k-clique percolation and clustering in directed and weighted networks

Gergely Palla[1] Dániel Ábel[2] Imre Derényi[2]
Illés Farkas[1] Péter Pollner[1] Tamás Vicsek[1,2]

[1]Statistical and Biological Physics Research Group,
Hungarian Acedmy of Sciences,

[2]Department of Biological Physics,
Eötvös University, Hungary

March 2008, Louvain-la-Neuve

Outline

- Introduction
  - The Clique Percolation Method (CPM)
  - Phase transition in the Erdős-Rényi graph
- Directed communities
  - Relative in- and out degree
  - Directed CPM
  - Results
- Weighted communities
  - Weights in the original CPM
  - Weighted CPM
  - Results

The Clique Percolation Method (CPM)

Definitions

- *k*-clique: a complete (fully connected) subgraph of k vertices.
- *k*-clique adjacency: two k-cliques are adjacent if they share $k-1$ vertices, *i.e.*, if they differ only in a single node.

The Clique Percolation Method (CPM)

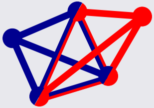
Definitions

- *k*-clique: a complete (fully connected) subgraph of k vertices.
- *k*-clique adjacency: two k-cliques are adjacent if they share $k-1$ vertices, *i.e.*, if they differ only in a single node.

The Clique Percolation Method (CPM)

Definitions

- *k*-clique: a complete (fully connected) subgraph of k vertices.
- *k*-clique adjacency: two k-cliques are adjacent if they share $k-1$ vertices, *i.e.*, if they differ only in a single node.



CPM (continued)

Definition

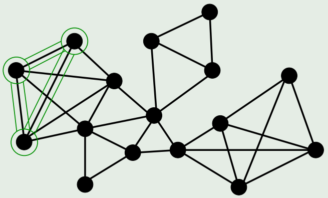
- k-clique community: the union of k-cliques that can be reached from one to the other through a sequence of adjacent k-cliques.

CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

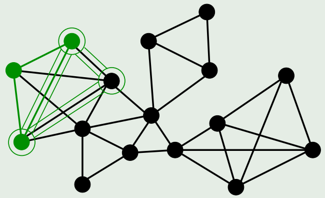


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

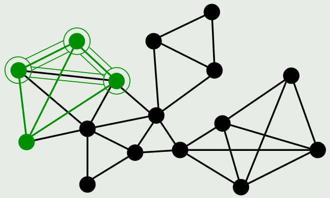


CPM
(continued)

Definition

- k-clique community: the union of k-cliques that can be reached from one to the other through a sequence of adjacent k-cliques.

Illustration:

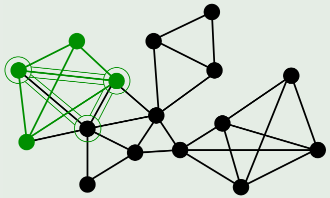


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

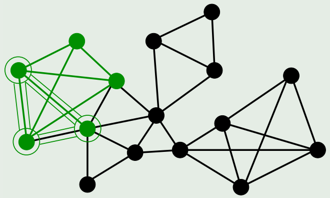


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

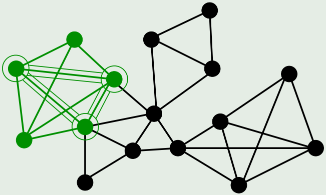


CPM
(continued)

Definition

- *k*-clique community: the union of k-cliques that can be reached from one to the other through a sequence of adjacent k-cliques.

