Une approche pour estimer l'influence dans les réseaux complexes : application au réseau social Twitter

Lobna Azaza

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UNIVERSITÉ DE TUNIS
INSTITUT SUPÉRIEUR
DE GESTION DE TUNIS

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE BOURGOGNE EN COTUTELLE INTERNATIONALE AVEC L'UNIVERSITÉ DE TUNIS

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

LOBNA AZAZA

Une approche pour estimer l'influence dans les réseaux complexes Application au réseau social Twitter

Thèse présentée et soutenue à Dijon, le 23 mai 2019

Composition du Jury :

BENSLIMANE DJAMAL	Professeur, Université de Lyon I	Président
COMYN-WATTIAU ISABELLE	Professeur, ESSEC	Rapporteur
LATIRI CHIRAZ	Professeur, Université de la Manouba	Rapporteur
CHERIFI HOCINE	Professeur, Université de Bourgogne	Examinateur
RALYTÉ JOLITA	Maître d'enseignement et de recherche, Université de Genève	Examinateur
FAIZ RIM	Professeur, Université de Carthage	Directeur de thèse
LECLERCQ ÉRIC	Maître de Conférences, Université de Bourgogne	Co-encadrant de thèse
SAVONNET MARINETTE	Maître de Conférences HDR, Université de Bourgogne	Directeur de thèse

école doctorale **sciences pour l'ingénieur et microtechniques**

Titre : Une approche pour estimer l'influence dans les réseaux complexes Application au réseau social Twitter

Mots-clés : Influence, Réseaux sociaux, *Twitter*, Réseaux multiplexes, Théorie des fonctions de croyance

Résumé :

L'étude de l'influence sur les réseaux sociaux et en particulier *Twitter* est un sujet de recherche intense. La détection des utilisateurs influents dans un réseau est une clé de succès pour parvenir à une diffusion d'information à large échelle et à faible coût, ce qui s'avère très utile dans le marketing ou les campagnes politiques. Dans cette thèse, nous proposons une nouvelle approche qui tient compte de la variété des relations entre utilisateurs afin d'estimer l'influence dans les réseaux sociaux tels que *Twitter*. Nous modélisons *Twitter* comme un réseau multiplexe hétérogène où les utilisateurs, les *tweets* et les objets représentent les nœuds, et les liens modélisent les différentes relations entre eux (par exemple, *retweets*, *mentions* et *réponses*). Le PageRank multiplexe est appliqué aux données issues de deux corpus relatifs au domaine politique pour classer les candidats selon leur influence. Si le classement des candidats reflète la réalité, les scores de PageRank multiplexe sont difficiles à interpréter car ils sont très proches les uns des autres. Ainsi, nous voulons aller au-delà d'une mesure quantitative et nous explorons comment les différentes relations entre les nœuds du réseau peuvent déterminer un degré d'influence pondéré par une estimation de la crédibilité. Nous proposons une approche, *TwitBelief*, basée sur la règle de combinaison conjonctive de la théorie des fonctions de croyance qui permet de combiner différents types de relations tout en exprimant l'incertitude sur leur importance relative. Nous expérimentons *TwitBelief* sur une grande quantité de données collectées lors des élections européennes de 2014 et de l'élection présidentielle française de 2017 et nous déterminons les candidats les plus influents. Les résultats montrent que notre modèle est suffisamment flexible pour répondre aux besoins des spécialistes en sciences sociales et que l'utilisation de la théorie des fonctions de croyances est pertinente pour traiter des relations multiples. Nous évaluons également l'approche sur l'ensemble de données CLEF RepLab 2014 et montrons que notre approche conduit à des résultats significatifs. Nous proposons aussi deux extensions de *TwitBelief* traitant le contenu des *tweets*. La première est l'estimation de la polarisation de l'influence sur le réseau *Twitter* en utilisant l'analyse des sentiments avec l'algorithme des forêts d'arbres décisionnels. La deuxième extension est la catégorisation des styles de communication dans *Twitter*, il s'agit de déterminer si le style de communication des utilisateurs de *Twitter* est informatif, interactif ou équilibré.

Title: Une approche pour estimer l'influence dans les réseaux complexes Application au réseau social Twitter

Keywords: Influence, Social Network, *Twitter*, Multiplex network, Belief functions theory

Abstract:

Influence in complex networks and in particular *Twitter* has become recently a hot research topic. Detecting most influential users leads to reach a large-scale information diffusion area at low cost, something very useful in marketing or political campaigns. In this thesis, we propose a new approach that considers the several relations between users in order to assess influence in complex networks such as *Twitter*. We model *Twitter* as a multiplex heterogeneous network where users, *tweets* and objects are represented by nodes, and links model the different relations between them (e.g., *retweets*, *mentions*, and *replies*).The multiplex PageRank is applied to data from two corpuses in the political field to rank candidates according to their influence. Even though the candidates' ranking reflects the reality, the multiplex PageRank scores are difficult to interpret because they are very close to each other. Thus, we want to go beyond a quantitative measure and we explore how relations between nodes in the network could reveal about the influence and propose *TwitBelief*, an approach to assess weighted influence of a certain node. This is based on the conjunctive combination rule from the belief functions theory that allow to combine different types of relations while expressing uncertainty about their importance weights. We experiment *TwitBelief* on a large amount of data gathered from *Twitter* during the European Elections 2014 and the French 2017 elections and deduce top influential candidates. The results show that our model is flexible enough to consider multiple interactions combination according to social scientists needs or requirements and that the numerical results of the belief theory are accurate. We also evaluate the approach over the CLEF RepLab 2014 data set and show that our approach leads to quite interesting results. We also propose two extensions of *TwitBelief* in order to consider the *tweets* content. The first is the estimation of polarized influence in *Twitter* network. In this extension, sentiment analysis of the *tweets* with the algorithm of forest decision trees allows to determine the influence polarity. The second extension is the categorization of communication styles in *Twitter*, it determines whether the communication style of *Twitter* users is informative, interactive or balanced.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besanon, France

REMERCIEMENTS

Je voudrais tout d'abord témoigner mes sincères remerciements et ma profonde reconnaissance à mes encadrants du laboratoire LIB, Madame Marinette Savonnet et Monsieur Éric Leclercq, qui m'ont accueilli au sein de l'équipe Science des Données, et qui m'ont encadré durant la réalisation de cette thèse. Madame Savonnet et Monsieur Leclercq ont été toujours présents pour écouter mes propositions et répondre à mes questions. Ils m'ont orienté constamment vers la bonne direction. Je les remercie beaucoup pour la qualité de leur encadrement, leur soutien moral, leurs conseils et les qualités humaines qu'ils possèdent. Je resterai toujours reconnaissante pour l'attention qu'ils ont apportée à mon regard.

Mes remerciements s'adressent également à mon encadrante du laboratoire LARODEC, Madame Rim Faiz, pour l'aide et l'encadrement qu'elle m'a apportés durant cette thèse.

J'adresse également mes remerciements les plus distingués à Messieurs les membres du jury qui ont bien accepté d'évaluer mon travail.

Durant cette thèse, j'ai partagé des moments exceptionnels avec les membres des laboratoires LIB et LARODEC, je leur exprime ma profonde sympathie et remerciement, en particulier à Monsieur Sergey Kirgizov. Aussi une pensée spéciale à Monsieur Malek Jebabli qui nous a quitté très tôt, paix à son âme.

Je tiens aussi à remercier toutes les personnes avec qui j'ai travaillé, et à tous ceux qui ont participé de près ou de loin à l'aboutissement de mon travail de thèse.

Finalement, je remercie mon mari, mes parents, mes frères et soeurs, ainsi que ma belle famille de m'avoir soutenue et de m'avoir fourni un environnement merveilleux dans lequel j'ai pu progresser. Je remercie également tous mes amis pour leur soutien, leur aide et leurs conseils lors de toutes les périodes difficiles.

Enfin, je dédie cette thèse à ma famille et à ma chère fille Jasmine.

SOMMAIRE

I

CONTEXTE ET PROBLÉMATIQUE

1

INTRODUCTION

Dans ce chapitre introductif, nous présentons tout d'abord le cadre de la thèse, son contexte général et ses motivations. Nous présentons ensuite nos objectifs et contributions ainsi que la description de la structure du manuscrit.

1.1/ CADRE DE LA THÈSE

1.1.1/ CONTEXTE GÉNÉRAL

Le Web évolue à un rythme exponentiel, aujourd'hui, il joue un rôle de plus en plus important en raison de ses ressources riches et variées. Dans son édition 2018 du Digital Economy Compass, Statista, un site de statistiques et études de marché [1], montre qu'une minute suffit pour que des millions de données à travers le monde soient générées. Par exemple, en une minute, 3,8 millions de requêtes *Google* sont enregistrées, 29 millions de messages *WhatsApp* traités, 350 000 *tweets* postés, 243 000 photos *Facebook* téléchargées, etc. Les applications Web telles que les plateformes de réseaux sociaux (comme *Twitter, Facebook, Instagram*) se développent pour rassembler des individus et renforcer leurs relations avec de nouvelles formes de coopération et de communication. Un réseau social est un concept théorique dans les sciences sociales se référant à une structure sociale composée d'individus ou d'organisations. Les réseaux sociaux en ligne sont utilisés par des millions de personnes liées entre elles par différents types de relations avec pour objectif de rencontrer d'autres utilisateurs afin de produire et partager des informations et des expériences sur divers sujets et intérêts. Par exemple, lors de l'élection présidentielle américaine de 2016, 40 876 345 *tweets* ont été émis entre le 8 et 9 novembre en réaction à l'élection de Donald Trump [2]. Le contenu de l'information partagée dépend du réseau lui-même, beaucoup de membres dans une communauté en ligne montrent des intérêts communs dans les loisirs, la religion, la politique ou les modes de vie alternatifs. Ainsi, les réseaux sociaux constituent un reflet important de la structure de la société du XXI^{e} siècle car il s'agit de réseaux de données portant sur une grande variété d'intérêts et de pratiques.

Les réseaux sociaux en ligne sont parfois utilisés comme une chambre d'écho, dans laquelle un échantillon restreint et/ou non représentatif est cité à plusieurs reprises

1. https ://www.statista.com/study/52194/digital-economy-compass/
2. Source Visibrain, une plateforme de veille des réseaux sociaux en ligne : https://www.visibrain.com/fr/blog/reseauxsociaux-antichambre-elections-americaines/

pour créer un élan autour d'un sujet ou d'une proposition politique [Goldie et al., 2014, Evans et al., 2018]. Comme par exemple les allégations, rendues publiques en octobre 2017, de harcèlements et d'agressions sexuelles attribués à Harvey Weinstein, personnalité influente de l'industrie du cinéma américain. Ces allégations ont débouché sur un mouvement féministe d'ampleur mondiale avec l'utilisation des hashtags `#MeToo` et sa déclinaison française `#balancetonporc`, ce qui a encouragé de nombreuses victimes à parler en Europe mais aussi en Asie, ou encore en Afrique. Récemment, en octobre 2018, une agression survenue dans un lycée de Créteil, dans le Val-de-Marne, a libéré la parole des enseignants sur les réseaux sociaux où à travers le hashtag `#PasDeVague`, ils dénoncent une absence de soutien de leur hiérarchie en cas d'agression.

En effet, l'augmentation spectaculaire des réseaux sociaux et du contenu généré par les utilisateurs a créé un nouveau phénomène, qui est l'interaction sociale et le « réseautage ». Des chercheurs de différentes disciplines ont uni leurs efforts pour étudier les réseaux sociaux. La recherche interdisciplinaire s'est développée dans une nouvelle direction appelée informatique sociale (*Computational Social Science* [Lazer et al., 2009]). La fouille des réseaux sociaux a émergé avec pour objectif de collecter et analyser les données sociales [Tang, 2017]. Ainsi, plusieurs thématiques de recherche ont été proposées. En général, les recherches sont autour des individus, de leurs comportements récurrents, de leurs interactions et de la structure topologique du réseau formé par les utilisateurs. L'exploitation de tels réseaux permet la détection de communautés, la prédiction de liens, l'analyse de sentiment des utilisateurs, la classification d'utilisateurs, l'analyse de l'influence, etc.

Les applications de l'informatique sociale citées ci-dessus peuvent être classées en trois catégories [Cohen, 2016]. La première est la fouille de structure du Web (*Web Structure Mining*) qui analyse des relations, inconnues *a priori*, entre les utilisateurs. Il y a plusieurs techniques de fouille telles que la classification et le classement. La détection de communauté et la recherche des utilisateurs influents font partie de cette catégorie. La deuxième catégorie est la fouille de contenus du Web (*Web Content Mining*) qui est le processus d'extraction d'informations contenues dans les documents stockés sur le Web. Parmi les applications de cette catégorie, nous trouvons l'analyse de sentiments, la recherche d'information et la classification thématique. La troisième catégorie est la fouille des usages du Web (*Web Usage Mining*), c'est un processus de découverte d'informations sur le comportement de l'utilisateur sur internet. Parmi les applications de cette catégorie, nous retrouvons la description des utilisateurs du réseau social sous la forme d'un profil détaillé.

La détection de communautés dans les réseaux sociaux est l'un des sujets les plus populaires de la fouille de la structure du Web. Les communautés sont des groupes d'individus ayant une probabilité plus élevée d'être reliés les uns aux autres qu'aux membres d'autres groupes [Newman, 2001, Newman, 2004, Girvan et al., 2002, Basaille et al., 2018]. Une autre branche de recherche dans la fouille de la structure de Web est de prédire et de recommander des liens inconnus dans les réseaux sociaux. Liben-Nowell *et al.* [Liben-Nowell et al., 2003] étudient le problème d'inférence de nouvelles interactions entre les individus. Ils développent plusieurs approches non supervisées pour traiter ce problème à partir de mesures d'analyse de la « proximité » des nœuds dans un réseau. Le principe est basé sur la similarité du contenu ou de la structure entre les individus. Leskovec *et al.* [Leskovec et al., 2010] utilisent un modèle de régression logistique pour prédire les liens positifs et négatifs dans les réseaux sociaux en ligne, où les liens positifs indiquent les relations telles que l'amitié, tandis que les liens négatifs indiquent l'opposition.

Dans le cadre de la fouille du contenu du Web, Tan *et al.* [Tan et al., 2011] ont étudié comment le sentiment des utilisateurs peut être inféré à partir du contenu des messages échangés dans un réseau social. En outre, en raison de problèmes de confidentialité, les utilisateurs de réseaux sociaux ont tendance à cacher leurs profils. Pour les fournisseurs de services de réseaux sociaux, les informations sur les profils des utilisateurs leur sont utiles pour personnaliser leurs services aux utilisateurs de diverses manières, telles que des recommandations d'amis et de contenus ou une recherche personnalisée [Tang et al., 2014]. Étant donné un réseau social et certaines informations sur des utilisateurs (attributs, préférences ou comportements), la classification des utilisateurs est conçue pour déduire les informations d'autres utilisateurs du même réseau [Getoor et al., 2005].

La modélisation du comportement et des caractéristiques des individus est un champ de recherche de la fouille de l'usage du Web dans laquelle se produit la construction d'un modèle/profil utilisateur pour caractériser ses compétences et connaissances. Firan *et al.* [Firan et al., 2007] construisent des profils utilisateur pour le site de musique `last.fm` en utilisant la musique stockée dans l'ordinateur de l'utilisateur et ses votes. Abel *et al.* [Abel et al., 2011] modélisent les intérêts des utilisateurs en analysant le contenu de leurs *tweets*. Basaille *et al.* [Basaille-Gahitte et al., 2013] proposent un profil thématique utilisateur dans le cadre de la relation client.

Récemment, les plateformes de micro-blogging attirent l'attention des utilisateurs et des chercheurs en raison de la facilité et de la rapidité du partage de l'information. Ces plateformes peuvent être considérées comme un très grand répertoire d'informations contenant des millions de messages textes généralement organisés en réseaux complexes impliquant des utilisateurs interagissant les uns avec les autres à des moments précis. En raison de son immense popularité, *Twitter* est considéré comme la plateforme de micro-blogging numéro un dans le monde entier, il offre des APIs permettant de collecter gratuitement les données servant de support pour développer des applications ou effectuer des analyses, c'est la raison pour laquelle nous l'avons choisi pour nos différentes expérimentations.

De nombreux travaux étudient les données issues de *Twitter* dans des domaines très différents : marketing, politique, catastrophes naturelles, etc. En effet, aujourd'hui, *Twitter* est une des « meilleures opportunités disponibles » pour une marque d'acquérir de la visibilité parmi des consommateurs potentiels. Les spécialistes du marketing prennent note des nombreuses possibilités offertes par *Twitter* et commencent à mettre en œuvre de nouvelles initiatives sociales à un rythme jamais atteint auparavant. Ainsi, les entreprises mondiales ont reconnu *Twitter* comme une plateforme de marketing à part entière, et l'utilisent avec des innovations pour alimenter leur campagne de publicité. *Twitter* est aussi exploité comme une plateforme pour les campagnes politiques [Ausserhofer et al., 2013] en devenant un média intégré au cœur de la stratégie de communication politique.

L'une des caractéristiques de *Twitter* est la diffusion d'information par l'utilisation d'opérateurs, tweeter, mentionner ou citer un utilisateur, utiliser un hashtag ou une URL par exemple. Les liens entre les utilisateurs déterminent le flux de l'information et conditionnent ainsi l'influence d'un utilisateur sur un autre. Certains utilisateurs, appelés influents, sont plus capables que d'autres de diffuser des informations à un grand nombre d'utilisateurs, d'influencer et de persuader les utilisateurs avec lesquels ils sont connectés. Par conséquent, la détection des utilisateurs influents dans un réseau est une clé de succès pour parvenir à une diffusion d'information à large échelle et à faible coût. La notion d'influence joue un rôle essentiel dans le fonctionnement des entreprises et le fonctionnement de la société. L'étude de l'influence sur *Twitter* peut nous aider à mieux comprendre pour-

quoi certaines tendances ou innovations sont adoptées plus rapidement que d'autres et comment nous pourrions aider les annonceurs et les spécialistes du marketing à concevoir des campagnes plus efficaces. L'étude de l'influence est cependant un sujet difficile. En effet, une telle étude ne se prête pas à une mesure facilement disponible, et des éléments essentiels tels que les choix humains et le fonctionnement de nos sociétés ne peuvent pas être reproduits dans les limites du laboratoire. Ainsi, l'étude de l'influence des utilisateurs sur *Twitter* est devenue un sujet de recherche de premier ordre pour les chercheurs en sciences de la communication et pose de nombreux problèmes de modélisation et d'estimation aux chercheurs en réseaux complexes.

1.1.2/ MOTIVATIONS

Cette thèse se place dans le cadre de l'analyse des données des réseaux sociaux. En effet, durant mon travail de recherche, j'ai eu l'occasion d'interagir et de collaborer avec des chercheurs des sciences de la communication dans le cadre des projets TEE'2014 et TEP'2017. TEE'2014 est un projet international interdisciplinaire réunissant près de quarante-cinq chercheurs (majoritairement des politologues, sociologues, chercheurs en communication) de dix laboratoires de recherche répartis dans six pays européens (France, Allemagne, Belgique, Italie, Espagne et Royaume-Uni). Il a pour objectif d'étudier la communication politique sur *Twitter* à l'occasion des élections au parlement européen de mai 2014. L'objectif scientifique général est de faire progresser les connaissances sur l'utilisation de ce média social à travers l'analyse de la construction du discours de la campagne, dans plusieurs pays. Dans ce cadre, j'ai pu assister à des réunions qui m'ont permis d'interagir avec les chercheurs, notamment à Dijon, Metz et Bonn (Allemagne) et valider de façon qualitative mon approche. TEP'2017 reprend les mêmes principes à travers l'étude de la campagne présidentielle française de 2017.

Récemment, l'analyse de l'influence a suscité beaucoup d'intérêt de la part des milieux de la recherche académique et industrielle. L'influence se produit lorsque les opinions, les émotions ou les comportements d'une personne sont affectés par d'autres, intentionnellement ou non [Kelman, 1958]. En général, les recherches existantes sur l'analyse de l'influence sociale peuvent être classées en trois catégories : test d'influence où il s'agit d'identifier l'existence de l'influence [Leavitt et al., 2009, Cha et al., 2010, Lee et al., 2010a], mesure d'influence [Sun et al., 2011, Romero et al., 2011, Chen et al., 2013a, Silva et al., 2013, Zhao et al., 2015, Jendoubi et al., 2017] et modèles de diffusion d'influence [Kempe et al., 2003, Leskovec et al., 2007, Zhang et al., 2014, Riquelme et al., 2016].

Plusieurs études d'influence ont été présentées dans la littérature. Souvent, le principal aspect sur lequel se sont focalisés les chercheurs fut le classement des utilisateurs de *Twitter* selon leur influence. Cependant, il est important de mesurer l'influence d'un certain utilisateur indépendemment des autres utilisateurs. La méthode de calcul de l'influence est généralement basée sur les interactions présentes dans le réseau, par exemple, un utilisateur de *Twitter* qui possède un grand nombre de *mentions* est généralement considéré comme plus influent que d'autres. Mais il est difficile de savoir quel type d'interaction permet d'indiquer le mieux l'influence des utilisateurs.

L'estimation de l'influence pose trois défis principaux. Le premier est la **diversité des interactions** sur lesquelles nous pouvons baser les calculs de l'influence. Il faut combiner l'influence respective des différentes interactions afin d'établir une mesure générale d'in-

fluence tout en prenant en compte l'**incertitude** lors de la combinaison des interactions. Dans le cas des réseaux multi-relationnels tel que *Twitter*, il est difficile d'attribuer des pondérations aux différentes interactions avant de les combiner. Le second défi est la **considération de l'influence indirecte**. L'influence est indirecte lorsqu'elle atteint un utilisateur à travers des utilisateurs intermédiaires. Par exemple, un utilisateur peut *retweeter* un *tweet* d'un autre utilisateur indirectement à travers un utilisateur intermédiaire. Il est donc nécessaire de mesurer l'influence en tenant compte des interactions directes et indirectes dans le réseau. Le troisième défi est la **polarité des *tweets***, il est important d'analyser le contenu des *tweets* afin de déduire si l'influence exercée est positive ou négative.

1.2/ OBJECTIFS ET PROBLÉMATIQUE DE LA THÈSE

La problématique principale de nos travaux de recherche est formulée de la manière suivante :

Comment exploiter les données d'un réseau social
pour estimer l'influence des utilisateurs ?

L'étude de l'influence des utilisateurs se base sur les interactions présente dans le réseau social. Notre cadre d'application est *Twitter* car il fournit des API permettant un accès facile aux données en grande quantité. Les données issues des *tweets* sont riches pour ce qui est des interactions, en effet *Twitter* offre plusieurs opérateurs qui établissent des liens entre les données.

Ainsi, nous classifions les questions de recherche abordées dans cette thèse selon le défi auquel nous devons répondre pour estimer l'influence d'un utilisateur sur *Twitter* :

— À partir des *tweets* collectés, il est difficile de déduire directement les interactions entre les utilisateurs. Ainsi, pour estimer l'influence d'un utilisateur, nous avons besoin de modéliser les données de *Twitter* afin de représenter les utilisateurs ainsi que toutes leurs interactions. Nous étudions alors :

> D_1 : Comment modéliser les données de *Twitter* afin de représenter toutes les interactions entre les utilisateurs ainsi que les *tweets* et leurs contenus ?

— Les interactions dans *Twitter* sont nombreuses et de différents types. Dans les travaux de recherche existants, la mesure de l'influence se base souvent sur un seul type d'interaction quelque soit la méthode utilisée. Or, les différents types d'interactions contribuent à l'influence des utilisateurs. Nous avons donc besoin de connaître le ou les types d'interactions sur lesquelles nous pouvons nous baser afin d'estimer l'influence. Ainsi, il est important de combiner ces interactions alors que nous ne disposons que d'informations incertaines sur l'importance des différents types d'interactions. Nous répondons ainsi au défi suivant :

> D_2 : Comment combiner les informations sur les interactions du réseau tout en exprimant l'incertitude que nous avons sur leur importance les unes par rapport aux autres ?

— Il est aussi important de connaître si l'influence exercée est positive, négative ou neutre. Le contenu des *tweets* doit être alors exploité dans le but d'analyser le sentiment exprimé dans ces derniers et de déduire ainsi la polarité de l'influence.

Nous répondons alors au défi :

D_3 : Comment analyser le contenu des *tweets* pour déduire la polarité de l'influence ?

— Nous nous intéressons aussi à connaître la manière dont certains utilisateurs interagissent avec les autres pour devenir influents. Pour ce faire, il est important d'étudier le style de communication des différents utilisateurs de *Twitter*. Nous étudions alors :

D_4 : Comment exploiter le contenu des *tweets* et l'usage des différents opérateurs pour déduire le style de communication utilisé ?

1.3/ CONTRIBUTIONS

Dans ce qui suit, nous présentons les principales contributions scientifiques de cette thèse. Notre objectif principal est de proposer une approche qui permet d'exploiter les données de *Twitter* afin d'estimer l'influence des utilisateurs. Les contributions suivantes sont des modèles que nous proposons en réponse aux différents défis. Ces contributions ont abouti à la publication de plusieurs articles scientifiques que nous citons dans l'annexe E.

C_1 *Modélisation de Twitter*

Pour répondre à notre premier défi, nous proposons une modélisation de *Twitter*. Dans la littérature, *Twitter* a été souvent modélisé sous forme d'un graphe où les nœuds représentent les utilisateurs, et les liens sont les différentes interactions existantes entre eux. Avec une telle représentation, nous ne parvenons pas à voir le contenu des *tweets* publiés et échangés dans le réseau. Pour ceci, nous proposons une nouvelle modélisation de *Twitter* comme un réseau multiplexe hétérogène. En effet, le réseau est représenté à travers plusieurs couches, chaque couche constitue une dimension (interactions entre utilisateurs, actions des utilisateurs et contenu des *tweets*). Les nœuds dans le réseau proposé sont hétérogènes puisque nous avons plusieurs types de nœuds possibles : utilisateurs, *tweets*, hashtags, URLs.

C_2 *Estimation de l'influence par la théorie des fonctions de croyance*

Cette contribution répond au deuxième défi. Afin d'estimer l'influence d'un utilisateur dans *Twitter*, il est important de considérer toutes les interactions se produisant dans le réseau. Comme il est difficile de connaître l'importance des différentes interactions les unes par rapport aux autres, nous employons la théorie des fonctions de croyance qui permet d'exprimer l'incertitude par la notion de masses de croyance portées par les différentes interactions. La théorie des fonctions de croyance permet la combinaison des masses de croyance à travers la règle de combinaison conjonctive. L'influence d'un utilisateur dans le réseau social est estimé avec un degré de certitude puis nous classons les utilisateurs selon leur influence.

C_3 *Polarité de l'influence sur Twitter*

Dans le but de déterminer la polarité de l'influence, nous exploitons le contenu des *tweets*

collectés. Nous étudions ainsi le sentiment exprimé à travers les *tweets* en utilisant un algorithme de machine learning (*random forest*). Une fois que nous avons déterminé la polarité des *tweets*, nous divisons le réseau *Twitter* en trois sous-réseaux, chaque sous-réseau représentant une polarité (positif, négatif ou neutre). La méthode d'estimation de l'influence proposée dans C_2 est utilisée dans chaque sous-réseau pour obtenir l'influence dans chaque sous-réseau. Enfin, les mesures d'influence des trois sous-réseaux sont combinées pour obtenir une estimation de l'influence globale polarisée.

C_4 Analyse des styles de communication sur Twitter

L'approche proposée dans C_2 peut être adaptée afin d'étudier les styles de communication dans *Twitter*. En effet, certains utilisateurs adaptent un style de communication spécifique pour atteindre un maximum de visibilité sur *Twitter* et devenir ainsi influent. Nous proposons trois styles de communication : 1) le style interactif à travers lequel l'utilisateur a plutôt tendance à interagir avec d'autres utilisateurs ; 2) le style informatif dans lequel l'utilisateur maximise la visibilité de son *tweet* en identifiant par exemple des utilisateurs populaires dans son *tweet* ; 3) enfin, le style balancé où l'utilisateur alterne entre les styles interactif et informatif.

1.4/ STRUCTURE DU MANUSCRIT

Ce manuscrit est composé de six chapitres en complément de cette introduction. Dans le chapitre 2, nous définissons tout d'abord les principaux concepts relatifs au sujet de la thèse. Nous présentons par la suite des problématiques connexes à l'étude de l'influence. Enfin, nous présentons les travaux de recherche qui étudient l'influence dans les réseaux sociaux et plus particulièrement *Twitter* afin de mieux se positionner par rapport aux travaux existants.

Dans le chapitre 3, nous décrivons la modélisation du réseau *Twitter* comme un réseau multiplexe hétérogène. Cette modélisation est basée sur les concepts des réseaux multiplexes. Elle prend en compte à la fois le contenu des *tweets* et les interactions entre les utilisateurs. Cette modélisation nous a permis d'exploiter l'algorithme du PageRank multiplexe afin de classer les utilisateurs selon leur influence.

Dans le chapitre 4, nous proposons *TwitBelief*, une approche d'estimation de l'influence dans *Twitter*. En se basant sur la théorie des fonctions de croyance, nous combinons les informations des différentes interactions du réseau. L'incertitude sur l'importance des différents critères considérés est exprimée. Le contenu des *tweets* n'est pas pris en compte dans le processus de l'estimation, nous considérons seulement les interactions présentes dans le réseau.

Dans le chapitre 5, nous proposons l'extension de la méthode de l'estimation de l'influence au contenu des *tweets*. Nous présentons d'abord une approche d'estimation de l'influence polarisée afin de considérer le sentiment exprimé à travers des *tweets*, nous déduisons ainsi la polarité de l'influence exercée. Ensuite, nous détaillons une deuxième extension, *Twitter* styles, qui détermine les styles de communication des utilisateurs en fonction de leurs *tweets*.

Dans le chapitre 6, nous présentons l'étude expérimentale des différentes contributions. Cette étude est effectuée sur trois jeux de données. Le premier s'appuie sur des données

Twitter collectées dans le cadre du projet inter-disciplinaire TEE'2014 lors de la campagne pour les élections européennes de 2014. Le deuxième est relatif aux données de l'élection présidentielle française de 2017. Nous menons également des expériences sur l'ensemble des données CLEF RepLab 2014 [3], qui sont des données *Twitter* contenant des utilisateurs *Twitter* manuellement annotés en fonction de leur influence. Ces dernières expériences nous permettent de comparer notre approche avec les annotations posées.

Dans le chapitre 7, nous concluons notre travail en revenant sur les principales contributions apportées dans cette thèse dans le domaine de l'estimation de l'influence dans *Twitter*. Ce chapitre contient également une partie discussions, présentant nos observations finales concernant notre travail. Nous expliquons ensuite la pertinence de nos contributions et leur capacité à être applicables/généralisables à d'autres domaines. Et enfin nous identifions les limites de notre approche et les perspectives de recherche ainsi que d'autres pistes prometteuses corrélées à notre travail.

3. https://www.damianospina.com/projects/the-replab-2014-dataset/

2

ÉTAT DE L'ART

L'objectif de ce chapitre est de définir tout d'abord les concepts fondamentaux utilisés dans cette thèse : la plateforme *Twitter* et les critères relatifs à son analyse puis les notions autour de l'influence. Nous présentons par la suite des problématiques connexes à l'influence puis nous détaillons l'état de l'art sur l'estimation de l'influence dans les réseaux sociaux et en particulier dans *Twitter* afin de les comparer et de positionner notre travail vis-à-vis de ces derniers.

2.1/ CONCEPTS FONDAMENTAUX

Dans cette section, nous présentons les concepts fondamentaux relatifs à notre travail à savoir la plateforme *Twitter* qui nous sert à valider nos résultats ainsi que les termes utilisés dans ce manuscrit afin de lever toute ambiguïté.

2.1.1/ LA PLATEFORME *Twitter*

Twitter est un réseau social et une plateforme de micro-blogging créé et lancé en 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass. Il permet à ses utilisateurs de partager des informations sous forme de messages de 280 caractères maximum appelés « *tweets* ». La syntaxe d'un *tweet* est assez simple : il peut contenir, en plus des caractères, des médias (images, vidéos), des hashtags (mots-clés commençant par #) qui permettent de signaler un sujet d'intérêt, des URLs (liens vers d'autres sites) ou bien d'autres comptes d'utilisateurs (préfixés par @) permettant ainsi d'interpeller, citer ou répondre à un autre utilisateur (voir la figure 2.1). Par défaut, les *tweets* sont publics, c'est-à-dire visibles par tous ; il existe cependant des cas où les *tweets* peuvent être privés.

Les interactions entre utilisateurs sont un point important de ce réseau social : un utilisateur peut *suivre* un autre utilisateur, lui permettant ainsi de voir les informations de l'utilisateur qu'il suit et d'être informé des nouveaux *tweets* de ce dernier. Les personnes qui suivent l'utilisateur sont appelées des *abonnés*, et les personnes que l'utilisateur suit sont appelées *abonnements*. Dans le but de relayer de l'information, un utilisateur peut *retweeter* un *tweet*, ce qui expose ce *tweet* à ses abonnés, qui à leur tour peuvent le *retweeter* (chaîne de *retweets*). Un utilisateur peut *mentionner* un autre utilisateur en utilisant le préfixe « @ » s'il veut lui adresser le *tweet*, ce même *tweet* pouvant être *retweeté* par un autre utilisateur. De plus, un utilisateur peut répondre à un *tweet* et créer ainsi une conversation avec l'utilisateur du *tweet* initial.

FIGURE 2.1 – Un exemple de *tweet*

Twitter est l'un des réseaux sociaux les plus utilisés : au troisième trimestre 2018, on dénombrait 326 millions d'utilisateurs actifs chaque mois, postant au total environ 504 millions de *tweets* par jour [1]. Aujourd'hui, pour les personnes publiques et les célébrités, *Twitter* est devenu incontournable pour parler de son actualité et exister dans les médias. Comme la drosophile, qui est l'espèce modèle dans la recherche en génétique, *Twitter* est le « modèle représentant » dans le domaine des Sciences Humaines et Sociales pour des recherches sur différents sujets d'intérêt. Ce phénomène a abouti à différentes études sur les données issues de *Twitter*.

Dans le domaine politique, les politiciens utilisent *Twitter* pour communiquer sur leurs campagnes, par exemple, pendant les élections européennes 2014, un grand nombre de *tweets*, émis par les candidats, ont été étudiés par des chercheurs afin de comprendre la circulation des discours politiques [Frame et al., 2015, Thimm et al., 2016]. Durant l'année 2016, le président américain Donald Trump s'est distingué pendant sa campagne par une utilisation de *Twitter* offensive et très différente de celle des autres personnalités politiques. Avec près de 1 800 *tweets* entre sa nomination en tant que candidat républicain, en mi-juillet 2016, et quelques jours avant son élection en tant que président, en mi-janvier 2017, Donald Trump a été très actif sur le réseau social, non seulement pour commenter les événements, mais aussi et surtout pour répondre à des critiques ou attaquer ses rivaux, écrivant jusqu'à 88 *tweets* en une journée [2]. Ainsi, des études explorent la manière dont le président américain Donald Trump utilise *Twitter* comme instrument stratégique dans la politique pour diffuser son discours [Ott, 2017, Kreis, 2017, Anderson, 2017]. Les citoyens utilisent aussi ce réseau pour émettre leurs opinions, par exemple, les chercheurs

1. http://www.journaldunet.com/ebusiness/le-net/1159246-nombre-d-utilisateurs-de-twitter-dans-le-monde/
2. https://urlz.fr/86eI

montrent que *Twitter* et les réseaux sociaux ont joué un rôle primordial dans la révolution Tunisienne et le printemps arabe [Dakhli, 2011, Huygue, 2011].

Dans le domaine du marketing, *Twitter* a constitué une plate-forme pour les études de la réputation des produits à travers l'analyse des sentiments exprimés dans les *tweets* émis par les consommateurs [Keller et al., 2007, Jansen et al., 2009, Vidya et al., 2015]. Ils existent également des études sur l'utilisation de *Twitter* pendant les catastrophes naturelles telles que la détection de l'épicentre de tremblement de terre [Earle et al., 2012] ou l'étude du comportement des utilisateurs suite à des situations d'urgence (crimes, accidents de voiture, etc.) [Li et al., 2012] ou l'étude de discours de haine après l'acte terroriste de Woolwich en Angleterre [Williams et al., 2015].

Nous retrouvons aussi des études sur l'utilisation de *Twitter* par des journalistes et des médias dans le but de diffuser les actualités [Álvarez, 2012, Debeaux, 2015]. Cependant, avec l'utilisation croissante de *Twitter* dans le domaine politique, les fausses informations (« *fake news* ») et la vérification des faits (« *fact-checking* ») ont pris une nouvelle signification. Le *fact-checking* désigne un mode de traitement journalistique, consistant à vérifier de manière systématique des affirmations et l'exactitude des faits, chiffres ou citations rapportés par les journalistes, les experts. Dans [Coddington et al., 2014], les auteurs examinent dans quelle mesure les techniques de *fact-checking* ont été utilisées sur *Twitter* à travers une analyse de contenu du discours sur *Twitter* des journalistes politiques sur les débats présidentiels de l'élection présidentielle américaine de 2012. Une typologie de *tweets* indique que le *fact-checking* a joué un rôle notable mais secondaire dans le discours des journalistes sur *Twitter*.

2.1.2/ CRITÈRES POUR L'ANALYSE DE *Twitter*

De nombreux travaux liés à *Twitter* visent à caractériser les utilisateurs de différentes manières, par exemple : son rôle dans *Twitter* (spammeurs, robots, organisations, etc.), sa nature (catégorie socioprofessionnelle, âge, etc.), ses sujets d'intérêt, etc. Cependant, pour un problème de classification d'utilisateur donné, il est très difficile de sélectionner un ensemble de critères appropriés, car, dans la littérature, les nombreux critères décrits sont très hétérogènes et apparaissent sous différents noms.

Dans [Cossu et al., 2016], les auteurs examinent les critères qui peuvent être extraits de *Twitter* dans le but de catégoriser les utilisateurs et de les appliquer pour détecter les utilisateurs influents dans la vie réelle sur la base de leur profil *Twitter*. Ils divisent les critères en sept catégories. La première catégorie est le profil de l'utilisateur, par exemple, s'assurer qu'il a une photographie de profil, que son compte est vérifié ou non, qu'il indique sa page Web dans son profil, etc. D'autres critères sont inclus dans cette catégorie comme le nombre de caractères dans le nom d'utilisateur, la description et l'âge du profil ainsi que le nombre d'URLs présentes dans la description du profil.

La deuxième catégorie est l'activité de l'utilisateur. L'activité est étudiée à travers le nombre de *tweets* publiés par l'utilisateur, de sources médias publiées, d'auto-*mentions*, de *tweets* géolocalisés et d'intervalle de temps entre deux *tweets* consécutifs.

Les connections locales de l'utilisateur représentent la troisième catégorie. Les critères de cette catégorie décrivent comment l'utilisateur est connecté avec le reste du réseau *Twitter*. Par exemple, le réseau des relations *abonnés-abonnements* où deux valeurs représentent un utilisateur : le nombre d'abonnés et le nombre d'abonnements. Nous

pouvons considérer aussi dans cette catégorie le nombre de *tweets* publiés par les *abonnés* et les *abonnements* d'un utilisateur, ce qui représente le niveau de l'activité de publication dans le voisinage direct de l'utilisateur.

La catégorie suivante représente les interactions de l'utilisateur avec les autres personnes. Les interactions sont représentées par le nombre de *retweet*, *mention* et *réponse*.

La cinquième catégorie couvre l'aspect lexical des *tweets* en prenant en compte le contenu produit par l'utilisateur. Les critères utilisés sont par exemple, la taille du lexique de l'utilisateur, c'est-à-dire le nombre de mots qu'il utilise, le nombre d'entités identifiées dans ses *tweets*. Les entités correspondent aux noms propres permettant d'identifier des personnes, des organisations, des lieux, des marques, etc.

La catégorie suivante représente les particularités d'écriture d'un *tweet*. Les critères sont alors le nombre de caractères par mot et par *tweet*, la longueur d'un *tweet* parfois exprimée en nombre de mots au lieu des caractères mais aussi le nombre de hashtags, d'URLs, l'auto-similarité des *tweets* de l'utilisateur. Ces critères peuvent aider à caractériser certains types d'utilisateurs, ainsi le contenu émis par certains spammeurs est formé d'un ensemble de mots sans structure grammaticale propre.

Enfin, la dernière catégorie représente les données externes, cette catégorie contient des critères correspondant aux données non récupérées directement depuis *Twitter*, par exemple, le nombre de résultats de recherches Web pour la page de l'utilisateur.

Le tableau 2.1 synthétise les sept catégories de critères sur lesquels les chercheurs de différents domaines peuvent baser leur analyse des données issues de *Twitter*. Des exemples de références de travaux de recherche sont donnés catégorie par catégorie.

TABLE 2.1 – Synthèse des critères pour l'analyse *Twitter* [Cossu et al., 2016]

Catégorie	**Description**	**Exemples de travaux de recherche**
Profil de l'utilisateur	Photo de profil	[Pennacchiotti et al., 2011, Vilares et al., 2014]
	Compte vérifié	[Zi et al., 2012, Lee et al., 2010b, Uddin et al., 2014, Vilares et al., 2014]
	Page Web	[Lee et al., 2013, Vilares et al., 2014]
	Age du profil	[Benevenuto et al., 2010, Pennacchiotti et al., 2011, Uddin et al., 2014]
	Nombre de caractères dans le nom d'utilisateur	[Lee et al., 2011, Lee et al., 2013]
	Description du profil	[Lee et al., 2011, Lee et al., 2013]
	Nombre d'URLs	[Ramírez-de-la Rosa et al., 2014]
Activité de l'utilisateur	Nombre de *tweets*	[Rao et al., 2010, Lee et al., 2011, Zi et al., 2012, Vilares et al., 2014]
	Nombre de sources médias	[Ramírez-de-la Rosa et al., 2014]
	Nombre d'auto-*mentions*	[Ramírez-de-la Rosa et al., 2014]
	Nombre de *tweets* géolocalisés	[Huang et al., 2014, Vilares et al., 2014]
	Intervalle de temps entre deux *tweets* consécutifs	[Benevenuto et al., 2010, Pennacchiotti et al., 2011]
		[Ramírez-de-la Rosa et al., 2014]
Connections locales	Relations abonnés/abonnements	[Ramírez-de-la Rosa et al., 2014, Vilares et al., 2014]
		[Tommasel et al., 2015]
	Nombre de *tweets* publiés par les *abonnés/abonnements*	[Benevenuto et al., 2010, Ramírez-de-la Rosa et al., 2014]
Interactions de l'utilisateur	Nombre de *retweets*	[Rao et al., 2010, Benevenuto et al., 2010, Cha et al., 2010]
	Nombre de *mentions*	[Cha et al., 2010, Uddin et al., 2014, Danisch et al., 2014]
	Nombre de *réponses*	[Pennacchiotti et al., 2011, Uddin et al., 2014]
Aspect lexical des *tweets*	Taille du lexique	[Benevenuto et al., 2010, Weren et al., 2014]
	Nombre d'entités	[Ramírez-de-la Rosa et al., 2014]
Traits stylistiques	Nombre de caractères par mot/*tweet*	[Ramírez-de-la Rosa et al., 2014]
	Longueur d'un *tweet*	[Benevenuto et al., 2010, Makazhanov et al., 2014, de Arruda et al., 2014]
	Nombre de hashtags	[Pennacchiotti et al., 2011, de Arruda et al., 2014, Uddin et al., 2014]
	Nombre d'URLs	[Zi et al., 2012, Lee et al., 2010b, Ramírez-de-la Rosa et al., 2014]
	Auto-similarité des *tweets*	[Wang, 2010, Lee et al., 2011, Lee et al., 2013]
Donées externes	Recherches Web	[Cossu et al., 2015]

Par conséquent, nous pouvons conclure que les critères d'analyse de *Twitter* sont nombreux rendant la tâche de leur sélection difficile. Un autre fait important concernant la sélection des critères est leur disponibilité, les données fournies pour l'étude pourraient être incomplètes par rapport aux critères à traiter. Par exemple, pour utiliser les critères issues de la catégorie connections locales, il faut remonter à partir de l'ensemble des *tweets* collectés à tous les comptes abonnés aux utilisateurs qui ont émis ces *tweets*. Les limitations de l'API *Twitter* peuvent empêcher d'accéder à ces données ou les comptes concernés peuvent ne plus exister ou avoir changé leur paramètre de confidentialité. Il existe également des contraintes liées au temps : les données collectées ne correspondent qu'à celles qui peuvent être obtenues dans un délai raisonnable. En outre, même si nous parvenons à collecter toutes les données nécessaires, le traitement de certains critères peut être très lent si l'ensemble des données à traiter est trop volumineux. De plus, les donnés *Twitter* ne peuvent pas être réduites à des propriétés statistiques mais demandent des outils riches pour les analyser.

2.1.3/ NOTIONS RELATIVES À L'INFLUENCE

Dans cette sous-section, nous commençons par définir l'influence sociale, ensuite, nous définissons quelques notions relatives à l'étude de l'influence et d'autres souvent confondues avec la notion d'influence.

Définition 1 : Influence sociale

Pouvoir social d'une personne qui amène les autres à se ranger à son avis [a].

a. Dictionnaire culturel en langue Française sous la direction d'Alain Rey.

En psychologie sociale, l'influence sociale est le changement dans les pensées, les sentiments, les attitudes ou les comportements d'un individu résultant de l'interaction avec un autre individu ou un groupe [Rashotte, 2007]. L'étude de l'influence sociale peut s'effectuer sur trois niveaux : le niveau d'analyse individuelle quand l'influence se produit entre deux individus, le niveau d'analyse communautaire quand l'influence est effectuée entre une communauté [3] et un individu, dans ce niveau, l'influence est visible à travers l'évolution de la communauté, et enfin, le niveau d'analyse d'influence dans le réseau, cette influence s'effectue entre les communautés ce qui entraînent l'évolution de l'ensemble du réseau. La plupart des changements et des évolutions des réseaux sociaux peuvent être considérés comme le résultat de l'influence sociale individuelle. L'influence sociale est distincte du conformisme, de l'innovation et de la soumission à l'autorité. Le conformisme ou influence majoritaire se produit lorsqu'un individu change son comportement pour le mettre en adéquation avec le comportement d'un groupe majoritaire. L'innovation est définie comme l'influence d'un individu ou d'une minorité de personnes sur une majorité. Contrairement au conformisme, c'est la minorité qui réussit à imposer son point de vue. On parle de soumission à l'autorité (l'obéissance) lorsqu'un individu change de comportement afin de se soumettre aux ordres émanant d'une autorité perçue comme légitime. Il existe également d'autres phénomènes comme la résistance qui s'oppose aux phénomènes précédents.

3. Une communauté est, dans le sens courant, un ensemble de personnes vivant ensemble ou ayant des interactions entre elles.

Définition 2 : Popularité

En sociologie, une personne populaire est une personne connue ou célèbre qui attire sur elle l'attention du public et parfois même des médias.

Un utilisateur est populaire lorsqu'il est reconnu par de nombreux autres utilisateurs dans le réseau [Riquelme et al., 2016]. Un exemple d'utilisateur populaire est une célébrité, qui n'a pas nécessairement un compte actif et influent. La mesure de popularité la plus connue est simplement le nombre d'*abonnés* de l'utilisateur en question. Le FollowerRank proposé par [Nagmoti et al., 2010] est la version normalisée de cette mesure de popularité, il s'agit de diviser le nombre d'abonnés d'un utilisateur par la somme du nombre d'abonnés et du nombre d'abonnements de ce même utilisateur, cette mesure est utilisée dans [Cha et al., 2010, Lee et al., 2011, Ramírez-de-la Rosa et al., 2014, Vilares et al., 2014, Tommasel et al., 2015]. L'avantage de cette mesure est de bannir les spammeurs qui sont des utilisateurs avec de nombreux abonnements et peu d'abonnés.

Définition 3 : Homophilie

L'homophilie est une mesure utilisée en analyse des réseaux sociaux pour mettre en évidence le degré d'attachement préférentiel [a] entre des utilisateurs partageant une ou des caractéristiques communes sur un ou des attributs particuliers (par exemple la situation géographique).

a. Un processus dynamique au cours duquel des individus arrivent l'un après l'autre et doivent choisir de rejoindre une "classe". La classe n'est pas choisie "au hasard" mais en fonction de la population présente.

L'homophilie est l'un des aspects les plus fondamentaux des réseaux sociaux. Dans le cadre de l'homophilie, un utilisateur dans le réseau social est similaire à ses voisins [McPherson et al., 2001]. C'est un résultat naturel car les voisins d'un utilisateur donné dans le réseau social ne sont pas un échantillon aléatoire de la population, en effet les individus se connectent car ils ont des points en commun comme l'âge, les occupations, les intérêts et les croyances.

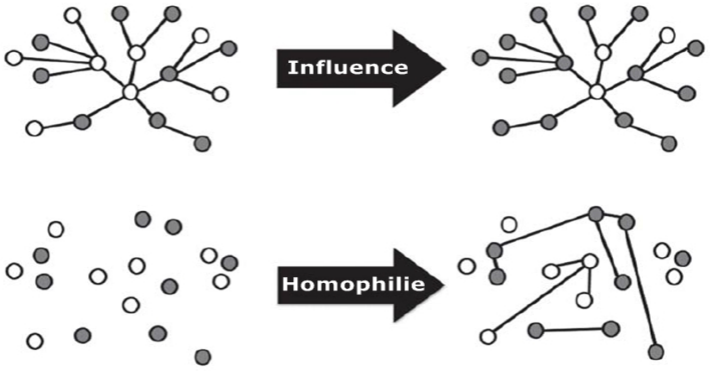


FIGURE 2.2 – Influence sociale vs homophilie

La figure 2.2 illustre la différence entre l'influence sociale et l'homophilie. Dans l'influence, des individus déjà liés ensemble changent les valeurs de leurs attributs et deviennent similaires à leurs voisins en adoptant leurs comportements. L'influence rend des individus connectés similaires tandis que l'homophilie trouve des individus similaires pour

les connecter, les connections sont donc dues à leur similarité. Intuitivement, les effets de l'homophilie et de l'influence sociale conduisent à des applications différentes dans l'analyse des données des réseaux sociaux. En particulier, les systèmes de recommandation [Candillier et al., 2012] sont basés sur l'homophilie, tandis que le marketing viral [Domingos et al., 2001, Richardson et al., 2002] est basé sur l'influence sociale.

Définition 4 : Assortativité

Dans un réseau social, les utilisateurs similaires sont reliés les uns aux autres plus souvent qu'avec des utilisateurs dissemblables.

Le phénomène de l'assortativité peut provenir de plusieurs mécanismes différents dont l'influence sociale, cela indique que les gens ont tendance à suivre les comportements de leurs amis. L'effet d'influence sociale conduit les gens à adopter les mêmes comportements que leurs voisins. L'assortativité peut être due aussi à l'homophilie (sélection), cela indique que les gens ont tendance à créer des relations avec d'autres personnes qui sont déjà similaires à eux. Et enfin, l'assortativité peut être le résultat du confounding, c'est-à-dire l'effet de l'environnement à rendre les individus similaires, dans ce cas des facteurs autres que les utilisateurs et les liens entre eux affectent la structure du réseau. Par exemple, l'influence sociale se produit quand les fumeurs influencent leurs amis non-fumeurs tandis que l'homophilie se manifeste quand les fumeurs deviennent amis. Le confounding peut se traduire par le fait qu'il y a beaucoup d'endroits où les gens peuvent fumer.

Dans ce travail, notre problématique principale est la détection des utilisateurs influents dans les réseaux sociaux, nous voyons dans le paragraphe suivant deux problématiques connexes : la diffusion et la maximisation de l'influence. Ces deux problématiques dépendent de la détection des utilisateurs influents qui forment l'ensemble de départ sur lequel ces deux processus reposent.

2.2/ PROBLÉMATIQUES CONNEXES : DIFFUSION ET MAXIMISATION D'INFLUENCE

L'analyse de l'influence sociale a diverses applications dans le monde réel. La diffusion et la maximisation d'influence dans le marketing viral sont des exemples de ces applications. Le problème de la maximisation d'influence remonte à la recherche sur le bouche-à-oreille et le marketing viral [Bass, 1969, Brown et al., 1987, Mahajan et al., 1991, Domingos et al., 2001, Goldenberg et al., 2001, Richardson et al., 2002]. Le problème est souvent motivé par la détermination de clients potentiels à des fins de marketing. L'objectif est de minimiser les coûts de commercialisation et de maximiser les bénéfices. Par exemple, une entreprise souhaite commercialiser un nouveau produit grâce à l'effet naturel du bouche-à-oreille découlant des interactions dans un réseau social. L'objectif est d'obtenir un petit nombre d'utilisateurs influents qui achètent et vantent le produit pour déclencher une grande cascade d'autres achats.

Les modèles de maximisation d'influence se basent sur les modèles de diffusion d'influence. À partir d'un modèle de diffusion M, d'un ensemble d'utilisateurs influents S et d'une fonction $f_M(S)$ qui estime la diffusion d'influence, l'objectif est de savoir comment sélectionner les individus de S pour maximiser $f_M(S)$. La diffusion de l'influence est le transfert et la propagation de l'information par le bouche-à-oreille entre les utilisateurs

dans le cadre d'un réseau social. Dans [Riquelme et al., 2016], les auteurs examinent les modèles classiques du problème de diffusion d'influence. Pour faciliter l'explication, ils associent à chaque utilisateur un statut : actif ou inactif. Initialement, tous les utilisateurs sont inactifs. Ensuite, des utilisateurs sont activés, il s'agit d'utilisateurs qui peuvent influencer davantage leurs amis (c'est-à-dire leurs voisins dans le réseau) qui deviennent à leur tour actifs. Il existe une grande variété de modèles que l'on peut catégoriser comme suit :

— **Modèle de seuil linéaire :** dans cette famille de modèles, un utilisateur est considéré actif si la proportion de ses voisins entrants déjà actifs dépasse un seuil spécifique à chaque utilisateur [Kempe et al., 2003]. Chaque utilisateur v reçoit un seuil θ_v qui lui est propre et v devient actif à l'étape t si $v(S) > \theta_v$, où S est l'ensemble des voisins de v qui sont actifs à l'étape $t-1$ et $v(S)$ est la somme des poids sur ses liens entrants venant de ses voisins (voir la figure 2.3).

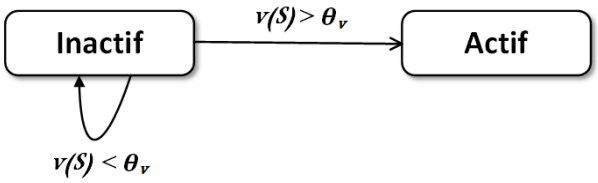


FIGURE 2.3 – Diagramme d'états-transitions d'un utilisateur dans le modèle de seuil linéaire

— **Modèle de cascade :** ce modèle est fondé sur le principe que dès qu'un utilisateur u est actif, il a une unique chance d'activer chacun de ses voisins directs v avec une probabilité de succès $P_{u,v} \in [0,1]$ [Goldenberg et al., 2001]. Que u réussisse ou échoue, u ne pourra plus influencer v par la suite. C'est un modèle chronologique qui procède par étape (voir la figure 2.4).

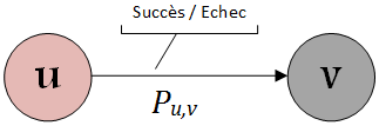


FIGURE 2.4 – Modèle de cascade

— **Modèle épidémique :** cette famille de modèles se base sur la métaphore de la transmission des maladies. La maladie commence lorsque quelques individus sont infectés. Les personnes qui ont eu des contacts avec des personnes infectées sont elles-mêmes infectées et à leur tour par contact propagent la maladie à leur entourage. Ce processus d'infection continue jusqu'à ce qu'il n'y ait plus de risque d'infection. Dans le modèle de diffusion d'influence épidémique, la transmission se fait d'un individu à l'autre avec une certaine probabilité dépendant du degré de réceptivité de l'individu. SIR est un des modèles appartenant à la famille des modèles épidémiques, il considère une population fixe qui se divise en trois classes distinctes : Susceptible (S), Infectieux (I) et Rétabli (R). La figure 2.5 montre le diagramme d'états-transitions, dans le modèle SIR, d'un nœud qui passe par les états successifs : $S \rightarrow I \rightarrow R$ avec sur les transitions α et β représentant respectivement le taux d'infection et le taux de retrait.

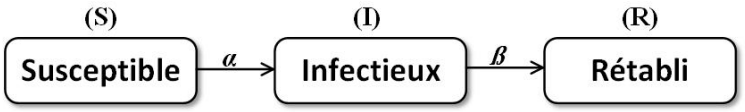


FIGURE 2.5 – Diagramme d'états-transitions d'un nœud dans le modèle SIR

Du point de vue empirique, Kempe et al. [Kempe et al., 2003] montrent que les modèles de diffusion de seuil linéaire et de cascade peuvent surpasser les heuristiques traditionnelles en termes de maximisation de l'influence. Mais ils ont prouvé théoriquement que le problème d'optimisation de la sélection des utilisateurs les plus influents dans les deux modèles est NP-Difficile. Des recherches se concentrent principalement sur l'amélioration de l'efficacité des modèles. Par exemple, Leskovec et al. [Leskovec et al., 2007] présentent une stratégie d'optimisation appelée CELF (*Cost-Effective Lazy Forward*) qui pourrait accélérer la procédure de diffusion jusqu'à 700 fois. D'autres recherches comme [Klemm et al., 2012] et [Yang et al., 2016] étudient l'influence dans les réseaux dynamiques, c'est-à-dire qu'ils prennent en compte l'ajout, la suppression de nœuds ainsi que la modification du poids d'un lien. Zhang et al. proposent un état de l'art sur la maximisation et diffusion de l'influence dans les réseaux sociaux dans [Zhang et al., 2014].

Les modèles de diffusion considèrent que le choix des utilisateurs influents est réalisé. Or la détection des utilisateurs influents est une tâche importante car ils forment l'ensemble de départ sur lequel repose le processus de diffusion et de maximisation de l'influence. Dans la section suivante, nous rappelons les principaux travaux de recherche sur l'estimation de l'influence dans les réseaux sociaux et dans *Twitter*.

2.3/ L'INFLUENCE DANS LES RÉSEAUX SOCIAUX ET EN PARTICULIER DANS *Twitter*

L'estimation de l'influence dans les réseaux sociaux est une problématique majeure qui a donné lieu à un grand nombre de recherches et différentes approches et applications ont été proposées.

Dans [Leavitt et al., 2009], l'influence dans le réseau social *Twitter* est définie comme la capacité d'un utilisateur à provoquer une action chez un autre utilisateur. Le terme « action » désigne les différentes relations possibles entre les nœuds (utilisateurs). Par conséquent, la mesure de l'influence dans *Twitter* est un problème complexe puisque *Twitter* offre plusieurs types de relations (*retweet*, *réponse*, *mention*, *suivre*), qui peuvent être combinés pour former différentes interactions. En fonction du domaine et de l'intention de l'utilisateur, la combinaison de ces relations induit une sémantique différente qui doit être prise en compte dans l'estimation de l'influence. Par exemple, l'utilisation du préfixe « @ » au début d'un *tweet* sert à interpeller un utilisateur alors que « @ » suivi d'un nom de média à la fin d'un *tweet* permet une large diffusion du *tweet*.

Dans [Neves et al., 2015, Riquelme et al., 2016, Lü et al., 2016, Al-Garadi et al., 2018], les auteurs ont réalisé des états de l'art sur l'estimation de l'influence dans les réseaux sociaux dont *Twitter*. Nous avons identifié trois grands types d'approches. Elles sont basées sur des **mesures de popularité**, sur la **topologie du réseau** et sur la **fusion d'informations**. Les approches basées sur la topologie incluent les algorithmes *PageRank* et *HITS* pour classer les utilisateurs les plus influents. L'approche fusion d'informations étend les

approches topologiques.

