Research Talk

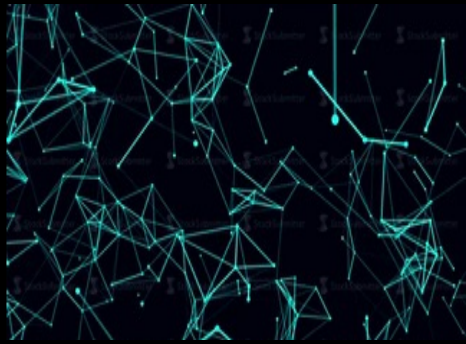
Graph Partatioing Clustering with User-Specfied Relative Density

February 1, 2023

Ms. Rohi Tariq (Ph.D student)

Research Advisors: Assoc. Prof. Dr. Kittichai Lavangnananda , Assoc. Prof . Dr. Pascal Bouvry, Assoc. Prof. Dr. Pornchai Mongkolnam

School of Information Technology
King Mongkut's University of Technology Thonburi



Outline

- Graph Clustering Algorithms Categorization
- Graph Clustering Quality Evaluation Metrics Categorization Based on Internal and External Connectivity
- Graph Clustering Algorithm Based on Clique Strategy
    1. Algorithm-1 Cliques Determination
    2. Algorithm-2 Clusters Determination
- Metric Mean Relative Density Deviation Coefficient (MRDDC)
- Experimental Results (Real-world Networks)
- Algorithm Effectiveness and Process Evaluation

Graph Clustering Algorithms Categorization

Graph Clustering Algorithms

Parameter-Free Algorithms

- **Fast-Greedy** (Modularity Optimization)
- **Info-Map** (Greedy Approach) (Information theory)
- **Label propagation** (Labels)
- **Walktrap** (Random Walk)
- **Spinglass** (Potts Model)
- **Leading eigenvector** (Modularity Matrix)
- **Graph-based k-means**

Parameterizable Algorithms

Girvan-Newman (Edge-Betweenness, Modularity Optimization)	**Level-Cut** Tunable parameters
Louvain (Modularity Optimization)	**Resolution-Limit (γ)** Tunable Parameter
Leiden (Modularity Optimization)	**Resolution-Limit (γ)** Tunable Parameter
Fluid-Communities.	**(k)** Number of Clusters
Spectral Methods	**ε-neighborhood, k-nearest** Connectivity parameter
k-means	**(k)** Number of Clusters
Expectation Maximization	**(k)** Number of Clusters

Graph Clustering Quality Metrics Categorization

Graph Clustering Quality Evaluation Metrics

Internal Connectivity parameter

Coverage:

$$\frac{\sum_{i=1}^{k} int_{deg}(C_i)}{e}$$

Internal-Edge Density

$$\delta_{int}(C_i) = \frac{int_{deg}(C_i)}{C_n(C_n-1)/2}$$

Average Embeddedness:

$$emb(G|C_1, C_2..C_k)\ \frac{1}{k}\sum_{i=1}^{k} emb(v, C_i)$$

Triangle Participation Ratio:

$$TPR\ (G|C_1, C_2..C_k) = \frac{1}{k}\sum_{i=1}^{k}(TPR\ (C_i))$$

FOMD (Fraction Over Median Degree)

$$FOMD\ (G|C_1, C_2..C_k) = \frac{1}{k}\sum_{i=1}^{k}(FOMD\ (C_i))$$

External Connectivity parameter

Conductance

$$\phi(G|C_1, C_2..C_k) = \frac{1}{k}\sum_{i=1}^{k}\phi(C_i)\ \phi(C_i) = \frac{deg_{ext}(C_i)}{min(deg_{tot}(C_i), 2m-deg_{tot}(C_i))}$$

Flake-Odf (Out Degree Fraction)

$$(G|C_1, C_2..C_k) = \frac{1}{k}\sum_{i=1}^{k}\text{Flake} - \text{ODF}\ (C_i)$$

Nor-Cut (Normalized Cut)

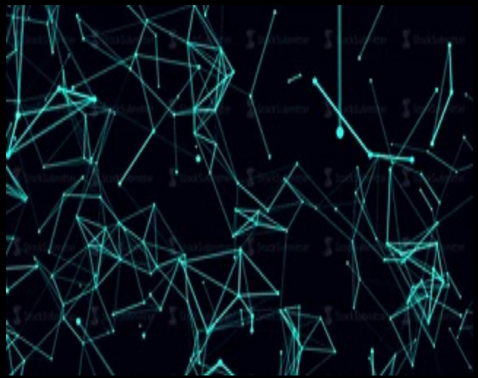
$$= \frac{ext_{deg}(C_i)}{2*int_{deg}(C_i)+ext_{deg}(C_i)} + \frac{ext_{deg}(C_i)}{2*\left(m-int_{deg}(C_i)\right)+ext_{deg}(C_i)}$$

$$(C_i)(G|C_1, C_2..C_k) = \frac{1}{k}\sum_{i=1}^{k}(NCut(C_i))$$

Cut-Ratio

$$CR(C_i) = \frac{ext_{deg}(C_i)}{C_n(n-C_n)}\ CR(G|C_1, C_2..C_k) = \frac{1}{k}\sum_{i=1}^{k}CR(C_i)$$

Graph Clustering Algorithm Based on Clique Strategy



Relative Density Metric/ User-Specified Relative Density $U(\delta_r)$

Equation = $C_i(\delta_r) = \frac{\sum_{v \in C_i} int_{deg}(v)}{\sum_{v \in C_i} int_{deg}(v) + ext_{deg}(C_i)}$ **(Value ranges from 0 to 1)**

Phase 1: Algorithm-1 → Cliques Determination in Graph (*G*)

Input : Connected Un-weighted and Un-directed Graph

- **Spanning Tree**
- **Fundamental Cycles**

Outcome → Cliques (Triangles) and their associated degrees detection

Phase 2: Algorithm-2 →Cluster Determination

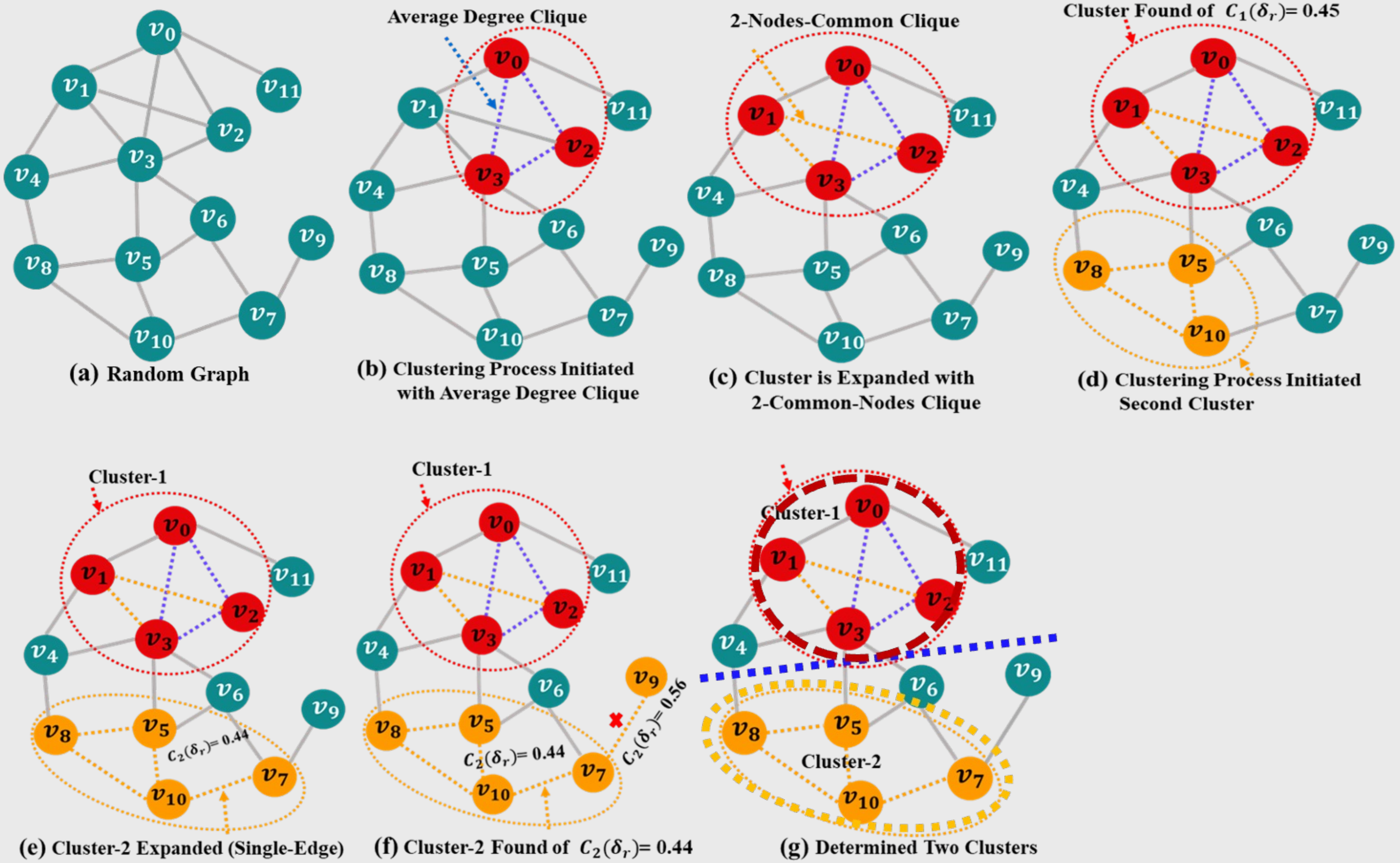
- **User-Specified Relative Density** $U(\delta_r)$
- **Connected Un-weighted and Un-directed Graph, Clique List**

