Week 12: Graph Mining (Roles & Centrality)

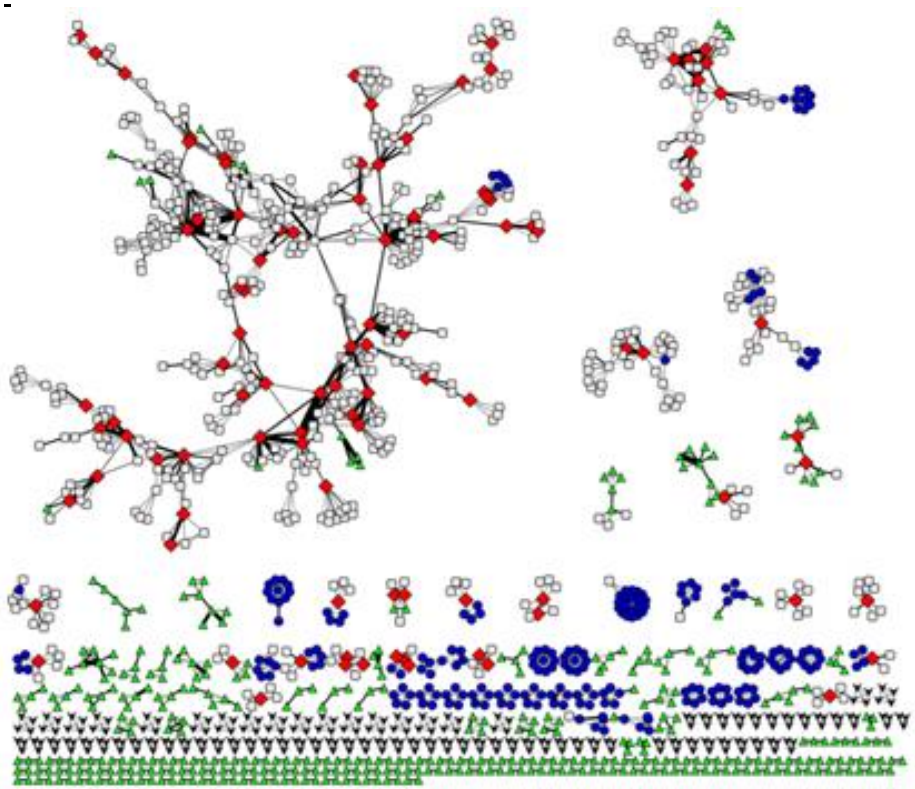
Instructor: Daejin Choi (djchoi@inu.ac.kr)

INCHEON NATIONAL UNIVERSITY

* Most slides are used from mmds.org (CS224W) with small updates

Plan for Today

- In the last
  - Subgraph
    - Defining
  - Computir

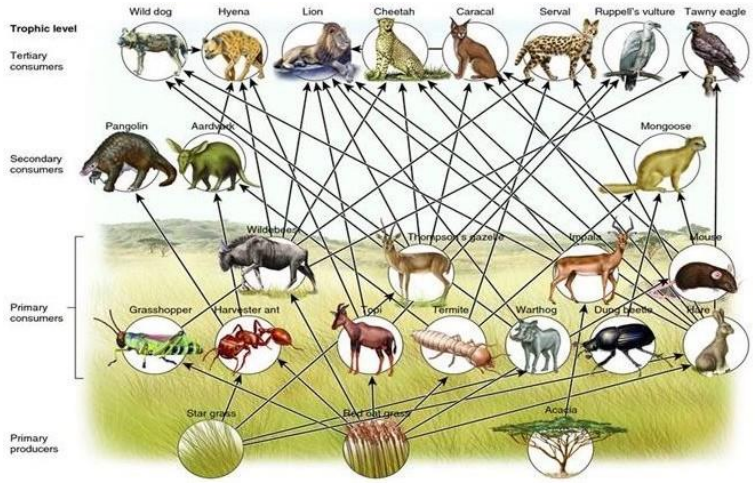


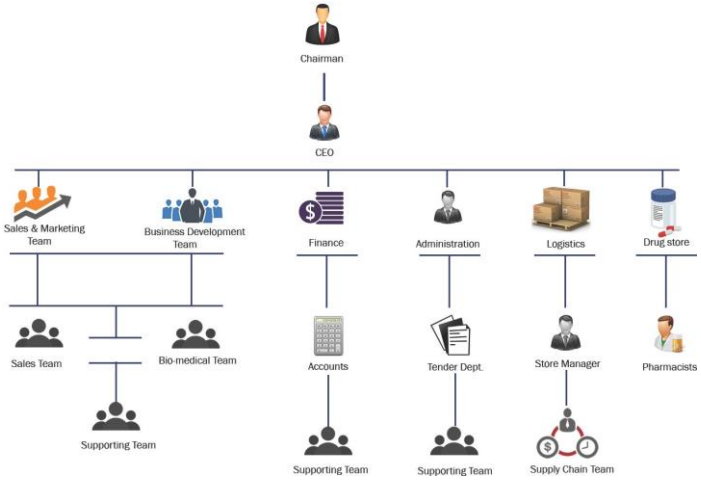
- Let's focus more on the "node itself"
  - Structural roles in networks
    - Discovering structural roles and its applications: RolX
  - Centrality: A Specific Role of Nodes in the Network

Structural Roles in Networks

What Are Roles?

