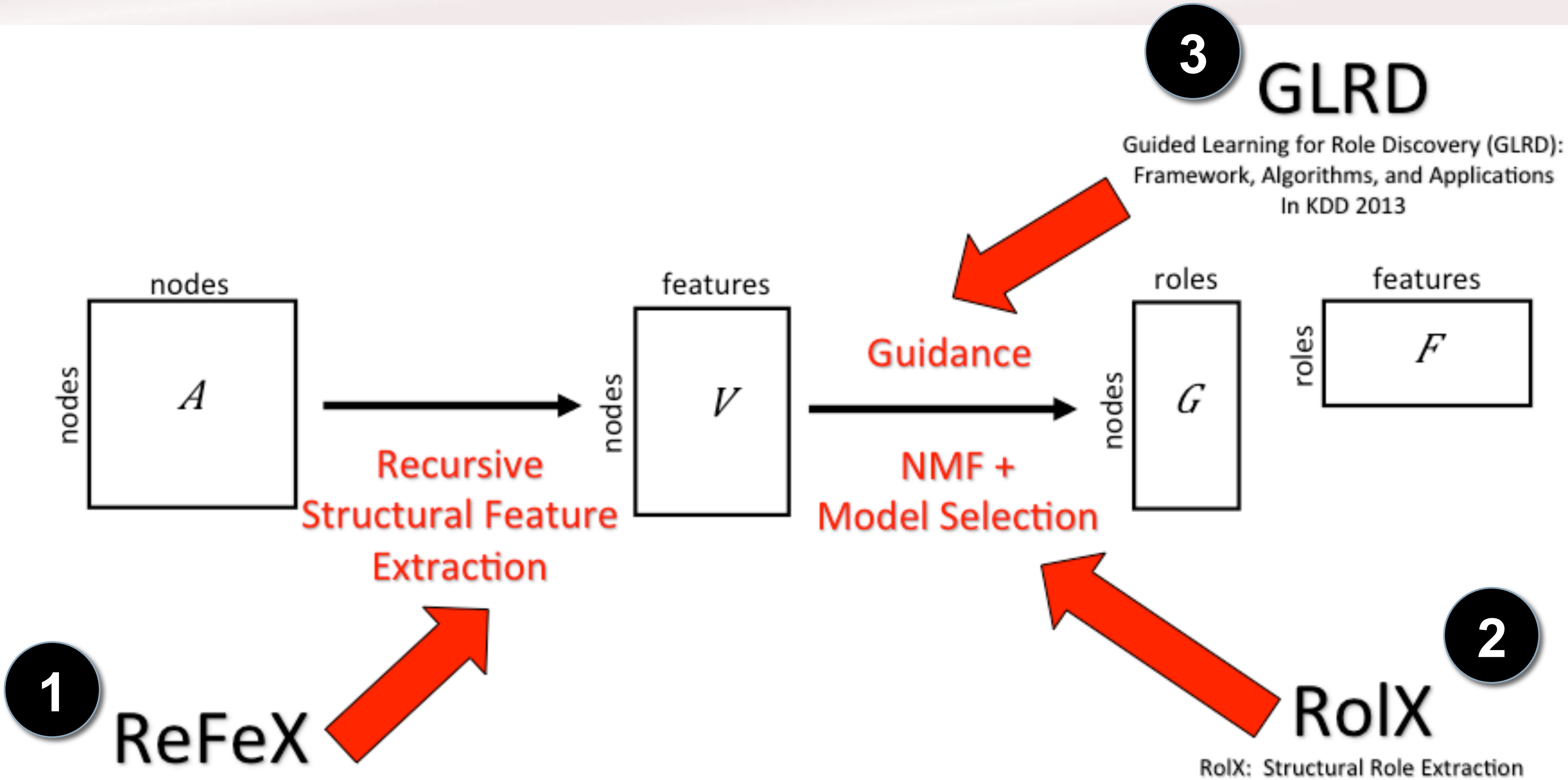
Tina Eliassi-Rad
tina@eliassi.org

Roadmap



- Part 1: Role discovery applied to re-identification
  - [KDD’11, KDD’12, KDD’13]
- Part 2: A relative view of privacy
  - [Work in Progress]

First Part of the Talk



Guidance

Recursive Structural Feature Extraction

NMF + Model Selection

GLRD

Guided Learning for Role Discovery (GLRD): Framework, Algorithms, and Applications
In KDD 2013

ReFeX

It's Who You Know: Graph Mining Using Recursive Structural Features
In KDD 2011

RolX

RolX: Structural Role Extraction & Mining in Large Graphs
In KDD 2012.

Cross-sectional Node Re-Identification

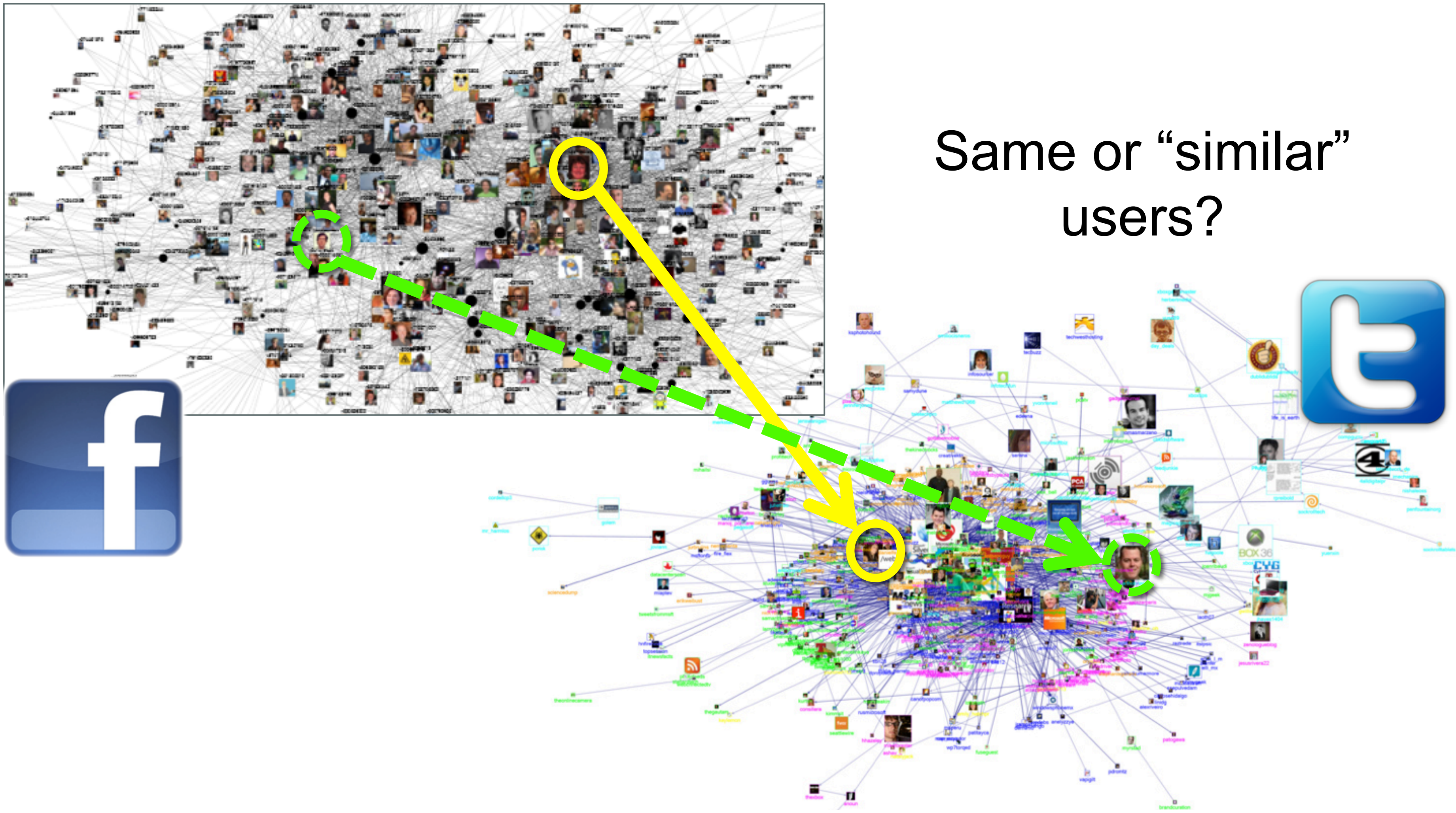
DBLP Co-authorship Networks from 2005-2009

Network	$\lvert V \rvert$	$\lvert E \rvert$	**k**	$\lvert LCC \rvert$	**#CC**
VLDB	1,306	3,224	4.94	769	112
SIGMOD	1,545	4,191	5.43	1,092	116
CIKM	2,367	4,388	3.71	890	361
SIGKDD	1,529	3,158	4.13	743	189
ICDM	1,651	2,883	3.49	458	281
SDM	915	1,501	3.28	243	165

Given a network, there are many behavioral questions we'd like to answer

Task	**Description**
Change detection	Identify unusual changes in behavior
Knowledge transfer	Use knowledge of one network to make predictions in another
Network similarity/ comparison	Determine network compatibility for knowledge transfer
Outlier detection	Identify individuals with unusual behavior
Re-identification	Identify individuals in an anonymized network
Similarity query	Identify individuals with similar behavior to a known target
…	…

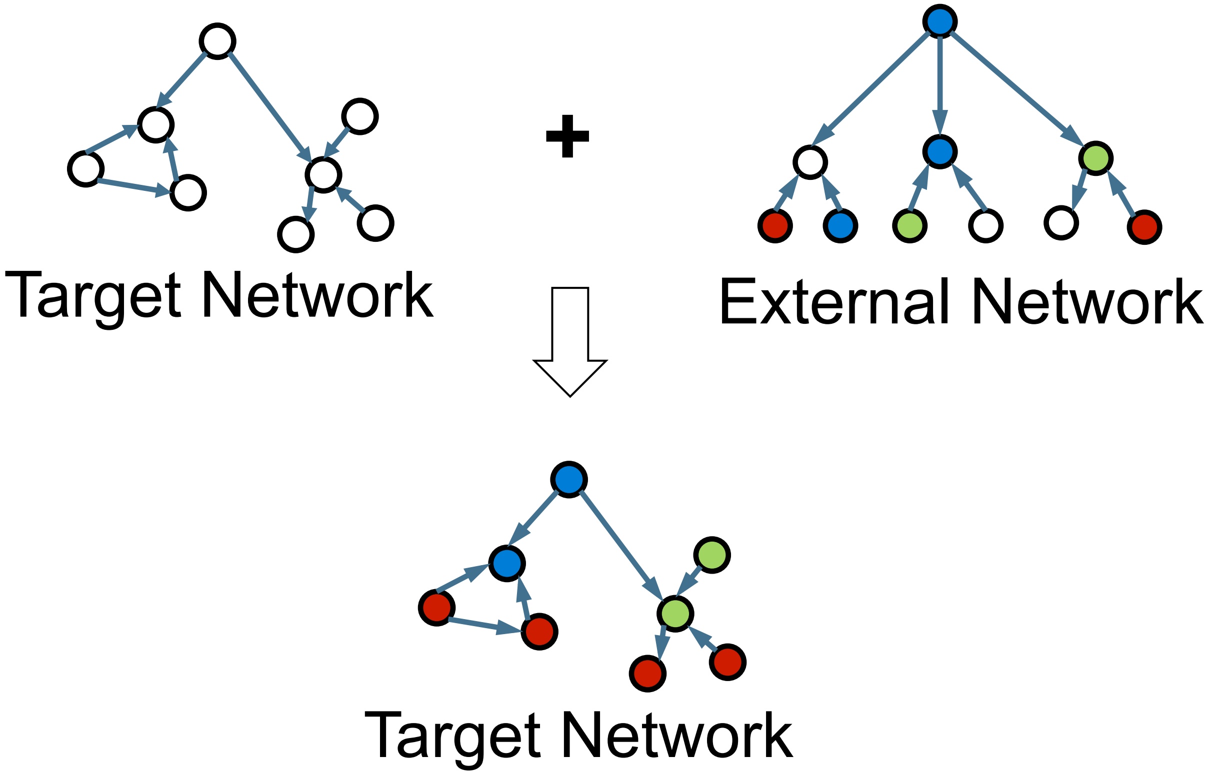
Example: Can we identify users across social graphs?



Same or "similar" users?

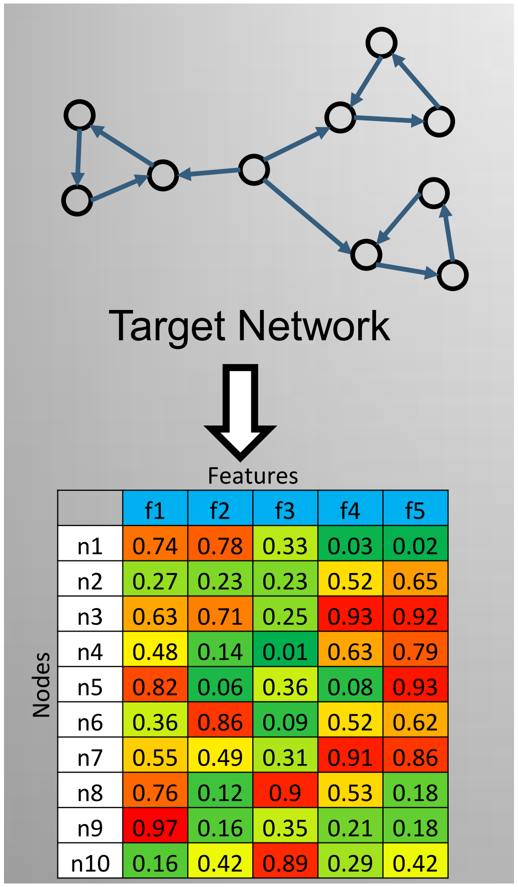
Example: Knowledge Transfer Query

- How can we use labels from an external source to predict labels on a network with no labels?



What features can we extract to do these tasks?

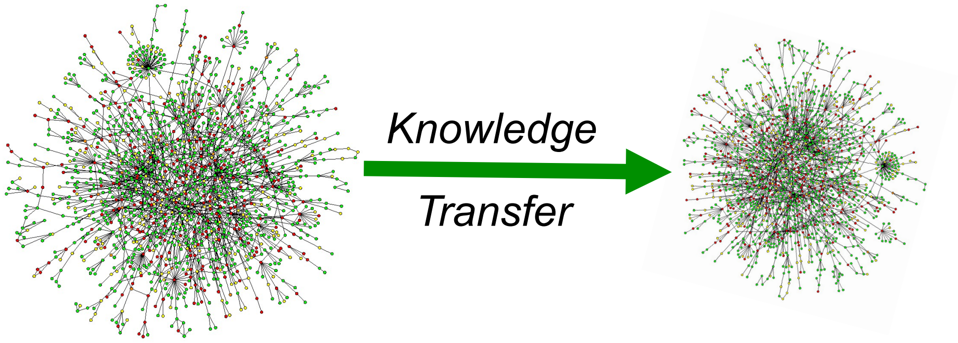
Task	**Description**
Change detection	Identify unusual changes in behavior
Knowledge transfer	Use knowledge of one network to make predictions in another
Network similarity/ comparison	Determine network compatibility for knowledge transfer
Outlier detection	Identify individuals with unusual behavior
Re-identification	Identify individuals in an anonymized network
Similarity Query	Identify individuals with similar behavior to a known target
…	…



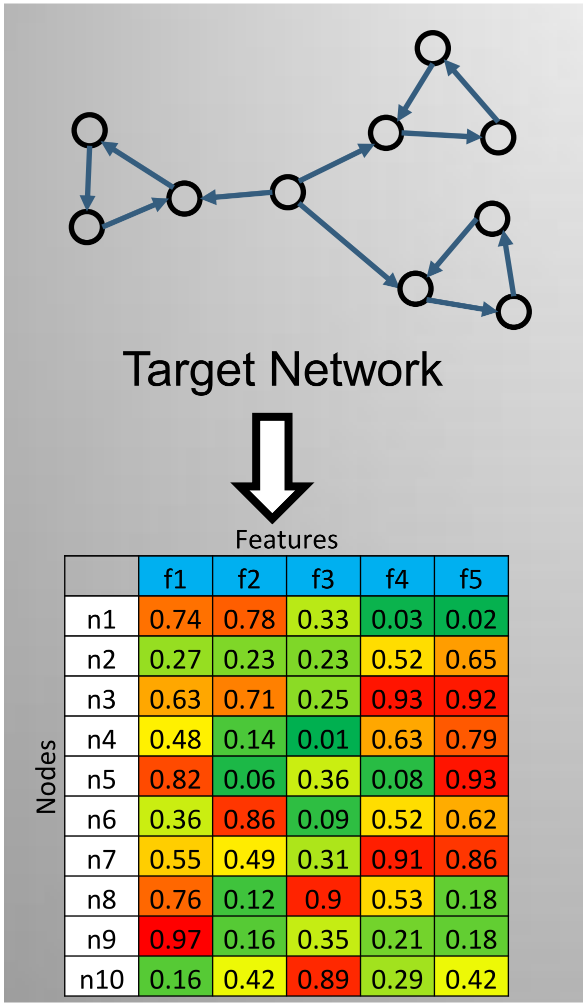
	f1	f2	f3	f4	f5
n1	0.74	0.78	0.33	0.03	0.02
n2	0.27	0.23	0.23	0.52	0.65
n3	0.63	0.71	0.25	0.93	0.92
n4	0.48	0.14	0.01	0.63	0.79
n5	0.82	0.06	0.36	0.08	0.93
n6	0.36	0.86	0.09	0.52	0.62
n7	0.55	0.49	0.31	0.91	0.86
n8	0.76	0.12	0.9	0.53	0.18
n9	0.97	0.16	0.35	0.21	0.18
n10	0.16	0.42	0.89	0.29	0.42

Feature Requirements

- Requirement 1: Effective
  - Features must be predictive and predictive models must transfer across graphs.



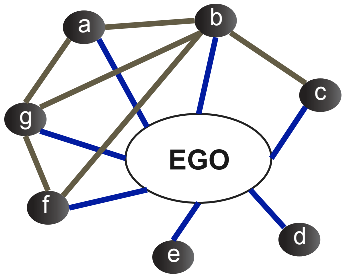
- Requirement 2: Structural
  - Features must not require additional attributes or identity maps.

