# Context-Sensitive Detection of Local Community Structure

Dr. L. Karl Branting
The MITRE Corporation
Annapolis Junction, MD

MITRE

# Outline

- The local community-detection task
- Algorithms for local community detection
  - Algorithm schema
  - Previous algorithms
  - New algorithms
- Empirical evaluation
  - Networks
  - Evaluation criterion
  - Experiments
  - Results

# Motivation

- Communities often correspond to significant system components
  - Functional units
  - Families
  - Organizations
- Community structure may reflect significant features of system of interest
  - Centrality
  - Cluster coefficient
  - Density
  - Diameter

# Motivation

- If the entire graph is accessible, community structure can be found by optimizing global criterion, e.g., the partition with
  - maximum modularity (Newman & Girvan 2004)
  - minimum description length (Rosvall & Bergstrom 2007, Chakrabarti & Faloutsos 2004)
  - maximum partition density (Ahn et al. 2010)
- Often, only a portion of the graph is accessible, e.g., when
  - Graph >> memory, or
  - Expensive to find neighbors of each vertex
- Often, only local structure is of interest, e.g.,
  - Unnecessary to cluster entire WWW to study structure of political blogosphere
  - Only interested in community structure that involves seed entities

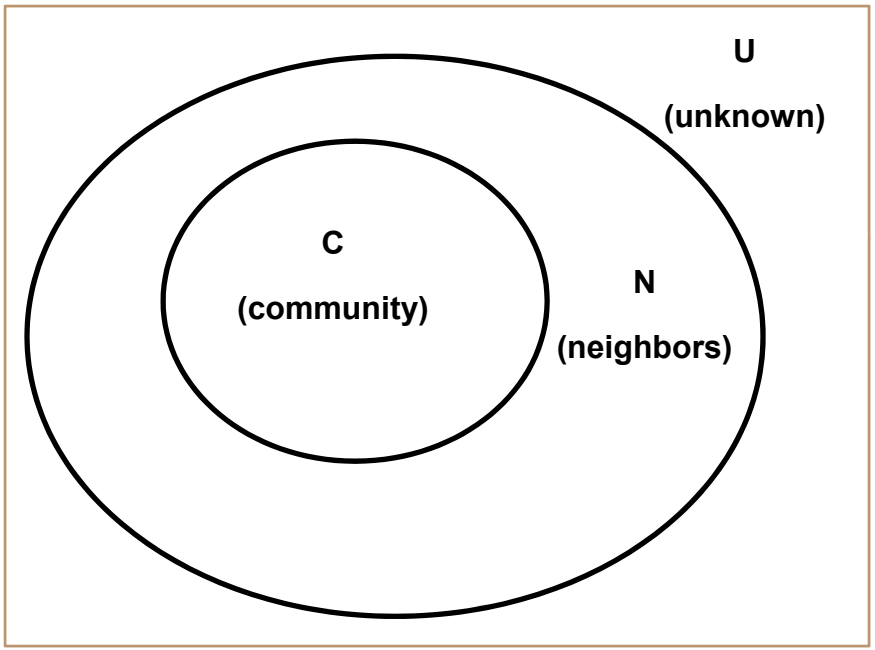
# Local Community Detection Task

- Given
  - Seed nodes
  - Graph
  - Maximum return set size *m**
- Find subgraph
  - Consisting of one or more communities
  - Containing seeds
  - Of size at most *m**

***Or other termination criterion, such as community boundary***

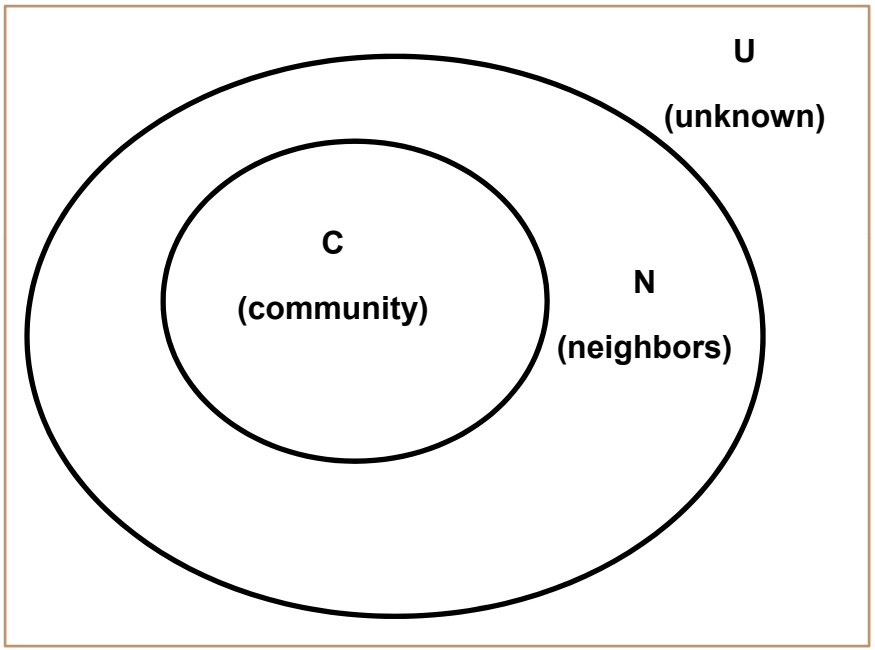
# Incremental Community Detection Schema

$C \leftarrow$ {queryVertex}
$N \leftarrow$ neighbors(queryVertex)
**while** (!terminationCriterion)
  select the 'best' vertex $n \in N$
  $C \leftarrow C \cup \{n\}$
  $N \leftarrow (N - n) \cup$ neighbors(n) – $C$
**end while**
**return** $C$



