UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR (UCAD)

ÉCOLE DOCTORALE MATHS-INFORMATIQUE (EDMI)

Année : 2017 N^0 ordre : xxx

THÈSE DE DOCTORAT UNIQUE

Spécialité : Informatique

Présentée par :
Ibrahima GAYE

Titre : Analyse des réseaux sociaux : contributions à la détection des semences dans la maximisation de l'influence

Soutenue publiquement le Mardi 23 Mai 2017 devant un jury composé de :

Président	**Cheikh Thiécoumba GUEYE, Professeur, UCAD de Dakar**
Rapporteurs	**Oumarou SIÉ, Professeur, Univ. Ouaga I-Pr. Joseph Ki-Zerbo**
	Ousmane THIARÉ, Professeur, UGB de Saint Louis
	Idrissa SARR, Maître de conférences, UCAD de Dakar
Examinateurs	**Cheikh SARR, Professeur, Université de Thiès**
	Samuel OUYA, Maître de conférences, UCAD de Dakar
Directeur	**Diaraf SECK, Professeur, UCAD de Dakar**
Co-Directeur	**Gervais MENDY, Maître de conférences, UCAD de Dakar**

Thèse effectuée au sein du :

Laboratoire d'Informatique, Réseaux-Télécoms (LIRT)

et

Laboratoire de Mathématiques de la Décision et de l'Analyse Numérique (LMDAN)

Dakar, Fann, UCAD

Sénégal

Résumé

Le nombre d'utilisateurs des réseaux sociaux qui font partie des systèmes les plus utilisés sur internet, ne cesse d'augmenter exponentiellement au jour le jour. Dans ces réseaux, plusieurs types d'information tels que des photos, du texte, des documents, des vidéos, etc, sont échangés entre les utilisateurs. Les entreprises, les gouvernements et diverses structures commencent à les utiliser à leurs propres fins. Beaucoup de recherches sont effectuées dans les réseaux sociaux. Dans nos travaux, nous traitons de la maximisation de l'influence qui consiste à sélectionner efficacement les semences (les utilisateurs initiateurs) et à trouver un modèle de diffusion optimal. Le thème principal de nos travaux est le premier point, à savoir trouver un ensemble de $k-$*éléments*, appelé les semences ou les graines [1], dans le réseau social qui vont maximiser l'influence.

Dans un premier temps, nous avons proposé une mesure de centralité appelée degré de diffusion ℓ-ième qui est notée par C_{dd}^{ℓ}. Dans cette mesure, nous donnons de l'importance aux utilisateurs qui ont le plus de voisins de niveau 1 (N^1) jusqu'au ℓ (N^{ℓ}) et qui sont influents sur leurs voisins. En d'autres mots, si nous considérons la propagation de l'influence comme une expansion de rayon ℓ à partir d'un utilisateur u, plus ses voisins de niveau 1 jusqu'à ℓ sont influents, plus u est important. Ansi, ce n'est pas le nombre de sommets qui est important dans la diffusion de l'influence mais le degré d'influence. Les $top-k$ utilisateurs sont considérés comme les plus influents et la maximisation de l'influence est initiée par eux.

Ensuite, nous avons étudié les Réseaux Sociaux MultiCouches (RSMC) [2] qui peuvent être, par exemple une agrégation de plusieurs réseaux sociaux, un réseau social qui

1. seeds en anglais
2. MultiLayer Social Networks (MLSN)

a plusieurs types de liens. Chaque réseau social constituant une nature de liens est considéré comme une couche. Nous utilisons les relations d'équivalence pour générer les matrices de mappage afin d'identifier les mêmes utilisateurs dans les différentes couches. Une fois la représentation des RSMC, nous avons proposé une mesure de centralité, appelée *degré de multi diffusion* et elle est notée par C_{dd}^{MLN}. Celle-ci accorde de l'importance aux utilisateurs qui ont plus de voisins de niveau 1 (N^1) dans toutes les couches du RSMC et qui sont influents sur leurs voisins. Les $top - k$ utilisateurs sont considérés comme les plus influents dans le RSMC et la maximisation de l'influence est initiée par eux.

Enfin, nous proposons une approche qui consiste à extraire un graphe partiel particulier appelé graphe couvrant de maximisation. Cette extraction se base sur le comportement des modèles de diffusion de base, à savoir les modèles cascades et les modèles seuil linéaires. Les modèles cascades donnent de l'importance aux utilisateurs qui détiennent l'information tandis que les modèles seuils linéaire donnent de l'importance aux utilisateurs qui ne la détiennent pas l'information. Dans ces deux modèles de diffusion un utilisateur est influencé une seule fois. A une itération, on traite seulement les utilisateurs non influencés sinon on parle de rétroaction. L'extraction de graphe partiel a pour but de d'empêcher la rétroaction vers les utilisateurs semences. Dans un premier temps, nous avons proposé un algorithme d'extraction, appelé SGC, qui prend e entrée un graphe non orienté connexe. Ce dernier est en deux versions : SGC_{v1} qui construit les descendants aléatoirement et SGC_{v2} qui les construit en se basant sur le nombre de leurs voisins dans le graphe initial. Ensuite, nous donnons une généralisation de chacun des deux versions. Elle prend tous les types de graphes et elle est appelée algorithme SG. Après l'extraction du graphe couvrant de maximisation, nous utilisons les heuristiques existantes pour déterminer les semences dans le graphe partiel.

Pour estimer les performances de nos approches, nous avons utilisé le logiciel open-source R et la librairie *igraph* pour faire les simulations. Nous avons mesuré la

propagation de l'influence donnée par nos approches et celle donnée par les approches références. Nous avons aussi mesuré la vitesse de propagation de l'influence de nos modèles.

Mots-clés :

Analyse des réseaux sociaux, graphe couvrant de maximisation, maximisation de l'influence, mesure de centralité, modèle de diffusion, niveau de voisinage, probabilité de diffusion, rétroaction, réseau social monoplex, réseau social multicouche.

Social networks analysis : contributions to the detection of seeds for maximizing the influence spread

Abstract

The number of the users of the social networks, some of the most used system on the internet, increases exponentially day by day. In these networks, several types of information like photos, text, documents, videos, etc are exchanged between users. Businesses, governments and others begin to use them it to their own purposes. Many works are carried out in these social networks. In our work, we treat the influence maximization that consists to select efficiently the seeds (the first diffusers) and to find an optimal diffusion model. It is very difficult to solve this two subproblems. The main theme of our works is the first point, namely to find a set of k-elements, called seeds or seeds, in the social network that will maximize influence.

At first step, we propose a centrality measure called $\ell - th$ degree diffusion and it denoted by C_{dd}^{ℓ}. In this measure, we give importance to individuals that have more neighbors of level 1 denoted by N^1, until level ℓ denoted by N^{ℓ} and that are influential on their neighbors. In other words, if we consider the influence propagation as an expansion of ℓ radius from a u users, more its neighbors from level 1 until ℓ spread the influence more u is important. Then, the number of neighbors is not important in the influence spread but the influence degree. The $k-top$ are considered the most influential and influence spread is initiated by them.

Then, We have proposed a representation of Monoplex Social Network (MSN) that can be, for example an aggregation of several social networks, a social network that has several types of links. Each social network or nature of links is considered as a layer. We use the equivalence relationships to generate the mapping matrices to identify the same users in the different layers. Once the representation of MLSN, we proposed a centrality measure, called the multi-diffusion degree and it is denoted by C_{dd}^{MLN}. It gives importance to users who have more neighbors of level 1 in all layers

of the MLSN and that are influential on their neighbors. The $k-top$ are considered most influential in MLSN and information spread is initiated by them.

Finally, we propose an approach that is to extract a particular partial graph called maximization spanning graph. This extraction is based on the behavior of benchmarks propagation models, namely the cascade models and linear threshold models. The Cascade models give importance to users who hold information, while linear threshold models give importance to users who do not hold the information. In these two diffusion models a user is influenced only once. At a given iteration, only uninfluenced users are treated, otherwise it is referred to as feedback. The purpose of partial graph extraction is to prevent feedback to seed users. Firstly, we propose an extraction algorithm, called SGC, that takes a connected graph. This latter is in two version, SGC_{v1} that builds randomly the neighbors of each node and SGC_{v2} that builds them by basing on their neighborhood in the initial graph. Secondly, we give a generalization for each of these two versions. It takes an arbitrary graph and it called SG. After extraction of maximization spanning graph, we use the existing heuristics to determine the seeds.

To estimate the pertinence of our approach, we compare the influence propagation of seeds given by the benchmarks and ours approach. The performances of these approach are very perceptible through the simulation carried out by the R software and the *igraph* package.

Keywords :

Centrality measure, diffusion model, diffusion probability, influence maximization, information feedback, maximization spanning graph, monoplex social network, multilayer social network, neighborhood level, social networks analysis, spread model.

Remerciements

Il y a prèsque trois ans et demi, lorsque que je quittais le ministère des Finances et de l'économie pour débuter une thèse à l'Université Cheikh Anta Diop de Dakar. Son aboutissement semblait être impossible devant ma personne. En octobre 2016, lorsque j'ai commencé la rédaction proprement dite, le point final de ce manuscrit semblait se trouver au bout d'un long tunnel interminable.

Aujourd'hui, à moi seul, je ne serais sans doute jamais arrivé à accomplir cette thèse. Après avoir rendu grâce à ALLAH Soubhanahou Wa Tahala (SWT) et son PROPHETE MOUHAMAD PSL, je tiens à remercier l'ensemble des personnes, qui par leurs conseils, leurs remarques, leurs encouragements et leurs accompagnements ont contribué, de près ou de loin, à l'aboutissement de ce travail :

Chiekh Thiécoumba GUEYE, Professeur Titulaire à l'Université Cheikh Anta Diop de Dakar, pour m'avoir fait l'honneur de présider le jury de cette thèse.

Diaraf SECK, Professeur Titulaire à l'Université Cheikh Anta Diop de Dakar, en tant que Directeur de thèse, pour sa rigueur intellectuelle et morale, qui est pour moi plus qu'un exemple à suivre, une référence.

Gervais MENDY, Professeur Assimilé à l'Université Cheikh Anta Diop de Dakar, mon Co-directeur de thèse, pour sa rigueur intellectuelle et morale, son accompagnement, qui est pour moi plus qu'un Co-directeur.

Ousmane THIARE, Professeur Titulaire à l'Université Gaston Berger de Saint Louis, qui fait parti des professeurs qui m'ont initié en informatique, d'avoir pris son temps pour rapporter ce travail.

Idrissa SARR, Professeur Assimilé a l'Université Cheikh Anta Diop de Dakar, d'avoir pris son temps pour rapporter ce travail, ses remarques et suggestions très perti-

nentes.

Oumarou SIE Professeur Titulaire à l'Université Ouaga I-Pr. Joseph Ki-Zerbo d'avoir pris son temps pour rapporter ce travail.

Cheikh SARR, Professeur Titulaire à l'Université de Thiès d'avoir accepter d'examiner ce travail.

Samuel OUYA, Professeur Assimilé à l'Université Cheikh Anta Diop de Dakar et Directeur du laboratoire Informatique, réseaux-télécom s(LIRT), de m'avoir guidé scientifiquement dans mes travaux, de m'avoir accepter dans son laboratoire et d'avoir accepter d'examiner ce travail.

Tegawendé FBISSYANDE, PhD de l'Université du Luxembourg, pour son soutient implicite, le temps consacré, ses pertinentes remarques pour que le document soit parfait.

Tous les chercheurs du département génie informatique de l'ESP, particulièrement au Dr Idy DIOP, Dr Ibrahima NGOM, Dr Ibra DIOUM et Dr Mandicou BA pour leurs conseils, leurs remarques scientifiques, ... Je remercie aussi la secrétaire Mme DIOUF que je considère comme une grande s{oeur, Adoulaye Sané qui m'a beaucoup soutenu et Caba SALL.

Également, les membres des laboratoires LIRT, LIMBI et LMDAN, en commençant par leurs Directeurs, tout particulièrement Dr Abdourahmane NDIAYE, Dr Madiagne DIOUF, Dr Massamba SECK et Dr Agnés NGOM qui m'ont guidé dans mes travaux et mes camarades de promotion Madiop DIOUF, Ousmane SADIO, Birahime DIOUF, Ousmane KHOUMA, Ndiaye DIOP, Ndiaga BA et Khassim MBODJI.

Enfin, je tiens à exprimer ma gratitude envers mes parents, qui depuis mon enfance n'ont menage aucun effort pour que je sois aujourd'hui à ce niveau, mon oncle mes frères et soeurs, ma femme mes neveux et ma petite Aminata Racine Gaye.

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

ARS Analyse des Réseaux Sociaux. 1, 2, 23, 34, 69, 103, 125

CELF Cost-Effective Lazy Forward. 62

IC Cascade Indépendant. 4, 41–43, 46, 49, 51, 65, 66, 69, 85, 87, 88, 102, 112, 113, 115, 120, 125, 132

LT Linéaire avec Seuil. 4, 41, 45, 46, 49, 51, 66, 85, 87, 88, 92, 102, 112, 120, 125, 132

MLSN MultiLayer Social Networks. iv, viii, ix, 4, 12, 30

MPI maximum influence paths. 59

MSN Monoplex Social Network. viii

RSM Réseau Sociau Monoplex. 11, 32, 47, 48, 53, 55, 56, 64, 76–78, 86, 109

RSMC Réseaux Sociaux MultiCouches. iv, v, 4, 12, 20, 23, 30–33, 47, 48, 53, 55, 56, 64, 65, 76–79, 84, 86, 109–112, 117, 118

SI Susceptible-Infected. 38

SIMPATH Simple Path. 59

SIR Susceptible-Infected-Recovered. 39

SNA Social Networks Analysis. 1, 2, 125

SP1 Shortest-path. 46

SPIN ShaPley value based Influential Nodes. 58

SPM Shortest-path Model. 46

UBLF Upper Bound basé sur Lazy Forward. 63

Liste des notations

C_c Centralité par proximité - Closeness centrality. 79, 84, 85

C_d Centralité degré - Degree centrality. 72

C_{dd}^{MLN} Centralité degré multi-diffision. iv, vii, 68, 96, 97, 106, 116, 130

C_{dd}^{ℓ} Centralité degré diffusion ℓ-ième. iii, vi, 68, 69, 72–75, 77, 78, 96, 106, 116, 129

E_k Ensemble de arêtes de couche k. 12

E_k Ensemble de sommets de couche k. 12

E_{SG} Ensemble des arêtes du graphe couvrant SG. 89, 90, 93

L_1 1-ième couche du graphe multicouche. 33, 34

L_2 2-ième couche du graphe multicouche. 33, 34

L_3 3-ième couche du graphe multicouche. 33, 34

L_k K-ième couche du graphe multicouche. 12

$MM_k^{k'}$ Matrice de mappage entre les couches k et k'. 13, 14

$N(\ell, v)$ Les voisins de niveau 1 à ℓ. 68, 70, 71, 76, 78

N^1 Les voisins de niveau 1. iii, iv, vi, vii, 8, 68–70, 72–75, 77, 104, 117–119

N^2 Les voisins de niveau 2. iii, vi, 70, 72–75, 117–119

N^{ℓ} Les voisins de niveau ℓ. iii, vi, 4, 72–74

S_k^* Les k sommets qui vont maximiser la propagation. 15, 16, 20, 45, 48, 77–79, 81, 129