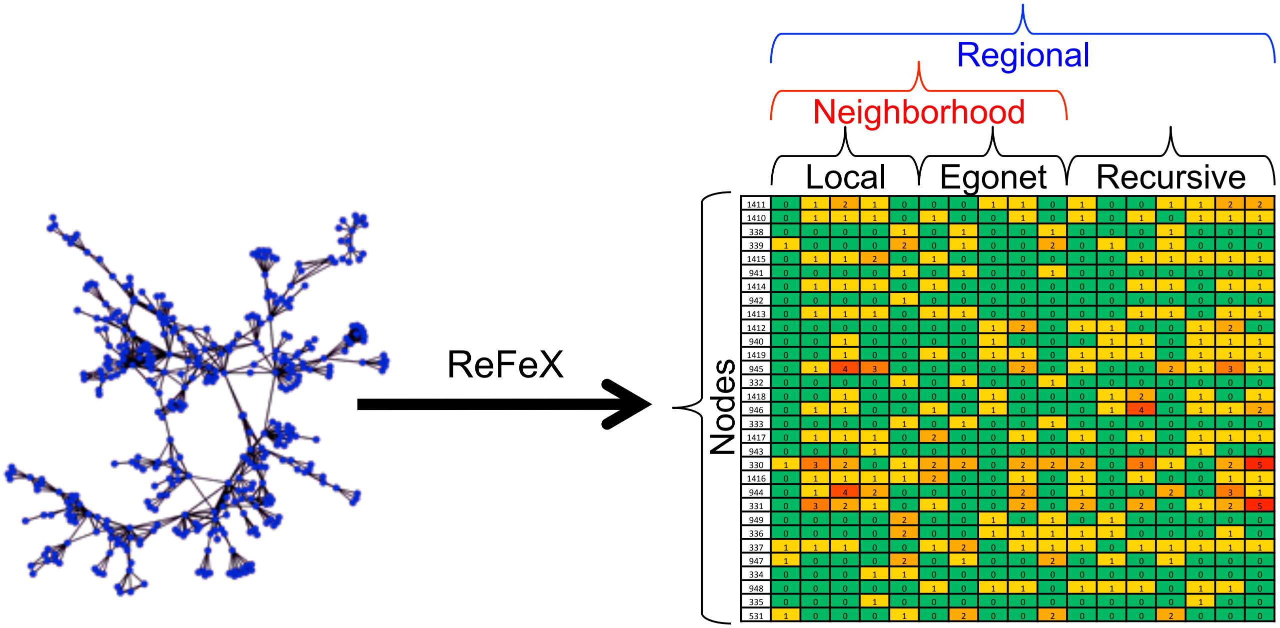


	f1	f2	f3	f4	f5
n1	0.74	0.78	0.33	0.03	0.02
n2	0.27	0.23	0.23	0.52	0.65
n3	0.63	0.71	0.25	0.93	0.92
n4	0.48	0.14	0.01	0.63	0.79
n5	0.82	0.06	0.36	0.08	0.93
n6	0.36	0.86	0.09	0.52	0.62
n7	0.55	0.49	0.31	0.91	0.86
n8	0.76	0.12	0.9	0.53	0.18
n9	0.97	0.16	0.35	0.21	0.18
n10	0.16	0.42	0.89	0.29	0.42

ReFeX: Recursive Feature Extraction

- [Henderson *et al*., KDD 2011]
- Recursively combines node-based features with egonet-based features; & outputs regional features





	Local					Egonet					Recursive						
1411	0	1	2	1	0	0	0	1	1	0	1	0	0	1	1	2	2
1410	0	1	1	1	0	1	0	0	1	0	1	0	1	0	1	1	1
338	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0
339	1	0	0	0	2	0	1	0	0	2	0	1	0	1	0	0	0
1415	0	1	1	2	0	1	0	0	0	0	0	0	1	1	1	1	1
941	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1414	0	1	1	1	0	1	0	0	0	0	0	0	1	1	0	1	1
942	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1413	0	1	1	1	0	1	1	0	0	0	0	0	1	1	0	1	1
1412	0	0	0	0	0	0	0	1	2	0	1	1	0	0	1	2	0
940	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	1	1
1419	0	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	1
945	0	1	4	3	0	0	0	0	2	0	1	0	0	2	1	3	1
332	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1418	0	0	1	0	0	0	0	1	0	0	0	1	2	0	1	0	1
946	0	1	1	0	0	1	0	1	0	0	0	1	4	0	1	1	2
333	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1417	0	1	1	1	0	2	0	0	1	0	1	0	1	0	1	1	1
943	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
330	1	3	2	0	1	2	2	0	2	2	2	0	3	1	0	2	5
1416	0	1	1	1	1	2	0	0	1	0	1	0	1	0	0	1	1
944	0	1	4	2	0	0	0	0	2	0	1	0	0	2	0	3	1
331	0	3	2	1	0	1	0	0	2	0	2	0	2	0	1	2	5
949	0	0	0	0	2	0	0	1	0	1	0	1	0	0	0	0	0
336	0	0	0	0	2	0	0	1	1	1	1	1	0	0	0	1	0
337	1	1	1	0	0	1	2	0	1	1	1	0	1	1	1	1	1
947	1	0	0	0	2	0	1	0	0	2	0	1	0	1	0	0	0
334	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
948	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	1	0
335	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
531	1	0	0	0	1	0	2	0	0	2	0	0	0	2	0	0	0

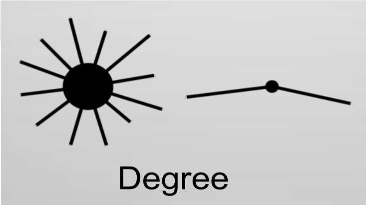
- Neighborhood features: What is your connectivity pattern?
- Recursive Features: To what kinds of nodes are you connected?

ReFeX: Structural Features

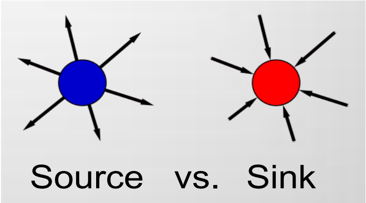
Regional

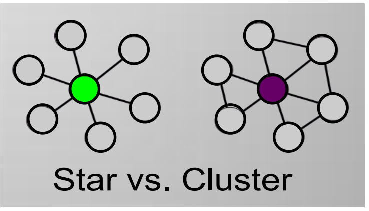
Neighborhood

- Local
  - Essentially measures of the node degree

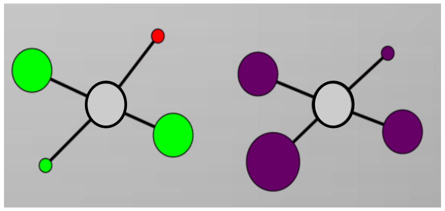

Degree

- Egonet
  - Computed based on each node’s ego network
  - Examples
    - # of within-egonet edges
    - # of edges entering & leaving the egonet

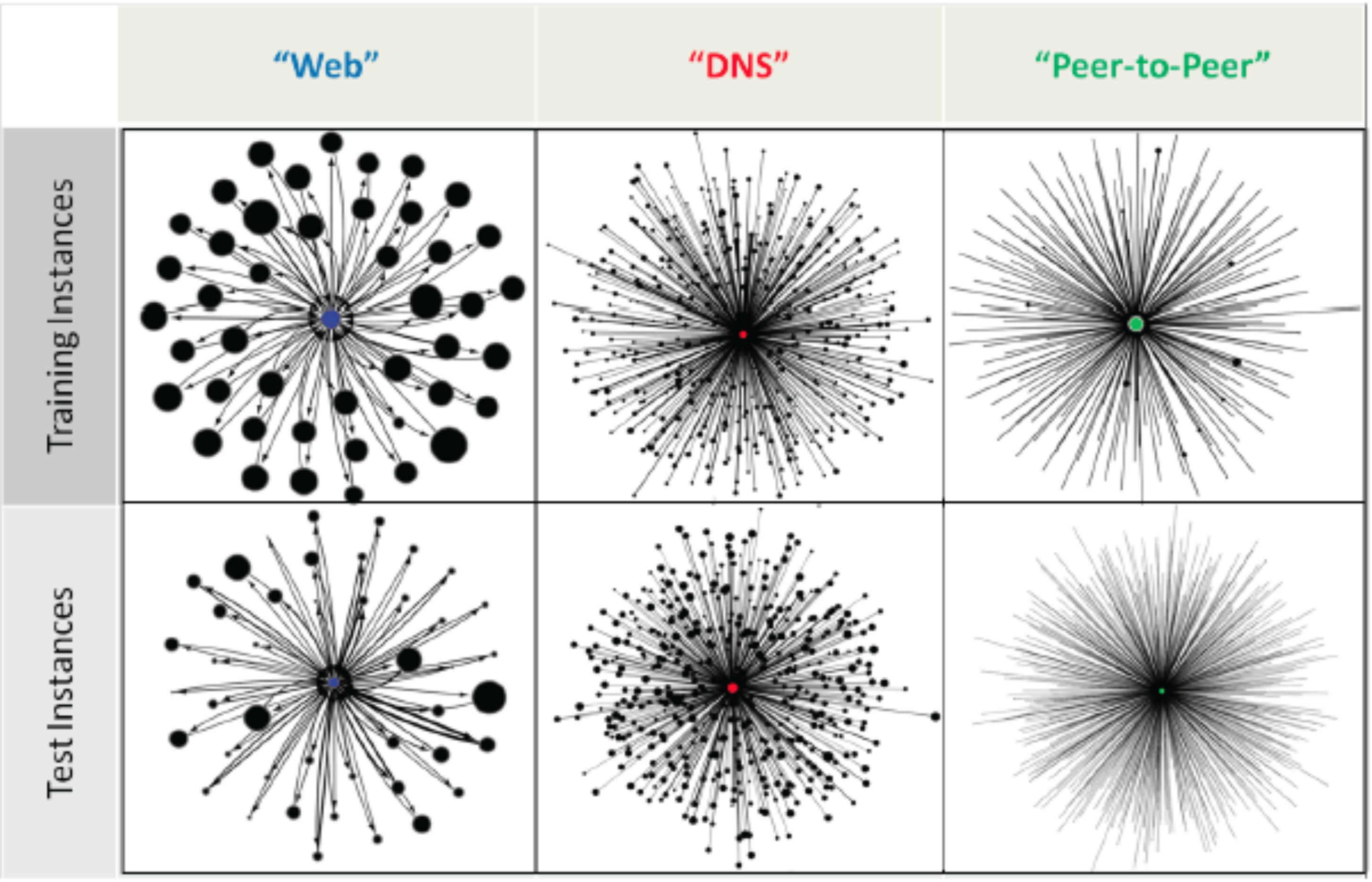

Source vs. Sink


Star vs. Cluster

- 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



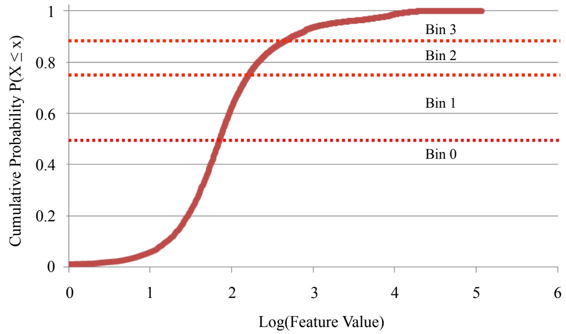
ReFex Intuition: Regional Structure Matters



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

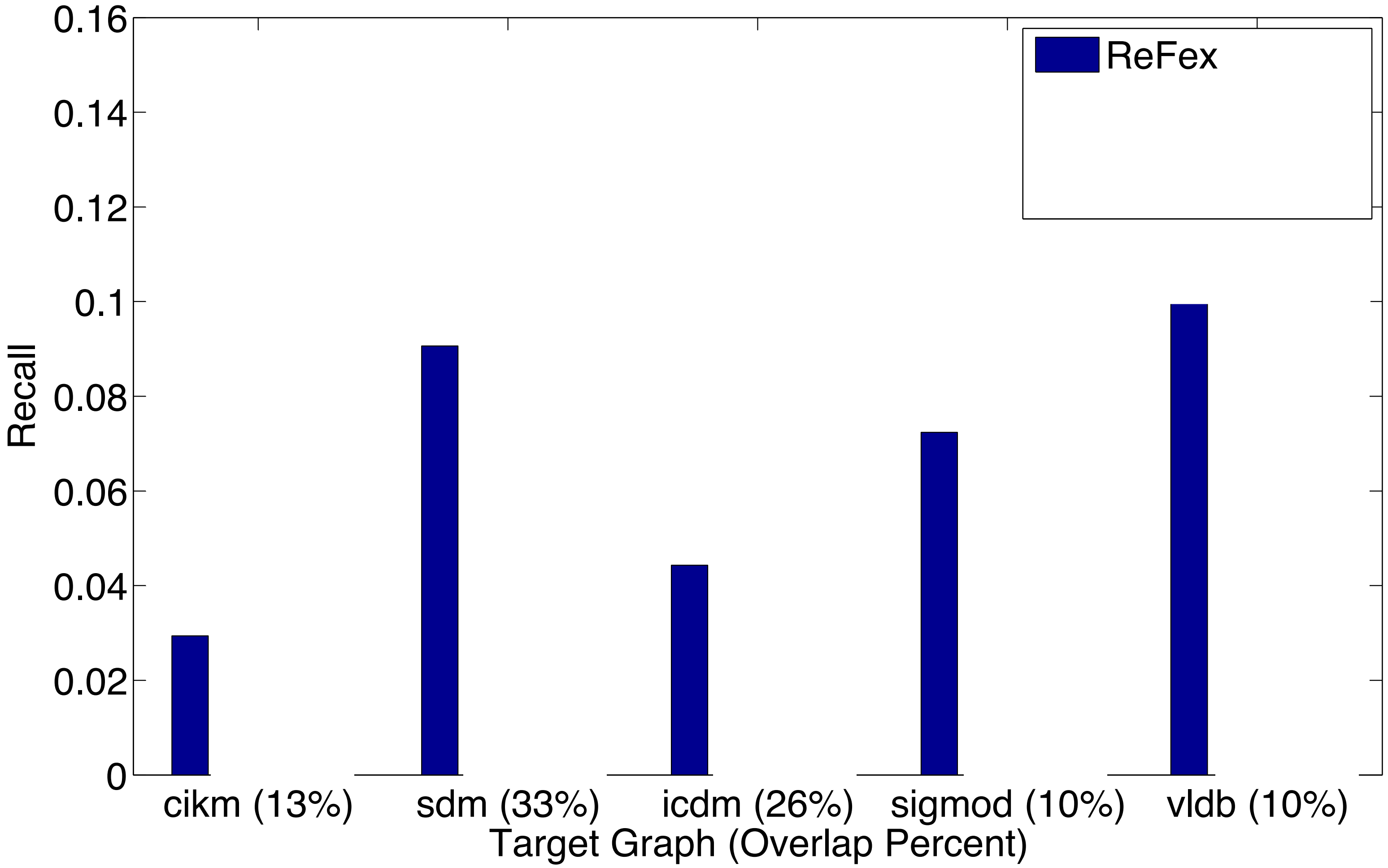
ReFeX (continued)

- 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



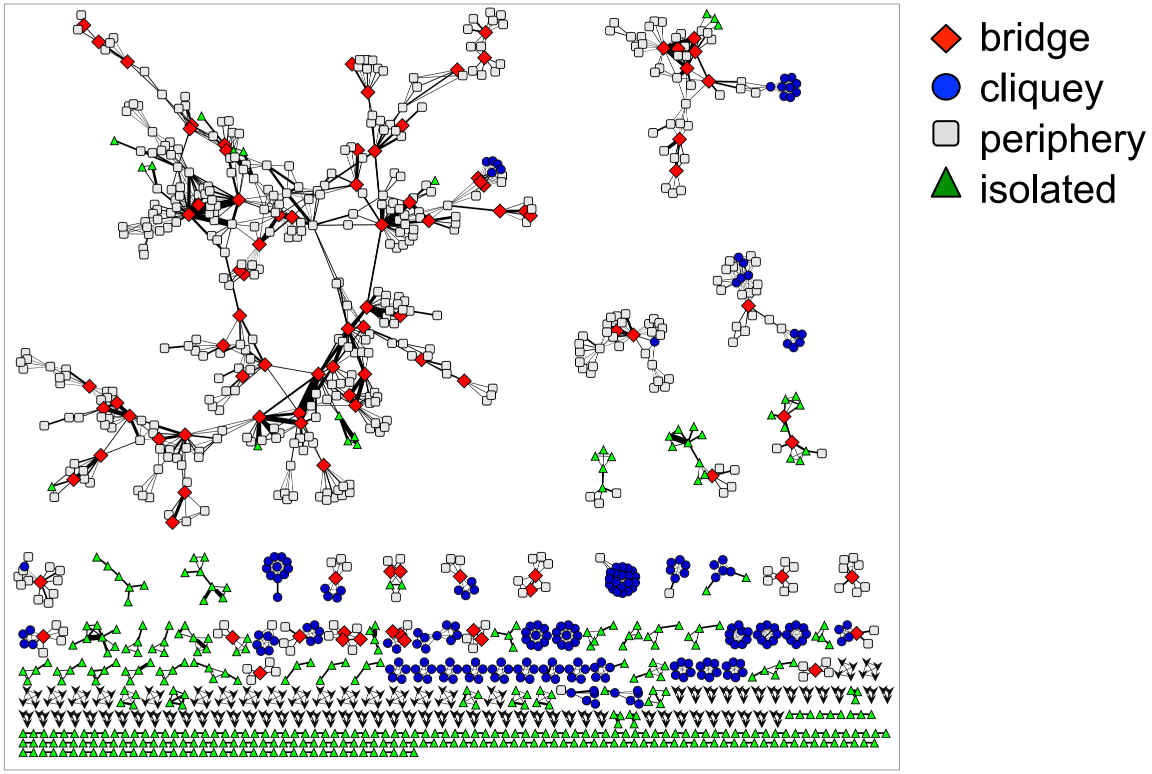
  - Look for pairs of features whose values never disagree by more than a threshold
    - A graph based approach
    - Threshold automatically set
    - Details in the KDD’11 paper

ReFeX on the DBLP Re-ID Task



What are Roles?

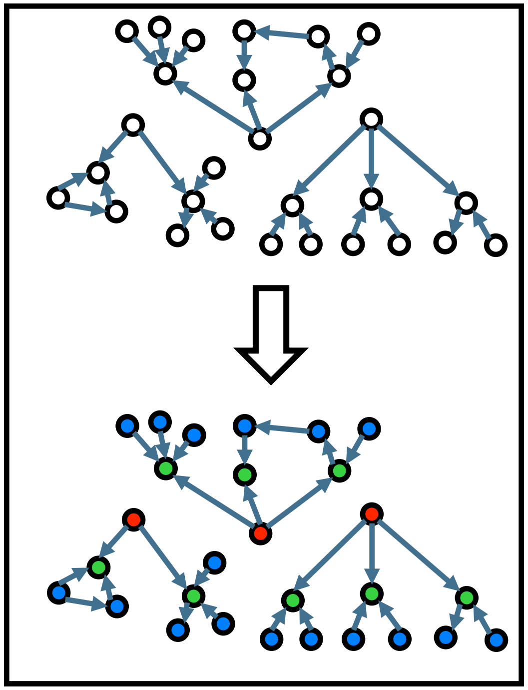
- Roles are “functions” of nodes in the network
  - Similar to functional roles of species in ecosystems
- Measured by structural behaviors



Network Science Co-authorship Network

Why are Roles Important?

