Uncovering the overlapping modular structure of complex networks

Tamás Vicsek
Dept. of Biological Physics, Eötvös University, Hungary

http://angel.elte.hu/~vicsek
http://cfinder.org
Why modules (densely interconnected parts)?

The internal organization of large networks is responsible for their function.

Complex systems/networks are typically *hierarchical*.

The units organize (become more closely connected) into groups which can themselves be regarded as units on a higher level.

We call these densely interconnected groups of nodes as modules/communities/coheseive groups/clusters etc. They are the “building blocks” of the complex networks on many scales.

For example:

Person→group→department→division→company→industrial sector

Letter→word→sentence→paragraph→section→chapter→book
Questions:

How can we recover the hierarchy of overlapping groups/modules/communities in the network if only a (very long) list of links between pairs of units is given?

What are their main characteristics?

Outline

• Basic facts and principles
• Community finding versus $k$-clique percolation
• Results for protein interaction, word association, phone calls, school friendship and collaboration networks
Basic observations:

A large complex network is bounded to be highly structured (has modules; function follows from structure)

The internal organization is typically hierarchical (i.e., displays some sort of self-similarity of the structure)

An important new aspect: Overlaps of modules are essential

“mess”, no function

Too constrained, limited function

Complexity is between randomness and regularity
Role of overlaps

Is this like a tree? (hierarchical methods)
Finding communities

Hierarchical methods

$4$-clique

Two nodes belong to the same community if they can be connected through adjacent $k$-cliques.

$k$-clique template rolling

Includes colleagues, friends, schoolmates, family members, etc.
Finding communities

Hierarchical methods

$k$-clique template rolling

Two nodes belong to the same community if they can be connected through adjacent $k$-cliques
Finding communities

Hierarchical methods

Two nodes belong to the same community if they can be connected through adjacent $k$-cliques
Finding communities

Hierarchical methods

Two nodes belong to the same community if they can be connected through adjacent $k$-cliques

$k$-clique template rolling

a 4-clique

Includes colleagues, friends, schoolmates, family members, etc.
Finding communities

Hierarchical methods

includes colleagues, friends, schoolmates, family members, etc.

Two nodes belong to the same community if they can be connected through adjacent k-cliques

k-clique template rolling

a 4-clique
Finding communities

Hierarchical methods

- all people
- includes colleagues, friends, schoolmates, family members, etc.

k-clique template rolling

- a 4-clique

Two nodes belong to the same community if they can be connected through adjacent $k$-cliques
Hierarchical versus template rolling clustering

Common clustering methods lead to a partitioning in which someone (a node) can belong to a single community at a time only.

For example, I can be located as a member of the community “physicists”, but not, at the same time, be found as a member of my community “family” or “friends”, etc.

$k$-clique template rolling allows large scale, systematic (deterministic) analysis of the network of overlapping communities
**$k$-CLIQUE PERCOLATION**

with I. Derényi and G. Palla

Definitions

*$k$-clique*: complete subgraph of $k$ vertices

*$k$-clique adjacency*: two $k$-cliques share a $k-1$ – clique

*$k$-clique walk*: series of steps to adjacent $k$-cliques

*$k$-clique cluster*: set of vertices of all $k$-clique walks from a given $k$-clique

(E-R percolation is the $k=2$ case)

The scaling of the relative size of the giant cluster of $k=3, 4$ and $5$-cliques at $p_c$.

**Order parameter for clique percolation, $k=4$**

$$p_c(k) = [(k-1)N]^{-1/(k-1)}$$

**Percolation threshold at**

For $k \leq 3$, $N_k^*/N_k(p_c) \sim N^{-k/6}$

For $k > 3$, $N_k^*/N_k(p_c) \sim N^{1-k/2}$

The scaling of the relative size of the giant cluster of $k=3, 4$ and $5$-cliques at $p_c$. 

