

Frontières des communautés polarisées : application à l'étude des théories complotistes autour des vaccins

Alexis Guyot, Annabelle Gillet, Éric Leclercq

Laboratoire d'Informatique de Bourgogne - EA 7534
Université de Bourgogne Franche-Comté
Dijon, France
{prenom}.{nom}@depinfo.u-bourgogne.fr

RÉSUMÉ. Les données des réseaux sociaux sont de plus en plus utilisées pour en extraire de la valeur, dans des domaines tels que le marketing, la politique ou la sociologie. Celles-ci peuvent être représentées sous forme de graphes, en modélisant précisément les interactions à travers des liens dirigés et pondérés. Dans l'analyse des données des réseaux sociaux, l'étude des communautés est une étape essentielle. Toutefois, pour une interprétation fine des phénomènes, il est également nécessaire d'étudier leurs interactions et de pouvoir détecter des traces de polarisation. Nous proposons une méthode qui permet d'évaluer l'antagonisme des communautés et d'identifier leurs frontières dans des réseaux pondérés et dirigés. Notre méthode s'accompagne d'une implémentation disponible en accès libre. Nous validons expérimentalement notre proposition par l'étude des théories complotistes autour de tweets liés à la vaccination contre la COVID-19.

ABSTRACT. Social network data are increasingly used to extract value from them, in different domains such as marketing, politics or sociology. These data can be represented by graphs which model the interactions between individuals through directed and weighted links. The detection and study of communities in online social networks are important tasks to understand the behavior of users. However, for a detailed interpretation of a phenomena, it is also necessary to study their interactions and to be able to detect and evaluate polarization of communities. We propose a method which allows to evaluate the antagonism of the communities and to identify their boundaries in weighted and directed networks. An implementation is available in open access. We experimentally validate our proposal by studying conspiracy theories around tweets related to COVID-19 vaccines.

MOTS-CLÉS : réseaux sociaux, graph mining, communautés, polarisation, frontières de communautés

KEYWORDS: social networks, graph mining, communities, polarization, communities boundaries

1. Introduction

Depuis leur apparition à la fin des années 90 et leur explosion dans la seconde moitié des années 2000, les réseaux sociaux numériques (RSN) se sont révélés comme une exceptionnelle source d'étude pour de très nombreux domaines de recherche. Avec plus de 2.7 milliards d'utilisateurs pour Facebook¹, 1.1 milliard pour Instagram ou encore 300 millions pour Twitter et une durée d'utilisation moyenne de 144 minutes par jour², les RSN captent à chaque instant un nombre conséquent d'interactions entre êtres humains. Ils les transforment en données qui peuvent être analysées en interne ou collectées par des tiers. Les données des RSN sont utilisées pour atteindre de nombreux objectifs, allant d'études à portée sociologique pour le monde de la recherche jusqu'à de l'analyse de marché pour orienter les stratégies marketing des entreprises. La polarisation des communautés est un exemple d'étude qui suscite un fort intérêt (Gillani *et al.*, 2018; Kearney, 2019; Lee *et al.*, 2014), afin de mieux comprendre l'organisation du réseau ainsi que les interactions qu'il traduit.

Toutes ces études profitent de façon directe ou indirecte des différentes propriétés induites par la représentation la plus courante de ces données : des graphes, dits sociaux, où les informations propres aux utilisateurs sont usuellement contenues dans les sommets de la structure et celles propres à leurs interactions dans les arêtes. Quand une interaction possède un émetteur et un destinataire, le lien est représenté sous la forme d'un arc. Pour simplifier le modèle de données, certains algorithmes de détection de communautés comme Louvain (Blondel *et al.*, 2008) font le choix d'ignorer ces informations contextuelles et de ne travailler qu'avec des arêtes. Cela entraîne la perte de beaucoup d'information, notamment concernant le rôle des utilisateurs aux extrémités. Souvent, une pondération est ajoutée aux liens pour indiquer le nombre d'interactions qui relie deux sommets et ainsi représenter la force de la connexion.

Une propriété intéressante des graphes sociaux concerne leur topologie. La distribution des degrés de leurs sommets suit bien souvent une loi de puissance, ce qui les catégorise comme graphes sans échelle (Barabási, Bonabeau, 2003). Cette propriété entraîne la possibilité de découvrir des zones localement denses, nommées communautés d'utilisateurs, qui peuvent elles aussi être analysées pour tirer de nouvelles conclusions. En plus de cela, s'ajoute la possibilité d'identifier des traces de polarisation entre les communautés détectées. Plusieurs travaux comme ceux de (Chitra, Musco, 2020; Interian, Ribeiro, 2018; Isenberg, 1986; Sunstein, 2002) définissent la polarisation comme le phénomène qui intervient lorsqu'un groupe de personnes peut en réalité être décomposé en deux sous-groupes qui possèdent des opinions contrastées et conflictuelles à propos d'un sujet particulier, avec éventuellement quelques individus en faible nombre qui restent neutres. Grâce à une méthode permettant de détecter des traces de polarisation entre deux communautés d'un graphe social, il devient alors possible de vérifier si les débats et interactions des personnes traduisent bien des

1. <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

2. <https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/>

formes d'antagonisme entre opinions autour de sujets présentés comme clivants. Les traces de polarisation peuvent ensuite être utilisées pour adapter une stratégie, par exemple pour déterminer la meilleure façon de communiquer. Pour s'affranchir de la barrière de la langue et des approximations orthographiques, des méthodes uniquement basées sur une analyse topologique du graphe sont à privilégier, sans faire intervenir de traitement du langage naturel ou d'analyse des sentiments.

