Understanding Network Concepts in Modules

Dong J, Horvath S (2007) BMC Systems Biology 2007, 1:24

Content

- Here we study network concepts in special types of networks, which we refer to as approximately factorizable networks. In these networks, the pairwise connection strength (adjacency) between 2 network nodes can be factored into node specific contributions, named node 'conformity'.
- *Scope: Our results apply to modules in gene co-expression networks and to special types of modules in protein-protein interaction networks*

Background

- Network concepts are also known as network statistics or network indices
  - Examples: connectivity (degree), clustering coefficient, topological overlap, etc
- Network concepts underlie network language and systems biological modeling.
- Dozens of potentially useful network concepts are known from graph theory.
- Question: How are seemingly disparate network concepts related to each other?

Review of *some* fundamental network concepts

Connectivity

- Gene connectivity = row sum of the adjacency matrix
  - For unweighted networks=number of direct neighbors
  - For weighted networks= sum of connection strengths to other nodes

$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

Density

- Density= mean adjacency
- Highly related to mean connectivity

$$Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{S_1(k)}{n(n-1)} = \frac{mean(k)}{n-1}$$

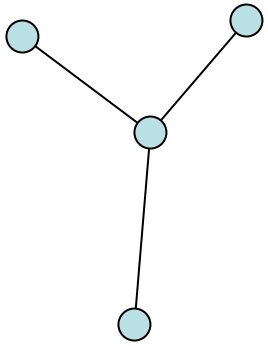
where n is the number of network nodes.

Centralization

$$Centralization = \frac{n}{n-2}\left(\frac{\max(k)}{n-1} - Density\right) \approx \frac{\max(k)}{n-1} - Density$$

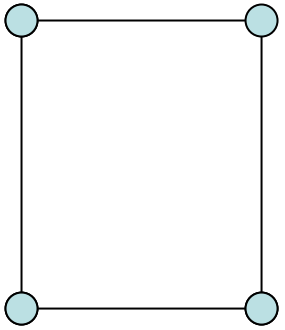
= 1 if the network has a star topology

= 0 if all nodes have the same connectivity



Centralization = 1

because it has a star topology



Centralization = 0

because all nodes have the same connectivity of 2

Heterogeneity

- Heterogeneity: coefficient of variation of the connectivity
- Highly heterogeneous networks exhibit hubs

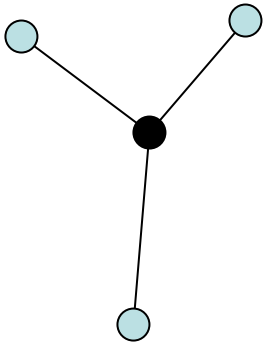
$$Heterogeneity = \frac{\sqrt{variance(k)}}{mean(k)}$$

Clustering Coefficient

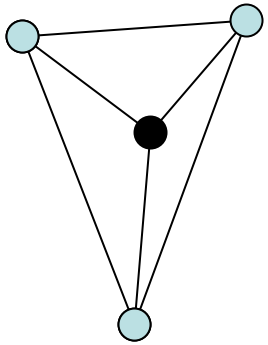
Measures the cliquishness of a particular node
« A node is cliquish if its neighbors know each other »

$$ClusterCoef_i = \frac{\sum_{l \neq i} \sum_{m \neq i,l} a_{il} a_{lm} a_{mi}}{\left(\sum_{l \neq i} a_{il} \right)^2 - \sum_{l \neq i} a_{il}^2}$$

This generalizes directly to weighted networks (Zhang and Horvath 2005)



Clustering Coef of the black node = 0



Clustering Coef = 1

The topological overlap dissimilarity is used as input of hierarchical clustering

$$TOM_{ij} = \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$DistTOM_{ij} = 1 - TOM_{ij}$$

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2007) to higher order interactions
- Generalized in Li and Horvath (2006) to multiple nodes

Question: What do all of these fundamental network concepts have in common?

Answer: They are tensor valued functions of the off-diagonal elements of the adjacency matrix A.

CHALLENGE

Challenge: Find relationships between these and other seemingly disparate network concepts.

- For general networks, this is a difficult problem.
- But a solution exists for a special subclass of networks: approximately factorizable networks
- Motivation:

  modules in larger networks are often approximately factorizable

Approximately factorizable networks and conformity

We define an adjacency matrix A to be exactly factorizable if, and only if, there exists a vector CF with non-negative elements such that

$a_{ij} = CF_i CF_j$ for all $i \neq j$

We also define the concept of conformity for a general, non-factorizable network.

Idea: approximate A with an exactly factorizable adjacency matrix

$A_{CF} = CFCF^{\tau} - diag(CF^2) + I$

We define the conformity as a maximizer of the factorizability function

$$F_A(v) = 1 - \frac{\sum_i \sum_{j \neq i} (a_{ij} - v_i v_j)^2}{\sum_i \sum_{j \neq i} (a_{ij})^2}$$

The conformity vector reduces the dimensionality of the adjacency matrix

- Note that the (symmetric) adjacency matrix contains n*(n-1)/2 parameters a(i,j).
- The conformity vector contains only n parameters CF(i)
- Thus, by focusing on the conformity based adjacency matrix, we effectively reduce the dimensionality of the adjacency matrix.
- This approximation is only valid if the network has high factorizability as defined on the next slide.

The higher $F(A)$, the better A_{CF} approximates A

- The factorizability $F(A)$ is normalized to take on values in the unit interval [0, 1].

$$F(A) = 1 - \frac{\| (A - I) - (A_{CF} - I) \|_F^2}{\| A - I \|_F^2}$$

Empirical observation: subnetworks comprised of module genes tend to have high factorizability F(A)>0.8

Applications: modules in
a) protein-protein networks
b) gene co-expression networks