S_k Une combinaison de k sommets. 16, 20, 48, 49

V_{SG} Ensemble des sommets du graphe couvrant SG. 89, 90, 93

λ Probabilité de propagation. 16, 38, 41, 43, 72, 97

ψ Modèle de diffusion. 15, 16, 34, 48, 65, 66, 120, 129

σ La fonction objective ou fonction d'influence. 15, 16, 45, 49, 50, 65, 66

BeginNode Sommet qui débute la construction du graphe couvrant SG. 82, 84, 90, 94

E Ensemble des arêtes d'un graphe social. 8, 9, 16, 20, 40, 43, 49, 59, 88, 89, 92, 93, 120

G Un graphe social. 8, 9, 16, 20, 40, 43, 49, 59, 88, 89, 92, 93, 120

m Nombre d'arêtes d'un graphe. 9, 83

MM Toutes les matrices de mappages. 12, 14

N Les voisins de niveau 1. 8, 42, 43, 70, 74

n Nombre de sommets d'un graphe. 9, 83

SG Graphe couvrant de maximisation. 9, 78, 79, 84, 85, 89, 93, 94

V Ensemble des sommets d'un graphe social. 8, 9, 16, 20, 40, 43, 49, 59, 88, 89, 92, 93, 120

Liste des Algorithmes

Introduction générale

Dans ces dernières décennies, le réseau internet devient de plus en plus indispensable dans nos activités quotidiennes. L'explosion des réseaux sociaux, tels que *Facebook*, *Twitter*, *Viadeo*, *Linkedin*, etc. offre de nouvelles opportunités d'établir de nouveaux contacts et de partager plusieurs types d'information. Ils peuvent changer donc la nature de la communication et de l'information. On se pose un certain nombre de questions telles que : à quels contacts envoyer un message sachant que la plupart des utilisateurs ont plusieurs centaines d'amis ? Quelles pages *Wikipédia* faut-il lire en priorité pour en apprendre le plus possible sur un sujet donné ? Comment prévenir la propagation d'une rumeur (influence négative) ? Comment propager le plus loin possible une information (influence positive) ? Dans tous ces contextes, la propagation de l'information entre le entités et leur évolution permet de mieux appréhender le fonctionnement global des systèmes. Par la suite, des outils méthodologiques et/ou algorithmiques adaptés aux problèmes rencontrés sont proposés. La science des réseaux sociaux tient ses origines en sociologie avec des travaux datant du début du vingtième siècle mais a pris un essor nouveau ces quinze dernières années et touche la plupart des disciplines. Ainsi l'Analyse des Réseaux Sociaux (ARS) [3] attire beaucoup d'attention grâce à ses domaines d'applications variés tels que minimiser/maximiser la diffusion d'une information, identifier des acteurs centraux d'un réseau, extraire de la connaissance des réseaux (apprentissage), détecter des groupes de personnes qui partagent le même centre d'intérêt etc. Dans ces dernières décennies, le réseau

3. Social Networks Analysis Social Networks Analysis (SNA) en anglais

internet devient de plus en plus indispensable dans nos activités quotidiennes. L'explosion des réseaux sociaux, tels que *Facebook*, *Twitter*, *Viadeo*, *Linkedin*, etc. offre de nouvelles opportunités d'établir de nouveaux contacts et de partager plusieurs types d'information. Ils peuvent changer donc la nature de la communication et de l'information. On se pose un certain nombre de questions telles que : à quels contacts envoyer un message sachant que la plupart des utilisateurs ont plusieurs centaines d'amis ? Quelles pages *Wikipédia* faut-il lire en priorité pour en apprendre le plus possible sur un sujet donné ? Comment prévenir la propagation d'un rumeur (influence négative) ? Comment propager le plus loins possible une information (influence positive) ? Dans tous ces contextes que l'information sur les relations entre entités et leur évolution permettent de mieux appréhender le fonctionnement global des systèmes. Par la suite, des outils méthodologiques et/ou algorithmiques adaptés aux problèmes rencontrés sont proposés. La science des réseaux sociaux tient ses origines en sociologie avec des travaux datant du début du vingtième siècle mais a pris un essor nouveau ces quinze dernières années et touche la plupart des disciplines. Ainsi l'ARS [4] attire beaucoup d'attention grâce à ses domaines d'applications variés tels que minimiser/maximiser la diffusion, identifier des acteurs centraux d'un réseau, extraire de la connaissance des réseaux (apprentissage), détecter des groupes de personnes qui partagent le même centre d'intérêt etc. De nombreux acteurs de la société (exemple : les entreprises, les services gouvernementaux, les journalistes et d'autres) cherchent à exploiter et analyser les réseaux sociaux à des fins diverses (exemple : analyser la réaction des consommateurs à propos de certains produits et les promouvoir, rendre visible un programme lors d'une campagne électoral, etc.). Depuis longtemps, la diffusion de l'information est observée et étudiée dans de nombreux domaines de la science : propagation des maladies [1] ou des virus informatiques [2], diffusion des innovations technologiques [3], déplacements humains [4], etc. Le phénomène de diffusion de l'information peut être défini comme l'action de propa-

4. Social Networks Analysis SNA en anglais

ger des éléments d'information auprès d'un public, suscite depuis plusieurs années un grand intérêt au sein de la communauté scientifique. Nos travaux portent sur la maximisation de l'influence dans les réseaux sociaux. Étudier se problème revient à bien sélectionner les semences (les diffuseurs initiaux) et à avoir un modèle de diffusion optimal. Vu ces deux problématiques, nos travaux se focalisent particulièrement sur la détection des semences qui est un problème combinatoire stochastique. Il consiste à trouver un ensemble de $k-$individus dans le réseau social qui va maximiser l'influence sous un modèle de diffusion optimal. Il est connu sous l'expression de maximisation de l'influence.

Actuellement, les thématiques dans les réseaux sociaux sont bien orientées, le nombre d'utilisateurs qui augmente exponentiellement et nous y trouvons les mêmes utilisateurs. Ces réseaux peuvent être vus comme une agrégation d'un seul réseau appelé réseau social multicouche et chacun d'eux est vu comme une couche. L'information peut circuler entre les couches via utilisateurs qui ont plusieurs comptes. Dans [5], Wang Wenjun *et al.* ont montré que l'influence circule facilement dans une communauté qui peuvent être vue comme une couche et toutes ces dernières comme un réseau multicouche. Le sexe, l'âge, etc [6], peuvent être des paramètres très important dans le problème de maximisation de l'influence. Chaque catégorie peut être vue comme une couche et l'ensemble forme un réseau multicouche. Dans la suite, nous parlons de réseaux sociaux monolex ou simplement réseaux sociaux. Si on a plusieurs couches, nous parlerons de réseaux sociaux multicouches.

Pour la représentation des réseaux sociaux (monoplex et multicouches), nous utilisons naturellement les graphes qui jouent un rôle très important dans la modélisation de beaucoup de problèmes pratiques et théoriques. Les graphes sont utilisés comme outil de representation pour les réseaux transports, les réseaux de communications, les architectures informatiques, les médias sociaux, etc. Dans le cas des réseaux sociaux multicouches, nous avons utilisé les graphes pour chaque couche et des matrices de mappages pour identifier les mêmes individus dans les différentes couches.

Nos contributions dans cette thèse peuvent être regroupées en deux points :

- (i) Nous avons proposé une heuristique appelée degré de diffusion ℓ-ième et qui est notée par C_{dd}^{ℓ}. Dans cette heuristique, nous donnons de l'importance aux utilisateurs qui ont le plus de voisins de niveau 1 (N^1) jusqu'au niveau ℓ (N^{ℓ}) qui acceptent la diffusion de l'information. Cette heuristique est efficace dans les réseaux sociaux monoplex. Nous avons proposé une représentation des RSMC [5] qui peuvent être, par exemple une agrégation de plusieurs réseaux sociaux qui ont plusieurs types de liens. Chaque nature de liens est considérée comme une couche (un réseau social simple ou monoplex). Nous utilisons les relations d'équivalence pour générer les matrices de mappage afin d'identifier les mêmes utilisateurs dans les différentes couches. Après une représentation, nous avons proposé une heuristique, appelée degré multi-diffusion et elle est notée par C_{dd}^{MLN}. Dans cette mesure, nous donnons donnons de l'importance aux utilisateurs qui ont de l'influence sur ses voisins de niveau 1 (N^1) dans toutes les couches qui ont aussi de l'influence sur leurs voisins.
- (ii) Après avoir étudié les deux principales familles de modèles de diffusion de base, à savoir les modèles cascades (Cascade Indépendant (IC)) et seuil linéaires (Linéaire avec Seuil (LT)). A un instant, les utilisateurs inactifs sont tenté d'être influencés par leurs voisins actifs. Par-ce-que, les utilisateurs sont influencés une seule fois. Sinon, on parlons de rétroaction. Nous avons proposé une approche qui est de faire un pré-traitement pour prévenir les rétroactions vers les utilisateurs. Dans un premier temps, nous avons proposé un algorithme d'extraction, appelé SGC, qui prend un graphe non orienté connexe. Ce dernier est en deux versions, SGC_{v1} qui construit les descendants aléatoirement et SGC_{v2} qui les construit en se basant sur le nombre de leurs voisins dans le graphe initial. Ensuite, nous donnons une généralisation de chacune des deux versions. Elle prend tous les types de graphes et elle est appelée algorithme SG. Après l'extraction du graphe couvrant

5. MLSN : MultiLayer Social Networks

de maximisation, nous utilisons les heuristiques existantes pour déterminer les semences dans le graphe partiel. Dans la détection des des semences en utilisant les mesures de centralité existantes, nous avons montré que les graphes couvrants donnent de meilleurs semences que les graphes initials.

Ce manuscrit est composé de quatre chapitres.

Dans le premier chapitre, nous fournirons toutes les définitions liées à la maximisation de l'influence. Dans le premier paragraphe, nous allons parler de quelques concepts, définitions et notations sur les graphes et nous allons proposer une representation des réseaux sociaux monoplex et multicouche. Dans le deuxième paragraphe, nous allons donner les définitions utilisées dans ce manuscrit, sur la maximisation de l'influence qui est le thème principal de nos travaux.

Dans le deuxième chapitre, nous parlerons premièrement, des réseaux sociaux en expliquant leur principe de fonctionnement et en montrant leur importance dans le web en fonction du nombre d'utilisateurs qui s'agrandit exponentiellement au jour le jour. Deuxièmement, nous exposerons différents modèles de propagation de l'influence, leurs propriétés, leurs modes de fonctionnement etc. et troisièmement, nous expliquerons.l'approche gloutonne pour la sélection des semences.

Au niveau du troisième chapitre, nous faisons d'abord l'état de l'art sur la détection de diffuseurs initiaux qui vont maximiser la diffusion de l'influence, en divisant les approches en trois catégories à savoir, les approches dont la sélection des semences est statique, dynamique et celles qu'utilisent l'approche gloutonne. Ensuite nous faisons une synthèse de l'état de l'art avant de parler de nos contributions en tenant en compte des paramètres très importants. Enfin, nous présentons nos différentes approches.

Dans le quatrième et dernier chapitre de ce document, nous présentons l'outil utilisé dans les simulations de nos modèles après avoir présenté quelques uns. Nous terminons cette partie par des tests afin de montrer que nos modèles sont plus performants que les existants après avoir expliqué les jeux de donnés utilisés.

Chapitre 1

Définitions, terminologie et notations

Dans ce chapitre, nous commençons par rappeler les termes et les concepts de base sur la maximisation de l'influence dans les réseaux sociaux monoplex et multi-couches. Ainsi nous aborderons quelques concepts tout en fournissant les définitions et notations utilisées sur les graphes qui nous servent d'outil de représentation des réseaux sociaux. Enfin, nous donnerons les définitions utilisées dans ce manuscrit sur la maximisation de la diffusion de l'influence qui est le thème principal de nos travaux.

1.1 Graphes et réseaux sociaux

Dans cette section, nous représentons des réseaux sociaux (monoplex et multi-couches) à l'aide des graphes après avoir rappelé quelques concepts de base sur ces derniers.

1.1.1 Concepts de base des graphe

Dans l'analyse des réseaux sociaux, les graphes jouent un rôle très important dans leurs représentations. Ils permettent de représenter les éléments et les relations qui les

lient. Un graphe peut être non orienté ou orienté. Dans ce manuscrit, nous parlerons simplement de graphes sauf mention expresse, nous préciserons que c'est orienté. Un graphe G = (V, E) est défini par l'ensemble fini $= \{v_1, v_2, \cdots, v_n\}$ dont les éléments sont appelés sommets ou nœuds et par l'ensemble fini E = $\{e_1, e_2, \cdots, e_m\}$ dont les éléments sont appelés arêtes. Une arête $e_i \in E$ est définie par une paire non-ordonnée de sommets, appelés extrémités de e_i. Si l'arête e_i relie les sommets a et b, on note ab et on dira que ces sommets sont adjacents, ou incidents avec e_i ou encore que l'arête e_i est incidente avec les sommets a et b. Les sommets a et b sont des voisins de niveau 1 (ou simplement des voisins) l'un de l'autre. L'ensemble des voisins d'un noeud a est appelé voisinage de niveau 1 (ou simplement voisinage) de a et noté par N(a) (ou N^1(a)). Le nombre de voisins d'un sommet a est le degré de a. On appelle ordre d'un graphe, ceque l'on note par n, le nombre de sommets de G. Le nombre d'arête du graphe est noté par m.

Dans le cas des graphes orientés, on parle d'arcs au lieu d'arêtes. Si $ab \in$E, alors nous avons un arc dirigé du sommet a vers le sommet b. a est le prédécesseur de b qui est le successeur de a. L'ensemble des successeurs du noeud a est appelé voisinage sortant du sommet a. Le nombre de voisins sortants est le degré sortant. L'ensemble des prédécesseurs du noeud a est appelé voisinage entrant. Le nombre de voisins entrants est le degré entrant du sommet a.

On appelle chaîne de G = (V, E) toute suite C alternée de sommets et d'arêtes de G = (V, E) : C = (x_0, a_1, x_1, $\cdots$, a_n, x_n) telle que $\forall i \in \{1, n\}$, $a_i = (x_{i-1}, x_i)$
C est une chaîne de longueur n (le nombre d'arêtes) joignant x_0 à x_n. Dans le cas d'un graphe simple, il est inutile de préciser les arêtes et nous noterons simplement : C = (x_0, x_1, $\cdots$, x_n).

Un graphe non vide est connexe si pour toute paire de sommets quelconques du graphe, il existe une chaîne qui les relie. Un sous-graphe de $G = (V, E)$ connexe est appelé une composante connexe ou simplement une composante. Soit k un entier naturel non nul. Un graphe G est $k-$*connexe* si $|V| \succ k$ et $G - X$ est connexe pour

tout sous-ensemble de sommets X de V vérifiant $|X| \prec k$. Le graphe représenté dans la figure 1.1, est connexe. Pour chaque paire de sommets qu'on choisit, on peut trouver une chaîne qui les relie.

