SCAN: A STRUCTURAL CLUSTERING ALGORITHM FOR NETWORKS

Mutlu Mete

Department of Computer Science and Information Systems

Outline

- Introduction
- Structural Clustering Algorithm for Networks (SCAN)
- Applications
- Conclusion

Complex Systems

- Made of many non-identical elements connected by diverse interactions

Complicated

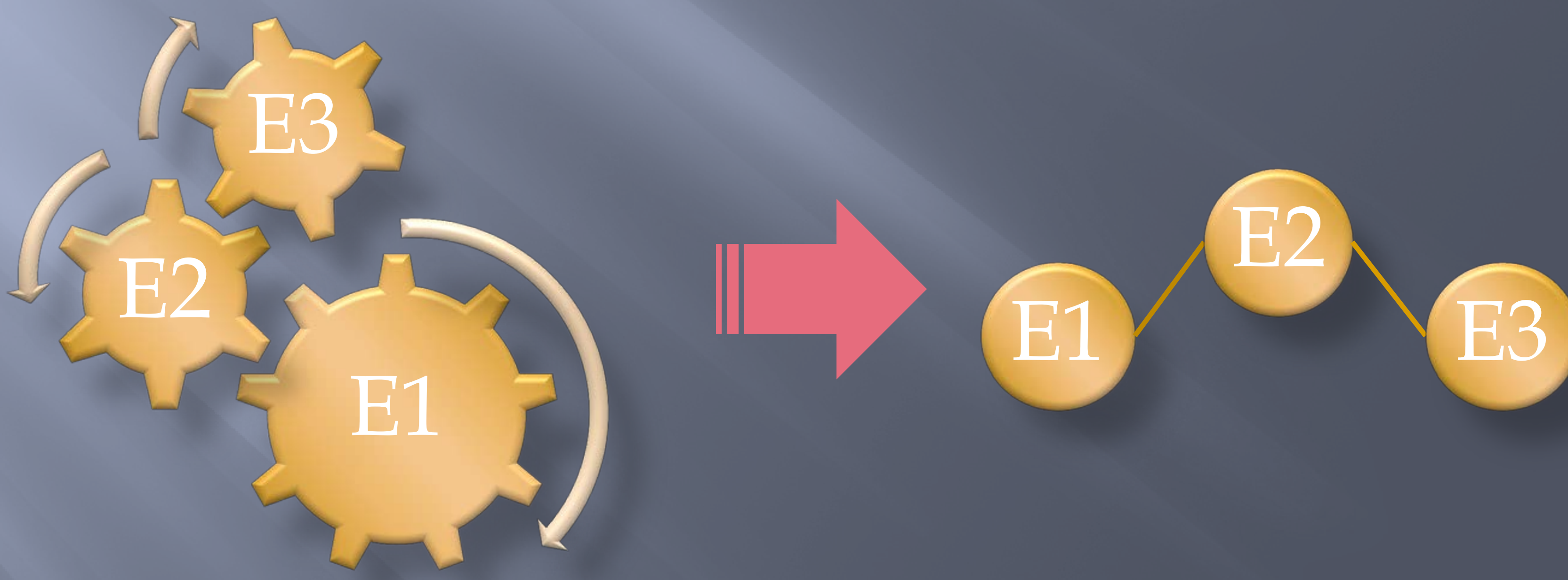
CC credit: jmiguel.rodriguez

Complex

CC credit: krayker

L.A.N. Amarala and J.M. Ottino. Complex networks. Eur. Phys. J. B 38, 147–162 (2004)

Complex Systems
Made of
Many non-identical elements
Connected by diverse interactions

