04-630
Data Structures and Algorithms for Engineers

David Vernon
Carnegie Mellon University Africa

vernon@cmu.edu
www.vernon.eu

Lecture 26

Complex Networks

- Communities
  - Fundamental Hypothesis & Connectedness and Density Hypothesis
  - Strong and weak communities
  - Graph partitioning & Community detection
    - Hierarchical clustering
    - Girvan-Newman Algorithm
    - Modularity
    - Random Hypothesis
    - Maximum Modularity Hypothesis
    - Greedy algorithm for community detection by maximizing modularity
  - Overlapping communities
    - Clique percolation algorithm and CFinder

This lecture is based on Chapters 1, 2, and 9 of *Network Science* by A.-L. Barabási (see http://barabasi.com/book/network-science)

Complex Networks

Community Detection

Modularity

H3: Random Hypothesis

Randomly wired networks lack an inherent community structure

In a randomly wired network the connection pattern between the nodes is expected to be uniform, independent of the network's degree distribution

Consequently these networks are not expected to display systematic local density fluctuations that we could interpret as communities

Complex Networks
Community Detection

Modularity

Systematic deviations from a random configuration allow us to define a quantity called ***modularity***, a measure of the quality of each partition

Modularity allows us to decide if a particular community partition is better than some other one

Complex Networks

Community Detection

Modularity

Consider a network with

N nodes
L links
a partition into n_c communities

each community having N_c nodes
connected to each other by L_c links
where $c = 1, \ldots, n_c$

Complex Networks

Community Detection

Modularity

If L_c is larger than the expected number of links between the N_c nodes,

the nodes of the subgraph C_c could indeed be part of a true community

(as expected based on the Density Hypothesis H2)

Complex Networks

Community Detection

Modularity

Measure the difference between the network's real wiring diagram A_{ij}

and the expected number of links between i and j if the network is randomly wired p_{ij}

$$M_c = \frac{1}{2L} \sum_{(i,j) \in C_c} (A_{ij} - p_{ij})$$

p_{ij} can be determined by randomizing the original network
(while keeping the expected degree of each node unchanged)

$$p_{ij} = \frac{k_i k_j}{2L}$$

Complex Networks

Community Detection

Modularity

$$M_c = \frac{1}{2L} \sum_{(i,j) \in C_c} (A_{ij} - p_{ij})$$

If $M_c > 0$

then the subgraph C_c has more links than expected by chance
hence it represents a potential community

If $M_c = 0$

then the connectivity between the N_c nodes is random

If $M_c < 0$

then the nodes of C_c do not form a community

Complex Networks

Community Detection

Modularity

Simpler form of modularity

$$M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2$$

L_c is the total number of links within the community C_c

k_c is the total degree of the nodes in this community

Complex Networks

Community Detection

Modularity

Generalize these ideas to a full network …

Consider a partition that breaks the network into n_c communities

To see if the local link density of the subgraphs defined by this partition differs from the expected density in a randomly wired network, we define the partition's *modularity* by summing over all n_c communities:

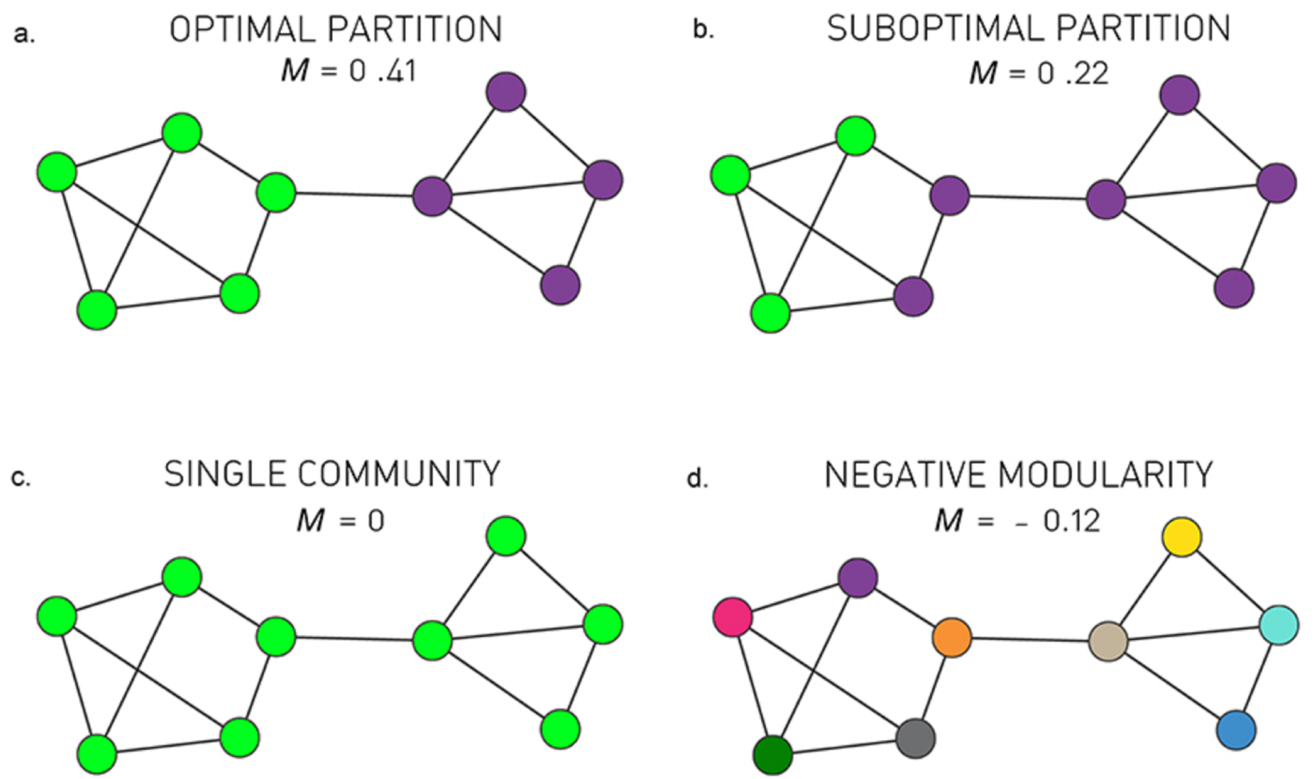
$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

Complex Networks

Community Detection

Modularity

Higher Modularity Implies Better Partition



Complex Networks

Community Detection

Modularity

H4: Maximal Modularity Hypothesis

For a given network, the partition with maximum modularity corresponds to the optimal community structure

Complex Networks

Community Detection

Modularity

Greedy Algorithm for Community Detection by Maximizing Modularity

The first modularity maximization algorithm, proposed by Newman

Iteratively joins pairs of communities if the move increases the partition's modularity

Complex Networks

Community Detection

Modularity

Greedy Algorithm for Community Detection by Maximizing Modularity

1. Assign each node to a community of its own, starting with N communities of single nodes

2. For each community pair connected by at least one link, compute the modularity difference ΔM obtained if we merge them.

   Merge the community pair for which ΔM is the largest

3. Repeat Step 2 until all nodes merge into a single community, recording M for each step

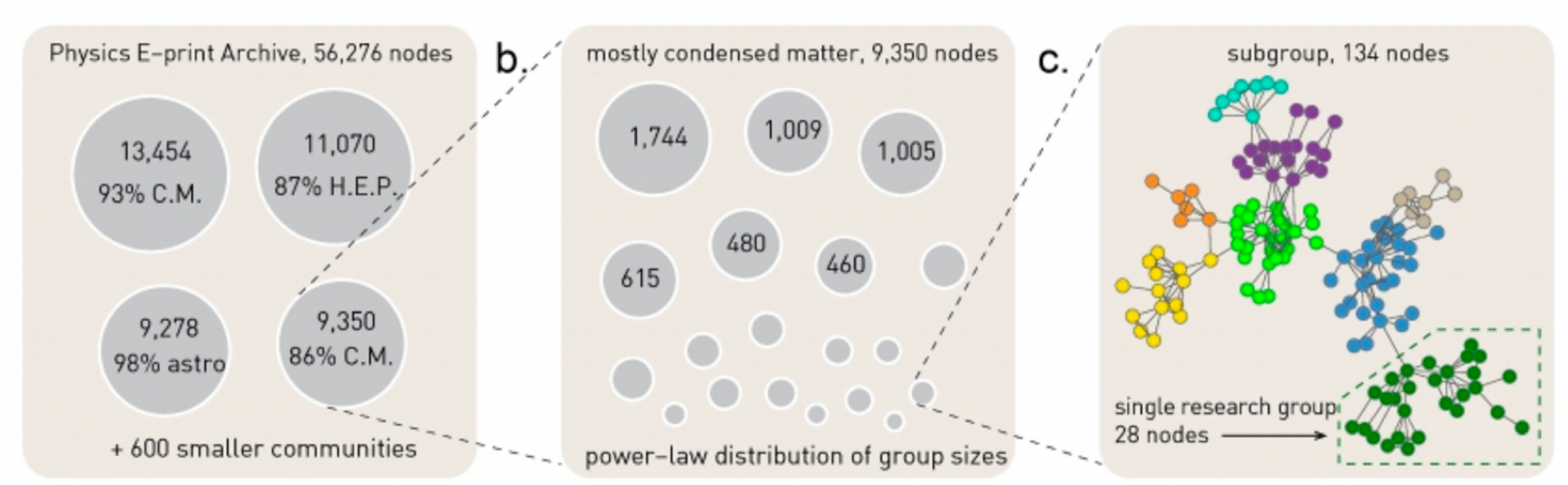
4. Select the partition for which M is maximal.