Un graphe couvrant [1] d'un graphe G, qu'on notera par SG, est un sous-graphe obtenu par suppression de quelques arêtes. Autrement dit, c'est un sous-graphe dont l'ensemble de sommets est exactement celui du graphe G et l'ensemble des arêtes est une partie de celui de de G. Si S est l'ensemble des arêtes supprimées, ce sous-graphe de G est noté $G\backslash S$. Un arbre couvrant [2] de G est un graphe couvant de G qui est connexe et qui n'a pas de cycle. La figure 1.2 est un graphe couvrant de G représenté par la figure 1.1. Comme ce graphe couvrant est connexe et n'a pas de cycle alors il est un arbre couvrant de G. Un graphe couvrant est dit forêt couvrante s'il n'est pas connexe et s'il n'a pas de cycle. Une forêt peut être vue aussi comme une réunion d'arbre couvrants.

Tout graphe connexe admet au moins un arbre couvrant. Si le graphe n'est pas connexe, il admet des composantes connexes qui admettent chacune un arbre couvrant et la réunion de ces derniers donne la forêt couvrante. Dans la figure 1.2, nous avons un des graphes couvrants du graphe social de la figure 1.1.

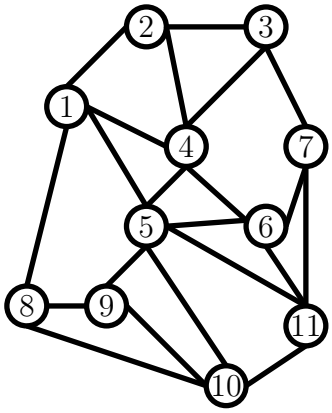

FIGURE 1.1: *Un graphe social connexe G*

1. spanning graph
2. spanning tree

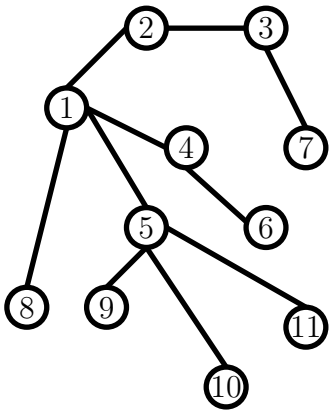

FIGURE 1.2: *Un graphe couvrant de G*

Dans le cas des graphes orientés, on parle de forte connexité dans la théorie des graphes. Cette forte connexité est définie par de la façon suivante : pour deux sommets a et b donnés, on peut trouver un chemin de a vers b. Dans nos travaux, nous utilisons les définitions suivantes :

Définition 1.1 ***Un graphe orienté connexe par rapport à un sommet***
Un graphe orienté est connexe par rapport à un sommet a si est seulement si, on peut accéder à tout sommet $b \in V - \{a\}$ en respectant l'orientation.

Définition 1.2 ***Un graphe orienté non connexe par rapport à un sommet***
Un graphe orienté est non connexe par rapport à un sommet a si est seulement si il existe un sommet $b \in V - \{a\}$ non accessible à partir de a en respectant l'orientation.

Après avoir donné quelques concepts et notations sur les graphes, nous allons voir comment représenter un réseau social monoplex ou multicouche à l'aide de ces derniers.

1.1.2 Graphes et réseaux sociaux

Les graphes sont utilisés pour représenter des problèmes réels. En 1736, le mathématicien suisse *Leonhard Euler* s'intéresse au problème des ponts de Konigsberg ([7])

"existe-t-il une promenade dans la ville prussienne de Konigsberg passant une et une seule fois par les sept ponts de la ville ?". Il représente chaque pont par une arête et les îles par des sommets. En 1856, le mathématicien irlandais *William Hamilton* utilise ce modèle ([7]) pour chercher un chemin autour du monde en passant une et une seule fois dans chaque ville. Depuis, le recours à un graphe pour représenter un système réel est devenu un outil classique des mathématiques discrètes et les propriétés des graphes ont été largement étudiées. Ils permettent de représenter et d'étudier les ensembles structurés complexes, les relations entre objets, l'évolution de systèmes dans le temps, les réseaux (informatiques, les médias sociaux, etc.). Dans cette même lancée, les réseaux sociaux représentent des utilisateurs et une ou plusieurs inter-actions entre eux. Naturellement, les graphes sont utilisés pour les représenter.

Réseau social monoplex

Un réseaux social monoplex, que nous noterons par Réseau Sociau Monoplex (RSM), est composé d'utilisateurs et d'une seule nature de lien entre eux. Formellement, un réseau social monoplex est représenté par un graphe étiqueté, où les sommets correspondent aux utilisateurs du service et où les liens représentent les connexions entre utilisateurs. Ce graphe social peut être orienté si le mode de connexion entre les utilisateurs du réseau social monoplex est unilatéral. Par contre, il est non orienté si le mode de connexion est bilatéral. Les sommets sont étiquetés avec les messages publiés par l'utilisateur correspondant. Un message est décrit par son auteur, son contenu et sa date de publication.

Réseau social multicouche

Un réseau social multicouche est composé d'un ensemble d'utilisateurs et plusieurs natures de relations entre eux. Chaque nature de relations est un réseau social monoplex qui représente une couche [3]. L'agrégation de tous ces réseaux forme un réseau

3. Layer en anglais

social multicouche RSMC [4]. Supposons un RSMC composé de η natures de relations et numérotés de de 1 à η, alors pour tout $k \in [1, ..., \eta]$, $L_k = (V_k, E_k)$ est un graphe monoplex représentant la k-ième couche où V_k est l'ensemble des utilisateurs et E_k est l'ensemble de liens entre eux. Un réseau social multicouche est la réunion de toutes les couches. Pour identifier les mêmes utilisateurs dans les différentes couches, en plus des L_k pour tout $k \in [1, ..., \eta]$, comme dans les études menées+ par Gaye Ibrahima *et al.* [8] et par Magnani Matteo *et al.* [9], nous définissons des matrices de mappage entre les différentes couches. Le réseau social multicouche, noté par RSMC, est défini par RSMC=$(L_1, L_2, L_3, \cdots, L_\eta, MM)$, où MM est la réunion des matrices de mappage entre les différentes couches. Dans la figure 1.3, nous avons un graphe qui représente un réseau social avec deux natures de relations (2 couches).

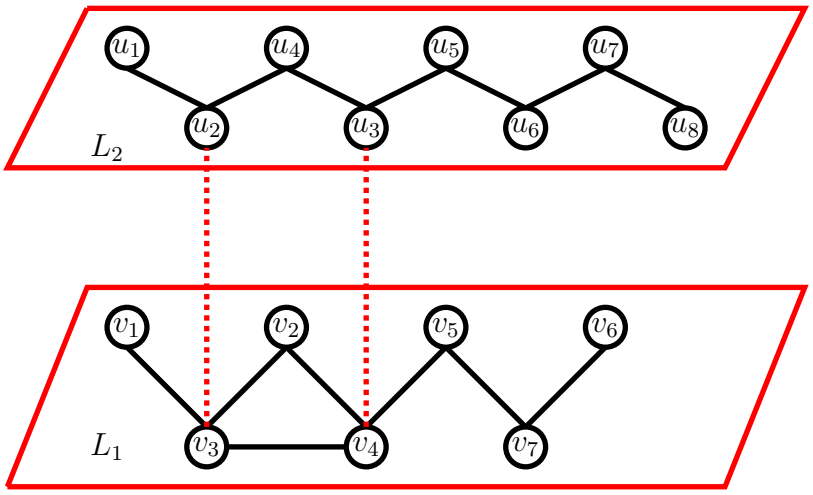

FIGURE 1.3: *Un réseau social 2-couches*

Pour construire la matrice de mappage entre les couches k et k' que l'on notera $MM_k^{k'}$, nous définissons une relation $\Re$ de la manière suivante :

- $V_k^i \in L_k$, $V_{k'}^j \in L_{k'}$
- $(V_k^i \ \Re V_{k'}^j)$ les mêmes utilisateurs dans les couches k et k'

4. MultiLayer social Network MLSN

$\Re$ est une relation d'équivalence.

Preuve :

Soit V_k^i un sommet de la couche k. Il représente le sommet i dans la même couche. Alors il est en relation avec lui même. Donc la relation est bien réflexive.

Soient deux sommets V_k^i et $V_{k'}^j$ des couches respectives k et k'. Supposons que V_k^i $\Re$ $V_{k'}^j$. Alors le sommet V_k^i dans la couche k est représenté par le sommet $V_{k'}^j$ dans la couche k'. Donc ces deux sommets représentent le même acteur dans ces deux couches différentes. Alors la relation est bien symétrique.

Soient trois sommets V_k^i, $V_{k'}^j$ et $V_{k''}^l$ dans les couches respectives k, k' et k''. Supposons que V_k^i et $V_{k'}^j$ représentent le même acteur dans les couches k et k'. Supposons aussi que $V_{k'}^j$ et $V_{k''}^l$ représentent le même acteur dans les couches k' et k''. Alors V_k^i et $V_{k''}^l$ représentent le même acteur dans les couches k et k''. Donc la relation est bien transitive.

Comme la relation $\Re$ est réflexive, symétrie et transitive alors elle est une relation d'équivalence.

Pour la construction de la matrice de mappage $MM_k^{k'}$, nous utilisons la relation d'équivalence $\Re$ entre tous les utilisateurs des deux couches. Cette matrice peut être définie comme :

$$MM_{k'}^{k} = \begin{array}{c} \\ V_k^1 \\ V_k^2 \\ V_k^3 \\ \vdots \\ V_k^{n_k} \end{array} \begin{array}{c} \begin{array}{ccccc} V_{k'}^1 & V_{k'}^2 & V_{k'}^3 & \cdots & V_{k'}^{n_{k'}} \end{array} \\ \begin{pmatrix} a_{1,1}^{k,k'} & a_{1,2}^{k,k'} & a_{1,3}^{k,k'} & \cdots & a_{1,n_{k'}}^{k,k'} \\ a_{2,1}^{k,k'} & a_{2,2}^{k,k'} & a_{2,3}^{k,k'} & \cdots & a_{2,n_{k'}}^{k,k'} \\ a_{3,1}^{k,k'} & a_{3,2}^{k,k'} & a_{3,3}^{k,k'} & \cdots & a_{3,n_{k'}}^{k,k'} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n_k,1}^{k,k'} & a_{n_k,2}^{k,k'} & a_{n_k,3}^{k,k'} & \cdots & a_{n_k,n_{k'}}^{k,k'} \end{pmatrix} \end{array}$$

où $a_{i,j}^{k,k'} = 1$ si V_k^i $\Re$ $V_{k'}^j$ 0 sinon

Comme la relation d'équivalence $\Re$ est symétrique, alors la matrice de mappage $MM_k^{k'}$ est la matrice transposée de $MM_{k'}^{k}$ donc il suffit de représenter une seule

de ces deux. Puisque la relation est réflexive, alors c'est inutile de représenter la matrice de mappage pour la même couche qui va se réduire à la matrice unité. En prenant en compte de la symétrie, la matrice de mappage entre la couche k et k' a les mêmes informations que la matrice de mappage entre k' et la couche k (i.e. $MM_{k'}^{k}$ et $MM_{k}^{k'}$ ont les mêmes informations). Comme ce qui nous intéresse est la circulation de l'information entre les différentes couches, alors si on prend en compte la réflexivité, la matrice de mappage MM_{k}^{k} qui se réduit la matrice de l'unité n'est pas importante. Donc l'ensemble MM sera réduit à la réunion de matrices vérifiant l'équation 1.1.

$$MM = \bigcup_{\substack{k,k' \in \{1 \cdots \eta\} \\ k \succ k'}} MM_{k'}^{k} \tag{1.1}$$

En guise d'exemple, considérons le réseau social de la figure 1.3 composé de deux couches (L_1 et L_2). On a un acteur représenté par les sommets $v3$ et $u2$ dans les couches respectives L_1 et L_2. Dans la matrice de mappage MM_1^2, le coefficient de $v3$ et $u2$ est égal à 1. Ci-après, nous représentons la matrice de mappage entre ces deux couches.

$$MM_1^2 = \begin{array}{c} \\ v1 \\ v2 \\ V3 \\ V4 \\ v5 \\ v6 \\ v7 \end{array} \begin{array}{c} \begin{array}{cccccccc} u1 & u2 & u3 & u4 & u5 & u6 & u7 & u8 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

La matrice MM, d'après l'équation 1.1, se réduit à MM_1^2. Donc le réseau social multicouche sera défini par $MLSN = (L_1, L_2, MM)$ où $MM = MM_1^2$.

Après une représentation des réseaux sociaux (monoplex et multicoucches) à l'aide des graphes, nous donnons quelques définissions liées au problème de maximisation de l'influence dans les réseaux sociaux.

1.2 Notion de maximisation de l'influence dans les réseaux sociaux

1.2.1 Influence

1.2.1.1 Généralité sur l'influence

Dans le Web comme dans d'autres domaines, le terme "influence" est très à la mode en ce moment. Selon le dictionnaire Lc'arousse, l'influene est définie comme :

- action, généralement continue, qu'exerce quelque chose sur quelque chose ou sur quelqu'un.
- ascendant de quelqu'un sur quelqu'un d'autre.
- pouvoir social et politique de quelqu'un, d'un groupe, qui leur permet d'agir sur le cours des événements, des décisions prises, etc.

Vu ces définitions, le phénomène d'influence est le fait de modifier le comportement, la pensée, etc. de quelqu'un ou de quelque chose. Il est observé dans plusieurs domaines dont certains vont être citer ci-dessous.

Dans le domaine de la politique : elle sert à désigner des phénomènes de pouvoir qui ne reposent pas sur la détention d'une autorité légale (l'influence des intellectuels, des médias ou des autorités morales, par exemple) ;

Dans le domaine de la géopolitique, on parle souvent de zones d'influence, politique d'influence, par opposition à politique de puissance ;

Dans le domaine de la sociologique, les groupes d'influence sont des organisations qui exercent une certaine emprise sur les décisions des autorités et les réorientent

dans un sens favorable à leurs intérêts ;

Dans le domaine de l'internet : on nomme "$e-$influents", les internautes qui ont la capacité d'attirer un grand nombre de visiteurs vers certains sites ou qui lancent des messages (écrits, des vidéos, etc.) que beaucoup d'autres internautes répercuteront.

Dans tous ces domaines, nous nous intéressons à l'influence dans le domaine de l'internet en particulier dans les réseaux sociaux.