Outcome → Clusters with equal or Closer to the user desired Density $U(\delta_r)$

Algorithm Illustration with a Simple Random Graph Example

Input: Random Graph

V = 12
E = 19
Traingles = 7
$U(\delta_r) = 0.45$



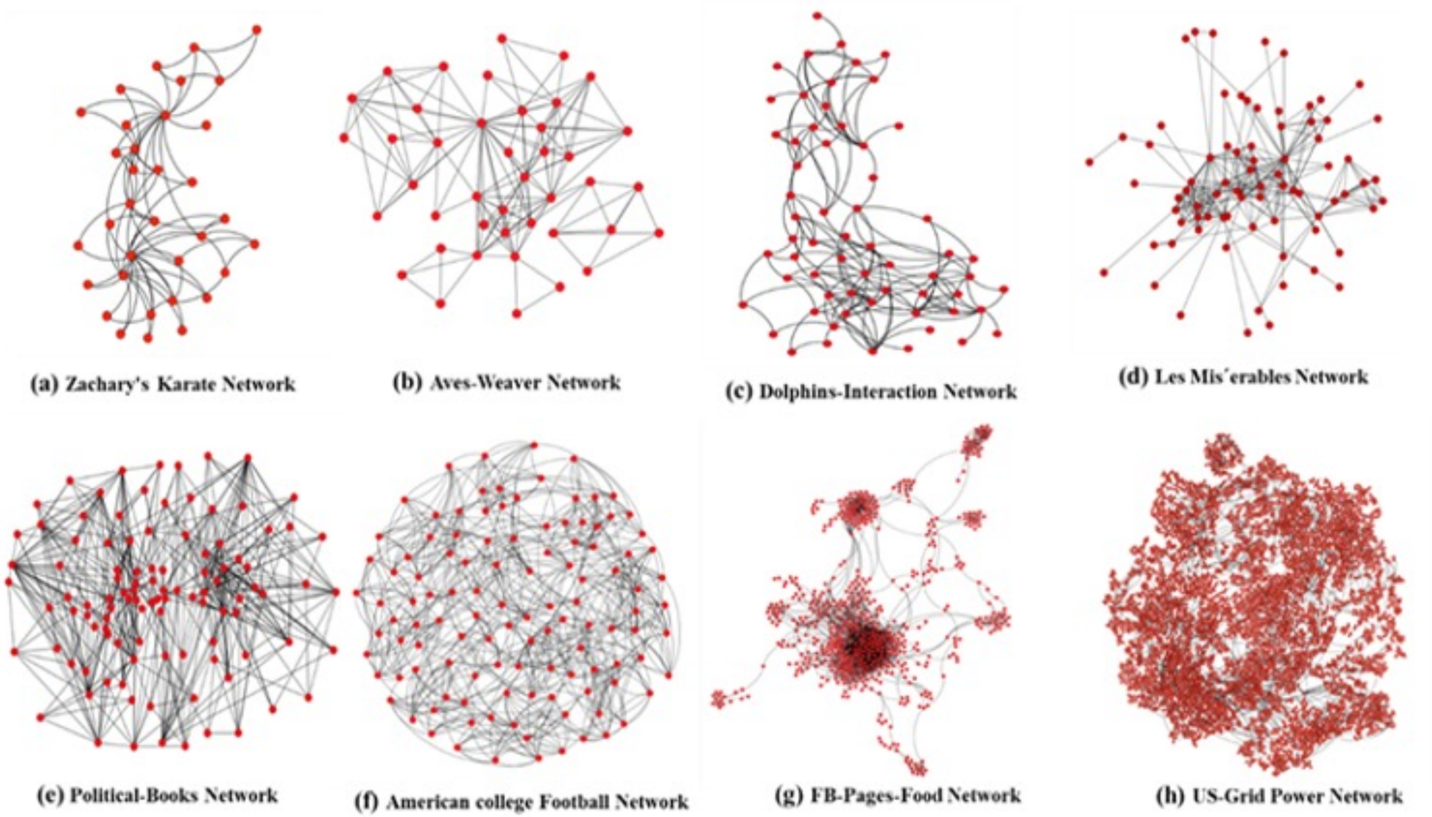
- **8 Real-world networks have been experimented with in this study**

