On the NP-Completeness of Some Graph Cluster Measures

Jiří Šíma† and Satu Elisa Schaeffer‡

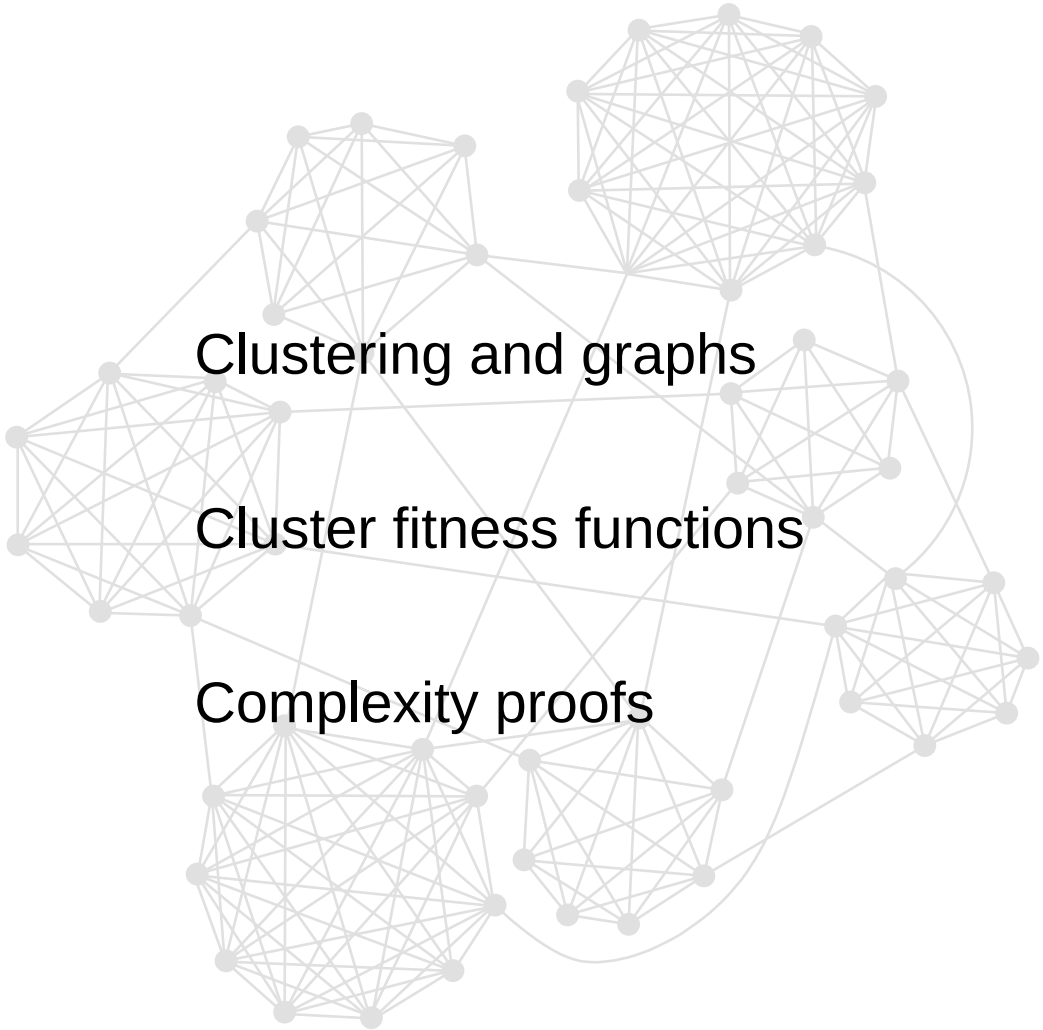
† Academy of Sciences of the Czech Republic

