On network clustering by modularity maximization with cohesion conditions

Sonia Cafieri

Laboratoire MAIAA

ENAC - École Nationale de l'Aviation Civile

University of Toulouse

France

Workshop on Clustering and Search techniques in large scale networks
Nizhny Novgorod, November 2014

Networks

Networks often used to represent complex systems

Mathematical representation: Graph $G = (V, E)$

V = *Vertices*, associated with the entities of the system under study

E = *Edges*, express that a relation defined on all pairs of vertices holds or not for each such pair

- social networks
- telecommunication networks
- transportation networks
- ...

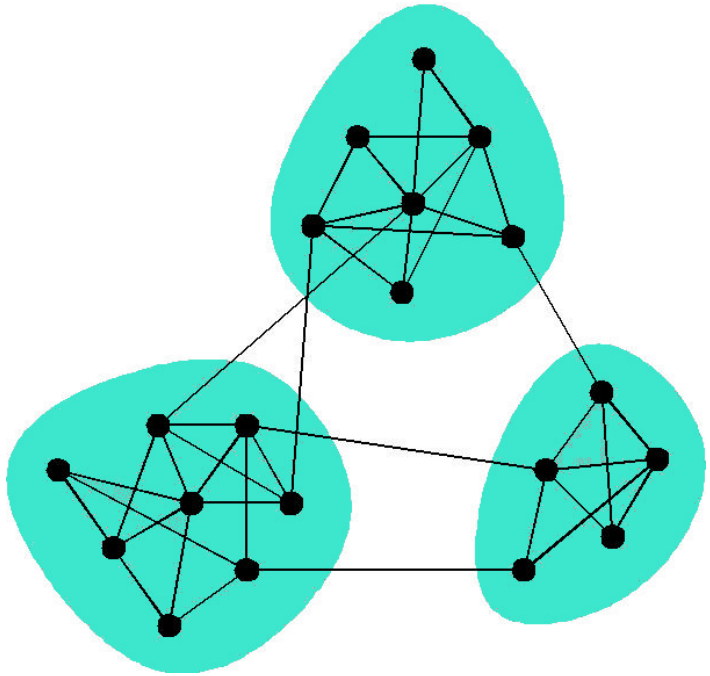
Network Clustering

Automatic analysis of complex systems represented as networks

$\Downarrow$

identification of communities

community (cluster) $\approx$ a subset of vertices that are more *densely* connected within the community while edges joining it to the outside are *sparse*



$\Rightarrow$ finding a partition of V into subgraphs induced by nonempty subsets

Outline

thanks to:

Alberto Costa (Singapore University of Technology and Design)

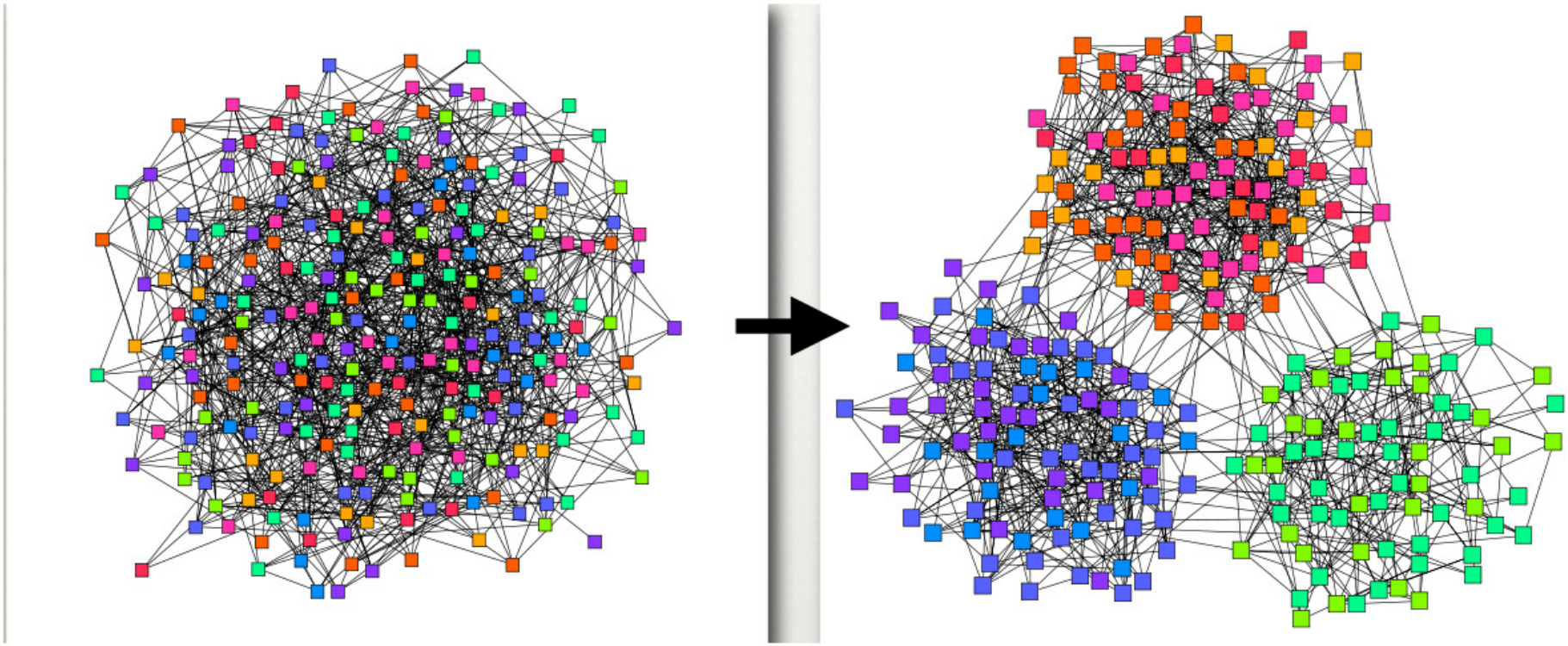
Pierre Hansen (GERAD, HEC Montréal, Canada)

ENAC

Outline

Clustering: finding communities

How to find and evaluate a partition?



We need

- a clustering criterion / definition of community
- a clustering algorithm

Evaluating a partition

(i) Use a heuristic

(ii) Choose a quality function, to be maximized or minimized

(iii) Specify conditions to be satisfied by a community

Evaluating a partition

(i) Use a heuristic

Example: edge removal heuristic (Girvan & Newman, 2002):

edges with maximum betweeness are iteratively removed, yielding partitions into an increasing number of communities.

The quality of the obtained results can only be judged a posteriori.

(ii) Choose a quality function, to be maximized or minimized

(iii) Specify conditions to be satisfied by a community

Evaluating a partition

(i) Use a heuristic

Example: edge removal heuristic (Girvan & Newman, 2002):

edges with maximum betweeness are iteratively removed, yielding partitions into an increasing number of communities.

The quality of the obtained results can only be judged a posteriori.

(ii) Choose a quality function, to be maximized or minimized

Example: Modularity (Newman & Girvan, 2004)

(iii) Specify conditions to be satisfied by a community

ENAC

Evaluating a partition

(i) Use a heuristic

Example: edge removal heuristic (Girvan & Newman, 2002):

edges with maximum betweeness are iteratively removed, yielding partitions into an increasing number of communities.

The quality of the obtained results can only be judged a posteriori.

(ii) Choose a quality function, to be maximized or minimized

Example: Modularity (Newman & Girvan, 2004)

(iii) Specify conditions to be satisfied by a community

Example: Strong and Weak conditions (Radicchi et al., 2004)

Semi-Strong and Extra-Weak conditions (Hu et al., 2008)

Almost-Strong condition (Cafieri et al., 2012)

Evaluating a partition

What is the *best* criterion to evaluate a partition of a network? – open question!