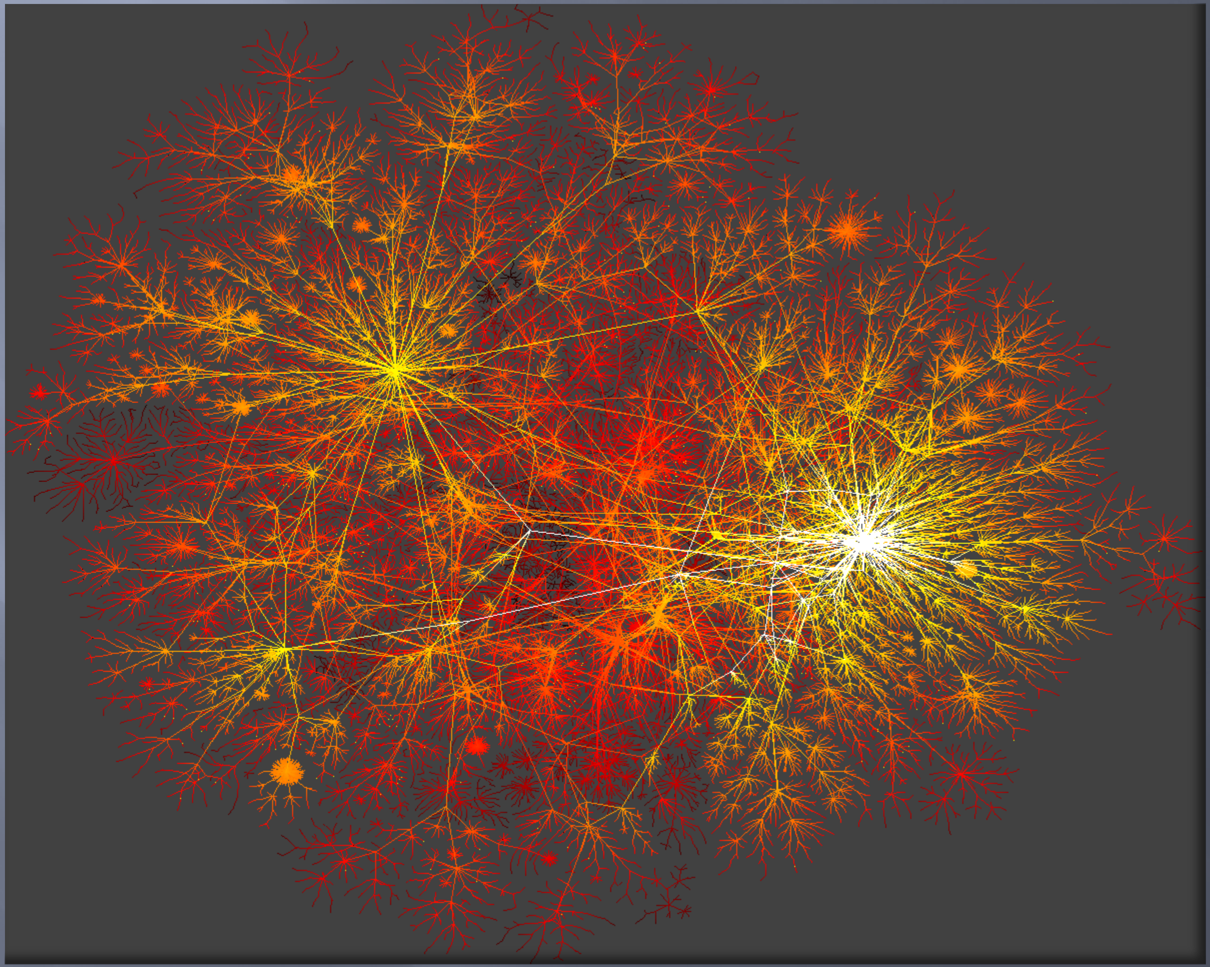


Complex System

Network/Graph

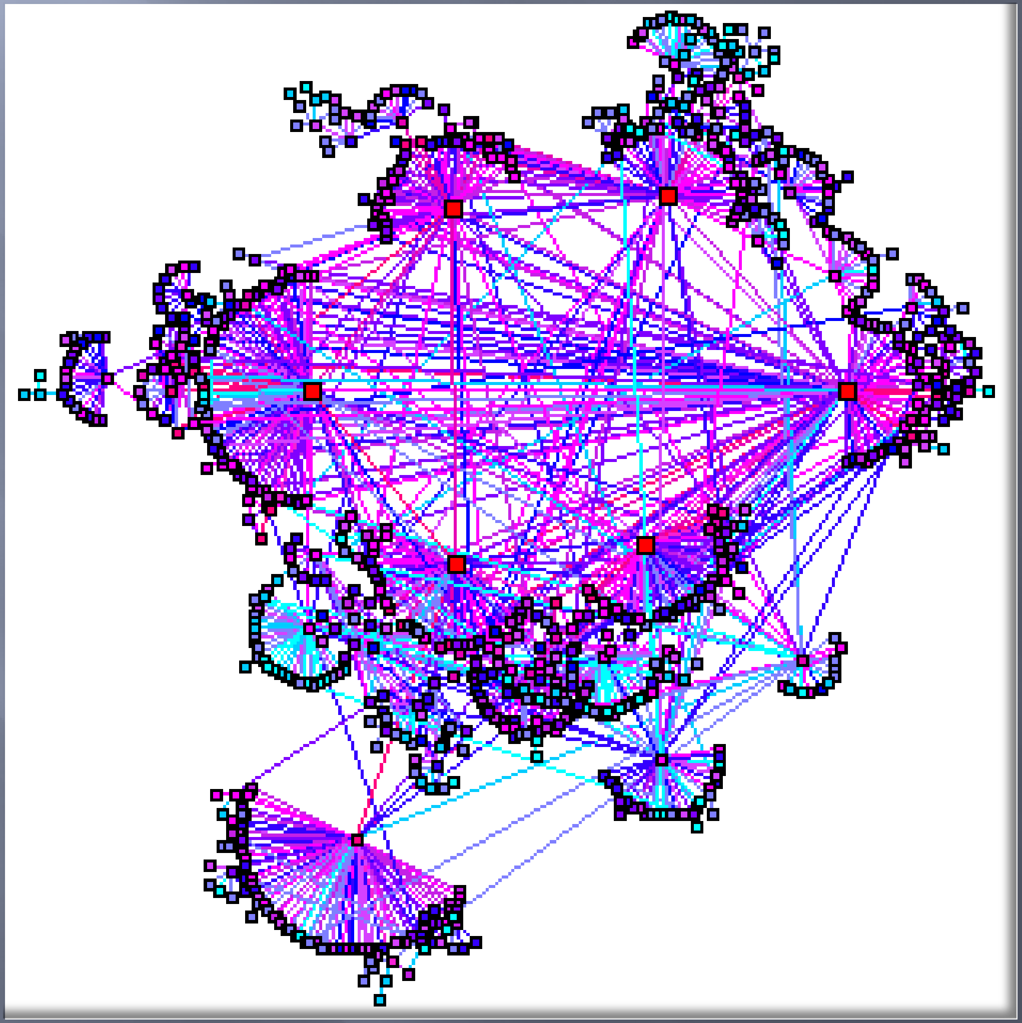
Internet

Nodes: computers
Links: connections



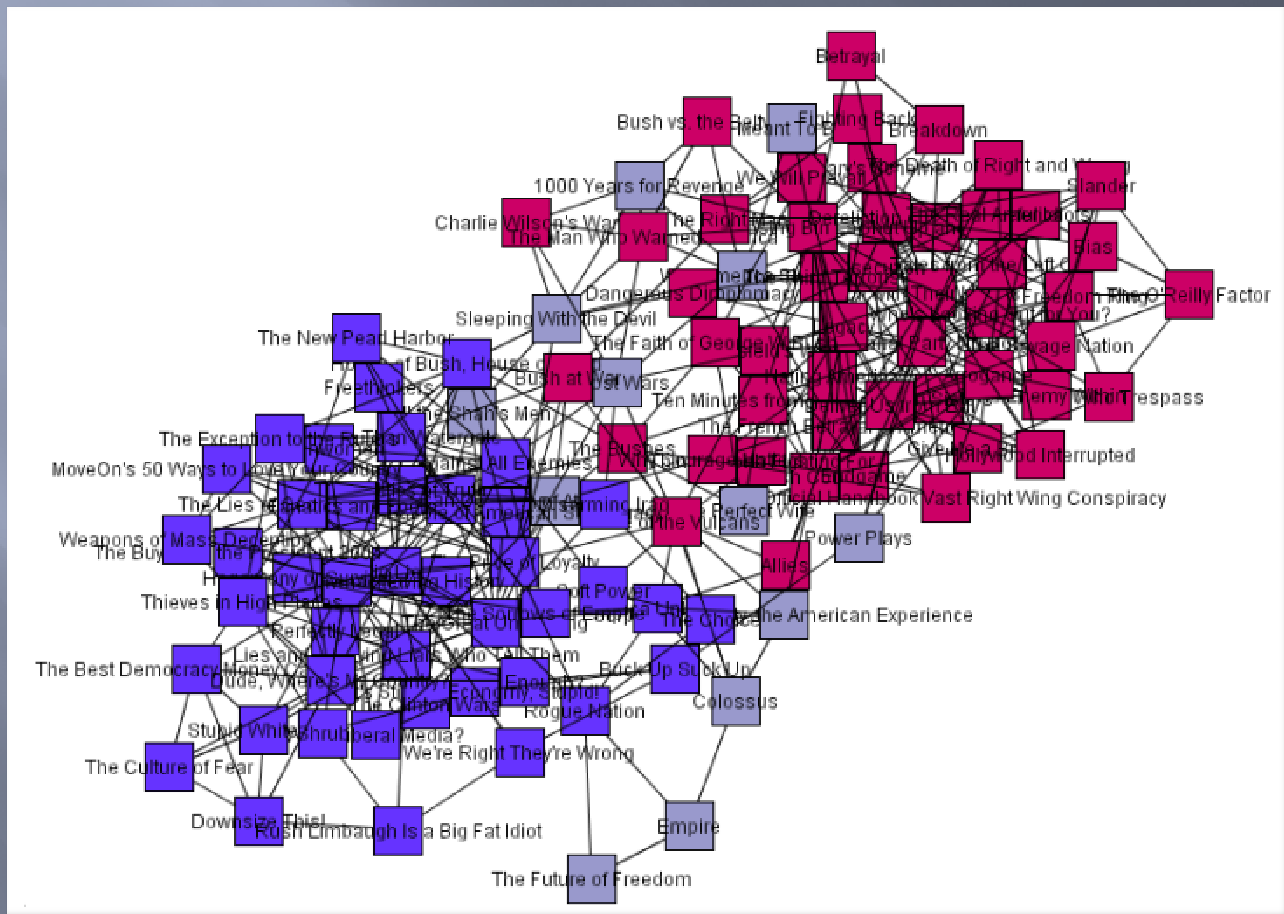
WWW

Nodes: webpages
Links: hyperlinks



Product Networks

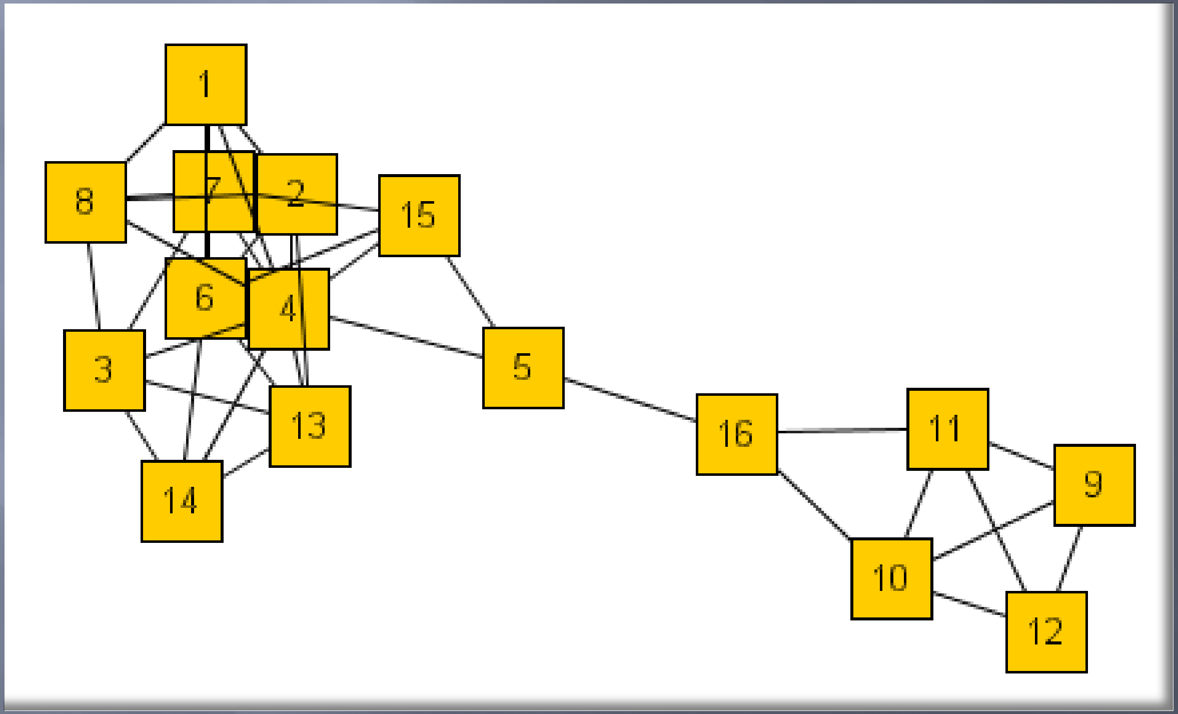
Nodes: products
Links: co-purchased



conservative: red - liberal: blue – neutral: grey

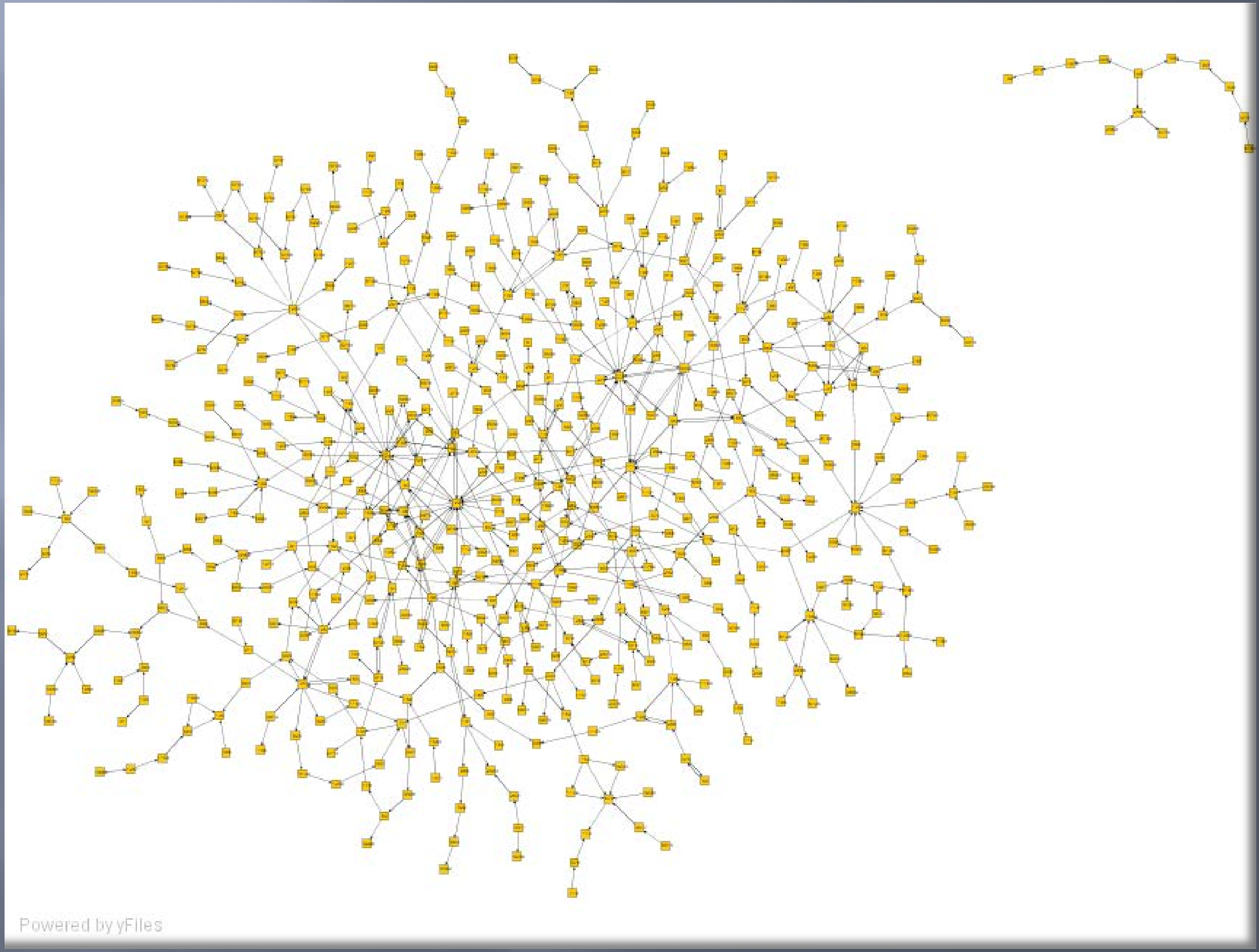
Customer Data Networks

Nodes: customer records
Links: match



Metabolic Network

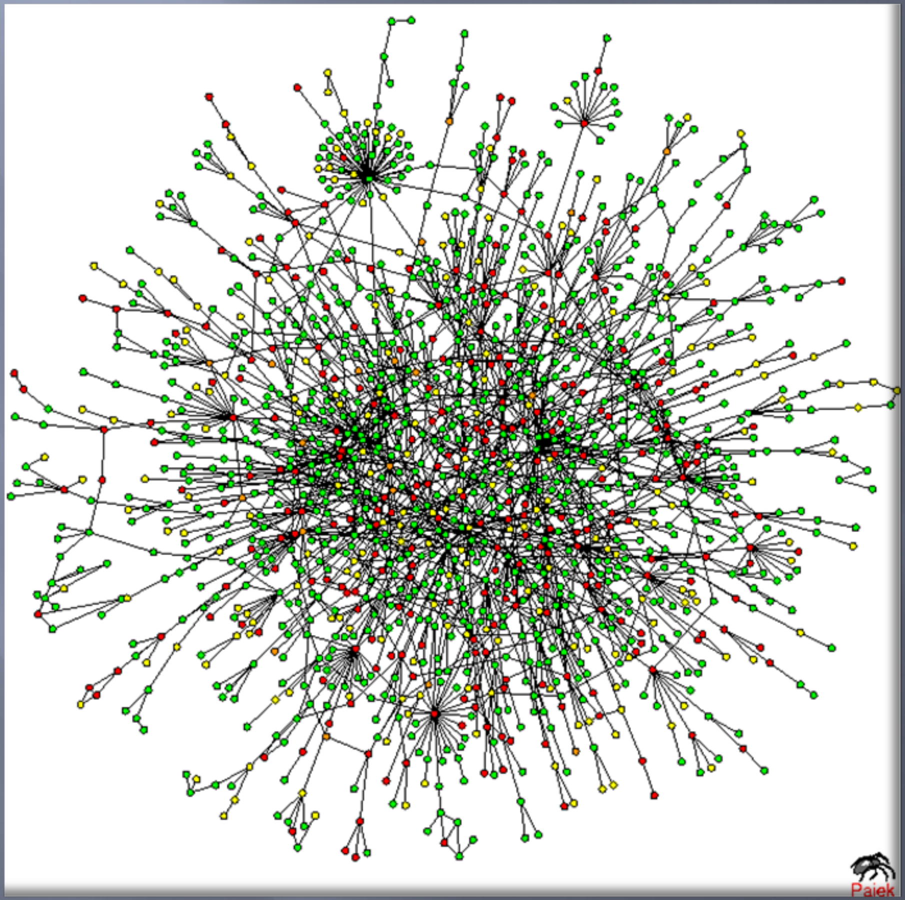
Nodes: chemicals (substrates)
Links: bio-chemical reactions



Protein Interaction Network

Nodes: proteins

Links: physical interactions (binding)



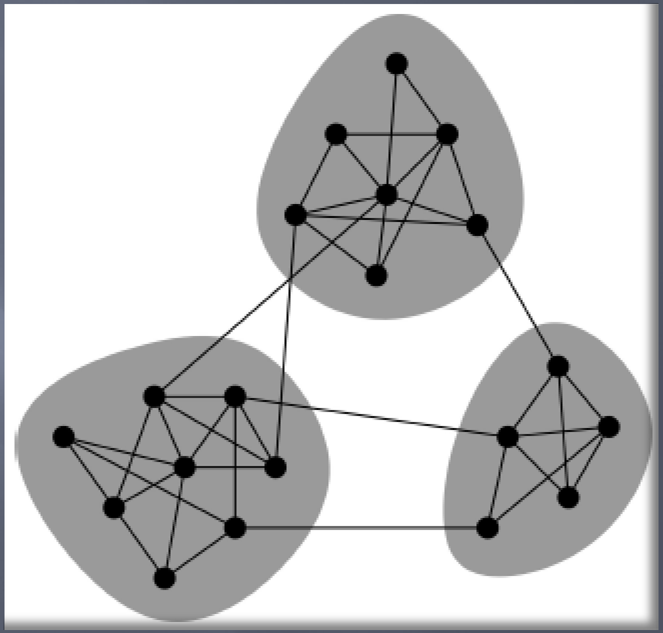
H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Nature 411, 41-42 (2001)

Network Clustering

- The elements form groups, e.g., communities, proteins of similar functions, web pages of similar topics, etc.
- Network clustering is aimed to find such groups or clusters in large networks

Traditional View of Network Clusters

- Network Clusters are densely connected groups of vertices, with only sparser connections between groups in networks
- Finding partition that maximizes intra-cluster links and minimize inter-cluster links is NP-hard problem

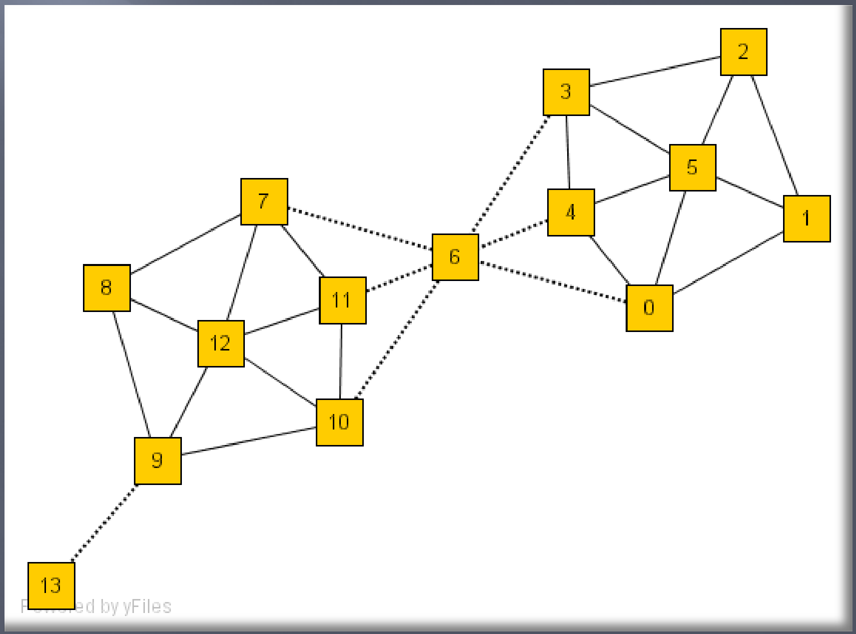


Traditional Algorithms

- They find tightly knit clusters by optimization either
  - Cut, or
  - Modularity

$$Q = \sum_{s=1}^{k} \left[\frac{ls}{L} - \left(\frac{ds}{2L} \right)^2 \right]$$

- They are not scalable
- They fail to identify
  - Hubs
  - Outliers



SCAN: A Structural Clustering Algorithm for Networks

- Structural view of network clusters
- SCAN algorithm
- Complexity
- Evaluation

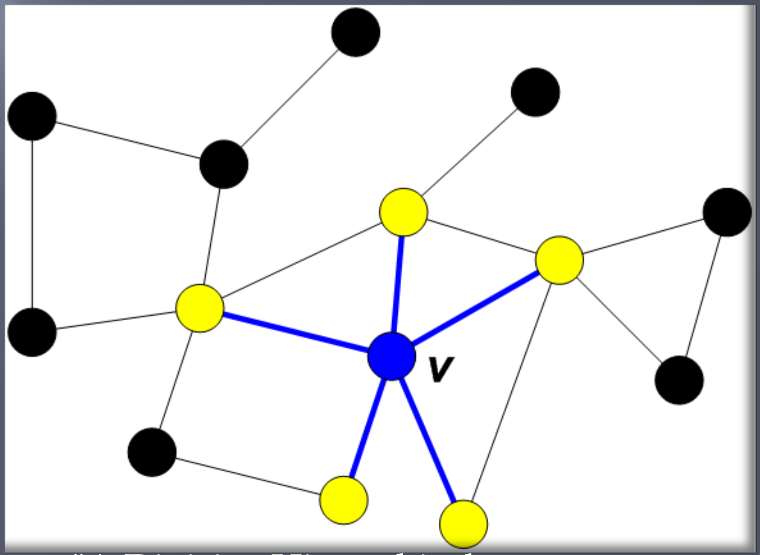
A Structural View of Network Clusters

- Individuals in a tight community (cluster) know many of the same people, regardless of the size of the group.
- Individuals who are hubs know many people in different groups but belong to no single group. Politicians, for example bridge multiple groups.
- Individuals who are outliers reside at the margins of society. Hermits, for example, know few people and belong to no group.

The Neighborhood of a Node

Define $\Gamma(v)$ as the immediate neighborhood of a node (i.e. the set of people that an individual knows).

$$\Gamma(v) = \{w \in V \mid (v, w) \in E\} \cup \{v\}$$



Nurcan Yuruk, **Mutlu Mete**, Xiaowei Xu, and Thomas Schweiger, "A Divisive Hierarchical Structural Clustering Algorithm for Networks", IEEE ICDM Workshop on Mining Graphs and Complex Structures

Structure Similarity

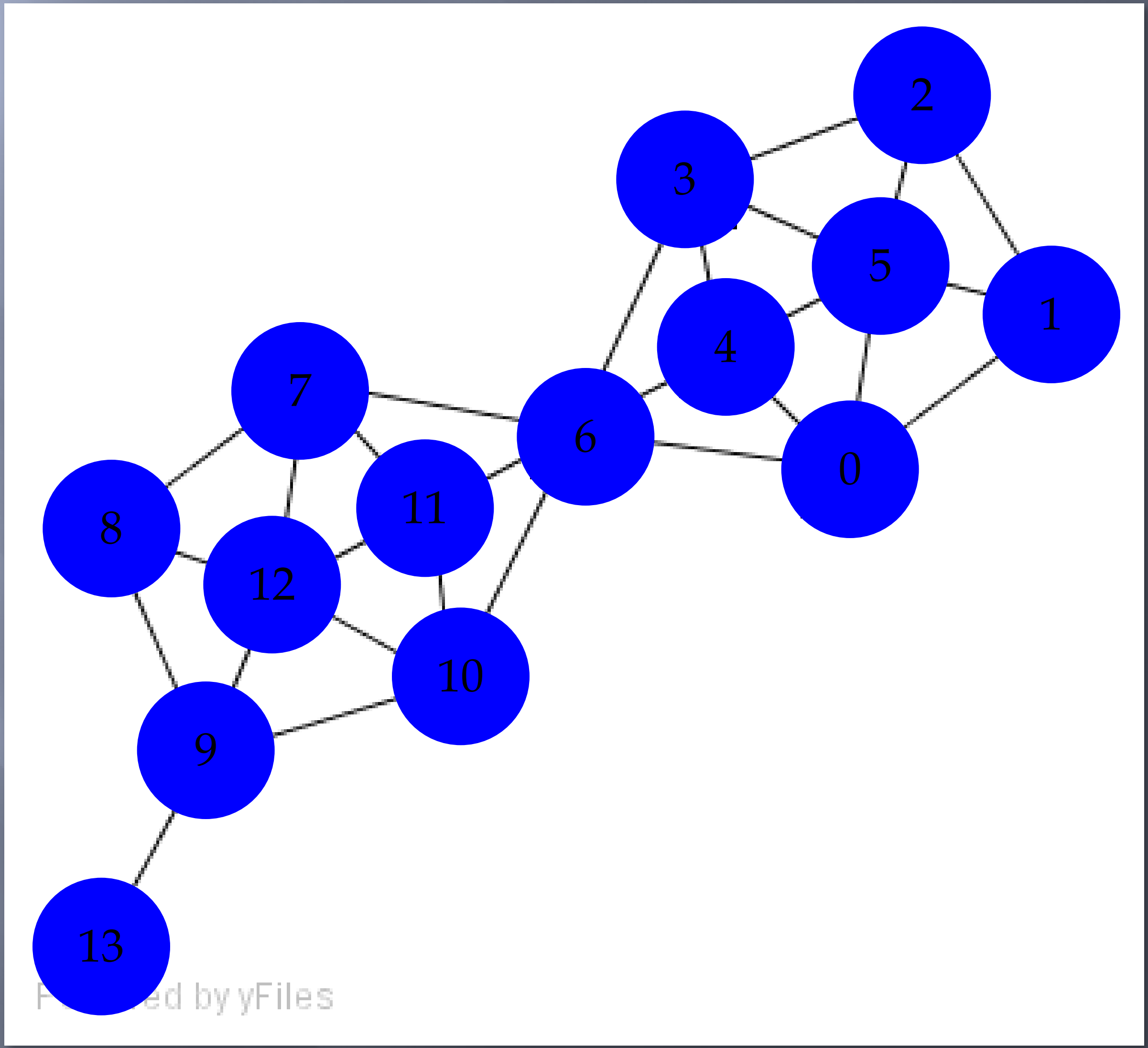
- The desired features tend to be captured by a measure we call Structural Similarity

$$\sigma(v,w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{\sqrt{|\Gamma(v)||\Gamma(w)|}}$$

- Structural similarity tends to be large for members of a cluster and small for hubs and outliers.
- We devised a novel algorithm SCAN (Structural Clustering Algorithm for Networks)