À côté de cette recherche académique, il existe également des outils disponibles en ligne pour estimer le score d'influence tels que Klout [4], Kred [5], SocialMention [6] et SocialBakers [7]. Ces outils restent des boîtes noires car ils n'exposent pas les méthodes utilisées pour estimer l'influence ce qui ne permet pas à un utilisateur de comprendre comment l'influence a été calculée. Dans la suite, nous présentons les principaux travaux académiques pour chacune des approches précédentes.

2.3.1/ Approches basées sur des mesures de popularité

Les méthodes basées sur les mesures de popularité exploitent le résumé statistique des relations et des attributs des utilisateurs de *Twitter*. Comme nous l'avons vu en section 2.1.2, si de nombreux critères peuvent être pris en considération lors de l'analyse de *Twitter*, les mesures de popularité se focalisent sur le nombre de relations comme le nombre de *réponses*, *retweets* et *mentions*. Ces mesures utilisent les critères correspondant aux catégories connections locales et interactions de l'utilisateur du tableau 2.1.

Dans [Leavitt et al., 2009], les auteurs utilisent quatre critères : le nombre de *réponses*, de *retweets* et de *mentions*, en plus du nombre d'*abonnés*. Dans leur conclusion, ils encouragent les autres chercheurs à poursuivre l'étude sur l'influence en se basant non seulement sur le nombre d'*abonnés*, mais aussi sur les interactions entre les utilisateurs dans *Twitter*. Cha et al. [Cha et al., 2010] définissent trois mesures d'influence dans *Twitter*, le nombre d'*abonnés*, indiquant la taille de l'audience d'un utilisateur ou sa popularité, le nombre de *retweets*, indiquant la capacité d'un utilisateur à écrire du contenu transmissible à d'autres et le nombre de *mentions*, indiquant sa capacité à engager avec les autres des conversations. Les auteurs calculent la valeur de chaque relation pour 6 millions d'utilisateurs puis ils les comparent. Pour ce faire, ils trient les utilisateurs en fonction de chaque relation, puis, ils quantifient comment le classement d'un utilisateur varie selon les différentes relations. La corrélation de Spearman [8] est utilisée comme une mesure de la force d'association entre deux variables traduisant le classement. À partir des résultats obtenus, ils font trois observations. D'abord, ils ont constaté que le nombre d'*abonnés* représente la popularité d'un utilisateur et non son influence, et ils concluent que le nombre d'*abonnés* seul révèle très peu sur l'influence d'un utilisateur. Ensuite, ils déduisent que les utilisateurs peuvent voir leur influence varier selon différents sujets. Et enfin, ils estiment que l'influence n'est pas acquise spontanément ou accidentellement, mais par des efforts concertés tels que la limitation des *tweets* à un seul sujet. Dans [Lee et al., 2010a], les auteurs calculent le nombre cumulé d'utilisateurs, nommés lecteurs potentiels, qui ont vu un *tweet* et étudient comment ce nombre évolue au cours du temps. Les utilisateurs influents sont déterminés en se basant sur leur nombre de lecteurs potentiels. Leurs résultats montrent que la majorité des utilisateurs influents sont des médias car ils ont une influence notable dans la diffusion d'informations à des lecteurs potentiels. Cependant, les auteurs ne montrent pas comment savoir qu'un *tweet* est vu.

4. https://klout.com/home
5. http://home.kred/
6. http://www.socialmention.com
7. https://www.socialbakers.com
8. En statistique, la corrélation de Spearman est une mesure de dépendance statistique non paramétrique entre deux variables.

D'autre part, ils existent des travaux qui utilisent le contenu des *tweets* afin d'étudier l'influence. Suh et al. [Suh et al., 2010] ont analysé les facteurs qui ont un impact positif sur le nombre de *retweets* donc sur l'influence : les URLs, les hashtags, l'ancienneté du compte, le nombre d'*abonnés/abonnements*. Ils constatent que, parmi les critères relatifs au contenu des *tweets*, les URLs et les hashtags augmentent significativement le nombre de *retweets*. Parmi les critères contextuels, le nombre d'*abonnés/abonnements* ainsi que l'âge du compte semblent affecter le nombre de *retweets*, en revanche, ils constatent que le nombre de *tweets* précédemment émis par l'utilisateur n'entre pas en compte sur l'influence d'un utilisateur. Leur étude a porté sur 74 millions de *tweets*. Bakshy et al. [Bakshy et al., 2011] travaillent avec les cascades de diffusion d'URLs raccourcis. Ils considèrent que les utilisateurs à la source des URLs qui produisent les cascades les plus longues sont les plus influents. Les résultats présentés sont obtenus à partir d'une enquête effectuée auprès de 1,6 million d'utilisateurs sur une période de deux mois en 2009. Dans ce travail, la définition de l'influence est limitée à la capacité d'être le premier à publier une URL qui sera ensuite *retweetée*.

Les recherches appartenant à la catégorie des méthodes basées sur les mesures de popularité sont les premiers travaux sur l'influence dans *Twitter* (2009-2011). L'inconvénient de ces méthodes est qu'elles donnent un classement des utilisateurs selon des statistiques relatives aux critères utilisés mais ne proposent pas un score global de l'influence se basant sur toutes les relations existantes. Elles ne fournissent qu'une indication qualitative de l'existence d'influence et non une mesure quantitative. De plus, la structure du réseau auquel l'utilisateur appartient n'est pas considérée. Dans la section suivante, nous présentons des approches basées sur la topologie du réseau.

2.3.2/ Approches basées sur la topologie du réseau

Les approches basées sur la topologie du réseau reposent sur l'analyse structurelle du réseau social. Il s'agit de considérer l'utilisateur comme un nœud dans le réseau social et d'étudier la structure du réseau auquel il appartient. Les différentes mesures de centralité (centralité de degré, de proximité, d'intermédiarité, etc.) répondent à la question « Comment caractériser l'importance d'un nœud selon la structure du réseau ? ». Une mesure de centralité donne une valeur à chaque nœud, les valeurs produites lui donnent un rang en fonction de son importance ou position dans le réseau. Ces mesures ont été utilisées pour identifier les utilisateurs influents [Sun et al., 2011, de Arruda et al., 2014].

2.3.2.1/ Utilisation du degré de centralité d'un nœud et de son voisinage

Afin d'estimer l'influence, des chercheurs ont proposé des méthodes se basant sur les mesures de centralité.

Dans un réseau simple non orienté $G(V,E)$ avec V et E étant l'ensemble de nœuds et de liens respectivement, le degré de centralité d'un nœud i, désigné par $DC(i)$, est défini comme le nombre de voisins directement connectés avec i. Formellement,

$$DC(i) = \sum_j a_{ij}$$

avec A la matrice d'adjacence telle que $a_{ij} = 1$ si i et j sont connectés, 0 sinon. Dans les réseaux orientés, chaque lien est dirigé, ainsi, le degré de centralité d'un nœud i a

deux formes : la centralité in-degree k_i^{in} est le nombre d'arcs ayant i comme extrémité et la centralité out-degree k_i^{out} est le nombre d'arcs ayant i comme origine.

Le degré de centralité est l'indice le plus simple pour estimer l'influence des nœuds : plus un nœud a des connexions, plus son influence est importante. La simplicité et la faible complexité de calcul du degré de centralité lui donnent une large gamme d'applications. Cependant, le degré de centralité est imprécis dans l'estimation de l'influence des nœuds puisqu'il utilise des informations très limitées. Or, la localisation d'un nœud dans le réseau joue un rôle plus important que son degré de centralité qui ne prend en compte que ses voisins directs [Kitsak et al., 2010, Chen et al., 2012a]. Par exemple, comme le montre la figure 2.6, bien que le nœud 1 ait le plus grand degré de centralité ($DC(1) = 8$) parmi les 23 nœuds, la propagation d'information à partir du nœud 1 est limité puisque tous les voisins du nœud 1 ont un degré de centralité très faible. En revanche, le nœud 23 peut avoir une influence plus élevée bien qu'il ait un degré de centralité ($DC(23) = 5$) inférieur à celui du nœud 1.

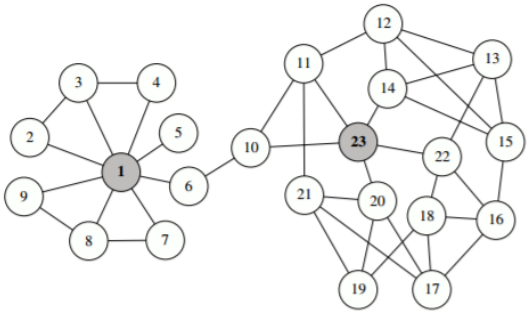


FIGURE 2.6 – Degré de centralité et influence (extrait de [Chen et al., 2012a])

D'autres mesures de centralité ont donc été proposées, telles que les centralités de proximité (*closeness centrality*) [Bavelas, 1950] et d'intermédiarité (*betweenness centrality*) [Freeman, 1977], qui permettent d'évaluer selon différents points de vue l'importance d'un nœud en fonction de ses relations avec les autres. Si ces mesures donnent de meilleurs résultats que le degré de centralité, elles présentent des problèmes de performance sur les réseaux réels. [Grando et al., 2018] trouvent que ces métriques ont des coûts de calcul élevés et des exigences qui entravent leurs applications dans les grands réseaux réels, ils proposent de les associer avec des techniques d'apprentissage automatique pour améliorer leurs performances. Ainsi, des extensions du degré de centralité sont proposées, nous les présentons dans la suite.

Chen et al. [Chen et al., 2012a] ont proposé LocalRank, une extension du degré de centralité qui tient compte du voisinage de profondeur 4 autour de chaque nœud. Le score LocalRank du nœud i, noté $LR(i)$ est défini par :

$$LR(i) = \sum_{j \in \tau_i} Q(j) \quad et \quad Q(j) = \sum_{k \in \tau_j} R(k)$$

avec τ_i l'ensemble des voisins directs de i et $R(k)$ la somme entre le degré de centralité de k et le nombre de voisins de voisins de k tel que $k \in \tau_j$ et $j \in \tau_i$. Si nous reprenons l'exemple de la figure 2.6, $LR(1) = 145$ et $LR(23) = 200$. Le calcul de LocalRank a une complexité en temps beaucoup plus faible que les centralités classiques ; en fait, la complexité de

LocalRank est $O(n < k >^2)$ avec $< k >$ le degré moyen dans le réseau. Cette mesure a été étendue aux réseaux pondérés par [Rezaei et al., 2015].

Après avoir proposé LocalRank, les mêmes auteurs [Chen et al., 2013a] ont proposé une méthode de classement local, nommée ClusterRank où non seulement le nombre de voisins les plus proches est considéré mais aussi leurs interactions grâce au cœfficient de clustering qui mesure à quel point les voisins d'un sommet sont connectés entre eux. Le score ClusterRank d'un nœud i est défini par :

$$CR(i) = f(c_i) \sum_{j \in \tau_i} (k_j^{out} + 1)$$

où c_i est le coefficient de clustering défini par :

$$c_i = \frac{\sum a_{jz}}{k_i^{out}(k_i^{out} - 1)}, j, z \in \tau_i$$

et $f(c_i) = 10^{-c_i}$ est une fonction qui permet de pénaliser les nœuds selon leur coefficient de clustering c_i. En effet, le cœfficient de clustering (voir la figure 2.7) joue généralement un rôle négatif dans les processus de diffusion [Eguiluz et al., 2002, Petermann et al., 2004]. En général, avec le même nombre de voisins, plus le coefficient de clustering d'un nœud est grand, plus son influence est faible. En effet, la propagation initiée à partir du nœud est alors plus susceptible d'être limitée dans une région locale puisque ses voisins interagissent étroitement entre eux plutôt qu'avec d'autres nœuds, au contraire, si ses voisins sont principalement liés avec des nœuds autres que les voisins du nœud initial, l'information se répandra rapidement sur une plus large échelle.

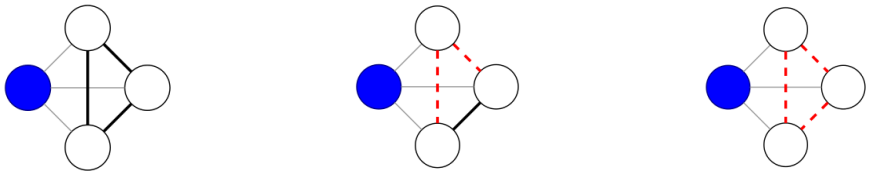


FIGURE 2.7 – Exemple de calcul du cœfficient de clustering pour le nœud bleu. Les liens noirs connectent les voisins du nœud bleu, et les liens rouges sont pour les liens non utilisés possibles. La définition standard du coefficient de clustering pour un nœud i est le nombre de liens reliant les voisins du nœud i (appelés triangles), divisé par le nombre total de liens possibles entre les voisins du nœud i. (Wikipédia)

Des travaux récents de [Zhao et al., 2017] proposent une mesure de Centralité Locale avec Coefficient (CLC) basée sur LocalRank et le coefficient de clustering utilisé dans ClusterRank. D'abord, ils calculent le nombre de nœuds voisins pour identifier les nœuds dans le centre du réseau et ceux qui présentent un "pont" entre deux sous-réseaux. Pour ces nœuds, ils calculent une fonction décroissante du coefficient de clustering c_i ($f(c_i) = e^{-c_i}$) et $CLC(i)$ comme suit :

$$CLC(i) = f(c_i) \times LR(i),$$

où $LR(i)$ est la mesure de LocalRank de i. Enfin, ils effectuent des expériences pour mesurer l'influence des nœuds sur des réseaux réels et d'autres générés par ordinateur en utilisant CLC et d'autres mesures telles que le degré de centralité, d'intermédiarité et de proximité. Les résultats montrent que les classements obtenus par la mesure proposée sont les plus semblables au classement réel, vérifiant ainsi que leur mesure reflète plus précisément l'influence des nœuds que les autres mesures.

En plus de LocalRank et ClusterRank, Chen et al. [Chen et al., 2013b] proposent l'algorithme KED. En effet, ils trouvent que les travaux antérieurs liés à l'étude de l'influence se concentrent énormément sur le nombre de chemins de propagation (c'est-à-dire sur le nombre de liens) sans tenir compte de la diversité des chemins qui peut améliorer la précision du classement. La figure 2.8 donne un exemple d'illustration de la diversité des chemins. Les nœuds rouges ont le même degré de centralité et le même nombre de seconds voisins les plus proches mais les distributions du degré de centralité de leurs voisins sont différentes. Dans ces deux réseaux, le nombre de chemin depuis le nœud rouge pour atteindre (influencer) chaque nœud est exactement le même et l'information peut se propager aux nœuds gris uniquement par un nœud bleu. Les chemins vers les nœuds gris dans la figure 2.8.a sont très diversifiés contrairement à la figure 2.8.b où la propagation de l'information aux nœuds gris s'effectuent à travers un unique nœud bleu. Intuitivement, l'information pourrait se propager plus facilement aux nœuds gris dans le réseau 2.8.a que dans le réseau 2.8.b.

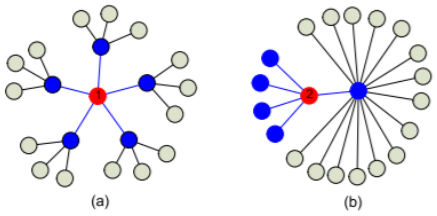


FIGURE 2.8 – Diversité des chemins et influence (extrait de [Chen et al., 2013b]).

La méthode KED combine l'information sur le nombre et la diversité des chemins. L'idée est que l'influence d'un nœud peut être réduite si la diffusion dépend de peu de chemins. La diversité des chemins est donné par le degré d'uniformité (*unevenness degree*) que l'on trouve en biologie dans l'étude des espèces [Hurlbert, 1971]. Mais ils se limitent à étudier la diversité des chemins de longueur 2 car il est difficile de tracer tous les chemins de diffusion à partir de chaque nœud.

Dans le même temps, le H-Index [Hirsch, 2005] a été exploité pour étudier l'influence, l'intention d'origine de cet index est de mesurer les impacts académiques des chercheurs ou des revues en fonction de leurs publications et citations. Il est défini comme la valeur maximale h tel qu'il existe au moins h publications, chacune avec des citations non inférieures à h. Le H-Index a été étendu pour quantifier l'influence des utilisateurs dans les réseaux sociaux [Korn et al., 2009]. Le H-Index d'un utilisateur i est défini comme le plus grand h tel que i a au moins h voisins ayant chacun un degré de centralité non inférieur à h. Mathématiquement, nous pouvons définir un opérateur H sur un nombre fini de variables réelles $\{x_1, x_2, ..., x_m\}$ qui renvoie l'entier maximal h tel que parmi $\{x_1, x_2, ..., x_m\}$ il y a au moins H éléments dont les valeurs ne sont pas inférieures à h. Par conséquent, le H-Index d'un nœud i, noté H_i dans un réseau social peut être écrit comme :

$$H_i = H(DC(j_1), DC(j_2), ..., DC(j_{k_i})), j \in \tau_i$$

Dans [Zhao et al., 2015], les auteurs introduisent un nouvel indice de centralité basé sur la structure communautaire du réseau pour identifier les nœuds influents. La notion de voisinage d'un nœud est élargie à celle de communauté. La centralité communautaire ($CbC(i)$) d'un nœud i considère à la fois le nombre et la taille des communautés qui sont directement liées par un nœud. Chaque nœud comporte deux types de liens : un lien fort est défini comme un lien entre les nœuds qui se trouvent dans la même communauté et un lien faible est défini comme un lien qui relie deux nœuds appartenant à des communautés différentes. Étant donné que les connexions sont beaucoup plus fortes dans une communauté que celles entre différentes communautés, l'influence des nœuds est calculée à la fois par les caractéristiques des liens et la taille des communautés. La centralité communautaire $CbC(i)$ d'un nœud i est définie par :

$$CbC_i = \sum_{w=1}^{c} d_{i_w} \frac{S_w}{N}$$

où c est le nombre de communautés dans le réseau, d_{i_w} est le nombre de liens entre le nœud i et les autres nœuds dans la communauté w, S_w est le nombre de nœuds dans la communauté w (sa taille), N est le nombre total de nœuds dans le réseau.

[Ghalmane et al., 2018] considèrent aussi la structure de la communauté afin d'étudier l'influence. Ils trouvent que les réseaux réels sont modulaires et que cette propriété doit être considérée lors de l'estimation de l'influence. Dans un réseau modulaire, un nœud a deux types d'influence : une influence locale (sur les nœuds de sa communauté) et une influence globale (sur les nœuds des autres communautés). Sur la base de cette idée, les auteurs proposent la Centralité Modulaire, une extension des mesures de centralité standard, tels que la centralité d'intermédiarité et de proximité, aux réseaux modulaires. La Centralité Modulaire est un vecteur à deux dimensions. Son premier composant quantifie l'influence locale d'un nœud dans sa communauté tandis que le second composant quantifie son influence globale sur les autres communautés du réseau. Les expériences sont menées sur des réseaux synthétiques modulaires. Des comparaisons approfondies avec les mesures de centralité standard montrent que les mesures de centralité modulaire fournissent des classements plus précis. Des simulations sur des réseaux réels ont également été réalisées. Comme leur structure de communauté est inconnue, un algorithme de détection de communauté a été utilisé. Les résultats confirment que les classements de nœuds basés sur la centralité modulaire sont plus précis que ceux établis par les mesures de centralité standard conçues pour des réseaux sans structure de communauté.

En conclusion, un des avantages des méthodes présentées précédemment est qu'elles sont basées sur des mesures de centralités faciles à mettre en œuvre, mais elles ne tiennent compte que d'une information locale correspondant à une partie du réseau, généralement les voisins des voisins. Cependant, l'influence d'un nœud n'est pas seulement déterminée par le nombre de ses voisins, mais aussi par l'influence de ses voisins, connue sous le nom du *mutual enhancement effect* (effet d'amélioration mutuelle) [Wittenbaum et al., 1999]. Ainsi, d'autres approches basées sur l'algorithme de décomposition k-shell ont été proposées. Nous les présentons dans la suite.

2.3.2.2/ UTILISATION DU RÉSEAU DANS SA GLOBALITÉ

Kitsak et al. [Kitsak et al., 2010] partent aussi du principe que la localisation d'un nœud dans le réseau peut déterminer son influence. Ainsi, un nœud, situé au centre du réseau et ayant peu de voisins très influents, peut avoir plus d'influence qu'un nœud ayant un plus grand nombre de voisins moins influents. Partant de ce postulat, l'algorithme de décomposition k-shell est utilisé [Seidman, 1983]. Son principe est d'attribuer un indice de référence ks pour chaque nœud tel que les nœuds ayant les valeurs les plus faibles sont situés à la périphérie du réseau tandis que les nœuds avec les valeurs les plus élevées se trouvent au centre du réseau, ce sont alors ces nœuds qui auront le plus d'influence, les nœuds les plus internes forment ainsi le noyau du réseau (voir la figure 2.9).

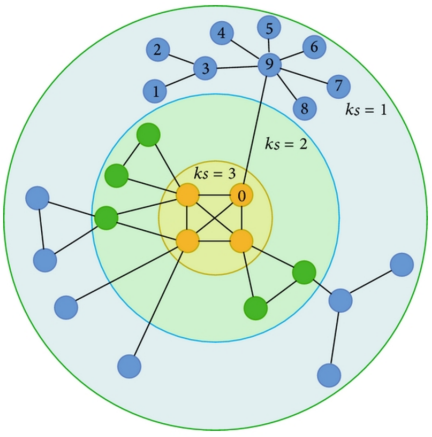


FIGURE 2.9 – Représentation schématique de la décomposition k-shell (extrait de [Jin et al., 2015])

De nombreuses extensions de l'algorithme de décomposition k-shell ont été proposées. Bae et al. dans [Bae et al., 2014] remarquent que le k-shell ne parvient pas à générer un classement pertinent des influenceurs car il attribue à trop de nœuds le même indice k-shell. Ainsi, ils proposent une nouvelle mesure pour estimer l'influence d'un nœud dans un réseau en utilisant les indices k-shell de ses voisins. Les résultats expérimentaux sur des réseaux réels et artificiels, par rapport à un modèle de propagation épidémique, montrent que la méthode proposée peut quantifier l'influence du nœud avec plus de précision et fournir un classement plus précis que les autres classements.

Liu et al. [Liu et al., 2015] présentent une version améliorée du k-shell en supprimant les liens redondants et en prenant en compte la diversité des liens pour trouver les "vrais" influenceurs. Ils ont constaté que la précision du classement s'est grandement améliorée.

Brown et al. [Brown et al., 2011] ont adapté l'algorithme de décomposition k-shell aux caractéristiques du réseau *Twitter* en utilisant une échelle logarithmique afin de produire des valeurs de k-shell moins nombreuses et plus significatives pour ce type de réseau. En effet, en utilisant l'algorithme k-shell classique pour analyser l'influence dans *Twitter*, ils ont observé que les résultats sont fortement biaisés.

Wei et al. [Wei et al., 2015] ont adapté la décomposition k-shell aux réseaux pondérés. Ils proposent une méthode de pondération des liens, le poids d'un lien est la somme des degrés de centralité des nœuds extrémités du lien. Pour le réseau pondéré construit, la

méthode de décomposition k-shell est adaptée. Enfin, les auteurs étudient leur méthode sur le processus de propagation épidémique des modèles SIR et SI dans des réseaux complexes réels et vérifient que leur méthode est efficace pour détecter l'influence des nœuds.

Ma et al. [Ma et al., 2016] prennent comme métaphore la formule de gravité de Newton où la masse est la valeur k-shell, et la plus petite distance entre deux nœuds est vue comme la distance, la formule de gravité est alors utilisée pour identifier les influenceurs. Ils utilisent également le modèle épidémique classique SIR pour vérifier les performances de leur méthode. La comparaison entre l'indice de centralité par gravité et certaines mesures bien connues, telles que le degré de centralité, la centralité d'intermédiarité, la centralité de proximité et le k-shell, indique que la méthode peut identifier efficacement les utilisateurs influents dans les réseaux réels ainsi que dans les réseaux artificiels.

Enfin, Basaras et al. [Basaras et al., 2013] constatent que le nombre de calculs important du k-shell rend inapproprié l'analyse des réseaux dynamiques. Ainsi, ils proposent μ-power community index (μ-PCI), une combinaison de k-shell et de la centralité d'intermédiarité dans le but de détecter des influenceurs dans des zones denses du réseau. D'après les auteurs, μ-PCI est adapté à tout type de réseau quelle que soit sa taille ou sa dynamique. Une évaluation expérimentale des deux algorithmes (k-shell et μ-PCI) ainsi que du degré de centralité démontre la supériorité de μ-PCI dans la détection d'utilisateurs influents.

En conclusion, les méthodes basées sur la centralité du nœud et de son voisinage ainsi que le *mutual enhancement effect* ont démontré leur efficacité dans la détection d'utilisateurs influents. Cependant, ces mesures ne concernent que les liens sortants d'un nœud lors de l'estimation de l'influence. Dans la section suivante, nous présentons les algorithmes de prestige dont la principale idée est de considérer les liens pointants vers le nœud.

2.3.2.3/ UTILISATION DES ALGORITHMES DE PRESTIGE

Les algorithmes de prestige font une distinction entre les liens sortants et les liens entrants. Un nœud prestigieux est défini comme cible de nombreux liens, en d'autres termes, pour calculer le prestige d'un nœud, il faut considérer les liens dirigés ou pointés vers le nœud. Afin d'estimer l'influence en se basant sur le prestige, des recherches proposent de classer les utilisateurs en utilisant des algorithmes basés sur le PageRank [Page et al., 1999] et le HITS [Kleinberg, 1999]. Dans la suite, nous présentons ces algorithmes et leurs extensions.

L'idée principale du PageRank est que « Les pages les plus importantes (des sites Web) sont susceptibles de recevoir plus de liens à partir d'autres pages ». PageRank suppose que l'importance d'une page Web est déterminée par la quantité et la qualité des pages qui s'y rattachent. Initialement, chaque nœud (c'est-à-dire, page) reçoit une valeur PR. Ensuite, chaque nœud distribue uniformément sa valeur PR à ses voisins à travers les liens sortants. Mathématiquement, la valeur PR du nœud i à l'étape t est :

$$PR_i(t) = \sum_{j=1}^{n} a_{ji} \frac{PR_j(t-1)}{k_j^{out}}$$

où n est le nombre total de nœuds dans le réseau. L'itération ci-dessus s'arrêtera si les

valeurs PR de tous les nœuds atteignent un état stationnaire. Un inconvénient majeur du processus de marche aléatoire du PR ci-dessus est que la valeur PR d'un nœud pendulaire (c'est-à-dire un nœud avec zéro pour out-degree) ne peut pas être redistribuée, ainsi la formule ne peut pas garantir la convergence. Pour résoudre ce problème, un facteur de téléportation a été introduit en supposant que l'utilisateur navigue entre les pages Web en suivant les liens avec la probabilité s, et quitte la page actuelle pour une page aléatoire avec une probabilité de $1-s$, $s \in [0,1]$ est généralement fixé empiriquement autour de $0,85$. En conséquence, la formule du PR est modifiée comme suit :

$$PR_i(t) = s\sum_{j=1}^{n} a_{ji}\frac{PR_j(t-1)}{k_j^{out}} + (1-s)\frac{1}{n}$$

PageRank a été utilisé bien au-delà de son intention initiale qui était de classer les pages Web [Gleich, 2015]. Il a été appliqué pour classer des objets selon leur structure de réseau : classer des images [Jing et al., 2008], des livres [Meng, 2009, Bollen et al., 2006, West et al., 2010] ou des gènes [Morrison et al., 2005] et des protéines [Morrison et al., 2005] en biologie et bioinformatique, classer des molécules [Mooney et al., 2012] en chimie, classer les régions cérébrales [Zuo et al., 2011] et les neurones [Crofts et al., 2011] en neuroscience, classer les noms d'hôtes [Arasu et al., 2002] et les interfaces de programmation [Kim et al., 2013] dans les systèmes d'information complexes, classer les nœuds influents [Weng et al., 2010] dans les réseaux sociaux, classer les scientifiques [Liu et al., 2005, Ding et al., 2009], les documents [Su et al., 2011, Sayyadi et al., 2009], classer les joueurs [Radicchi, 2011] et les équipes [Govan et al., 2008] dans les sports, etc.

Dans le cadre de l'influence, l'hypothèse est qu'un utilisateur influent doit être en relation avec de nombreux voisins très influents. Ainsi, plusieurs adaptations de l'algorithme PageRank ont été proposées afin de classer les utilisateurs selon leur influence dans les réseaux sociaux et pour certains dans *Twitter*, [Riquelme et al., 2016] en recense dix-sept. Dans la suite nous présentons les extensions de base et leurs adaptations aux caractéristiques des réseaux sociaux.

Daniel Tunkelang a proposé TunkRank [Tunkelang, 2009] pour classer un utilisateur i à partir de l'influence de ses *abonnés* en prenant en compte les faits suivants :

— si i appartient aux *abonnés* de j, alors il y a $\frac{1}{|abonnement(i)|}$ probabilité que i lise un *tweet* émis par j où $abonnement(i)$ est l'ensemble des utilisateurs que i suit ;

— si i lit un *tweet* émis par j, il y a une probabilité de p pour que i le *retweete*.

L'influence de i est alors donnée par :

$$TunkRank(i) = \sum_{j \in abonnés(i)} \frac{1 + p * TunkRank(j)}{|abonnement(j)|}$$

[Kwak et al., 2010, Ashwini et al., 2015] proposent des approches similaires mais travaillent sur la relation *retweet*.

Silva et al. [Silva et al., 2013] étudient le problème de l'identification des utilisateurs influents et du contenu pertinent dans *Twitter*. Ils proposent ProfileRank, un algorithme inspiré de PageRank, qui exploite le principe selon lequel le contenu pertinent est créé et propagé par des utilisateurs influents et des utilisateurs influents créent un contenu pertinent. ProfileRank travaille sur un graphe biparti utilisateur-*tweet* où ils existent deux types de lien : un lien qui lie le *tweet* à l'utilisateur qui l'a écrit et un lien qui lie l'utilisateur

au *tweet* qu'il a écrit ou *retweeté*. Ainsi, le graphe est représenté par deux matrices *tweet*-utilisateur et utilisateur-*tweet*, ensuite, les auteurs appliquent PageRank sur les deux matrices pour obtenir deux classements : un sur les *tweets* et un sur les utilisateurs. Les résultats des expérimentations montrent que l'influence de l'utilisateur calculée par ProfileRank est fortement corrélée avec le nombre de *retweets* des utilisateurs.

NodeRanking [Pujol et al., 2002] est une autre variante du PageRank avec deux différences : il travaille sur des graphes pondérés et le facteur de téléportation n'est pas fixé pour tout le graphe mais calculé pour chaque nœud et dépend du out-degree du nœud. Selon les auteurs, cette caractéristique permet à NodeRanking de s'adapter dynamiquement aux graphes avec différentes topologies. Ainsi, les équations suivantes illustrent l'algorithme Noderanking.

$$P_{choose}(i \to j) = \frac{w(i \to j)}{\sum w(i \to \tau_i^{out})} \qquad Inf(i) = Inf(i) + \frac{P_{choose}(i \to j) \times Inf(j)}{F_i}$$

$P_{choose}(i \to j)$ renvoie le prochain nœud j à visiter à partir du nœud i, j est sélectionné avec une probabilité calculée en fonction du poids w du lien entre i et j, τ_i^{out} est l'ensemble des nœuds pointés par i. Les auteurs introduisent un facteur de correction F_i qui dépend à la fois de l'influence de i et de l'influence totale de l'ensemble des nœuds du graphe. F_i est un facteur pour contrôler la croissance de l'influence et maintenir sa valeur dans un intervalle limité de valeurs. Sans ce facteur, les valeurs calculées tendent vers l'infini car l'influence d'un nœud devient de plus en plus grande lors du processus. Pour chaque nœud i, F_i est initialisé avec la somme de l'influence de tous les nœuds du graphe au moment où l'influence du nœud i est calculée. L'influence initiale d'un nœud i doit être positive et le facteur F_i doit être supérieur ou égal à 1.

Ghosh et al. [Ghosh et al., 2012] proposent CollusionRank, une approche basée sur PageRank mais où les nœuds identifiés comme spammeurs sont initialisés avec une valeur négative. Ainsi, un utilisateur est pénalisé pour avoir suivi des spammeurs et non pour être suivi par des spammeurs, le score CollusionRank d'un nœud est calculé en fonction du score de ses *abonnements* (et non de ses *abonnés*). Par conséquent, les utilisateurs qui suivent un plus grand nombre de spammeurs ou qui suivent ceux qui à leur tour suivent des spammeurs, reçoivent un score négatif et sont poussés vers le bas du classement.

Dans [Lü et al., 2011], les auteurs proposent l'algorithme LeaderRank qui se base sur la relation *abonnés*. LeaderRank est basé sur PageRank mais le réseau est rendu fortement connecté par l'introduction d'un nœud g ayant deux arcs orientés e_{gi} et e_{ig} vers chaque nœud i du réseau d'origine, de sorte que le réseau devient fortement connecté ce qui permet à l'algorithme de converger plus rapidement et de donner de meilleurs résultats que PageRank en termes d'efficacité de classement et de robustesse. Li et al. [Li et al., 2014] améliorent LeaderRank en introduisant un mécanisme de pondération ; les nœuds pondérés avec leurs différents nombres d'*abonnés* obtiennent des rangs différents.

Dans [Weng et al., 2010], les auteurs proposent TwitterRank afin d'estimer l'influence des utilisateurs en tenant compte de la similarité des sujets (*via* les hashtags associés aux *tweets*) entre les utilisateurs et de la structure des liens, dans l'article le lien *abonnés* est exploité. Ils construisent un graphe pondéré d'utilisateurs où le poids des liens indique la similarité des sujets entre deux utilisateurs. Ensuite, ils exécutent une variante de l'algorithme PageRank sur le graphe pondéré orienté, l'algorithme est exécuté séparément

pour chaque sujet afin de trouver par sujet les utilisateurs influents. Dans un premier temps, les auteurs appliquent la LDA (*Latent Dirichlet Allocation*) pour extraire les sujets, DT_{it} contient le nombre de fois où un mot apparaît dans les *tweets* de l'utilisateur i pour le sujet t. Ensuite pour chaque sujet t, ils construisent une matrice, équivalente à la matrice d'adjacence utilisée dans le PageRank, cette matrice est définie par :

$$P_t(i,j) = \frac{|\beta_j|}{|\beta_{\tau_i}|} \times sim_t(i,j)$$

où $|\beta_j|$ est le nombre de *tweets* publiés par j, $|\beta_{\tau_i}|$ est le nombre de *tweets* publiés par tous les *abonnés* de i. $Sim_t(i,j)$ est la similarité entre i et j pour le sujet t avec $Sim_t(i,j) = |DT_{it} - DT_{jt}|$. Après cette étape l'algorithme PageRank est utilisé. Deux autres travaux ([Chen et al., 2012c, Katsimpras et al., 2015]) se sont aussi basés sur l'utilisation du PageRank et des sujets pour estimer l'influence sujet par sujet.

Bien que l'idée de TwitterRank soit prometteuse, les résultats expérimentaux montrent qu'il existe des utilisateurs qui *suivent* d'autres utilisateurs sans présence de thématiques similaires entre eux. La méthode a aussi ignoré d'autres critères importants tels que les *mentions* et les *réponses*.

L'algorithme PageRank a toujours été comparé à l'algorithme HITS (*Hyperlink-Induced Topic Search*), un algorithme qui permet de mesurer l'autorité d'une page Web par rapport à d'autres à travers l'analyse de liens. HITS a été développé en 1999 par Jon Kleinberg [Kleinberg, 1999] et il est parfois considéré comme précurseur de l'algorithme PageRank. HITS attribue deux scores pour chaque page : un score d'autorité qui estime la valeur du contenu de la page, et un score de hub qui estime la valeur de ses liens vers d'autres pages. L'algorithme HITS s'appuie sur un principe simple : tous les sites Web n'ont pas la même importance, et ne jouent pas le même rôle. Certains sites sont des « sites de référence », leurs pages sont souvent citées dans d'autres sites. Ces sites de référence sont appelés « authorities ». Alors que les « authorities » sont les véritables sites qui contiennent de l'information, d'autres sites appelés « hubs » jouent un rôle tout aussi important, bien qu'ils ne contiennent pas, à proprement parler, de contenu informatif. Il s'agit des sites qui contiennent des liens vers les « authorities », et qui permettent de « structurer » le Web en indiquant où sont les pages intéressantes sur un sujet donné.

Le principe du HITS est simple, pour commencer le classement, les scores de hub et d'autorité de tous les nœuds sont initialisés à 1 : $\forall i, auth(i) = 1$ et $hub(i) = 1$. Ensuite des itérations de deux types de mises à jour sont exécutées : la règle de mise à jour de l'autorité et la règle de mise à jour du hub. L'application $k-$étape de l'algorithme HITS implique d'appliquer d'abord k fois la règle de mise à jour de l'autorité, puis la règle de mise à jour du hub. Elles sont calculées de la manière suivante :

— Règle de mise à jour de l'autorité : $auth(i) = \sum_{j=1}^{n} hub(j)$
— Règle de mise à jour du hub : $hub(i) = \sum_{j=1}^{n} auth(j)$

où n est le nombre total de nœuds connectés à i et j est un nœud connecté à i. Comme l'application directe et itérative des règles de mise à jour conduit à des valeurs divergentes, il est nécessaire de normaliser les valeurs des scores après chaque itération. Ainsi, les valeurs de hub et d'autorité obtenues à partir de ce processus finiront par converger vers des valeurs stables.

Romero et al. [Romero et al., 2011] travaillent avec la relation *retweets* de *Twitter* et considèrent l'influence comme le niveau de propagation d'un contenu dans le réseau

social et ils estiment que l'influence d'un utilisateur dépend non seulement de la taille de son audience mais aussi de sa passivité qui correspond au fait qu'il ne transmet pas l'information au réseau. Ils proposent IP-Algorithm qui assignent à chaque utilisateur i un score d'influence I_i et un score de passivité P_i qui respectivement correspondent aux scores de hub et d'autorité dans le HITS. Pour calculer ces deux scores, les auteurs utilisent E et w_{ij} (respectivement l'ensemble des relations et le poids sur l'arc (i, j)) et définissent :

— un taux d'acceptation u_{ij} qui représente l'influence acceptée par j de i normalisée par l'influence totale que j accepte de tous les autres utilisateurs :

$$u_{ij} = \frac{w_{ij}}{\sum_{k|(k,j)\in E} w_{kj}}$$

— un taux de rejet, opposé du taux d'acceptation, qui représente le rejet normalisé de i par j :

$$v_{ji} = \frac{1 - w_{ji}}{\sum_{k|(j,k)\in E}(1 - w_{jk})}$$

À partir de ces deux taux, les scores d'influence et de passivité sont respectivement calculés par :

$$I_i = \sum_{j|(i,j)\in E} u_{ij} P_j$$

$$P_i = \sum_{j|(j,i)\in E} v_{ji} I_j$$

Les auteurs montrent que cet algorithme a une meilleure précision que d'autres mesures d'influence telles que PageRank, le nombre d'*abonnés* et le nombre de *mentions*. Ce travail est le premier à prendre en compte la passivité d'un nœud dans le calcul de l'influence.

En conclusion, l'inconvénient des algorithmes basés sur la topologie du réseau est de considérer les informations du nœud, c'est-à-dire les liens des nœuds, sans considérer les interactions complexes entre les nœuds à travers des séquences de liens. Les méthodes basées sur les algorithmes PageRank et HITS ont pour principale lacune qu'elles ne traitent que d'une seule relation, c'est-à-dire un seul type de liens à la fois.

2.3.3/ APPROCHES BASÉES SUR LA FUSION D'INFORMATION

Dans des travaux récents, la fusion d'information est considérée afin de contourner les limitations des approches précédentes.

Dans [Simmie et al., 2013], les auteurs proposent la combinaison de deux modèles pour classer les nœuds influents dans *Twitter* : l'algorithme PageRank et un HMM (*Hidden Markov Model*) [9]. Le modèle permet d'observer l'évolution de l'influence à travers le temps

9. Un modèle de Markov caché est un modèle statistique dans lequel le système modélisé est supposé être un processus markovien de paramètres inconnus. Contrairement à une chaîne de Markov classique, où les transitions prises sont inconnues des nœuds mais où les états d'une exécution sont connus, dans un modèle de Markov caché, les états d'une exécution sont inconnus du nœud (seuls certains paramètres sont connus).

en utilisant les trois relations *retweet*, *mention* et *réponse*. Le modèle est évalué sur une enquête considérée comme une réalité du terrain. Cette enquête permet de valider le modèle et les résultats montrent une amélioration de la précision du classement par rapport aux méthodes basées uniquement sur la topologie du réseau pour la zone du réseau que les auteurs ont échantillonnée. Enfin, en utilisant l'aspect évolutif du HMM, ils ont essayé de prévoir les états futurs en utilisant les preuves actuelles. L'algorithme de prédiction surpasse significativement une collection de modèles, en particulier à court terme (1-3 semaines).

Dans [Muruganantham et al., 2015], les auteurs combinent les mesures de centralité grâce à la méthode TOPSIS afin d'identifier les nœuds influents. La méthode TOPSIS (*Technique for Order Preference by Similarity to an Ideal Solution*) est présentée dans [Chen et al., 2012b], il s'agit d'une méthode de décision multi-critères pour trouver une solution dans un ensemble fini d'alternatives. Le principe de base est que l'alternative choisie doit avoir la distance la plus courte de la solution idéale positive et la distance la plus éloignée de la solution idéale négative. D'abord, pour chaque nœud, les différentes mesures de centralité sont calculées : degré de centralité, centralité d'intermédiarité, centralité de proximité et centralité de vecteur propre. TOPSIS est ensuite appliquée pour combiner ces différentes mesures sur un sous-ensemble de nœuds sélectionnés sur des caractéristiques d'homophilie (âge, travail, niveau d'éducation). Les utilisateurs sont classés d'après le rang fourni par TOPSIS. Enfin, la méthode proposée est testée sur un ensemble de données réelles du réseau *Facebook* et a montré son efficacité selon les auteurs.

Ainsi, les modèles proposés par [Simmie et al., 2013] et [Muruganantham et al., 2015] diffèrent des autres approches par la combinaison de deux modèles (respectivement PageRank et HMM ; mesures de centralité et TOPSIS). Le but était de classer les nœuds selon leur influence. Toutefois, les méthodes ne donnent pas une mesure d'influence pour les nœuds donnés. De plus, lors de la combinaison, ils ne représentent pas l'incertitude sur l'importance des informations à combiner les unes par rapport aux autres. Pour ceci, des approches ont eu recours à la théorie des fonctions de croyances afin d'assurer la fusion d'information tout en exprimant l'incertitude.

Wei et al. [Wei et al., 2013] proposent une nouvelle mesure de centralité dans les réseaux pondérés, appelée centralité évidentielle (EVC), qui combine le degré de centralité d'un nœud et son importance dans le réseau mesurée par la somme des poids des liens incidents. Pour combiner ces deux informations, Wei et al. utilisent la théorie des fonctions de croyance [Shafer, 1976]. La théorie des fonctions de croyance permet d'exprimer une croyance sur un élément de l'univers étudié. Dans ce travail, pour chaque nœud on donne la part de croyance dans une influence forte (respectivement faible) pour le degré de centralité et la part de croyance dans une influence forte (respectivement faible) pour l'importance du nœud. La théorie des fonctions de croyance permet de combiner ces différentes parts de croyance. Des exemples illustrent l'efficacité de la méthode proposée.

Deux points à améliorer dans le travail de Wei et al. ont été soulevés par les travaux de Gao et al. [Gao et al., 2013] et Ren et al. [Ren et al., 2015] qui ont proposé respectivement la centralité évidentielle semi locale (ESC) et la centralité evidentielle à structure locale (ELSC). D'abord, un paramètre de correction est associée à la part de croyance dans l'influence forte (respectivement faible) pour la mesure sur le nœud afin de tenir compte

de la distribution des degrés [10] au lieu de suivre une distribution uniforme. Deuxièmement, pour pallier les défauts du degré de centralité à déterminer précisément l'influence d'un nœud, celui-ci est remplacé par des mesures qui prennent en compte pour Gao et al. la structure autour du nœud avec l'utilisation de LocalRank et pour Ren et al. les connections entre les voisins d'un nœud *via* le coefficient de clustering. La théorie des fonctions de croyance est toujours utilisée pour combiner la mesure sur le nœud et son importance. Enfin, afin d'évaluer la performance de la méthode proposée, les deux articles utilisent les modèles épidémiques pour simuler le processus de diffusion dans des réseaux réels. Les résultats des expérimentations montrent l'efficacité des méthodes dans l'identification des nœuds influents. Ces trois recherches (EVC, ESC et ELSC) travaillent sur des réseaux pondérés, sur un réseau non pondéré, Mo et al. [Mo et al., 2015] proposent la centralité évidentielle compréhensive (CEC) qui combinent, selon le même principe que les autres travaux, plusieurs mesures sur les nœuds : degré de centralité, centralité de proximité et centralité d'intermediarité. Pour évaluer la performance de la méthode proposée, le modèle susceptible infecté (SI) est adopté pour examiner la diffusion de l'influence des nœuds classés par différentes mesures de centralité. Les résultats indiquent que la méthode proposée est légèrement meilleure que le degré de centralité, la centralité de proximité et la centralité d'intermediarité.

Jendoubi et al. [Jendoubi et al., 2017] utilisent aussi la théorie des fonctions de croyance dans leur approche afin de prendre en compte les imperfections des données issues de *Twitter*. Afin d'estimer l'influence, les auteurs considèrent trois types de relation : *suivre*, *retweet* et *mention*. Ensuite, pour chaque lien, ils attribuent un poids w_f, w_m et w_r représentent respectivement les poids des relations *suivre*, *mention* et *retweet* en utilisant les formules suivantes :

$$w_f(i,j) = \frac{|\tau_i^{out} \cap \tau_j^{in}| + 1}{|\tau_{max}^{out}|} \qquad w_m(i,j) = \frac{|M_j(i)|}{|M_{max}|} \qquad w_r(i,j) = \frac{|R_i(j)|}{|T_{max}|}$$

où τ_i^{out} représente l'ensemble des nœuds pointés par i (successeurs de i), τ_j^{in} est l'ensemble des nœuds qui ont pointé vers j (prédécesseurs de j), $\tau_{max}^{out} = max_{i \in V}\ \tau_i$, V étant l'ensemble des nœuds du réseau, $M_j(i)$ est l'ensemble des *tweets* de j dans lesquels i est mentionné, $M_{max} = max_{i \in V}\ M_i$, $R_i(j)$ est l'ensemble des *tweets* de i qui sont *retweetés* par j et $T_{max} = max_{i \in V}\ T_i$, T_i étant l'ensemble des *tweets* de i.

Ensuite, les poids des nœuds sont déduits en additionnant les poids de leurs liens sortants. Le processus d'estimation d'influence contient trois étapes. La première étape consiste à définir des parts de croyance sur l'importance (I) et la passivité (P) de chaque nœud par rapport aux trois types de relations considérées, ensuite ces parts de croyances sont combinées. Dans la deuxième étape, ils introduisent un nouvel aspect d'influence résumée par le fait que «je suis plus influent si je suis connecté aux utilisateurs influents». Par conséquent, pour chaque nœud, ils utilisent les parts de croyance combinées obtenues de la première étape pour mettre à jour les poids des liens entrants. Dans la dernière étape, les poids des liens mis à jour sont utilisés pour estimer l'influence des nœuds. En effet, pour estimer l'influence d'un nœud sur un autre, les poids mis à jour des liens existants entre les deux nœuds sont combinés. L'approche proposée par [Jendoubi et al., 2017] considère de nombreux aspects d'influence telle que la combinaison de plusieurs types de liens qui ont été utilisés séparément dans les travaux d'états de l'art. En plus du modèle

10. Dans la théorie des graphes, la distribution des degrés donne la quantité de sommets par nombre de connexions.

d'estimation d'influence, les auteurs proposent un modèle de maximisation d'influence évidentiel dont le but est de trouver un ensemble de nœuds qui maximise une fonction objective. Cependant, les expérimentations ont démontré que la maximisation d'influence selon leur modèle est NP-Difficile. De plus, l'influence tient compte d'un seul niveau alors que l'influence peut être exercée sur plusieurs niveaux (par exemple, un utilisateur peut influencer un autre utilisateur à travers des utilisateurs intermédiaires). En outre, ils existent d'autres relations dans *Twitter* que les auteurs n'ont pas considérées telle que la *réponse*. Et enfin, le contenu des *tweets* n'est pas exploité afin d'étudier la polarité de l'influence (positive ou négative).

En conclusion, les méthodes de centralité évidentielle et ses extensions permettent d'identifier les nœuds influents tout en répondant aux limites des méthodes précédentes, à travers la fusion des différentes informations, la représentation de l'incertitude grâce à la théorie des fonctions de croyance et l'estimation de l'influence sous forme d'un score exploitable pour le classement d'utilisateurs. Cependant, vu la diversité des liens du réseau, la considération des importances des types de liens les uns par rapport aux autres n'est pas prise en compte. Or, il est important de considérer l'importance du type de lien lors de l'estimation de l'influence car ils représentent des sémantiques différentes et n'ont pas la même importance dans l'estimation de l'influence. Par exemple, dans le réseau *Twitter*, le lien *retweet* n'a pas la même importance que le lien *réponse* car le premier permet une meilleure diffusion d'information.

2.4/ SYNTHÈSE ET CONCLUSION

Enfin, pour conclure les travaux d'état de l'art sur l'étude de l'influence des différentes catégories, la figure 2.10 résume les approches d'estimation de l'influence dans les réseaux sociaux et représentent leurs limites. Le tableau 2.2 représente la synthèse des travaux de recherche cités dans chacune de ces catégories. Ces recherches ne travaillent pas sur le même type de réseau. Certaines approches travaillent sur des réseaux simples, or, vu la complexité des réseaux réels, d'autres approches travaillent sur des réseaux orientés, pondérés ou les deux à la fois. De plus, ils ne travaillent pas sur le même réseau social, la majorité des recherches travaillent sur le réseau *Twitter* vu la facilité d'accès à ses données, certains travaillent sur d'autres réseaux tels que *Facebook*, d'autres ne spécifient pas le réseau utilisé. En outre, certaines approches prennent en compte la sémantique des liens du réseau étudié c'est-à-dire le réseau est représenté par différents types de liens qui sont pris en compte lors de l'estimation de l'influence. Par ailleurs, le contenu de l'information diffusée (par exemple les *tweets*) a été considéré par quelques travaux. Enfin, de nouveaux critères ont été proposés par certaines approches, tels que le nombre de lecteurs potentiels et la passivité. Ainsi, les critères pris en compte dans les différents travaux de recherches sont nombreux et dépendent du réseau étudié. Cependant, vu la diversité des critères que l'on peut considérer lors de l'analyse de *Twitter* (voit le tableau 2.1), ces travaux de recherche nous aident à déterminer les critères à prendre en compte dans l'étude de l'influence.

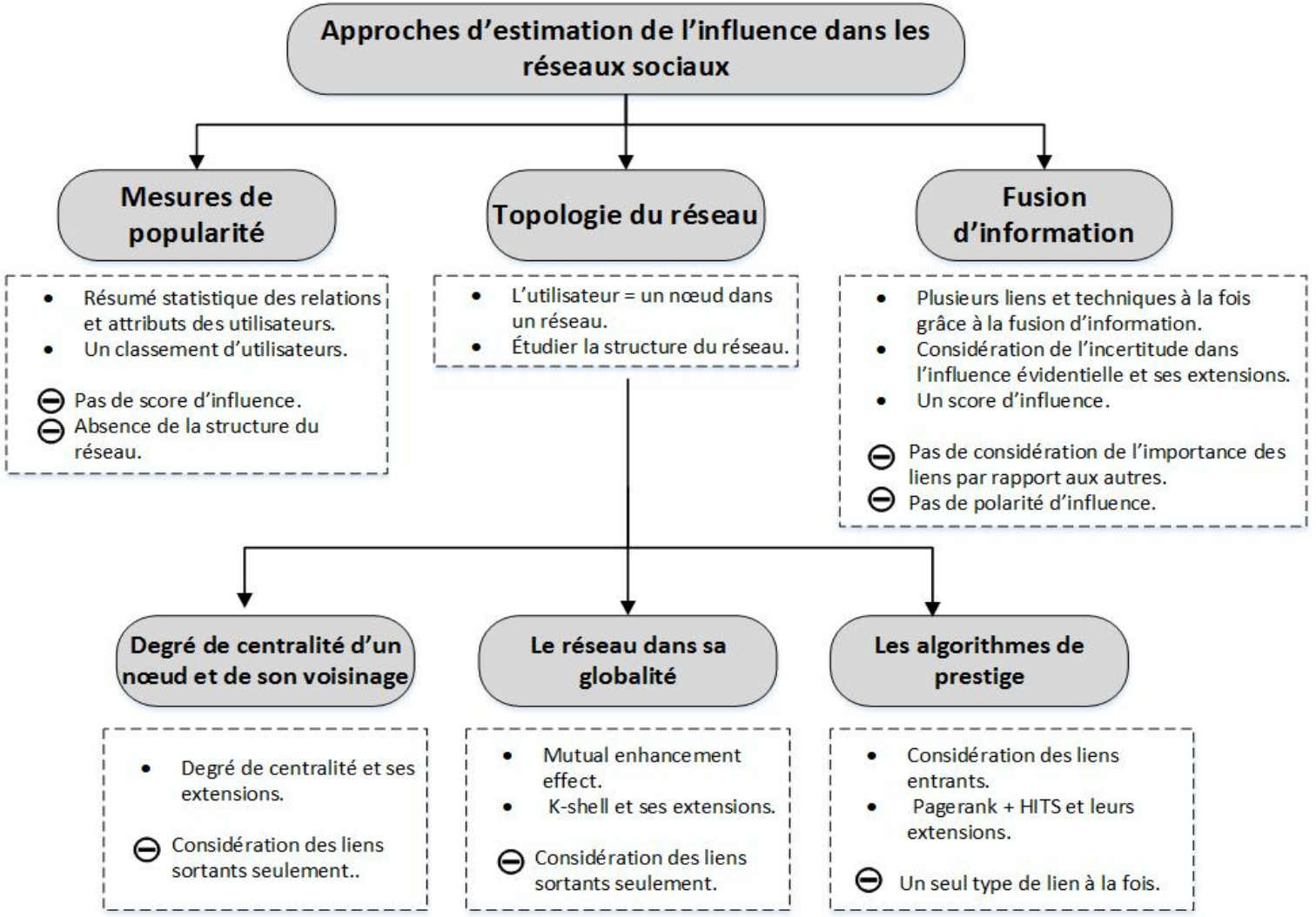


FIGURE 2.10 – Taxonomie des approches d'estimation de l'influence dans les réseaux sociaux

TABLE 2.2 – Synthèse des travaux de recherche sur l'influence

	Article	Type de réseau	Sémantique des liens	Contenu	Autre critère	Algorithme/Mesure
Popularité	[Leavitt et al., 2009]	*Twitter*	*Abonnés, Retweet, Mention, Réponse*	✗	✗	Nombre de liens par type
	[Cha et al., 2010]	*Twitter*	*Abonnés, Retweet, Mention*	✗	✗	Nombre de liens par type
	[Lee et al., 2010a]	*Twitter*	✗	✗	Lecteurs potentiel	Nombre de lecteurs potentiels
	[Suh et al., 2010]	*Twitter*	*Abonnés, Retweet*	URL + Hashtag	Age du compte	Nombre de *retweets*
	[Bakshy et al., 2011]	*Twitter*	*Retweet*	URL	✗	Cascades d'URLs
Topologie du réseau	[Chen et al., 2012a]	Simple	✗	✗	✗	LocalRank : Degré de centralité des voisins
	[Chen et al., 2013a]	Orienté	✗	✗	✗	ClusterRank : LocalRank + cœfficient de clustering
	[Zhao et al., 2017]	Orienté	✗	✗	✗	LCL : LocalRank + ClusterRank
	[Chen et al., 2013b]	Orienté	✗	✗	✗	KED : Nombre de liens + diversité de chemins
	[Korn et al., 2009]	Simple	✗	✗	✗	H-Index
	[Zhao et al., 2015]	Simple	✗	✗	✗	CbC : centralité communautaire
	[Ghalmane et al., 2018]	Modulaire	✗	✗	✗	Centralité modulaire
	[Kitsak et al., 2010]	Simple	✗	✗	✗	K-shell
	[Basaras et al., 2013]	Simple	✗	✗	✗	μ-power community index
	[Bae et al., 2014]	Simple	✗	✗	✗	K-shell + K-shell des voisins
	[Liu et al., 2015]	Simple	✗	✗	✗	K-shell + diversité des liens
	[Wei et al., 2015]	Pondéré	✗	✗	✗	K-shell pondéré (Wks)
	[Ma et al., 2016]	Simple	✗	✗	✗	indice de centralité par gravité
	[Brown et al., 2011]	Simple	✗	✗	✗	Echelle logarithmique de K-shell
	[Tunkelang, 2009]	*Twitter*	*Abonnés*	✗	✗	TunkRank : Prise en compte de l'influence des *abonnés*
	[Silva et al., 2013]	*Twitter*	*Retweet*	Analyse de sentiment	✗	ProfileRank : Contenu pertinent
	[Pujol et al., 2002]	Pondéré, Orienté	✗	✗	✗	NodeRanking : PageRank sur graphes pondérés
	[Ghosh et al., 2012]	*Twitter*	*Abonnements*	✗	✗	CollusionRank : Pénaliser l'*abonnement* aux spammeurs
	[Lü et al., 2011]	*Twitter*	*Abonnés*	✗	✗	LeaderRank : Pagerank + introduction nœud g
	[Weng et al., 2010]	*Twitter*	*Abonnés*	Hashtag	✗	TwitterRank : PageRank + Similarité de sujet
	[Romero et al., 2011]	*Twitter*	*Retweet*	✗	Passivité	IP-Algorithm : HITS Influence et Passivité
Fusion	[Simmie et al., 2013]	*Twitter*	*Retweet, Mentions, Réponses*	✗	✗	PageRank + HMM
	[Muruganantham et al., 2015]	Simple *Facebook*	✗	✗	✗	Mesures de centralité + TOPSIS
	[Wei et al., 2013]	Pondéré	✗	✗	✗	EVC : Degré de centralité + Importance du nœud
	[Gao et al., 2013]	Pondéré	✗	✗	✗	ESC : EVC + Centralité semi locale
	[Ren et al., 2015]	Pondéré	✗	✗	✗	ELSC : ESC + Connections topologiques entre voisins
	[Mo et al., 2015]	Simple	✗	✗	✗	CEC : EVC, ESC et ELSC adaptés aux réseaux simples
	[Jendoubi et al., 2017]	*Twitter*	*Abonnés, Retweets, Mentions*	✗	✗	Combinaison avec la théorie des fonctions de croyance

Dans *Twitter*, de nombreux critères de classification des utilisateurs peuvent être utilisés. Ainsi, afin d'étudier l'influence, il est difficile de choisir les critères à prendre en compte. Dans l'état de l'art, plusieurs approches sont proposées dans le but d'étudier l'influence. L'étude de ces différentes approches montre qu'elles sont nombreuses et que les critères utilisés diffèrent d'une approche à une autre.é Certains travaux se sont basés sur des mesures de popularité, c'est-à-dire le résumé statique des liens de *Twitter*. Cependant, cette catégorie ne prend pas en compte la structure du réseau étudié. Afin de contourner cette limite, des approches se sont basées sur la topologie du réseau qui repose sur l'analyse structurelle du réseau incluant les mesures de centralité, l'algorithme k-shell et les algorithmes de prestige tels que PageRank et HITS. L'inconvénient de cette catégorie est d'ignorer la diversité des liens du réseau et les interactions complexes entre les nœuds à travers des séquences de liens. Une autre catégorie étend les approches topologiques pour assurer la fusion d'informations issues des différents critères. Par conséquent, les méthodes suivies afin d'étudier l'influence sont nombreuses. Néanmoins, ces travaux de recherche nous aident à déterminer les critères à prendre en compte dans l'étude de l'influence.