Complex Networks

Community Detection

Modularity

Greedy Algorithm for Community Detection by Maximizing Modularity



Greedy Algorithm applied network of physicists

Greedy Algorithm applied sub-network

Greedy Algorithm applied sub-sub-network

Complex Networks

Community Detection

Modularity

Limitations

- Resolution limit: modularity maximization cannot detect communities that are smaller than the resolution limit

$$k \leq \sqrt{2L}$$

k is the total degree of the community

For example, if L=1,497,134 modularity maximization will have difficulties resolving communities with total degree $k_C \lesssim$ 1,730

Real networks contain numerous small communities

Complex Networks
Community Detection

Modularity

Limitations

- Modularity maxima:

  All algorithms assume that a network with a clear community structure has an optimal partition with a maximal M

  In practice, however, there may be a large number of close to optimal partitions

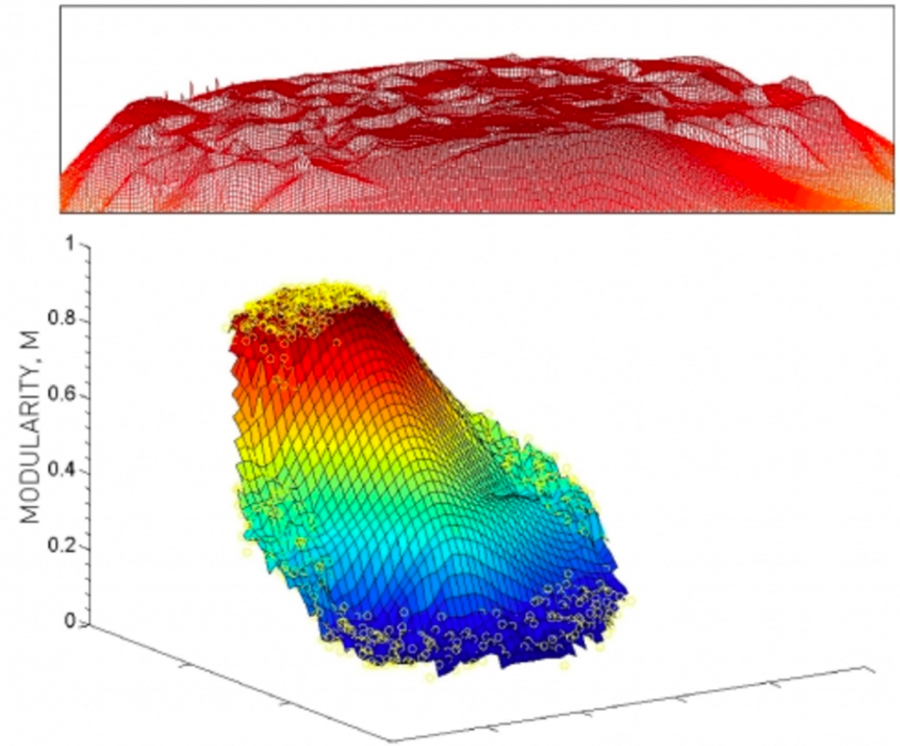
Complex Networks

Community Detection

Modularity

Limitations

- Modularity maxima:

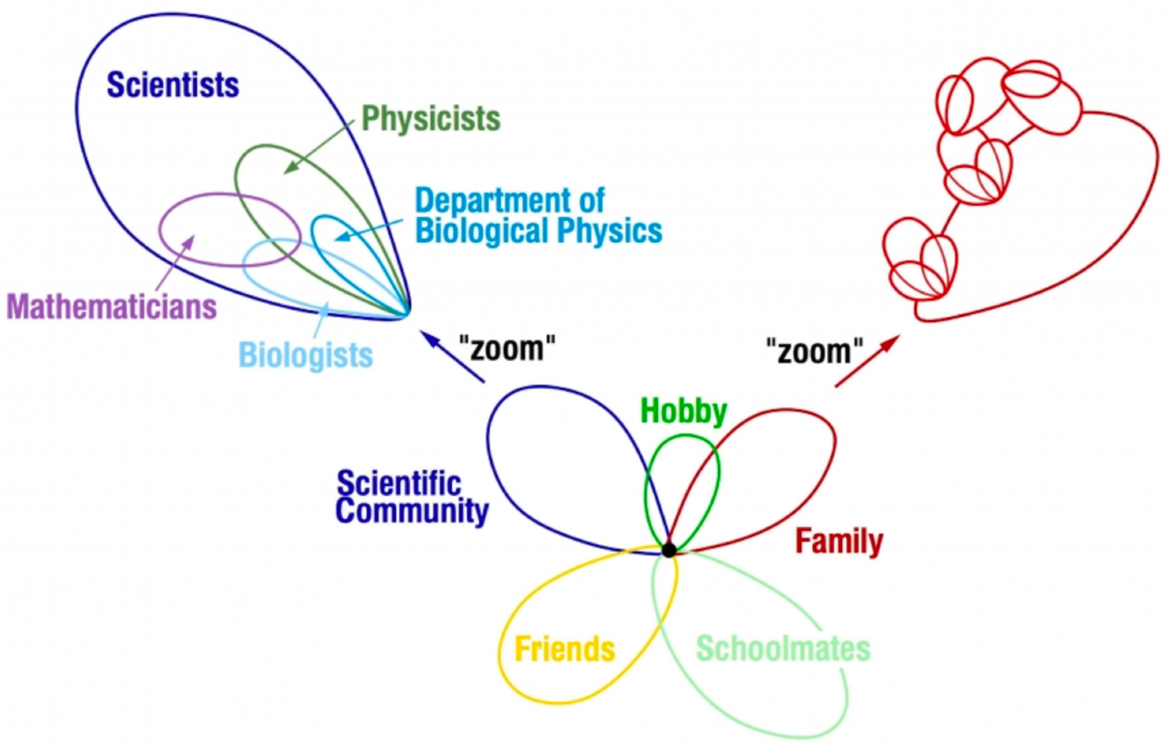


Complex Networks

Community Detection

Overlapping Communities

A node is rarely confined to a single community



Complex Networks

Community Detection

Overlapping Communities

A node is rarely confined to a single community

Clique Percolation Algorithm: CFinder (*)

- views a community as the union of overlapping cliques

(*) The CFinder software can be downloaded from www.cfinder.org

Complex Networks

Community Detection

CURRENT VERSION:
CFinder 2.0.6
Mar 21, 2014

Requires Java >>

Clusters & Communities

overlapping dense groups in networks

Download CFinder | Manual | Network Data | Publications | WebCFinder

Home

Download
- Software
- Network Data

Publications

Support
- FAQ
- Manual
- Contact us

People

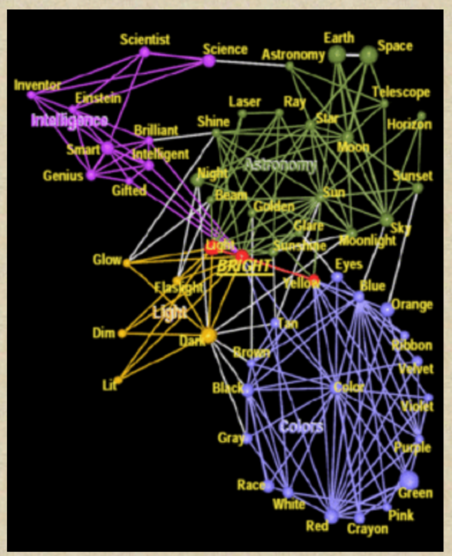
Links

CFinder is a free software for finding and visualizing overlapping dense groups of nodes in networks, based on the Clique Percolation Method (CPM) of Palla et. al., *Nature* **435**, 814-818 (2005). CFinder was recently applied to the quantitative description of the evolution of social groups: Palla et. al., *Nature* **446**, 664-667 (2007).

