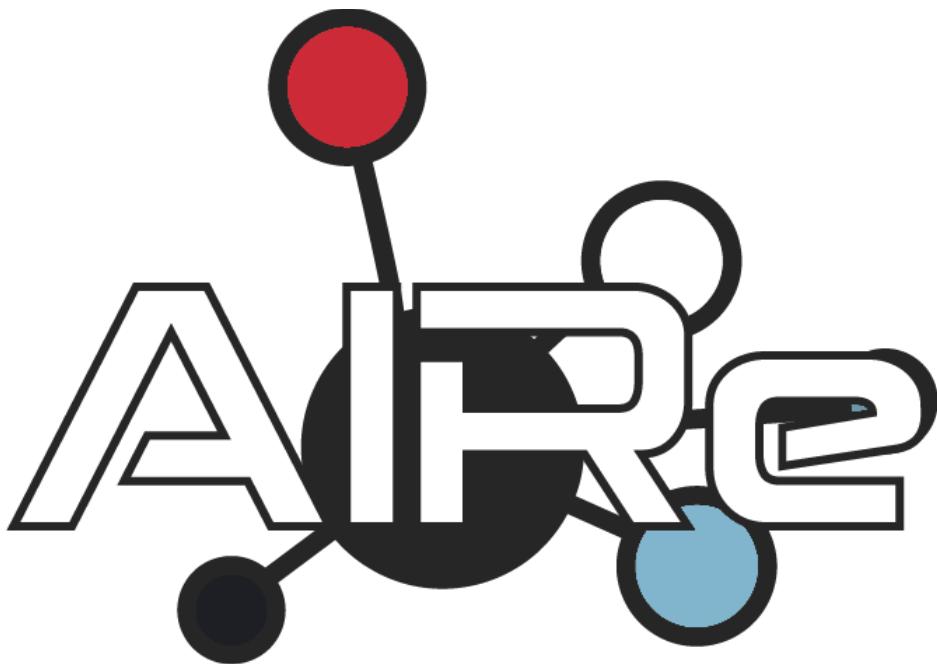


Rapport final



Mise en lumière des phénomènes de polarisation politique dans le réseau des médias à visée politique

Référence	Groupe 8
Projet	Projet transpromotion
Date de début	19/10/20

Equipe
1A : Arnaud Bascop, Mathis Gendron, Benjamin Haté, Lottie Rocuet
2A : Albane Arthuis, Justine Bulteau, Tom Pelletreau-Duris

TABLE DES MATIÈRES

INTRODUCTION	3
I. Objectifs	4
II. Démarche	6
Etat de l'art	6
API de Twitter	9
Étapes de la démarche	9
Discussion autour des limites rencontrées	10
III. Résultats	11
Premiers résultats	11
Résultats intermédiaires	14
Résultats finaux	18
IV. GESP	26
Répartition du travail et coordination	26
Outils	28
Risques	28
Planning	29
Biais	31
Difficultés rencontrées et adaptation à la crise sanitaire	33
CONCLUSION	34
ANNEXE	35
BIBLIOGRAPHIE	36
WEBOGRAPHIE	38

INTRODUCTION

Le projet AIRe est né de la collaboration d'étudiants de l'ENSC avec la volonté de faire lien entre les disciplines qualitatives et quantitatives, faire dialoguer deux approches souvent mises en concurrence pour enrichir notre compréhension du monde politique et des relations humaines en générales. Notre étude porte en particulier sur deux médias politiques populaires en France, le Figaro et Libération, et le réseau d'abonnement sur la plateforme Twitter qu'ils ont en commun. Avec la théorie sociologique des réseaux sociaux d'une part et la théorie des graphes de l'autre, nous avions pour objectif de mettre en lumière les phénomènes de pouvoirs de réseaux, d'asymétrie d'information et de mise en communauté.

L'idée de ce projet a germé à partir d'observations faites dans les débats qui nous entourent au quotidien. Bien souvent, les désaccords et autres disputes semblent bien plus liés à des malentendus voire à des problèmes de définition axiologique qu'à des problèmes d'oppositions construits (et constructifs). Or, le système de croyance de chacun qui d'une certaine manière peut s'apparenter à un système axiologique de pensée (ensemble de proposition acquises comme vraies pour élaborer des conclusions grâce à la logique), se construit sur la base du contact avec l'environnement informationnel proche (parents, école, médias, amis, etc). En ce qui concerne les idées politiques, elles semblent se forger d'abord par l'influence parentale avant d'être affinées ou cultivées tout au long de la vie par la lecture de contenus journalistiques ou par des échanges verbaux. Dans cette logique, un simple *follow* sur Twitter est un marqueur socio-politique de premier choix. Les réseaux sont ainsi à même de dresser un profil de plus en plus précis sur leurs utilisateurs et leurs idées. Et en même temps, un individu est lui-même influencé et ses choix façonnés par les suggestions et les recommandations. On retrouve dans cette analyse l'opposition

Individualisme/Structuralisme des sciences sociales.

En effet, les utilisateurs des réseaux sociaux numériques (RSN) sont sujets aux biais cognitifs et particulièrement celui de confirmation. Lorsqu'une personne navigue sur les RSN, elle voit inconsciemment ses convictions renforcées et est moins soumise à des opinions contraires. De fait, nous pouvons observer la formation de regroupement d'individus pensant de plus en plus uniformément au sein d'un groupe donné et de plus en plus différemment d'autres groupes d'individus. Avec l'incorporation du deep-learning aux algorithmes de recommandations il semble que l'accès à l'information est devenu de plus en plus tautologique et polarise d'autant plus les opinions. Les malentendus semblent donc, pour la grande majorité, être le résultat de l'interaction entre deux mondes informationnels qui ne discutent pas à partir du même contenu idéologique, du même système de croyance. L'ensemble des médias à visée politique auxquels un individu est abonné est donc un marqueur socio-politique de la couleur politique de son monde informationnel. Cela reflète son système de croyance. Ainsi, l'étude de réseaux des médias politiques français permettrait de mettre au jour ces communautés qui fonctionnent autant comme de "petits" mondes idéologiques, s'auto-alimentant et se renforçant.

I. Objectifs

D'une manière générale, nous observons dans les réseaux que les informations sortent rarement de la communauté de laquelle elles sont issues et donc que l'emplacement dans le réseau détermine l'accès à l'information et donc oriente une certaine vision du monde. En mettant en lumière les acteurs centraux (hubs et méga-hubs) avec de forts pouvoirs de réseaux et les phénomènes de mise en communautés (clusters) autour de ces hubs, nous voulons montrer que la structure du réseau n'est pas neutre.

Pour lutter contre cet effet de polarisation des opinions, seule la prise de conscience des individus participant au réseau peut permettre de faire diminuer la polarisation des opinions. En effet, le chercheur Akihiro Nishi et son équipe ont par exemple montré, avec l'aide de la théorie des jeux, que la transparence informationnelle dans un réseau avait pour effet de limiter l'accroissement des inégalités de ressources en provoquant une sorte de contrôle collectif tacite. De ce fait, l'accaparement des ressources économiques par les plus riches peut s'apparenter à l'accaparement de l'information par ceux en possédant le plus. Les médias possédant le plus de pouvoir de réseau pouvant tout aussi bien véhiculer le plus leurs informations.

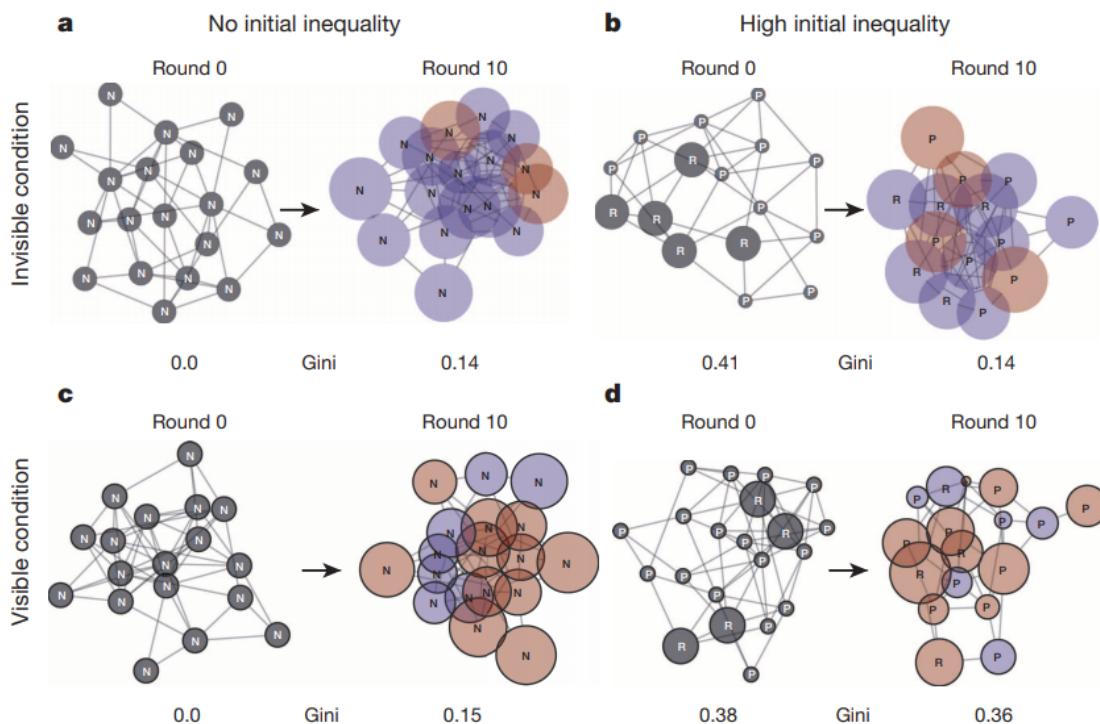


Schéma n°1 : illustration de l'effet de la transparence informationnelle sur l'allocation des ressources au sein d'un réseau (en haut à gauche : sans inégalité, en haut à droite : avec inégalité forte et sans transparence, augmentation du Gini, en bas à droite : sans inégalité et avec transparence, en bas à gauche : avec inégalité et avec transparence, légère diminution du Gini)

Le fait que les conditions initiales du jeu soient connues par tous dès le début à permis d'éviter les phénomènes de regroupement d'intérêt. Nous pouvons donc légitimer notre démarche en mettant en avant que la conscientisation des phénomènes de polarisation informationnelle par les médias chez les lecteurs/abonnés participe à la réduction des inégalités informationnelles au sein du réseau, certains abonnés peuvent eux aussi, par leur capital social, influencer l'accès à l'information des autres.

Le but final serait alors de diffuser cette analyse accompagnée de visualisations parlantes et d'analyses chiffrées de manière à faire prendre conscience de ces effets de structure. On peut résumer cela par la volonté de diffuser une certaine "ouverture d'esprit".

II. Démarche

a) Etat de l'art

Le premier travail devait être collectif. En tant qu'équipe il était nécessaire d'avoir en commun les connaissances et la terminologie appropriée. En conséquence, nous avons mis en place une bibliographie catégorisée en 3 parties : *Science des réseaux*, *Sociologie des réseaux sociaux* et *Technique*. À partir de ces ressources, nous avons constitué un document avec des définitions et des fiches pour faciliter la compréhension de notre sujet par chacun et éclaircir sa dimension transdisciplinaire. Vous trouverez dans la bibliographie de ce rapport ses sources. Il nous fallait en particulier définir la notion de "pouvoir de réseau" afin d'analyser correctement nos graphes.