1.2.1.2 Influence dans les réseaux sociaux

L'influence sociale appelée aussi imitation ou pression sociale est un phénomène social exercé par un individu ou par un groupe sur chacun de ses membres dont le résultat est d'imposer des normes dominantes en matière d'attitude et de comportement. Ce terme d'influence social est défini par Anagnostopoulos *et al.* [10] comme suit :

Définition 1.3 ***Influence sociale***
Elle traduit le fait que les actions d'un utilisateur peuvent induire ses voisins à se comporter d'une manière similaire. L'influence se manifeste parfois explicitement dans les réseaux sociaux, par exemple sur facebook, lorsqu'un utilisateur « re-publie» un message. C'est-à-dire lorsqu'il recopie un message en créditant l'auteur original.

On distingue plusieurs types d'influence sociale, dites typologies d'influence, tels que le *conformisme*, l'*innovation*, la *soumission à l'autorité* ou l'*obéissance* et la *normalisation*. Il existe également d'autres phénomènes qui peuvent s'expliquer en termes d'influence sociale comme la résistance qui s'oppose aux phénomènes précédents.

↬ *Le conformisme ou la conformité* est l'idée qu'une personne modifie sa position et ou ses idées dans la direction de celle du groupe. Un des aspects centraux de l'influence sociale est que nous ne nous basons pas sur notre propre expérience seulement pour définir ce qu'est la réalité. Nous nous reposons également beaucoup sur les idées des autres. De façon générale, nous sommes sensibles aux points de

vue des hommes, et donc nous sommes dans une situation difficile quand nous n'avons pas les mêmes points de vue.

↬ *L'obéissance ou la soumission* est l'idée qu'un individu modifie son comportement afin de se soumettre à l'ordre direct d'une autorité.

↬ *L'innovation*, appelée aussi l'influence des minorités, est l'influence exercée par un individu ou un petit groupe à l'encontre de l'ensemble du groupe auquel il appartient et avec lequel il manifeste un conflit. Si l'influence s'exerce, on parle de minorité agissante. De façon générale, on est tenté de concevoir le but de processus d'influence sociale comme étant essentiellement de réduire les divergences entre les individus. L'influence sociale s'exercerait surtout dans un sens unique du groupe vers l'individu.

↬ *La normalisation* est l'interaction d'individus ou de groupe qui aboutit à un compromis et à un nivellement des positions respectives. C'est une conception qui diffère des deux processus précédents d'influence sociale dans le sens où le conformisme et l'obéissance font référence à la pression plus ou moins explicite d'un individu ou d'un groupe sur un autre individu. Alors qu'avec la normalisation, on a une pression réciproque qui s'exerce au cours des échanges entre les individus et qui vise à dégager une norme de jugement acceptable par tous.

1.2.2 Le problème de maximisation

L'optimisation est une branche mathématique cherchant à modéliser, à analyser et à résoudre analytiquement ou numériquement les problèmes qui consistent à minimiser ou maximiser une fonction sur un ensemble. Elle est définie par une fonction de $A \longrightarrow f$. On cherche à trouver un élément α de A tel que $f(\alpha) \geq f(\beta)$ pour tout β de A. On dit que l'on cherche à maximiser f sur l'ensemble A. f est appelée fonction objective ou fonction objectif. Cette maximisation est notée par l'équation 1.2 ou

1.3

$$\max\{f(X) : X \in A\} \tag{1.2}$$

$$\arg \max_{X \in A} f(X) \tag{1.3}$$

L'optimisation est découpée en sous-disciplines qui se chevauchent, suivant la forme de la fonction objectif et celle des contraintes : l'optimisation en dimension finie ou infinie (on parle ici de la dimension de l'espace vectoriel des variables à optimiser), l'optimisation continue ou combinatoire (les variables à optimiser sont discrètes dans ce dernier cas), l'optimisation différentiable ou non lisse (on qualifie ici la régularité des fonctions définissant le problème), l'optimisation linéaire (fonctions affines), quadratique (objectif quadratique et contraintes affines), semi-définie positive (la variable à optimiser est une matrice dont on requiert la semi-définie positivité), copositive (la variable à optimiser est une matrice dont on requiert la copositivité), conique (généralisation des disciplines précédentes, dans laquelle on minimise une fonction linéaire sur l'intersection d'un cône et d'un sous-espace affine), convexe (fonctions convexes), non linéaire, la commande optimale, l'optimisation stochastique et robuste (présence d'aléas), l'optimisation multicritère (un compromis entre plusieurs objectifs contradictoires est recherché), l'optimisation algébrique (fonctions polynomiales), l'optimisation bi-niveaux, l'optimisation sous contraintes de complémentarité, l'optimisation disjonctive (l'ensemble admissible est une réunion d'ensembles), etc. Cette abondance de disciplines provient du fait que pratiquement toute classe de problèmes modélisables peut conduire à un problème d'optimisation, pourvu que l'on y introduise des paramètres à optimiser. Par ailleurs, les conditions d'optimalité de ces problèmes d'optimisation apportent parfois des expressions mathématiques originales qui, par le mécanisme précédent, conduisent à leur tour à de nouveaux problèmes d'optimisation.

Dans nos travaux, nous traitons la maximisation de l'influence dans les réseaux qui est une optimisation stochastique.

1.2.3 Le problème de maximisation de l'influence

Kempe David et al. [11] montre que le problème de la maximisation de la diffusion de l'influence est $NP-$difficile. Elle est définie sur un graphe social monoplex et multicouche. Résoudre ce problème revient à :

- (i) avoir un modèle de diffusion optimal ψ,
- (ii) Trouver un ensemble de k *utilisateurs*, appelés les semences ou les graines, qui vont initier la propagation.

En partant de k utilisateurs du réseau social et un modèle de diffusion ψ, le but est savoir lesquels des k utilisateurs et lesquels des modèles ψ entraîneront une diffusion maximale. On définit alors la fonction objective appelée aussi fonction d'influence σ qu'on cherche à optimiser. Dans le cas de propagation d'une bonne information, on cherche à maximiser la diffusion et on parle de influence positive. Dans le cas de propagation d'une mauvaise information, on cherche à minimiser la diffusion et on parle d'influence négative. En d'autres mots, il s'agit de déterminer le nombre d'individus actifs à la fin d'une diffusion. Dans la pratique, un certain nombre de modèles étant stochastiques comme dans [11], $\sigma()$ correspond à l'espérance mathématique du nombre d'utilisateurs ayant diffusé le contenu. Dans nos travaux, nous nous sommes focalisés sur l'influence positive. L'objectif est simplement de maximiser la fonction σ qui prend en paramètre les k diffuseurs initiaux et un modèle de propagation ψ. Le problème est énoncé ci dessous.

Problème 1.1 *Le problème de la maximisation de l'influence peut être posé comme le problème d'optimisation stochastique suivant :*
Pour un graphe social G=(V,E) donné, un modèle de diffusion ψ, un budget k et un ensemble de semences $S_k \in V$, trouver un S_k^* qui va maximiser la diffusion de

l'information sous le modèle ψ.
Mathématiquement, le problème de maximisation de l'influence peut être définie par l'équation 1.4. Il cherche un ensemble de k utilisateurs ($S_k \subseteq V$) qui vont maximiser la fonction objectif aussi appelée la fonction d'influence $\sigma(S)$.

$$S_k^* = \arg \max_{S_k \subseteq V, |S_k|=k} \sigma(S) \tag{1.4}$$

où V est l'ensemble des utilisateur, k le nombre de semences, S_k^* les graines qui vont maximiser la fonction objective $\sigma()$

Résoudre cette équation revient à résoudre deux sous problèmes. D'abord, on cherche à trouver un ensemble S_k^* qui représente les diffuseurs initiaux (les semences). Puis, on cherche un modèle de propagation ψ optimal. Ces modèles peuvent être regroupés en deux familles : les modèles cascades et les modèles seuils linéaire. Dans ces modèles de diffusion, les utilisateurs vont subir une influence sociale qui amène un changement de comportement. Dans le chapitre 2, nous développons les modèles de diffusion et dans le chapitre 3, nous donnerons des modèles de détection des semences qui costituent le thème phare de nos travaux.

Conclusion

Dans ce chapitre, nous avons rappelé quelques termes et concepts de base sur la maximisation de l'influence dans les réseaux sociaux monoplex et multicouches. Nous avons donné quelques concepts, définitions et notations sur les graphes qu'on a utilisé dans ce manuscrit. Avec les graphes, nous avons donné une représentation les réseaux sociaux. Dans les réseaux sociaux multicouches, que nous notons par RSMC en plus de la représentation naturelle, on a montré comment identifier les mêmes utilisateurs dans les différentes couches en utilisant les matrices de mappages. Dans le chapitre

suivant, nous allons voir le principe de fonctionnement des réseaux sociaux et les modèles de diffusion utilisés.

Chapitre 2

Les réseaux sociaux et la diffusion de l'information

Depuis des décennies, l'internet est devenu indispensable dans la vie de l'homme. Les réseaux sociaux, dont les logos de certains des plus populaires sont dans la figure 2.1, deviennent de plus en plus un moyen de partager beaucoup de données tels que des photos, du texte, de la vidéo, etc. Ils occupent une place très importante dans le développement de l'internet. Plusieurs acteurs comme les entreprises, les gouvernements, et d'autres commencent à s'y intéresser. L'analyse de réseaux sociaux (ARS) est devenue un axe de recherche très important. Dans ce chapitre, nous expliquerons les réseaux sociaux et la diffusion de l'information dans ces derniers. Il est divisé en trois parties. Dans la première partie, nous exposerons les généralités sur les réseaux sociaux. Avant d'entrer dans le vif du sujet, nous rappelons des principes de fonctionnement. Ensuite, nous montrerons leur importance dans l'évolution de l'internet via leurs nombres d'utilisateurs qui ne s'arrêtent de s'agrandir exponentiellement et nous détaillerons les deux réseaux sociaux phares, à savoir *Facebook* et *Twitter*. Enfin, nous donnerons des exemples de RSMC dont la plupart des travaux de recherche restent théoriques. Dans la partie suivante, nous présenterons des différents modèles de propagation de l'information. Nous présenterons les modèles épidémiologies et

nous terminons cette partie en expliquant les modèles centrés sur la topologie du réseau qui sont les principaux modèles utilisés dans nos approches de détection de semences. Dans la dernière partie, nous présenterons le problème de la maximisation de l'influence en montrant sa complexité et en expliquant l'approche gloutonne.

FIGURE 2.1: *Logos de quelques réseaux sociaux*

2.1 Généralité sur les réseaux sociaux

De nos jours, il existe plusieurs réseaux sociaux qui ont une ou plusieurs thématiques permettant des liens d'amitiés, professionnels, de rencontres, etc. Ils peuvent être différents selon leur principe de fonctionnement et les modules qu'ils proposent. Dans cette section, nous ferons une étude générale des réseaux sociaux les plus connus. Nous commencerons par définir et par donner les différents principes de fonctionnement de ces réseaux. Ensuite, nous fournirons quelques chiffres sur les réseaux les plus populaires et nous parlerons de deux réseaux phares, à savoir *Facebook* et *Twitter*. Enfin, nous présenterons les réseaux sociaux multicouches.

2.1.1 Définition et principe de fonctionnement

Selon Wikipedia, un réseau social est un ensemble d'utilisateurs ou d'organisations reliés par des interactions sociales régulières. Car oui, si aujourd'hui, en parlant de réseau social nous avons plutôt tendance à penser web, il faut savoir que par définition, un réseau social est un groupe de personne qui maintiennent des liens. Souvent, il s'agit d'utilisateurs partageant des valeurs, des intérêts communs. C'est autour de ces éléments qu'ils échangent pour maintenir des liens dans la durée. La nature de ces liens dépend du réseau social qui va servir de support au réseau : amis *Facebook*, relations *LinkedIn*, followers *Twitter*, etc. De proche en proche, un utilisateur peut se lier aux amis de ses amis et le réseau social se transforme en une gigantesque toile comme on le voit dans la figure 2.2.

FIGURE 2.2: *Un réseau social sous forme de toile*

D'une manière générale, un réseau social peut être vu comme une application en ligne, on parle d'application sociale, dont les utilisateurs vont s'inscrire pour avoir un compte utlisateur. Ils peuvent faire essentiellement deux choses :

- (i) créer une page de profil sur laquelle l'utilisateur peut publier des messages (photos, du texte, etc.),
- (ii) se connecter à d'autres utilisateurs afin de suivre leurs publications, de partager des informations.

Cette définition générale est similaire à celle proposée par Ellison Nicole *et al.* [12]. On peut citer : *On définit les réseaux sociaux comme des services basés sur le web qui permettent de (i) construire un profil public ou semi-public dans un système borné,*

(ii) articuler une liste d'autres utilisateurs avec lesquels ils partagent une connexion, et (iii) afficher et parcourir leur liste de connexions et celles faites par d'autres au sein du système. La nature et la nomenclature de ces connexions peuvent varier d'un site à l'autre.

Beaucoup de réseaux sociaux respectent cette définition et chacun présente ses spécificités. Les différents réseaux sociaux se distinguent d'une part en fonction de la visibilité et de l'accessibilité des pages de profil de leurs utilisateurs. D'autre part, les réseaux sociaux peuvent se différencier par la manière dont les utilisateurs se connectent entre eux. Cette connexion peut être soit unilatéralement, soit bilatéralement. Ils se différencient par leurs thématiques. Nous pouvons avoir des réseaux sociaux de partage de photos (*FlickR*[1]), des réseaux sociaux professionnelles (*viadéo*[2], *linkedIn*[3], ...), des réseaux sociaux de rencontre (*twoo*[4], *badoo*[5], *hi5*[6], ...). NOus notons des réseaux sociaux généraux tels que *facebook*[7], *twitter*[8]. Il existe également d'autres types de réseaux sociaux qui ont leur particularités. Parmi ceux-là, on citer les réseaux sociaux de proximité (*proxiigen*[9], ...) qui permettent de se connecter avec tous les utilisateurs qui sont aux alentours. Dans la diffusion de l'influence, il est très important de connaître la thématique du réseau social qui est un facteur très important. Par exemple, si nous voulons diffuser une information qui parle de politique dans un réseau social rencontre, une diffusion large sera quasi impossible.

1. https ://www.flickr.com/
2. https ://fr.viadeo.com/
3. https ://fr.linkedin.com/
4. https ://www.twoo.com/
5. https ://badoo.com/fr/
6. www.hi5.com/
7. https ://fr-fr.facebook.com/
8. https ://twitter.com/ ?lang=fr
9. https ://proxiigen.com/

2.1.2 Quelques chiffres

Ces dernières années, le nombre d'utilisateurs des réseaux sociaux ne cesse de grandir exponentiellement. Certains chefs d'états, des entreprises et d'autres structures utilisent comme un moyen de communication et de partage de l'information. Des hommes politiques dont des présidents disent qu'ils ne font plus confiance aux médias et préfèrent communiquer directement avec les populations via les réseaux sociaux. Howard Philip *et al.* [13] ont étudié le rôle de *Facebook* dans le printemps arabe. Hughes A. Lee et Palen Leysia [14] montrent le rôle de *Twitter* dans les élections présidentielles 2008 des USA. Avec le développement des téléphones permettant de se connecter à internet, plusieurs de ces réseaux sociaux ont développés des plateformes mobiles. Cela participe aussi à l'augmentation des utilisateurs. Actuellement, les réseaux sociaux font partie des systèmes les plus utilisés sur internet. Dans les tableaux 2.1 et 2.2, nous avons quelques chiffres en 2016 de *Facebook* et de *Twitter* selon *leptidigital*[10].