The Topological Overlap Matrix Can Be Considered as Adjacency Matrix

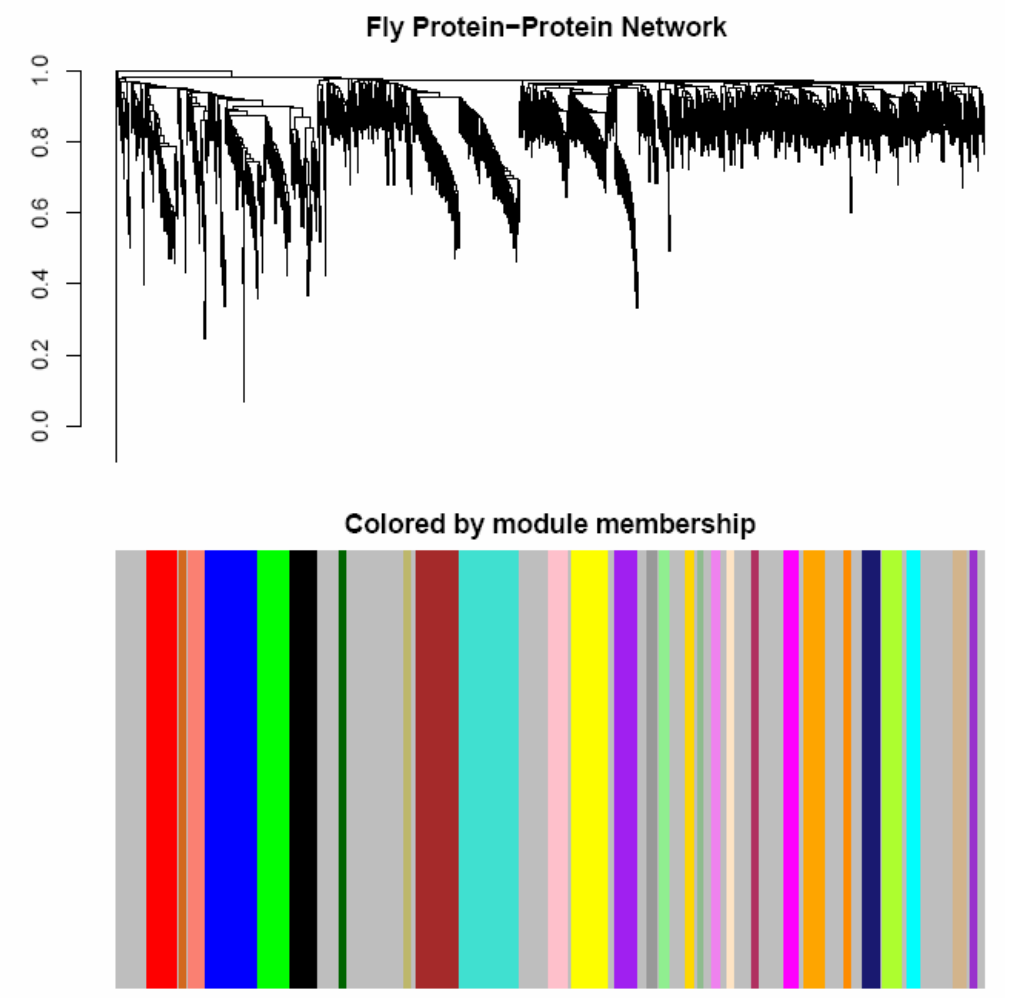
- Important insight for protein-protein interaction (PPI) networks:
- Since the matrix *TopOverlap[i,j]* is symmetric and its entries lie in [0, 1], it satisfies our assumptions on an adjacency matrix.
- Since the adjacency matrices of our PPI networks are very sparse, we replaced them by the corresponding topological overlap matrices.
- Roughly speaking, the topological overlap matrix can be considered as a 'smoothed out' version of the adjacency matrix.

Hierarchical clustering dendrogram and module definition.

Drosophila PPI network.

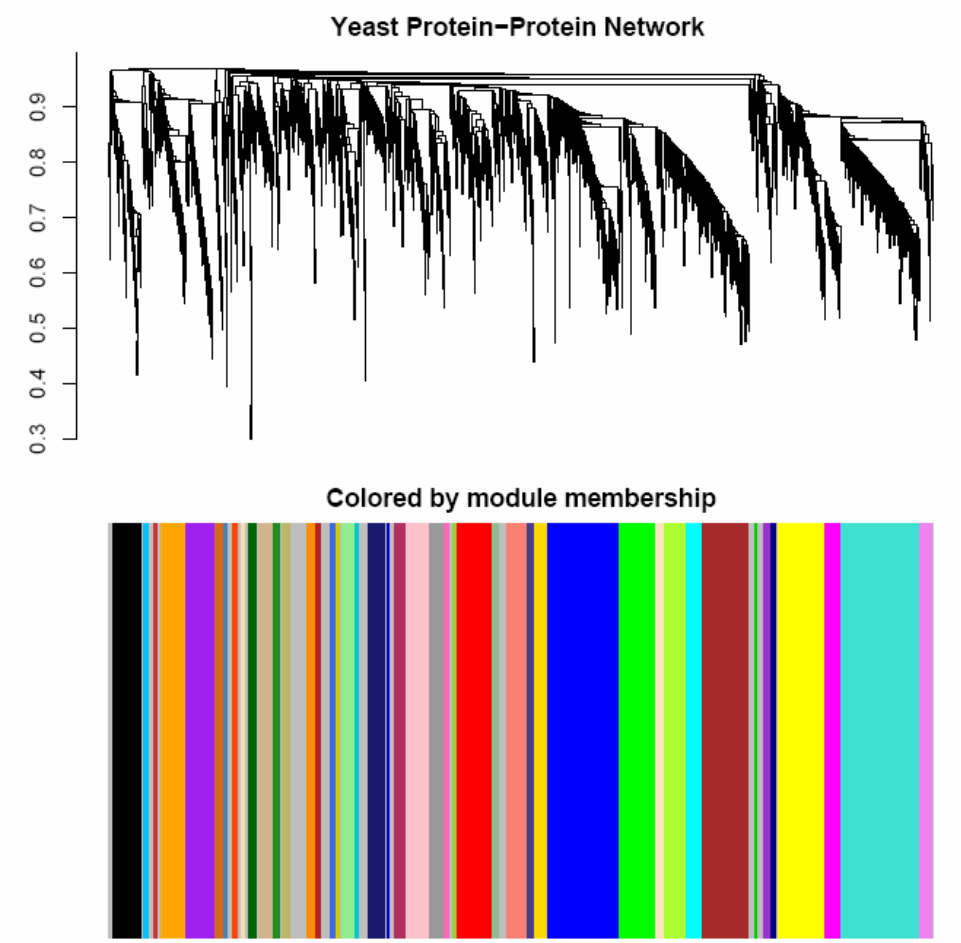
The color-band below the dendrogram denotes the modules, which are defined as branches in the dendrogram.
Of the 1371 proteins, 862 were clustered into 28 proper modules, and the remaining proteins are colored in grey;

