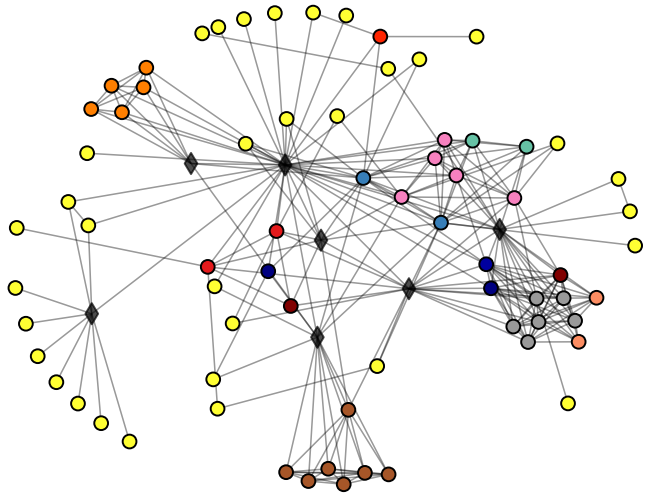
Role Discovery in Graphs using Global Features: Algorithms, Applications and a Novel Evaluation Strategy

Pratik Vinay Gupte (Netradyne Technology)
Balaraman Ravindran (IIT Madras)
Srinivasan Parthasarathy (The Ohio State University)

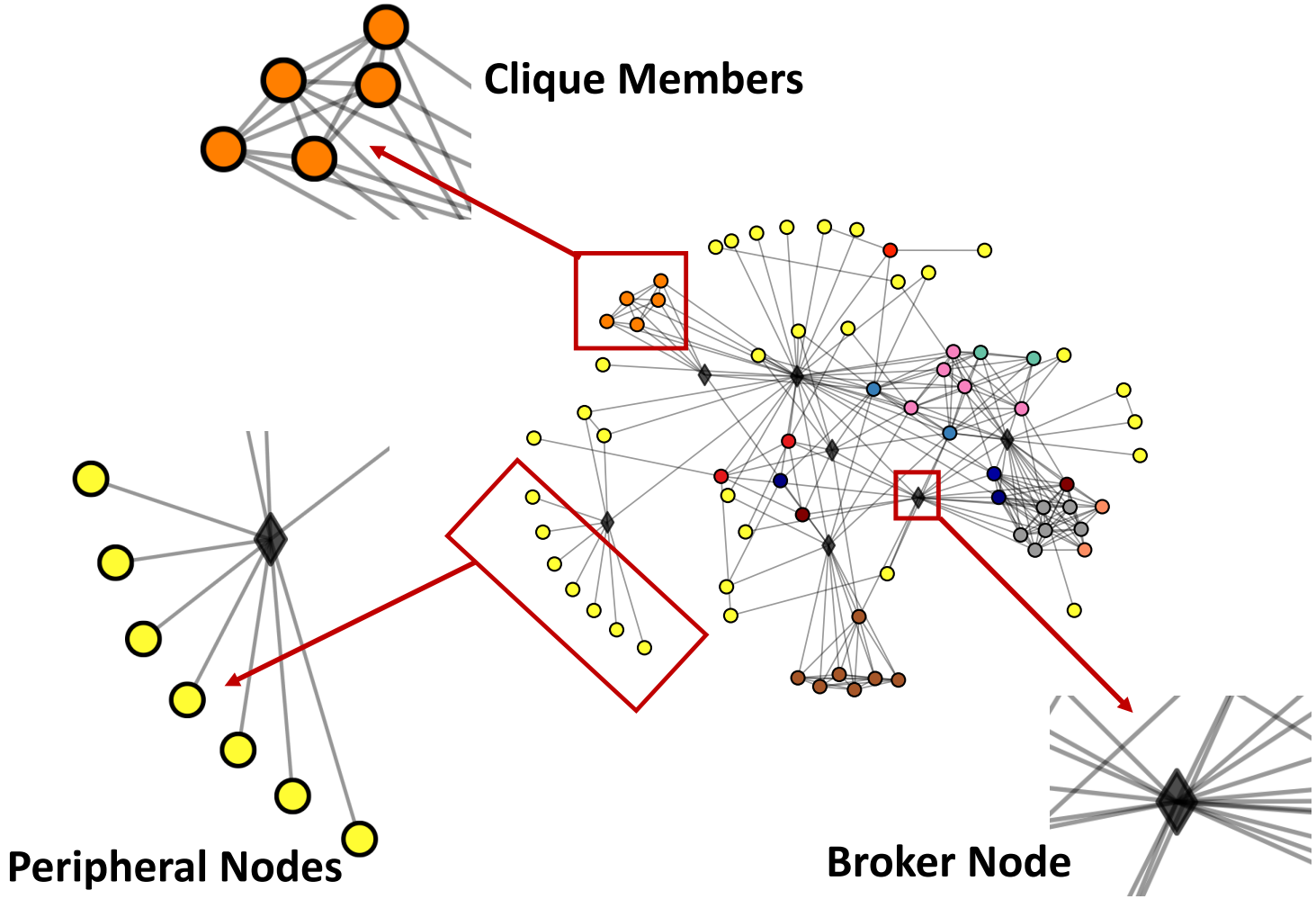
April 20, 2017

What are Roles?