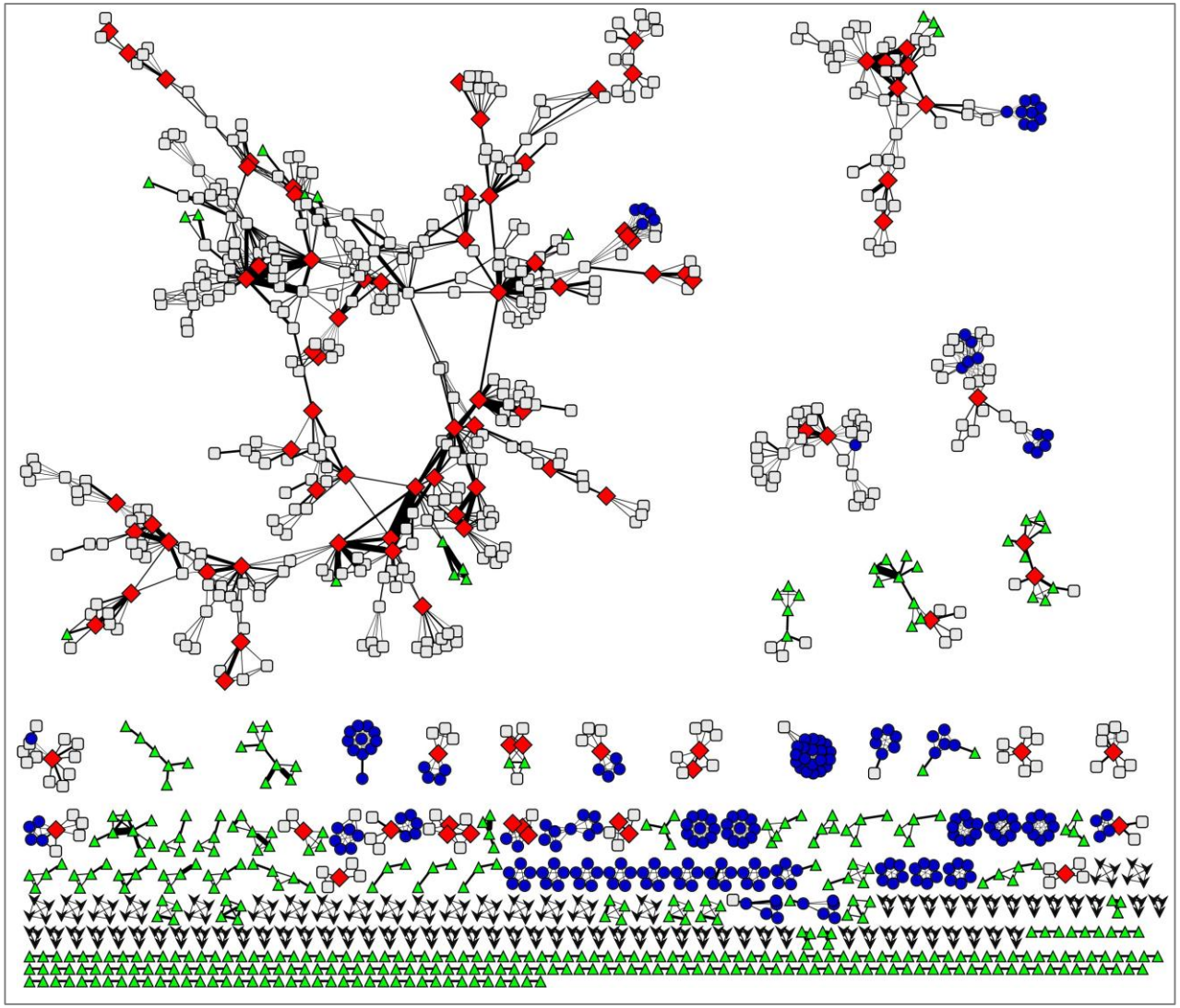
- Roles are "**functions**" of nodes in a network:
  - Roles of species in ecosystems
  - Roles of individuals in companies





- Roles are measured by structural behaviors:
  - centers of stars (hub)
  - members of cliques (community)
  - peripheral nodes, etc. (outliers, …)

Example of Roles



- centers of stars
- members of cliques
- peripheral nodes

Network Science
Co-authorship network
[Newman 2006]

Roles vs. Groups in Networks

- Role: A collection of **nodes** which have **similar positions** in a network:
  - Roles are based on the similarity of ties between subsets of nodes

- **Different** from groups/communities!
  - Group is formed based on adjacency, proximity or reachability
  - This is typically adopted in current data mining

Nodes with the same role need not be in direct, or even indirect interaction with each other

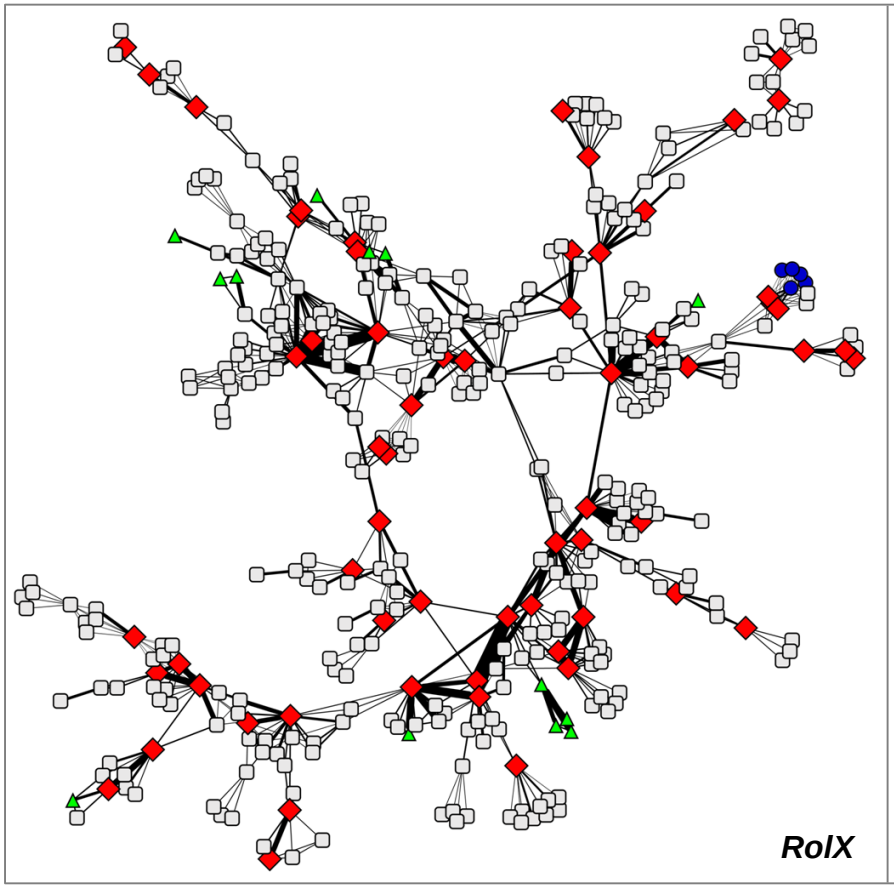
Roles vs. Groups in Networks

- Roles:
  - A group of nodes with similar structural properties
- Communities/Groups:
  - A group of nodes that are **well-connected** to each other
- Roles and communities are complementary

- Consider the social network of a CSE Dept:
  - Roles: Faculty, Staff, Students
  - Communities: AI Lab, Network Lab, System Lab