- **Diverse properties and connectivity structures.**

1. Zachary's Karate Club
2. American College Football
3. US-Grid Power Network

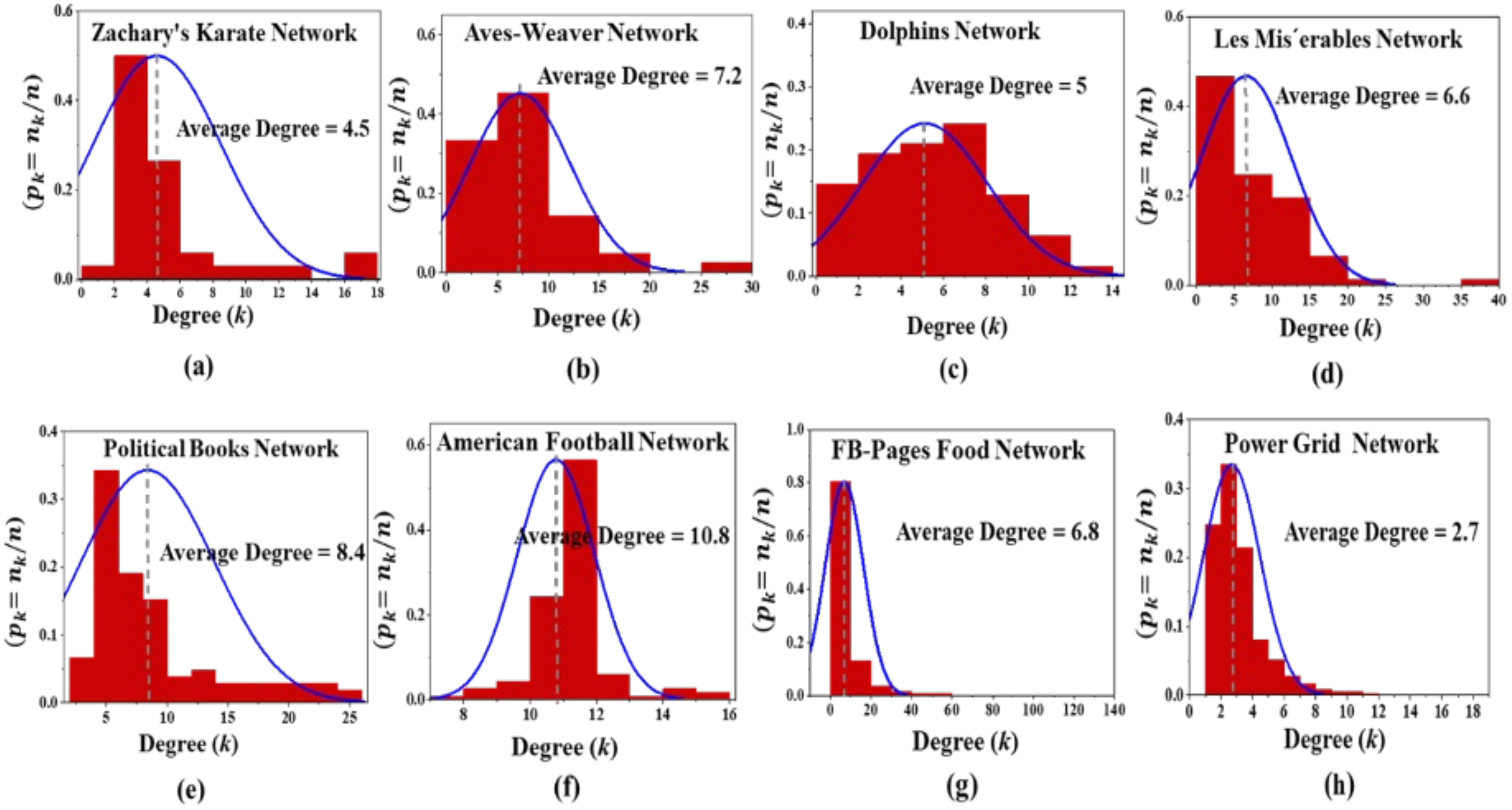
Table 1: Summarized descriptive statistical characteristic of the real-world networks used to assess the performance of the proposed algorithm

Network	$\delta(G)$	V	E	Min_{deg}	Max_{deg}	Avg_{deg}	(σ)	$G(K)$
Zachary's Karate	0.139	34	78	1	17	4.5	3.8	45
Aves-Weaver	0.176	42	152	2	27	7.2	4.9	287
Dolphins-Interaction	0.084	62	159	1	12	5.1	2.9	95
Les Mis´erables	0.086	77	254	1	36	6.6	6.0	467
Political-Books	0.080	105	441	2	25	8.4	5.4	560
American College Football	0.093	115	613	7	12	10.8	0.8	810
FB-Pages-Food	0.010	620	2102	1	134	6.8	9.4	2935
US-Grid Power	0.0005	4941	6594	1	19	2.7	1.7	651

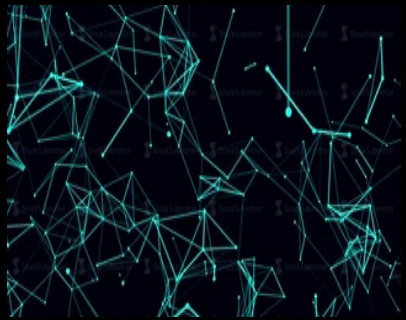

(a) Zachary's Karate Network (b) Aves-Weaver Network (c) Dolphins-Interaction Network (d) Les Mis´erables Network
(e) Political-Books Network (f) American college Football Network (g) FB-Pages-Food Network (h) US-Grid Power Network

Connectivity structures and Degree Distribution of the networks

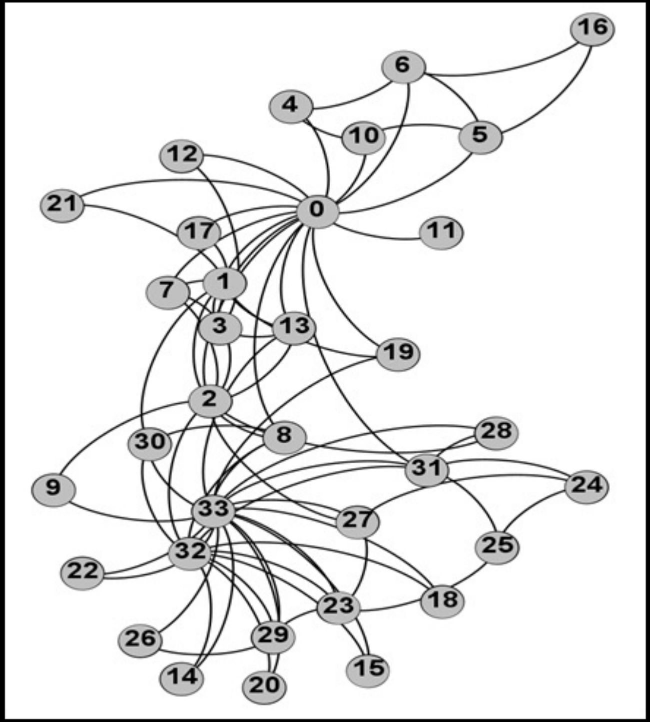
Depicts the relative frequency distribution of 8 real-world networks which tell us about the characteristic's comprehension. (p_k) is relative frequency and can also be thought of as a probability, the likelihood that a node has a degree of exactly k. n_k represents the number of nodes with degree (k), and n is the count of nodes in a graph.



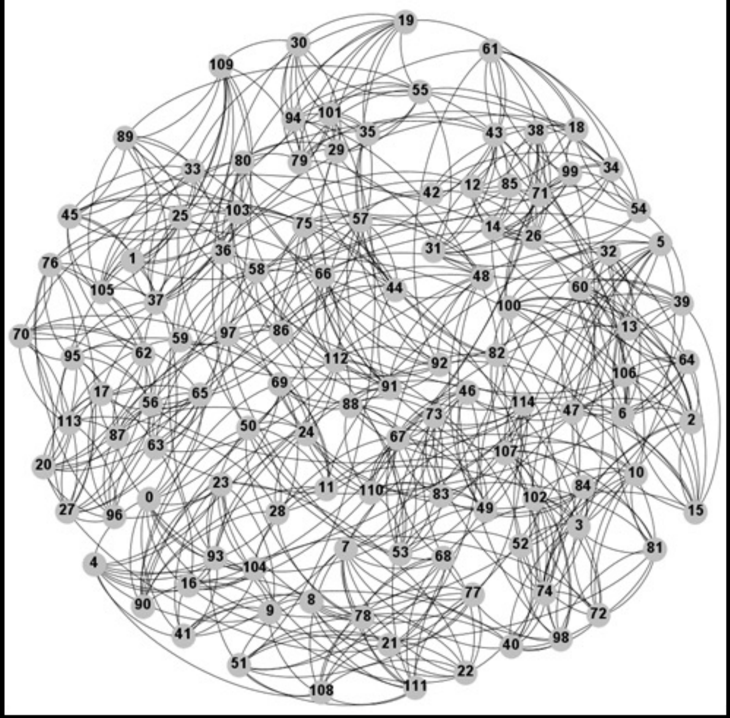
Results (Real-world Networks)