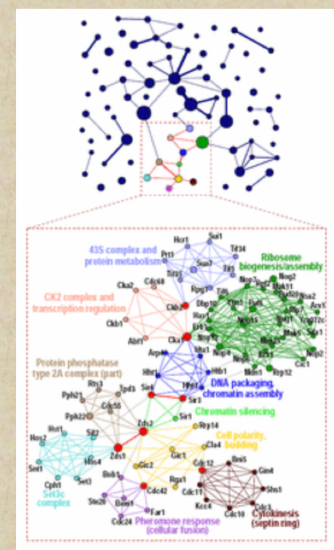
CFinder offers a fast and efficient method for clustering data represented by large graphs, such as genetic or social networks and microarray data. CFinder is also very efficient for locating the cliques of large sparse graphs.

Download: Software | Manual | Publications

A cluster -- also called a community or module -- in a network is a group of nodes more densely connected to each other than to nodes outside the group. In real networks clusters often overlap.

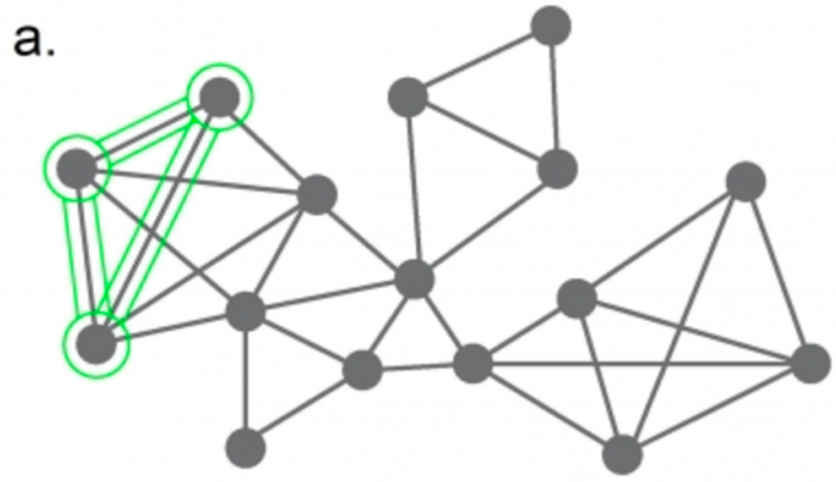


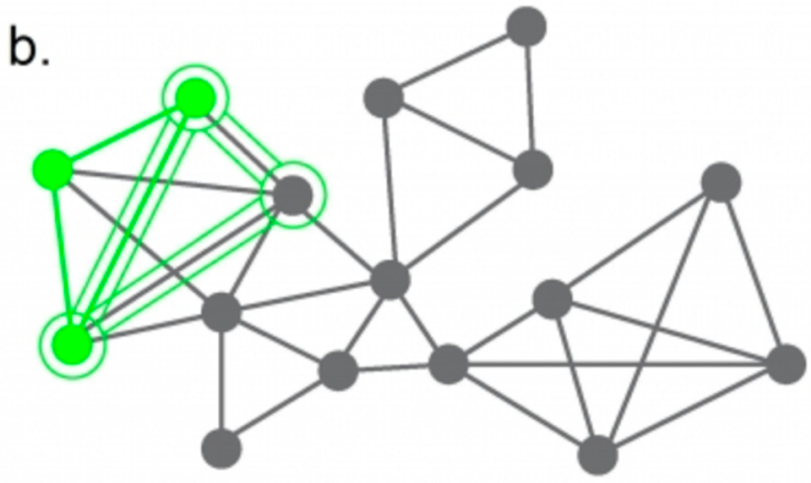
The **overlapping network modules** of the word "bright" in a word association network represent the different meanings of this word. From Palla et. al., *Nature* **435**, 814-818 (2005).