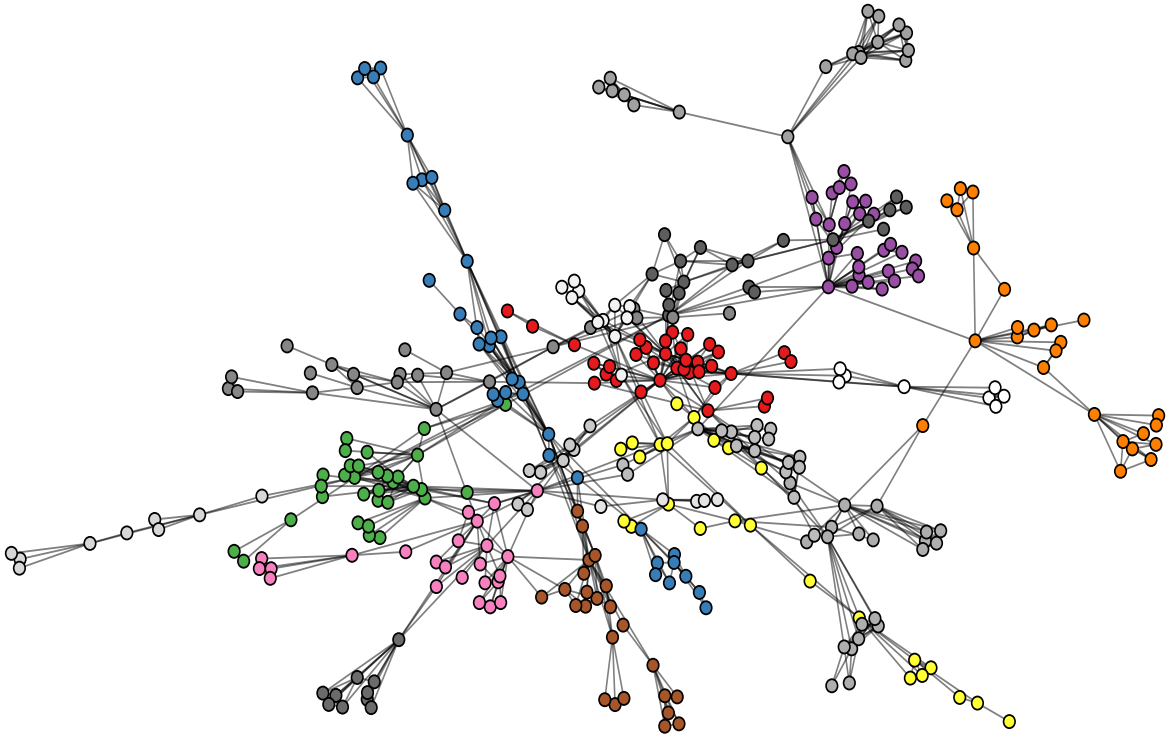
- Discover actors who have similar structural signatures
- “Functions” of nodes in the network
  - Roles of team members in a football team
  - Roles of people in a University
  - Roles of countries in trade