‡ Helsinki University of Technology, Finland

elisa.schaeffer@tkk.fi

SofSem 06, Merin, Czech Republic

OUTLINE



Clustering and graphs

Cluster fitness functions

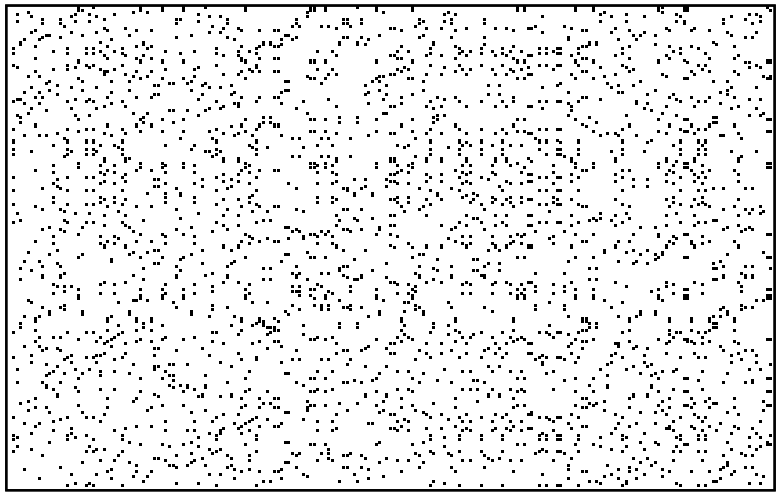
Complexity proofs

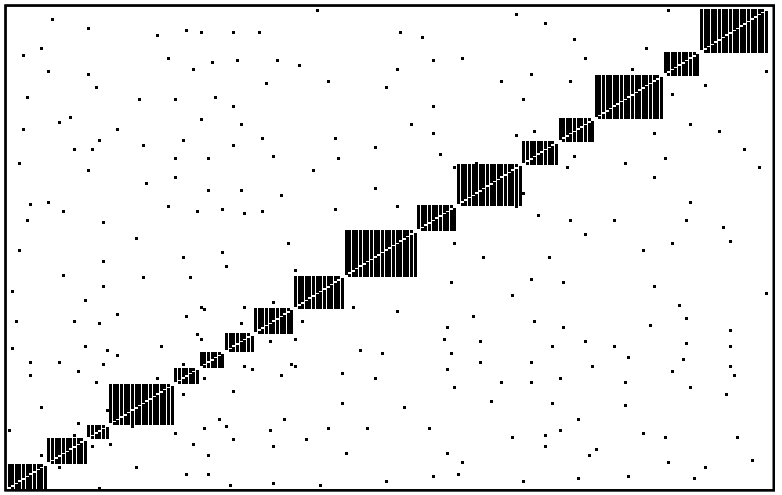
CLUSTERING

- Tool for *analysis* and *exploration* of data; discovering **natural groups** (clusters) of *similar* elements in a data set
- **Applications**: data mining, VLSI design, parallel computing, web searching, relevance queries, software engineering, computer graphics, pattern recognition, gene analysis
- Massive input data sets $\Rightarrow$ complexity research to study *scalability*

GRAPH CLUSTERING

Cluster $\approx$ a *connected* subgraph induced by a vertex set S with *many* internal edges and *few* edges to outside vertices in $V \setminus S$.





NOTATION & TERMINOLOGY

$G = (V, E)$	an *undirected*, *unweighted* graph with no self-loops
$G(S) = (S, E(S))$	a *subgraph induced* by $S \subseteq V$
	$E(S) = \{\{u, v\} \in E \mid u, v \in S\}$
Clique	a *fully connected* subgraph
Degree	$d_G(v) = \lvert\{u \in V \,;\, \{u, v\} \in E\}\rvert$
Cubic graph	$d_G(v) = 3 \, \forall v \in V$

CONDUCTANCE

$S \subset V$ creates a **cut** of $G \triangleq$ a partition of V into non-empty **disjoint** sets S and $V \setminus S$

The **size** of the cut is

$$c_G(S) = |\{\{u, v\} \in E \,;\, u \in S \,, v \in V \setminus S\}| \,.$$

The **conductance** of a cut $\emptyset \neq S \subset V$ is

$$\Phi_G(S) = \frac{c_G(S)}{\min(d_G(S), d_G(V \setminus S))},$$

where $d_G(S) = \sum_{v \in S} d_G(v)$.