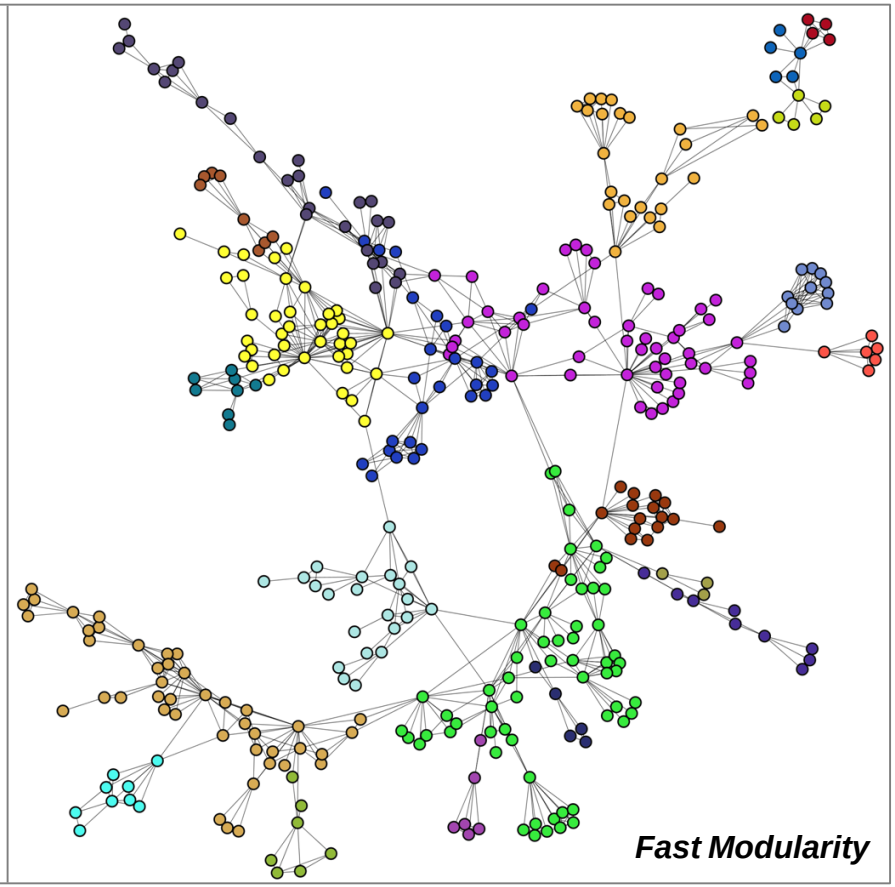
Roles vs. Groups in Networks

Roles



Henderson, *et al.,* KDD 2012

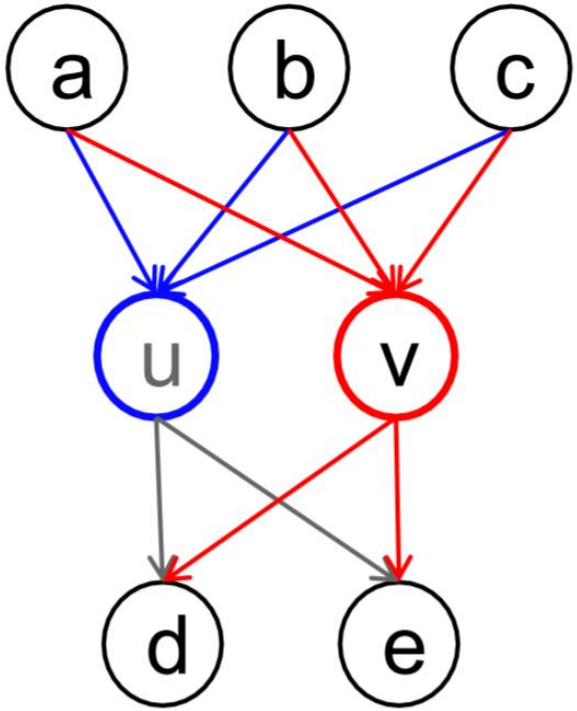
Communities



Clauset, *et al.,* Phys. Rev. E 2004

Roles: More Formally

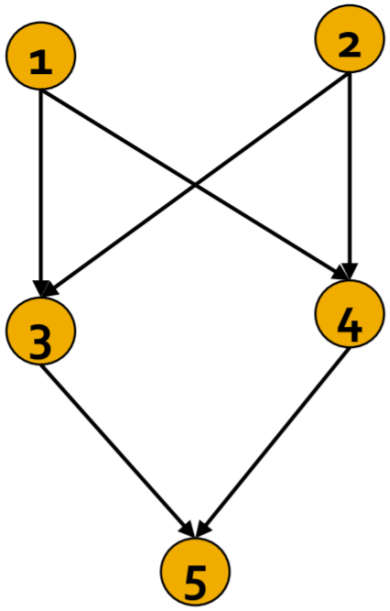
- **Structural equivalence**: Nodes u and v are **structurally equivalent** if they have the same relations **to all other nodes** [Lorrain & White 1971]
  - Structurally equivalent nodes are likely to be similar in other ways – i.e., friendships in social networks



Structural Equivalence: Example

- Nodes u and v are structurally equivalent:
  - For all the other nodes k, node u has tie to k iff node v has tie to k

Example:



Adjacency matrix

	1	2	3	4	5
1	-	0	1	1	0
2	0	-	1	1	0
3	0	0	-	0	1
4	0	0	0	-	1
5	0	0	0	0	-

E.g., nodes 3 and 4 are structurally equivalent

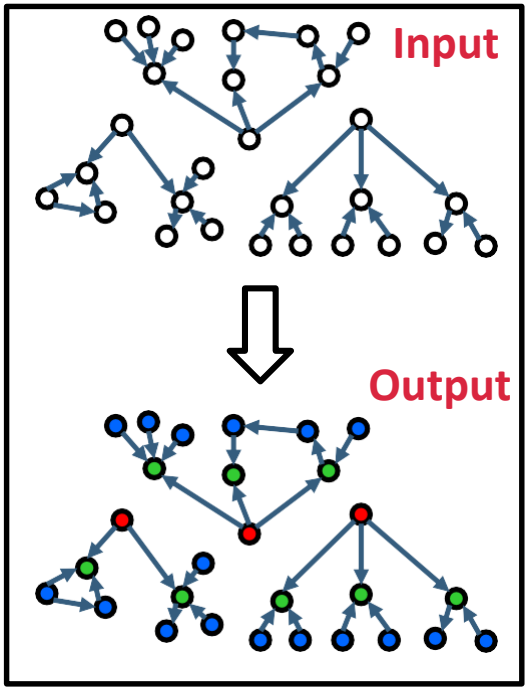
Discovering Structural Roles in Networks

Method to Discover Structural Roles

- **RolX**: Automatic discovery of nodes' structural roles in networks [Henderson, et al. 2011b]

  - Unsupervised learning approach
  - No prior knowledge required
  - Assigns a mixed-membership of roles to each node
  - Scales linearly in #(edges)