❖ **sommet ou nœud** : unité fondamentale du graphe.

- ❖ **arcs ou arêtes** : liaison entre deux sommets. Deux sommets sont **adjacents** s'il existe un arc ou une arête entre eux.
- ❖ **hub** : Noeud du réseau avec fort pouvoir de réseau (voir def plus bas).
- ❖ **méga-hub** : Noeud du réseau avec très fort pouvoir de réseau.
- ❖ **réseau invariant d'échelle** : Plus il y a de nœuds au sein de ce type de réseau, plus la probabilité de se connecter à un nœud qui est déjà très connecté augmente. Phénomène de polarisation aussi appelé "winner takes all".
- ❖ **réseau "petit monde"** : caractérisé par une distance entre les nœuds très faible et un coefficient d'agglomération (ou clustering) fort.
- ❖ **degré d'un sommet** : nombre d'arcs qui en partent ou arrivent (mesure de la taille de son voisinage). Le degré d'un sommet peut être interprété comme un indicateur de l'intégration ou de l'isolement du nœud dans l'ensemble du réseau. Ou un indicateur de centralité.
- ❖ **diamètre** : plus longue distance possible entre deux nœuds du réseau.
- ❖ **densité d'un graphe** : Rapport entre le nombre d'arêtes observé et le nombre maximal d'arêtes possibles. Lorsque l'ordre d'un graphe est de n , le nombre maximal d'arêtes est de $n(n-1)$. Si k est le nombre d'arcs alors densité = $k/n(n-1)$.
- ❖ **pouvoir de réseau** : mélange de centralité, bridging et réputation.
Le pouvoir d'un individu est proportionnel à la fois au nombre de ses relations et au nombre de trous structuraux dans son environnement relationnel.

- ❖ **centralité** : Linton Freeman (1979) propose de distinguer en réalité trois formes fondamentales de centralité dans les réseaux sociaux : la **centralité de degré** (degré sur Gephi, nombre de contacts d'un individu), la **centralité de proximité** (Closeness Centrality sur Gephi, la somme des distances d'un sommet à tous les autres sommets du graphe) et la **centralité d'intermédiairité** (Betweenness Centrality sur Gephi, le nombre de chemins du réseau auxquels il appartient, c'est-à-dire le nombre de chemins qui passent par lui).
- ❖ **réputation**: à la fois la force des liens entre eux et la densité des liens autour d'eux qui permettent d'exercer des contrôles indirects. (Centrality Eigenvector sur Gephi).
- ❖ **bridging** : liens faibles connectant des composantes qui autrement seraient déconnectées les unes des autres, on parle aussi de *trou structuraux*. Plus les liens de type bridging sont nombreux, plus la contrainte est faible, plus le capital social du nœud faisant bridge est important.

Tableau n°1 : lexique utile de notre bibliographie

En conséquence, nous nous proposons d'étudier le pouvoir de réseau des médias en se concentrant sur trois critères : **la centralité de degré** (le nombre d'abonné d'un compte), **la réputation** au sein du réseau (calculée par la *centralité d'Eigenvector*) et la **capacité à fédérer** un cluster (bridging).

Pour aborder le problème de manière pratique, il nous a fallu définir notre domaine de recherche, le caractériser et le limiter. Sur quelle communauté est-il le plus pertinent de se baser pour lancer la recherche d'utilisateurs ? Pourquoi ? Avec quelles méthodes ? Après la lecture de textes de référence, nous avons fait le choix de procéder

par l'élaboration d'un réseau par profondeur. Nous avons d'abord sélectionné deux journaux majeurs : l'un plutôt de droite et l'un plutôt de gauche sur l'échiquier politique. Ces deux journaux nous semblaient équivalents dans leur visibilité et semblaient soulever des intérêts relativement divergents ce qui nous permettait de mettre en commun des communautés distinctes. Ce choix, bien que subjectif, nous permettait de former un réseau relativement représentatif de la sphère politico-média (voir la section biais pour plus de détails concernant la représentativité de notre étude). Il nous fallait ensuite récupérer les abonnés de ces deux journaux, puis les abonnés des abonnés. Nous avons donc un réseau de diamètre 2.

b) API de Twitter

Afin de récupérer les informations nécessaires à l'élaboration de notre réseau, nous devions utiliser l'API de Twitter. L'API est une interface mise à disposition par l'entreprise, contenant des méthodes informatiques permettant de récupérer les données disponibles sur ses serveurs. Aussi, nous savions que le langage python est reconnu pour la manipulation de données et le data mining. Nous avons aussi utilisé GitHub pour la facilité d'échange de nos algorithmes. La bibliothèque d'accès à Twitter que nous avons utilisé s'appelle "Tweepy".

Tweepy permet d'utiliser l'API de Twitter avec le langage Python et nous facilitait la mise en place de requêtes de récupération de données et l'application de modification aux données récoltées (filtrage notamment).

c) Étapes de la démarche

Ainsi, après la découverte et la maîtrise de ces outils, la démarche expérimentale s'est établie :

1. Sélection de hubs de départs

2. Récupération d'un échantillon de leurs abonné.e.s (10%)
3. Récupération des abonnements de leurs abonné.e.s
4. Génération des liens par notre algorithme : un lien est un abonnement
5. Classification des hubs selon notre algorithme : plus un utilisateur reçoit de liens, plus il est un hub important (son poids).
6. Génération de fichier tableur contenant les hubs, les poids, les liens, et les utilisateurs
7. Restructuration des fichier tableur pour la constitution d'un graphe avec Gephi.
8. Génération des représentations avec un graphe important
9. Filtrage
10. Graphe final et représentations finales.

d) Discussion autour des limites rencontrées

L'utilisation de Tweepy a été une évidence en vue des nombreuses recherches effectuées. En effet, nous avons testé d'autres outils tels que "Twimp", qui est aussi une bibliothèque python pour interagir avec Twitter, mais sans avoir accès à leur API. Elle s'est finalement avérée inefficace comparée aux performances de récupération de Tweepy. C'est en effet les performances de récupération qui ont été notre problème majeur tout au long de notre démarche. Twitter imposant des grandes limites de temps aux développeurs (comme nous) leur demandant des données, nous tombons vite sous la contrainte.

La première étape de récupération est plutôt très accessible (environ 4h pour obtenir 600 000 individus de différents hubs de départ) et va nous permettre d'établir un premier graphe simple pour imager la répartition des utilisateurs autour des deux hubs de départ (soit l'utilisateur est abonné à l'un, soit à l'autre, soit aux deux). Mais la 3ème étape est beaucoup longue (environ 4h pour parcourir les abonnements de seulement 300 utilisateurs).

Cette difficulté a remis en question notre démarche et nous a poussé à explorer d'autres solutions, qui ont toutes été un échec, comme notre essai d'utiliser une banque libre de données déjà disponibles en ligne (multivacplatform.org, le site ne nous a jamais répondu), ou de changer de bibliothèque python, etc... Ne pouvant pas accéder à des services payants pour arriver à nos fins, nous avons finalement décidé de rester sur notre démarche afin de récupérer nos données. Nous découvrirons de toutes manières par la suite que les abonnements de 370 abonnés nous feraient déjà obtenir pas moins de 20 millions de liens.

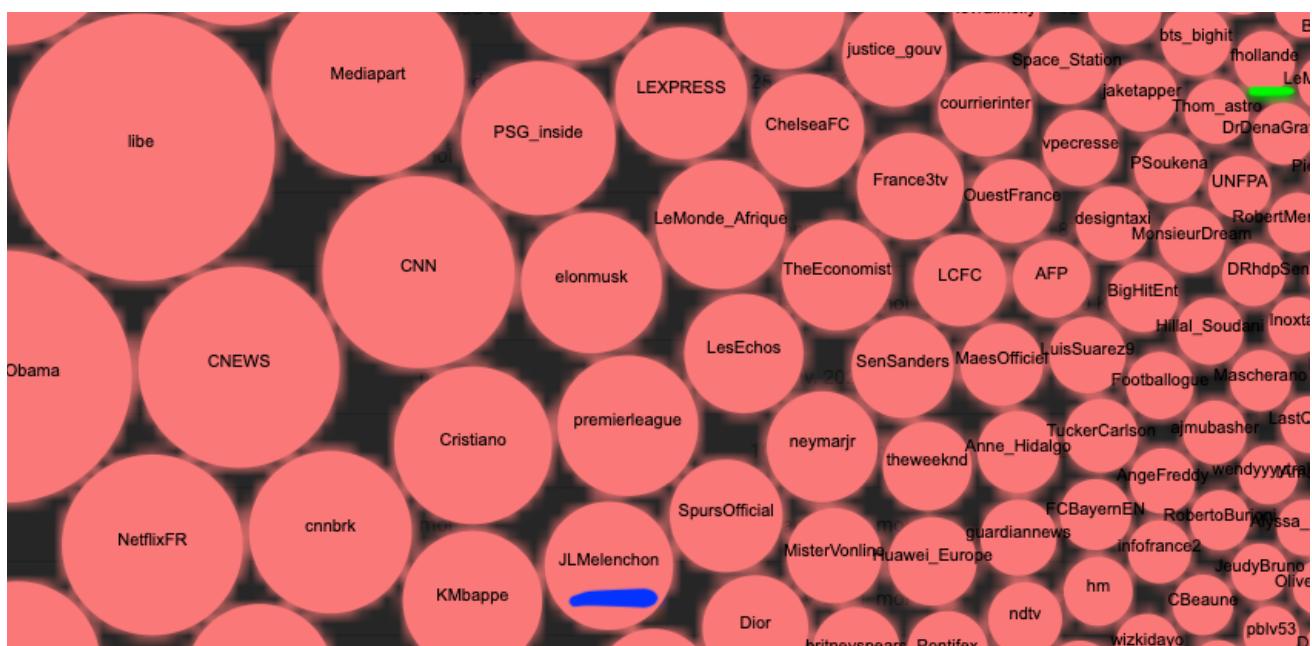
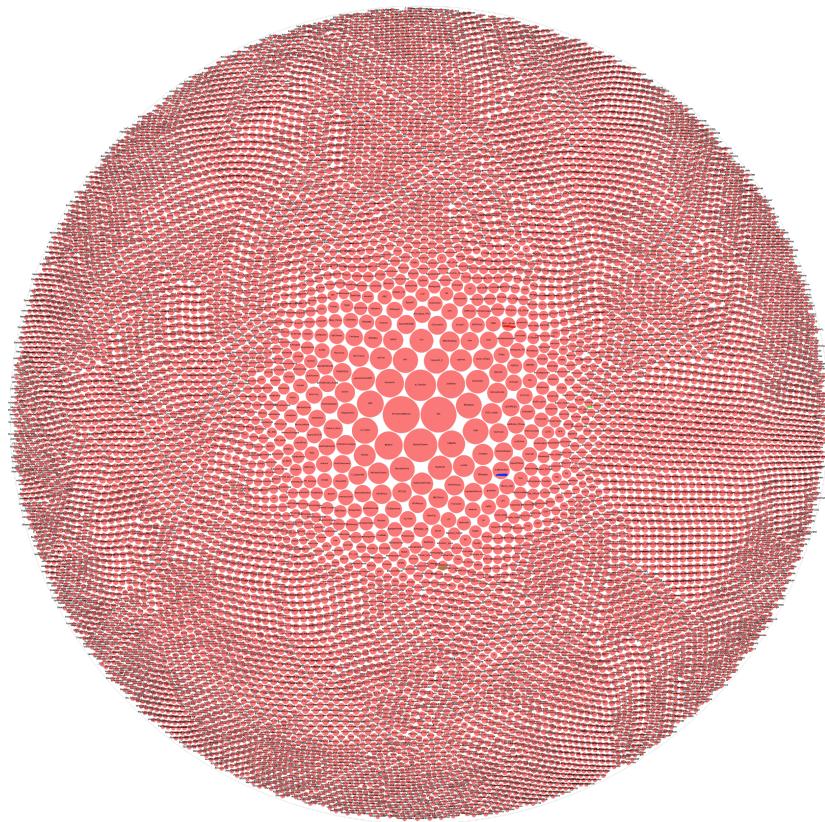
Nous avons alors rencontré une nouvelle difficulté : nos limites techniques à la génération de si gros réseaux sur le logiciel libre Gephi. En tournant suffisamment longtemps sur nos ordinateurs et en appliquant des tris sur les hubs et les liens pour les rendre toujours plus pertinents et moins nombreux, nous avons finalement obtenu des représentations analysables.