Example Structural Roles



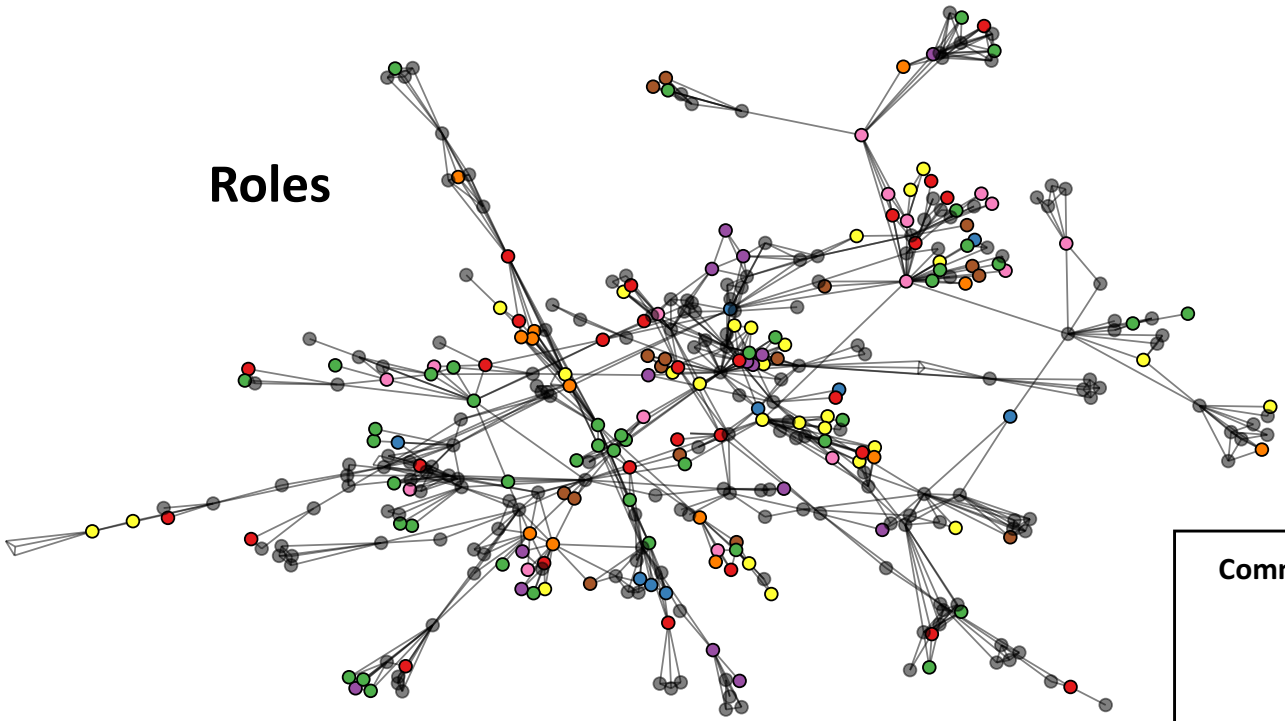
Roles and Communities

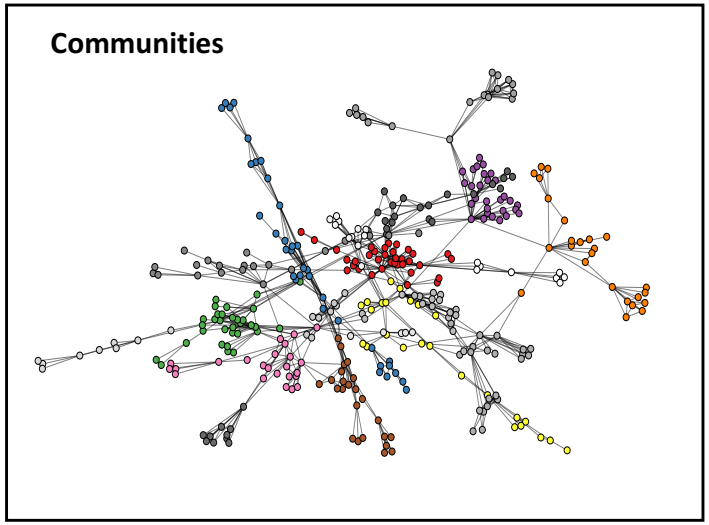


Communities in Network Sciences Co-Authorships

Blondel *et. al.*, “Fast unfolding of communities in large networks”

Roles and Communities





Roles and Communities

- Roles
  - Faculty
  - Research Scholars
  - Staff
  - ...
- Communities
  - AI Lab
  - Architecture Lab
  - Networks Lab
  - ...

- Roles: Structural Signatures
- Communities: Cohesive Subgroups
- **Both are Complementary!**

Applications of Roles

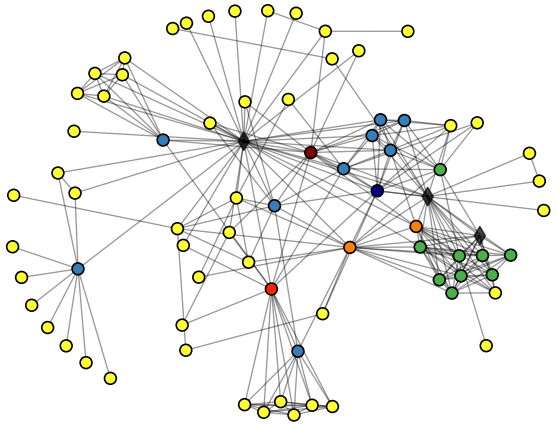
Application	**Description**
Query/Search	Given a query node, find the structurally most similar node/nodes
Network Sampling	Finding a representative smaller graph of a larger network
Classification	Role based features in learning algorithms
Anomaly Detection	Rapid role changes in temporal networks signifies anomalous behaviour

Key Challenges in Role Discovery

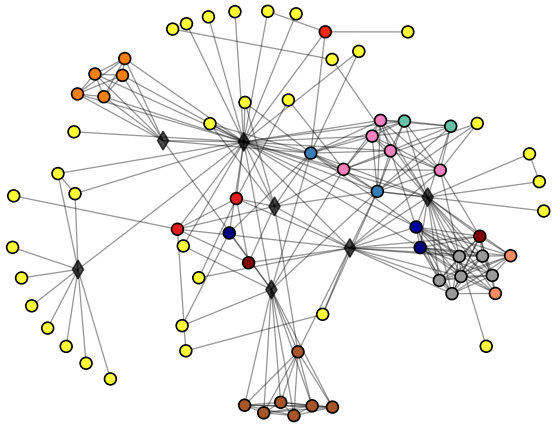
Task	**Challenge**
Characterization of graph structure	Graph represents a system, capturing complete structural view is non-trivial
Validation of role discovery algorithms	Absence of role labeled ground-truth datasets

RIDεRs: Role Identification and Discovery using ε-equitable Refinements

- We combine multiple ε-Equitable Refinements (εER) of a graph into a single node-feature representation
  - εER captures the complete structural view of the graph
  - εER role discovery algorithm depends on the choice of ε
- Perform role discovery using Non-negative Matrix Factorization