![Graph showing the scaling of the giant cluster size](image)
Definitions
An order $k$ community is a $k$-clique percolation cluster

Such communities/clusters obviously can overlap
This is why a lot of new interesting questions can be posed

New fundamental quantities (cumulative distributions) defined:

- $P(d^{\text{com}})$: community degree distribution
- $P(m)$: membership number distribution
- $P(s^{\text{ov}})$: community overlap distribution
- $P(s)$: community size distribution (not new)
DATA

cond-mat (electronic preprints, about 30,000 authors)
protein-protein (DIP database, yeast, 2,600 nodes)
word association (sets of words associated with given words, questionnaire, 10,600 words)
mobile phone (~ 4,000,000 users calling each other)
school friendship (84 schools from USA)

large data sets: efficient algorithm is needed! Our method is the fastest known to us for these type of data

Steps:
- determine: cliques (not k-cliques!)
- clique overlap matrix
- components of the corresponding adjacency matrix

Do this for “optimal” k and w, where optimal corresponds to the “richest” (most widely distributed cluster sizes) community structure
Visualization of the communities of a node.

You can download the program and check your own communities.
“Web of networks”

Each node is a community

Nodes are weighted for community size
Links are weighted for overlap size

DIP “core” data base of protein interactions (S. cerevisiae, a yeast)

The other networks we analysed are much larger!!
Community size distribution

Community degree distribution

Combination of exponential and power law!

Emergence of a new feature as going “up” to the next level
Community overlap size     membership number
Case studies + dynamics

Protein interaction (prediction of function)

School friendship (disassortativity of communities, role of races)

Social group evolution in a co-authorship and a mobile phone network
network of yeast PPI modules

node: module of proteins, link: overlap of modules
enlarged portions of the network of modules

Marked:

- single proteins (function prediction) and groups (anticipated new modules)

- Zds1,2: chromatin silencing, cell polarity
- prot. phosph. type 2A complex + putative member (Rts3)
- Yck1: casein kinase, phosphorilation
- cAMP-dep. protein kinase complex and its regulator
- histone deacet. complex (part)
- function: negative reg. of meiosis
  (5 of 6 total in the genome included here)
- Rts3
- Zds2
- Sif2
- Kap60,95: protein carriers in the nucleus
- Kap60
- Zds1
- Cdc12
- Cdc42
- establishment of cell polarity (10 of 103 total in the genome) and Csm1 (DNA repl.)
- Cdc11
- Gln4
- Cdc10
- Cdc3
- Cdc12
- septin ring (part) and its assembly (part)
- predicted cellular sub-process
- established of cell polarity (6 of 103 total in the genome) and Far1 (cell-cycle arrest)
- common biological process of Rsr1 and Sec15: bipolar bud site selection
- vesicle-mediated transp. except: Snx41 (transport at Golgi), Eeb1 (function: unknown)
Three schools from the Add-Health school friendship data set

Grades 7-12
Network of school friendship communities

with M. Gonzalez, J. Kertész and H Herrmann

$k=3$ (looser)

Minorities tend to form more densely interconnected groups

$k=4$ (more dense)
Distribution functions (for $k=3$)

- $P(k)$ – degree distribution
- $C(k)$ – clustering coefficient
- $<k_n>(k)$ – degree of neighbour (individuals: assortative
  communities: diassortative)
Quantifying social group evolution

Small part of the phone call network (surrounding the circled yellow node up to the fourth neighbour)

Small part of the collaboration network (surrounding the circled green node up to the fourth neighbour)
Callers with the same zip code or age are over-represented in the communities we find.
Examples for tracking individual communities.
Lifetime ($\tau$) of a social group as a function of stability (steadiness, $\zeta$) and size ($s$)

Thus, a large group is around longer if it is less steady (and the opposite is true for small groups)
Probability of disintegrating ($p_d$) and the lifetime ($\tau^*$) of a community whose members have a total amount of “commitments” to other communities equal to $W_{\text{out}}$. 

\[ b) \]

![Graph showing the relationship between $P_d$ and $\tau^*$ over $W_{\text{out}} / (W_{\text{in}} + W_{\text{out}})$](image)

CFinder has been recently applied to quantifying the evolution of social groups: Palla et al. *Nature* **446**, 664 (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.