Illustration:

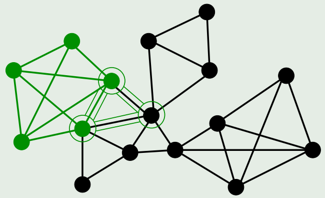


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

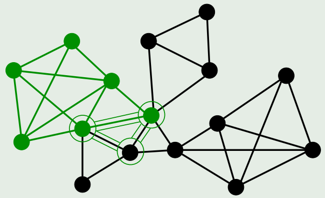


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

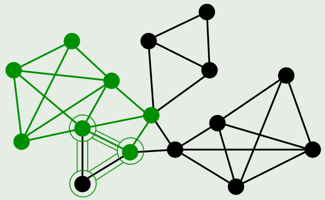


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

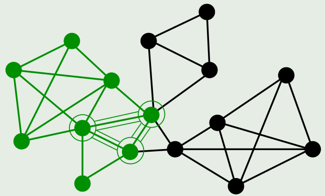


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

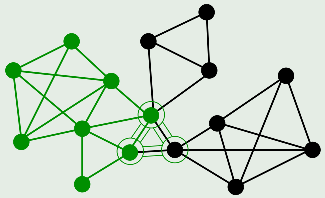


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

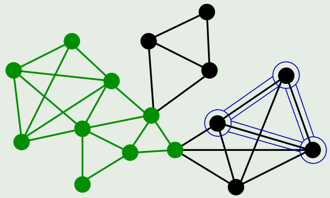


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

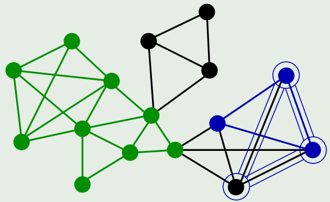


CPM (continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

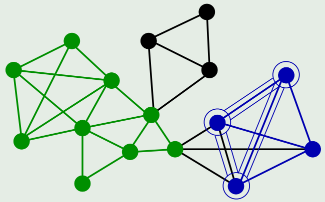


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

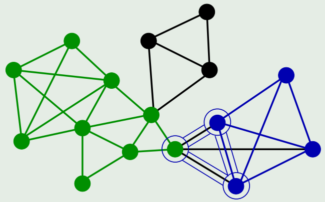


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

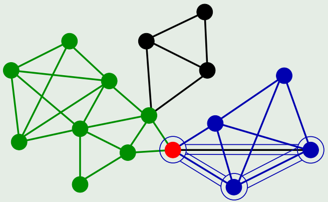


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

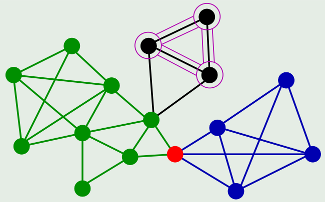


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

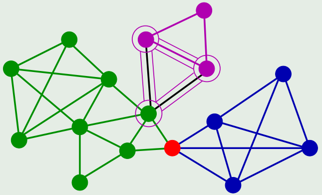


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

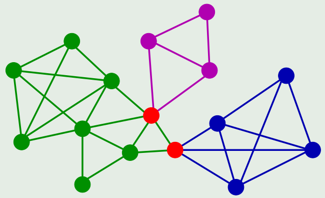


CPM
(continued)

Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:

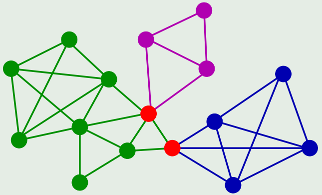


CPM
(continued)

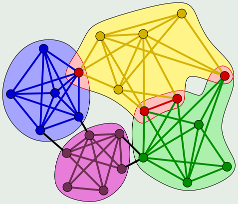
Definition

- *k*-clique community: the union of *k*-cliques that can be reached from one to the other through a sequence of adjacent *k*-cliques.

Illustration:



same at $k = 4$:



Advantages of the CPM

The main advantages of the CPM:

- Allows overlaps between the communities.
- The definition is based on the density of the links.
- It is local. (No resolution limit).

The order parameters

The Erdős-Rényi graph:

- N nodes,
- every pair is independently linked with probability p.

A giant k-clique percolation cluster can be found if $p \geq p_c(k)$.

The order parameter of the phase transition is the size of the giant cluster:

The number of nodes, N^* $\longrightarrow$ $\Phi \equiv N^*/N$,
The number of k-cliques, $\mathcal{N}^*$ $\longrightarrow$ $\Psi \equiv \mathcal{N}^*/\mathcal{N}$.