La méthode qui sera présentée dans cet article est une extension de celle développée par (Guerra *et al.*, 2013). Nos contributions sont les suivantes : 1) une amélioration de la méthode pour prendre en considération les informations contextuelles contenues dans la pondération et dans la direction des liens ; 2) un algorithme pour identifier les zones internes et frontières des communautés et calculer leur antagonisme, ainsi qu'une implémentation dans le langage R³ ; 3) une application de la méthode sur des données réelles provenant de Twitter, collectées pour une étude liée au complotisme et à la vaccination.

La suite de l'article est organisée de la manière suivante : la section 2 présente les travaux connexes majeurs sur la polarisation et la mesure de l'antagonisme. La section 3 détaille l'approche proposée avec la méthode, l'algorithme et son implémentation. La section 4 est un retour d'expérience sur des données réelles. Enfin la section 5 conclut l'article.

2. Travaux connexes

Dans cette partie, nous commencerons par présenter la méthode spécifiée par Guerra *et al.* (2013), qui permet d'évaluer l'antagonisme entre communautés. Dans un second temps, nous discuterons de quelques solutions alternatives.

Les premiers travaux portant sur la polarisation des communautés sur Twitter remontent à 2011 avec (Conover *et al.*, 2011). L'article de Guerra *et al.* (2013) propose une méthode qui permet d'évaluer finement la polarisation. Celle-ci utilise un graphe G découpé en communautés par un algorithme classique. Elle identifie des zones frontières et définit quelques indicateurs pour évaluer la polarisation. Ces indicateurs peuvent être utilisés en complément de la modularité (Newman, 2006). Il s'agit d'une mesure de densité dont une forte valeur indique la présence de communautés assez fermées sur elles-mêmes et ayant peu d'interactions avec les autres communautés, ce qui peut être interprété comme une forme de clivage. Bien que tout à fait utilisable, la modularité seule présente quelques limites. La principale réside dans le fait qu'une grande valeur de modularité est une condition nécessaire mais pas suffisante pour conclure sur une potentielle polarisation des communautés. Il est effectivement tout à fait possible de découvrir une valeur élevée dans des graphes dont on sait les communautés non-polarisées.

3. <https://github.com/AlexisGuyot/CommunityBoundaries>

La méthode de (Guerra *et al.*, 2013) se concentre sur les interactions entre N communautés $G_n, n = 1, \dots, N$, et permet de calculer une valeur d'antagonisme entre chaque paire de communautés. Deux types d'ensembles d'utilisateurs sont ainsi définis pour chaque paire G_i et G_j (voir figure 1) : 1) les membres de la zone interne I_{ij} , qui appartiennent à G_i mais n'interagissent pas avec G_j ; 2) les membres de la zone frontière B_{ij} , qui appartiennent à G_i et interagissent à la fois avec I_{ij} et B_{ji} . Plus les membres des zones frontières sont impliqués au sein de leur communauté, plus ils sont susceptibles de prendre à cœur le point de vue qu'ils défendent et donc de présenter un fort antagonisme envers ceux qui ne le partagent pas. En partant de ce postulat, l'évaluation de l'antagonisme est faite en mesurant la proportion d'interactions des utilisateurs de la zone frontière vers l'intérieur de leur communauté (E_{int}) par rapport à l'ensemble de leurs interactions, y compris celles avec l'autre communauté (E_B). Ces différentes notions sont formalisées dans les définitions suivantes.

$$B_{i,j} = \{v_i : v_i \in G_i, \exists e_{ik} \mid v_k \in G_j, \exists e_{ik} \mid (v_k \in G_i \mid \nexists e_{kl} \mid v_l \in G_j), i \neq j\} \quad (1)$$

$$I_i = G_i - B_{ij} \quad (2)$$

$$E_B = \{e_{mn} : v_m \in B_{i,j} \wedge v_n \in B_{j,i}\} \quad (3)$$

$$E_{int} = \{e_{mn} : v_m \in (B_{i,j} \cup B_{j,i}) \wedge v_n \in (I_i \cup I_j)\} \quad (4)$$

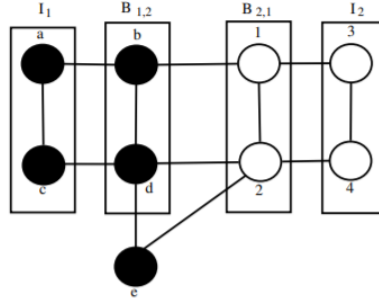


FIGURE 1. Un exemple simple pour comprendre les différentes zones des communautés, tiré de l'article de (Guerra *et al.*, 2013).

À partir de ces différents sous-ensembles, il est ensuite possible de calculer l'antagonisme entre les communautés G_i et G_j grâce à l'équation suivante :

$$P = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \right] \quad (5)$$

avec $d_i(v)$ le nombre d'arêtes du sous-ensemble E_{int} liées au sommet v , et $d_b(v)$ le nombre d'arêtes du sous-ensemble E_B liées au sommet v . La proportion entre les deux types d'interactions est comparée à chaque fois avec l'hypothèse nulle suivante : chaque sommet possède autant de connexions avec l'intérieur de la zone interne de sa communauté qu'avec la zone frontière de l'autre communauté. La valeur de P est comprise entre les valeurs -0.5 et 0.5 . Une valeur négative est signe d'une absence de polarisation entre deux communautés, et inversement. Le cas où l'ensemble B est vide ne nous intéresse pas, car il correspond à une situation où il est impossible de savoir si les deux communautés ont conscience de l'existence de l'une et de l'autre.