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. Examples for overlapping clusters obtained by CFinder in a protein-protein interaction network and a word association graph are reproduced here from Palla et al. (2005). Click on the images to view them enlarged in a separate window.
Social network of the 3000 employees of an European company determined from an on-line survey. Visualization of the betweenness centrality.
Visualization of the communities for the same company shown here using an adaptation of our CFinder-Firmnet software. Theridion provides organizational development services based on network analysis.
Outlook:

Networks of networks

- hierarchical aspects

- correlations, clustering, etc.,
  i.e., everything you can do for vertices

- applications, such as protein function prediction or organizational development
This will also become a commercial product by Firmlinks with GORDIO, a Budapest based HR company.
Internal organization of large complex networks in terms of their modular structure

- Research on modules/communities is a very active field (Amaral, Barabási, Newman)
- How does a large complex network may look like? + many further groups

**Random tree**

**Deterministic, loops**

**Random “blobs”**
To find overlapping communities we
consider: connected groups (clusters) of motifs e.g. a 4-clique

define: a cluster of adjacent complete subgraphs (cliques) is a community (simple assumption)

Two aspects
I) $\kappa$-clique percolation
II) communities in large real networks:
    overlaps and their statistics
Evolution of a single large community of collaborators

$s$ - size (number of authors), $t$ - time (in months)
Dedicated home page (software, papers, data)
http://angel.elte.hu/clustering/

Home

Screen shots
Basic observations:
A large complex network is bounded to be highly structured (has modules; function follows from structure)
The internal organization is typically hierarchical (i.e., displays some sort of self-similarity of the structure)
An important new aspect: Overlaps of modules are essential
Communities in a “tiny” part of a phone calls network of 4 million users (with A-L Barabási and G. Palla, *Nature, April 5 2007*)
Information about the age distribution of users in communities of size $s$ (Ratio of the standard deviation in a randomized set over actual)

Information about the Zip code (spatial) distribution of users in communities of size $s$ (Ratio of the standard deviation in a randomized set over actual)
The number of vertices in the largest component grows as $N$ increases, and the width of the quickly growing region decays as $1/N^{1/2}$. 

For example, in graph (a) with $k=4$, we observe this behavior for different values of $N$: 100, 200, 500, 1000, 2000, and 5000. The transition point is marked at $p/p_c(k) = 1$, where $p_c(k)$ is the critical probability for the largest component to appear.

In graph (b), we see similar behavior for different values of $k$: 3, 4, and 5, with $\alpha = 0.45$ and $\alpha = 0.5$. The curves indicate how the fraction of vertices in the largest component, $\Phi$, changes with the rescaled parameter $[p/p_c(k) - 1] N^\alpha$.
Evolution of the social network of scientific collaborations

The Erdős graph and the Erdős number
(Ei=2,W=8,BG=4)

Data: collaboration graphs in (M) Mathematics and (NS) Neuroscience
Collaboration network

Cumulative data, 1991 - 98

Degree distribution:

power-law with

\[ \gamma_M = 2.1, \quad \gamma_{NS} = 2.4 \]

due to growth and preferential attachment
Internal preferential attachment:

$$\kappa(k_1, k_2) = \int_1^{k_1} \Pi(k_1, k_2) d(k_1, k_2)$$

Cumulative attachment rate:

Measured data shows:

- $$\kappa(k_1, k_2)$$ is quadratic in $$k_1 k_2$$
- $$\Pi(k_1, k_2)$$ is linear in $$k_1 k_2$$

Attachment rate

Communities of collaborators are formed
The scaling of the relative size of the giant cluster of $k$-cliques at $p_c$ for $k \leq 3$.

For $k \leq 3$, $\frac{N_k^*}{N_k(p_c)} \sim N^{-k/6}$

For $k > 3$, $\frac{N_k^*}{N_k(p_c)} \sim N^{1-k/2}$
Distribution of community sizes

Over-representation of the usage of a given service as a function of the number of users in a community
Community dynamics

with P. Pollner and G. Palla

Dynamics of community growth: the preferential attachment principle applies on the level of communities as well

The probability that a previously unlinked community joins a community larger than \( s \) grows approximately linearly (for the cond-mat coauthorship network)