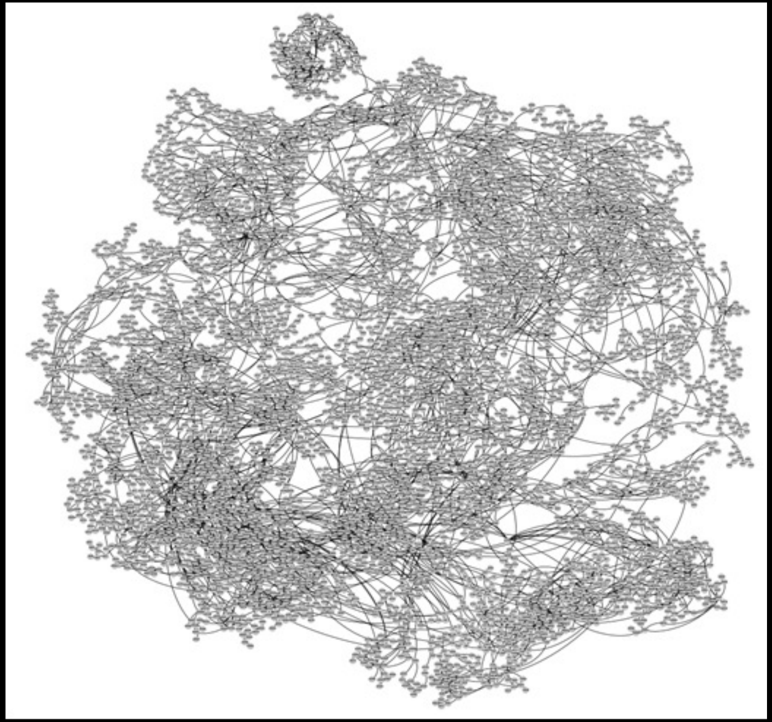
- Real structures of three networks: (a) Zachary's Karate Club, (b) American College Football, and (c) US Grid Power.
- The average degree of connectivity is 4.5, 10.8, and 2.7 respectively.



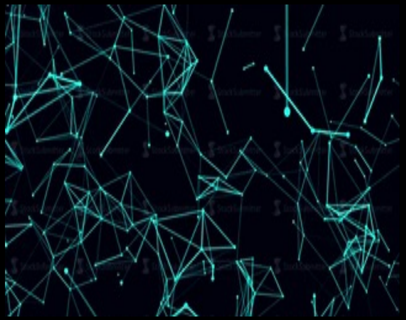
(b) Zachary's Karate Real Network



(b) American college Football Real- Network

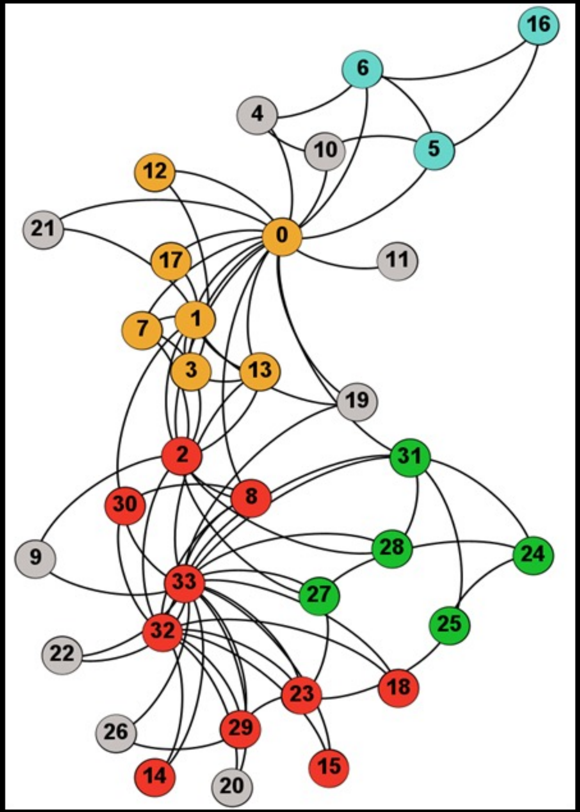


(c) US-Grid Power Real Network



Results (Real-world Networks):

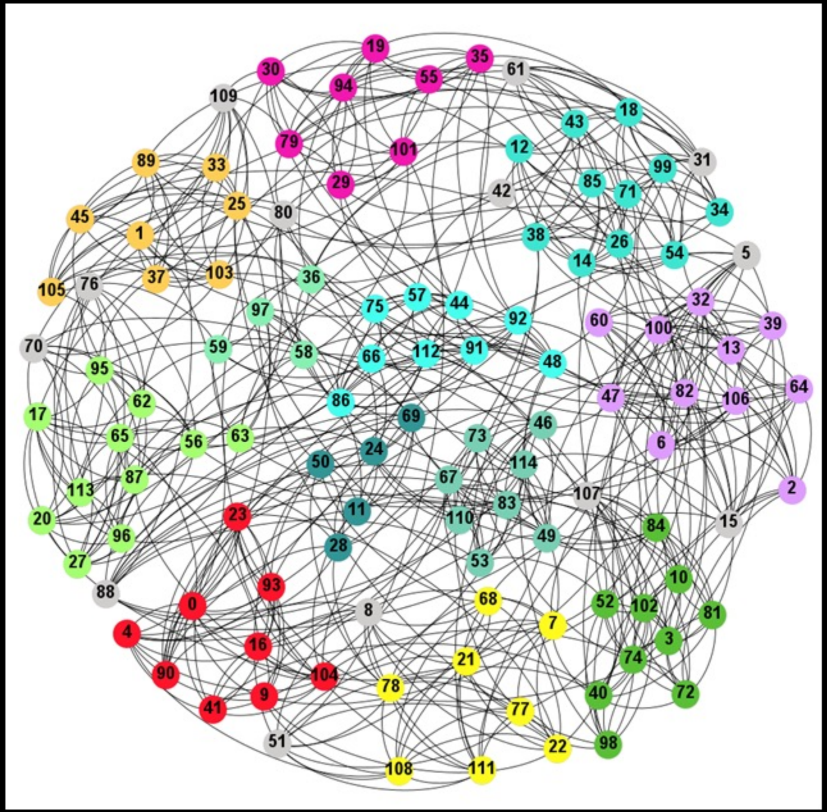
Clustering results with a $U(\delta_r)$ value of 0.4



