Extraction de communautés ego-centrées par apprentissage supervisé d'espaces prétopologiques.

Gaëtan Caillaut, Guillaume Cleuziou, Nicolas Dugué

▶ **To cite this version:**

Gaëtan Caillaut, Guillaume Cleuziou, Nicolas Dugué. Extraction de communautés ego-centrées par apprentissage supervisé d'espaces prétopologiques.. 19èmes Journées Extraction et Gestion des Connaissances (EGC 2019), Jan 2019, Metz, France. pp.117-128. hal-02003133

HAL Id: hal-02003133
https://hal.science/hal-02003133

Submitted on 30 Mar 2022

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.

Extraction de communautés ego-centrées par apprentissage supervisé d'espaces prétopologiques

Gaëtan Caillaut*, Guillaume Cleuziou*, Nicolas Dugué**

*Université d'Orléans, INSA Centre Val de Loire, LIFO EA 4022, Orléans, France
prénom.nom@univ-orleans.fr
**Le Mans Université, LIUM, EA 4023, Le Mans, France
prénom.nom@univ-lemans.fr

Résumé. Nous proposons une méthode d'extraction de communautés ego-centrées reposant sur l'apprentissage d'un modèle de propagation prétopologique. Là où les méthodes classiques ne considèrent souvent qu'un aspect de la structuration du réseau pour en extraire ses communautés, la prétopologie permet une analyse multi-critères du réseau. Notre démarche consiste à apprendre de façon supervisée un espace prétopologique défini par une combinaison logique de descripteurs du réseau. Une communauté locale à chaque nœud peut alors être extraite par une opération définie sur l'espace prétopologique appris.La qualité de chaque communauté locale est ensuite évaluée selon une communauté de référence. Nous avons comparé notre approche aux approches existantes sur des réseaux synthétiques et du réel et montrons ainsi sa pertinence.

1 Introduction

La théorie des réseaux complexes a mis en avant l'existence de propriétés communes aux réseaux modélisant des systèmes réels. En particulier, la plupart de ces réseaux possèdent une structure communautaire, i.e. une partition de l'ensemble des nœuds telle que les nœuds de chaque partie sont plus connectés entre eux qu'avec l'extérieur (Newman, 2006). Un cas typique est celui des réseaux sociaux, où les utilisateurs se regroupent autour de thèmes. La structure de communautés est particulièrement importante pour l'étude du réseau puisqu'elle permet de se placer à un niveau intermédiaire (mésoscopique) entre le niveau local (voisinage uniquement) et le niveau global (la totalité du réseau).

Une façon courante d'extraire les communautés d'un réseau consiste à trouver une partition de ses nœuds maximisant la modularité (Newman, 2006), c'est à dire qui maximise la densité des liens au sein des communautés en minimisant le nombre de liens entre communautés. Il est dans ce cas nécessaire de connaître la totalité du réseau afin d'en déterminer ses communautés. Dans le cas des très grands réseaux tels que l'internet ou les réseaux sociaux en ligne, cette condition est parfois impossible à remplir, soit parce qu'on ne connaît pas le réseau complet, soit parce qu'il est difficile de le stocker en mémoire. De plus, définir l'ensemble des communautés d'un réseau comme une partition stricte de ses nœuds est souvent éloigné de la réalité, puisque cela empêche un nœud d'appartenir à plusieurs communautés (Palla et al., 2005).

Ainsi, nous nous concentrons sur la notion de communauté locale à un (ou plusieurs) nœud(s) d'intérêt. On cherche dans ce cas à détecter la ou les communautés de ce nœud, on parle de communautés *ego-centrées* (Chen et al., 2009; Danisch et al., 2013). Cette approche permet de faire une optimisation locale, bien moins gourmande que l'approche globale, et réalisable même dans le cas où la totalité du graphe n'est pas connue. Par ailleurs, cette approche est adaptée à la mise en évidence de communautés chevauchantes, aboutissant ainsi à des résultats plus en accord avec la réalité.

Cet article a pour objectif de présenter une méthode prétopologique d'extraction de communautés locales. La théorie de la prétopologie est une généralisation de la théorie des graphes (Dalud-Vincent, 2017) et est de ce fait particulièrement adaptée à l'étude et à la modélisation de réseaux. Elle permet notamment de représenter des relations de natures différentes entre ensembles d'éléments, là où un graphe ne décrit des relations qu'entre paires d'éléments. Il est par exemple possible d'imaginer une prétopologie sociale définie à partir de différents types de relations entre ses utilisateurs (amis, collègues, familles, ...).

Après une présentation de travaux liés à l'extraction de communautés (Section 2), nous introduisons (Section 3) les concepts clés de la prétopologie ainsi que la classe (générique) des espaces prétopologiques définis par une fonction d'adhérence logique. Cette dernière formalisation offre la possibilité d'apprendre un espace prétopologique adapté aux caractéristiques du réseau. La Section 4 établit le cadre de nos expérimentations et décrit les prédicats qui composent nos règles logiques. La Section 5 expose les résultats obtenus sur quatre jeux de données réels et synthétiques. Une comparaison entre les méthodes classiques et les méthodes prétopologiques est faite.