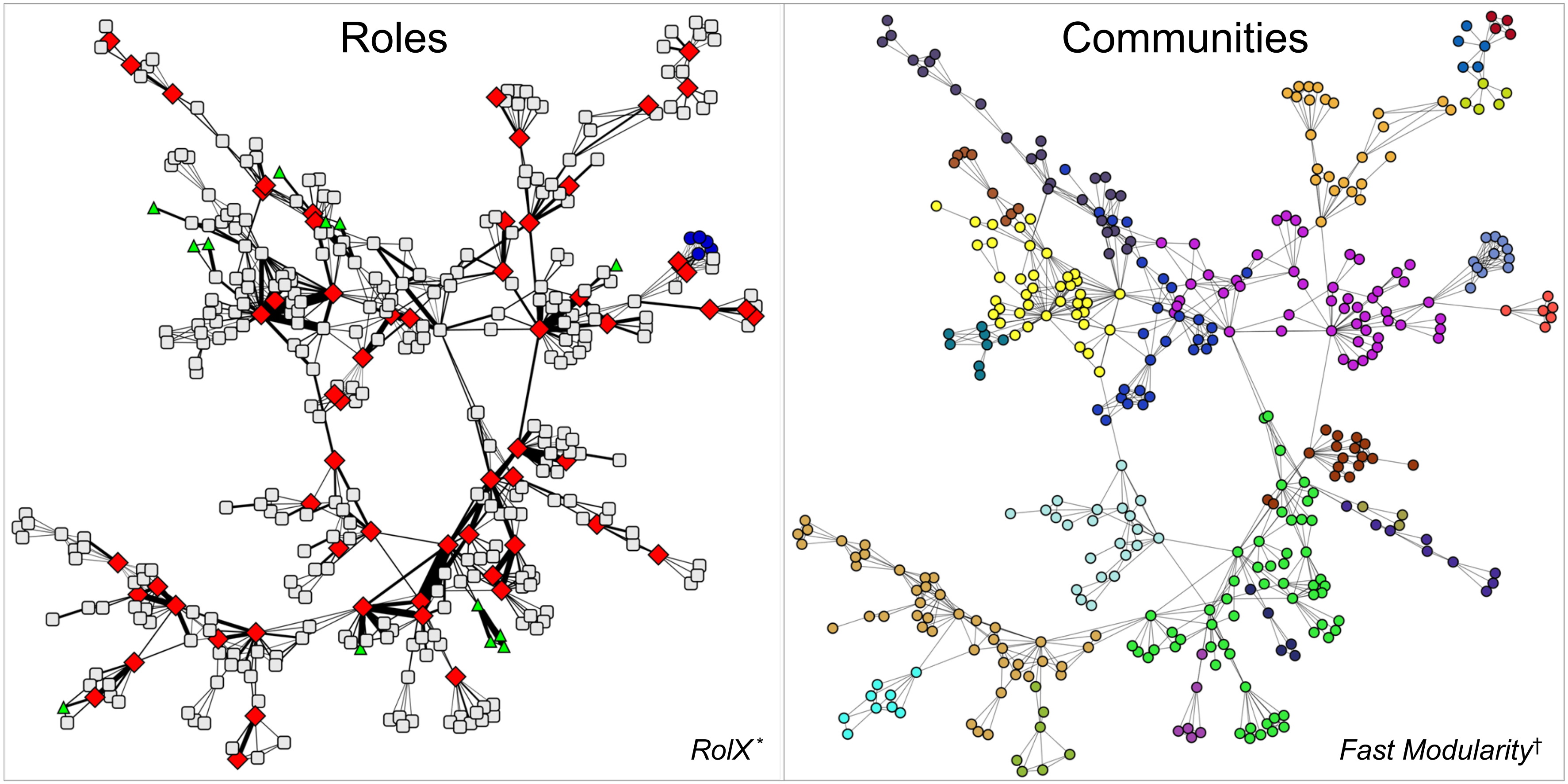
Role Discovery



- ✓ Automated discovery
- ✓ Behavioral roles
- ✓ Roles generalize

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

Roles and Communities are Complementary



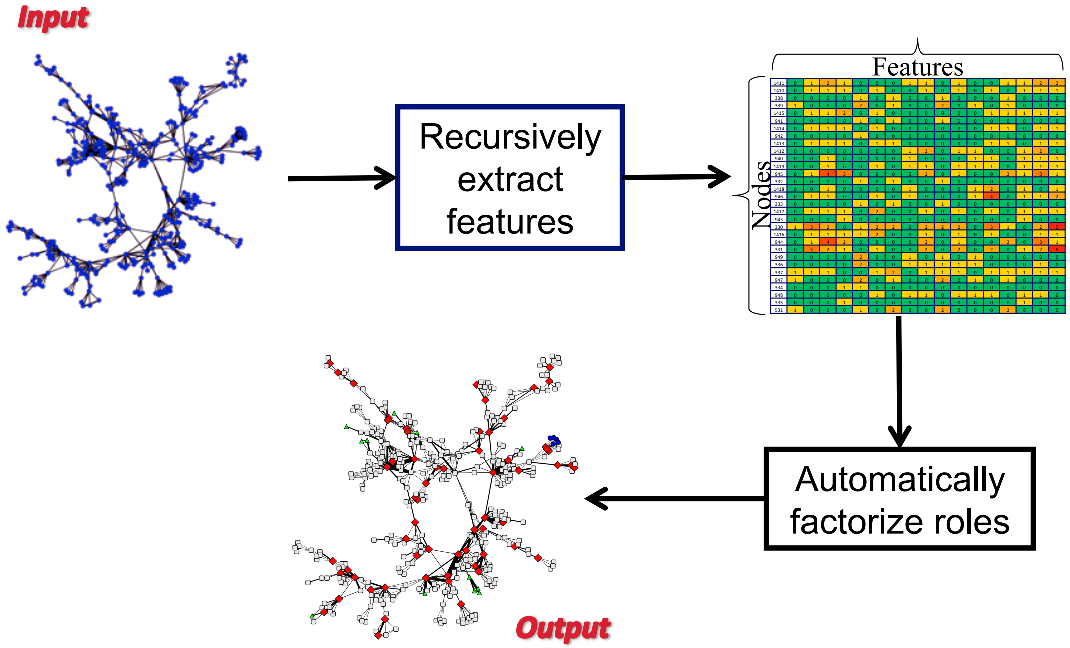
Roles

Communities

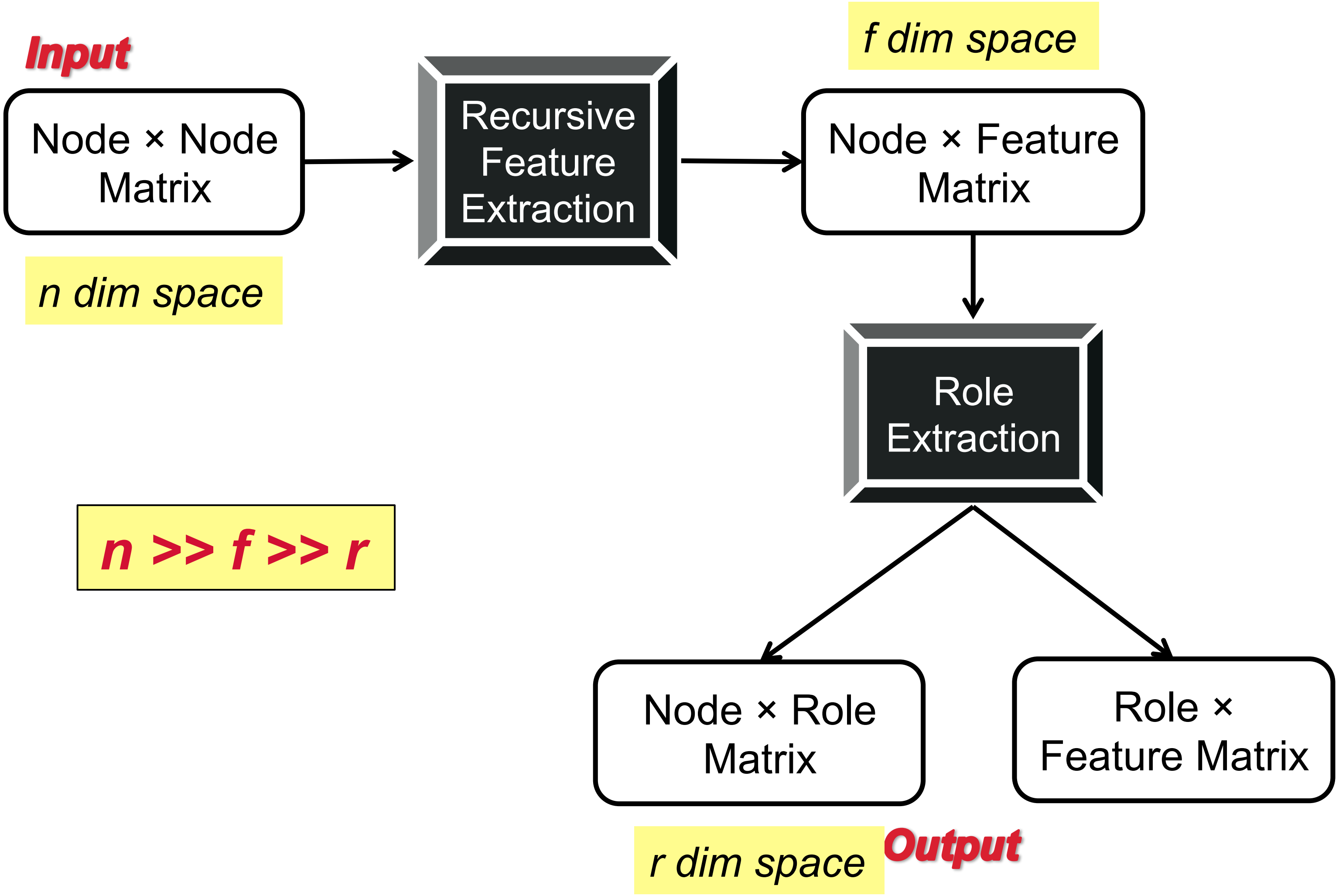
- Roles group nodes with similar structural properties
- Communities group nodes that are well-connected to each other

RolX: Role eXtraction

- [Henderson *et al.*, KDD 2012]
- Automatically extracts the underlying roles in a network
- Determines the number of roles automatically
- Assigns a mixed-membership of roles to each node
- Scales linearly on the number of edges



RolX: Flowchart



Input

Node × Node Matrix

n dim space

Recursive Feature Extraction

f dim space

Node × Feature Matrix

Role Extraction

Node × Role Matrix

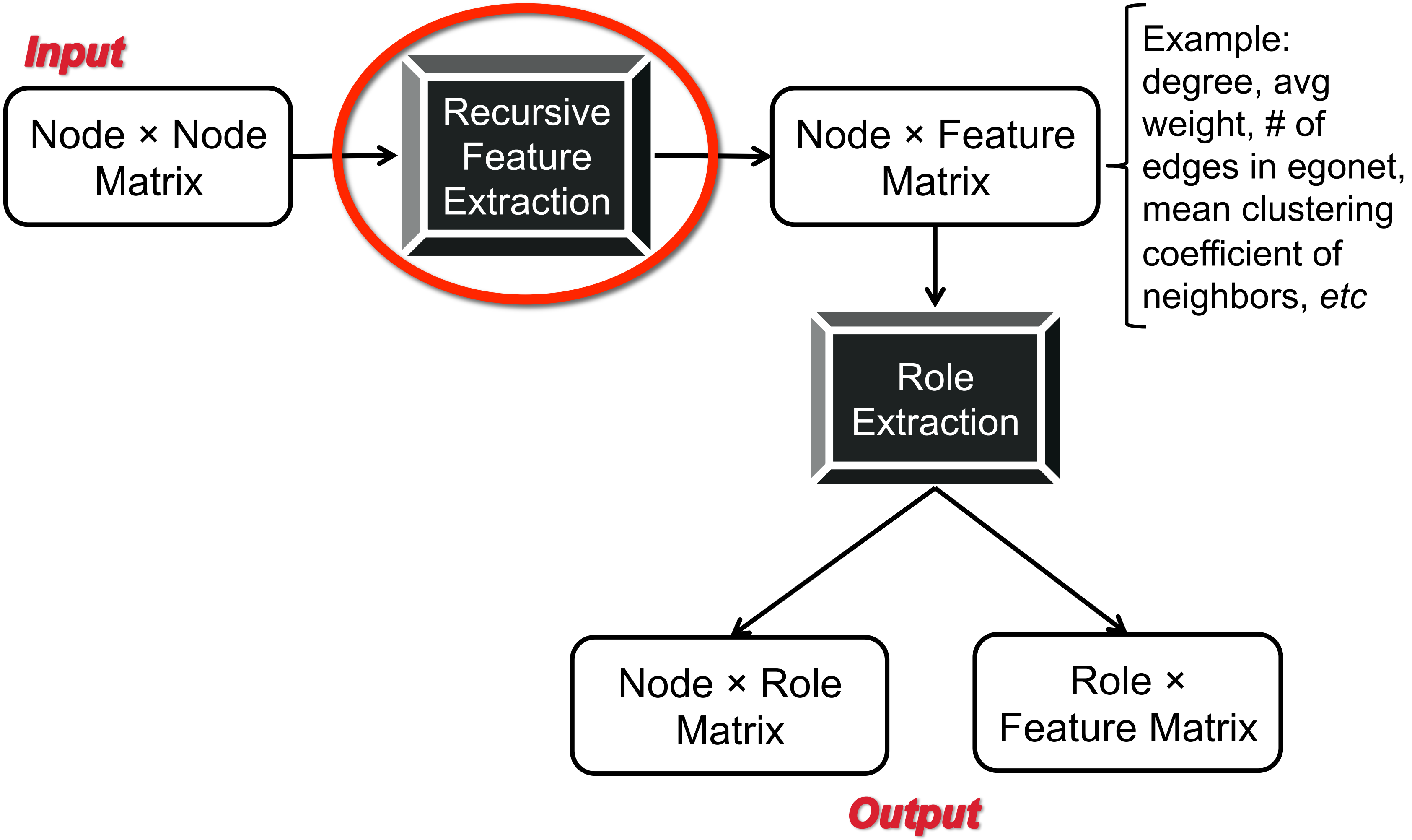
r dim space

Role × Feature Matrix

Output

n >> f >> r

RolX: Flowchart



Input

Node × Node Matrix

Recursive Feature Extraction

Node × Feature Matrix

Example: degree, avg weight, # of edges in egonet, mean clustering coefficient of neighbors, *etc*