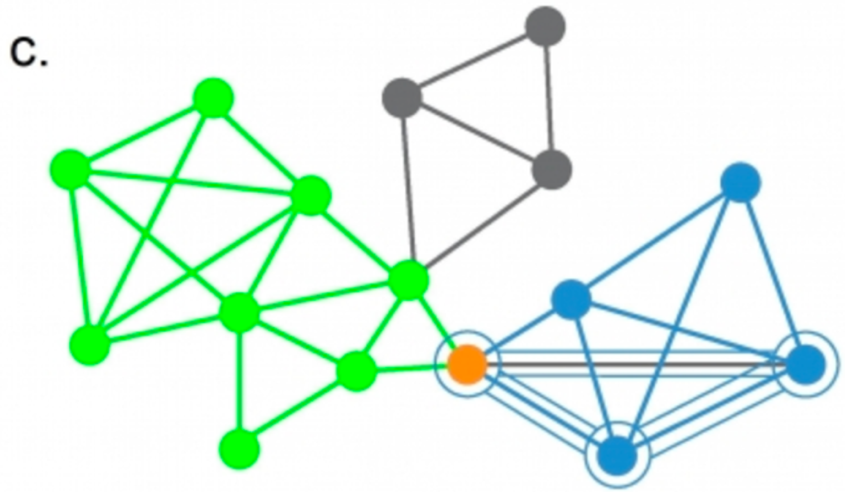


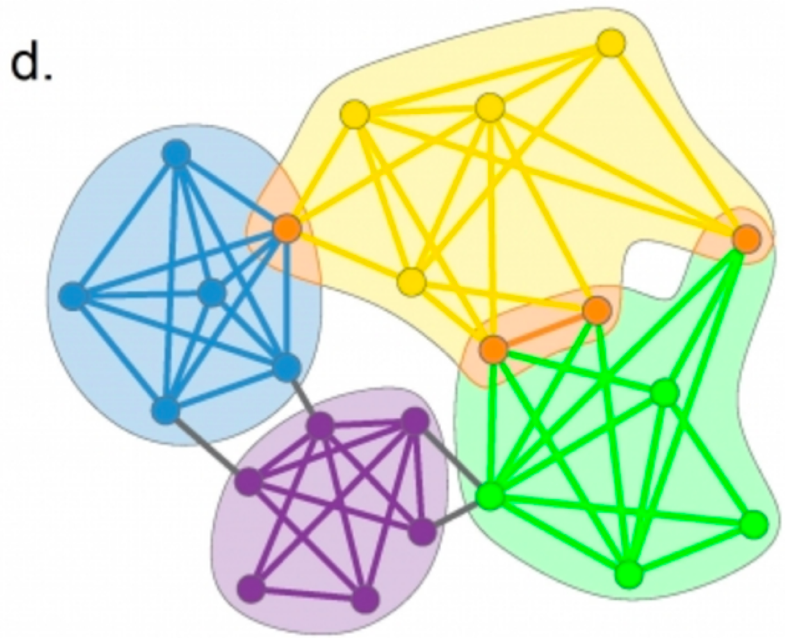
The **network of protein modules** in the protein-protein interaction network of yeast. Overlaps between the communities are shown in red. From Adamcsek et. al., *Bioinformatics* **22**, 1021 (2006).

Two k-cliques are considered adjacent if they share $k-1$ nodes









A k-clique community is the largest connected subgraph obtained by the union of all adjacent k-cliques

k-cliques that can not be reached from a particular k-clique belong to other k-clique communities

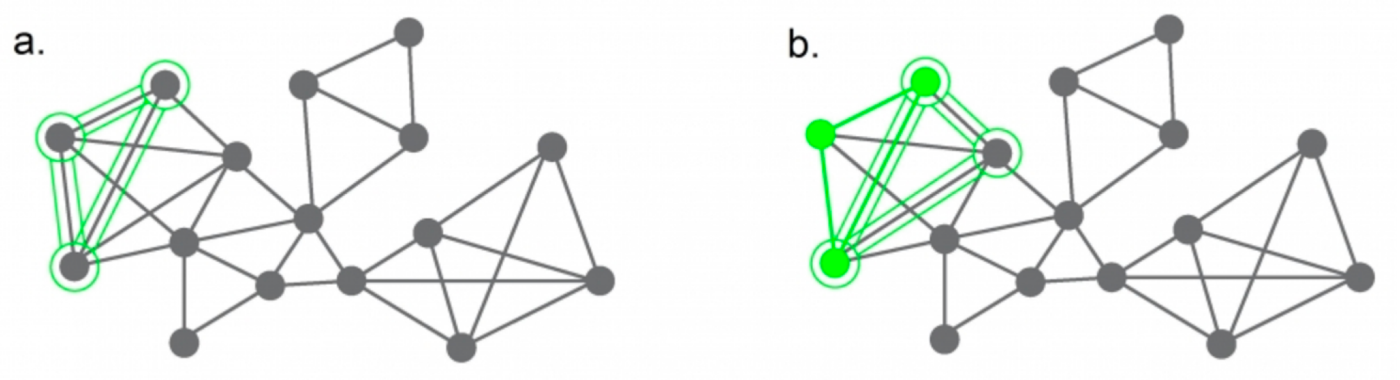
Complex Networks

Community Detection

Clique Percolation Algorithm (CFinder)

