Algorithmique des réseaux sociaux **2013/2014**

Cours 3 — 29 Octobre

Enseignant: Marc Lelarge *Scribe: Baptiste Lefebvre*

Pour information

- Page web du cours
  `http://www.di.ens.fr/~lelarge/soc.html`

3.1 Détection de communautés

3.1.1 Motivation

On définit la notion de communauté au travers de la notion d''homophilie' (homophily en anglais), par exemple, pour le Web, les pages de sujets similaires sont fortement connectées. Une définition alternative issue des sciences sociales se concentre d'avantage sur la notion de "liens faibles" (ou "weak ties") introduite par Mark Granovetter.

Remarque 3.1.1 *Dans les travaux de Mark Granovetter, les communautés sont considérées comme les lieux où l'information est homogène et les liens faibles comme les responsables de la diffusion de l'information entre les communautés.*

Définition 3.1.1 *Soit $G = (V, E)$ un graphe. On définit les quantités suivantes :*

- $e(S, T) = E \cap (S \times T)$
  ensemble des arêtes avec une extrémité dans S et une dans T
- $e(S) = e(S, S)$
  ensemble des arêtes internes à S
- $d_s(v) = |e(\{v\}, S)|$
  degré de v dans l'ensemble S
- $D(S) = \frac{|e(S)|}{|S|}$
  densité d'arêtes d'un ensemble de sommets S

Pour préciser notre modèle de communauté, nous pouvons nous poser les questions suivantes :

1. Est-ce qu'un noeud peut appartenir à plusieurs communautés ?
2. Est-ce que chaque noeud doit être dans au moins une communauté ?

Dans une première partie de ce cours, nous verrons des modèles où les réponses seront respectivement négative et affirmative ce qui rapproche le problème de celui de la détection de sous-graphes denses. Dans une seconde partie, nous étudierons des modèles où les réponses seront affirmative et négative avec un rapprochement du problème de partitionnement de graphe.

3.1.2 Sous-graphes denses

Le problème est de trouver dans un graphe G le sous-graphe S maximisant $D(S)$. Le problème de décision associé est : étant donné un paramètre α, déterminer s'il existe un sous-graphe S tel que $D(S) \geq \alpha$. Ce probléme de décision peut se ramener à un problème de recherche de coupe minimale, en effet on sait que :

$$\sum_{v \in S} d(v) = 2|e(S)| + \left|e\left(S, \overline{S}\right)\right|$$

par conséquent $|e(S)| \geq \alpha |S|$ se traduit en :

$$\begin{aligned} \sum_{v \in S} d(v) - \left|e\left(S, \overline{S}\right)\right| - 2\alpha |S| &\geq 0 \\ \sum_{v \in V} d(v) - \left(\sum_{v \in \overline{S}} d(v) + \left|e\left(S, \overline{S}\right)\right| + 2\alpha |S|\right) &\geq 0 \\ 2|E| - \beta(S) &\geq 0 \end{aligned}$$

Le problème se réduit donc à minimiser $\beta(S) := \sum_{v \in \overline{S}} d(v) + \left|e\left(S, \overline{S}\right)\right| + 2\alpha |S|$.

3.1.3 Problème de coupe minimale

Nous introduisons maintenant un graphe G' tel que $\beta(S)$ corresponde à la capacité d'une coupe $(S, \overline{S})$.

Soit $G' = (V', E', c')$ tel que :

$$\begin{aligned} V' &= V \cup \{s, t\} \\ E' &= E \cup (\{s\} \times V) \cup (V \times \{t\}) \\ c'(e') &= \begin{cases} 1 & \text{si } e' \in E \\ d(v) & \text{si } e' = (s, v) \in (\{s\} \times V) \\ 2\alpha & \text{si } e' \in (V \times \{t\}) \end{cases} \end{aligned}$$

En résumé G' s'obtient en considérant que :
- la capacité de $e \in E$ dans le graphe initial est 1 ;
- s est connecté à tous les noeuds $v \in V$ avec capacité $d(v)$, i.e. le degré dans le graphe originel ;
- t est connecté à tous les noeuds $v \in V$ avec capacité 2α

La capacité de la coupe $\left(\{s\} \cup S, \overline{S} \cup \{t\}\right)$ dans G' vaut alors $\beta(S)$.