III. Résultats

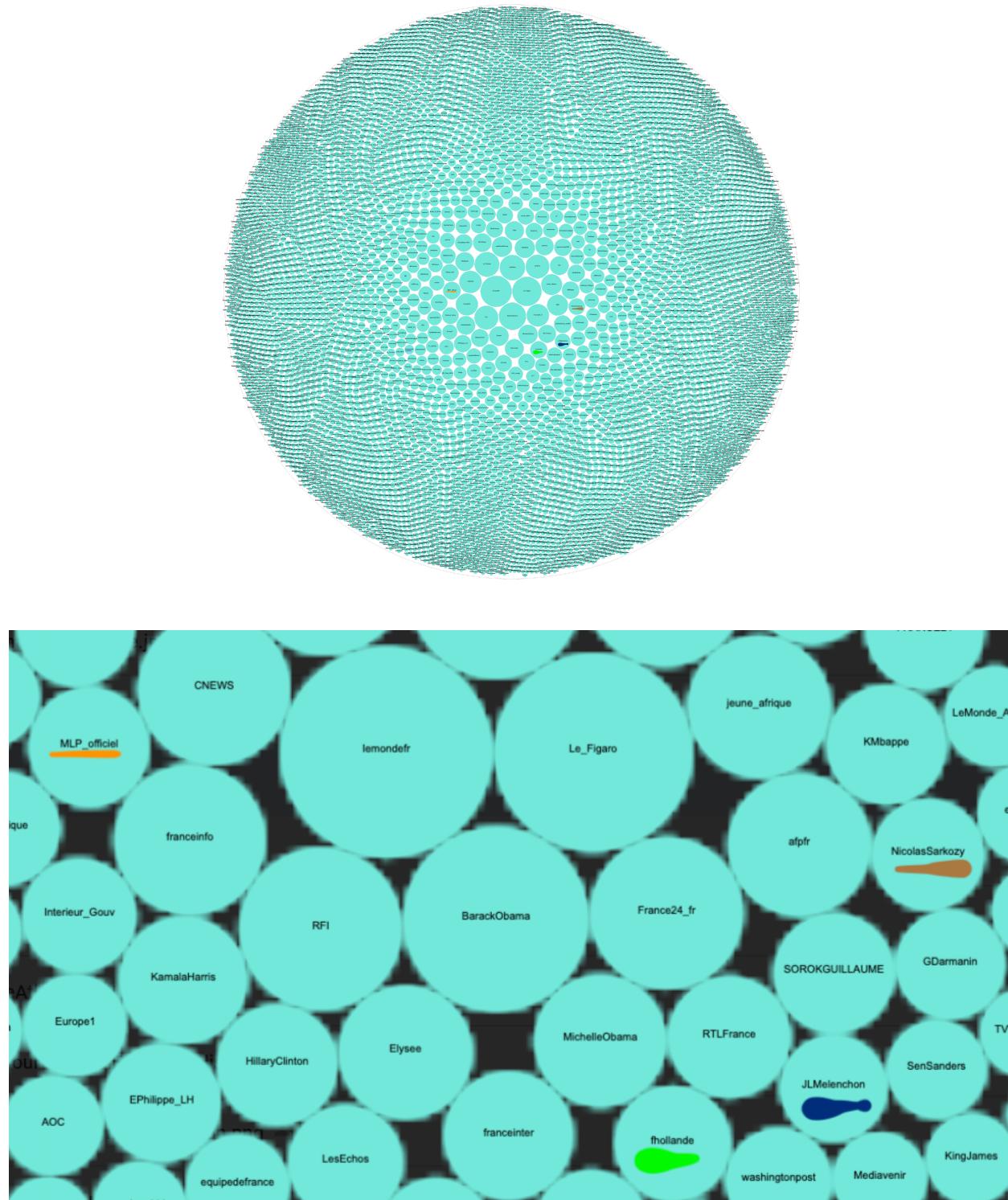
a) Premiers résultats

La première étape a été une simple mise en forme des nœuds obtenus. Les données ont été stockées en fichier csv, constituant une énorme liste de nœuds et de leur poids.

Nous pouvons, par exemple, présenter les plus gros hubs que nous avons déniché pour la communauté de Libération (rouge) et la communauté de Figaro (bleu). On remarque alors des distinctions entre les différents hubs principaux (au centre). Par exemple, on peut citer une popularité forte pour Emmanuel Macron et Jean Luc Mélenchon chez les abonnés du Libération, alors que les abonnés du Figaro ont une tendance similaire pour Nicolas Sarkozy et Marine Le Pen.



Graphique 1 : Hubs principaux concernant la communauté de Libération et un zoom central



Graphique 2 : Hubs principaux concernant la communauté de Figaro et un zoom central

b) Résultats intermédiaires

Pour se rapprocher de nos objectifs finaux, il a été nécessaire d'effectuer des traitements sur toutes les données que nous avons réussi à récupérer. Il nous fallait recouper les redondances pour lier les deux communautés et formater les données d'une manière à pouvoir les rentrer sur le logiciel Gephi. Pour cela, nous avons mis au point un algorithme en C# afin de correctement configurer les données. Les nœuds devaient être de la forme *Label;Pondération* et les liens de la forme *Source;Target*. Le code de cet algorithme est disponible en annexe. L'enjeu principal était de transformer les liens entre les noeuds d'un format

LabelDuNoeud;LabelNoeudLien1;LabelNoeudLien2;etc en autant de ligne

LabelDuNoeud;LabelNoeudLien1 et *LabelDuNoeud;LabelNoeudLien2, etc.*

```
static void LectureEcriture(string fichier_source, string fichier_cible)
{
    // Création d'une instance de StreamReader pour permettre la lecture de notre fichier source fichierDico
    System.Text.Encoding encoding = System.Text.Encoding.GetEncoding("iso-8859-1");
    StreamReader monStreamReader = new StreamReader(fichier_source, encoding);

    // Création d'une instance de StreamWriter pour permettre l'écriture de notre fichier cible
    Streamwriter monStreamWriter = File.CreateText(fichier_cible);

    string ligne = monStreamReader.ReadLine();
    monStreamWriter.WriteLine("Source" + "\t" + "Target");

    while (ligne != null)
    {
        //Pour chaque ligne,
        // Pour une ligne on veut prendre le premier pseudo, le garder en mémoire quelque part pour le réécrire à chaque ligne avant d'écrire, après une tab
        //Pour différencier le premier des autres on met en place un compteur qui, égale à zéro, indique qu'on est bien sur le premier pseudo. C'est avant d'.
        int nbPseudo = 0; //compte le nombre de mot

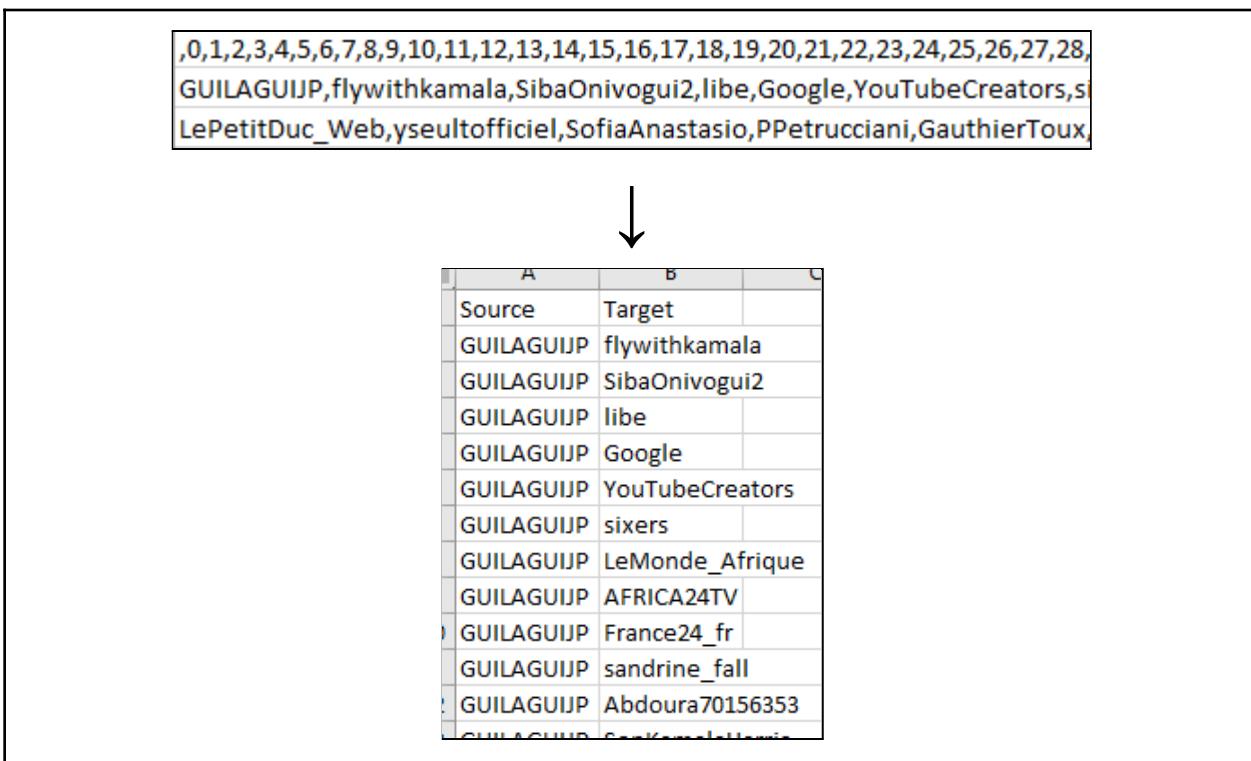
        //Ensuite, à chaque fois qu'on rencontre un point virgule, on saute une ligne et on réécrit notre premier pseudo avant d'écrire ce qu'on trouve.
        string premierPseudo = "";
        string autrePseudo = "";

        for (int i = 0; i < ligne.Length; i++)
        {
            //on initialise le premierPseudo pour chaque ligne

            char lettresPseudo = ligne[i];
            //tant qu'on est sur le premier pseudo, on l'enregistre quelque part
            if (nbPseudo == 0)
            {
                if (lettresPseudo.Equals(',') == false)
                {
                    premierPseudo += lettresPseudo;
                }
                else if (lettresPseudo.Equals(',') == true)
                {
                    nbPseudo++;
                }
            }
            //On arrive au cas de la première target et de toutes les autres
            else
            {
            }
        }
    }
}
```

Capture d'écran n°1 : Extrait de code utilisant le streamReader de façon à ce que les données soient correctement mise en forme pour le logiciel Gephi

Finalement les données étaient bien mises en forme :



The diagram illustrates the process of data transformation. At the top, a rectangular box contains raw data in CSV format. An arrow points downwards to a second rectangular box containing a structured CSV table.