2 Travaux connexes

De nombreux travaux se sont déjà attelés à la tâche de détection de communautés ego-centrées. On peut en distinguer au moins trois types : les méthodes guidées par une mesure locale inspirée de la modularité (Newman, 2006), celles basées sur des algorithmes de propagations ou encore des méthodes reposant sur des algorithmes d'apprentissage de plongements de graphes (Grover et Leskovec, 2016).

Les méthodes guidées par une mesure de modularité proposent de construire une communauté ego-centrée en ajoutant successivement des nœuds à un ensemble initial de nœuds d'intérêts. La fonction objectif de ces méthodes est une variante de la modularité adaptée au cas des communautés locales. À chaque itération de l'algorithme, le nœud apportant le plus grand gain au score de modularité locale est inséré dans la communauté. L'algorithme s'arrête quand il n'est plus possible d'améliorer le critère ou lorsque la communauté détectée est suffisamment grande (Clauset, 2005). Certaines méthodes ajoutent une étape d'élagage afin de corriger de potentielles erreurs (Chen et al., 2009; Luo et al., 2008).

Par ailleurs, Danisch et al. (2013) définissent une méthode inspirée de la propagation de l'opinion ou de la chaleur dans un graphe. Si l'on considère le nœud d'intérêt comme une source de chaleur, cette chaleur se transmet en suivant les liens du réseau, aboutissant à un score de température pour chaque nœud indiquant la proximité entre le nœud d'intérêt et le reste du réseau. Une communauté ego-centrée peut alors être extraite en ne conservant que les nœuds dont le score dépasse un seuil fixé. Cette méthode offre la possibilité d'extraire des communautés à différents niveaux de granularité.

Enfin, certains travaux démontrent la pertinence de considérer des méthodes d'apprentissage de plongements lexicaux dans le contexte de l'étude des graphes (Figueiredo et al., 2017; Grover et Leskovec, 2016). Pour faire une analogie, si l'on considère qu'un sommet représente un mot, on peut générer des phrases décrivant les chemins empruntés par des marches aléatoires. Ces phrases permettent alors l'apprentissage de plongements pour chaque nœud du réseau, plongements ensuite utilisés par des méthodes d'apprentissage pour en extraire des communautés.

On observe ainsi que de nombreuses approches ont été proposées pour résoudre le problème de détection de communautés. Il est probable qu'il n'existe pas de méthode générique pour réaliser cette tâche de façon optimale. Chaque méthode permet toutefois d'extraire une portion d'information qu'il serait regrettable de ne pas exploiter. C'est pourquoi nous proposons une méthode alternative, basée sur la théorie de la prétopologie, permettant de tirer profit des qualités des différentes approches en les combinant. La prétopologie permet de décrire un processus d'expansion à partir d'une combinaison de plusieurs "sources d'informations". C'est par ce processus d'expansion que nous suggérons d'extraire les communautés d'un réseau.

3 Éléments de Prétopologie

Un espace prétopologique est défini par un couple (E, a) avec E un ensemble fini non-vide d'éléments et $a : \mathcal{P}(E) \to \mathcal{P}(E)$ sa fonction d'adhérence vérifiant les propriétés 1 et 2.

$$\forall A \in \mathcal{P}(E), A \subseteq a(A) \tag{1}$$

$$a(\emptyset) = \emptyset \tag{2}$$

L'opérateur d'adhérence modélise ainsi un processus d'expansion d'une partie A de E. Il est usuellement défini par un ensemble $\mathcal{V}$ de voisinages sur E (Belmandt, 1993) où $V \in \mathcal{V}$ est une application réflexive de E vers $\mathcal{P}(E)$.

$$\forall A \in \mathcal{P}(E),\ a(A) = \{x \in E \mid \forall V \in \mathcal{V},\ V(x) \cap A \neq \emptyset\} \tag{3}$$

Contrairement aux opérateurs de la topologie, l'opérateur d'adhérence prétopologique n'est pas nécessairement idempotent, on peut alors l'appliquer de façon successive sur un ensemble $A \in \mathcal{P}(E)$ jusqu'à obtenir un ensemble K tel que $A \subseteq K \subseteq E$ et $a(K) = K$. On appelle K le fermé de A et on le note $F(A)$. Si $|A| = 1$ alors on appelle $F(A)$ un fermé élémentaire.

Caillaut et Cleuziou (2018) introduisent une nouvelle classe d'espaces prétopologiques dont la fonction d'adhérence est définie par une formule logique Q en forme normale disjonctive (DNF).

$$\forall A \in \mathcal{P}(E),\ a_Q(A) = \{x \in E \mid Q(A, x)\} \tag{4}$$

Cette définition de l'opérateur d'adhérence possède d'une part l'avantage d'être plus générale que la définition précédente, mais surtout permet d'envisager l'apprentissage de règles de combinaison de voisinages, et donc d'espaces prétopologiques. Cleuziou et Dias (2015) proposent la méthode LPS (*Learning Pretopological Spaces*) qui consiste à apprendre une fonction numérique pour définir l'opérateur d'adhérence. Cette fonction impose quelques restrictions, notamment le fait d'être nécessairement linéaire. C'est pourquoi Caillaut et Cleuziou (2018) proposent la méthode LPSMI (*Learning Pretopological Spaces Multi-Instance*) consistant en l'apprentissage d'une règle logique, plus souple qu'un modèle linéaire.