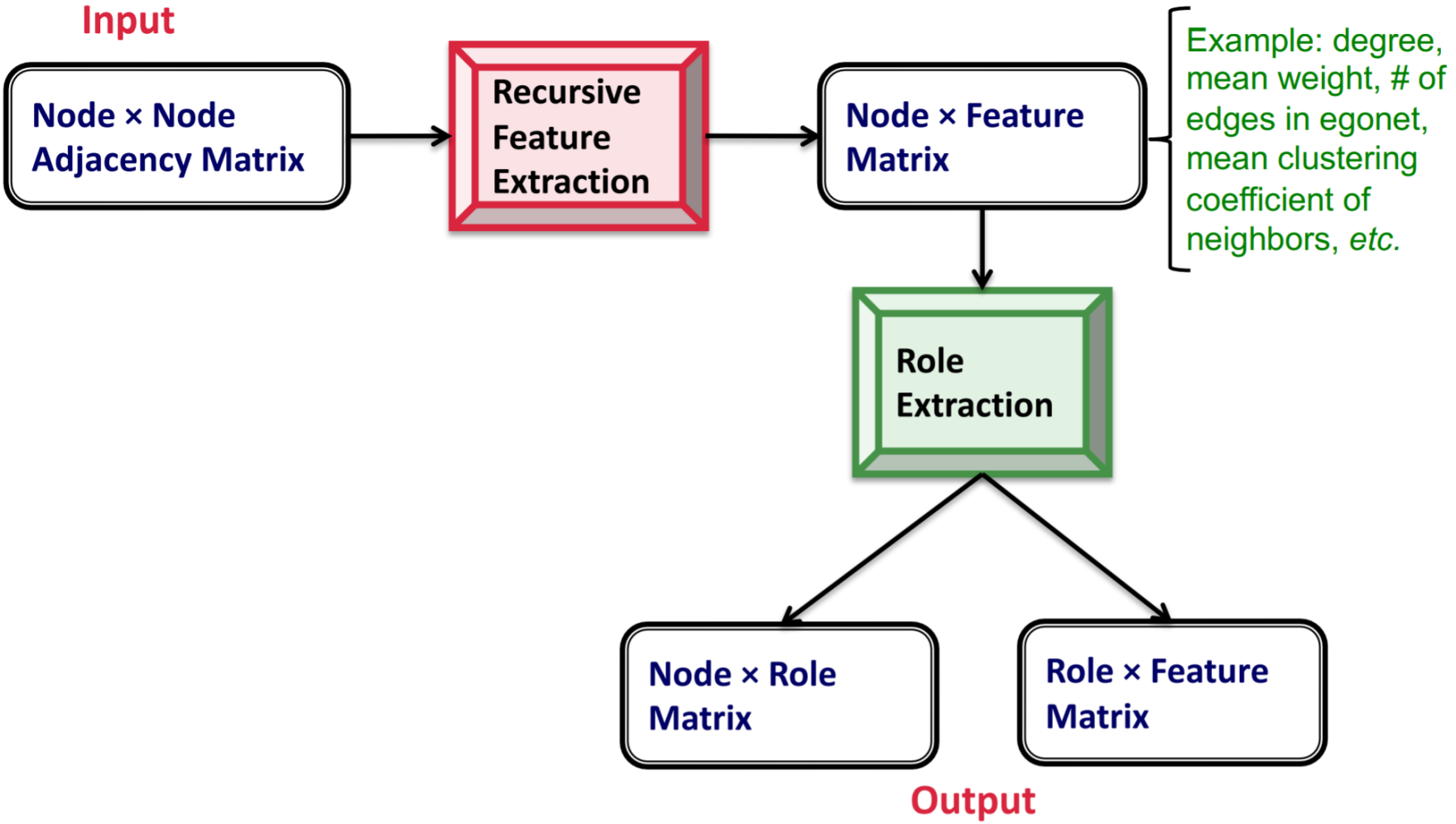
Role Discovery



✓ Automated discovery
✓ Behavioral roles
✓ Roles generalize

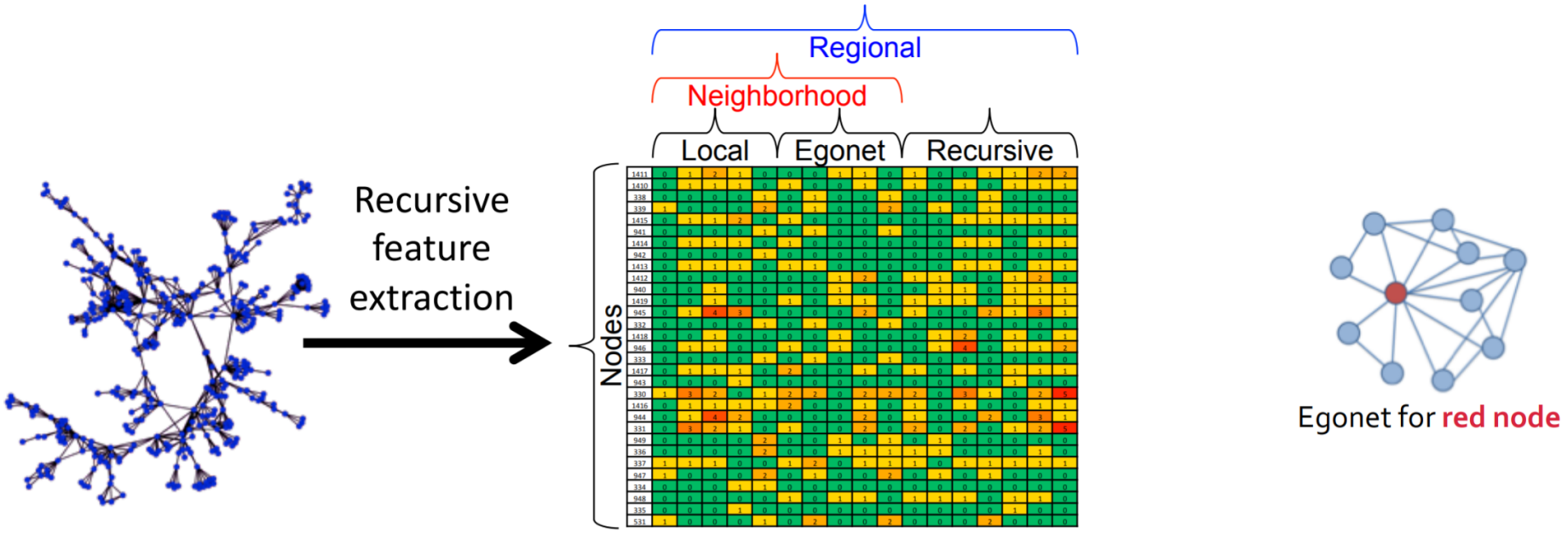
https://github.com/benedekrozemberczki/RolX

RolX: Approach Overview



Recursive Feature Extraction

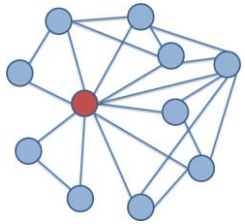
- Recursive feature extraction [Henderson, et al. 2011a] turns network connectivity into structural features



- Neighborhood features: What is a node's **connectivity pattern**?
- Recursive features: To **what kinds of nodes** is a node **connected**?

Recursive Feature Extraction

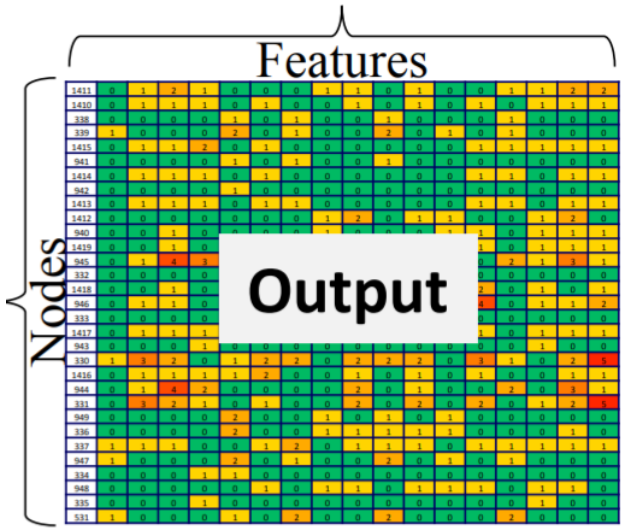
- **Idea**: Aggregate features of a node and use them to generate new recursive features
- Base set of a node's neighborhood features:
  - **Local features**: All measures of the node degree:
    - If network is directed, include in- and out-degree, total degree
    - If network is weighted, include weighted feature versions
  - **Egonet features**: Computed on the node's egonet:
    - Egonet includes the node, its neighbors, and any edges in the **induced subgraph** on these nodes
    - #(within-egonet edges),
    - #(edges entering/leaving egonet)