# Incremental Community Detection Schema

$C \leftarrow$ {queryVertex}
$N \leftarrow$ neighbors(queryVertex)
**while** (!terminationCriterion)
  select the 'best' vertex $n \in N$
  $C \leftarrow C \cup \{n\}$
  $N \leftarrow (N - n) \cup$ neighbors(n) – $C$
**end while**
**return** $C$

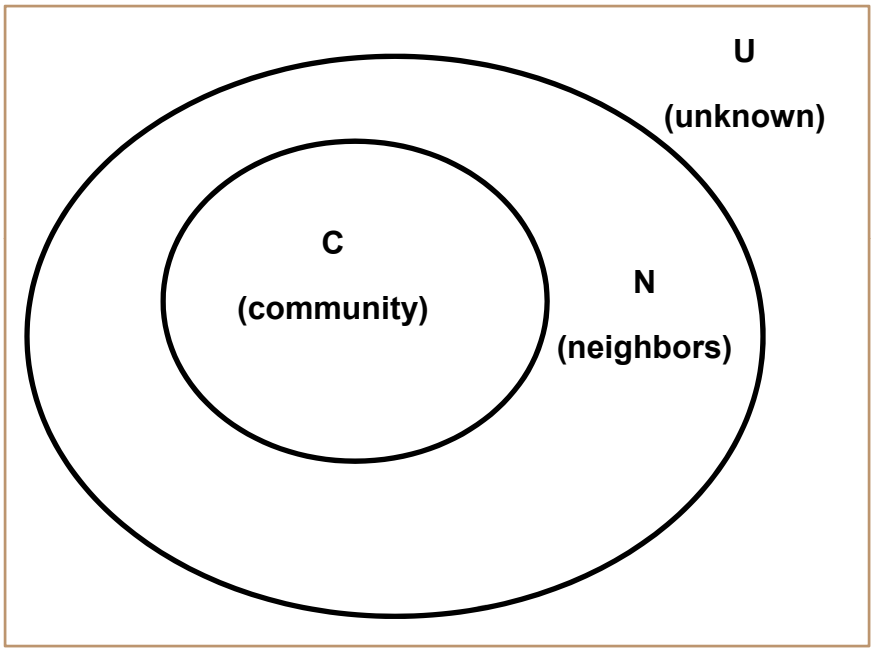


# Incremental Community Detection Schema

- **Implicit assumption: 'best' vertex criterion should favor vertices:**
  - **Most likely to be in the community**
  - **Of those equally likely to be in community, those most central to the community**
    - **All else being equal, higher-centrality vertices are better than lower-centrality vertex**

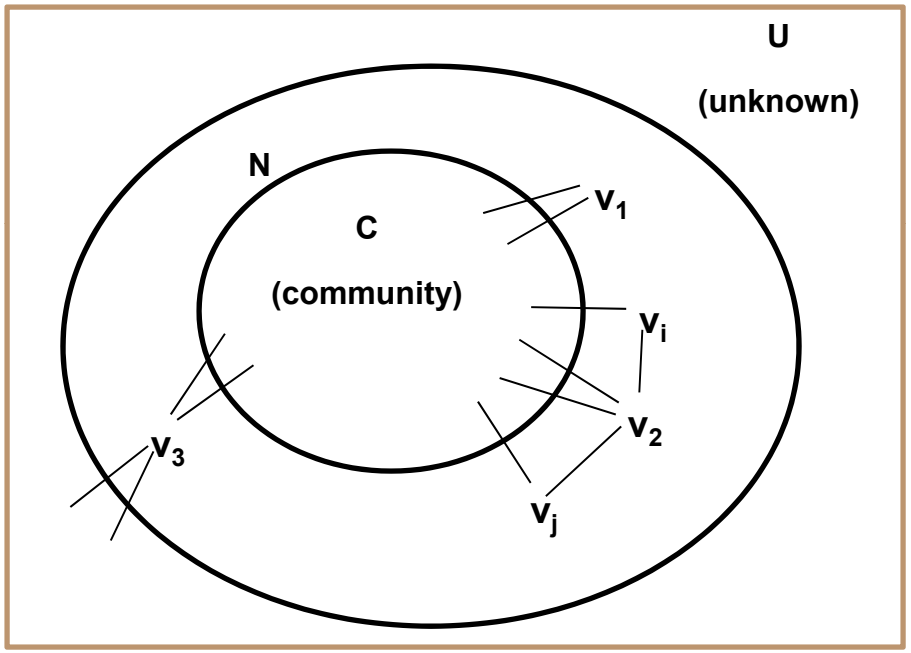
# Criteria for 'Best' Vertex in N

- **Previous**
  - **Maximize the ratio of edges inside C to edges from C to N (Luo et al. 2008): *MaxM***
  - **Minimize 'outwardness' =** $\frac{k_v^{out} - k_v^{in}}{k_v}$ **(Bagrow 2008): *MinOmega***
  - **Maximize boundary sharpness (proportion of community edges internal to C) (Clauset 2006): *MaxR***
- **Implicit assumptions**
  - **N-to-U edges are bad**
  - **N-to-N edges are as bad as N-to-U edges**



# Situation Where Assumptions Might Be False

- For MaxR, MinOmega, and MaxM
  - $v_1$ preferable to $v_2$ and $v_3$
  - v2 and v3 indistinguishable
- In practice
  - $v_2$ might have highest centrality in community, because it is more highly connected
  - $v_1$ might have lowest centrality, because it has lowest degree