Ces méthodes proposent d'apprendre un espace prétopologique en se basant sur ses fermés élémentaires. Étant donné un ensemble S^* de fermés élémentaires cibles et une liste de prédicats, LPSMI apprend une DNF Q composée des prédicats donnés en entrée et telle que les fermés élémentaires S^* puissent être obtenus par la fonction d'adhérence logique $a_Q(.)$.

Nous proposons d'appliquer LPSMI pour apprendre un espace prétopologique dont les fermés correspondraient à des communautés locales. Cependant, LPSMI utilise une fonction d'optimisation spécialisée pour l'apprentissage d'espaces prétopologiques de type V. Ces espaces sont définis par une fonction d'adhérence possédant la propriété d'isotonie.

$$\forall A, B \in \mathcal{P}(E),\ A \subseteq B \Rightarrow a(A) \subseteq a(B) \tag{5}$$

Un espace de type V impose une contrainte forte sur la façon dont les communautés peuvent se chevaucher [1]. Soient trois points $x, y, z \in E$, si $y \in F(\{x\})$ et $y \in F(\{z\})$, alors $F(\{y\}) \subseteq F(\{x\}) \cap F(\{z\})$ [2]. Par conséquent, la communauté centrée sur y sera, selon les cas, comprimée entre celle de x et celle z, ou alors elle imposera la présence de nœuds indésirables dans les communautés centrées sur x et sur z. Ce formalisme interdit donc de nombreuses formes de structures qui se retrouvent pourtant dans des cas réels.

C'est pourquoi nous prétendons qu'une prétopologie de type V n'est probablement pas adaptée à la modélisation des communautés à partir de fermés élémentaires. Nous proposons alors de recourir à une variante de LPSMI plus simple et qui n'impose pas ces contraintes de formes sur les communautés.

4 Méthode d'extraction de communautés

L'algorithme d'apprentissage LPSMI reprend le principe de l'apprentissage de concepts et consiste à construire une formule logique Q en forme normale disjonctive de manière gloutonne, c'est-à-dire par ajouts successifs de littéraux. La construction de la règle logique est guidée par un critère objectif exploitant la propriété d'isotonie des espaces de type V. Dans le but de s'affranchir de cette contrainte nous proposons la méthode LPSFM dont la seule différence avec LPSMI est son critère objectif. La construction d'une règle logique par LPSFM est guidée par le score de F-mesure entre les communautés réelles (cibles) et les communautés prétopologiques obtenues. Nous noterons *critère MI* le critère utilisé par LPSMI et *critère FM* celui utilisé par LPSFM. Nous détaillons dans cette section les différences entre ces deux critères ainsi que les prédicats utilisés pour l'apprentissage des règles logiques.

4.1 Fonctions objectives pour l'apprentissage

Étant donné un ensemble de fermés élémentaires cibles S^* et un espace prétopologique (E, a), les deux critères MI et FM proposent d'évaluer la qualité de l'espace prétopologique en mesurant la correspondance entre ses fermés élémentaires et S^*. Les critères MI et FM partagent donc le même objectif, ils sont toutefois fondamentalement différents.

Le critère MI s'appuie sur les propriétés structurelles des espaces prétopologiques de type V (Éq. 5) pour évaluer finement non seulement la *qualité* d'un prédicat au regard des fermés

1. Si on considère qu'un fermé élémentaire exprime une communauté ego-centrée.
2. Car $F(\{y\}) \subseteq F(\{x\})$ et $F(\{y\}) \subseteq F(\{z\})$.

élémentaires qu'il génère mais aussi son *potentiel* à travers ses fermés non-élémentaires. Il en résulte que l'espace appris sera nécessairement de type V [3]. Il est donc primordial que les prédicats composant la DNF Q respectent les propriétés des espaces de type V, c'est à dire lorsqu'un ensemble $A \in \mathcal{P}(E)$ se propage par l'opérateur d'adhérence à un élément $x \in E$, tout sur-ensemble de A doit aussi se propager à x. Soit q un prédicat défini sur $\mathcal{P}(E) \times E$, q est de type V s'il respecte :

$$\forall A, B \in \mathcal{P}(E), x \in E\ ,\ A \subseteq B \Rightarrow [q(A, x) \Rightarrow q(B, x)] \tag{6}$$