Egonet for red node

Recursive Feature Extraction

- Start with the base set of node features
- Use the set of current node features to generate additional features:
  - Two types of **aggregate** functions: **mean** and **sum**
    - E.g., mean value of "unweighted degree" feature between all neighbors of a node
    - Compute means and sums over all current features, including other recursive features
  - Repeat

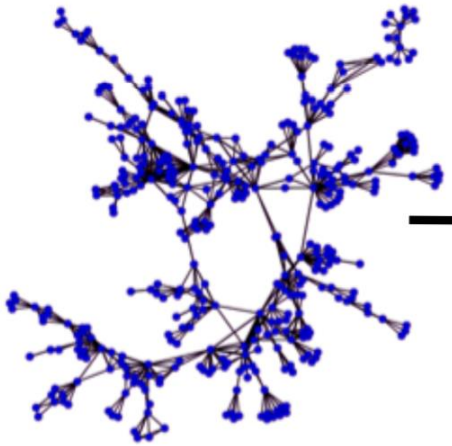


Recursive Feature Extraction

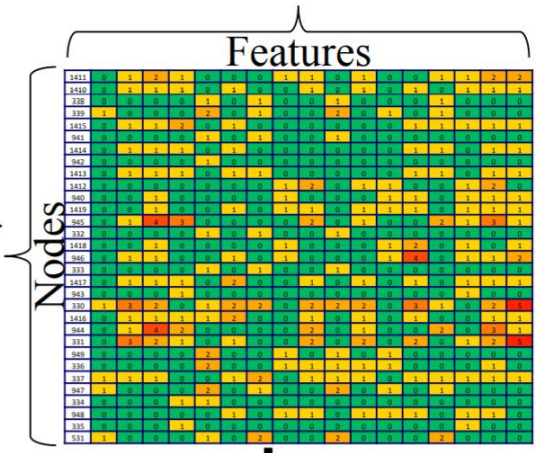
- The number of possible recursive features grows exponentially with each recursive iteration:
  - Reduce the number of features using a **pruning** technique:
  - **Look for pairs of features** that are **highly correlated**
  - **Eliminate one of the features** whenever two features are correlated above a user-defined threshold

Role Extraction

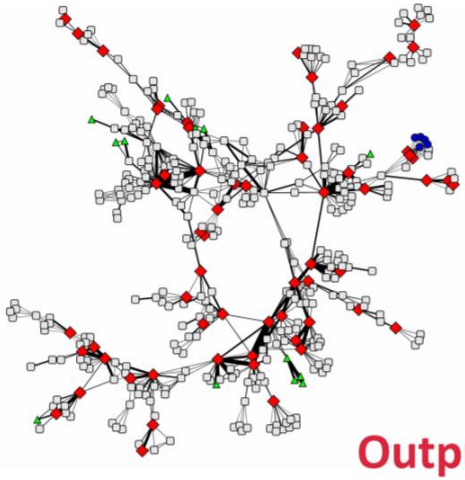
Input



Recursively extract features



Cluster nodes based on extracted features



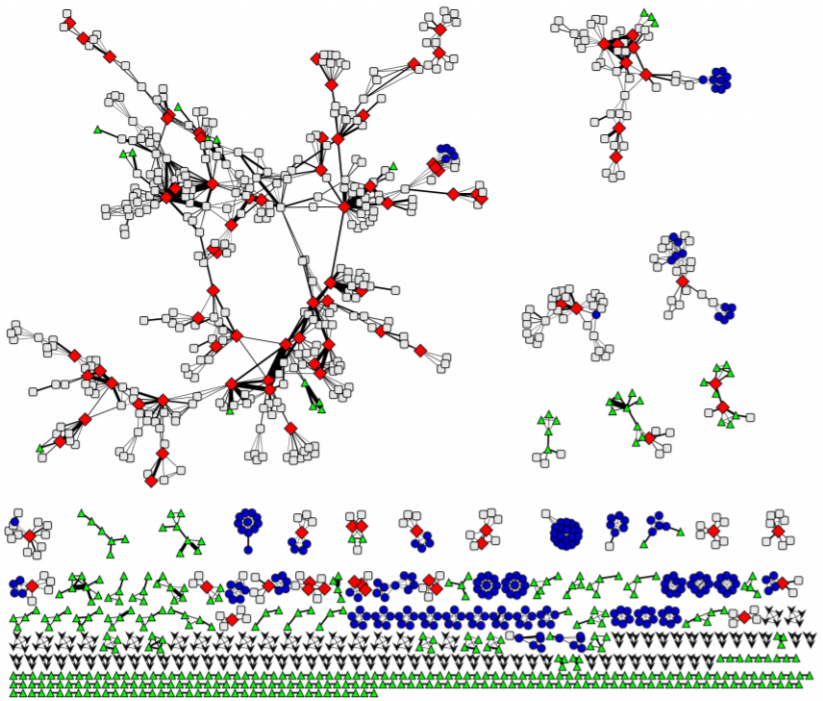
Output

RolX uses non negative matrix factorization for clustering, MDL for model selection, and KL divergence to measure likelihood

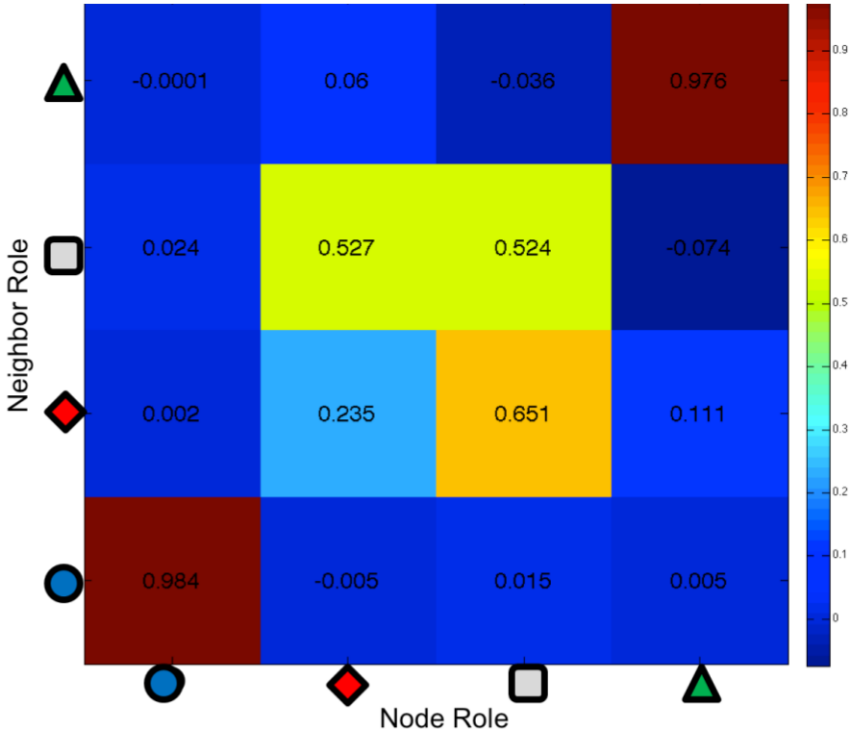
Application: Structural Similarity

- Task: Cluster nodes based on their structural similarity
- Two networks:
  - **Network science co-authorship** network:
    - Nodes: Network scientists;
    - Edges: The number of co-authored papers
  - **Political books co-purchasing** network:
    - Nodes: Political books on Amazon;
    - Edges: Frequent co-purchasing of books by the same buyers
- Setup: For each network,
  1. Use RolX to assign each node a distribution over the set of discovered, structural roles
  2. Determine similarity between nodes by comparing their role distributions

Structural Sim: Co-authorship Net.