TABLEAU 2.1: *Chiffres clés de Facebook en 2016*

Nombre de	Chiffre
inscrits par jour	518 400
utilisateurs via les mobiles par jour	172 800
messages échangés par jour	216 millions
liens partagés par jour	72 millions
status partagés par jour	422 millions
demandes d'ajout en amis par jour	144 millions
photos partagées par jour	196 millions

10. http ://www.leptidigital.fr/reseaux-sociaux/chiffres-reseaux-sociaux-temps-reel-6335/

TABLEAU 2.2: *Chiffres clés de Twitter en 2016*

Nombre de	Chiffre
inscrits par jour	44 410
utilisateurs via les mobiles par jour	35 597
tweets partagés par jour	500 millions
recherches par jour	2 milliards

Les chiffres dans les tableaux ci-dessus montrent que les réseaux *Twitter* et *Facebook* sont très visités par les internautes. Ils sont un outil très puissant dans le domaine de la communication et du partage de l'information.

2.1.3 Exemples

Les deux réseaux sociaux les plus populaires du monde pour l'instant, sont *Facebook* et *Twitter* dont certains chiffres sont donnés dans la section précédente. Ils ont un principe de fonctionnement un peu différent. Mais il existe d'autres réseaux sociaux moins populaires que les deux cités ci-haut.

2.1.3.1 *Facebook*

Il est le plus large réseau social au monde. Il est devenu pour beaucoup de personnes, la porte d'entrée sur le web. Il permet de découvrir de nouveaux contenus, de suivre la vie de ses proches, de chatter et de partager des photos et des vidéos auprès de ses amis. Il propose également des solutions efficaces pour aider les entreprises à toucher leurs clientèles. Les profils créés sur *Facebook* sont privés sauf si son créateur en décide autrement. Il permet à la personne ayant initié la connexion de recevoir automatiquement les messages publiés par l'utilisateur ciblé. *Facebook* se base sur un mode de connexion bilatéral, ce qui signifie que les deux utilisateurs doivent autoriser la création du lien. L'information circule alors dans les deux sens et ce lien

est appelé « lien d'amitié ». Le terme amitié est employé par de nombreux réseaux sociaux pour désigner les connexions entre utilisateurs. Néanmoins, comme l'observe Danah Michele BOYD *et al.* [15], cette appellation est trompeuse et les utilisateurs se connectent entre eux pour de nombreuses raisons sans pour autant être amis au sens commun du terme.

2.1.3.2 *Twitter*

À la différence de *Facebook* centré sur le réseau d'amis proches, Twitter permet de suivre librement n'importe quel utilisateur (ami, marque, personnalité) et de partager des messages courts limités à 140 caractères. Tandis que certains individus l'utilisent quasiment comme un chat public pour interagir avec un noyau dur d'amis, entreprises, marques, médias, journalistes et créateurs, d'autres s'en servent également comme d'un outil de promotion efficace. Tous les profils créés sur *Twitter* sont, par défaut, publics et indexés par les moteurs de recherche traditionnels, ce qui les rend accessibles à tout un chacun sans nécessairement posséder un compte *Twitter*. Par exemple *Twitter* propose un mode de connexion unilatéral qui permet à tout utilisateur de se connecter à n'importe quel autre utilisateur. Ce lien est appelé sur Twitter un lien d'abonnement [11].

2.1.3.3 Autres réseaux sociaux

Il existe d'autres réseaux sociaux qui ont un principe de fonctionnement plus moins identique aux deux réseaux détaillés plus haut, avec un nombre d'utilisateurs moyen grand. Parmi eux, on peut citer Google Plus, Tumblr, Medium, etc.

Google Plus

Google+ est un service à mi-chemin entre *Facebook* et *Twitter*. En effet, il permet

11. following en anglais

de communiquer avec ses amis et sa famille en limitant la visibilité de vos messages et des photos à un groupe défini de personnes grâce aux cercles. Pour autant, des utilisateurs pourront vous suivre sans que vous ayez besoin de les accepter en tant qu'amis au préalable. Les pages des entreprises permettent aux marques de communiquer vers leurs clients.

Tumblr

Sur *Tumblr*, le cupcake, le Gif animé, les blagues de gestionnaires de communautaires et les "boobs" règnent en maître. Tumblr est une plate-forme permettant de publier des textes, citations, liens, photos, sons et vidéos de manière simple sans passer par la création fastidieuse d'un blog. Mais avec des possibilités intéressantes de personnalisation au niveau du design de votre page.

Medium

Medium est un réseau social pour les écrivains, les experts et les penseurs qui souhaiteraient laisser traîner leur plume en longueur. Medium est une interface incroyablement belle et simple pour publier des articles ou des histoires sans limite de longueur puis de les publier dans des « collections » classées par thème et au sein desquels chacun peut venir collaborer et tenter d'enrichir le flux de nouvelles.

2.1.4 Réseaux sociaux multicouches

Les réseaux sociaux tels que *Facebook, Twitter, Viadeo, Linkedin* deviennent de plus en plus populaires. Souvent ces réseaux sont utilisés par les mêmes individus. Plusieurs systèmes qui ont des natures de liens différentes, peuvent être vus comme un seul système. Dans le cas des réseaux sociaux, on parle de réseau social multicouche RSMC [12]. Sur ces types de réseaux l'essentiel des travaux demeurent théoriques. Les RSMC peuvent être utilisés dans beaucoup de circonstances.

12. Mutilayer social Network MLSN

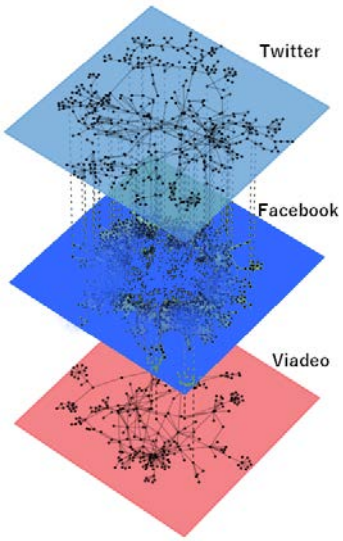

FIGURE 2.3: *Un réseau social multicouche (facebook, twitter, viadeo)*

2.1.4.1 Agrégation de plusieurs réseaux sociaux

Un utilisateur peut avoir un compte dans plusieurs réseaux sociaux. Le même utilisateur peut être détecté grâce à :

- aux adresses e-mails qui sont uniques à la personne,
- l'authentification unifiée qui est le fait d'utiliser le même compte pour se connecter dans plusieurs réseaux différents,
- l'hémophilie qui a été étudiée par Crandall David *et al.* [16] et par Aiello Luca Maria *et al.* [17]. L'idée de l'homophilie. Cependant, étant donné que l'homophilie est également utilisée de plus en plus pour désigner le simple fait cumulatif que les utilisateurs des réseaux sociaux sont similaires. L'hémophilie peut être utiliser pour détecter les utilisateurs qui partagent le même centre d'intérêt (détection des communautés) ou aussi les utilisateurs similaires de deux réseaux différents.
- etc.

Ainsi des utilisateurs identiques dans plusieurs réseaux sociaux peuvent être détectés. Donc plusieurs réseaux sociaux peuvent être vus comme une agrégation d'un seul réseau social. Une interconnexion de réseaux sociaux sera appelée RSMC. Car-

dillo Alessio *et al.* [18] et Magnani Matteo *et al.* [19] montrent qu'un RSMC peut être une agrégation de plusieurs de réseaux sociaux en ligne, hors ligne et hybride. Ils montrent aussi qu'il peut être une agrégation de plusieurs réseaux sociaux qui ont différents types de liens tels que contact, communication. Ces RSMC sont observés dans plusieurs domaines tels que : transport aérien étudié par Cardillo, Alessio *et al.* [18], théorie des jeux en lignes étudiée par Szell Michael *et al.* [20]. Kivelä, Mikko *et al.* [21] proposent des exemples complets de RSMC. Dans la figure 2.3, nous avons un exemple de réseau social à trois couches, composé de trois RSM.

2.1.4.2 Un réseau social avec plusieurs natures de relations

Un autre exemple est la propagation d'idées entre des utilisateurs connectés par plusieurs types de relation (Membres de la famille, amis, voisins, relations de travail ou d'affaires, etc.). Pour ajouter plus de réalisme dans l'étude de réseaux sociaux, il est nécessaire de s'intéresser à toutes les relations y compris les relations simultanées existant au sein d'un même ensemble social et à l'interaction entre ces relations. Par exemple, dans le cas d'une entreprise, les employés peuvent avoir plusieurs relations simultanées comme des relations d'amitié, de collaboration et de conseil. Il faut donc des outils permettant de manipuler ces réseaux (ensembles de relations) dans lesquels deux sommets (par exemple deux employés) peuvent être reliés par plusieurs natures de relations. Chaque couche du RSMC correspond à un type de relations entre les utilisateurs. Différentes relations peuvent résulter du caractère de la connexion, des types du canal de communication ou des types de diverses activités de collaboration que les agents (par exemple, les utilisateurs de divers services informatiques) peuvent effectuer dans un système ou dans un environnement donnés. Les exemples de différentes relations peuvent être : l'amitié, la famille et le travail. A titre d'exemple, sur la figure 2.4, les paires (a, L_1), (a, L_2) et (a, L_3) identifient des sommets spécifiques dans les différentes couches.

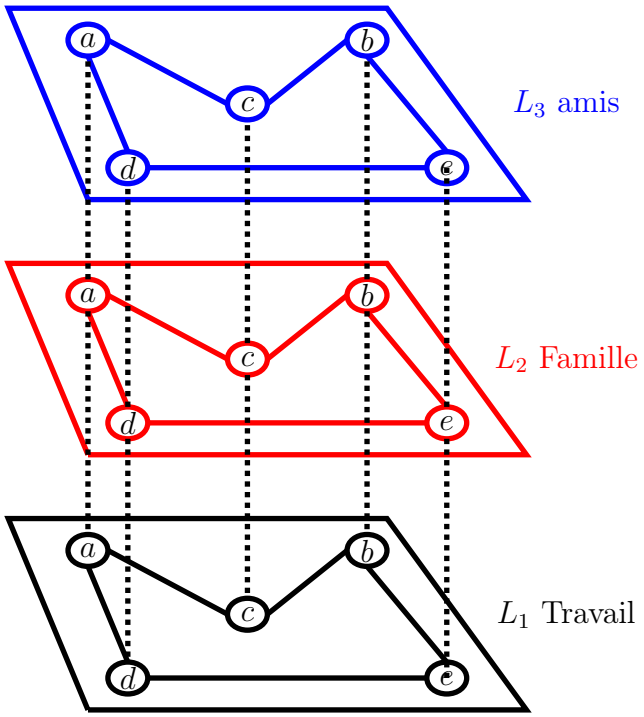

FIGURE 2.4: *Un réseau social multicouche (famille, amis, travail)*

2.1.4.3 Autres exemples de réseaux multicouches

Dans le problème de maximisation de l'influence, les tranches d'âges, le sexe, le comportement, etc comme le montre Li Chao *et al.* [6] peuvent être importants. Si nous prenons l'exemple des tranches d'âge, un RSMC peut être construit et chaque couche représente une tranche d'âge. Les utilisateurs dont leur âge coïncide avec les bornes vont être dans deux couches. L'information va circuler entre les couches via ces utilisateurs. Considérons les tranches d'âge suivantes : 0 à 20 ans, 20 à 30 ans et plus de 30 ans. Ici nous avons les trois tranches qui représentent une couche. L'intervalle 0 à 20 ans va représenter la couche L_1, l'intervalle 20 à 30 ans va représenter la couche L_2 et plus 30 ans va être représenté par la couche L_3. Les utilisateurs qui ont 20 ans vont être présents dans les couches L_1 et L_2. Les utilisateurs qui ont 30 ans vont être dans les couches L_2 et L_3. Dans la maximisation de l'influence dans les réseaux sociaux, les utilisateurs qui appartiennent à deux

couches peuvent communiquer avec les utilisateurs de chaque couche.

2.2 Analyse des réseaux sociaux

Vu le nombre important d'utilisateurs, l'analyse des réseaux sociaux (ARS) [13] devient une nouvelle branche de recherche. Elle repose sur la théorie des graphes et d'algorithmes d'optimisation permettant d'évaluer la position relative de chaque entité au sein d'un réseau et la densité des liens entre ces dernières. On mesure la centralité des sommets. Cet indicateur permet de connaître le rôle d'une entité et d'évaluer leur proximité pour déterminer des communautés. On peut répondre ainsi à des questions telles que : qui sont les intermédiaires, qui partagent le même centre d'intérêt, les plus influents, les leaders, les points isolés, où sont les grappes et qui en fait partie, qui est au coeur du réseau, et qui en est à la périphérie ... ? Ces méthodes sont utilisées dans beaucoup de domaines tels que les telecoms, le marketing, la fraude, la surveillance. L'analyse des réseaux sociaux comporte plusieurs axes de recherche, comme le montre la figure 2.5, qui peuvent être regroupés en quatre : les processus dynamiques, exploitation, la fouille et la préparation des données. Dans les processus dynamiques, on peut étudier le comportement de la diffusion, l'influence, l'évolution des réseaux sociaux au cours du temps. Dans l'exploitation, on peut avoir des recherches sur la représentation, sur des mesures qui donnent la position, l'importance d'un utilisateur dans le réseau. Dans la préparation des données, on peut étudier les tranformations du réseau par exemple pour une optimisation afin de rendre efficace un processus, on peut aussi étudier les données manquantes, etc. Dans le domaine de la l'apprentissage, on peut prédire les liens, détecter les communautés, etc.

Nos travaux se focalise la maximisation de l'influence pour une diffusion large de l'information en utilisant les mesures de centralité.