Le critère FM est quant à lui beaucoup plus simple puisqu'il ne tient pas compte du *potentiel* d'un prédicat, il ne s'appuie que sur les fermés élémentaires générés. Caillaut et Cleuziou (2018) montrent qu'en pratique ce critère est moins efficace que le critère MI lorsque la tâche consiste spécifiquement à apprendre des espaces de type V. Il reste cependant un recours précieux pour guider l'apprentissage d'espaces prétopologiques non contraints (non V). Toute DNF qui induit une fonction d'adhérence respectant les deux propriétés décrites par les équations 1 et 2 est alors autorisée. Les seules propriétés à satisfaire pour les prédicats considérés sont alors les propriétés 7 et 8 suivantes :

$$\forall A \in \mathcal{P}(E),\ \forall x \in A, q(A, x) = 1 \tag{7}$$

$$\forall x \in E, q(\emptyset, x) = 0 \tag{8}$$

4.2 Construction des prédicats/descripteurs d'un réseau

Nous proposons un ensemble de prédicats spécifiquement dédiés à la tâche d'extraction de communautés ego-centrées. Chaque prédicat peut être vu comme un descripteur, il permet de capturer une caractéristique du réseau. Dans la suite, nous notons E l'ensemble des éléments du réseau (qui peut ne pas être connu en totalité), A un sous-ensemble de E et x un élément de E. L'ensemble des prédicats que nous proposons se décompose en trois catégories décrites ci-après. La diversité de cet ensemble de prédicat est un bel exemple illustrant la capacité d'analyse multi-critères offerte par le formalisme prétopologique.

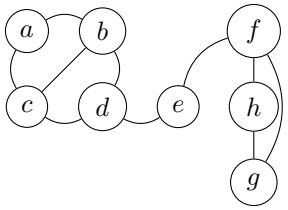
Les prédicats topologiques. Soient $V(x)$ les voisins du nœud x dans le réseau et $V(A) = \bigcup_{x \in A} V(x)$ l'union des voisinages de chaque élément de A. Nous considérons que ces voisinages sont réflexifs, tels que $x \in V(x)$ (et par conséquent $A \subseteq V(A)$).

Un premier prédicat de base est défini à partir de la matrice d'adjacence du réseau. On le note $q_{adj}(A, x)$ et il est vrai lorsqu'un élément de A est connecté à x.

$$q_{adj}(A, x) = x \in V(A) \tag{9}$$

Nous proposons quatre prédicats supplémentaires définis par les voisinages de A et de x permettant de capturer différentes variantes d'intensités d'interactions entre A et x.

3. Les fermés cibles S^* peuvent cependant ne pas correspondre à ceux d'un espace de type V. Les fermés de l'espace prétopologiques résultant ne seront donc jamais similaires à S^*.



(a) Un réseau possédant deux communautés.

	a	b	c	d	e	f	g	h
a	1,00	0,77	0,77	0,59	0,30	0,07	0,00	0,00
b	0,91	1,00	0,86	0,70	0,36	0,07	0,00	0,00
c	0,85	0,79	1,00	0,63	0,30	0,05	0,00	0,00
d	0,68	0,76	0,76	1,00	0,53	0,13	0,00	0,00
e	0,00	0,06	0,06	0,33	1,00	0,44	0,26	0,26
f	0,00	0,04	0,04	0,19	0,58	1,00	1,00	1,00
g	0,00	0,03	0,03	0,14	0,43	0,76	1,00	0,88
h	0,00	0,03	0,03	0,12	0,40	0,73	0,85	1,00

(b) Matrice de proximités *carryover-opinion*.

FIG. 1: Exemple

— $q_{r1}(A,x,k) = \frac{|A \cap V(x)|}{|A|} \geq k$

— $q_{r2}(A,x,k) = \frac{|A \cap V(x)|}{|V(x)|} \geq k$

— $q_{r3}(A,x,k) = \frac{|A \cap V(x)|}{|A \cup V(x)|} \geq k$

— $q_{r4}(A,x,k) = \frac{|V(A) \cap V(x)|}{|V(A) \cup V(x)|} \geq k$

avec k un paramètre de seuil dans $[0,1]$. Parmi ces quatre prédicats, on peut montrer que seul q_{r2} est un prédicat de type V.

Prédicats basés sur la modularité. Trois prédicats sont construits autour des définitions de modularité locale de Clauset (2005), Luo et al. (2008) et Chen et al. (2009) respectivement. On note ces prédicats $q_X(A,x)$ avec $X \in \{clauset, luo, chen\}$. Le prédicat $q_X(A,x)$ est vrai lorsque l'ajout de x à la communauté A améliore sa modularité, notée $mod_X(A)$.

$$\forall A \in \mathcal{P}(E),\, \forall x \in E\,,\; q_X(A,x) = mod_X(A \cup \{x\}) > mod_X(A) \tag{10}$$

Ces prédicats ne respectent pas les propriétés des espaces de type V et ne peuvent donc pas être exploités par l'approche LPSMI.

Prédicats définis par une mesure de proximité. Le prédicat $q_{danisch}(A,x,k)$ est défini à partir de la mesure de proximité *carryover-opinion* (Danisch et al., 2013). Ce prédicat est vrai lorsqu'il existe un élément de A dont la proximité avec x est supérieure à un seuil k dans $[0,1]$.

$$\forall A \in \mathcal{P}(E), \forall x \in E\,,\; q_{danisch}(A,x,k) = \max_{y \in A} \{carryover(x,y)\} \geq k \tag{11}$$

D'autres prédicats pourraient être envisagés, à partir des approches récentes de représentations vectorielles des nœuds d'un graphe (e.g. node2vec). Cependant celles-ci nécessitent de connaître l'intégralité du réseau et ne tiennent pas compte du caractère *local* de la tâche considérée dans cette étude. En outre, les expérimentations que nous avons menées sont venues confirmer l'absence d'efficacité de ce type de prédicat pour l'extraction de communautés ego-centrées.