Role-colored graph: each node is colored by the primary role that RolX finds

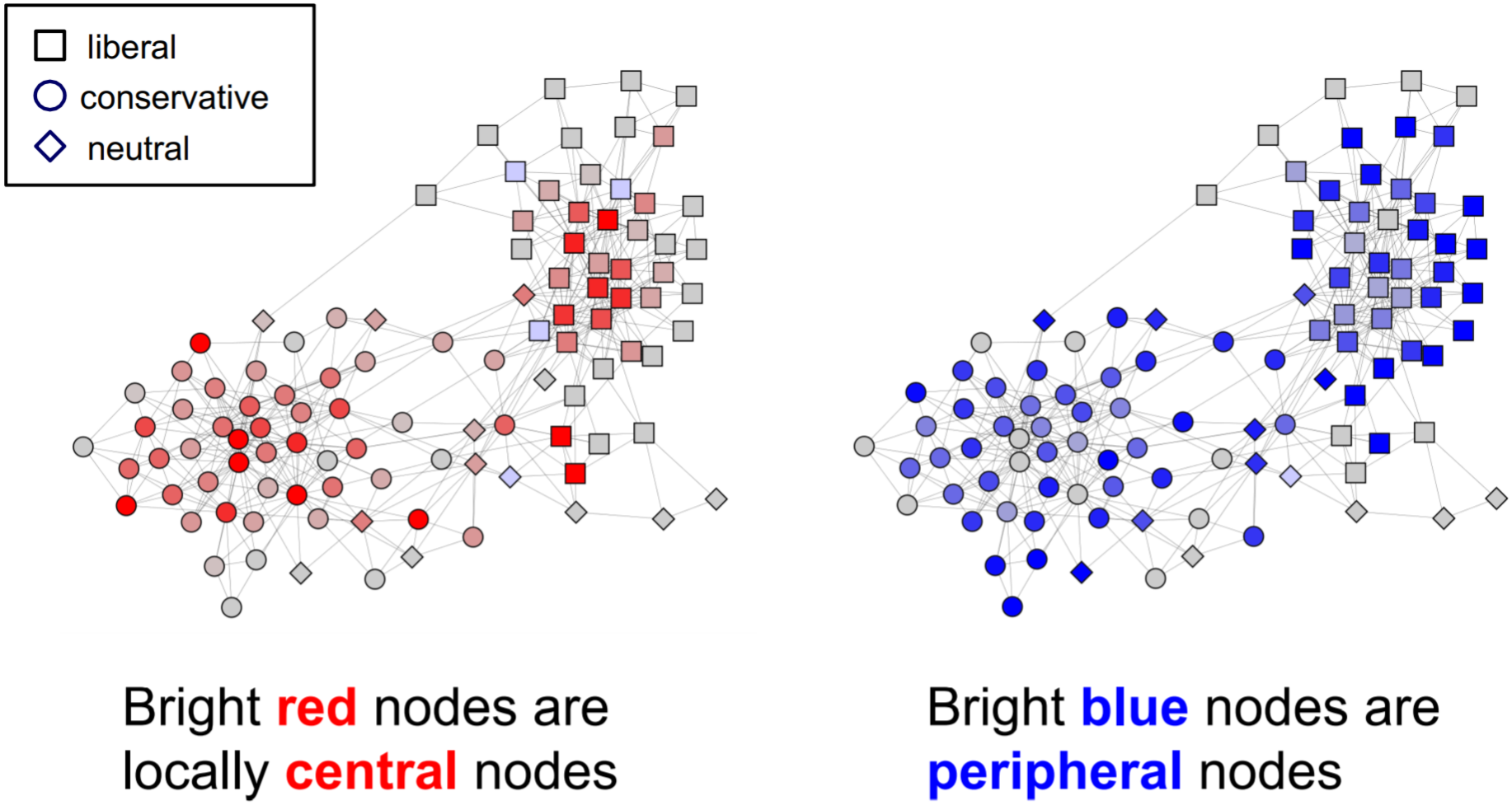


Role affinity heat-map

Making sense of roles:

- **Blue circle: Tightly knit**, nodes that participate in tightly-coupled groups
- **Red diamond: Bridge nodes**, that connect groups of nodes
- **Gray rectangle: Main-stream**, majority of nodes, neither a clique, nor a chain
- **Green triangle: Pathy**, nodes that belong to elongated clusters

Structural Sim: Co-authorship Net.



Bright **red** nodes are locally **central** nodes

Bright **blue** nodes are **peripheral** nodes

Book labels (i.e., liberal, conservative, neutral) were not given to role discovery algorithm

Summary

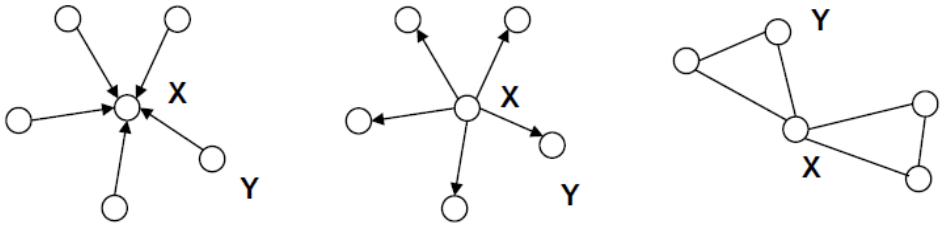
- Structural roles in networks
  - **RolX**: Structural Role Discovery Method

- Discovering structural roles and its applications:
  - Structural similarity
  - Role generalization and transfer learning
  - Making sense of roles

Centrality: A Specific Role of Nodes in the Network

Centrality (or Prestige)

- Some nodes are more important than others



- "Important" nodes?
  - Can be differently identified depending on the context
  - Exchange, spread of information, brokerage opportunities, etc.