# New Criteria for 'Best' Vertex in N

- **MaxActivation (Spreading activation)**
  - **Pick $n \in N$ that receives the greatest activation from the query vertex through the network**
    - **Activation propagates strictly outward, concurrently at each ply, attenuated along each edge**
    - **Differs from heat flow in that outgoing edges don't diminish activation**
    - **Roughly equivalent to Katz similarity**
- **MaxDensity**
  - **Pick $n \in N$ having the most connections to vertices in C**
  - **Break ties by choosing $n \in N$ with most edges to vertices in $N$**
  - **Break ties by choosing $n \in N$ with shortest path to query vertex**
- **Assumptions**
  - **N-to-N edges are good, not bad**
  - **N-to-U edges are not informative**

# Evaluation Criterion

- **Goal**
  - **Compare results to 'ground-truth' community selected by global criterion**
  - **Evaluate *vertex-selection* in isolation from termination criterion**
    - **Best termination criterion may depend on accuracy of vertex-selection**
- **Intuition**
  - ***Ceteris paribus,* higher centrality vertices better than lower centrality vertices**
- **Formalization**
  - **Normalized Utility-Weighted Recall (NUWR) [similar to R-precision (Baeza-Yates & Ribeiro-Neto 1999)]**
  - **For *m*-member return set, sum of utilities of all members, divided by the sum of utilities of the *m* most central members of the query vertex's community**

# Evaluation Criterion

- Implementation
  - Return set members not in the query vertex's community have utility 0.0
  - NUWR =1.0 means that return set is exactly equal to actual community
  - Utility is *node betweenness centrality* in globally optimal community (from oracle)
  - Two global criteria used in evaluation
    - Modularity (Newman 2004)
    - Partition density (Ahn et al. 2010)

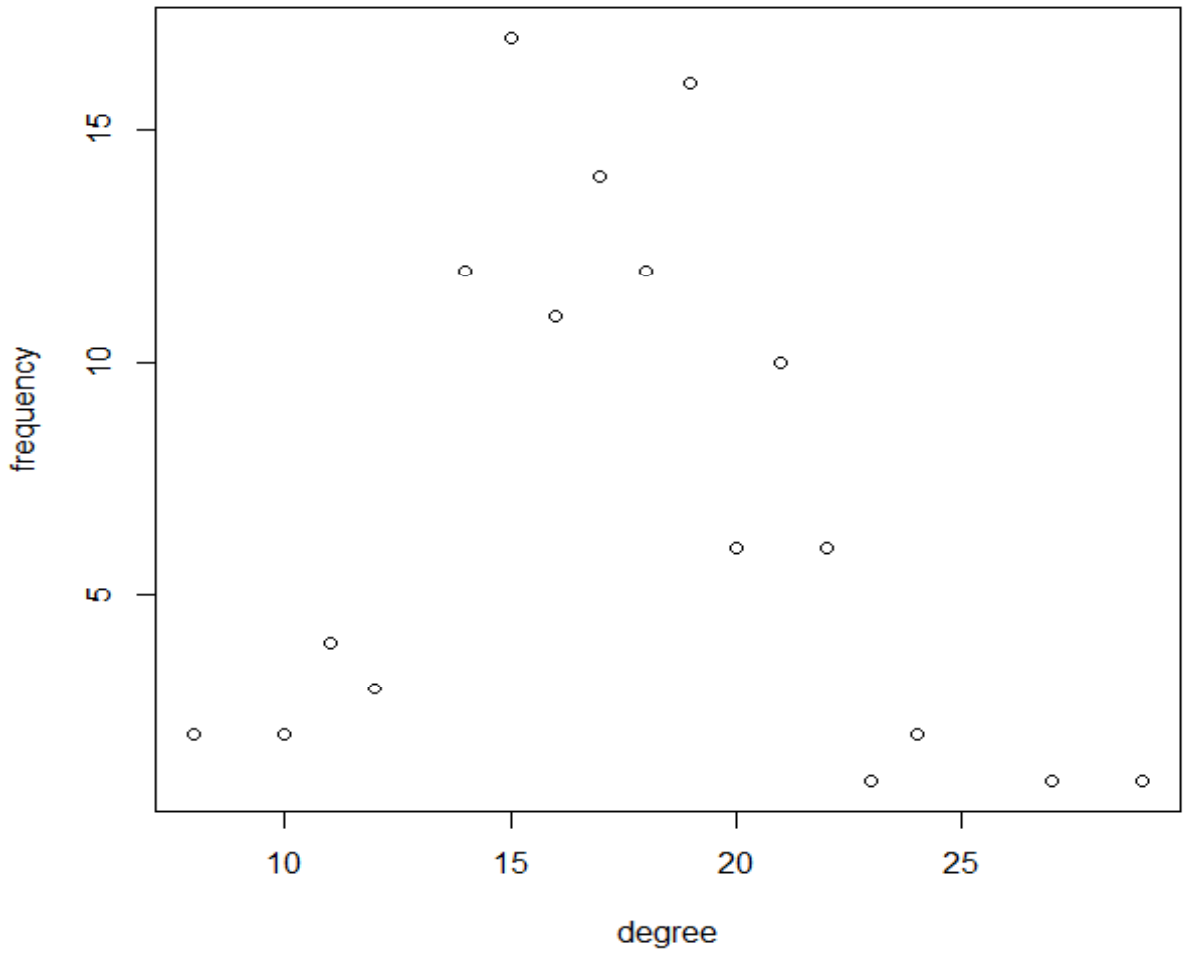
# Evaluation Set: Social, Cultural, and Natural Networks