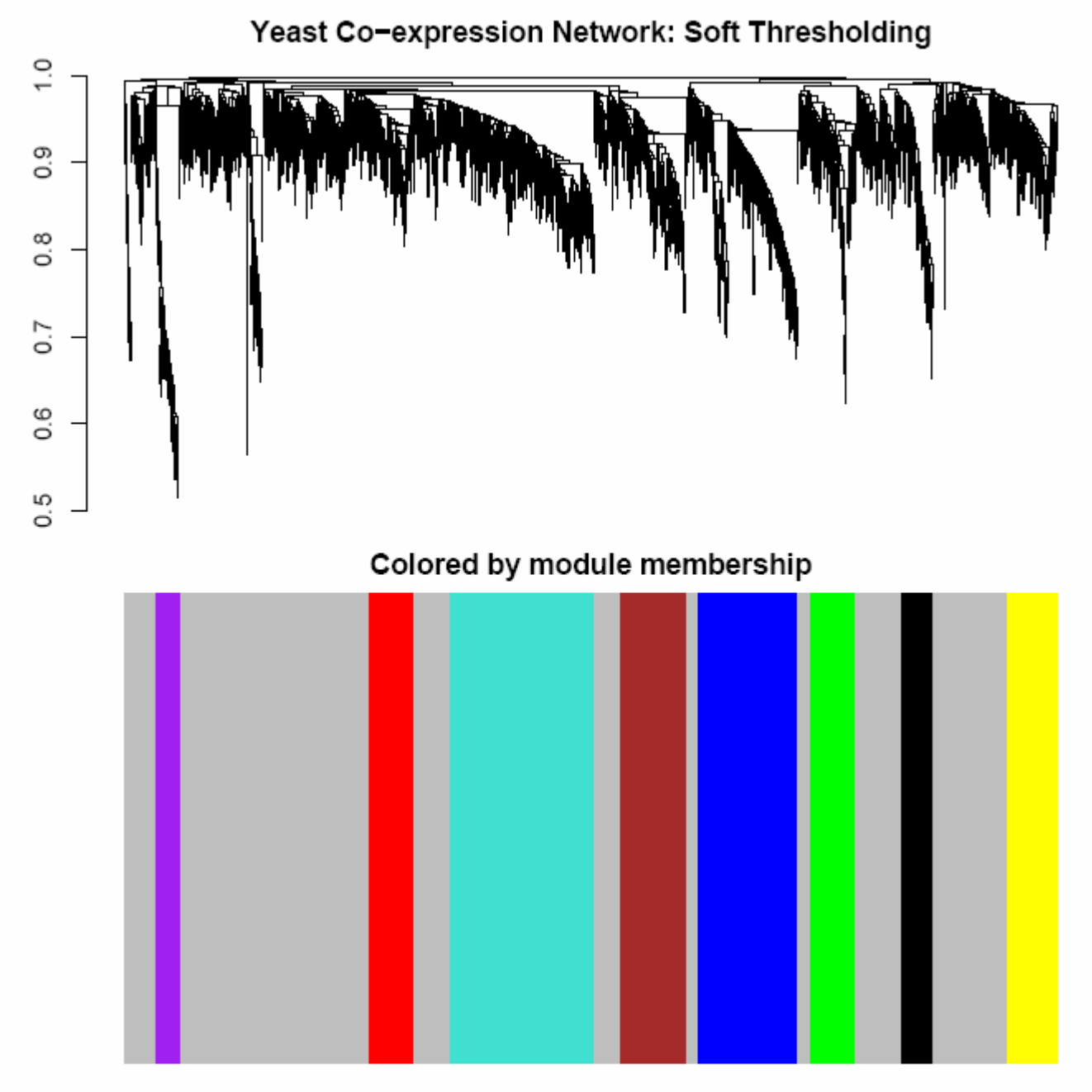
Recall that we used TOM instead of the original adjacency matrix as weighted network between the proteins

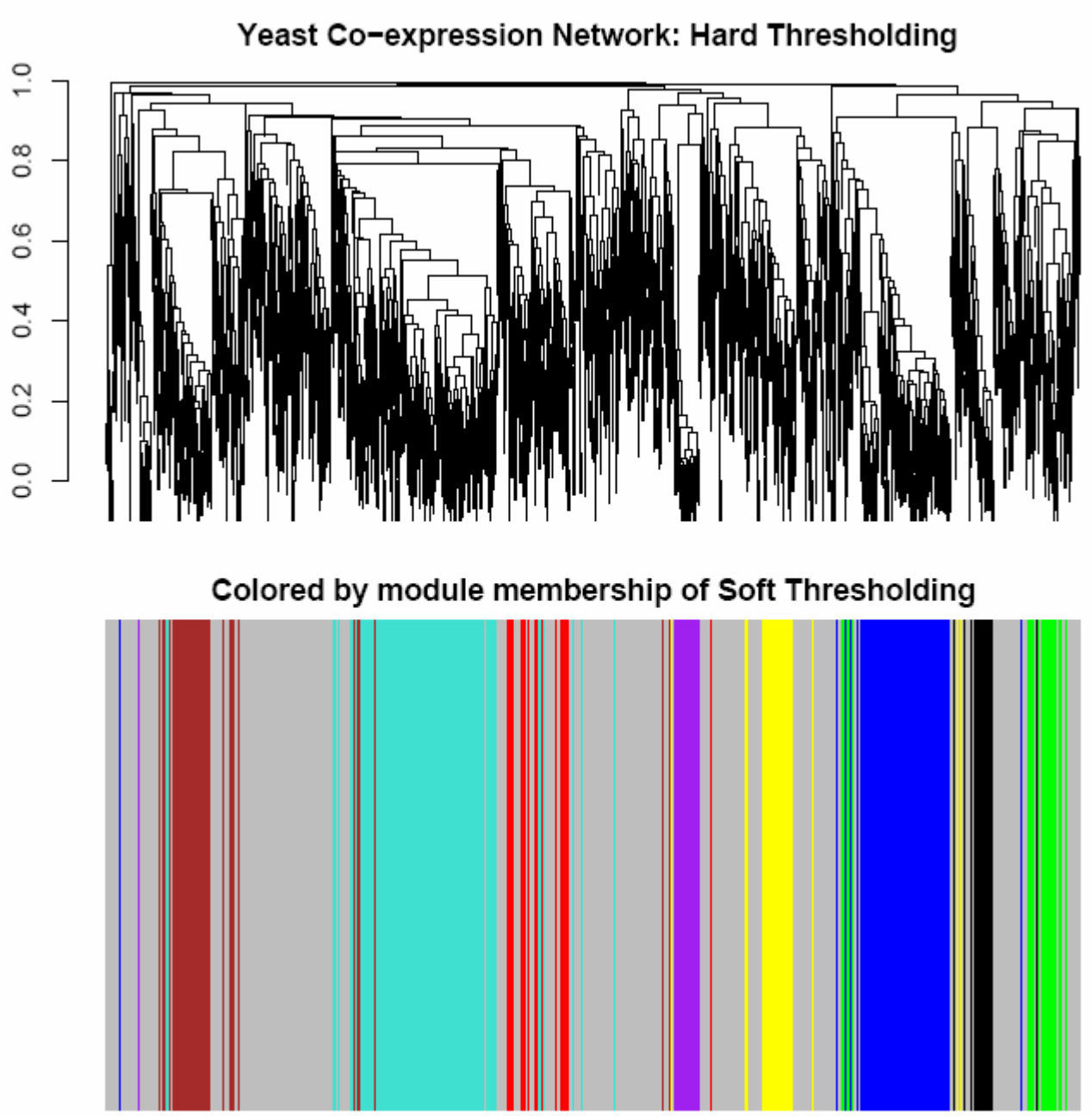


Hierarchical clustering dendrogram and module definition.

Yeast PPI network




Yeast Co−expression Network: Soft Thresholding
1.0
0.9
0.8
0.7
0.6
0.5
Colored by module membership


Yeast Co−expression Network: Hard Thresholding
1.0
0.8
0.6
0.4
0.2
0.0
Colored by module membership of Soft Thresholding

Observation 1

- Sub-networks comprised of module nodes tend to be approximately factorizable.
- Specifically, they have high factorizability F(A)

Table 1: Summary of fundamental network concepts in real network applications.