Bien que très intéressante à bien des égards, la méthode proposée par (Guerra *et al.*, 2013) comporte quelques limitations et faiblesses. Par rapport à notre problématique, son plus grand défaut réside dans le fait qu'elle soit spécifiée pour des graphes non-orientés et non-pondérés. Puisque caractériser le rôle des utilisateurs du graphe est au centre de la méthode pour évaluer l'antagonisme et donc la polarisation, il est important de ne pas négliger l'information apportée par la direction. En effet, la relation d'antagonisme n'est pas forcément symétrique. Certains utilisateurs d'une communauté peuvent adresser des messages à d'autres utilisateurs d'une autre communauté en les mentionnant sans que la réciproque existe. Il en est de même pour la pondération, puisqu'il n'est pas pertinent de considérer que seul un lien sur trois sort de la frontière en direction de l'autre communauté si celui-ci représente en réalité 80 interactions alors que les 2 autres seulement 10 en cumulé. La méthode présente également une petite inconsistance dans la spécification des zones internes. En effet, avec une telle définition, le sommet e de la figure 1 devrait normalement être considéré comme membre de la zone interne de G_1 . Or, ce n'est pas le cas. Ces trois raisons nous ont motivés à proposer une extension de la méthode.

Plusieurs autres approches ont également été développées pour mesurer la polarisation au sein d'un graphe social. Morales *et al.* (2015) utilisent par exemple le concept de propagation d'influence. Pour cela, ils définissent deux groupes d'utilisateurs, les élites et les auditeurs, qui se voient attribuer un score qui évalue leur degré d'approbation par rapport à une problématique. Les utilisateurs élites correspondent alors aux utilisateurs influents du réseau, les points de départ de l'information. Une propagation par label (Zhu, Ghahramani, 2002) depuis ceux-ci est effectuée pour construire deux pôles, puis une valeur de polarisation est obtenue en prenant en considération leur taille et la distance qui sépare leur centre de gravité. La méthode est spécifiée sur des graphes dirigés mais non-pondérés. Une autre contrainte de leur système est la phase d'initialisation des degrés d'approbation des sommets élites. Une valeur fixe doit être attribuée à chaque noeud de ce type, sans qu'une méthode particulière ne soit précisée pour la déterminer.

Une méthode s'appuyant sur la factorisation de matrices a été proposée par Al Amin *et al.* (2017). Le graphe social de base est converti en un graphe biparti contenant d'un côté les utilisateurs du réseau (les sources), et de l'autre les faits relayés dans le graphe et qui sont en faveur, en défaveur ou neutres par rapport à la problématique de l'étude (les assertions). Ce graphe est obtenu après une phase de nettoyage et de transforma-

tion des données, où les messages similaires sont transformés en assertions. La matrice d'incidence de ce graphe biparti est ensuite décomposée en composantes principales grâce à un algorithme de descente de gradient, pour aboutir à une évaluation de la polarisation. Le choix de la mesure de similarité entre deux messages, pour déterminer s'ils appartiennent à la même assertion, constitue une limite à cette méthode. Il implique une connaissance précise de la nature de l'interaction et doit être adapté en fonction de si le lien représente un partage, une réponse ou une réaction par exemple.

Même s'il est plus courant de vouloir se concentrer sur une analyse structurelle des graphes sociaux pour s'affranchir de la barrière de la langue, certaines méthodes comme (Alamsyah, Adityawarman, 2017) proposent une approche hybride basée à la fois sur une analyse de la topologie du réseau et sur une analyse de sentiments à l'aide de techniques de *machine learning*. Une première phase de traitement du langage naturel permet de classifier les utilisateurs comme pro, anti ou neutres, puis des communautés d'utilisateurs sont détectées grâce à l'algorithme de Louvain. Les membres des communautés sont comparés à leur classification par l'algorithme de *machine learning*. Les communautés sont ensuite analysées selon plusieurs angles pour pouvoir conclure sur la polarisation. Le bon fonctionnement de ce type de méthodes se base toutefois quasiment entièrement sur l'efficacité de l'algorithme de traitement du langage naturel utilisé, qui peut en plus être impacté par les abréviations, fautes d'orthographe, de syntaxe, etc.

Guerra *et al.* (2017) ont par la suite continué d'affiner leur méthode pour mieux prendre en considération la sémantique des interactions de Twitter pour détecter des traces de polarisation, notamment pour les *retweets*. De nouveaux éléments d'analyse sont avancés, comme le temps écoulé depuis la diffusion du message original ou l'utilisation de l'interaction de *quote* comme outil pour exprimer de l'antagonisme à travers du sarcasme ou de la moquerie. Ils montrent également que plus le degré de polarisation entre deux communautés augmente, plus la proportion d'interactions entre elles par rapport à celles avec les autres communautés augmente.