- The Western US Power Grid [4941 vertices, 6594 edges]
- Network Science coauthorship [1589 vertices, 2742
- David Copperfield word adjacencies [112 vertices, 425 edges]
- Les Miserables co-appearance network [77 vertices, 254 edges]
- C. Elegans neural network [297 vertices, 2359 edges]
- Zachary's karate club [34 vertices, 78 edges]
- Dolphin social network [62 vertices, 159 edges]
- Jazz musician co-performances [198 vertices, 2742 edges]
- American college football, Division IA, Fall 2000 [115 vertices, 616 edges]

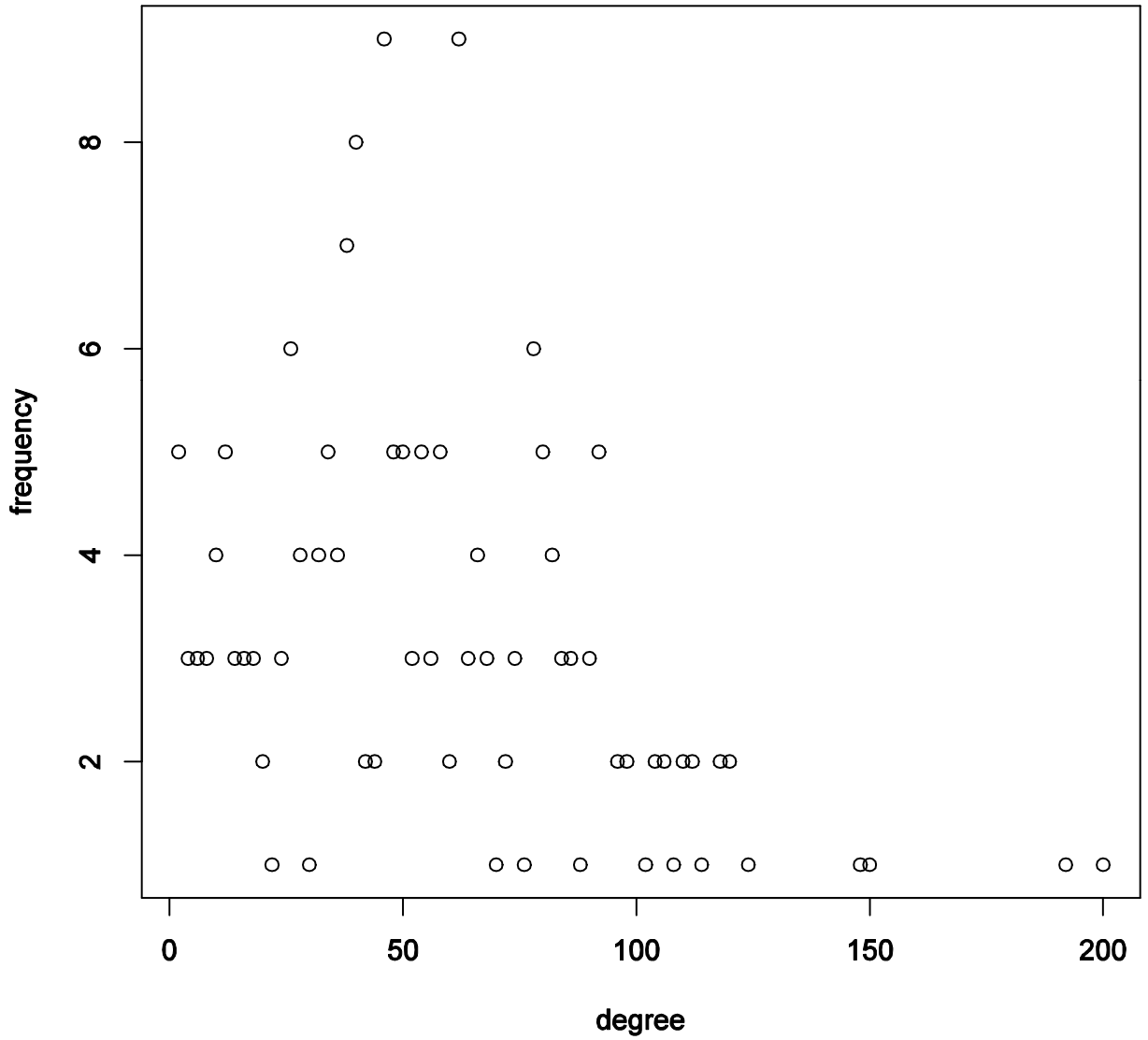
# Evaluation Set: Girvan-Newman Benchmarks

- Random networks used in many community-detection papers
- 128 vertices
- 4 communities
- Average vertex degree 16
- Variable proportion of edges internal to communities
  - 0.67 – weak community structure
  - 0.83 – moderate community structure
  - 0.90 – strong community structure