- Centrality measures can quantify the **important** nodes from various perspectives

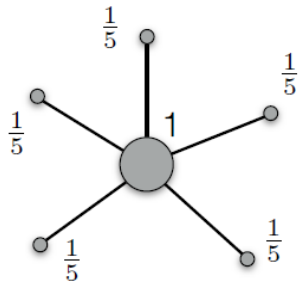
Node Centrality

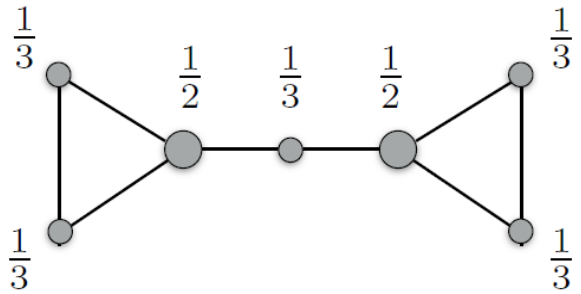
- We can identify hub nodes (or important nodes) from various perspectives
  - **Degree centrality**: Nodes with many neighbors
  - **Closeness centrality**: Close (higher weights or lower costs) to other nodes
  - **Betweenness centrality**: Nodes which connect to more "sub-networks"
  - **Eigenvector Centrality** (similar to **PageRank** Centrality): Nodes which connects to larger network

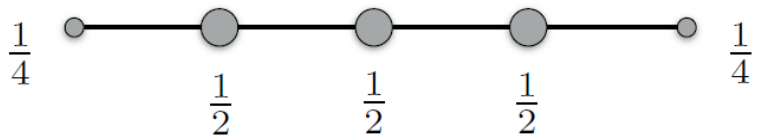
Degree Centrality

- The node with the most connections is most important
  - e.g., #supporters, audience size, #trading partners, #direct reports

$$C_d(v_i) = \frac{1}{N-1} d_i$$







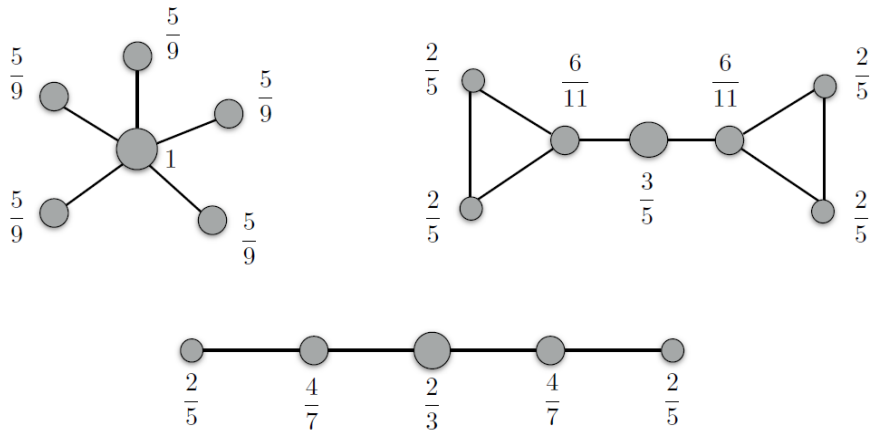
Closeness Centrality

- The node in the middle of the action is most central
  - Node with highest closeness centrality has the **shortest distance to the other nodes**, on average
  - e.g., access to information, opinion formation, spread of disease, adoption of new technology

$$C_C(v_i) = \frac{(N-1)}{\sum_{v_j \in G} d(v_i, v_j)}$$

Normalization (min possible distance to the N-1 other nodes)

Total distance to the other nodes



Betweenness Centrality

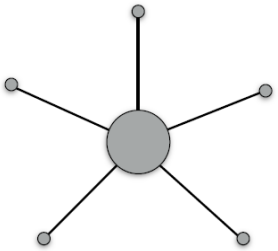
- The node you have to go through is most central
  - e.g., **brokering between groups**, control of information, innovation, collaboration

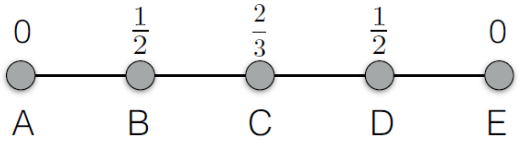
$$C_B(v_i) = \frac{\sum_{j<k} \frac{g_{jk}(v_i)}{g_{jk}}}{(N-1)(N-2)/2}$$

Fraction of geodesics going through the node

Normalization (number of pairs of nodes)

where g_{jk} is the number of geodesics between j and k and $g_{jk}(v_i)$ is the number that go through i

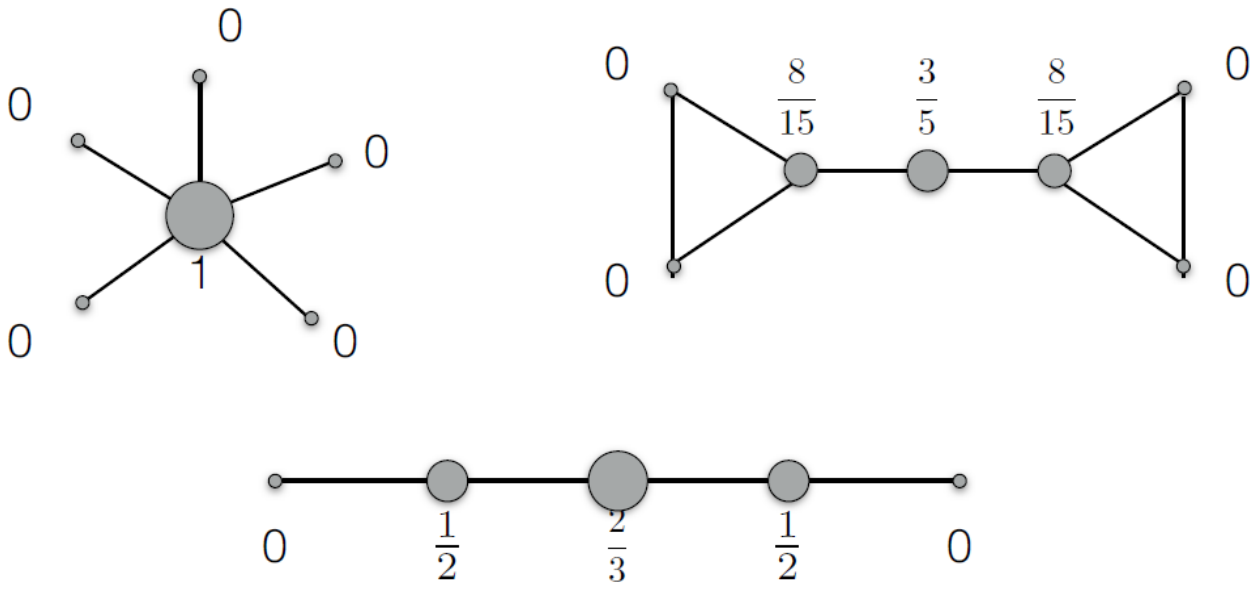




- A and E are not on any shortest paths
- B and D are both on 3 shortest paths
- C is on 4 shortest paths
- There are (N-1)(N-2)/2 = 6 total paths