En particulier, nous concluons qu'il est d'abord important de se baser sur la topologie du réseau, c'est-à-dire de considérer l'utilisateur comme un nœud et d'étudier la structure du réseau auquel il appartient. Il s'agit d'exploiter ses relations avec les voisins avec lesquels il est connecté directement ou indirectement à travers les interactions complexes vues comme une séquence de liens. En plus, la fusion des informations issues des différents critères permet une meilleure étude d'influence. Il est important aussi de représenter l'incertitude lors de la fusion pour exprimer l'importance des critères combinés les uns par rapport aux autres, pour ceci la théorie des fonctions de croyance a démontré son efficacité. Enfin, il est important d'estimer l'influence sous forme d'un score global qui peut être exploitable pour le classement d'utilisateurs.

Ainsi, les travaux existants nous ont permis de tirer des conclusions intéressantes. Cependant, la diversité et la sémantique des liens du réseau n'ont pas été prises en compte dans les approches proposées. Or, il est important de considérer les types de lien car ils ne représentent pas la même importance dans l'estimation de l'influence. Par exemple, dans le réseau *Twitter*, le lien *retweet* n'a pas la même importance que le lien *réponse* car le premier permet une meilleure diffusion d'information. Par ailleurs, la combinaison de plusieurs types de liens avec de l'incertitude n'a pas été considérée. Or, il nous parait important, pour mesurer l'influence, de tenir compte des degrés d'incertitude sur les poids attribués aux différents liens selon leur importance. Dans des recherches existantes, la théorie des fonctions de croyance est exploitée pour mesurer l'influence dans des réseaux pondérés et complexes avec l'objectif commun de modifier les mesures de centralité existantes. Néanmoins, la théorie des fonctions de croyance n'a pas été exploitée pour mesurer l'influence sur le réseau *Twitter* avec des motifs d'interactions au lieu des mesures de centralité. De plus, aucun travail de recherche existant n'étudie la **polarité de l'influence**. En effet, il est important d'analyser le sentiment exprimé dans les *tweets* afin de déterminer si l'influence d'un nœud est positive ou négative.

II

CONTRIBUTIONS

3

MODÉLISATION DES RÉSEAUX SOCIAUX : DES MODÈLES DE GRAPHES THÉORIQUES AUX RÉSEAUX MULTIPLEXES – APPLICATION À *Twitter*

Dans ce chapitre, nous introduisons les notions fondamentales sur lesquelles est basée notre modélisation de *Twitter*. Nous rappelons d'abord les modèles théoriques utilisés pour étudier les propriétés des réseaux issus de données réelles tels que *Twitter*, *Facebook*, etc., appelés réseaux complexes [1] ou graphes de terrain. Afin de mieux comprendre les caractéristiques de ce type de réseaux, il est important d'étudier les recherches théoriques menées depuis les années 1960 sur des réseaux qui ne correspondent pas exactement à des réseaux complexes mais qui ont permis de mettre en avant des caractéristiques que l'on peut trouver dans les réseaux complexes. Ensuite, nous présentons les différentes modélisations des réseaux complexes dans la théorie des graphes. Nous détaillons la modélisation que nous proposons pour *Twitter* c'est-à-dire un réseau multiplexe hétérogène. À ce niveau notre contribution consiste, dans un premier temps, à déterminer le modèle le plus adapté et à le spécialiser pour spécifier les relations engendrées par les interactions entre les utilisateurs de *Twitter*. Dans un second temps, nous exploitons ce réseau multiplexe afin de mesurer l'influence des utilisateurs. Pour ce faire nous utilisons une extension pour les réseaux multiplexes de l'algorithme PageRank. Notre contribution sur ce point consiste à représenter les données extraites de *Twitter* dans le modèle de réseau multiplexe que nous avons défini et d'en exploiter la richesse des relations traduisant l'influence et enfin d'étudier les paramètres qui modifient le comportement de l'algorithme.

3.1/ MODÈLES THÉORIQUES DES RÉSEAUX COMPLEXES

Dans de nombreux contextes applicatifs, de grands graphes ne montrant pas de structure forte apparente comme des graphes réguliers, des arbres, sont appelés réseaux com-

1. Dans la suite, nous utiliserons la terminologie de réseau complexe pour désigner des données du monde réel et le terme de graphe pour désigner les objet mathématiques/théoriques associés.

plexes ou graphes de terrain, par opposition aux graphes explicitement construits par un modèle ou une théorie. Par exemple, des réseaux biologiques (interactions protéiques, topologie du cerveau ou circulation sanguine), des réseaux technologiques (Internet, voies routières, ferrées et aériennes), des réseaux sociaux comme *Twitter* ou Facebook [Lesne, 2006]. Dans un réseau complexe, le comportement global ne peut pas être déduit des comportements individuels de ses entités. Les propriétés de ces réseaux émergent au niveau du réseau pris dans son ensemble à partir de l'évolution et de l'interaction non planifiées des nombreux éléments du réseau (entités et liens), ce qui justifie l'adjectif « complexe » [Barrat, 2013].

Différents modèles théoriques ont été proposés et éclairent les propriétés des réseaux complexes. L'un des plus populaires est le modèle de **graphe aléatoire**. Le premier modèle de graphe aléatoire a été proposé par deux mathématiciens Paul Erdös et Alfréd Rényi dans [Erdős et al., 1960]. Le modèle consiste à établir un lien entre chaque paire de nœuds avec une probabilité p, l'existence de chaque lien est indépendant de celle des autres. Une caractéristique importante du modèle est la distribution de probabilité $P(K)$ qu'un nœud soit connecté par un lien direct avec K autres nœuds, ce qu'on appelle la distribution des degrés. Deux propriétés ont été déterminées à partir du modèle : si le graphe est grand, la distribution des degrés suit une loi de Poisson [2], ainsi malgré des liens aléatoires, beaucoup de sommets auront à peu près le même degré ; une composante connexe de grande taille ($n^{2/3}$) se forme quand le degré moyen est égal à 1.

Les propriétés d'un réseau complexe diffèrent d'un graphe aléatoire et d'autres caractéristiques doivent être prises en compte lors de sa modélisation. Un réseau complexe de N nœuds et M liens peut être comparé à un graphe aléatoire ayant en moyenne le même nombre de liens, c'est-à-dire un graphe de N nœuds défini par la probabilité $p = 2M/N(N-1)$. Malgré tout, grâce à ses concepts (distribution des degrés, diamètre), le modèle d'Erdös et Rényi a fourni un paradigme permettant d'étudier les réseaux aléatoires et les phénomènes qui peuvent s'y produire [Bollobás, 1985].

Le modèle de réseaux **petit monde** (*small world networks*) proposé par Watts et Strogatz [Watts et al., 1998] trouve son origine dans les travaux de Stanley Milgram [Milgram et al., 1967]. L'expérience de Milgram consiste à demander à des habitants de Omaha (dans le Middle West) de faire parvenir une lettre à un destinataire de Boston (sur la Côte Est), qu'ils ne connaissent pas, en utilisant comme intermédiaires des personnes de leur entourage. Milgram a constaté que la moyenne des chaînes parvenues au destinataire n'était que de 5 à 6. Cette expérience a permis de confirmer la thèse de Frigyes Karinthy [Karinthy, 1929] selon laquelle toutes les personnes du globe sont reliées par une chaîne d'au plus 5 maillons, devenue dans sa version populaire les six degrés de séparation [Kleinfeld, 2002]. Dans le modèle de Watts et Strogatz, le graphe de départ est un graphe k-régulier, c'est-à-dire tous les nœuds ont le même degré k. L'idée est de présenter une façon simple de transformer ce graphe k-régulier en graphe aléatoire. À chaque étape, un lien est supprimé de façon aléatoire avec une probabilité p, et un lien est ajouté aléatoirement. Watts et Strogatz proposent deux mesures : 1) $L(p)$ qui désigne la longueur moyenne du plus court chemin entre les paires de nœuds lorsque p varie ; 2) $C(p)$ qui désigne le cœfficient de clustering. Ce cœfficient est en rapport avec la notion de transitivité dans le graphe dont l'idée est traduite par le fait que les voisins des voisins

2. En théorie des probabilités et en statistiques, la loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé, si ces événements se produisent avec une fréquence moyenne ou espérance connue et indépendamment du temps écoulé depuis l'événement précédent (source Wikipédia).

d'un nœud sont souvent ses voisins. Une forte transitivité dans le graphe se traduit par le fait que, du point de vue topologique, on trouve beaucoup de triangles. Les indicateurs $C(p)$ et $L(p)$ évoluent de façon différente. La distance moyenne entre les nœuds diminue rapidement tandis que celle du coefficient de clustering reste stable puis décroît plus rapidement. Watts et Strogatz estiment que pour des valeurs intermédiaires de p, les réseaux restent assez structurés, mais avec une faible longueur moyenne de chemins. Les graphes petit monde sont caractérisés par un grand nombre de nœuds, un nombre de liens loin de la saturation, un degré important de clustering et une faible distance moyenne.

Le modèle de réseau **sans échelle** (*scale-free network*) représente des réseaux dans lesquels quelques sommets sont très fortement connectés et un très grand nombre de sommets très faiblement connectés. Ils sont caractérisés par une distribution des degrés suivant une loi de puissance [3] [Barabási et al., 1999].

Les différences observées entre les réseaux sociaux, les pages Web et les réseaux sans échelle suggèrent que ces réseaux complexes se sont construits non pas par hasard mais suivant des principes organisateurs et soumis à divers mécanismes de sélection et d'attachement préférentiel qui est un processus dynamique au cours duquel les nœuds choisissent de rejoindre une "classe". La classe n'est pas choisie "au hasard" mais en fonction de la population présente entre les nœuds partageant une ou des caractéristiques communes sur un ou des attributs particuliers (par exemple la situation géographique).

Les trois modèles théoriques des réseaux que nous venons de présenter ne répondent pas de façon satisfaisante à la description des réseaux complexes. Empiriquement, on constate que les réseaux complexes possèdent plusieurs propriétés :

- étudiés globalement, ils sont peu denses. En effet, le nombre de liens est faible par rapport au nombre total de liens pouvant exister. Cependant, à un niveau de détail plus fin, les réseaux complexes révèlent des structures denses, avec la présence de nombreux liens formant des triangles ;
- la distribution du nombre de liens est hétérogène. On retrouve des nœuds très connectés, mais aussi beaucoup de nœuds peu connectés ;
- la distance moyenne entre les différents nœuds est faible. En effet, le nombre moyen de liens à suivre pour aller d'un nœud choisi aléatoirement à un autre est peu élevé.

Afin de représenter des réseaux complexes, des modélisations sous forme de graphe ont été proposées. Dans la section suivante, nous détaillons ces différentes modélisations, en prenant comme exemple *Twitter*.

3.2/ MODÉLISATION DES RÉSEAUX SOCIAUX SOUS FORME DE GRAPHE – LE CAS PARTICULIER DE *Twitter*

Les réseaux sociaux sont constitués de plusieurs entités interagissant dans des modèles complexes. Par exemple, dans *Twitter*, plusieurs types de nœuds sont possibles (utilisateurs, *tweets*, hashtags, url, etc.) ainsi que différentes relations (*retweet*, *mention*, *suivre*, etc.). Afin d'étudier ce réseau, différentes dimensions peuvent être exploitées, par exemple

3. La loi de puissance est une relation mathématique entre deux quantités. Si une quantité est la fréquence d'un évènement et l'autre la taille d'un évènement, alors la relation est une distribution de la loi de puissance si les fréquences diminuent très lentement lorsque la taille de l'évènement augmente (source Wikipédia).

les interactions entre utilisateurs (*suivre*, *retweet*, etc.), les actions des utilisateurs (*écrire*, etc.), et la structure des *tweets* (hashtag, URL, etc.). Afin de représenter les réseaux complexes tels que *Twitter* tout en ayant la possibilité de visualiser leurs différentes entités, plusieurs modèles de graphe ont été proposés allant du graphe simple à des formes plus complexes.

3.2.1/ GRAPHE SIMPLE ET MULTI-GRAPHE

Un graphe simple est défini comme $G = (V, E)$ avec V un ensemble non vide de nœuds et E un ensemble non vide d'arêtes reliant deux éléments de V, E est un sous-ensemble de $V \times V$ [Barnes, 1969]. Dans un graphe simple, une seule arête peut exister entre deux nœuds alors que dans un multi-graphe il peut exister plusieurs arêtes reliant une même paire de nœuds.

Dans [Drakopoulos et al., 2016], les auteurs modélisent le réseau *Twitter* comme un graphe simple avec le modèle de données de Neo4j[4] : les nœuds sont les utilisateurs et les liens représentent la relation *suivre*. Dans le modèle de Neo4j, les nœuds et les relations peuvent porter des propriétés. Pour un utilisateur, ses propriétés sont son compte, le nombre de *tweets* émis et *retweetés*, le nombre d'*abonnés*, de hashtags utilisés et la fréquence d'émission des *tweets*. Les auteurs utilisent des requêtes Cypher[5] pour calculer un résumé statistique proche des mesures de popularité présentées en section 2.3.1 du chapitre état de l'art. La modélisation proposée par [Drakopoulos et al., 2016] semble intéressante, mais la seule relation présente dans le graphe est la relation *suivre*. Or, dans les réseaux sociaux réels tels que *Twitter*, nous pouvons avoir plusieurs types de relations entre deux mêmes nœuds. Ainsi, une modélisation sous forme d'un **multi-graphe avec labels** est plus proche de la réalité. La figure 3.1 présente l'implémentation de *Twitter* avec Neo4j dans la plateforme SNFreezer[6] [Basaille et al., 2016]. Les objets (*tweets*, utilisateurs, hashtags, etc.) sont les nœuds et les relations sont décrites par les liens. Cependant, ces modèles prennent mal en compte les dimensions temporelles et spatiales que l'on trouve dans *Twitter*.

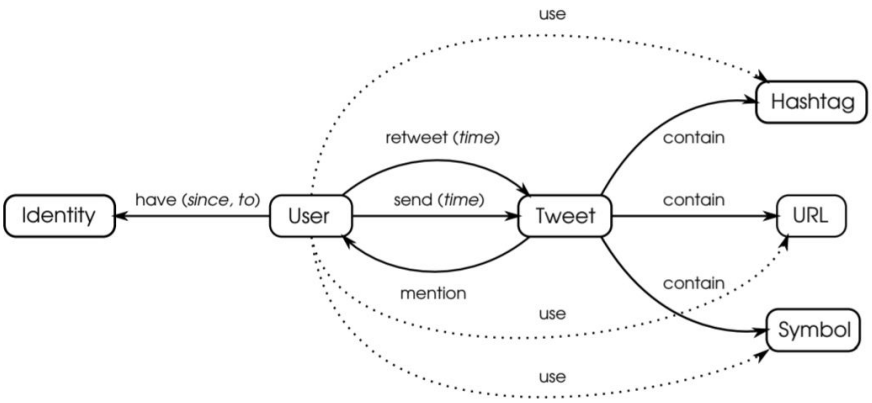


FIGURE 3.1 – Représentation multi-graphe labellé de *Twitter* [Basaille et al., 2016]

4. Neo4j est un système de gestion de base de données graphe NoSQL au code source libre, développé en Java par la société suédo-américaine Neo technology.
5. http://neo4j.org/learn/cypher
6. SNFreezer est un outil de collecte et d'analyses de *tweets* développé par l'équipe de recherche avec laquelle je travaille, https://github.com/EricLeclercq/SNFreezer.

La modélisation traditionnelle des réseaux complexes sous forme de graphe simple ou multi-graphe n'est pas suffisante, un nouveau cadre plus général est fourni par les hypergraphes.

3.2.2/ HYPERGRAPHE

Un hypergraphe H est un couple (V, E) où V est un ensemble de nœuds et E est un ensemble d'hyperliens. Chaque hyperlien $e \in E$ peut contenir arbitrairement plusieurs nœuds, E est donc défini comme un sous-ensemble de 2^V. Un hypergraphe est constitué de plusieurs graphes avec des hyperliens qui les relient [Ghoshal et al., 2009, Zlatić et al., 2009].

Dans [Fang et al., 2014], les auteurs développent TSIM (*Topic-Sensitive Influencer Mining*), une approche de fouille de nœuds influents dans les réseaux sociaux en se basant sur une modélisation hypergraphe. Plus précisément, ils prennent *Flickr* comme plate-forme d'étude. Les utilisateurs de *Flickr* interagissent les uns avec les autres par le biais d'images. L'estimation de l'influence est déterminée avec une approche d'apprentissage basée sur la modélisation hypergraphe. Dans l'hypergraphe, les nœuds représentent les utilisateurs et les images, et les hyperliens sont utilisés pour décrire des relations multi-types, notamment des relations de contenu visuel-textuel entre images, ainsi que des liens sociaux entre les utilisateurs et les images. TSIM commence par répartir les sujets en exploitant les images fournies par les utilisateurs, puis déduit l'influence dans différents sujets pour chaque nœud de l'hypergraphe. Dans leur scénario, les utilisateurs s'influencent mutuellement à travers les images, ce qui se reflète dans les comportements indirects des liens favoris ou de commentaires des utilisateurs. Des expériences sur un ensemble de données du monde réel de *Flickr* de plus de 50 000 images et 70 000 liens de type commentaire et favori ont démontré l'efficacité de l'approche. Ils démontrent également que TSIM peut améliorer considérablement les performances des applications de suggestion d'amis et de recommandation de photo.

[Amato et al., 2016] proposent un nouveau modèle de données pour les réseaux sociaux qui s'appuie sur la structure de données des hypergraphes pour représenter de manière simple tous les types de relations des réseaux sociaux. La modélisation permet de combiner des informations sur les utilisateurs et le contenu. Ils présentent également certaines fonctions de classement des utilisateurs. Les expériences concernant l'efficacité de l'approche pour soutenir des activités de recherche d'informations multimédia sont rapportées et discutées.

Twitter peut être modélisé comme un hypergraphe orienté avec V l'ensemble des nœuds qui représentent les utilisateurs V_U, les *tweets* V_T et les objets V_O tels que les hashtags, les URLs et les symboles (emoji), les intersections des ensembles pris deux à deux sont vides, et un ensemble de relations R entre ces nœuds. Ces deux ensembles sont définis comme suit : $V = \{V_U \cup V_T \cup V_O\}$ et $R = \{R_i, i = 1, \ldots, m\}$ avec $R_i : V^p \rightarrow \mathbb{N}$.

Par exemple nous pouvons considérer les relations *Écrire*, *Retweet*, *Mention*, *Suivre*, *Réponse* et *Contenir*. Ces relations ne s'appliquent pas toujours entre le même type de nœuds (voir le tableau 3.1), par exemple la relation *Suivre* peut exister entre les nœuds utilisateurs, la relation *Mention* peut s'appliquer entre un *tweet* et un utilisateur mais aussi entre deux utilisateurs lorsque l'on agrège le nombre de mentions d'un utilisateur fait par l'utilisateur source ($Mention_U$) ou la relation $Mention_{ctx}$ qui agrège les relations *Écrire* et *Mention*. Les relations n-aire sont aussi nécessaires pour décrire l'usage des différents opérateurs, par exemple l'utilisation conjointe de RT et @ dans le même *tweet* ou pour

détailler les objets contenus dans un *tweet*, par exemple le *tweet* t_{1234} peut contenir une URL et deux hashtags.

TABLE 3.1 – Exemple de relations possibles dans *Twitter*

Relation	**Signature**
Écrire	*utilisateur* $\times$ *tweet* $\rightarrow \mathbb{B}$
Suivre	*utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{B}$
$Retweet_T$	*tweet* $\times$ *tweet* $\rightarrow \mathbb{B}$
$Retweet_U$	*utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{N}$
$Retweet_{ctx}$	*utilisateur* $\times$ *tweet* $\times$ *utilisateur* $\times$ *tweet* $\rightarrow \mathbb{B}$
Mention	*tweet* $\times$ *utilisateur* $\rightarrow \mathbb{B}$
$Mention_U$	*utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{N}$
$Mention_{ctx}$	*utilisateur* $\times$ *tweet* $\times$ (*utilisateur*)$^n \rightarrow \mathbb{B}$
Réponse	*tweet* $\times$ *tweet* $\rightarrow \mathbb{B}$
$Réponse_U$	*utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{N}$
$Réponse_{ctx}$	*utilisateur* $\times$ *tweet* $\times$ *utilisateur* $\times$ *tweet* $\rightarrow \mathbb{B}$
Contenir	*tweet* $\times$ *objet* $\rightarrow \mathbb{B}$
$Contenir_U$	*utilisateur* $\times$ *objet* $\rightarrow \mathbb{N}$
$Contenir_{ctx}$	*utilisateur* $\times$ *tweet* $\times$ (*objet*)$^n \rightarrow \mathbb{B}$

Prenons l'exemple suivant d'un réseau *Twitter* avec les nœuds et relations décrites ci-dessous :

$\mathbf{R}_1$ **:** ***Suivre*** $: (u_1, u_2) \rightarrow 1$
$\mathbf{R}_2$ **:** $\boldsymbol{Retweet}_{ctx} : (u_1, t_1, u_2, t_2) \rightarrow 1$
$\mathbf{R}_3$ **:** $\boldsymbol{Mention}_{ctx} : (u_2, t_2, u_3) \rightarrow 1$
$\mathbf{R}_4$ **:** $\boldsymbol{Mention}_{ctx} : (u_2, t_5, u_3) \rightarrow 1$
$\mathbf{R}_5$ **:** $\boldsymbol{Mention}_U : (u_2, u_3) \rightarrow 2$
$\mathbf{R}_6$ **:** $\boldsymbol{Réponse}_{ctx} : (u_3, t_3, u_4, t_4) \rightarrow 1$
$\mathbf{R}_7$ **:** $\boldsymbol{Contenir}_{ctx} : (u_4, t_4, h_1, h_2) \rightarrow 1$

Ce réseau est modélisé par l'hypergraphe donné en figure 3.2 avec :
$V = \{u_1, u_2, u_3, u_4, t_1, t_2, t_3, t_4, t_5, h_1, h_2\}$,
$R = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7\} = \{\{u_1, u_2\}, \{u_1, t_1, u_2, t_2\}, \{u_2, t_2, u_3\}, \{u_2, t_5, u_3\}, \{u_2, u_3\}, \{u_3, t_3, u_4, t_4\}, \{u_4, t_4, h_1, h_2\}\}$.

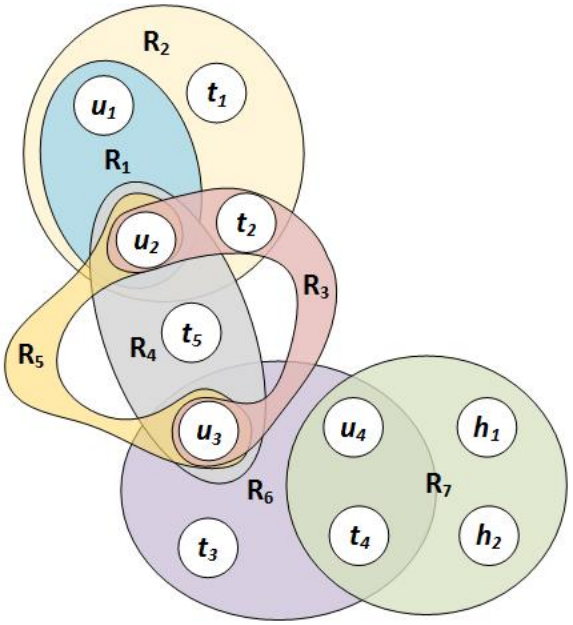


FIGURE 3.2 – Exemple d'une modélisation sous la forme d'un hypergaphe de *Twitter*

L'hypergraphe peut être représenté par sa matrice d'incidence A_{ir} à coefficients entiers appartenant à l'ensemble $\{0, +1, -1\}$ telle que chaque colonne correspond à un hyperlien r, et chaque ligne à un nœud i de H. Si un hyperlien r arrive au nœud i, alors $A_{ir} = 1$, si un hyperlien r sort du nœud i, alors $A_{ir} = -1$, $A_{ir} = 0$ sinon.

$$
\begin{array}{c}
 \\
u_1 \\
u_2 \\
u_3 \\
u_4 \\
t_1 \\
t_2 \\
t_3 \\
t_4 \\
t_5 \\
h_1 \\
h_2
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 \end{array} \\
\begin{pmatrix}
-1 & -1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -1 & -1 & -1 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\end{array}
$$

Bien que les hypergraphes offrent un moyen de modéliser *Twitter* selon plusieurs points de vue sur une même relation, cette modélisation reste limitée puisque les hyperliens n'ont pas de sémantique explicite et que certaines informations sont perdues. Par exemple, les mentions peuvent être utilisées plusieurs fois dans un même *tweet*, ainsi l'arité n'est pas fixée. Nous pouvons les modéliser par un ensemble de relations *Mention* ou $\textit{Mention}_{ctx}$ mais l'ordre dans lequel les utilisateurs sont mentionnés est perdu. De plus, dans des réseaux réels tels que *Twitter*, nous pouvons atteindre des milliers d'interactions, avec une telle modélisation, nous seront confrontés à plusieurs types d'hyperliens, ce qui peut compliquer la lisibilité des interactions. Par ailleurs, comparé au graphe simple, peu d'algorithmes ont été développés pour les hypergraphes.

HyperGraphDB [7] est un système de gestion de base de données graphe dont le modèle repose sur les hypergraphes avec Java comme langage de requêtes. Il a été utilisé dans le cadre d'un projet opérationnel pour la modélisation de réseaux sociaux multimédia [Amato et al., 2016].

3.3/ MODÉLISATION DES RÉSEAUX SOCIAUX SOUS FORME DE RÉSEAUX MULTI-COUCHES

3.3.1/ TYPOLOGIE DES RÉSEAUX MULTI-COUCHES

Récemment, plusieurs travaux de recherche ont proposé d'améliorer et de généraliser les approches existantes pour prendre en compte des réseaux incluant plusieurs sous-systèmes et différents niveaux de connectivité entre les systèmes. Bien que la terminologie varie largement, en fonction de chaque cas spécifique, de tels réseaux peuvent généralement être appelés réseaux multi-couches. Une couche représente un aspect ou une caractéristique, par exemple, une catégorie de liens comme des liens professionnels ou familiaux. Les notations utilisées pour des graphes simples doivent être étendues pour permettre de représenter les structures qui ont des couches en plus des nœuds et des liens.

Dans [Kivelä et al., 2014], les auteurs synthétisent et discutent des travaux existants sur la modélisation et l'analyse des réseaux multi-couches. Un réseau multi-couches est défini comme un quadruplet $\mathcal{M} = (V, L, V_M, E_M)$ où :

- V est l'ensemble de nœuds ;
- $L : (L_a)_{a=1}^{d}$ est une suite d'ensembles de couches élémentaires définies pour chacun des d aspects ;
- $V_M \subseteq V \times L_1 \times \cdots \times L_d$ est l'ensemble des combinaisons nœud-couches, c'est-à-dire les couches dans lesquelles un nœud $v \in V$ est présent ;
- $E_M \subseteq V_M \times V_M$ est l'ensemble de liens contenant l'ensemble de paires de combinaisons possibles de nœuds et de couches élémentaires.

Un nœud peut exister dans plusieurs couches et il apparaît dans au moins une couche. Ainsi, un n-uplet nœud-couches $(i, \alpha_1, \cdots, \alpha_d)$ indique dans quelle couche appartient le nœud i. De plus, il est à la fois pratique et naturel de distinguer les liens qui relient des nœuds de différentes couches et les liens reliant les nœuds d'une même couche. Ces liens sont appelés respectivement des liens inter-couches et intra-couches (voir la figure 3.3). Le couplage entre les couches est représenté par les liens inter-couches. Il peut s'agir par exemple de liens signifiant que deux nœuds représentent la même entité dans différentes couches. La modélisation proposée dans [Kivelä et al., 2014] avec la notion d'aspect complexifie la modélisation.

G. Bianconi dans [Bianconi, 2018] propose une modélisation plus simple en considérant un réseau multi-couches comme un triplet $\mathcal{M} = (Y, \overrightarrow{G}, \mathcal{G})$ où

- Y est l'ensemble des couches, $Y = \{\alpha, \alpha \in \{1, 2, \ldots, M\}\}$ et M indique le nombre total de couches ($M = |Y|$) ;
- $\overrightarrow{G}$ une liste ordonnée de réseaux (graphes) caractérisant les interactions à l'intérieur de chaque couche $\alpha = 1, 2, \ldots, M$ soit $\overrightarrow{G} = (G_1, G_2, \ldots, G_\alpha, \ldots, G_M)$ pour lesquels $G_\alpha = (V_\alpha, E_\alpha)$;

7. http://www.hypergraphdb.org/

— $\mathcal{G}$ est une liste ordonnée de taille au plus $M \times M$ de graphes bi-parti caractérisant les interactions entre des paires de nœuds issues de couches différentes. Un élément $\mathcal{G}_{\alpha,\beta}$ de cette liste est défini par $\mathcal{G}_{\alpha,\beta} = (V_\alpha, V_\beta, E_{\alpha,\beta})$ pour $\alpha < \beta$.

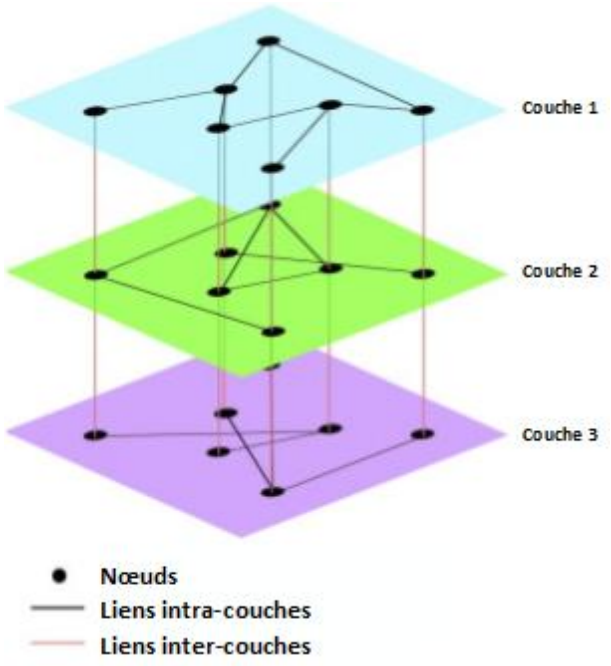


FIGURE 3.3 – Exemple d'un réseau multi-couches (extrait de [Tewarie et al., 2016])

Il existe plusieurs catégories de réseaux multi-couches, comme les réseaux multi-relationnels, hétérogènes, temporels, multiplexes. Dans la suite de cette section, nous . L'annexe C présente brièvement des modèles de graphes théoriques permettant d'approcher la modélisation des réseaux multi-couches.

Réseau multi-relationnel. Dans cette catégorie de réseaux, les liens sont étiquetés avec un certain type, chaque type représente une relation spécifique entre les nœuds dans le réseau et se traduit en une couche [Stroele et al., 2009]. L'ensemble des liens E est divisé en classes disjointes : $E = \bigcup_{r \in R} E_r$, où R est l'ensemble de relations possibles. En conséquence, les couches sont implicites et dans chaque couche l'ensemble des nœuds est homogène, les liens également. Dans la littérature, plusieurs travaux ont étudié les réseaux multi-relationnels. [De Domenico et al., 2013] ont transposé sur un réseau multi-relationnel les outils courants comme les mesures de centralité, le cœfficients de clustering et les algorithmes de marches aléatoires ; [Cai et al., 2005] et [Wu et al., 2013] ont étudié la détection de communautés.

Pour les réseaux sociaux tels que *Twitter* et afin de modéliser l'hétérogénéité des relations, [Dai et al., 2012, Kivelä et al., 2014] proposent une modélisation sous la forme de réseau multi-relationnel. Cependant, cette modélisation n'est pas complètement satisfaisante car elle fixe une contrainte forte : l'ensemble des nœuds est homogène. Or, dans *Twitter*, les relations peuvent exister entre différents types de nœuds comme l'illustre le tableau 3.1.

Réseau hétérogène. Ils sont aussi appelés réseau multi-types et possèdent des nœuds étiquetés avec un certain type qui peuvent être adjacents à des nœuds étiquetés avec le même type ou un type différent [Vazquez, 2006, Allard et al., 2009]. *Twitter* peut être vu comme un réseau hétérogène puisque les nœuds peuvent être de types différents comme des utilisateurs, des *tweets*, des hashtags, etc. Mais cette modélisation reste insuffisante puisque l'hétérogénéité des liens n'est pas prise en compte.

Dans la littérature, il y a eu des propositions de modélisation de réseaux complexes sous forme de réseau multi-relationnels hétérogène pour certaines tâches (détection de communautés, filtrage d'opinions, etc.). Ainsi, dans [Liu et al., 2014], les auteurs proposent une méthode de détection de communautés dans les réseaux multi-relationnels hétérogènes, ils utilisent le réseaux *Flickr* comme exemple de ce type de réseau. Or, dans la modélisation présentée, le type de relation entre les nœuds de même type est unique, de plus, une couche contient un seul type de nœud. Dans [Azaza et al., 2015], un réseau multi-relationnel hétérogène est utilisé pour représenter des opinions exprimées en ligne par des utilisateurs sur des services ou produits. De même que le travail de [Liu et al., 2014], dans le réseau présenté par [Azaza et al., 2015], le type de relation entre le même type de nœuds est unique.

Réseau temporel. Chaque couche correspond à un évènement, à un timestamp ou un intervalle de temps. Ainsi, les réseaux temporels permettent de représenter un ensemble d'évènements consécutifs comme une séquence ordonnée de réseaux [Holme et al., 2012]. Dans le cas d'un ensemble d'évènements, les évènements peuvent être représentés comme des triplets $e = (i, j, t)$ où $i, j \in V$ sont des nœuds et $t \in T$ est le timestamp d'un évènement. [Bianconi, 2018] utilise le formalisme défini pour les réseaux multi-couches pour spécifier les réseaux temporels, qualifiés de réseaux *multi-slices* ou multi-tranches. Un réseau multi-tranches est un réseau multi-couches dans lequel il existe une correspondance un à un entre les nœuds des différentes couches. Les tranches sont des instantanés du réseau temporel pris sur des fenêtres temporelles δt et ordonnées. Chaque tranche capture les interactions dans un intervalle de temps spécifié par $[(\alpha - 1)\delta t, \alpha\delta t[$. Le découpage peut aussi ne pas être régulier mais centré autour d'évènements successifs importants. Les liens inter-couches sont uniquement les liens qui établissent les correspondances entre nœuds ainsi un réseau temporel se réduit à un couple $\mathcal{M} = (Y, \overrightarrow{G})$ avec : $\overrightarrow{G} = (G_1, G_2, \ldots, G_\alpha, \ldots, G_M)$. Chaque graphe $G_\alpha = (V_\alpha, E_\alpha)$ modélise les interactions qui ont lieu entre les éléments de V dans la couche α c'est-à-dire $V_\alpha = V = \{i, i \in \{1, 2, \ldots, N\}\}$, pour une fenêtre temporelle $\alpha \in \{1, 2, \ldots, M\}$. Chaque graphe du réseau temporel peut être représenté par une matrice d'adjacence $A^{[\alpha]}$ de taille $N \times N$.

Twitter peut être modélisé comme un réseau temporel puisque chaque *tweet* possède un timestamp. Esteban Moro et son équipe ont proposé une modélisation de *Twitter* sous forme d'un réseau temporel afin d'étudier la diffusion de l'information [8]. Ils ont recueilli les *tweets* (750k) concernant la grève générale du 29 mars 2014 en Espagne, et ils ont construit un réseau temporel. Ils ont constaté une augmentation de l'activité lors de la soirée et la nuit du 29 mars, qui a conduit à une explosion de *retweets* au cours de cette journée.

8. http://sonic.northwestern.edu/temporal-network-of-information-diffusion-in-twitter-2/

Réseau multiplexe. Il s'agit d'une spécialisation de réseau multi-couches qui impose au moins un nœud en commun entre les couches et qui permet que les nœuds n'existent pas dans toutes les couches [Kanawati, 2015]. [Bianconi, 2018] propose une modélisation issue d'une simplification de la structure des réseaux multi-couches généraux respectant les propriétés suivantes :

— les réseaux multiplexes sont des réseaux multi-couches pour lesquels il existe une correspondance un à un entre les nœuds des différentes couches (nommés nœuds réplicas). Les correspondances peuvent être omises si tous les nœuds sont présents dans toutes les couches ;
— si il existe des liens entre les différentes couches (*interlinks*), ils expriment uniquement des correspondances entre nœuds réplicas.

Les réseaux multiplexes sont utilisés pour modéliser les réseaux sociaux en représentant différents types d'interactions pour le même ensemble d'utilisateurs. Ils peuvent aussi représenter des interactions entre différents types de nœuds en respectant la contrainte de correspondance 1-1 uniquement. La représentation des réseaux multiplexes sans liens inter-couches est similaire à celle des réseaux temporels : $\mathcal{M} = (Y, \overrightarrow{G})$. D'un point de vue algébrique, un réseau multiplexe pondéré et orienté peut être représenté par une suite de matrices d'adjacences $(A^{[\alpha]})_{\alpha \in \{1,2,\ldots,M\}}$.

Dans la plupart des réseaux multiplexes, les nœuds des différentes couches sont en correspondance 1-1 pour indiquer qu'il s'agit de la même entité. Dans certains cas, il peut être utile de distinguer l'identité des nœuds dans les différentes couches. On a alors recours à des liens inter-couches qui peuvent porter un poids spécifiant, comme dans les réseaux de transports, le coût associé à un changement de couche. Pour cela, on adopte une notation complémentaire pour les N nœuds et les M couches. Chaque nœud i peut avoir M réplicas représentés par des couples (i, α) avec $i = 1, 2, \ldots, N$ et $\alpha = 1, 2, \ldots, M$. Dans ce cas, $V = \{i, i \in \{1, 2, \ldots, N\}\}$ et pour α fixé, $V_\alpha = \{(i, \alpha), i \in \{1, 2, \ldots, N\}\}$. Un réseau multiplexe avec des liens inter-couches est alors modélisé par un triplet $\mathcal{M} = (Y, \overrightarrow{G}, \mathcal{G})$ avec $\overrightarrow{G} \neq \varnothing$ et $\mathcal{G} \neq \varnothing$. Chaque couche α est représentée par un graphe $G_\alpha = (V_\alpha, E_\alpha)$ qui lui-même peut être représenté par sa matrice d'adjacence, $A^{[\alpha]}$ de taille $N \times N$ incluant éventuellement des liens dirigés et/ou pondérés. Le réseau de couplage représentant les liens entre les couches $\mathcal{G}_{\alpha,\beta} = (V_\alpha, V_\beta, E_{\alpha,\beta})$ connecte les nœuds réplicas avec les liens définis par $E_{\alpha,\beta} = \{((i, \alpha), (i, \beta)), i \in \{1, 2, \ldots, N\}\}$ et représenté par une matrice d'adjacence $\mathscr{C}^{[\alpha,\beta]}$, nommée alors matrice de couplage de taille $N \times N$, où les éléments non diagonaux sont nuls.

La figure 3.4 montre un exemple de réseau multiplexe ayant 4 nœuds et 2 couches.

Afin d'exploiter un réseau multiplexe, celui-ci doit être dans un format que les algorithmes sont capables de traiter directement ou bien dans un format à partir duquel il est possible de dériver les structures de données nécessaires aux algorithmes. Dans le cadre général des réseaux multi-couches, la notion de matrice de supra-adjacence est souvent utilisée. Elle peut être obtenue à partir des différentes matrices d'adjacence intra-couche (composante $\overrightarrow{G}$) et des matrices de couplage inter-couches (composante $\mathcal{G}$).

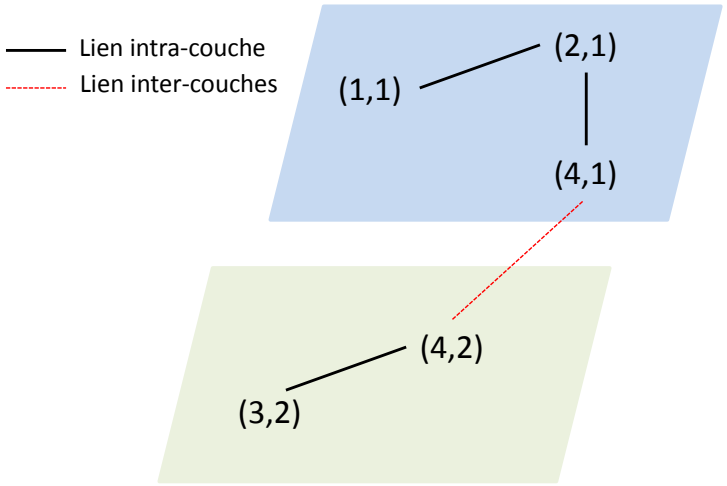


FIGURE 3.4 – Exemple de réseau multiplexe

La matrice de supra-adjacence de taille $(N \times M) \times (N \times M)$ $\mathscr{S}$, est définie de la manière suivante :

$$\mathscr{S}_{(i,\alpha),(j,\beta)} = \begin{cases} A_{ij}^{[\alpha]} & \text{si } \alpha = \beta \\ \mathscr{C}_{ij}^{[\alpha,\beta]} & \text{si } \alpha \neq \beta \end{cases}$$

avec $A^{[\alpha]}$ la matrice d'adjacence de la couche α et $\mathscr{C}^{[\alpha,\beta]}$ la matrice de couplage entre les couches α et β.

La forme de la matrice de supra-adjacence est la suivante :

$$\mathscr{S} = \left(\begin{array}{c|c|c|c} A^{[1]} & \mathscr{C}^{[1,2]} & \dots & \mathscr{C}^{[1,M]} \\ \hline \mathscr{C}^{[2,1]} & A^{[2]} & \dots & \mathscr{C}^{[2,M]} \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline \mathscr{C}^{[M,1]} & \dots & \dots & A^{[M]} \end{array}\right)$$

La figure 3.5 est la matrice de supra-adjacence du réseau multiplexe décrit en figure 3.4.

	$(1,1)$	$(2,1)$	$(3,1)$	$(4,1)$	$(1,2)$	$(2,2)$	$(3,2)$	$(4,2)$
$(1,1)$	0	1	0	0	0	0	0	0
$(2,1)$	1	0	0	1	0	0	0	0
$(3,1)$	0	0	0	0	0	0	0	0
$(4,1)$	0	1	0	0	0	0	0	1
$(1,2)$	0	0	0	0	0	0	0	0
$(2,2)$	0	0	0	0	0	0	0	0
$(3,2)$	0	0	0	0	0	0	0	1
$(4,2)$	0	0	0	1	0	0	1	0

FIGURE 3.5 – Matrice de supra-adjacence de l'exemple donné en figure 3.4

3.3.2/ MODÉLISATION DE *Twitter* SOUS FORME DE RÉSEAU MULTIPLEXE HÉTÉROGÈNE

Un modèle adapté aux données de *Twitter* est le modèle de réseaux multiplexes puisque les différentes couches peuvent être exploitées pour représenter les différents types de relations. Bruno Gonçalves modélise *Twitter* comme un réseau multiplexe [9], les nœuds représentent les utilisateurs et les couches représentent les relations *Retweet*, *Mention* et *Suivre* ainsi qu'une couche supplémentaire pour représenter la localisation géographique. Or, ces relations peuvent exister entre différents types de nœuds (voir le tableau 3.1). Comme le modèle de réseaux multiplexes ne contraint pas les nœuds à avoir le même type, ni à tous exister dans toutes les couches ce modèle est le plus pertinent. Comme nous l'avons montré, *Twitter* peut être modélisé sous plusieurs formes de réseaux multi-couches. Dans le cadre de notre étude, nous ne nous intéressons pas à l'aspect temporel, il est donc inutile pour nous de modéliser *Twitter* sous forme d'un réseau temporel. En revanche, nous souhaitons présenter les différentes entités du réseau telles que les *tweets* et les utilisateurs par différents types de nœuds, or, dans un réseau multi-relationnel, les nœuds sont homogènes. Par conséquent, nous proposons de modéliser *Twitter* sous forme d'un réseau multiplexe hétérogène.

Nous précisons le modèle de réseau multiplexe défini par [Bianconi, 2018] et nous proposons un modèle pour *Twitter* sous forme d'un réseau multiplexe hétérogène $\mathcal{M} = (Y, \overrightarrow{G}, \mathcal{G})$:

— Y est l'ensemble des couches correspondant aux différentes relations binaires du tableau 3.1 :

$$Y = \{\acute{E}crire, Suivre, Retweet_T, Retweet_U, Mention, Mention_U, \\ R\acute{e}ponse, R\acute{e}ponse_U, Contenir, Contenir_U\}$$

— $\overrightarrow{G}$ est la liste des graphes représentant les couches définies à partir de l'ensemble des nœuds $V = \{V_U \cup V_T \cup V_O\}$ pour lesquels V_U est l'ensemble des utilisateurs, V_T l'ensemble des tweets et V_O l'ensemble des objets c'est-à-dire un hashtag, une URL ou un symbole (emoji par exemple). Les graphes de $\overrightarrow{G}$ sont :
  — $G_{\acute{E}crire} = (V_U \cup V_T, E_{\acute{E}crire})$ est un graphe orienté avec $E_{\acute{E}crire} \subseteq V_U \times V_T$;
  — $G_{Suivre} = (V_U, E_{Suivre})$ est un graphe orienté avec $E_{Suivre} \subseteq V_U \times V_U$;
  — $G_{Retweet_T} = (V_T, E_{Retweet_T})$ est un graphe orienté qui représente les retweets avec $E_{Retweet_T} \subseteq V_T \times V_T$;

9. https://www.youtube.com/watch?v=8at3cMhaKUE

- $G_{Retweet_U} = (V_U, E_{Retweet_U})$ est un graphe orienté pondéré avec $E_{Retweet_U} \subseteq V_U \times V_U$, dans ce graphe les liens représentent le nombre de retweets d'un utilisateur donné effectués par un autre ;
- $G_{Mention} = (V_U \cup V_T, E_{Mention})$ est un graphe orienté qui représente les mentions des utilisateurs dans les *tweets*, $E_{Mention} \subseteq V_U \times V_T$;
- $G_{Mention_U} = (V_U, E_{Mention_U})$ est un graphe orienté pondéré avec $E_{Mention_U} \subseteq V_U \times V_U$, dans ce graphe, les liens représentent le nombre de mentions d'un utilisateur dans les *tweets* émis par un autre utilisateur ;
- $G_{Réponse} = (V_T, E_{Réponse})$ est un graphe orienté qui représente les *tweets* qui sont des réponses à d'autres *tweets*, $E_{Réponse} \subseteq V_T \times V_T$;
- $G_{Réponse_U} = (V_U, E_{Réponse_U})$ est un graphe orienté pondéré avec $E_{Réponse_U} \subseteq V_U \times V_U$, dans ce graphe, les liens représentent le nombre de réponses d'un utilisateur à un autre, tous *tweets* confondus ;
- $G_{Contenir} = (V_T \cup V_O, E_{Contenir})$ est un graphe orienté qui représente les objets contenus (hashtag, emoji, URL, images) dans les *tweets*, $E_{Contenir} \subseteq V_T \times V_O$;
- $G_{Contenir_U} = (V_U \cup V_O, E_{Contenir_U})$ est un graphe orienté pondéré avec $E_{Contenir_U} \subseteq V_U \times V_O$, dans ce graphe, les liens représentent le nombre de fois où un utilisateur a utilisé un objet, tous *tweets* confondus.

— $\mathcal{G}$ représente les liens inter-couches modélisés par les trois types de graphes :
- correspondances d'identités entre utilisateurs : $\{\mathcal{G}_{\alpha,\beta}\}$ avec α et $\beta \in \{Écrire, Suivre, Retweet_U, Mention, Mention_U, Réponse_U, Contenir_U\}$;
- correspondances d'identités entre tweets : $\{\mathcal{G}_{\alpha,\beta}\}$ avec α et $\beta \in \{Écrire, Retweet_T, Mention, Réponse, Contenir\}$;
- correspondances d'identités entre objets : $\{\mathcal{G}_{\alpha,\beta}\}$ avec α et $\beta \in \{Contenir, Contenir_U\}$.

Remarque :
On peut définir C comme l'ensemble des aspects regroupant les différentes couches. Pour notre modélisation de *Twitter*, on retient trois aspects : $C = \{C_1, C_2, C_3\}$ où les C_i représentent respectivement les interactions entre utilisateurs, les actions des utilisateurs et la structure des *tweets*, soit en utilisant les couches élémentaires :
- $C_1 = \{Suivre, Retweet_U, Mention_U, Réponse_U\}$;
- $C_2 = \{Écrire\}$;
- $C_3 = \{Retweet_T, Mention, Réponse, Contenir, Contenir_U\}$.

La figure 3.6 est la représentation en réseau multiplexe correspondante aux données modélisées par l'hypergraphe de la figure 3.2, elle illustre les différentes couches organisées selon trois aspects. Dans notre modélisation du réseau *Twitter*, l'aspect **Interactions entre utilisateurs** regroupe quatre couches qui ne contiennent qu'un seul type de nœud V_U. Les autres couches sont hétérogènes et peuvent contenir différents types de nœuds comme par exemple dans l'aspect **Actions des utilisateurs** qui regroupe les relations entre les nœuds utilisateurs V_U et *tweets* V_T dans cinq couches élémentaires. Par ailleurs, différents types de relations peuvent exister entre les mêmes types de nœuds, par exemple, les relations *Réponse* et $Retweet_T$ peuvent exister entre les nœuds de type *tweet* V_T.

Les liens inter-couches sont représentés par des traits pointillés, ils correspondent au couplage qui relient les mêmes nœuds à travers les différentes couches.

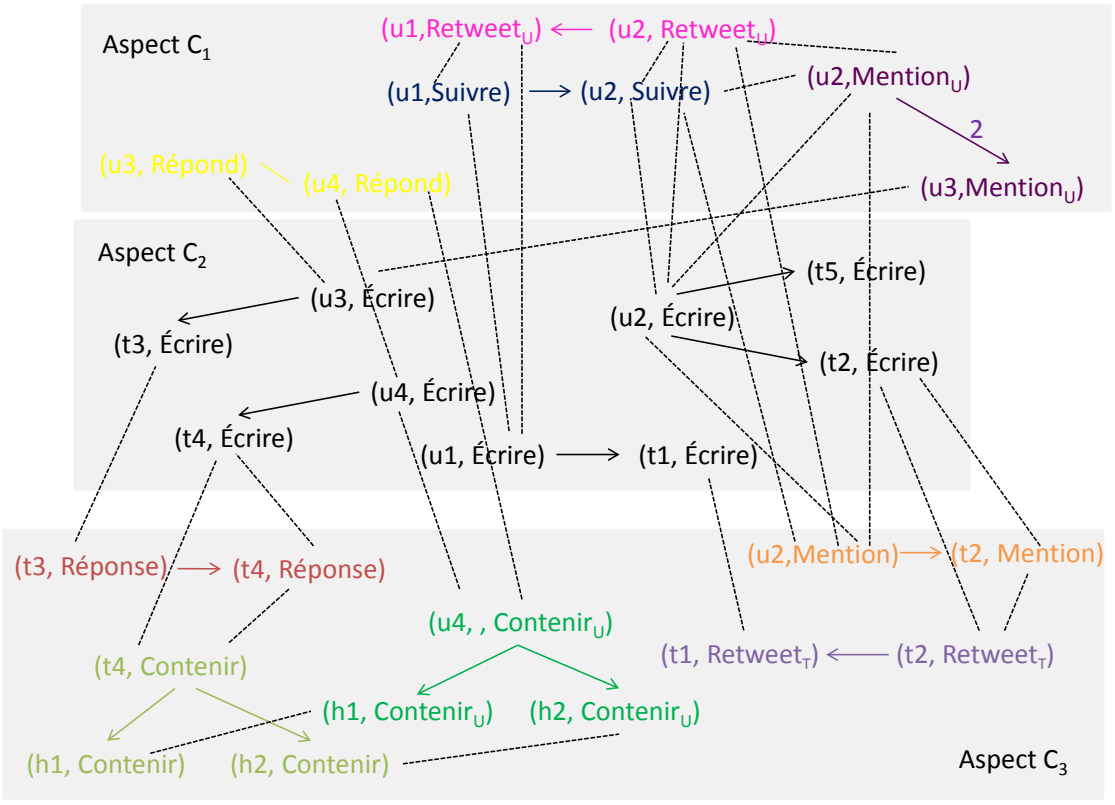


FIGURE 3.6 – Représentation sous la forme d'un réseau multiplexe des données de l'hypergraphe de la figure 3.2 incluant dix couches regroupées selon trois aspects

3.3.3/ EXPLOITATION DE LA MODÉLISATION SOUS FORME D'UN RÉSEAU MULTIPLEXE HÉTÉROGÈNE DE *Twitter via* L'UTILISATION D'ALGORITHME PAGERANK ÉTENDU À UN RÉSEAU MULTIPLEXE

Afin d'exploiter la richesse des réseaux multiplexes dans l'estimation de l'influence, des mesures populaires tels que les mesures de centralité et l'algorithme de PageRank ont été transposés sur les réseaux multiplexes. [Solé-Ribalta et al., 2014] proposent une mesure de centralité d'intermédiarité qui prend en compte la structure inhérente des réseaux multiplexes et proposent un algorithme pour la calculer de manière efficace. Pour montrer la nécessité et l'avantage de la mesure proposée, ils analysent les mesures de centralité obtenues pour deux réseaux multiplexes réels, un réseau multiplexe social à deux couches obtenu à partir de *Twitter* et *Instagram* et un autre réseau de co-auteurs à quatre couches obtenu à partir d'*arXiv*[10]. Les résultats montrent que la mesure proposée fournirait des résultats plus significatifs que l'approche d'évaluation de la centralité sur le réseau agrégé, en particulier pour les nœuds de rang moyen.

[Spatocco et al., 2018] ont présenté une méthodologie générale de calcul de mesures de centralité de manière itérative, basée sur le théorème de Perron-Frobenius[11], permettant les classements sur un ensemble de matrices d'un réseau multiplexe.

10. *arXiv* est une archive de prépublications électroniques d'articles scientifiques accessible gratuitement.
11. http://www.bibmath.net/dico/index.php?action=affiche&quoi=./p/perron-frobenius.html

Dans [Tu et al., 2018], les auteurs exploitent le concept de migration aléatoire des populations dans un réseau de transport multiplexe (train et avion). Ils proposent une nouvelle mesure de PageRank multiplexe, dans laquelle les effets d'influence et de rétroaction entre les couches sur la centralité des nœuds sont considérés. Ils appliquent la mesure proposée à un réseau artificiel. Leurs résultats indiquent que le fait de considérer le réseau avec des couches donne un classement des nœuds différent du classement dans un réseau simple.

PAGERANK MULTIPLEXE

Dans [Halu et al., 2013] et [Iacovacci et al., 2016], les auteurs proposent une modification de l'algorithme PageRank pour l'adapter aux réseaux multiplexes. Nous reprenons cet algorithme, et nous l'adaptons à notre formalisation du réseau multiplexe décrivant *Twitter*. Le principe de base consiste à choisir une couche principale α avec sa matrice d'adjacence $A^{[\alpha]}$ et de calculer le PageRank $x_i^{[\alpha]}$, $i = 1, \ldots, N_\alpha$ pour chaque nœud présent dans cette couche par la formule habituelle soit :

$$x_i^{[\alpha]} = d_\alpha \sum_{j=1}^{N_\alpha} A_{ji}^{[\alpha]} \frac{x_j^{[\alpha]}}{g_j} + (1 - d_\alpha) \frac{1}{N_\alpha}$$

où d_α est le facteur d'amortissement de la couche α, $g_j = max(1, \sum_{r=1}^{N_\alpha} A_{jr})$.

L'ensemble des valeurs constitué par chaque $x_i^{[\alpha]}$ est représenté par le vecteur $X^{[\alpha]}$ pour la couche α. Ensuite, l'algorithme PageRank multiplexe mesure une forme de centralité pour tous les nœuds (i,β) présents dans une autre couche β en utilisant la matrice d'adjacence $A^{[\beta]}$ et prenant en compte les mesures effectuées sur la couche α.

Les mesures sur la couche β sont influencées par la couche α et par le voisinage du nœud (i,β) et des nœuds réplicas dans la couche α. On obtient ainsi une formule générale du PageRank :

$$x_i^{[\beta]} = d_\beta \sum_{j=1}^{N_\beta} x_i^{[\alpha]} A_{ji}^{[\beta]} \frac{x_j^{[\beta]}}{G_j} + (1 - d_\beta) \frac{1}{N_\beta} \frac{x_i^{[\alpha]}}{\langle X^{[\alpha]} \rangle} \tag{3.1}$$

où $G_j = \sum_{r=1}^{N_\beta} A_{jr}^{[\beta]} x_r^{[\alpha]} + \delta(0, \sum_{r=1}^{N_\beta} A_{jr}^{[\beta]} x_r^{[\alpha]})$ et $\delta(a,b)$ est la fonction Kronecker, c'est une fonction à deux variables qui est égale à 1 si celles-ci sont égales, et 0 sinon.

Le nœud (i,β) est directement influencé par son réplica (i,α), cette influence est prise en compte en biaisant le premier terme de l'équation du PageRank au niveau du nœud lui-même et de ses voisins. La contribution de chaque voisin (j,β) de (i,β) est atténuée en divisant la centralité de (i,β) par la somme des centralités que les voisins de (j,β) ont dans la couche α.

Le second terme de l'équation (3.1) traduit la contribution de la centralité du nœud (i,α) au nœud (i,β). En effet, même si un nœud de β n'a pas la capacité à attirer des voisins importants, il peut quand même bénéficier de sa centralité acquise dans la couche α. Une

autre manière de formuler cette propriété est : l'importance d'un nœud dans une couche est affectée de manière bénéfique par l'importance qu'il possède dans une autre couche indépendamment de sa capacité à attirer d'autres nœuds. Pour ce faire on ajoute un facteur multiplicateur sur la partie de la formule liée au saut aléatoire qui est la valeur du PageRank de (i, α) dans la couche α divisée par la moyenne des valeurs des PageRank des nœuds de la couche α soit $\frac{x_i^{[\alpha]}}{\langle X^{[\alpha]} \rangle}$

Ainsi, les effets de l'interaction entre les différentes couches du réseau sur la centralité de type PageRank sont directement pris en compte dans la marche aléatoire. En particulier, en fonction de l'intensité de l'interaction entre les couches, [Iacovacci et al., 2016] définissent différentes versions du PageRank multiplexe : additif, multiplicatif et combiné.

$$x_i^{[\beta]} = d_\beta \sum_j^{N_\beta} (x_i^{[\alpha]})^b A_{ji}^{[\beta]} \frac{x_j^{[\beta]}}{G_j} + (1 - d_\beta) \frac{1}{N_\beta} \left(\frac{x_i^{[\alpha]}}{< X^{[\alpha]} >} \right)^a \tag{3.2}$$

où $G_j = \sum_{r=1}^{N_\beta} A_{jr}^{[\beta]} \left(x_r^{[\alpha]} \right)^b + \delta(0, \sum_{r=1}^{N_\beta} A_{jr}^{[\beta]} \left(x_r^{[\alpha]} \right)^b)$

Les valeurs de a et b permettent d'ajuster les deux parties de la formule. On peut identifier quatre cas limites dont trois correspondent aux PageRank multiplexe [Halu et al., 2013] puisque la cas $(a = 0, b = 0)$ correspond à un PageRank simple.