More measures

$$\delta_G(S) = \frac{|E(S)|}{\binom{|S|}{2}} = \frac{2|E(S)|}{|S|(|S|-1)}$$ **Local density**

$(\delta_G(S) = 0 \text{ for } |S| = 1)$

$$\varrho_G(S) = \frac{|E(S)|}{|E(S)| + c_G(S)}$$ **Relative density**

$$\varepsilon_G(S) = \binom{|S|}{2} - |E(S)| + c_G(S)$$ **Single cluster editing**

ALGORITHMS

Algorithms usually construct clusters somehow optimizing one or more fitness measures.

We prove that the decision problems corresponding to thresholding $\Phi_G(S)$, $\delta_G(S)$, $\varrho_G(S)$, and $\varepsilon_G(S)$ are **NP**-complete.

DECISION PROBLEM: CONDUCTANCE

Minimum Conductance (CONDUCTANCE)

Instance: A graph $G = (V, E)$ and a rational number $\phi \in [0, 1]$.

Question: Is there a cut $S \subset V$ such that $\Phi_G(S) \leq \phi$?

Theorem: CONDUCTANCE is **NP**-complete.

PROOF

CONDUCTANCE $\in \mathbf{NP}$ (guess $S \subset V$ and verify $\Phi_G(S) \leq \phi$ in polyn. time)

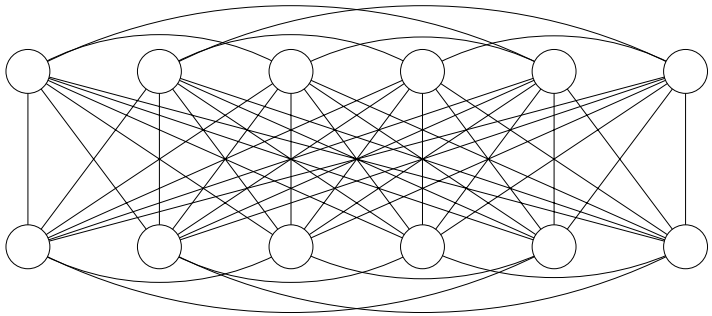
NP-hardness: the following problem is reduced to CONDUCTANCE in polynomial time

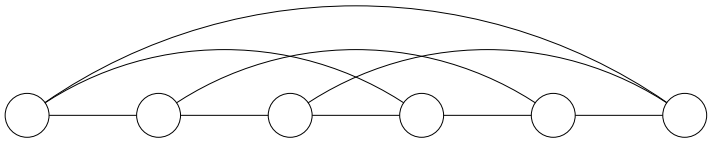
Maximum Cut for Cubic Graphs (MAX CUT–3)
Instance: A cubic graph $G = (V, E)$ and an integer $a > 0$.
Question: Is there a cut $A \subset V$ s.t. $c_G(A) \geq a$?

Reduction from Max Cut–3

Max Cut–3 **instance:**

a cubic graph $G = (V, E)$ with $n = |V|$ and an integer $a > 0$.





Conductance **instance:**

$G' = (V', E')$ composed of two fully interconnected copies of the *complement* of G

CONSTRUCTION DETAILS

$V' = V_1 \cup V_2$ $\quad V_i = \{v^i \,;\, v \in V\}$ for $i = 1, 2$

$E' = E_1 \cup E_2 \cup E_3$ $\quad E_3 = \{\{u^1, v^2\} \,;\, u, v \in V\},$

$E_i = \{\{u^i, v^i\} \,;\, u, v \in V, u \neq v, \{u, v\} \notin E\}$

for $i = 1, 2$

Conductance bound: $\quad \phi = \frac{1}{2n-4}\left(n - \frac{2a}{n}\right)$

G cubic $\Rightarrow$ $\quad d_{G'}(v) = 2n - 4 \,\forall v \in V'$

Polynomiality: $\quad |V'| = 2n$ and $|E'| = (2n-4)n$

CONDUCTANCE IN G'

A cut $\emptyset \neq S \subset V'$ in G' with $k = |S| \leq 2n$

$$c_{G'}(S) = c_{G'}(V' \setminus S) \Rightarrow \Phi_{G'}(S) = \Phi_{G'}(V' \setminus S)$$

($k \leq n$ w.l.o.g.)