To identify k=3 clique-communities we roll a triangle across the network, such that each subsequent triangle shares one link (two nodes) with the previous triangle

(a)-(b) Rolling Cliques
Starting from the triangle shown in green in (a), (b) illustrates the second step of the algorithm.



Complex Networks

Community Detection

Clique Percolation Algorithm (CFinder)

(c) Clique Communities for k=3
The algorithm pauses when the final triangle of the green community is added.

As no more triangles share a link with the green triangles, the green community has been completed.

Note that there can be multiple *k*-clique communities in the same network (see second community in blue)

The figure highlights the moment when we add the last triangle of the blue community. The blue and green communities overlap, sharing the orange node.

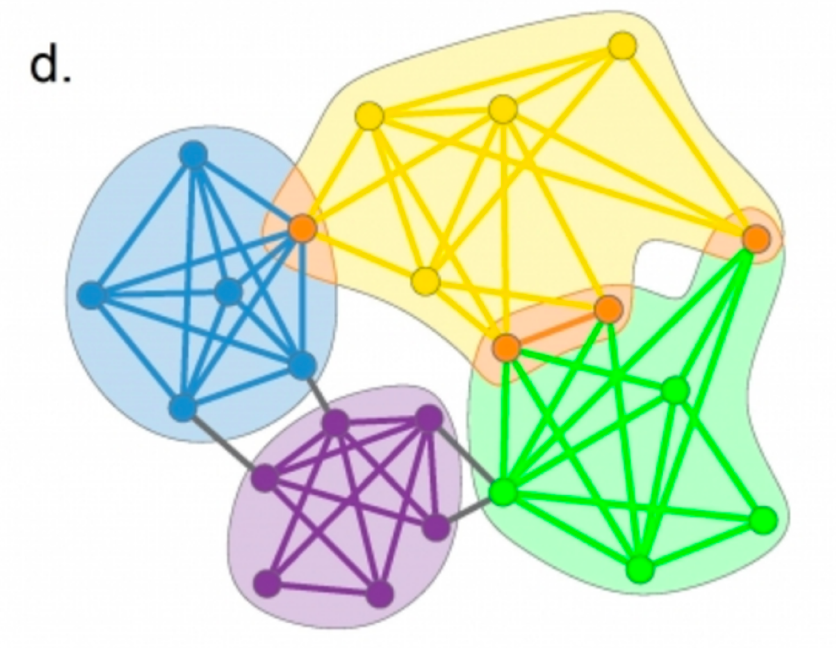
Complex Networks

Community Detection

Clique Percolation Algorithm (CFinder)

(d) Clique Communities for k=4

k=4 community structure of a small network, consisting of complete four node subgraphs that share at least three nodes. Orange nodes belong to multiple communities.



Complex Networks

Community Detection

Name	Nature	Comp.
Ravasz	Hierarchical Agglomerative	$O(N^2)$
Girvan-Newman	Hierarchical Divisive	$O(N^2)$
Greedy Modularity	Modularity Optimization	$O(N^2)$
Greedy Modularity (Optimized)	Modularity Optimization	$O(N\log^2 N)$
Louvain	Modularity Optimization	$O(L)$
Infomap	Flow Optimization	$O(N\log N)$
Clique Percolation (CFinder)	Overlapping Communities	$Exp(N)$
Link Clustering	Hierarchical Agglomerative; Overlapping Communities	$O(N^2)$

Complex Networks

Community Detection

At a Glance: Communities

Community identification rests on several hypotheses, pertaining to the nature of communities:

Fundamental Hypothesis
Communities are uniquely encoded in a network's wiring diagram. They represent a grand truth that remains to be discovered using appropriate algorithms.

Connectedness and Density Hypothesis
A community corresponds to a locally dense connected subgraph.

Random Hypothesis
Randomly wired networks do not have communities.

Maximal Modularity Hypothesis
The partition with the maximum modularity offers the best community structure, where modularity is given by

$$M = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

Community Finding: a Brief History