The order parameters

The Erdős-Rényi graph:

- N nodes,
- every pair is independently linked with probability p.

A giant k-clique percolation cluster can be found if $p \geq p_c(k)$.

The order parameter of the phase transition is the size of the giant cluster:

$$\begin{array}{lcl} \text{The number of nodes, } N^* & \longrightarrow & \Phi \equiv N^*/N, \\ \text{The number of } k\text{-cliques, } \mathcal{N}^* & \longrightarrow & \Psi \equiv \mathcal{N}^*/\mathcal{N}. \end{array}$$

The order parameters

The Erdős-Rényi graph:

- N nodes,
- every pair is independently linked with probability p.

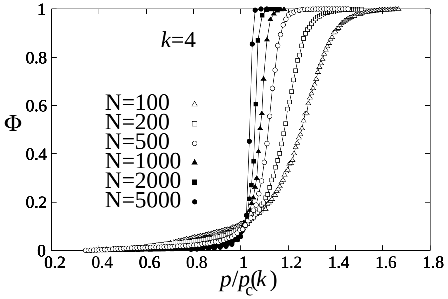
A giant k-clique percolation cluster can be found if $p \geq p_c(k)$.

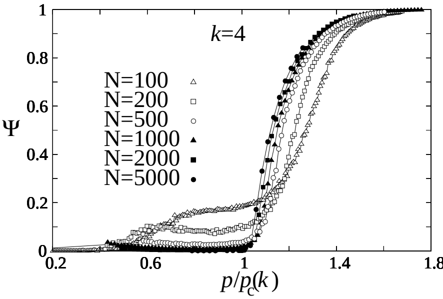
The order parameter of the phase transition is the size of the giant cluster:

$$\begin{array}{lcl} \text{The number of nodes, } N^* & \longrightarrow & \Phi \equiv N^*/N, \\ \text{The number of } k\text{-cliques, } \mathcal{N}^* & \longrightarrow & \Psi \equiv \mathcal{N}^*/\mathcal{N}. \end{array}$$

Results

Numerical results:





$$p_c(k) = \frac{1}{[N(k-1)]^{\frac{1}{k-1}}}.$$

Directed links

Direction of the links:

- Direction of some kind of flow (e.g. information, energy).
- Asymmetrical relation (e.g. superior-inferior).

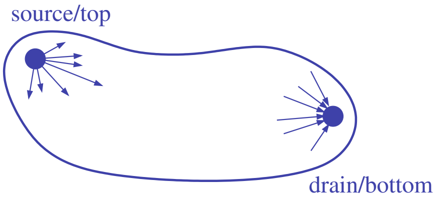
Out-hubs in communities represent “sources”, whereas in-hubs correspond to “drains”:

Directed links

Direction of the links:

- Direction of some kind of flow (e.g. information, energy).
- Asymmetrical relation (e.g. superior-inferior).

Out-hubs in communities represent "sources", whereas in-hubs correspond to "drains":

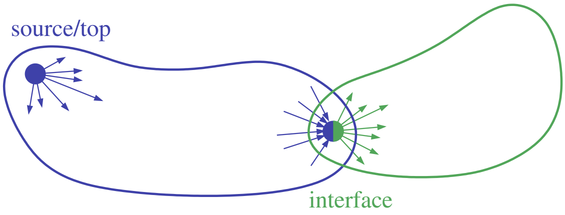


Directed links

Direction of the links:

- Direction of some kind of flow (e.g. information, energy).
- Asymmetrical relation (e.g. superior-inferior).