$\varepsilon = 6$



$\varepsilon = 2$

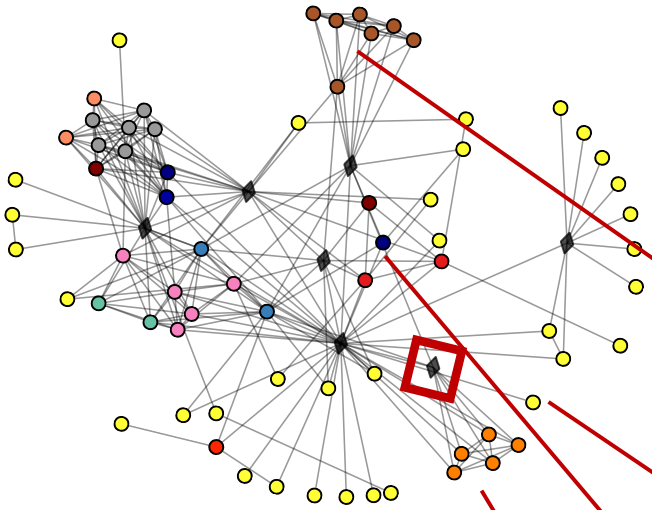
RIDεRs: Node Features

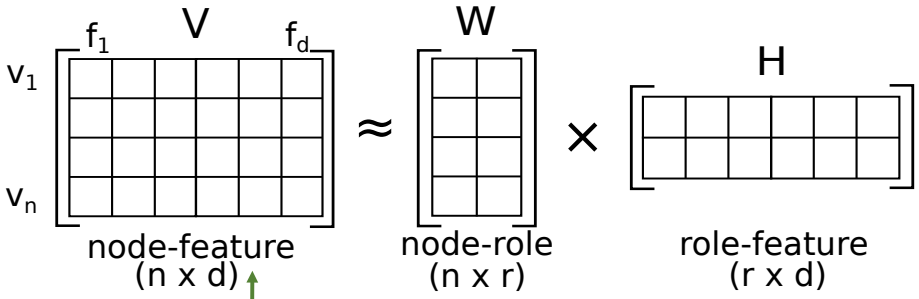
- Given, G, the vertex set $V = \{v_1, v_2, ..., v_n\}$ and its ε-equitable partition $\pi = \{C_1, C_2, ..., C_k\}$
- For an ε
  - Create a feature vector $\mathbf{f_j} = [f_{1j}, f_{2j}, ..., f_{kj}]$ corresponding to each role/cell C_j
  - For each node v_i, the value of a feature f_{ij} is calculated as:

$$\begin{aligned} &\forall v_i \in V \text{ and } \forall C_j \in \pi, \\ &f_{ij} = |adj\_list(v_i) \cap C_j|, \end{aligned} \tag{1}$$

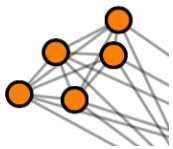
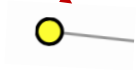
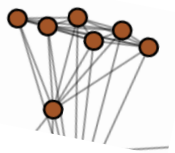
  - where,
    $adj\_list(v)$ is the adjacency list of vertex v
- Graph based node feature matrix is $\mathbf{V}_\varepsilon = [\mathbf{f_1^T}|\mathbf{f_2^T}|...|\mathbf{f_k^T}]$

RIDεRs: Node Features and NMF





Non-negative Matrix Factorization

Role ->					...
Node-Feature Value	**5**	**1**	**1**	**0**	

Node features for 2-equitable refinement ($\varepsilon = 2$)

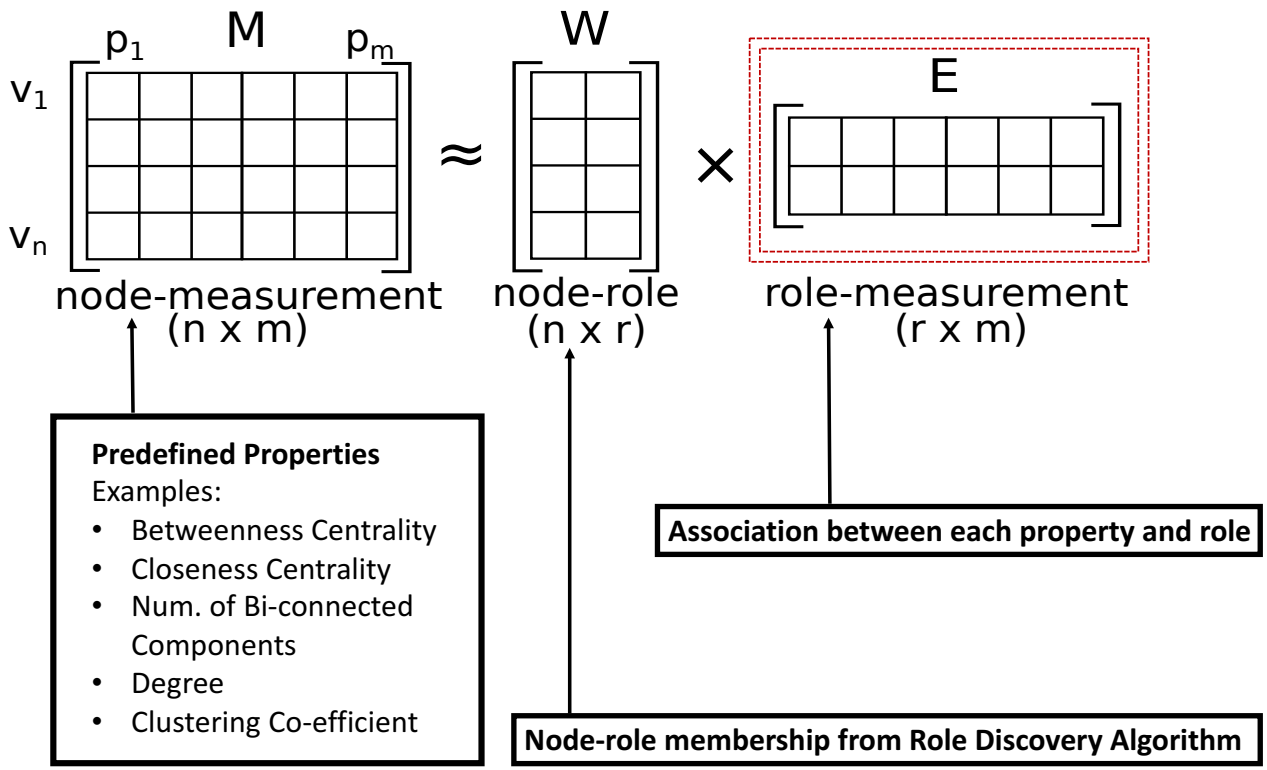
Novel Evaluation Framework: NodeSense[♯]