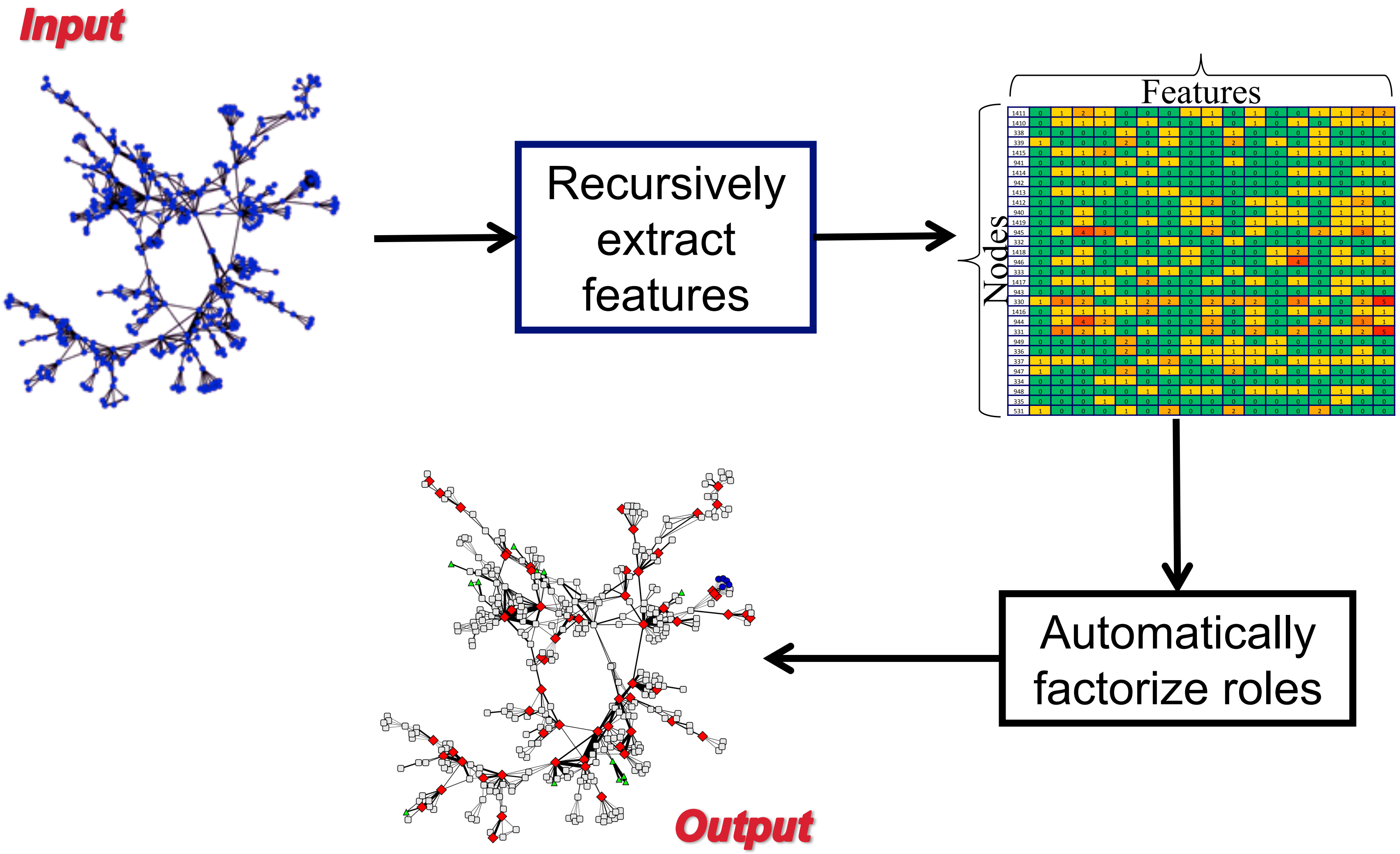
Role Extraction

Node × Role Matrix

Role × Feature Matrix

Output

RolX



Input

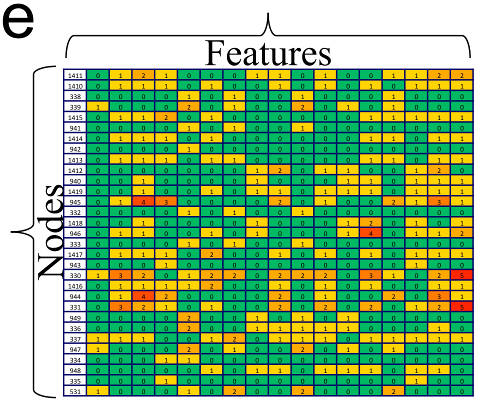
Recursively extract features

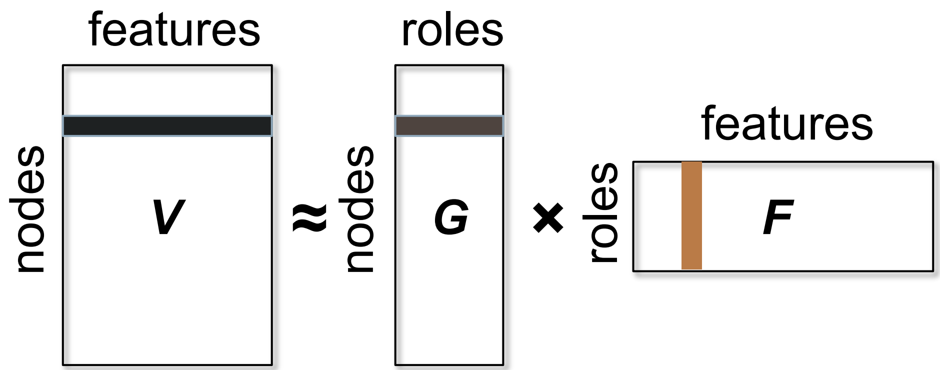
Automatically factorize roles

Output

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

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

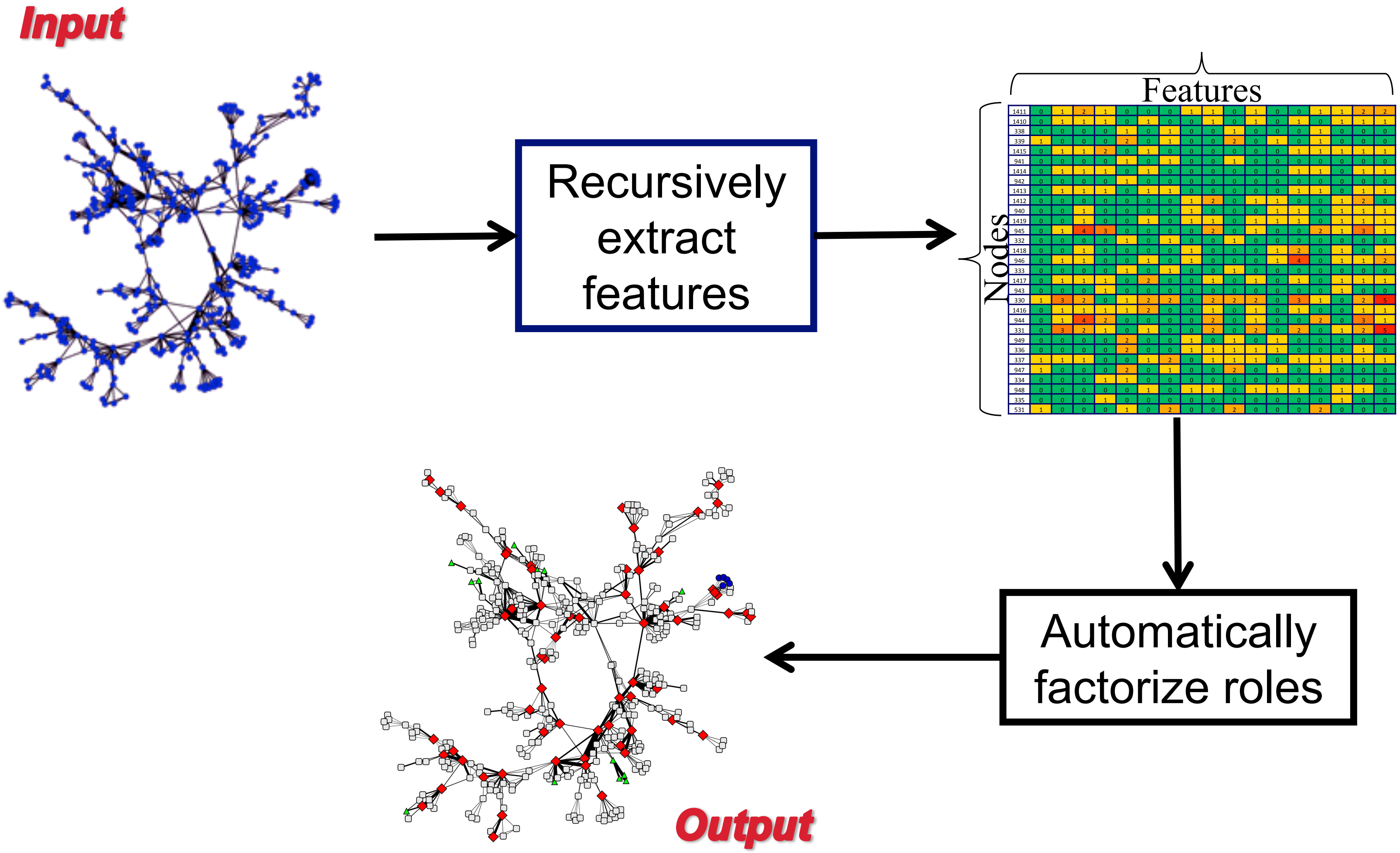
Role Extraction: Model Selection

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

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

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

RolX



Input

Recursively extract features

Features

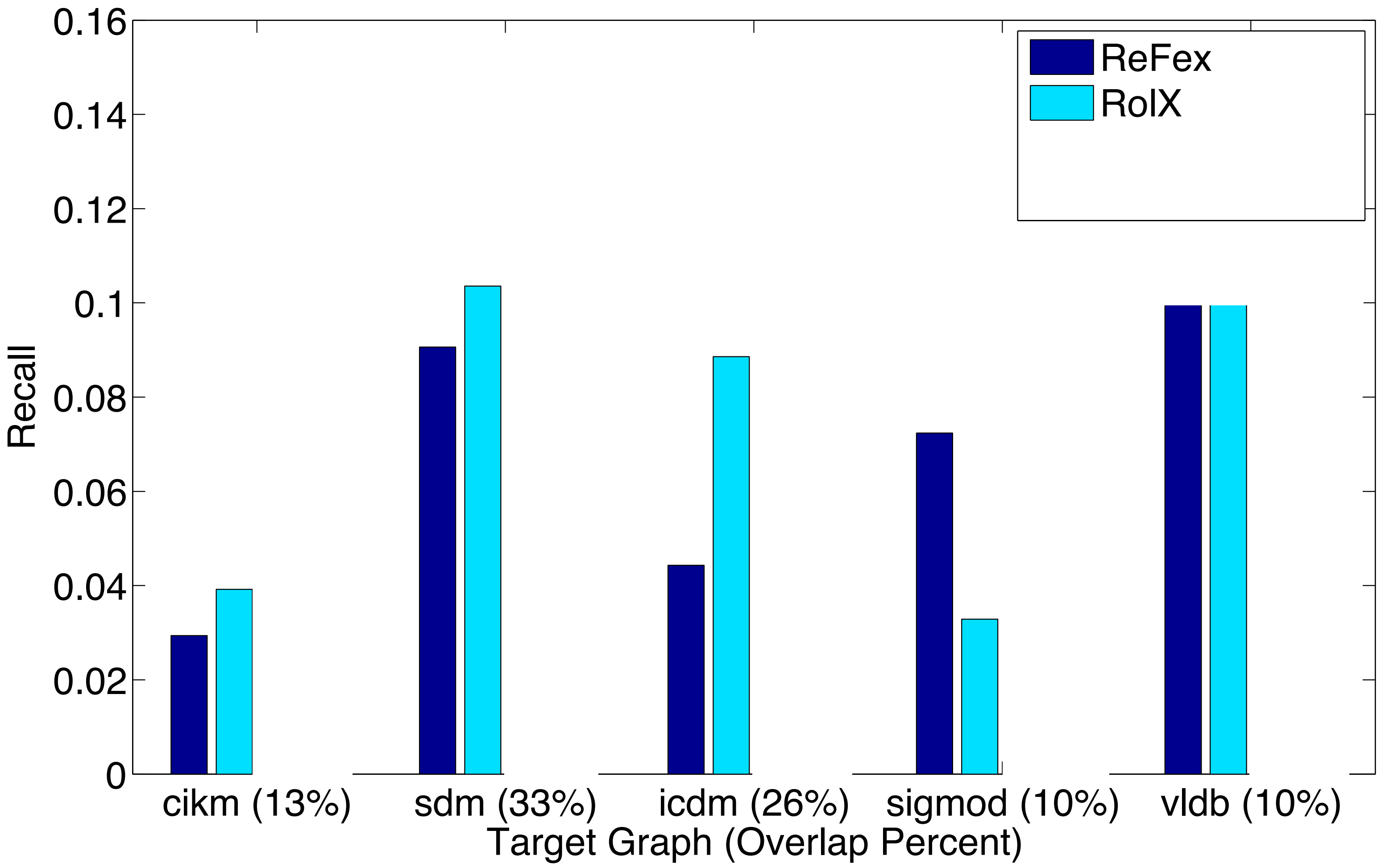
Nodes

1411	0	1	2	1	0	0	0	1	1	0	1	0	0	1	1	2	2
1410	0	1	1	1	0	1	0	0	1	0	1	0	1	0	1	1	1
338	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0
339	1	0	0	0	2	0	1	0	0	2	0	1	0	1	0	0	0
1415	0	1	1	2	0	1	0	0	0	0	0	0	1	1	1	1	1
941	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1414	0	1	1	1	0	1	0	0	0	0	0	0	1	1	0	1	1
942	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1413	0	1	1	1	0	1	1	0	0	0	0	0	1	1	0	1	1
1412	0	0	0	0	0	0	0	1	2	0	1	1	0	0	1	2	0
940	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	1	1
1419	0	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	1
945	0	1	4	3	0	0	0	0	2	0	1	0	0	2	1	3	1
332	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1418	0	0	1	0	0	0	0	1	0	0	0	1	2	0	1	0	1
946	0	1	1	0	0	1	0	1	0	0	0	1	4	0	1	1	2
333	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1417	0	1	1	1	0	2	0	0	1	0	1	0	1	0	1	1	1
943	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
330	1	3	2	0	1	2	2	0	2	2	2	0	3	1	0	2	5
1416	0	1	1	1	1	2	0	0	1	0	1	0	1	0	0	1	1
944	0	1	4	2	0	0	0	0	2	0	1	0	0	2	0	3	1
331	0	3	2	1	0	1	0	0	2	0	2	0	2	0	1	2	5
949	0	0	0	0	2	0	0	1	0	1	0	1	0	0	0	0	0
336	0	0	0	0	2	0	0	1	1	1	1	1	0	0	0	1	0
337	1	1	1	0	0	1	2	0	1	1	1	0	1	1	1	1	1
947	1	0	0	0	2	0	1	0	0	2	0	1	0	1	0	0	0
334	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
948	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	1	0
335	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
531	1	0	0	0	1	0	2	0	0	2	0	0	0	2	0	0	0

Automatically factorize roles

Output

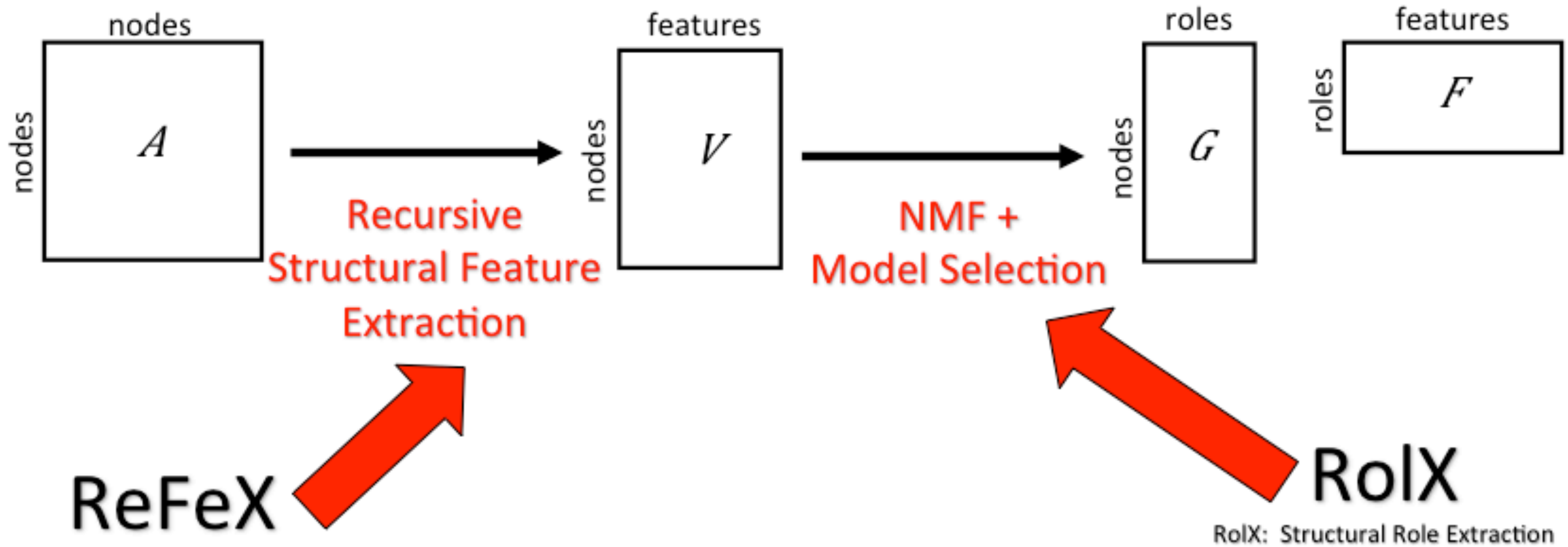
RolX on the DBLP Re-ID Task



GLRD: Guided Learning for Role Discovery

- [KDD’13] with Sean Gilpin and Ian Davidson
- RolX is unsupervised
- What if we had guidance on roles?
  - Guidance as in weak supervision encoded as constraints
- Types of guidance
  - Sparse roles
  - Diverse roles
  - Alternative roles, given a set of existing roles

GLRD



Recursive Structural Feature Extraction

NMF + Model Selection

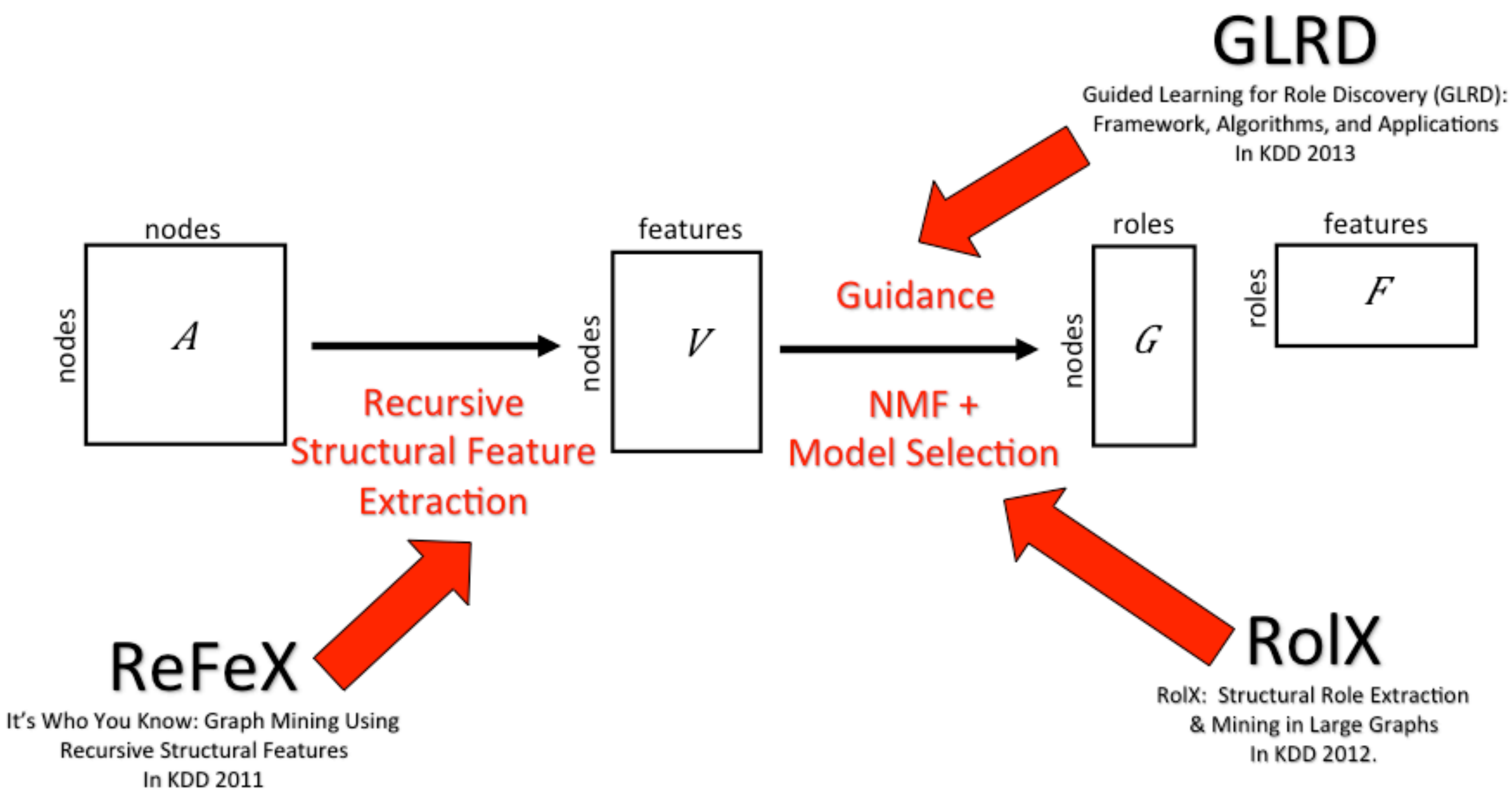
ReFeX

It's Who You Know: Graph Mining Using Recursive Structural Features
In KDD 2011

RolX

RolX: Structural Role Extraction & Mining in Large Graphs
In KDD 2012.

GLRD



GLRD

Guided Learning for Role Discovery (GLRD): Framework, Algorithms, and Applications
In KDD 2013

Guidance

Recursive Structural Feature Extraction

NMF + Model Selection

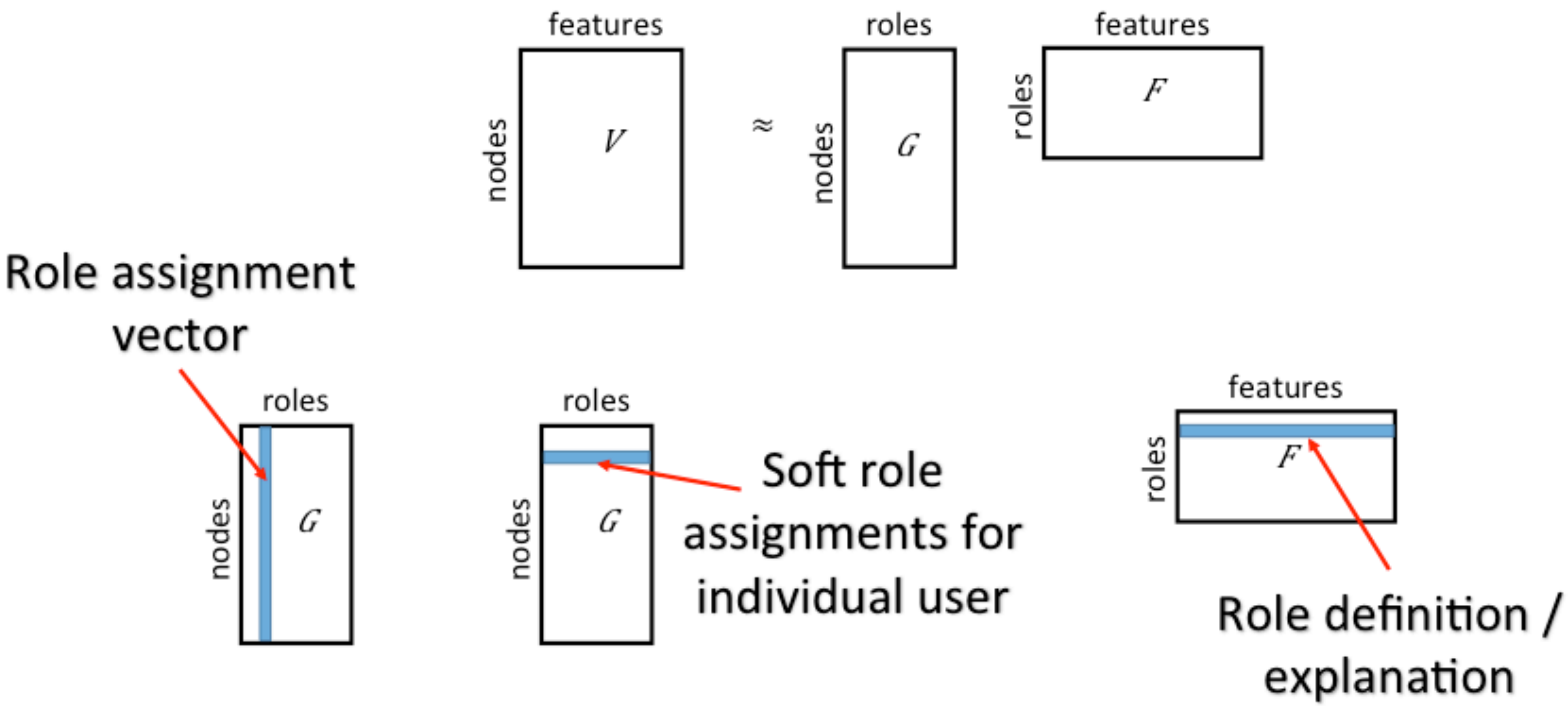
ReFeX

It's Who You Know: Graph Mining Using Recursive Structural Features
In KDD 2011

RolX

RolX: Structural Role Extraction & Mining in Large Graphs
In KDD 2012.