Par rapport aux trois méthodes de (Morales *et al.*, 2015), (Al Amin *et al.*, 2017) et de (Alamsyah, Adityawarman, 2017), celle de (Guerra *et al.*, 2013) présente l'avantage de ne pas nécessiter de connaissance métier supplémentaire sur le graphe pour pouvoir l'appliquer. Son seul pré-requis est la présence de communautés identifiées, qui peuvent être détectées et construites automatiquement à partir d'algorithmes d'analyse de la topologie du graphe. Nous proposons d'étendre la méthode pour prendre en compte les liens dirigés et pondérés, sans entrer en contradiction avec les éléments supplémentaires apportés dans (Guerra *et al.*, 2017).

3. Approche proposée

Dans cette section, nous commencerons par décrire l'extension que nous proposons à la méthode de Guerra *et al.* (2013). Dans le cadre du projet interdisciplinaire

COCKTAIL⁴ qui a pour objectif la création d'un observatoire en temps réel des tendances et des signaux faibles sur Twitter pour aider les acteurs du domaine agro-alimentaire à la prise de décisions, nous avons également créé une implémentation en R. Nous décrirons son fonctionnement dans un second temps.

3.1. Présentation de la méthode

Pour prendre en compte la direction et la pondération des liens, nous proposons les définitions suivantes pour les sous-ensembles des zones internes et frontières, ainsi que pour ceux des arêtes intra et inter-communautés :

$$I_{i,j} = \{v : v \in G_i, \nexists e_{vn} \mid n \in G_j, i \neq j\} \quad (6)$$

$$B_{i,j} = \{v : v \in G_i, \exists e_{vn_1} \mid n_1 \in G_j, \exists e_{vn_2} \mid n_2 \in I_{i,j}, i \neq j\} \quad (7)$$

$$E_B = \{e_{sd} : s \in B_{i,j} \wedge d \in G_j\} \quad (8)$$

$$E_{int} = \{e_{sd} : s \in B_{i,j} \wedge d \in I_{i,j}\} \quad (9)$$

Il est important de noter qu'avec cette nouvelle spécification, certains sommets peuvent donc se trouver ni dans la zone interne ni dans la zone frontière de leur communauté, ce qui corrige le problème posé par la précédente définition de I par Guerra *et al.* (2013). Il s'agit dans ce cas de sommets dont tous les voisins se trouvent soit dans l'autre communauté ou soit dans la zone frontière de la leur. Dans le cadre de notre recherche d'antagonisme, les utilisateurs associés à ce type de sommets ne nous intéressent pas. En effet, ceux-ci ne présentent pas de traces évidentes d'implication au sein de leur communauté et pourraient très bien être des utilisateurs neutres qui se sont retrouvés mêlés à des interactions entre communautés.

E_B est toujours défini comme le sous-ensemble des arêtes inter-communautés dont l'origine est un sommet s appartenant à $B_{i,j}$ et dont la destination est cette fois-ci un sommet d appartenant à G_j . E_{int} est le sous-ensemble contenant les arêtes intra-communauté qui ont pour origine un sommet s appartenant à $B_{i,j}$ et pour destination un sommet d appartenant à $I_{i,j}$. La nouvelle formule à utiliser pour prendre en charge la pondération du graphe en plus de l'orientation de ses arêtes est la suivante :

$$P = \frac{1}{|B_{i,j}|} \sum_{v \in B_{i,j}} \left[\frac{\sum_{e \in E_{iv}} weight(e)}{\sum_{e \in E_{iv}} weight(e) + \sum_{e \in E_{bv}} weight(e)} - 0.5 \right] \quad (10)$$

$$E_{iv} = \{e_{vd} : e_{vd} \in E_{int}\} \quad (11)$$

$$E_{bv} = \{e_{vd} : e_{vd} \in E_B\} \quad (12)$$

Par rapport à la formule 5, on ne se contente pas de compter les arêtes liées à v , mais on additionne leur poids. E_{iv} et E_{bv} représentent respectivement l'ensemble des arêtes de

4. <https://projet-cocktail.fr/>

E_B ou de E_{int} dont la source est le sommet v . Pour chaque utilisateur membre d'une zone frontière, on calcule la proportion que représente la somme des poids de ses interactions vers la zone interne de sa communauté par rapport à la somme des poids de ses interactions vers l'extérieur. Le score d'antagonisme du sommet est obtenu par comparaison avec l'hypothèse nulle, et celui de la zone frontière par calcul d'une moyenne.

La prise en charge de l'orientation nous conduit à étudier individuellement chaque paire de communautés, à la fois (G_i, G_j) et (G_j, G_i) . Cette perte en efficacité est compensée par le gain en précision par rapport à la version de Guerra, puisque cette définition de l'antagonisme permet de remarquer si l'une des deux communautés polarisées exprime un plus fort degré d'antagonisme que l'autre.

3.2. De la méthode à l'algorithme

Nous proposons un algorithme s'appuyant sur une structure de donnée adaptée, la *structural matrix*, afin de réduire la complexité assez importante de la méthode naïve, qui consiste à construire les différents ensembles en parcourant tous les nœuds du graphe pour vérifier récursivement les conditions d'appartenance.

La *structural matrix* est en réalité un tableau qui indique pour un sommet v , qu'on sait appartenir à la communauté G_i , la zone dans laquelle il se situe lorsqu'on étudie ses liens avec chaque autre communauté. Pour chaque cellule du tableau située à la ligne i et à la colonne j , on utilise le système de code suivant :

- le code 0 signifie que le sommet v_i fait partie de la zone interne de sa communauté lorsqu'on étudie la paire formée par celle-ci et la communauté G_j ;
- le code 1 signifie que le sommet v_i ne fait partie d'aucune zone particulière dans sa communauté lorsqu'on étudie la paire formée par celle-ci et la communauté G_j ;
- le code 2 signifie que le sommet v_i fait partie de la zone frontière de sa communauté lorsqu'on étudie la paire formée par celle-ci et la communauté G_j ;
- le code 3 signifie que le sommet v_i fait partie de la communauté G_j .