13. social network analysisARS

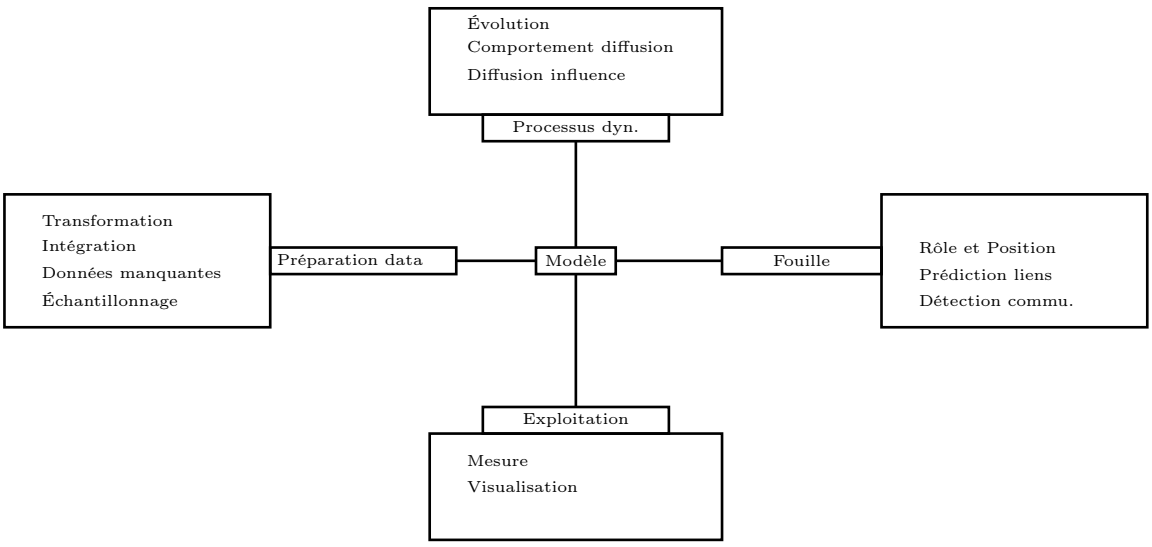

FIGURE 2.5: *ARS : axes de recherches dans les réseaux sociaux monoplex ou multi-couches*

2.3 Diffusion de l'information

Le problème de maximisation de la diffusion l'influence est $NP-Difficile$. Résoudre ce problème revient à bien choisir les diffuseurs initiaux les plus influents (appelés aussi semences ou graines) qui peuvent dépendre de plusieurs paramètres. Il faut aussi avoir un modèle de diffusion optimal ψ qui est la manière de diffuser l'information dans le réseau social à partir de ces semences. Il faut faire la différence entre la diffusion et la percolation. Cette dernière consiste à savoir si la structure permet la propagation l'information. Les modèles de diffusion pour la prévision de la diffusion de l'information dans les réseaux sociaux reposent sur de nombreux travaux menés dans divers domaines. Ils s'inspirent en particulier des travaux menés en épidémiologie, dans le but d'anticiper la propagation des maladies au sein d'une population, et en marketing, dans le but de prédire l'adoption d'un produit ou d'une technologie parmi un groupe de consommateurs. Du fait de la nature différente des problèmes abordés dans ces domaines, les modèles développés pour les traiter diffèrent de par

les hypothèses sur lesquelles ils se basent, et également de par la manière dont ils caractérisent la diffusion. En effet, les modèles épidémiologies classiques ne supposent pas l'existence d'un réseau explicite interconnectant la population d'individus étudiée, ce qui est le cas pour les modèles développés en marketing. Par ailleurs, les modèles épidémiologiques se concentrent sur l'évolution temporelle du processus de diffusion, tandis que les modèles développés en marketing s'intéressent plutôt à l'évolution structurelle de la diffusion.

2.3.1 La percolation dans les réseaux sociaux

La théorie de la percolation a été introduite en 1957 par Broadbent Simon R *et al.* [22], pour analyser la pénétration d'un gaz dans un labyrinthe formé de passages ouverts ou fermés. À l'origine, l'objectif était de comprendre comment les masques à gaz des soldats devenaient inefficaces. Ces masques sont en effet constitués de petites particules de carbone poreuses qui forment un réseau aléatoire de tunnels inter-connectés. Si les pores sont assez larges et suffisamment connectés, le gaz passe à travers les particules. En revanche, si les pores sont trop petits ou s'ils sont imparfaitement connectés, les émanations ne peuvent plus traverser le filtre. L'efficacité de la solution dépend donc d'un seuil critique qui est caractéristique du phénomène de percolation. D'une façon générale, l'étude de la percolation vise à mettre en évidence les phases de transitions sur des structures aléatoires. Ces transitions sont généralement liées à la valeur critique d'un paramètre clé, appelé seuil de transition ou seuil de percolation, à partir duquel un système subit un changement brutal d'état qui permet la pénétration d'un élément.

Dans le contexte de l'étude des phénomènes de propagations sur les réseaux sociaux, la théorie de percolation a été utilisée pour répondre à des questions telles que : la structure du réseau permet-elle une propagation du phénomène ? ou quel pourcentage d'individus vont potentiellement être affectés ? Il s'agit d'évaluer la probabilité d'existence d'un ensemble de liens, permettant la connexion directe ou indirecte

entre deux entités du réseau. Plus précisément, on s'intéresse au seuil de paramètres critiques qui garantissent la connexité de la structure, c'est-à-dire le maintien d'une composante principale géante capable de supporter un phénomène de propagation et d'affecter un maximum de sommets. Quand une telle composante est maintenue, on dit que le réseau *"percole"*. Dans le tableau 2.3, nous avons la dualité entre le phénomène de percolation et de diffusion dans les réseaux sociaux.

TABLEAU 2.3: *Dualité entre le problème de Percolation et de Diffusion*

	Percolation	**Diffusion**
Propagation	Déterministe	Aléatoire
Topologie du réseau	Aléatoire	Déterministe

2.3.2 Les modèles épidémiques

Depuis longtemps, la diffusion est observée et étudiée dans de nombreux domaines. Kermack William O *et al.* [1] science étudient la diffusion des maladies, l'informatique. Serazzi Giuseppe *et al.* [2], dans le domaine de l'informatique, étudient la propagation des virus et Rogers Everett M [3] étudie la diffusion des innovations technologiques. Brockmann Dirk *et al.* [4] étudient les déplacements humains. Le phénomène de diffusion de l'information en particulier qui peut être défini comme l'action de propager des éléments d'information auprès d'un public, suscite depuis plusieurs années un grand intérêt au sein de la communauté scientifique. Nous avons des modèles épidémiques qui sont des modèles mathématiques précurseurs basés sur le concept de mélange homogène. Nous avons aussi des modèles plus récents basés sur les réseaux, qui visent à intégrer la complexité des interactions humaines impliquées. Les modèles épidémiques considèrent une population comme un ensemble de groupes (c'est-à-dire des compartiments), caractérisés chacun par l'état des individus au regard de l'épidémie. Dans ce type de modèle, on suppose que les individus changent de compartiments de façon homogène. On parle de mélange homogène ou

d'action de masse. D'une certaine façon, cette approche suppose que les individus au sein d'un même compartiment entretiennent une structure relationnelle régulière avec les individus des autres compartiments.

2.3.2.1 Le modèle SI (Susceptible-Infected)

Le modèle Susceptible-Infected (SI), proposé par Kermack William O *et al.* [1], est l'un des modèles épidémiques les plus simples. Dans ce modèle, deux groupes d'individus peuvent être identifiés : les susceptibles (S) et les infectés (I). Les susceptibles sont les individus qui peuvent contracter la maladie s'ils entrent en contact avec des individus infectés. Les infectés sont, eux, des individus porteurs de la maladie, qui peuvent la transmettre lors de contacts avec des susceptibles. Les individus infectés ont une probabilité α d'établir un contact avec un individu susceptible. Cette approche simpliste modélise la propagation de la maladie par le passage de l'état susceptible à l'état infecté. Ce modèle suppose qu'un individu infecté reste dans cet état. Dans la figure 2.6, nous avons représenté les deux états de ce modèle. Une personne dans le groupe S peut devenir I via une probabilité. La figure 2.7 donne l'allure des deux ensembles par rapport au temps. A $t = 0$, la maladie commence à atteindre la population. Le temps passe, le nombre de personnes qui ne sont pas atteinte, représenté par l'ensemble S, diminue progressivement. Le virus finit par infecter toute la population. Après infection, on peut avoir des personnes guéries de la maladie. Ce phénomène n'est pas pris en compte par le modèle SI. Il sera corrigé par le modèle SIR.

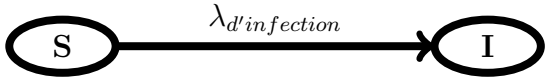

FIGURE 2.6: *Le modèle SI (Susceptible-Infected)*

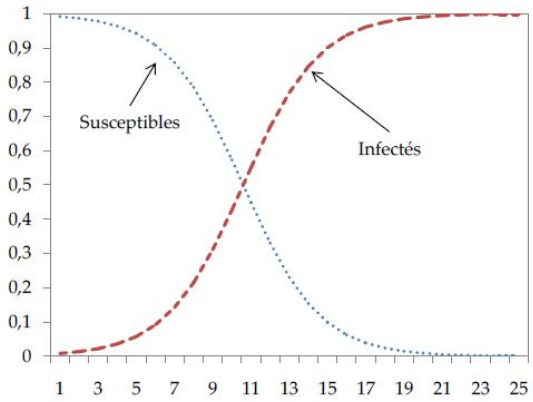

FIGURE 2.7: *Courbe d'évolution du modèle SI (Susceptible-Infected)*

2.3.2.2 Le modèle *SIR* (Susceptible-Infected-Recovered)

Le modèle Susceptible-Infected-Recovered (SIR), proposé aussi par Kermack William O *et al.* [1], est également l'un des modèles de diffusion épidémique les plus simples et les plus fréquemment retrouvés dans la littérature. Il constitue une évolution directe du modèle *SI*. Dans le modèle *SIR*, les deux états du modèle SI sont conservés (Susceptibles et Infected), auquel est ajouté le troisième état R (Recovered), qui n'est atteint que par les individus infectés selon une certaine probabilité λ. Dans ce modèle, le terme "Recovered" indique que l'individu est sorti de son état d'infection et ne peut plus contracter la maladie, soit parce qu'il devient immunisé, soit à la suite d'un décès. Ce modèle suppose qu'un individu dans l'état R conserve son immunité. Par conséquent, un individu recovered ne peut pas être de nouveau susceptible ou infecté. Dans la figure 2.8, nous avons représenté les trois états du modèle *SIR* et les probabilités pour qu'une personne dans le groupe S soit infectée et pour qu'une personne infectée soit guérie. La figure 2.9 est la courbe d'évolution des trois états au fil du temps. Au début toutes les personnes sont susceptibles d'être infectées par le virus. Elles seront dans le groupe S. Au fil du temps, certaines personnes de S deviennent infectées et elles seront dans le groupe I. Ils introduisent une probabilité

d'infection λ_I. Le nombre de personnes, dans le groupe S, commence à diminuer. A un certain moment, le virus peut être maîtrisé. Certaines personnes infectées guérissent et sont dans le groupe R. Ils mettent une probabilité de guérissons λ_G

FIGURE 2.8: *Le modèle SIR (Susceptible-Infected-Recovered)*

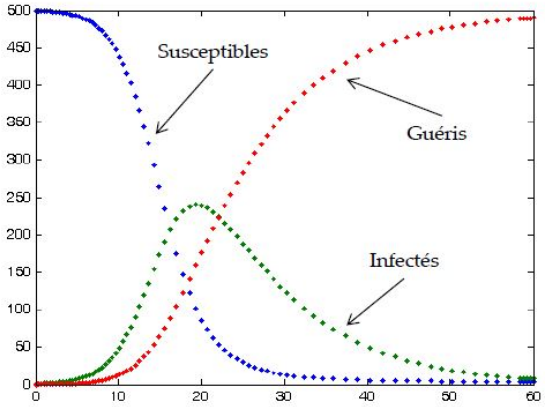

FIGURE 2.9: *Courbe d'évolution du modèle SIR (Susceptible-Infected-Recovered)*

Dans ce modèle, par exemple, une personne qui n'est pas infectée (dans le groupe S) peut prendre un vaccin et devient directement R. Pour palier à ces genres de manquement, d'autres modèles de propagation beaucoup plus complexes sont proposés.

2.3.2.3 Autre modèles épidémiques

Des modèles beaucoup plus élaborés ont également été proposés, tels que les modèles *SEIS, SEIR, MSIR*, ... qui représentent des situations plus complexes, comme celle des individus qui ne sont pas totalement guéris, mais qui continuent de propager la maladie. Une présentation détaillée de ces différents modèles mathématiques peut être trouvée dans les travaux de Easley David *et al.* [23] et de M. Newman. [24].

2.3.3 La diffusion dans les réseaux sociaux

Avec le développement des TIC [14] (smartphones, tablettes, lunettes à réalité augmentée, etc.), l'émergence de nouveaux réseaux sur l'Internet (blog, site d'échanges et de partages, sites communautaires, etc.), et le large accès aux transports en commun, qui permettent aujourd'hui de diffuser une information à très grande échelle et en très peu de temps, l'étude des phénomènes de diffusion est devenue un enjeu majeur dans de multiples contextes. Typiquement, il est crucial pour une entreprise de comprendre et de maîtriser comment une nouvelle ou un récit peut se propager et affecter son image. De même, dans le domaine du marketing, il devient essentiel de savoir quels sont les individus à cibler pour maximiser les ventes par du marketing viral. Les modèles de diffusion épidémiques sont ceux qui ont reçu le plus d'attention dans la littérature pour leur intérêt dans de nombreux domaines. Bien que les modèles épidémiques aient été largement utilisés dans l'étude des phénomènes de diffusion, l'hypothèse selon laquelle les utilisateurs ont une même probabilité d'établir des contacts s'avère être irréaliste. En effet, dans la réalité les contacts qu'entretiennent les utilisateurs sont souvent hétérogènes, puisque les individus ne sont généralement connectés qu'à une petite proportion d'utilisateurs et cette proportion n'est jamais choisie aléatoirement. Tous les modèles centrés sur la topologie du réseau ont deux modèles de référence : le modèle cascade indépendante [15] et le modèle seuil linéaire [16]. Dans nos travaux, nous avons utilisé ces deux comme modèle de diffusion pour calculer la propagation de l'information.

2.3.3.1 Les modèles de base : Cascade indépendante et Seuil linéaire

Ces travaux, menés initialement dans le domaine du marketing, modélisent le processus de diffusion au sein d'une population constante de N utilisateurs interconnectés

14. Technologies de l'Information et de la Communication
15. Independent Cascade (IC)
16. Linear Threshold (LT)

par un réseau statique décrit par un graphe orienté G(V,E). Ces modèles supposent que l'information ne peut se propager que le long des liens de ce réseau. Il existe deux manières de modéliser ce type de processus de diffusion.

- Elle est centrée sur l'utilisateur qui doit recevoir l'information : on parle alors de "modèle seuil", tel que le modèle de seuil linéaire développé par Granovetter Mark [25]. Dans ces modèles, l'utilisateur est influencé par ses voisins actifs,
- Ou bien elle est centrée sur l'utilisateur qui va diffuser l'information : on parle alors de "modèle cascade", tel que le modèle des cascades indépendantes proposé par Goldenberg Jacob *et al.* [26]. Dans ces modèles, l'utilisateur va influencer ses voisins inactifs.

Dans les deux cas, on considère que chaque membre du réseau peut être soit inactif, soit actif ; un utilisateur actif étant un utilisateur ayant reçu l'information et participent à sa propagation. Dans ces deux modèles, un utilisateur ne peut pas nier avoir diffusé l'information. En d'autres mots, un utilisateur actif le reste à jamais. Ces modèles caractérisent un processus de diffusion par une séquence d'activation le long d'un axe temporel discret, puisqu'ils modélisent la diffusion comme un processus itératif où les membres du réseau changent d'état de façon monôtone (i.e. les membres actifs ne peuvent pas redevenir inactifs) et synchrone. Par conséquent, et contrairement aux modèles épidémiques, ces modèles se concentrent sur l'aspect structurel de la diffusion. L'arbre de la diffusion peut même être généré afin d'avoir toutes les informations sur le processus de la diffusion.

