

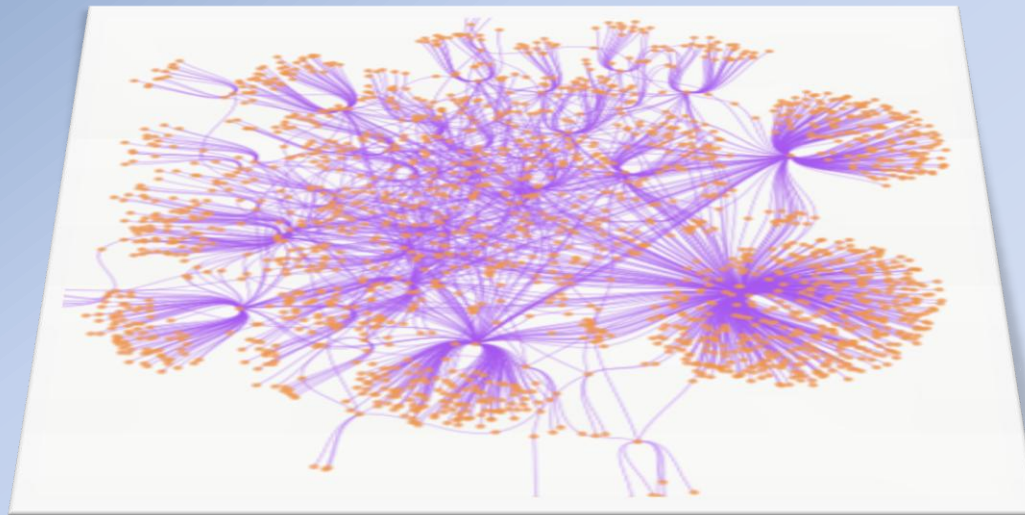
Modularity Detection in Protein-Protein Interaction Networks

Tejaswini Narayanan

PhD Candidate

Department of Electrical and Computer Engineering

Bioinformatics and Systems Biology Lab



11 Feb 2011

Agenda



- Introduction to modularity and community structures in networks
 - Fundamentals of Graph theory
 - Community structures in networks
- Existing betweenness algorithms
 - Random-walk betweenness
 - Current-flow betweenness
 - Edge-betweenness
- Insight into Edge-betweenness algorithm
 - Modularity factor
- Novel stopping criterion
 - *Geometric mean* approach
 - Comparison with NG algorithm
- A Variational Bayes approach to modularity detection
- Applications
 - Biological Interpretation of the Yeast Network
- Conclusions

Introduction to modularity and community structures in networks

Fundamentals of Graph theory

Network:

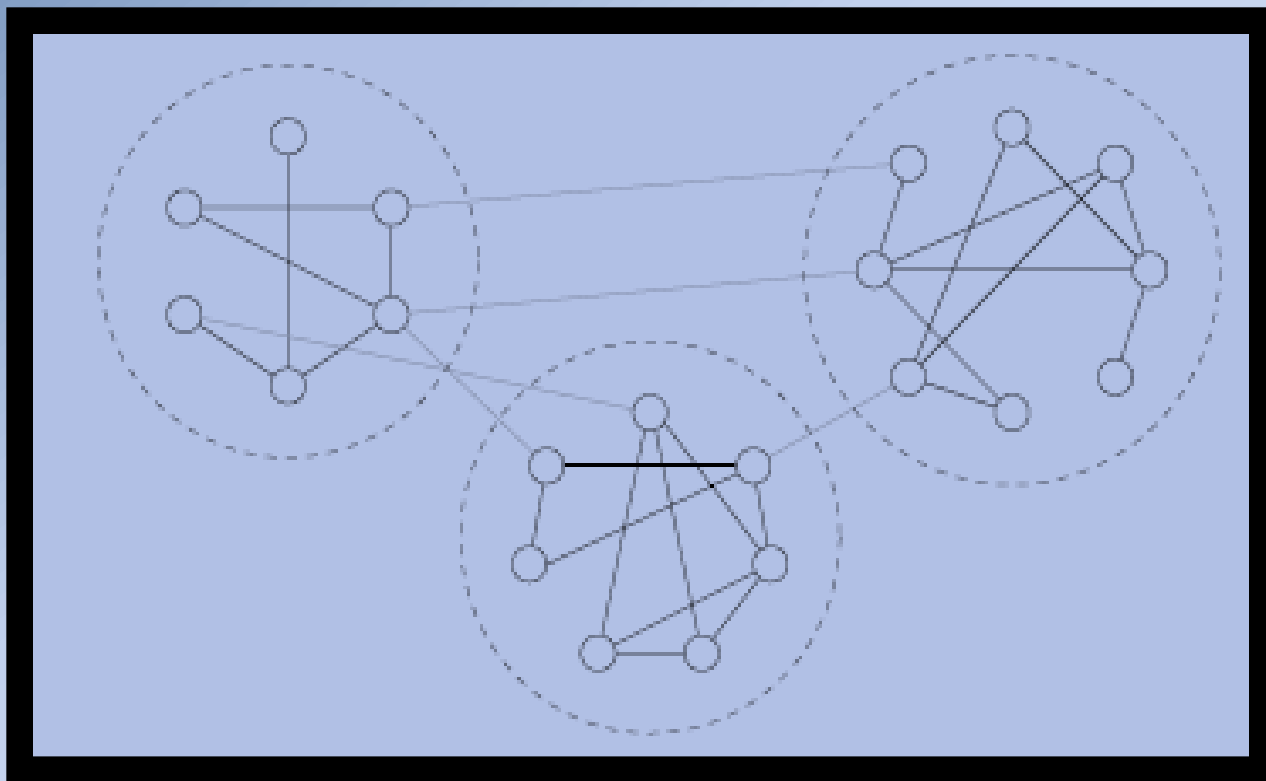
Collection of *nodes*, logically connected to each other by *edges*, giving some information about the nodes' relationships.