(a) Zachary's Karate Network

$U(\delta_r) = 0.4$

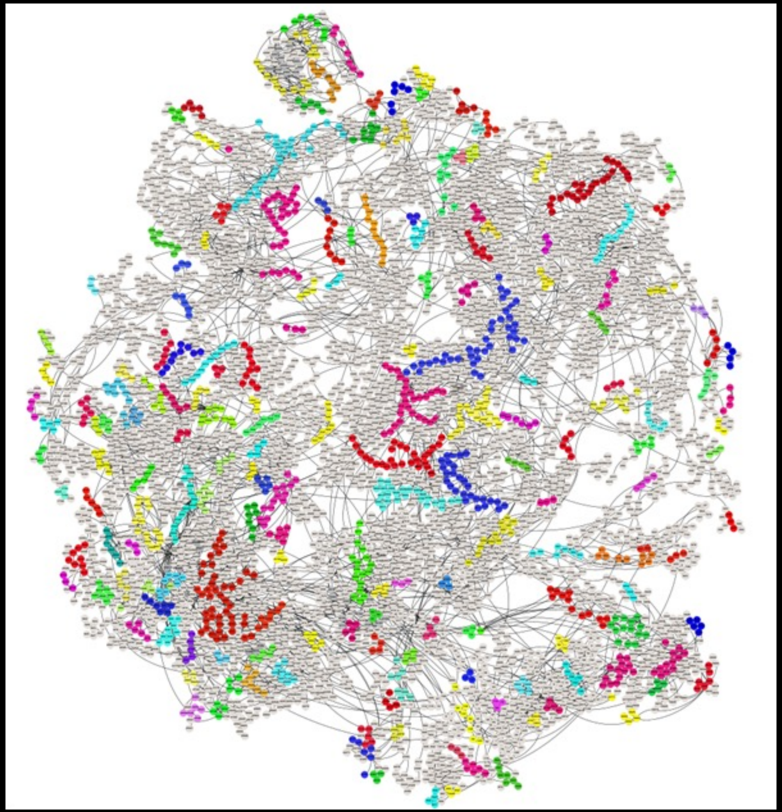
Identified 4 Clusters



(b) American college Football Network

$U(\delta_r) = 0.4$

Identified 12 Clusters



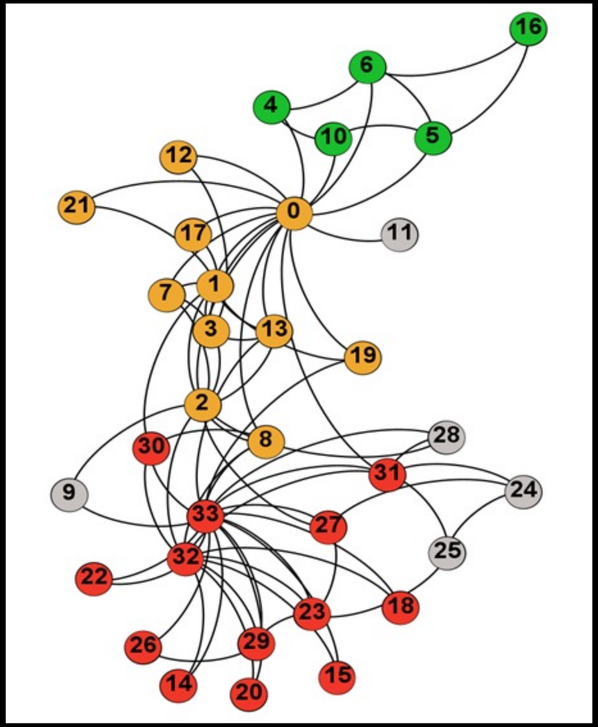
(c) US-Grid Power Network

$U(\delta_r) = 0.4$

Identified 180 Clusters

Results (Real-world Networks):

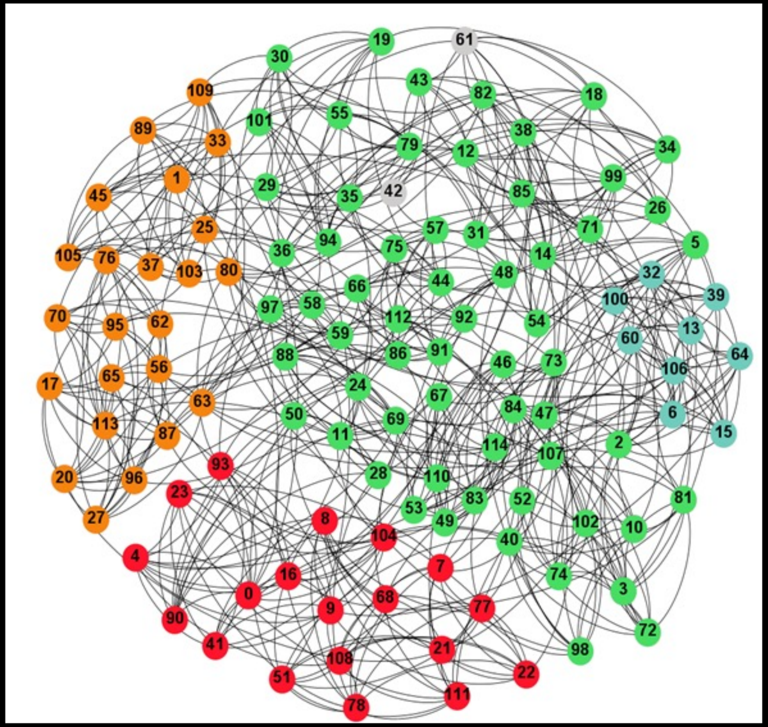
Clustering results with a $U(\delta_r)$ value of 0.6