Given,

- $n \times r$ node-role matrix $\mathbf{W}$
- $n \times m$ matrix of node measurements $\mathbf{M}$

Estimate,

- $r \times m$ matrix $\mathbf{E}$, the role contribution to each measurement
- $\mathbf{N}_{ij} = \frac{\mathbf{E}_{ij}}{\mathbf{E}_i^d}$



[♯]Henderson *et. al.*, "RolX: Structural Role Extraction & Mining in Large Graphs," KDD 2012, pp. 1231 – 1239.

Novel Evaluation Framework: RandomSense

- Generate a *random* node-role assignment $\mathbf{W^r}$ and compute the corresponding *NodeSense*.
- For a role i and measurement j, $|\mathbf{N^r}_{ij} - \mathbf{N}_{ij}|$ is the absolute deviation from the baseline
- The *mean of the absolute deviations from the random role assignments* for a node measurement j can be computed as follows:

$$\frac{1}{r}\|\mathbf{N^r}_{\bullet j} - \mathbf{N}_{\bullet j}\|_1 \tag{2}$$

where,
r is the number of roles, $\mathbf{N^r}_{\bullet j}$ is the *RandomSense* vector of the j^{th} property and $\mathbf{N}_{\bullet j}$ is the *NodeSense* vector of the j^{th} property of the role discovery method being evaluated

Novel Evaluation Framework: Two Baselines

- Power-law random role assignments
- Uniform random role assignments

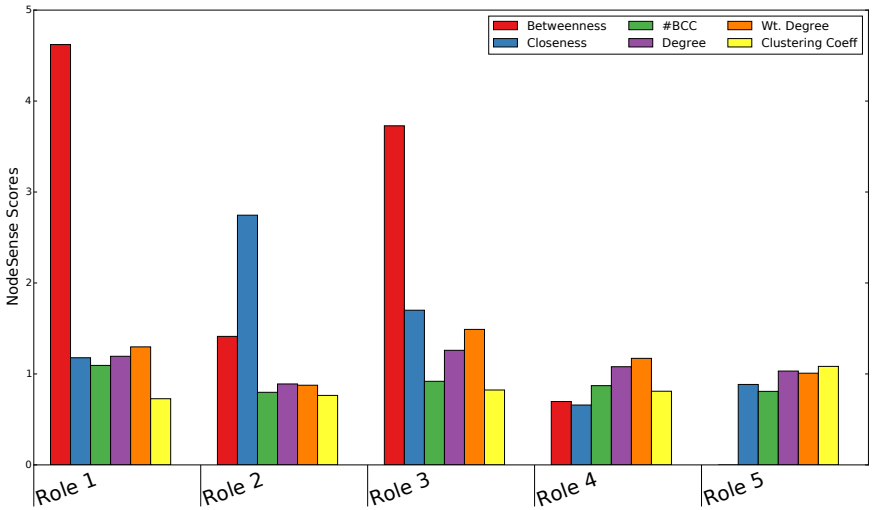
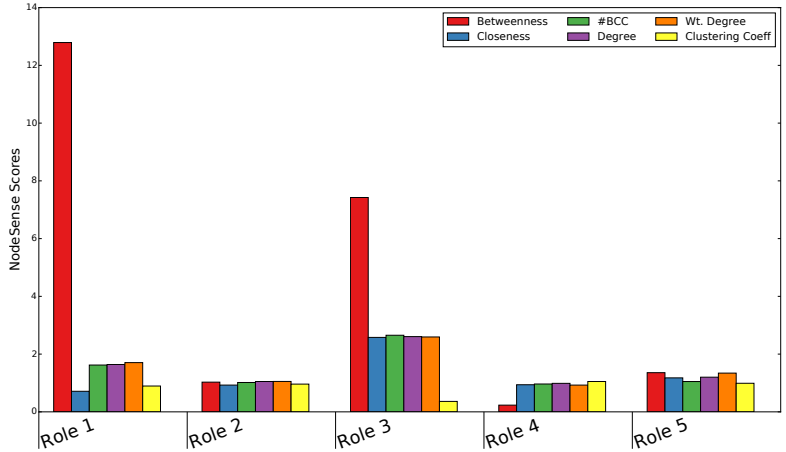
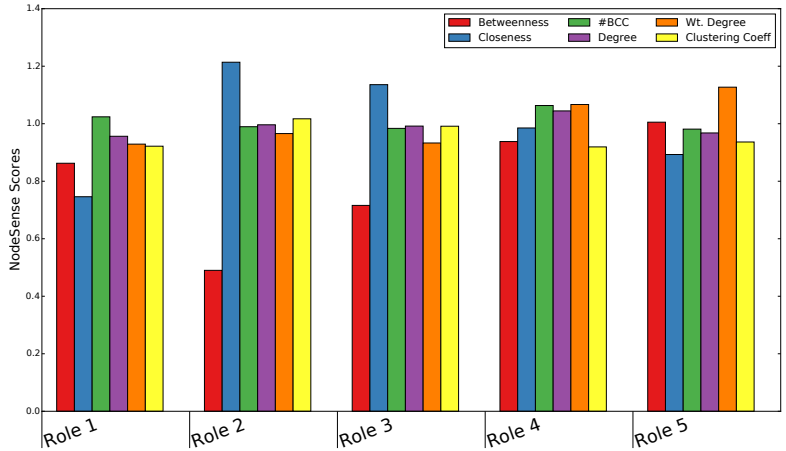