Tableau 1. Exemple de *structural matrix*. Le sommet v appartient à la communauté G_1 et fait partie de sa zone interne pour la paire (G_1, G_2) , d'aucune zone pour (G_1, G_3) et de sa zone frontière pour (G_1, G_4) .

Structural matrix				
	G_1	G_2	G_3	G_4
v_{i-1}
v	3	0	1	2
v_{i+1}

Dans un premier temps, l'algorithme 1 est exécuté pour construire la *structural matrix*. Deux passages complets dans la structure sont effectués. Entre les lignes 3 à 11, on recense les sommets qui ne seront pas membres de zones internes pour chaque

Algorithm 1 Build Structural Matrix

Require: adjacency_list, community_membership, community_count and vertex_count

Ensure: structural_matrix $\in \mathbb{N}^{\text{vertex_count} \times \text{community_count}}$

```
1: Initialize structural_matrix, with structural_matrix  $\in \mathbb{N}^{\text{vertex\_count} \times \text{community\_count}}$ 
2: Fill structural_matrix with 0
3: for  $v = 1, \dots, \text{length}(\text{adjacency\_list})$  do
4:   com_v  $\leftarrow$  community_membership[v]
5:   for  $n = 1, \dots, \text{length}(\text{adjacency\_list}[v])$  do
6:     com_n  $\leftarrow$  community_membership[v]
7:     if com_v  $\neq$  com_n then
8:       structural_matrix[v, com_n]  $\leftarrow$  1
9:     end if
10:   end for
11: end for
12: neighboring_com  $\leftarrow$  []
13: com_to_internals  $\leftarrow$  []
14: for  $v = 1, \dots, \text{length}(\text{adjacency\_list})$  do
15:   com_v  $\leftarrow$  community_membership[v]
16:   for  $n = 1, \dots, \text{length}(\text{adjacency\_list}[v])$  do
17:     com_n  $\leftarrow$  community_membership[v]
18:     if com_v  $\neq$  com_n then
19:       neighboring_com  $\leftarrow$  com_n
20:     else
21:       for  $c = 1, \dots, \text{ncol}(\text{structural\_matrix}[n])$  do
22:         if structural_matrix[n, c] = 0 then
23:           com_to_internals  $\leftarrow$  c
24:         end if
25:       end for
26:     end if
27:   end for
28:   is_boundary_with  $\leftarrow$  neighboring_com  $\cap$  com_to_internals
29:   for  $c \in \text{is\_boundary\_with}$  do
30:     structural_matrix[v, c]  $\leftarrow$  2
31:   end for
32: end for
```

Algorithm 2 Build Antagonism Matrix

Require: structural_matrix, community_membership, community_count, adjacency_list

Ensure: antagonism_matrix $\in \mathbb{R}^{community_count \times community_count}$

```
1: Initialize antagonism_matrix, with antagonism_matrix  $\in \mathbb{R}^{community\_count \times community\_count}$ 
2: Initialize count_matrix, with count_matrix  $\in \mathbb{N}^{community\_count \times community\_count}$ 
3: for  $v = 1, \dots, \text{nrow}(\text{structural\_matrix})$  do
4:   com_i  $\leftarrow$  community_membership[v]
5:   ens_com_j  $\leftarrow$  []
6:   for  $j = 1, \dots, \text{ncol}(\text{structural\_matrix})$  do
7:     if structural_matrix[v, j] = 2 then
8:       ens_com_j  $\leftarrow$  j
9:     end if
10:  end for
11:  for com_j  $\in$  ens_com_j do
12:    ebv = 0
13:    eiv = 0
14:    for  $n = 1, \dots, \text{length}(\text{adjacency\_list}[v])$  do
15:      if community_membership[n] = com_j then
16:        ebv  $\leftarrow$  ebv + weight(v,n)
17:      else
18:        if community_membership[n] = com_i AND structural_matrix[n,com_j] = 0 then
19:          eiv  $\leftarrow$  eiv + weight(v,n)
20:        end if
21:      end if
22:    end for
23:    antagonism_matrix[com_i, com_j]  $\leftarrow$  antagonism_matrix[com_i, com_j] +
      ( $\frac{eiv}{eiv+ebv} - 0.5$ )
24:    count_matrix[com_i, com_j]  $\leftarrow$  count_matrix[com_i, com_j] + 1
25:  end for
26: end for
27: antagonism_matrix  $\leftarrow$  antagonism_matrix / count_matrix
```

communauté. La deuxième boucle, entre les lignes 14 et 32, permet de différencier les sommets qui constitueront les zones frontières de ceux à ignorer. Pour cela, on détecte les sommets qui possèdent à la fois un premier voisin dans une autre communauté G_2 et un second situé dans la zone interne de sa communauté avec G_2 .