Network	Nodes	Edges
Electrical n/w	V/I source/sink	Resistance
Specific intra & internet	Computer terminals	Cables/ connections
Protein interaction	Proteins	Protein-protein interaction
Metabolic reactions	Metabolites	Reaction/ interaction between metabolites
Neuronal connectivity	Neurons	Axons
Social	Individuals	Social interactions

Community structures in networks

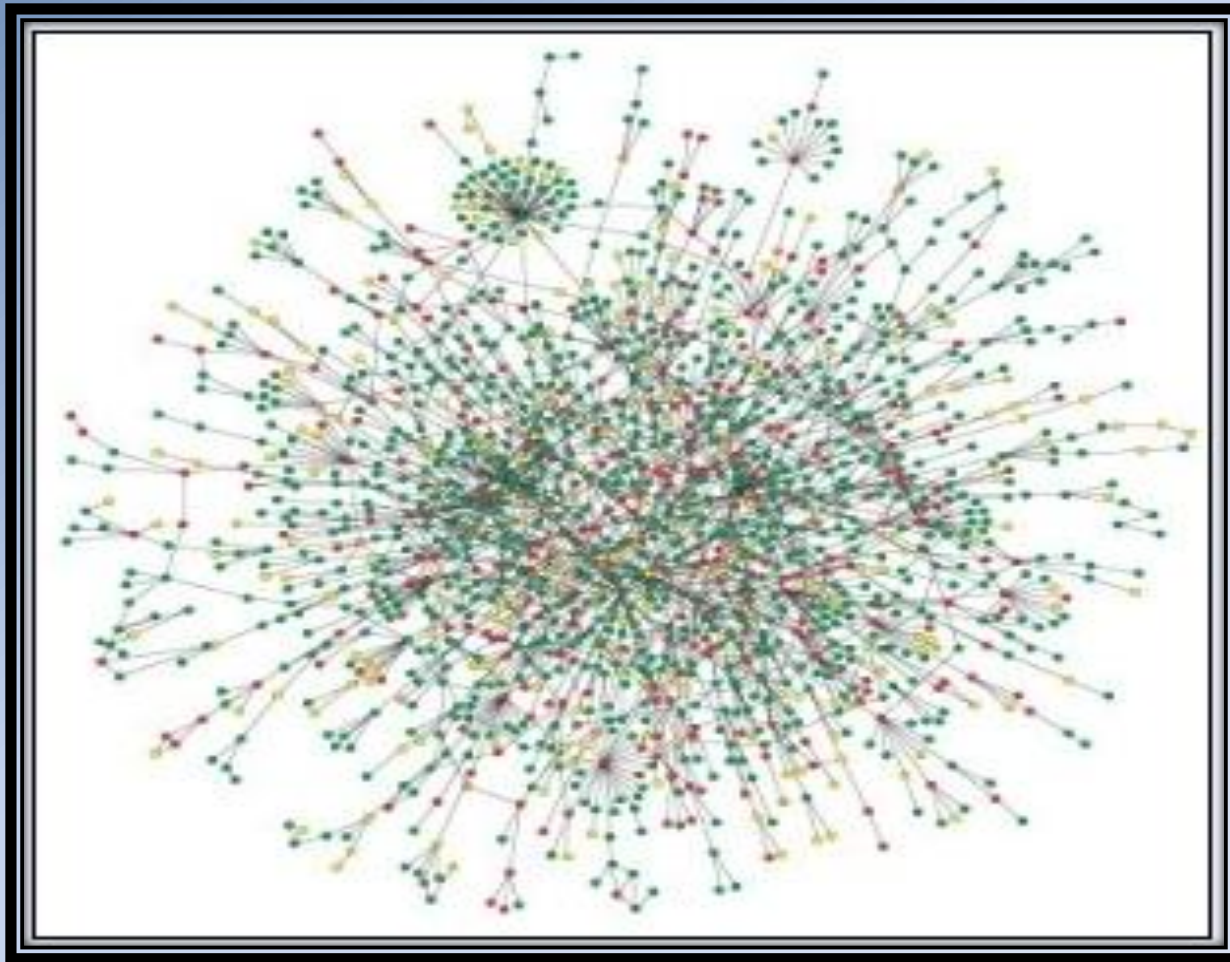
- **Definition**

“The division of network nodes into groups within which the network connections are dense, but between which they are sparser”