Idea: combine different criteria

- study to what extent optimal partitions for modularity maximization satisfy the cohesion conditions
- examine the effect of imposing these conditions, one at a time, as constraints in an optimization model for modularity maximization

Outline

Optimizing a quality function: Modularity

Newman and Girvan, 2004:

compare the fraction of edges falling within communities to the expected fraction of such edges

Modularity:

$$Q = \sum_s [a_s - e_s]$$

a_s = fraction of all edges that lie within module s

e_s = expected value of the same quantity in a graph in which the vertices have the same degrees but edges are placed at random.

Optimizing a quality function: Modularity

Newman and Girvan, 2004:

compare the fraction of edges falling within communities to the expected fraction of such edges

Modularity:

$$Q = \sum_s [a_s - e_s]$$

a_s = fraction of all edges that lie within module s

e_s = expected value of the same quantity in a graph in which the vertices have the same degrees but edges are placed at random.

- $Q \approx 0$: the network is equivalent to a random network (barring fluctuations);
- $Q \approx 1$: the network has a strong community structure;
- in practice, the maximum modularity Q is often between 0.3 and 0.7.

Maximizing modularity gives an optimal partition with the optimal number of clusters

Modularity maximization methods

- Exact algorithms for modularity maximization
  - proposed only in a few papers
  - can only solve small instances (with a few hundred entities) in reasonable time
  - provide an optimal solution together with the proof of its optimality
- Heuristics for modularity maximization
  - widely used
  - can solve approximately very large instances with up to thousand entities
  - do not have either an a priori performance guarantee
    (finding always a solution with a value which is at least a given percentage of the optimal one),
    nor an a posteriori performance guarantee
    (that the obtained solution is at least a computable percentage of the optimal one)

ENAC

Outline

Cohesion conditions

a priori conditions to have a community

- Strong condition
- Almost-strong condition
- Semi-strong condition
- Weak condition
- Extra-weak condition

$G = (V, E)$ graph, $A = (A_{ij})$ adjacency matrix

k_i = degree of vertex v_i

$k_i^{in}(S)$ = number of neighbors of v_i inside $S \subseteq V$

$k_i^{out}(S)$ = number of neighbors of v_i outside $S \subseteq V$

Cohesion *strong* conditions

- **Strong Cohesion Condition (SCC):**

  S community in the *strong sense* if and only if
  every one of its vertices has more neighbors within the community than outside:

$$\forall v_i \in S \quad k_i^{in}(S) > k_i^{out}(S)$$

Cohesion *strong* conditions

- **Strong Cohesion Condition (SCC):**

  S community in the *strong sense* if and only if
  every one of its vertices has more neighbors within the community than outside:

$$\forall v_i \in S \quad k_i^{in}(S) > k_i^{out}(S)$$

- **Almost-Strong Cohesion Condition (ASCC):**

  S community in the *almost-strong sense* if and only if
  every one of its vertices with degree different from 2 has more neighbors within the community than outside, and
  every vertex with degree 2 has at least one neighbor in the same community:

$$\forall v_i \in S \,|\, k_i \neq 2 \quad k_i^{in}(S) > k_i^{out}(S)$$
$$\forall v_i \in S \,|\, k_i = 2 \quad k_i^{in}(S) > 0$$

Cohesion *strong* conditions

- **Semi-Strong Cohesion Condition (SSCC):**

  S community in the *semi-strong sense* if and only if
  every one of its vertices has more neighbors within the community than the maximum number of neighbors within any other community:

$$\forall v_i \in S \quad k_i^{in}(S) > \max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_j \in S_t} A_{ij}$$

Cohesion *weak* conditions

- **Weak Cohesion Condition (WCC):**

  S community in the *weak sense* if and only if
  the sum of internal degrees within S is larger than the sum of external degrees, that is the number of edges joining S to the rest of the network $V \setminus S$:

$$\sum_{v_i \in S} k_i^{in}(S) > \sum_{v_i \in S} k_i^{out}(S)$$

Cohesion *weak* conditions

- **Weak Cohesion Condition (WCC):**

  S community in the *weak sense* if and only if
  the sum of internal degrees within S is larger than the sum of external degrees, that is the number of edges joining S to the rest of the network $V \setminus S$:

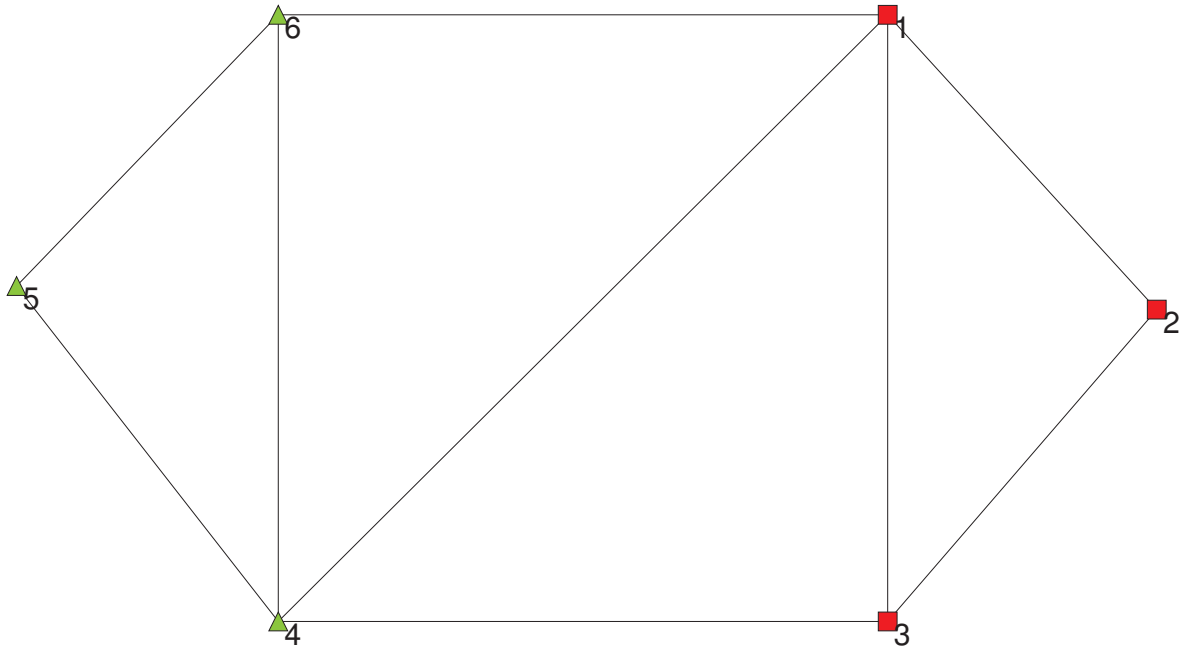
$$\sum_{v_i \in S} k_i^{in}(S) > \sum_{v_i \in S} k_i^{out}(S)$$

- **Extra-Weak Cohesion Condition (EWCC):**

  S community in the *extra-weak sense* if and only if
  the sum of internal degrees within S is larger than the maximum number of edges joining a vertex of S to a vertex in some other community in the rest of the network:

$$\sum_{v_i \in S} k_i^{in}(S) > \max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_i \in S} \sum_{v_j \in S_t} A_{ij}$$

Cohesion conditions: Example



WCC and EWCC satisfied

SCC and SSCC not satisfied

Outline

Cohesion conditions in modularity maximization

Do optimal solutions obtained by modularity maximization satisfy, and to which degree, the five cohesion conditions?