	Fly Protein		Yeast Protein		Yeast (Weighted)		Yeast (Unweighted)	
Concept	Proper	Grey	Proper	Grey	Proper	Grey	Proper	Grey
Factorizability	.82 (.086)	.170	.85 (.100)	.200	.73 (.084)	.180	.62 (.130)	.110
Density	.21 (.074)	.017	.28 (.120)	.026	.08 (.056)	.005	.40 (.150)	.024
Centralization	.18 (.091)	.052	.20 (.055)	.036	.10 (.026)	.021	.41 (.110)	.140
Heterogeneity	.35 (.130)	.460	.36 (.140)	.430	.56 (.066)	.580	.51 (.097)	.830
Mean Cluster Coef.	.28 (.110)	.050	.36 (.120)	.093	.13 (.072)	.032	.72 (.087)	.370
Mean Conformity	.45 (.076)	.130	.51 (.120)	.150	.26 (.084)	.062	.63 (.100)	.120

We use both PPI and gene co-expression network data to show empirically that subnetworks comprised of module nodes are often approximately factorizable.

CAVEATS

- Approximate factorizability is a very stringent structural assumption that is not satisfied in general networks.
- Modules in gene co-expression networks tend to be approximately factorizable if the corresponding expression profiles are highly correlated,
- the situation is more complicated for modules in PPI networks: only after replacing the original adjacency matrix by a 'smoothed out' version (the topological overlap matrix), do we find that the resulting modules are approximately factorizable.

To reveal relationships between network concepts, we use a trick.

We focus attention to the approximate conformity based adjacency matrix.

$$A_{CF,app} = CFCF^{\tau} = [CF_i CF_j]$$

- Strictly speaking it violates our assumption on an adjacency matrix since its diagonal elements are not 1.
- It is very useful for defining *approximate conformity based network concepts.*
- Approximately conformity based network concepts have several theoretical advantages as we detail below.

Network Concept Functions

Abstract definition:

tensor-valued function of a general *n* × *n matrix M* = [*mij*] a general matrix.

Examples

$$Connectivity_i(M) = \sum_j m_{ij} = e_i^\tau M 1,$$

$$Density(M) = \frac{\sum_i \sum_j m_{ij}}{n(n-1)},$$

$$Centralization(M) = \frac{n}{n-2}\left(\frac{\max(M1)}{n-1} - Density(M) \right),$$

$$Heterogeneity(M) = \sqrt{\frac{n(1^\tau MM1)}{(1^\tau M1)^2} - 1},$$

$$TopOverlap_{ij}(M) = \frac{e_i^\tau MMe_j + e_i^\tau Me_j}{\min\{e_i^\tau M1, e_j^\tau M1\} + 1 - e_i^\tau Me_j},$$

$$ClusterCoef_i(M) = \frac{e_i^\tau MMMe_i}{e_i^\tau MB_M Me_i},$$

Table 2: Brief overview of different types of network concepts.

Input Matrix	Type of Concept	Example: Connectivity
$A - I$	fundamental	$Connectivity_i(A - I) = \sum_{j \neq i} a_{ij}$
$A_{CF} - I = \boldsymbol{CF}\,\boldsymbol{CF}^{\tau} - \mathrm{diag}(\boldsymbol{CF}^2)$	CF-based	$Connectivity_i(A_{CF} - I) = CF_i \sum_{j \neq i} CF_j$
$A_{CF,app} = \boldsymbol{CF}\,\boldsymbol{CF}^{\tau}$	approximate CF-based	$Connectivity_i(A_{CF,app}) = CF_i \sum_j CF_j$

A network concept arises by evaluating a *network concept function* on a special type of input matrix. We assume that the diagonal elements of the matrix $A - I$ are 0.

Question:

Find simple relationships between approximate CF based network concepts

$$k_{CF,app,i} = CF_i S_1(CF),$$

$$Density_{CF,app} = \frac{S_1(CF)^2}{n(n-1)} \approx \left(\frac{S_1(CF)}{n}\right)^2,$$

$$Centralization_{CF,app} = \frac{nS_1(CF)}{(n-1)(n-2)}\left(\max(CF) - \frac{S_1(CF)}{n}\right)$$

$$\approx \frac{S_1(CF)}{n}\left(\max(CF) - \frac{S_1(CF)}{n}\right),$$

$$Heterogeneity_{CF,app} = \sqrt{\frac{nS_2(CF)}{\left(S_1(CF)\right)^2} - 1},$$

$$ClusterCoef_{CF,app,i} = \left(\frac{S_2(CF)}{S_1(CF)}\right)^2,$$

$$TopOverlap_{CF,app,ij} \approx \frac{CF_i CF_j\left(S_2(CF)+1\right)}{\min(CF_i, CF_j)S_1(CF) + 1 - CF_i CF_j}$$

Observation 1

Major advantage of approximate CF-based network concepts: they exhibit simple relationships