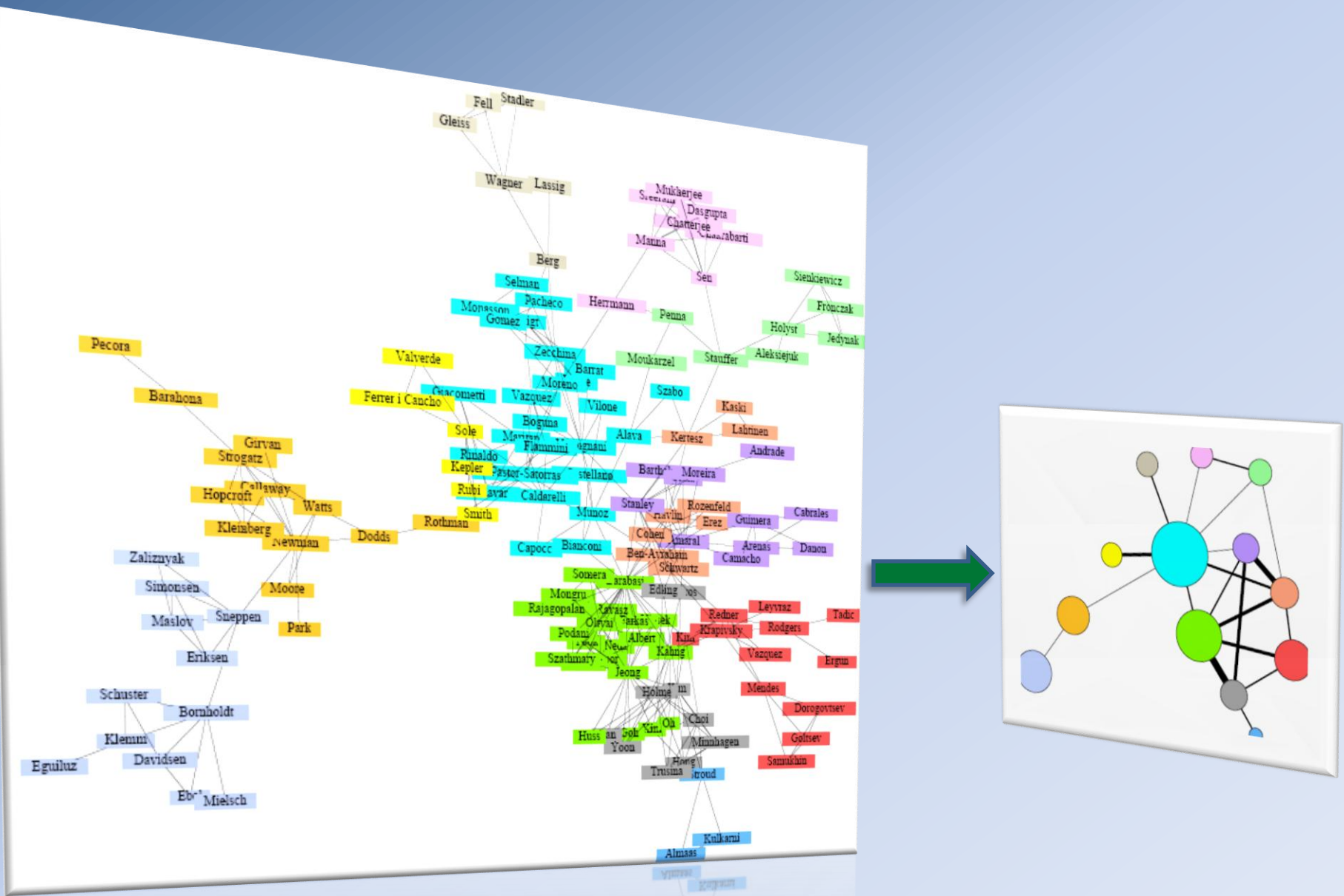


- **Necessity for definition:**

A typical human metabolic reaction graphs have about 88K edges!