Cohesion conditions in modularity maximization

Do optimal solutions obtained by modularity maximization satisfy, and to which degree, the five cohesion conditions?

dataset	*n*	*m*	*M*	*M_strong*	*M_almost strong*	*M_semi strong*	*M_weak*	*M_extra weak*
strike	24	38	4	2	3	2	4	4
karate	34	78	4	1	2	2	4	4
Korea1	35	69	5	2	2	3	5	5
Korea2	35	84	5	3	4	3	5	5
sawmill	36	62	4	4	4	4	4	4
dolphins small	40	70	6	3	6	3	6	6
graph	60	114	7	0	2	3	7	7
dolphins	62	159	5	2	2	3	4	5
Les Misérables	77	254	6	2	2	3	6	6
p53 protein	104	226	7	1	2	2	6	7
political books	105	441	5	2	2	2	4	4
percentage of communities satisfying the condition				37.93%	53.45%	51.72%	94.83%	98.28%

Outline

Modularity maximization formulations

Mathematical Programming formulations:

- ⋆ reduction of modularity maximization to clique partitioning

  $\Rightarrow$ linear optimization problem (LP) in 0-1 variables

- ⋆ direct formulation

  $\Rightarrow$ mixed 0-1 quadratic optimization problem (MIQP)

Modularity maximization formulations

Mathematical Programming formulations:

- ⋆ reduction of modularity maximization to clique partitioning
  $\Rightarrow$ linear optimization problem (LP) in 0-1 variables

- ⋆ direct formulation
  $\Rightarrow$ mixed 0-1 quadratic optimization problem (MIQP)

- Clique partitioning: assignment of entities to communities is not explicitly considered, it only appears as a consequence of the optimal solution

- MIQP formulation: uses variables to denote assignment of entities to communities

Modularity maximization formulations

Mathematical Programming formulations:

- ⋆ reduction of modularity maximization to clique partitioning
  $\Rightarrow$ linear optimization problem (LP) in 0-1 variables

- ⋆ direct formulation
  $\Rightarrow$ mixed 0-1 quadratic optimization problem (MIQP)

- Clique partitioning: assignment of entities to communities is not explicitly considered, it only appears as a consequence of the optimal solution
  $\rightarrow$ adding cohesion conditions not easy

- MIQP formulation: uses variables to denote assignment of entities to communities
  $\rightarrow$ adding cohesion conditions easier

Modularity maximization: MIQP (Xu, Tsoka and Papageorgiou, 2007)

Variables used to identify to which community each vertex and each edge belongs:

$$X_{rs} = \begin{cases} 1 & \text{if edge } r \text{ belongs to community } s \\ 0 & \text{otherwise} \end{cases} \qquad \forall r = 1, 2, \ldots m,\ s = 1, 2, \ldots M$$

$$Y_{is} = \begin{cases} 1 & \text{if vertex } i \text{ belongs to community } s \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i = 1, 2, \ldots n,\ s = 1, 2, \ldots M$$

$$\max\ Q = \sum_s [a_s - e_s] = \sum_s \left[\frac{m_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right]$$

m_s = number of edges in community s
d_S = sum of degrees k_i of vertices in s

- $m_s = \sum_r X_{rs}$ and $d_S = \sum_i k_i Y_{is}$
- $\sum_s Y_{is} = 1 \quad \forall i = 1, 2, \ldots n$ — each vertex belongs to one community
- $X_{rs} \le Y_{is} \quad \forall r = \{v_i, v_j\} \in E$
  $X_{rs} \le Y_{js} \quad \forall r = \{v_i, v_j\} \in E$
  any edge $r = \{v_i, v_j\}$ belongs to community s $\Leftrightarrow$ both of its end vertices i,j belong to s
- $u_s \le u_{s-1}$ — community s nonempty $\Leftrightarrow$ $s-1$ is so ($u_s = 1$ if module s nonempty, 0 otherwise)
- symmetry-breaking constraints

Modularity maximization: MIQP (Xu, Tsoka and Papageorgiou, 2007)

Variables used to identify to which community each vertex and each edge belongs:

$$X_{rs} = \begin{cases} 1 & \text{if edge } r \text{ belongs to community } s \\ 0 & \text{otherwise} \end{cases} \qquad \forall r = 1, 2, \ldots m,\ s = 1, 2, \ldots M$$

$$Y_{is} = \begin{cases} 1 & \text{if vertex } i \text{ belongs to community } s \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i = 1, 2, \ldots n,\ s = 1, 2, \ldots M$$

$$\max\ Q = \sum_s [a_s - e_s] = \sum_s \left[\frac{m_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right]$$

m_s = number of edges in community s
d_S = sum of degrees k_i of vertices in s

- $m_s = \sum_r X_{rs}$ and $d_S = \sum_i k_i Y_{is}$
- $\sum_s Y_{is} = 1 \quad \forall i = 1, 2, \ldots n$
- $X_{rs} \le Y_{is} \quad \forall r = \{v_i, v_j\} \in E$
  $X_{rs} \le Y_{js} \quad \forall r = \{v_i, v_j\} \in E$
- $u_s \le u_{s-1}$
- symmetry-breaking constraints

$\Downarrow$

Mixed-Integer Quadratic Program
with a convex continuous relaxation

Adding cohesion conditions in the MIQP (1/5)

- SCC:

  S community in the *strong sense* $\Leftrightarrow$ every one of its vertices has more neighbors within the community than outside:

$$\forall s \in \{1, \ldots, M\},\ \forall v_i \in V \qquad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is} \left(\lfloor \frac{k_i}{2} \rfloor + 1 \right).$$

Adding cohesion conditions in the MIQP (1/5)

- SCC:

  S community in the *strong sense* $\Leftrightarrow$ every one of its vertices has more neighbors within the community than outside:

$$\forall s \in \{1,\dots,M\},\ \forall v_i \in V \qquad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is} \left(\lfloor \frac{k_i}{2} \rfloor + 1 \right).$$

Indeed, from the definition of SCC:

$$\forall s \in \{1,\dots,M\},\ \forall v_i \in V \qquad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq k_i - \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} + 1,$$

i.e. the in-degree ($\sum_{v_j \in V : j \neq i} A_{ij} Y_{js}$) of vertex v_i is strictly greater than the out-degree.

$\Longrightarrow$ (algebraic manipulations)

$$\forall s \in \{1,\dots,M\},\ \forall v_i \in V \qquad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq \lfloor \frac{k_i}{2} \rfloor - (1 - Y_{is}) \lfloor \frac{k_i}{2} \rfloor + Y_{is}$$

(easily checked for both $Y_{is} = 1$ and $Y_{is} = 0$).

Adding cohesion conditions in the MIQP (2/5)

- ASCC:
  S community in the *almost-strong sense* $\Leftrightarrow$ every one of its vertices with degree different from 2 has more neighbors within the community than outside, and
  every vertex with degree 2 has at least one neighbor in the same community:

$$\forall s \in \{1,\dots,M\},\ \forall v_i \in V \,|\, k_i \neq 2 \qquad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is} \left(\lfloor \frac{k_i}{2} \rfloor + 1 \right)$$

$$\forall s \in \{1,\dots,M\},\ \forall v_i \in V \,|\, k_i = 2 \qquad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is}$$

Adding cohesion conditions in the MIQP (3/5)

- SSCC:

  S community in the *semi-strong sense* $\Leftrightarrow$ every one of its vertices has more neighbors within the community than the max number of neighbors within any other community:

$$\forall s,t \in \{1,\dots,M\} \,|\, s \neq t,\ \forall v_i \in V \sum_{j \in V : j \neq i} A_{ij} Y_{js} \geq \sum_{v_j \in V : j \neq i} A_{ij} Y_{jt} + 1 - (1 - Y_{is})(k_i + 1)$$

Adding cohesion conditions in the MIQP (3/5)

- SSCC:

  S community in the *semi-strong sense* $\Leftrightarrow$ every one of its vertices has more neighbors within the community than the max number of neighbors within any other community:

$$\forall s, t \in \{1, \ldots, M\} \,|\, s \neq t,\ \forall v_i \in V \sum_{j \in V : j \neq i} A_{ij} Y_{js} \geq \sum_{v_j \in V : j \neq i} A_{ij} Y_{jt} + 1 - (1 - Y_{is})(k_i + 1)$$

  Indeed:

  (i) $Y_{is} = 1 \Rightarrow$
  - the *lhs* term = in-degree of v_i,
  - the first term of the *rhs* = part of the out-degree of v_i corresponding to edges with extremities in s and $t \neq s$.