4.3 Extraction de communautés à partir d'un espace prétopologique

Afin d'illustrer le principe d'extraction d'une communauté ego-centrée par un fermé élémentaire dans un espace prétopologique, nous considérons le réseau de la figure 1a. Soit l'espace prétopologique (E, a_Q) avec E l'ensemble des nœuds du réseau et Q la DNF définie

par $Q = q_{danisch}(A, x, 0.5) \wedge q_{r1}(A, x, 0.5)$; soit la matrice de proximités *carryover-opinion* donnée par la table 1b ; la communauté ego-centrée issue du nœud a est obtenue par le fermé élémentaire $F_Q(\{a\})$ dans l'espace prétopologique (E, a_Q) :

$$\begin{aligned} a_Q(\{a\}) &= \{a, b, c\} \\ a_Q(\{a, b, c\}) &= \{a, b, c, d\} \\ a_Q(\{a, b, c, d\}) &= \{a, b, c, d\} = F_Q(\{a\}) \end{aligned}$$

Le fermé obtenu correspond effectivement à une communauté identifiable intuitivement sur le réseau. L'obtention du fermé résulte de deux applications successives de l'adhérence a_Q. Par définition de Q, l'expansion d'un sous-ensemble A à un nouvel élément x nécessite que les deux prédicats $q_{danisch}(A, x, 0.5)$ et $q_{r1}(A, x, 0.5)$ soient satisfaits. Ainsi, le singleton $\{a\}$, s'étend aux éléments b et c par une première application de l'adhérence car :

— d'une part $carryover(a, b) \geq 0.5$ et $\frac{|\{a\} \cap V(b)|}{|\{a\}|} \geq 0.5$

— et d'autre part $carryover(a, c) \geq 0.5$ et $\frac{|\{a\} \cap V(c)|}{|\{a\}|} \geq 0.5$.

mais le nœud d n'est pas atteint par la première application d'adhérence (bien que $q_{danisch}(\{a\}, d, 0.5)$ soit vrai) car $\frac{|\{a\} \cap V(d)|}{|\{a\}|} = 0$. Il le sera lors de la second application de l'adhérence grâce aux éléments b et c précédemment inclus puisque $\frac{|\{a,b,c\} \cap V(d)|}{|\{a,b,c\}|} \geq 0.5$.

Intéressons nous à présent à la communauté locale issue de d :

$$\begin{aligned} a_Q(\{d\}) &= \{b, c, d, e\} \\ a_Q(\{b, c, d, e\}) &= \{a, b, c, d, e\} \\ a_Q(\{a, b, c, d, e\}) &= \{a, b, c, d, e\} = F_Q(\{d\}) \end{aligned}$$

On retrouve encore une fois la communauté identifiable $\{a, b, c, d\}$ à laquelle s'est ajouté l'élément e ce qui est tout à fait cohérent du point de vue local au nœud d.

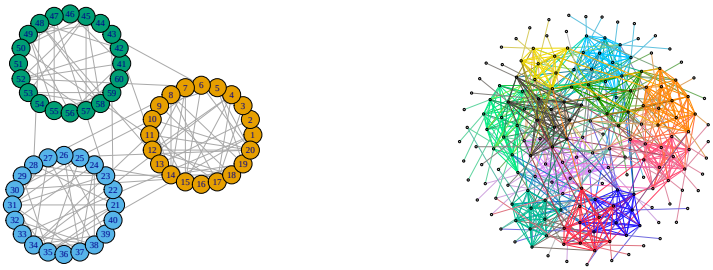
Cet exemple montre qu'un espace prétopologique correctement défini permet d'extraire la structure complexe latente d'un réseau. Cette notion d'espace prétopologique "correctement défini" nécessite que la DNF définissant l'espace prétopologique soit pertinente. C'est ce problème les méthodes d'apprentissage LPSMI et LPSFM tentent de résoudre.

5 Expérimentations

Il n'existe pas à notre connaissance de travaux visant à résoudre de manière supervisée le problème d'extraction de communautés égo-centrées. C'est pourquoi nous positionnons ces nouvelles contributions supervisées (LPSMI et LPSFM) par rapport aux approches non-supervisées existantes, en toute objectivité.

5.1 Jeux de données

Le premier réseau synthétique est composé de 60 nœuds répartis dans trois communautés de tailles égales. Il est construit sur un modèle aléatoire très simple : tout d'abord, chaque



(a) Un réseau composé de trois graphes aléatoires formant chacun une communauté.

(b) Réseau des confrontations entre les équipes de football américain.

FIG. 2: Deux exemples de réseaux synthétiques et réels.

communauté est générée suivant un modèle d'Erdős–Rényi avec une probabilité de 0.2 ; ensuite des arcs entre chaque paire de nœuds de communautés différentes sont ajoutés avec une probabilité de 0.01. La figure 2a montre un exemple de réseau ainsi généré.