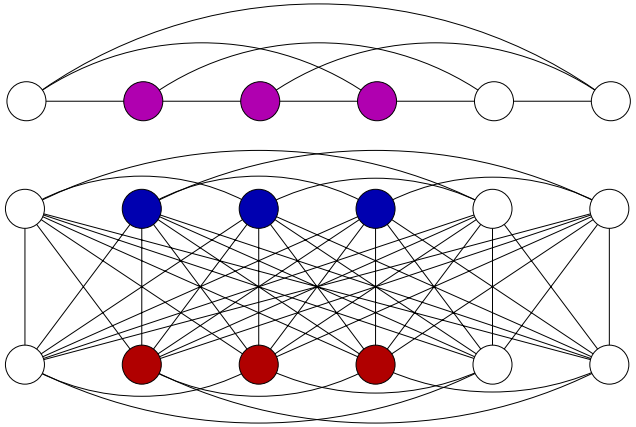
$$\begin{aligned} \Phi_{G'}(S) &= \frac{|S| \cdot |V' \setminus S| - c_G(S_1) - c_G(S_2)}{(2n-4) \cdot |S|} \\ &= \frac{1}{2n-4}\left(2n - k - \frac{c_G(S_1) + c_G(S_2)}{k}\right) \end{aligned}$$

CORRECTNESS ($\Rightarrow$)

The MAX CUT–3 instance has a solution iff the CONDUCTANCE instance is solvable.

$A \subset V$ in G s.t. $c_G(A) \geq a$

$S^A \subset V'$ in G' s.t.



$$S^A = \{v^1 \in V_1 \,;\, v \in A\} \cup \{v^2 \in V_2 \,;\, v \in V \setminus A\}$$

CORRECTNESS $(\Rightarrow, cont.)$

Since $|S^A| = n$ and $c_G(A) = c_G(V \setminus A)$,

$$\Phi_{G'}\left(S^A\right) = \frac{1}{2n-4}\left(n - \frac{2c_G(A)}{n}\right) \leq \frac{1}{2n-4}\left(n - \frac{2a}{n}\right) = \phi,$$

$\Rightarrow S^A$ is a solution of the CONDUCTANCE instance

Correctness ($\Leftarrow$)

$\emptyset \neq S \subset V'$ in G' s.t. $\Phi_{G'}(S) \leq \phi$.

Let $A \subset V$ be a maximum cut in G.

$$\Phi_{G'}\left(S^A\right) \leq \Phi_{G'}(S)$$

$$\frac{1}{2n-4}\left(n - \frac{2c_G(A)}{n}\right) \leq \frac{1}{2n-4}\left(2n - k - \frac{c_G(S_1) + c_G(S_2)}{k}\right)$$

CORRECTNESS ($\Leftarrow$, CONT.)

A is a *maximum* cut in $G \Rightarrow 2c_G(A) \geq c_G(S_1) + c_G(S_2)$

$$\left(\tfrac{1}{n} - \tfrac{1}{k} \leq 0\right) \wedge \left(c_G(S_1) + c_G(S_2) \leq |S_1| \cdot n + |S_2| \cdot n = kn\right)$$

$$\Rightarrow \quad n - k + \left(\tfrac{1}{n} - \tfrac{1}{k}\right)\left(c_G(S_1) + c_G(S_2)\right) \geq 0 \quad \Rightarrow$$

$$\tfrac{1}{2n-4}\left(n - \tfrac{2c_G(A)}{n}\right) = \Phi_{G'}\left(S^A\right) \leq \Phi_{G'}(S) \leq \phi = \tfrac{1}{2n-4}\left(n - \tfrac{2a}{n}\right)$$

$\Rightarrow c_G(A) \geq a \Rightarrow A$ solves the MAX CUT–3 instance

DECISION PROBLEM: DENSITY

Maximum Density (DENSITY)
Instance: A graph $G = (V, E)$, an integer $0 < k \leq |V|$, and a rational number $0 \leq r \leq 1$.
Question: Is there a subset $S \subseteq V$ s.t. $|S| = k$ and the density of S in G is at least r ?