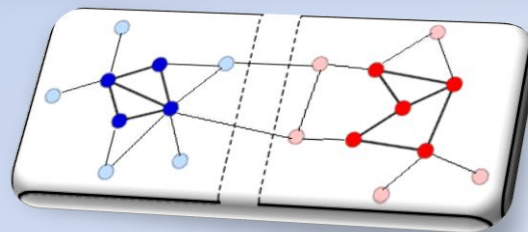
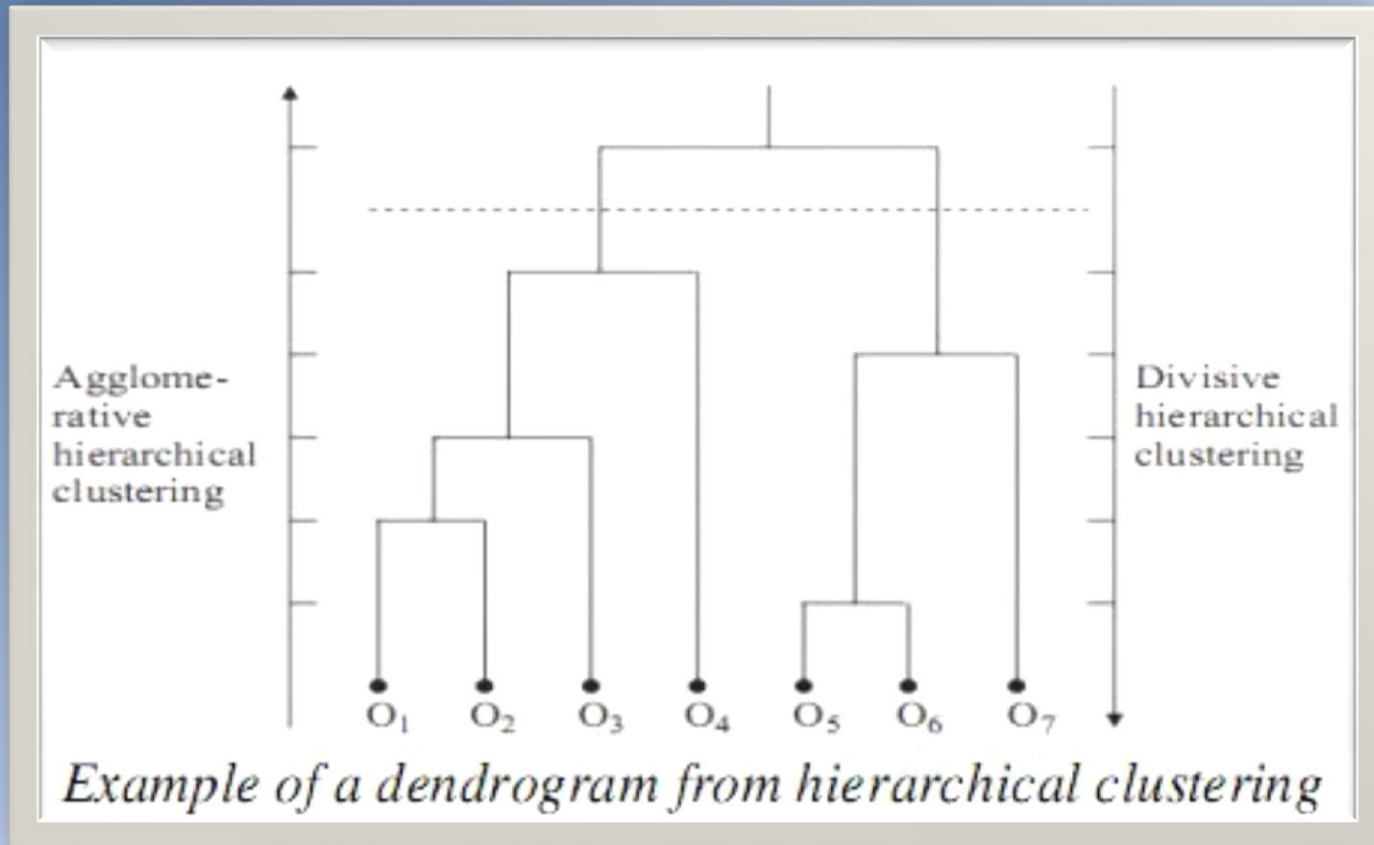


A simple 'moduled' representation of the graph is easier to work with.



Hierarchical Clustering

“Aimed at discovering *natural divisions* of networks into groups, using metrics of similarity / strength of connection between vertices”



- Addition of edges
- Similarity between vertex pairs
- Disadvantage:

- Removal of edges
- Least similar connected pair of vertices removed

Existing betweenness algorithms

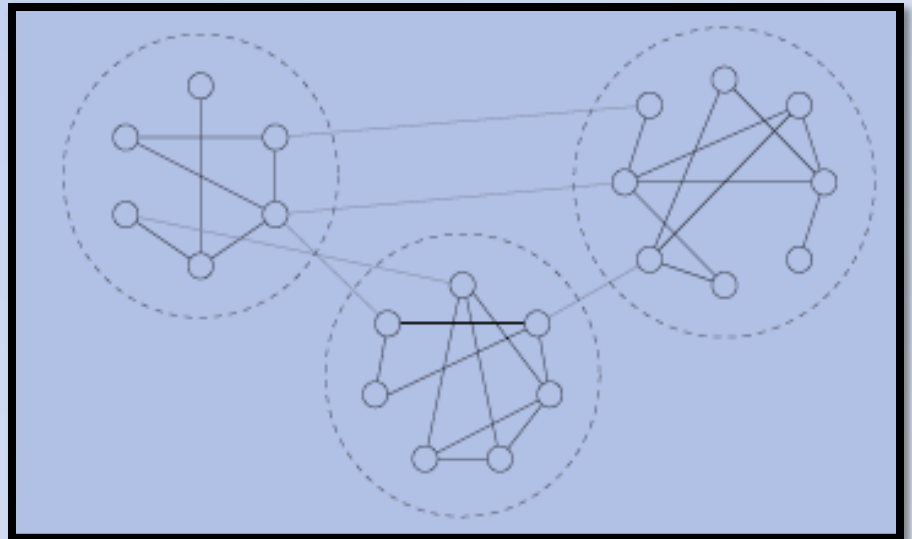
[1] Random-walk betweenness:

[2] Current-flow betweenness

- The least resistance path.
- $| \text{current} |$ along an edge summed

[3] Edge-betweenness:

- Edge-Betweenness -> '*rush*' -> 'shortest path betweenness'



Insight into *Edge-betweenness* algorithm

- Paths between inter-community vertices must pass through the relatively fewer edges
- Expected to be largest for intercommunity edges.

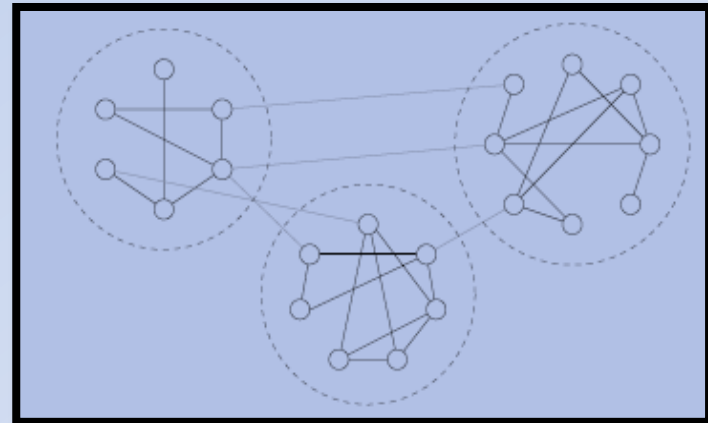
Newman and Girvan's two step algorithm:

1. Iterative removal of edges
2. Recalculation step

Modularity Factor: measure of quality of a particular division of network

- Output is a dendrogram
- Q = fraction of within-community edges – E[same quantity in a network with the same community divisions], but random connections between the vertices.
- Range of Q

$$Q = \sum_i (e_{ii} - a_i^2) = \text{Tr } e - \|e^2\|,$$



Novel stopping criterion

- Need for an effective stopping criterion
 - Time and memory constraints

Geometric Mean approach

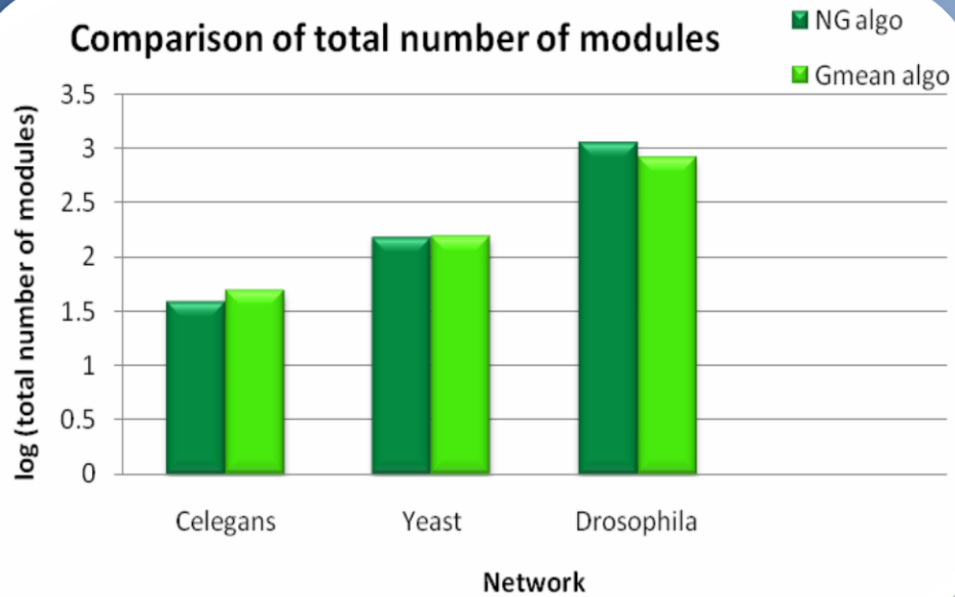
1. EB for edges in original network.
2. Gmean calculated.
3. while(value of EB of edge to be removed $<$ Gmean)
 - {
 - 4. Edge with the highest betweenness removed.
 - 5. Betweenness recalculated.
 - }