  The last term disappears $\rightarrow$ this partial out-degree must be strictly smaller than the in-degree of v_i.

  Similar conditions hold for all other communities $\rightarrow$ such a relation holds for the community for which the partial out-degree of v_i is largest.

  (ii) $Y_{is} = 0 \Rightarrow$ the rhs is non-positive and the condition is verified.

Adding cohesion conditions in the MIQP (4/5)

- WCC:

  S community in the *weak sense* $\Leftrightarrow$ the sum of internal degrees within S is larger than the sum of external degrees, that is the number of edges joining S to the rest of the network :

$$\forall s \in \{1, \ldots, M\} \quad 4 \sum_{r \in E} X_{rs} \geq \sum_{v_i \in V} k_i Y_{is} + 1$$

  Indeed:

  - the sum of in-degrees for community s may be written as $2 \sum_{r \in E} X_{rs}$
  - the sum of out-degrees of s = sum of all the degrees minus the sum of in-degrees for vertices of that community: $\sum_{v_i \in V} k_i Y_{is} - 2 \sum_{r \in E} X_{rs}$.

Adding cohesion conditions in the MIQP (5/5)

- EWCC:

  S community in the *extra-weak sense* $\Leftrightarrow$ the sum of internal degrees within S is larger than the max number of edges joining a vertex of S to a vertex in some other community:

$$\forall s,t \in \{1,\dots,M\} \,|\, s \neq t \quad 2\sum_{r\in E} X_{rs} \geq \sum_{r=\{v_i,v_j\}\in E} \left(Y_{is}Y_{jt} + Y_{js}Y_{it}\right) + 1.$$

Linearization:

introduce $\forall r = \{v_i, v_j\} \in E$ non-negative variables $Z_{rst} = Y_{is}Y_{jt}$ and $Z'_{rst} = Y_{js}Y_{it}$:

$$\forall s,t \in \{1,\dots,M\} \,|\, s \neq t \qquad 2\sum_{r\in E} X_{rs} \geq \sum_{r\in E} (Z_{rst} + Z'_{rst}) + 1$$

and add linearization constraints $\forall s,t \in \{1,\dots,M\} \,|\, s \neq t$:

$$\begin{aligned}
Z_{rst} &\leq Y_{is} \\
Z_{rst} &\leq Y_{jt} \\
Z_{rst} &\geq Y_{is} + Y_{jt} - 1 \\
Z'_{rst} &\leq Y_{js} \\
Z'_{rst} &\leq Y_{it} \\
Z'_{rst} &\geq Y_{js} + Y_{it} - 1
\end{aligned}$$

Mathematical Programming models using cohesion conditions

Modularity maximization with cohesion constraints:

New mathematical models:

- MIQP + SCC
- MIQP + SSCC
- MIQP + ASCC
- MIQP + WCC
- MIQP + EWCC

Outline

Solving the optimization problems by an exact method

The proposed MIQP problems solved exactly using CPLEX

Why exact methods?

- having an exact solution solves the problem of separating possible inadequacies of the model from eventual errors resulting from the use of heuristics
  $\Rightarrow$ communities may be interpreted with more confidence
- an exact algorithm can provide a benchmark of exactly solved instances which can be used to compare heuristics and fine tune them
- an exact algorithm may be stopped and the best solution found considered as a heuristic one

Inconvenient: cannot solve large-scale problems

ENAC

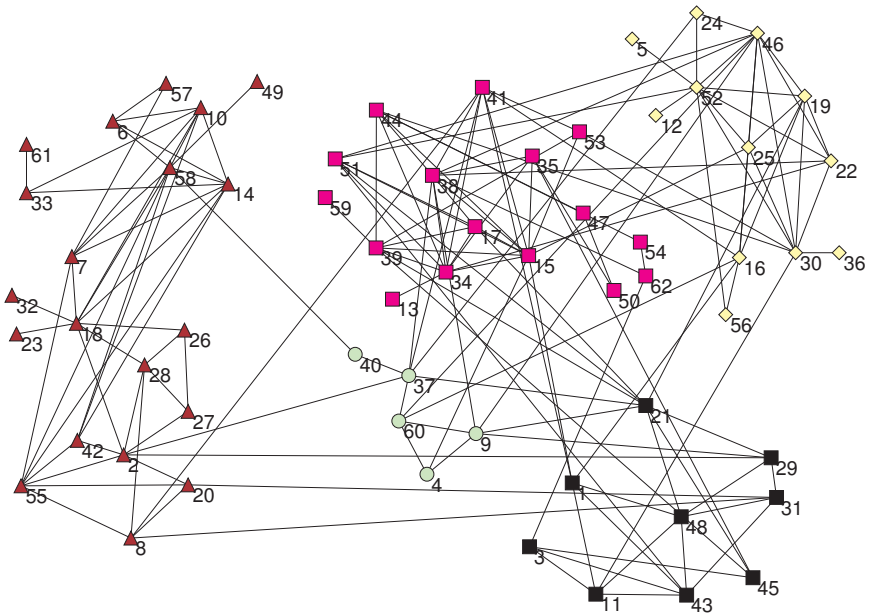
Outline

Results: Modularity maximization + weak constraints

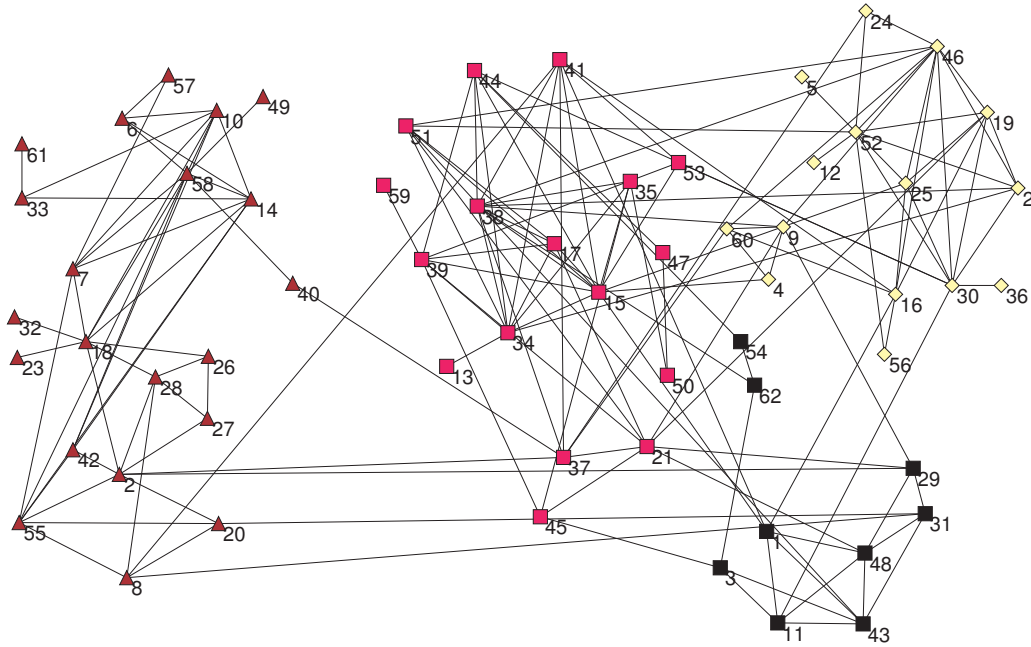
network			modularity maximization		weak		extra-weak	
dataset	n	m	M	Q	M_w	Q_w	M_{ew}	Q_{ew}
strike	24	38	4	0.561981	4	0.561981	4	0.561981
karate	34	78	4	0.41979	4	0.41979	4	0.41979
Korea1	35	69	5	0.477736	5	0.477736	5	0.477736
Korea2	35	84	5	0.450822	5	0.450822	5	0.450822
sawmill	36	62	4	0.550078	4	0.550078	4	0.550078
dolphins small	40	70	4	0.620714	4	0.620714	4	0.620714
graph	60	114	7	0.502655	7	0.502655	7	0.502655
dolphins	62	159	5	0.528519	**4**	**0.526799**	5	0.528519
Les Misérables	77	254	6	0.560008	6	0.560008	6	0.560008
p53 protein	104	226	7	0.535134	**6**	**0.534488**	7	0.535134
political books	105	441	5	0.527237	**4**	**0.526938**	**4**	**0.526938**
average			5.090909	0.521334	4.818182	0.521092	5	0.521307