Zachary's Karate Network

$U(\delta_r) = 0.6$

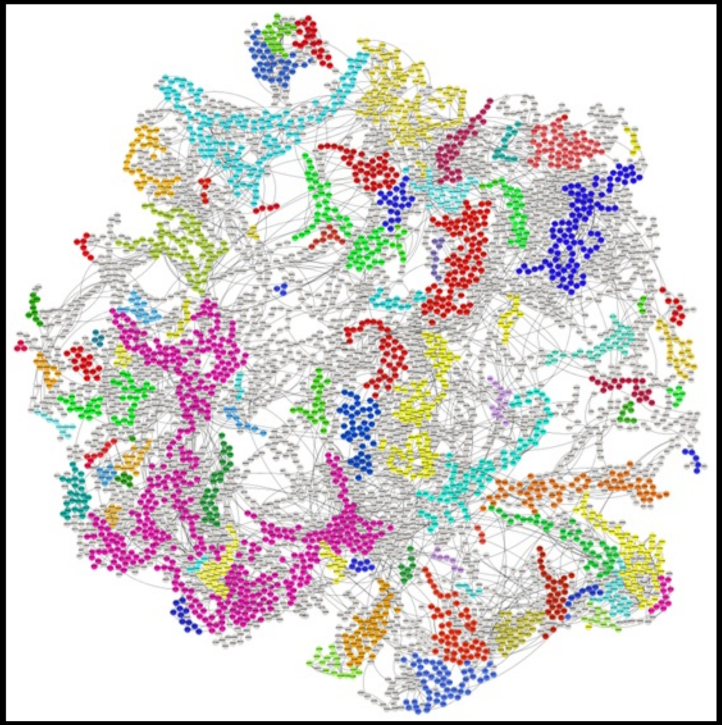
Identified 3 Clusters



American college Football Network

$U(\delta_r) = 0.6$

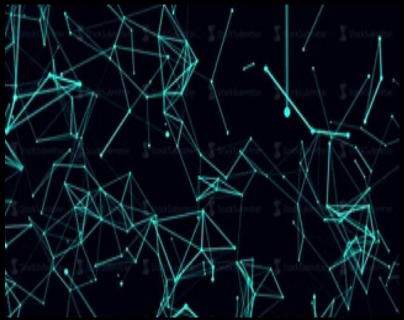
Identified 4 Clusters



US-Grid Power Network

$U(\delta_r) = 0.6$

Identified 84 Clusters

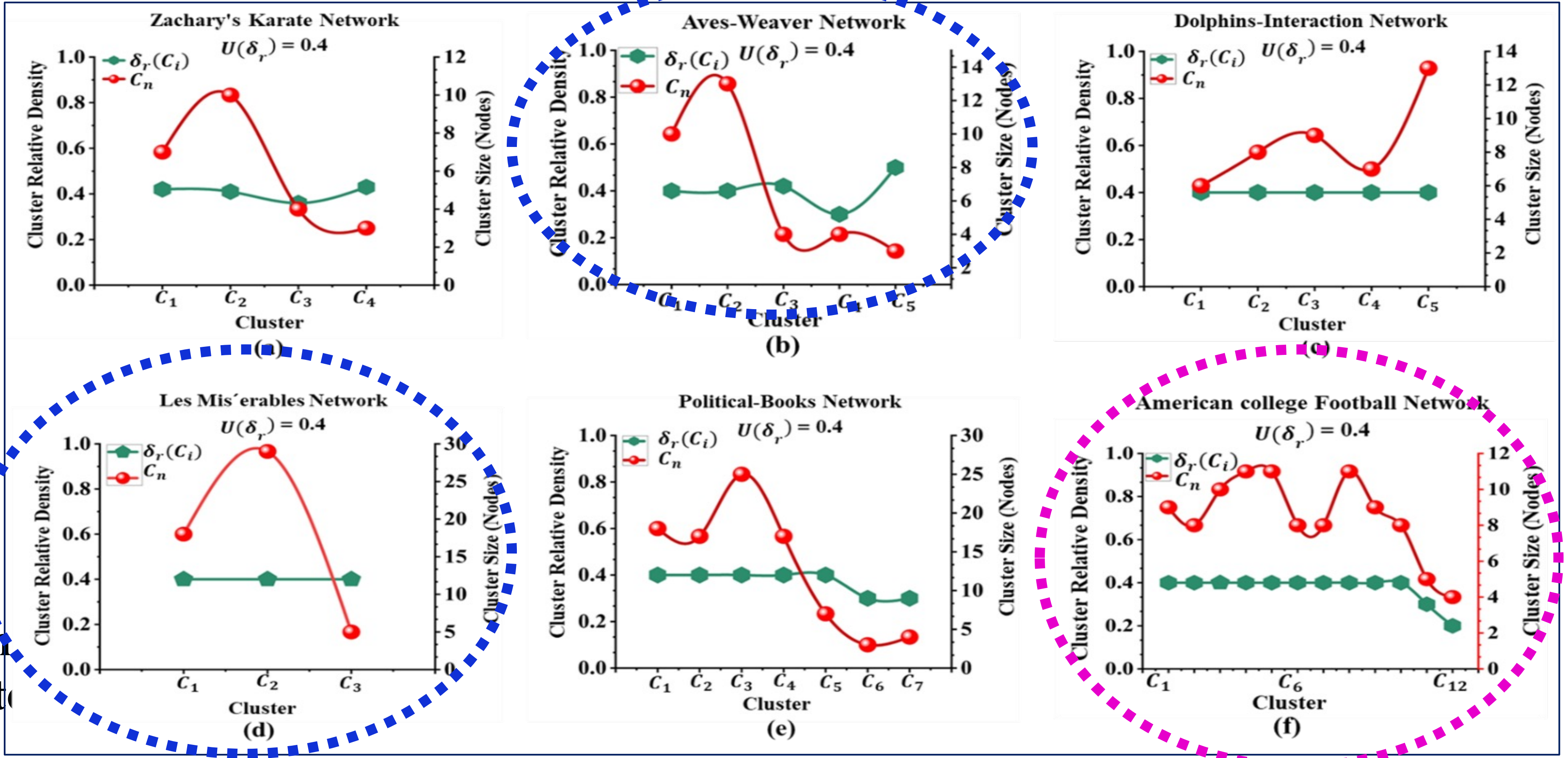


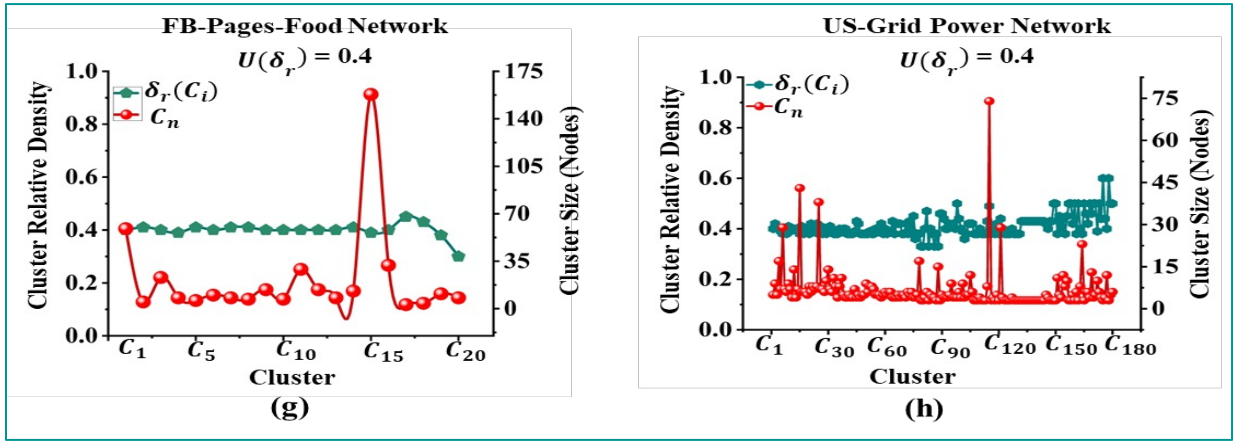
Results (Real-world Networks)

	User-Specified Relative Density $U(\delta_r)$ of 0.4		
	Dataset	**No of Clusters (C_k)**	**Average Relative Density $Avg\_C(\delta_r)$**
1	Zachary's Karate	4	0.4
2	American College Football	12	0.4
3	US-Grid Power	180	0.4
	User-Specified Relative Density $U(\delta_r)$ of 0.6		
1	Zachary's Karate	3	0.6
2	American College Football	4	0.6
3	US-Grid Power	84	0.6

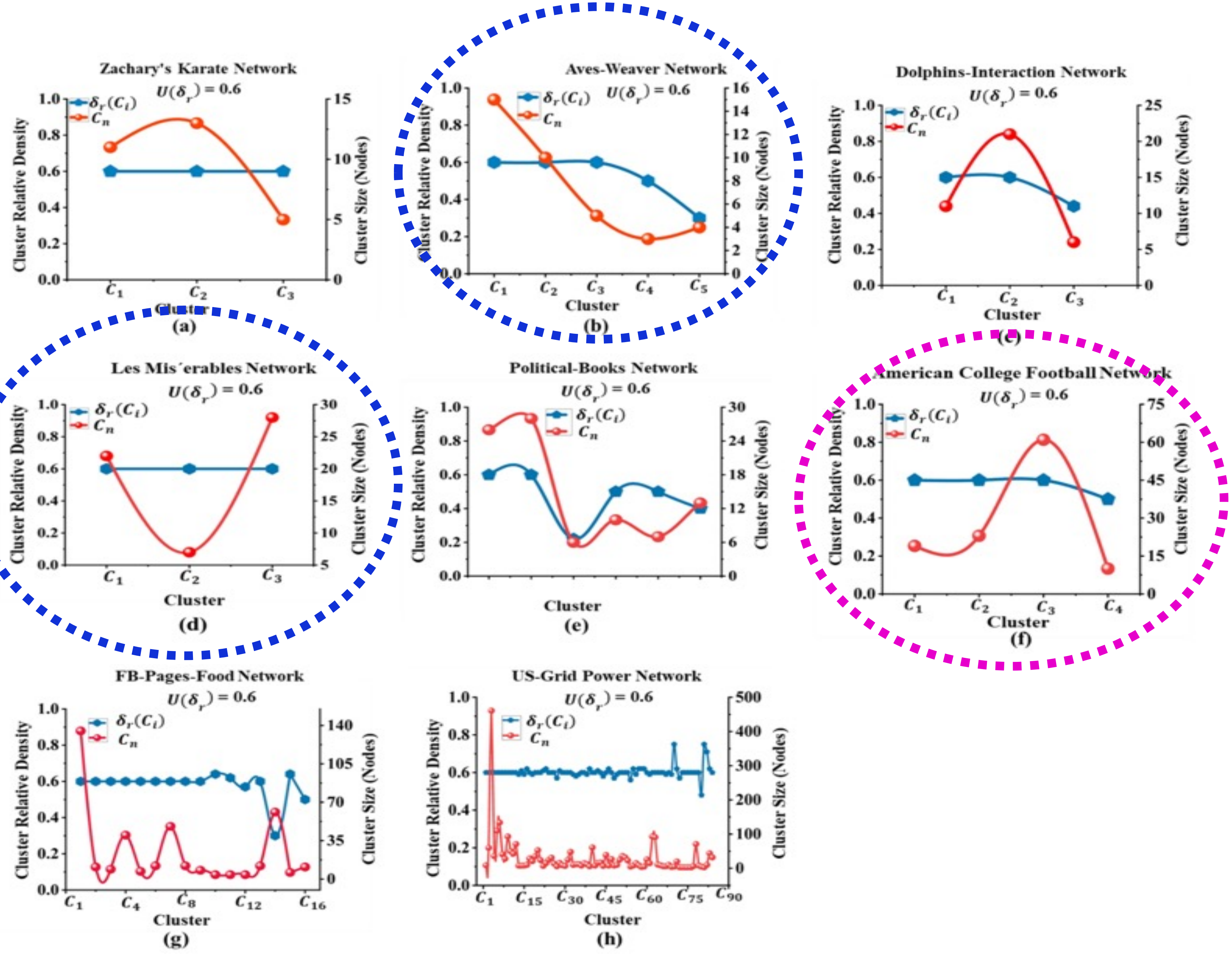
- $U(\delta_r)$ value of 0.4

- Graph structural and clustering type impact on cluster size with respect to number of nodes.





