Maximizing the Cohesion is NP-hard

Adrien Friggeri — Eric Fleury

N° 7734 — version 2

initial version 8 September 2011 — revised version 3 October 2011
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, coésion, np-complet, compléxité
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 $\text{NP}$-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 $d(u)$. 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 $G$:

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

An example is given on Figure 1. The cohesion of a given set $S$ in $G$ can naively be computed in $O(n^3)$ 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 $C(S) = \frac{1}{4}$](image)

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

Outline

We now proceed to prove that finding a set of vertices with maximum cohesion in $G$ is $\text{NP}$-hard. We will first show in Section [1] that this problem is equivalent

\footnote{Here, 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:

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

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

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

\[
\frac{i(S_1)^2}{|S_1|^3} + \frac{i(S_2)^2}{|S_2|^3} \leq (i(S_1) + o(S_1) + i(S_2) + o(S_2))\mathcal{C}(S_1 \cup S_2)
\]

\[
\leq (i(S_1 \cup S_2) + o(S_1 \cup S_2))\mathcal{C}(S_1 \cup S_2)
\]

\[
\leq \frac{(i(S_1) + i(S_2))^2}{|S_1|^3 + |S_2|^3}
\]

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:

\[
\frac{i(S_1)^2}{|S_1|^3} + \frac{i(S_2)^2}{|S_2|^3} < \frac{(i(S_1) + i(S_2))^2}{(|S_1|^3 + |S_2|^3)}
\]

Which simplifies to:

\[
\left(\frac{|S_2|}{3}\right)i(S_1) - \left(\frac{|S_1|}{3}\right)i(S_2) < 0
\]

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

\[
C(S_1) \leq C(S_1 \cup S_2) \Rightarrow C(S_2) > C(S_1 \cup S_2)
\]

**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 disconnect subsets \(S_1, S_2 \subseteq S\) such that \(S = S_1 \cup S_2\). Given that \(S\) has maximum cohesion, we have \(C(S) \geq C(S_1)\). Thus per Lemma 1.1, \(C(S) < C(S_2)\) and \(S\) does not have the highest cohesion, hence the contradiction.

**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-COHESEIVE 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-COHESEIVE by computing its cohesion, its size, its connectivity and the minimum degree of its vertices, all in polynomial time. Therefore CONNECTED-COHESEIVE is in NP.

**Algorithm 1** Transforms an instance of CLIQUE in an instance of CONNECTED-COHESEIVE

| 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 \(2|V|^3\) |
| 5: \(W' \leftarrow W \cup K\) |
| 6: \(E' \leftarrow E' \cup \{uv\} \cup \{\{u, v\} \times K\}\) |
| 7: end for |
| 8: return \(G' = (V \cup W, E'), \lambda = \frac{\binom{t}{3}}{\binom{t}{2} + \binom{u-k}{2}}\) |

Let us now reduce CLIQUE to CONNECTED-COHESEIVE. 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.
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 $2\binom{n}{3}$ 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:

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

Conversely, let $S \subseteq V'$ be a connected set of vertices such that $C_{G'}(S) \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n-k)}$. 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 $|S| < 3$, $C_{G'}(S) = 0$ 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 $uv \notin E$. 

\footnotesize

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

\normalsize

INRIA
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:

$$i_{G'}(S) + o_{G'}(S) \geq K$$

$$C_{G'}(S) \leq \frac{i_{G'}(S)^2}{2\binom{\binom{n}{3}^4}{3}}$$

$$\leq \frac{1}{2\binom{n}{3}^2}$$

$$< \frac{\binom{k}{3}}{\binom{n}{3} + \binom{k}{3}(n - k)}$$

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

$$C_{G'}(S) = \frac{\binom{k'}{3}^2}{\binom{k'}{3} + \binom{k'}{2}(n - k')$$

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

$$C_{G'}(S) \geq \frac{\binom{k}{3}}{\binom{k}{3} + \binom{k}{2}(n - k)} \Leftrightarrow \frac{(k')^2(n - k')}{\binom{k'}{3} + \binom{k'}{2}(n - k)} \leq \frac{(\binom{k}{3})^3}{\binom{k}{3}}$$

$$\Leftrightarrow \frac{n - k'}{k' - 3} \leq \frac{n - k}{k - 3}$$

$$\Leftrightarrow k' \geq k$$

Therefore, we can now conclude that if there exist a connected set $S$ in $G'$ with cohesion $C_{G'}(S) \geq \frac{(\binom{k}{3})^3}{\binom{k}{3} + \binom{k}{2}(n - k)}$, 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 $G$.

**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 $\lambda \geq \frac{(\binom{k}{3})^3}{\binom{k}{3} + \binom{k}{2}(n - k)}$ and the transformation from $G, k$ to $G', \lambda$ 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.

3 Conclusion

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

---

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.