Out-hubs in communities represent “sources”, whereas in-hubs correspond to “drains”:



Relative in- and out-degree

We define the relative in-degree and relative out-degree of node i in community α as

$$D_{i,\mathrm{in}}^{\alpha} \equiv \frac{d_{i,\mathrm{in}}^{\alpha}}{d_{i,\mathrm{in}}^{\alpha}+d_{i,\mathrm{out}}^{\alpha}},$$

$$D_{i,\mathrm{out}}^{\alpha} \equiv \frac{d_{i,\mathrm{out}}^{\alpha}}{d_{i,\mathrm{in}}^{\alpha}+d_{i,\mathrm{out}}^{\alpha}},$$

For weighted networks these can be replaced by the relative in-strength and relative out-strength:

$$W_{i,\mathrm{in}}^{\alpha} \equiv \frac{w_{i,\mathrm{in}}^{\alpha}}{w_{i,\mathrm{in}}^{\alpha}+w_{i,\mathrm{out}}^{\alpha}},$$

$$W_{i,\mathrm{out}}^{\alpha} \equiv \frac{w_{i,\mathrm{out}}^{\alpha}}{w_{i,\mathrm{in}}^{\alpha}+w_{i,\mathrm{out}}^{\alpha}},$$

Relative in- and out-degree

We define the relative in-degree and relative out-degree of node i in community α as

$$D^{\alpha}_{i,\mathrm{in}} \equiv \frac{d^{\alpha}_{i,\mathrm{in}}}{d^{\alpha}_{i,\mathrm{in}} + d^{\alpha}_{i,\mathrm{out}}},$$

$$D^{\alpha}_{i,\mathrm{out}} \equiv \frac{d^{\alpha}_{i,\mathrm{out}}}{d^{\alpha}_{i,\mathrm{in}} + d^{\alpha}_{i,\mathrm{out}}},$$

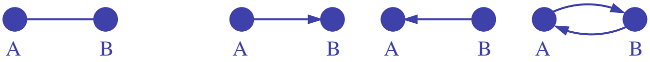
For weighted networks these can be replaced by the relative in-strength and relative out-strength:

$$W^{\alpha}_{i,\mathrm{in}} \equiv \frac{w^{\alpha}_{i,\mathrm{in}}}{w^{\alpha}_{i,\mathrm{in}} + w^{\alpha}_{i,\mathrm{out}}},$$

$$W^{\alpha}_{i,\mathrm{out}} \equiv \frac{w^{\alpha}_{i,\mathrm{out}}}{w^{\alpha}_{i,\mathrm{in}} + w^{\alpha}_{i,\mathrm{out}}},$$

Directed k-cliques?

Comparing undirected and directed connections:

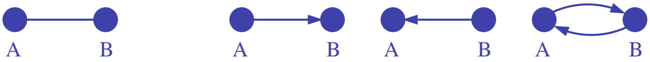


In case of k-cliques:

- $k(k-1)/2$ links $\longrightarrow 3^{k(k-1)/2}$ possible configurations.
- However, we would like the k-clique to have some kind of directionality as a whole as well.

Directed k-cliques?

Comparing undirected and directed connections:



In case of k-cliques:

- $k(k-1)/2$ links $\longrightarrow 3^{k(k-1)/2}$ possible configurations.
- However, we would like the k-clique to have some kind of directionality as a whole as well.

Definition

A directed k-clique has to fulfil the following conditions:

In the absence of double links:

- Any directed link in the k-clique points from a node with a higher order (larger restricted out-degree) to a node with a lower order.
- The k-clique contains no directed loops.
- The restricted out-degree of each node in the k-clique is different.

If double links are present:

It is possible to eliminate the double links in such a way that the single links fulfil the above conditions.

Definition

A directed k-clique has to fulfil the following conditions:

In the absence of double links:

- Any directed link in the k-clique points from a node with a higher order (larger restricted out-degree) to a node with a lower order.
- The k-clique contains no directed loops.
- The restricted out-degree of each node in the k-clique is different.

If double links are present:

It is possible to eliminate the double links in such a way that the single links fulfil the above conditions.

Definition

A directed k-clique has to fulfil the following conditions:

In the absence of double links:

- Any directed link in the k-clique points from a node with a higher order (larger restricted out-degree) to a node with a lower order.
- The k-clique contains no directed loops.
- The restricted out-degree of each node in the k-clique is different.

If double links are present:

It is possible to eliminate the double links in such a way that the single links fulfil the above conditions.