Adding Constraints



Role assignment vector

Soft role assignments for individual user

Role definition / explanation

GLRD Framework

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

$$
\begin{array}{ll}
\underset{\mathbf{G},\mathbf{F}}{\text{minimize}} & ||\mathbf{V}-\mathbf{GF}||_2 \\
\text{subject to} & g_i(\mathbf{G}) \leq d_{Gi}, \; i = 1, \ldots, t_G \\
 & f_i(\mathbf{F}) \leq d_{Fi}, \; i = 1, \ldots, t_F
\end{array}
$$

$$\text{where } \; g_i \text{ and } f_i \; \text{ are convex functions.}$$

- 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

Guidance Overview

Guidance Type	Effect of **increasing** guidance	
	on role assignment (*G*)	on role definition (*F*)
Sparsity	Reduces the number of nodes with minority memberships in roles	Decreases likelihood that features with small explanatory benefit are included
Diversity	Limits the amount of allowable overlap in assignments	Roles must be explained with completely different sets of features
Alternative	Decreases the allowable similarity between the two sets of role assignments	Ensures that role definitions are very dissimilar between the two sets of role assignments

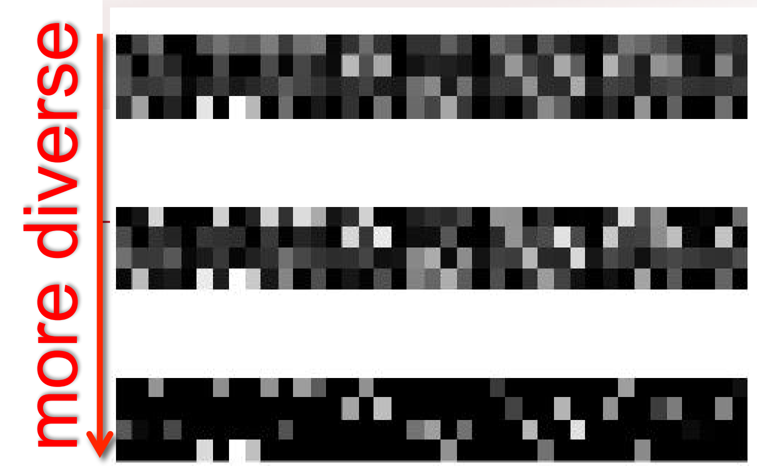
Sparsity

$$\begin{aligned}
\underset{\mathbf{G},\mathbf{F}}{\text{argmin}} \quad & ||\mathbf{V}-\mathbf{GF}||_2 \\
\text{subject to:} \quad & \mathbf{G} \geq 0, \mathbf{F} \geq 0 \\
& \forall i \quad ||\mathbf{G}_{\bullet\mathbf{i}}||_1 \leq \epsilon_G \\
& \forall i \quad ||\mathbf{F}_{\mathbf{i}\bullet}||_1 \leq \epsilon_F
\end{aligned}$$

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

Diversity

Goal: Find role assignments or definitions that are very different from each other



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

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

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

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

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

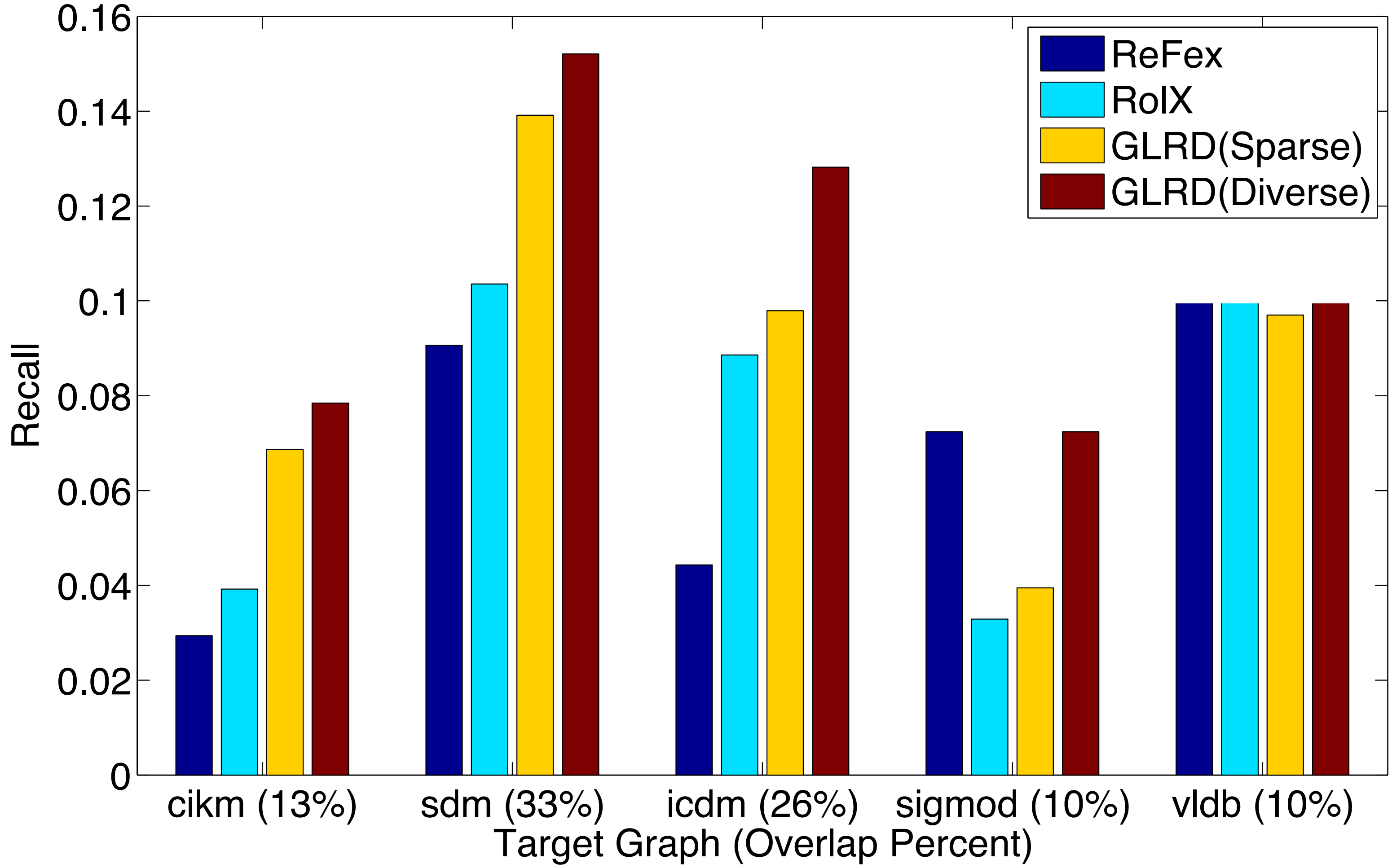
Diverse Roles and Sparse Roles

- 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 graph mining results using various methods for role discovery

Network	**|V|**	**|E|**	**k**	**|LCC|**	**#CC**
VLDB	1,306	3,224	4.94	769	112
SIGMOD	1,545	4,191	5.43	1,092	116
CIKM	2,367	4,388	3.71	890	361
SIGKDD	1,529	3,158	4.13	743	189
ICDM	1,651	2,883	3.49	458	281
SDM	915	1,501	3.28	243	165

DBLP Co-authorship Networks from 2005-2009

GLRD on the DBLP Re-ID Task



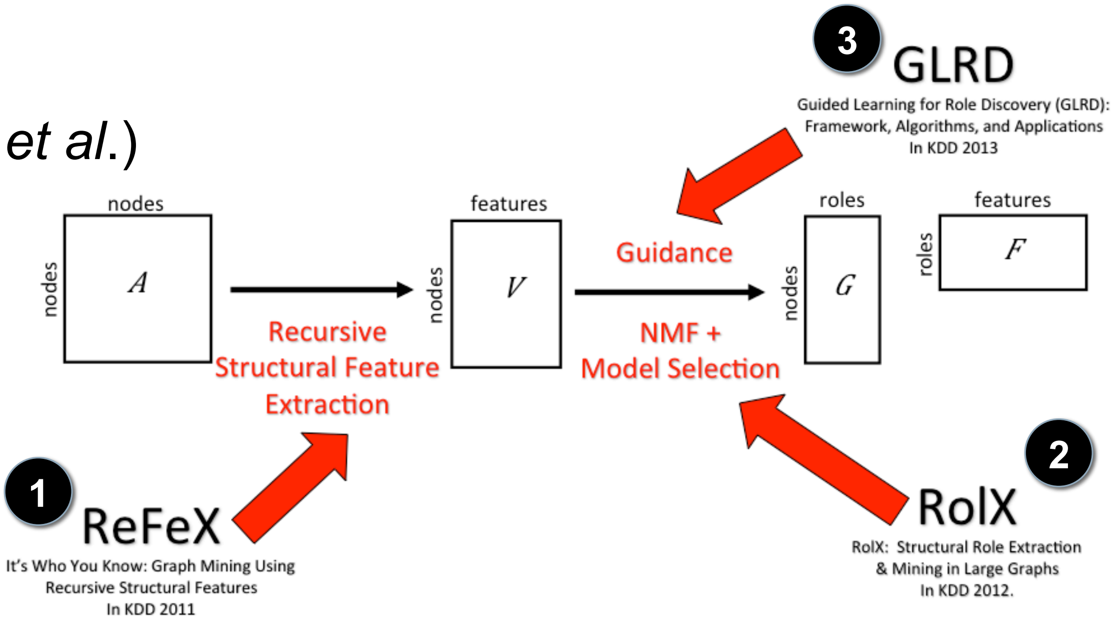
See KDD'11, KDD'12, and KDD'13 papers for details: http://eliassi.org/pubs.html

Recap Part 1: Role Discovery

- **ReFeX** automatically extracts regional structural features
  - Neighborhood features: What is your connectivity pattern?
  - Recursive features: To what kinds of nodes are you connected?
- **Roles** are structural behavior (“function”) of nodes and are complementary to communities
- **RolX**
  - Maps nodes in a graph to a lower-dimensional *role space*
  - Each node has a mixed-membership over roles
  - Automatically selects the best model
  - Roles generalize across disjoint graphs
  - Has many applications in graph mining: transfer learning, affecting dissemination, re-ID, node dynamics, *etc*
- **GLRD** can incorporate guidance in role discovery
- All are scalable (linear on # of edges)

Recap Part 1: Role Discovery

- Several tutorials on this work are available (http://eliassi.org)
- Previous work mostly in sociology under positions and regular equivalences
- Joint work with
  - LLNL (Keith Henderson & Brian Gallagher)
  - CMU (Christos Faloutsos *et al*.)
  - Google (Sugato Basu)
  - UC Davis (Ian Davidson *et al*.)
  - Rutgers (Long T. Le)

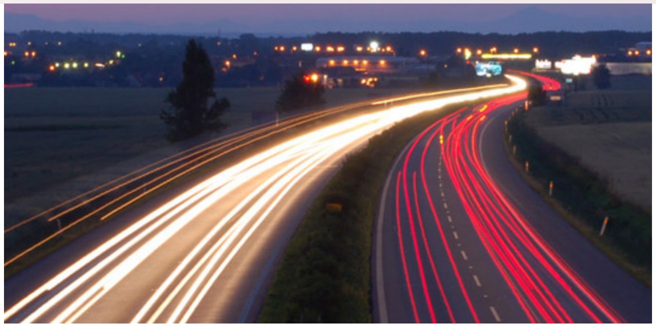


Guided Learning for Role Discovery (GLRD): Framework, Algorithms, and Applications
In KDD 2013

It's Who You Know: Graph Mining Using Recursive Structural Features
In KDD 2011

RolX: Structural Role Extraction & Mining in Large Graphs
In KDD 2012.

Roadmap



- Part 1: Role discovery applied to re-identification
  - [KDD’11, KDD’12, KDD’13]
- Part 2: A relative view of privacy
  - [Work in Progress]
  - Joint with Priya Govindan (Rutgers), Shawndra Hill & Jin Xu (UPenn Wharton), and Chris Volinsky (AT&T Research)

Motivation

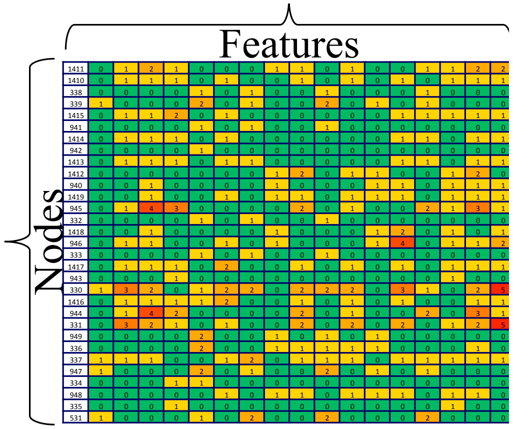
- 87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}[1]
- Releasing anonymized graphs, with a small partial matching can reveal identities.[2]

[1] L. Sweeney. Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper, 2000.

[2] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, 2009.

Motivation

- 87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}[1]
- Releasing anonymized graphs, with a small partial matching can reveal identities.[2]
- **Can a <u>handful</u> of anonymized structural features "break privacy"?**



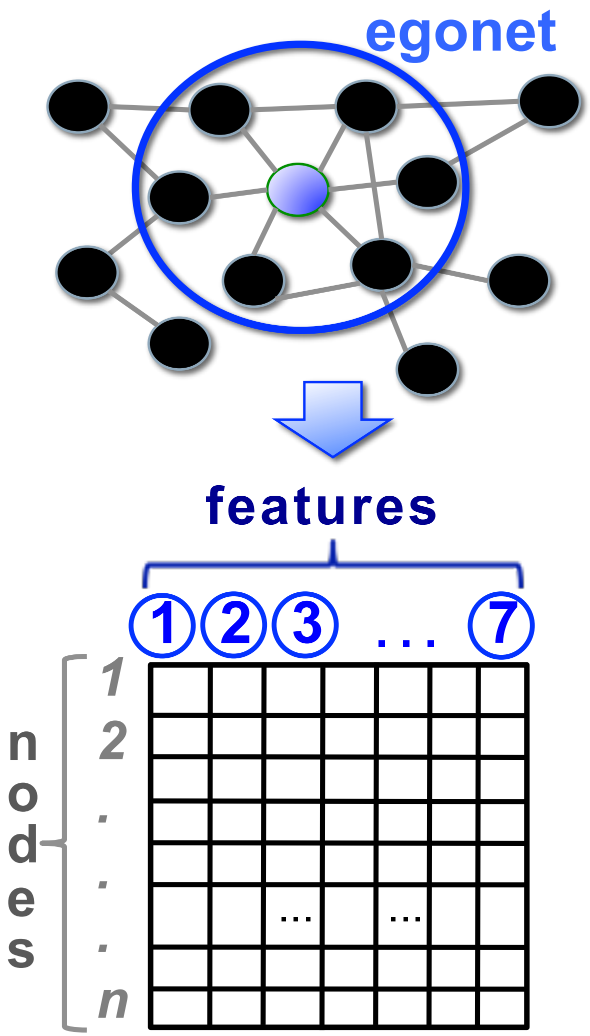
1411	0	1	2	1	0	0	0	1	1	0	1	0	0	1	1	2	2
1410	0	1	1	1	0	1	0	0	1	0	1	0	1	0	1	1	1
338	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0
339	1	0	0	0	2	0	1	0	0	2	0	1	0	1	0	0	0
1415	0	1	1	2	0	1	0	0	0	0	0	0	1	1	1	1	1
941	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1414	0	1	1	1	0	1	0	0	0	0	0	0	1	1	0	1	1
942	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1413	0	1	1	1	0	1	1	0	0	0	0	0	1	1	0	1	1
1412	0	0	0	0	0	0	0	1	2	0	1	1	0	0	1	2	0
940	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	1	1
1419	0	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	1
945	0	1	4	3	0	0	0	0	2	0	1	0	0	2	1	3	1
332	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1418	0	0	1	0	0	0	0	1	0	0	0	1	2	0	1	0	1
946	0	1	1	0	0	1	0	1	0	0	0	1	4	0	1	1	2
333	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
1417	0	1	1	1	0	2	0	0	1	0	1	0	1	0	1	1	1
943	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
330	1	3	2	0	1	2	2	0	2	2	2	0	3	1	0	2	5
1416	0	1	1	1	1	2	0	0	1	0	1	0	1	0	0	1	1
944	0	1	4	2	0	0	0	0	2	0	1	0	0	2	0	3	1
331	0	3	2	1	0	1	0	0	2	0	2	0	2	0	1	2	5
949	0	0	0	0	2	0	0	1	0	1	0	1	0	0	0	0	0
336	0	0	0	0	2	0	0	1	1	1	1	1	0	0	0	1	0
337	1	1	1	0	0	1	2	0	1	1	1	0	1	1	1	1	1
947	1	0	0	0	2	0	1	0	0	2	0	1	0	1	0	0	0
334	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
948	0	0	0	0	0	1	0	1	1	0	1	1	1	0	1	1	0
335	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
531	1	0	0	0	1	0	2	0	0	2	0	0	0	2	0	0	0

[1] L. Sweeney. Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper, 2000.

[2] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, 2009.