Le second réseau synthétique provient du benchmark LFR (Lancichinetti et Fortunato, 2009) paramétré de sorte à obtenir un réseau de 200 nœuds. La moyenne des degrés des nœuds est 15 et un nœud peut avoir au maximum 30 voisins. Le paramètre de mélange vaut 0.3 et 40 nœuds appartiennent à trois communautés différentes parmi les 15 communautés qui composent ce réseau. Les autres paramètres sont laissés à leur valeur par défaut.

Le premier réseau réel que nous avons utilisé est le célèbre Zachary's karate club (Zachary, 1977). Il modélise les interactions entre les 34 membres d'un club de karaté et est composé de deux communautés connues.

Le second réseau issu de données réelles (figure 2b) représente les interactions entre les équipes universitaires de la division 1-A [4] qui se sont rencontrées lors des matchs de football américains sur la saison 2006. Il est identique au réseau utilisé par Chen et al. (2009). Le réseau est composé de 179 nœuds (équipes) et 787 arcs (matchs) ; 115 équipes sont réparties dans 11 communautés pré-identifiées et 64 équipes ne sont dans aucune communauté.

5.2 Protocole expérimental et résultats

Nous proposons de mesurer la qualité d'une méthode d'extraction de communautés ego-centrées en calculant, via la F-mesure, la correspondance entre les communautés extraites et les communautés réelles ou connues. Soit E l'ensemble des nœuds d'un réseau, pour chaque nœud $x \in E$, une communauté ego-centrée $C(x)$ est extraite puis comparée à la communauté attendue $C^*(x)$. Pour les méthodes prétopologiques proposées, c'est le fermé élémentaire qui défini la communauté ego-centrée extraite ($C(x) = F(\{x\})$). Les scores de précision (P),

4. http://www.espn.com/college-football/standings/_/season/2006

rappel (R) et leur moyenne harmonique (FM ou F-mesure) sont rappelées en (12).

$$P = \frac{\sum_{x \in E} |C(x) \cap C^*(x)|}{\sum_{x \in E} |C(x)|} \;\; ; \;\; R = \frac{\sum_{x \in E} |C(x) \cap C^*(x)|}{\sum_{x \in E} |C^*(x)|} \;\; ; \;\; FM = 2 \cdot \frac{P \times R}{P + R} \tag{12}$$

Nous comparons les scores obtenus par les méthodes prétopologiques avec les scores obtenus en utilisant les méthodes de Clauset, Luo, Chen (Clauset, 2005; Luo et al., 2008; Chen et al., 2009) et Danisch (Danisch et al., 2013). Les méthode de Clauset, Luo et Chen construisent les communautés ego-centrées par accumulation des noeuds maximisant une mesure de modularité. La méthode de Danisch repose sur l'idée qu'il existe une forte différence entre la proximité d'un nœud à sa communauté locale par rapport à cette proximité aux autres sommets du réseau. Danisch montre que la courbe de la proximité *carryover-opinion* pour un nœud donné est une succession de plateaux suivis de brusques décroissances. Des communautés locales à différents niveaux de granularité peuvent alors être obtenues suivant la pente que l'on considère comme marqueur de fin de la communauté. Nous avons calculé les performances de cette méthode en considérant qu'une communauté regroupe les 2, 3 ou 4 premiers plateaux, respectivement notés "Danisch2", "Danisch3" et "Danisch4" dans la suite.

La méthode LPSMI ne peut combiner que les prédicats de type V tandis que LPSFM dispose de l'ensemble des prédicats décrits dans la section précédente pour apprendre les espaces prétopologiques. Nous avons fixé arbitrairement $k = 0.3$ pour les prédicats topologiques et avons construit deux prédicats $q_{danisch}$ avec les seuils $k = 0.15$ et $k = 0.3$, nous les noterons respectivement $q_{danisch}(k = 0.15)$ et $q_{danisch}(k = 0.3)$.

Les résultats provenant des méthodes LPSMI et LPSFM ont été obtenus par validation croisée à cinq plis (5-*fold* cross-validation). Afin de comparer équitablement toutes les méthodes entre elles, les scores présentés dans la table 1 sont ceux obtenus en évaluant les méthodes sur les cinq mêmes jeux de test représentant chacun 20% des communautés à retrouver. Il est important de noter que les communautés de référence utilisées dans ces expérimentations sont déduites des partitions des nœuds des réseaux (hormis le réseau LFR qui contient des chevauchements). De ce fait, les communautés cibles correspondent à des approximations de communautés ego-centrées, les scores peuvent donc ne pas refléter la qualité réelle de chaque modèle ; ils restent cependant un bon indicateur. Enfin, les meilleures règles logiques apprises par LPSFM sont présentées dans la table 2.