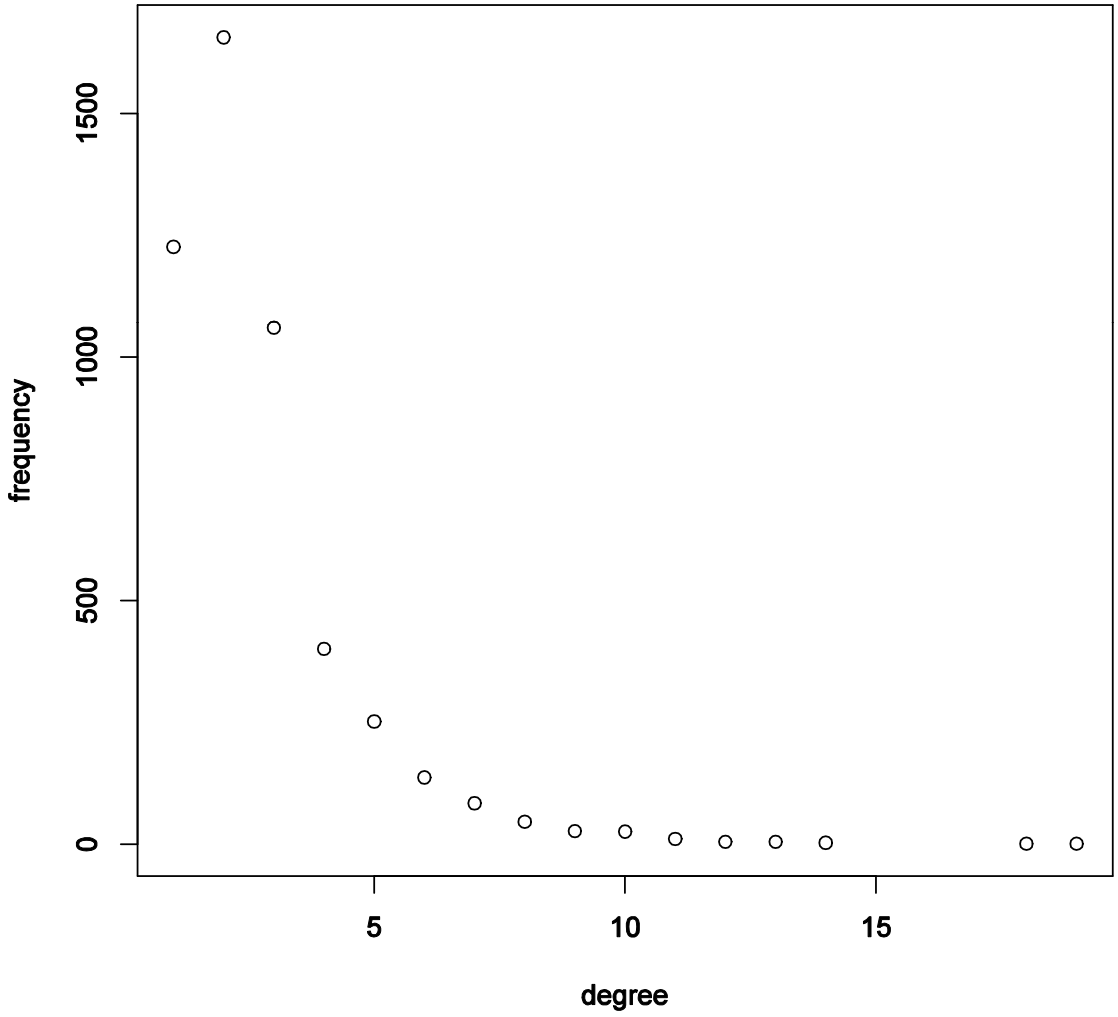
# Degree Distribution: GN-0.67



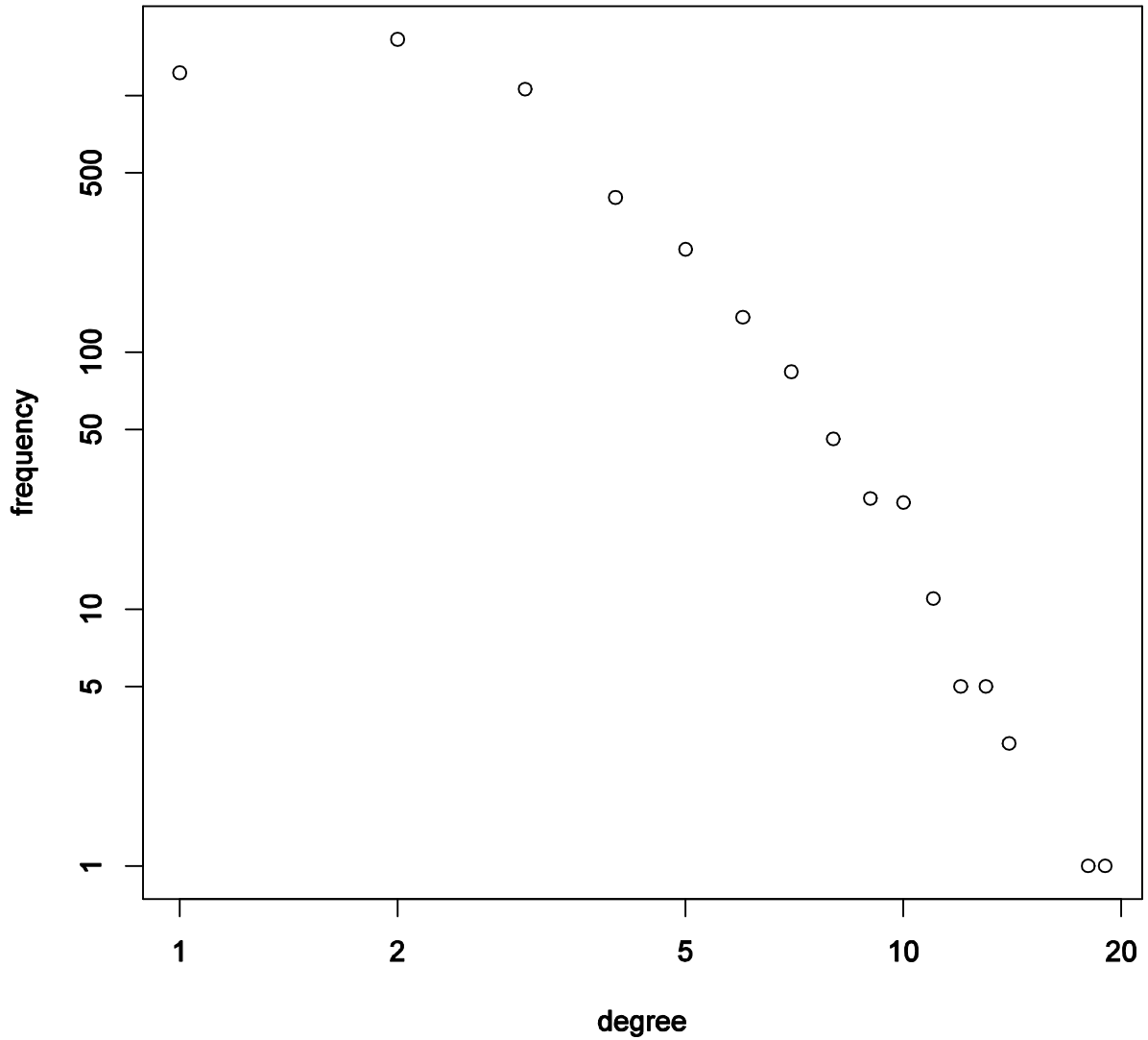
# Degree Distribution: Jazz Collaborations