2.3.3.2 Cascade indépendante

Goldenberg Jacob *et al.* [26] proposent le modèle cascade indépendante IC. Ils supposent que chaque sommet u nouvellement actif influence, indépendamment des autres, chacun de ses voisins inactifs. Le modèle IC qui met l'accent sur les utilisateurs qui vont diffuser l'information, requiert que l'on définisse pour chaque arête uv la probabilité λ_{uv} que l'utilisateur u influence son voisin inactif v de sorte que

celui-ci passe à l'état actif. Étant donné un ensemble S d'utilisateurs du réseau social initialement actifs, le processus de diffusion se déroule d'une manière itérative. A une itération t, chaque utilisateur v a un seul sens d'activer ses voisins w avec une probabilité λ_{vw}. A la prochaine itération $(t+1)$, l'utilisateur v ne participera au processus de diffusion, par contre l'utilisateur w y participera. Le processus s'achève lorsqu'aucune nouvelle activation n'est plus possible. Pour illustrer le modèle IC, nous prenons le graphe social de la figure 2.10 composée de 11 utilisateurs. La figure 2.11 simule le processus de propagation du modèle IC. Nous avons choisi comme semences, les sommets 4 et 10. Chacun de ces sommets peut activer ses voisins via une probabilité définie. Le sommet 4 a un seul sens d'activer chaque élément de l'ensemble N(4). Dans la simulation, le sommet 4 est parvenu avec une pression suffisante à activer ses voisins 1, 5 et 6. La pression sociale n'est pas suffisante pour activer ses autres voisins 3 et 7. Une même scenario pour le sommet semence 10 qui a activer avec une pression sociale suffisante ses voisins 7 et 11. A la prochaine itération, tous les sommets activés par les semences vont à leur tour essayer d'activer leurs voisins nos actifs. A l'itération 3, les sommets activés à l'itération 2 n'ont pas de pression sur leurs voisins non actifs. Alors il n'y a pas de sommets nouvellement activés. Le processus de diffusion pour le modèle IC s'arrête là. On voit que le sommet 2 n'est pas activé. Aucun de ses voisins n'a de pression sur lui.

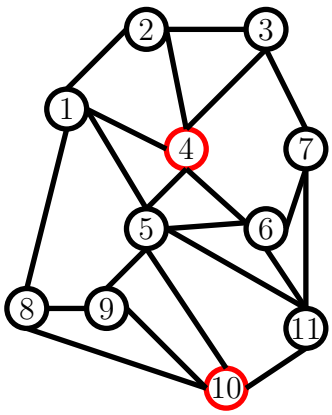

FIGURE 2.10: *Un graphe social avec deux semences en rouge*

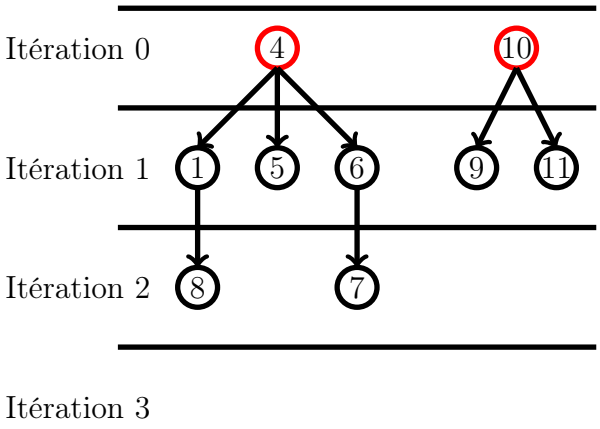

FIGURE 2.11: *Processus de diffusion selon le modèle IC*

2.3.3.3 Seuil linéaire

Granovetter, Mark [25] propose le modèle seuil linéaire fondé sur le principe selon lequel le passage de l'état inactif à l'état actif d'un utilisateur du réseau social dépend de l'influence exercée par ses voisins actifs dans le réseau social. Ce modèle met l'accent sur les utilisateurs qui vont être activés. Chaque arête uv du graphe G(V,E) est associé à un paramètre λ_{uv} représentant la pression sociale (ou le degré d'influence) que l'utilisateur u exerce sur son voisin v. Chaque utilisateur v du réseau social est associé à un seuil d'influence (une résistance sociale) θ_v. Par ailleurs, pour chaque sommet v du graphe G(V,E), la somme des pressions sociales de l'ensemble des voisins N(v) est déterminée. Si cette somme est supérieure à la résistance sur sommet v, il devient actif à un instant t donné. Ce dernier va contribuer à l'activation de ses voisins à l'instant $t+1$. Mathématiquement, l'utilisateur v sera actif si l'équation 2.2 est vérifiée avec la condition de l'équation 2.1.

$$\sum_{u \ actif \ et \ u \in N(v)} p_{(u,v)} \prec 1 \tag{2.1}$$

$$\sum_{u \ actif \ et \ u \in N(v)} p_{(u,v)} \succ \theta_v \tag{2.2}$$

Les figures 2.12 et 2.13 simulent le processus de propagation du modèle LT. Dans la figure 2.12, nous considérons deux utilisateurs 4 et 10 comme étant les semences. Dans ce modèle, on donne de l'importance que utilisateurs qui vont recevoir l'information. A la première itération, L'utilisateur 1 a comme voisin actif seulement l'utilisateur 4 qui est une semence. La somme des probabilités (la pression sociale ou le degré d'influence) de tous ses voisins actifs se réduit à la probabilité de l'utilisateur 4. Comme on le voit avec la figure 2.13, le degré d'influence de l'utilisateur semence 4 est inférieur au seuil (la resistance sociale) de l'utilisateur 1, alors ce dernier ne sera pas activé. Par contre l'utilisateur semence a réussi à influencer l'utilisateur 5. Le seuil développé par l'utilisateur 5 est inférieur à la somme des degrés d'influence de tous ses voisins actifs. Dans une autre itération dans la figure 2.13, les voisins de l'utilisateur 1 ne se résument plus à la semence 4, mais il y a aussi l'utilisateur nouvellement activé par la semence 10 qui est l'utilisateur 5. Pour l'activation de l'utilisateur 1 qui a deux voisins actifs, la somme de leur pression sociale est comparée avec le seuil d'influence développé par l'utilisateur 1. Ce dernier est actif par-ce-que son seuil est inférieur à la somme des pressions sociales de ses voisins actifs. Ainsi, Le processus continue jusqu'à ce qu'il n'y ait plus possibilité d'activer un sommet.

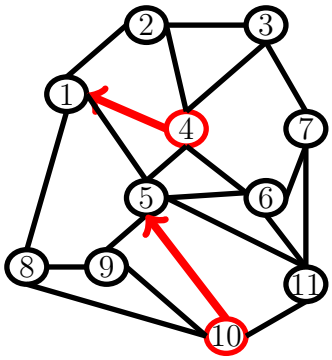

FIGURE 2.12: *Un graphe social avec deux semences*

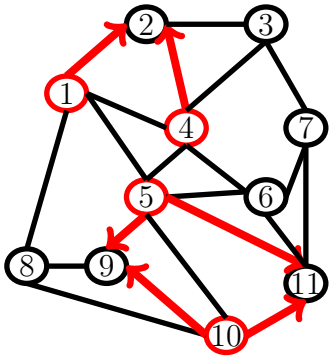

FIGURE 2.13: *Processus d'activation d'un sommet sous le modèle LT*

2.3.3.4 Autres modèles de propagation

Il existe d'autres modèles de propagation dérivant du modèle IC ou du modèle LT. Kimura Masahiro *et al.* [27] ont proposé en 2006 un modèle de propagation qui se base sur le modèle LT. Ce modèle utilise la plus petite distance entre un sommet u inactif est l'ensemble S_k^*. Ce modèle est appelé Shortest-path Model (SPM). La distance entre un sommet u et l'ensemble des semences est définie par l'équation 2.3.

$$d(S_k^*, u) = Min_{v \in S_k^*} d(v, u) \tag{2.3}$$

Si u n'est pas accessible via S_k^* alors sa distance sera infinie. Dans ce même papier, ils proposent une généralité de SPM qu'ils appellent modèle Shortest-path (SP1). Cette généralité ne se base plus sur le modèle LT seulement. Chaque sommet inactif a la chance d'être activé seulement à $t = d(S_k^*, v)$ et à $t = d(S_k^*, v) + 1$. Lagnier, Cédric*et al.* [28] proposent aussi un modèle centré sur l'utilisateur. Ce modèle prend en compte le profile de l'utilisateur et le contenu de l'information à diffuser. Des similarités entre le profil de la personne et celui du message sont évaluées via un seuil. Ils montrent aussi que si la fonction d'influence σ de l'algorithme Greedy hill climbing (définie dans le paragraphe 2.4) n'est pas sous-modulaire, on peut aussi

avoir une bonne approximation mais elle n'est pas garantie.

2.3.4 Modèles IC et LT dans les réseaux sociaux multicouches

Un RSMC est une agrégation de plusieurs RSM qui représentent chacun une couche.qui peut être une nature de relation, un groupe d'âge, etc. Les travaux effectués dans les RSM ne sont pas applicable dans les RSMC. Li Chao *et al.* [6] proposent une heuristique appelée $k-shell$ qui détermine les utilisateurs les plus influents dans les RSM. Kitsak, Maksim *et al.* montrent dans [29] que l'heuristique $k-shell$ n'est pas efficace dans les RSMC (interconnection de réseaux). Ils introduisent une nouvelle heuristique utilisant les caractéristiques des RSMC. Cela a motivé d'autres chercheurs à redéfinir les modèles proposés dans la maximisation de l'influence dans les RSM.

2.3.4.1 Le modèle seuil linéaire

Le principe est le même que dans les RSM. Chaque utilisateur va développer un seuil d'activation (une résistance sociale). Ses voisins vont développer une pression sociale (degré d'influence). Dans les RSMC, un utilisateur peut participer dans plusieurs couches. Li *et al.* **(author?)** [6] montrent que l'utilisateur va développer une résistance globale θ^G et une résistance locale (dans chaque couche L_v où il se trouve) θ_v^L. La résistance d'un utilisateur sera définie par l'équation 2.4

$$Th(v) = a(\theta^G + \theta_v^L) \tag{2.4}$$

On désigne par a le facteur d'ajustement du seuil. Mathématiquement, un utilisateur u dans une couche donnée, sera influencé si la condition de l'équation 2.5 est vérifiée et la somme des pressions sociales de ses voisins est supérieure à la résistante $Th(u)$

développée par le sommet u (voir équation 2.6).

$$\sum_{v \ active \ et \ v \in N(u)} p_{(v,u)} \prec 1 \tag{2.5}$$

$$\sum_{v \ active \ et \ v \in N(u)} p_{(v,u)} \succ Th(u) \tag{2.6}$$

2.3.4.2 Le modèle cascade

Dans les RSM, les modèles cascades donnent de l'importance aux sommets qui ne sont pas encore activés. L'influence se fait d'un sommet vers ses voisins directs. Salehi Mostafa *et al.* [30] proposent un modèle de propagation cascade dans les RSMC qui conserve toujours l'idée générale dans les RSM. Dans ce modèle, l'information peut circuler entre les couches et dans la même couche. Mais il ne faut pas perdre le fait qu'un sommet peut avoir des représentants dans les autres couches. Comme on l'a représentée dans le tableau 2.4, l'information peut circuler dans la même couche avec la proximité et entre les couches par la représentativité. Dans la même couche, un utilisateur va essayer d'influencer ces voisins de niveau 1. Dans une autre couche, l'influence se propage via ses représentants et ces derniers vont essayer d'influencer leurs voisins dans la couche où ils se trouvent.

TABLEAU 2.4: *Les différentes possibilités de propagation d'une couche à une autre dans un réseau multicouche*

	Même sommet	Autre sommet
Même couche	Rien	diffusion vers un voisin de la même couche
Autre couche	diffusion vers un représentant dans une autre couche	diffusion vers un autre voisin dans une autre couche

2.4 Maximisation de l'influence et algorithme glouton

Domingos Pedro *et al.* [31] sont les premiers à attaquer le problème de maximisation de l'influence dans le domaine de marketing. Ils le voient comme un problème d'optimisation stochastique. Leurs travaux sont repris par Kempe David *et al.* [11] qui proposent une approche basée sur les modèles IC et LT pour le choix des semences. Ils choisissent un ensemble S_0, comme l'ensembles des diffuseurs initiaux (sommets actifs), qui va commencer le processus de la diffusion de l'information. Ils définissent l'influence d'un ensemble d'utilisateurs S, noté par $\sigma(S)$, comme le nombre d'utilisateurs influencés à la fin du processus de propagation. Dans leur approche, ils font des combinaisons de S_0 jusqu'à celle qui va maximiser la fonction σ. Kempe David *et al.* [11] essayent de résoudre le problème en proposant un algorithme glouton ou Greedy hill-climbing algorithm (algorithme 1). Il commence par choisir le meilleur initiateur. Ensuite, il choisit le second qui offre le meilleur gain marginal par rapport au premier utilisateur déjà choisi. L'algorithme continue ensuite jusqu'à sélectionner un ensemble des $K-$*utilisateurs*. Cet algorithme offre une bonne approximation quand la fonction que l'on veut maximiser respecte certaines propriétés. Les premiers résultats montrent que sous les modèles de base IC et LT, une approximation de 63% est garantie. L'algorithme qui réalise cette garantie de performance est une stratégie naturelle de recherche d'un optimum local (Hill-Climbing). Il se base sur une approche liée à celle des travaux de Domingos Pedro *et al.* [31]. Les travaux sont repris par les mêmes auteurs et le fait assez surprenant est que l'algorithme glouton donne une approximation inférieur à 63%. Des études très poussées sur la fonction d'influence sont faites par Kempe David *et al.* [11]. Ils utilisent des techniques de la théorie des fonctions sous-modulaires, détaillées par Cornuejols Gerard *et al.* [32] et Nemhauser George *et al.* [33], qui se révèlent fournir un contexte naturel pour raisonner sur les deux modèles de diffusion et des algorithmes gloutons.