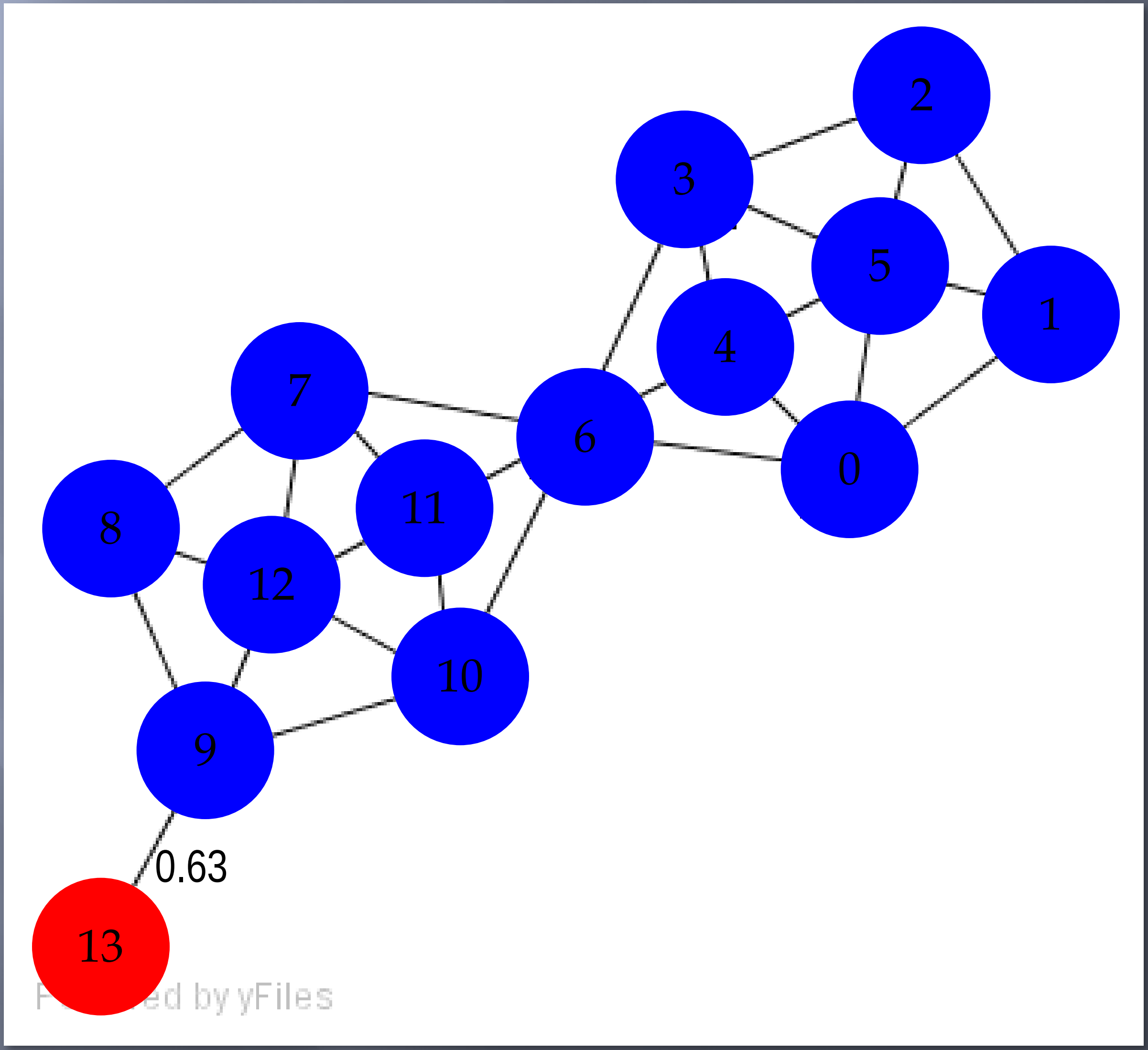
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



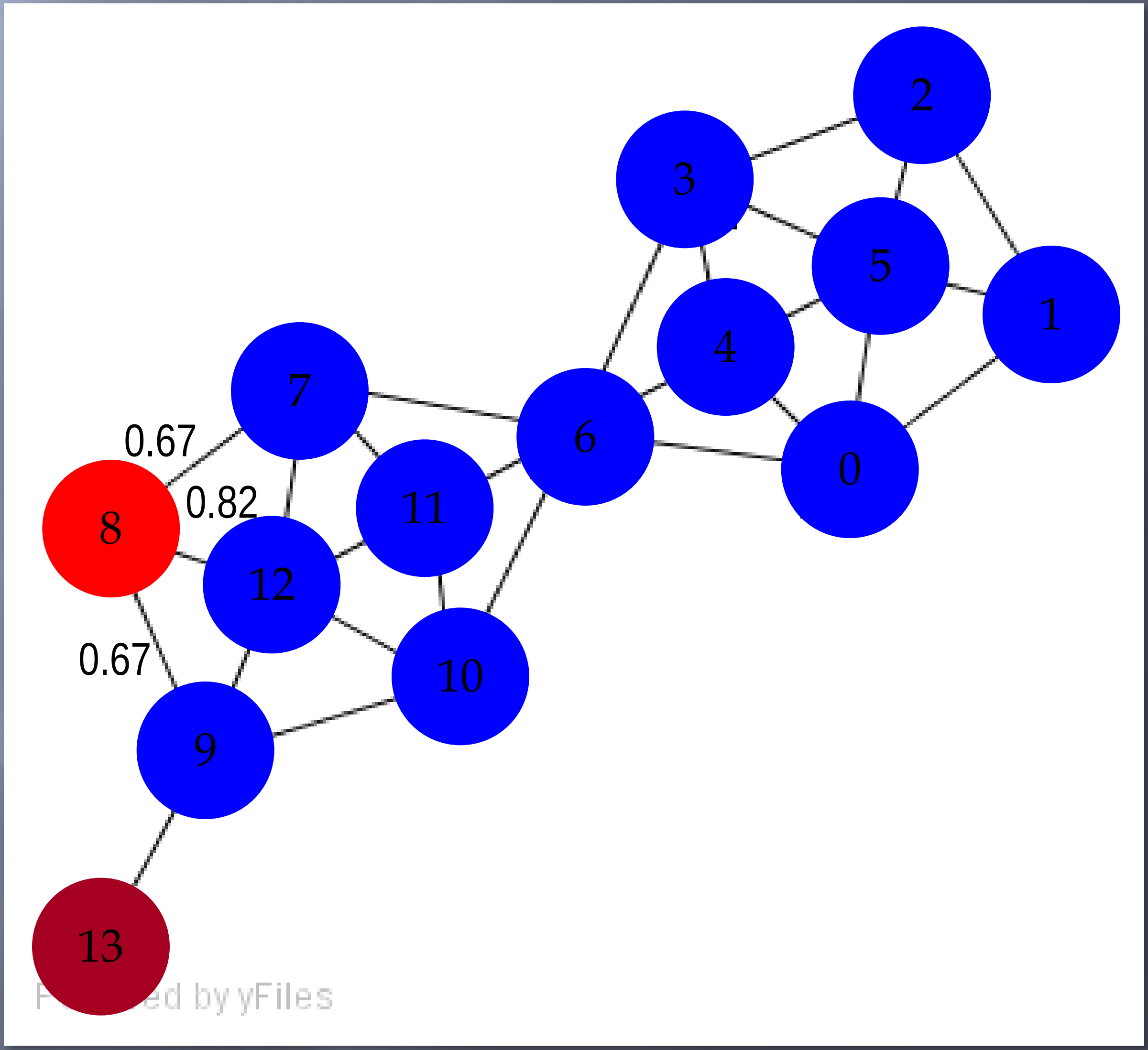
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