Definition

A directed k-clique has to fulfil the following conditions:

In the absence of double links:

- Any directed link in the k-clique points from a node with a higher order (larger restricted out-degree) to a node with a lower order.
- The k-clique contains no directed loops.
- The restricted out-degree of each node in the k-clique is different.

If double links are present:

It is possible to eliminate the double links in such a way that the single links fulfil the above conditions.

Definition

A directed k-clique has to fulfil the following conditions:

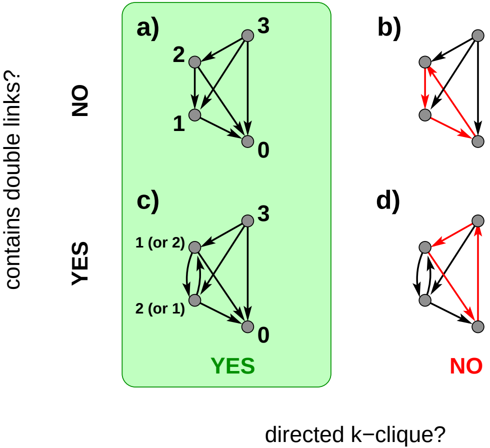
In the absence of double links:

- Any directed link in the k-clique points from a node with a higher order (larger restricted out-degree) to a node with a lower order.
- The k-clique contains no directed loops.
- The restricted out-degree of each node in the k-clique is different.

If double links are present:

It is possible to eliminate the double links in such a way that the single links fulfil the above conditions.

Illustration



Phase transition in the directed E-R graph

The directed E-R graph:

- N nodes,
- The $N(N-1)$ possible “places” for the directed links are filled independently with probability p.

Theoretical prediction of the critical point for the appearance of a giant directed k-clique percolation cluster:

$$p_c^{\mathrm{theor}} = \frac{1}{[Nk(k-1)]^{\frac{1}{k-1}}}.$$

Order parameters: Φ, Ψ (same as in the undirected case).

Phase transition in the directed E-R graph

The directed E-R graph:

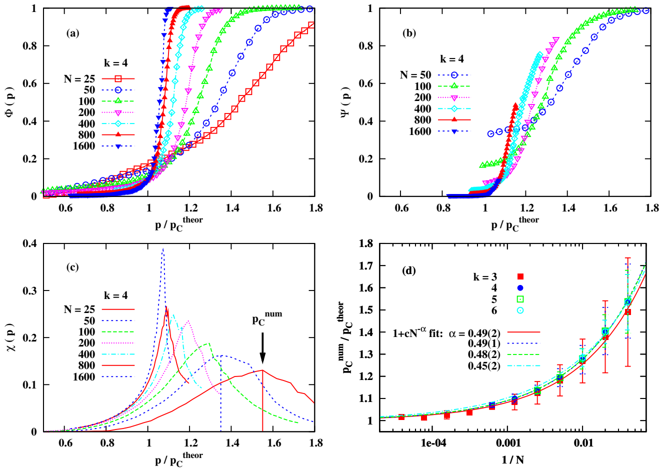
- N nodes,
- The $N(N-1)$ possible “places” for the directed links are filled independently with probability p.

Theoretical prediction of the critical point for the appearance of a giant directed k-clique percolation cluster:

$$p_c^{\mathrm{theor}} = \frac{1}{[Nk(k-1)]^{\frac{1}{k-1}}}.$$

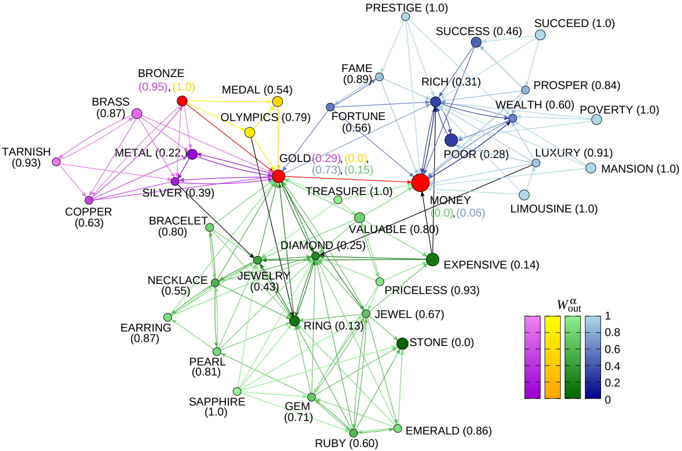
Order parameters: Φ, Ψ (same as in the undirected case).