Proposition 3.1.1 *Si la coupe minimale* $\left(\{s\} \cup S^*, \overline{S^*} \cup \{t\}\right)$ *satisfait* $2|E| \geq \beta(S^*)$ *alors* S^* *est un ensemble de densité* $\geq \alpha$ *sinon aucun tel ensemble n'existe.*

Démonstration. Il suffit de vérifier que $\beta(S)$ est bien la capacité de la coupe $S, \overline{S}$). □

Remarque 3.1.2 *Si on recherche la valeur minimale de α, la première solution est d'exécuter l'algorithme sur des valeurs différentes de α tout en effectuant une recherche par dichotomie. La seconde solution, plus appropriée, correspond au théorème suivant.*

Théorème 3.1.1 *(Gallo, Grigoriadis, Tarjan '89) Soit $\alpha \in \mathbf{R}$, $G = (V, E)$ ayant des capacités $c_e \geq 0$ qui dépendent de α.*

- *capacités des arêtes sortantes de s croissante avec α*
- *capacités des arêtes entrantes de t décroissante avec α*
- *capacités des autres arêtes constante*

alors les flots maximum / coupes minimales pour toutes les valeurs $\alpha_1 \leq \cdots \leq \alpha_l$ peuvent être calculés en temps $O\left(n^2\left(l+m\right)\right)$ et les coupes minimales $\left(S_\alpha, \overline{S_\alpha}\right)$ sont tel que $S_\alpha \subseteq S_{\alpha'}$ si $\alpha \leq \alpha'$.

Remarque 3.1.3 *Sur les réseaux de moins de 1000 voire 10000 noeuds, un temps en $O\left(n^2\left(l+m\right)\right)$ n'est pas prohibitif. En pratique les réseaux intéressants sont beaucoup plus grands, cet algorithme n'est donc pas utilisé.*

Remarque 3.1.4 *Trouver le sous-graphe le plus dense de taille au moins / exactement k noeuds est NP-difficile. La preuve se fait par réduction au problème de recherche de clique de taille k dans un graphe.*

Une extension naturelle du problème précédent est : étant donné G, trouver l'ensemble S le plus dense contenant $X \subseteq V$, c'est-à-dire :

$$max\left(\frac{|e\left(S\right)|}{|S|}\right) \quad \text{tel que } X \subseteq S$$

Ce problème se ramène au précédent en ajoutant des arêtes de capacité infinie (très grande) entre la source s et les sommets $v \in X$.

3.1.4 Une $\frac{1}{2}$-approximation

Le remplacement de la recherche d'une solution exacte par celle d'une solution approchée permet de dépasser la difficulté inhérente au problème. Comme souvent un algorithme glouton atteint cet objectif.

Soit S^* l'ensemble maximisant $D\left(S\right)$, l'ensemble $\hat{S}$ est une $\frac{1}{2}$-approximation s'il vérifie :

$$\frac{\left|e\left(\hat{S}\right)\right|}{\left|\hat{S}\right|} \quad \geq \quad \frac{1}{2}\frac{|e\left(S^*\right)|}{|S^*|}$$

L'algorithme est le suivant :

Données : $G = (V, E)$
Résultat : sous-graphe le plus dense parmi $G_n, \ldots, G_{|X|}$
$G_n \leftarrow G$;
for $k \leftarrow n$ **downto** $|X| + 1$ **do**
  soit $v \notin X$ de degré minimal dans $G_k \backslash X$;
  $G_{k-1} \leftarrow G_k \backslash \{v\}$;
end