- PageRank multiplexe additif $(a = 1, b = 0)$, dans ce cas, $G_j = max(1, \sum_{r=1}^{N_\beta} A_{jr}^{[\beta]})$ ajoute une part de centralité aux nœuds de la couche β en fonction de la centralité dans la couche α. Ici, chaque nœud de la couche β tire un avantage supplémentaire du fait qu'il est central dans la couche α, quelle que soit la pertinence des nœuds qui le désignent dans la couche β.
- PageRank multiplexe multiplicatif $(a = 0, b = 1)$ l'effet de la couche α se produit sur le voisinage des nœuds de la couche β. Ici, tous les avantages d'être central dans la couche α dépendent des connexions qu'un nœud reçoit des nœuds centraux de la couche β.
- PageRank multiplexe combiné $(a = 1, b = 1)$ ajoute les deux effets précédents.

Les auteurs fournissent une fonction Matlab pour calculer le PageRank multiplexe [12]. Nous avons utilisé cette fonction avec les données du projet TEE'2014 et celles de l'élection présidentielle Française TEP'2017 afin d'évaluer le PageRank multiplexe des candidats. Pour ce faire, nous avons considéré les candidats et les utilisateurs ayant interagit avec eux selon les trois relations *Retweet*, *Mention* et *Réponse*, chaque relation forme une couche. Les couches sont représentées par leur matrice d'adjacence. La couche *Retweet* est la couche principale puisqu'elle est plus significative en terme d'influence, les autres calculs sont faits à partir de cette couche. Le tableau 3.2 résume les données utilisées pour chacun des jeux de données.

12. https://github.com/ginestrab/Multiplex-PageRank

TABLE 3.2 – Paramètres des données des deux corpus

Jeux de données	TEE'2014	TEP'2017 tour 1	TEP'2017 tour 2
Nombre de comptes	73 264	320 803	488 281
Nombre de candidats	616	11	2
Nombre de relations	614 191	3 088 066	2 355 394
Nombre de *retweets*	145 633	2 708 751	2 040 214
Nombre de *mentions*	434 207	29 652	18 243
Nombre de *réponses*	34 351	349 663	296 937

Ensuite, les trois couches sont données en entrée de l'algorithme du PageRank multiplexe, nous calculons les scores pour les cas limites correspondant aux PageRanks multiplicatif, additif et combiné. Le facteur de téléportation utilisé a été fixé à $0,85$. Pour chacun des deux corpus, les résultats sont synthétisés en figures 3.7 et 3.9.

À des fins de comparaison, nous avons aussi effectué le PageRank classique selon chacune des relations choisies (figures 3.8 et 3.10) et classer les candidats en fonction des scores PageRank obtenus. Les figures détaillant chacun des PageRanks multiplexes et chacun des PageRanks classiques ainsi que les classements sont donnés en annexe D.

Nous observons que sur les dix candidats classés pour le corpus TEE'2014 six candidats sont en communs dans les résultats renvoyés par les PagesRanks classiques et les PageRanks multiplexes multiplicatif et combiné, seul le PageRank multiplexe additif fait apparaître de nouveaux candidats. Cependant, le PageRank multiplexe indique clairement les candidats pouvant jouer un rôle important dans le réseau (nœuds indiqués par des flèches dans les figures). Nous retrouvons classé premier le candidat Marine Le Pen avec une valeur de PageRank multiplexe combiné de 0.1216 suivie par Jean-Luc Mélenchon qui a 0.0515 comme valeur de PageRank.

Pour le premier tour de la présidentielle française (corpus TEP'2017), nous retrouvons François Fillon (score $0,1761$) au sommet du classement du PageRank multiplexe multiplicatif suivi par Emmanuel Macron (score $0,1195$). Concernant le PageRank multiplexe additif, Jacques Cheminade est classé en premier avec un score de $0,9994$ contre des valeurs de PageRank multiplexe additif presque nulles pour les autres candidats. Enfin, le PageRank multiplexe combiné classe Emmanuel Macron en premier avec un score de $0,4696$ suivi de François Fillon avec un score de $0,2045$. Quand au corpus TEP'2017 du deuxième tour, Emmanuel Macron est classé premier selon le PageRank simple de la relation *retweet* et *réponse* ainsi que les PageRank multiplexe multiplicatif et additif.

Les résultats obtenus ont été appréciés par les spécialistes en science de la communication et politologues des projets TEE'2014 et TEP'2017. Les PageRanks multiplexe multiplicatif et combiné sont capables de distinguer correctement les personnes qui ont effectivement une influence importante dans le réseau *Twitter* et dans la vie réelle. Seuls les résultats du PageRank multiplexe additif sont très différents des autres résultats faisant apparaître des "petits" candidats. Ces résultats mettant en avant de tels candidats sont certainement les plus intéressants pour nos collègues de SHS car ils exposent des singularités. Cependant, bien que les PageRanks multiplexes permettent d'obtenir un "bon classement" d'utilisateurs, le score obtenu pour chaque utilisateur pris individuellement n'indique pas son influence et il est inexploitable si on ne le compare pas avec le score des autres utilisateurs.

En conclusion, les différentes versions de PageRank multiplexe reflètent l'intérêt de la combinaison des importances des nœuds dans les différentes couches. Les résultats montrent que la prise en compte de la nature multiplexe du réseau permet d'obtenir des classements de nœuds différents de ceux obtenus en considérant qu'une seule couche. Néanmoins, bien que le PageRank multiplexe permet d'obtenir un classement d'utilisateurs significatif, le score PageRank multiplexe obtenu à travers ses trois versions pour chaque utilisateur pris individuellement n'indique pas son degré d'influence et il est inexploitable si on ne le compare pas avec le score des autres utilisateurs.

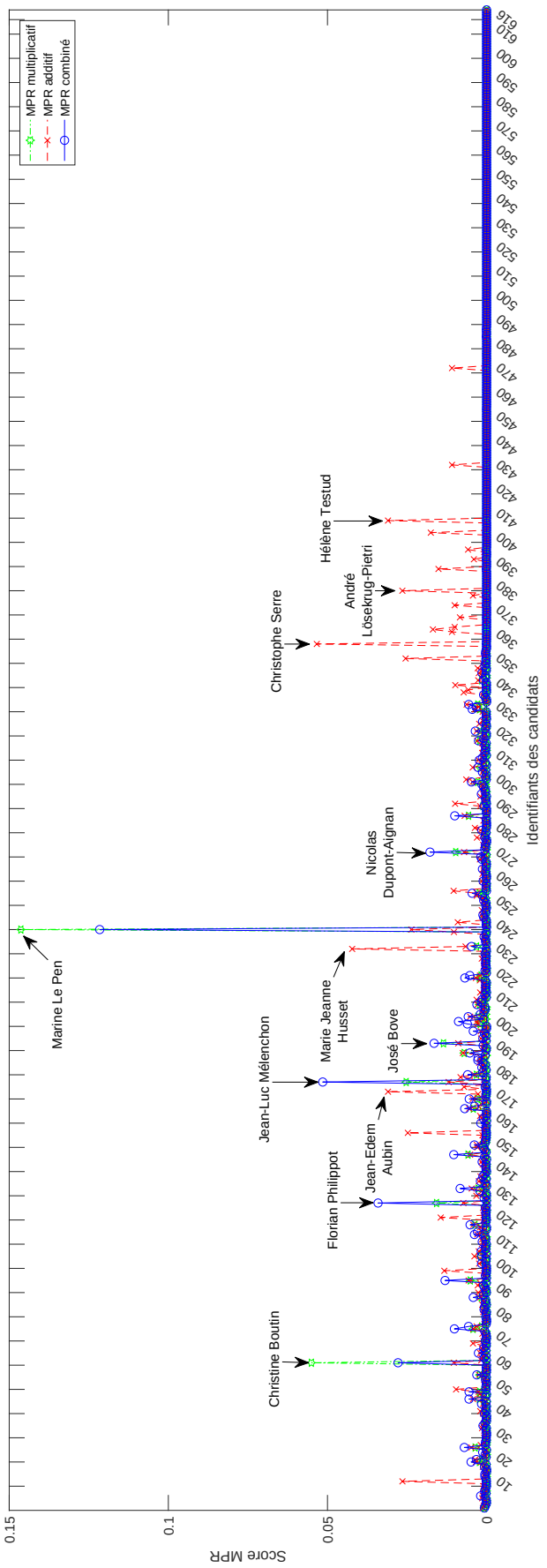


FIGURE 3.7 – PageRank multiplexe multiplicatif, additif et combiné des candidats français du corpus TEE'2014

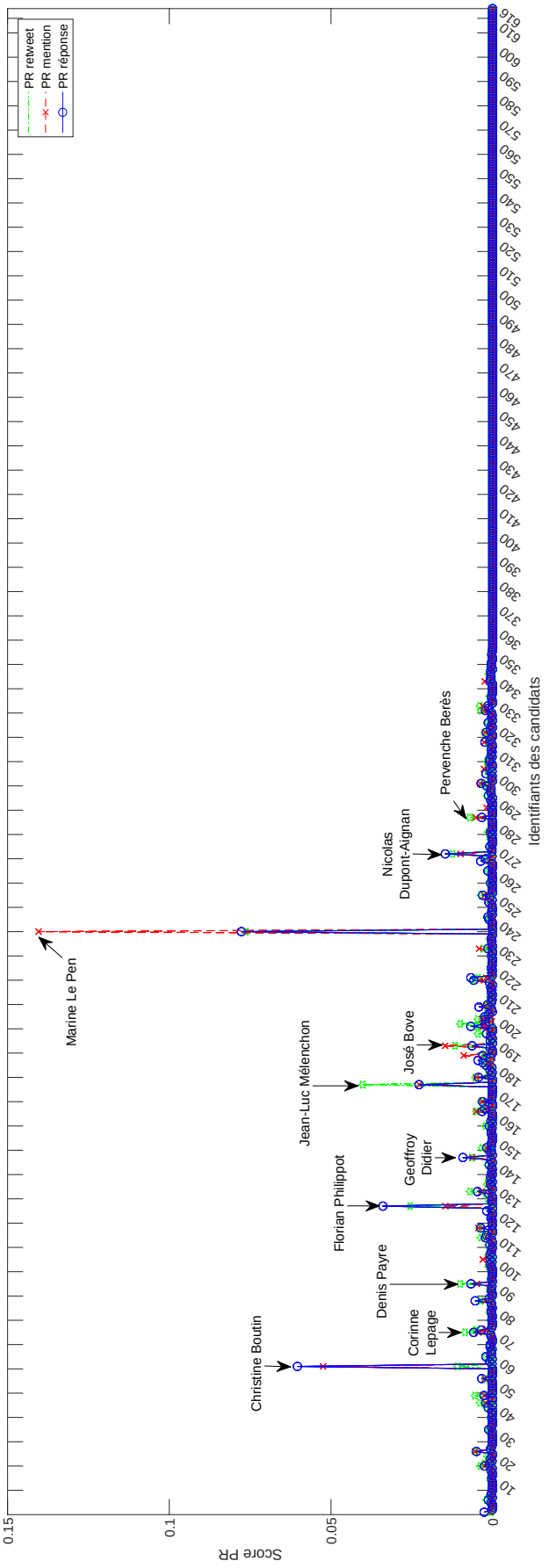


FIGURE 3.8 – PageRank des candidats français du corpus TEE'2014 selon les relations *retweet*, *mention* et *réponse*

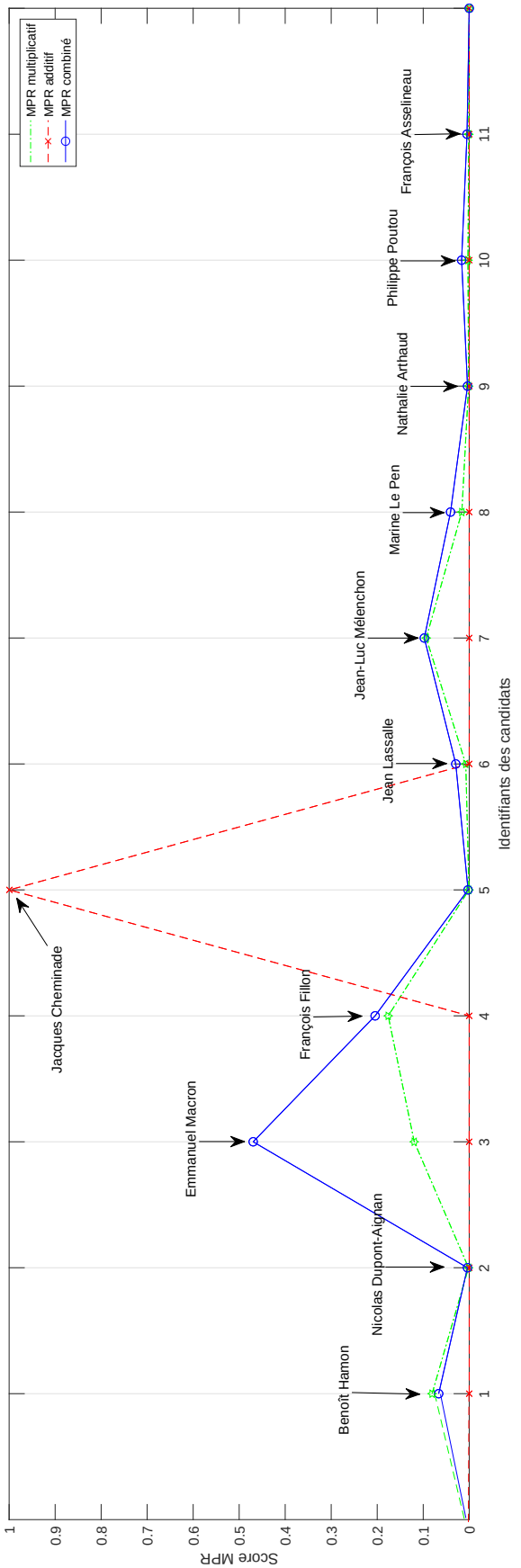


FIGURE 3.9 – PageRank multiplexe multiplicatif, additif et combiné des candidats du corpus TEP 2017

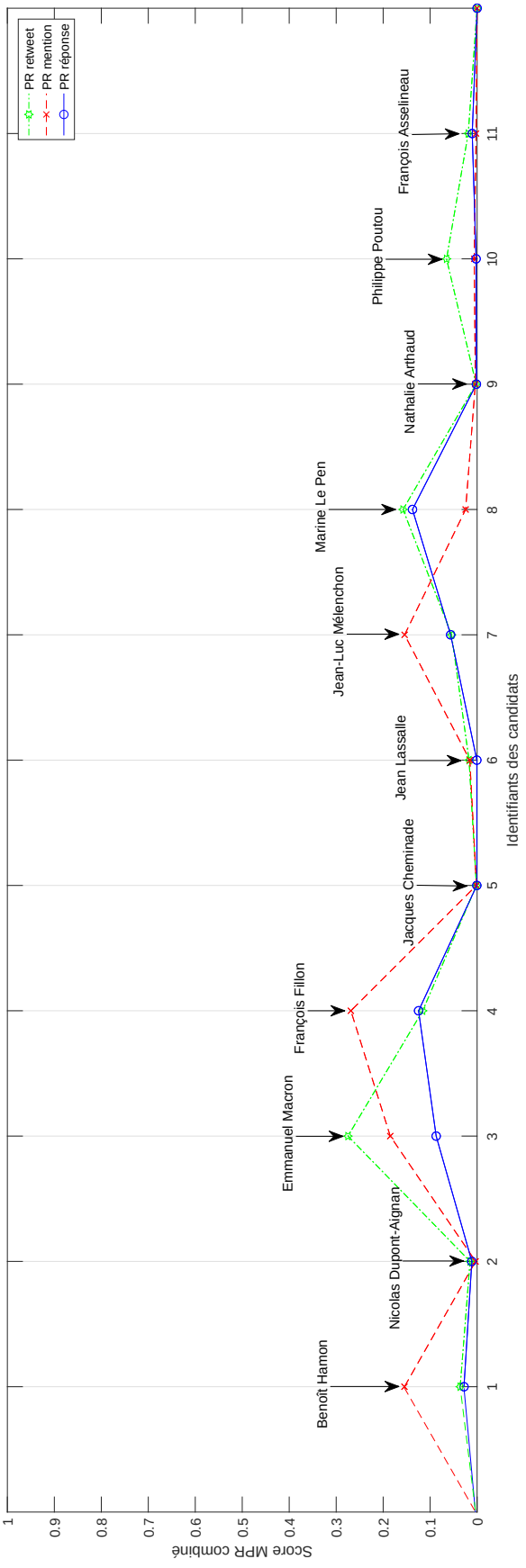


FIGURE 3.10 – PageRank des candidats du corpus TEP 2017 selon les relations *retweet*, *mention* et *réponse*

3.4/ CONCLUSION

Dans ce chapitre, nous avons d'abord rappelé les modèles théoriques de modélisation des réseaux complexes. Ensuite, nous avons présenté les différentes modélisation des réseaux sous forme de graphe.

Par la suite, nous avons proposé une nouvelle modélisation de *Twitter* sous forme d'un réseau multiplexe hétérogène. Nous avons formalisé et spécialisé ce modèle pour spécifier les relations engendrées par les interactions entre les utilisateurs de *Twitter* et issues de l'utilisation des différents opérateurs. Cette modélisation permet de visualiser, à travers les couches du réseau, les différentes relations présentes dans le réseau *Twitter*, les nœuds hétérogènes permettent à leur tour de présenter les différents types de nœuds tels que les utilisateurs et les *tweets*. Cette diversité de liens et de nœuds est primordiale afin d'étudier l'influence des utilisateurs de *Twitter* car les relations ne représentent pas la même importance dans l'estimation de l'influence. Ainsi, il est important de combiner les informations des différentes relations. Enfin, nous avons exploité ce réseau multiplexe hétérogène en utilisant une extension du PageRank pour les réseaux multiplexes. Nous avons appliqué l'algorithme du PageRank multiplexe aux données du projet TEE'2014 et à celles de l'élection présidentielle française TEP'2017. Nous avons réalisé différentes expériences en faisant varier les paramètres qui modifient le comportement de l'algorithme.

Si le classement des candidats obtenu reflète la réalité, les scores de PageRank multiplexe sont difficiles à interpréter car ils sont très proches les uns des autres. Dans le chapitre suivant nous voulons aller au-delà d'une mesure quantitative et nous proposons *TwitBelief*, une approche qui toujours en exploitant les différentes relations entre les nœuds du réseau détermine pour chaque nœud un degré d'influence pondéré par une estimation de la crédibilité.

4

ESTIMATION DE L'INFLUENCE DANS TWITTER : *TwitBelief*

Dans ce chapitre, nous présentons *TwitBelief*, une approche d'estimation de l'influence des utilisateurs dans *Twitter* en se basant sur les relations présentes dans le réseau multiplexe hétérogène de *Twitter*. La théorie des fonctions de croyance est utilisée afin de combiner les relations tout en exprimant l'incertitude sur leurs importances les unes par rapport aux autres.

4.1/ INTRODUCTION

L'étude de l'influence des utilisateurs dans *Twitter* est devenue un sujet de recherche important [Riquelme et al., 2016, Al-Garadi et al., 2018]. Cependant, pour estimer l'influence, il est important d'utiliser les différentes relations sur lesquelles nous pouvons baser les estimations de l'influence. Ces relations doivent être combinées afin d'établir une mesure générale d'influence. En raison de la diversité des relations et de la façon d'utiliser les opérateurs de *Twitter* (par exemple, l'utilisation de l'opérateur « @ » au début d'un *tweet* est différente de son utilisation au milieu ou à la fin, de plus, @nom-utilisateur permet d'interpeller une personne alors que @media est utilisé pour diffuser le *tweet* largement), il est difficile de quantifier l'importance des différentes relations et opérateurs les uns par rapport aux autres et c'est d'autant plus difficile lorsqu'ils sont combinés. Pour ces raisons, nous utilisons la théorie des fonctions de croyance [Shafer, 1976], un outil général pour le raisonnement avec incertitude. Cette théorie nous permet de combiner les informations des différentes relations tout en exprimant notre incertitude sur leurs importances. Dans la suite, nous présentons la théorie des fonctions de croyance puis notre approche *TwitBelief*.

4.2/ THÉORIE DES FONCTIONS DE CROYANCE

Dans le domaine des sciences appliquées, nous sommes très souvent amenés à raisonner à partir de cas ayant différentes alternatives entre lesquelles il n'est pas possible de trancher avec certitude. L'incertitude peut être due au hasard ou à la connaissance imparfaite (données en provenance de différentes sources). Le cadre classique connu pour raisonner sur l'incertitude est la théorie des probabilités. Formellement, un modèle probabiliste est un couple (Ω, p) tel que Ω représente l'ensemble de réponses possibles

(mutuellement disjointes) à une certaine question. p est la fonction de la probabilité $p : \Omega \rightarrow [0, 1]$ telle que :

$$\sum_{\omega \in \Omega} p(\omega) = 1 \tag{4.1}$$

La théorie des probabilités classique définit la probabilité d'une réponse comme le rapport entre le nombre de résultats élémentaires favorables et le nombre total de résultats élémentaires. Par exemple, pour le lancer de deux pièces de monnaie indiquant Pile (P) ou Face (F), nous pouvons prendre comme ensemble de réponses possibles $\Omega = \{PP, PF, FP, FF\}$ et alors $p(\omega) = 1/4 \; \forall \; \omega \in \Omega$. Mais, comme nous ne savons pas distinguer une pièce de l'autre, nous aurions pu choisir $\Omega = \{PP, PF, FF\}$, alors $p(PP) = p(FF) = 1/4$ et $p(PF) = 1/2$. Nous voyons que Ω n'est pas uniquement déterminé par le phénomène que nous étudions mais aussi par le choix que nous faisons du degré de finesse avec lequel le phénomène doit être décrit.

Considérons maintenant le cas d'une course hippique avec deux experts donnant leurs opinions, nous avons Ω qui représente trois chevaux C_1, C_2 et C_3 et nous souhaitons parier sur le cheval susceptible de gagner la course. Le premier expert pense que les trois chevaux sont de même niveau, la modélisation du problème dans un cadre probabiliste nous donne une équiprobabilité :

Expert 1 : $p_1(C_1) = p_1(C_2) = p_1(C_3) = 1/3$

Le deuxième expert affirme qu'il n'a aucune idée sur le cheval qui va gagner la course, en absence d'information (principe de raisonnement insuffisant), le choix le plus courant est la distribution uniforme sur l'ensemble des réponses :

Expert 2 : $p_2(C_1) = p_2(C_2) = p_2(C_3) = 1/3$

Les deux experts ont donné deux opinions différentes, or, la modélisation du problème dans un cadre probabiliste donne le même résultat pour les deux opinions : l'équiprobabilité et l'incertitude totale sont représentées de la même façon. D'où le besoin d'une théorie plus riche et plus flexible pour modéliser l'ignorance.

L'approche Bayésienne est également utilisée dans le cadre du raisonnement avec incertitude [Howson et al., 2006]. Elle est basée sur la théorie des probabilités et permet de combiner les informations en utilisant principalement le théorème de Bayes représentant et traitant des probabilités conditionnelles. Le théorème de Bayes est utilisé dans l'inférence statistique [1] pour mettre à jour ou actualiser les estimations d'une probabilité ou d'un paramètre quelconque, à partir des observations. Il énonce des probabilités conditionnelles : étant donné deux événements A et B, il permet de déterminer la probabilité de A sachant B, si l'on connaît les probabilités de A, de B et de B sachant A :

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{4.2}$$

Le terme $P(A)$ est la probabilité *a priori* de A, elle est « antérieure » au sens où elle précède toute information sur B. Le terme $P(A|B)$, appelé probabilité conditionnelle de A sachant B

1. L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population.

ou encore de A sous condition B, est « postérieure », au sens où 'elle dépend directement de B.

Malgré l'efficacité de l'approche Bayésienne dans le raisonnement incertain, elle présente quelques limites. En revanche, la théorie des fonctions de croyance est de plus en plus utilisée pour représenter l'incertitude. Ses avantages par rapport aux approches probabilistes et Bayésiennes sont nombreux [Mira, 2014]. Premièrement, l'approche Bayésienne fait intervenir la notion des probabilités *a priori*, ce qui est gênant lorsque nous ne pouvons pas l'évaluer, la nécessité de connaissances a priori est très contraignante et l'utilisation de l'approche est en conséquence parfois impossible [Dempster, 1968]. La théorie des fonctions de croyance nous permet d'éviter cette notion. Elle est plus flexible et permet aux sources de fournir des informations avec différents niveaux de détail. De plus, les approches probabiliste et Bayésienne sont parfois ambiguës car elles ne permettent pas de représenter la connaissance partielle, l'ignorance totale est remplacée par l'équiprobabilité ce qui peut engendrer une incohérence dans les résultats. Dans [Fiche et al., 2009], les auteurs montrent l'intérêt des fonctions de croyance face à une approche Bayésienne pour répondre à des problèmes de classification pour la caractérisation des fonds marins à partir d'images sonar. Les résultats de classification sont significativement meilleurs avec l'approche fondée sur les fonctions de croyance.

Ainsi, la théorie des fonctions de croyance est considérée comme un cadre général pour le raisonnement avec incertitude, et a été reliée à d'autres cadres tels que les théories des probabilités, des possibilités et des probabilités imprécises [Shafer, 1976]. La théorie des fonctions de croyance, aussi connue comme la théorie de l'évidence ou théorie de Dempster-Shafer, a d'abord été introduite par A. Dempster dans le contexte de l'inférence statistique, et a été développée plus tard par G. Shafer comme un outil général pour la modélisation de l'incertitude épistémique, c'est-à-dire une incertitude due à un manque de connaissance [Kotz et al., 1982].

Dans les paragraphes suivants, nous rappelons les concepts de base de la théorie des fonctions de croyance. Soient Ω un ensemble fini de réponses à une question et 2^{Ω} l'ensemble de tous les sous-ensembles de Ω, dans le contexte de la théorie des fonctions de croyance, Ω est appelé un cadre de discernement. Cette théorie est basée sur l'utilisation des fonctions de croyance qui représentent le degré avec lequel l'ensemble des informations disponibles accrédite l'hypothèse. La masse de croyance m est une fonction $m : 2^{\Omega} \rightarrow [0, 1]$ telle que :

$$\sum_{X \in 2^{\Omega}} m(X) = 1 \text{ et } m(\varnothing) = 0 \tag{4.3}$$

La masse $m(X)$ exprime l'état de connaissance sur le sous-ensemble X de Ω, c'est-à-dire la part de la croyance qui accrédite X, la masse $m(\Omega)$ représente le degré d'ignorance, $m(\varnothing) = 0$ car Ω est exhaustif, c'est-à-dire que nous connaissons toutes les réponses possibles. En conclusion, la masse de croyance est une opinion pondérée et à chaque alternative du monde est associé un nombre compris entre 0 et 1.

Par exemple, la modélisation du problème de la course hippique dans le cadre de la théorie des fonctions de croyance nous donne les masses suivantes correspondantes aux croyances des deux experts :

Expert 1 : $m_1(\{C_1\}) = m_1(\{C_2\}) = m_1(\{C_3\}) = 1/3$

Expert 2 : $m_2(\{C_1, C_2, C_3\}) = m_2(\Omega) = 1$

La théorie des fonctions de croyance permet, non seulement la représentation de la connaissance partielle et de l'incertitude, mais aussi la fusion d'information [Nimier et al., 1995]. Considérons différentes sources d'information qui s'expriment sur le même cadre de discernement, nous cherchons à combiner ces informations à travers une seule masse de croyance. La fusion d'information est réalisée par la règle de combinaison conjonctive [Smets, 1997], elle suppose que toutes les sources sont fiables et consistantes (non contradictoires). Considérant deux fonctions de masse m_1 et m_2, la règle de combinaison conjonctive est définie par :

$$(m_1 \otimes m_2)(C) = \sum_{A \cap B = C} m_1(A) m_2(B), \quad A, B, C \in 2^{\Omega} \tag{4.4}$$

Afin d'illustrer le fonctionnement de la règle de combinaison conjonctive, nous reprenons l'exemple de la course hippique avec de nouvelles connaissances en provenance des deux experts. L'expert 1 pense que les chevaux C_1 et C_2 ont une chance de 70% de gagner la course car ils sont plus jeunes que C_3. L'expert 2 pense que les chevaux C_1 et C_3 ont une chance de 60% de gagner la course car ils sont de sexe mâle. Nous considérons alors les fonctions de masse suivantes associées à la croyance des deux experts :

$$\text{Expert 1} \mapsto \begin{cases} m_1(\{C_1, C_2\}) = 0.7 \\ m_1(\Omega) = 0.3 \end{cases} \qquad \text{Expert 2} \mapsto \begin{cases} m_2(\{C_1, C_3\}) = 0.6 \\ m_2(\Omega) = 0.4 \end{cases}$$

Pour l'expert 1, la masse restante de 0.3 n'est pas mise sur $\{C_3\}$ car l'expert 1 ne dit rien sur le cheval C_3, le même raisonnement est fait pour l'expert 2. Les masses de croyance $m_1(\Omega)$ et $m_2(\Omega)$ représentent l'ignorance partielle. Nous utilisons alors la règle de combinaison conjonctive (équation 4.4) pour déterminer quel cheval va gagner la course. La fonction de masse combinée des deux experts est donnée dans le tableau 4.1 :

TABLE 4.1 – Combinaison des connaissances des deux experts

$\otimes$	$\{C_1, C_2\}$ 0.7	$\boldsymbol{\Omega}$ 0.3
$\{C_1, C_3\}$ 0.6	$\{C_1\}$ 0.42	$\{C_1, C_3\}$ 0.18
$\boldsymbol{\Omega}$ 0.4	$\{C_1, C_2\}$ 0.28	Ω 0.12

Nous obtenons : $m(\{C_1\}) = 0.42$; $m(\{C_1, C_2\}) = 0.28$; $m(\{C_1, C_3\}) = 0.18$; $m(\Omega) = 0.12$

Afin de prendre une décision, il est nécessaire de sélectionner l'hypothèse la plus probable, ce qui peut être difficile à réaliser directement avec la théorie des fonctions de croyance où les fonctions de masse sont données, non seulement pour les singletons, mais aussi pour les sous-ensembles du cadre de discernement. Ils existent plusieurs solutions pour assurer la prise de décision au sein de la théorie des fonctions de croyance, la plus connue est la probabilité pignistique [Smets, 1989]. Contrairement aux fonctions de masse qui sont définies sur 2^{Ω}, la probabilité pignistique est une mesure de probabilité définie sur

Ω. La probabilité pignistique a été proposée dans le modèle des croyances transférables [Smets, 1989]. Ce modèle est basé sur deux niveaux : le « niveau crédal » où les croyances sont représentées par des fonctions de masse et le « niveau pignistique » où les croyances sont utilisées pour prendre la décision et représentées comme des fonctions de probabilité, appelées probabilités pignistiques et notées *bet*, définies par :

$$\text{bet}(x) = \sum_{X \subseteq 2^{\Omega} | x \in X} \frac{m(X)}{|X|} + \frac{1}{1 - m(\varnothing)} \tag{4.5}$$

La probabilité pignistique consiste à répartir équitablement chaque masse de croyance entre les singletons x composant X, $|X|$ est la cardinalité de l'ensemble X. Revenons à l'exemple de la course hippique, finalement, pour prendre la décision sur le cheval qui va gagner la course, nous calculons la probabilité pignistique en utilisant l'équation 4.5. Par exemple,

$$\text{bet}(C_1) = m(\{C_1\}) + \frac{m(\{C_1, C_2\})}{|\{C_1, C_2\}|} + \frac{m(\{C_1, C_3\})}{|\{C_1, C_3\}|} + \frac{m(\Omega)}{|\Omega|}$$

$$\text{bet}(C_1) = 0.42 + \frac{0.28}{2} + \frac{0.18}{2} + \frac{0.12}{3} = 0.69$$

De la même manière, nous calculons la probabilité pignistique relative aux chevaux C_2 et C_3 et nous obtenons les résultats suivants : $\text{bet}(C_2) = 0.18$ et $\text{bet}(C_3) = 0.13$

On peut conclure avec une certitude de 69% que le cheval qui va gagner la course est C_1 puisqu'il a la plus grande probabilité pignistique soit 0.69.

La figure 4.1 résume le processus de la synthèse de connaissance en utilisant la théorie des fonctions de croyance. La première étape est la modélisation des informations qui dépend du domaine étudié, il s'agit de choisir le cadre de discernement et d'initialiser les masses de croyance des différentes sources d'information. Ensuite, dans le niveau crédal, la combinaison conjonctive est utilisée afin de combiner les masses de croyance des différentes sources d'information. Enfin, dans le niveau pignistique pour prendre la décision, la masse de croyance combinée est transformée en probabilité pignistique.

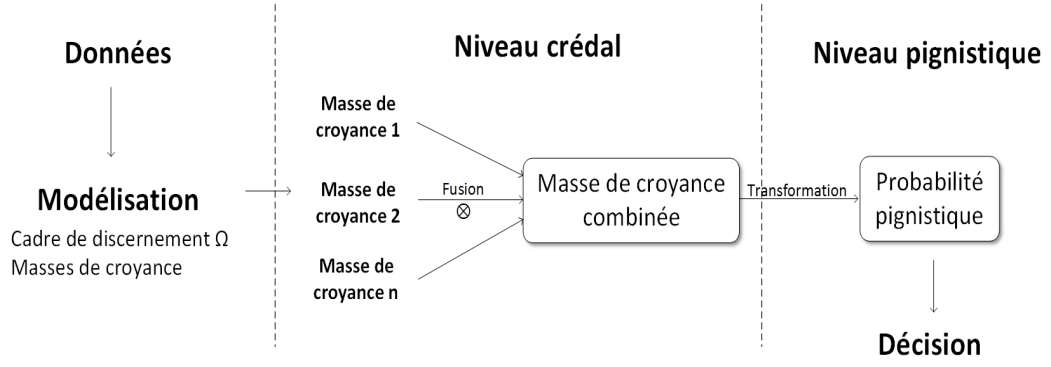


FIGURE 4.1 – Processus de synthèse de connaissance avec la théorie des fonctions de croyance

La théorie des fonctions de croyance a été largement utilisée dans de nombreux domaines. De nombreux travaux ont utilisés la théorie des fonctions de croyance dans le domaine du traitement d'images comme la segmentation d'images [Vannoorenberghe et al., 2003], la reconnaissance de forme et son application au diagnostic des circuits de voies ferroviaires [Debiolles, 2007] et la segmentation spatiale des tumeurs pulmonaires dans des images 3D [Chunfeng et al., 2017]. La théorie des fonctions de croyance a aussi été appliquée dans le domaine des réseaux routiers, par exemple, la gestion des informations imparfaites dans les réseaux de véhicules [Mira, 2014], la fusion d'informations pour la compréhension de scènes routières [Xu et al., 2014] et de scènes urbaines [Xu et al., 2016] pour des systèmes d'aide à la conduite. Nous retrouvons aussi plusieurs applications de la théorie des fonctions de croyance dans le domaine de la classification, par exemple, la catégorisation des messages dans les réseaux sociaux [Ben Dhaou et al., 2014], la classification d'états physiologiques dans un bioprocédé fermentaire [Régis et al., 2007] et la reconnaissance d'adresse postale [Mercier et al., 2009]. La théorie des fonctions de croyances a été aussi employée dans le domaine de la prédiction tel que la prédiction du résultat de traitement du cancer [Lian et al., 2016] et la prédiction des risques naturels [Demotier et al., 2006, Tacnet et al., 2010]. Enfin, cette théorie a été exploitée pour localiser des dommages sur une structure d'aéronef, notamment sur une aile d'avion [Worden et al., 2009].

4.3/ *TwitBelief* : ESTIMATION DE L'INFLUENCE DES UTILISATEURS DE *Twitter* EN UTILISANT LA THÉORIE DES FONCTIONS DE CROYANCE

L'une des caractéristiques de *Twitter* est la diffusion de l'information par l'utilisation d'opérateurs. Les relations entre les utilisateurs déterminent le flux de l'information et conditionnent ainsi l'influence d'un utilisateur sur un autre. Leavitt et al. [Leavitt et al., 2009] ont défini l'influence sur *Twitter* comme la capacité d'un utilisateur à provoquer une action chez un autre utilisateur. Le terme « action » désigne les différentes interactions entre les utilisateurs au moyen des opérateurs *retweet*, *mention*, *réponse*, *suivre*.

Les *retweets* permettent d'atteindre les abonnés de l'utilisateur qui retweete, qui à leur tour peuvent les retweeter. les *mentions* permettent en revanche d'atteindre n'importe quel utilisateur directement en utilisant le préfixe « @ » suivi par le nom de l'utilisateur qu'on souhaite mentionner, ainsi, les *mentions* peuvent rendre l'information contenue dans les *tweets* plus visible en ciblant les utilisateurs les plus appropriés. Une *réponse* à un *tweet* permet de créer une conversation avec l'utilisateur du *tweet* initial à laquelle ses autres abonnés peuvent prendre part. Enfin, la relation *suivre* permet de s'abonner aux autres utilisateurs et de pouvoir ainsi voir leurs *tweets*.

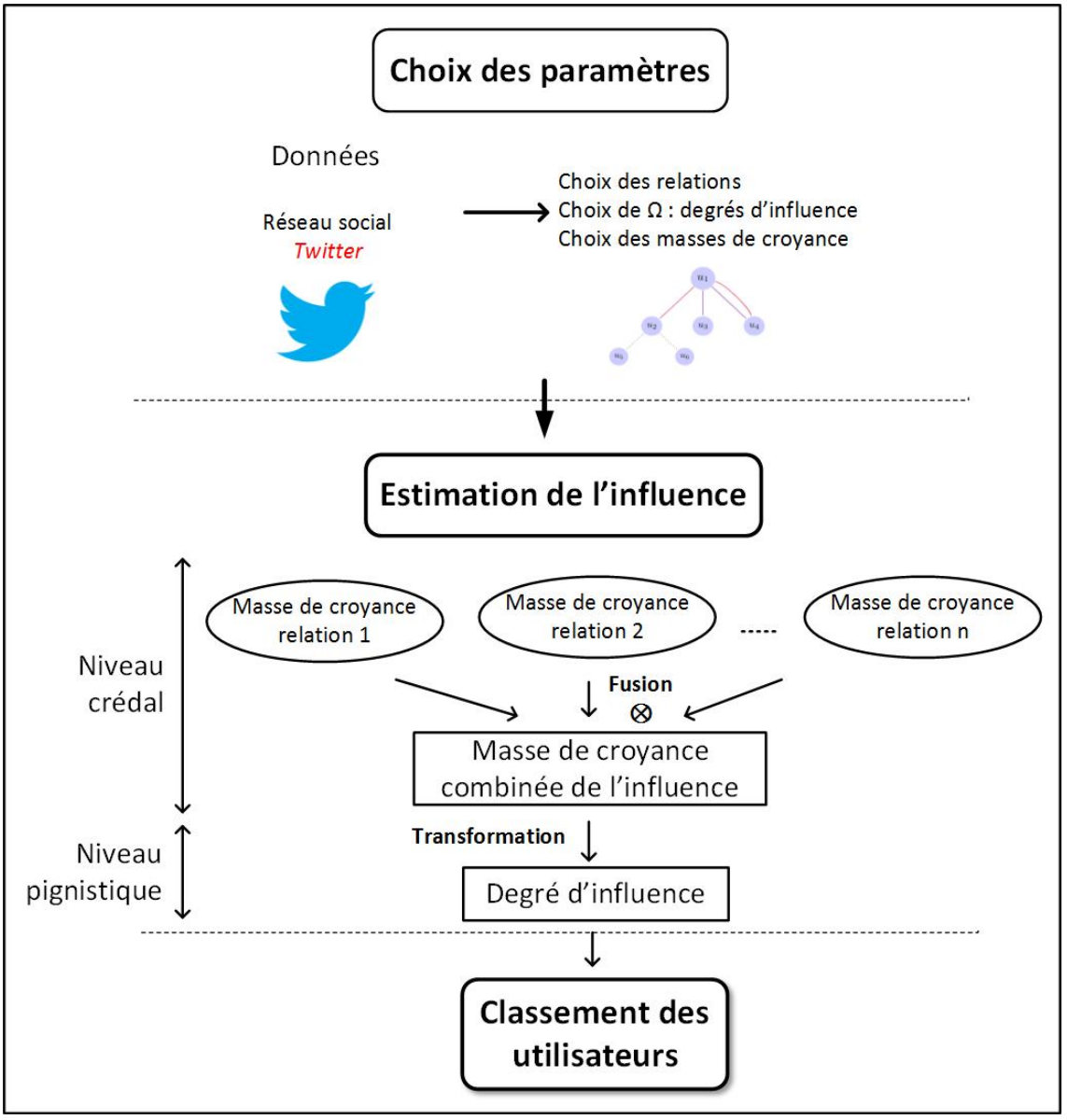


FIGURE 4.2 – Étapes de l'approche proposée, *TwitBelief*

Par conséquent, l'estimation de l'influence sur *Twitter* est un problème complexe puisque *Twitter* offre plusieurs opérateurs qui peuvent être combinés pour former différentes relations entre les utilisateurs, de plus chaque utilisateur en fonction du domaine auquel il appartient à sa façon d'écrire un *tweet*. Afin d'estimer l'influence d'un utilisateur dans le réseau social *Twitter*, nous utilisons la théorie des fonctions de croyance pour effectuer la fusion des informations issues de l'utilisation des différentes relations. La figure 4.2 donne un aperçu des étapes principales de l'approche proposée, *TwitBelief*. Tout d'abord, les experts du domaine (sociologues, politologues, chargés de communication, consultants marketing, etc.) exploitent les données de *Twitter* : sélection des relations pertinentes, choix du cadre de discernement et initialisation des masses de croyance. Ensuite, dans le niveau crédal, nous combinons les différentes masses de croyance associées à chaque relation considérée pour obtenir la masse de croyance combinée pour chaque utilisateur. Dans le niveau pignistique, la masse de croyance combinée est transformée en probabilité

pignistique pour estimer le degré d'influence de l'utilisateur. Enfin, nous calculons le classement de tous les utilisateurs selon leur influence. Dans les paragraphes suivants, nous détaillons chaque étape de l'approche.

4.3.1/ CHOIX DES PARAMÈTRES

Nous définissons un *motif d'interaction (motif)* p comme une séquence de relations, par exemple, un *retweet* contenant une *mention* ou le *retweet* d'une *réponse*. Soit P l'ensemble des *motifs d'interaction* possibles, notons par $R = R \cup P$ l'ensemble des relations y compris les *motifs d'interaction*. Par exemple, dans *Twitter*, nous pouvons considérer :
R = {retweet, mention, réponse, suivre, retweet + réponse, retweet + mention, mention + mention, réponse + mention }.

Dans ce chapitre, nous nous intéressons à l'étude de l'influence entre les utilisateurs indépendemment du contenu des *tweets*. Ainsi, les relations *retweet, mention* et *réponse* que nous utilisons ici désignent les relations $retweet_U$, $mention_U$ et $réponse_U$ définies dans le chapitre 3.

Les figures 4.3 et 4.4 représentent des exemples de relations observées lors des élections européennes 2014. La figure 4.3 représente un exemple de la relation *mention* entre le compte « BourdinDirect » et le compte « F_Philippot » qui signifie que le *tweet* émis par « BourdinDirec » mentionne « F_Philippot », ce *tweet* est vu par les abonnés de « BourdinDirect ». La figure 4.4 représente le *motif d'interaction réponse + mention entre le compte « Philippe Platon » et le compte « F_Philippot » correspondant au tweet de « Philippe Platon » qui est une réponse au tweet de « BourdinDirect » qui mentionnait « F_Philippot », ce qui représente le motif d'interaction réponse d'une mention entre les comptes « Philippe Platon » et « F_Philippot » effectué à travers le compte « BourdinDirect ».*

FIGURE 4.3 – Exemple de relation *mention*

FIGURE 4.4 – Exemple de *motif d'interaction réponse + mention*

Nous parlons d'influence indirecte lorsqu'elle atteint un utilisateur à travers des utilisateurs intermédiaires comme dans la figure 4.4. Ainsi, les relations sont les critères de manifesta-

tion de l'influence directe d'un utilisateur tandis que les *motifs d'interaction* représentent l'influence indirecte.

Le choix des relations (y compris les *motifs d'interaction*) sur lesquelles nous basons les estimations de l'influence dépend de plusieurs facteurs. Généralement, il dépend du domaine étudié, par exemple, dans le domaine du marketing, les experts du domaine des sciences sociales affirment qu'il est nécessaire de considérer les relations *mention* et *réponse* afin d'estimer l'influence des comptes *Twitter* de certaines marques [Jansen et al., 2009, Vidya et al., 2015]. Dans les études politiques, une *mention* ou une *réponse* peuvent être moins intéressantes qu'un *retweet* [Stieglitz et al., 2012, Mustafaraj et al., 2011, Wong et al., 2016], de plus, une *réponse* suivie par un *retweet* est un motif d'interaction très important. Mais le choix des relations est aussi contraint par les APIs de *Twitter* à travers lesquelles nous collectons les données, comme discuté dans le chapitre 3, certains *motifs d'interaction* ne peuvent pas être obtenus, d'autres relations ne sont pas accessibles à cause des droits d'accès quand le compte n'est pas public par exemple.

L'influence d'un utilisateur est déterminée par l'importance des relations R qui lui sont associées. Chaque relation est associée à un **degré d'influence** d_r pour $r \in R$, par exemple, la relation *retweet* est associée au degré d'influence $d_{retweet} =$ Très Faible. Soit Ω l'ensemble des différentes réponses possibles à notre question : *quel est le degré d'influence d'un certain utilisateur ?*, Ω représente alors l'ensemble ordonné de tous les degrés d'influence possibles :

$$\Omega = \{\text{Très Faible, Faible, Assez Moyenne, Moyenne, Assez forte, Forte, Très Forte, Extrêmement Forte}\} \tag{4.6}$$

Dans la théorie générale des fonctions de croyance, 2^{Ω} est utilisé comme domaine des fonctions de masse, dans notre approche, nous utilisons seulement un sous-ensemble Ω_{Inf} de 2^{Ω} car nous voulons traduire la certitude de l'expert du domaine étudié par une fonction de masse m_r sur une relation, précisément :

$$\Omega_{Inf} = \{\text{Très Faible, Faible, Assez Moyenne, Moyenne, Assez Forte, Forte, Très Forte, Extrêmement Forte}, \Omega\} \tag{4.7}$$

Les fonctions de masse expriment un lien entre les différentes relations qui jouent sur l'influence d'un utilisateur, elles représentent l'importance des relations. Une fonction de masse est associée pour chaque relation, les fonctions de masse sont définies comme suit : $m_r : \Omega_{Inf} \rightarrow [0, 1]$. Alors, pour chaque relation $r \in R$, en plus du degré d'influence d_r, une fonction de masse m_r est associée, d_r et m_r dépendent de la relation et du domaine étudié.

Dans ce contexte, nous introduisons le graphe d'influence (voir la figure 4.5) comme un graphe multiplexe étiqueté $G = (U, E)$, où U est l'ensemble de nœuds représentés par les utilisateurs, et E est l'ensemble de liens qui modélisent les différentes relations $r \in R$ entre les nœuds. Les liens sont étiquetés par leur type de relation r (dans la figure 4.5, chaque type de relation est représenté par une couleur), leur degré d'influence d_r et leur masse de croyance m_r. Certaines recherches récentes ont introduit des graphes incertains dont les liens sont étiquetés par leurs probabilités d'existence

[Khan et al., 2014, Parchas et al., 2014]. Dans notre cas, l'incertitude ne concerne pas la présence ou l'absence de liens mais concerne notre croyance dans l'importance du lien. Les nœuds sont étiquetés par le nom du compte (par exemple, u_1, u_2).

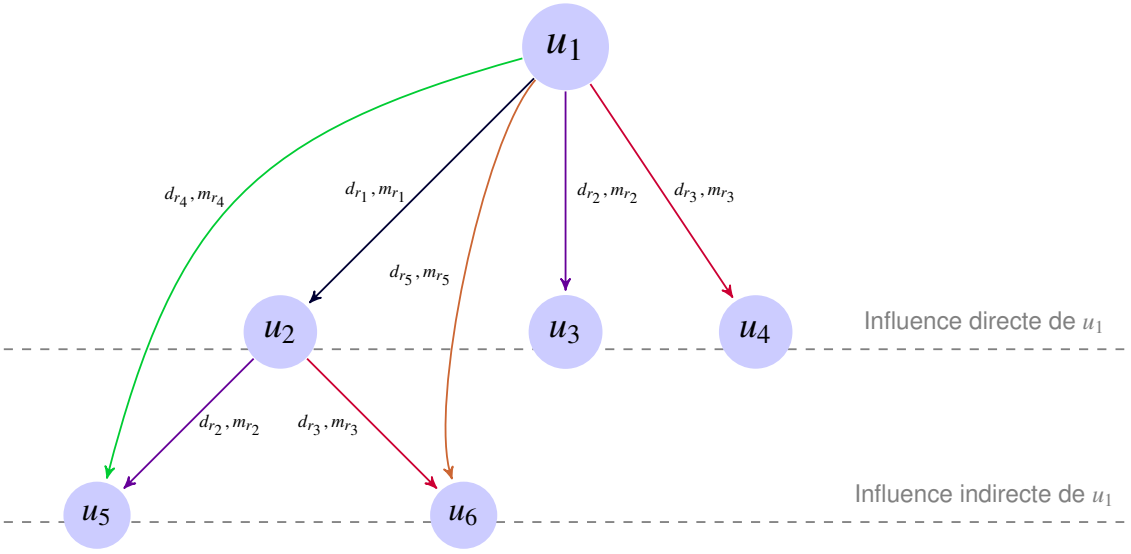


FIGURE 4.5 – Graphe d'influence centré sur l'utilisateur u_1

4.3.2/ ESTIMATION DU DEGRÉ D'INFLUENCE D'UN UTILISATEUR

En se basant sur la théorie des fonctions de croyance présentée dans la section 4.2, nous fusionnons les différentes fonctions de masse définies dans le graphe d'influence. Afin d'estimer le degré d'influence d'un utilisateur spécifique u, nous prenons en compte la structure locale du graphe autour du nœud représentant l'utilisateur u et nous combinons les fonctions de masses de croyance des liens incidents en utilisant une version modifiée de la règle de combinaison conjonctive (équation 4.4) :

$$(m_{r_1} \otimes m_{r_2})(z) = \sum_{x \textcircled{@} y = z} m_{r_1}(x) m_{r_2}(y), \quad x, y, z \in \Omega_{Inf} \tag{4.8}$$

$\textcircled{@}$ est une opération symétrique, $\textcircled{@} : \Omega_{Inf} \times \Omega_{Inf} \rightarrow \Omega_{Inf}$, le tableau 4.2 présente un exemple de l'opération $\textcircled{@}$. Cette fonction assure notre hypothèse : plus nous combinons des relations relatives à un utilisateur, plus son influence devient importante.

TABLE 4.2 – Exemple de l'opération $@$

$@$	**T.Faible**	**Faible**	**A.Moyenne**	**Moyenne**	**A.Forte**	**Forte**	**T.Forte**	**E.Forte**	Ω
T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	E.Forte	T.Faible
Faible	A.Moyenne	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	E.Forte	Faible
A.Moyenne	Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	T.Forte	E.Forte	A.Moyenne
Moyenne	A.Forte	A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte	E.Forte	Moyenne
A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	A.Forte
Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	E.Forte	E.Forte	Forte
T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	E.Forte	E.Forte	T.Forte
E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte
Ω	T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	E.Forte	Ω

Comme plusieurs relations peuvent exister plusieurs fois entre un utilisateur u et ses voisins, l'ordre des combinaisons peut affecter nos résultats, nous devons alors choisir un ordre pour être consistant (voir l'annexe A). Pour une relation de type r_i, la masse initialisée est représentée comme suit : $\{m_{r,i} : r \in R, i \in |R|\}$. Nous désignons par ℓ_{u,r_i} l'ensemble de relations de type r_i pour l'utilisateur u. Nous combinons les fonctions de masse afin d'obtenir une masse de croyance globale correspondante au degré d'influence de l'utilisateur u (voir la propriété 1 dans l'annexe A). Afin de simplifier les expressions, nous écrivons $\bigotimes_{i\in\{1,2,3\}}$ au lieu de $m_1 \otimes m_2 \otimes m_3$. Ainsi, nous considérons l'ordre des combinaisons suivant :

1. Pour un type de relation donné r_i, nous combinons les masses des relations de type r_i afin d'obtenir un r_i-pré-résultat avec M_{u,r_i} défini comme suit : $M_{u,r_i} = \bigotimes_{i\in\ell_{u,r_i}} m_{r,i}$
2. Après nous combinons tous les r_i-pré-résultats en utilisant : $\bigotimes_{r_i\in R} M_{u,r_i}$

En fonction de l'opération $@$, une telle procédure peut finalement converger vers une certaine masse stationnaire.

Dans l'annexe A, nous présentons les démonstrations mathématiques relatives à la nouvelle règle de combinaison. D'abord, nous traitons ses propriétés, à savoir que la combinaison de deux fonctions de masse est une autre fonction de masse et qu'en général $\otimes$ est non-associative. Puis nous démontrons, en utilisant la théorie des chaînes de Markov, que l'opération $@$ converge.

Une fois que nous avons la masse de croyance globale pour un utilisateur donné, nous utilisons une version modifiée de la probabilité pignistique définie dans l'équation 4.5 afin de prendre la décision à propos du degré d'influence de l'utilisateur. Dans notre cas les masses de croyance sont définies sur Ω_{Inf} et la probabilité pignistique est calculée en répartissant uniformément la masse de Ω sur tous les autres éléments de Ω_{Inf} :

$$\text{bet}(x) = m(x) + \frac{m(\Omega)}{|\Omega|}, \quad x \in \Omega_{Inf}\backslash\{\Omega\} \tag{4.9}$$

L'algorithme 1 formalise l'étape de l'estimation de l'influence d'un utilisateur, il requiert en entrée le graphe d'influence G, l'initialisation des masses et des degrés d'influence pour les différentes relations $m_r,\ r \in R$, et la fonction $@$. Pour chaque utilisateur, l'algorithme

commence par calculer le nombre d'occurrences pour chaque type de relation ou *motif d'interaction* (définit par $|\ell_{u,r_i}|$). Puis, pour chaque type de relation r, en utilisant la formule (4.8), la combinaison des masses de croyance est calculée. Ensuite, l'équation (4.8) est utilisée à nouveau pour combiner les masses de croyance de toutes les relations. Et enfin, en utilisant l'équation (4.9), les distributions de masses de croyance sont transformées en probabilité pignistique. L'algorithme retourne au final le degré d'influence qui est le degré ayant la probabilité pignistique maximale. L'ensemble des degrés d'influence finaux $\{\mathrm{Inf}_u : u \in U\}$ est noté par Inf.

Le code source est disponible sur github : https://github.com/kerzol/Influence-assessment-in-twitter. Il s'agit du code R[2] générique qui peut être spécialisé en fonction du domaine étudié et des relations utilisées.

Algorithme 1 : TwitBelief

Entrées : G le graphe d'influence sur U
L'ensemble des relations $R = r_1, r_2, \ldots$
Initialisation des masses $m_r, r \in R$
Opération $@$

Sorties : Degrés d'influence Inf_u, distribution de probabilité pignistique Bet_u pour chaque utilisateur $u \in U$

pour $u \in U$ **faire**
  pour $i \in [1..|R|]$ **faire**
    $|\ell_{u,r_i}|$:= nombre de relations ou motifs pour le type de relation r_i pour l'utilisateur u dans le graphe G;
    $M_{u,r_i} := m_{r_i}$;
    pour $i \in [2..|\ell_{u,r_i}|]$ **faire**
      $M_{u,r_i} := M_{u,r_i} \otimes m_{r_i}$;
      // Notons que $\otimes$ dépend de $@$, voir équation 4.8.
    fin
  fin
  $M_u := M_{u,r_1}$;
  pour $i \in [2..|R|]$ **faire**
    $M_u := M_u \otimes M_{u,r_i}$
  fin
  Bet_u := distribution de probabilité pignistique obtenue en utilisant l'équation 4.9;
  Inf_u := degré d'influence ayant la probabilité pignistique maximale ;
fin
retour $\mathrm{Inf}_u, \mathrm{Bet}_u, u \in U$;

Afin de discuter la complexité de l'algorithme 1, nous devons déterminer la complexité de la règle de combinaison (équation 4.8) et de la probabilité pignistique (équation 4.9). La complexité de l'opérateur de fusion $@$ est $O(|\Omega_{Inf}|^2)$ en général, parce qu'il correspond à la multiplication d'un vecteur par une matrice. Soient $|U|$ le nombre d'utilisateurs et d_u le nombre de relations quelque soit le type impliquant un utilisateur u et $\Delta = \max(d_u),\ u \in U$. Ainsi, le nombre maximal de combinaisons pour calculer Inf_u est $\Delta - 1$ pour tout utilisateur u. La complexité de la règle de combinaison est donc de $O(|U|\Delta|\Omega_{Inf}|^2)$. La distribution de probabilité pignistique Bet_u est calculée en $O(\Omega_{Inf})$ opérations et le degré d'influence Inf_u

2. R est à la fois un logiciel de statistique et un langage de programmation. Il est dédié aux statistiques et à la science des données. https://www.r-project.org/

est également calculé en $O(\Omega_{Inf})$ opérations, la complexité est alors de $O(|U|\Delta|\Omega_{Inf}|^2)$. La complexité de l'algorithme 1 est $O(|U|\Delta|\Omega_{Inf}|^4)$. Dans notre cas, $|\Omega_{Inf}|$ étant fixe, on peut alors conclure que la complexité de l'algorithme est de $O(|U|\Delta)$.

4.3.2.1/ EXEMPLE D'ILLUSTRATION

Afin de mieux comprendre l'étape de l'estimation de l'influence, nous présentons un exemple d'illustration dans lequel nous considérons les fonctions de masse suivantes associées aux relations *retweet* et *mention* :

$$\text{Retweet} \mapsto \begin{cases} m_{\text{retweet}}(\text{Faible}) = 0.4 \\ m_{\text{retweet}}(\Omega) = 0.6 \end{cases} \qquad \text{Mention} \mapsto \begin{cases} m_{\text{mention}}(\text{T.Faible}) = 0.3 \\ m_{\text{mention}}(\Omega) = 0.7 \end{cases}$$

Les masses de croyance $m_{\text{retweet}}(\Omega)$ et $m_{\text{mention}}(\Omega)$ représentent l'ignorance partielle.

Cas 1 : Deux retweets

Pour effectuer la combinaison de deux *retweets*, nous utilisons d'abord l'opération $@$ donnant les correspondances entre les degrés d'influence (tableau 4.2), après nous calculons la combinaison conjonctive en utilisant l'équation 4.8. La fonction de masse combinée des deux relations est donnée dans le tableau 4.3 :

TABLE 4.3 – Combinaison de deux *retweets*

$\otimes$	Faible 0.4	Ω 0.6
Faible 0.4	Assez Moyenne 0.16	Faible 0.24
Ω 0.6	Faible 0.24	Ω 0.36

Nous obtenons alors :

$m(\text{Faible}) = 0.24 + 0.24 = 0.48$

$m(\text{Assez Moyenne}) = 0.16$

$m(\Omega) = 0.36$

Enfin, afin de prendre une décision sur le degré d'influence, nous calculons la probabilité pignistique en utilisant l'équation (4.9) (voir le tableau 4.4). Par exemple, pour le degré Faible, Nous procédons comme suit :

$$\text{bet}(\text{Faible}) = m(\text{Faible}) + \frac{m(\Omega)}{|\Omega|} = 0.48 + \frac{0.36}{8} = 0.525$$

TABLE 4.4 – Probabilité pignistique pour deux *retweets*

Très Faible	0.045
Faible	**0.525**
Assez Moyenne	0.205
Moyenne	0.045
Assez Forte	0.045
Forte	0.045
Très Forte	0.045
Extrêmement Forte	0.045

Nous en concluons que le degré d'influence est Faible puisqu'il a la probabilité pignistique maximale 0.525. Ce degré avait une masse de 0.4 avant de considérer la combinaison de deux *retweets*.