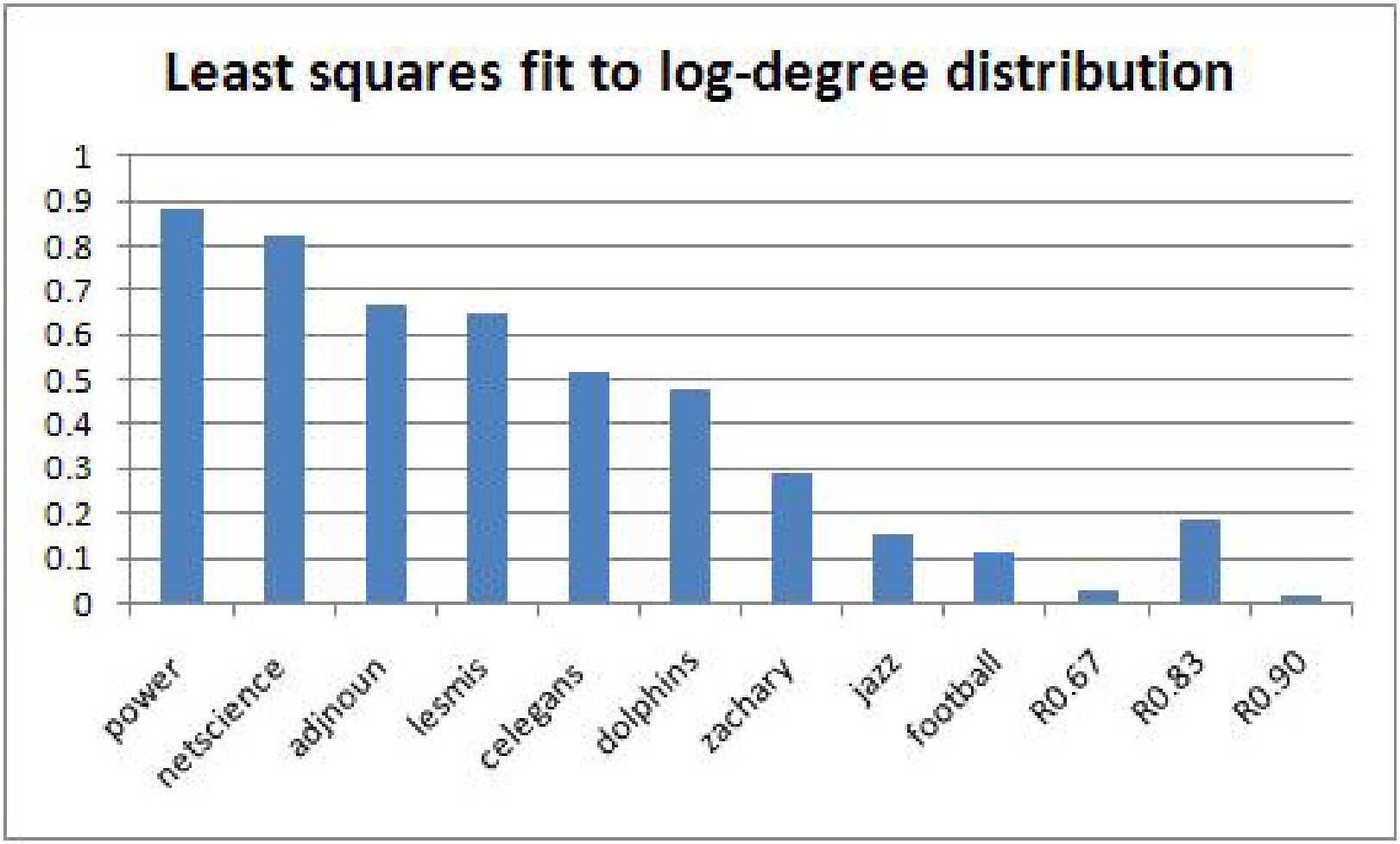
# Vertex Degree Distribution: Power Grid



# Log-log Degree Distribution: Power Grid



# Networks Ranked by Fit to Power-Law Distribution



# Experimental Design

- **Preparation**
  - **'Actual' community structure of each network determined using global criterion**
    - **Maximum modularity, found by hierarchical agglomerative clustering (Newman 2004)**
    - **Maximum partition density, found by link clustering (Ahn et al. 2010)**
  - **Node-betweenness centrality calculated for each community member (if node is in multiple communities, choose largest)**