Remarque 3.1.5 *Cet algorithme peut être implémenté en temps $O(n^2)$ pour un graphe avec n noeuds et m arêtes.*

Proposition 3.1.2 *Cet algorithme est une $\frac{1}{2}$-approximation.*

Démonstration. Soit S le sous-graphe le plus dense de G. Si l'algorithme ne retourne pas S alors à un moment de son exécution l'algorithme a supprimé un sommet $v \in S$. Soit G_k le graphe juste avant que le premier $v \in S$ soit retiré. Comme S est optimal,

$$\begin{aligned} \frac{|e(S)|}{|S|} &\geq \frac{|e(S-v)|}{|S-v|} \\ &= \frac{|e(S)| - d_S(v)}{|S|-1} \end{aligned}$$

donc :

$$d_S(v) \geq \frac{|e(S)|}{|S|}$$

comme G_k contient S :

$$d_{G_k}(v) \geq d_S(v)$$

par définition du choix de v, on a :

$$d_{G_k}(u) \geq d_{G_k}(v) \quad \forall u \in G_k \backslash X$$

on cherche alors à minorer la densité de G_k.

$$\begin{aligned} \frac{|e(G_k)|}{|G_k|} &\geq \frac{\sum_{u \in S} d_S(u) + \sum_{u \in G_k \backslash S} \frac{|e(S)|}{|S|}}{2|G_k|} \\ &= \frac{2|e(S)| + |G_k \backslash S| \frac{|e(S)|}{|S|}}{2|G_k|} \\ &\geq \frac{|e(S)|}{|S|} \left(\frac{|S| + |G_k \backslash S|}{2|G_k|} \right) \\ &= \frac{1}{2} \frac{|e(S)|}{|S|} \end{aligned}$$

□

3.1.5 Communautés fortes

Les algorithmes précédents ont le défaut d'inclure systématiquement les noeuds qui possédent un fort degré. Pour y remédier la notion de communauté doit être redéfinie avec plus d'attention sur ce qui sépare les comunautés les unes des autres.

Cette partie se base sur le postulat que les communautés doivent être séparées.

Définition 3.1.2 *Soit un graphe G et $\alpha \in [0;1]$. Un ensemble $S \subseteq G$ est appelé α-communauté si et seulement si :*

$$d_s(v) \geq \alpha d(v) \quad \forall v \in S$$

Remarque 3.1.6 *En anglais on parle de "strong α-community" et par défaut de "strong community" pour le cas $\alpha = \frac{1}{2}$.*

Remarque 3.1.7 *Pour $\alpha = 1$ on obtient les composantes connexes du graphe (i.e.le graphe dans la plupart des cas).*

Nous admetrons :

Théorème 3.1.2 *Il est NP-complet de décider si un graphe a une α-communauté de taille $\geq k$.*

Soit $\left(S, \overline{S}\right)$ une coupe minimale $s-t$ de G alors S est 'presque' une $\frac{1}{2}$-communauté. En effet si $d_S(v) < \frac{1}{2}d(v)$ alors en déplaçant v de S à $\overline{S}$ on diminue la coupe donc :

$$d_S(v) \quad \geq \quad \frac{1}{2}d(v) \quad \forall v \notin \{s,t\}$$

Remarque 3.1.8 *Calcul à effectuer pour tout couple (s,t).*

Théorème 3.1.3 *(Gomory-Hu tree '61) Soit $G = (V,E)$. Pour toute paire $(i,j) \in V \times V$ soit $f_{i,j}$ le flot maximum / coupe minimale entre i et j. Soit G' le graphe complet sur V dont les coûts des arêtes correspondent à $f_{i,j}$. Soit T un arbre couvrant maximal de G. Pour chaque $(i,j) \in V \times V$, soit $P_{i,j}$ le chemin entre i et j dans T alors l'arbre T a la propriété :*

$$f_{i,j} \quad = \quad \min_{e \in P_{i,j}} (f_e) \quad \forall i,j$$

Si e est l'arête atteignant le minimum alors les deux composantes de $T \backslash \{e\}$ définissent une coupe minimale entre i et j dans G.