Numerical results



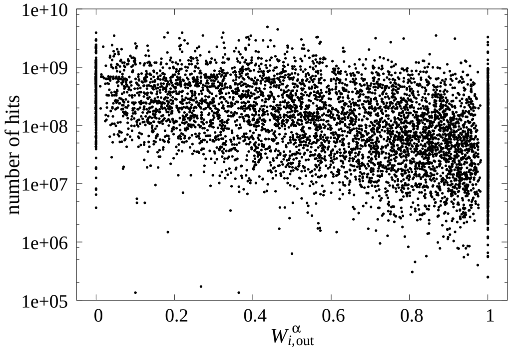
Word association network

Local picture of the communities:



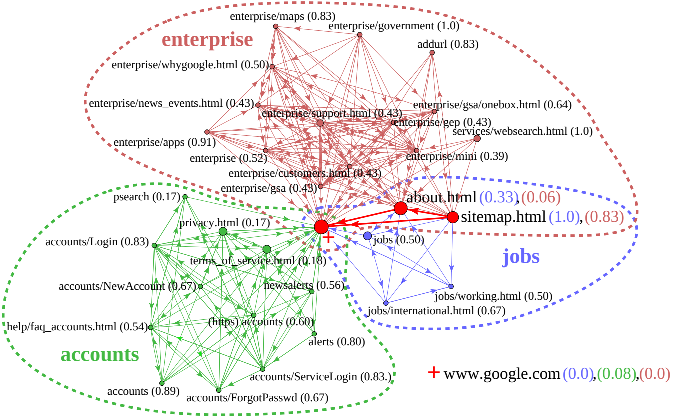
Relative out-degree and number of hits

The number of hits in Google as a function of $W_{i,\mathrm{out}}^{\alpha}$:



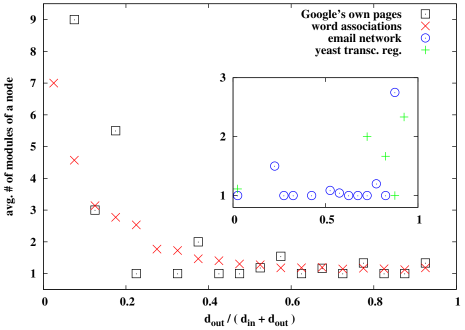
Google's on web pages

Local picture of the communities:



Comparing overlaps

Membership number in function of $D^{\alpha}_{i,\mathrm{out}}$:



Community overlaps:

- word association net, Google's web pages $\longrightarrow$ in-hubs,
- e-mail net, transcription regulatory network $\longrightarrow$ out-hubs.

Link weights in the original CPM

In the original CPM we can take into account the weights by ignoring links weaker than a certain threshold w^*.

Changing w^* and k is similar to changing the resolution in a microscope.

Optimal k-clique size and w^*

Where the community structure is as highly structured as possible: just below the critical point of the appearance of a giant k-clique community.

Link weights in the original CPM

In the original CPM we can take into account the weights by ignoring links weaker than a certain threshold w^*.

Changing w^* and k is similar to changing the resolution in a microscope.

Optimal k-clique size and w^*

Where the community structure is as highly structured as possible: just below the critical point of the appearance of a giant k-clique community.

Link weights in the original CPM

In the original CPM we can take into account the weights by ignoring links weaker than a certain threshold w^*.

Changing w^* and k is similar to changing the resolution in a microscope.