Cas 2 : Deux retweets + deux mentions

Dans le second cas, nous considérons deux *mentions* additionnelles existantes entre les mêmes utilisateurs du cas 1. Afin de mesurer l'influence, nous utilisons notre processus proposé pour combiner les masses des deux *mentions* puis nous combinons les masses obtenues avec les résultats du cas 1. La combinaison conjonctive sur les deux *mentions* donne :

TABLE 4.5 – Combinaison de deux *mentions*

$\otimes$	Très Faible 0.3	Ω 0.7
Très Faible 0.3	Faible 0.09	Très Faible 0.21
Ω 0.7	Très Faible 0.21	Ω 0.49

Nous obtenons :

$m(\text{Très Faible}) = 0.42$

$m(\text{Faible}) = 0.09$

$m(\Omega) = 0.49$

Maintenant, nous combinons les masses obtenues avec les résultats du cas 1 :

TABLE 4.6 – Cas 2 : deux *retweets* + deux *mentions*

$\otimes$	Faible 0.48	Assez Moyenne 0.16	Ω 0.36
Très Faible 0.42	Assez Moyenne 0.2016	Moyenne 0.0672	Très Faible 0.1512
Faible 0.09	Assez Moyenne 0.0432	Moyenne 0.0144	Faible 0.0324
Ω 0.49	Faible 0.2352	Assez Moyenne 0.0784	Ω 0.1764

Nous obtenons :

$m(\text{Très Faible}) = 0.1512$; $m(\text{Faible}) = 0.2676$

$m(\text{Assez Moyenne}) = 0.3232$; $m(\text{Moyenne}) = 0.0816$; $m(\Omega) = 0.1764$

Nous notons que, en combinant les quatre relations, la masse de croyance sur le degré Faible a diminué par rapport au premier cas, cela est dû au fait que la masse du degré Assez Moyenne a augmenté et devenue égale à 0.3232. Nous notons également que le degré Moyenne est apparu avec une masse égale à 0.0816. Nous pouvons conclure que plus nous avons de relations et plus nous les combinons, plus l'influence augmente. Maintenant, pour prendre une décision sur le degré d'influence, nous calculons la probabilité pignistique (table 4.7). Nous concluons que le degré d'influence pour deux *retweets* et deux *mentions* est Assez Moyenne avec une probabilité pignistique de 0.34525. Avant d'envisager les deux *mentions*, le degré d'influence Assez Moyenne était de 0.205.

TABLE 4.7 – Probabilité pignistique pour deux *retweets* + deux *mentions*

Très Faible	0.17325
Faible	0.28965
Assez Moyenne	**0.34525**
Moyenne	0.10365
Assez Forte	0.02205
Forte	0.02205
Très Forte	0.02205
Extrêmement Forte	0.02205

4.3.3/ CLASSEMENT DES UTILISATEURS

Dans cette étape, nous exploitons les résultats de l'estimation de l'influence pour classer les utilisateurs en fonction de leur degré influence. Tout d'abord, pour chaque utilisateur, nous prenons le degré d'influence ayant la probabilité pignistique maximale . Ensuite, nous classons les utilisateurs en utilisant ce « degré d'influence maximal ». Lorsque deux utilisateurs ont le même « degré d'influence maximal », nous les classons selon le degré d'influence supérieur suivant avec l'ordre de classement des degrés d'influence suivant :

$\Omega <$ T.Faible $<$ Faible $<$ A.Moyenne $<$ Moyenne $<$ A.Forte $<$ Forte $<$ T.Forte $<$ E.Forte

Nous procédons ainsi, car il faut tenir compte de l'ensemble des probabilités pignistiques que les utilisateurs ont sur les différents degrés. Pendant le processus de fusion des masses, l'influence d'un utilisateur augmente et il passe d'un degré d'influence au degré d'influence supérieur suivant, etc. Ainsi, pour les utilisateurs qui ont de nombreuses relations et donc où de nombreuses combinaisons sont faites, ils partent d'un degré d'influence donné jusqu'à atteindre des degrés d'influence plus élevés. La masse sur un degré est au début du processus faible puis, par combinaisons, elle devient plus importante, pour ensuite diminuer afin que la masse sur le prochain degré d'influence supérieur suivant augmente à son tour. Pour cette raison, afin de classer les utilisateurs qui ont une probabilité pignistique maximale sur le même degré (par exemple, deux utilisateurs qui ont le degré Très forte comme degré maximal), nous ne considérons pas la

probabilité pignistique qu'ils ont sur ce degré, parce que, nous pouvons avoir un utilisateur plus influent que l'autre bien qu'il ait une probabilité pignistique plus faible que lui sur le même degré. Ceci est dû au fait que la masse de croyance du degré plus fort suivant a augmenté et est devenue importante. En conséquent, on peut établir la régle suivante : pour comparer les utilisateurs qui possèdent une probabilité pignistique maximale sur le même degré, nous considérons la probabilité pignistique sur le degré supérieur suivant.

L'algorithme 2 décrit la méthode utilisée pour classer les utilisateurs.

Algorithme 2 : Classement des utilisateurs

Entrées : Ensemble d'utilisateurs U.
Distribution de probabilité pignistique Bet_u pour chaque utilisateur $u \in U$
Sorties : Classement d'utilisateurs R

1 **pour** $u \in U$ **faire**
2 $MaxInf_u$:= Degré d'influence ayant la probabilité pignistique maximale;
3 $SecM_u$:= Degré d'influence supérieur suivant de $MaxInf_u$;
4 **fin**
5 R = Classement de U selon $MaxInf$, $SecM$;
6 // d'abord nous classons les utilisateurs selon $MaxInf$,
7 // dans le cas d'égalité, nous les classons selon $SecM$.
8 retour R ;

Pour calculer $MaxInf_u$ et $SecM_u$ pour un utilisateur u nous effectuons $O(|\Omega_{Inf}|)$ opérations. Nous avons $|U|$ utilisateurs, ainsi, la complexité des lignes 1-3 est $O(|U||\Omega_{Inf}|)$. Après, nous classons l'ensemble des utilisateurs, la complexité de l'algorithme 2 est $O\big(|U|(|\Omega_{Inf}| + log|U|)\big)$.

4.3.3.1/ EXEMPLE D'ILLUSTRATION

Afin d'illustrer le principe de classement des utilisateurs selon leur influence dans *TwitBelief*, nous présentons un exemple dans lequel nous considérons trois utilisateurs ayant les distributions de probabilité pignistique suivantes :

TABLE 4.8 – Distribution de probabilité pignistique pour un exemple de trois utilisateurs

	U_1	U_2	U_3
T.Faible	0	0.000011065	0.000030278
Faible	0	0.00007295998	0.0001832843
A.Moyenne	0	0.0007035528	0.001403947
Moyenne	0	0.003033557	0.004954501
A.Forte	0	0.008340205	0.01247841
Forte	0	0.02191526	0.02977818
T.Forte	0.1826552	0.5830090	0.7960571
E.Forte	0.8173448	0.3829144	0.1551143

Comme décrit dans la section précédente, nous prenons en premier, pour chaque uti-

lisateur, le degré d'influence avec la probabilité pignistique maximale (par exemple, $Inf(U_1)$ = E.Forte). Après, nous classons les utilisateurs selon leur degré d'influence maximal. Si deux utilisateurs ont le même degré d'influence maximal :

$$Inf(U_2) = Inf(U_3) = \text{T.Forte}$$

nous comparons les probabilité pignistiques sur le degré d'influence supérieur suivant : $\text{bet}_{U_2}(\text{E.Forte}) > \text{bet}_{U_3}(\text{E.Forte})$ ainsi, nous obtenons le classement suivant :

TABLE 4.9 – Classement des utilisateurs

Classement	Utilisateurs	Degré d'influence	Probabilité pignistique
1	U_1	E.Forte	0.8173448
2	U_2	T.Forte	0.5830090
3	U_3	T.Forte	0.7960571

Le chapitre 6 présente nos résultats sur trois corpus de données.

4.4/ CONCLUSION

Dans ce chapitre, nous avons proposé *TwitBelief*, une approche pour l'estimation de l'influence des utilisateurs dans un réseau social. Le réseau social est modélisé par un graphe d'influence vu comme un graphe multiplexe étiqueté où les différentes relations entre les utilisateurs sont représentées. *Twitter* est utilisé comme exemple de réseau social. *TwitBelief* est une approche topologique qui prend en considération la diversité des relations du réseau et leur importance dans le domaine étudié. La théorie des fonctions de croyance est adaptée, elle permet d'établir pour chaque utilisateur une estimation de son degré d'influence dans le réseau avec un degré de certitude. Le degré d'influence est une échelle fixée par les experts du domaine. Notre approche *TwitBelief* donne des résultats plus précis, contrairement à la majorité des travaux existants qui se contentent de classer les utilisateurs selon leur influence ce que permet aussi *TwitBelief*. L'estimation de l'influence d'un utilisateur considère aussi bien les relations directes entre utilisateurs que les relations indirectes qui passent par des enchaînements de relations (*motifs d'interactions*) *via* des utilisateurs intermédiaires.

Cependant, dans *TwitBelief*, seules les relations et leur sémantique sont exploitées dans le processus d'estimation de l'influence d'un utilisateur. Dans le chapitre suivant, nous étendons *TwitBelief* au contenu des *tweets*. Dans un premier temps, la polarité des tweets est exploitée afin d'étudier si l'influence exercée par un utilisateur est positive, neutre ou négative. La polarité des *tweets* est déterminée en utilisant un modèle d'analyse de sentiments. La deuxième extension porte sur le style de communication adopté par les utilisateurs. Il s'agit de savoir comment un utilisateur utilise les différents opérateurs (leur position, leur association) mais aussi les hashtags, les URLs dans ses *tweets* de façon à déterminer si il a un style interactif ou informatif. Le chapitre suivant présente ces deux extensions.

5

EXTENSION DE *TwitBelief* AU CONTENU DES TWEETS

Dans ce chapitre, nous étendons *TwitBelief* afin de prendre en compte le contenu des *tweets*. D'abord, nous décrivons la démarche suivie pour l'analyse de sentiment exprimé dans les *tweets* puis nous montrons comment nous adaptons *TwitBelief* afin d'étudier la polarité de l'influence. Il s'agit de déterminer si l'influence est positive, négative ou neutre. Ensuite, nous présentons la deuxième extension dans laquelle nous étudions le style de communication adopté par les utilisateurs dans leurs *tweets*. Cette extension vise à décrire la manière avec laquelle les utilisateurs de *Twitter* interagissent avec les autres. Contrairement à *TwitBelief*, les extensions que nous présentons dans ce chapitre se placent dans la fouille du contenu du Web « *Web content mining* » puisque nous exploitons le contenu des *tweets*.

5.1/ INFLUENCE POLARISÉE

Dans cette section, le contenu des *tweets* est exploité afin d'analyser le sentiment exprimé par les utilisateurs, pour déterminer, en étendant *TwitBelief*, si leur influence est positive, négative ou neutre.

5.1.1/ ANALYSE DE SENTIMENT DANS *Twitter*

L'analyse du langage subjectif a été largement appliquée à la classification des opinions et des émotions dans le texte [Wiebe et al., 2005]. En effet, l'analyse de sentiment, qui vise à annoter le texte à l'aide d'une échelle mesurant le degré de sentiment négatif et positif dans le texte, est considérée comme l'un des axes de recherche les plus importants pour les chercheurs dans les domaines de recherche d'information, fouille de données et apprentissage automatique.

Dans ce contexte, *Twitter* a constitué le terrain de jeu le plus utilisé pour les solutions d'analyse de sentiments, les entreprises et les scientifiques tentent de comprendre l'enthousiasme des utilisateurs partageant leurs opinions publiquement en ligne. Une des difficultés inhérentes dans l'analyse de sentiment est la traduction des données textuelles dans un format que l'ordinateur peut comprendre et traiter. Pour cette raison, un certain nombre de méthodes de Traitement Automatique de Langage (TAL) ont été développées au fil des années. Les plus populaires sont le sac de mots et les N-grammes. Le sac de

mots peut être considéré comme la méthode la plus simple. Selon cette approche, les phrases du document (ou texte) dont la machine a besoin pour juger le sentiment sont divisées en un ensemble de mots en utilisant l'espace ou les caractères de ponctuation [Pak et al., 2010]. Un document ou un texte particulier est représenté par les occurrences des mots le composant. Les N-grammes sont très semblables au sac de mots mais la différence réside dans le fait que le texte est divisé en pseudo-mots consécutifs de longueur égale [Pang et al., 2008]. La longueur N dépend de la nature des documents ou des textes d'entrée et du problème à résoudre. Généralement, 2-grammes, 3-grammes et 4-grammes sont utilisés. Les N-grammes permettent par exemple d'attacher la négation avec le mot qui la suit (par exemple : je n'aime pas). Une telle procédure permet d'améliorer l'analyse de sentiments dans des textes puisque la négation joue un rôle particulier dans une expression d'opinion et de sentiment.

La popularité des réseaux sociaux a rendu la tâche de l'analyse de sentiment difficile. En effet, les textes à analyser sont devenus courts, contenant de nombreuses abréviations, ainsi que de nombreuses erreurs de syntaxe et de grammaire. Il devient impératif de filtrer et de nettoyer les textes avant toute étape d'analyse de sentiments. Plusieurs possibilités s'offrent telles que l'élimination des mots vides (ou *stop words*) qui sont des mots qui sont tellement communs qu'il est inutile de les utiliser dans une recherche. En français, des mots vides évidents sont « le », « la », « de », « du », « ce ». La racinisation des mots (*stemming*) est aussi pratiquée afin de filtrer les textes, c'est-à-dire la transformation/réduction des mots en leur racine qui correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s).

Une fois les textes préparés sous forme de sac de mots et/ou N-grammes puis filtrés, l'étape suivante est leur analyse en utilisant différents algorithmes d'apprentissage automatique (classifieurs). Dans [Psomakelis et al., 2015], les auteurs présentent une revue des algorithmes les plus populaires en analyse de sentiments dans *Twitter*. Ils ont testé plusieurs algorithmes de classification : les machines à vecteurs de support, la classification naïve bayésienne, la régression logistique, le perceptron multi-couches et les arbres de décision. Les résultats ont montré la supériorité des performance de l'algorithme de régression logistique en utilisant 5-grammes.

Burnap et al. [Burnap et al., 2015b] ont proposé un classifieur de *tweets* par rapport au discours de haine. La méthode est basée sur le sac de mots, un dictionnaire de termes et phrases de discours de haine de Wikipedia et les dépendances typées [1] [Marneffe et al., 2006] afin de représenter les relations grammaticales entre les mots dans une phrase. Les auteurs utilisent les algorithmes des machines à vecteurs de support, les forêts d'arbres décisionnels et la régression logistique bayésienne. Dans les résultats des expérimentations, les performances des différents algorithmes utilisés étaient similaires et les critères les plus efficaces sont les N-grammes combinés avec le dictionnaire des termes relatifs à la haine. Dans [Burnap et al., 2016], les auteurs proposent un modèle de classification supervisé pour la détection du discours de haine associé à différents sujets qui sont la race, le handicap et l'orientation sexuelle. Ils se sont basés sur quatre critères : le sac de mots, les N-grammes, des termes et phrases de discours de haine de Wikipedia et les dépendances typées. Deux algorithmes de classification ont été utilisés : les machines à vecteurs de support et les forêts d'arbres décisionnels. Les résultats ont montré que l'utilisation des dépendances typées est très intéressante contrairement aux

1. La représentation de dépendances typées a été conçue pour fournir une description simple des relations grammaticales dans une phrase qui peut être facilement comprise et utilisée efficacement par des personnes sans expertise linguistique qui souhaitent extraire des relations textuelles.

termes de discours de haine qui sont des indicateurs faibles. Dans [Burnap et al., 2015a], Burnap et al. développent une application de fouille d'opinion capable de classer les *tweets* publics en tenant compte du niveau de tension. Les algorithmes de machines à vecteurs de support et la classification naïve bayésienne ont été utilisés en se basant sur les N-grammes et sur un ensemble de mots préalablement classés en tant que mots qui expriment la tension. Les résultats indiquent que le dictionnaire de mots utilisé est un fort indicateur de tension. Il existe également des sites disponibles en ligne pour analyser les sentiments dans *Twitter*[2] en introduisant le nom de la personne ou l'entité à propos de laquelle nous souhaitons connaître le sentiment. Ces outils sont des boîtes noires et ne montrent pas l'approche utilisée pour analyser les sentiments.

En conclusion, dans les recherches d'analyse de sentiments, la méthode dépend du domaine étudié, certains algorithmes ou méthodes peuvent donner de bons résultats dans un domaine et échouer dans d'autres. Le principal paramètre à prendre en compte est l'adéquation entre les caractéristiques retenues pour la modélisation du texte et la question à laquelle nous souhaitons répondre.

5.1.2/ MÉTHODE SUIVIE POUR L'ANALYSE DE SENTIMENTS DES *Tweets*

L'objectif est d'analyser la polarité des *tweets*, c'est-à-dire déterminer si le contenu des *tweets* est positif, négatif ou neutre. Pour construire un modèle d'analyse de sentiments qui sera capable de classer les *tweets*, nous prenons un ensemble de *tweets* manuellement annotés par des experts selon leur polarité et nous essayons de déduire les critères, présents dans les *tweets*, qui permettent de déterminer leurs polarités.

5.1.2.1/ PRÉPARATION DES TWEETS

La première étape de la construction du modèle d'analyse de sentiments est la préparation des *tweets*. Il s'agit de présenter les *tweets* à analyser sous une forme compréhensible par les différents algorithmes d'analyse de sentiments. Pour ceci, chaque *tweet* est transformé en sac de mots. Le processus d'obtention de N-grammes à partir d'un *tweet* est le suivant :

- **Le filtrage :** nous supprimons les mots vides, les liens URL, les noms d'utilisateur *Twitter* (avec le symbole @ indiquant un nom d'utilisateur), les mots spéciaux de *Twitter* (tel que « RT »).
- **Le sac de mots :** les *tweets* sont divisés en un ensemble de mots en utilisant l'espace entre les mots.
- **Les N-grammes :** les *tweets* sont représentés aussi sous forme de 3-grammes et 2-grammes de mots consécutifs mais nous avons constaté que les résultats sont meilleurs sans transformation en N-grammes (Voir le tableau 5.1).

5.1.2.2/ ALGORITHME DES FORÊTS D'ARBRES DÉCISIONNELS

L'étape suivante est l'utilisation d'un algorithme d'analyse de sentiments. Nous construisons un classifieur en utilisant l'algorithme des forêts d'arbres décisionnels. Cet algorithme a été formellement proposé en 2001 par Leo Breiman [Breiman, 2001] et fait partie des techniques d'apprentissage supervisé. La proposition de Breiman vise à corriger plusieurs

2. http://socialmouths.com/2010/03/31/6-tools-for-twitter-sentiment-tracking/

inconvénients connus des arbres de décision dont la principale limite est la dépendance des performances à l'échantillon de départ. Il s'agit précisément de l'*overfitting* qui se produit quand un modèle est excessivement complexe, comme avoir trop de paramètres par rapport au nombre d'observations. L'*overfitting* d'un modèle se traduit par de mauvaises performances prédictives sur des données autres que les jeux d'apprentissage ou de tests. Le nom **forêt** d'arbres vient de la construction de nombreux arbres de décision. Pour éviter d'avoir des arbres semblables, chaque arbre construit dispose d'une vision parcellaire du problème, conditionnée par un double tirage **aléatoire** : tirage aléatoire sur les observations et tirage aléatoire sur les variables. Dans le principe des forêts d'arbres décisionnels, plutôt que d'avoir un modèle d'arbre décisionnel complexe, il s'agit de construire de nombreux modèles d'arbres décisionnels moins performants individuellement ayant sa vision du problème et faisant au mieux pour le résoudre avec les données dont il dispose. Pour construire chaque arbre, les critères de Gini [3] et d'entropie [4] sont utilisés. Les arbres créés sont ensuite unis pour donner une vision globale du problème, ce qui rend les forêts d'arbres décisionnels très efficaces.

5.1.2.3/ MODÈLE D'ANALYSE DE SENTIMENTS

Pour construire le modèle d'analyse de sentiments, nous avons choisi aléatoirement un échantillon de 2000 *tweets* composé de 1000 *tweets* contenant des *mentions* et 1000 *réponses*. Ensuite, les *tweets* ont été manuellement annotés par des sociologues selon leurs polarités : positif, neutre ou négatif. Cet ensemble de *tweets* est divisé en deux, le premier est un échantillon à partir duquel nous construisons le modèle de classification, ensuite le modèle construit est utilisé pour prédire la polarité du deuxième échantillon de *tweets*. Enfin, nous comparons les prédictions avec leur polarité spécifiée pour évaluer les performances du modèle construit.

La première étape de la construction du modèle d'analyse de sentiments est le filtrage des *tweets* et la préparation des données. Après le filtrage, les *tweets* sont nettoyés et prêts à être utilisés par l'algorithme d'analyse de sentiments. Nous avons obtenu 861 termes (mots) différents.

Nous procédons ensuite à une approche de validation croisée, une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. L'échantillon est divisé en deux, 1700 *tweets* comme ensemble de validation et les 300 *tweets* restants constitueront l'ensemble d'apprentissage. Nous appliquons l'algorithme des forêts d'arbres décisionnels pour construire le modèle de classification des *tweets* à partir de l'échantillon d'apprentissage. Le choix de cet algorithme est justifié par sa forte performance dans les travaux de l'état de l'art. De plus, l'algorithme permet d'obtenir un modèle lisible et facilement interprétable. En se basant sur leurs polarités réelles, l'algorithme cherche à identifier les termes présents dans les *tweets* qui permettent d'identifier leurs polarités. Le nombre d'arbres générés est fixé à 500 arbres décisionnels, paramètre généralement utilisé dans les tâches de classification. Ensuite, le modèle construit est testé sur l'échantillon de validation (1700 *tweets*) pour étudier sa performance en comparant les prédictions du modèle avec leurs polarités réelles. La métrique d'évaluation utilisée est le F-mesure, basé sur la Précision P et le Rappel R calculés pour chaque polarité, ce qui est classique dans les tâches de classification. Le F-Score est calculé comme suit :

3. Le critère de Gini se focalise sur la séparation de la classe la plus représentée.
4. Le critère d'entropie vise à maximiser le gain d'information à chaque étape de réalisation de l'algorithme.

$$F = 2 \times \frac{(P \times R)}{P + R} \tag{5.1}$$

$P = VP/VP + FP$; $R = VP/VP + FN$
où $VP =$ vrais positifs, $FP =$ faux positifs, et $FN =$ faux négatifs.

Le tableau 5.1 représente les différentes valeurs de F-mesures avec la variation des paramètres N-grammes et le pourcentage des termes sparses, il s'agit d'ignorer les termes qui ont une sparsité supérieure à un seuil donné (la sparsité = 1 - fréquence), ce qui peut aider à prévenir l'*overfitting*. Par exemple, si sparse est égal à 0.8, cela supprimera chaque terme qui apparaît dans moins de 20% de documents. Au contraire, si sparse est égal à 0.01, seuls les termes qui apparaissent dans presque chaque document seront conservés. En langage naturel, des mots communs comme «le» sont susceptibles de se produire dans chaque texte et donc ne seront jamais sparses.

TABLE 5.1 – F-mesure en fonction des variations des différents paramètres

N-grammes	3	2	Non	Non	**Non**
Termes sparses	0.997	0.997	0.991	0.993	**0.997**
Nombre de termes	59	67	20	45	**117**
Précision	0.66	0.76	0.78	0.78	**0.79**
Rappel	0.61	0.64	0.72	0.75	**0.76**
F-mesure	0.63	0.69	0.74	0.76	**0.77**

Nous constatons que la meilleure valeur de F-mesure a été obtenue en utilisant une valeur sparse égale à 0.997 sans transformation en N-grammes. Le nombre de termes utilisés est 117. Ces paramètres constituent notre modèle qui sera appliqué dans les expérimentations. La figure 5.1 montre les 20 mots les plus utilisés dans les *tweets* filtrés. La taille des mots représentent l'importance de leur utilisation dans les *tweets*, par exemple, le mot « pas » est le plus utilisé.

FIGURE 5.1 – Nuage de mots des *tweets* filtrés

5.1.3/ ESTIMATION DE L'INFLUENCE POLARISÉE ET RÉSULTATS

Après avoir analysé le sentiment exprimé dans les *tweets* en utilisant l'algorithme des forêts d'arbres décisionnels, nous procédons à l'estimation de l'influence polarisée. D'abord, pour chaque utilisateur, en fonction de la polarité des *tweets* obtenus, le corpus est divisé en trois chacun représentant une polarité. *TwitBelief* est alors appliqué dans chaque graphe d'influence construit. Ainsi, nous obtenons pour chaque utilisateur et pour au plus trois polarités, une masse et un degré d'influence. Ensuite nous combinons ces informations afin d'estimer utilisateur par utilisateur son influence polarisée. La figure 5.2 résume le processus suivi.

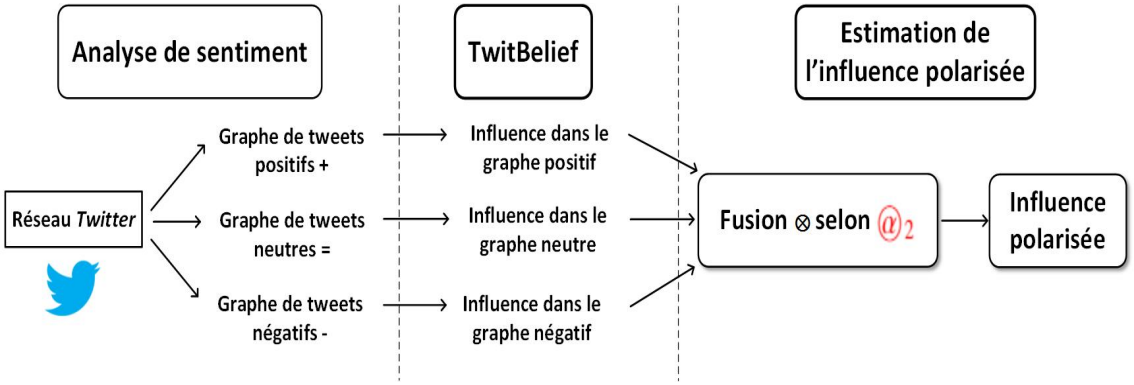


FIGURE 5.2 – Principe de l'estimation de l'influence polarisée

Afin d'estimer l'influence polarisée, nous utilisons un nouveau cadre de discernement Ω_{Pol} qui représente les différentes réponses possibles à notre question : "Quelle est l'influence polarisée d'un certain utilisateur ?" Soit Ω_{Pol} l'ensemble des degrés d'influence polarisée possibles. Les masses obtenues à partir des trois graphes d'influence polarisés sont combinées en utilisant le principe de *TwitBelief* mais avec une nouvelle fonction $@_2$. Cette fonction, donnée dans le tableau 5.2, attribue les correspondances entre les combinaisons des différents couples polarité/degré d'influence. Le principe à travers ces correspondances est d'obtenir pour un utilisateur donné, la polarité dominante dans ses *tweets*. La polarité dominante est celle qui concerne le graphe d'influence ayant la taille la plus importante par rapport aux deux autres graphes. Or, la taille d'un graphe (d'une polarité) est exprimée à travers les degrés d'influence, puisque dans *TwitBelief*, plus la taille du graphe est importante, plus nous effectuons des combinaisons, et plus le degré d'influence augmente. Par exemple, combiner le degré d'influence très faible trouvé dans le graphe de polarité positive (+T.F) au degré d'influence assez moyenne du graphe négatif (-A.Moy) donne le couple (-F), la polarité retenue est négative car sa taille est plus importante que celle de la polarité positive (A.Moy $>$ T.F). Cependant, le degré d'influence obtenu est moins important (F $<$ A.Moy) car, avec la présence de la polarité positive, la croyance dans la polarité négative diminue.

TABLE 5.2 – Définition de l'opération $@_2$

$@_2$	**+T.F**	**=T.F**	**-T.F**	**+F**	**=F**	**-F**	**+A.Moy**	**=A.Moy**	**-A.Moy**	**+Moy**	**=Moy**	**-Moy**	**+A.Fo**	**=A.Fo**	**-A.Fo**	**+Fo**	**=Fo**	**-Fo**	**+T.Fo**	**=T.Fo**	**-T.Fo**	**+E.Fo**	**=E.Fo**	**-E.Fo**	Ω_{Pol}
+T.F	+F	+T.F	=T.F	+A.Moy	+T.F	-T.F	+Moy	+T.F	-F	+A.Fo	+T.F	-A.Moy	+Fo	=A.Moy	-A.Moy	+T.Fo	=A.Moy	-A.Moy	+E.Fo	=Moy	-Mo	+E.Fo	=Moy	-Moy	+T.F
=T.F	+T.F	=T.F	-T.F	+T.F	=T.F	-T.F	+F	=F	-F	+A.Moy	=A.Moy	-A.Moy	+A.Moy	=A.Moy	-A.Moy	+A.Moy	=A.Moy	-A.Moy	+Moy	=Moy	-Moy	+Moy	=Moy	-Moy	=T.F
-T.F	=T.F	-T.F	-F	+T.F	-T.F	-A.moy	+F	-T.F	-Moy	+A.Moy	-T.F	-A.Fo	+A.Moy	=A.Moy	-Fo	+A.Moy	=A.Moy	-T.Fo	+Moy	=Moy	-E.Fo	+Moy	=Moy	-E.Fo	-T.F
+F	+A.Moy	+T.F	+T.F	+A.Moy	+T.F	=F	+Moy	+F	-F	+A.Fo	+T.F	-F	+Fo	=A.Moy	-A.Moy	+T.Fo	=Moy	-Moy	+E.Fo	=Moy	-Moy	+E.Fo	=A.Fo	-A.Fo	+F
=F	+T.F	=T.F	-T.F	+T.F	=F	-T.F	+F	=F	-F	+F	=F	-F	+A.Moy	=A.Moy	-A.Moy	+Moy	=Moy	-Moy	+Moy	=Moy	-Moy	+A.Fo	=A.Fo	-A.Fo	=F
-F	-T.F	-T.F	-A.moy	=F	-T.F	-A.Moy	+F	-T.F	-Moy	+F	-T.F	-A.Fo	+A.Moy	=A.Moy	-Fo	+Moy	=Moy	-T.Fo	+Moy	=Moy	-E.Fo	+A.Fo	=A.Fo	-E.Fo	-F
+A.Moy	+Moy	+F	+F	+Moy	+F	+F	+Moy	+F	=A.Moy	+A.Fo	+F	-A.Moy	+Fo	=Moy	-Moy	+T.Fo	=Moy	-Moy	+E.Fo	=A.Fo	-A.Fo	+E.Fo	=A.Fo	-A.Fo	+A.Moy
=A.Moy	+T.F	=F	-T.F	+T.F	=F	-T.F	+F	=A.Moy	-F	+A.Moy	=A.Moy	-A.Moy	+Moy	=Moy	-Moy	+Moy	=Moy	-Moy	+A.Fo	=A.Fo	-A.Fo	+A.Fo	=A.Fo	-A.Fo	=A.Moy
-A.Moy	-F	-F	-Moy	-F	-F	-Moy	=A.Moy	-F	-Moy	+A.Moy	-F	-A.Fo	+Moy	=Moy	-Fo	+Moy	=Moy	-T.fo	+A.Fo	=A.Fo	-E.Fo	+A.Fo	=A.Fo	-E.Fo	-A.Moy
+Moy	+A.Fo	+A.Moy	+A.Moy	+A.Fo	+F	+F	+A.Fo	+A.Moy	+A.Moy	+A.Fo	+F	=Moy	+Fo	=Moy	-Moy	T.Fo	=A.Fo	-A.Fo	+E.Fo	=A.Fo	-A.Fo	+E.Fo	=A.Fo	-A.Fo	+Moy
=Moy	+T.F	=A.Moy	-T.F	+T.F	=F	-T.F	+F	=A.Moy	-F	+F	=Moy	-F	+Moy	=Moy	-Moy	+A.Fo	=A.Fo	- A.Fo	+A.Fo	=A.Fo	-A.Fo	+A.Fo	=A.Fo	-A.Fo	=Moy
-Moy	-A.Moy	-A.Moy	-A.Fo	-F	-F	-A.Fo	-A.Moy	-A.Moy	-A.Fo	=Moy	-F	-A.Fo	+Moy	=Moy	-Fo	+A.Fo	=A.Fo	-T.Fo	+A.Fo	=A.Fo	-E.Fo	+A.Fo	=A.Fo	-E.Fo	-Moy
+A.Fo	+Fo	+A.Moy	+A.Moy	+Fo	+A.Moy	+A.Moy	+Fo	+Moy	+Moy	+Fo	+Moy	+Moy	+Fo	+A.Moy	=A.Fo	+T.Fo	=A.Fo	-A.Fo	+E.Fo	=A.Fo	-Fo	+E.Fo	=Fo	-Fo	+A.Fo
=A.Fo	=A.Moy	=A.Moy	=A.Moy	=A.Moy	=A.Moy	=A.Moy	=Moy	+Moy	+Moy	=Moy	=Moy	=Moy	+A.Moy	=A.Fo	-A.Moy	+A.Fo	=A.Fo	-A.Fo	+Fo	=Fo	-Fo	+Fo	=Fo	-Fo	=A.Fo
-A.Fo	-A.Moy	-A.Moy	-Fo	-A.Moy	-A.Moy	-Fo	-Moy	-Moy	-Fo	-Moy	-Moy	-Fo	=A.Fo	-A.Moy	-Fo	+A.Fo	=A.Fo	-T.Fo	+Fo	=Fo	-E.Fo	+Fo	=Fo	-E.Fo	-A.Fo
+Fo	+T.Fo	+A.Moy	+A.Moy	+T.Fo	+Moy	+Moy	+T.Fo	+Moy	+Fo	+T.Fo	+A.Fo	+A.Fo	+T.Fo	+A.Fo	+A.Fo	+T.Fo	+A.Moy	=Fo	+E.Fo	=A.Fo	-Fo	+E.Fo	=T.Fo	-T.Fo	+Fo
=Fo	=A.Moy	=A.Moy	=A.Moy	=Moy	=Moy	=Moy	=Moy	=Moy	=Moy	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	+A.Moy	=Fo	-A.Moy	+Fo	=Fo	-Fo	+T.Fo	=T.Fo	-T.Fo	=Fo
-Fo	-A.Moy	-A.Moy	-T.Fo	-Moy	-Moy	-T.Fo	-Moy	-T.Fo	-A.Fo	-A.Fo	-A.Fo	-T.Fo	- A.Fo	-A.Fo	-T.Fo	=Fo	-A.Moy	-T.Fo	+Fo	=Fo	-E.Fo	+T.Fo	=T.Fo	-E.Fo	-Fo
+T.Fo	+E.Fo	+Moy	+Moy	+E.Fo	+Moy	+E.Fo	+A.Fo	+A.Fo	+E.Fo	+E.Fo	+A.Fo	+A.Fo	+E.Fo	+Fo	+Fo	+E.Fo	+Fo	+Fo	+E.Fo	+Moy	=T.Fo	+E.Fo	=T.Fo	-T.Fo	+T.Fo
=T.Fo	=Moy	=Moy	=Moy	=Moy	=Moy	=Moy	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=Fo	=Fo	=Fo	=Fo	=Fo	=Fo	+Moy	=T.Fo	-Moy	+T.Fo	=E.Fo	-T.Fo	=T.Fo
-T.Fo	-Moy	-Moy	-E.Fo	-Moy	-Moy	-E.Fo	-A.Fo	-E.Fo	-A.Fo	-A.Fo	-E.Fo	-E.Fo	-Fo	-Fo	-E.Fo	-Fo	-Fo	-E.Fo	=T.Fo	-Moy	-E.Fo	+T.Fo	=T.Fo	-E.Fo	-T.Fo
+E.Fo	+E.Fo	+Moy	+Moy	+E.Fo	+A.Fo	+A.Fo	+E.Fo	+A.Fo	+A.Fo	+E.Fo	+A.Fo	+A.Fo	+E.Fo	+Fo	+Fo	+E.Fo	+T.Fo	+T.Fo	+E.Fo	+T.Fo	+T.Fo	+E.Fo	+Moy	=E.Fo	+E.Fo
=E.Fo	=Moy	=Moy	=Moy	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=A.Fo	=Fo	=Fo	=Fo	=T.Fo	T.Fo	=T.Fo	=E.Fo	=T.Fo	=T.Fo	=Moy	=E.Fo	-Moy	=E.Fo
-E.Fo	-Moy	-Moy	-E.Fo	-A.Fo	-A.Fo	-E.Fo	-A.Fo	-E.Fo	-A.Fo	-E.Fo	-Fo	-E.Fo	-Fo	-E.Fo	-TFo	-T.Fo	-T.Fo	-E.Fo	-TFo	-T.Fo	-E.Fo	-Moy	-E.Fo	-E.Fo	-E.Fo
Ω_{Pol}	+T.F	=T.F	-T.F	+F	=F	-F	+A.Moy	=A.Moy	-A.Moy	+Moy	=Moy	-Moy	+A.Fo	=A.Fo	-A.Fo	+Fo	=Fo	-Fo	+T.Fo	=T.Fo	-T.Fo	+E.Fo	=E.Fo	-E.Fo	Ω_{Pol}

Le choix de l'ordre de combinaison des trois polarités d'un utilisateur est important. Ainsi, nous choisissons l'ordre de combinaison suivant : (Positive $\otimes_{@_2}$ Négative) $\otimes_{@_2}$ Neutre. Ce choix est justifié par le fait que nous privilégions les polarités positive et négative par rapport à la polarité neutre. En effet, dans la littérature, la polarité neutre a été souvent négligée. À travers le choix de cet ordre de combinaison, nous souhaitons obtenir en premier lieu la combinaison des deux polarités positive et négative pour avoir la polarité dominante entre les deux. Puis nous combinons ce résultat avec la polarité neutre tout en privilégiant la polarité positive ou négative. Ceci est assuré à travers la fonction $@_2$ où la polarité neutre ne domine une autre polarité que si la taille de son graphe est importante ($>$ A.Forte). Par exemple, (+F) $@_2$ (=A.Moy) donne (+T.F), la polarité neutre dans ce cas ne domine pas la polarité positive car sa taille n'est pas importante. Toutefois, dans le cas de (+F) $@_2$ (=Fo), la polarité neutre domine et nous obtenons (=Moy). Nous limitons ainsi l'obtention de la polarité neutre puisqu'elle n'est pas très significative.

D'autre part, dans le cas où les deux polarités à combiner ont la même taille (c'est-à-dire le même degré d'influence), nous procédons comme suit :
+ $@_2$ - $\Longrightarrow$ = et nous gardons le même degré
+ $@_2$ = $\Longrightarrow$ + et nous diminuons le degré
- $@_2$ = $\Longrightarrow$ - et nous diminuons le degré.
Enfin, dans le cas de la combinaison de la même polarité, nous obtenons la même polarité avec augmentation du degré d'influence. Par exemple, (+T.F) $@_2$ (+Fo) donne (+T.Fo). Ceci signifie que la croyance que nous avons dans l'importance de la polarité positive augmente.

Après avoir combiné les informations issues des trois polarités, les masses de croyance sont transformées en probabilités pignistiques. Enfin, l'influence polarisée est représentée par la polarité ayant le degré d'influence le plus élevé ainsi que le degré d'influence et la probabilité pignistique correspondante. Des détails supplémentaires et des exemples d'illustration sont donnés dans l'annexe B.

Nous avons appliqué l'approche de l'estimation de l'influence polarisée sur le corpus français des *tweets* des élections européennes 2014 comprenant 616 candidats et 4 millions de *tweets*. Le tableau 5.3 montre la répartition de la polarité des *tweets* pour trois candidats.

TABLE 5.3 – Nombre de *tweets* par polarité pour trois candidats

Polarité	**Marine Le Pen**	**Florian Philippot**	**Jean-Luc Mélenchon**
Positive	538	429	149
Neutre	2358	1719	760
Négative	669	716	340

Le tableau 5.3 montre que les candidats ont un très grand nombre de relations. L'initialisation des masses de croyance dépend de la taille du graphe d'influence. En effet, comme expliqué dans l'annexe A, l'influence converge en fonction des masses de croyances et du nombre de relations. Si nous utilisons les masses de la sous-section 4.3.2.1, l'influence converge rapidement vers le plus haut degré et nous obtenons le même degré d'influence

pour les trois candidatss. Ainsi, pour chaque relation, nous utilisons les masses suivantes :

$$\textit{Relation} \mapsto \begin{cases} m_{\text{relation}}(\text{T.Faible}) = 0.005 \\ m_{\text{relation}}(\Omega) = 0,995 \end{cases}$$

TwitBelief est appliqué à chaque graphe d'influence. Ainsi, pour chaque polarité, chaque candidat est représenté par un degré d'influence et une masse de croyance. Par exemple, le candidat Marine Le Pen est représenté de la manière suivante :

$$\text{Marine Le Pen} \left\{ \begin{array}{l} \text{Positive, A.Moyenne, m = 0.24594492} \\ \text{Neutre, T.Forte, m = 0.9770573} \\ \text{Négative, Moyenne, m = 0.2204240} \end{array} \right\}$$

Les résultats de l'influence polarisée pour les trois candidats sont présentés dans le tableau 5.4. Les résultats fournissent non seulement le degré d'influence d'un candidat, mais aussi indiquent sa polarité et donnent par les masses une indication de la croyance que nous avons dans les résultats donnés.

TABLE 5.4 – Résultats de l'influence polarisée pour trois candidats

Candidat	**Polarité**	**Degré d'influence**	**Masse de croyance**
Marine Le Pen	Neutre	Forte	0.98155287
Florian Philippot	Neutre	Forte	0.885226444
Jean-Luc Mélenchon	Neutre	Assez Moyenne	0.362760929

5.2/ STYLES DE COMMUNICATION DANS *Twitter*

Dans cette section, nous présentons une deuxième extension de *TwitBelief* qui consiste à déterminer le style de communication adopté par les utilisateurs de *Twitter*. Cette extension diffère de *TwitBelief* qui utilise les opérateurs pour caractériser les relations entre utilisateurs alors qu'ici les opérateurs sont utilisés comme élément techno-discursif dans un *tweet*.

De manière générale, le style de communication décrit la manière dont une personne interagit en présence d'autrui non seulement sur le plan verbal mais aussi en prenant compte des composantes paraverbale et non verbale. La composante paraverbale désigne le ton, la hauteur (aiguë ou grave), l'intensité (forte ou faible), les intonations, le débit de la parole, etc. Quant aux comportements non verbaux, ils comprennent l'expression faciale, le contact visuel, la posture, les gestes, la distance, etc. Des stratégies de communication sont aussi mises en place dans les réseaux sociaux. Pour rejoindre une communauté particulière de *Twitter*, appelée *Twittersphere* dans [Ausserhofer et al., 2013], les utilisateurs suivent des règles pour être perçus comme faisant partie de cette communauté [Thimm et al., 2012a]. Cela signifie que les interactions interpersonnelles sociales [Goffman, 1967] doivent s'adapter à un dispositif technique qui possède sa langue particulière [Bays, 1998, Coutant et al., 2010], appelé « twittécriture et twittérature » dans le cas de *Twitter* [Paveau, 2012].

5.2.1/ RÔLE DES OPÉRATEURS DE *Twitter* DANS LE DISCOURS

Les premiers travaux portant sur la définition des styles de communication dans *Twitter* avaient pour objectif de déterminer les fonctions des opérateurs de communication, c'est-à-dire @, #, RT et URL, selon leur rôle dans les discours. En s'inspirant de la classification proposée par [Thimm et al., 2012b, Dang-Anh et al., 2013], le tableau 5.5 résume les rôles communicationnels de chaque opérateur :

TABLE 5.5 – Fonctions des principaux opérateurs de communication dans un *tweet*

Opérateur	**Rôle communicationnel**
@	Adresser : référence personnelle, interaction, contact
#	Indexation : contextualisation, référence à un sujet, étiquetage
URL	Hyperlien : diffusion d'information, argumentation, illustration
RT	Redistribution : diffusion, référence, citation

Sur la base de cette classification, les principaux styles de communication *Twitter* ont été déterminés en exploitant le nombre de fois où un opérateur de communication est utilisé dans les *tweets*. Le style est *personnel interactif* si le nombre de @ et RT est supérieur au nombre de # et URLs, sinon le style est dit *thématique informatif*. Cependant, cette approche quantitative a montré des limites. Ainsi, des améliorations ont été proposées afin d'affiner la classification en ajoutant d'autres critères tels que le degré d'implication dans un dialogue, la fonction de communication dominante, et l'étude de l'audience concernée par les messages. Dans l'objectif de mesurer l'interactivité des utilisateurs de *Twitter* et de catégoriser leur style de communication, Kondrashova et al. [Kondrashova et al., 2015] proposent le modèle *I to I*.

5.2.2/ LE MODÈLE *I to I*

Le modèle *I to I* regroupe deux styles extrêmes de *Twitter* : informatif et interactif, ce qui explique le nom acronyme du modèle (Informatif-to-Interactif). *I to I* permet de juger le degré d'interactivité ou de dialogicité de chaque utilisateur en fonction des éléments suivants : 1) la définition des fonctions principales des opérateurs de communication ; 2) le public ciblé et 3) le potentiel de diffusion d'un *tweet*.
L'idée est que chaque énoncé, même un *tweet*, possède, d'après Bahktine [Bakhtine, 1996], un potentiel dialogique. Si nous considérons que chaque *tweet* est initialement adressé à un public ciblé, nous devons être en mesure de mesurer le degré d'interactivité de chaque *tweet* en fonction du public ciblé, par exemple : @nom_personne semble être une partie d'un dialogue entre deux personnes, alors que @bbc cible toute la communauté des *abonnés* à la BBC. Trois types de degré d'interactivité ont été définis dans le cas de *Twitter* : informatif, interactif et équilibré en fonction du potentiel dialogique de chacun d'eux. À travers ces deux exemples, nous considérons que l'interactivité dépend de deux facteurs principaux : 1) la nature et le paramètre de l'opérateur de communication, puisque, par exemple, @+nom_personne ne révèle pas le même potentiel dialogique dans un *tweet* que @+source_média (par exemple BBC) ; et 2) les éléments voisins qui accompagnent l'opérateur de communication, le sens de @nom_personne change radicalement s'il est accompagné de RT (RT + @nom_personne). En se basant sur ces deux

facteurs, nous attribuons des rôles communicationnels dans un *tweet* à des combinaisons d'opérateurs et d'éléments voisins. Nous avons choisi environ quarante combinaisons qui semblent être pertinentes, limitées dans un premier temps aux combinaisons de deux opérateurs de communication, le tableau 5.6 montre des exemples de combinaisons d'opérateurs.

Les éléments de gratifications et les sources médias dépendent du corpus et du domaine étudiés. Par exemple, dans le domaine politique, nous pouvons considérer les termes « merci » et « félicitations » comme éléments de gratification et le terme « BBC-News » comme une source média.

TABLE 5.6 – Exemples de combinaisons d'opérateurs et d'éléments voisins

Combinaisons d'opérateurs et d'éléments	**Rôle communicationnel**
@nom_personne + #element_gratification	Interactif
@nom_personne + photo/vidéo	Interactif
@nom_personne + URL	Interactif
@nom_personne + #source_média	Équilibré
@source_média + URL	Informatif
#source_média + URL	Informatif
RT source_média + URL	Informatif

5.2.3/ CATÉGORISATION DES STYLES DE COMMUNICATION DANS *Twitter* ET RÉSULTATS

Nous présentons dans cette sous-section l'extension de *TwitBelief* afin de catégoriser les styles de communication dans *Twitter*. La méthode consiste à modifier *TwitBelief* selon deux aspects : 1) les relations ou motifs sont remplacés par les combinaisons d'opérateurs et d'éléments voisins et 2) le rôle communicationnel devient le cadre de discernement Ω_{Style} et l'opération $@$ est adaptée (voir le tableau 5.7) pour déterminer le style de communication.

TABLE 5.7 – Définition de l'opération $@_3$

$@_3$	**Interactif**	**Équilibré**	**Informatif**	Ω_{Style}
Interactif	Interactif	Interactif	Équilibré	Interactif
Équilibré	Interactif	Équilibré	Informatif	Équilibré
Informatif	Équilibré	Informatif	Informatif	Informatif
Ω_{Style}	Interactif	Équilibré	Informatif	Ω_{Style}

Ainsi, afin de catégoriser les styles de communication dans *Twitter*, nous combinons les informations sur les différents opérateurs et éléments utilisés dans les *tweets*. La méthode peut être appliquée sur un *tweet* en tenant compte des différents opérateurs de communication et des éléments utilisés dans ce *tweet* afin de déterminer son style ou il peut également catégoriser le style global d'un utilisateur en considérant toutes les combinaisons d'opérateurs communicationnels et d'éléments qu'il a utilisé dans ses *tweets*. Cependant, afin de fusionner les combinaisons d'opérateurs, il est intéressant

de représenter l'incertitude sur l'importance des combinaisons d'opérateurs les unes par rapport aux autres et donc utiliser la démarche de *TwitBelief*.

Nous avons appliqué notre méthodologie au corpus anglais constitué lors des élections européennes de 2014. Nous avons 309 candidats qui ont émis près de 73 000 tweets sur une période de quatre semaines. Pour étudier leur style de communication, nous avons en premier calculé pour chaque candidat le nombre d'occurrences de chacune des combinaisons. Le tableau 5.8 montre les résultats pour les trois candidats "DavidoOrr", "YesEdinSouth" et "LindaMcAvanMEP" et trois combinaisons qui correspondent à un des trois styles de communication.

TABLE 5.8 – Occurences pour trois candidats anglais

Combinaison d'opérateurs	DavidoOrr	YesEdinSouth	LindaMcAvanMEP
@nom_personne + photo/vidéo	8	31	39
@nom_personne + #source_média	130	0	1
#autre + #source_média	244	0	1

Le tableau 5.9 montre avec une masse de croyance le style de communication des trois candidats. Par exemple, le style de communication de "DavidoOrr" est Informatif avec une croyance de $48,25\%$. Ainsi, la catégorisation des styles de communication dans *Twitter* en utilisant le principe de *TwitBelief* est un travail complémentaire à celui de [Kondrashova et al., 2015] puisqu'il permet de déduire, de façon automatique, le style de communication global des utilisateurs du réseau, ce qui est difficile à déduire directement par les sociologues.

TABLE 5.9 – Style de communication des 3 candidats

	DavidoOrr	YesEdinSouth	LindaMcAvanMEP
Interactif	0.1943147	0.47523382331	0.43164746
Équilibré	0.323154	0.2400061633	0.25460243
Informatif	0.4825313	0.2847600133	0.31375011

Jackson et al. [Jackson et al., 2011] montrent que les politiciens préfèrent la communication « un-à-plusieurs », se rapprochant d'une simple distribution d'information (style informatif) plutôt que d'une communication interactive « un-à-un ». Ceci n'est pas confirmé avec nos résultats puisque la majorité des candidats des élections européennes de 2014 utilisent plutôt un style interactif.

5.3/ CONCLUSION

Dans ce chapitre, nous avons proposé deux extensions de *TwitBelief* afin de considérer le contenu des *tweets*. La première est l'estimation de l'influence polarisée sur le réseau *Twitter*. Dans cette extension, l'analyse des sentiments des *tweets* avec l'algorithme des forêts d'arbres décisionnels permet de déterminer la polarité de l'influence. Il s'agit de diviser le corpus en trois sous-ensembles, chacun représente une polarité. *TwitBelief*

est appliqué dans les trois sous-ensembles pour obtenir l'influence polarisée de chaque utilisateur. Enfin, les mesures d'influence des trois polarités sont combinées en utilisant le principe de *TwitBelief* mais en tenant compte des combinaisons des couples polarité/degré d'influence. La deuxième extension est la catégorisation des styles de communication dans *Twitter*, il s'agit d'étudier les styles de communication des utilisateurs de *Twitter* en se basant sur le modèle *I to I*. Cette extension utilise le principe de *TwitBelief* en modifiant deux aspects : le cadre de discernement et les combinaisons d'opérateurs qui remplacent les relations et motifs. L'annexe B donne des exemples d'illustration des deux extensions de *TwitBelief*

Ainsi, les extensions présentées dans ce chapitre montrent la flexibilité de la méthode *TwitBelief* et son pouvoir d'adaptation aux différents aspects du réseau *Twitter*. Par exemple, dans *TwitBelief*, l'algorithme se base uniquement sur l'aspect multiplexe du réseau et utilise les liens et leur diversité dans les estimations alors que dans les extensions présentées dans ce chapitre, *TwitBelief* s'adapte pour considérer l'aspect hétérogène du réseau en exploitant les *tweets* et leurs contenus.

Afin d'adapter *TwitBelief* à des problématiques différentes de son intention initiale, l'idée clé est de formaliser le problème et la question à laquelle nous souhaitons répondre puis de définir le cadre de discernement qui représente les réponses possibles à cette question. En plus du cadre de discernement, il est aussi important d'identifier les critères de manifestation de chaque problème. Par exemple, afin d'estimer l'influence, les relations et les motifs représentent les critères de manifestation de l'influence alors que les combinaisons d'opérateurs représentent les critères de manifestation des styles de *Twitter*.

6

ÉTUDE EXPÉRIMENTALE

Le but de ce chapitre est d'évaluer l'approche *TwitBelief*, la contribution principale apportée pendant cette thèse, et de la mettre en œuvre à travers un prototype d'évaluation. Les expérimentations portent sur trois jeux de données. Le premier représente les données collectées dans le cadre du projet TEE'2014, le deuxième jeu a été collecté lors de l'élection présidentielle française de 2017 et le troisième concerne l'ensemble de données CLEF RepLab 2014 qui a été conçu pour un défi d'influence organisé dans le contexte de la conférence CLEF. Nous détaillons dans la suite notre outil de collecte *SNFreezer*, puis pour chacun des trois jeux de données nous présentons les différentes expérimentations effectuées ainsi que les résultats obtenus.

6.1/ COLLECTE ET DESCRIPTION DES DONNÉES

Pour collecter les informations de *Twitter*, nous avons utilisé l'outil *SNFreezer*[1] développé par l'équipe de recherche [Leclercq et al., 2015]. Le module de collecte connecté aux *Streaming* et *Search* APIs de Twitter réalise la récupération des données d'une collecte. Pour traiter le problème de stockage des *tweets* aussi bien en termes de performance et d'interopérabilité (connexions facilitées avec des outils tiers) que d'adéquation entre les algorithmes et les structures de données, une couche de stockage de type polystore a été développée. La couche de stockage inclut une base de données relationnelles, une base de données graphe et un système de stockage de documents qui peuvent être utilisés conjointement et simultanément. À réception de *tweets* au format JSON, la couche de stockage répartit les *tweets* dans les différents systèmes en fonction des analyses prévues et éventuellement les dupliquent sur plusieurs systèmes de stockage.

Trois types d'informations, généralisées sous le terme « source », peuvent être pris en paramètre lors de la collecte : des comptes utilisateurs, des *hashtags* et des mots ou phrases. Ces différentes sources ont été choisies par les politologues, et nous retrouvons parmi elles les noms des principaux candidats, leurs comptes *Twitter*, et les *hashtags* relatifs à ces candidats, leurs partis, ou plus généralement à l'élection étudiée. L'objectif de la collecte est de capter les *tweets* mentionnant les utilisateurs désignés, ceux contenant un certain *hashtag*, mot ou phrase, ou encore les *tweets* émis par les utilisateurs spécifiés. En plus, des informations sur ces *tweets* sont collectées tels que les *tweets retweetés*, les utilisateurs *mentionnés* dans les *tweets* et les *réponses* aux *tweets*. La figure 6.1 montre le schéma relationnel représentant les données collectées. Les attributs soulignés sont les

1. https://github.com/SNFreezer

clés primaires de chaque table, et les attributs en gras les clés étrangères (le lien indique l'attribut référencé). La table *Tweet* contient les *tweets*, les *tweets retweetés* ainsi que les *réponses* c'est-à-dire tous les *tweets*. *Tweet* est connecté par des clés étrangères avec l'utilisateur qui l'a émis (table User) à travers l'attribut from_user_id, les hashtags (Tweet_Hashtag), les URLs (Tweets_URL), les symboles (Tweet_Symbol) et les sources (Tweet_Source). Cette dernière relation fait exception puisqu'elle est un exemple de table définie spécifiquement pour répondre aux besoins d'analyse (traçabilité des données collectées). Elle contient la référence au *tweet* et à la source, c'est-à-dire le critère qui a déclenché sa collecte. Les tables Retweet et Mention représentent les relations entre les utilisateurs et les *tweets*. La table Retweet contient l'utilisateur qui a retweeté, l'ID du *tweet* retweeté et la date. La table Mention lie le *tweet* avec les utilisateurs qu'ils mentionnent. La relation *favori* n'est pas obtenue pendant la collecte.

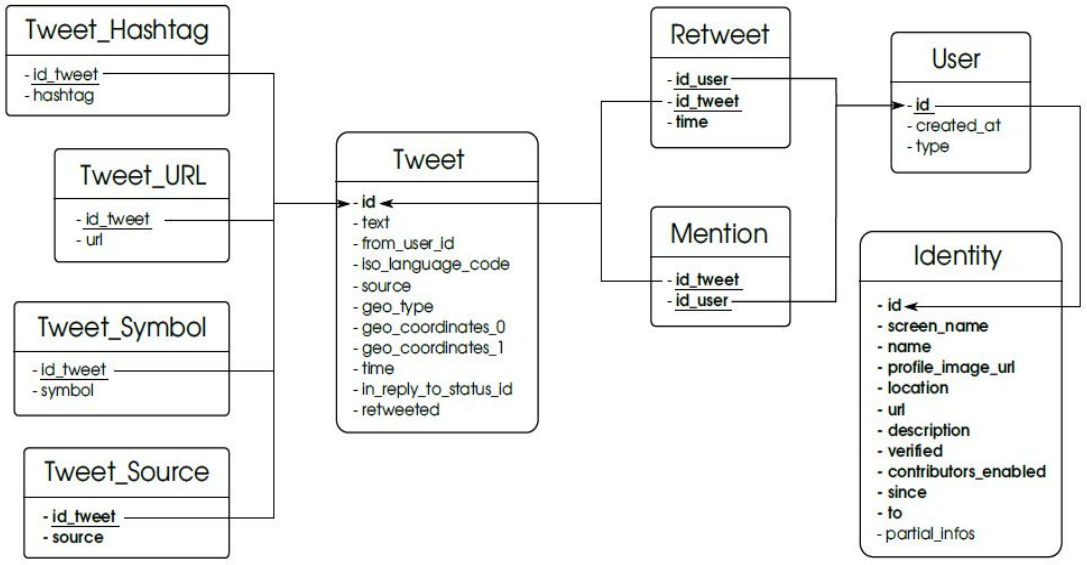


FIGURE 6.1 – Modèle relationnel représentant les données *Twitter* collectées

La base de données relationnelle est implémentée sous PostgreSQL puisque ce SGBD est stable et optimisé pour contenir de grandes quantités de données. De plus, sa manière d'optimiser les chaînes de caractères de taille variable est assez intéressante, et se prête bien aux attributs des *tweets*.