Démonstration. On montre que :

$$f_{i,j} \quad = \quad \min_{e \in P_{i,j}} (f_e) \quad \forall i,j$$

Si :

$$f_{i,j} > \min_{e \in P_{i,j}} (f_e)$$

on peut construire un nouvel arbre couvrant en retirant l'arête e et en ajoutant l'arête (i, j) qui contredit le fait que T soit un arbre couvrant maximal.

Si :

$$f_{i,j} < \min_{e \in P_{i,j}} (f_e)$$

$\left(S, \overline{S}\right)$ coupe entre i et j de capacité $f_{i,j}$, $e = (u, v) \in P_{i,j}$ avec $u \in S$ et $v \in \overline{S}$. $\left(S, \overline{S}\right)$ est aussi une coupe minimale entre u et v donc $f_{u,v} \leq f_{i,j}$, ce qui est contradictoire. □

Remarque 3.1.9 *A priori, parmi les n^2 couples de sommets il n'y a que $n - 1$ valeurs différentes à mémoriser / calculer. Le théorème 3.1.3 est à l'origine d'un gain d'un facteur n.*

Remarque 3.1.10 *Le cas des "strong communities" ($\alpha = \frac{1}{2}$) a été étudié dans ce cours, en ce qui concerne l'étude du cas général (α quelconque) la marche à suivre est de se ramener au cas $\alpha = \frac{1}{2}$ en jouant sur l'augmentation de la taille du graphe et le choix des capacités des arêtes.*

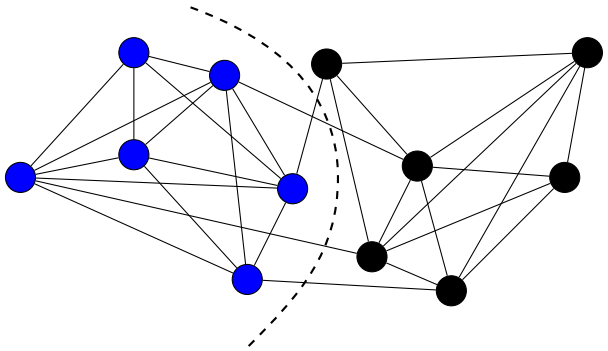

FIGURE 3.1. Illustration de la détection d'un sous-graphe dense

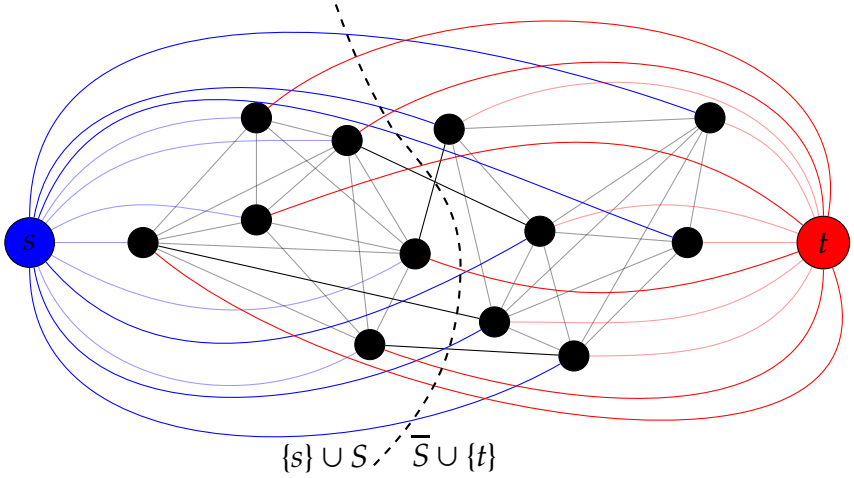

FIGURE 3.2. Illustration de la réduction au probème de coupe minimale