Les résultats obtenus viennent d'abord confirmer l'analyse selon laquelle les espaces prétopologiques de type V (LPSMI) sont inadaptées à la modélisation des structures de communautés locales. S'affranchir des contraintes des espaces de type V permet en revanche de construire de espaces prétopologiques tout à fait pertinents comme le montrent les scores obtenus par LPSFM. La nouvelle approche proposée obtient globalement, et de manière significative, de meilleurs scores que les méthodes existantes. Cela montre l'intérêt de la supervision pour la tâche d'extraction de communautés. En effet, cela permet de prendre en considération les caractéristiques d'un réseau donné, et donc de produire un modèle adapté à celui-ci. À l'inverse, les performances des approches non-supervisées sont fortement liées au réseau sur lequel elles s'appliquent. Les algorithmes de Luo et Chen en sont un très bon exemple. Les communautés extraites sur le réseau *Foot* par Chen sont de bonne qualité, contrairement à celle de Luo. Ce

Méthode	Erdős–Rényi	Karaté	Foot	LFR
Clauset	0.45 ± 0.08	0.68 ± 0.11	0.53 ± 0.07	0.50 ± 0.06
Luo	0.74 ± 0.06	0.82 ± 0.08	0.57 ± 0.07	0.57 ± 0.05
Chen	0.39 ± 0.11	0.37 ± 0.05	0.88 ± 0.04	0.46 ± 0.06
Danisch2	0.70 ± 0.01	0.79 ± 0.05	0.63 ± 0.03	0.43 ± 0.06
Danisch3	0.80 ± 0.03	$\mathbf{0.89 \pm 0.03}$	0.65 ± 0.03	0.51 ± 0.06
Danisch4	0.82 ± 0.04	0.88 ± 0.01	0.52 ± 0.05	0.56 ± 0.06
LPSMI*	0.50 ± 0.00	0.67 ± 0.01	0.31 ± 0.07	0.34 ± 0.09
LPSFM*	$\mathbf{0.85 \pm 0.02}$	0.80 ± 0.07	$\mathbf{0.96 \pm 0.03}$	$\mathbf{0.65 \pm 0.04}$

TAB. 1: Scores de F-mesure obtenus par différentes approches d'extraction de communautés ego-centrées sur 4 réseaux. Les méthodes supervisées sont identifiées par le symbole *.

Réseaux	Règles logiques
Erdős–Rényi	$(q_{luo} \wedge q_{danisch}(k = 0.3)) \vee (q_{r2} \wedge q_{luo}) \vee (q_{adj} \wedge q_{r4} \wedge q_{danisch}(k = 0.3))$
Karaté	$q_{luo} \wedge q_{danisch}(k = 0.15)$
Foot	$(q_{r4}) \vee (q_{r3} \wedge q_{danisch}(k = 0.15))$
LFR	$q_{r1} \wedge q_{luo}$

TAB. 2: Exemples de règles apprises par l'algorithme LPSFM. Les clauses sont affichées dans l'ordre dans lequel elles sont ajoutées dans la DNF.

phénomène s'inverse lorsque l'on considère le réseau *Karaté*. Danisch semble proposer une approche qui ne souffre pas de ce problème mais elle reste sensible aux seuils.

Les modèles prétopologiques appris par LPSFM ont été appliqués sur des jeux de données différents du jeu d'apprentissage de sorte à évaluer leur capacité de généralisation. Les résultats sont présentés dans la table 3. Les mesures ont été effectuées sur l'ensemble du réseau et non plus sur des jeux de test ; c'est pourquoi les valeurs de la diagonale diffèrent des scores de LPSFM présentés dans la table 1.

De toute évidence, les modèles prétopologiques parviennent mal à se généraliser. Les règles apprises (table 2) ne se ressemblent pas, ce qui semble suggérer que les réseaux eux mêmes et leur structure communautaire en particulier ne se ressemblent pas. Il n'est donc pas étonnant qu'un modèle appris spécifiquement pour un réseau ne convienne pour un autre. Nous verrons néanmoins en perspective que nous disposons d'une piste de travail pour lever ce verrou.

Enfin notre méthode d'apprentissage d'espaces prétopologiques ouvre la voie vers des approches exploitant l'intégralité des aspects d'un réseau en permettant de combiner tous types de descripteurs entre eux. La règle apprise sur le réseau Erdős–Rényi propose par exemple

	Erdős–Rényi	Karaté	Foot	LFR
Erdős–Rényi	$0,85 \pm 0,01$	$0,62 \pm 0,07$	$0,29 \pm 0,21$	$0,07 \pm 0,00$
Karaté	$0,75 \pm 0,03$	$0,80 \pm 0,04$	$0,47 \pm 0,14$	$0,34 \pm 0,11$
Foot	$0,41 \pm 0,00$	$0,59 \pm 0,00$	$0,97 \pm 0,00$	$0,41 \pm 0,00$
LFR	$0,54 \pm 0,01$	$0,74 \pm 0,00$	$0,60 \pm 0,00$	$0,65 \pm 0,00$

TAB. 3: Scores de généralisation pour les modèles prétopologiques (LPSFM).

de combiner des descripteurs topologiques de bas niveau (q_{r2}) avec des descripteurs de plus haut niveau (q_{luo}). Cette formulation logique apporte une compréhension précise du modèle appris : il est clair que la règle apprise sur le graphe Erdős–Rényi est guidée par les deux prédicats q_{luo} et $q_{danisch}(k = 0.3)$. Cette règle montre que, sur ce réseau précis, ces deux prédicats sont (1) complémentaires puisqu'ils n'apparaissent pas ensemble dans les clauses 2 et 3 et (2) trop permissifs puisqu'ils doivent être restreints par d'autres prédicats. D'autre part, l'importance des clauses conjonctives est déterminé par l'ordre dans lequel elles apparaissent dans la formule logique apprise : la clause $q_{luo} \wedge q_{danisch}(k = 0.3)$ est ainsi celle apportant le plus d'informations utiles à la détection des communautés du réseau Erdős–Rényi.