# Experimental Design

- In each trial, for each network
  - Query vertex v selected randomly from graph
  - Corresponding community C for which v$\in$ C retrieved
  - Each algorithm run with seed v and return-set size |C|
  - Betweenness centralities of all return-set members summed
  - Sum divided by sum of betweenness centralities of all community members
- 1000 trials for each target-community/algorithm/network triple

# NUWR in Natural Graphs Relative to Globally Maximal Modularity

| | power | netscience | adjnoun | lesmis | celegans | dolphins | zachary | jazz | football |
|---|---|---|---|---|---|---|---|---|---|
| *$R^2$* | *0.881* | *0.821* | *0.669* | *0.646* | *0.515* | *0.478* | *0.291* | *0.153* | *0.116* |
| MaxM | 0.636 | 0.846 | 0.445 | 0.706 | **0.776** | **0.837** | **0.89** | **0.818** | 0.738 |
| MaxR | 0.324 | 0.800 | 0.380 | 0.708 | 0.66 | 0.614 | 0.606 | 0.722 | 0.292 |
| MinOmega | 0.492 | 0.830 | 0.290 | 0.539 | 0.359 | 0.545 | 0.527 | 0.349 | 0.331 |
| MaxDensity | 0.647 | 0.856 | 0.419 | 0.635 | 0.576 | 0.768 | 0.766 | 0.807 | **0.826** |
| MaxActivation | **0.702** | **0.885** | **0.538** | **0.727** | 0.669 | 0.824 | 0.826 | 0.803 | 0.733 |
| | | | | | | | | | |

# NUWR in Artificial Graphs Relative to Globally Maximal Modularity

| | R0.67 | R0.83 | R0.90 |
|---|---|---|---|
| $R^2$ | *0.030* | *0.184* | *0.014* |
| MaxM | 0.789 | 0.892 | 0.936 |
| MaxR | 0.413 | 0.345 | 0.355 |
| MinOmega | 0.300 | 0.300 | 0.322 |
| MaxDensity | **0.927** | **0.985** | **1.000** |
| MaxActivation | 0.769 | 0.912 | 0.942 |

# NUWR in Natural Graphs Relative to Globally Maximal Partition Density

| | power | netscience | adjnoun | lesmis | celegans | dolphins | zachary | jazz | football |
|---|---|---|---|---|---|---|---|---|---|
| *$R^2$* | *0.881* | *0.821* | *0.669* | *0.646* | *0.515* | *0.478* | *0.291* | *0.153* | *0.116* |
| MaxM | 0.81 | 0.48 | 0.74 | 0.78 | 0.60 | 0.76 | 0.86 | 0.68 | 0.65 |
| MaxR | 0.74 | 0.48 | 0.69 | 0.77 | 0.52 | 0.65 | 0.68 | 0.68 | 0.37 |
| MinOmega | 0.70 | 0.39 | 0.61 | 0.59 | 0.28 | 0.54 | 0.49 | 0.36 | 0.34 |
| MaxDensity | 0.89 | **0.49** | 0.73 | 0.85 | 0.45 | 0.76 | 0.69 | 0.79 | 0.78 |
| MaxActivation | **0.96** | **0.49** | **0.87** | **0.91** | **0.75** | **0.94** | **0.82** | **0.84** | **0.98** |
| | | | | | | | | | |

# NUWR in Artificial Graphs Relative to Globally Maximal Partition Density

| | R0.67 | R0.83 | R0.90 |
|---|---|---|---|
| $R^2$ | *0.03* | *0.18* | *0.01* |
| MaxM | 0.45 | 0.43 | 0.43 |
| MaxR | 0.40 | 0.34 | 0.35 |
| MinOmega | 0.36 | 0.33 | 0.37 |
| MaxDensity | 0.78 | 0.88 | 0.91 |
| MaxActivation | **0.85** | **0.93** | **0.96** |

# Discussion

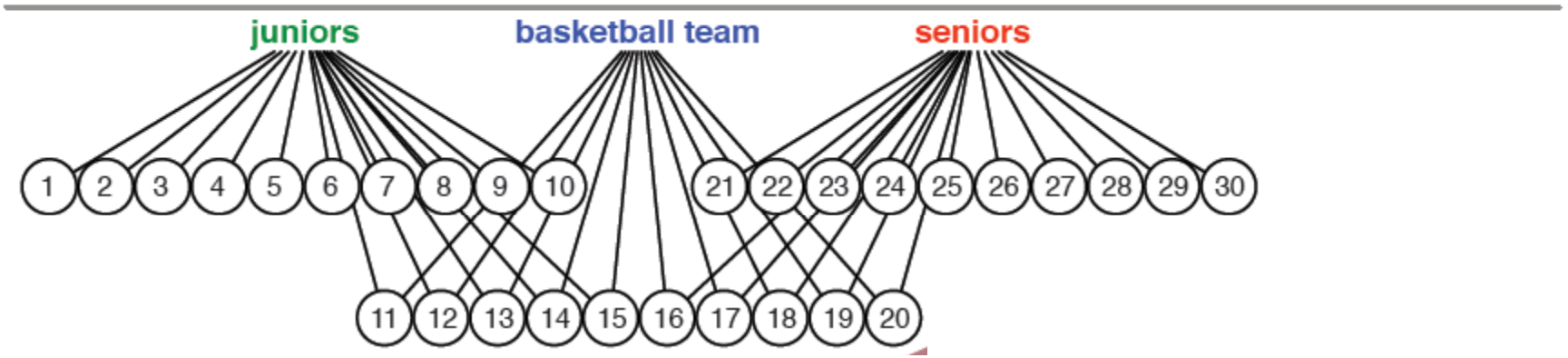
- The best vertex selection criterion depends on the target criterion and the degree distribution of the graph
  - Modularity
    - Fat tailed (as in classic scale-free graphs) – MaxActivation
    - Normally distributed (as in random graphs) – MaxDensity
    - All others – MaxM
  - Partition density
    - All degree distributions - MaxActivation
- N-to-U edges appear to be uninformative in fat-tailed and normal distributions, but informative for networks with other distributions