Raw Data (Top Box):

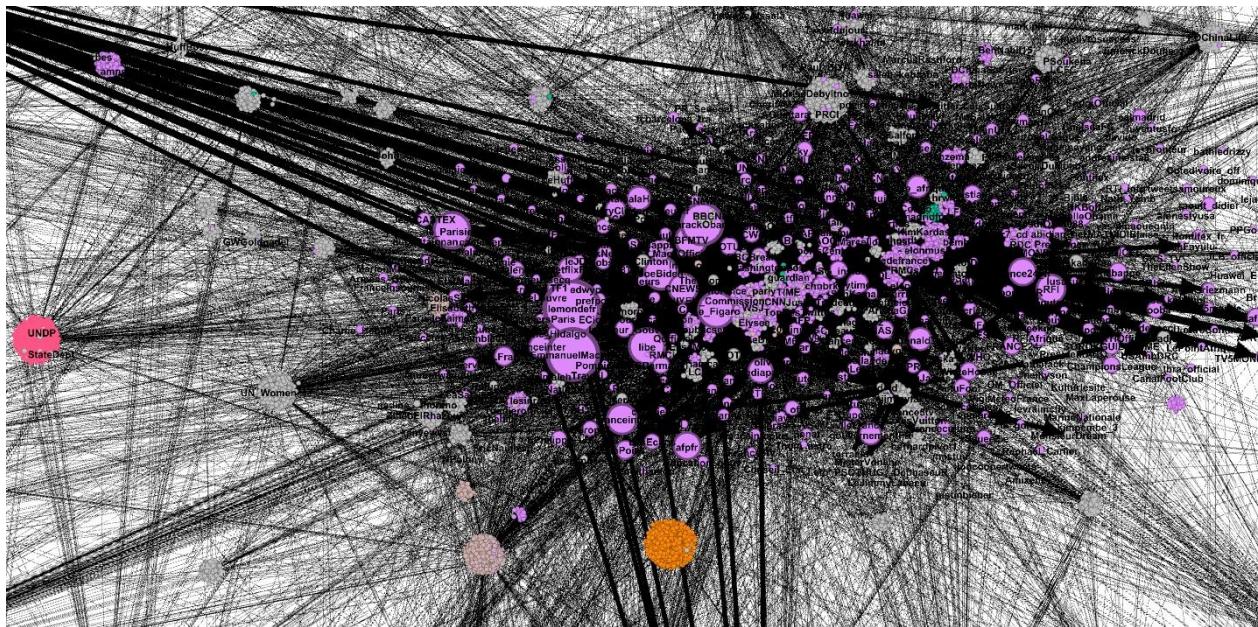
```
,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,
GUILAGUIJP,flywithkamala,SibaOnivogui2,libe,Google,YouTubeCreators,s
LePetitDuc_Web,yseultofficiel,SofiaAnastasio,PPetrucciani,GauthierToux,
```

Transformed Data (Bottom Box):

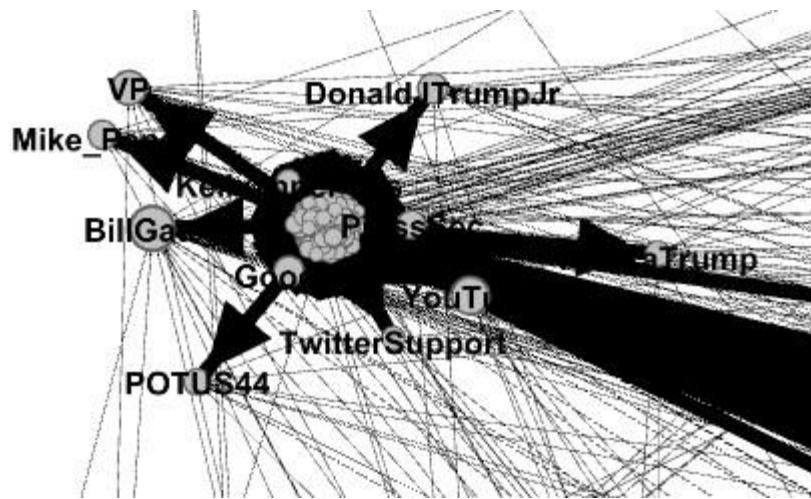
A	B
Source	Target
GUILAGUIJP	flywithkamala
GUILAGUIJP	SibaOnivogui2
GUILAGUIJP	libe
GUILAGUIJP	Google
GUILAGUIJP	YouTubeCreators
GUILAGUIJP	sixers
GUILAGUIJP	LeMonde_Afrique
GUILAGUIJP	AFRICA24TV
GUILAGUIJP	France24_fr
GUILAGUIJP	sandrine_fall
GUILAGUIJP	Abdoura70156353

Schéma 2 : Illustration de la mise en forme des données

Pour rappel, avec les abonnements de 370 utilisateurs abonnés à deux médias légèrement politisés, nous avons obtenu environ 20 millions de connexions. Résultait de cette quantité importante de données, un traitement particulièrement long sur nos machines pour la génération d'un seul graphe, avec des résultats peu pertinents et très peu lisibles et qui sortaient du cadre de notre étude tel qu'on peut le voir sur les graphiques ci-dessous.



Graphique 3 : zoom sur le graphe obtenu avec un nombre (trop) important de noeuds



Graphique 4 : Cluster de personnalités politiques américaines (notamment Mike Pence, Donald Trump et Bill Gates)

Pour essayer d'enlever les communautés non politiques et non françaises de notre graphe, nous avons alors appliqué une nouvelle méthodologie :

1. Nous avons réalisé un script qui récupère des comptes depuis le site YMobActus. YMobActus est un site internet tenu par Olivier Duprez pour "disposer d'outils de veille de la presse francophone". En analysant les activités d'un grand nombre de comptes Twitter, il dresse automatiquement des tops 50 selon la popularité de personnalités sur Twitter. Le script récupère ainsi 150 utilisateurs : le top 50 des politiques, le top 50 des journalistes politiques et le top 50 des médias en France.
2. Nous avons trié nos CSVs contenant les nœuds par ordre de poids pour récupérer 30 comptes qui ne font pas encore partie des 150 récupérés, tout en étant le plus en tête du classement possible afin de s'adapter au mieux aux utilisateurs collectés.
3. Nous avons ensuite appliqué un algorithme de tri en Python qui nous permet de retirer de nos fichiers tous les nœuds et les liens n'étant pas dans les 180 comptes d'influence politique récupérés.
4. Le fonctionnement de l'algorithme est comme tel : si une personne est abonné à un compte présent dans la liste, on le récupère, on pondère et on l'ajoute à un nouveau fichier csv.

Nous passons alors d'une grande base de données de plus de 20 000 000 de connexions, à une base de données beaucoup plus restreinte et exploitable dans le domaine de la politique et des médias spécialisés à seulement 1 206 000 connexions. De même, nous passons de 38 000 nœuds à 154 nœuds, et de 55 000 liens à 4048 liens. Ce qui nous permet d'exploiter plusieurs graphiques afin d'appréhender une réponse à notre problématique.

c) Résultats finaux

Tout d'abord nous avons étudié les différents algorithmes de spatialisation car leur interprétation dépend de la manière dont ils ont été conçus. Nous nous concentrerons sur les algorithmes Force Atlas 2 et Open Ord aux vues du nombre important de nœuds que nous avons et de l'utilisation que nous voulons en faire.

Nom de l'algorithme de spatialisation	Utilisation	Auteur	Taille du graphe supportée
Force Atlas	Analyse de complémentarité. Sert à spatialiser les nœuds d'une manière à mettre en évidence les réseaux "petit monde", et de manière plus générale, les réseaux invariants d'échelle.	Mathieu Jacomy	0 - 10 000 de noeuds
Force Atlas 2	Même utilisation que le 1, mais beaucoup plus rapide grâce à des approximations mathématiques efficaces.	Mathieu Jacomy	1 à 1 000 000 de nœuds
Yifan Hu Proportional	Utiliser pour les grands réseaux, pour une modélisation très rapide - analyse de complémentarités (si le temps n'est pas un paramètre	Yifan Hu	100 à 100 000 nœuds