Results: Modularity maximization + weak constraints - Details

`dolphins` dataset



unconstrained modularity maximization



modularity maximization + weak cohesion constraint

Results: Modularity maximization + weak constraints - Details

dolphins dataset

Partition obtained with unconstrained modularity maximization

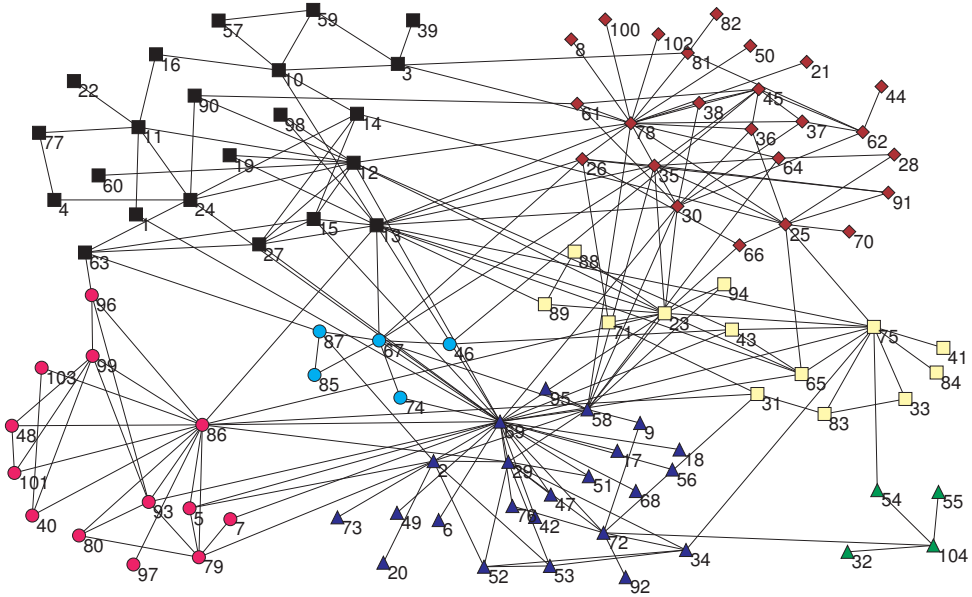
C_1	C_2	C_3	C_4	C_5
1, 3, 11, 21 29, 31, 43, 45, 48	2, 6, 7, 8, 10 14, 18, 20, 23 26, 27, 28, 32 33, 42, 49, 55 57, 58, 61	4, 9, 37, 40 60	5, 12, 16, 19 22, 24, 25, 30 36, 46, 52, 56	13, 15, 17, 34 35, 38, 39, 41 44, 47, 50, 51 53, 54, 59, 62

Partition obtained with modularity maximization + weak cohesion constraint

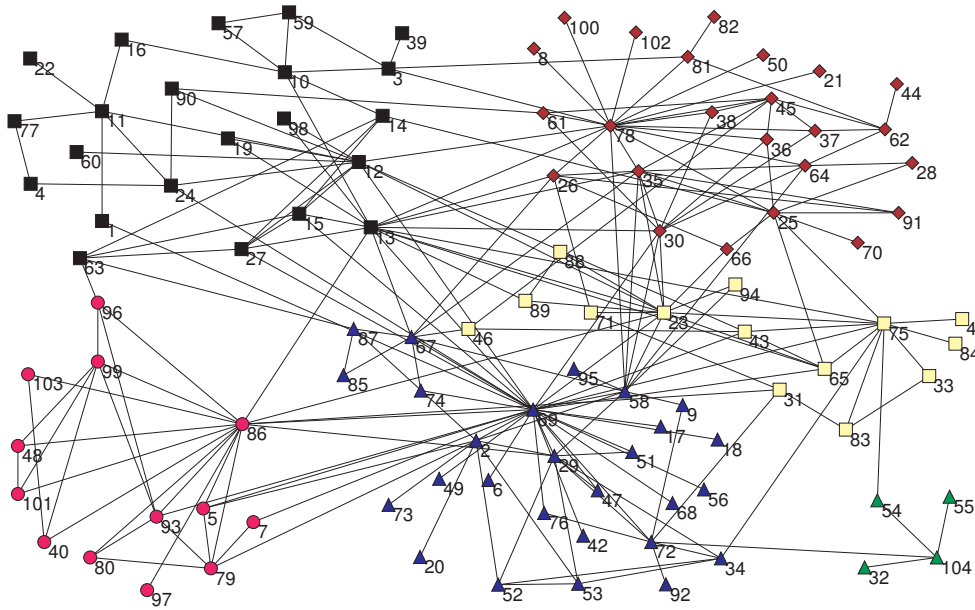
C_1^w	C_2^w	C_3^w	C_4^w
1, 3, 11, 29 31, 43, 48, 54 62	2, 6, 7, 8, 10 14, 18, 20, 23 26, 27, 28, 32 33, 40, 42, 49 55, 57, 58, 61	4, 5, 9, 12, 16 19, 22, 24, 25 30, 36, 46, 52 56, 60	13, 15, 17, 21 34, 35, 37, 38 39, 41, 44, 45 47, 50, 51, 53 59