Une fois cette première phase terminée, la *structural matrix* est retournée puis utilisée par l'algorithme 2. Celui-ci lit la structure précédente pour alimenter la formule 10 d'antagonisme. Pour chaque code égal à 2, on parcourt la liste des voisins de v (lignes 14 à 22). En fonction de leur communauté et de leur rôle au sein de cette der-

nière, on incrémente les compteurs d'interactions inter-communautés (ebv) et intra-communauté (eiv) de la valeur du poids de l'arc qui les sépare. Une fois les compteurs à jour pour le sommet v , on calcule à la ligne 23 le ratio des interactions comme indiqué dans la formule 10. À chaque fois qu'un ratio est ajouté pour une paire de communautés, on le notifie dans une matrice de compteurs pour, à la ligne 27, pouvoir calculer les valeurs moyennes attendues dans la matrice d'antagonisme.

4. Application sur des données réelles

Nous proposons maintenant d'appliquer notre méthode sur des données réelles issues de Twitter. On cherche, à partir des communautés identifiées, à appliquer notre méthode pour mesurer l'antagonisme entre les communautés et détecter leur polarisation.

La collecte des *tweets* s'est faite du 18 novembre 2020 au 26 janvier 2021 sur une détection de mots-clés liés aux vaccins et à la COVID-19 et à partir d'une liste de comptes d'utilisateurs⁵. Le corpus complet contient plus de 9 millions de *tweets* en langue française. À partir de celui-ci, nous extrayons le graphe des mentions, interaction qui capture le mieux les discussions et débats. Celui-ci contient 6 450 sommets et 19 939 arcs, une fois les utilisateurs peu actifs retirés. Après application de l'algorithme de détection de communautés Louvain, on identifie 9 communautés significatives. La modularité du graphe après un tel découpage est de 0.5, ce qui atteste de la présence d'une structure communautaire. Pour pouvoir catégoriser les communautés détectées, nous avons recherché les *hashtags* les plus représentatifs utilisés par leurs membres. Pour cela, nous avons construit un graphe biparti avec d'un côté un ensemble de sommets qui représentent les utilisateurs de la communauté et de l'autre un deuxième ensemble où cette fois-ci un sommet représente un *hashtag*. Sur celui-ci, nous avons calculé la centralité *PageRank* (Page *et al.*, 1999) pour identifier les nœuds, et donc les *hashtags*, les plus influents. À partir de leur étude, nous proposons dans le tableau 2 une catégorisation manuelle des 9 communautés extraites du graphe des mentions, obtenue par identification des thèmes les plus récurrents dans les différents groupes.

Quelques points intéressants sont à commenter concernant ces premiers résultats. Tout d'abord, on peut noter l'absence de communauté définie pour les anti-vaccins, contrairement aux pro-vaccins. Une hypothèse est que ces utilisateurs ont tendance à se disperser au sein d'autres communautés, ou alors qu'ils ne discutent pas ou peu de leur opinion entre eux. Le deuxième point intéressant à commenter est la présence en nombre non-négligeable de *hashtags* liés au complotisme au sein des différentes communautés. On retrouve par exemple *#blanquement* ("Blanquer ment"), *#plandemie*, *#greatreset*, *#complotvaccinobligatoire* ou encore *#jenesaispasjedemande*. Alors que

5. vaccin, antivax, covid, sinopharm, bigpharma, corona, santepublique, astazeneca, spoutnik, sputnik, pfizer, biontech, moderna, vaxxie, @BioNTech_Group, @SinopharmIntl, @Pfizer_France, @moderna_tx, @sputnikvaccine

Tableau 2. Catégorisation des communautés significatives.

ID	Taille	Exemples hashtags	Catégorie
2	887	#blanquerment #parentsencolere #blanquerdemission	Anti-Blanquer
7	1173	#polqc #plandemie #dictaturesanitaire #caq #polcan	Québécois
14	1110	#raoult #ivermectine #dictaturesanitaire #plandemie #greatreset	Pro-traitements alternatifs
89	282	#dictaturesanitaire #jenemeconfineraipas #enmarche #complotvaccinobligatoire	Anti-gouvernement
114	1230	#ggrmc #afp #cnews #lci #dictaturesanitaire	Réactions médias
117	539	#familyisnottourism #dictaturesanitaire #familyiseverything	Famille
146	300	#avecxb #ladroitequipeutgagner #dictaturesanitaire #lr	Droite
161	670	#rtlmatin #antivax #jenesaispasjedemande	Auditeurs radio
163	259	#stopdictaturesanitaire #jemefaisvacciner #standwithscience	Pro-vaccin

le corpus n'est pas du tout construit autour de cette thématique complotiste, on remarque alors que celle-ci revient souvent lors des discussions autour de la vaccination et de la COVID-19. Il en est de même pour le hashtag #dictaturesanitaire, qui traduit une contestation de la gestion de la crise sanitaire. Ce hashtag est présent dans pratiquement toutes les communautés, qu'elles soient anti-gouvernement ou non, comme le montre la communauté pro-vaccin dont le hashtag #stopdictaturesanitaire est l'un des plus représentatifs.

Tableau 3. Valeurs d'antagonisme entre les différentes communautés.