Comparison with NG

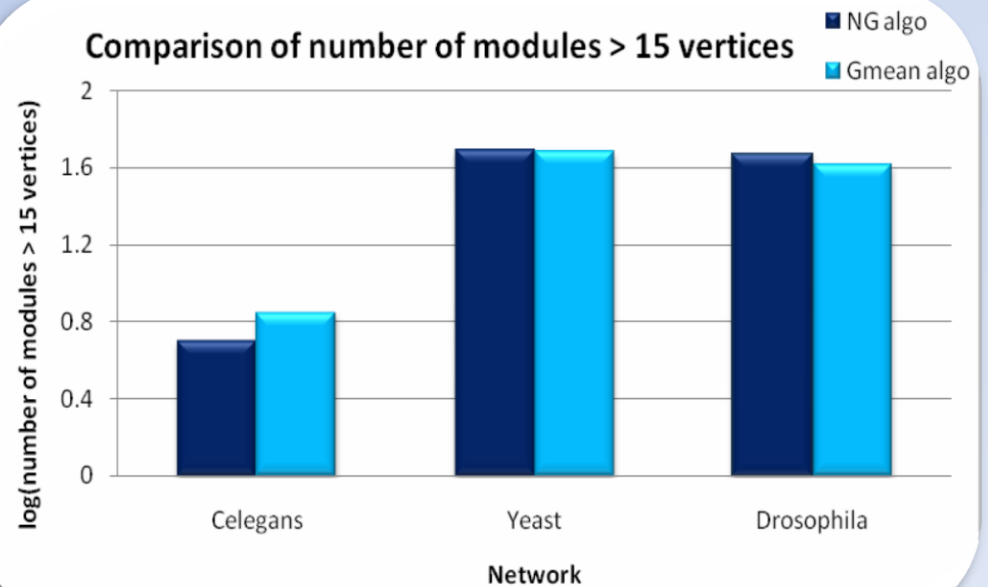
- Data set :
- Data preprocessing

Network	Source	#Vertices [Original network]	# Edges	
			Original network	Network considered
Celegans	[16]	453	4,596	2,025
Yeast*	<i>Biogrid</i>	3,654	15,316	9,946
Drosophila	<i>Biogrid</i>	7,666	25,649	25,433