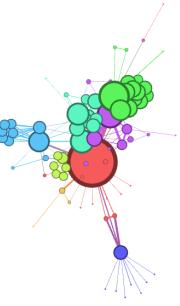
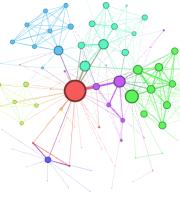
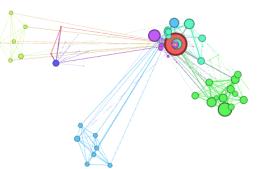
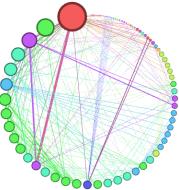
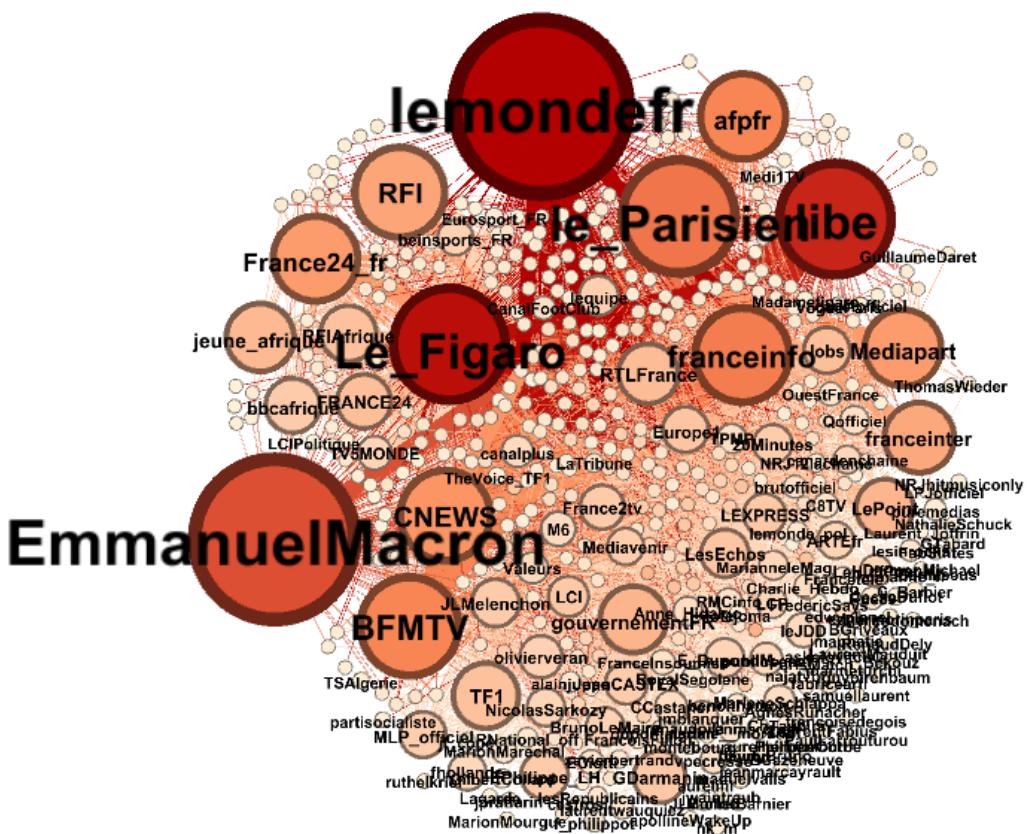
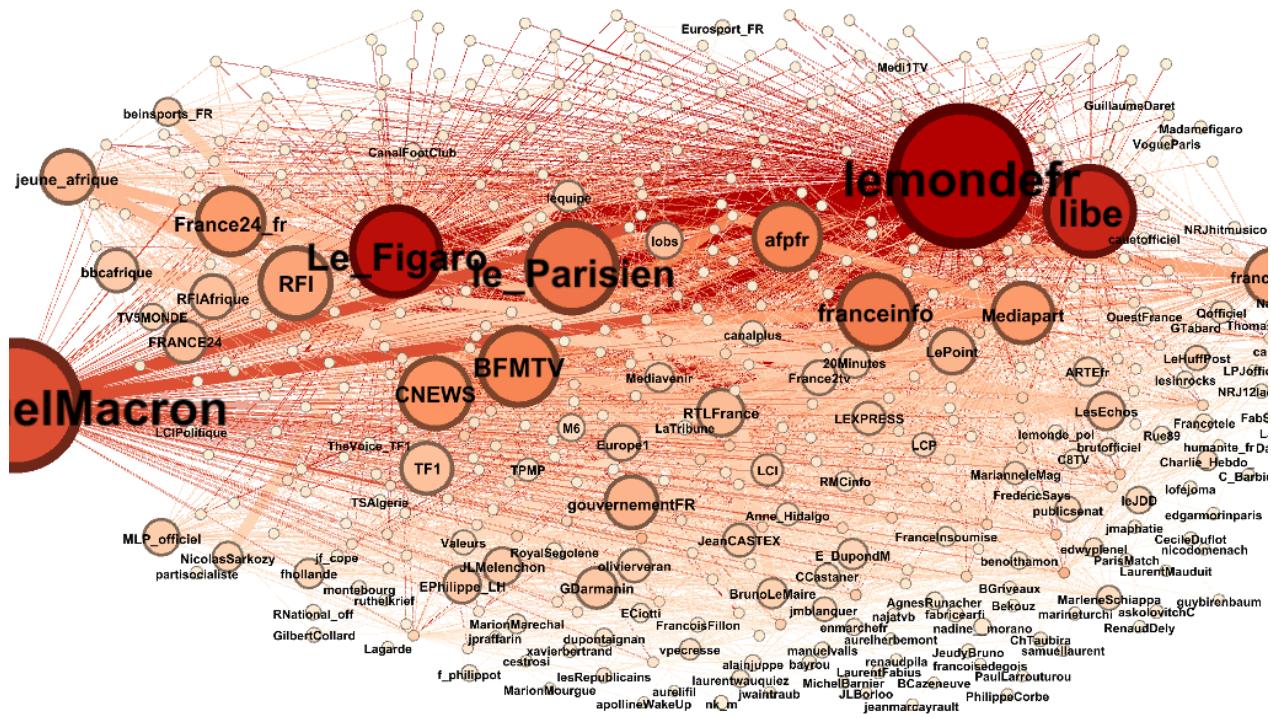
	important, Force Atlas 2 est plus adapté).		
Fruchterman Reingold 	Une visualisation devenue classique, mais un algorithme d'une complexité importante, qui la disqualifie pour l'étude de grands réseaux.	Thomas Fruchterman & Edward Reingold	1 à 1 000 nœuds
Open Ord 	Permet d'insister sur les divisions, de distinguer clairement les clusters. Open Ord est un algorithme pensé pour décrire des réseaux qui ne seraient pas des petits mondes, et dont on voudrait mettre en avant les différents groupes (clusters) constituant	S. Martin, W. M. Brown, R. Klavans et K. Boyack	100 à 1 000 000 de nœuds
Circular Layout 	Une visualisation plutôt simple pour mettre en valeur les plus gros nœuds et les liens qui relient les nœuds.	Matt Groeninger	1 à 1 000 000 de nœuds

Tableau 3 : Présentation et utilisation des différents algorithmes de spatialisation

Finalement, nous avons donc pu représenter nos données filtrées sous deux formes différentes : Open Ord et Force Atlas 2. Dans un premier temps, avec Force Atlas 2 nous avons simplement affiché notre réseau en pondérant la grosseur des noeuds par l'intensité de couleur en fonction de leur **centralité de degré** (le nombre d'abonné d'un compte).



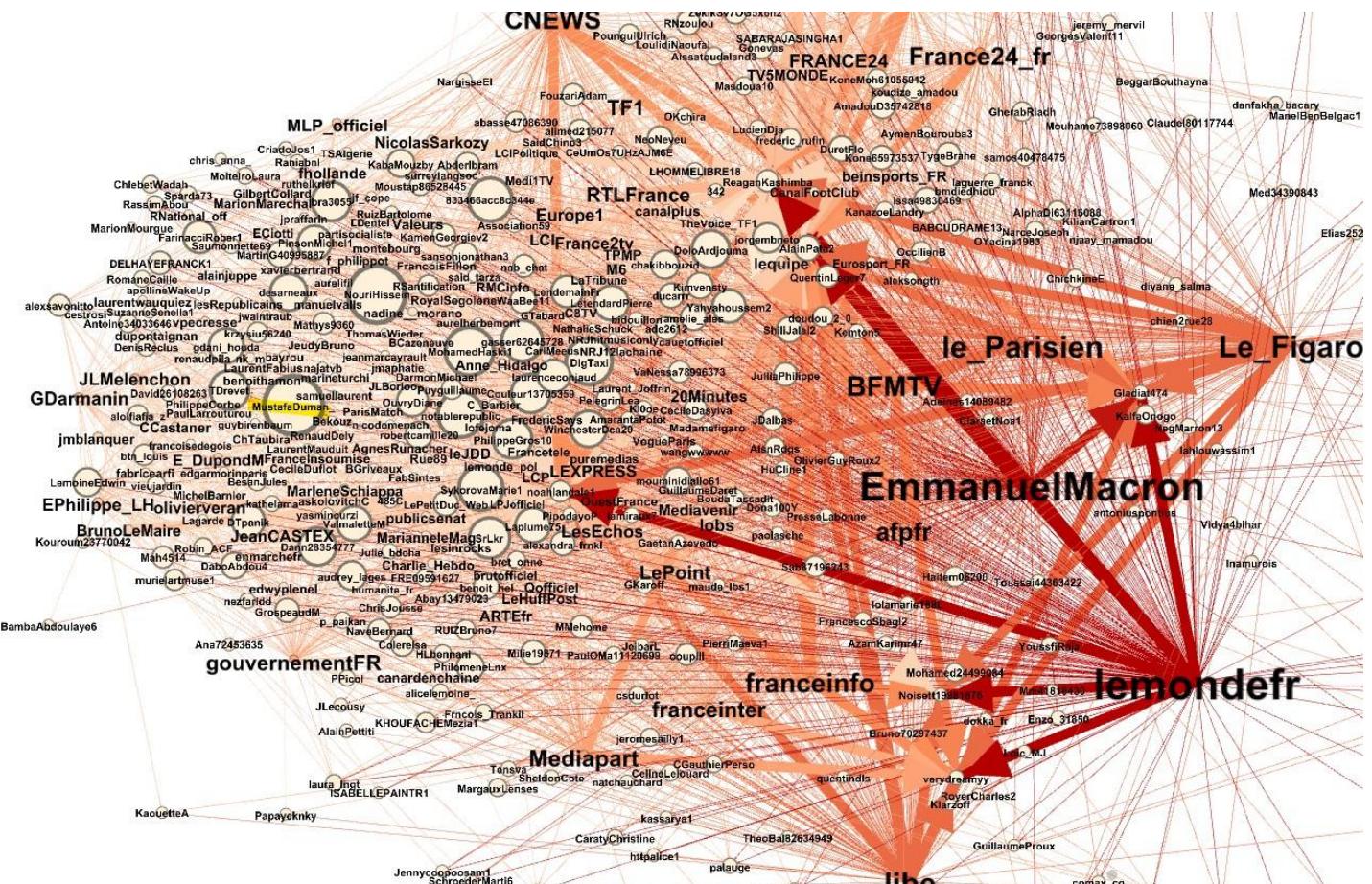
Graphique 5 : ForceAtlas2 avec centralité de degré



Graphique 6 : ForceAtlas2 avec centralité de degré et une gravité plus faible

Nous pouvons voir que le réseau étudié est particulièrement soumis à l'information venant du Monde, d'Emmanuel Macron, du Figaro et de Libération. Ces 4 Mega-hubs centralisent les abonnements et ont donc un très fort impact sur le réseau en général. Ce sont des comptes mainstream très fortement suivis. Lorsqu'ils publient quelque chose, l'information est directement communiquée au plus grand nombre.