SCAN Algorithm

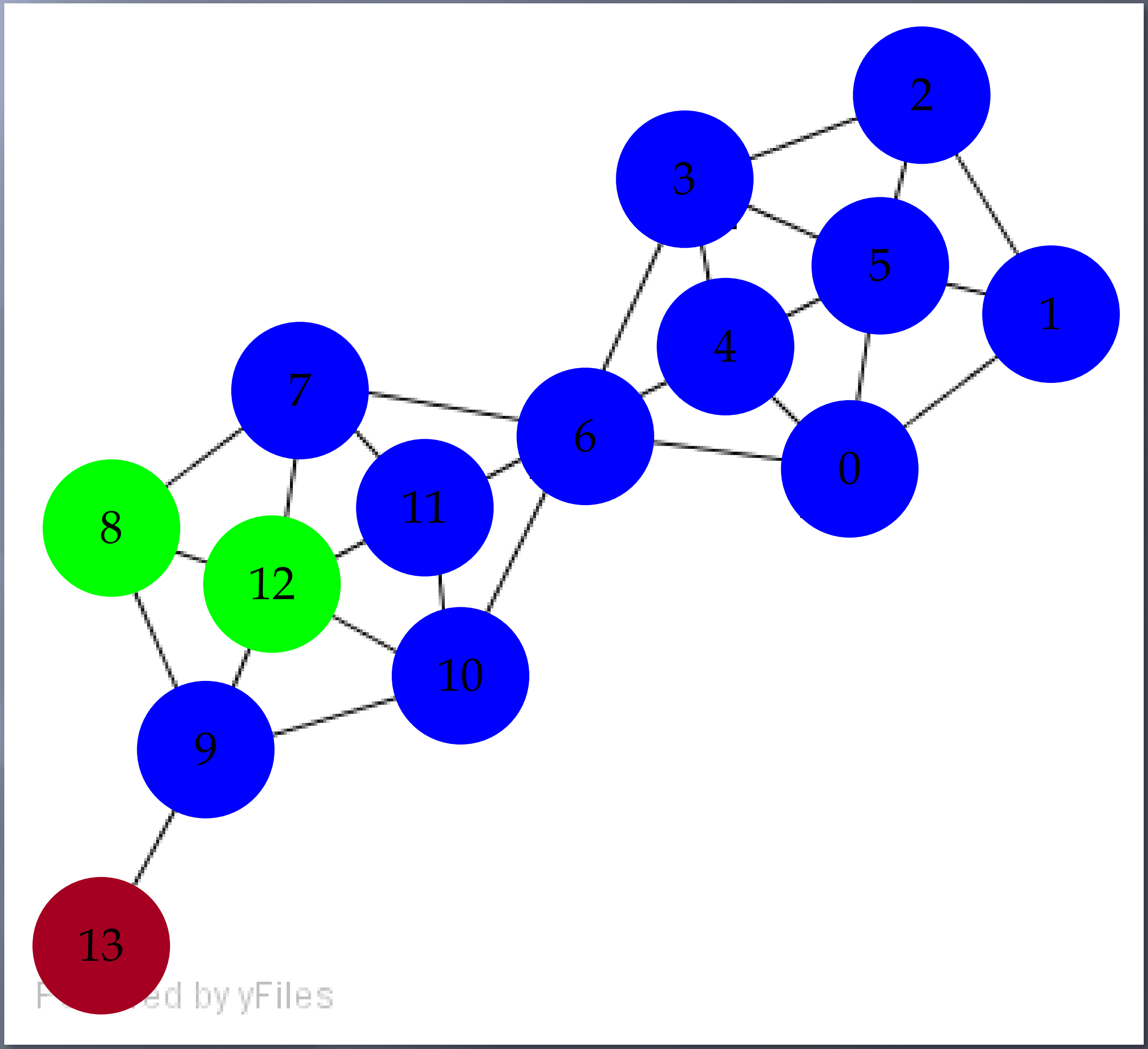
$\mu = 2$
$\varepsilon = 0.7$



SCAN Algorithm

$\mu = 2$

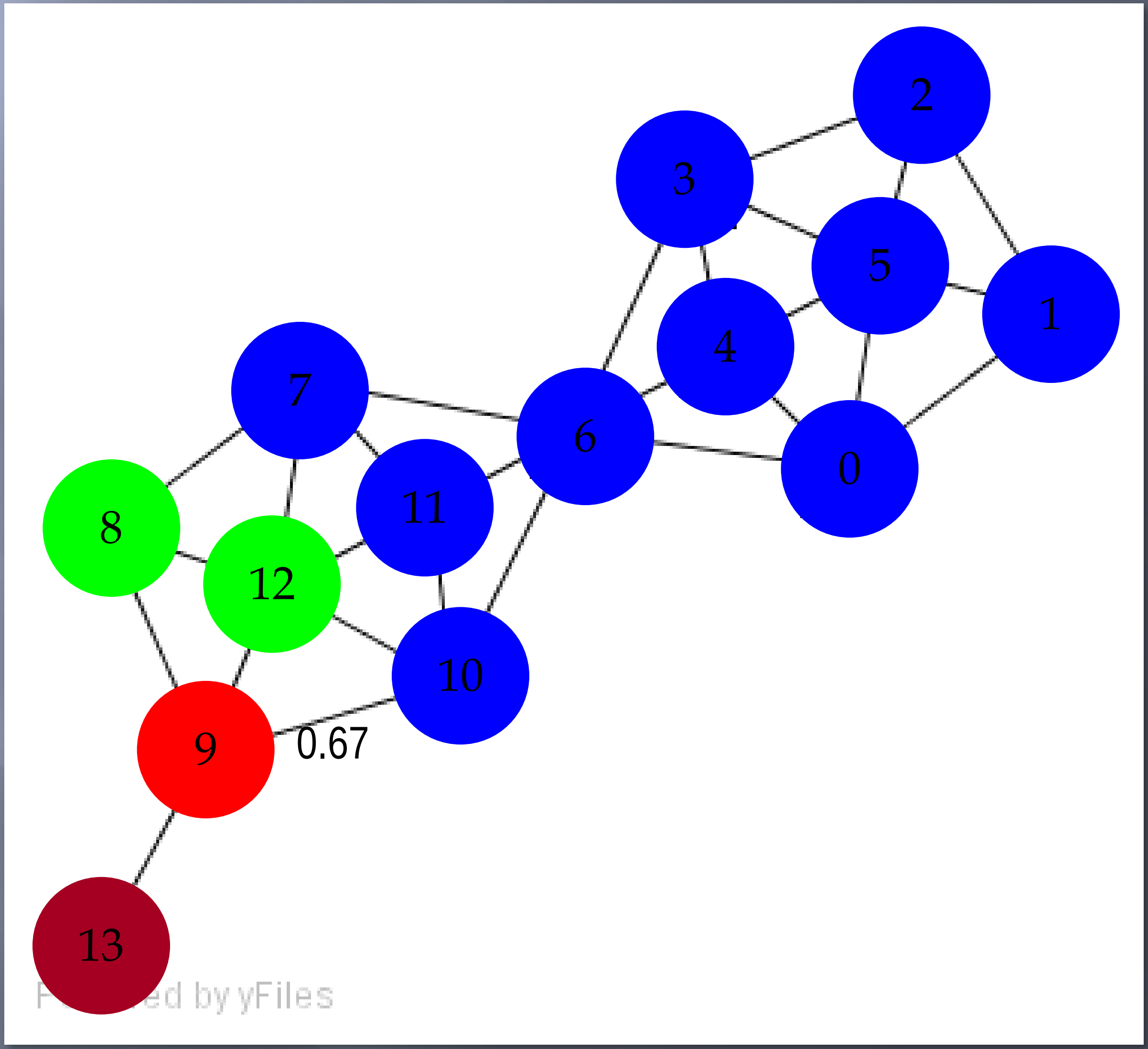
$\varepsilon = 0.7$



SCAN Algorithm

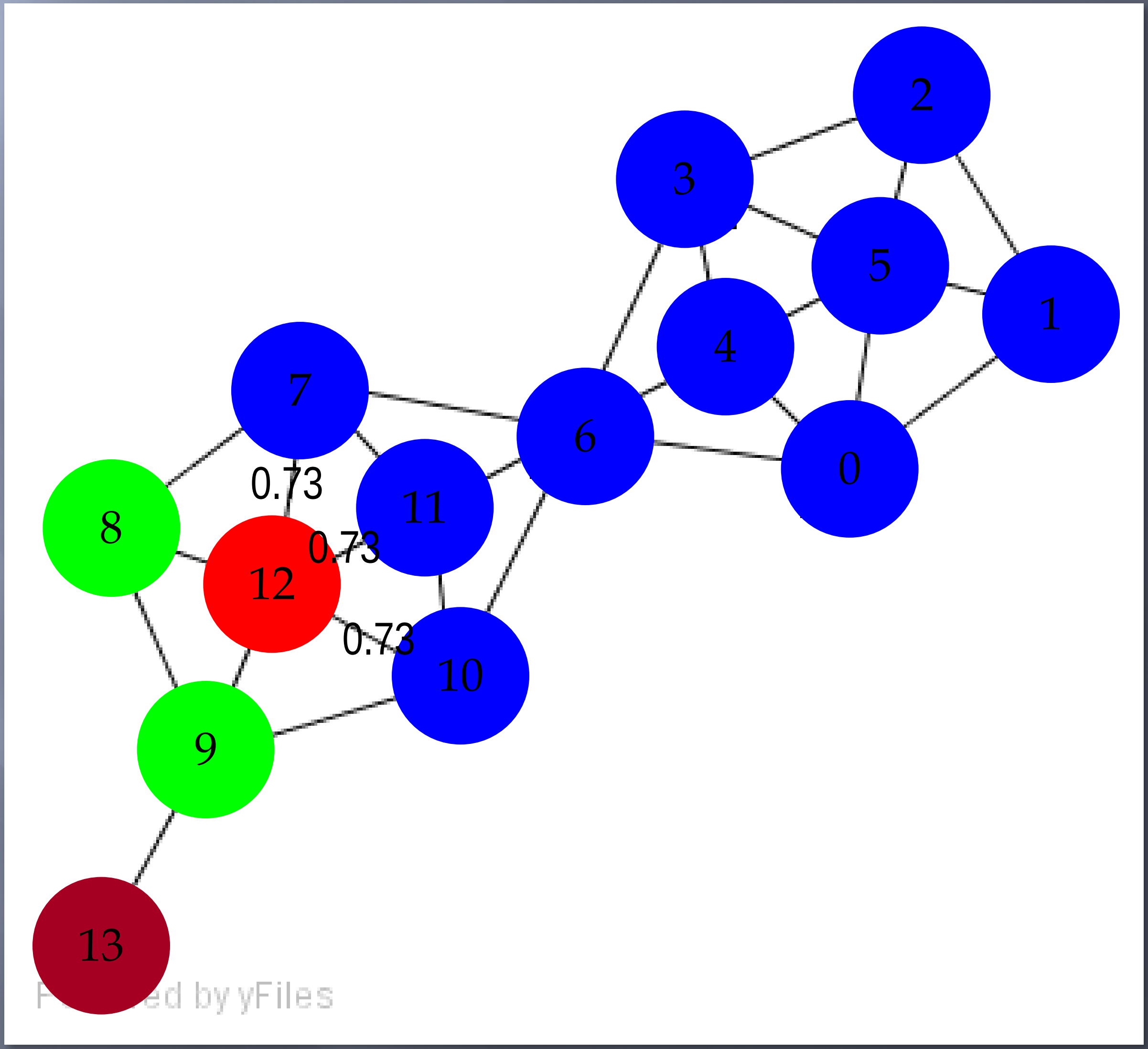
$\mu = 2$

$\varepsilon = 0.7$



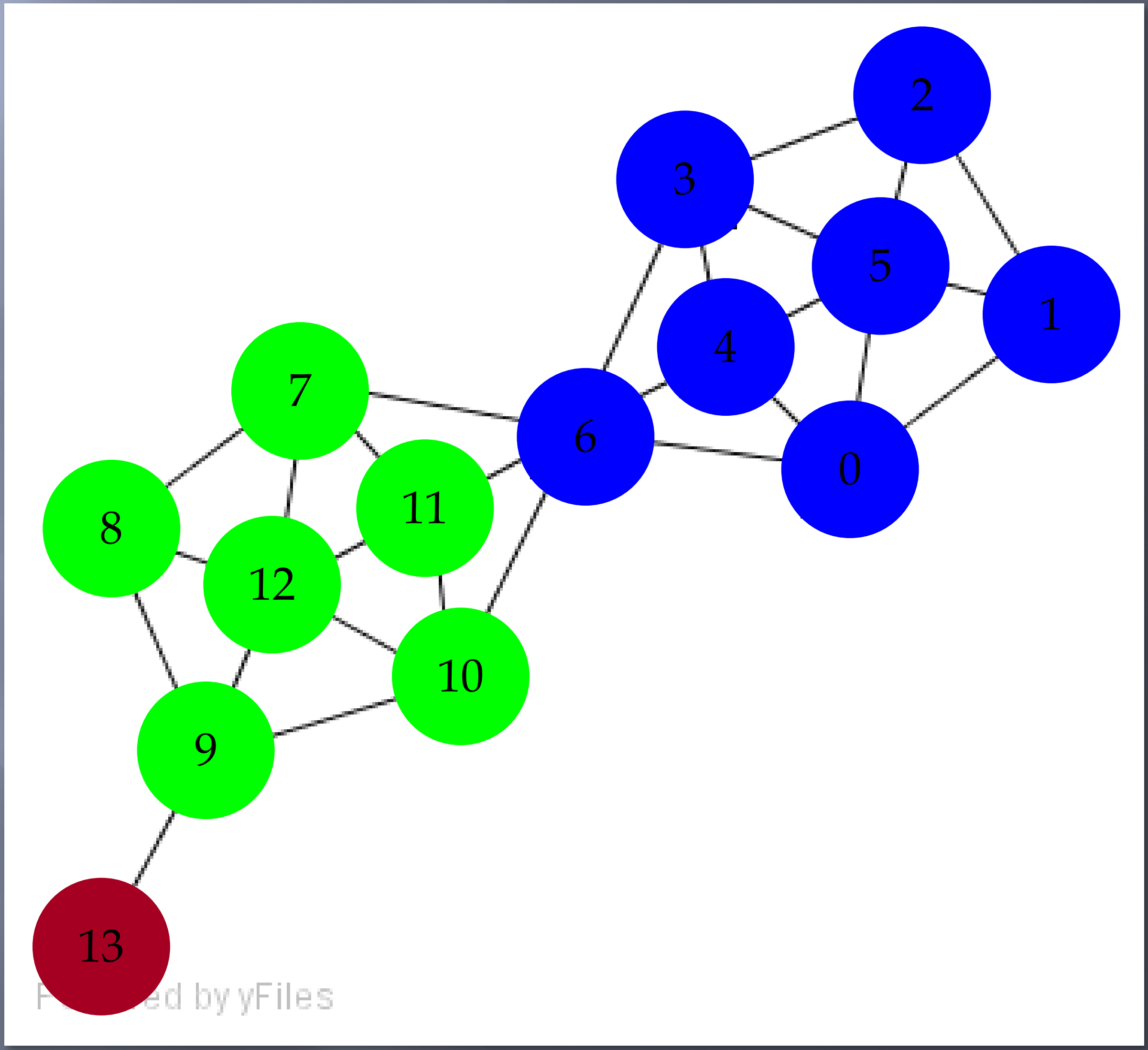
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



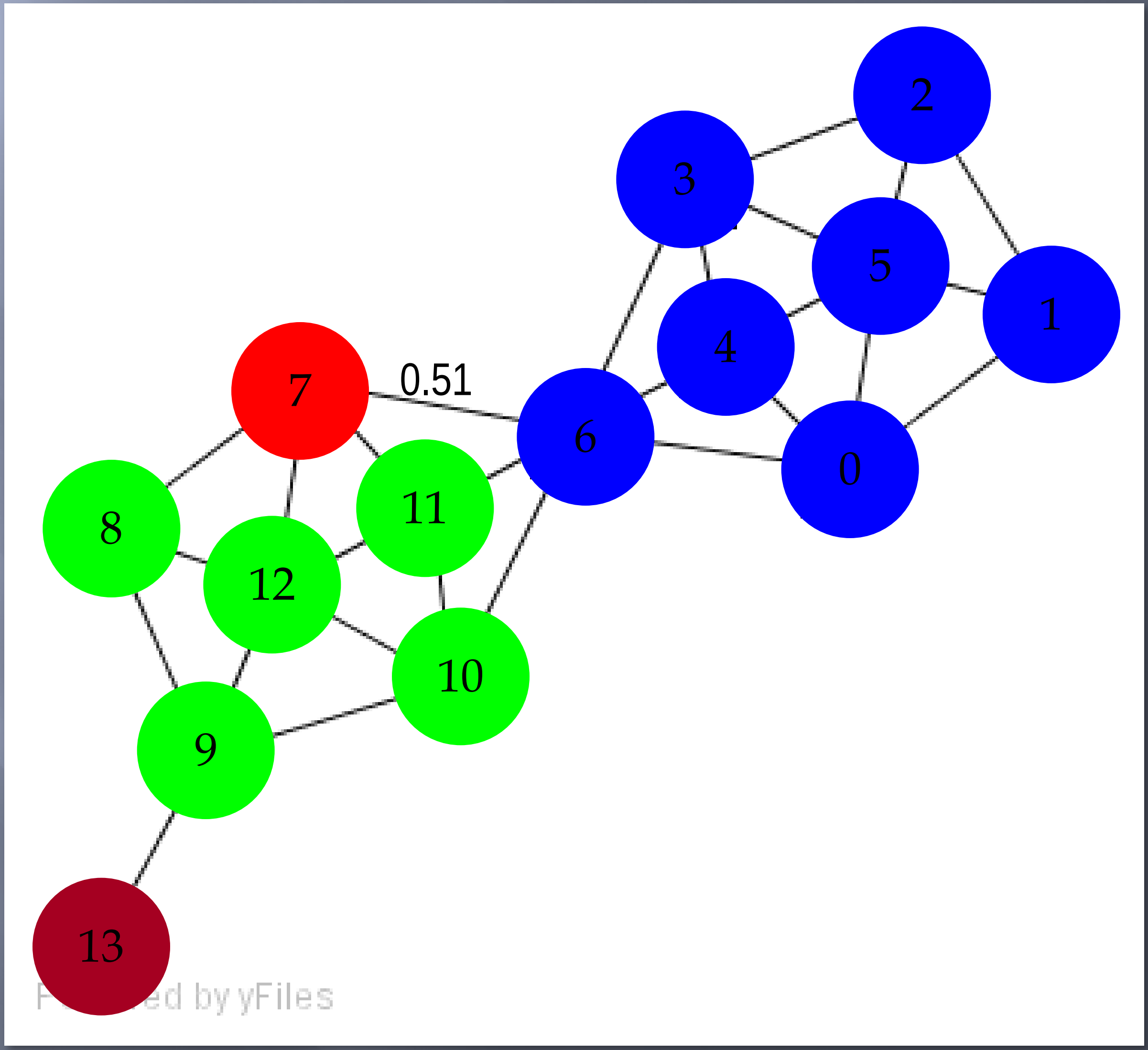
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



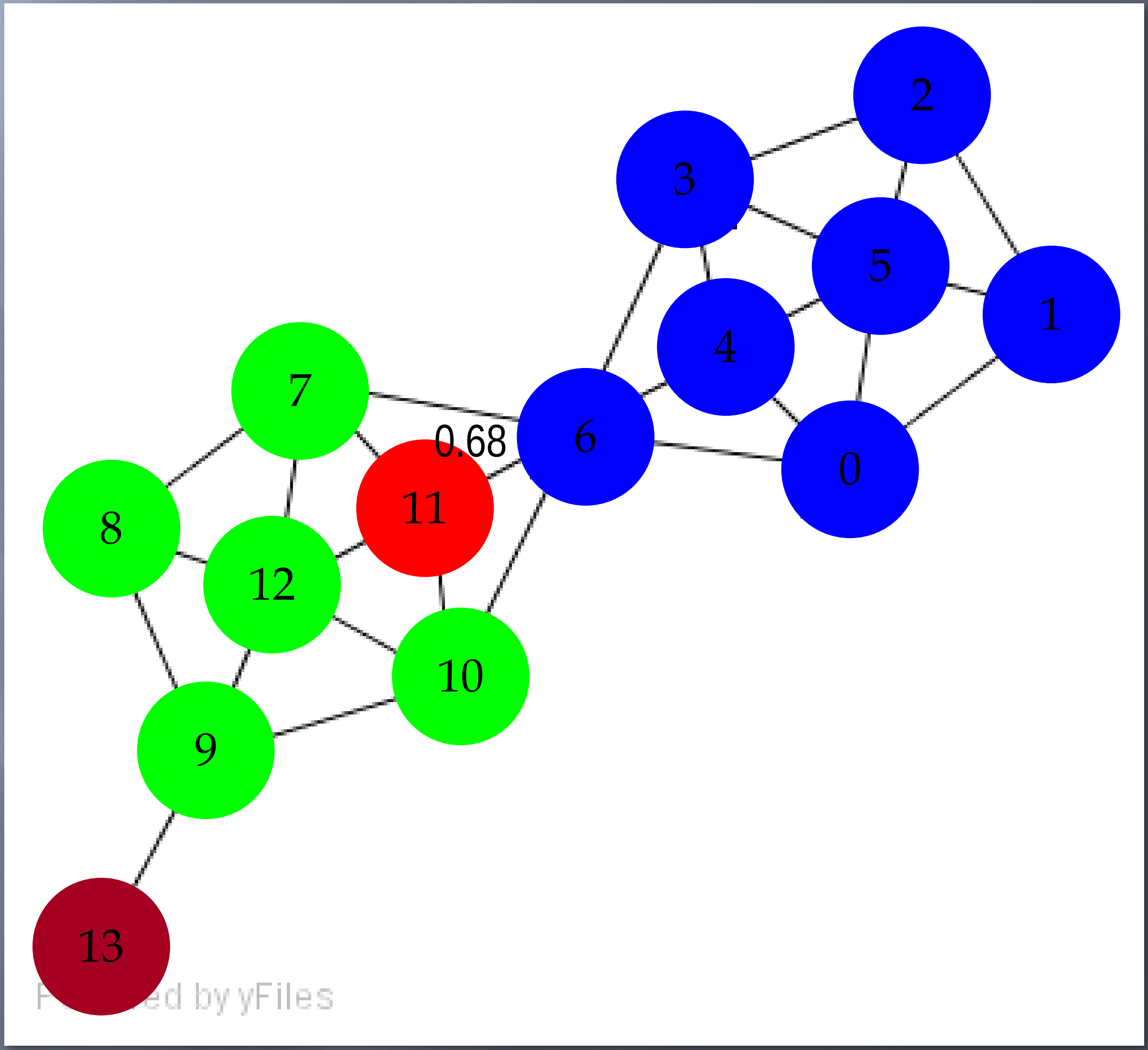
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



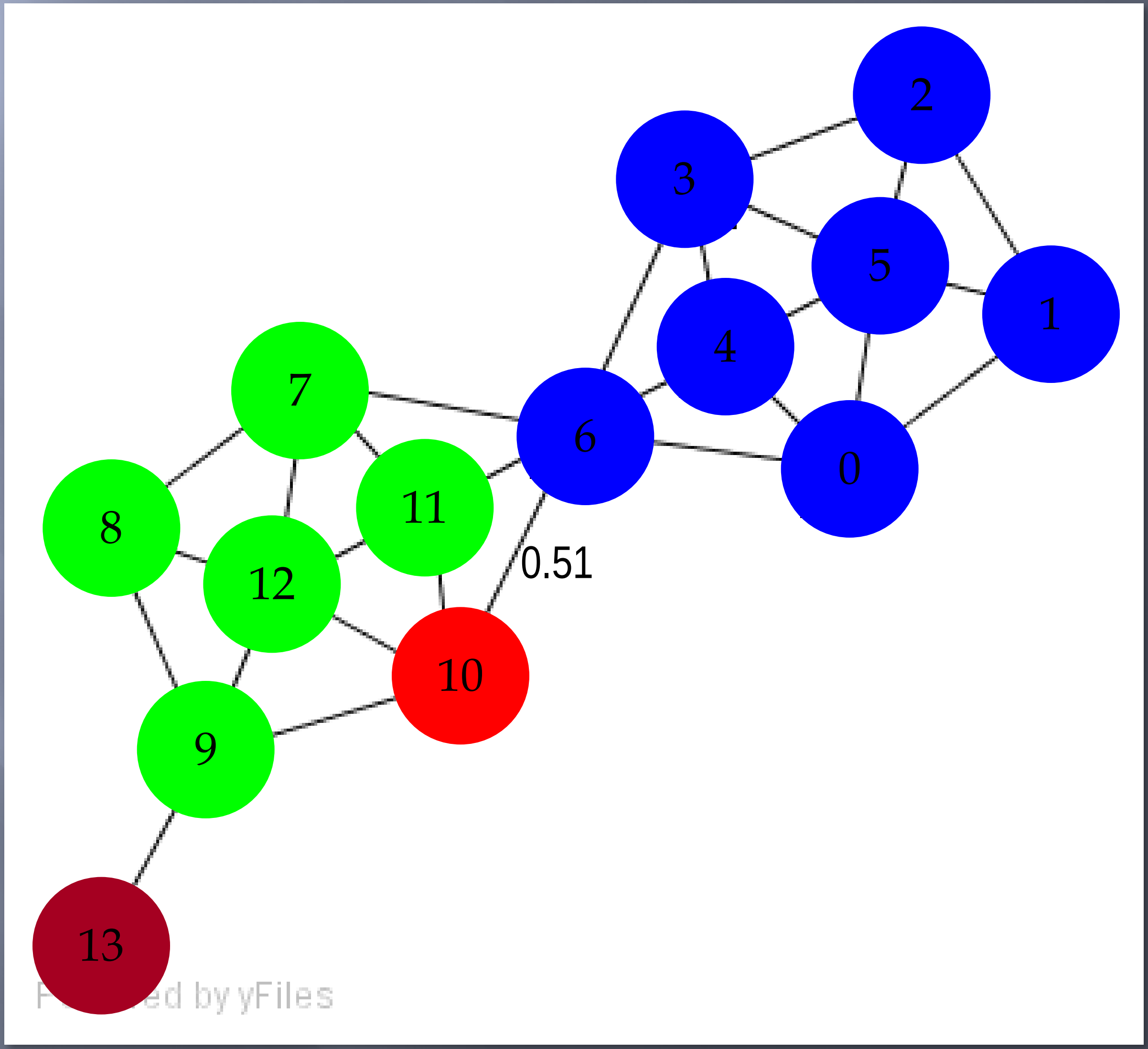
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



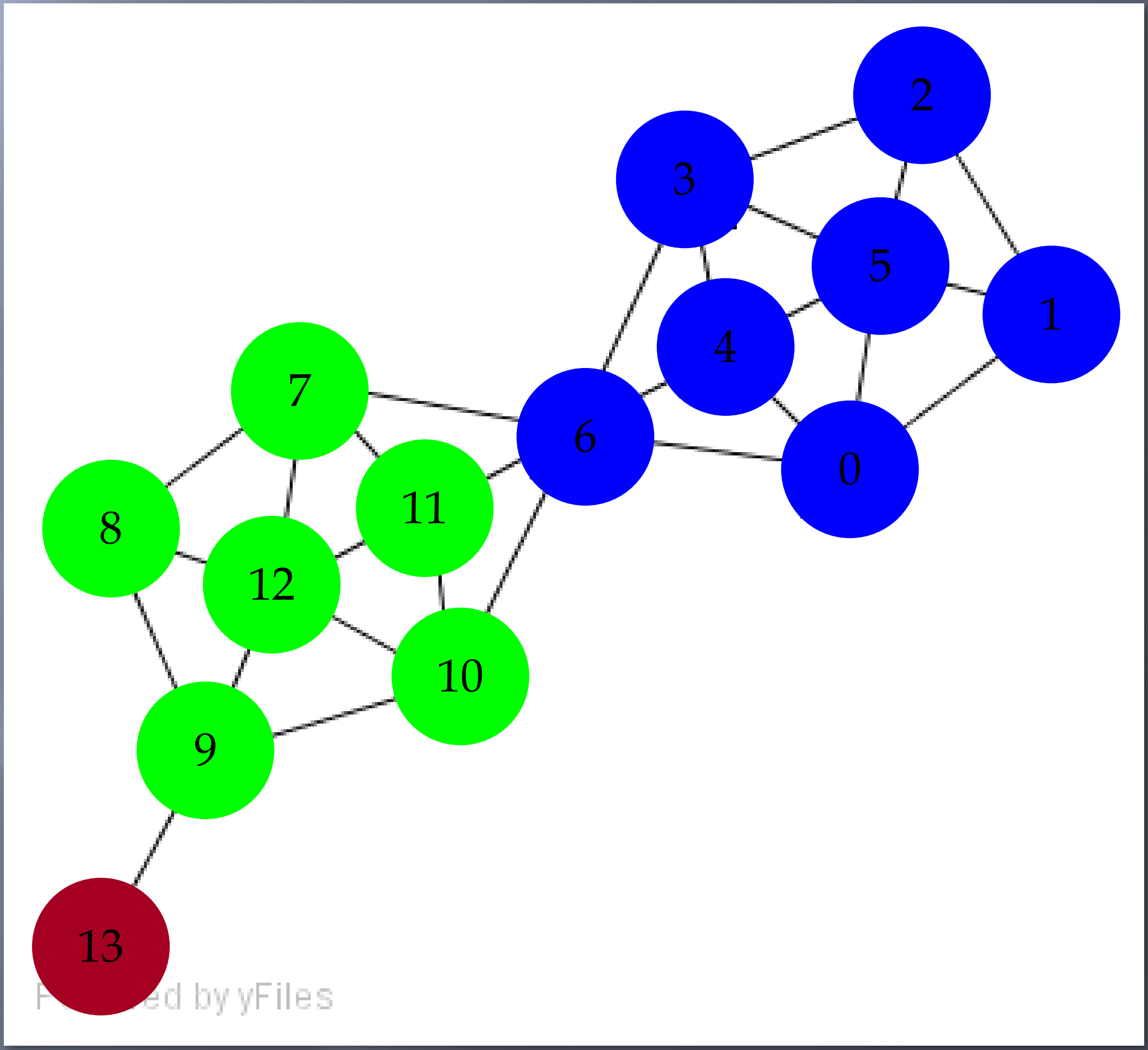
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