Results: Modularity maximization + weak constraints - Details

`p53 protein` dataset



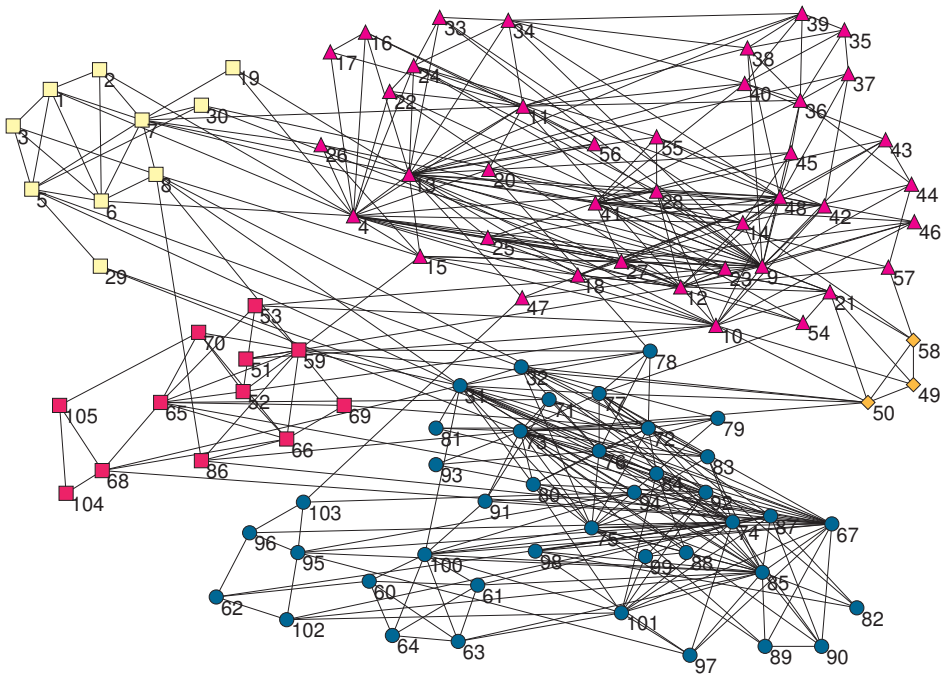
unconstrained modularity maximization



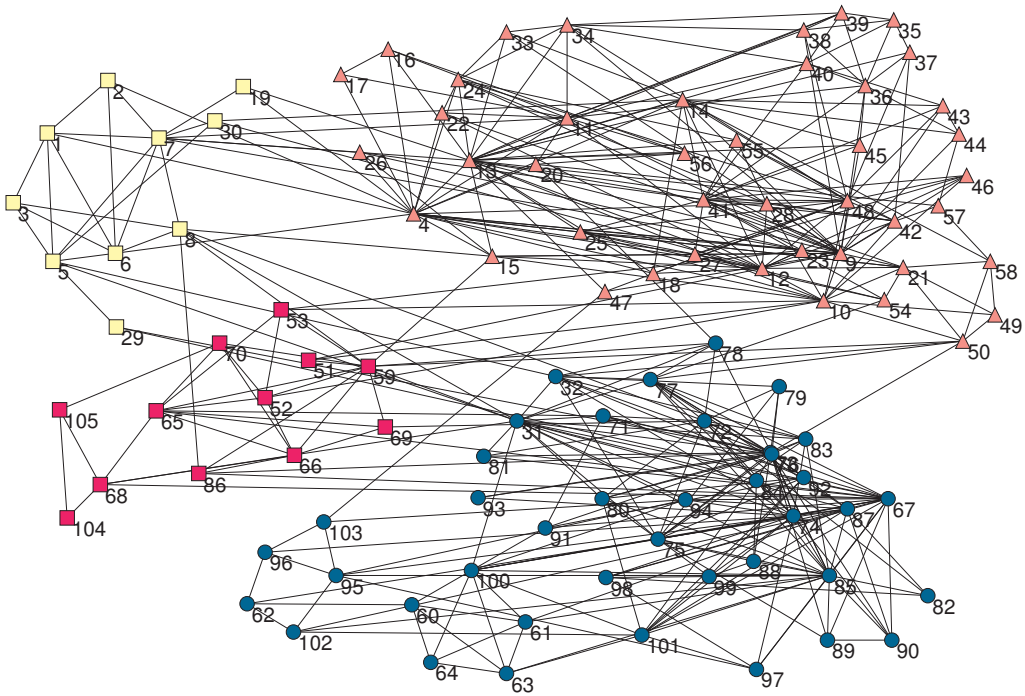
modularity maximization + weak cohesion constraint

Results: Modularity maximization + weak constraints - Details

`polbooks` dataset



original modularity maximization



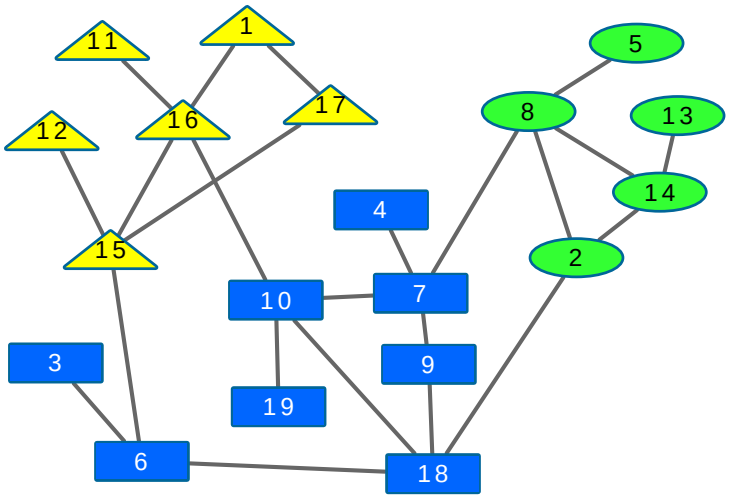
modularity maximization + weak cohesion constraint

Results: Modularity maximization + strong constraints

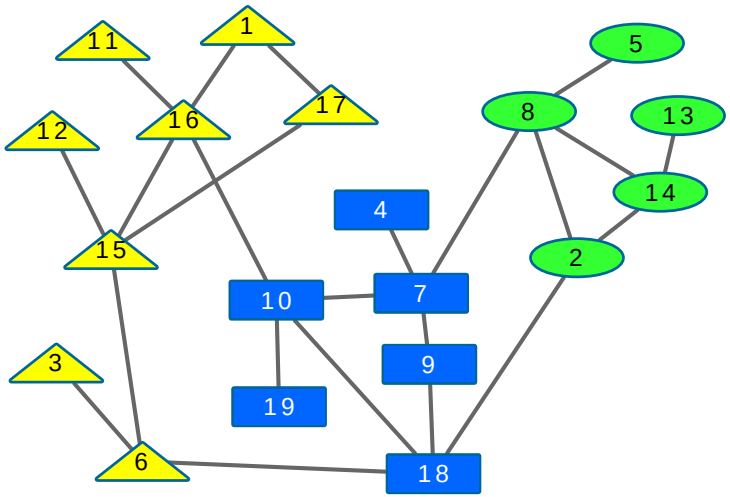
network			modularity max		strong		almost-strong		semi-strong	
dataset	n	m	M	Q	M_s	Q_s	M_{as}	Q_{as}	M_{ss}	Q_{ss}
strike	24	38	4	0.561981	2	0.257271	3	0.54813	2	0.257271
karate	34	78	4	0.41979	2	0.132807	4	0.402038	2	0.132807
Korea1	35	69	5	0.477736	4	0.383638	4	0.383638	4	0.383638
Korea2	35	84	5	0.450822	3	0.424036	4	0.432469	3	0.424036
sawmill	36	62	4	0.550078	4	0.550078	4	0.550078	4	0.550078
dolphins small	40	70	4	0.620714	3	0.573571	4	0.620714	3	0.573571
graph	60	114	7	0.502655	1	0	4	0.438135	1	0
dolphins	62	159	5	0.528519	2	0.359242	3	0.480598	2	0.359242
Les Misérables	77	254	6	0.560008	4	0.437868	6	0.52921	4	0.437868
p53 protein	104	226	7	0.535134	2	0.284204	4	0.472502	2	0.284204
political books	105	441	5	0.527237	3	0.497969	3	0.497969	3	0.497969
average			5.09091	0.521334	2.727273	0.354608	3.909091	0.486862	2.727273	0.354608

Results: Modularity maximization + strong constraints

modularity with the strong and the semi-strong conditions yields different results:



modularity = 0.476371



modularity = 0.483932

Vertex 18 in the semi-strong partition does not respect the strong condition, since it has two neighbors inside its own community (i.e., vertices 9 and 10) and two neighbors outside (i.e., vertices 2 and 6).

In the strong partition all the neighbors of vertex 18 belong to its own community.