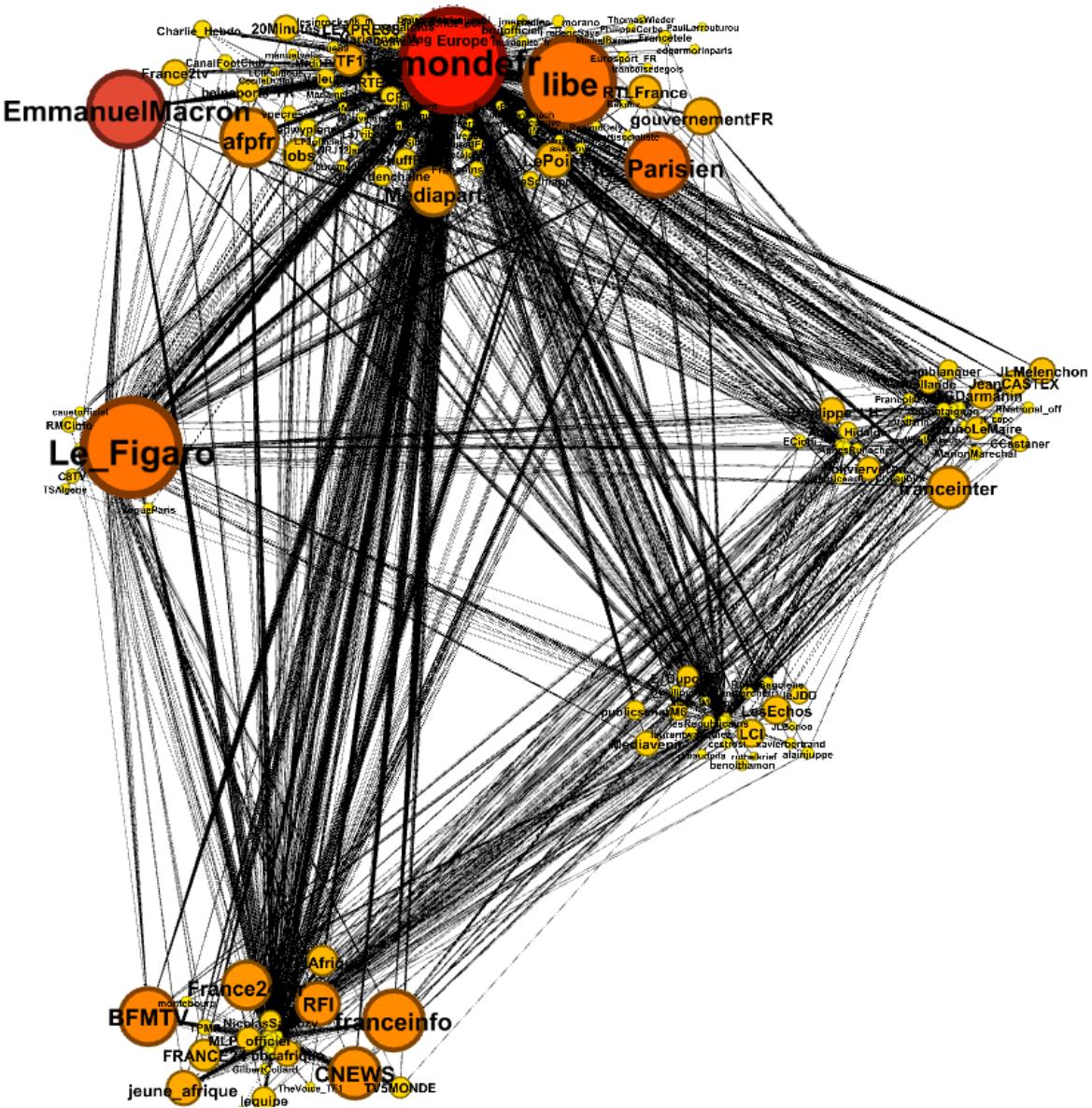
Afin de mesurer la **réputation** au sein du réseau (calculée par la *centralité d'Eigenvector*) nous avons fait un second graphe nous permettant de voir les comptes les plus centraux, c'est-à-dire ayant une forte densité de liens autour d'eux, ce sont alors des comptes secondaires, souvent méconnus, mais qui sont cruciaux dans les échanges d'informations aux seins de réseaux "petit-monde".



Graphique 7 : ForceAtlas2 avec centralité d'Eigenvector

En particulier, nous pouvons noter les comptes de Mustafa Duman (surligné) et Nourri Hissen. Ce sont des acteurs importants au sein de la twittosphère politico-média-tique. MustafaDuman_ est par exemple en L3 double-licence Droit & Histoire à la Sorbonne et, de part sa formation et son engagement politique, véhicule un nombre important de tweets politiquement marqués. Il est suivi par plus de 1000 abonnés (voir : https://twitter.com/MustafaDuman_). Lorsqu'on regarde en détail, les abonnés de ce compte appartiennent à un monde universitaire et ont des connaissances fines en politique, cela pourrait justifier sa réputation sur le réseau.

Afin d'analyser la **capacité à fédérer** un cluster (bridging), nous avons aussi représenté nos données avec l'algorithme Open Ord.

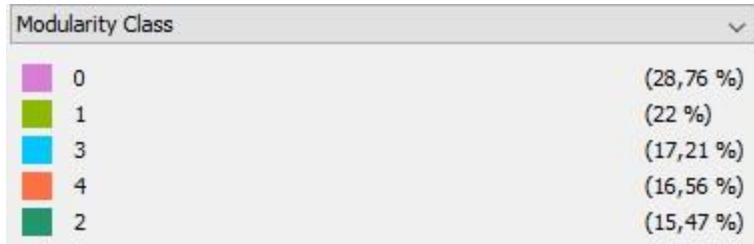


Graphique 8 : Open Ord avec centralité de degré

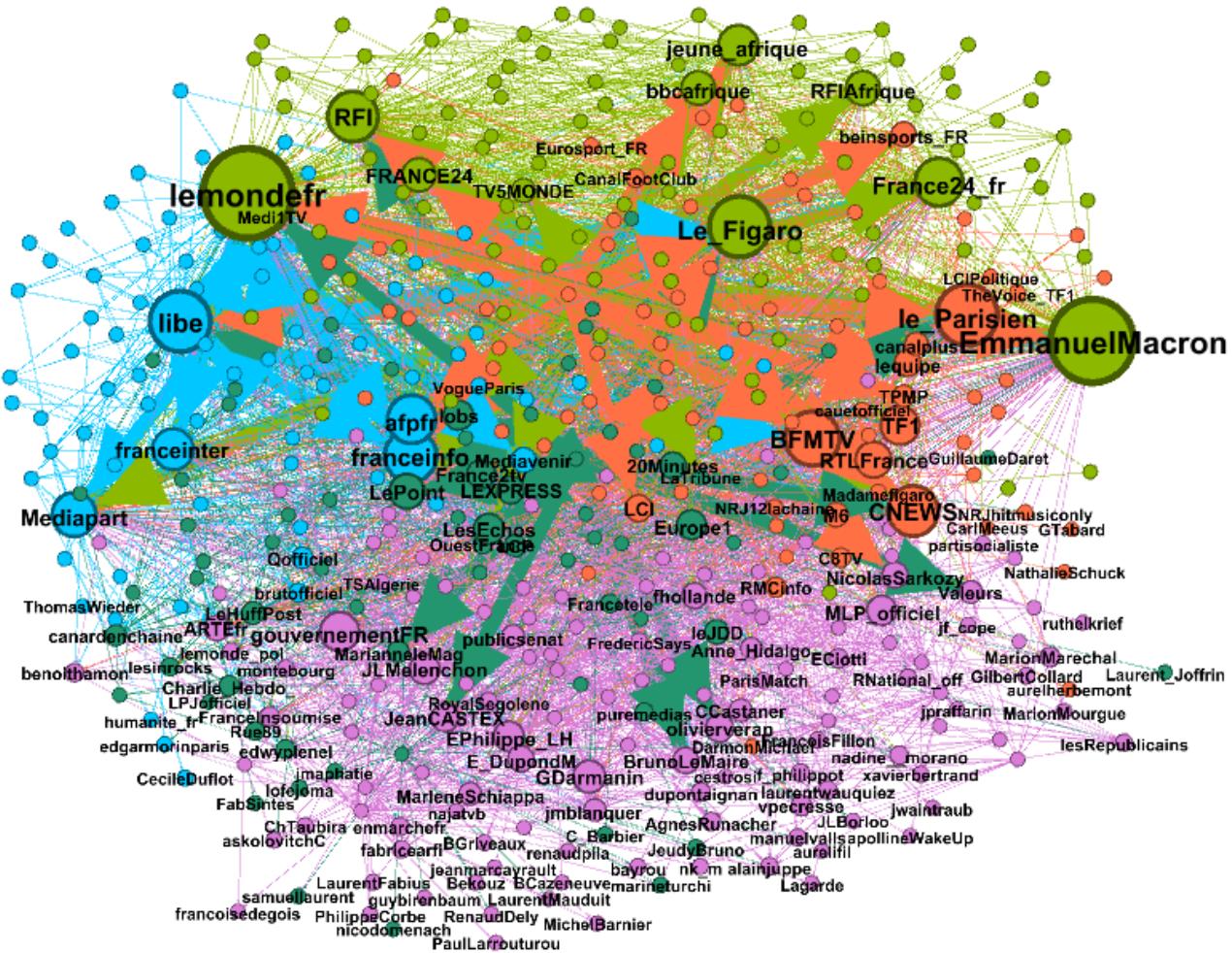
Nous y avons identifié 5 clusters différents avec des comptes particulièrement importants dans chacun des clusters. On observe par exemple que le Figaro est au centre d'un petit cluster signifiant sa particularité sur la scène médiatique, tandis que Libération est en périphérie du journal Le Monde. Les autres clusters ont pour dominants

: France Inter / Les Echos / France info. Chacun de ces médias étant suffisamment différent des autres. Il est intéressant de ne pas observer de répartition en fonction de la couleur politique (notamment en observant les personnalités politiques gravitant autour des noyaux des clusters). Cela peut être interprété comme la manifestation de certains liens de régularité entre certaines personnalités et certains médias. Par exemple, il est notable qu'autour de France Inter nous puissions retrouver les têtes de partis, souvent mis dans l'actualité de ce média (François Fillon, Jean-Luc Mélenchon, Benoît Hamon). Tandis qu'au contraire Sarkozy et Marine le Pen sont près de la sphère BFMTV, ce qui peut être lié aux récentes mais nombreuses polémiques, souvent débattues sur BFMTV.

Enfin, nous avons procédé à la mise en place d'un filtre de couleur nous permettant de mettre au jour les **modularités**, c'est à dire les appartenances de groupes sur une spatialisation Force Atlas 2. Il en ressort 5 grandes classes (ou familles).



Graphique 9 :Pourcentage de répartition des familles



Graphique 10 : Force Atlas avec modularité et centralité de degré

Nous pouvons identifier :

- **Violet** : la famille des personnalités politiques (très suivie par la famille des vert foncé). Ils ne suivent pas particulièrement les autres médias.
- **Vert foncé** : les médias spécialisés (de niche) comme Le Canard enchaîné, le JDD, Public Sénat, Rue89. Les comptes affiliés à cette famille suivent surtout les comptes des personnalités politiques, ce qui est logique compte tenu de leur volonté de précision, mais aussi les médias neutres et mainstream.
- **Orange** : la famille des médias plutôt axés sur le divertissement comme TF1, CNEWS, TPMP, l'Equipe, TheVoice. Les comptes affiliés aux oranges sont très

pointés vers les verts clairs, les médias mainstream avec une grande portée internationale.

- **Vert clair :** les médias mainstream avec une grande portée internationale comme le Monde, France 24, TV5 Monde, RFI Afrique. Ce sont eux qui sont suivis. On observe que les comptes affiliés à cette famille ont peu de contact avec l'extérieur et sont donc très mono-source.
- **Bleu :** Les médias de référence journalistique qu'on peut qualifier de neutres comme France Info, France Inter ou l'AFP. Généralement, les comptes affiliés à la famille bleu sont eux aussi relativement mono-source.

Nous pouvons résumer cette dernière analyse en affirmant que, dans le réseau étudié, les individus suivant des médias de références journalistiques, ou des médias à portée internationale sont souvent peu portés vers les autres sources d'information. Au contraire, les individus et comptes suivant les médias spécialisés sont particulièrement attentifs à la diversification de leurs sources d'information comme le montre les flèches vertes foncées. Pour ce qui est des compte suivant des médias plus axés sur le divertissement, il sont aussi très proches des médias dits *mainstreams*. Il est alors possible de lire le graphe en termes de flux d'information et imaginer une chaîne de Markov d'un tweet passant de famille en famille en étant retweeté : personnalité politique → médias spécialisés → médias neutres → médias mainstream/médias axé sur le divertissement.