Features Tied to Popular Social Theories

- Tied to four social theories
  - *Social capital* (connectivity)
  - *Social exchange* (reciprocity)
  - *Balance* (transitivity)
  - *Structural hole* (control of info flow)
- Local and egonet features [Berlingerio et al. ASONAM'13]:
  - ① # of neighbors
  - ② clustering coefficient
  - ③ avg. # of neighbors' neighbors
  - ④ avg. clustering coeff. of neighbors
  - ⑤ edges in egonet
  - ⑥ outgoing edges from egonet
  - ⑦ # of neighbors of egonet



Problem Definition

Releases

Structural Feature Table, *F*

	Degree	Clust. Coeff.	Avg Degree of Nbrs	Avg Clust. Coeff. of Nbrs	# Edges in Egonet	# Outgoing Edges from Egonet	# Nbrs of Egonet
1							
2							
3							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
n							

Nodes

Local and Egonet based Features

Problem Setting

Releases

Structural Feature Table, *F*

	Degree	Clust. Coeff.	Avg Degree of Nbrs	Avg Clust. Coeff. of Nbrs	# Edges in Egonet	# Outgoing Edges from Egonet	# Nbrs of Egonet
1							
2							
3							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
n							

Nodes

Local and Egonet based Features

Problem Setting

Releases

Structural Feature Table, F

	Degree	Clust. Coeff.	Avg Degree of Nbrs	Avg Clust. Coeff. of Nbrs	# Edges in Egonet	# Outgoing Edges from Egonet	# Nbrs of Egonet
1							
2							
3							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
n							

Nodes

Local and Egonet based Features

Structural Feature Table, F^{aux}

	Degree	Clust. Coeff.	Avg Degree of Nbrs	Avg Clust. Coeff. of Nbrs	# Edges in Egonet	# Outgoing Edges from Egonet	# Nbrs of Egonet
1							
2							
3							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
…							
n^{aux}							

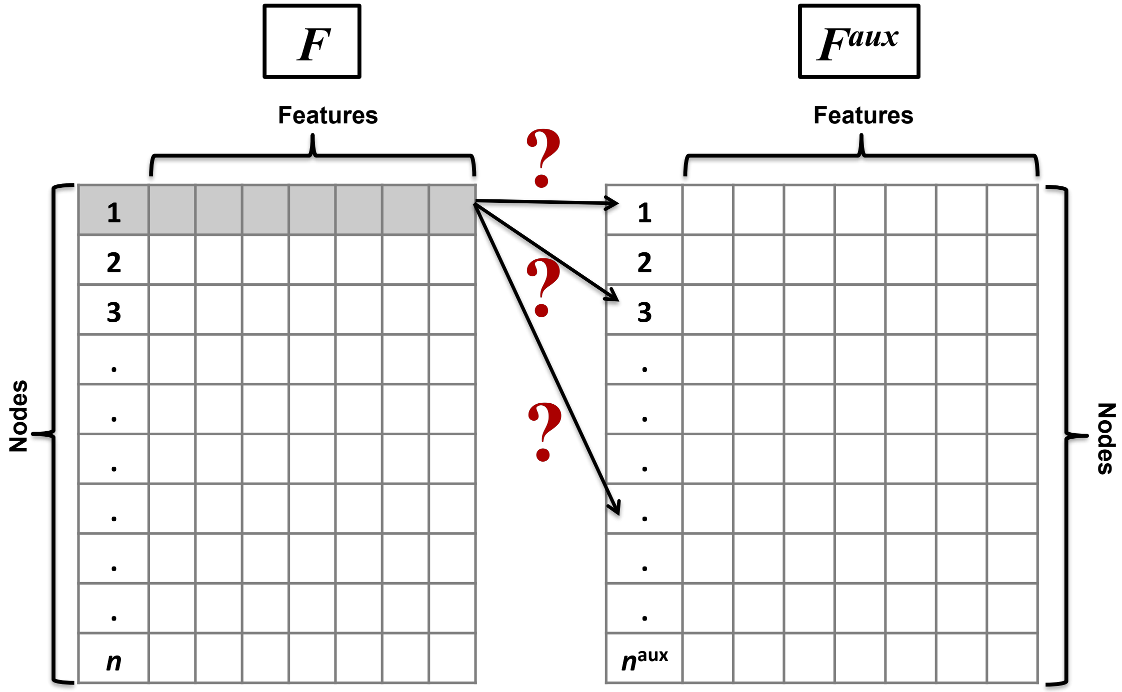
Nodes

Local and Egonet based Features

Finds G^{aux} & Computes F^{aux}

Problem Setting

- Adversary's algorithm
  - For each node v in F,
    - automatically find the smallest set of nodes in F^{aux} that are most likely to be v



The size of the "re-ID" set varies from node to node.

Why is this interesting?

- Defines *threatening privacy* as a *relative* concept
- R_i = the smallest set of known individuals that is most likely to include an anonymized individual *i*
- If $|R_i| << |R_j|$ then individual *i* is more "distinguishable" than individual *j*
- Example
  - In DBLP co-authorship graphs, we observe super-stars having smaller *R* sets than recent graduates

How should we evaluate this slightly different problem setting?

- *Recall*: Is node *v*'s match present in the matched cluster?

$$Recall(v, \mathcal{G}^{aux}) = \begin{cases} 1, & \text{if } v \in C_j^{i,aux} \in \mathcal{G}^{aux} \\ 0, & \text{otherwise.} \end{cases}$$

How should we evaluate this slightly different problem setting?

- *Recall*: Is node *v*'s match present in the matched cluster?

$$Recall(v, \mathcal{G}^{aux}) = \begin{cases} 1, & \text{if } v \in C_j^{i,aux} \in \mathcal{G}^{aux} \\ 0, & \text{otherwise.} \end{cases}$$

- *Precision*: How much of *v*'s uncertainty was reduced?

$$Precision(v, \mathcal{G}^{aux}) = \begin{cases} 1 - \frac{|C_j^{i,aux}|}{n^{aux}}, & \text{if } v \in C_j^{i,aux} \in \mathcal{G}^{aux} \\ 0, & \text{otherwise.} \end{cases}$$

$$= \left(1 - \frac{|C_j^{i,aux}|}{n^{aux}}\right) Recall(v, G^{aux})$$

Evaluation Metrics

- *Recall*: Is node *v*'s match present in the matched cluster?

$$Recall(v, \mathcal{G}^{aux}) = \begin{cases} 1, & \text{if } v \in C_j^{i,aux} \in \mathcal{G}^{aux} \\ 0, & \text{otherwise.} \end{cases}$$

- *Precision*: How much of *v*'s uncertainty was reduced?

$$Precision(v, \mathcal{G}^{aux}) = \begin{cases} 1 - \frac{|C_j^{i,aux}|}{n^{aux}}, & \text{if } v \in C_j^{i,aux} \in \mathcal{G}^{aux} \\ 0, & \text{otherwise.} \end{cases}$$

$$= \left(1 - \frac{|C_j^{i,aux}|}{n^{aux}}\right) Recall(v, G^{aux})$$

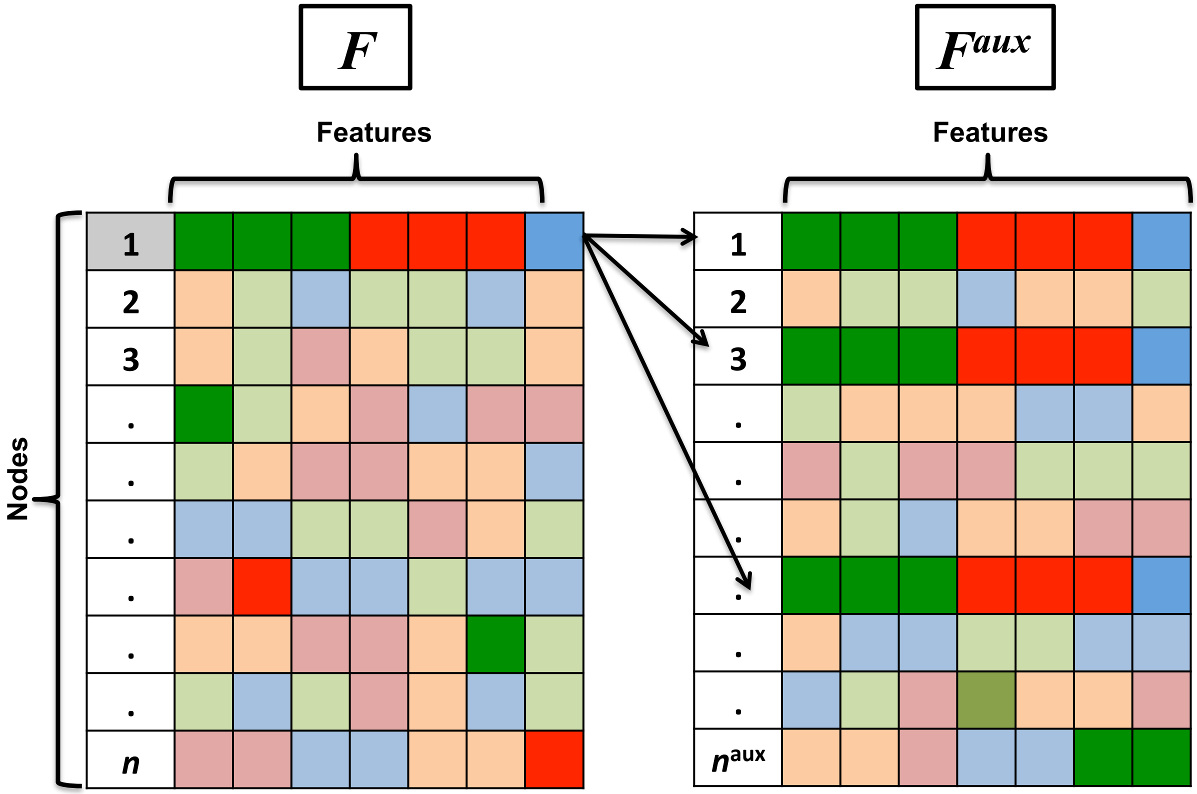
Objective: Maximize Precision to narrow down the set of likely matches for each node in F

Challenges

1. No link structure

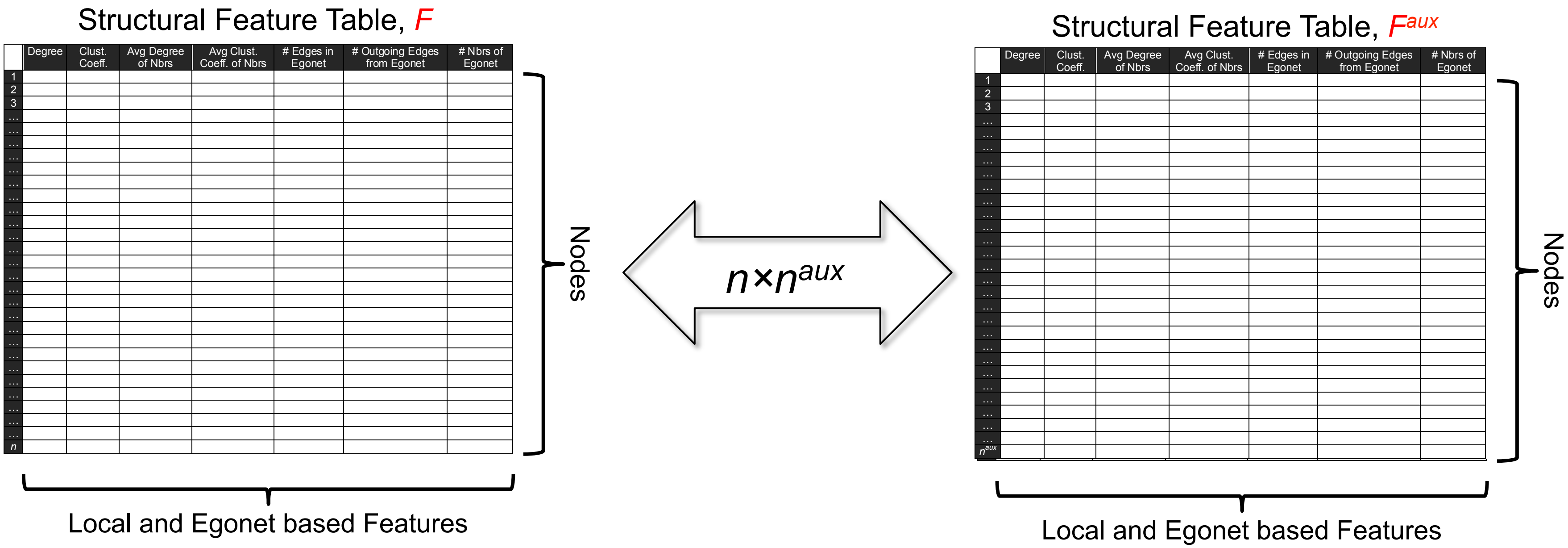
Challenges

1. No link structure
2. Nodes have many lookalikes



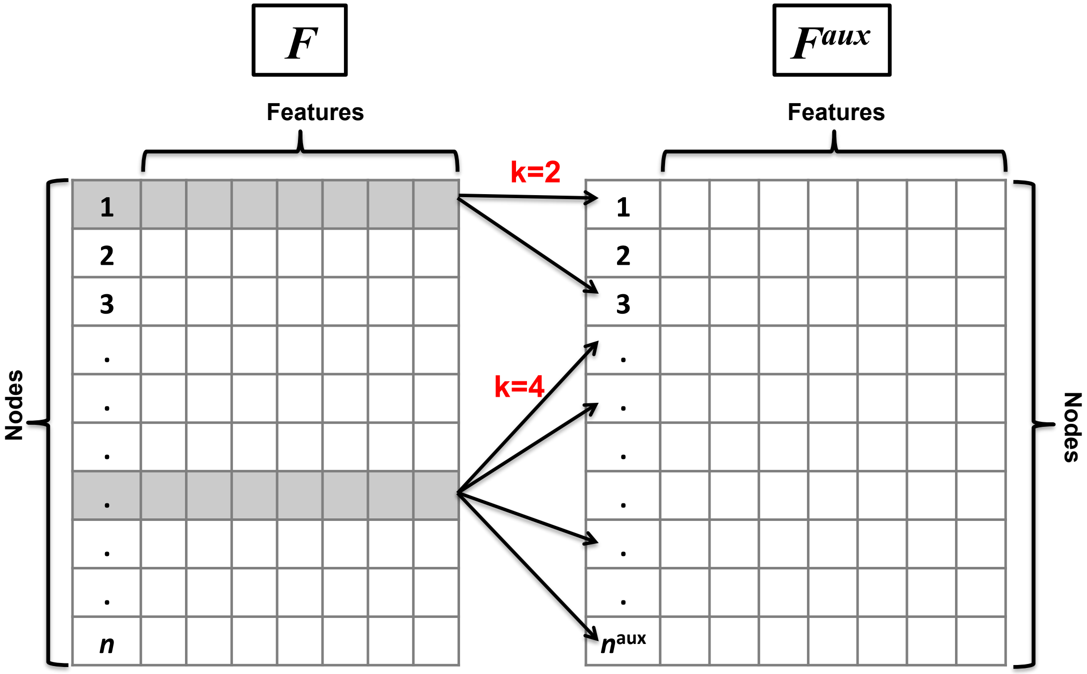
Challenges

1. No link structure
2. Nodes have many Lookalikes
3. Trivial $n \times n^{aux}$ comparisons not feasible



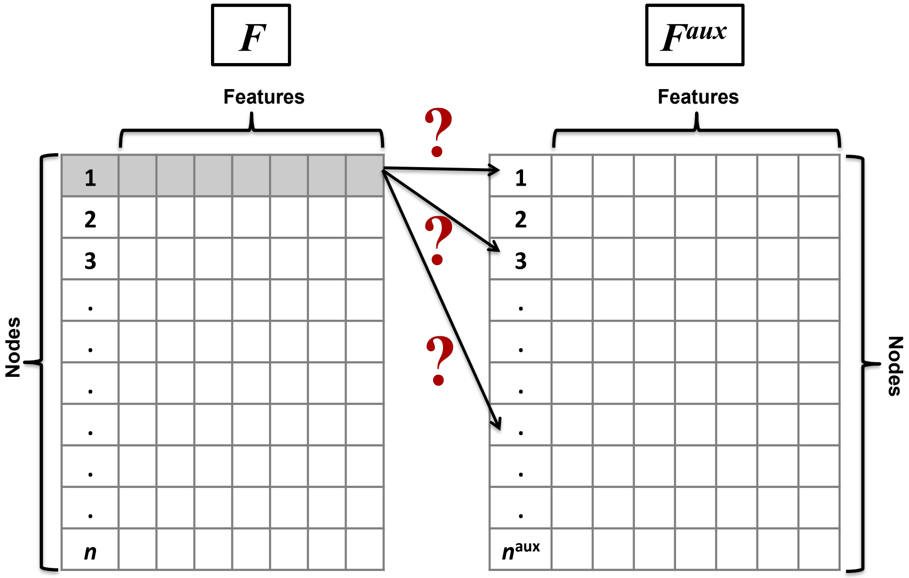
Challenges

1. No link structure
2. Nodes have many Lookalikes
3. Trivial n^2 comparisons not feasible
4. No *k* given so need to automatically find the most likely *k* nodes



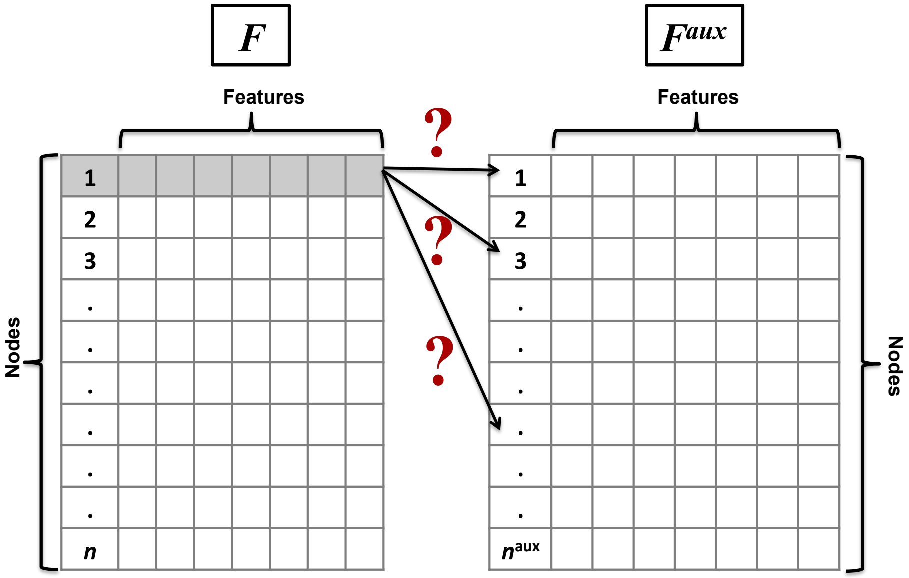
RRID^{+}: Cluster, Match, Repeat

- **Goal**
  - Narrow down the set of likely matches for each node in *F*
- **Approach**
  - Recursively match sets of similar nodes in *F* with sets of nodes in F^{aux}