Concernant les motifs d'interaction, les seuls motifs obtenus sont le *retweet* de *réponse* et le *retweet* de *mention*. En effet, la collecte permet d'obtenir le *retweet* sur un seul niveau. De cette façon, nous ne pouvons pas tracer les cascades de *retweets* et obtenir les *retweets* de *retweets* ainsi que les *retweet* de *retweet* de *mention* par exemple. Cela est dû au fait que les APIs de *Twitter* renvoie un *retweet* (quelque soit son niveau) en se référant uniquement au *tweet* initial. Reprenons l'exemple de la relation *mention* présenté dans la figure 4.3 du chapitre 4.

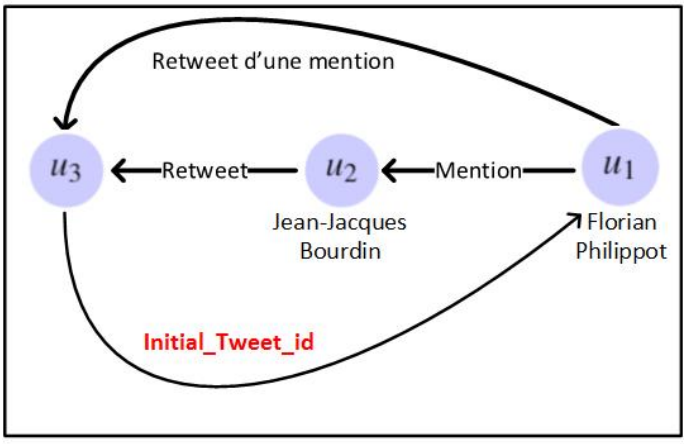


FIGURE 6.2 – Cas d'un *retweet* d'une *mention*

La figure 6.2 illustre le cas où un utilisateur U_3 *retweete* la *mention* effectué par U_2. Lors de la récupération de ce *retweet*, le champ **Initial-tweet-id** fourni par les APIs de *Twitter* se réfère au *tweet* initial contenant la *mention*. Quand un utilisateur U_4 *retweete* le *retweet* de U_3 (voir la figure 6.3), le champ **Initial-tweet-id** renvoie toujours au *tweet* intial. Ainsi, même si les APIs de *Twitter* fournissent beaucoup d'informations, nous ne pouvons pas tout collecter. Cependant, *Twitter* reste le terrain de jeu le plus favorable par rapport à d'autres réseaux sociaux puisque ses APIs permettent d'obtenir gratuitement une variété d'information assez importante.

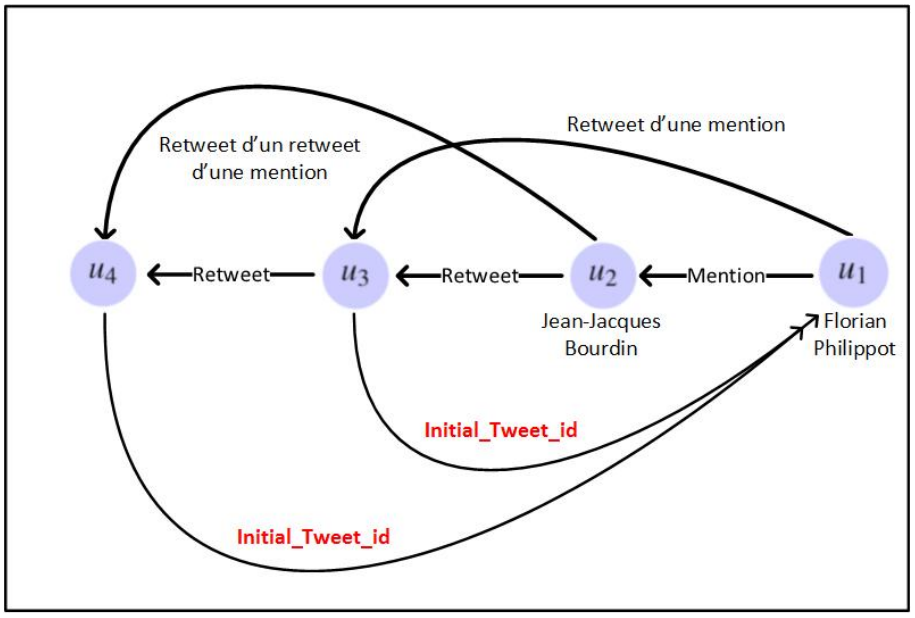


FIGURE 6.3 – Cas d'un *retweet* d'un *retweet* d'une *mention*

6.2/ CORPUS TEE'2014

Dans cette section, les travaux de recherche menés se déroulent dans le cadre du projet TEE'2014 dont l'intitulé exact est « *Twitter* aux élections européennes : une étude contrastive internationale des utilisations de *Twitter* par les candidats aux élections au Parlement Européen en mai 2014 ». Ce projet international, mené par la Maison des Sciences de l'Homme (MSH) de Dijon, a réuni près de quarante-cinq chercheurs (majoritairement des politologues, sociologues, chercheurs en communication) de dix laboratoires de recherche répartis dans six pays européens (France, Allemagne, Belgique, Italie, Espagne et Royaume-Uni). L'objectif global de ce projet était d'observer et d'analyser la communication des politiques sur *Twitter* durant les élections européennes de mai 2014 dans les six pays couverts par l'étude. Les résultats de cette étude ont été publiés dans deux ouvrages collectifs.

La collecte avec *SNFreezer* sur deux mois a fourni 50 millions de *tweets* pour un volume de 50 Go environ. Dans nos expérimentations, nous nous concentrons sur le corpus français et anglais. Le tableau 6.1 présente les paramètres des données utilisées.

TABLE 6.1 – Paramètres des données relatives aux corpus français et anglais

	Corpus français	**Corpus anglais**
Nombre de *tweets*	4 593 665	4 207 157
Nombre de comptes	937 860	743 248
Nombre de candidats	616	754
Nombre de relations	2 922 566	2 990 211
Nombre de *retweets*	639 531	1 541 294
Nombre de *mentions*	1 945 773	1 110 739
Nombre de *réponses*	337 262	338 178

6.2.1/ APPLICATION DE *TwitBelief*

L'objectif des expérimentations est d'estimer l'influence des candidats sur le réseau de *Twitter* et d'en tirer un classement. Contrairement aux illustrations données dans le chapitre 4 à la section 4.3.2.1, nous ne voulons pas estimer l'influence entre deux utilisateurs mais l'influence globale des candidats dans le réseau.

6.2.1.1/ CHOIX DES PARAMÈTRES

Les relations choisies pour représenter l'influence sont le *retweet*, la *mention* et la *réponse* et les motifs d'interactions considérés sont le *retweet* de *réponse* et le *retweet* de *mention*.

L'affectation des masses dans l'étape d'initialisation est une question importante pour traiter des données réelles. Dans certains domaines tels que la politique, les utilisateurs ont un très grand nombre de relations. Avec des masses initialisées comme dans la sous-section de l'illustration 4.3.2.1 du chapitre 4, l'influence converge rapidement vers le plus haut degré possible E.Forte après seulement 40 combinaisons de *retweets*. La figure 6.4 illustre cette convergence rapide quand la masse de croyance d'un *retweet* est définie comme $m_{\text{retweet}}(\text{Faible}) = 0.4$, $m_{\text{retweet}}(\Omega) = 0.6$. Dans [Kirgizov et al., 2016] et

l'annexe A, nous étudions en détail plusieurs questions théoriques sur la convergence de l'influence en utilisant la théorie des chaînes de Markov.

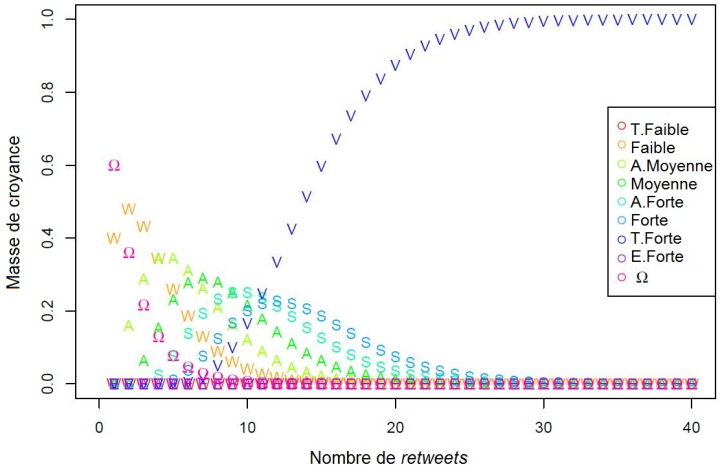


FIGURE 6.4 – Convergence de l'influence en fonction du nombre de *retweets*

De cette manière, nous ne pouvons pas comparer l'influence des candidats puisque nous obtenons le même degré d'influence avec des masses similaires pour la plupart d'entre eux. Pour faire face à cela, nous effectuons un ré-échelonnement (*rescaling*) et utilisons l'initialisation des masses suivantes :

$$\textit{Retweet} \mapsto \begin{cases} m_{\text{retweet}}(\text{T.Faible}) = 0.55 \cdot 10^{-3} \\ m_{\text{retweet}}(\Omega) = 1 - 0.55 \cdot 10^{-3} \end{cases}$$

$$\textit{Mention} \mapsto \begin{cases} m_{\text{mention}}(\text{T.Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{mention}}(\Omega) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

$$\textit{Réponse} \mapsto \begin{cases} m_{\text{réponse}}(\text{T.Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{réponse}}(\Omega) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

$$\textit{Retweet de Réponse} \mapsto \begin{cases} m_{\text{retweetDEréponse}}(\text{T.Faible}) = 0.75 \cdot 10^{-3} \\ m_{\text{retweetDEréponse}}(\Omega) = 1 - 0.75 \cdot 10^{-3} \end{cases}$$

$$\textit{Retweet de Mention} \mapsto \begin{cases} m_{\text{retweetDEmention}}(\text{T.Faible}) = 0.65 \cdot 10^{-3} \\ m_{\text{retweetDEmention}}(\Omega) = 1 - 0.65 \cdot 10^{-3} \end{cases}$$

La masse de la relation *retweet* est légèrement plus importante que les deux autres relations car nous considérons que c'est un meilleur indicateur d'influence que les autres relations puisqu'elle permet de diffuser les *tweets*. Les masses des motifs d'interaction sont aussi un peu plus importantes que celles posées sur les relations d'influence directe car nous considérons que les motifs sont un bon indicateur d'influence, cela montre que certains utilisateurs sont capables de diffuser les *tweets* à plusieurs niveaux en

exerçant une influence même sur les utilisateurs avec lesquels ils ne sont pas directement connectés.

6.2.1.2/ ESTIMATION DE L'INFLUENCE DIRECTE

Afin de montrer l'importance de la considération des motifs d'interaction dans l'estimation de l'influence, nous commençons par appliquer *TwitBelief* en se basant seulement sur les relations directes entre les utilisateurs puis nous introduisons les motifs d'interaction dans l'estimation. Pour estimer l'influence directe, nous prenons le nombre de *retweets*, *mentions* et *réponses* de chaque candidat et combinons leurs masses. Le tableau 6.2 montre les résultats pour les candidats français "Marine Le Pen", "Florian Philippot" et "Jean-Luc Mélenchon". Par exemple, nous concluons que le degré d'influence du candidat "Marine Le Pen" qui a 14 678 *retweets*, 66 798 *mentions* et 4003 *réponses*, est E.Forte avec une certitude de près de 82%. (probabilité pignistique à 0.8173448). Les résultats obtenus fournissent non seulement le degré d'influence, mais donnent également une indication de notre croyance dans les résultats grâce aux probabilités pignistiques sur les différents degrés.

TABLE 6.2 – Résultats pour trois candidats français

	Marine Le Pen	**Florian Philippot**	**Jean-Luc Mélenchon**
T.Faible	0	0.000011065	0.000030278
Faible	0	0.00007295998	0.0001832843
A.Moyenne	0	0.0007035528	0.001403947
Moyenne	0	0.003033557	0.004954501
A.Forte	0	0.008340205	0.01247841
Forte	0	0.02191526	0.02977818
T.Forte	0.1826552	0.5830090	0.7960571
E.Forte	0.8173448	0.3829144	0.1551143

La figure 6.5 offre une représentation visuelle de l'influence pour dix candidats français. Afin de contourner la complexité visuelle de l'ensemble du graphe, nous n'utilisons que 1% de toutes les données du graphe. Les nœuds de grande taille correspondent aux comptes des candidats, les autres nœuds représentent les autres utilisateurs. La taille et les couleurs des nœuds correspondent à leur degré d'influence : la couleur rouge correspond à E.Forte, l'orange à T.Forte et le jaune à Moyenne et A.Moyenne.

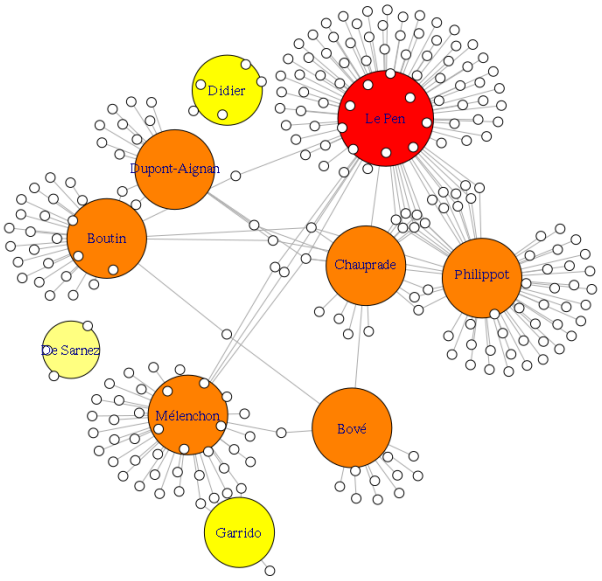


FIGURE 6.5 – Représentation de l'influence pour dix candidats français

Nous avons procédé de même sur le corpus anglais, le tableau 6.3 présente les résultats des candidats anglais "Katie Hopkins", "Nigel Farage" et "Patrick O'Flynn".

TABLE 6.3 – Résultats pour trois candidats anglais

	Katie Hopkins	**Nigel Farage**	**Patrick O'Flynn**
T.Faible	0	0	0.0022350807
Faible	0	0	0.0067817003
A.Moyenne	0	0	0.0292114368
Moyenne	0	0	0.0707907792
A.Forte	0	0	0.1140117014
Forte	0	0	0.1635190125
T.Forte	0.0260529	0.2663876	0.5804666241
E.Forte	0.9739471	0.7336124	0.0329836651

6.2.1.3/ CLASSEMENT DES UTILISATEURS

Dans cette section, l'objectif expérimental est de classer les candidats selon leur influence dans le réseau en se basant sur *TwitBelief*. Comme décrit dans la section 4.3.3, nous prenons d'abord, pour chaque candidat, l'influence avec la probabilité pignistique maximale (par exemple, Inf('Marine Le Pen" = E.Forte). Après, nous classons les candidats selon leur « degré d'influence maximal ». Lorsque deux candidats ont le même « degré d'influence maximal » alors nous comparons les probabilités pignistiques sur le plus haut degré d'influence suivant. Par exemple,

Inf(“Florian Philippot”) $= Inf$(“Jean-Luc Mélenchon”) = T.Forte et $\text{bet}_{\text{Philippot}}(\text{E.Forte}) > \text{bet}_{\text{Mélenchon}}(\text{E.Forte})$

Comme expliqué dans la section 4.3.3, nous procédons de cette façon car il est injuste de classer les utilisateurs selon les probabilités pignistiques maximales qu'ils ont sur les degrés. Par exemple, le candidat “Florian Philippot” a une probabilité pignistique sur le degré T.Forte plus faible que celle de “Jean-Luc Mélenchon” sur le même degré comme nous pouvons le voir dans le tableau 6.2. Malgré cela, il est classé avant “Jean-Luc Mélenchon” (voir le tableau 6.4) car il a une plus grande probabilité pignistique sur le degré E.Forte. Nous déterminons ainsi le classement de tous les candidats. Les résultats sont présentés dans les tableaux 6.4 et 6.5.

TABLE 6.4 – Classement des candidats français les plus influents

Rang	Candidats	Degré d'influence	Probabilité pignistique
1	Marine Le Pen	E.Forte	0.8173448
2	Florian Philippot	T.Forte	0.5830090
3	Jean-Luc Mélenchon	T.Forte	0.7960571
4	Christine Boutin	T.Forte	0.9796956
5	Aymeric Chauprade	T.Forte	0.4171324655
6	Nicolas Dupont-Aignan	T.Forte	0.5293170700
7	José Bové	T.Forte	0.2925722297
8	Geoffroy Didier	Moyenne	0.2092645352
9	Raquel Garrido	Moyenne	0.2048485
10	Marielle De Sarnez	A.Moyenne	0.2074260

Nous avons procédé de même pour les candidats anglais :

TABLE 6.5 – Classement des candidats anglais les plus influents

Rang	Candidats	Degré d'influence	Probabilité pignistique
1	Katie Hopkins	E.Forte	0.9739471
2	Nigel Farage	E.Forte	0.7336124
3	Patrick O'Flynn	T.Forte	0.5804666241
4	Daniel Hannan	T.Forte	0.423098905
5	Roger Helmer	T.Forte	0.427644003
6	Nick Griffin	T.Forte	0.449170270
7	Marcus Chown	T.Forte	0.420214325
8	Mickael Heaver	T.Forte	0.421310071
9	Rufus Hound	T.Forte	0.212697476
10	Janice Atkinson	Moyenne	0.2116610

6.2.1.4/ PRISE EN COMPTE DE L'INFLUENCE INDIRECTE

Dans les résultats précédents, nous avons seulement considéré l'influence directe des candidats. Notre approche est flexible et peut être étendue à l'influence indirecte en réalisant une combinaison de croyances sur plusieurs niveaux grâce à l'utilisation des motif d'interaction. Afin de prendre en compte l'influence indirecte, nous estimons l'influence indirecte, puis nous combinons les résultats avec ceux obtenus par l'influence directe.

Les tableaux 6.6 et 6.7 représentent les résultats en tenant compte de l'influence indirecte pour les candidats français et anglais. Ils montrent que, pour les candidats considérés, l'influence est devenue plus importante après avoir considéré l'influence indirecte. Par exemple, pour le candidat "Marine Le Pen", nous avons trouvé qu'elle a 4003 *retweets de réponses* et 37 715 *retweets de mentions* , le degré d'influence obtenu après fusion à plusieurs niveaux est toujours le même degré E.Forte mais la probabilité pignistique est devenue plus importante et a atteint 0.99. En outre, nous constatons un cas de certitude totale pour le candidat "Katie Hopkins" qui a obtenu une influence E.Forte avec la probabilité pignistique 1.

TABLE 6.6 – Résultats pour trois candidats français en considérant leur influence indirecte

	Marine Le Pen	**Christine Boutin**	**Jean-Luc Mélenchon**
T.Faible	0	0	0
Faible	0	0	0
A.Moyenne	0	0	0.00001235114
Moyenne	0	0	0.00009165825
A.Forte	0	0	0.0003897511
Forte	0	0	0.001821524
T.Forte	0.004328	0.0667283	$0,345777616$
E.Forte	0.995672	0.9332717	0.6519071

TABLE 6.7 – Résultats pour trois candidats anglais en considérant leur influence indirecte

	Katie Hopkins	**Nigel Farage**	**Patrick O'Flynn**
T.Faible	0	0	0
Faible	0	0	0
A.Moyenne	0	0	0
Moyenne	0	0	0
A.Forte	0	0	0
Forte	0	0	0
T.Forte	0	0.0367487	0.8519071
E.Forte	1	0.9632513	0.1480929

Nous avons également classé les candidats français en considérant leur influence indirecte.

Le tableau 6.8 montre ce nouveau classement. Par rapport au classement donné dans le tableau 6.4, le classement de certains candidats a changé. Par exemple, le candidat “Christine Boutin” est devenu le deuxième candidat le plus influent puisque son degré d’influence a augmenté pour devenir E.Forte avec la probabilité pignistique de 0.93, ce qui prouve l’importance de l’influence indirecte.

TABLE 6.8 – Classement des candidats français en prenant en compte l’influence indirecte

Rang	Candidats	Degré d’influence	Probabilité pignistique
1	Marine Le Pen	E.Forte	0.995672
2	Christine Boutin	E.Forte	0.9332717
3	Jean-Luc Mélenchon	E.Forte	0.6519071
4	Florian Philippot	E.Forte	0.6021768
5	Aymeric Chauprade	T.Forte	0.66411504
6	Nicolas Dupont-Aignan	T.Forte	0.6615197
7	José Bové	T.Forte	0.6003759
8	Raquel Garrido	T.Forte	0.4805315962
9	Geoffroy Didier	T.Forte	0.416560612
10	Marielle De Sarnez	Forte	0.789653

6.2.2/ COMPARAISON AVEC LES TRAVAUX EXISTANTS

Le tableau 6.9 présente le classement obtenu en utilisant les critères utilisés par [Cha et al., 2009]. Ces critères sont le nombre de *retweets, mentions et réponses*. La dernière colonne montre le classement des candidats selon leur degré de centralité calculé en utilisant le nombre de voisins de chaque candidat dans le réseau c’est-à-dire le nombre total de *retweets, mentions et réponses*. Ce degré de centralité permet, pour chaque candidat, de prendre en compte l’ensemble des relations qu’ils possèdent avec ses voisins comme nous le faisons dans l’estimation de l’influence directe avec *TwitBelief*. Mais le classement obtenu ne comporte aucune indication sur le degré d’influence et la certitude dans les résultats contrairement à nos résultats présentés dans le tableau 6.8. Nous voyons que sur dix candidats neuf sont communs avec un ordre un peu différent, seul le dixième est différent.

Le tableau 6.10 présente le classement des candidats selon l’algorithme du HITS décrit dans la section 2.3.2.3 du chapitre 2. Afin d’estimer l’influence des candidat, nous nous basons sur le score d’autorité du HITS en utilisant les relations *réponse, retweet* et *mention*. Les résultats présentés ne montrent pas l’influence globale dans le réseau puisque nous trouvons différents classements pour chaque type de relation. Alors que notre méthode (Tableau 6.8) nous permet d’avoir un classement unique qui tient compte de toutes les relations considérées.

TABLE 6.9 – Candidats français les plus influents selon les différentes relations et le degré de centralité

Rang	*Retweet*	*Mention*	*Réponse*	*Degré de centralité*
1	Marine Le Pen	Marine Le Pen	Christine Boutin	Marine Le Pen
2	Florian Philippot	Christine Boutin	Marine Le Pen	Christine Boutin
3	J.L Mélenchon	J.L Mélenchon	Florian Philippot	Florian Philippot
4	Aymeric Chauparde	Florian Philippot	J.L Mélenchon	J.L Mélenchon
5	François Asselineau	N. Dupont-Aignan	L. de Gouyon Matigon	N. Dupont-Aignan
6	C. Morel-Darleux	José Bové	N. Dupont-Aignan	Aymeric Chauparde
7	N. Dupont-Aignan	Aymeric Chauparde	Jean-Sébastien Herpin	José Bové
8	Louis Aliot	Raquel Garrido	Julien Rochedy	Geoffroy Didier
9	Denis Payre	Jérome Lavrilleux	Geoffroy Didier	Raquel Garrido
10	Yannick Jadot	Marielle de Sarnez	Louis Aliot	Yannick Jadot

TABLE 6.10 – Candidats français les plus influents selon l'algorithme du HITS

Rang	*HITS-Réponse*	*HITS-Retweet*	*HITS-Mention*
1	Marine Le Pen	Marine Le Pen	Marine Le Pen
2	Christine Boutin	Aymeric Chauprade	Aymeric Chauparde
3	Florian Filippot	Bernard Monot	Florian Philippot
4	Jean-Luc Mélenchon	Florian Philippot	Jean-Marie Le Pen
5	Nicolas Dupont-Aignan	Nicolas Bay	Louis Aliot
6	Aymeric Chauparde	Bruno Gollnisch	Bernard Monot
7	José Bové	Audrey Guibert	Geoffroy Didier
8	Geoffroy Didier	Gilles Lebreton	Julien Rochedy
9	Raquel Garrido	Jean-Marie Le Pen	Gilles Lebreton
10	Yannick Jadot	Karim Ouchikh	Bruno Gollnisch

6.2.3/ DISCUSSION

Les expériences sur l'ensemble du corpus TEE'2014 montrent que l'utilisation de *TwitBelief* conduit à des résultats intéressants, la méthode prend en compte différentes relations et motifs d'interaction et fournit une estimation d'influence globale dans le réseau par un degré d'influence associé à une masse de croyance. L'approche proposée est également flexible car elle tient compte du réseau dans sa globalité en utilisant des motifs d'interaction ce qui nous permet d'estimer l'influence indirecte d'un nœud dans le réseau. L'influence indirecte donne des résultats différents de l'estimation de l'influence directe. Ces résultats ont été appréciés par les sociologues et les spécialistes politiques du projet TEE'2014 et l'influence indirecte a été jugée plus pertinente que l'influence directe.

6.3/ CORPUS DE L'ÉLECTION PRÉSIDENTIELLE FRANÇAISE DE 2017

6.3.1/ DESCRIPTION DES DONNÉES

Les données de l'élection présidentielle française de 2017 ont été collectées de la même manière que les données du corpus TEE'2014 en utilisant *SNFreezer*. La collecte s'est effectuée sur une période de plus de trois mois découpée en trois phases : avant le premier tour de l'élection du 27 février 2017 jusqu'au 23 avril 2017 ; entre les deux tours du 24 avril 2017 au 7 mai 2017 inclus et ensuite après le 7 mai jusqu'au 2 juillet 2017. Le tableau 6.11 présente les paramètres des données collectées.

TABLE 6.11 – Paramètres des données relatives aux corpus de l'élection présidentielle française de 2017

Nombre de *tweets*	58 994 106
Nombre de comptes	2 051 981
Nombre de candidats	11
Nombre de relations	72 120 974
Nombre de *retweets*	43 521 835
Nombre de *mentions*	22 362 822
Nombre de *réponses*	6 236 317

6.3.2/ APPLICATION DE *TwitBelief* ET DISCUSSION

Dans nos expérimentations, nous nous concentrons sur les données des deux premières périodes de collecte relatives aux premier et deuxième tours de l'élection. L'objectif est d'estimer l'influence des candidats à l'élection présidentielle française de 2017 sur le réseau de *Twitter* puis de déduire leur classement selon leur influence. Les relations et les motifs d'interactions choisis pour représenter l'influence sont les mêmes que ceux choisies pour le corpus TEE'2014 : *retweet*, *mention*, *réponse*, *retweet* de *réponse* et *retweet* de *mention*. Mais nous avons diminué les masses accordées à ces relations car les utilisateurs ont un plus grand nombre de relations par rapport aux données du corpus TEE'2014. Ainsi, nous utilisons les masses suivantes :

$$\textit{Retweet} \mapsto \begin{cases} m_{\text{retweet}}(\text{T.Faible}) = 0.5 \cdot 10^{-4} \\ m_{\text{retweet}}(\Omega) = 1 - 0.5 \cdot 10^{-4} \end{cases}$$

$$\textit{Mention} \mapsto \begin{cases} m_{\text{mention}}(\text{T.Faible}) = 0.4 \cdot 10^{-4} \\ m_{\text{mention}}(\Omega) = 1 - 0.4 \cdot 10^{-4} \end{cases}$$

$$\textit{Réponse} \mapsto \begin{cases} m_{\text{réponse}}(\text{T.Faible}) = 0.4 \cdot 10^{-4} \\ m_{\text{réponse}}(\Omega) = 1 - 0.4 \cdot 10^{-4} \end{cases}$$

$$\textit{Retweet de Réponse} \mapsto \begin{cases} m_{\text{retweetDEréponse}}(\text{T.Faible}) = 0.6 \cdot 10^{-4} \\ m_{\text{retweetDEréponse}}(\Omega) = 1 - 0.4 \cdot 10^{-4} \end{cases}$$

$$\textit{Retweet de Mention} \mapsto \begin{cases} m_{\text{retweetDEmention}}(\text{T.Faible}) = 0.5 \cdot 10^{-4} \\ m_{\text{retweetDEmention}}(\Omega) = 1 - 0.5 \cdot 10^{-3} \end{cases}$$

Les résultats de classement du premier et deuxième tour sont présentés dans les tableaux 6.12 et 6.13. Les classements selon les différentes relations sont présentés dans le tableau 6.14, ce classement est difficile à interpréter car il diffère d'une relation à une autre. Cependant, *TwitBelief* permet d'obtenir une estimation d'influence en tenant compte des relations et motifs possibles les dans la même mesure.

TABLE 6.12 – Classement des candidats à l'élection présidentielle française de 2017 les plus influents avant le premier tour

Rang	Candidats	Degré d'influence	Probabilité pignistique
1	François Fillon	E.Forte	0.7831437
2	Emmanuel Macron	T.Forte	0.9302146
3	Marine Le Pen	T.Forte	0.9179830
4	Jean-Luc Mélenchon	T.Forte	0.8916434
5	Benoît Hamon	T.Forte	0.8785880
6	Philippe Poutou	A.Moyenne	0.2345158
7	François Asselineau	A.Moyenne	0.2118752
8	Nicolas Dupont-Aignan	Faible	0.2629045
9	Jean Lassalle	Faible	0.2468728
10	Jacques Cheminade	T.Faible	0.3090487
11	Nathalie Arthaud	T.Faible	0.2237506

Nous présentons les résultats pour les deux candidats qualifiés au second tour de la présidentielle.

TABLE 6.13 – Classement des candidats à l'élection présidentielle française de 2017 les plus influents entre les deux tours

Rang	Candidats	Degré d'influence	Probabilité pignistique
1	Emmanuel Macron	E.Forte	0.4426158
2	Marine Le Pen	T.Forte	0.9464662

TABLE 6.14 – Classement des candidats aux élections françaises 2017 selon les différentes relations avant le premier tour

Rang	***Retweet***	***Mention***	***Réponse***
1	Jean-Luc Mélenchon	Marine Le Pen	François Fillon
2	François Fillon	François Fillon	Emmanuel Macron
3	Marine Le Pen	Emmanuel Macron	Marine Le Pen
4	Benoît Hamon	Benoît Hamon	Benoît Hamon
5	Emmanuel Macron	Jean-Luc Mélenchon	Jean-Luc Mélenchon
6	François Asselineau	François Asselineau	Philippe Poutou
7	Philippe Poutou	Nicolas Dupont-Aignan	Nicolas Dupont-Aignan
8	Jean Lassalle	Nathalie Arthaud	François Asselineau
9	Nicolas Dupont-Aignan	Philippe Poutou	Jean Lassalle
10	Jacques Cheminade	Jean Lassalle	Nathalie Arthaud
11	Nathalie Arthaud	Jacques Cheminade	Jacques Cheminade

Les expérimentations sur les données à l'élection présidentielle française de 2017 permettent d'obtenir des classements avec lesquels nous pouvons avoir une idée sur les tendances des internautes les dernières semaines qui précèdent les élections. Nous constatons à travers le classement obtenu pour le premier tour des élections qu'il y a deux groupes de candidats, ceux avec une forte influence (top 5 des candidats) et d'autres avec une influence plus ou moins faible. De plus, les degrés d'influence ainsi que les probabilités pignistiques des candidats de chacun de ces groupes sont très rapprochés. L'écart important entre ces deux groupes reflète le résultat du premier tour de la présidentielle.

6.4/ CORPUS REPLAB 2014

L'ensemble de données CLEF RepLab 2014 a été conçu pour un défi d'influence organisé dans le contexte de la conférence CLEF (Conference and Labs of the Evaluation Forum) [2]. Dans cette section, nous utilisons cet ensemble de données pour nos propres expériences. L'objectif est de comparer *TwitBelief* avec une mesure de F-score à des méthodes d'estimation d'influence qui ont utilisé les données REPLAB 2014 pour leurs expérimentations [Ramírez-de-la Rosa et al., 2014, Cossu et al., 2015, Cossu et al., 2016].

6.4.1/ DESCRIPTION DES DONNÉES

L'ensemble de données RepLab contient des utilisateurs étiquetés manuellement par des spécialistes de Llorente & Cuenca [3], une entreprise espagnole leader dans le domaine de l'e-réputation. Ces utilisateurs ont été annotés en fonction de leur influence réelle perçue, et non en considérant spécifiquement leurs comptes *Twitter*. L'annotation est binaire : un utilisateur est influent ou non influent. L'ensemble de données contient un ensemble de données de formation (*training set*) de 2500 utilisateurs, dont 796 utilisateurs influents, et un ensemble de données de test de 5900 utilisateurs, dont 1563 utilisateurs influents.

2. http://clef2014.clef-initiative.eu/
3. http://www.llorenteycuenca.com/

Il contient également les 600 derniers identifiants des *tweets* de chaque utilisateur au moment de l'analyse. Ces *tweets* peuvent être écrits en anglais ou en espagnol. L'ensemble de données est accessible au public [4].

Afin d'évaluer notre approche sur l'ensemble de données RepLab, nous devons choisir les relations qui seront prises en compte dans l'estimation, mais l'ensemble de données RepLab ne fournit pas ce type d'information. Nous avons seulement les noms des utilisateurs avec pour chacun d'entre eux 600 identifiants de leurs *tweets*. Donc, nous devions d'abord rassembler les informations nécessaires sur ces *tweets*, pour cela, nous avons utilisé Twurl [5], un outil qui permet de collecter des informations sur les données des *tweets* à partir de l'API *Twitter*. Cependant, l'API *Twitter* limite la collecte à 180 *tweets* par 15 minutes. Il a fallu donc 174 jours pour collecter les informations sur tous les *tweets*. Les relations que nous pouvons extraire des informations collectées sont : le *retweets*, le *favori* et l'*abonnement*. Le tableau 6.15 montre les paramètres de l'ensemble de données utilisées.

TABLE 6.15 – Paramètres du jeu de données RepLab 2014

Nombre de *tweets*	5 040 000
Nombre de comptes	8 400
Nombre de liens	78 997 702
Nombre de *retweets*	20 922 694
Nombres de *favoris*	7 441 174
Nombre d'*abonnés*	50 633 834

6.4.2/ APPLICATION DE *TwitBelief* ET DISCUSSION

Nous présentons maintenant nos expériences sur le corpus RepLab. L'initialisation des masses est présentée comme suit :

$$\textit{Retweet} \mapsto \begin{cases} m_{\text{retweet}}(\text{T.Faible}) = 0.55 \cdot 10^{-3} \\ m_{\text{retweet}}(\Omega) = 1 - 0.55 \cdot 10^{-3} \end{cases}$$

$$\textit{Favori} \mapsto \begin{cases} m_{\text{favori}}(\text{T.Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{favori}}(\Omega) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

$$\textit{Abonné} \mapsto \begin{cases} m_{\text{abonné}}(\text{T.Faible}) = 0.2 \cdot 10^{-3} \\ m_{\text{abonné}}(\Omega) = 1 - 0.2 \cdot 10^{-3} \end{cases}$$

La masse la plus importante est donnée à la relation *retweet* car nous croyons que cette relation est un bon indicateur d'influence. Comme l'a montré [Cha et al., 2010], la relation *abonné* est un indicateur de popularité, or les utilisateurs populaires ne sont pas nécessairement les plus influents. Nous ne lui accordons donc pas une masse de croyance très importante. La relation *Favori* est un indicateur d'influence assez important

4. http://nlp.uned.es/replab2014/
5. https://github.com/twitter/twurl

aussi, c'est pour cela que nous lui avons attribué une masse de croyance plus importante que celle de la relation *Abonné*.

Un des défis de RepLab peut être vue comme un problème de classification binaire, consistant à décider si un utilisateur est influent ou non. Dans notre approche, les degrés d'influence sont représentés par huit classes (allant de T.Faible jusqu'à E.Forte), ainsi pour pouvoir comparer notre approche avec l'influence réelle donnée par Replab, nous considérons que les utilisateurs ayant des degrés d'influence en dessous de T.Forte sont non influents, et ceux ayant le degré T.Forte et plus sont considérés comme influents.

Lorsque nous avons appliqué *TwitBelief* sur l'ensemble de données Replab, nous avons constaté que notre approche détecte 1184 utilisateurs parmi les 1563 utilisateurs influents (75,75%) ainsi que 3416 parmi les 4337 utilisateurs non influents (78,76%). Le tableau 6.16 présente les valeurs de la précision, le rappel et le F-score (voir l'équation 5.1 dans le chapitre 5) . Ces résultats sont assez satisfaisants et montrent que *TwitBelief* est précis et performant.

TABLE 6.16 – Résultats de F-score de *TwitBelief sur les données Replab*

Précision	**Rappel**	**F-score**
.739	.770	.754

Le tableau 6.17 montre la comparaison des résultats de F-mesure entre *TwitBelief* et des recherches qui utilisent les données REPLAB 2014. Afin d'étudier le niveau d'influence des utilisateurs, [Ramírez-de-la Rosa et al., 2014] utilisent un ensemble de critères tels que l'âge du compte *Twitter*, le contenu des *tweets* (medias, urls, auto-*mentions*), le nombre de *retweets*, etc. Les résultats indiquent qu'il est possible d'identifier automatiquement les utilisateurs influents et d'obtenir des résultats intéressants (F-score $0,694$). [Cossu et al., 2015] obtiennent une meilleure mesure de F-score $(0,781)$ en exploitant le contenu des *tweets* et les données externes tels que les recherches du compte sur le Web. Partant de l'hypothèse que les utilisateurs influents ont tendance à utiliser des termes spécifiques dans leurs *tweets*, [Cossu et al., 2016] obtiennent la meilleure mesure de F-score $(0,792)$ en utilisant l'aspect lexical du contenu des *tweets* et se concentrent sur les occurrences des termes dans les *tweets*. Avec la valeur de F-score de $0,754$ *TwitBelief*, se place en 3ème position et est légèrement inférieur aux deux premières mesures obtenues à partir des méthodes de l'état de l'art.

TABLE 6.17 – Tableau de comparaison des résultats de F-score

	F-score
Cossu *et al.* [Cossu et al., 2016]	.792
Cossu *et al.* [Cossu et al., 2015]	.781
TwitBelief	.754
Ramírez *et al.* [Ramírez-de-la Rosa et al., 2014]	.694

Ce résultat peut s'expliquer par différentes raisons, tout d'abord, lors de la phase d'enrichissement des données Replab, nous avons constaté que 36% des *tweets* ont été supprimés, ainsi, les informations sur ces *tweets* n'étant plus disponibles cela peut biaiser

les résultats. Un aspect positif de nos résultats est que les utilisateurs ayant la majorité ou tous leurs *tweets* supprimés ont été assignés à la classe Ω, ce qui signifie que notre approche est capable d'exprimer correctement son ignorance sur le degré d'influence des utilisateurs sur lesquels il n'a pas assez d'informations. Une autre raison pour laquelle le F-score de *TwitBelief* est inférieur aux deux autres est que nous ne considérons pas le domaine étudié, en effet notre approche détecte les utilisateurs influents dans le réseau quel que soit le domaine étudié, beaucoup d'utilisateurs considérés comme influents par notre approche ne le sont pas par Replab car l'influence est déterminée par rapport à deux domaines (Banque et Automobile). Lorsque nous avons étudié manuellement ces comptes, nous avons constaté que ces comptes semblent être influents dans la vie réelle mais ne sont pas influents dans les domaines étudiés.

6.5/ CONCLUSION

Dans ce chapitre, nous avons expérimenté *TwitBelief* sur des données réelles collectées sur le réseau *Twitter* dans les contextes du projet TEE'2014 et de la présidentielle française de 2017 ainsi que pour le challenge Replab 2014. Les expériences montrent que la combinaison de relations sous incertitude conduit à des résultats assez intéressants. Dans le premier jeu de données TEE'2014, les résultats ont été validés par les sociologues du projet. Concernant les résultats sur la présidentielle de 2017, les résultats sont corrélés avec le résultat du vote. Quand au troisième jeu de données, le F-score $(0,754)$ montre que la classification à travers *TwitBelief* est intéressante.

Des perspectives intéressantes émergent pour renforcer davantage *TwitBelief*. Nous les détaillons dans le chapitre suivant.

III

CONCLUSION

7

CONCLUSION GÉNÉRALE

Pour conclure ce manuscrit, nous présenterons un bilan des travaux que nous avons présentés puis nous terminerons par une description des perspectives de recherche.

7.1/ BILAN

Cette thèse s'inscrit dans le domaine de l'analyse de données issues des réseaux sociaux. Notre principale contribution est d'estimer l'influence des utilisateurs dans les réseaux sociaux et en particulier le réseau *Twitter*. Le choix de *Twitter* est motivé par le fait qu'il est considéré comme la plateforme de micro-blogging numéro un dans le monde entier et qu'il offre des APIs permettant de collecter gratuitement les données pour nos différentes expérimentations. Toutes les expérimentations ont été menées dans le cadre de projets impliquant des chercheurs en science de la communication.

L'une des caractéristiques de *Twitter* est la diffusion d'information par l'utilisation d'opérateurs, tweeter, mentionner ou citer un utilisateur, utiliser un hashtag ou une URL par exemple. Les liens entre les utilisateurs déterminent le flux de l'information et conditionnent ainsi l'influence d'un utilisateur sur un autre. Certains utilisateurs, appelés influents, sont plus capables que d'autres de diffuser des informations à un grand nombre d'utilisateurs, d'influencer et de persuader les utilisateurs avec lesquels ils sont connectés. Par de la diffusion d'information à large échelle et à faible coût ou a contrario pour stopper les fake news.

Ainsi, l'étude de l'influence des utilisateurs sur *Twitter* est devenue un sujet de recherche de premier ordre pour les chercheurs en sciences de la communication. Les problématiques abordées dans cette thèse correspondent à deux orientations prioritaires. La première orientation est la **modélisation des données issues des réseaux sociaux** et en particulier *Twitter* et la deuxième orientation est l'**estimation de l'influence**. Les contributions ont été développées dans les chapitres 3, 4 et 5 et les expérimentations sont décrites dans le chapitre 6.

La plupart des travaux sur les réseaux complexes ont été abordés sous l'angle de la théorie ou de l'algorithmique des graphes. En revanche, très peu de travaux de modélisation de tels réseaux complexes existent. Notre contribution, dans le chapitre 3 a consisté, dans un premier temps, à déterminer le modèle le plus adapté et à le spécialiser pour spécifier les relations engendrées par les interactions entre les utilisateurs de *Twitter*. Le modèle choisi est un réseau multiplexe hétérogène où chaque couche représente une

relation (écrire, mention, retweet, etc.) entre utilisateurs ou entre utilisateurs et *tweets* ou encore entre les *tweets* et les objets (URLs, hahstags), . . . , les nœuds étant hétérogènes à l'intérieur de chaque couche. Dans un second temps, le réseau multiplexe hétérogène est exploité afin d'estimer l'influence des nœuds en utilisant une extension de l'algorithme PageRank, nommée Multiplex PageRank. Nous étudions aussi les paramètres qui modifie le comportement du Multiplex PageRank. Si le classement des nœuds obtenu reflète la réalité, les scores du PageRank multiplexe sne permettent qu'un classemnt des utilisateurs.

Afin de dépasser une mesure quantitative, nous proposons dans le chapitre 4, *TwitBelief*, une approche qui exploite aussi les différentes relations entre les nœuds du réseau. Elle utilise la théorie des fonctions de croyance afin de déterminer, pour chaque nœud un degré d'influence pondéré par une estimation de la crédibilité. En effet, la théorie des fonctions de croyance permet de combiner les différentes interactions du réseau *Twitter* tout en exprimant l'incertitude sur l'importance des différentes interactions à travers les masses de croyance. *TwitBelief* répond au principal inconvénient des approches présentées dans le chapitre 2 où nous avons montré que l'estimation de l'influence se focalise sur un seul type de relation négligeant les autres.

Dans le chapitre 5, nous proposons deux extensions de *TwitBelief*. La première extension permet d'exprimer si l'influence est positive ou négative en analysant le sentiment exprimé dans les *tweets*. Nous exploitons le contenu des *tweets* afin de déterminer le sentiment exprimé à travers eux en utilisant l'algorithme des forêts d'arbres décisionnels. Ensuite, nous divisons le réseau de *Twitter* en trois sous-réseaux, chacun représentant une polarité (positif, négatif ou neutre), *TwitBelief* est alors appliqué dans chaque sous-réseau pour obtenir l'influence de chacun. Enfin, les mesures d'influence des trois sous-réseaux sont combinées pour obtenir une estimation de l'influence polarisée. La seconde extension porte sur le style de communication utilisé par les utilisateurs de *Twitter*. Il s'agit de catégoriser leur style de communication en se basant sur le principe de *TwitBelief* en exploitant les différentes combinaisons d'opérateurs utilisés dans les *tweets*. Les chercheurs en sciences de la communication avec lesquels nous travaillons ont établi les différentes combinaisons à prendre en compte.

Finalement, nous avons effectué des expérimentations présentées en chapitre 6. Les expérimentations portent sur trois jeux de données. Le premier jeu représente les données du projet TEE'2014 relatif aux élections européennes de 2014, le deuxième concerne les données collectées dans le cadre de l'élection présidentielle française de 2017 et le troisième est l'ensemble de données CLEF RepLab 2014 qui a été conçu pour un défi d'influence organisé dans le contexte de la conférence CLEF. Chaque contribution est implémentée, les données CLEF RepLab 2014 permettent de nous positionner et les expériences ont conduit à des résultats significatifs.

7.2/ PERSPECTIVES

Dans cette thèse, nous avons répondu aux limites des approches existantes et nous avons obtenu de bons résultats en comparaison ceux fournis par les recherches existantes. Cependant, des perspectives intéressantes émergent pour renforcer davantage nos propositions. Dans la suite, nous présentons quelques perspectives pour nos travaux futurs :

— L'élection européenne de mai 2019 nous offre la possibilité de reconduire notre approche afin de mener une étude comparative et longitudinale [1] 2014-2019. Les participants aux projet TEE'2014 ont accepté de participer à cette nouvelle étude.

— L'approche *TwitBelief* doit être enrichie avec l'introduction de nouveaux motifs d'interaction contenant les *hashtags*. Certains *hashtags* bien choisis peuvent offrir une meilleure visibilité.
Puis nous voulons étendre l'algorithme du PageRank multiplexe avec les *hashtags* afin de proposer un Hashtag-Sensitive Multiplex PageRank. Pour cela, nous voulons exploiter l'influence de la couche des *hashtags* (graphe de *hashtags* corrélés) sur les autres couches.

— Une autre piste de recherche consiste à étudier l'influence au sein des communautés ou entre communautés. Une communauté est définie comme un ensemble d'utilisateurs ou de nœuds connectés les uns aux autres plus que les autres utilisateurs d'autres communautés [Schaub et al., 2017]. Les personnes d'une même communauté ont généralement des propriétés communes. Par exemple, ils peuvent être des amis qui ont fréquenté la même école ou sont originaires de la même ville. Une telle étude peut être utile dans des applications de marketing. En effet, l'estimation de l'influence à l'échelle de la communauté permet d'identifier les personnes qui ont une influence sur les clients potentiels. Ainsi, les activités marketing sont orientées autour de ces personnes influentes plutôt que sur le marché cible dans son ensemble. De plus, le temps passé à identifier les personnes influentes est réduit puisqu'il est évident que la communauté est plus petite que le réseau social. Il est aussi intéressant de connaître les personnes d'une communauté qui peuvent influencer les personnes appartenant à une autre communauté, ces personnes jouant un rôle charnière dans la diffusion d'information d'une communauté vers une autre.

— Une autre perspective intéressante consiste à transposer nos contributions à d'autres réseaux sociaux tels que *Facebook, Instagram, forums*, etc. En effet, chaque réseau social présente des caractéristiques spécifiques qui le distinguent des autres. Par exemple, *Facebook* permet à ses utilisateurs d'exprimer leurs sentiments ou réactions par rapport à un message donné. Par exemple, un utilisateur peut aimer un message ou le trouver amusant. Toutes ces spécificités peuvent être très instructives pour l'étude de l'influence.

— En outre, nous souhaitons adapter la méthode pour prendre en compte les aspects temporels, à savoir l'évolution de l'influence plutôt que l'influence à un instant donné. En effet, les réseaux sociaux en ligne collectent chaque jour une énorme quantité de données pouvant contenir de nombreuses informations nouvelles. Un utilisateur détecté influent aujourd'hui, peut ne plus l'être après une période de temps ou voir la polarité de son influence changer, et il peut apparaître d'autres utilisateurs influents. D'autre part, le processus d'estimation de l'influence peut être coûteux et lent dans certains cas. Par conséquent, une approche de mise à jour pour la détection des utilisateurs influents peut être très intéressante. Son objectif est de mettre à jour l'ensemble des résultats en

1. En sociologie, l'analyse longitudinale est l'analyse des données portant sur des échantillons de personnes interrogées plusieurs fois au cours du temps.

ajoutant de nouveaux utilisateurs influents et/ou en supprimant ceux qui ne le sont plus. La mise à jour se fera sans relancer tout le processus d'estimation de l'influence. Nous pouvons nous baser sur la modélisation des réseaux temporels.

— Et enfin, nous prévoyons d'appliquer le principe de l'approche proposée sur d'autres mesures qui nécessitent elles aussi la fusion d'information comme l'estimation de la crédibilité ou la réputation des utilisateurs dans *Twitter* ou dans des réseaux complexes. En effet, de telles mesures dépendent de plusieurs critères à fusionner, par exemple, afin d'estimer la crédibilité d'un certain utilisateur, plusieurs facteurs peuvent être exploités, un utilisateur est plus crédible que d'autres si l'image de son compte est une photo personnelle, s'il est souvent *mentionné* ou *retweeté*, s'il a un compte vérifié par les autorités de *Twitter* ou si ses *tweets* contiennent des URLs.

BIBLIOGRAPHIE

[Abel et al., 2011] Abel, F., Gao, Q., Houben, G.-J., et Tao, K. (2011). **Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web**. Dans *The Semantic Web : Research and Applications, 8th Extended Semantic Web Conference (ESWC)*, pages 375–389.

[Al-Garadi et al., 2018] Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., Mujtaba, G., Khan, M. U. S., et Khan, S. U. (2018). **Analysis of online social network connections for identification of influential users : Survey and open research issues**. *ACM Comput. Surv.*, 51(1) :16 :1–16 :37.

[Allard et al., 2009] Allard, A., Noël, P.-A., Dubé, L. J., et Pourbohloul, B. (2009). **Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics**. *Physical Review E*, 79(3) :036113.

[Álvarez, 2012] Álvarez, P. C. (2012). **Journalism and social media : How spanish journalists are using twitter/periodismo y social media : cómo están usando twitter los periodistas españoles**. *Estudios sobre el mensaje periodístico*, 18(1) :31–53.

[Amato et al., 2016] Amato, F., Moscato, V., Picariello, A., et Sperlì, G. (2016). **Modeling User-Content Interaction in Multimedia Social Networks Using Hypergraphs**. Dans *12th International Conference on Signal-Image Technology & Internet-Based Systems, SITIS*, pages 343–350.

[Anderson, 2017] Anderson, B. (2017). **Tweeter-in-chief : A content analysis of president trump's tweeting habits**. *ELON JOURNAL*, page 36.

[Arasu et al., 2002] Arasu, A., Novak, J., Tomkins, A., et Tomlin, J. (2002). **Pagerank computation and the structure of the web : Experiments and algorithms**. Dans *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*, pages 107–117.

[Ashwini et al., 2015] Ashwini, S. S., et Sindhu, M. (2015). **Profile Ranking Using User Influence and Content Relevance with Classification Using Sentiment Analysis**. *International Journal of Computer Science and Mobile Computing*, 4(6) :1075–1080.

[Ausserhofer et al., 2013] Ausserhofer, J., et Maireder, A. (2013). **National politics on twitter : Structures and topics of a networked public sphere**. *Information, Communication & Society*, 16(3) :291–314.

[Azaza et al., 2015] Azaza, L., Faiz, R., et Benslimane, D. (2015). **Une approche de filtrage d'opinions à base de crédibilité dans un contexte de réseaux sociaux**. *ISI*, 20(4).

[Bae et al., 2014] Bae, J., et Kim, S. (2014). **Identifying and ranking influential spreaders in complex networks by neighborhood coreness**. *Physica A : Statistical Mechanics and its Applications*, 395 :549–559.

[Bakhtine, 1996] Bakhtine, M. (1996). **Collected Works. T. 5 : Travaux des années 1940 et début des années 1960**, volume 5. Moscou :Dictionnaires russes.

[Bakshy et al., 2011] Bakshy, E., Hofman, J. M., Mason, W. A., et Watts, D. J. (2011). **Everyone's an Influencer : Quantifying Influence on Twitter**. Dans *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA. ACM.

[Barabási et al., 1999] Barabási, A.-L., et Albert, R. (1999). **Emergence of scaling in random networks**. *science*, 286(5439) :509–512.

[Barnes, 1969] Barnes, J. A. (1969). **Graph Theory and Social Networks : A Technical Comment on Connectedness and Connectivity**. *Sociology*, pages 215–232.

[Barrat, 2013] Barrat, A. (2013). **La notion de réseau complexe : du réseau comme abstraction et outil à la masse de données des réseaux sociaux en ligne**. *Communication et organisation*, (43) :15–24.

[Basaille et al., 2016] Basaille, I., Kirgizov, S., Leclercq, E., Savonnet, M., et Cullot, N. (2016). **Towards a twitter observatory : A multi-paradigm framework for collecting, storing and analysing tweets**. Dans *Tenth IEEE International Conference on Research Challenges in Information Science, (RCIS)*, pages 1–10.

[Basaille et al., 2018] Basaille, I., et Leclercq, É. (2018). **Détection de communautés dans les réseaux sociaux, approches, algorithmes, interprétations et limites**. Dans Brachotte, G., et Frame, A., éditeurs, *L'usage de Twitter par les candidats*, pages 189–216. EMS.

[Basaille-Gahitte et al., 2013] Basaille-Gahitte, I., Abrouk, L., Cullot, N., et Leclercq, E. (2013). **Using social networks to enhance customer relationship management**. Dans *Fifth International Conference on Management of Emergent Digital EcoSystems, MEDES*, pages 169–176.

[Basaras et al., 2013] Basaras, P., Katsaros, D., et Tassiulas, L. (2013). **Detecting Influential Spreaders in Complex, Dynamic Networks**. *Computer*, 46(4) :24–29.

[Bass, 1969] Bass, F. M. (1969). **A new product growth for model consumer durables**. *Management science*, 15(5) :215–227.

[Bavelas, 1950] Bavelas, A. (1950). **Communication patterns in task-oriented groups**. *The Journal of the Acoustical Society of America*, 22(6) :725–730.

[Bays, 1998] Bays, H. (1998). **Framing and face in internet exchanges : A socio-cognitive approach**. *Linguistik Online*, 1(1).

[Ben Dhaou et al., 2014] Ben Dhaou, S., Kharoune, M., Martin, A., et Yaghlane, B. B. (2014). **Belief Approach for Social Networks**. Dans *Belief Functions : Theory and Applications*, pages 115–123. Springer.

[Benevenuto et al., 2010] Benevenuto, F., Magno, G., Rodrigues, T., et Almeida, V. (2010). **Detecting spammers on twitter**. Dans *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12.

[Bianconi, 2018] Bianconi, G. (2018). **Multilayer Networks : Structure and Function**. Oxford University Press.

[Bollen et al., 2006] Bollen, J., Rodriquez, M. A., et Van de Sompel, H. (2006). **Journal status**. *Scientometrics*, 69(3) :669–687.

[Bollobás, 1985] Bollobás, B. (1985). **Random graphs**. Academic Press.

[Breiman, 2001] Breiman, L. (2001). **Random forests**. *Machine Learning*, 45(1) :5–32.

[Brown et al., 1987] Brown, J. J., et Reingen, P. H. (1987). **Social ties and word-of-mouth referral behavior**. *Journal of Consumer research*, 14(3) :350–362.

[Brown et al., 2011] Brown, P. E., et Feng, J. (2011). **Measuring user influence on twitter using modified k-shell decomposition**. Dans *Fifth International AAAI Conference on Weblogs and Social Media*, pages 18–23.

[Buldyrev et al., 2010] Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., et Havlin, S. (2010). **Catastrophic cascade of failures in interdependent networks**. *Nature*, 464(7291) :1025.

[Burnap et al., 2015a] Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., et Sloan, L. (2015a). **Detecting tension in online communities with computational twitter analysis**. *Technological Forecasting and Social Change*, 95 :96–108.

[Burnap et al., 2015b] Burnap, P., et Williams, M. L. (2015b). **Cyber hate speech on twitter : An application of machine classification and statistical modeling for policy and decision making**. *Policy and Internet*, 7(2) :223–242.

[Burnap et al., 2016] Burnap, P., et Williams, M. L. (2016). **Us and them : identifying cyber hate on twitter across multiple protected characteristics**. *EPJ Data Science*, 5(1) :11.

[Cai et al., 2005] Cai, D., Shao, Z., He, X., Yan, X., et Han, J. (2005). **Community mining from multi-relational networks**. Dans *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 445–452. Springer.

[Candillier et al., 2012] Candillier, L., Chai, E., et Delpech, E. (2012). **Systèmes de recommandation et Recherche d'Information**. Dans Ghislaine Chartron, Imad Saleh, G. K., éditeur, *Journée d'étude " Systèmes de recommandation " (CNAM)*, Collection information, hypermédias et communication, Paris, France. Hermès Sciences.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., et Gummadi, P. K. (2010). **Measuring user influence in twitter : The million follower fallacy.** *ICWSM*, 10(30) :10–17.

[Cha et al., 2009] Cha, M., Mislove, A., et Gummadi, K. P. (2009). **A Measurement-driven Analysis of Information Propagation in the Flickr Social Network**. Dans *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 721–730.

[Chen et al., 2012a] Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., et Zhou, T. (2012a). **Identifying influential nodes in complex networks**. *Physica a : Statistical mechanics and its applications*, 391(4) :1777–1787.

[Chen et al., 2013a] Chen, D.-B., Gao, H., Lü, L., et Zhou, T. (2013a). **Identifying Influential Nodes in Large-Scale Directed Networks : The Role of Clustering**. *PLoS ONE*, 8(10) :1–10.

[Chen et al., 2013b] Chen, D.-B., Xiao, R., Zeng, A., et Zhang, Y.-C. (2013b). **Path diversity improves the identification of influential spreaders**. *EPL (Europhysics Letters)*, 104(6) :68006.

[Chen et al., 2012b] Chen, S.-J., et Hwang, C.-L. (2012b). **Fuzzy multiple attribute decision making : methods and applications**, volume 375. Springer Science & Business Media.

[Chen et al., 2012c] Chen, W., Cheng, S., He, X., et Jiang, F. (2012c). **InfluenceRank : An Efficient Social Influence Measurement for Millions of Users in Microblog**. Dans *Second International Conference on Cloud and Green Computing*, pages 563–570.

[Chunfeng et al., 2017] Chunfeng, L., Su, R., Thierry, D., Hua, L., et Pierre, V. (2017). **Spatial evidential clustering with adaptive distance metric for tumor segmentation in fdg-pet images**. *IEEE Transactions on Biomedical Engineering*, 1(99).

[Coddington et al., 2014] Coddington, M., Molyneux, L., et Lawrence, R. G. (2014). **Fact checking the campaign : How political reporters use twitter to set the record straight (or not)**. *The International Journal of Press/Politics*, 19(4) :391–409.

[Cohen, 2016] Cohen, S. (2016). **Data Management for Social Networking**. Dans *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '16, pages 165–177. ACM.