IV. GESP

a) Répartition du travail et coordination

Nous avons opté pour une gestion de projet dite Agile. Ainsi, tout le monde a pu être porteur d'initiative au fur et à mesure de l'avancée du projet. Les élèves de deuxième année ont notamment mis les cours de gestion de projet Agile suivis au deuxième

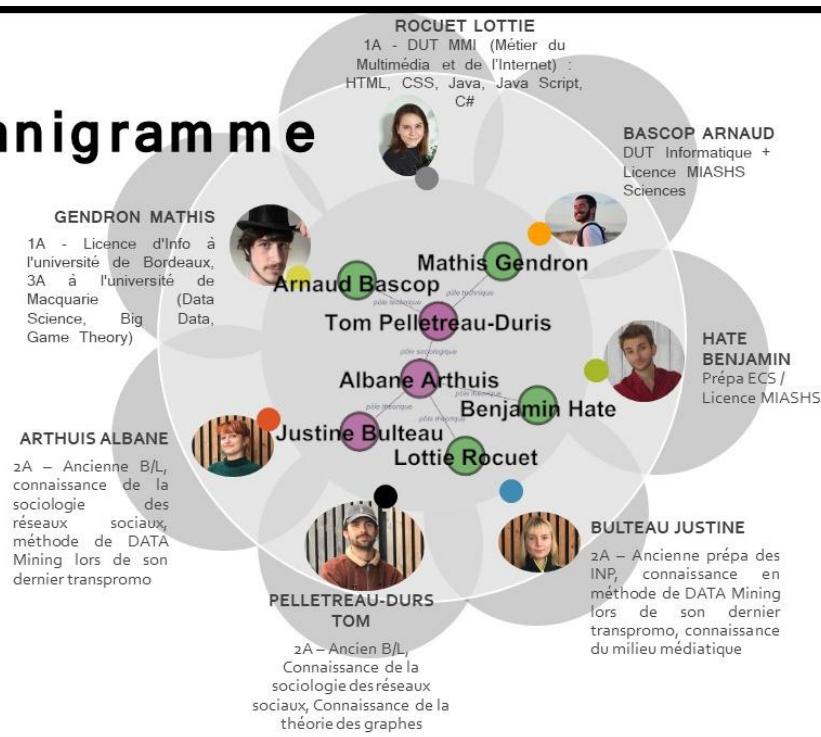
semestre de première année à disposition. Chacun a ainsi pu se former à ces techniques de gestion de projet.

Néanmoins, deux pôles se sont dessinés en fonction des affinités de chacun et chacune envers les techniques déployées. Il y a donc un pôle technique, s'occupant notamment du code pour le *data mining*, et un pôle théorique, s'intéressant aux définitions inhérentes à la compréhension du sujet à l'établissement des critères pour la sélection des entités ou individus pour le *data mining* et à la production d'éléments de diffusion et de supports pour les *reporting*.

Le pôle théorique est tenu au courant à toutes les réunions des avancées du code et les comprend (possibilité de se référer à des MOOC au besoin). De plus, l'entièreté du groupe a débattu des techniques, langages et logiciels à employer. Chacun a ainsi pu s'initier ou développer ses compétences en Python et Jupiter Notebook.

De plus, les idées et décisions prises pendant les réunions aboutissent à une liste de tâches réparties à la fin de chacune des réunions. Cette répartition se fait selon le désir ou la disponibilité de chacun des membres du groupe.

Organigramme



Organigramme de l'équipe AIRe

b) Outils

L'organisation interne dépend de la maîtrise et l'utilisation de différents outils par les différents membres. Bien que le déroulement du projet était établi avant le recrutement des élèves de première année, leurs savoir-faire nous ont permis d'orienter le sujet.

Afin d'assurer une bonne communication, nous utilisons Zoom pour les réunions et Messenger pour discuter entre celles-ci.

Pour organiser notre projet et assurer son suivi, nous utilisons Trello qui recense les tâches à faire, celles en cours de traitement et celles achevées. Nous stockons nos ressources et documents sur Google Drive : compte-rendus des réunions, les livrables, etc. GitHub nous a également permis de stocker et partager le code, constituant la partie technique du projet.

c) Risques

Pour ce projet, voici les risques rencontrables :

- Les données et formations nécessaires aux différentes étapes ne sont pas accessibles gratuitement en ligne. Si les membres du projet ne disposent pas des ressources nécessaires au développement de celui-ci, le projet est compromis.
- Accentuation des mesures de protection contre la Covid-19. Ces mesures nous empêcheraient d'avoir accès au matériel de l'école, dont les performances sont plus élevées que les outils dont disposent les membres du groupe.
- Les données dont nous avons besoin pour construire nos graphes ne sont pas accessibles en quantité suffisante. Cela peut ralentir, voire faire changer radicalement le projet. Ici nous avons décidé de prélever les données Twitter car

elles sont accessibles, un changement de la politique de mise à disposition ou de condition d'exploitation de ces données rentre dans le cadre de ce risque.

- Le logiciel de construction de graphe : Gephi, devient payant ou inaccessible. Cela demanderait une adaptation rapide, il faudrait trouver un logiciel effectuant le même service et se former dessus en peu de temps.
- Personne ne veut relayer notre travail. Un des objectifs de ce projet est de partager la théorie de la clusterisation sur les réseaux sociaux en montrant un exemple de celle-ci. Si le projet n'intéresse personne en dehors de notre équipe, le projet perd de son sens.

d) Planning

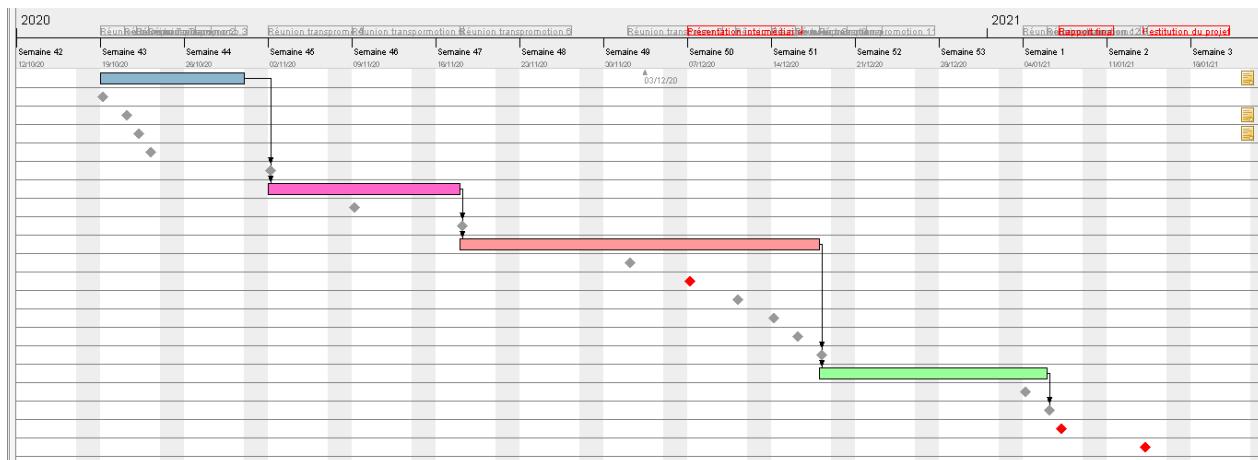
Au début du projet nous avons déterminé quatre étapes successives qui devaient découper ce projet en grands axes : la réparation des savoirs et de l'équipe, le data mining, la production de graphe et enfin l'analyse de ceux-ci et le partage de nos conclusions. Détailons ces étapes :

- **L'étape 1** s'est déroulée sur 2 semaines, aussi bien en théorie qu'en pratique. Elle correspond au *forming* et à la constitution de l'état de l'art. Nous avons pris en main les théories gravitant autour du projet et décidé des techniques à employer. À savoir : la science des réseaux et les principes du *data mining*, le choix d'un langage, ici Python et Jupiter Notebook et le logiciel Gephi.
- **L'étape 2** constitue la construction de la base de données et le *data mining* (soit la récupération des données en français). La date finale initialement prévue a été reportée quelque peu lors de cette étape. En effet, nous avons rencontré des limites dans notre récupération de données : Twitter qui limite le nombre de données récupérables par jour à 1000, et par heure à 150. De plus, nous sommes dépendants de cette étape pour la suite du projet. La collecte de données a donc

pris plus de temps que prévu, nous avons alors ajouté deux semaines supplémentaires à cette étape afin d'avoir un nombre minimum décent de données.

- **L'étape 3** correspond à la construction de graphes et à leur analyse. Nous avons eu des résultats dès le 25/11, nous avons donc commencé à constituer des réseaux peu de temps après cette date.
- **L'étape 4** consiste en le partage de nos résultats et la production d'un rapport en tant que livrable final.

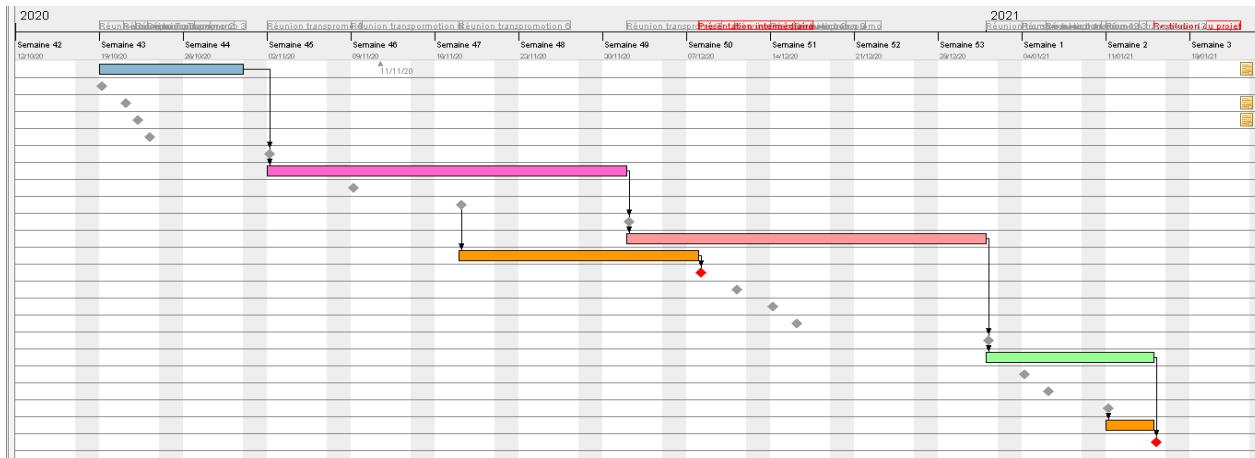
Conjointement, nous avions établi un premier planning pour avoir une idée de quels objectifs nous semblaient atteignables à telle ou telle date. Ayant l'habitude des diagrammes de GANTT, nous avons utilisé cet outil, présenté ci-dessous :



GANTT du planning initial

Les quatre différentes barres représentent les étapes du projet. Les losanges gris représentent les créneaux de transpromotion définis dans l'emploi du temps, nous les avons utilisés pour faire des réunions afin de discuter de l'orientation du projet ou pour que chacun présente son travail aux autres. Les losanges rouges représentent les jalons

de reporting, soit la présentation intermédiaire, la livraison du livrable final et sa présentation. Comparons ces étapes au planning réel :



GANTT du planning réel

Initialement, nous avions pensé dédié la majeure partie de notre temps à l'étape 3 : la construction et l'analyse des graphes. Toutefois, l'étape 2 : le data mining, essentielle à l'étape 3, nous a pris plus de temps que prévu. En effet, nous avons été limité dans notre récolte de données par les conditions d'extraction de Twitter et par la puissance de calcul de nos ordinateurs. En raison de la pandémie, nous n'avions pas accès aux machines de l'école.

