

Maximizing the Cohesion is NP-hard

Adrien Friggeri, Eric Fleury

► **To cite this version:**

Adrien Friggeri, Eric Fleury. Maximizing the Cohesion is NP-hard. [Research Report] RR-7734, INRIA. 2011. <inria-00621065v2>

HAL Id: inria-00621065

<https://hal.inria.fr/inria-00621065v2>

Submitted on 9 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Maximizing the Cohesion is NP-hard

Adrien Friggeri — Eric Fleury

N° 7734 — version 2

initial version 8 September 2011 — revised version 3 October 2011

A large, light grey stylized 'R' logo is positioned to the left of the text 'Rapport de recherche'.

*Rapport
de recherche*

Maximizing the Cohesion is NP-hard

Adrien Friggeri , Eric Fleury

Thème : Réseaux et télécommunications
Équipe-Projet DNET

Rapport de recherche n° 7734 — version 2 — initial version 8 September 2011
— revised version 3 October 2011 — 8 pages

Abstract: We show that the problem of finding a set with maximum cohesion in an undirected network is **NP**-hard.

Key-words: social networks, complex networks, cohesion, np-complete, complexity

Maximiser la Cohésion est NP-dur

Résumé : Nous montrons que le problème de trouver un ensemble de cohésion maximum dans un graphe non orienté est **NP**-dur.

Mots-clés : réseaux sociaux, réseaux complexes, cohésion, np-complet, complexité

Introduction

In [1], we have introduced a new metric called the *cohesion* which rates the community-ness of a group of people in a social network from a sociological point of view. Through a large scale experiment on Facebook, we have established that the cohesion is highly correlated to the subjective user perception of the communities. In this article, we show that finding a set of vertices with maximum cohesion is **NP**-hard.

Notations

Let $G = (V, E)$ be a graph with vertex set V and edge set E of size $n = |V| \geq 4$. For all vertices $u \in V$, we write $d_G(u)$ the degree of u , or more simply $d(u)$ ¹. A *triangle* in G is a triplet of pairwise connected vertices.

For all sets of vertices $S \subseteq V$, let $G[S] = (S, E_S)$ be the subgraph induced by S on G . We write $m(S) = |E_S|$ the number of edges in $G[S]$, and $i(S) = |\{(u, v, w) \in S^3 : (uv, vw, uw) \in E^3\}|$ the number of triangles in $G[S]$. We define $o(S) = |\{(u, v, w), (u, v) \in S^2, w \in V \setminus S : (uv, vw, uw) \in E^3\}|$, the number of *outbound* triangles of S , that is: triangles in G which have exactly two vertices in S .

Moreover, for all (u, v) in E , let $\Delta(uv) = |\{w \in V : (uw, vw) \in E^2\}|$ be the number of triangles the edge uv belongs to in G .

Finally, we recall the definition of the cohesion of a set S in G :

$$\mathcal{C}(S) = \frac{i(S)^2}{\binom{|S|}{3}(i(S) + o(S))}$$

An example is given on Figure 1. The cohesion of a given set S in G can naively be computed in $\mathcal{O}(n^3)$ by listing all triangles in G and counting those inside and outbound to S .

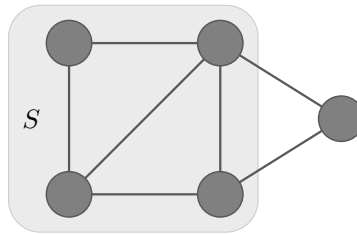


Figure 1: In this example, $i(S) = 2$ and $o(S) = 1$, thus $\mathcal{C}(S) = \frac{1}{6}$

In this article we examine the problem of finding a set of vertices $S \subseteq V$ of maximum cohesion, i.e. for all subset $S' \subseteq V$, $\mathcal{C}(S') \leq \mathcal{C}(S)$.

Outline

We now proceed to prove that finding a set of vertices with maximum cohesion in G is **NP**-hard. We will first show in Section 1 that this problem is equivalent

¹Here, as elsewhere, we drop the index referring to the underlying graph if the reference is clear.

to that of finding a connected set of vertices with maximum cohesion in G . The decision problem associated to the latter is **CONNECTED-COHESIVE**.

Then, we shall prove that **CONNECTED-COHESIVE** is **NP**-complete by reducing **CLIQUE** (problem GT19 in [2]). From there we deduce that the optimization problem of finding a set of vertices with maximum cohesion is **NP**-hard.

Problems

1. **CONNECTED-COHESIVE**:

Input A graph $G = (V, E)$, $\lambda \in \mathbb{Q}$, $\lambda \in [0, 1]$

Question Is there a subset connected S of V such that $\mathcal{C}(S) \geq \lambda$?

2. **CLIQUE**:

Input A graph $G = (V, E)$, $k \in \mathbb{N}$, $k \leq |V|$

Question Is there a subset S of V such that $|S| = k$ and the subgraph induced by S is a clique?