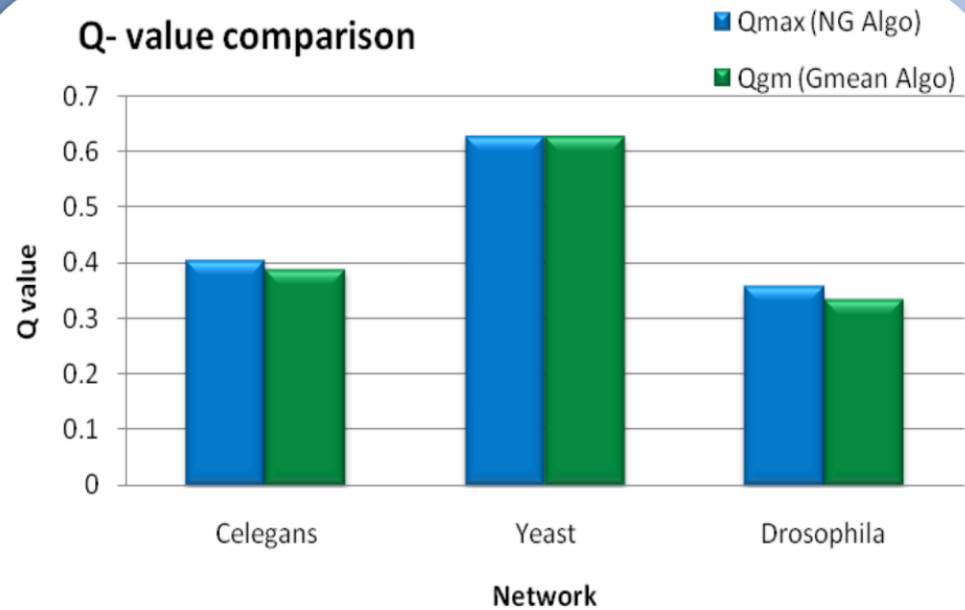
Comparison of total number of modules



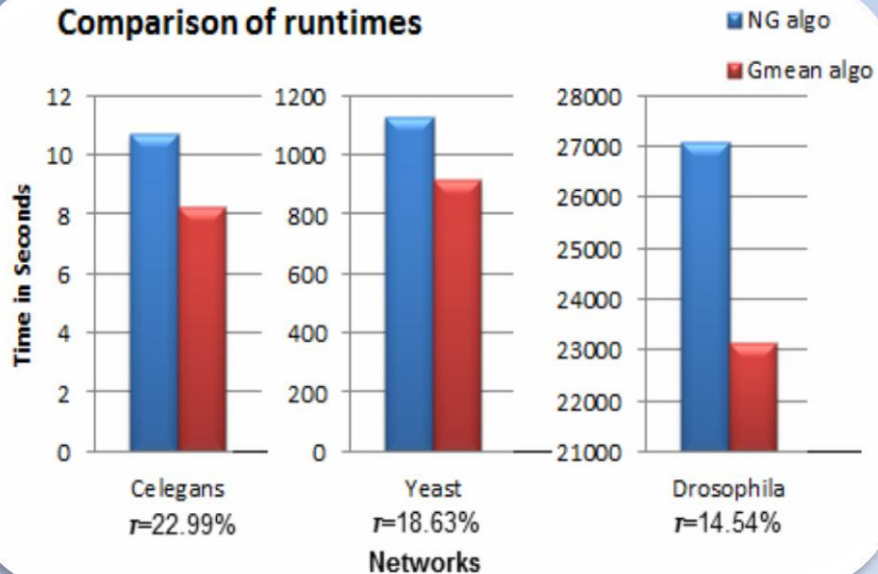
Comparison of number of modules > 15 vertices



Q- value comparison



Comparison of runtimes

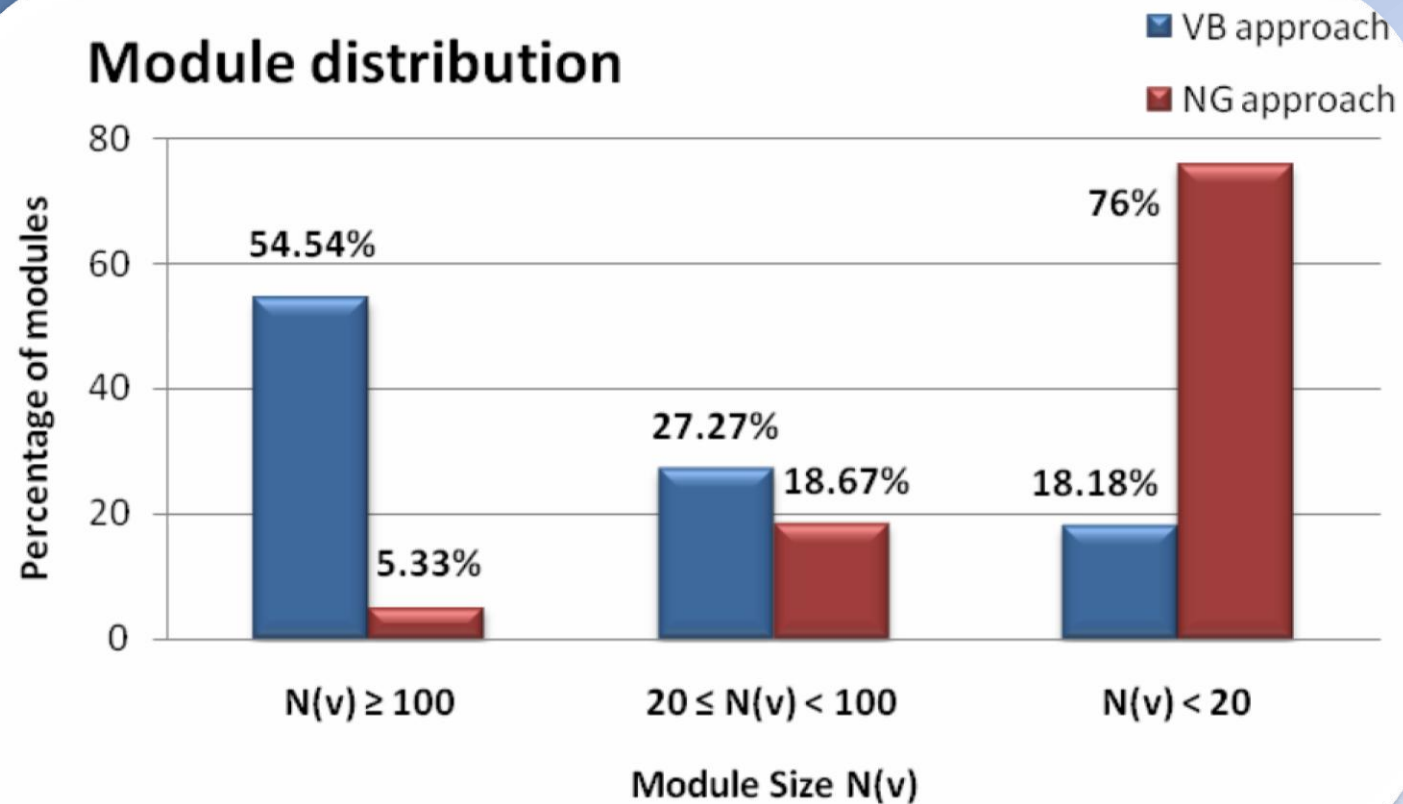


A Variational Bayes approach to modularity detection

- VB model is a framework for *inferring* the:
 - number of modules,
 - model parameters, and
 - module assignments.
- Module detection is posed as *inference* of a *latent variable* within a *probabilistic* model.
 - Given \mathbf{A} , K^* is determined
 - $K^* = \operatorname{argmax}_K p(K|\mathbf{A})$
 - Infer posterior distributions over model parameters and latent module assignments.
 - $p(K|\mathbf{A})$ is *evidence*.