Relationship between heterogeneity, density, and clustering coefficient

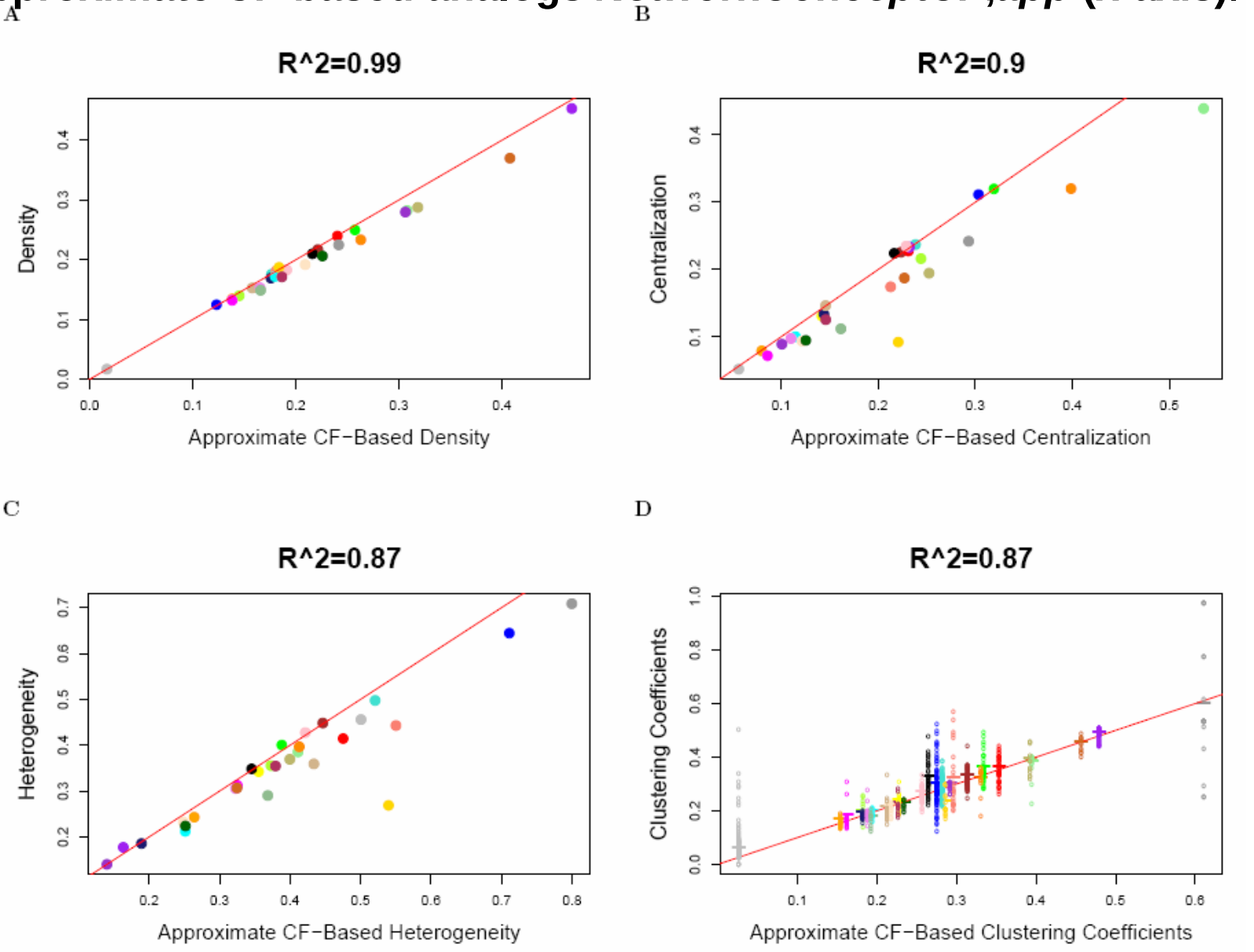
$$Heterogeneity_{CF,app} \approx \sqrt{\sqrt{\frac{ClusterCoef_{CF,app}}{Density_{CF,app}}} - 1}$$

$$ClusterCoef_{CF,app,i} \approx \left(1 + Heterogeneity_{CF,app}^{2}\right)^{2} \times Density_{CF,app}$$

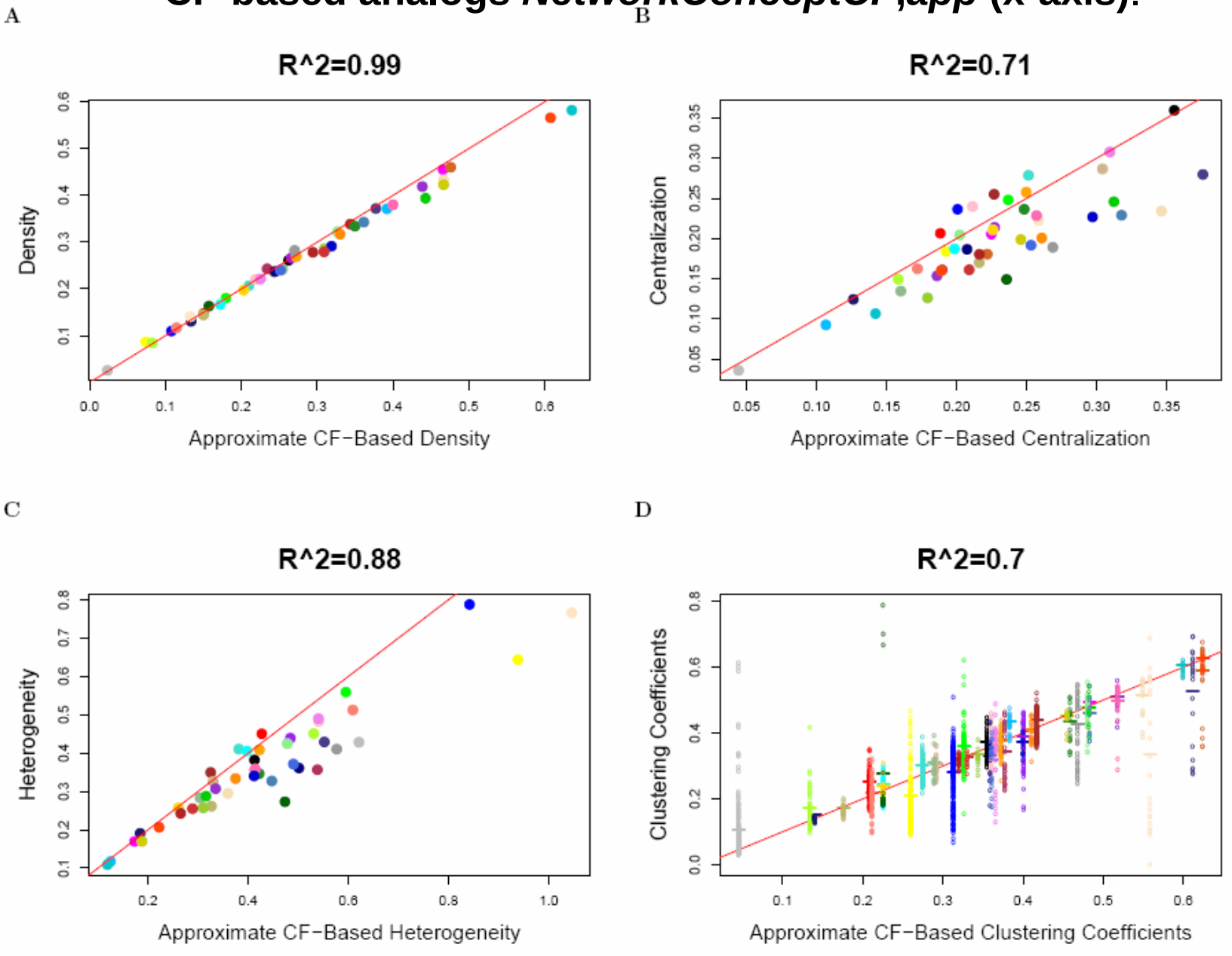
Observation 2

- *Fundamental network concepts are approximately equal to their approximate CF-based analogs in approximately factorizable networks*
- Recall that fundamental network concepts are defined with respect to the adjacency matrix
- Approximate CF-based network concepts are defined with respect to the conformity vector.