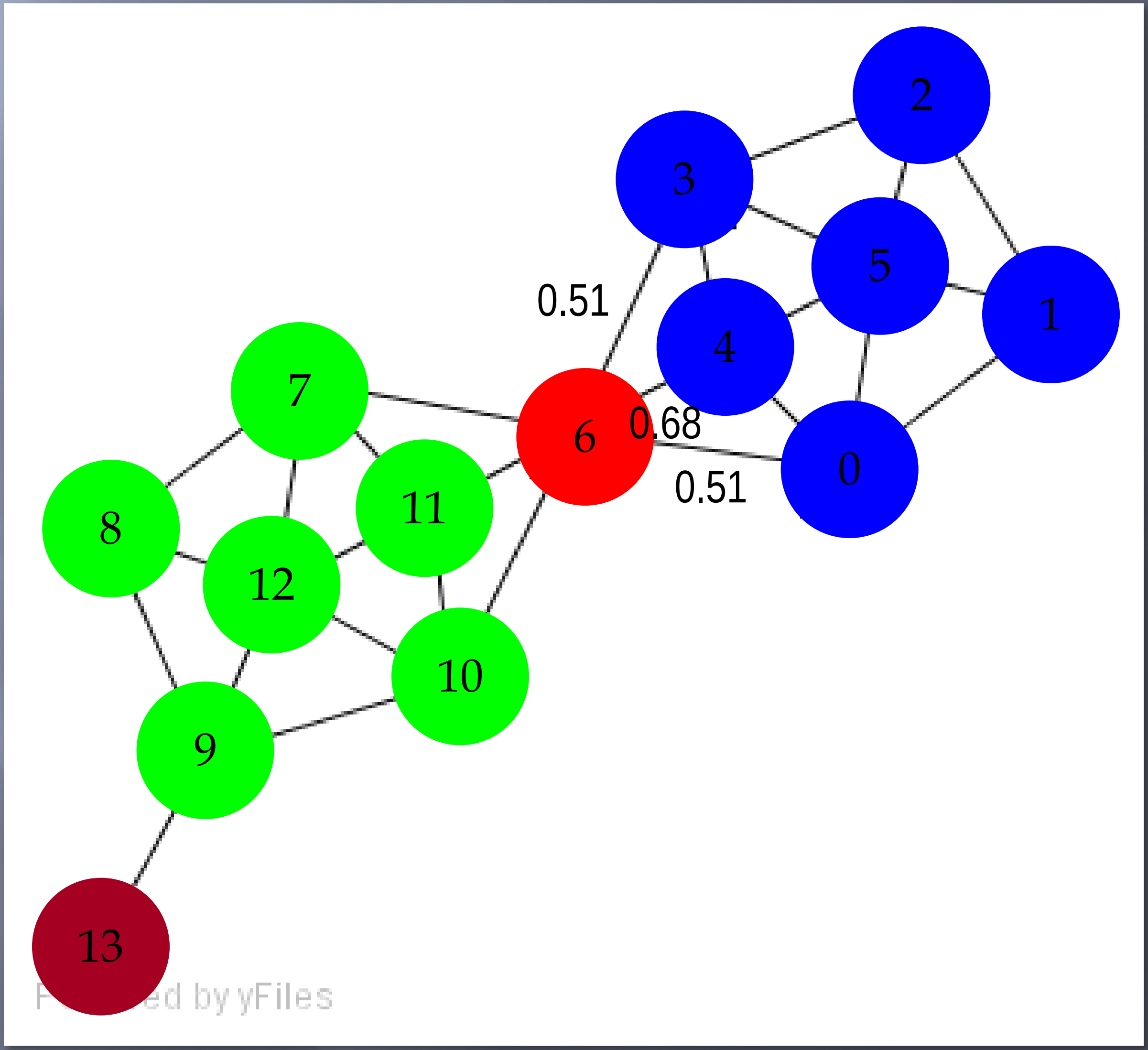
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$



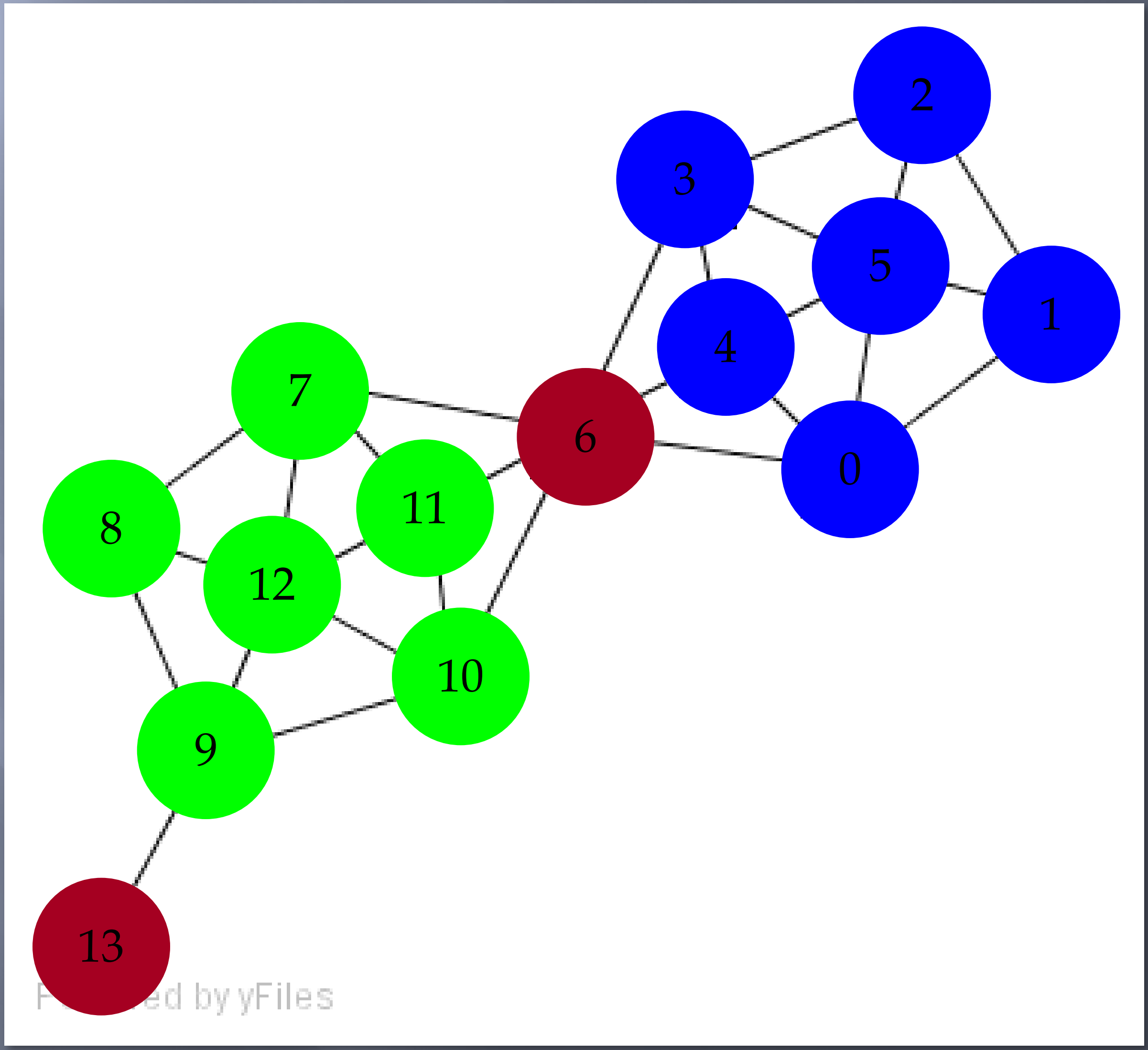
SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$

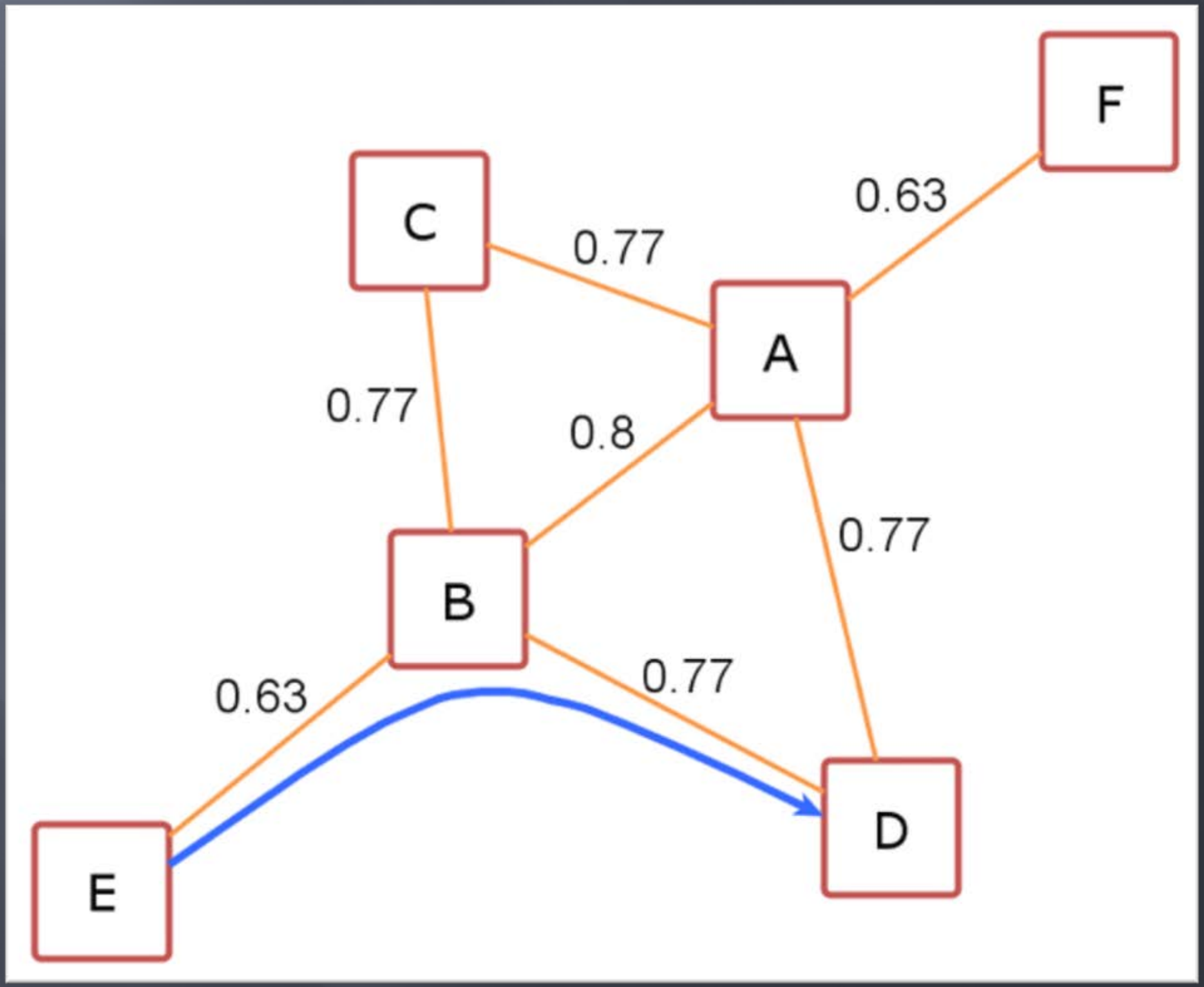


SCAN Algorithm

$\mu = 2$
$\varepsilon = 0.7$

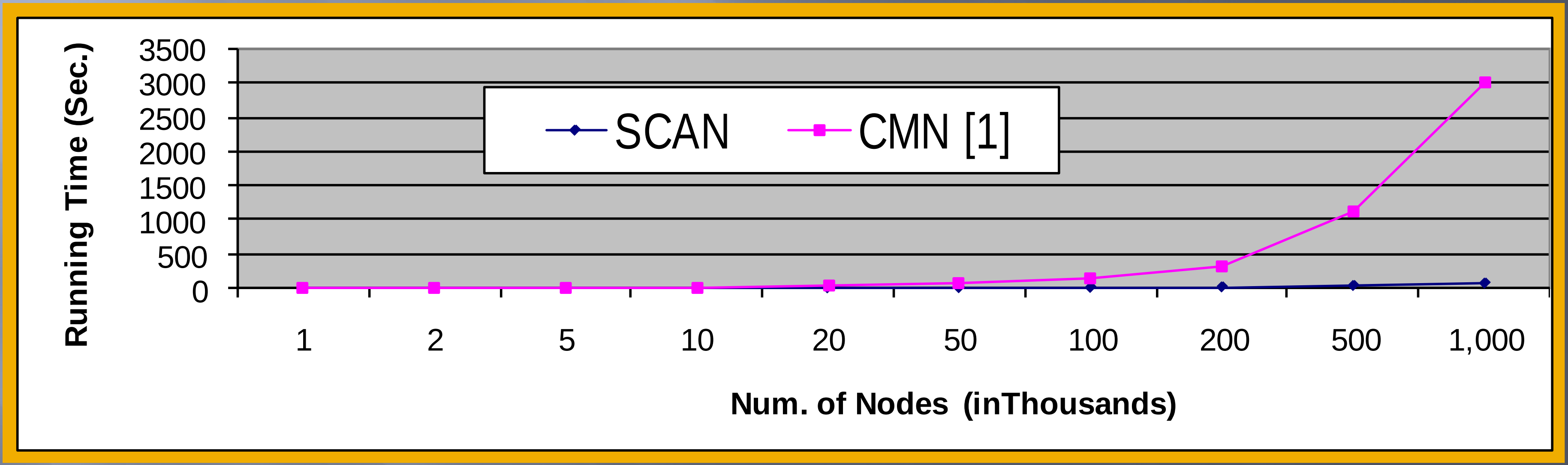


SCAN Algorithm



Running Time

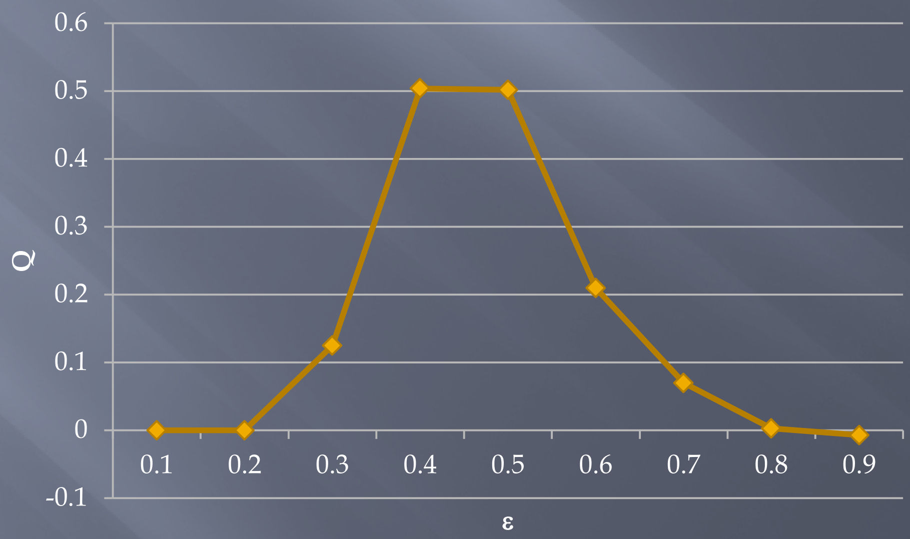
- Running time = $O(|E|)$
- For sparse networks = $O(|V|)$



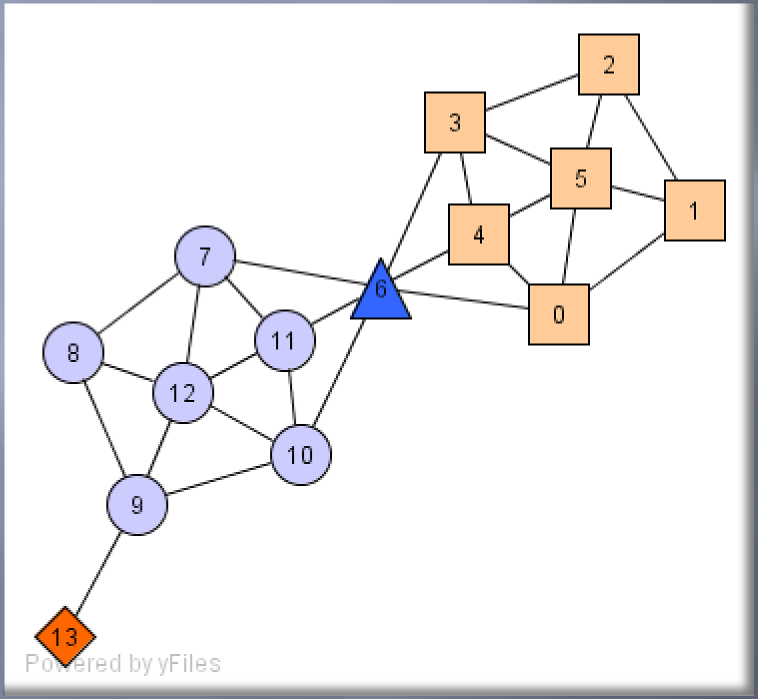
[1] A. Clauset, M. E. J. Newman, & C. Moore, *Phys. Rev. E* **70**, 066111 (2004).