[Cossu et al., 2015] Cossu, J., Dugué, N., et Labatut, V. (2015). **Detecting real-world influence through twitter**. Dans *2015 Second European Network Intelligence Conference*, pages 83–90.

[Cossu et al., 2016] Cossu, J.-V., Labatut, V., et Dugué, N. (2016). **A review of features for the discrimination of twitter users : application to the prediction of offline influence**. *Social Network Analysis and Mining*, 6(1) :25.

[Coutant et al., 2010] Coutant, A., et Stenger, T. (2010). **Processus identitaire et ordre de l'interaction sur les réseaux socionumériques**. *Les Enjeux de l'information et de la communication*, 1 :45–64.

[Crofts et al., 2011] Crofts, J. J., et Higham, D. J. (2011). **Googling the brain : Discovering hierarchical and asymmetric network structures, with applications in neuroscience**. *Internet Mathematics*, 7(4) :233–254.

[Dai et al., 2012] Dai, B. T., Chua, F. C. T., et Lim, E.-P. (2012). **Structural Analysis in Multi-Relational Social Networks**. Dans *Proceedings of the SIAM International Conference on Data Mining*, chapitre 38, pages 451–462.

[Dakhli, 2011] Dakhli, L. (2011). **Une lecture de la révolution tunisienne**. *Le mouvement social*, (3) :89–103.

[Dang-Anh et al., 2013] Dang-Anh, M., Einspänner, J., et Thimm, C. (2013). **Mediatisierung und medialität in social media : Das diskurssystem "twitter"**. Dans *Sprache und Kommunikation im technischen Zeitalter. Wieviel Internet (v)erträgt unsere Gesellschaft ?*, page 326. de Gruyter.

[Danisch et al., 2014] Danisch, M., Dugué, N., et Perez, A. (2014). **On the importance of considering social capitalism when measuring influence on Twitter**. Dans *Behavioral, Economic, and Socio-Cultural Computing*, pages 1–7, Shanghai, China.

[Davis et al., 2011] Davis, D., Lichtenwalter, R., et Chawla, N. V. (2011). **Multi-relational link prediction in heterogeneous information networks**. Dans *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 281–288. IEEE.

[de Arruda et al., 2014] de Arruda, G. F., Barbieri, A. L., Rodríguez, P. M., Rodrigues, F. A., Moreno, Y., et da Fontoura Costa, L. (2014). **Role of centrality for the identification of influential spreaders in complex networks**. *Physical Review E*, 90(3) :032812.

[De Domenico et al., 2013] De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., Gómez, S., et Arenas, A. (2013). **Mathematical Formulation of Multilayer Networks**. *Phys. Rev. X*, 3 :041022.

[Debeaux, 2015] Debeaux, M. (2015). **Immersion de Twitter dans la sphère journalistique : quelles conséquences et quelles pratiques ?** Master's thesis, Université Stendhal - Grenoble 3.

[Debiolles, 2007] Debiolles, A. (2007). **Diagnostic de systèmes complexes à base de modèle interne, reconnaissance des formes et fusion d'informations**. PhD thesis, Université de Technologie de Compiègne, Compiègne, France.

[Demotier et al., 2006] Demotier, S., Schon, W., et Denoeux, T. (2006). **Risk assessment based on weak information using belief functions : a case study in water treatment**. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 36(3) :382–396.

[Dempster, 1968] Dempster, A. P. (1968). **A generalization of bayesian inference**. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2) :205–247.

[Dickison et al., 2012] Dickison, M., Havlin, S., et Stanley, H. E. (2012). **Epidemics on interconnected networks**. *Physical Review E*, 85(6) :066109.

[Ding et al., 2009] Ding, Y., Yan, E., Frazho, A., et Caverlee, J. (2009). **Pagerank for ranking authors in co-citation networks**. *Journal of the American Society for Information Science and Technology*, 60(11) :2229–2243.

[Domingos et al., 2001] Domingos, P., et Richardson, M. (2001). **Mining the network value of customers**. Dans *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 57–66, New York, NY, USA. ACM.

[Donges et al., 2011] Donges, J. F., Schultz, H. C., Marwan, N., Zou, Y., et Kurths, J. (2011). **Investigating the topology of interacting networks**. *The European Physical Journal B*, 84(4) :635–651.

[Drakopoulos et al., 2016] Drakopoulos, G., Kanavos, A., et Tsakalidis, A. K. (2016). **Evaluating Twitter Influence Ranking with System Theory**. Dans *Proceedings of the 12th International Conference on Web Information Systems and Technologies, (WEBIST)*, pages 113–120.

[Earle et al., 2012] Earle, P. S., Bowden, D. C., et Guy, M. (2012). **Twitter earthquake detection : earthquake monitoring in a social world**. *Annals of Geophysics*, 54(6).

[Eguiluz et al., 2002] Eguiluz, V. M., et Klemm, K. (2002). **Epidemic threshold in structured scale-free networks**. *Physical Review Letters*, 89(10) :108701.

[Erdős et al., 1960] Erdős, P., et Rényi, A. (1960). **On the evolution of random graphs**. Dans *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61.

[Evans et al., 2018] Evans, T., et Fu, F. (2018). **Opinion formation on dynamic networks : identifying conditions for the emergence of partisan echo chambers**. *Open Science*, 5(10) :181122.

[Fang et al., 2014] Fang, Q., Sang, J., Xu, C., Rui, Y., et others (2014). **Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning.** *IEEE Trans. Multimedia*, 16(3) :796–812.

[Fiche et al., 2009] Fiche, A., et Martin, A. (2009). **Approche bayésienne et fonctions de croyance continues pour la classification**.

[Firan et al., 2007] Firan, C. S., Nejdl, W., et Paiu, R. (2007). **The Benefit of Using Tag-Based Profiles**. Dans *Fifth Latin American Web Congress (LA-Web 2007)*, pages 32–41.

[Frame et al., 2015] Frame, A., et Brachotte, G. (2015). **Citizen participation and political communication in a digital world**.

[Freeman, 1977] Freeman, L. C. (1977). **A set of measures of centrality based on betweenness**. *Sociometry*, pages 35–41.

[Gallager, 2013] Gallager, R. G. (2013). **Stochastic processes : theory for applications**. Cambridge University Press.

[Gao et al., 2013] Gao, C., Wei, D., Hu, Y., Mahadevan, S., et Deng, Y. (2013). **A modified evidential methodology of identifying influential nodes in weighted networks**. *Physica A : Statistical Mechanics and its Applications*, 392(21) :5490–5500.

[Gao et al., 2011] Gao, J., Buldyrev, S. V., Havlin, S., et Stanley, H. E. (2011). **Robustness of a network of networks**. *Physical Review Letters*, 107(19) :195701.

[Getoor et al., 2005] Getoor, L., et Diehl, C. P. (2005). **Link mining : A survey**. *SIGKDD Explor. Newsl.*, 7(2) :3–12.

[Ghalmane et al., 2018] Ghalmane, Z., Hassouni, M. E., Cherifi, C., et Cherifi, H. (2018). **Centrality in modular networks**. *arXiv preprint arXiv :1810.05101*.

[Ghosh et al., 2012] Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., et Gummadi, K. P. (2012). **Understanding and combating link farming in the twitter social network**. Dans *Proceedings of the 21st international conference on World Wide Web*, pages 61–70. ACM.

[Ghoshal et al., 2009] Ghoshal, G., Zlatić, V., Caldarelli, G., et Newman, M. (2009). **Random hypergraphs and their applications**. *Physical Review E*, 79(6) :066118.

[Girvan et al., 2002] Girvan, M., et Newman, M. E. J. (2002). **Community structure in social and biological networks**. *PNAS*, 99(12) :7821–7826.

[Gleich, 2015] Gleich, D. F. (2015). **Pagerank beyond the web**. *SIAM Review*, 57(3) :321–363.

[Goffman, 1967] Goffman, E. (1967). **On face-work**. *Interaction ritual*, pages 5–45.

[Goldenberg et al., 2001] Goldenberg, J., Libai, B., et Muller, E. (2001). **Talk of the network : A complex systems look at the underlying process of word-of-mouth**. *Marketing letters*, 12(3) :211–223.

[Goldie et al., 2014] Goldie, D., Linick, M., Jabbar, H., et Lubienski, C. (2014). **Using bibliometric and social media analyses to explore the "echo chamber" hypothesis**. *Educational Policy*, 28(2) :281–305.

[Govan et al., 2008] Govan, A. Y., Meyer, C. D., et Albright, R. (2008). **Generalizing google's pagerank to rank national football league teams**. Dans *Proceedings of the SAS Global Forum*, volume 2008.

[Grando et al., 2018] Grando, F., Granville, L. Z., et Lamb, L. C. (2018). **Machine learning in network centrality measures : Tutorial and outlook**. *ACM Computing Surveys (CSUR)*, 51(5) :102.

[Halu et al., 2013] Halu, A., Mondragón, R. J., Panzarasa, P., et Bianconi, G. (2013). **Multiplex pagerank**. *PloS one*, 8(10) :e78293.

[Hirsch, 2005] Hirsch, J. E. (2005). **An index to quantify an individual's scientific research output**. *Proceedings of the National academy of Sciences of the United States of America*, pages 16569–16572.

[Holme et al., 2012] Holme, P., et Saramäki, J. (2012). **Temporal networks**. *Physics reports*, 519(3) :97–125.

[Horvát et al., 2012] Horvát, E.-A., et Zweig, K. A. (2012). **One-mode projection of multiplex bipartite graphs**. Dans *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 599–606. IEEE Computer Society.

[Horvát et al., 2013] Horvát, E.-Á., et Zweig, K. A. (2013). **A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs**. *Social Network Analysis and Mining*, 3(4) :1209–1224.

[Howson et al., 2006] Howson, C., et Urbach, P. (2006). **Scientific Reasoning : The Bayesian Approach**. Philosophy Series. Open Court.

[Huang et al., 2014] Huang, W., Weber, I., et Vieweg, S. (2014). **Inferring nationalities of twitter users and studying inter-national linking**. Dans *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 237–242, New York, NY, USA. ACM.

[Hurlbert, 1971] Hurlbert, S. H. (1971). **The nonconcept of species diversity : a critique and alternative parameters**. *Ecology*, 52(4) :577–586.

[Huygue, 2011] Huygue, F. B. (2011). **Facebook, twitter, al-jazeera et le "printemps arabe"**. *Observatoire Géostratégique de l'Information, IRIS*.

[Iacovacci et al., 2016] Iacovacci, J., et Bianconi, G. (2016). **Extracting information from multiplex networks**. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 26(6) :065306.

[Jackson et al., 2011] Jackson, N., et Lilleker, D. (2011). **Microblogging, constituency service and impression management : Uk mps and the use of twitter**. *Legislative Studies*, 17 :86–105.

[Jansen et al., 2009] Jansen, B. J., Zhang, M., Sobel, K., et Chowdury, A. (2009). **Twitter power : Tweets as electronic word of mouth**. *J. Am. Soc. Inf. Sci. Technol.*, 60(11) :2169–2188.

[Jendoubi et al., 2017] Jendoubi, S., Martin, A., Liétard, L., Hadji, H. B., et Yaghlane, B. B. (2017). **Two Evidential Data Based Models for Influence Maximization in Twitter**. *Knowledge-Based Systems*, 121 :58–70.

[Jin et al., 2015] Jin, S., Yu, P. S., Li, S., et Yang, S. (2015). **A parallel community structure mining method in big social networks**. *Mathematical Problems in Engineering*, 2015.

[Jing et al., 2008] Jing, Y., et Baluja, S. (2008). **Pagerank for product image search**. Dans *Proceedings of the 17th international conference on World Wide Web*, pages 307–316. ACM.

[Kanawati, 2015] Kanawati, R. (2015). **Multiplex Network mining : a brief survey**. *IEEE Intelligent Informatics Bulletin*, 16(1) :24–27.

[Karinthy, 1929] Karinthy, F. (1929). **Chain-links**. *Everything is different*.

[Katsimpras et al., 2015] Katsimpras, G., Vogiatzis, D., et Paliouras, G. (2015). **Determining Influential Users with Supervised Random Walks**. Dans *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 787–792, New York, NY, USA. ACM.

[Keller et al., 2007] Keller, E., Fay, B., et Berry, J. (2007). **Leading the conversation : Influencers' impact on word of mouth and the brand conversation**. Dans *The Keller Fay Group, Word of Mouth Marketing Research Symposium*.

[Kelman, 1958] Kelman, H. C. (1958). **Compliance, identification, and internalization three processes of attitude change**. *Journal of conflict resolution*, 2(1) :51–60.

[Kempe et al., 2003] Kempe, D., Kleinberg, J., et Tardos, E. (2003). **Maximizing the spread of influence through a social network**. Dans *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA. ACM.

[Khan et al., 2014] Khan, A., Bonchi, F., Gionis, A., et Gullo, F. (2014). **Fast Reliability Search in Uncertain Graphs**. Dans *Proceedings of Proceedings of the International Conference on Extending Database Technology (EDBT '14)*, pages 535–546.

[Kim et al., 2013] Kim, M., Sumbaly, R., et Shah, S. (2013). **Root cause detection in a service-oriented architecture**. Dans *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 93–104. ACM.

[Kirgizov et al., 2016] Kirgizov, S., Gastineau, N., et Azaza, L. (2016). **Limit of generalized belief fusion operator**. Preprint : http://kirgizov.link/publications/azaza-limit/doc/limit.pdf.

[Kitsak et al., 2010] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., et Makse, H. A. (2010). **Identification of influential spreaders in complex networks**. *Nature physics*, 6(11) :888–893.

[Kivelä et al., 2014] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., et Porter, M. A. (2014). **Multilayer networks**. *Journal of complex networks*, 2(3) :203–271.

[Kleinberg, 1999] Kleinberg, J. M. (1999). **Authoritative Sources in a Hyperlinked Environment**. *Journal ACM*, 46(5) :604–632.

[Kleinfeld, 2002] Kleinfeld, J. (2002). **Could it be a Big World After All ? The 'Six Degrees of Separation' Myth**. *Society*.

[Klemm et al., 2012] Klemm, K., Serrano, M. Á., Eguiluz, V. M., et San Miguel, M. (2012). **A measure of individual role in collective dynamics**. *Scientific reports*, 2 :292.

[Kondrashova et al., 2015] Kondrashova, T., et Frame, A. (2015). **Exploring the dialogical dimension of political tweets : A qualitative analysis of 'twitter styles' of uk candidates during the 2014 eu parliamentary elections**. *Peter Lang*.

[Korn et al., 2009] Korn, A., Schubert, A., et Telcs, A. (2009). **Lobby index in networks**. *Physica A : Statistical Mechanics and its Applications*, 388(11) :2221–2226.

[Kotz et al., 1982] Kotz, S., et N. L. Johnson eds., W. (1982). **Belief functions**. *Encyclopedia of Statistical Sciences 1 209*.

[Kreis, 2017] Kreis, R. (2017). **The "tweet politics" of president trump**. *Journal of Language and Politics*, 16(4) :607–618.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., et Moon, S. (2010). **What is Twitter, a Social Network or a News Media ?** Dans *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600.

[Lazer et al., 2009] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., et Van Alstyne, M. (2009). **Computational Social Science**. *Science*, 323(5915) :721–723.

[Leavitt et al., 2009] Leavitt, A., Burchard, E., Fisher, D., et Gilbert, S. (2009). **The Influentials : New Approaches for Analyzing Influence on Twitter**. *Webecology Project*.

[Leclercq et al., 2015] Leclercq, E., Savonnet, M., Grison, T., Kirgizov, S., et Basaille, I. (2015). **SNFreezer : a Platform for Harvesting and Storing Tweets in a Big Data Context**. Dans Frame, A., Mercier, A., Brachotte, G., et Thimm, C., éditeurs, *Twitter and the European Parliamentary Elections : researching political uses of microblogging*, pages 1–16. Peter Lang, DE.

[Lee et al., 2010a] Lee, C., Kwak, H., Park, H., et Moon, S. (2010a). **Finding influentials based on the temporal order of information adoption in twitter**. Dans *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 1137–1138.

[Lee et al., 2010b] Lee, K., Caverlee, J., et Webb, S. (2010b). **Uncovering social spammers : Social honeypots + machine learning**. Dans *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 435–442, New York, NY, USA. ACM.

[Lee et al., 2011] Lee, K., Eoff, B. D., et Caverlee, J. (2011). **Seven months with the devils : A long-term study of content polluters on twitter.** Dans *Proceedings of the Fifth International Conference on Weblogs and Social Media : ICWSM*, pages 185–192, Barcelona, Catalonia, Spain. The AAAI Press 2011.

[Lee et al., 2013] Lee, K., Tamilarasan, P., et Caverlee, J. (2013). **Crowdturfers, campaigns, and social media : Tracking and revealing crowdsourced manipulation of social media.** Dans Kiciman, E., Ellison, N. B., Hogan, B., Resnick, P., et Soboroff, I., éditeurs, *ICWSM*. The AAAI Press.

[Leicht et al., 2009] Leicht, E., et D'Souza, R. M. (2009). **Percolation on interacting networks**. *arXiv preprint arXiv :0907.0894*.

[Leskovec et al., 2010] Leskovec, J., Huttenlocher, D., et Kleinberg, J. (2010). **Predicting positive and negative links in online social networks**. Dans *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 641–650, New York, NY, USA. ACM.

[Leskovec et al., 2007] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., et Glance, N. (2007). **Cost-effective outbreak detection in networks**. Dans *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM.

[Lesne, 2006] Lesne, A. (2006). **Complex networks : from graph theory to biology**. *Letters in Mathematical Physics*, 78(3) :235–262.

[Li et al., 2014] Li, Q., Zhou, T., Li, L., et Chen, D. (2014). **Identifying influential spreaders by weighted LeaderRank**. *Physica A : Statistical Mechanics and its Applications*, 404 :47–55.

[Li et al., 2012] Li, R., Lei, K. H., Khadiwala, R., et Chang, K. C.-C. (2012). **Tedas : A twitter-based event detection and analysis system**. Dans *IEEE 28th international conference on Data engineering (ICDE)*, pages 1273–1276. IEEE.

[Lian et al., 2016] Lian, C., Ruan, S., Denœux, T., Jardin, F., et Vera, P. (2016). **Selecting radiomic features from fdg-pet images for cancer treatment outcome prediction**. *Medical Image Analysis*, 32 :257 – 268.

[Liben-Nowell et al., 2003] Liben-Nowell, D., et Kleinberg, J. (2003). **The link prediction problem for social networks**. Dans *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA. ACM.

[Liu et al., 2005] Liu, X., Bollen, J., Nelson, M. L., et Van de Sompel, H. (2005). **Co-authorship networks in the digital library research community**. *Information processing & management*, 41(6) :1462–1480.

[Liu et al., 2014] Liu, X., Liu, W., Murata, T., et Wakita, K. (2014). **A framework for community detection in heterogeneous multi-relational networks**. *Advances in Complex Systems*, 17(06) :1450018.

[Liu et al., 2015] Liu, Y., Tang, M., Zhou, T., et Do, Y. (2015). **Improving the accuracy of the k-shell method by removing redundant links : From a perspective of spreading dynamics**. *Scientific reports*, 5 :13172.

[Lü et al., 2011] Lü, L., Zhang, Y.-C., Yeung, C. H., et Zhou, T. (2011). **Leaders in social networks, the delicious case**. *PloS one*, 6(6) :e21202.

[Lü et al., 2016] Lü, L., Zhou, T., Zhang, Q.-M., et Stanley, H. E. (2016). **The h-index of a network node and its relation to degree and coreness**. *Nature communications*, 7.

[Ma et al., 2016] Ma, L.-l., Ma, C., Zhang, H.-F., et Wang, B.-H. (2016). **Identifying influential spreaders in complex networks based on gravity formula**. *Physica A : Statistical Mechanics and its Applications*, 451(C) :205–212.

[Mahajan et al., 1991] Mahajan, V., Muller, E., et Bass, F. M. (1991). **New product diffusion models in marketing : A review and directions for research**. Dans *Diffusion of technologies and social behavior*, pages 125–177. Springer.

[Makazhanov et al., 2014] Makazhanov, A., Rafiei, D., et Waqar, M. (2014). **Predicting political preference of twitter users**. *Social Network Analysis and Mining*, 4(1) :193.

[Marneffe et al., 2006] Marneffe, M., Maccartney, B., et Manning, C. (2006). **Generating typed dependency parses from phrase structure parses**. Dans *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy. European Language Resources Association (ELRA). ACL Anthology Identifier : L06-1260.

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., et Cook, J. M. (2001). **Birds of a feather : Homophily in social networks**. *Annual review of sociology*, 27(1) :415–444.

[Meng, 2009] Meng, X. (2009). **Computing bookrank via social cataloging**. Dans *Web slides for CADS conference, February*, volume 22.

[Mercier et al., 2009] Mercier, D., Cron, G., Denœux, T., et Masson, M.-H. (2009). **Decision fusion for postal address recognition using belief functions**. *Expert Systems with Applications*, 36(3, Part 1) :5643 – 5653.

[Milgram et al., 1967] Milgram, S., et Travers, J. (1967). **The small world problem**. *Phychology Today*, 1(1) :61–67.

[Mira, 2014] Mira, B. F. (2014). **Méthodes utilisant des fonctions de croyance pour la gestion des informations imparfaites dans les réseaux de véhicules.** PhD thesis, Université d'Artois, France.

[Mo et al., 2015] Mo, H., Gao, C., et Deng, Y. (2015). **Evidential method to identify influential nodes in complex networks**. *Systems Engineering and Electronics, Journal of*, 26(2) :381–387.

[Mooney et al., 2012] Mooney, B. L., Corrales, L. R., et Clark, A. E. (2012). **Molecularnetworks : An integrated graph theoretic and data mining tool to explore solvent organization in molecular simulation**. *Journal of computational chemistry*, 33(8) :853–860.

[Morrison et al., 2005] Morrison, J. L., Breitling, R., Higham, D. J., et Gilbert, D. R. (2005). **Generank : using search engine technology for the analysis of microarray experiments**. *BMC bioinformatics*, 6(1) :233.

[Muruganantham et al., 2015] Muruganantham, A., Gandhi, D. G. M., Fathima, Y. A., Muthumani, D., Al-Dossari, H., Mahmoud, A. S., Chandu, P., Bohanudin, S., Ismail, M., et Abdulah, M. (2015). **Ranking the influence of users in a social networking site using an improved topsis method**. *Journal of Theoretical and Applied Information Technology*, 73(1).

[Mustafaraj et al., 2011] Mustafaraj, E., et Metaxas, P. T. (2011). **What edited retweets reveal about online political discourse**. Dans *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, AAAIWS'11-05, pages 38–43. AAAI Press.

[Nagmoti et al., 2010] Nagmoti, R., Teredesai, A., et De Cock, M. (2010). **Ranking approaches for microblog search**. Dans *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 153–157, Washington, DC, USA. IEEE Computer Society.

[Neves et al., 2015] Neves, A., Vieira, R., Mourão, F., et Rocha, L. (2015). **Quantifying Complementarity among Strategies for Influencers' Detection on Twitter**. *Procedia Computer Science*, 51 :2435–2444.

[Newman, 2001] Newman, M. E. J. (2001). **Clustering and preferential attachment in growing networks**. *Physical Review Letters E*, 64 :025–102.

[Newman, 2004] Newman, M. E. J. (2004). **Fast algorithm for detecting community structure in networks**. *Physical Review Letters E*, 69 :066–133.

[Nimier et al., 1995] Nimier, V., et Appriou, A. (1995). **Utilisation de la théorie de dempster-shafer pour la fusion d'informations**. *GRETSI, Groupe d'Etudes du Traitement du Signal et des Images*, pages 137–140.

[Ott, 2017] Ott, B. L. (2017). **The age of Twitter : Donald J. Trump and the politics of debasement**. *Critical Studies in Media Communication*, 34(1) :59–68.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., et Winograd, T. (1999). **The pagerank citation ranking : Bringing order to the web**. Dans *Proceedings of the 7th International World Wide Web Conference*, pages 161–172.

[Pak et al., 2010] Pak, A., et Paroubek, P. (2010). **Twitter as a Corpus for Sentiment Analysis and Opinion Mining**. 10(2010) :1320–1326.

[Pang et al., 2008] Pang, B., et Lee, L. (2008). **Opinion Mining and Sentiment Analysis**. *Found. Trends Inf. Retr.*, 2(1-2) :1–135.

[Parchas et al., 2014] Parchas, P., Gullo, F., Papadias, D., et Bonchi, F. (2014). **The Pursuit of a Good Possible World : Extracting Representative Instances of Uncertain Graphs**. Dans *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 967–978.

[Parshani et al., 2010] Parshani, R., Buldyrev, S. V., et Havlin, S. (2010). **Interdependent networks : Reducing the coupling strength leads to a change from a first to second order percolation transition**. *Physical review letters*, 105(4) :048701.

[Paveau, 2012] Paveau, M.-A. (2012). **Genre de discours et technologie discursive. Tweet, twittécriture et twittérature**. *Pratique*, pages 7–30.

[Pennacchiotti et al., 2011] Pennacchiotti, M., et Popescu, A.-M. (2011). **A machine learning approach to twitter user classification.** *ICWSM*, 11(1) :281–288.

[Petermann et al., 2004] Petermann, T., et De Los Rios, P. (2004). **Role of clustering and gridlike ordering in epidemic spreading**. *Physical Review E*, 69(6) :066116.

[Psomakelis et al., 2015] Psomakelis, E., Tserpes, K., Anagnostopoulos, D., et Varvarigou, T. A. (2015). **Comparing methods for Twitter Sentiment Analysis**. *CoRR*, abs/1505.02973.

[Pujol et al., 2002] Pujol, J. M., Sangüesa, R., et Delgado, J. (2002). **Extracting Reputation in Multi Agent Systems by Means of Social Network Topology**. Dans *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems : Part 1*, AAMAS '02, pages 467–474. ACM.

[Radicchi, 2011] Radicchi, F. (2011). **Who is the best player ever ? a complex network analysis of the history of professional tennis**. *PloS one*, 6(2) :e17249.

[Ramírez-de-la Rosa et al., 2014] Ramírez-de-la Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H., et Sánchez-Sánchez, C. (2014). **Towards automatic detection of user influence in twitter by means of stylistic and behavioral features**. Dans Gelbukh, A., Espinoza, F. C., et Galicia-Haro, S. N., éditeurs, *Human-Inspired Computing and Its Applications, 13th Mexican International Conference on Artificial Intelligence, MICAI, Tuxtla Gutiérrez,, Proceedings, Part I*, pages 245–256, Cham, Mexico. Springer International Publishing.

[Rao et al., 2010] Rao, D., Yarowsky, D., Shreevats, A., et Gupta, M. (2010). **Classifying latent user attributes in twitter**. Dans *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

[Rashotte, 2007] Rashotte, L. (2007). **Social influence**. *The blackwell encyclopedia of social psychology*, 9 :562–563.

[Ren et al., 2015] Ren, J., Wang, C., He, H., et Dong, J. (2015). **Identifying influential nodes in weighted network based on evidence theory and local structure**. *International journal of innovative computing information and control*, 11(5) :1765–1777.

[Rezaei et al., 2015] Rezaei, Z., et Tarokh, M. J. (2015). **Discovering Influencers for Spreading in Weighted Networks**. *International Journal of Information and Communication Technology (IJICT)*, 7(3) :43–51.

[Richardson et al., 2002] Richardson, M., et Domingos, P. (2002). **Mining knowledge-sharing sites for viral marketing**. Dans *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 61–70, New York, NY, USA. ACM.

[Riquelme et al., 2016] Riquelme, F., et González-Cantergiani, P. (2016). **Measuring user influence on twitter : A survey**. *Information Processing & Management*, 52(5) :949–975.

[Romero et al., 2011] Romero, D. M., Galuba, W., Asur, S., et Huberman, B. A. (2011). **Influence and passivity in social media**. Dans *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 113–114.

[Régis et al., 2007] Régis, S., Doncescu, A., et Desachy, J. (2007). **Théorie des fonctions de croyance pour la fusion et l'évaluation de la pertinence des sources d'informations : application à un bioprocédé fermentaire**. *Traitement du signal. Numéro spécial La théorie des fonctions de croyance*, 24.

[Saumell-Mendiola et al., 2012] Saumell-Mendiola, A., Serrano, M. Á., et Boguná, M. (2012). **Epidemic spreading on interconnected networks**. *Physical Review E*, 86(2) :026106.

[Sayyadi et al., 2009] Sayyadi, H., et Getoor, L. (2009). **Futurerank : Ranking scientific articles by predicting their future pagerank**. Dans *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 533–544. SIAM.

[Schaub et al., 2017] Schaub, M. T., Delvenne, J.-C., Rosvall, M., et Lambiotte, R. (2017). **The many facets of community detection in complex networks**. *Applied Network Science*, 2(1) :4.

[Seidman, 1983] Seidman, S. B. (1983). **Network structure and minimum degree**. *Social Networks*, 5(3) :269 – 287.

[Shafer, 1976] Shafer, G. (1976). **A Mathematical Theory of Evidence**. Princeton University Press, Princeton.

[Silva et al., 2013] Silva, A., Guimarães, S., Meira, Jr., W., et Zaki, M. (2013). **ProfileRank : Finding Relevant Content and Influential Users Based on Information Diffusion**. Dans *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, SNAKDD '13, pages 2 :1–2 :9, New York, NY, USA. ACM.

[Simmie et al., 2013] Simmie, D., Vigliotti, M., et Hankin, C. (2013). **Ranking twitter influence by combining network centrality and influence observables in an evolutionary model**. Dans *International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 491–498.

[Smets, 1989] Smets, P. (1989). **Constructing the Pignistic Probability Function in a Context of Uncertainty**. 89(1989) :29–40.

[Smets, 1997] Smets, P. (1997). **Imperfect Information : Imprecision and Uncertainty**. Dans Motro, A., et Smets, P., éditeurs, *Uncertainty Management in Information Systems*, pages 225–254.

[Solé-Ribalta et al., 2014] Solé-Ribalta, A., De Domenico, M., Gómez, S., et Arenas, A. (2014). **Centrality rankings in multiplex networks**. Dans *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 149–155, New York, NY, USA. ACM.

[Spatocco et al., 2018] Spatocco, C., D'Andrea, A., Domeniconi, C., et Stilo, G. (2018). **A New Framework for Centrality Measures in Multiplex Networks**. *CoRR*, abs/1801.08026.

[Stieglitz et al., 2012] Stieglitz, S., et Dang-Xuan, L. (2012). **Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior**. Dans *45th Hawaii International Conference on System Sciences*, pages 3500–3509.

[Stroele et al., 2009] Stroele, V., Oliveira, J., Zimbrao, G., et Souza, J. M. (2009). **Mining and analyzing multirelational social networks**. Dans *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 711–716. IEEE.

[Su et al., 2011] Su, C., Pan, Y., Zhen, Y., Ma, Z., Yuan, J., Guo, H., Yu, Z., Ma, C., et Wu, Y. (2011). **Prestigerank : A new evaluation method for papers and journals**. *Journal of Informetrics*, 5(1) :1–13.

[Suh et al., 2010] Suh, B., Hong, L., Pirolli, P., et Chi, E. H. (2010). **Want to Be Retweeted ? Large Scale Analytics on Factors Impacting Retweet in Twitter Network**. Dans *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, pages 177–184, Washington, DC, USA. IEEE Computer Society.

[Sun et al., 2011] Sun, J., et Tang, J. (2011). **A Survey of Models and Algorithms for Social Influence Analysis**. Dans *Social Network Data Analytics*, pages 177–214.

[Sun et al., 2013] Sun, Y., et Han, J. (2013). **Mining heterogeneous information networks : a structural analysis approach**. *Acm Sigkdd Explorations Newsletter*, 14(2) :20–28.

[Sun et al., 2009] Sun, Y., Yu, Y., et Han, J. (2009). **Ranking-based clustering of heterogeneous information networks with star network schema**. Dans *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 797–806. ACM.

[Tacnet et al., 2010] Tacnet, J., Batton-Hubert, M., et Dezert, J. (2010). **Analyse multi-critères et fusion d'information pour l'expertise et la gestion intégrée des risques naturels en montagne**. Dans *Colloque LambdaMu 17*, page 10.

[Tan et al., 2011] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., et Li, P. (2011). **User-level sentiment analysis incorporating social networks**. Dans *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1397–1405, New York, NY, USA. ACM.

[Tang, 2017] Tang, J. (2017). **Computational Models for Social Network Analysis : A Brief Survey**. Dans *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 921–925. International World Wide Web Conferences Steering Committee.

[Tang et al., 2014] Tang, J., Chang, Y., et Liu, H. (2014). **Mining Social Media with Social Theories : A Survey**. *SIGKDD Explor. Newsl.*, 15(2) :20–29.

[Tewarie et al., 2016] Tewarie, P., Hillebrand, A., van Dijk, B. W., Stam, C. J., O'Neill, G. C., Mieghem, P. V., Meier, J. M., Woolrich, M. W., Morris, P. G., et Brookes, M. J. (2016). **Integrating cross-frequency and within band functional networks in resting-state meg : A multi-layer network approach**. *NeuroImage*, 142 :324 – 336.

[Thimm et al., 2012a] Thimm, C., Einspänner, J., et Dang-Anh, M. (2012a). **Politische deliberation online – twitter als element des politischen diskurses**. Dans Krotz, F., et Hepp, A., éditeurs, *Mediatisierte Welten : Forschungsfelder und Beschreibungsansätze*, pages 283–305. VS Verlag für Sozialwissenschaften, Wiesbaden.

[Thimm et al., 2012b] Thimm, C., Einspänner, J., et Dang-Anh, M. (2012b). **Twitter as a medium in election campaigns**. *Publizistik*, 57(3) :293–313.

[Thimm et al., 2016] Thimm, C., Frame, A., Einspänner-Pflock, J., Leclercq, E., et Anastasiadis, M. (2016). **The eu-election on twitter : Comparison of german and french candidates' tweeting styles**. Dans *Europawahlkampf 2014*, pages 175–204. Springer Fachmedien Wiesbaden.

[Tommasel et al., 2015] Tommasel, A., et Godoy, D. (2015). **A novel metric for assessing user influence based on user behaviour**. Dans *Proceedings of the 1st International Conference on Social Influence Analysis - Volume 1398*, SocInf'15, pages 15–21, Aachen, Germany, Germany. CEUR-WS.org.

[Tu et al., 2018] Tu, X., Jiang, G.-P., Song, Y., et Zhang, X. (2018). **Novel Multiplex PageRank in Multilayer Networks**. *IEEE Access*, 6 :12530–12538.

[Tunkelang, 2009] Tunkelang, D. (2009). **A Twitter Analog to PageRank**. http ://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank.

[Uddin et al., 2014] Uddin, M. M., Imran, M., et Sajjad, H. (2014). **Understanding types of users on twitter**. *CoRR*, abs/1406.1335.

[Vannoorenberghe et al., 2003] Vannoorenberghe, P., Lefevre, E., et Colot, O. (2003). **Traitement d'images et théoorie des fonctions de croyance**. *Rencontres Francophones sur la Logique Floue et Ses Applications, LFA*, pages 26–27.

[Vazquez, 2006] Vazquez, A. (2006). **Spreading dynamics on heterogeneous populations : multitype network approach**. *Physical Review E*, 74(6) :066114.

[Vidya et al., 2015] Vidya, N. A., Fanany, M. I., et Budi, I. (2015). **Twitter sentiment to analyze net brand reputation of mobile phone providers**. *Procedia Computer Science*, 72 :519 – 526. The Third Information Systems International Conference 2015.

[Vilares et al., 2014] Vilares, D., Hermo, M., Alonso, M. A., Gómez-Rodríguez, C., et Vilares, J. (2014). **Lys at clef replab 2014 : Creating the state of the art in author influence ranking and reputation classification on twitter.** Dans *CLEF (Working Notes)*, pages 1468–1478.

[Wang, 2010] Wang, A. H. (2010). **Don't follow me : Spam detection in twitter**. Dans *2010 International Conference on Security and Cryptography (SECRYPT)*, pages 1–10.

[Watts et al., 1998] Watts, D. J., et Strogatz, S. H. (1998). **Collective dynamics of 'small-world'networks**. *nature*, 393(6684) :440.

[Wei et al., 2015] Wei, B., Liu, J., Wei, D., Gao, C., et Deng, Y. (2015). **Weighted k-shell decomposition for complex networks based on potential edge weights**. *Physica A : Statistical Mechanics and its Applications*, 420 :277–283.

[Wei et al., 2013] Wei, D., Deng, X., Zhang, X., Deng, Y., et Mahadevan, S. (2013). **Identifying influential nodes in weighted networks based on evidence theory**. *Physica A : Statistical Mechanics and its Applications*, 392(10) :2564–2575.

[Weng et al., 2010] Weng, J., Lim, E.-P., Jiang, J., et He, Q. (2010). **TwitterRank : Finding Topic-sensitive Influential Twitterers**. Dans *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA. ACM.

[Weren et al., 2014] Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., et Wives, L. K. (2014). **Examining multiple features for author profiling.** *JIDM*, 5(3) :266–279.

[West et al., 2010] West, J. D., Bergstrom, T. C., et Bergstrom, C. T. (2010). **The eigenfactor metricstm : A network approach to assessing scholarly journals**. *College & Research Libraries*, 71(3) :236–244.

[Wiebe et al., 2005] Wiebe, J., Wilson, T., et Cardie, C. (2005). **Annotating Expressions of Opinions and Emotions in Language**. *Language Resources and Evaluation*, 39(2) :165–210.

[Williams et al., 2015] Williams, M. L., et Burnap, P. (2015). **Cyberhate on social media in the aftermath of woolwich : A case study in computational criminology and big data**. *British Journal of Criminology*, page azv059.

[Wittenbaum et al., 1999] Wittenbaum, G. M., Hubbell, A. P., et Zuckerman, C. (1999). **Mutual enhancement : Toward an understanding of the collective preference for shared information.** *Journal of Personality and Social Psychology*, 77(5) :967.

[Wong et al., 2016] Wong, F. M. F., Tan, C. W., Sen, S., et Chiang, M. (2016). **Quantifying political leaning from tweets, retweets, and retweeters**. *IEEE Transactions on Knowledge and Data Engineering*, 28(8) :2158–2172.

[Worden et al., 2009] Worden, K., Manson, G., et Denœux, T. (2009). **An evidence-based approach to damage location on an aircraft structure**. *Mechanical Systems and Signal Processing*, 23(6) :1792 – 1804. Special Issue : Inverse Problems.

[Wu et al., 2013] Wu, Z., Yin, W., Cao, J., Xu, G., et Cuzzocrea, A. (2013). **Web information systems engineering – wise 2013 : 14th international conference, nanjing, china, october 13-15, 2013, proceedings, part ii**. chapitreCommunity Detection in Multi-relational Social Networks, pages 43–56. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Xu et al., 2014] Xu, P., Davoine, F., Bordes, J.-B., et Denoeux, T. (2014). **Fusion d'informations pour la compréhension de scènes**. *Traitement du Signal*, 31(1-2) :57–80.

[Xu et al., 2016] Xu, P., Davoine, F., Bordes, J.-B., Zhao, H., et Denœux, T. (2016). **Multimodal information fusion for urban scene understanding**. *Machine Vision and Applications*, 27(3) :331–349.

[Yang et al., 2016] Yang, Y., Wang, Z., Pei, J., et Chen, E. (2016). **Tracking Influential Nodes in Dynamic Networks**. *CoRR*, abs/1602.04490.

[Zhang et al., 2014] Zhang, H., Mishra, S., Thai, M. T., Wu, J., et Wang, Y. (2014). **Recent advances in information diffusion and influence maximization in complex social networks**. Dans *Opportunistic Mobile Social Networks*, chapitre 2, pages 37–69. CRC Press.

[Zhao et al., 2017] Zhao, X., Liu, F., Wang, J., et Li, T. (2017). **Evaluating Influential Nodes in Social Networks by Local Centrality with a Coefficient**. *International Journal of Geo-Information (ISPRS)*, 6(2).

[Zhao et al., 2015] Zhao, Z., Wang, X., Zhang, W., et Zhu, Z. (2015). **A community-based approach to identifying influential spreaders**. *Entropy*, 17(4) :2228–2252.

[Zhou et al., 2007] Zhou, D., Orshanskiy, S. A., Zha, H., et Giles, C. L. (2007). **Co-ranking authors and documents in a heterogeneous network**. Dans *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 739–744. IEEE.

[Zi et al., 2012] Zi, C., Steven, G., Haining, W., et Sushil, J. (2012). **Detecting automation of twitter accounts : Are you a human, bot, or cyborg ?** *IEEE Transactions on Dependable and Secure Computing*, 9(6) :811–824.

[Zlatić et al., 2009] Zlatić, V., Ghoshal, G., et Caldarelli, G. (2009). **Hypergraph topological quantities for tagged social networks**. *Physical Review E*, 80(3) :036118.

[Zuo et al., 2011] Zuo, X.-N., Ehmke, R., Mennes, M., Imperati, D., Castellanos, F. X., Sporns, O., et Milham, M. P. (2011). **Network centrality in the human functional connectome**. *Cerebral cortex*, 22(8) :1862–1875.

TABLE DES FIGURES

LISTE DES TABLES

LISTE DES DÉFINITIONS

IV

ANNEXES

A

DÉMONSTRATIONS MATHÉMATIQUES

Cette annexe est dédiée aux démonstrations mathématiques, d'abord, nous traitons les propriétés liées à la règle de combinaison, ensuite, nous présentons l'étude de la convergence de l'influence dans *TwitBelief*.

A.1/ PROPRIÉTÉS LIÉES À LA RÈGLE DE COMBINAISON

Dans ce qui suit, nous démontrons deux propriétés importantes de la règle de combinaison $\otimes$ avec la fonction symétrique $@$ qui remplace l'opérateur d'intersection dans la règle de combinaison conjonctive classique (équation 4.4).

Propriété 1 La combinaison de deux fonctions de masse est une autre fonction de masse.

Preuve. Notons $(m \otimes m')$ comme m''. Il est facile de voir que pour tout x nous avons $m''(x) \geqslant 0$, parce que nous calculons m'' en utilisant seulement la multiplication et l'addition de nombres non-négatifs. Après, nous montrons que $\sum_{z \in \Omega_{Inf}} m''(z) = 1$.

Soit $\Omega_{Inf}{}^2_z = \{(x,y) \in \Omega_{Inf} : x@y = z\}$ et nous procédons comme suit :

$$\begin{aligned}\sum_z m''(z) &= \sum_z \sum_{x@y=z} m(x)m'(y) \\ &= \sum_z \sum_{(x,y)\in\Omega_{Inf}{}^2_z} m(x)m'(y)\,.\end{aligned}$$

Notons que $\Omega_{Inf}{}^2_z \neq \Omega_{Inf}{}^2_{z'} \iff z \neq z'$, et $\bigcup_{z\in\Omega_{Inf}} \Omega_{Inf}{}^2_z = \Omega^2_{Inf}$. Donc, nous pouvons omettre $\sum_z$ et réécrivons comme suit :

$$\begin{aligned}&= \sum_{(x,y)\in\Omega^2_{Inf}} m(x)m'(y) \\ &= \sum_x \sum_y m(x)m'(y) \\ &= \sum_x m(x) \sum_y m'(y)\end{aligned}$$

m et m' sont des fonctions de masse : $\sum_x m(x) = \sum_y m'(y) = 1$, alors $\sum_x m(x) \sum_y m'(y) = 1$

□.

Propriété 2 En général $\otimes$ est non-associative : $(m \otimes m') \otimes m'' \neq m \otimes (m' \otimes m'')$

Preuve. Considérons $\Omega = \{A, B, C\}$, et l'opération $@$ suivante :

$@$	***A***	***B***	***C***	Ω
A	B	B	C	A
B	B	C	C	B
C	C	C	C	C
Ω	A	B	C	Ω

$$m = m' = \begin{array}{|c|c|c|c|} \hline A & B & C & \Omega \\ \hline 1 & 0 & 0 & 0 \\ \hline \end{array}$$

$$m'' = \begin{array}{|c|c|c|c|} \hline A & B & C & \Omega \\ \hline 0 & 1 & 0 & 0 \\ \hline \end{array}$$

Il est facile de voir que :

$$(m \otimes m') \otimes m'' = \begin{array}{|c|c|c|c|} \hline A & B & C & \Omega \\ \hline 0 & 0 & 1 & 0 \\ \hline \end{array}$$

$$m \otimes (m' \otimes m'') = \begin{array}{|c|c|c|c|} \hline A & B & C & \Omega \\ \hline 0 & 1 & 0 & 0 \\ \hline \end{array}$$

Ainsi, en général :

$$(m \otimes m') \otimes m'' \neq m \otimes (m' \otimes m'')$$

□.

A.2/ ÉTUDE DE LA CONVERGENCE DE L'OPÉRATION @

Dans les travaux de recherche utilisant la théorie des fonctions de croyance, la question de la convergence de l'application itérative de l'opérateur $\otimes$ n'a pas encore été envisagée, malgré son importance pour les recherches contemporaines où les réseaux sont complexes et présentent plusieurs aspects à combiner selon la problématique étudiée. Dans le cas de *TwitBelief*, le nombre cumulé de relations tels que les retweets, *réponses* et *mentions* augmente avec le temps. Même si nous nous limitons à une période de temps spécifique, ce nombre peut être très important (des centaines de milliers) pour certains utilisateurs. Dans cette démonstration, en utilisant la théorie des chaînes de Markov, nous étudions la convergence de l'influence suite à l'application itérative de l'opérateur $\otimes$.

A.2.1/ LA FONCTION DE CROYANCE GÉNÉRALISÉE

La *règle de combinaison généralisée* ($\otimes$) est une version modifiée de la règle de combinaison classique proposée dans *TwitBelief* (voir section 4.2 du chapitre 4) :

au lieu d'utiliser l'opérateur d'intersection, nous considérons un opérateur symétrique $@ : \Omega_{Inf} \times \Omega_{Inf} \to \Omega_{Inf}$.

$$(m_{r_1} \otimes m_{r_2})(z) = \sum_{y @ x = z} m_{r_1}(x) m_{r_2}(y), \quad x, y, z \in \Omega_{Inf} \qquad \text{(règle de combinaison modifiée)}$$

Soit $m^{\otimes n} = \underbrace{\Big(((m \otimes m) \otimes m) \otimes \cdots \otimes m\Big)}_{n \text{ fois}}$, l'expression étant évaluée de la gauche vers la droite, alors nous pouvons omettre les parenthèses dans ce qui suit.

Considérons la limite $\lim_{n\to\infty} m^{\otimes n} : \Omega_{Inf} \to [0,1]$ définit par :

$$\lim_{n\to\infty} m^{\otimes n} = K \iff \exists K : \Omega_{Inf} \to [0,1],\ \forall x \in \Omega_{Inf},\ \lim_{n\to\infty} m^{\otimes n}(x) = K(x)$$

Dans ce qui suit, nous montrons comment l'existence de cette limite dépend de m et $@$.

A.2.2/ CHAÎNES DE MARKOV

Dans cette section, nous rappelons les notions de base de la théorie des chaînes de Markov. Des informations et des preuves supplémentaires peuvent être trouvées dans le chapitre 4 de [Gallager, 2013].

Soit I un *ensemble d'états*. Une *distribution initiale* $\tau : I \to [0,1]$ est une distribution de probabilité définie sur l'ensemble des états, c'est-à-dire $\sum_{x \in I} \tau(x) = 1$. Notons $x \to y$ une *transition* de l'état x à l'état y, nous remarquons que $(x \to y) \in I^2$. Nous notons $x \circlearrowleft$ au lieu de $x \to x$.

Une *distribution de transition* $T : I^2 \to [0,1]$ est une distribution de probabilité définie sur toutes les transitions, telle que pour tout $x \in I$ nous avons $\sum_{y \in I} T(x \to y) = 1$.

Une *chaîne de Markov* est un triplet (I, τ, T). Un processus de Markov peut être décrit par :

- Cas de base : Le processus commence à l'état x avec la probabilité $\tau(x)$.
- Étape inductive : Le processus passe de l'état y à l'état z avec la probabilité $T(y \to z)$.

La distribution de transition peut être vue comme une matrice stochastique T [1]. Nous définissons $T_{i,j} = T(i \to j)$ et représentons la distribution initiale τ par un vecteur. En utilisant la notation matricielle, le processus de Markov décrit ci-dessus, peut être écrit comme τT^n. L'étude de la convergence de la chaîne de Markov se traduit naturellement par l'étude de $\lim_{n\to\infty} \tau T^n$.

Désignons par $x \rightsquigarrow y$ la séquence de transitions avec des probabilités strictement positives qui commence à x et se termine à y. Nous appelons cette séquence de transitions un *chemin*. En général, il existe plusieurs chemins de x à y, si nous voulons les distinguer, nous utilisons la notion suivante : $x \rightsquigarrow_1 y$, $x \rightsquigarrow_2 y$, etc. Afin de désigner l'ensemble de tous les chemins nous utilisons $P_{x \rightsquigarrow y}$.

1. Une matrice est stochastique si la somme de chaque ligne est égale à 1

Notons par $|p|$ le nombre de transitions dans un chemin p. La *période* $d(x)$ d'un état x est définie comme $d(x) = \gcd(\{|p| : p \in P_{x \rightsquigarrow x}\})$. Un état x est dit *périodique* si $d(x) > 1$. Un état x est appelé *apériodique* si $d(x) = 1$. Parfois, il n y a pas de chemins entre x et x, dans ce cas $d(x) = \gcd(\varnothing)$ est indéfinie, nous disons aussi que cet état x est apériodique. On dit qu'une chaîne de Markov est apériodique lorsque tous ses états sont apériodiques.

Un *état absorbant* t, noté par $\boxed{\text{t}}$, est un état sans chemins sortants, sauf l'auto-boucle $\boxed{\text{t}} \circlearrowleft$. Tous les états absorbants sont apériodiques, mais l'inverse n'est pas vrai en général.

Lorsque nous avons $x \rightsquigarrow y$ et $y \rightsquigarrow x$, nous disons que x et y *communiquent*. De plus, nous disons que x communique avec lui-même. Une *classe de communication* est un sous-ensemble maximal d'états $J \subseteq I$ tel que n'importe quel paire d'états de J communique. De cette façon, l'ensemble de tous les états est partitionné en plusieurs sous-ensembles disjoints mutuellement (classes de communication) $I = \bigcup_{i=1}^{k} J_i$, où k est le nombre de classes de communication. Nous désignons par $J[x]$ la classe de communication de x.

Une chaîne de Markov est dite *irréductible* s'il n'y a qu'une seule classe de communication, c'est-à-dire tous les paires d'états $x, y \in I$ communiquent. Autrement, une chaîne de Markov est appelée *réductible* et peut être réduite en plusieurs classes de communication irréductibles. Il est utile de représenter la structure d'une chaîne de Markov réductible comme un graphe acyclique orienté, où les nœuds sont les classes de communication, et où il y a un lien orienté $J_i \to J_k$ entre deux communications de classes différentes J_i et J_k si et seulement s'il existe une transition avec une probabilité strictement positive $x \to y$ pour certains $x \in J_i$ et $y \in J_k$. Une classe de communication qui n'a pas de liens sortants est appelée une *classe absorbante*. Quand une classe absorbante contient un seul état, cet état est absorbant. Si nous quittons une classe, nous ne pouvons pas revenir en arrière. Ainsi, après un assez grand nombre de transitions, seules les classes absorbantes sont importantes.

Une chaîne de Markov avec une distribution de transition T et une distribution initiale τ est dite *convergente* lorsque $\lim_{n \to \infty} \tau T^n$ existe. Quand cette limite existe et ne dépend pas de τ, nous disons que la chaîne de Markov possède une *distribution limite unique*.

Une *distribution stationnaire* π est un vecteur tel que $\pi T = \pi$. La distribution limite unique est stationnaire. Toutes les chaînes de Markov finies ont une distribution stationnaire. Mais certaines chaînes de Markov n'ont pas la distribution limite unique.

Des résultats classiques sur la convergence des chaînes de Markov finies [Gallager, 2013] peuvent être résumés de la manière suivante :

Finie irréductible apériodique	$\Longrightarrow$	Converge. La limite *ne dépend pas* de la distribution initiale.
Finie apériodique	$\Longrightarrow$	Converge. La limite *dépend* de la distribution initiale. Seuls les états absorbants peuvent avoir des masses positives à la limite.
Tous les états sont périodiques	$\Longrightarrow$	Ne converge pas *en général* (mais converge quand la distribution initiale est stationnaire).
Seulement quelques états périodiques	$\Longrightarrow$	Il faut examiner la structure du graphe acyclique dirigé des classes de communication et voir ce qui se passe dans les classes absorbantes : certaines classes absorbantes peuvent être périodiques, d'autres peuvent ne pas l'être ; selon la distribution initiale, certaines classes absorbantes peuvent avoir une masse nulle à la limite.

A.2.3/ QUESTION DE CONVERGENCE

Dans cette section, nous répondons à la question de l'existence de $\lim_{n\to\infty} m^{\otimes n}$.

A.2.3.1/ DE @ VERS UNE CHAÎNE DE MARKOV

En utilisant $@$, $\otimes$ et une fonction de masse m, nous construisons une chaîne de Markov, notée $M_@$, comme suit :

- Les éléments de Ω_{Inf} sont les états de la chaîne
- m est la distribution initiale
- La probabilité de transition $T(x \to z)$ est définie comme $T(x \to z) = \sum_{x@y=z, y\in\Omega_{Inf}} m(y)$.

  Par exemple, $x@y = z$ et $x@y' = z$ correspondent à la partie de la chaîne de Markov présentée dans la figure A.1. Notons que $\sum_{z\in\Omega_{Inf}} T(x \to z) = \sum_{x@y=z;\, y,z\in\Omega_{Inf}} m(y) = 1$.

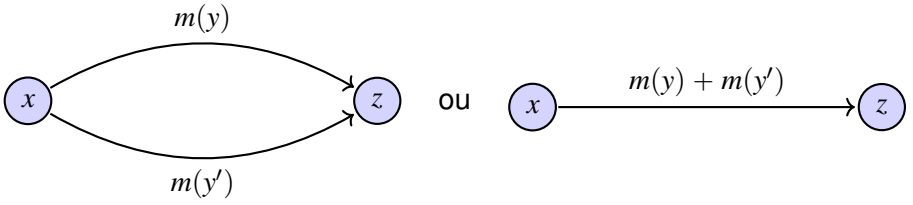


FIGURE A.1 – Une partie de la chaîne de Markov obtenue dans le cas de $x@y = z$ et $x@y' = z$.

Observation A.2.1. *L'existence de* $x \xrightarrow{m(y)} z$ *implique l'existence de* $y \xrightarrow{m(x)} z$ *parce que* $@$ *est symétrique. Dans le cas particulier* $x = y$*, ces deux transitions coïncident et nous avons seulement une transition* $x \xrightarrow{m(x)} z$.

Rappelons que $\otimes$ est associative à gauche, et observons que

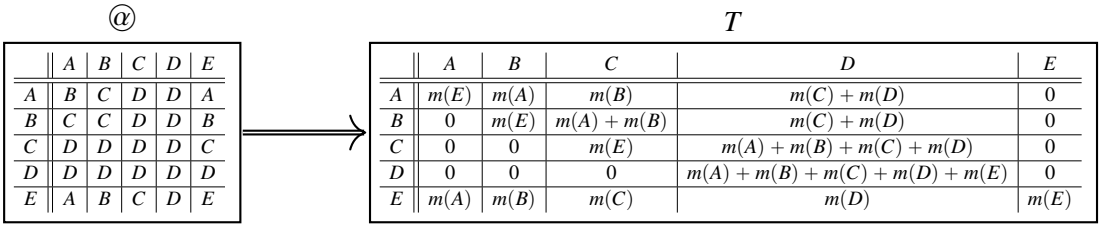
$$m \underbrace{\overbrace{\otimes m}^{T}\ \overbrace{\otimes m}^{T} \cdots \overbrace{\otimes m}^{T}}_{\text{n fois}} = mT^n$$

Ainsi, toute question portant sur $\lim_{n\to\infty} m^{\otimes n}$ se traduit naturellement par une question sur $\lim_{n\to\infty} mT^n$.

Exemple. Considérons $\Omega_{Inf} = \{A, B, C, D, E\}$, soit $@$ définit comme suit :

	A	B	C	D	E
A	B	C	D	D	A
B	C	C	D	D	B
C	D	D	D	D	C
D	D	D	D	D	D
E	A	B	C	D	E

La figure A.2a représente la matrice de transition construite à partir de $@$. Une construction partielle de la chaîne de Markov est représentée sur la figure A.2b. La chaîne de Markov complète peut être visualisée dans la figure A.2c.



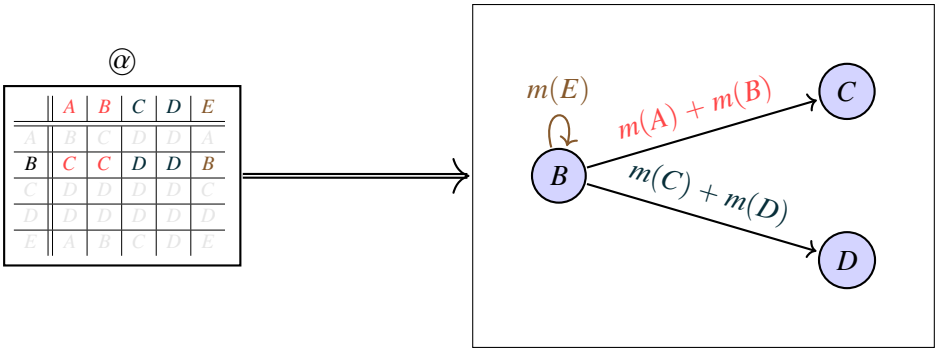
$@$

	A	B	C	D	E
A	B	C	D	D	A
B	C	C	D	D	B
C	D	D	D	D	C
D	D	D	D	D	D
E	A	B	C	D	E

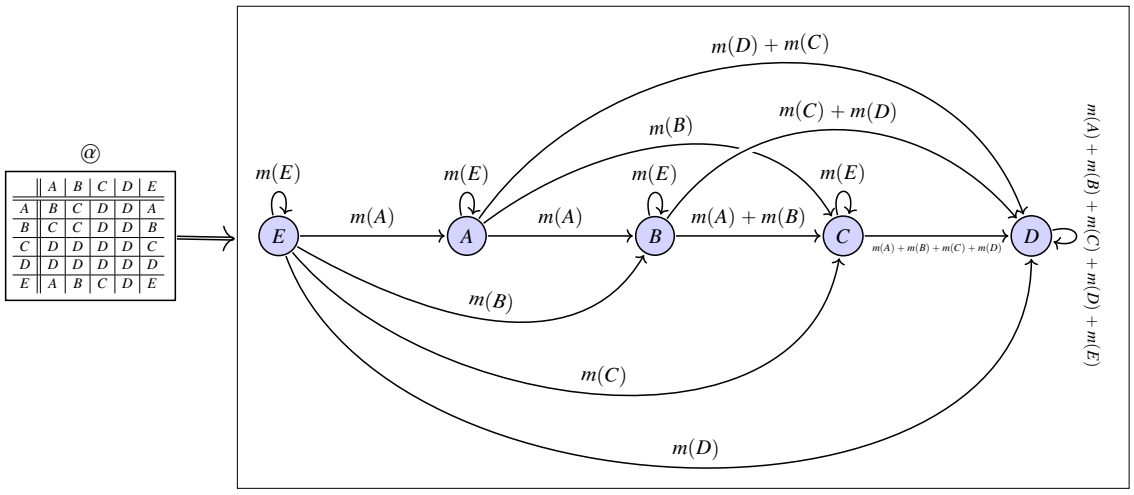
T

	A	B	C	D	E
A	$m(E)$	$m(A)$	$m(B)$	$m(C)+m(D)$	0
B	0	$m(E)$	$m(A)+m(B)$	$m(C)+m(D)$	0
C	0	0	$m(E)$	$m(A)+m(B)+m(C)+m(D)$	0
D	0	0	0	$m(A)+m(B)+m(C)+m(D)+m(E)$	0
E	$m(A)$	$m(B)$	$m(C)$	$m(D)$	$m(E)$

(a) Matrice de transition construite à partir de $@$



(b) Vue partielle de la chaîne de Markov



(c) Vue complète de la chaîne de Markov

FIGURE A.2 – Une chaîne de Markov construite à partir de $@$

A.2.3.2/ PROPRIÉTÉS DE LA CHAÎNE DE MARKOV CONSTRUITE

En utilisant $Ⓐ : \Omega_{Inf}^2 \to \Omega_{Inf}$, nous définissons une relation binaire transitive $\preceq$ sur Ω_{Inf} comme suit :

$$x Ⓐ y = z \quad \Rightarrow \quad x \lhd z \text{ et } y \lhd z \qquad (\preceq)$$
$$\text{avec } \preceq \text{ une fermeture transitive de } \lhd$$

Observation A.2.2. *Supposons que toutes les masses soient positives, c'est-à-dire* $\forall x \in \Omega_{Inf} : m(x) > 0$ *et considérons une paire d'états* $x, y \in \Omega_{Inf}$*, il n'est pas difficile de voir qu'il existe un chemin* $x \rightsquigarrow y$ *si et seulement si* $x \preceq y$*.*

La chaîne de Markov est dite *faiblement connectée* lorsque son graphe sous-jacent non orienté est connecté. Deux états x et y sont faiblement connectés, s'il existe un chemin entre eux dans le graphe sous-jacent non orienté.