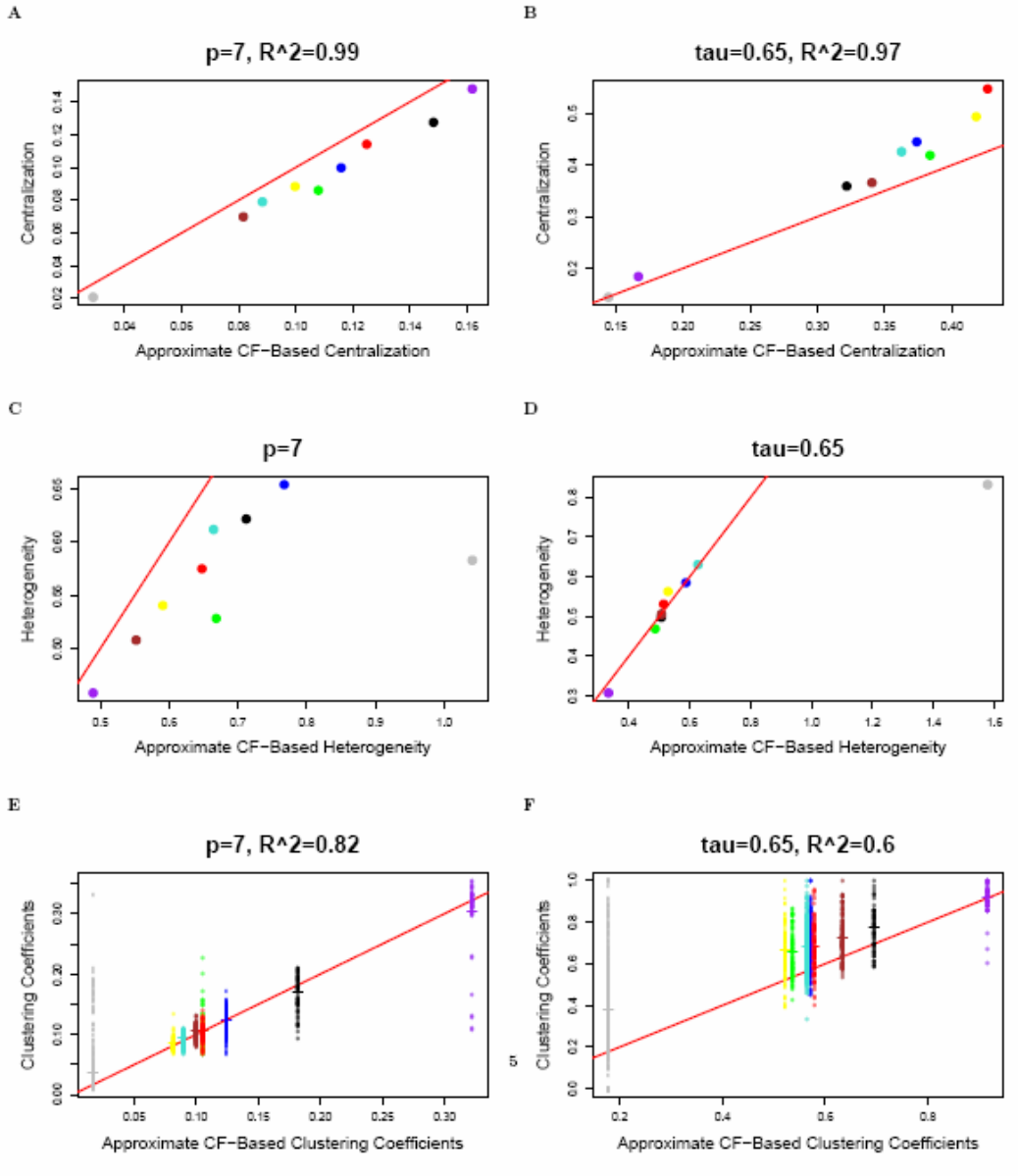
Drosophila PPI module networks: the relationship between fundamental network concepts *NetworkConcep* (y-axis) and their approximate CF-based analogs *NetworkConceptCF*,*app* (x-axis).



Yeast PPI module networks: the relationship between fundamental network concepts *NetworkConcep* (y-axis) and their approximate CF-based analogs *NetworkConceptCF*,*app* (x-axis).



Yeast gene co-expression module networks: the relationship between fundamental network concepts *NetworkConcept*(*A* - *I*) (y-axis) and their approximate CF-based analogs *NetworkConceptCF*,*app* (x-axis).



Observation 3

Approximate relationships between network concepts in modules

$$mean\left(ClusterCoef\right) \approx \left(1+Heterogeneity^2\right)^2 \times Density$$

$$TopOverlap_{ij} \approx \frac{\max(k_i, k_j)}{n} \times \left(1+Heterogeneity^2\right)$$

$$TopOverlap_{[1]j} \approx \frac{k_{[1]}}{n} \times \left(1+Heterogeneity^2\right)$$

$$\approx \left(Centralization + Density\right) \times \left(1+Heterogeneity^2\right)$$

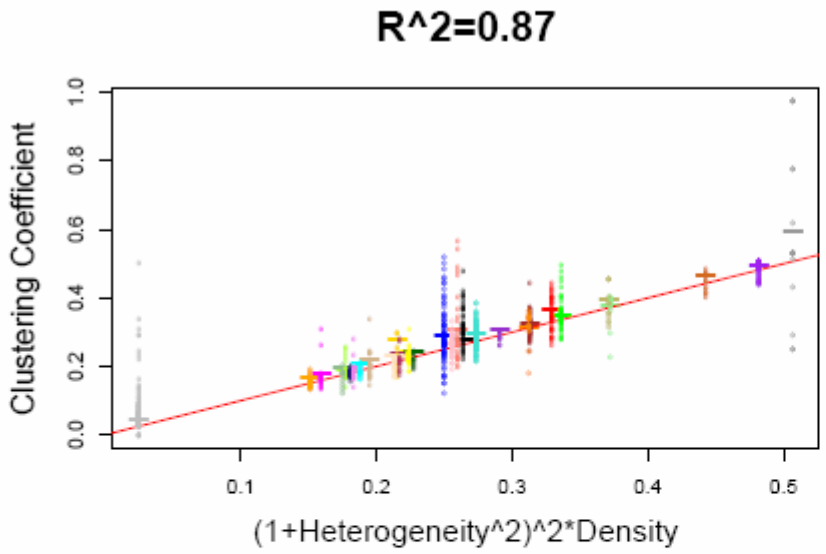
The topological overlap between two nodes is determined by the maximum of their respective connectivities and the heterogeneity.

Observation 3 (cont'd)

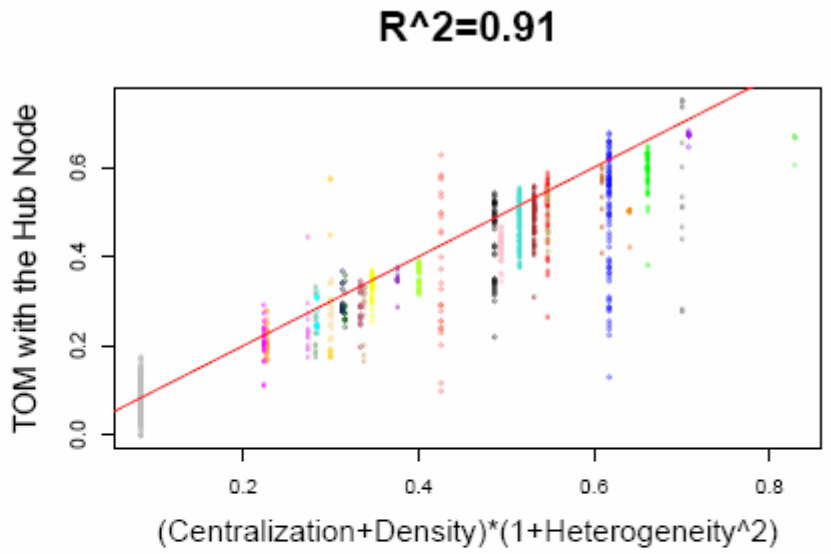
- The mean clustering coefficient is determined by the density and the network heterogeneity in approximately factorizable networks.
- Other examples involve the topological overlap
- Thus, seemingly disparate network concepts satisfy simple and intuitive relationships in these special but biologically important types of networks.