Determine Parameters

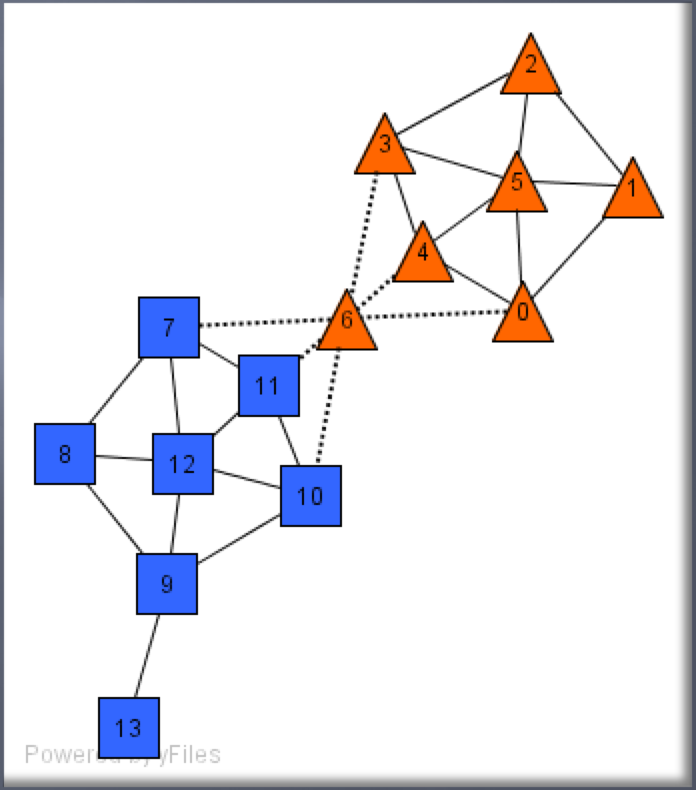
- Fix μ=2
- Run SCAN for ε=0.1,0.2,0.3,…,1
- Choose optimal ε, which maximize *Q*



SCAN vs. CMN



SCAN



CMN

Applications

- Social networks
- Product networks
- Customer data networks
- Biological networks
  - Metabolic networks
  - Protein-protein interaction networks

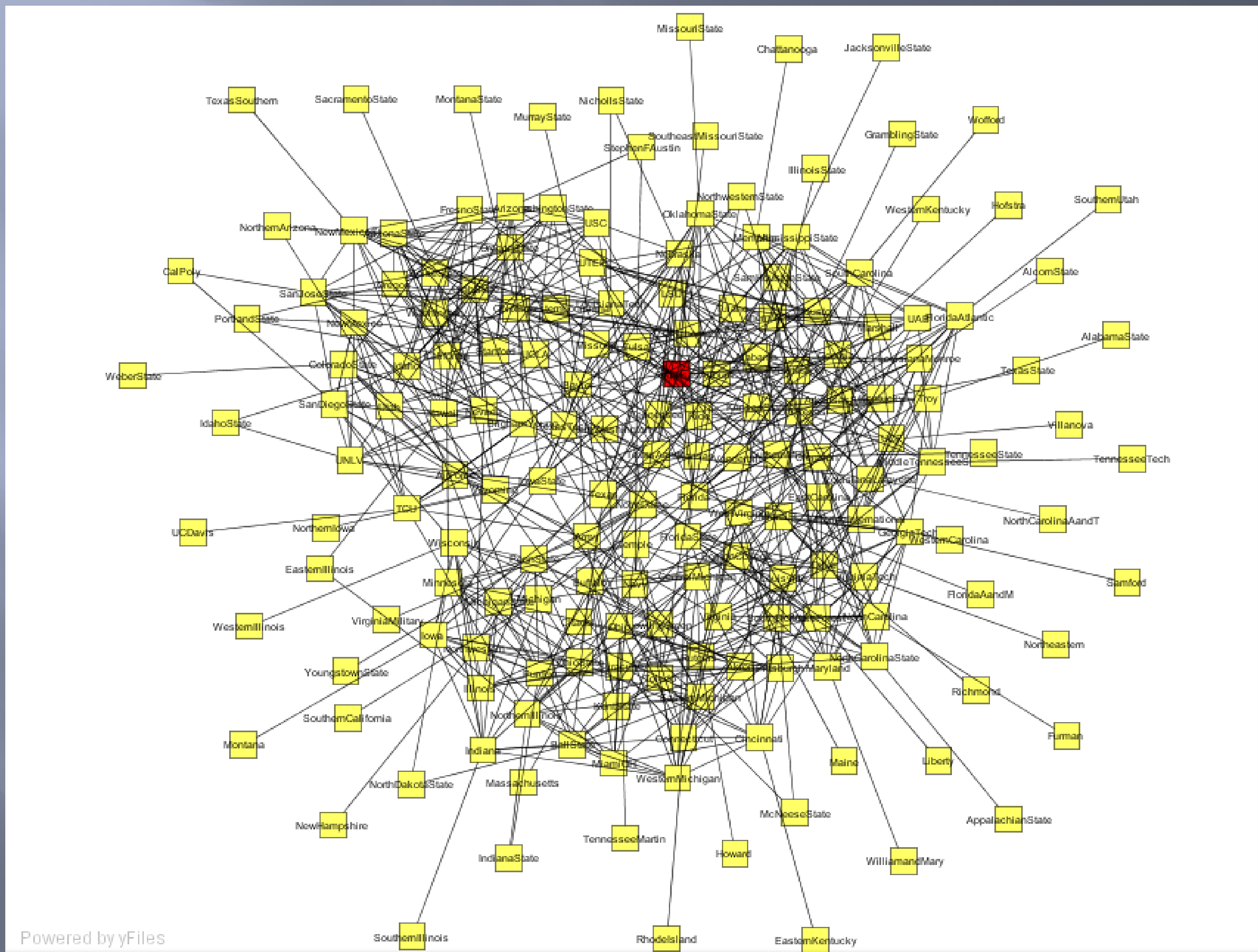
Are you ready for some football?

- Given only the 2006 schedule of what schools each NCAA Division 1A team met on a football field, what underlying structures could one discover?

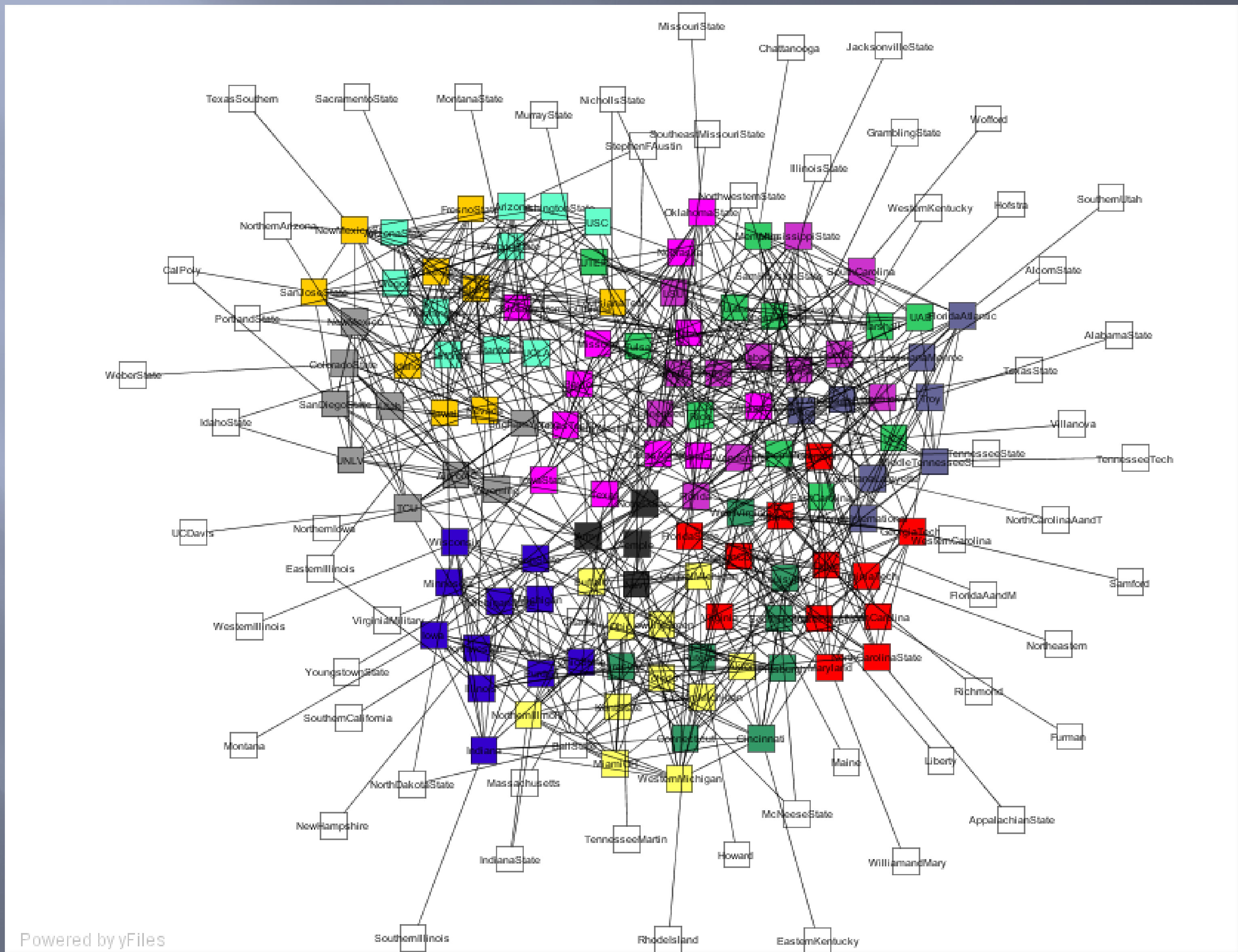
789 Contests

- 119 Division 1A school who play:
  - schools in their conference
  - schools in other 1A conferences
  - independent 1A schools (eg. Army)
  - schools in sub-1A conferences (eg. Maine)