	2	7	14	89	114	163
2	0	0.182	0.237	0.294	0.398	0
7	0.218	0	0.292	0.213	0	0
14	0.204	0.153	0	0.333	0	0.208
89	0.159	0	0.310	0	0	0
114	0.159	0.251	0.167	0.241	0	0
163	0.196	0.242	0.250	0.167	0.025	0

Avec le graphe des mentions découpé en communautés, on peut appliquer notre méthode d'évaluation de l'antagonisme par construction des zones frontières. L'algorithme permet de produire la matrice présentée dans le tableau 3. Toutes les communautés ne sont pas représentées. Nous proposons d'illustrer la manière d'utiliser ces résultats en se concentrant sur la communauté 163, les pro-vaccins. Dans un premier temps, la colonne correspondante dans la matrice permet d'obtenir des informations sur la façon dont la communauté reçoit l'antagonisme de la part des autres. Ici, on voit

qu'une seule valeur est non nulle, 0.208, qui correspond au degré d'antagonisme reçu par la communauté 163 de la part de la communauté 14. On peut alors conclure que dans notre modèle de données, seuls les utilisateurs en faveur des traitements alternatifs supposés contre la COVID-19, comme l'hydroxychloroquine ou l'ivermectine, sont porteurs d'antagonisme envers les défenseurs de la vaccination.

En étudiant cette fois-ci la ligne associée à la communauté 163, on identifie les communautés envers lesquelles ses membres expriment de l'antagonisme. Les communautés non représentées dans la matrice reçoivent toutes une valeur nulle de la part des pro-vaccins. À part celles-ci, on remarque que cette communauté exprime une forme d'antagonisme envers toutes les autres, avec une valeur plus importante à l'égard des membres de la communauté québécoise et en retour aux attaques des partisans des traitements alternatifs. Si la rivalité entre les partisans de la vaccination et ceux de la médecine alternative est compréhensible, on peut s'interroger sur la provenance de l'antagonisme envers les québécois. Ceux-ci ont-ils exprimés une plus grande méfiance envers les vaccins contre la COVID-19 ou envers la gestion de la crise sanitaire de leur pays ? Dans tous les cas, il est intéressant de noter que malgré le fait que la communauté 163 soit peu attaquée par les autres, celle-ci présente un comportement assez offensif. Les membres de cette communauté semblent bien investis dans la défense de leur opinion et s'opposent à la plupart des autres communautés.

En exploitant notre méthode, on peut isoler les utilisateurs membres des différentes zones internes et frontières. La méthode peut alors également être utilisée pour mesurer la porosité des frontières de communautés. Une frontière est considérée comme poreuse quand elle est constituée d'un grand nombre d'utilisateurs qui interagissent autant voire plus avec l'extérieur de la communauté qu'avec sa zone interne. Ceux-ci sont alors moins impliqués et investis dans le groupe, ce qui traduit une plus faible cohésion au sein de la communauté. Avec notre formule d'antagonisme, un sommet qui possède une valeur intermédiaire⁶ négative ou nulle participe à la porosité de la frontière à laquelle il appartient. En calculant le pourcentage de sommets des frontières pour lesquelles la valeur intermédiaire d'antagonisme est négative, on peut estimer la porosité de chaque communauté.

Tableau 4. Porosité des frontières des différentes communautés.

Communauté	2	7	14	89	114	117	146	161	163
Porosité	16%	13%	17%	24%	13%	50%	5%	10%	20%

On peut interpréter la porosité comme un indicateur de l'adéquation d'une communauté au sens informatique du terme avec une communauté au sens social du terme. Quand les algorithmes de détection de communautés identifient des groupes d'utilisateurs, ils découvrent surtout des zones localement denses où les sommets sont plus reliés entre eux qu'avec le reste du graphe. Cela traduit seulement le fait que les utilisateurs correspondants communiquent plus entre eux qu'avec le reste de la population,

6. Valeurs calculées à l'intérieur de la somme et propres à chaque membre d'une zone frontière.

mais pas forcément qu'ils appartiennent à un même groupe social dont les membres partagent une même vision ou un même but commun. Quand on étudie l'entière des *hashtags* significatifs dans la communauté 117, dont les frontières sont très poreuses, on remarque la présence de plusieurs thèmes pré-dominants. Dans le tableau 2, nous avons identifié la thématique familiale comme principale car ses *hashtags* associés sont légèrement plus nombreux, mais on retrouve aussi des sujets liés par exemple à la politique (gilets jaunes, regroupement familial) ou à l'actualité médiatique (affaire Duhamel, élections américaines). On remarque alors que même si du point de vue de la topologie du réseau, les membres de la communauté interagissent entre eux, ils ne forment pas en réalité un groupe social particulier. En revanche, la communauté 146, qui possède une porosité de seulement 5%, regroupe les partisans du parti politique républicain, qui lui est un groupe social identifié.

D'autres interprétations des résultats retournés par notre méthode sont possibles. Il est par exemple envisageable d'étudier les caractéristiques (centralité, degré, etc.) des sommets des différentes zones internes et frontières des communautés, pour identifier le profil des utilisateurs de ces zones. On peut également catégoriser les zones comme nous l'avons fait avec les communautés, pour identifier les sujets qui sont débattus au sein de la communauté, ceux qui sont défendus vers l'extérieur, etc. Toutes ces connaissances acquises sur la structure des communautés et leur façon d'interagir entre elles permettent ensuite de conclure sur des traces de polarisation au sein du graphe.