Outline

Results: qualitative analysis

For some real world problems, the behaviour of the system is known
$\Rightarrow$ compare obtained partitions against the actual outcomes

Results: qualitative analysis

- `strike` dataset

  informal communications among the 24 employees of a wood processing facility concerning a strike.

  vertices = employees

  edges = frequent discussions beetween employees about the strike

3 categories of employees:

- spanish-speaking
- young (below 30 years old) english-speaking
- old english-speaking

$\Rightarrow$ the correct partition consists of 3 communities

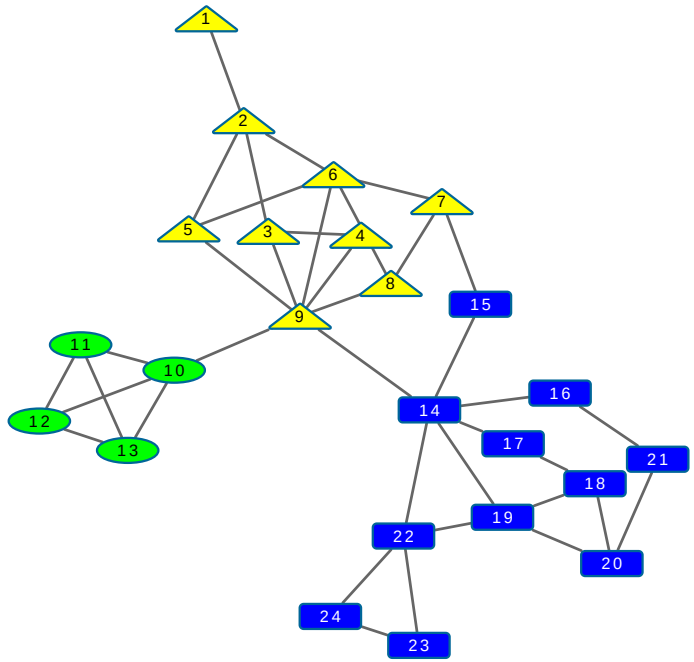
Results: qualitative analysis

- `strike` dataset

  informal communications among the 24 employees of a wood processing facility concerning a strike.

  vertices = employees

  edges = frequent discussions beetween employees about the strike



3 categories of employees:

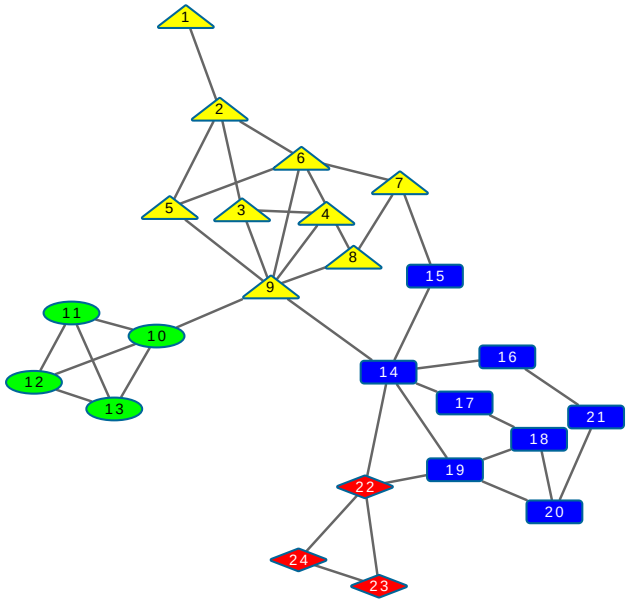
- spanish-speaking
- young (below 30 years old) english-speaking
- old english-speaking

$\Rightarrow$ the correct partition consists of 3 communities

modularity + almost-strong condition: correct partition

Results: qualitative analysis

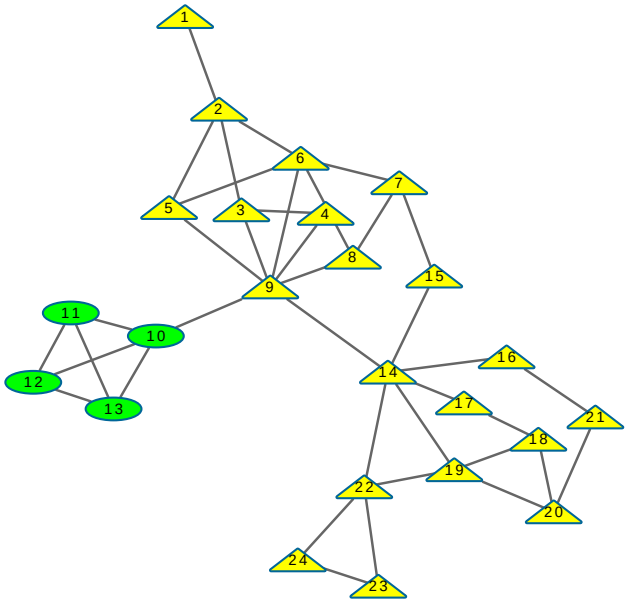
- `strike` dataset



modularity maximization alone and
modularity + weak and extra-weak conditions

4 communities:

the new one (red) does not seem to be related to any particular tie between the workers



modularity + strong and semi-strong conditions

2 communities:

spanish-speaking employees, english-speaking employees

$\Rightarrow$ strong and semi-strong cond. have got the effect of breaking the hierarchical structure of the english-speaking community

Results: qualitative analysis

For some real world problems, the behaviour of the system is known
$\Rightarrow$ compare obtained partitions against the actual outcomes

Results: qualitative analysis

- `political books` dataset

  vertices = books about politics in US

  edges = two vertices are connected if they are often bought by the same readers

3 main types of books:

- liberal
- conservative
- centrist or unaligned

$\Rightarrow$ we would expect 3 communities

Results: qualitative analysis

- `political books` dataset

  vertices = books about politics in US
  edges = two vertices are connected if they are often bought by the same readers

3 main types of books:

- liberal
- conservative
- centrist or unaligned

$\Rightarrow$ we would expect 3 communities

- modularity maximization: 5 communities
- modularity + weak and extra-weak conditions: 4 communities
- modularity + strong, almost-strong, and semi-strong conditions: 3 communities

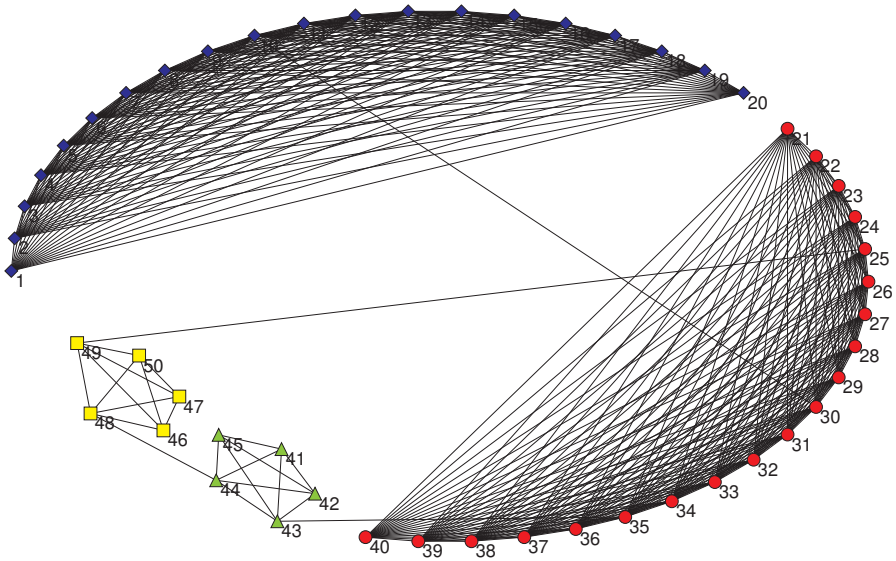
  Average number of vertices classified correctly: 60.8%
  Books belonging to the 3rd category (i.e., centrist or unaligned) are not densely connected between each other and have got many neighbors in other communities

Outline

Impact on modularity resolution limit

Modularity **resolution limit**:
in some cases small clusters may not be detected, and they remain hidden within other clusters