Reference: Jake M. Hofman and Chris H. Wiggins, *A Bayesian Approach to Network Modularity*, *Phys. Rev. Lett.* 100, 258701 (2008)

Comparison with NG algorithm:



Q value comparison:

VB approach: 0.5431

Qmax of the NG algorithm: 0.6254

Applications

Biological Interpretation of Yeast Network

Method:

Simulating the probability of expression of a given *functional category* (FC) in a module of a given size.

Eg: Module 1 / 109 / transcription regulator activity / 32

Observations / Interpretation:

1. Resulting modules correspond to functional units (*GO Slim mapping*)

2. Biological Interpretation for Yeast modules

Distribution of function is *non-random* across structure

- Enriched FC 90.91% of modules.
- 9 FCs are uniquely / highly expressed in exactly 1 module
- 5/11 clusters have 1 FC, uniquely expressed

	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	Key
ligase activity												$p < 10^{-5}$
hydrolase activity												$10^{-5} \leq p < 10^{-1}$
protein kinase activity												$10^{-1} \leq p$
transferase activity												
transporter activity												
DNA binding												
transcription regulator activity												
phosphoprotein phosphatase activity												
molecular_function												
other												
enzyme regulator activity												
oxidoreductase activity												
protein binding												
structural molecule activity												
lipid binding												
RNA binding												
peptidase activity												
nucleotidyltransferase activity												
lyase activity												
isomerase activity												
signal transducer activity												
helicase activity												
motor activity												
not_yet_annotated												
translation regulatory activity												

Biological Interpretation for Yeast modules using GO slim functional annotation

2. Supports specificity in *functional enrichment* across modules.

- For all 17 occurrences of enhanced FC representation (FCR)

$$\frac{\# \text{ vertices (in M)} \equiv \text{functional activity A}}{\text{Vertex cardinality (of M)}}$$

- All enhanced FCR (except for one) have at least 10 % of the total module size constituted by vertices \equiv enhanced FC

Module	Module size	Enhanced GO slim functional category	Vertex Cardinality	Percentage	Key
					$r > 30$
M1	109	transcription regulator activity	32		$30 \geq r > 10$
M2	21	oxidoreductase activity	10		$10 \geq r$
	21	structural molecule activity	7		
M3	363	RNA binding	62		
M4	10	hydrolase activity	10		
M5	29	signal transducer activity	10		
	29	protein binding	4		
M6	308	DNA binding	53		
	308	transcription regulator activity	80		
	308	nucleotidyltransferase activity	30		
M8	8	DNA binding	7		
M9	739	DNA binding	88		
	739	protein binding	161		
	739	structural molecule activity	82		
M10	499	transporter activity	154		
M11	34	DNA binding	11		
	34	nucleotidyltransferase activity	8		

Percentage of module size constituting the enhanced GO slim functional annotation

3. Relative distribution of vertices in a given GO Slim FC across modules :

- For each of the 10 highly expressed FCs,

$$\frac{\text{\# vertices (in M) where F is highly expressed}}{\text{total \# vertices} \equiv F}$$

- For all enhanced FCR, (except for one) there is at least 1 module which contains a minimum 10 % of vertices \equiv that FC across all modules.

GO slim Functional Activity	Vertex Cardinality (across all modules)	Module	Vertex Cardinality (per module)	Percentage	Key
					$r > 30$
					$30 \geq r > 10$
					$10 \geq r$
hydrolase activity	330	M4	10		
transporter activity	199	M10	154		
DNA binding	193	M6	53		
	193	M8	7		
	193	M9	88		
	193	M11	11		
transcription regulator activity	169	M1	32		
	169	M6	80		
oxidoreductase activity	70	M2	10		
protein binding	336	M9	161		
	336	M5	4		
structural molecule activity	145	M9	82		
	145	M2	7		
RNA binding	135	M3	62		
nucleotidyltransferase activity	47	M6	30		
signal transducer activity	26	M5	10		

Percentage of module size constituting the enhanced GO slim functional annotation

Conclusions

- NG's edge-betweenness algorithm is state-of-art, but too complex
- Novel stopping criterion (Gmean) increases efficiency and usability
- A Variational Bayes approach to modularity detection
- Applications- Biological Interpretation of the Yeast Network
- Future Work:
 - Time complexities and run times comparison (VB and NG)
 - VB's applicability to other PPI networks (eg. the H. Sapiens network)
 - A comprehensive study of *performance characteristics* across other modularity detection algorithms.

Questions?

Email:

tnarayan@ucsd.edu