- $U(\delta_r)$ value of 0.6
- **Graph Structural and clustering type impact on cluster size with respect t[o] number of nodes.**



Algorithm Effectiveness and Process Evaluation with the Existing Quality Metrics:

- **Clustering Results Evaluation based on Internal Connectivity**

Resultant Clusters External Connectivity Evaluation

Network	Conductance		Cut-Ratio		Nor-Cut		Flake-Odf	
	$U(\delta_r)$		$U(\delta_r)$		$U(\delta_r)$		$U(\delta_r)$	
	0.4	**0.6**	**0.4**	**0.6**	**0.4**	**0.6**	**0.4**	**0.6**
Zachary's Karate	0.41	0.25	0.08	0.04	0.5	0.3	0.2	0.05
Aves-weaver	0.42	0.34	0.09	0.05	0.5	0.4	0.3	0.1
Dolphins-interaction	0.4	0.27	0.05	0.03	0.4	0.3	0.4	0.05
Les Misérables	0.42	0.22	0.06	0.03	0.5	0.28	0.2	0.04
Political-Books	0.47	0.33	0.03	0.02	0.5	0.4	0.3	0.2
American college Football	0.42	0.25	0.04	0.02	0.4	0.3	0.2	0.04
FB-Pages-Food	0.43	0.26	0.006	0.005	0.45	0.2	0.18	0.07
US-Grid power	0.41	0.37	0.003	0.0015	0.4	0.23	0.1	0.05

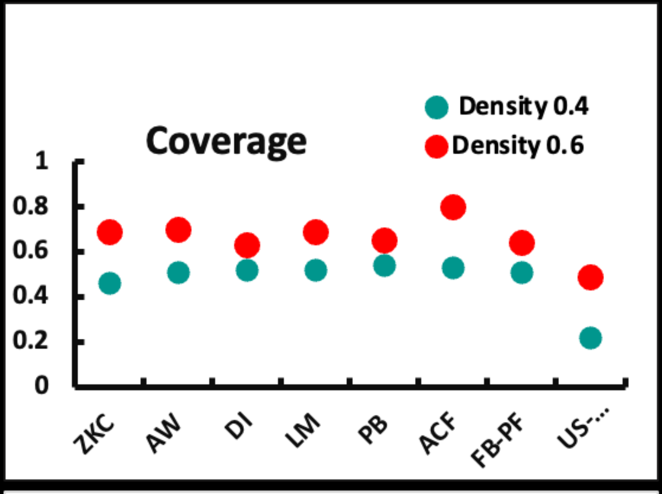
- **Clustering Results Evaluation based on External Connectivity**

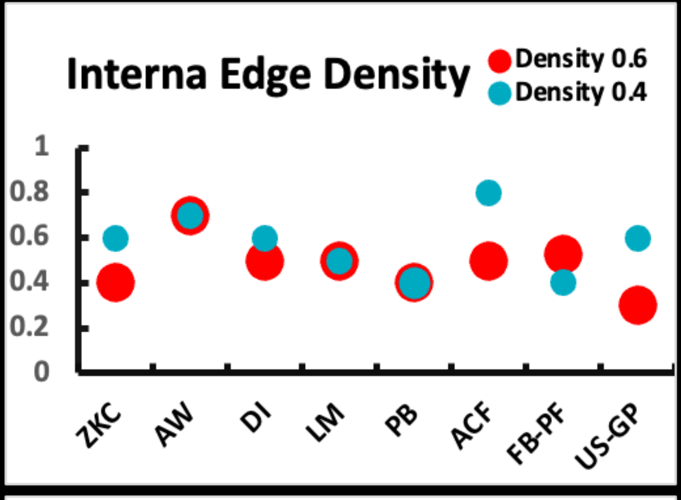
Resultant Clusters Internal Connectivity Evaluation

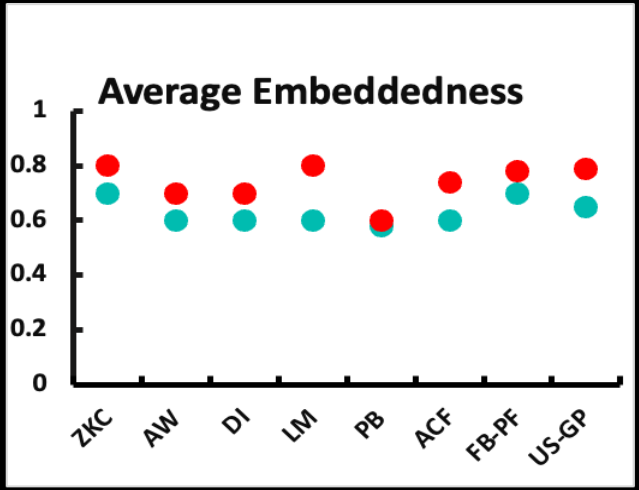
Network	Coverage		Average Embeddedness		FOMD		Average TPR		Internal Edge Density	
	$U(\delta_r)$									
	0.4	**0.6**	**0.4**	**0.6**	**0.4**	**0.6**	**0.4**	**0.6**	**0.4**	**0.6**
Zachary's Karate	0.46	0.69	0.7	0.8	0.3	0.4	0.91	0.9	0.6	0.4
Aves-weaver	0.51	0.7	0.6	0.7	0.2	0.4	0.95	0.95	0.7	0.7
Dolphins-interaction	0.52	0.63	0.6	0.7	0.3	0.3	0.9	1	0.6	0.5
Les Misérables	0.52	0.69	0.6	0.8	0.29	0.3	0.9	0.8	0.5	0.5
Political-Books	0.54	0.65	0.58	0.6	0.4	0.29	0.9	0.9	0.4	0.4
American college Football	0.53	0.8	0.6	0.74	0.17	0.33	1	1	0.8	0.5
FB-Pages-Food	0.51	0.64	0.7	0.78	0.35	0.31	0.8	0.8	0.4	0.53
US-Grid power	0.22	0.49	0.65	0.79	0.3	0.36	0.8	0.5	0.6	0.3

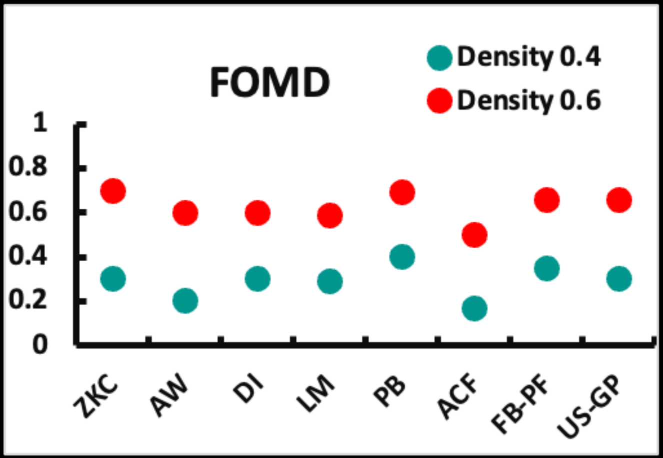
Algorithm Effectiveness and Process Evaluation:

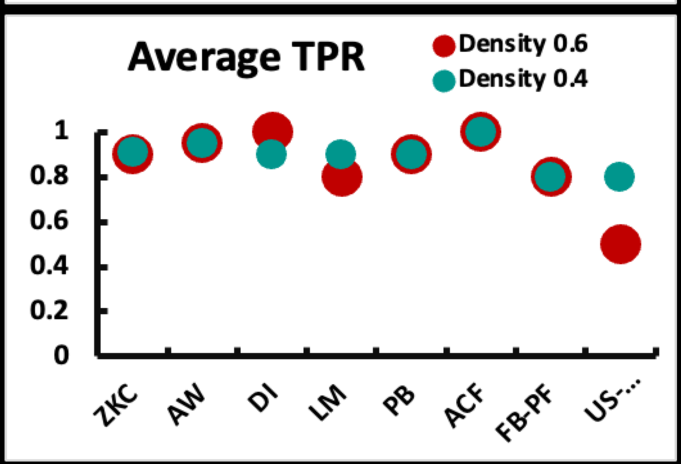
- **Clustering Results Evaluation based on Internal Connectivity**