6 Conclusion

Dans cette étude nous avons proposé une formalisation du problème d'extraction de communautés ego-centrées fondée sur les techniques récentes d'apprentissage supervisé d'espaces prétopologiques. Nous avons défini une première collection de descripteurs locaux, s'appuyant sur les principales méthodologies existantes pour cette tâche et montré expérimentalement sur des données réelles ou simulées, d'une part qu'il existe des modèles de structuration prétopologiques adaptés aux réseaux étudiés et d'autre part que ces modèles peuvent être appris de façon supervisée.

Nos travaux démontrent la pertinence de la prétopologie dans l'extraction de communautés dans les réseaux puisqu'elle permet de gagner en performance par rapport aux méthodes classiques de détection de communautés ego-centrées. Les outils mis à disposition par la prétopologie permettent d'exprimer naturellement et élégamment des interactions de différentes natures entre ensembles d'éléments. Cette capacité se révèle indispensable lorsqu'il s'agit d'exploiter différents niveaux d'informations présentes dans un réseau. Cette étude ouvre la voie à de nombreuses perspectives de recherche dont nous dégageons quelques pistes de travail.

Notre étude a permis de mettre en évidence qu'une prétopologie de type V n'est pas adaptée à la tâche d'extraction de communautés ego-centrées au moyen des fermés élémentaires. Cependant, il est toutefois possible qu'une définition différente permette l'extraction de communautés de bonne qualité depuis un espace prétopologique de type V.

D'autre part, si les modèles prétopologiques permettent l'extraction de bonnes communautés, ces modèles semblent difficiles à généraliser aux réseaux sur lesquels ils n'ont pas été entraînés, or cet entraînement représente un coût évident et potentiellement rédhibitoire [5]. Apprendre un modèle prétopologique de façon non-supervisée permettrait de lever ce verrou. Dans cet objectif, nous avons observé les bonnes capacités de généralisation du modèle appris sur le réseau LFR. Ce résultat ainsi que de récents travaux (Lu et al., 2018) semblent montrer qu'il est possible d'utiliser des réseaux générés artificiellement pour obtenir des données étiquetées. Une nouvelle approche consisterait alors à générer automatiquement des données d'entraînement à partir d'un réseau synthétique structurellement proche du réseau réel ciblé.

Références

Belmandt, Z. (1993). Manuel de prétopologie et ses applications.

5. Bien qu'en pratique, on peut observer que peu d'exemples suffisent à guider efficacement l'apprentissage.

Caillaut, G. et G. Cleuziou (2018). Learning pretopological spaces to model complex propagation phenomena : A multiple instance learning approach based on a logical modeling. *arXiv preprint arXiv :1805.01278*.

Chen, J., O. R. Zaïane, et R. Goebel (2009). Local community identification in social networks. *ASONAM*, 237–242.

Clauset, A. (2005). Finding local community structure in networks. *Physical review E 72*(2), 026132.

Cleuziou, G. et G. Dias (2015). Learning pretopological spaces for lexical taxonomy acquisition. In *ECML/PKDD (2)*, Volume 9285 of *LNCS*, pp. 493–508. Springer.

Dalud-Vincent, M. (2017). Une autre manière de modéliser les réseaux sociaux. applications à l'étude de co-publications. *Nouvelles perspectives en sciences sociales 12*(2), 41–68.

Danisch, M., J. Guillaume, et B. L. Grand (2013). Towards multi-ego-centred communities : a node similarity approach. *IJWBC 9*(3), 299–322.

Figueiredo, D. R., L. F. R. Ribeiro, et P. H. P. Saverese (2017). struc2vec : Learning node representations from structural identity. In *KDD*.

Grover, A. et J. Leskovec (2016). node2vec : Scalable feature learning for networks. In *KDD*, pp. 855–864. ACM.

Lancichinetti, A. et S. Fortunato (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E 80*(1), 016118.

Lu, X., K. Kuzmin, M. Chen, et B. K. Szymanski (2018). Adaptive modularity maximization via edge weighting scheme. *Inf. Sci. 424*(C), 55–68.

Luo, F., J. Z. Wang, et E. Promislow (2008). Exploring local community structures in large networks. *Web Intelligence and Agent Systems 6*(4), 387–400.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences 103*(23), 8577–8582.

Palla, G., I. Derényi, I. Farkas, et T. Vicsek (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*(7043), 814.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research 33*(4), 452–473.

Summary

We present a pretopological based approach to extract ego-centered communities. Classical methods often consider only one structural feature of the the network, whereas pretopology enables to do multi-criteria analysis. Our approach consists in learning a logical combination of a network's descriptors to define a pretopological space. Ego-centered communities are extracted by computing the elementary closure of each node. The quality of such communities is evaluated against the ground truth communities. We show the benefits of our method by comparing it to others on both real and synthetic networks.