Drosophila PPI module networks: the relationship between fundamental network concepts.

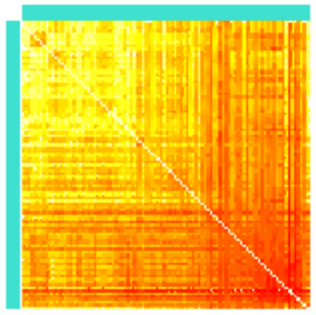
A



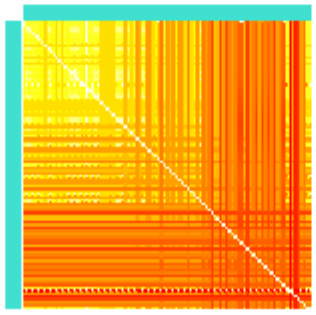
B



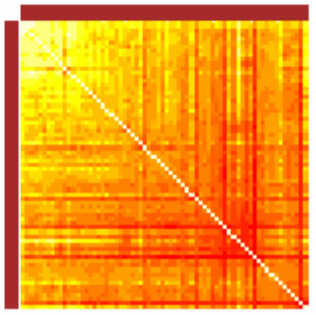
C



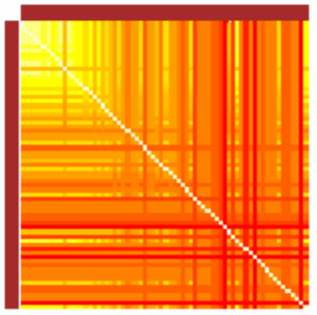
D



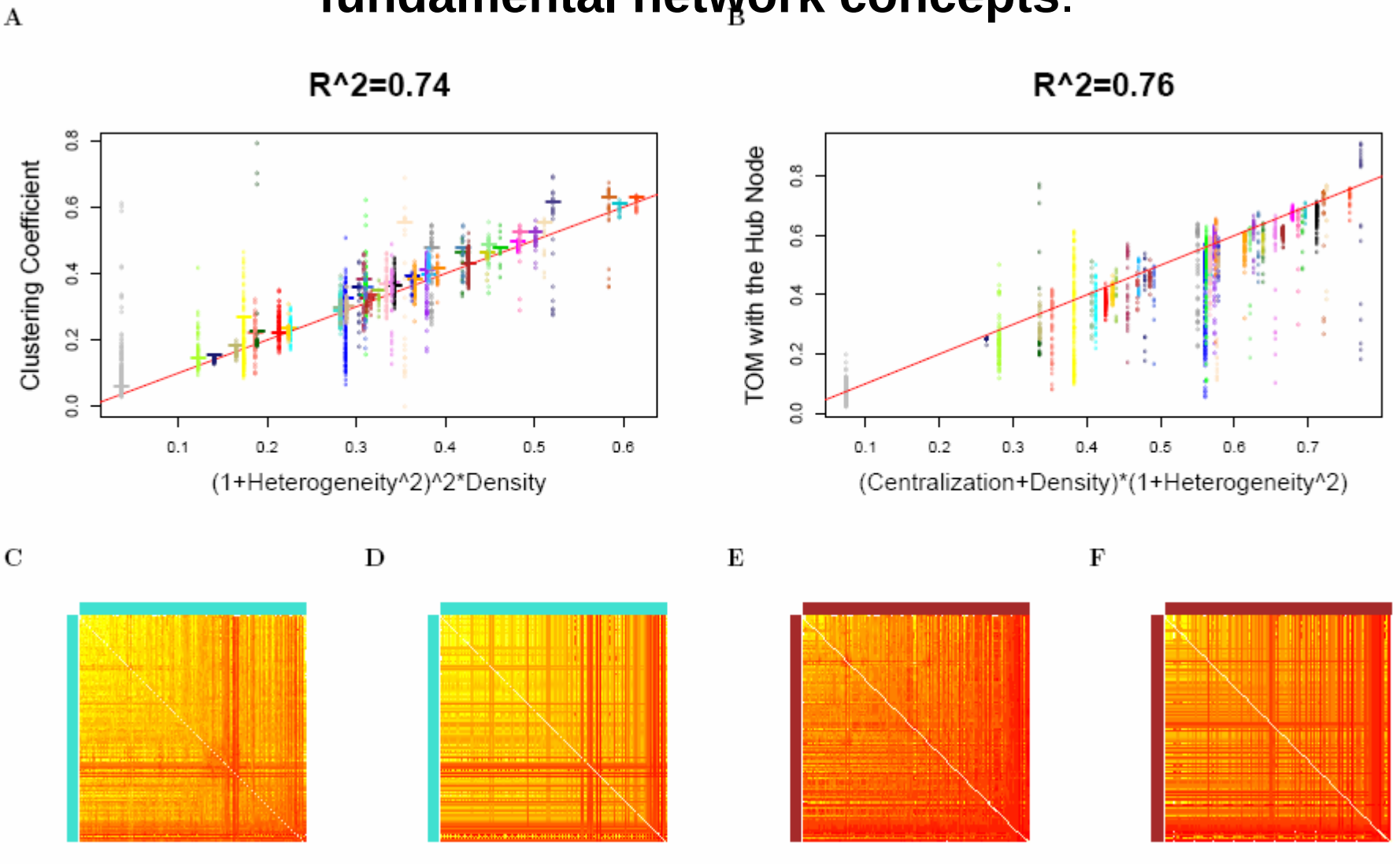
E



F

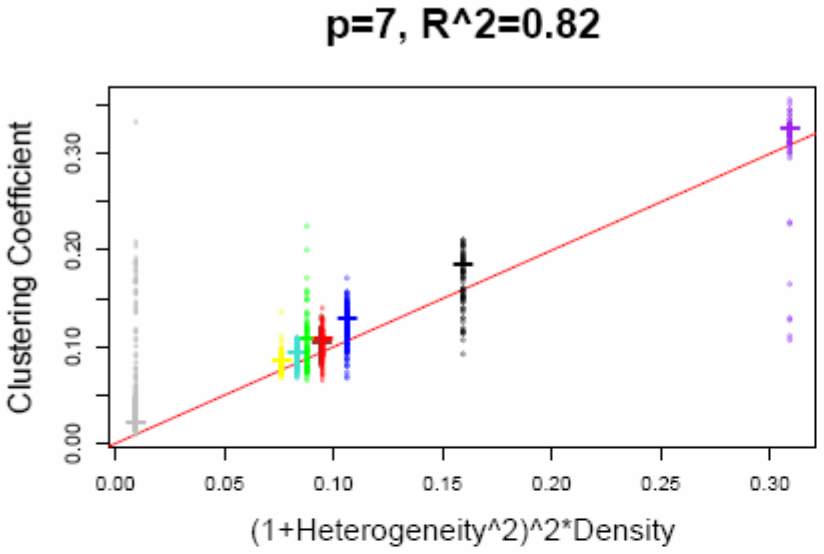


Yeast PPI module networks: the relationship between fundamental network concepts.