Figure: NodeSense on the ICDM Co-Authorship Network (Year 2005-10)



(a) RandomSense using *power-law* random distribution, $\gamma = 3$



(b) RandomSense using *uniform* random distribution

Novel Evaluation Framework: Results

	Betweenness	Closeness	#BCC	Ego_0_Deg	Ego_1_Deg	Ego_0_Wt	Ego_1_Wt	Degree	Wt_Deg	Clus_Coeff
RIDεRs-R	1.61 (0.04)	0.87 (0.02)	0.71 (0.0)	0.97 (0.01)	0.98 (0.02)	1.01 (0.02)	1.02 (0.02)	0.92 (0.01)	0.99 (0.01)	0.7 (0.0)
RIDεRs	2.04 (0.06)	0.71 (0.02)	0.67 (0.0)	0.8 (0.02)	0.85 (0.02)	0.84 (0.02)	0.91 (0.02)	0.75 (0.01)	0.84 (0.01)	0.64 (0.01)
RolX	1.56 (0.08)	0.66 (0.02)	0.72 (0.0)	0.73 (0.02)	0.66 (0.02)	0.74 (0.02)	0.72 (0.02)	0.7 (0.01)	0.75 (0.02)	0.76 (0.0)
GLRD-S	1.38 (0.06)	0.73 (0.02)	0.69 (0.0)	0.73 (0.02)	0.85 (0.02)	0.72 (0.02)	0.8 (0.02)	0.7 (0.01)	0.72 (0.02)	0.6 (0.0)
GLRD-D	1.41 (0.09)	0.52 (0.02)	0.41 (0.01)	0.49 (0.02)	0.46 (0.02)	0.41 (0.02)	0.46 (0.03)	0.58 (0.01)	0.64 (0.02)	0.42 (0.0)

(a) Baseline Evaluation on the CIKM Co-Authorship Network

	Betweenness	Closeness	#BCC	Ego_0_Deg	Ego_1_Deg	Ego_0_Wt	Ego_1_Wt	Degree	Wt_Deg	Clus_Coeff
RIDεRs-R	1.6 (0.09)	0.97 (0.02)	0.66 (0.01)	0.97 (0.02)	1.05 (0.03)	1.12 (0.02)	1.15 (0.03)	0.83 (0.01)	0.9 (0.02)	0.63 (0.0)
RIDεRs	2.65 (0.1)	0.83 (0.03)	0.7 (0.0)	0.98 (0.02)	1.16 (0.03)	1.06 (0.03)	1.24 (0.04)	0.81 (0.01)	0.93 (0.02)	0.7 (0.01)
RolX	2.12 (0.09)	0.89 (0.03)	0.76 (0.01)	0.92 (0.02)	1.18 (0.03)	1.02 (0.03)	1.18 (0.03)	0.76 (0.01)	0.88 (0.02)	0.74 (0.0)
GLRD-S	1.81 (0.14)	0.93 (0.03)	0.65 (0.01)	0.85 (0.03)	1.09 (0.03)	0.92 (0.03)	1.16 (0.03)	0.76 (0.02)	0.81 (0.02)	0.54 (0.0)
GLRD-D	2.28 (0.14)	0.51 (0.03)	0.29 (0.01)	0.47 (0.03)	0.81 (0.04)	0.48 (0.03)	0.83 (0.04)	0.49 (0.02)	0.58 (0.02)	0.32 (0.0)

(b) Baseline Evaluation on the ICDM Co-Authorship Network

	Betweenness	Closeness	#BCC	Ego_0_Deg	Ego_1_Deg	Ego_0_Wt	Ego_1_Wt	Degree	Wt_Deg	Clus_Coeff
RIDεRs-R	1.56 (0.06)	0.85 (0.02)	0.7 (0.01)	1.11 (0.02)	1.08 (0.02)	1.13 (0.02)	1.12 (0.02)	1.01 (0.02)	1.05 (0.02)	0.68 (0.0)
RIDεRs	2.21 (0.09)	0.7 (0.02)	0.7 (0.0)	1.0 (0.02)	1.05 (0.03)	1.06 (0.03)	1.08 (0.03)	0.86 (0.02)	0.98 (0.02)	0.65 (0.0)
RolX	1.78 (0.1)	0.67 (0.02)	0.71 (0.0)	0.94 (0.02)	0.92 (0.03)	0.94 (0.02)	0.95 (0.03)	0.84 (0.01)	0.91 (0.02)	0.71 (0.0)
GLRD-S	1.31 (0.13)	0.7 (0.02)	0.45 (0.0)	0.7 (0.02)	0.86 (0.03)	0.71 (0.02)	0.83 (0.03)	0.56 (0.02)	0.63 (0.03)	0.45 (0.0)
GLRD-D	1.73 (0.14)	0.61 (0.02)	0.46 (0.01)	0.6 (0.02)	0.86 (0.03)	0.63 (0.03)	0.89 (0.03)	0.59 (0.02)	0.7 (0.02)	0.42 (0.0)