Dans les faits l'étape 3 nous a pris un mois, soit la durée exacte prévue dans le planning initial. C'est donc l'étape 4 : la préparation d'un compte rendu et le partage de l'information qui a été réduite par le retard pris lors de l'étape 2. Ainsi, ce sont nos objectifs de diffusion de l'information qui n'ont pas été totalement remplis.

De plus, nous avons rajouté des tâches : les vidéos de présentation intermédiaire et finale. Elles sont représentées par les barres oranges sur le diagramme de GANTT.

e) Biais

Lors de notre analyse nous courons le risque d'être soumis à certains biais cognitifs. Bien que notre sujet prête à la vigilance sur ce point, certains biais sont

inévitables, notamment pour la constitution du jeu de données et notre rapport à Twitter et au sujet en général.

Certains d'entre nous sont des utilisateurs de Twitter, cela porte un avantage en termes de connaissances de notre sujet d'étude mais peut également être dangereux car chacun a une expérience subjective du réseau et ne peut se porter expert des comportements entrepris sur cette plateforme.

De plus, les réseaux sociaux et leur influence sont un sujet fréquemment développé par des médias plus ou moins objectifs. Il a fallu que chacun d'entre nous fasse en sorte de déconstruire ses présupposés sur le sujet afin de ne pas biaiser les méthodes ou l'analyse. Partir de fausses informations nous aurait porté à un décalage de notre sujet de la mise en évidence d'un effet bulle déjà démontré par xx, à une nouvelle étude choc. Nous ne prétendons pas produire un papier de recherche scientifique mais bien sélectionner un échantillon et mettre en avant des théories préétablies.

La sélection des données, de fait, est une étape sensible en termes de biais. D'abord parce que l'équation *twittosphère politico-média*française ↔ *sphère politico-média*française n'est pas évidente. Nous faisons le choix d'assumer cette représentativité approximative. Mais aussi parce qu'il nous a fallu choisir une liste de comptes de base. N'ayant pas trouvé de catégories "Politique" qui aurait pu être construite par Twitter pour signaler les comptes politiques individuels ou de médias, ce fut à nous de choisir vers quels comptes nous tourner. Nous avons alors fait le choix de sélectionner deux journaux déjà existants au format papier, l'un plutôt de droite et l'un plutôt de gauche sur l'échiquier politique. Ces deux journaux nous semblent équivalents et semblent soulever des intérêts divergents et être consultés par des communautés distinctes. Il nous a semblé que ce choix, même s'il est relativement subjectif, est l'un des plus pertinents que nous pouvions faire. Ici, nous avons évoqué l'idée d'effectuer une étude statistique afin de savoir si notre jeu de données (les followers de ces deux

journaux) était représentatif. Toutefois, notre objectif étant de réaliser un graphe à partir d'informations issues de Twitter et dans l'idéal d'illustrer, de présenter un exemple de théories déjà établies, il ne nous a pas semblé nécessaire d'effectuer cette étude statistique.

f) Difficultés rencontrées et adaptation à la crise sanitaire

La première difficulté que nous avons rencontré fut que Twitter limite la quantité de données extractibles. Cela a considérablement ralenti notre data mining (étape 2).

Par la suite, l'accentuation des mesures gouvernementales nous a également ralenti. En effet, l'extraction de données et la production de graphe demande une certaine puissance de calcul. Nos ordinateurs personnels nous ont permis d'avoir des résultats mais nous serions allés plus vite avec des outils plus performants : ordinateurs de l'école (en puissance mais aussi en nombre).

En ce qui concerne la GESP, la crise sanitaire n'a pas eu un grand impact sur notre dynamique de groupe. Les réunions ont permis à chacun de rester en lien et d'être à jour sur le projet. La production de deux vidéos en lieu et place de présentation en présentiel n'est toutefois pas à négliger, nous avons dû ajouter ces tâches à notre planning.

Enfin, le temps pris par les difficultés explicitées précédemment nous ont amenés à réduire le temps dévolu à la diffusion du projet.

CONCLUSION

Les différentes analyses produites mettent en évidence le pouvoir de réseau de méga-hubs bien connus. Le Monde, Emmanuel Macron, Libération, Le Figaro, Le Parisien, France Info, autant de médias traitant de politique, concentrant les regards et les abonnements. Ces méga-hubs ont une capacité à émettre de l'information à un très grand nombre de personnes et sont dès lors des producteurs d'opinion de premier ordre. De plus, les graphes mettent en évidence l'apparition de clusters. Le Figaro par exemple semble au centre d'un cluster, possédant une communauté différenciée des autres médias, à la fois plus restreinte et plus fidèle. Nos interprétations nous ont aussi permis de dégager des comptes particulièrement réputés et centraux dans les échanges d'informations. Ce sont ces comptes qui font le lien entre plusieurs autres comptes et fournissent des informations souvent précises (*bridging*). Enfin, nous avons pu mettre en lumière les différentes familles de la twittosphère politico-médiatique et voir les liens qu'ils entretenaient entre eux. Il en ressort que les individus d'abord abonnés aux grands médias sont fortement mono-sourcés, alors que les individus abonnés aux médias spécialisés et aux comptes des personnalités politiques sont plus pluri-sourcés.

Nous avons donc rempli notre objectif de mise en évidence des effets de polarisation sur le RSN Twitter. Toutefois, la volonté de diffusion de nos travaux n'a pas été satisfaite. Nous n'écartons donc pas la possibilité de partager et faire diffuser notre projet afin de remplir cet objectif, bien que cela se fasse hors délai.

D'un point de vue pédagogique, chacun a pu découvrir les théories liées aux graphes, à la science des réseaux et au data mining. Ce projet Transpromotion fut le cadre idéal pour un partage de connaissances des deuxièmes années vers les premières années comme des premières années vers les deuxièmes années en ce qui concerne l'utilisation de logiciels, de langages et de pratiques.

ANNEXE

- Lexique commun et fiche de travail :
<https://docs.google.com/document/d/1FrCP9r3lOK3DIrb7d97CCIQ5riBYZ4C9Z6gHYTbNJMc/edit?usp=sharing>
- Infographie de la méthodologie de récupération de données :
<https://drive.google.com/file/d/1XSJZDR4BUYq0mKlxOU7yMADyP8gzfIGA/view?usp=sharing>
- Document Jupiter NoteBook de Data Mining :
https://drive.google.com/file/d/1rg7vDNP-HZSQT28V_AEATxsynQxP4E1/view?usp=sharing
- Document Jupiter Notebook de Tri des CSV :
https://drive.google.com/file/d/1tQ3UOUP9WKLZd8_VvQ9St6Qv_HaZiZHq/view?usp=sharing
- Code de l'algorithme de restructuration des données :
<https://drive.google.com/drive/folders/1Fb1px6pPlDHvbOBHDL6rsIxL692yza8z?usp=sharing>

Lien vers le GitHub du projet :

https://github.com/MatGendron/Transpromo_AIRe

BIBLIOGRAPHIE

Repère sur la sociologie des réseaux sociaux :

Sociologie des réseaux sociaux, (2004), Pierre Mercklé, La découverte, Repère 2011

Article de Nishi sur l'Inégalité et la visibilité de la richesse dans les réseaux sociaux, c'est à dire que la connaissance de la structure du réseau par l'ensemble de ses participants auto-régule les inégalités d'accès aux ressources :

Nishi, A., Shirado, H., Rand, D. et al. Inequality and visibility of wealth in experimental social networks. *Nature* 526, 426–429 (2015).
<https://doi.org/10.1038/nature15392>

Biais de l'opinion par une polarisation accentuée par les réseaux sociaux:

Sîrbu A, Pedreschi D, Giannotti F, Kertész J (2019) Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. PLoS ONE 14(3): e0213246. <https://doi.org/10.1371/journal.pone.0213246>

Le réseau d'influences politiques :

Gaumont, N., Panahi, M., & Chavalarias, D. (2018). PloS one, 13(9).

L'étude du réseau d'une organisation terroriste :

Yang, C. C., Liu, N., & Sageman, M. (2006). Springer, Berlin, Heidelberg.

L'article de Paul Erdős sur les réseaux aléatoires :

Erdős, P.; Rényi, A. (1959). *Publicationes Mathematicae*. 6: 290–297.

La publication de Milgram sur les six degrés de séparation :

Travers, Jeffrey, and Stanley Milgram, *Sociometry* 32(4, Dec. 1969):425–443

L'article de Duncan Watts sur le réseau d'acteurs :

Watts, D. J., & Strogatz, S. H. (1998). *Nature*, 393(6684), 440.

Les deux livres de Barabasi qui ont été la source d'information principal sur la science des réseaux :

Barabási, A. L. (2003), *Linked: The new science of networks*.

Barabási, A. L. (2010), *Bursts*

La publication de Barabasi sur le mécanisme qui conduit à la formation des hubs :

Barabási, A. L., & Albert, R. (1999). *Science*, 286(5439), 509-512.

L'étude du réseau des échanges entre les musées :

Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C., & Barabási, A. L. (2018).
Science, 362(6416), 825-829.

WEBOGRAPHIE

Chaîne youtube *Fouloscopie* : épisode sur la science des réseaux sociaux avec l'exemple de l'analyse des chaînes youtube :

https://www.youtube.com/watch?v=UX7YQ6m2r_o&t=1s

L'étude du réseau des découvertes scientifiques :

<https://www.nature.com/articles/d41586-019-03308-7>

Une splendide illustration ici : <https://youtu.be/GW4s58u8PZo>

Sociologie des réseaux sociaux : les auteurs et concepts les plus importants

<https://pierremerckle.fr/2011/11/%C2%AB-carte-blanche-%C2%BB-du-monde-les-reseaux-sociaux-contre-les-classes-sociales-pour-en-savoir-un-peu-plus%E2%80%A6/>

Algorithme de spatialisation :

Liste des algorithmes existants :

http://www-igm.univ-mly.fr/~dr/XPOSE2012/visualisation_de_graphes/algorithmes.html

Explication mathématique et histoire des algorithmes de spatialisation :

<https://cedric.cnam.fr/vertigo/Cours/RCP216/coursVisuGraphes.html>

Démarche de l'analyse des graphes :

<https://graal.hypotheses.org/716>

<https://graal.hypotheses.org/729>

<https://graal.hypotheses.org/758>

Méthodologie analytique :

<https://master-iesc-angers.com/utilisation-du-logiciel-gephi-pour-lanalyse-cartographique/>

What do we see when we look at networks ? (de Mathieu Jacomy)

<https://arxiv.org/ftp/arxiv/papers/1905/1905.02202.pdf>

Site internet ymobactus pour recenser les comptes politiques les plus suivis sur twitter

<http://ymobactus.miaouw.net/>