Dans notre cas d'étude, on remarque des communautés avec une bonne cohésion, comme le montrent la valeur de modularité et les différentes valeurs de porosité en majorité inférieures à 20%. En revanche, le découpage ne fait pas apparaître deux pôles pro et anti-vaccins. La seconde catégorie d'utilisateurs est dispersée au sein de différentes communautés, et la première, bien que identifiée et présente dans le corpus, reste en faible nombre comparé à la taille du graphe. On peut donc en conclure que la thématique des vaccins contre la COVID-19 en tant que telle n'est pas source de polarisation. Cela est dû au fait que les partisans anti-vaccins ne se regroupent pas sous une même bannière, et se différencient plutôt les uns des autres par les idées annexes qu'ils partagent conjointement à leur avis sur le sujet, notamment des idées complotistes et anti-gouvernement. En revanche, les différentes valeurs d'antagonisme nous permettent d'identifier d'autres sujets qui eux sont vecteurs de polarisation. C'est notamment le cas de la thématique plus globale des types de traitements possibles contre la maladie, où on remarque une polarisation entre les personnes en faveur de la vaccination et celles en faveur de solutions médicamenteuses.

5. Conclusion

Dans cet article, nous avons présenté notre contribution à la détection de polarisation au sein de graphes sociaux à travers une extension de la méthode de Guerra *et al.* (2013) et une application sur des données réelles. Notre nouvelle spécification propose une prise en charge de deux nouvelles propriétés intrinsèquement liées aux

graphes issus de réseaux sociaux numériques, la pondération et la direction des liens. Celles-ci permettent de bénéficier d'une bien plus grande précision lors de l'étude des interactions entre communautés. Notre méthode permet d'atteindre plusieurs finalités : catégorisation de rôles pour les membres de communautés (zones internes et frontières), mesure de l'antagonisme, mesure de la porosité des frontières de communautés, etc. Nous avons également détaillé le fonctionnement d'une implémentation de la méthode, qui utilise une structure intermédiaire appelée la *structural matrix*. Pour finir, nous avons montré l'utilité et le bon fonctionnement de notre méthode au travers de l'étude d'un cas pratique lié aux débats sur le RSN Twitter autour de la vaccination contre la COVID-19. Nous avons développé deux manières d'interpréter les résultats obtenus : la détection de polarisation à travers l'évaluation de l'antagonisme et l'analyse structurelle intra-communautaire à travers l'évaluation de la porosité des zones frontières. À partir de ces différents résultats, nous avons pu conclure sur la polarisation des débats et discussions autour de cette thématique.

Les premières utilisations réelles de l'algorithme semblent indiquer un temps d'exécution linéaire. Cette tendance est à confirmer en réalisant une étude expérimentale de la complexité. Du côté des analyses, la porosité permet d'entrevoir la possibilité d'établir des liens entre les notions informatiques et de sciences humaines et sociales de communautés. Nous prévoyons de développer ce concept à l'avenir, en s'appuyant sur le projet interdisciplinaire COCKTAIL.

Remerciements

Ce travail est soutenu par le programme « Investissements d'Avenir », projet ISITE-BFC (contrat ANR-15-IDEX-0003). Le projet Cocktail est piloté scientifiquement par Gilles Brachotte, laboratoire CIMEOS EA-4177, Université de Bourgogne.

Bibliographie

- Al Amin M. T., Aggarwal C., Yao S., Abdelzaher T., Kaplan L. (2017). Unveiling polarization in social networks: A matrix factorization approach. In *IEEE Conference on Computer Communications (INFOCOM)*, p. 1–9.
- Alamsyah A., Adityawarman F. (2017). Hybrid sentiment and network analysis of social opinion polarization. In *5th International Conference on Information and Communication Technology (ICoICT7)*, p. 1–6.
- Barabási A.-L., Bonabeau E. (2003). Scale-free networks. *Scientific american*, vol. 288, n° 5, p. 60–69.
- Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, vol. 2008, n° 10, p. P10008.
- Chitra U., Musco C. (2020). Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th international conference on web search and data mining*, p. 115–123.

- Conover M., Ratkiewicz J., Francisco M., Gonçalves B., Menczer F., Flammini A. (2011). Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, vol. 5.
- Gillani N., Yuan A., Saveski M., Vosoughi S., Roy D. (2018). Me, my echo chamber, and I: introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, p. 823–831.
- Guerra P., Meira Jr W., Cardie C., Kleinberg R. (2013). A measure of polarization on social media networks based on community boundaries. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7.
- Guerra P., Nalon R., Assunção R., Meira Jr W. (2017). Antagonism also flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In *Proceedings of the international aaai conference on web and social media*, vol. 11.
- Interian R., Ribeiro C. C. (2018). An empirical investigation of network polarization. *Applied Mathematics and Computation*, vol. 339, p. 651–662.
- Isenberg D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, vol. 50, nº 6, p. 1141.
- Kearney M. W. (2019). Analyzing change in network polarization. *new media & society*, vol. 21, nº 6, p. 1380–1402.
- Lee J. K., Choi J., Kim C., Kim Y. (2014). Social media, network heterogeneity, and opinion polarization. *Journal of communication*, vol. 64, nº 4, p. 702–722.
- Morales A. J., Borondo J., Losada J. C., Benito R. M. (2015). Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, nº 3, p. 033114.
- Newman M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, vol. 103, nº 23, p. 8577–8582.
- Page L., Brin S., Motwani R., Winograd T. (1999). *The pagerank citation ranking: Bringing order to the web.*. Rapport technique. Stanford InfoLab.
- Sunstein C. R. (2002). The law of group polarization, 10 j. *Pol. Phil*, vol. 175, p. 177.
- Zhu X., Ghahramani Z. (2002). Learning from labeled and unlabeled data with label propagation.