(c) Baseline Evaluation on the KDD Co-Authorship Network

Applications: Top-k Structurally Similar Nodes

- Find the *top-k* nodes, which are structurally most similar to a query node
- Given, the node-role matrix $\mathbf{W}$ and a *query* node u. Find a node v which has the k^{th} closest node-role vector $\mathbf{W_{v\bullet}}$ to $\mathbf{W_{u\bullet}}$

CIKM (2005-13)	*Closest to* **Jiawei Han**	*Betweenness*	*Closeness*	*#BCC*	*Ego_1_Wt*	*Degree*	*Wt.Degree*	*Clus_Coeff*
RIDεRs	C. Lee Giles	0.03	0.00	0.01	0.01	0.00	0.05	0.00
RIDεRs-R	Jie Tang	0.02	0.00	0.00	0.01	0.03	0.01	0.00
RolX	Mounia Lalmas	0.05	0.00	0.01	0.03	0.03	0.03	0.00
GLRD-S	Philip S. Yu	0.06	0.00	0.03	0.01	0.00	0.00	0.00
GLRD-D	Wei Wang (0011)	0.18	0.00	0.02	0.06	0.08	0.08	0.01

Table: **Top-k Illustrative Example**.

Top-k Structurally Similar Nodes: Results

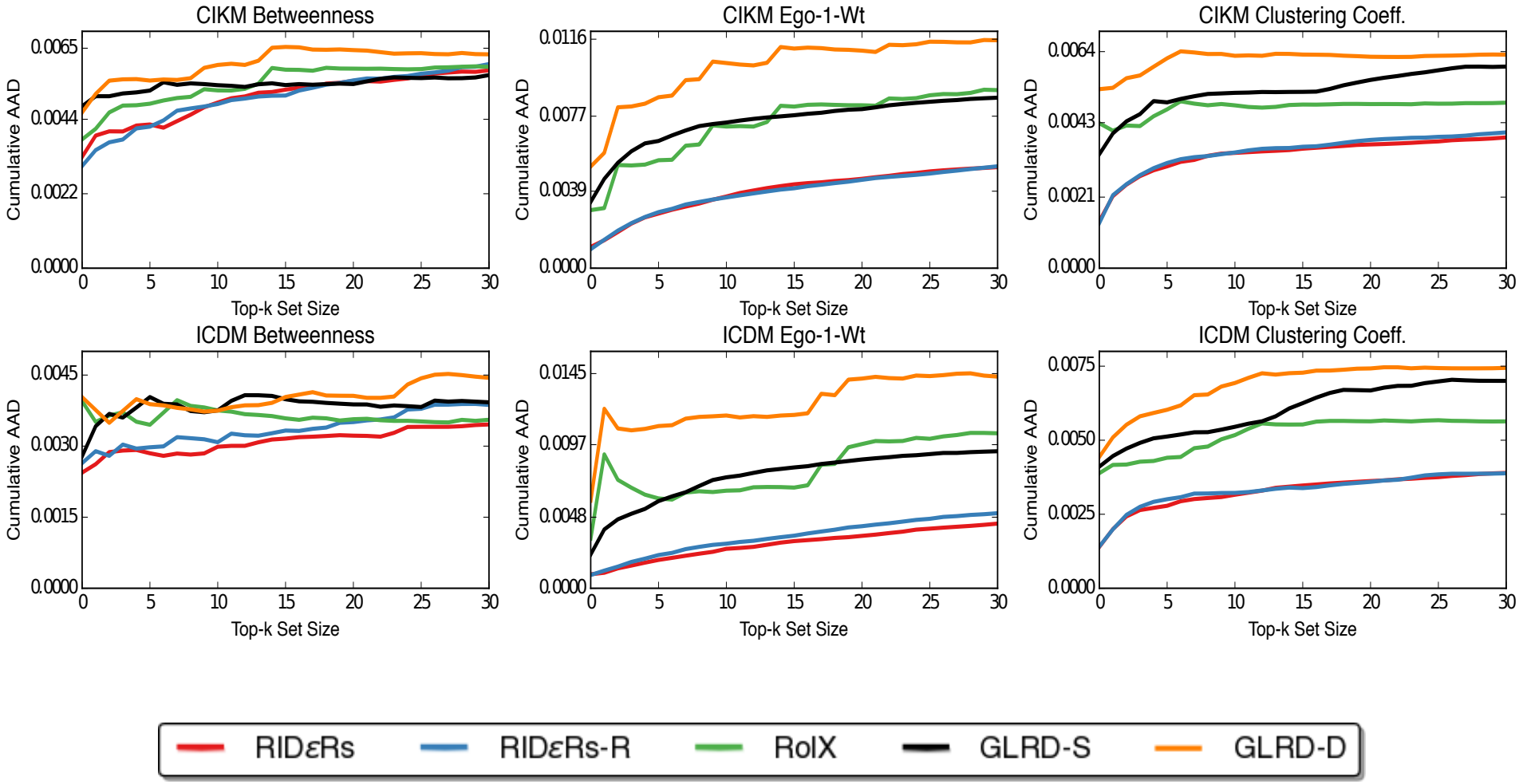


Figure: Top-k Evaluation Results: CIKM and ICDM Co-Authorship Year 2005-2013 Networks