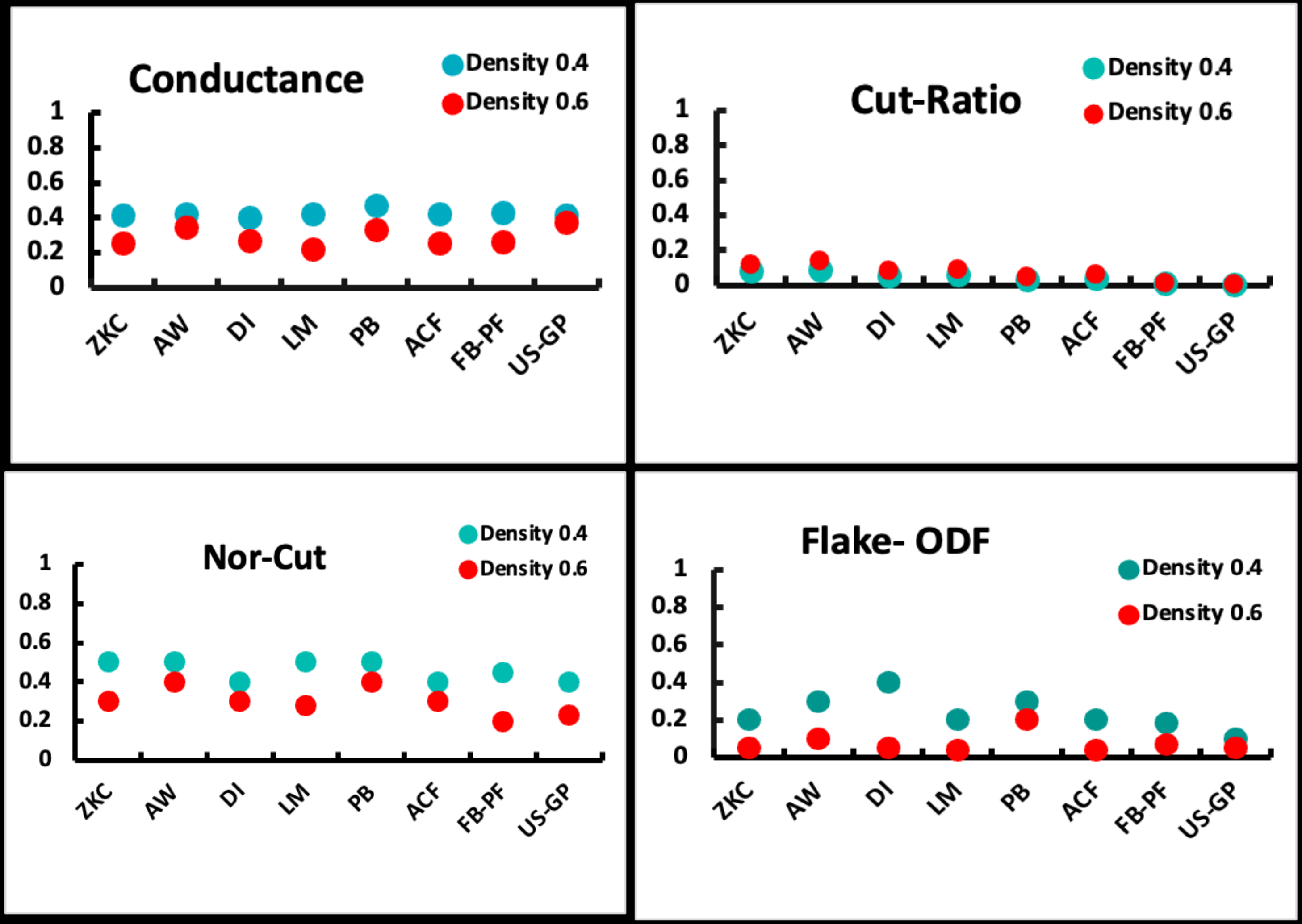






Algorithm Effectiveness and Process Evaluation:

- **Clustering Results Evaluation based on External Connectivity**



Proposed Quality Metric Mean Relative Density Deviation Coefficient (MRDDC):

- The notion of "Mean Relative Density Deviation Coefficient" is defined as a metric (MRDDC) for assessing the proximity between the relative density specified by the user and the cluster calculated density discover by the algorithm.
- The equation below represents the metric's formal mathematical definition.

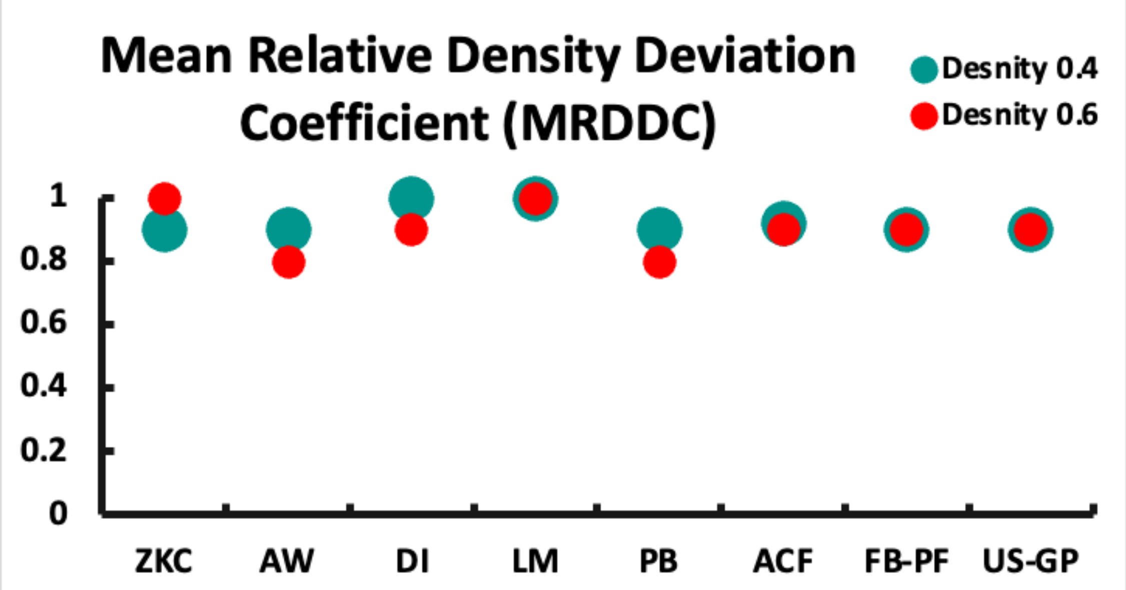
$$\text{MRDDC} = 1 - \frac{1}{k}\sum_{C_i=1}^{C_k} \left| \frac{[U(\delta_r) - C_i(\delta_r)}{U(\delta_r)} \right|$$

- The optimal deviation value for each cluster is zero, while the ideal value for this mean coefficient is 1.

Proposed Quality Metric Mean Relative Density Deviation Coefficient (MRDDC):

The algorithm is not sensitive to structural characteristics variations in the inputs and can still produce meaningful results under a wide range of conditions.

Mean Relative Density Deviation Coefficient (MRDDC)			
	Dataset	$U(\delta_r) = 0.4$ **MRDDC**	$U(\delta_r) = 0.6$ **MRDDC**
1	Zachary's Karate	0.9	1
2	Aves-Weaver	0.9	0.8
3	Dolphins-Interaction	1	0.9
4	Les Misérables	1	1
5	Political-Books	0.9	0.8
6	American College Football	0.9	0.9
7	FB-Pages-Food	0.9	0.9
8	US-Grid Power	0.9	0.9



Continuation of Prior Work

- Overlapping Clustering
- Hierarchal Clustering

2: Edge-based Clustering

- Partitioning Clustering
- Overlapping Clustering
- Hierarchal Clustering

3: Algorithm Sensitivity

Thank You

Q/A