1 A maximum cohesive group is connected

In order to prove that a set of vertices with maximum cohesion in a given network is connected, we need the following lemma:

Lemma 1.1. *Let $S_1 \subseteq V$ and $S_2 \subseteq V$ be two disconnected sets of vertices ($(S_1 \times S_2) \cap E = \emptyset$). If $\mathcal{C}(S_1) \leq \mathcal{C}(S_1 \cup S_2)$ then $\mathcal{C}(S_2) > \mathcal{C}(S_1 \cup S_2)$.*

Proof. Suppose $\mathcal{C}(S_1) \leq \mathcal{C}(S_1 \cup S_2)$ and $\mathcal{C}(S_2) \leq \mathcal{C}(S_1 \cup S_2)$. Given that S_1 and S_2 are disconnected, $i(S_1 \cup S_2) = i(S_1) + i(S_2)$ and $o(S_1 \cup S_2) = o(S_1) + o(S_2)$. We can then write:

$$\frac{i(S_1)^2}{\binom{|S_1|}{3}} \leq (i(S_1) + o(S_1))\mathcal{C}(S_1 \cup S_2) \quad (1)$$

$$\frac{i(S_2)^2}{\binom{|S_2|}{3}} \leq (i(S_2) + o(S_2))\mathcal{C}(S_1 \cup S_2) \quad (2)$$

By summing (1) and (2), we obtain:

$$\begin{aligned} \frac{i(S_1)^2}{\binom{|S_1|}{3}} + \frac{i(S_2)^2}{\binom{|S_2|}{3}} &\leq (i(S_1) + o(S_1) + i(S_2) + o(S_2))\mathcal{C}(S_1 \cup S_2) \\ &\leq (i(S_1 \cup S_2) + o(S_1 \cup S_2))\mathcal{C}(S_1 \cup S_2) \\ &\leq \frac{(i(S_1) + i(S_2))^2}{\binom{|S_1| + |S_2|}{3}} \end{aligned}$$

Furthermore, given that $|S_1|, |S_2| > 1$,

$$\binom{|S_1|}{3} + \binom{|S_2|}{3} < \binom{|S_1| + |S_2|}{3}$$

We then have:

$$\frac{i(S_1)^2}{\binom{|S_1|}{3}} + \frac{i(S_2)^2}{\binom{|S_2|}{3}} < \frac{(i(S_1) + i(S_2))^2}{\binom{|S_1|}{3} + \binom{|S_2|}{3}}$$

Which simplifies to:

$$\left(\binom{|S_2|}{3} i(S_1) - \binom{|S_1|}{3} i(S_2) \right)^2 < 0$$

Hence the contradiction. Therefore, for all $S_1, S_2 \subseteq V$, disconnected:

$$\mathcal{C}(S_1) \leq \mathcal{C}(S_1 \cup S_2) \Rightarrow \mathcal{C}(S_2) > \mathcal{C}(S_1 \cup S_2) \quad \square$$

Theorem 1.2. *Let S be the set of vertices of G with the highest cohesion, S is connected.*

Proof. Suppose S is not connected, then there exist two disconnected subsets $S_1, S_2 \subseteq S$ such that $S = S_1 \cup S_2$. Given that S has maximum cohesion, we have $\mathcal{C}(S) \geq \mathcal{C}(S_1)$. Thus per Lemma 1.1: $\mathcal{C}(S) < \mathcal{C}(S_2)$ and S does not have the highest cohesion, hence the contradiction. \square

Corollary 1.3. *Per Theorem 1.2, the problem of searching for a set of vertices with maximum cohesion is strictly equivalent to that of searching a set of connected vertices with maximum cohesion.*

2 CONNECTED-COHESIVE is NP-complete

First note that given a set S of vertices of G , it is possible to verify that S is a solution of CONNECTED-COHESIVE by computing its cohesion, its size, its connectivity and the minimum degree of its vertices, all in polynomial time. Therefore CONNECTED-COHESIVE is in **NP**.

Algorithm 1 Transforms an instance of CLIQUE in an instance of CONNECTED-COHESIVE

Require: $G = (V, E), k \in \mathbb{N}$

- 1: $W := \emptyset$
 - 2: $E' := E$
 - 3: **for** $uv \in V^2 \setminus E$ **do**
 - 4: let K be a clique of size $2\binom{n}{3}^4$
 - 5: $W \leftarrow W \cup K$
 - 6: $E' \leftarrow E' \cup \{uv\} \cup (\{u, v\} \times K)$
 - 7: **end for**
 - 8: **return** $G' = (V \cup W, E'), \lambda = \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$
-

Let us now reduce CLIQUE to CONNECTED-COHESIVE. Let $(G = (V, E), k \in \mathbb{N})$ be an instance of CLIQUE². We can assume that G is connected (if not, we

²We consider here that $|G| > 2$ and $k > 2$, although this is not exactly CLIQUE, this problem is clearly **NP**-complete, given that the complexity of CLIQUE does not arise from those small values.