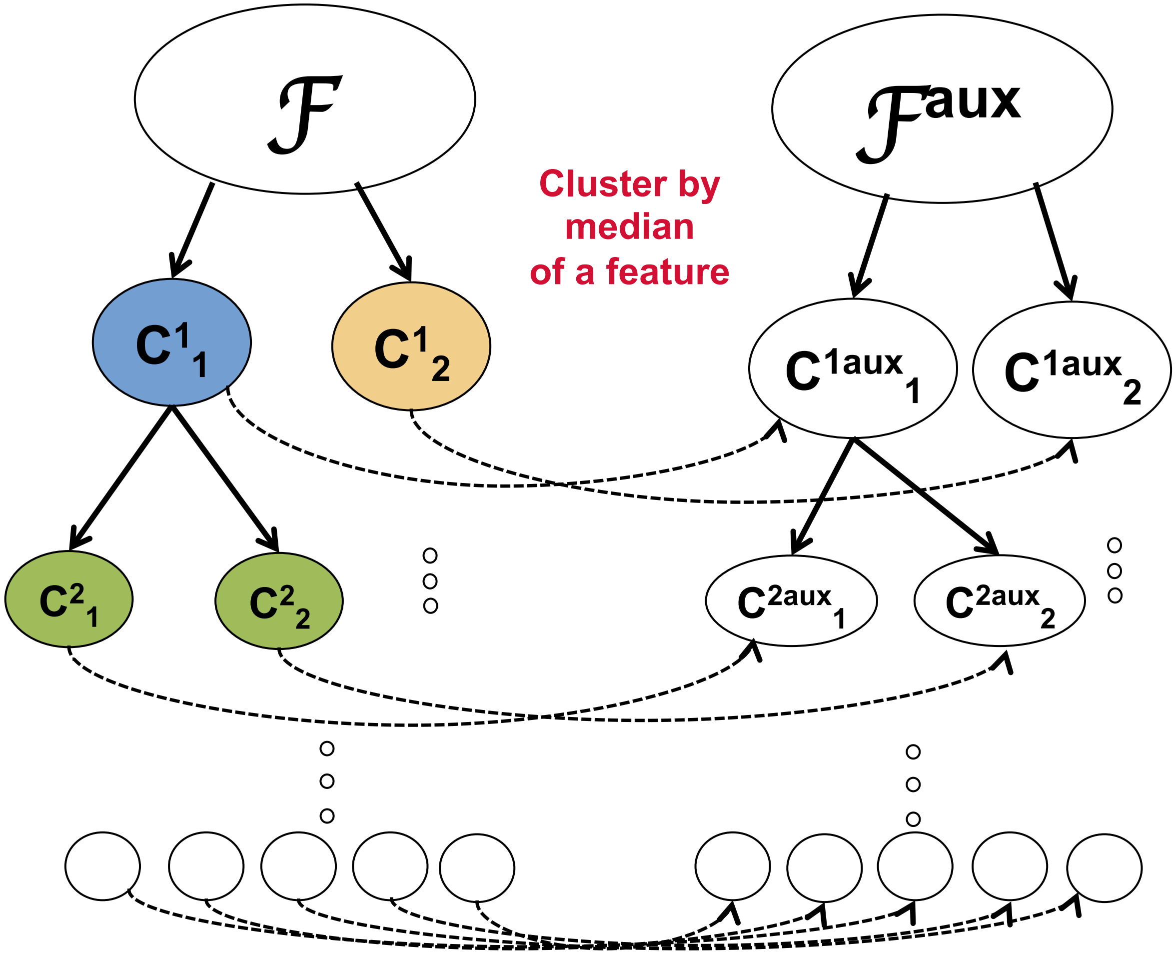
RRID$^+$: Cluster, Match, Repeat

- **Goal**
  - Narrow down the set of likely matches for each node in F
- **Approach**
  - Recursively match sets of similar nodes in F with sets of nodes in F^{aux}



- **Assumption**
  - If a node $v \in G$ has a corresponding node $v^{aux} \in G^{aux}$, then v and v^{aux} are structurally similar

RRID$^+$: Cluster, Match, Repeat

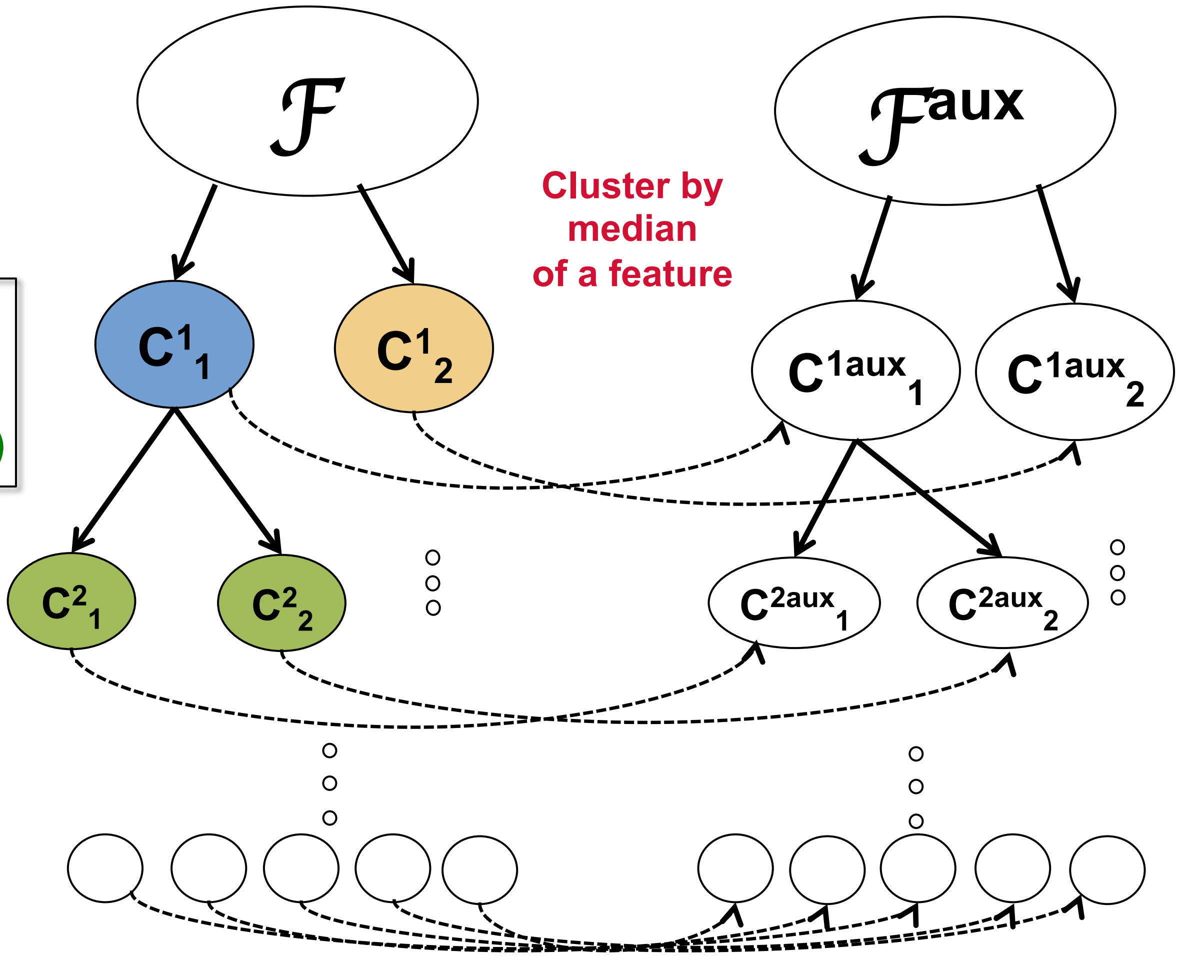


RRID$^+$: Cluster, Match, Repeat

old Precision (C^1_1)

$\leq$

Precision (C^2_1) +Precision (C^2_2)



Cluster by median of a feature

RRID$^+$: Cluster, Match, Repeat

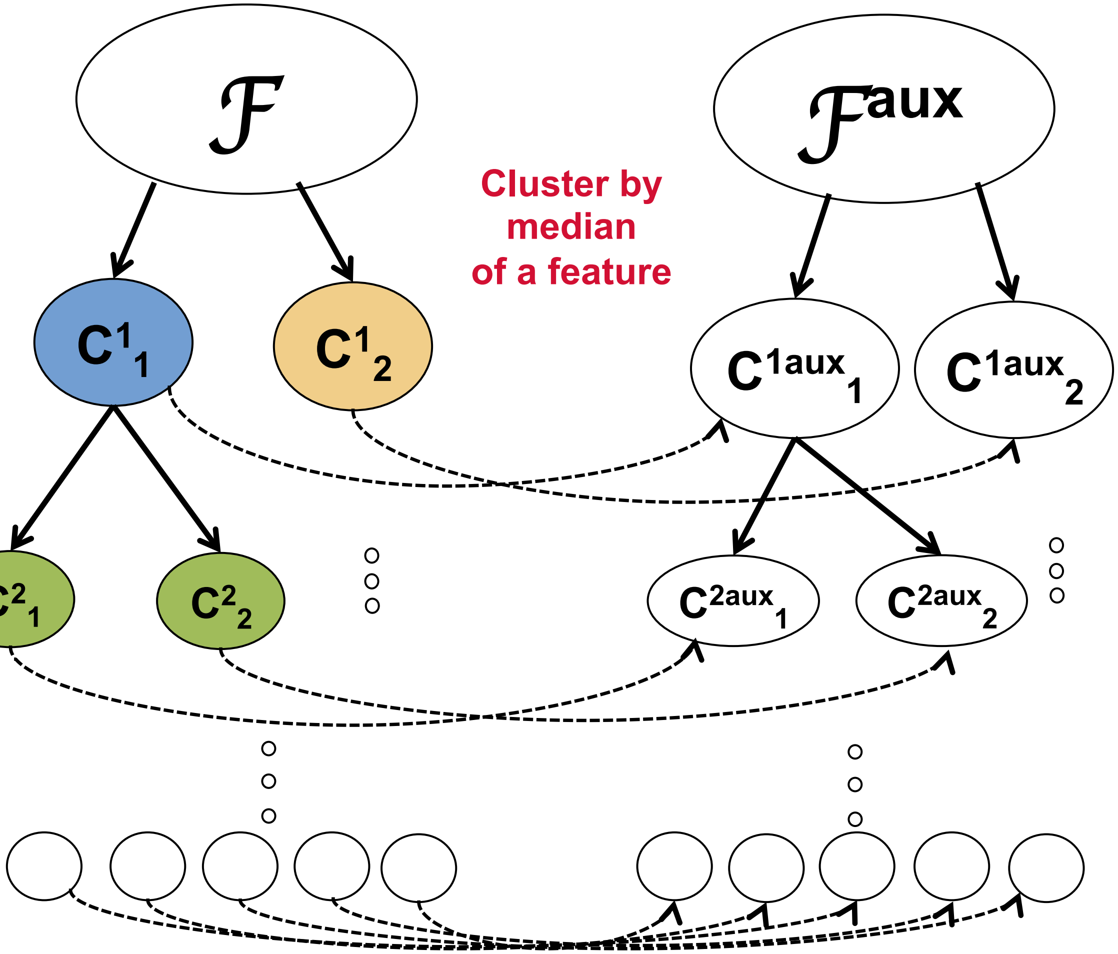
Precision (C^1_1)

≤

Precision (C^2_1) + Precision (C^2_2)

Yes:
Continue on C^2_1 C^2_2

No:
Continue DFS on C^1_2



Cluster by median of a feature

Runtime complexity: O(n log n), n = # of nodes.

Experiments: Graph Data

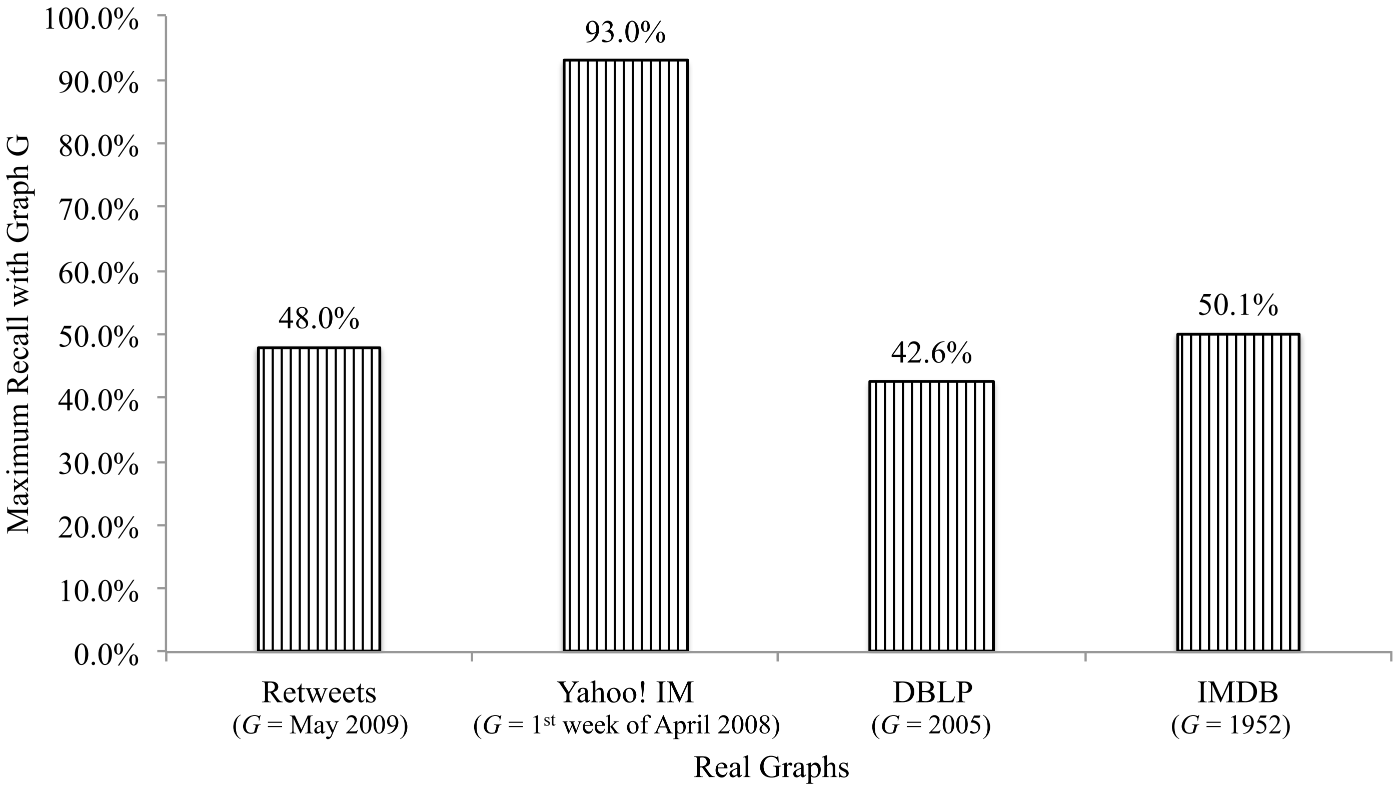
Real Graphs	**Avg. Number of Nodes**	**Avg. Number of Edges**
Twitter Retweet Monthly	64,072	81,906
Yahoo! IM Weekly	84,992	261,167
DBLP Co-authorship Yearly	2,045	4,024
IMDB Collaboration Yearly	10,887	236,132
Synthetic Graphs	**Number of Nodes**	**Number of Edges**
Barabási-Albert Graph	5,000	124,375
Erdös-Rényi Random Graph	5,000	125,021
Forest Fire Graph	5,000	116,135
Watts-Strogatz Graph	5,000	125,000

Auxiliary Graphs

- Various noise models generate auxiliary graphs
  1. Edge rewiring while keeping degree distribution the same
  2. Edge deletion
  3. Node deletion
- Noise parameter tested at 5%, 10%, 20%

Real Graphs	**Avg. Number of Nodes**	**Avg. Number of Edges**
Twitter Retweet Monthly	64,072	81,906
Yahoo! IM Weekly	84,992	261,167
DBLP Co-authorship Yearly	2,045	4,024
IMDB Collaboration Yearly	10,887	236,132
Synthetic Graphs	**Number of Nodes**	**Number of Edges**
Barabási-Albert Graph	5,000	124,375
Erdös-Rényi Random Graph	5,000	125,021
Forest Fire Graph	5,000	116,135
Watts-Strogatz Graph	5,000	125,000

Maximum Recall Varied In Real Graph Pairs



$$\text{Maximum } Recall \text{ (in real graphs)} = \frac{\text{number of overlapping nodes}}{\text{number of nodes in } G}$$

Comparison with Baselines

Average F1 Score

	Real Graph + Real Noise	**Real Graph + Synthetic Noise**	**Synthetic Graph + Synthetic Noise**
RRID$^+$ (Our method)	0.543	0.78	0.74
Paired hierarchical random clustering	0.30	0.35	0.38
K-means clustering	0.21	0.36	0.36
Random clustering	0.31	0.28	0.30

$$\overline{\text{F1 Score}} = \frac{2 \times \overline{\text{Recall}} \times \overline{\text{Precision on Recalled Nodes}}}{\overline{\text{Recall}} + \overline{\text{Precision on Recalled Nodes}}}$$

Comparison with Baselines

Average F1 Score
(Recall; Precision on Recalled Nodes)

	Real Graph + Real Noise	**Real Graph + Synthetic Noise**	**Synthetic Graph + Synthetic Noise**
RRID$^+$ (Our method)	0.543 (R = 0.44; P= 0.71)	0.78 (R = 0.89; P = 0.70)	0.74 (R = 0.80; P = 0.68)
Paired hierarchical random clustering	0.30 (R = 0.19; P = 0.74)	0.35 (R = 0.23; P = 0.71)	0.38 (R = 0.26; P = 0.70)
K-means clustering	0.21 (R = 0.12; P = 0.74)	0.36 (R = 0.25; P = 0.66)	0.36 (R = 0.25; P = 0.66)
Random clustering	0.31 (R = 0.21 P = 0.61)	0.28 (R = 0.18; P = 0.63)	0.30 (R = 0.19; P = 0.68)

Comparison with KD-Tree and LSH

- KD-tree and LSH require *k*, the size of cluster to be specified *a priori*
- In KD-tree and LSH, number of queries is *N* (= size of graph)

Comparison with KD-Tree and LSH

- KD-tree and LSH require *k*, the size of cluster to be specified *a priori*
- In KD-tree and LSH, number of queries is N (= size of graph)

Average F1 Score

	Real Graph + Real Noise	**Real Graph + Synthetic Noise**	**Synthetic Graph + Synthetic Noise**
RRID$^+$ (Our method)	0.54	0.78	0.74
KD-Tree$^+$	0.55	0.78	0.68
LSH$^+$	0.55	0.79	0.67

$$\overline{\text{F1 Score}} = \frac{2 \times \overline{\text{Recall}} \times \overline{\text{Precision on Recalled Nodes}}}{\overline{\text{Recall}} + \overline{\text{Precision on Recalled Nodes}}}$$