Betweenness Centrality: Examples

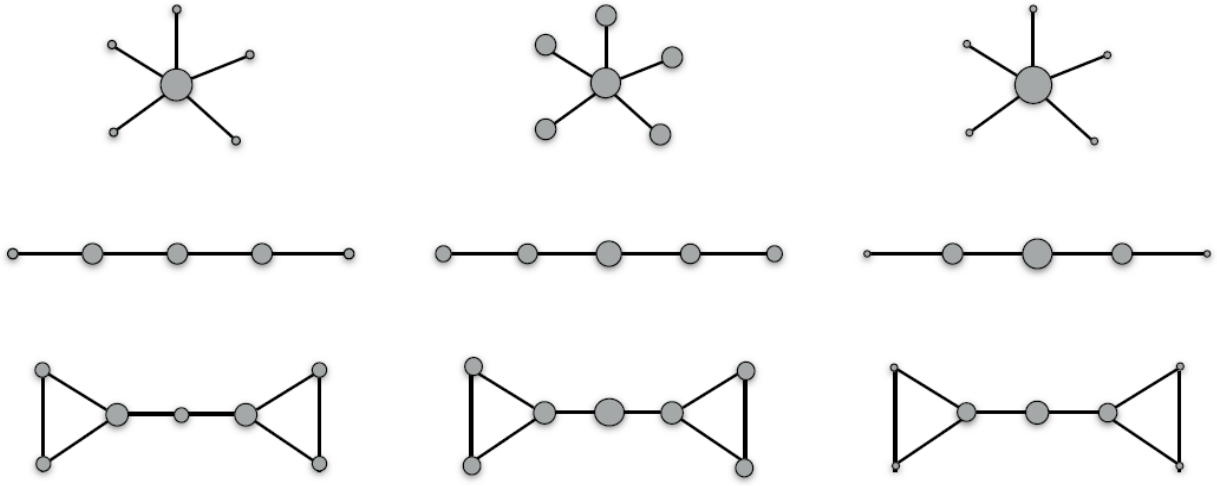
$$C_B(v_i) = \frac{\sum_{j<k} \frac{g_{jk}(v_i)}{g_{jk}}}{(N-1)(N-2)/2}$$



Eigenvector Centrality

- The node that **connects to the important nodes is important**
  - e.g., a twitter account followed by someone with a large audience, a entrepreneur who knows Jack Dorsey, a senator's barer
- A node's eigenvector centrality is proportional to the centrality of it's neighbors
- A node can have higher eigenvector centrality because:
  - They have more connections
  - They have more important connections

Comparing Three Centralities



The three measures are clearly related, but they each get at something slightly different!

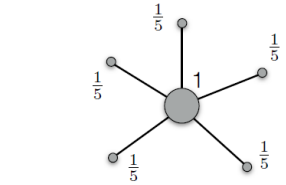
CF: Network Centralization

- A measure how centrality is distributed in the network

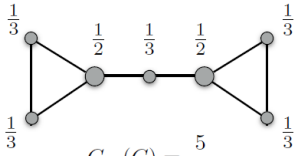
Difference between a node's centrality and the maximum centrality in the network

$$C_D(G) = \frac{\sum_{v_i \in G} [C_D(v^*) - C_D(v_i)]}{(N-1)}$$

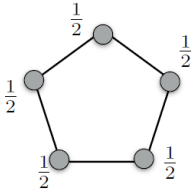
Normalization



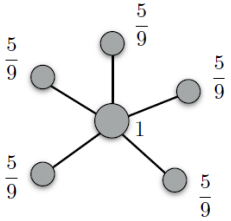
$C_D(G) = \frac{4}{5}$



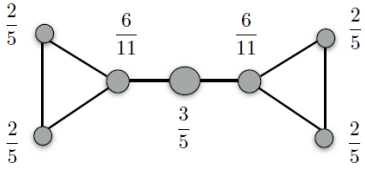
$C_D(G) = \frac{5}{36}$



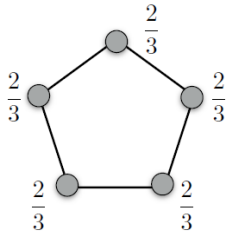
$C_D(G) = 0$



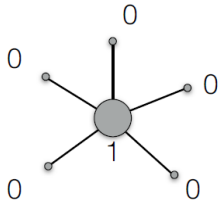
$C_C(G) = \frac{4}{9}$



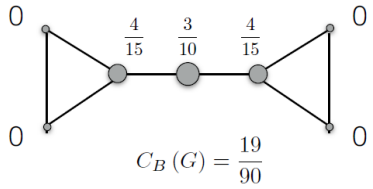
$C_C(G) = \frac{10}{66}$



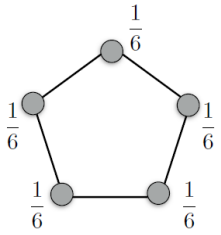
$C_C(G) = 0$



$C_B(G) = 1$



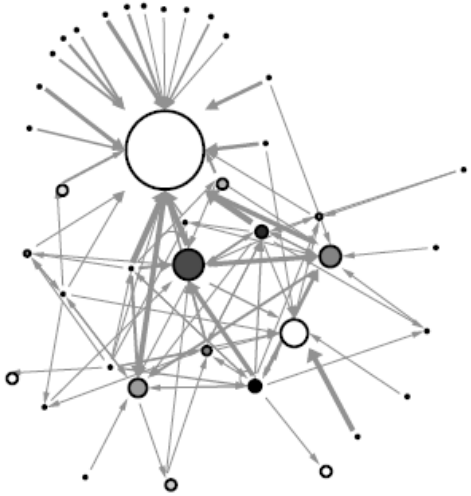
$C_B(G) = \frac{19}{90}$



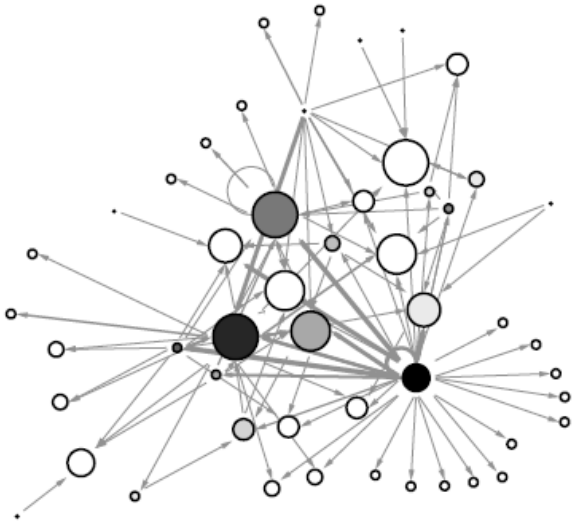
$C_B(G) = 0$

Network Centralization

- Centralization tells us about "how influence (or importance) is spread across the network"
  - e.g., Financial trading networks



High centralization: one node dominates the network



Low centralization: trades are more evenly distributed

Prestige

- Centrality in **directed** networks is called "**prestige**"
- This is sometimes a fine name:
  - admiration or trust
  - influence
  - friendship
  - trade
- But depending on the type of link, it might be misleading:
  - money lending
  - giving advice
  - hatred or distrust
- **Lesson**: **Context matters**! Always consider the **interpretation of a measure** in a particular context

Computing A Few Types of Prestige

- In-degree
  - A website that is linked to often has high prestige
  - A person who is frequently nominated for a reward has high prestige
  - How about out-degree?
- Closeness analogue: proximity
  - Uses shortest directed path length: directed geodesic
- Directed betweeness
  - Almost exactly the same as (undirected) betweenness, but with directed geodesics and normalized in a directed way
- Influence range
  - The influence range is what fraction of the nodes in the network can reach you via directed paths

Thank you!

Instructor: Daejin Choi (djchoi@inu.ac.kr)