# Conclusion

- In local community structure algorithms, vertex-selection criteria should depend on structure of interest and properties of network:
  - Desired community structure
    - Modularity
    - Minimum Description Length
    - Partition density
  - Degree distribution
  - Others?
- Long term goal
  - Adapt community detection to any given community-structure model or vertex utility function provided by user:
    - Vertex-selection criterion
    - Termination criterion
    - Target community

# Link Clustering: Motivation

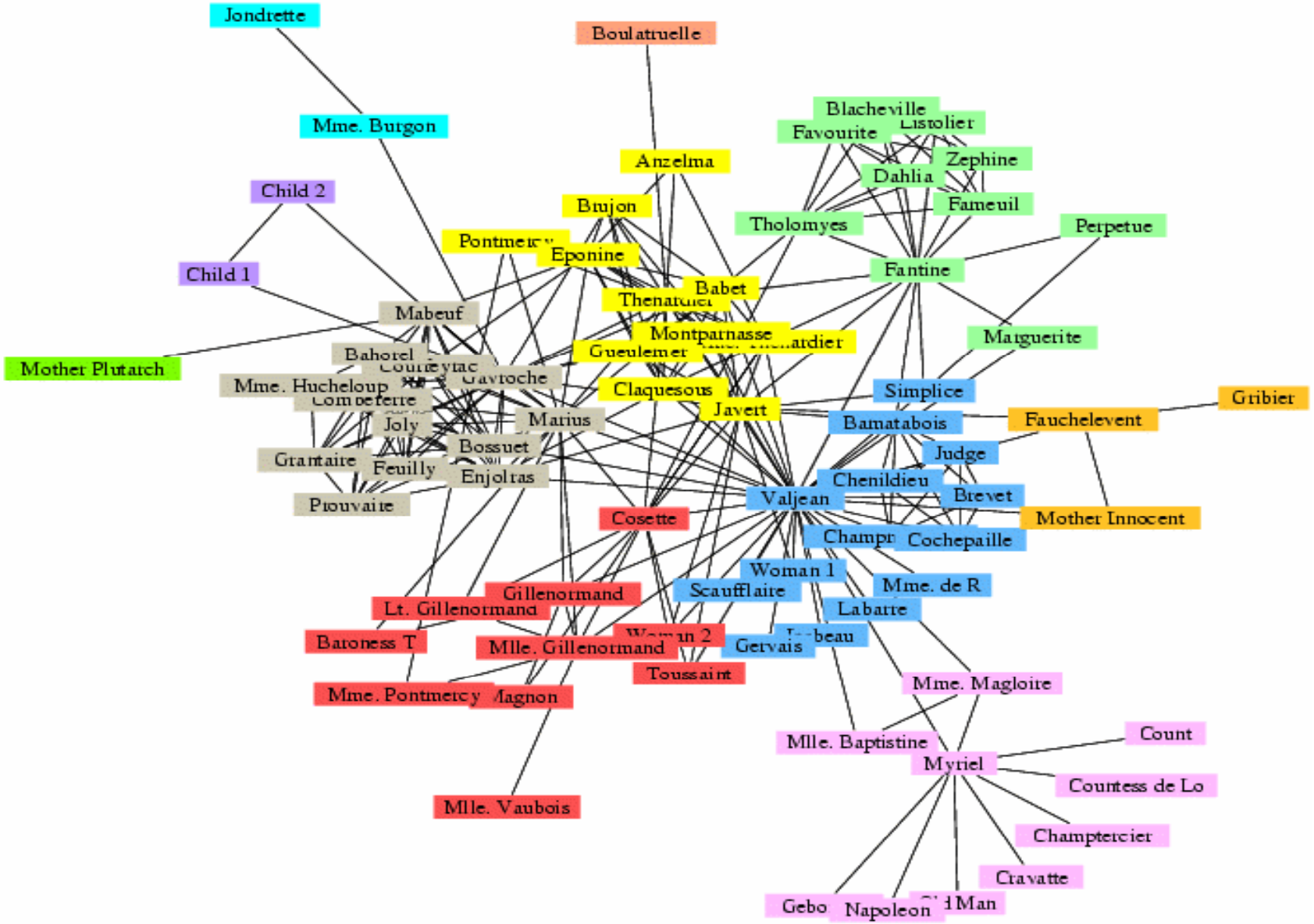
- Most individuals belong to many different groups
- Conventional partitional clustering assigns each node to just one community
- Link clusters permitting overlapping communities



# Link Clustering: Algorithm

- Link clustering algorithm proposed by Ahn et al., *Link communities reveal multiscale complexity in networks*, Nature 466:761-764 (August 2010).
- Elements
  - Distance between pairs of edges based on variant of Jaccard
  - Single-link agglomerative clustering
  - Link cluster quality measured by partition density
  - Best community structure found by maximizing partition density

# Modularity Maximization: Les Miserable Characters



# Link Clustering Example: Les Miserable Characters