Comparison with KD-Tree and LSH

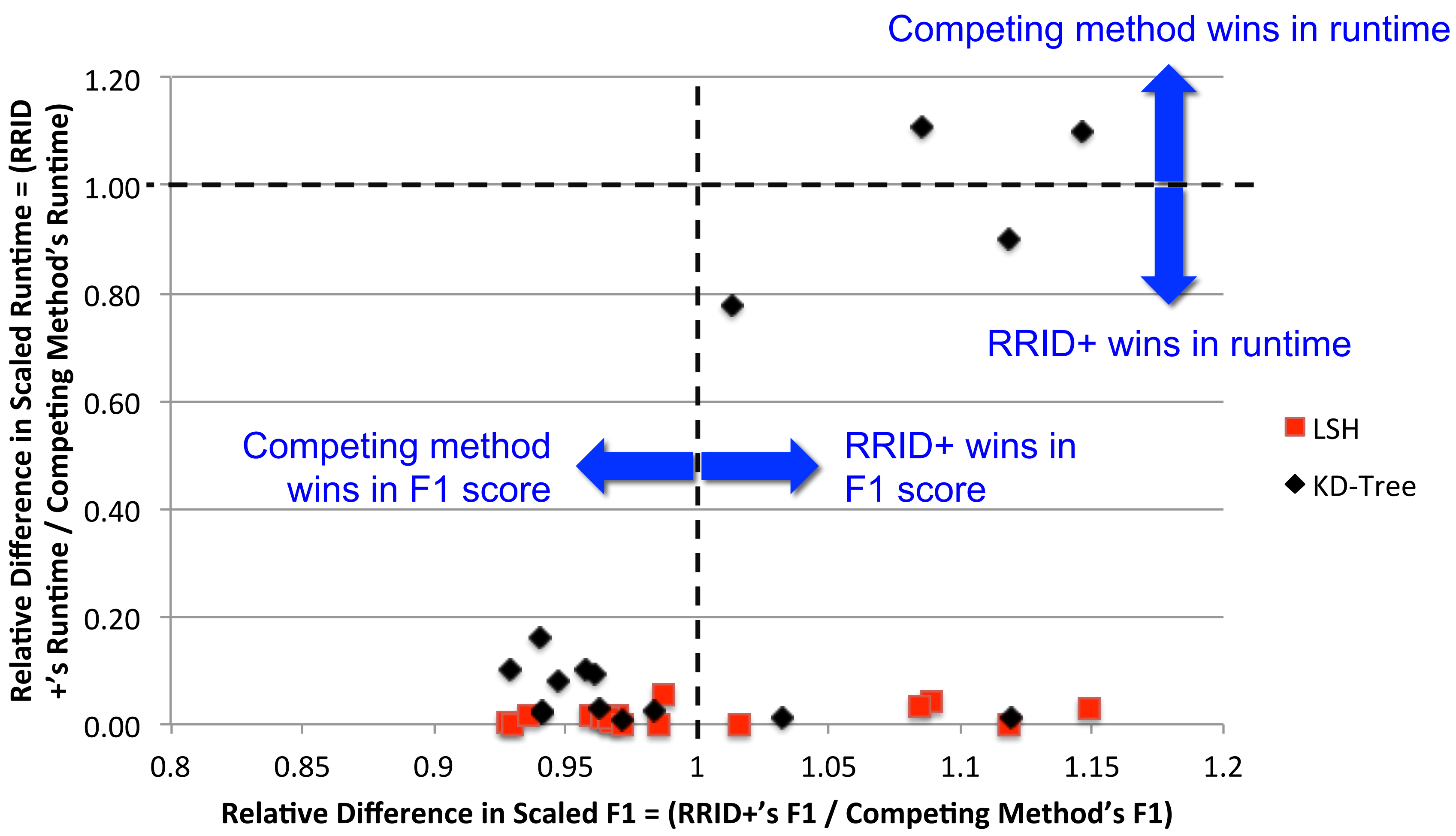
- KD-tree and LSH require *k*, the size of cluster to be specified *a priori*
- In KD-tree and LSH, number of queries is N, size of graph

Average F1 Score

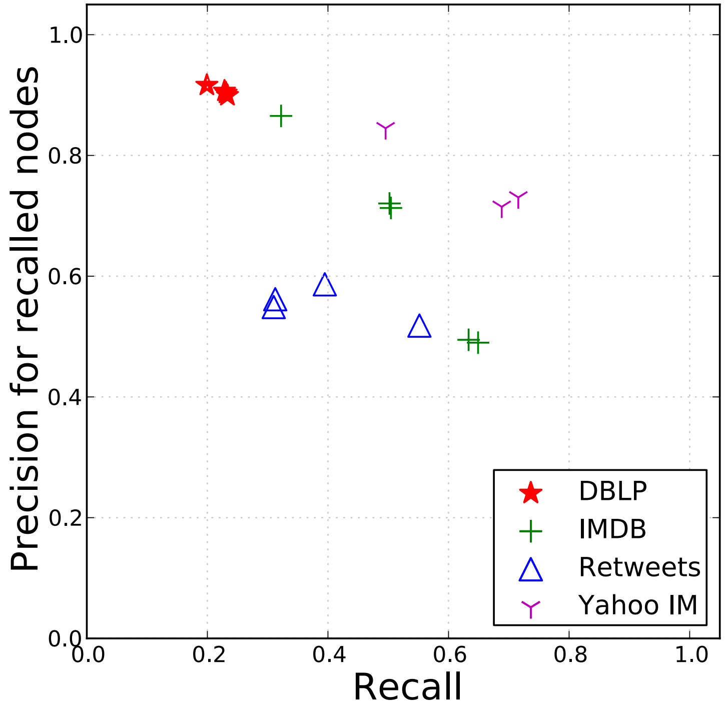
(Recall; Precision on Recalled Nodes)

	Real Graph + Real Noise	**Real Graph + Synthetic Noise**	**Synthetic Graph + Synthetic Noise**
RRID$^+$ (Our method)	0.54 (R = 0.44; P = 0.71)	0.78 (R = 0.89; P = 0.70)	0.74 (R = 0.80; P = 0.68)
KD-Tree$^+$	0.55 (R = 0.45; P = 0.70)	0.78 (R = 0.90; P = 0.69)	0.68 (R = 0.69; R = 0.67)
LSH$^+$	0.55 (R = 0.45; P = 0.70)	0.79 (R = 0.90; P = 0.70)	0.67 (R = 0.68; R = 0.67)

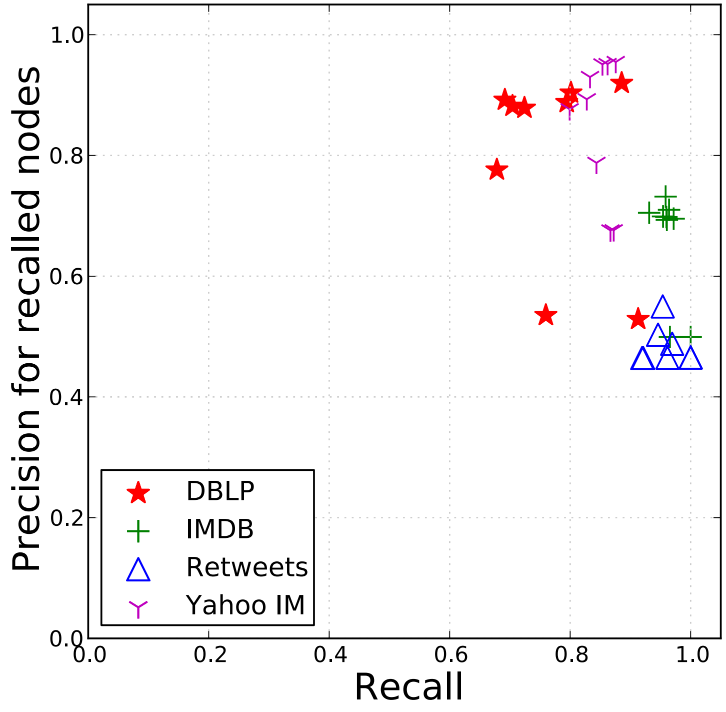
Runtime Performance vs. F1 Score of KD-Tree & LSH Relative to RRID^{+}



RRID+ on Real Graphs: Precision on Recalled Nodes *vs.* Recall



Real noise



Synthetic Noise added to G^{aux}

$$\text{Precision of recalled nodes}(\mathcal{G},\mathcal{G}_{aux}) = \frac{\sum_{v\in RecalledNodes} Precision(v,\mathcal{G}_{aux})}{|RecalledNodes|}$$

$$\text{where } RecalledNodes = \{\forall v : Recall(v,G) = 1\}$$

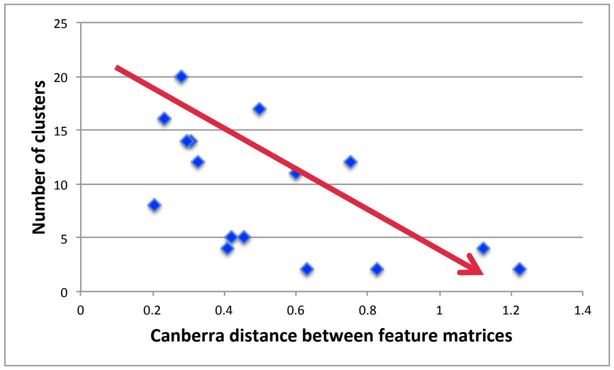
Insights into the Performance

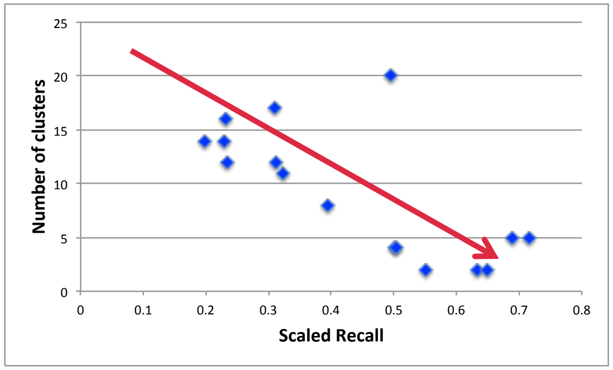
- As distance between feature matrices increases
  - Number of clusters decreases
    - Recall increases
    - Precision of recalled nodes decreases

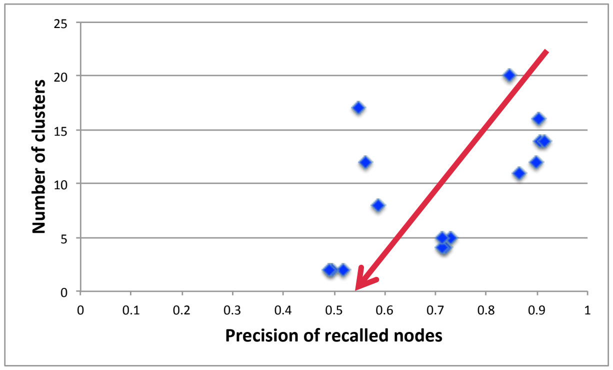
Insights into the Performance

- As distance between feature matrices increases
    - Number of clusters decreases
        - Recall increases
        - Precision of recalled nodes decreases

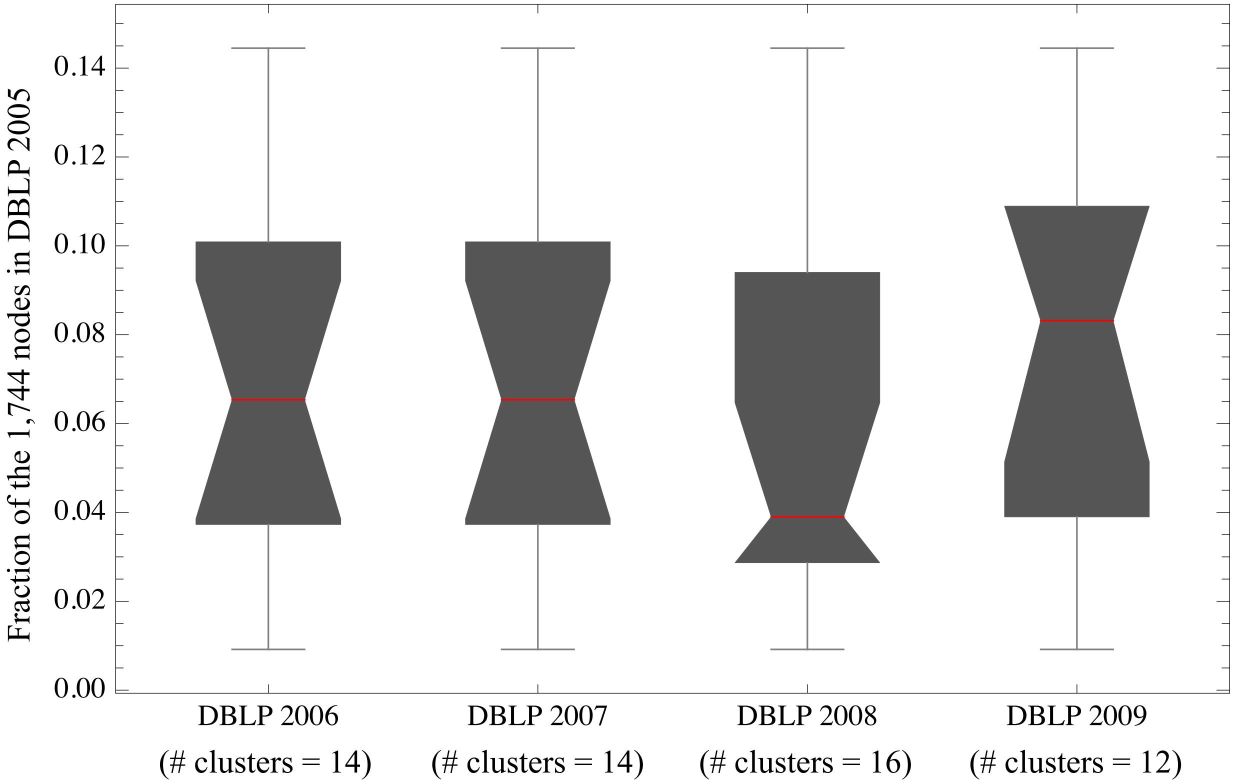
Real Graphs + Real Noise





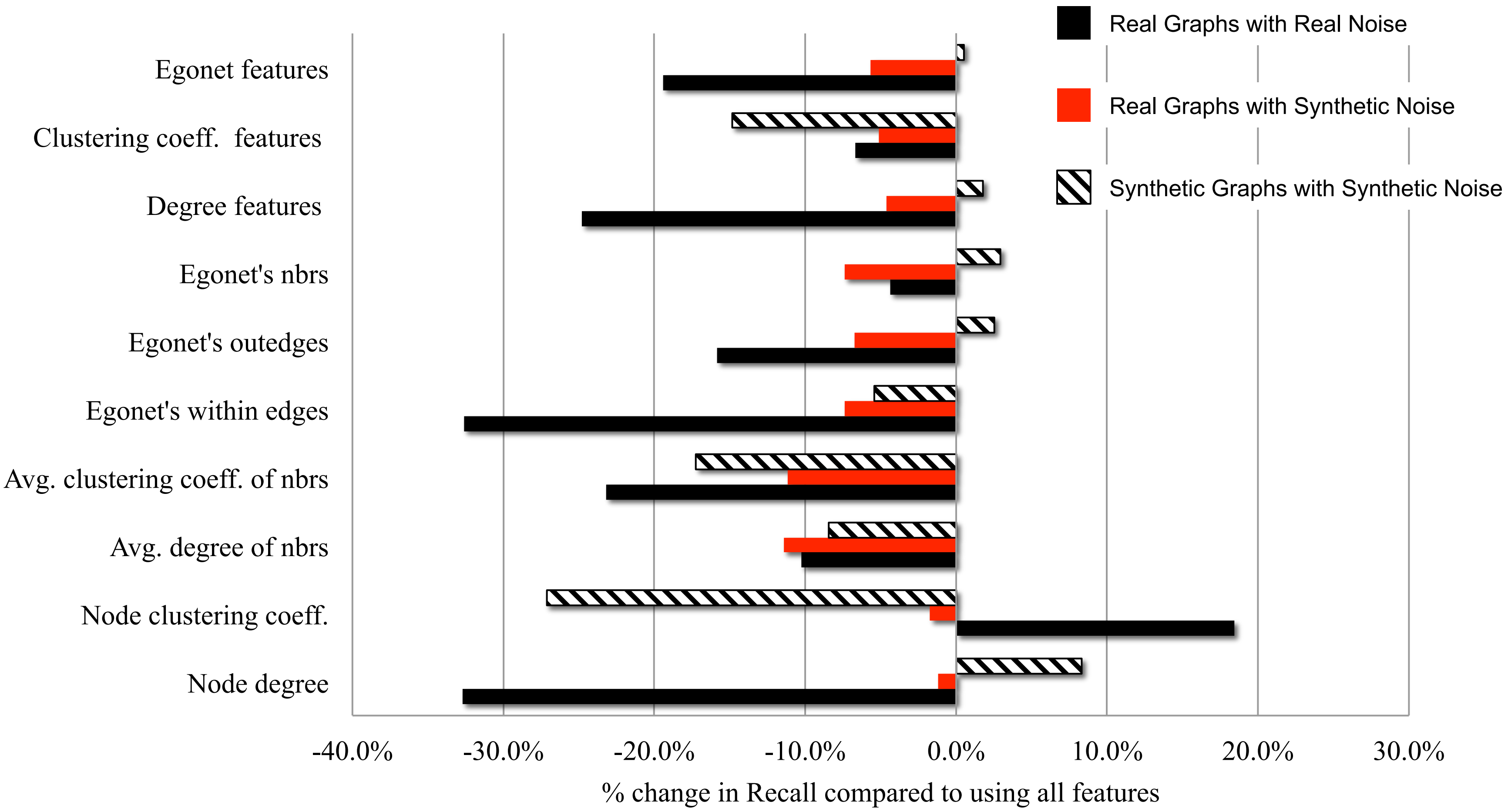


RRID$^+$ Outputs Varying Sized Clusters



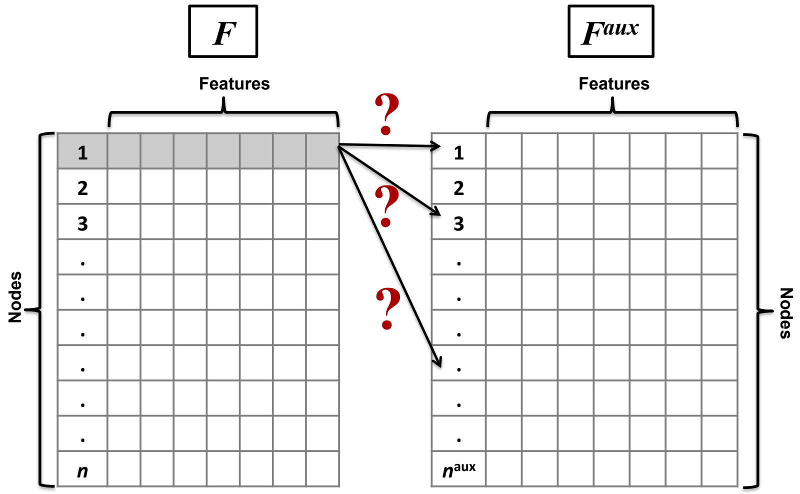
Individuals in smaller clusters are more ‘distinguishable’

Effects of Various Subsets of Structural Features on Recall



Recap Part 2: A Different View of Privacy

1. A new way of looking at the re-identification problem
2. Defining a threat to privacy as a relative concept
3. A novel *collective* solution
4. Performance on real graphs with real noise
   - Average *Recall* = 0.44
   - Average Precision on Recalled Nodes = 0.71
5. An examination of re-identification performance based on feature selection, cluster sizes, and runtime.



Future work: Quantifying noise in real social graphs

Summary

- Structural features and roles threaten privacy in social graphs
- Threats are w.r.t.
  - one-to-one mappings between nodes
  - personalized one-to-many mappings between nodes

Supported by NSF, LLNL, DTRA, DARPA, and IARPA.

Thank You! (http://eliassi.org)