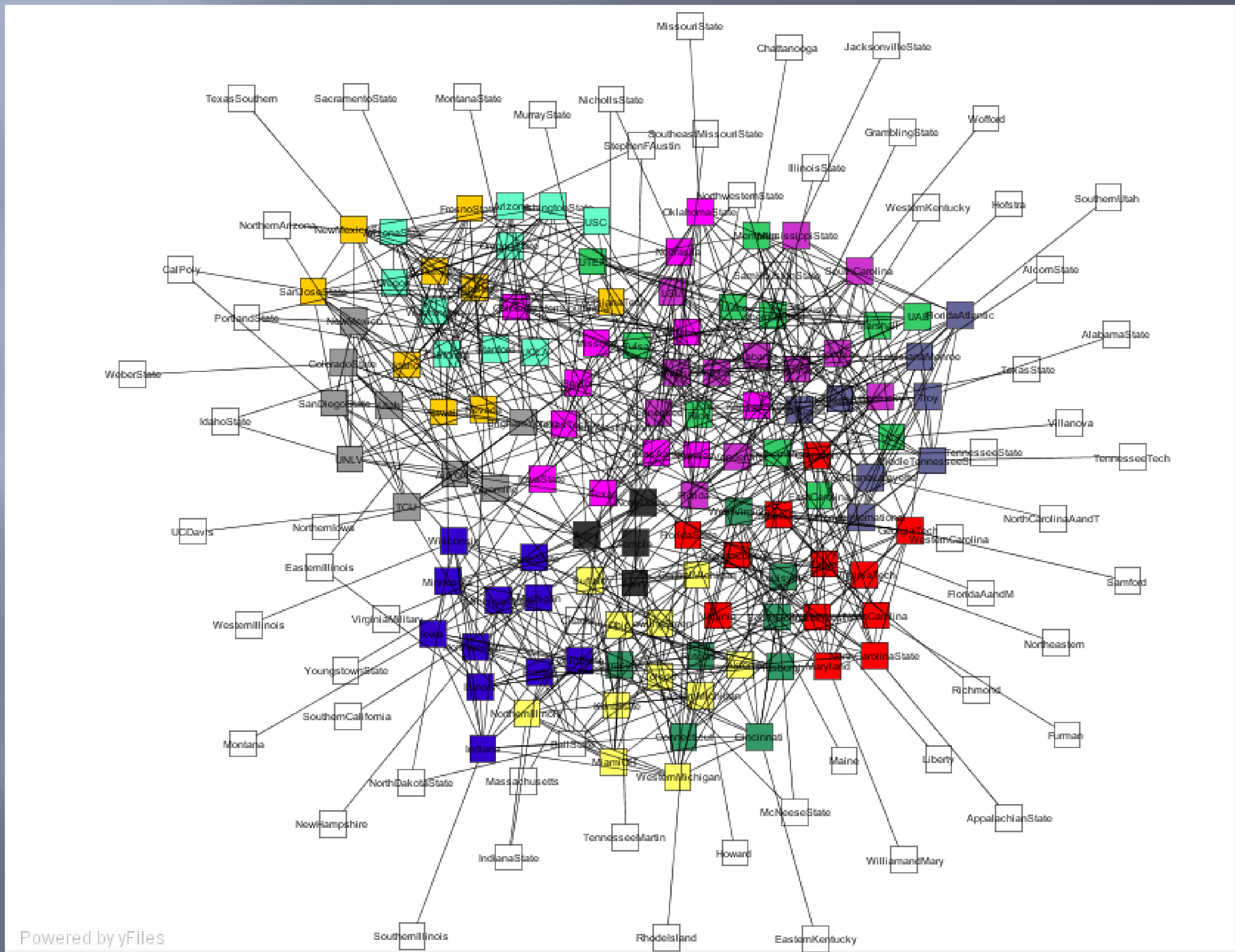
College Football Team Network



1A Conferences

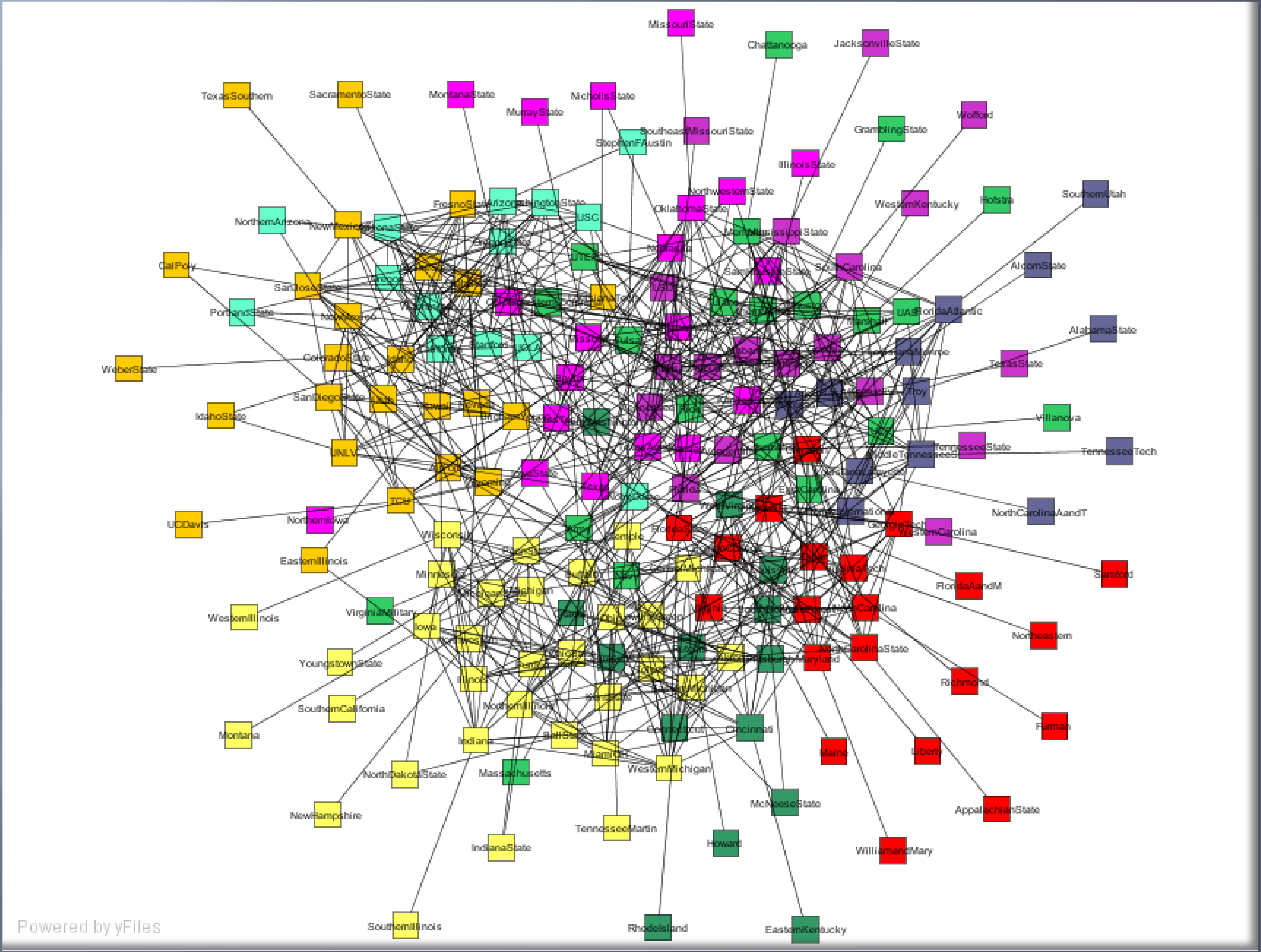


SCAN Result



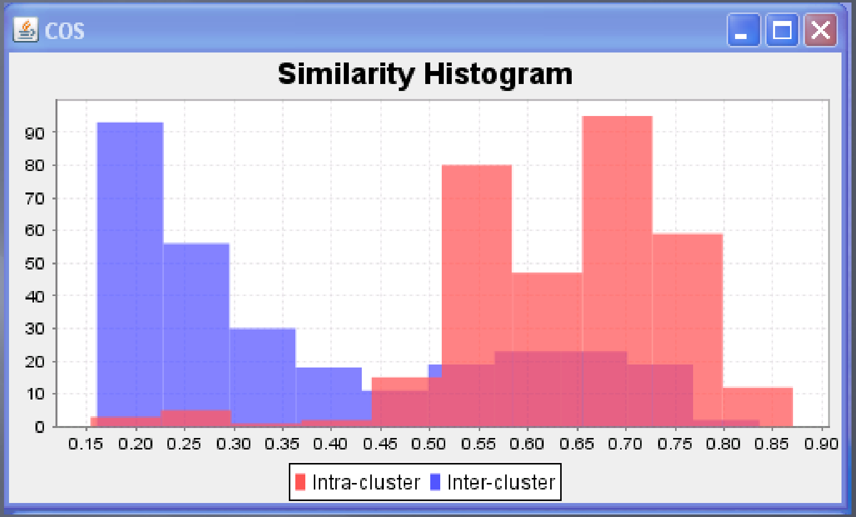
MissouriState
Chattanooga
JacksonvilleState
TexasSouthern
SacramentoState
MontanaState
NichollsState
MurrayState
SoutheastMissouriState
StephenFAustin
GramblingState
Wofford
IllinoisState
NorthwesternState
WesternKentucky
Hofstra
SouthernUtah
NorthernArizona
FresnoState
USC
CalPoly
AlcornState
PortlandState
AlabamaState
WeberState
ColoradoState
TexasState
SanDiegoState
IdahoState
Villanova
UNLV
TennesseeState
TennesseeTech
NorthCarolinaAandT
UCDavis
TCU
NorthernIowa
WesternCarolina
EasternIllinois
Samford
FloridaAandM
WesternIllinois
VirginiaMilitary
Northeastern
YoungstownState
Richmond
SouthernCalifornia
Furman
Montana
Liberty
Maine
NorthDakotaState
Massachusetts
McNeeseState
AppalachianState
NewHampshire
TennesseeMartin
Howard
WilliamandMary
IndianaState
SouthernIllinois
RhodeIsland
EasternKentucky
Powered by yFiles

CMN Result


MissouriState
Chattanooga
JacksonvilleState
TexasSouthern
SacramentoState
MontanaState
MurrayState
NichollsState
Wofford
GramblingState
IllinoisState
WesternKentucky
Hofstra
SouthernUtah
NorthernArizona
CalPoly
AlcornState
PortlandState
AlabamaState
TexasState
WeberState
IdahoState
Villanova
TennesseeTech
NorthCarolinaAandT
UCDavis
NorthernIowa
WesternCarolina
Samford
EasternIllinois
FloridaAandM
WesternIllinois
VirginiaMilitary
Northeastern
YoungstownState
Richmond
SouthernCalifornia
Furman
Montana
Maine
Liberty
NorthDakotaState
Massachusetts
WesternMichigan
NewHampshire
McNeeseState
AppalachianState
TennesseeMartin
Howard
IndianaState
WilliamandMary
Powered by yFiles
SouthernIllinois
RhodeIsland
EasternKentucky

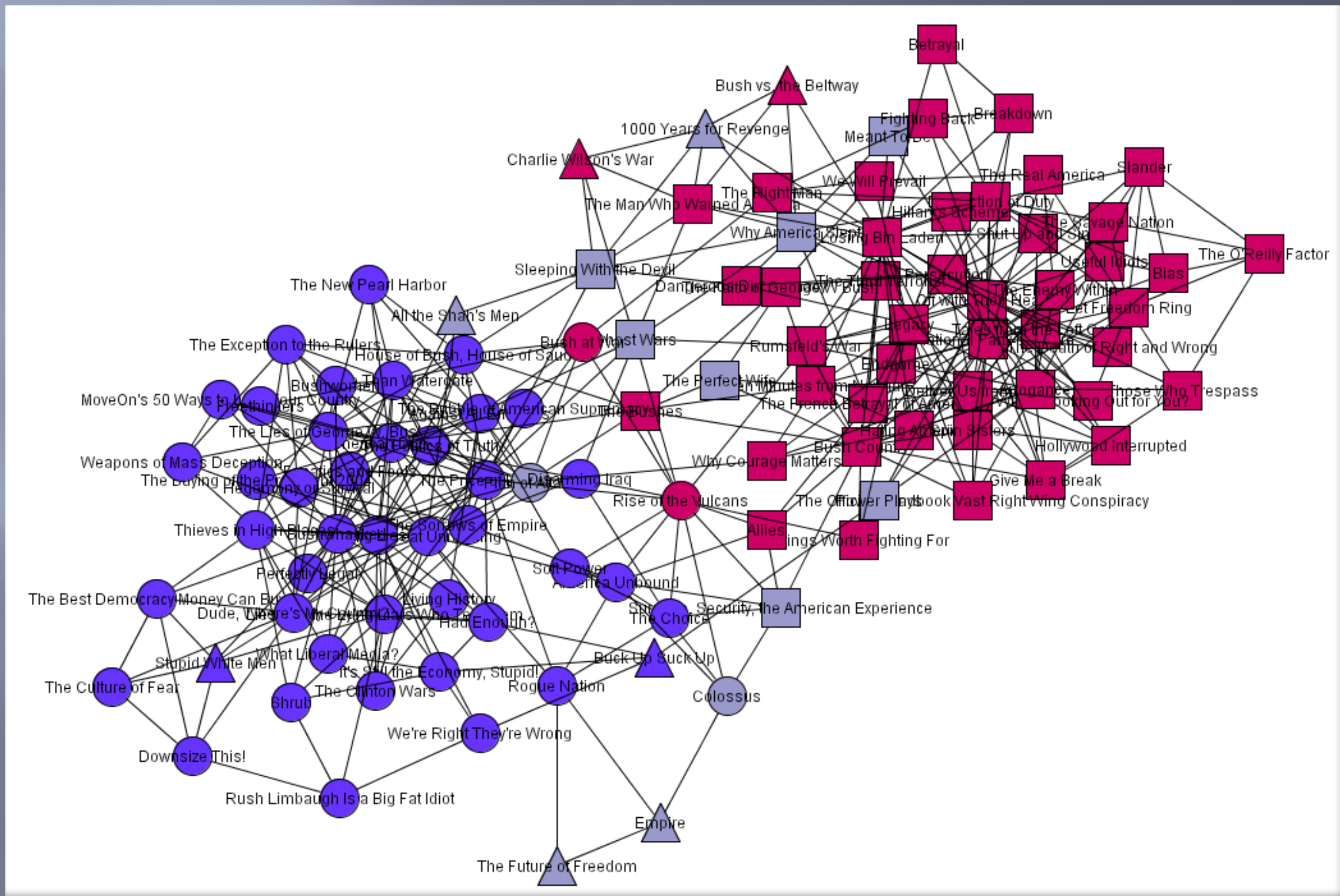
Why SCAN works better?

	INTRA	INTER	SUM
Count	380	233	613
Per.	62%	38%	100%

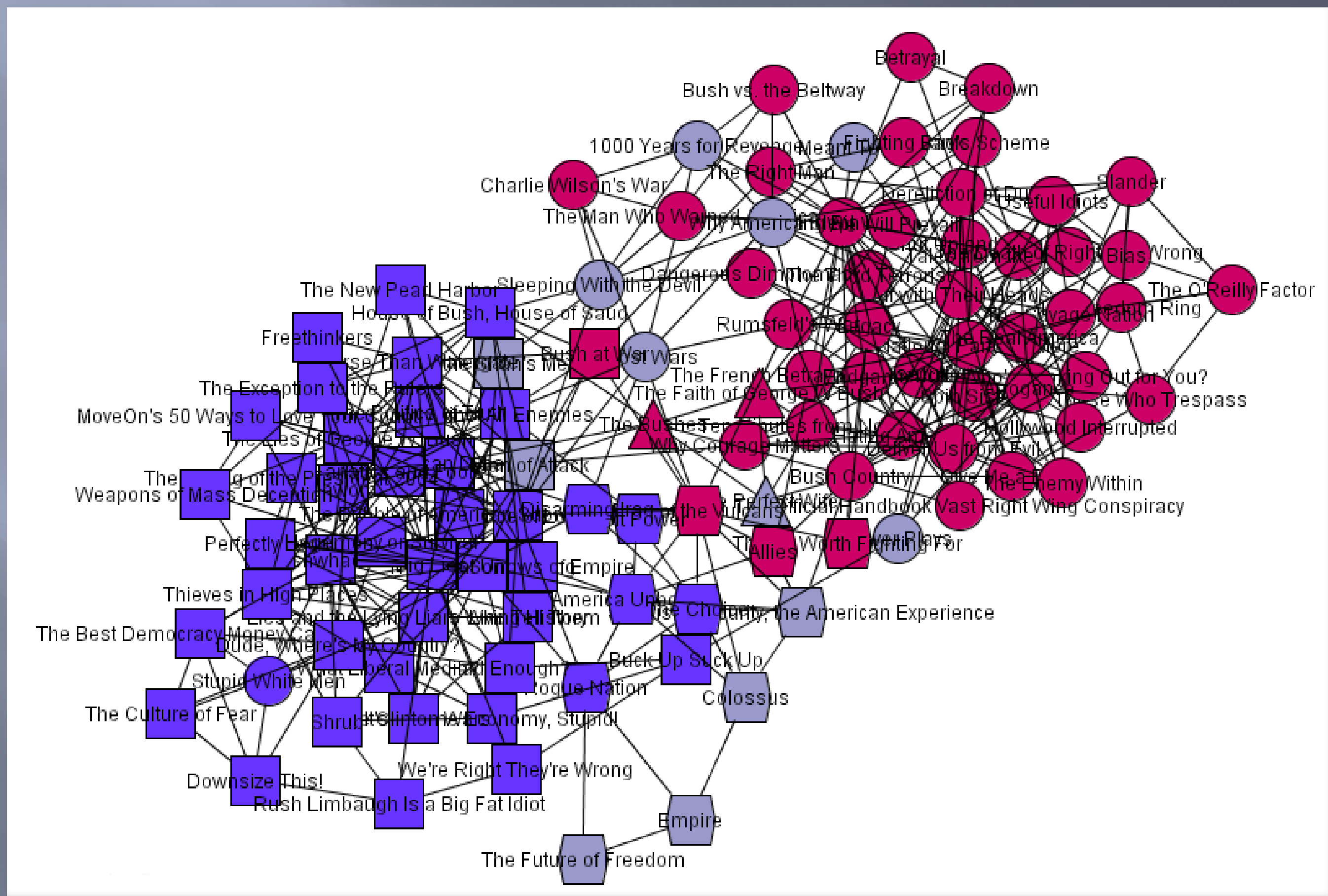


- The assumption that there should be much fewer inter-cluster links does not hold for college-football dataset and many other networks
- Structural-similarity is obviously more discriminative

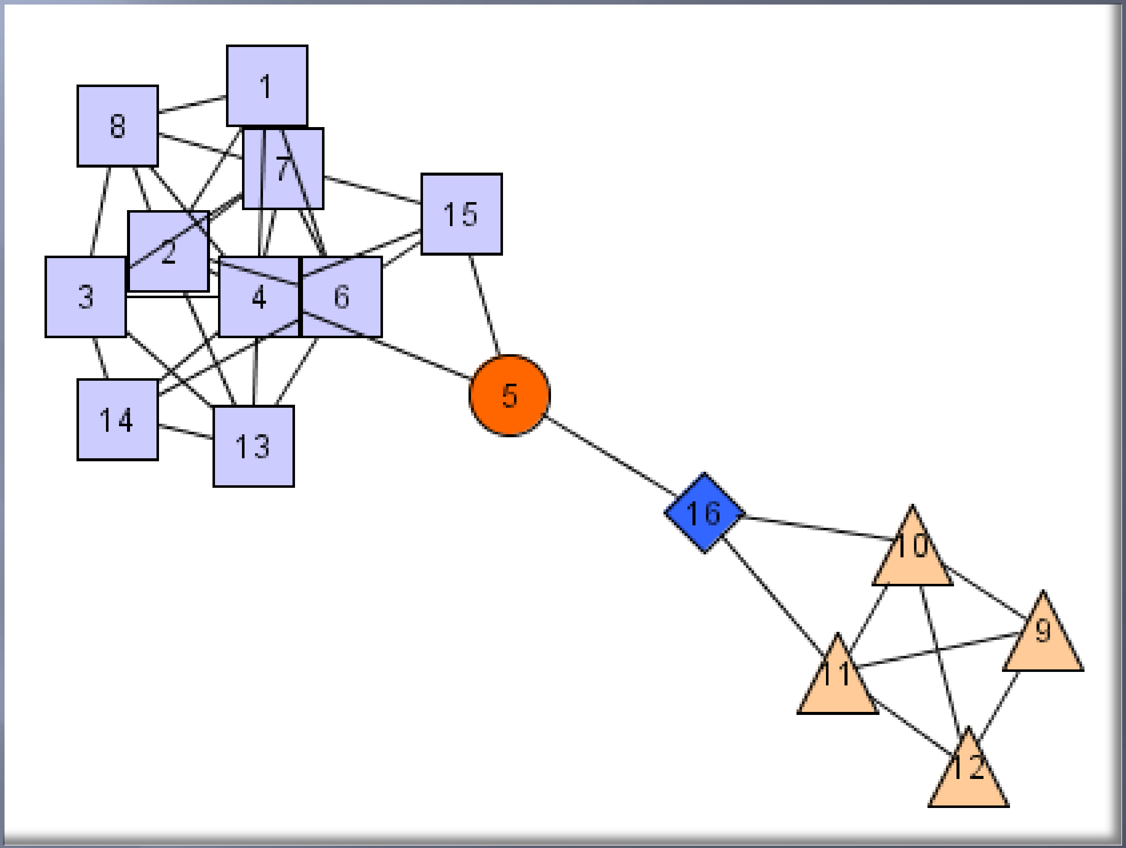
Political Books (SCAN)



Political Books (CMN)

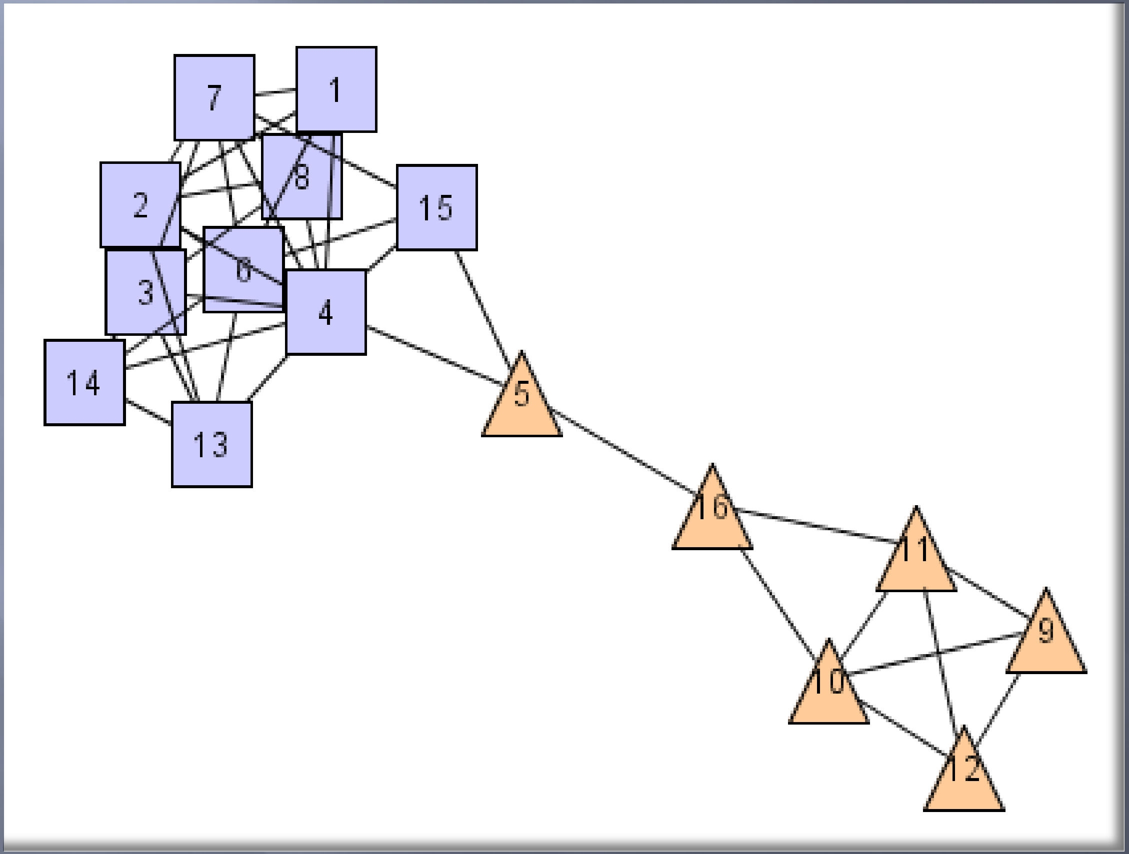


Customer Data (SCAN)