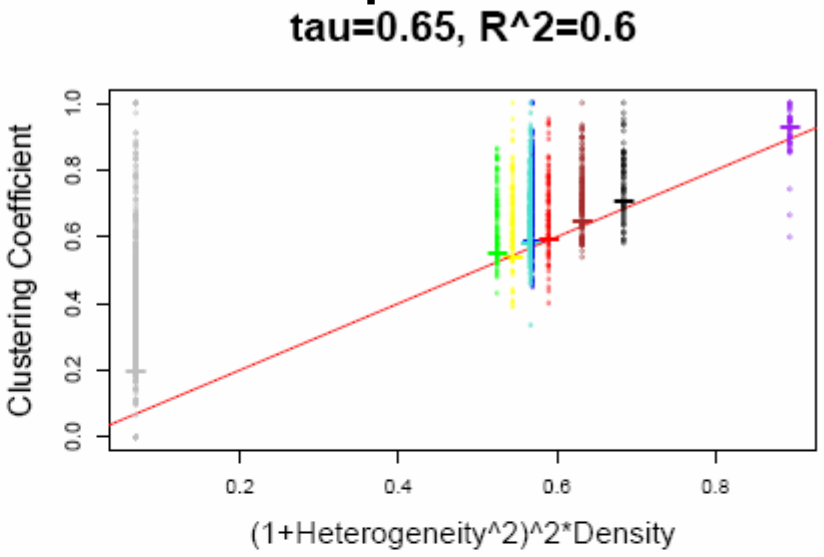


Yeast gene co-expression module networks: the relationship between fundamental network concepts.

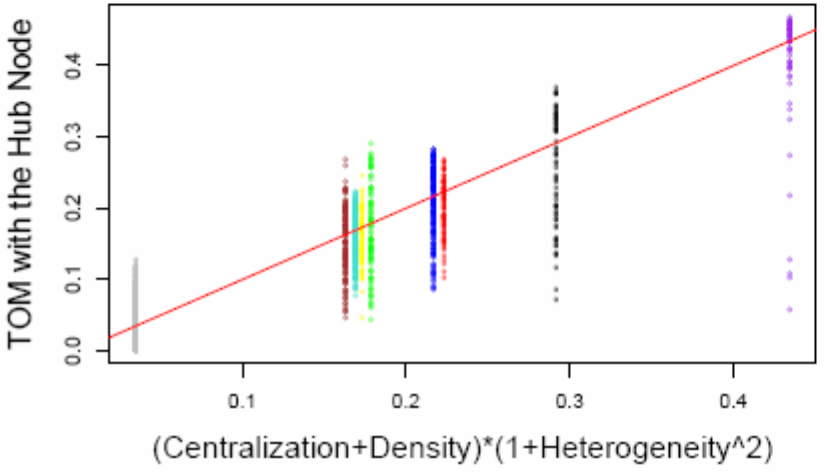
A



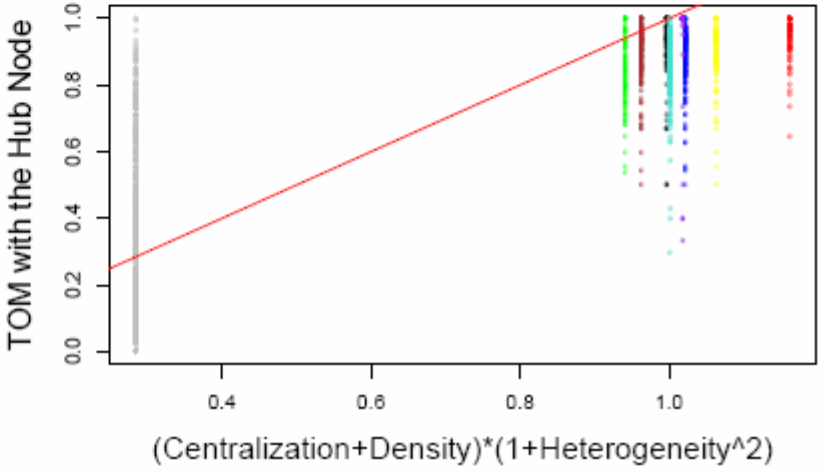
B



C



D



Observation 4: network concepts are simple function of the connectivity in approximately factorizable networks

$$ClusterCoef_i \approx \frac{\left(S_2(k)\right)^2}{\left(S_1(k)\right)^3},$$

$$TopOverlap_{ij} \approx \frac{\max(k_i, k_j)}{n} \times \frac{S_2(k)}{S_1(k)},$$

where the last approximation assumes

$$\frac{S_1(k)}{S_2(k)} \approx 0 \quad \text{and} \quad \frac{S_1(k) - k_i k_j}{\min(k_i, k_j) S_1(k)} \approx 0$$

Robustness to module definition

- In our applications, we define modules as branches of an average linkage hierarchical clustering tree based which uses the topological overlap measure as input.
- But our theoretical results are applicable to any approximately factorizable network.
- We find that the theoretical results are quite robust with respect to the underlying assumptions and are highly robust with respect to the module definition.

Summary

- We study network concepts in special types of networks, which we refer to as approximately factorizable networks.
- To provide a formalism for relating network concepts to each other, we define three types of network concepts: fundamental-, conformity-based-, and approximate conformity-based concepts.
- The approximate conformity-based analogs of fundamental network concepts have several theoretical advantages.
  1. they allow one to derive simple relationships between seemingly disparate networks concepts.

     For example, we derive simple relationships between the clustering coefficient, the heterogeneity, the density, the centralization, and the topological overlap.
  2. Approximate conformity-based network concepts is that they allow one to show that fundamental network concepts can be approximated by simple functions of the connectivity in module networks.

Appendix

What is the conformity?

We find that for most real networks, the conformity is highly related to the first eigenvector of the adjacency matrix, i.e.

$$CF(i) \propto \sqrt{d_1} u_1(i)$$

where

d_1 is the largest singular value of A

u_1 is the corresponding unit length eigenvector with positive components.

This insight leads to an iterative algorithm for computing CF, see the next slide

Monotonic algorithm for computing the conformity

$$\hat{A}(i-1) = A - I + diag\left(CF(i-1)^2\right)$$

$$CF(i) = \sqrt{d_1(i-1)} \times u_1(i-1)$$

$$F_A\left(CF(i)\right) \geq F_A\left(CF(i-1)\right)$$