Scalability

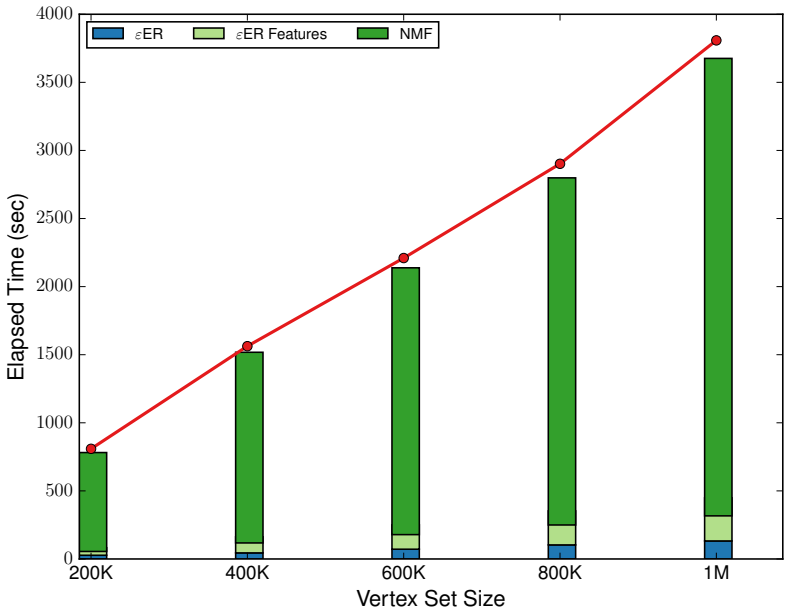


Figure: Running time of RIDεERs algorithm on random power-law graphs, $\gamma = 2.5$.

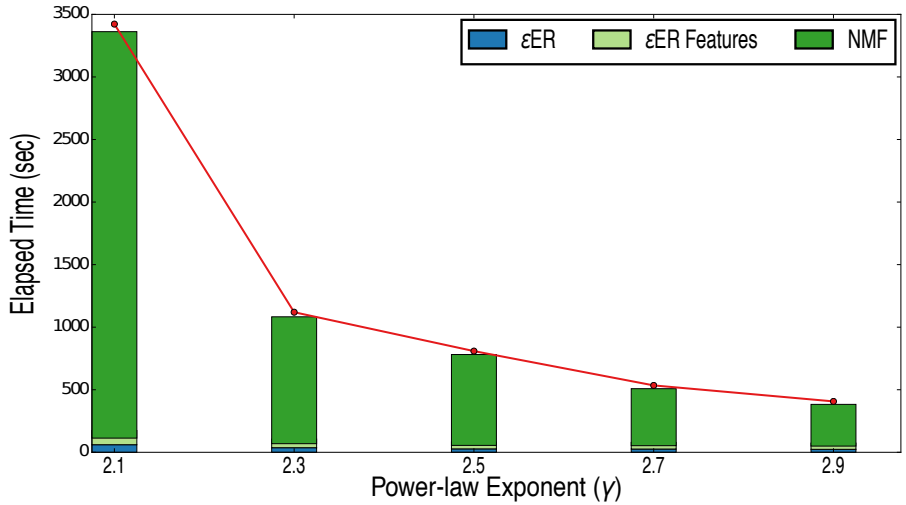


Figure: Effect of power-law exponent $\gamma \in (2, 3)$ on running time of RIDεERs algorithm. Lower values of exponent signify denser graphs. The number of nodes in each of the graphs is $200,000$.

Summary

- RID$\mathcal{E}$Rs encapsulates the complete structural view of a graph
- Novel evaluation framework helps in validation and comparison of role discovery algorithms

Acknowledgments

- Ericsson Global Research Center - Chennai, India
- Tina Eliassi-Rad, Northeastern University
- Interdisciplinary Laboratory for Data Sciences, IIT Madras, India
- NSF

- Thank you.

References

- [*ReFeX*] K. Henderson, B. Gallagher, L. Li, L. Akoglu, T. Eliassi-Rad, H. Tong, and C. Faloutsos, "It's who you know: graph mining using recursive structural features," in Proceedings of the 17th ACM SIGKDD. ACM, 2011, pp. 663-671.
- [*RolX*] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "RolX: structural role extraction & mining in large graphs," in Proceedings of the 18th ACM SIGKDD. ACM, 2012, pp. 1231-1239.
- [*NMF*] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in NIPS, 2001, pp. 556-562.
- [*SNMF*] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," Bioinformatics, vol. 23, no. 12, pp. 1495-1502, 2007.
- [*GLRD*] S. Gilpin, T. Eliassi-Rad, and I. Davidson, "Guided learning for role discovery (glrd): framework, algorithms, and applications," in Proceedings of the19th ACM SIGKDD. ACM,2013,pp.113-121.
- [*SϵEP*] P. V. Gupte and B. Ravindran, "Scalable positional analysis for studying evolution of nodes in networks," SIAM Data Mining workshop on Mining Networks and Graphs (SDM), 2014.

Thank You

- Thank you.