Proposition A.2.3. *Lorsque toutes les masses sont positives, la chaîne de Markov* $M_Ⓐ$ *est faiblement connectée.*

Démonstration. Considérons n'importe quelle paire d'états $x, y \in \Omega_{Inf}$. Rappelons que $Ⓐ : \Omega_{Inf}^2 \to \Omega_{Inf}$, alors il existe $z \in \Omega_{Inf}$ tel que $x Ⓐ y = z$. Par la définition de $\preceq$, nous avons $x \preceq z$ and $y \preceq z$. En utilisant l'observation A.2.2, nous obtenons $x \rightsquigarrow z$ et $y \rightsquigarrow z$. Ainsi, x et y sont faiblement connectés via z. □

La situation est légèrement plus compliquée si certaines masses sont nulles. Considérons, par exemple, la chaîne de Markov de la figure A.2c : soit $m(E) = 1$, et voyons qu'il n'y a pas de chemins entre les différents états. Pour étudier les cas où seules quelques masses sont positives, nous procédons comme suit :

Soit $A(x)$ l'ensemble des états accessibles à partir de l'état x, c'est-à-dire $A(x) = \{y \in \Omega_{Inf} \text{ t.q. } \exists x \rightsquigarrow y\} \cup \{x\}$. Soit $\mathcal{A}(\Upsilon) = \bigcup_{x \in \Upsilon} A(x)$, notons par $M_Ⓐ^\Upsilon$ la chaîne de Markov créée à partir de $M_Ⓐ$ en supprimant tous les états qui ne sont pas dans $\mathcal{A}(\Upsilon)$ avec toutes les transitions correspondantes (c'est-à-dire que nous supprimons une transition $x \to y$ if $x \notin \mathcal{A}(\Upsilon)$ ou $y \notin \mathcal{A}(\Upsilon)$), et en fixant $m(x) = 0$ pour tout $x \in \Omega_{Inf} \backslash \Upsilon$. Cette sous-chaîne se produit lorsque tous les états de $\Omega_{Inf} \backslash \Upsilon$ ont des masses nulles.

Avant de prouver que la sous-chaîne $M_Ⓐ^\Upsilon$ est faiblement connectée pour tout $\Upsilon \subseteq \Omega_{Inf}$, nous prouvons le lemme suivant :

Lemma A.2.4. *Toute paire d'états* $x, y \in \Omega_{Inf}$ *avec des masses positives est faiblement connecté.*

Démonstration. Nous avons $x Ⓐ y = z$ pour $z \in \Omega_{Inf}$. Par la construction de la chaîne de Markov nous avons $x \xrightarrow{m(y)} z \xleftarrow{m(x)} y$. Nous avons aussi $m(x) > 0$ and $m(y) > 0$, alors ils existent $x \rightsquigarrow z$ et $y \rightsquigarrow z$. Donc, x et y sont faiblement connectés. □

Proposition A.2.5. *Pour tout* $\Upsilon \subseteq \Omega_{Inf}$*, la sous-chaîne* $M_Ⓐ^\Upsilon$ *est faiblement connectée.*

Démonstration. Considérons n'importe quelle paire d'états différents x, y de $M_Ⓐ^\Upsilon$. Quatre possibilités existent :

— Lorsque $m(x) > 0$ et $m(y) > 0$, nous utilisons le Lemme A.2.4 pour montrer que x et y sont faiblement connectés.

— Lorsque $m(x) > 0$ et $m(y) = 0$, il y a deux sous-cas :
  — Si $y \in A(x)$, nous avons $x \rightsquigarrow y$.
  — Sinon, il existe un état $z \neq x$ avec une masse positive, tel que $y \in A(z)$ et $z \in \Upsilon$, puisque la chaîne de Markov $M_{@}^{\Upsilon}$ contient seulement des états de $\bigcup_{u \in \Upsilon} A(u)$. Alors, nous avons $z \rightsquigarrow y$ par le Lemme A.2.4 et x et z sont faiblement connectés. Ainsi, x et y sont faiblement connectés.
— Le cas où $m(y) > 0$ et $m(x) = 0$ est symétriquement équivalent au précédent.
— Lorsque $m(x) = m(y) = 0$ nous avons deux sous-cas :
  — Il existe un état z avec une masse positive tel que $x, y \in A(z)$. Dans ce cas, nous avons $z \rightsquigarrow x$ et $z \rightsquigarrow y$. Ainsi, x et y sont faiblement connectés.
  — Ils existent deux états différents z et w avec des masses positives tels que $x \in A(z)$ et $y \in A(w)$. Alors, nous avons $z \rightsquigarrow x$, $w \rightsquigarrow y$. Par le Lemme A.2.4, z et w sont faiblement connectés. Ainsi, x et y sont faiblement connectés.

□

A.2.3.3/ POSET D'ÉTATS NON NÉCESSAIREMENT RÉFLEXIF

Par définition, $\preceq$ est une relation binaire transitive définie sur Ω_{Inf}. Parfois, cette relation est antisymétrique, c'est-à-dire si $x \preceq y$ et $y \preceq x$ alors $x = y$. Un ensemble avec une relation binaire transitive antisymétrique est appelé *poset non nécessairement réflexif* : pour quelques x la relation est réflexive, c'est-à-dire $x \preceq x$, mais pour certain x, elle est irréflexive, $x \npreceq x$.

Proposition A.2.6. *Quand $(\Omega_{Inf}, \preceq)$ est un poset non nécessairement réflexif, alors il existe un élément maximal unique dans ce poset.*

Démonstration. Supposons qu'il y a deux éléments maximaux différents x et y. Rappelons que $@ : \Omega_{Inf}^2 \to \Omega_{Inf}$, alors il y a $z \in \Omega_{Inf}$ tel que $x@y = z$. Selon la définition de $\preceq$, nous avons $x \preceq z$. Puisque, l'élément x est maximal, alors $z = x$, mais dans ce cas nous avons $x@y = x$ et $y \preceq x$, ainsi, y ne peut pas être maximal. □

Proposition A.2.7. *Quand $(\Omega_{Inf}, \preceq)$ est un poset non nécessairement réflexif, la chaîne de Markov $M_{@}$ converge toujours.*

Démonstration. Rappelons que $x \preceq z$, obtenu à partir de $x@y = z$, correspond à $x \xrightarrow{m(y)} z$ dans la chaîne de Markov. Quand $(\Omega_{Inf}, \preceq)$ est un poset non nécessairement réflexif, il n ya pas de cycles dans la chaîne de Markov en plus des auto-boucles, parce que s'il y a un cycle $x \to z \to \ldots \to x$ alors nous devons avoir

1. $x \preceq z \preceq \ldots \preceq x$ (par transitivité)
2. $x = z$ (par antisymétrie)

Ainsi, tout état de la chaîne de Markov est apériodique. Les chaînes apériodiques finies convergent toujours. □

Quand $(\Omega_{Inf}, \preceq)$ est un poset non nécessairement réflexif, la limite $\lim_{n \to \infty} \tau T^n$ peut dépendre de la distribution initiale τ, mais seuls les états absorbants (c'est-à-dire avec auto-boucles et sans transitions sortantes de masse positive) peuvent avoir des masses positives à la limite. Considérons un état x avec une auto-boucle $x \circlearrowleft_{m(x)}$. Soit $m(x) = 1$: toute la

masse restera à l'état x pour toujours. S'il n'y a qu'un seul état t avec une auto-boucle $t \circlearrowleft_{m(t)}$, alors nous avons $(\lim_{n\to\infty} \tau T^n)(t) = 1$ pour toute distribution initiale τ. La chaîne de Markov de la figure A.2c peut avoir seulement deux états convergents E et D. De plus, $(\lim_{n\to\infty} \tau T^n)(E) = 1$ si et seulement si $m(E) = 1$.

A.2.3.4/ POSET STRICT DES CLASSES DE COMMUNICATION

Notons par $\mathcal{J}$ l'ensemble de toutes les classes de communication. Rappelons qu'une chaîne de Markov peut être représentée comme un graphe acyclique orienté (DAG), où les nœuds sont les classes de communication, et où il y a un lien orienté $J_i \to J_k$ entre deux classes de communication différentes J_i et J_k si et seulement s'il y a une transition avec une probabilité strictement positive $x \to y$ pour certains $x \in J_i$ et $y \in J_k$. Supposons que toutes les masses sont positives et définissons une relation binaire $<$ sur l'ensemble de toutes les classes de communication $\mathcal{J}$:

$$\begin{gathered} J_i \to J_k \quad \Rightarrow \quad J_i \lessdot J_k \\ \text{Soit } < \text{ une fermeture transitive de } \lessdot \end{gathered} \tag{$<$}$$

L'ensemble de toutes les classes de communication $\mathcal{J}$ avec $<$ forme un poset strict (irréflexif).

Proposition A.2.8. *Lorsque toutes les masses sur les états sont positives, c'est-à-dire $\forall x \in \Omega_{Inf} : m(x) > 0$, le DAG des classes de communication d'une chaîne de Markov $M_@$ est faiblement connecté.*

Démonstration. Le DAG des classes de communication est créé par la fusion de certains états, de sorte que la connectivité faible du DAG découle de la faible connectivité de la chaîne de Markov (voir la Proposition A.2.3). □

Proposition A.2.9. *Pour tout $\Upsilon \subseteq \Omega_{Inf}$ le DAG des classes de communication de la chaîne de Markov $M_@^\Upsilon$ est faiblement connecté.*

Démonstration. Cette preuve est identique à la preuve précédente, mais nous utilisons la Proposition A.2.5 au lieu de la Proposition A.2.3. □

Lemma A.2.10. *Considérons deux états x et y, si il existe un chemin $x \rightsquigarrow y$ alors nous avons seulement deux possibilités : $J[x] = J[y]$ ou $J[x] < J[y]$.*

Démonstration. Si il existe un chemin $x \rightsquigarrow y$ et si $J[x] \neq J[y]$, alors nous avons $J[x] < J[y]$ selon la définition du DAG des classes de communication. □

Proposition A.2.11. *Lorsque toutes les masses sur les états sont positives, le DAG des classes de communication de la chaîne de Markov $M_@$ a une seule classe absorbante. De manière équivalente, le poset $(\mathcal{J}, <)$ a un seul élément maximal.*

Démonstration. Supposons qu'il y a deux éléments maximum J et J' et prenons certains $x \in J$ et $y \in J'$. Considérons $x @ y = z$, par la construction de $M_@$ nous avons $x \rightsquigarrow z$. Alors, selon le Lemme A.2.10 nous obtenons $J[x] = J[z]$ ou $J[x] < J[z]$. Dans ce dernier cas, nous avons une contradiction (car nous supposons que J est maximal). Symétriquement, à partir de $y @ x = z$ nous obtenons $J[y] = J[z]$. Ainsi, $J[y] = J[z] = J[x]$, ce qui signifie que $J = J'$. □

S'il y a des états avec des masses nulles, le DAG des classes de communication de la chaîne de Markov construit à partir de Ω_{Inf} et $@$ peut avoir plusieurs états absorbants. Considérons la figure A.3, et soit $m(X) = 0.5, m(Y) = 0.5$. Dan ce cas, l'état Z aura la masse 0.5 à la limite, et, de plus, il y a deux classes de communication périodiques : $\{W, W'\}$ et $\{V, V'\}$. Il est facile de voir (en regardant seulement les transitions noires épaisses), que cette chaîne de Markov a trois classes absorbantes : Z, $\{W, W'\}$ et $\{V, V'\}$.

Le théorème suivant et ses corollaires découlent naturellement de nos propositions :

Theorem A.2.12. *$\lim_{n\to\infty} m^{\otimes n}$ existe si et seulement si toutes les classes absorbantes de la chaine de Markov $M_{@}^{\Upsilon}$ sont apériodiques, où $\Upsilon = \{x \in \Omega_{Inf} \text{ t.q.} m(x) > 0\}$. Seuls les états des classes absorbantes peuvent avoir des masses positives à la limite.*

Selon la proposition A.2.11, le DAG des classes de communication de la chaîne de Markov $M_{@}^{\Upsilon} = M_{@}$ contient seulement une classe absorbante quand toutes les masses sont positives. Ainsi, nous obtenons le corollaire suivant :

Corollary A.2.13. *Quand toutes les masses sont positives, c'est-à-dire $\Upsilon = \Omega_{Inf}$, $\lim_{n\to\infty} m^{\otimes n}$ existe si et seulement si la classe absorbante unique est apériodique.*

Corollary A.2.14. *Les affirmations suivantes sont équivalentes :*

- *$\lim_{n\to\infty} m^{\otimes n}$ existe pour tout m.*
- *Pour tout $\Upsilon \subseteq \Omega_{Inf}$ le DAG des classes de communication de la chaîne de Markov $M_{@}^{\Upsilon}$ a seulement des classes absorbantes apériodiques.*

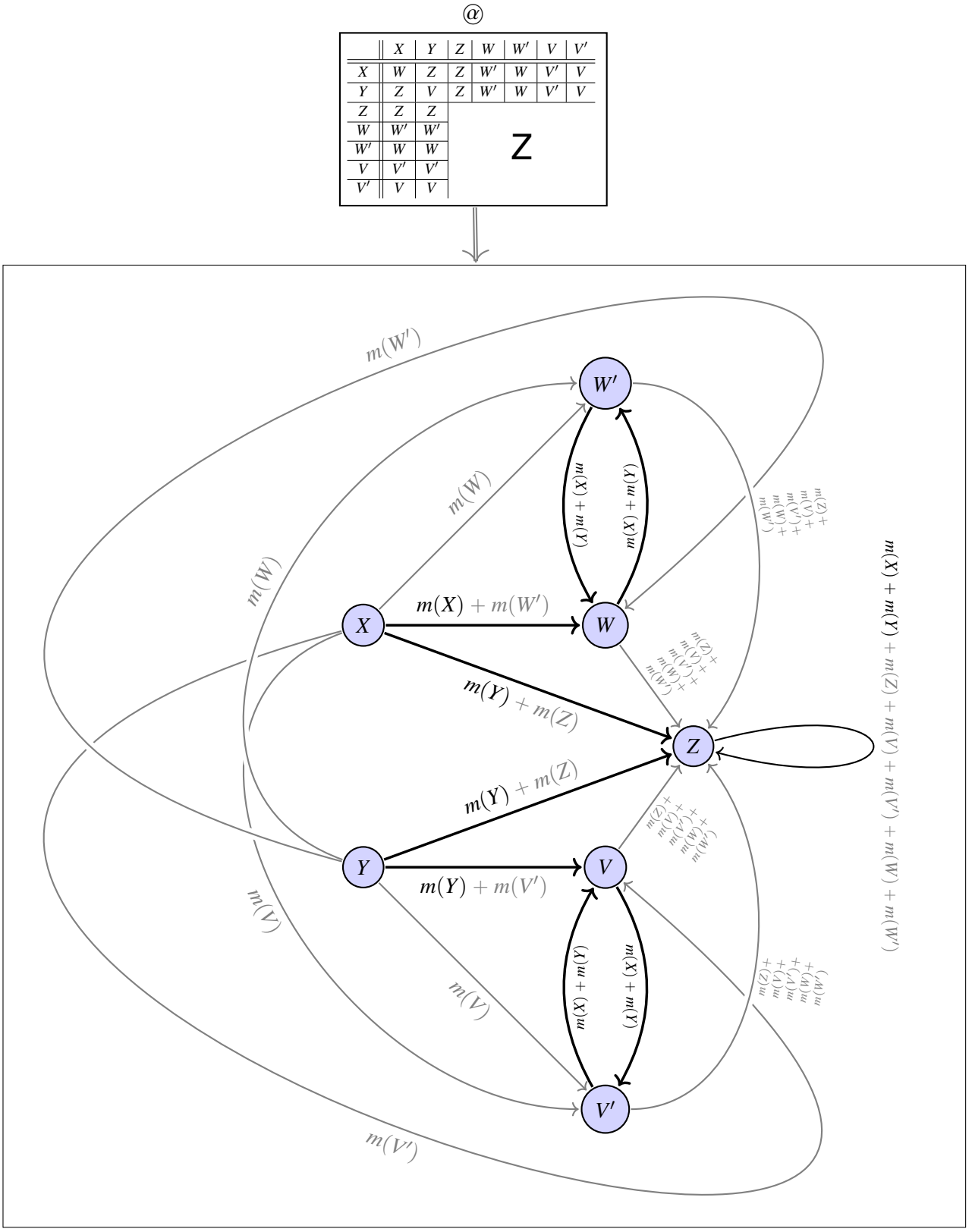


FIGURE A.3 – Une chaîne de Markov semblable à un papillon construite à partir de @. Lorsque tous les états ont des masses positives, Z est l'unique état absorbant. Mais si seulement $m(X) > 0$ et $m(Y) > 0$, il y a trois états absorbants : $Z, \{W, W'\}, \{V, V'\}$.

B

DÉTAILS ET EXEMPLES D'ILLUSTRATIONS DES EXTENSIONS DE *TwitBelief*

Cette annexe détaille les deux extensions présentées dans le chapitre 5. Nous expliquons les étapes suivies et nous représentons des exemples d'illustrations pour chaque extension.

B.1/ INFLUENCE POLARISÉE

Dans cette section, nous détaillons les étapes suivies pour adapter et modifier *TwitBelief* afin d'étudier l'influence polarisée. Les étapes suivies sont les suivantes :

1. **Modélisation du réseau :** D'abord, après avoir analysé le sentiment des *tweets*, pour chaque utilisateur, le réseau de *Twitter* est divisé en trois sous-réseaux $g_n \subset G, n \in 1,2,3$, chaque sous-réseau modélise les *tweets* d'une polarité : positive, négative et neutre. Après, l'initialisation des masses est effectuée pour chaque relation ou pattern à combiner dans l'estimation.
2. **Fusion des masses dans chaque sous-réseau :** Dans cette étape, *TwitBelief* est appliqué sur les trois sous-réseaux. Les masses de croyance sont combinées dans chaque sous-réseau en utilisant la formule 4.8 selon $@$ défini dans *TwitBelief*. Ainsi, après cette étape, chaque utilisateur u est représenté par une masse de croyance et un degré d'influence pour chaque sous-réseau.
3. **Choix des paramètres :** Pour chaque utilisateur, nous avons deux éléments à fusionner : une polarité et une influence (degré d'influence et masse de croyance). Ainsi, nous utilisons un nouveau cadre de discernement, Ω_{Pol} qui représente les différentes réponses possibles à notre question : "Quelle est l'influence polarisée d'un certain utilisateur ?" Soit Ω_{Pol} l'ensemble des degrés d'influences polarisées possibles :
$\Omega_{Pol} = \{$PositiveT.Faible, NeutreT.Faible, NégativeT.Faible, PositiveFaible,
NeutreFaible, NégativeFaible, PositiveA.Moyenne, NeutreA.Moyenne,
NégativeA.Moyenne, PositiveMoyenne, NeutreMoyenne, NégativeMoyenne, ...
PositiveE.Forte, NeutreE.Forte, NégativeE.Forte$\}$

Nous utilisons un sous-ensemble Ω_P de $2^{\Omega_{Pol}}$, précisément :
$\Omega_P = \{$PositiveT.Faible, NeutreT.Faible, NégativeT.Faible, PositiveFaible,
NeutreFaible, NégativeFaible, PositiveA.Moyenne, NeutreA.Moyenne,
NégativeA.Moyenne, PositiveMoyenne, NeutreMoyenne, NégativeMoyenne, ...
PositiveE.Forte, NeutreE.Forte, NégativeE.Forte, $\Omega_{Pol}\}$

4. **Fusion des masses des trois sous-réseaux :** Dans cette étape, les masses de croyances obtenues pour chaque sous-réseau sont fusionnées en utilisant la formule 4.8 selon la fonction $@_2$ (Tableau 5.2). Cette fonction donne l'intersection entre les couples polarités/degré d'influence possibles, ici, nous n'utilisons pas l'opération $@$ comme dans l'étape précédente car $@$ assure l'hypothèse que plus nous combinons des relations, plus l'influence est importante. Mais dans cette étape, les relations sont déjà combinées dans les trois sous-réseaux et l'objectif est d'obtenir l'influence polarisée globale.
5. **Probabilité pignistique :** Une fois que nous obtenons la distribution de masses de croyance de l'influence fusionnée des trois sous-réseaux de l'utilisateur, nous utilisons une version modifiée de la probabilité pignistique définie dans la formule 4.5 afin de prendre la décision à propos du degré d'influence globale polarisée. Dans notre cas, les masses de croyance sont définies sur Ω_P et la probabilité pignistique est calculée en répartissant uniformément la masse de Ω_{Pol} sur tous les autres éléments de Ω_P :

$$\mathrm{bet}_2(x) = m(x) + \frac{m(\Omega_{Pol})}{|\Omega_P|}, \quad x \in \Omega_P \tag{B.1}$$

6. **Influence polarisée :** Enfin, l'influence polarisée est représentée par la polarité Pol_u ayant le degré d'influence maximal, en plus du degré d'influence Inf_u et de la probabilité pignistique Pg_u correspondants.

L'algorithme 3 formalise les étapes suivies pour estimer l'influence polarisée, il requiert en entrée, les trois sous-réseaux $g_n \subset G, n \in 1,2,3$, l'ensemble des relations $r \in R$ et l'initialisation des masses pour les différentes relations m_r ainsi que $@$ et $@_2$. Pour chaque utilisateur, l'algorithme commence par mesurer l'influence relative à chaque sous-réseau, c'est-à-dire l'influence relative à chaque polarité en utilisant le principe de *TwitBelief*. Enfin, en utilisant la formule 4.8 mais avec l'opération $@_2$, l'algorithme calcule l'influence polarisée globale en fusionnant les masses d'influence résultant de chaque sous-réseau. L'algorithme renvoie l'influence polarisée finale : la polarité Pol_u qui est la polarité ayant le degré d'influence le plus élevé ; le degré d'influence Inf_u et la probabilité pignistique correspondante Pg_u.

Algorithme 3 : Estimation de l'influence polarisée

Input : $g_n \subset G, n \in 1,2,3$, les sous-réseaux
L'ensemble des relations $R = r_1, r_2, \ldots$
Initialisation des masses $m_r, r \in R$
Fonctions $@$, $@_2$

Output : Polarité Pol_u, degré d'influence Inf_u, et la probabilité pignistique Pg_u pour chaque utilisateur $u \in U$

1 **pour** $u \in U$ **faire**
2 **pour** $g_n \subset G$ **faire**
3 **pour** $i \in [1..|R|]$ **faire**
4 ℓ_{u,r_i,g_n} := nombre de relations ou patterns de type r_i pour l'utilisateur u, dans le sous-réseau g_n ;
5 $M_{u,r_i,g_n} := m_{r_i}$;
6 **pour** $i \in [2..|\ell_{u,r_i,g_n}|]$ **faire**
7 $M_{u,r_i,g_n} := M_{u,r_i,g_n} \otimes m_{r_i}$; //selon $@$
8 **fin**
9 **fin**
10 $M_{u,g_n} := M_{u,r_1,g_n}$;
11 **pour** $i \in [2..|R|]$ **faire**
12 $M_{u,g_n} := M_{u,g_n} \otimes M_{u,r_i,g_n}$; //selon $@$
13 **fin**
14 **fin**
15 $M_{u1,2,3} := M_{u,g_1} \otimes M_{u,g_2} \otimes M_{u,g_3}$; // selon $@_2$
16 Bet_{2_u} := distribution de la probabilité pignistique obtenue en utilisant l'équation B.1 ;
17 Pol_u := polarité ayant le degré d'influence le plus élevé ;
18 Inf_u := degré d'influence correspondant à Pol_u ;
19 Pg_u := probabilité pignistique correspondante à Pol_u ;
20 **fin**
21 retour $Pol_u, \mathrm{Inf}_u, Pg_u, u \in U$;

Exemple illustration : Pour illustrer la fusion des masses des trois sous-réseaux représentant les polarités (Lignes 15-20 de l'algorithme 3), nous présentons un exemple d'un utilisateur ayant les masses suivantes :

$$\text{Utilisateur} \left\{ \begin{array}{r} \text{Positive, Moyenne, m = 0.3} \\ \text{Neutre, A.Moyenne, m = 0.7} \\ \text{Négative, Faible, m = 0.9} \end{array} \right\}$$

Comme expliqué dans la section 5.1.3, nous commençons par combiner les polarités positives et négatives. Le tableau B.1 montre la combinaison de masses des deux sous-réseaux représentant les polarités Positive et Négative. Nous commençons par combiner les couples polarités/influence en utilisant $@_2$, par exemple, l'intersection entre les couples Positive, Moyenne(+Moy) et Négative, Faible(-F) est Positive, Faible (+F).

TABLE B.1 – Combinaison des masses des sous-réseaux des polarités Positive et Négative

$\otimes$ selon $@_2$	Positive, Moyenne (+Moy) 0.3	Ω_{Pol} 0.7
Négative, Faible (-F) 0.9	Positive, Faible (+F) 0.27	Négative, Faible (-F) 0.63
Ω_{Pol} 0.1	Positive, Moyenne (+Moy) 0.03	Ω_{Pol} 0.07

Nous déduisons ainsi le degré et la masse correspondants à chaque polarité, par exemple, pour la polarité Positive, nous avons obtenu deux degrés d'influence avec des masses différentes dans le tableau B.1, pour les fusionner, nous regardons l'intersection entre les deux polarités/influences +A.Moy et +F dans $@_2$, puis nous effectuons la somme des deux masses 0.27 + 0.03 = 0.3. Nous obtenons alors :

$$\left\{\begin{array}{r}\text{Positive, Moyenne(+A.Fo), m = 0.3}\\ \text{Négative, Faible(-F), m = 0.63}\\ \Omega_{Pol}\text{ , m = 0.07}\end{array}\right\}$$

Maintenant, nous combinons les masses de la fusion des polarités Positive et Négative avec les masses de la polarité Neutre, les résultats sont donnés dans le tableau B.2.

TABLE B.2 – Combinaison des masses des trois sous-réseaux

$\otimes$ selon $@_2$	Positive, A.Forte (+A.Fo) 0.3	Négative, Faible (-F) 0.63	Ω_{Pol} 0.07
Neutre, A.Moyenne (=A.Moy) 0.7	Positive, Moyenne (+Moy) 0.21	Négative, T.Faible (-T.F) 0.441	Neutre, A.Moyenne (=A.Moy) 0.049
Ω_{Pol} 0.3	Positive, A.Forte (+A.Fo) 0.09	Négative, Faible (-F) 0.189	Ω_{Pol} 0.021

Nous obtenons alors : $\left\{\begin{array}{r}\text{Positive, Forte (+Fo), m = 0.3}\\ \text{Neutre, A.Moyenne(=A.Moy), m = 0.049}\\ \text{Négative, A.Moyenne(-A.Moy), m = 0.63}\\ \Omega_{Pol}\text{ ,m = 0.021}\end{array}\right\}$

Finalement, pour prendre la décision sur le degré d'influence polarisée, nous calculons la probabilité pignistique en utilisant la formule (B.1) (Tableau B.3). Par exemple, pour la polarité Positive, nous procédons comme suit pour obtenir la probabilité pignistique :

$$\text{bet}_2(\text{Positive, Forte}) = m(\text{Positive, Forte}) + \frac{m(\Omega_{Pol})}{|\Omega_P|} = 0.3 + \frac{0.021}{3} = 0.307$$

TABLE B.3 – Probabilité pignistique

Positive, Forte	Neutre, A.Moyenne	Négative, A.Moyenne
0.307	0.056	0.637

Nous pouvons conclure que l'influence polarisée est Positive avec le degré Forte et la probabilité pignistique 0.307. L'influence est positive car cette polarité a le degré d'influence le plus élevé (Forte).

B.2/ LES STYLES DE COMMUNICATION DANS *Twitter*

B.2.1/ ÉTAPES DE LA MÉTHODE

Dans cette sous-section, nous détaillons les étapes suivies pour adapter et modifier *TwitBelief* afin d'étudier le style de communication dans *Twitter*. Les étapes suivies sont les suivantes :

1. **Ensemble de combinaisons d'opérateurs :** La première étape de la catégorisation des styles de communication dans *Twitter* est la préparation de l'ensemble des combinaisons d'opérateurs que l'on peut considérer. Ces combinaisons dépendent du domaine étudié.
2. **Choix des paramètres :** Soit Ω_{Style} l'ensemble des réponses possibles à notre question : "Quel est le style de communication dans *Twitter* d'un certain utilisateur pour le modèle I-to-I ?" $\Omega_{Style} = \{\text{Interactif}, \text{Équilibré}, \text{Informatif}\}$. Nous utilisons un sous-ensemble Ω_S de 2^{Ω}_{Style}, précisément : $\Omega_S = \{\text{Interactif}, \text{Équilibré}, \text{Informatif}, \Omega_{Style}\}$. Après, nous associons des masses de croyance pour chaque combinaison d'opérateurs. Ces masses représentent l'importance des combinaisons d'opérateurs dans la catégorisation des styles de *Twitter*.
3. **Fusion des masses :** Pour chaque utilisateur, nous calculons le nombre d'occurrences des combinaisons d'opérateurs dans ses *tweets*. Ensuite, de la même manière que dans *TwitBelief*, les masses de croyance sont combinées en utilisant la formule 4.8 mais selon $@_3$, l'opération qui donne l'intersection entre les éléments de Ω_S (voir tableau 5.7). Nous obtenons ainsi la masse de croyance fusionnée du style de communication dans *Twitter* de l'utilisateur.
4. **Probabilité pignistique :** Dans cette étape, nous procédons comme dans *TwitBelief* pour prendre la décision sur le style de *Twitter* adapté par l'utilisateur. Nous utilisons une version modifiée de la probabilité pignistique bet_3 répartissant uniformément la masse de Ω_{Style} sur tous les autres éléments de Ω_S :

$$\text{bet}_3(x) = m(x) + \frac{m(\Omega_{Style})}{|\Omega_S|}, \quad x \in \Omega_S \tag{B.2}$$

5. **Catégorisation du style :** Enfin, le style de *Twitter* de l'utilisateur est le style ayant la probabilité pignistique la plus élevée.

B.2.2/ ILLUSTRATIONS

Afin d'illustrer la méthode, nous considérons les fonctions de masse suivantes associées à un exemple de deux opérateurs :

$$\text{Opérateur interactif} \mapsto \begin{cases} m_{\text{Opérateur interactif}}(\text{Interactif}) = 0.3 \\ m_{\text{Opérateur interactif}}(\Omega_{Style}) = 0.7 \end{cases}$$

$$\text{Opérateur Informatif} \mapsto \begin{cases} m_{\text{Opérateur Informatif}}(\text{Informatif}) = 0.4 \\ m_{\text{Opérateur Informatif}}(\Omega_{Style}) = 0.6 \end{cases}$$

Cas 1 : Deux opérateurs interactifs

Après initialisation des masses de croyances sur les opérateurs, nous suivons le processus d'approche proposé pour déterminer le style *Twitter* résultant de la fusion de deux opérateurs interactifs utilisés par un utilisateur dans son *tweet*. Nous utilisons d'abord l'opération $@_3$ donnant les correspondances entre les styles, puis nous calculons la combinaison des masses. La fonction de masse résultante des deux opérateurs interactifs est montrée dans le tableau B.4 :

TABLE B.4 – Fusion des deux opérateurs interactifs

$\otimes$	**Interactif** 0.3	$\boldsymbol{\Omega_{Style}}$ 0.7
Interactif 0.3	Interactif 0.09	Interactif 0.21
$\boldsymbol{\Omega_{Style}}$ 0.7	Interactif 0.21	Ω 0.49

Nous obtenons ainsi :

$m(\text{Interactif}) = 0.09 + 0.21 + 0.21 = 0.51$

$m(\Omega_{Style}) = 0.49$

Cas 2 : Deux opérateurs interactifs + Un Opérateur informatif

Dans le second cas, nous considérons un opérateur informatif supplémentaire utilisé par le même utilisateur du cas 1. Afin d'estimer son style *Twitter*, nous combinons les masses de l'opérateur informatif avec les résultats du cas précédent relatif à la combinaisons de deux opérateurs interactifs :

TABLE B.5 – Cas 2 : 2 opérateurs interactifs + 1 opérateur informatif

$\otimes$	**Informatif** 0.4	$\boldsymbol{\Omega_{Style}}$ 0.6
Interactif 0.51	Équilibré 0.204	Interactif 0.306
$\boldsymbol{\Omega_{Style}}$ 0.49	Informatif 0.196	Ω_{Style} 0.294

Nous obtenons :

$m(\text{Interactif}) = 0.306 \quad m(\text{Équilibré}) = 0.204 \quad m(\text{Informatif}) = 0.196 \quad m(\Omega_{Style}) = 0.294$

Notons qu'en combinant les trois opérateurs, la masse de croyance sur le style interactif a diminué par rapport au premier cas, ceci est dû au fait que les masses des styles informatif et équilibré sont apparues.

Enfin, pour prendre la décision sur le style de communication dans *Twitter*, nous calculons la probabilité pignistique :

TABLE B.6 – Probabilité pignistique pour le cas 2

Interactif	Équilibré	Informatif
0.404	0.302	0.294

Nous concluons que le degré d'influence est interactif avec une probabilité pignistique de $0,404$.

C

AUTRES GRAPHES MULTI-COUCHES

Dans cette annexe, nous présentons brièvement d'autres types de graphe multi-couches qui n'ont pas été présentés dans la section 3.3.1.

C.1/ GRAPHES DE NŒUDS COLORÉS

Les graphes de nœuds colorés sont des graphes dans lesquels chaque nœud a exactement une couleur : $G_c = (V_C, E_C, C, X)$. V_C et E_C sont les nœuds et les liens, C est l'ensemble de couleurs possibles où chaque couleur est une catégorie/un type possible pour les nœuds, et $X : V_C \rightarrow C$ est une fonction assignant la couleur à chaque nœud. Le mot "couleur" est utilisé dans un sens très général, en particulier, deux nœuds de la même couleur peuvent être adjacents, c'est-à-dire que la couleur est en fait une catégorie. Les graphes de nœuds colorés peuvent être représentés en utilisant le cadre de réseau multi-couche avec $d = 1$ et en considérant chaque couche comme une couleur. Chaque nœud n'a qu'une seule couleur possible. Les graphes de nœuds colorés peuvent être utilisés pour modéliser les types de graphes multi-couches suivants proposés dans la littérature :

Graphe interdépendant Dans ce type de graphe, les nœuds dans deux ou plusieurs couches sont adjacents les uns aux autres via des liens appelées liens de dépendance [Parshani et al., 2010], c'est-à-dire la présence d'un lien de dépendance signifie qu'un nœud dépend de l'autre et vice versa. [Buldyrev et al., 2010] introduisent un modèle et un cadre analytique pour étudier les réseaux interdépendants.

Graphe inter-connecté Dans un graphe inter-connecté, les nœuds des différentes couches sont adjacents les uns aux autres, mais les liens qui connectent différentes couches n'ont pas besoin d'indiquer des relations de dépendance [Dickison et al., 2012, Saumell-Mendiola et al., 2012]. Ces graphes sont aussi appelés réseaux en interaction [Leicht et al., 2009, Donges et al., 2011] ou réseaux de réseaux [Gao et al., 2011].

Graphe d'information hétérogène Il s'agit de graphes dans lesquels chaque nœud a un type distinct [Sun et al., 2009, Davis et al., 2011, Sun et al., 2013]. Dans [Zhou et al., 2007], les auteurs utilisent un graphe d'information hétérogène afin de

représenter un réseaux de co-citations reliant des auteurs, leurs publications et les citations.

C.2/ GRAPHES DE LIENS COLORÉS

Les graphes de liens colorés sont des graphes avec plusieurs types de liens, ils sont définis comme un triplet : $G_e = (V, E, C)$ où V est l'ensemble de nœuds, C est l'ensemble de couleurs utilisées pour marquer le type d'un lien, et $E \subseteq V \times V \times C$ est l'ensemble des liens. La couleur a de nouveau la signification générale d'une étiquette, ainsi les liens qui sont incidents au même nœud peuvent avoir la même couleur. Nous distinguons deux types de graphes de liens colorés : les graphes multi-relationnels et les graphes multiplexes.

C.3/ GRAPHE K-PARTI

Un graphe de k-parti est un graphe dont les nœuds peuvent être partitionnés en k ensembles disjoints de sorte que deux nœuds d'un même ensemble ne soient pas adjacents. Ainsi, ce graphe comprend différents types de nœuds et les liens ne sont pas autorisés entre le même type de nœud. C'est un cas particulier de graphes de nœuds colorés. Chaque type de nœud correspond à une couleur, et la coloration est une coloration de nœud propre, c'est-à-dire que deux nœuds de même couleur ne peuvent pas être incidents sur le même lien [Horvát et al., 2012]. Dans [Horvát et al., 2013], les auteurs proposent une méthode de projection des graphes K-parti afin de prendre en compte plusieurs types de liens.

D

Figures et tableaux supplémentaires du Pagerank multiplexe

Dans cette annexe, nous présentons des figures et tableaux supplémentaires relatifs aux résultats des expérimentations sur le Pagerank multiplexe présenté dans la section 3.3.3.

D.1/ Classements des candidats Français du corpus TEE 2014 selon les Pagerank des relations *retweet*, *mention* et *réponse*

Table D.1 – Classement des candidats Français du corpus TEE 2014 selon le score PageRank de la relation *retweet*

Rang	Candidat	Score PageRank *retweet*
1	Marine Le Pen	0.0763
2	Jean-Luc Mélenchon	0.0402
3	Florian Philippot	0.0255
4	Nicolas Dupont-Aignan	0.0124
5	José Bove	0.0116
6	Christine Boutin	0.0108
7	Denis Payre	0.0100
8	Julien Salingue	0.0100
9	Corinne Lepage	0.0085
10	Pervenche Berès	0.0070

TABLE D.2 – Classement des candidats Français du corpus TEE 2014 selon le score PageRank de la relation *mention*

Rang	**Candidat**	**Score PageRank *mention***
1	Marine Le Pen	0.1404
2	Christine Boutin	0.0524
3	Jean-Luc Mélenchon	0.0227
4	José Bove	0.0147
5	Florian Philippot	0.0146
6	Nicolas Dupont-Aignan	0.0100
7	Jérôme Lavrilleux	0.0090
8	Geoffroy Didier	0.0063
9	Aymeric Chauprade	0.0055
10	Pervenche Berès	0.0053

TABLE D.3 – Classement des candidats Français du corpus TEE 2014 selon le score PageRank de la relation *réponse*

Rang	**Candidat**	**Score PageRank *réponse***
1	Marine Le Pen	0.0777
2	Christine Boutin	0.0604
3	Florian Philippot	0.0340
4	Jean-Luc Mélenchon	0.0228
5	Nicolas Dupont-Aignan	0.0147
6	Geoffroy Didier	0.0092
7	Julien Rochedy	0.0068
8	Louis de Gouyon Matignon	0.0067
9	Denis Payre	0.0067
10	José Bove	0.0064

D.2/ DÉTAILS DES RÉSULTATS DU PAGERANK MULTIPLEXE MULTIPLICATIF, ADDITIF ET COMBINÉ DES CANDIDATS FRANÇAIS DU CORPUS TEE 2014

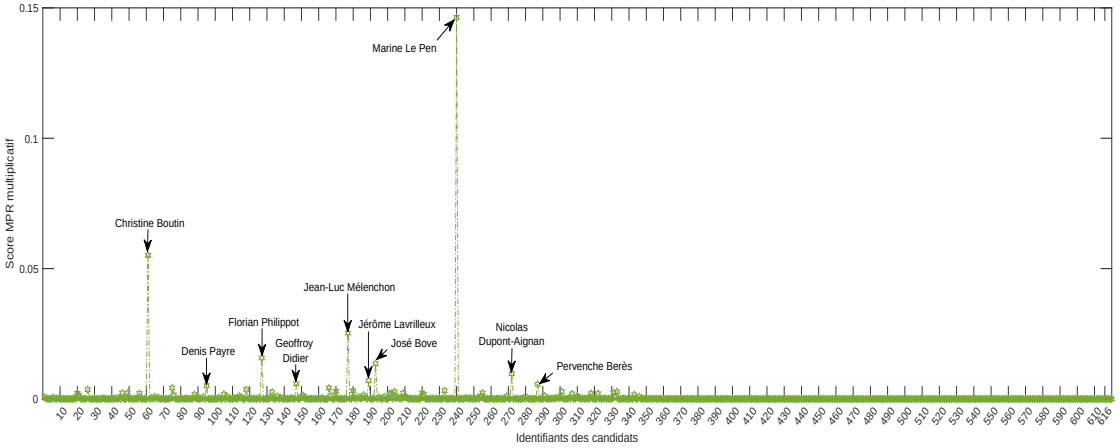


FIGURE D.1 – PageRank multiplexe multiplicatif des candidats français du corpus TEE 2014

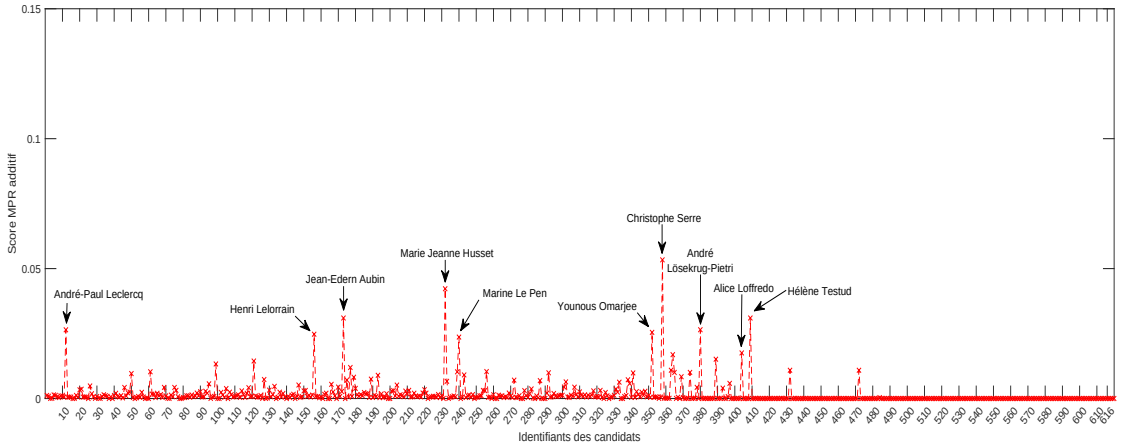


FIGURE D.2 – PageRank multiplexe additif des candidats français du corpus TEE 2014

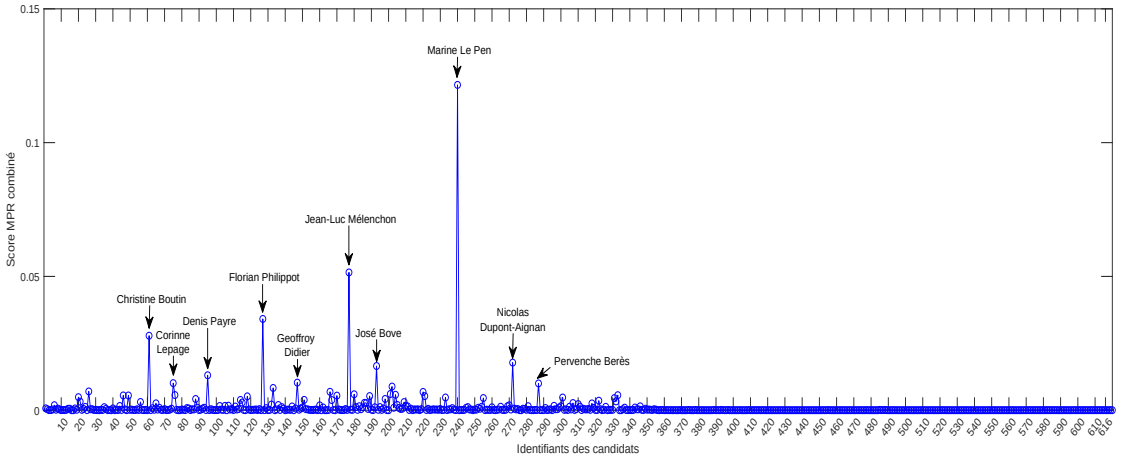


FIGURE D.3 – PageRank multiplexe combiné des candidats français du corpus TEE 2014

TABLE D.4 – Classement des candidats Français du corpus TEE 2014 selon le score PageRank multiplexe multiplicatif

Rang	Candidat	Score PageRank multiplexe multiplicatif
1	Marine Le Pen	0.1463
2	Christine Boutin	0.0551
3	Jean-Luc Mélenchon	0.0253
4	Florian Philippot	0.0157
5	José Bove	0.0136
6	Nicolas Dupont-Aignan	0.0097
7	Jérôme Lavrilleux	0.0070
8	Geoffroy Didier	0.0058
9	Pervenche Berès	0.0057
10	Denis Payre	0.0051

TABLE D.5 – Classement des candidats Français du corpus TEE 2014 selon le score PageRank multiplexe additif

Rang	Candidat	Score PageRank multiplexe additif
1	Christophe Serre	0.0534
2	Marie Jeanne Husset	0.0422
3	Hélène Testud	0.0310
4	Jean-Edern Aubin	0.0310
5	André Lösekrug-Pietri	0.0265
6	André-Paul Leclercq	0.0264
7	Younous Omarjee	0.0255
8	Henri Lelorrain	0.0247
9	Marine Le Pen	0.0236
10	Alice Loffredo	0.0175

TABLE D.6 – Classement des candidats Français du corpus TEE 2014 selon le score PageRank multiplexe combiné

Rang	Candidat	Score PageRank multiplexe combiné
1	Marine Le Pen	0.1216
2	Jean-Luc Mélenchon	0.0515
3	Florian Philippot	0.0341
4	Christine Boutin	0.0278
5	Nicolas Dupont-Aignan	0.0178
6	José Bove	0.0165
7	Denis Payre	0.0131
8	Geoffroy Didier	0.0103
9	Corinne Lepage	0.0101
10	Pervenche Berès	0.0100

D.3/ CLASSEMENTS DES CANDIDATS DU CORPUS TEP 2014 SELON LES PAGERANK DES RELATIONS *retweet*, *mention* ET *réponse*

TABLE D.7 – Classement des candidats du corpus TEP 2017 selon le score PageRank de la relation *retweet*

Rang	Candidat	Score PageRank *retweet*
1	Emmanuel Macron	0.2749
2	Marine Le Pen	0.1577
3	François Fillon	0.1158
4	Philippe Poutou	0.0651
5	Jean-Luc Mélenchon	0.0543
6	Benoît Hamon	0.0360
7	François Asselineau	0.0189
8	Jean Lassalle	0.0172
9	Nicolas Dupont-Aignan	0.0150
10	Nathalie Arthaud	0.0022
11	Jacques Cheminade	0.0015

TABLE D.8 – Classement des candidats du corpus TEP 2017 selon le score PageRank de la relation *mention*

Rang	**Candidat**	**Score PageRank *mention***
1	François Fillon	0.2687
2	Emmanuel Macron	0.1849
3	Benoît Hamon	0.1546
4	Jean-Luc Mélenchon	0.1543
5	Marine Le Pen	0.0244
6	Jean Lassalle	0.0155
7	Philippe Poutou	0.0061
8	Nathalie Arthaud	0.0038
9	Nicolas Dupont-Aignan	0.0032
10	François Asselineau	0.0022
11	Jacques Cheminade	0.0014

TABLE D.9 – Classement des candidats du corpus TEP 2017 selon le score PageRank de la relation *réponse*

Rang	**Candidat**	**Score PageRank *réponse***
1	Marine Le Pen	0.1376
2	François Fillon	0.1246
3	Emmanuel Macron	0.0871
4	Jean-Luc Mélenchon	0.0563
5	Benoît Hamon	0.0280
6	Nicolas Dupont-Aignan	0.0112
7	François Asselineau	0.0104
8	Philippe Poutou	0.0022
9	Nathalie Arthaud	0.0015
10	Jacques Cheminade	0.0006
11	Jean Lassalle	0.0006

D.4/ DÉTAILS DES RÉSULTATS DU PAGERANK MULTIPLEXE MULTIPLICATIF, ADDITIF ET COMBINÉ DES CANDIDATS DU CORPUS TET 2017

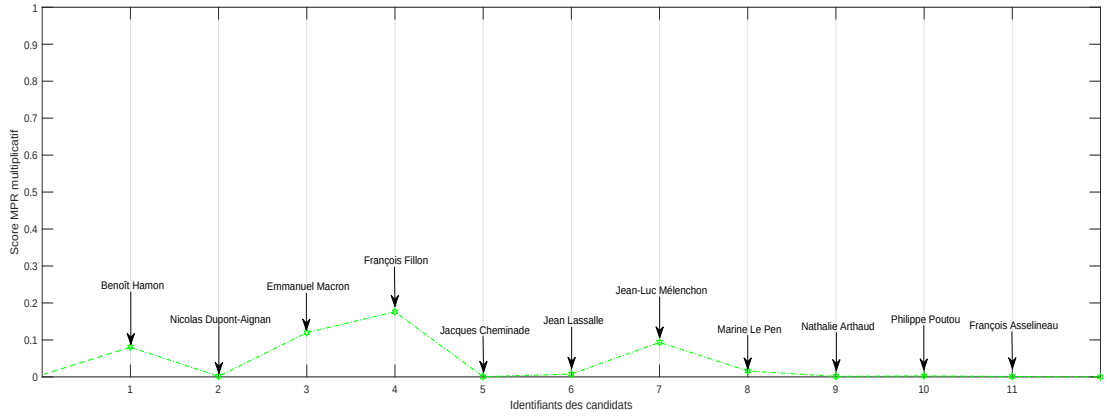


FIGURE D.4 – PageRank multiplexe multiplicatif des candidats du corpus TEP 2017

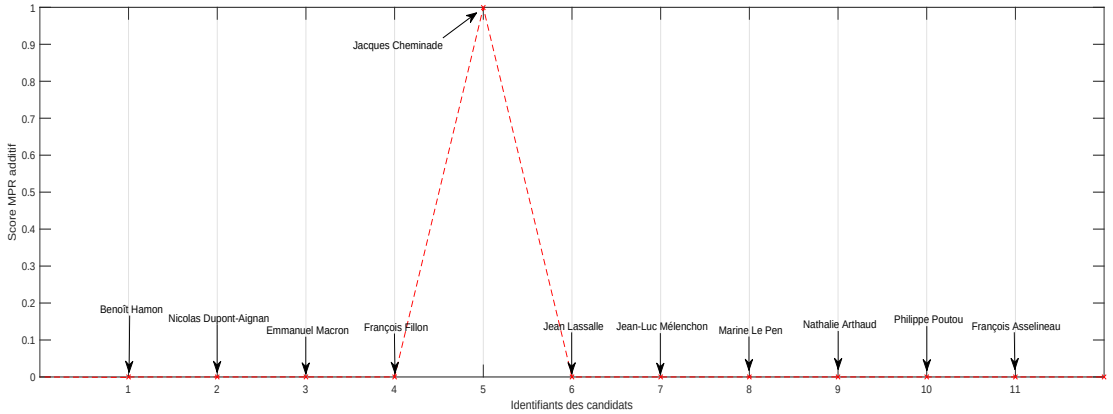


FIGURE D.5 – PageRank multiplexe additif des candidats du corpus TEP 2017

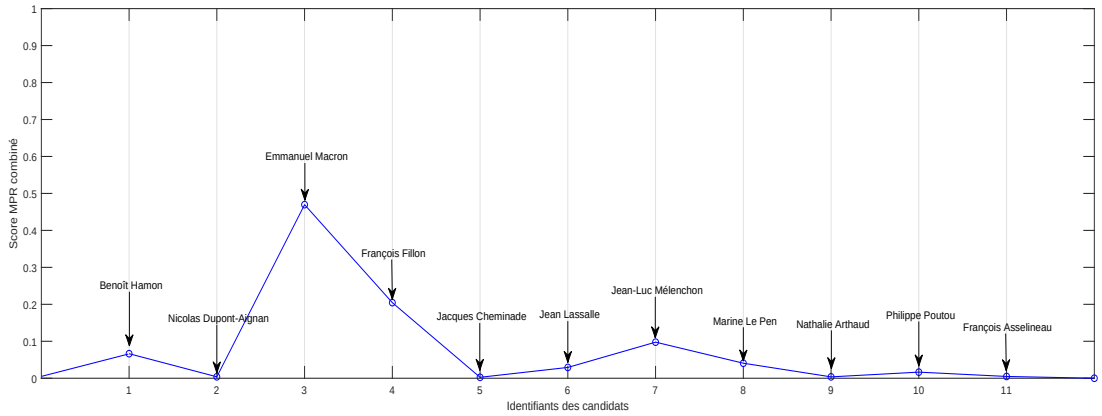


FIGURE D.6 – PageRank multiplexe combiné des candidats du corpus TEP 2017

TABLE D.10 – Classement des candidats du corpus TEP 2017 selon le score PageRank multiplexe multiplicatif durant le premier tour

Rang	Candidat	Score PageRank multiplexe multiplicatif
1	François Fillon	0.1761
2	Emmanuel Macron	0.1195
3	Jean-Luc Mélenchon	0.0934
4	Benoît Hamon	0.0807
5	Marine Le Pen	0.0161
6	Jean Lassalle	0.0075
7	Philippe Poutou	0.0028
8	Nicolas Dupont-Aignan	0.0015
9	Nathalie Arthaud	0.0014
10	François Asselineau	0.0011
11	Jacques Cheminade	0.0006

TABLE D.11 – Classement des candidats du corpus TEP 2017 selon le score PageRank multiplexe additif durant le premier tour

Rang	Candidat	Score PageRank multiplexe additif
1	Jacques Cheminade	0.9994
2	Emmanuel Macron	0.0001
3	Nathalie Arthaud	0.0001
4	Marine Le Pen	0.0001
5	François Fillon	0.0001
6	Philippe Poutou	0
7	Jean-Luc Mélenchon	0
8	Benoît Hamon	0
9	François Asselineau	0
10	Jean Lassalle	0
11	Nicolas Dupont-Aignan	0

TABLE D.12 – Classement des candidatsdu corpus TEP 2017 selon le score PageRank multiplexe combiné durant le premier tour

Rang	Candidat	Score PageRank multiplexe combiné
1	Emmanuel Macron	0.4696
2	François Fillon	0.2045
3	Jean-Luc Mélenchon	0.0976
4	Benoît Hamon	0.0664
5	Marine Le Pen	0.0406
6	Jean Lassalle	0.0292
7	Philippe Poutou	0.0165
8	François Asselineau	0.0048
9	Nicolas Dupont-Aignan	0.0039
10	Nathalie Arthaud	0.0037
11	Jacques Cheminade	0.0025

TABLE D.13 – Scores des candidats du corpus TEP 2017 lors du deuxième tour

Candidat	Emmanuel Macron	Marine Le Pen
Score PageRank selon la relation *retweet*	0.2641	0.1838
Score PageRank selon la relation *mention*	0.0331	0.4893
Score PageRank selon la relation *réponse*	0.1493	0.1242
Score PageRank multiplexe multiplicatif	0.2678	0.0190
Score PageRank multiplexe additif	0.4518	0.0928
Score PageRank multiplexe combiné	0.0001	0.0001

E

PUBLICATIONS SCIENTIFIQUES

Les contributions proposées dans cette thèse ont abouti aux publications suivantes :

1. Lobna Azaza, Sergey Kirgizov, Marinette Savonnet, Eric Leclercq et Alex Frame. *Evaluation de l'influence sur Twitter : Application au projet 'Twitter aux Elections Européennes 2014'*. Workshop Webpol 2015 « Etudier le Web politique : Regards croisés ». Lyon, France.
2. Lobna Azaza, Sergey Kirgizov, Marinette Savonnet, Eric Leclercq et Rim Faiz. *Influence assessment in Twitter multi-relational network*. SITIS 2015, 11th International Conference on Signal Image Technology and Internet Based Systems. Bangkok, Thaïlande. 23-27 Novembre 2015. pp 436-443.
3. Lobna Azaza, Marinette Savonnet et Eric Leclercq. *Candidates' influence assessment in Twitter during the 2014 European Elections*. TEE 2014, Twitter at the European Elections 2014 : International Perspectives on a Political Communication Tool. Dijon, France. 26-27 Novembre 2015.
4. Lobna Azaza, Sergey Kirgizov, Marinette Savonnet, Eric Leclercq et Rim Faiz. *Évaluation de l'influence dans un réseau multirelationnel : le cas de Twitter*. INFORSID 2016, INFormatique des ORganisation et Systèmes d'Information et de Décision. Grenoble, France. 31 Mai - 3 Juin, 2016. pp 131-146
5. Lobna Azaza, Marinette Savonnet et Éric Leclercq. *Évaluation de l'influence des candidats sur Twitter durant les élections européennes de 2014*. Chapitre dans l'ouvrage collectif "Twitter aux éléctions européennes". TEE 2014 Editions L'Harmattan.
6. Lobna Azaza, Fatima Zohra Ennaji, Zakaria Maamar, Abdelaziz El Fazziki, Marinette Savonnet, Mohamed Sadgal, Eric Leclercq et Djamal Benslimane. *A credibility and classification-based approach for opinion analysis in social networks*. MEDI 2016, 6th International Conference on Model and Data Engineering. Aguadulce, Almeria, Espagne. 21-23 Septembre 2016. pp 303-316.
7. Lobna Azaza, Sergey Kirgizov, Marinette Savonnet, Eric Leclercq, Nicolas Gastineau et Rim Faiz. *Information fusion based approach for studying influence on Twitter using belief theory*. Revue Computational Social Networks 2016. Volume 3, pp 5-31.
8. Lobna Azaza, Sergey Kirgizov, Marinette Savonnet, Eric Leclercq et Rim Faiz. *Évaluation de l'influence polarisée dans un réseau multi-relationnel : Application à Twitter*. Revue Document Numérique 2017. Volume 20, pp 67-100.
9. Fatima Zohra Ennaji, Lobna Azaza, Zakaria Maamar, Abdelaziz El Fazziki, Marinette Savonnet, Mohammed Sadgal, Eric Leclercq, Idir Amine Amarouche et Djamal Benslimane : *Impact of Credibility on Opinion Analysis in Social Media.* Fundamenta Informaticae 2018, Volume 162, pp 259-281.