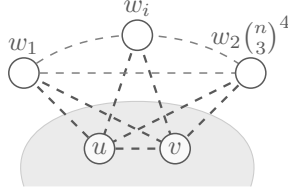


Figure 2: Illustration of Algorithm 1. At this step, we join u and v , add a clique of size $2\binom{n}{3}^4$ to the network, and join u and v to all vertices in the added clique.

use the following reasoning separately on each connected component of G). We construct an instance $(G' = (V', E'), \lambda)$ of CONNECTED-COHESIVE by adding an edge between all non connected vertices u and v in G and then linking those two vertices to all vertices in a clique of size $2\binom{n}{3}^4$ which we add to the network, as described in Algorithm 1 and illustrated by Figure 2.

Theorem 2.1. *There exist a clique of size k in G iff there exist a connected group of vertices of G' with cohesion $\lambda \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$.*

Proof. Let $K \subseteq V$, be a clique of size $|K| = k$ in G . Given that no node or edges are deleted when constructing G' , G is a subgraph of G' and thus K is a clique in G' and $i_{G'}(K) = \binom{k}{3}$.

Moreover, by construction, $G'[V]$ is a clique and for all u in K , the neighbors of u are also in V . Therefore, each edge in K forms one triangle with each vertex in $V \setminus K$, which leads to $o_{G'}(K) = \binom{k}{2}(n-k)$. Finally, this gives a cohesion:

$$\mathcal{C}_{G'}(K) = \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$$

Conversely, let $S \subseteq V'$ be a connected set of vertices such that $\mathcal{C}_{G'}(S) \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$. We will show that S is a clique of size larger than k and that $S \subseteq V$. First note that $|S| \geq 3$, because by definition, if $|S| < 3$, $\mathcal{C}_{G'}(S) = 0$ which would lead to a contradiction.

First, suppose that S is not a clique in G , then let us distinguish two cases:

1. If $S \subseteq V$ and S is not a clique, then S contains two vertices $u, v \in V^2$ such that $uv \notin E$.
2. If $S \not\subseteq V$, then $\exists u \in S \setminus V$, and S being connected, there exist $v \in V'$ such that $uv \notin E$.

Therefore, if S is not a clique in G , it contains an edge $uv \notin E$ and by construction, this edge belongs to at least $2\binom{n}{3}^4$ triangles, which leads to:

$$\begin{aligned} i_{G'}(S) + o_{G'}(S) &\geq K \\ \mathcal{C}_{G'}(S) &\leq \frac{i_{G'}(S)^2}{2\binom{|S|}{3}\binom{n}{3}^4} \\ &\leq \frac{1}{2\binom{n}{3}^2} \\ &< \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)} \end{aligned}$$

Hence the contradiction, therefore S must be a clique in G . From there it comes that:

$$\mathcal{C}_{G'}(S) = \frac{\binom{k'}{3}}{\binom{k'}{3} + \binom{k'}{2}(n-k')}$$

where $k' = |S|$. Therefore:

$$\begin{aligned} \mathcal{C}_{G'}(S) \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)} &\Leftrightarrow \frac{\binom{k'}{2}(n-k')}{\binom{k'}{3}} \leq \frac{\binom{k}{2}(n-k)}{\binom{k}{3}} \\ &\Leftrightarrow \frac{n-k'}{k'-3} \leq \frac{n-k}{k-3} \\ &\Leftrightarrow k' \geq k \end{aligned}$$

Therefore, we can now conclude that if there exist a connected set S in G' with cohesion $\mathcal{C}_{G'}(S) \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$, then S is a clique of size at least k in G , and thus there exist a clique $K \subseteq S$ of size k in G . \square

Theorem 2.2. CONNECTED-COHESIVE is NP-complete.

Proof. Per Theorem 2.1, there exist a clique of size k in G iff there exist a connected subset of vertices of G' of cohesion $\lambda \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$ and the transformation from G, k to G', λ runs in polynomial time. Thus CLIQUE is reducible to CONNECTED-COHESIVE and CONNECTED-COHESIVE is NP-hard.

Given that CONNECTED-COHESIVE is in NP, the problem is thus NP-complete. \square

3 Conclusion

The associated decision problem being NP-complete, the problem of finding a set of vertices with maximum cohesion is NP-hard³.

³Note that the problem of finding a set of vertices of maximum cohesion containing a set of predefined vertices is also NP-hard, by an immediate reduction

References

- [1] Adrien Friggeri, Guillaume Chelius, and Eric Fleury. Triangles to Capture Social Cohesion. In *Third IEEE International Conference on Social Computing*, Cambridge, United States, September 2011.
- [2] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman, San Francisco, 1979.



Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
<http://www.inria.fr>
ISSN 0249-6399