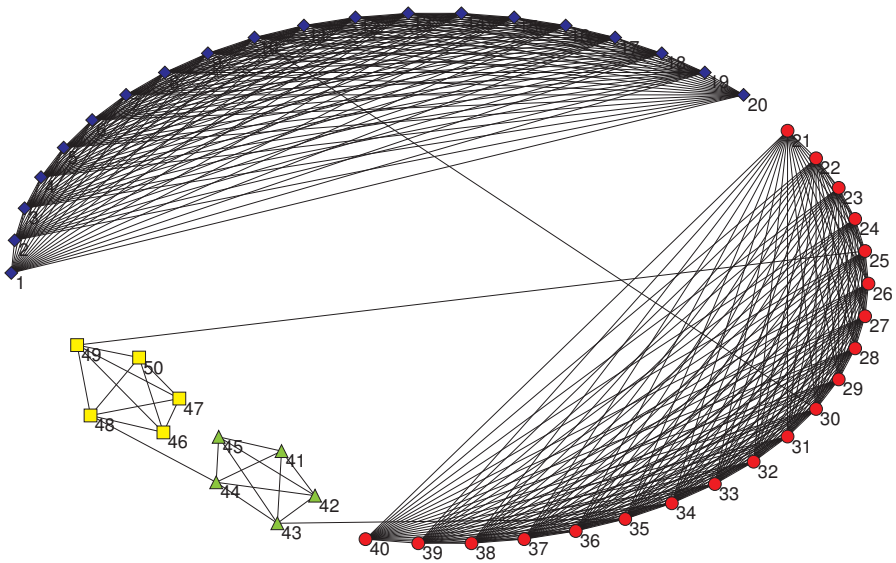
Example (Fortunato & Barthelemy, 2007)



Impact on modularity resolution limit

Modularity **resolution limit**:
in some cases small clusters may not be detected, and they remain hidden within other clusters

Example (Fortunato & Barthelemy, 2007)



- modularity without cohesion conditions:
  3 communities (the two large cliques + the union of the small ones)
- modularity + weak and exra-weak cond.:
  3 communities
- modularity + strong, almost-strong, and semi-strong conditions:
  correct partition with 4 cliques

strong, semi-strong and almost-strong cohesion conditions overcome the resolution limit

Outline

Relation with detectability (1/3)

Theory of **detectability** of communities:
There is a sharp *phase transition* s.t.
community detection appears to be possible above a certain threshold,
while below this threshold methods to detect communities are expected to fail.

In case of

Poissonian degrees distribution

2 communities

the detection of a modular structure is possible when

$$c_{in} - c_{out} \geq \sqrt{c_{in} + c_{out}}$$

c_{in} = internal node degrees averages
c_{out} = external node degrees averages

Relation with detectability (2/3)

Can we relate the detectability of communities to the strength of cohesion conditions?

Numerical tests:

- $\sqrt{c_{in} + c_{out}}$ constantly equal to $2\sqrt{2}$ $\Rightarrow$ threshold at $c_{in} = 5.4$ and $c_{out} = 2.6$
- c_{in} increased from 4 to 7, c_{out} decreased from 4 to 1, step size 0.2
- for each one of these 16 combinations, 10 random instances generated
  $\Rightarrow$ 80 instances below the detectability threshold, and 80 above
- quality metric: average number of vertices that are classified correctly on the 2 communities, averaged over the 10 random instances

Relation with detectability (2/3)

Can we relate the detectability of communities to the strength of cohesion conditions?

Numerical tests:

- $\sqrt{c_{in} + c_{out}}$ constantly equal to $2\sqrt{2}$ $\Rightarrow$ threshold at $c_{in} = 5.4$ and $c_{out} = 2.6$
- c_{in} increased from 4 to 7, c_{out} decreased from 4 to 1, step size 0.2
- for each one of these 16 combinations, 10 random instances generated
  $\Rightarrow$ 80 instances below the detectability threshold, and 80 above
- quality metric: average number of vertices that are classified correctly on the 2 communities, averaged over the 10 random instances

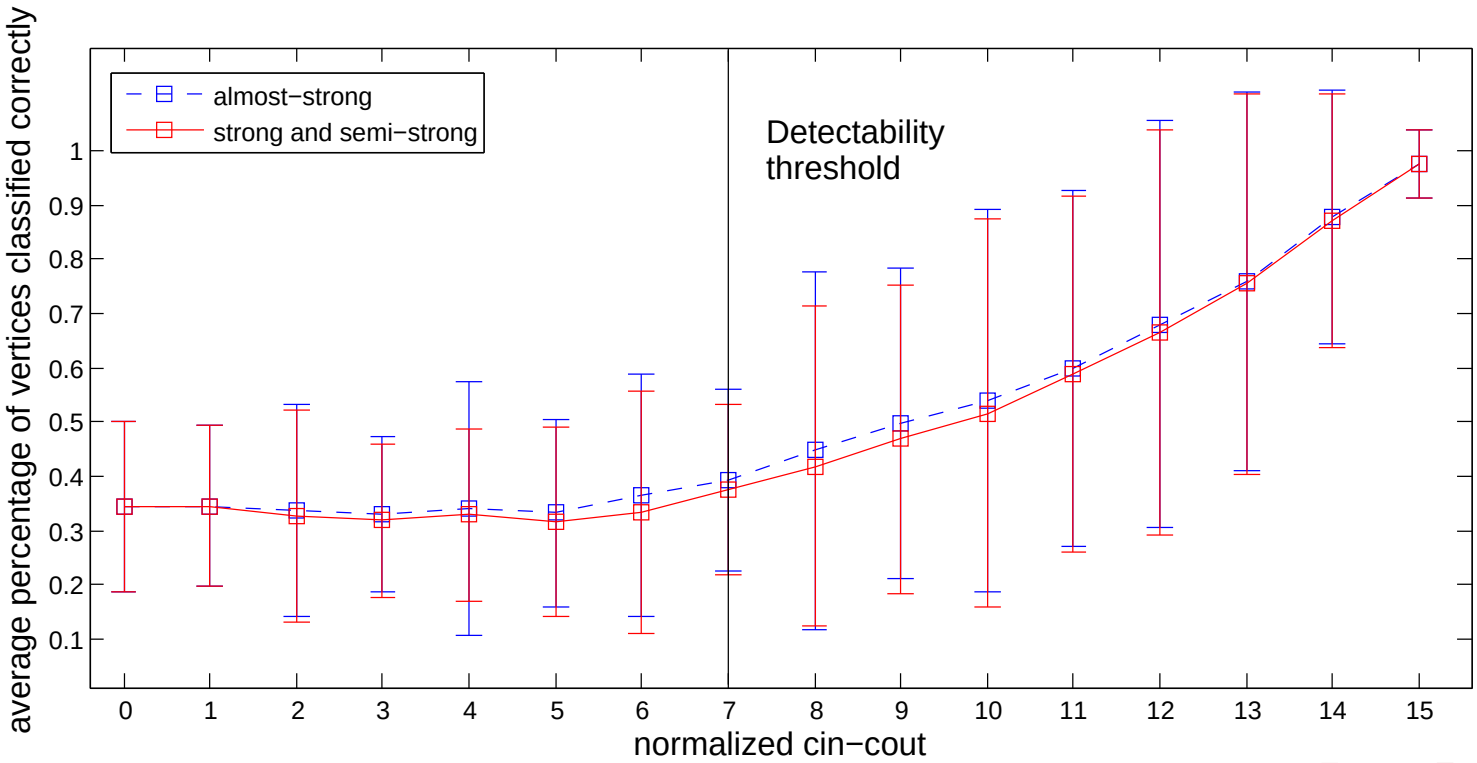
The behaviour of modularity maximization subject to cohesion constraints appears to be coherent with the detectability of the considered network structures

ENAC

Relation with detectability (3/3)

Strict cohesion conditions (SCC, SSCC, ASCC):

- for instances below the detectability threshold
  (community structure intrinsically difficult to detect)
  → low percentage of correctly classified vertices
- for instances above the threshold
  → a significantly higher precision even with such strict conditions



Outline

Conclusions

- Five kinds of cohesion conditions
- Some of them are quite strict, the weak one is more intuitive
- Added to a modularity maximization (MIQP) model, yield interesting results

Future work:

- Solution of large-scale datasets:

  $\Rightarrow$ heuristics tailored on the problem
- Hierarchical network clustering using cohesion conditions

The end

Thank you!