Optimal k-clique size and w^*

Where the community structure is as highly structured as possible: just below the critical point of the appearance of a giant k-clique community.

k-clique intensity

The intensity I of a sub-graph is defined as the geometrical mean of its link weights.

$$\text{For a } k\text{-clique } \mathcal{C}\text{: } I(\mathcal{C}) = \left(\prod_{\substack{i<j \\ i,j\in\mathcal{C}}} w_{ij} \right)^{2/k/(k-1)}$$

Weighted k-clique

A k-clique with an intensity greater or equal to a given intensity threshold I^*.

Percolation transition in the E-R graph

A weighted E-R graph:

- N nodes,
- every pair is linked independently with uniform probability p,
- each link is assigned a weight chosen randomly from a uniform distribution on the $(0, 1]$ interval.

The critical linking probability is a function of the intensity threshold. At $I = 0$ we recover $p_c(I = 0) = [N(k-1)]^{-1/(k-1)}$.

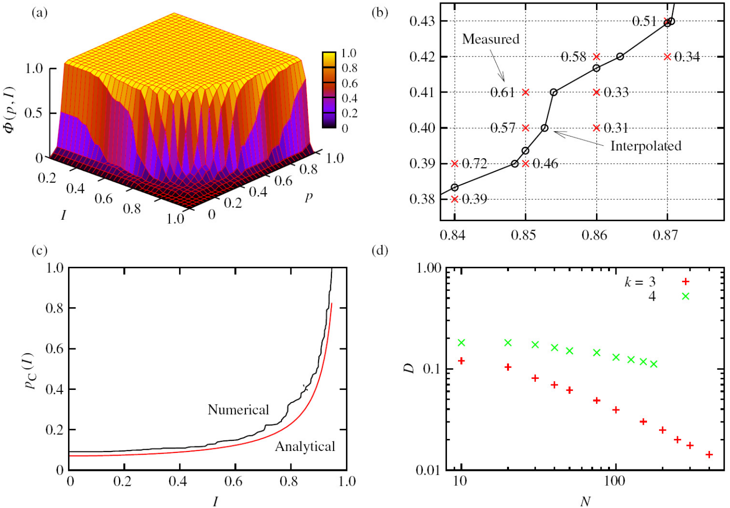
Percolation transition in the E-R graph

A weighted E-R graph:

- N nodes,
- every pair is linked independently with uniform probability p,
- each link is assigned a weight chosen randomly from a uniform distribution on the $(0, 1]$ interval.

The critical linking probability is a function of the intensity threshold. At $I = 0$ we recover $p_c(I = 0) = [N(k-1)]^{-1/(k-1)}$.

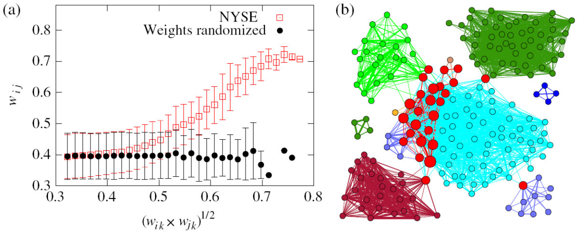
Results


(a)
Φ(p, I)
1.0
0.5
0
0.2
0.4
0.6
0.8
1.0
I
0
0.2
0.4
0.6
0.8
1.0
p
1.0
0.8
0.6
0.4
0.2
0
(b)
0.43
0.42
0.41
0.40
0.39
0.38
0.84
0.85
0.86
0.87
Measured
Interpolated
0.51
0.58
0.34
0.61
0.33
0.57
0.31
0.72
0.46
0.39
(c)
$p_C(I)$
1.0
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
I
Numerical
Analytical
(d)
D
1.00
0.1
0.01
10
100
N
k = 3
4

NYSE graph

New York Stock Exchange graph:

- We studied the pre-computed stock correlation matrix containing the averaged correlation between the daily logarithmic returns.
- The correlation coefficients can be used as link weights. We kept only the strongest 3%.



Summary

- Directed communities:
  - Relative in- and out-degree,
  - Directed *k*-cliques.
- Weighted communities:
  - *k*-clique intensity.
- Publications:
  - New Journal of Physics **9**, 180 (2007),
  - New Journal of Physics **9**, 186 (2007).
- Downloadable community finding software:
  - `http://cfinder.org`