Hubs tend to be damaged data
SCAN can be used as a data
quality or entity resolution tool

Customer Data (CMN)

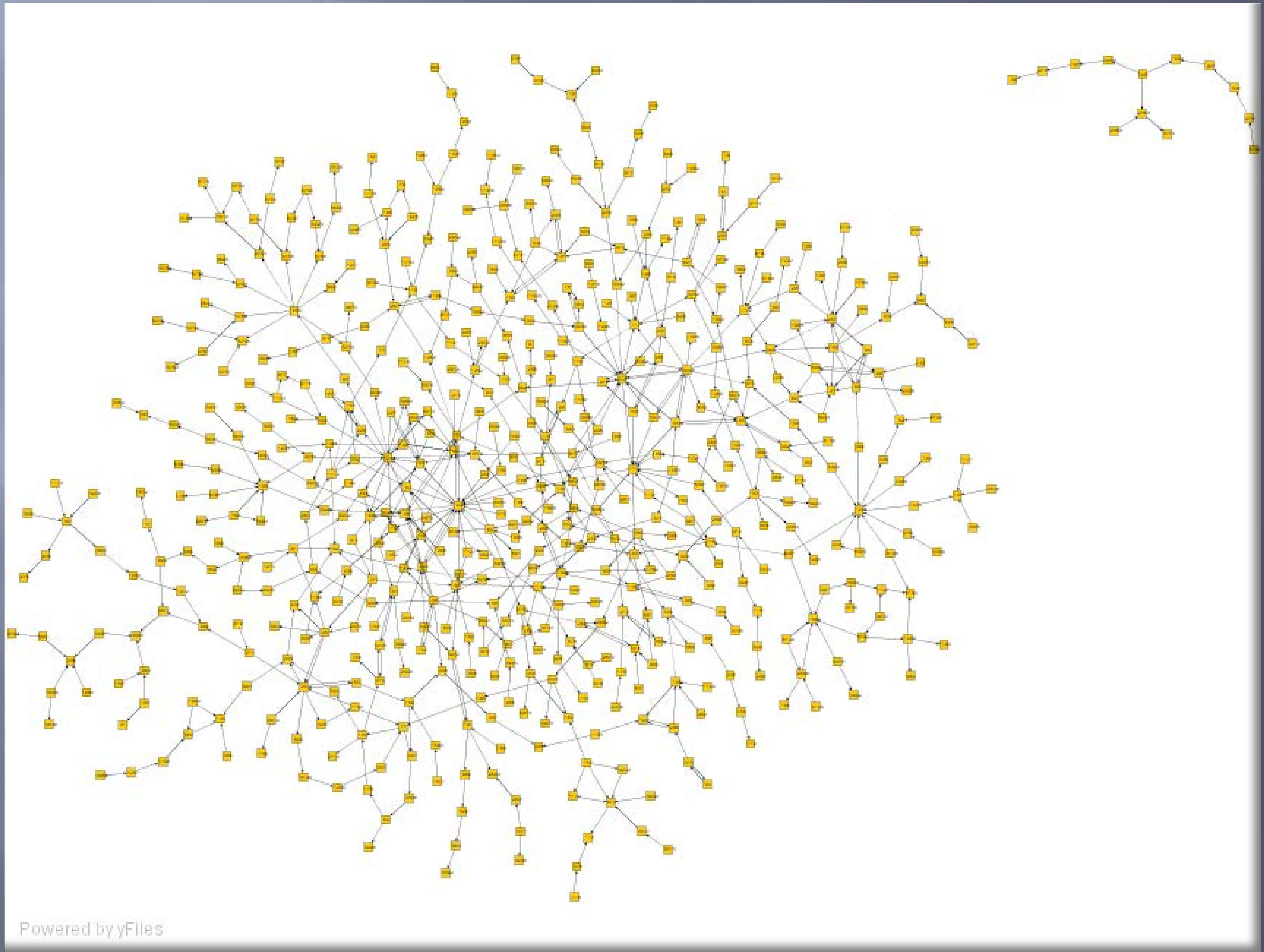


Adjusted Rand Index

	SCAN	CMN
College football	**1**	0.24
Political books	**0.71**	0.64
Customer data	**1**	0.85

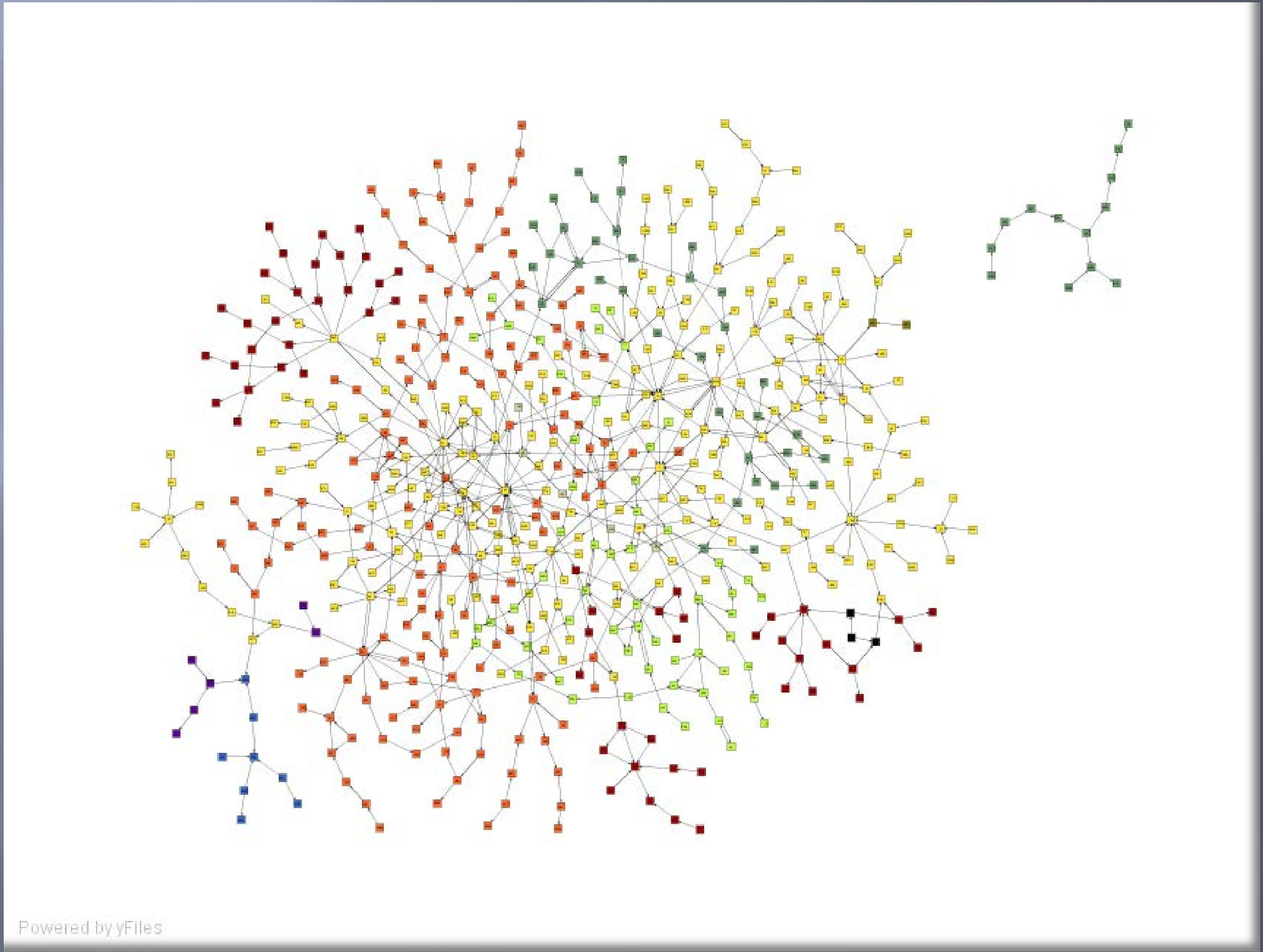
Metabolic Network for E.coli

513 nodes
750 links



Obtained Clusters

Colors: cluster



Finding Functional Modules in Protein–Protein Interaction Networks

- The Protein-Protein Interactions (PPI) network of budding yeast consists of 26,571 interactions between 4,030 proteins [1].
- We compare new algorithm with well-known *CMN* algorithm [2].
- *Validation* through GO annotations is a domain-based method.

[1] http://www.yeastgenome.org/
[2] Aaron Clauset, M. E. J. Newman, and Christopher Moore, Phys. Rev. E 70, 066111 (2004).

Clustering Score

$$p-value=\sum_{m}^{n}\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \quad [2]$$

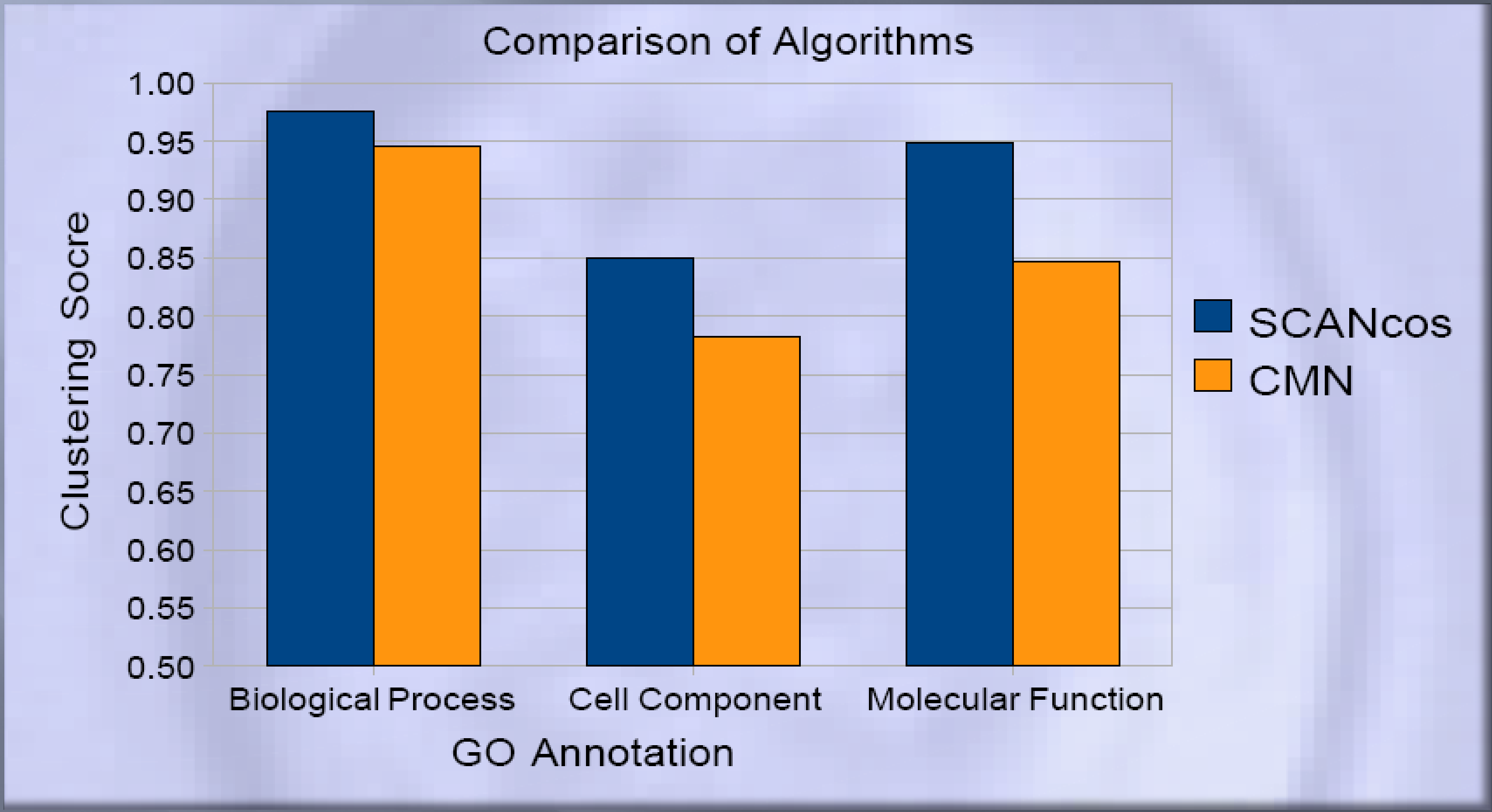
N: Number of proteins in PPI network
M: Number of GO term g in PPI network
n: Number of protein in cluster c
m: Number of GO term g in cluster c

$$Clustering\ Score=1-\frac{\sum_{i=1}^{n_s} min(p_i)+n_i*cutoff}{(n_s+n_i)*cutoff} \quad [3]$$

n_s: Number of significant clusters, $min(p_i) < cutoff$
n_i: Number of insignificant clusters, $min(p_i) > cutoff$
cutoff: threshold of 0.05

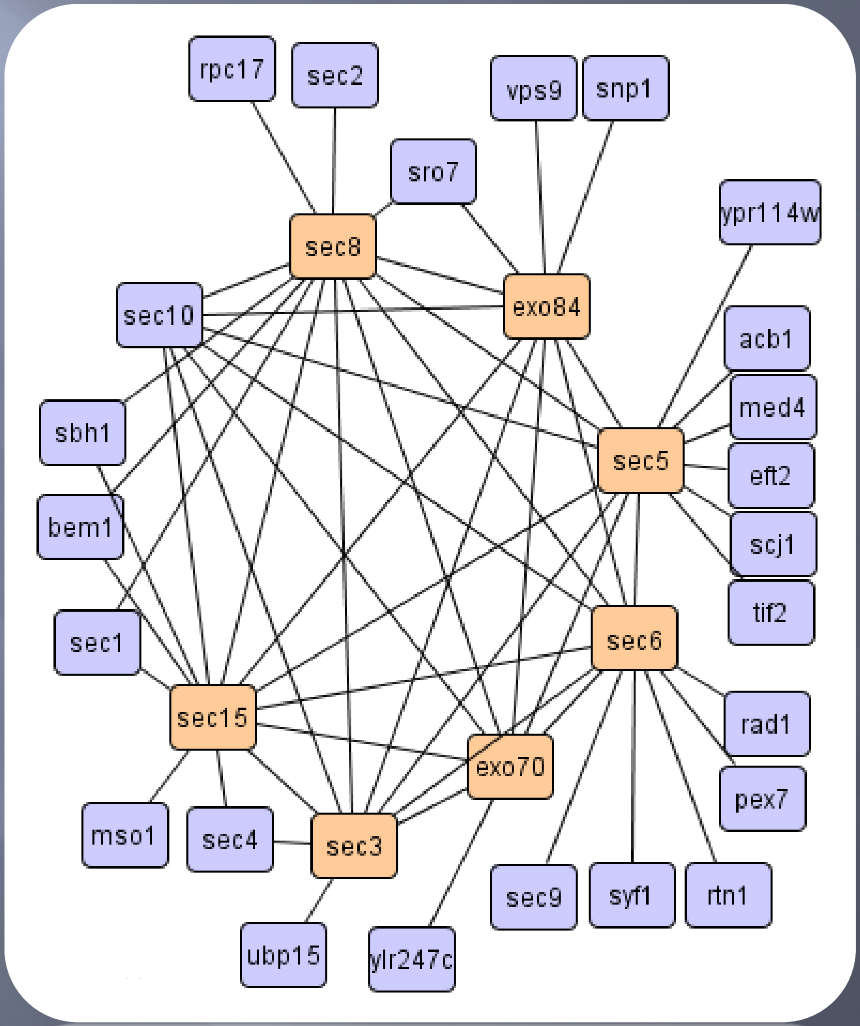
[2] Spirin and Mirny, 2003 V. Spirin and L.A. Mirny, Protein complexes and functional modules in molecular networks, Proc. Natl Acad. Sci. U.S.A. 100 (21) (2003), pp. 12123–12126.
[3] S. Asur, D. Ucar, S. Parthasarathy, An ensemble framework for clustering protein–protein interaction networks, Bioinformatics 2007 23: i29-i40; doi:10.1093/bioinformatics/btm212

Comparison of Algorithms



Results

Exocyst complex



Conclusion

- SCAN algorithm:
  - It is fast $O(|E|)$, for scale free networks: $O(|V|)$
  - It can find clusters, as well as hubs and outliers
- Applications of SCAN
  - Organizational networks (NCAA College Football)
  - Product networks (Political Books)
  - Customer data networks (Customer Records)
  - Biological networks (Metabolic, PPI Networks)

Future Work

- Cluster structures in dynamic networks
  - Evolution of cluster structures
  - Bi-directional pairs
  - Weighted edges
- Roles of nodes in terms of clusters
  - Leaders, followers, mediators, etc.
- Hierarchical cluster structures
  - Clusters of parent a cluster

Thank You

Dank U wel

Thank you

谢谢

Merci

Благодаря

Tesekkurler

Danke

आभारी हुँ

شكرا

多謝

תודה