LOCAL DENSITY is **NP**-complete since this problem for $r = 1$ coincides with the **NP**-complete CLIQUE problem.

Theorem: RELATIVE DENSITY is **NP**-complete.

AN NP-COMPLETE PROBLEM: MIN BISECTION–3

Minimum Bisection for Cubic Graphs (MIN BISECTION–3)

Instance: A cubic graph $G = (V, E)$ with $n = |V|$ and an integer $a > 0$.

Question: Is there a cut $S \subset V$ s.t. $|S| = \frac{n}{2}$ and $c_G(S) \leq a$?

Reduction to RELATIVE DENSITY:

MIN BISECTION–3 instance: a cubic graph $G = (V, E)$ with $n = |V|$ and an integer $a > 0$

RELATIVE DENSITY instance: the same graph G with parameters $k = \frac{n}{2}$ and $r = \frac{3n-2a}{3n+2a}$

REDUCTION

For any $S \subset V$ s.t. $|S| = k = \frac{n}{2}$

$$|E(S)| = \frac{3|S| - c_G(S)}{2} = \frac{3n - 2c_G(S)}{4} \quad \text{(G cubic)}$$

$$\varrho_G(S) = \frac{3n - 2c_G(S)}{3n + 2c_G(S)} \quad \text{(by def.)}$$

$$\Rightarrow \varrho_G(S) \geq r \text{ iff } c_G(S) \leq a$$

DECISION PROBLEM: SINGLE CLUSTER EDITING

Minimum Single Cluster Editing (1–CLUSTER EDITING)
Instance: A graph $G = (V, E)$, integers $0 < k \leq |V|$ and $m > 0$.
Question: Is there a subset $S \subseteq V$ s.t. $|S| = k$ and $\varepsilon_G(S) \leq m$?

Theorem: 1–CLUSTER EDITING is **NP**-complete.

PROOF (AT A GLANCE)

1–CLUSTER EDITING belongs to **NP** (guess $S \subseteq V$ s.t. $|S| = k$ and verify $\varepsilon_G(S) \leq m$ in polym. time)

NP-hardness: MIN BISECTION–3 is reduced to 1–CLUSTER EDITING in polynomial time.

MIN BISECTION–3 instance: a cubic graph $G = (V, E)$ with $n = |V|$ and an integer $a > 0$

1–CLUSTER EDITING instance: the same graph G with parameters $k = \frac{n}{2}$ and $m = \frac{12a + n(n-8)}{8}$

Proof (cont.)

For any $S \subset V$ s.t. $|S| = k = \frac{n}{2}$

$$\begin{aligned} \varepsilon_G(S) &= \frac{|S| \cdot (|S| - 1)}{2} - \frac{3|S| - c_G(S)}{2} + c_G(S) \\ &= \frac{12c_G(S) + n(n-8)}{8} \end{aligned}$$

Combined with $|E(S)| = \dfrac{3|S| - c_G(S)}{2} = \dfrac{3n - 2c_G(S)}{4}$

$$\Rightarrow \varepsilon_G(S) \leq m \text{ iff } c_G(S) \leq a$$

CONCLUSIONS

We have presented **NP**-completeness proofs for the **decision problems** associated with the optimization of four possible graph cluster measures.

In clustering algorithms, *combinations* of fitness measures are recommended as only optimizing one may result in anomalies such as small cliques or sparse connected components as clusters.

Further work

An open problem is the complexity of such thresholding of the *product* of the local and relative densities (the sum of which closely related to the edge operation count for the single cluster editing problem).

Another important area for further research is the complexity of finding related **approximate solutions**.

THANK YOU FOR YOUR ATTENTION

Questions and comments are welcome both now and during the conference, as well as later on by e-mail.

elisa.schaeffer@tkk.fi

More info at http://www.tcs.hut.fi/~satu