Une fonction f est sous-modulaire si elle satisfait la propriété suivante. Nous définissons deux ensembles S et T tel que $S \subset T$. Tous les éléments de S se trouvent dans T. Soit u un élément de l'ensemble de départ qui n'est pas dans T (évidement n'est pas dans S aussi). L'équation 2.7, détaillée dans la figure 2.14, doit être satisfaite.

$$f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T) \tag{2.7}$$

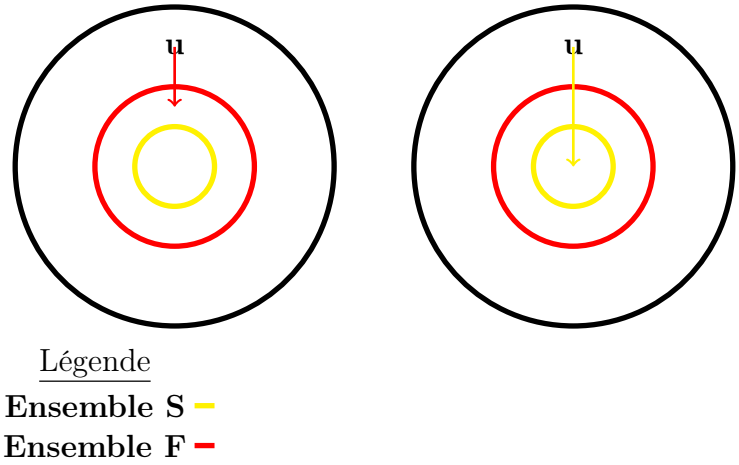

FIGURE 2.14: *Une fonction sous-modulaire f*

Nemhauser George *et al.* [33] montrent que l'ensemble des semences données par l'algorithme de glouton a une bonne approximation si la fonction d'activation est monôtone et sous-modulaire.

Théorème 2.1 *Tiré dans [33]*

Pour une fonction non négative, monôtone[17] *et sous-modulaire f, soit S un ensemble de K−utilisateurs obtenus par l'algorithme "Greedy Hill Climbing" maximisant la fonction f. Soit S^* l'ensemble qui maximise la valeur de f pour K−utilisateurs. Alors $f(S) \geq (1 - 1/\epsilon)f(S^*)$, en d'autres termes, A est une $(1 - 1/\epsilon)$−approximation.*

17. Les fonctions de diffusion des modèles étudiés sont non négatives et monotones de par la définition des modèles

Le théorème 2.1 a été utilisé dans de nombreuses applications d'optimisation discrètes comme le montre Nemhauser George *et al.* [34]. Du Ding-Zhu *et al.* **(author?)** [35] proposent une généralisation de cet algorithme qui consiste à ne pas choisir les utilisateurs un par un mais par bloc de n. L'avantage est que théoriquement on obtient une meilleure approximation (voir l'optimal si on choisit $n = K$), l'inconvénient étant que la complexité augmente drastiquement.

Données : $G(V, E)$, K (le nombre de diffuseurs initiaux)

Résultat : S optimal

1 Initialisation : $S \longleftarrow \phi$;

2 **tant que** *(*$|S| \prec K$*)* **faire**

3 $\quad u \leftarrow \arg \max_{v \in V \text{ et } v \notin S} \sigma(S \cup \{v\})$;

4 $\quad S \leftarrow S \cup \{u\}$;

5 **fin**

6 retourner S;

Algorithme 1 : *Algorithme glouton (Greedy Algorithm)*

Kempe David *et al.* [11] montrent que si la fonction σ est sous-modulaire, une approximation de 63% est garantie par l'algorithme de greedy hill-climbing (algorithme 1) sous les deux modèles IC et LT.

Conclusion

Ce chapitre s'articule sur trois points. D'abord, nous avons présenté les réseaux sociaux en montrant leur principe de fonctionnement et en montrant, avec des chiffres, la position des ces réseaux dans le web. Ensuite, nous avons rappelé les modèles de propagation épidémiologique en faisant un peu d'historique sur dans la maximisation de la diffusion de l'information où la manière de diffuser l'information est très importante. Nous avons parlé aussi des modèles qui ne tiennent pas en compte la topologie du réseau et des modèles de base qui prennent en compte la topologie du réseau. Enfin, nous avons parlé du principe de la maximisation de l'influence en

expliquant sa complexité et l'approche gloutonne. Comme dans tous nos travaux, dans la problématique posée, nous donnons de l'importance aux diffuseurs initiaux. Dans le chapitre suivant, nous allons faire une étude générale sur la détection des semences et donner notre contribution dans la maximisation de l'influence.

Chapitre 3

Choix des semences dans la maximisation de l'influence

Dans le processus de maximisation de l'influence, il est important d'avoir un bon modèle de propagation tout comme il est très important de bien choisir les utilisateurs qui vont commencer la diffusion de l'information. Dans cette thèse, nos travaux se focalisent essentiellement sur la détection des diffuseurs initiaux appelés aussi les semences [1]. Déterminer ces derniers est un problème $NP-Difficile$ selon Kempe David [11]. Il est un problème combinatoire très difficile. Il n'est pas facile de combiner $k-$utilisateurs qui vont maximiser la propagation de l'influence. Tous les modèles proposés sont des heuristiques car il n'existe pas une combinaison de semences qui va atteindre tous les utilisateurs du réseau social. Donc plus on a une bonne approximation plus l'heuristique proposée est optimale. La plupart des travaux effectués dans les RSM sont redéfinis dans les RSMC. Ce chapitre est composé de quatre parties. Dans la première partie, nous présenterons un état de l'art sur la détection des semences en les divisant en trois approches et en montrant les limites de ces dernières. Dans les trois dernières parties, nous présenterons nos travaux effectués dans la détection des semences. D'abord, nous mettrons en

1. seeds ou les bonnes graines

limière la mesure de centralité degré de diffusion ℓ-ième, ensuite la centralité *degré multi-diffusion* applicable dans les réseaux sociaux mutilicouches et enfin le graphe couvrant de maximisation.

3.1 Etat de l'art sur la détection des semences

Comme nous l'avions annoncé dans le chapitre 2, résoudre le problème de maximisation de l'influence pour une diffusion large de l'information revient à détecter les semences et à trouver un modèle de propagation optimal. Dans nos travaux, le thème principal est le premier point qui est la détection des semences. Plusieurs travaux qui peuvent être divisés en trois approches, ont été effectués. On peut avoir :

- une sélection de semences statique qui consiste à donner à chaque utilisateur son taux d'influence (une mesure de centralité) d'une manière fixe,
- une sélection dynamique qui consiste à donner à chaque utilisateur son taux d'influence en le modifiant en fonction des sélections ou à prédire le chemin de l'influence initié par un utilisateur.
- une sélection gloutonne qui est de choisir les semences une à une.

Certains travaux, nous avons une combinaisons des approches dans le but d'utiliser les avantages des chacune.

3.1.1 Sélection statique

Dans beaucoup de travaux, l'approche basée sur la centralité de l'utilisateur est utilisée. Cette centralité est une mesure qui est le degré de connexion de l'utilisateur par rapport aux autres. Les $k - top$, autrement dit les k utilisateurs qui ont la plus grande mesure de centralité, sont considérés comme plus influents. D'abord, Kempe David *et al.* dans [11] en 2003 proposent une *centralité de degré* (*degree centrality*). Ils affirment que les individus plus influents du réseau sont les utilisateurs qui ont plus de voisins. Ils considèrent les k utilisateurs qui ont la plus grande mesure de cen-

tralité comme les utilisateurs les plus influents. Cette mesure est efficace uniquement dans les RSM. Elle sera redéfinie par Bródka Piotr *et al.* [36] et Magnani Matteo *et al.* [9] en 2011 pour les RSMC. Ils considèrent les voisinages dans toutes couches en sachant qu'un utilisateur peut avoir des représentants dans plusieurs couches. Les mesures de centralité de degré [11] et *multi Degré* [36] se basent sur le nombre de voisinage. En 2010 Maksim Kitsak *et al.* [29] proposent une heuristique qui se base sur la centralité degré et un regroupement. Leur heuristique est appelée $k-shell$ (les k coquilles) et pour détecter les utilisateurs les plus influent pour une diffusion large de l'information, ils utilisent la décomposition du réseau en *"k-shells"*. L'heuristique détermine, pour chaque sommet du réseau, à quel point il est bien connecté, et à quel point ses voisins sont bien connectés. Ils montrent que, quel que soit le graphe social considéré, les sommets les plus influents dans la diffusion sont ceux qui ont un indice de *k-shell* très élevé. Pour cela, ils choisissent des sommets sources de la diffusion appartenant à des *k-shells* aux indices différents, et observent la vitesse de diffusion et le nombre de sommets atteints dans chaque cas. Il fonctionne de la manière suivante. Les sommets sont affectés au groupe k (la coquille k) selon leur degré restant, ce qui est obtenu par la taille successive des sommets avec un degré inférieur à la valeur de k_s du groupe courant. Ils commencent par enlever tous les sommets de degré k = 1. Après avoir enlevé tous ces sommets, certains peuvent être laissés avec un lien, donc nous continuons l'extraction de manière itérative jusqu'à ce qu'il n'y ait plus de sommets avec k = 1 dans le réseau. Les sommets supprimés, ainsi que les liens correspondants, forment une couche k avec index $k_s = 1$. De la même façon, on extrait par itération la prochaine couche k, $k_s = 2$ et continuer à extraire les groupes supérieurs à k jusqu'à ce que tous les sommets sont supprimés. Ainsi, chaque sommet est associé à un indice de k_s, et le réseau peut être vu comme l'union de tous les groupes. Comme le problème est d'avoir une diffusion large d'une information, les heuristiques proposées ne tiennent en compte le fait qu'un utilisateur puisse influencer ou non ses voisins. En 2011, Kundu Suman *et al.* [37] proposent

une mesure de centralité qui donne les individus plus influents dans le réseau social. Cette centralité se basse sur une probabilité de diffuser l'information. Ils montrent qu'un sommet n'est seulement pas important par son nombres de voisins élevés mais par la probabilité qu'ils diffusent l'information. Cette heuristique utilise le nombre de voisins d'un utilisateur et le taux d'influence de ce dernier. Cette heuristique prend en compte seulement son voisinage de niveau 1. Ils supposent que l'utilisateur u diffuse une information δ et qu'il est important que la portée de la diffusion de u soit grande. En 2012, De Meo Pasquale *et al.* [38] proposent $k-$*path edge*. Cette heuristique évalue le chemin de la diffusion en se basant sur le nombre d'arêtes. Ils donnent la définition de $k-$*path edge* ci-après.

Définition 3.1 ***Centralité*** $k-$***path edge***

Pour chaque arête du graphe G=(V,E), la centralité $k-$*path edge de e, notée par* $L^k(e)$*, est définie comme la somme, sur tous les sommets sources possibles s, de la fréquence à laquelle un message provenant de s traverse e, en supposant que les parcours de message sont seulement les chemins simples aléatoires qui ont au plus* k *arêtes.*

Jusque là, la plupart des travaux évoqués sont effectués dans les RSM. Zhao Dawei *et al.* [39] en 2014 montrent que l'heuristique $k-shell$ ne donne pas de bons résultats si on l'applique dans ces réseaux. Ils introduisent une nouvelle mesure qui considère les propriétés structurales de la propagation dans plusieurs couches. Ils redéfinissent plusieurs mesures de base pour les RSMC. Dans le chapitre 2, nous avons montré avec des chiffres l'évolution des réseaux sociaux. Chaque jour, plusieurs comptes sont crées. Le réseau évolue rapidement. Les heuristiques proposées ne tiennent pas compte de cette evolution. En 2014, Viard Jordan *et al.* [40], définissent une nouvelle notion de densité pour les graphes dynamiques. Au lieu d'étudier la densité topologique d'un graphe statique, ils définissent une mesure prenant en compte à la fois des aspects structurels et temporels. Dans la détection des semences, certains utilisent la centralité *PageRank* proposée par Brin Sergey *et al.* [41] en 1998 qui est

un outil d'analyse des liens concourant au système de classement des pages Web que le moteur de recherche *Google*[2] utilise. Si un internaute utilise le moteur de recherche *google*, il sélectionne les pages webs les plus importantes pour l'information cherchée en utilisant la mesure *PageRank*. Comme ces pages webs peuvent être vues comme un graphe alors la mesure *PageRank* est utilisée pour donner les utilisateurs les plus influents dans un réseau social. La mesure centralité *PageRank* est définie par l'équation 3.1.

$$PageRank(A) = (1-d) + d(PageRank(T_1)/C(T_1) + \cdots + PageRank(T_n)/C(T_n)) \tag{3.1}$$

où *PageRank(A)* représente la *PageRank* de la page A, *PageRank(t_i)* est la mesure de *PageRank* des pages liées à la page A, $C(t_n)$ est le nombre de liens partant de la page T_i et enfin d le facteur d'amortissement qui vatie entre 0 et 1.

Dans la détection des semences dans la maximisation de l'influence, plusieurs travaux font une sélection statique. Ils ne tiennent pas compte de la semence sélectionnée au moment de la sélection des autres. Dans d'autres travaux, la sélection se fait dynamiquement. Un sommet est choisit comme une semence en mesurant la portée qu'il donne à l'information.

3.1.2 Sélection dynamique

Dans beaucoup de travaux, les semences sont choisies dynamiquement en mettant à jour les centralités à chaque selection de semence ou en regardant le chemin le l'influence initié par un sommet. Les travaux de Kempe David *et al* [11] sont repris par Chen Wei *et al.* [42] en 2009. Au lieu de donner à chaque utilisateurs un taux d'influence fixe, ils les modifient en fonction des semences sélectionnées. Ils proposent une heuristique appelée centralité de *degré discontinu* qui est une amé-

2. www.google.com

lioration de mesure de centralité définie dans [11]. Elle fonctionne de la manière suivante : soit v un voisin du sommet u. Si le sommet v est choisi comme un élément dans l'ensemble des semences, lors de la mise à jour du sommet u, l'arête uv ne sera pas examinée. Dans l'algorithme 2, ils ont implémenté l'heuristique de degré Discontinu. Cet algorithme se limite aux premiers voisins. R. Narayanam *et al.*

Données : $G(V, E)$ et k
Résultat : S_k^*

Initialisation : $S_k^* \longleftarrow \phi$;
pour *each sommet* $v \in V$ **faire**
    calculer $C_d(v)$;
    $dd_v \longleftarrow C_d(v)$;
    initialiser $t_v \longleftarrow 0$;
fin
pour $i = 1$ *à* k **faire**
    $v \longleftarrow \arg\max_u \{dd_u,\ u \in V \text{ et } u \notin S_k^*\}$;
    $S_k^* \longleftarrow S_k^* \bigcup \{v\}$;
    pour *tout voisin* u *de* v *et* $u \in V$ *et* $u \notin S_k^*$ **faire**
        $t_u \longleftarrow t_u + 1$;
        $dd_u \longleftarrow d_u - 2t_u - (d_v - t_u)t_v p$;
    fin
fin
returner S_k^*;

Algorithme 2 : *Algorithme de l'heuristique de Degré Discontinu*

proposent ShaPley value based Influential Nodes (SPIN) [43] en 2010. Dans leur approche, deux problèmes sont résolus, à savoir, trouver les $k - top$ qui vont maximiser l'influence et trouver le plus petit ensemble qui va couvrir $\alpha\%$ les sommets du réseau qu'il appelle $\alpha-coverage$. Dans le premier point, ils mappent le processus de diffusion de l'information dans un réseau social sur la formation des jeux coopératifs [44] convenablement définis. Les valeurs de "Shapley" [45] des sommets dans ce jeu représentent la contribution marginale des sommets vers le processus de diffusion des informations. Ce fait leur permet de concevoir un algorithme pour découvrir les sommets les plus influents dans le réseau social.

Un jeu coopératif peut être analysé à l'aide d'un concept de solutions, qui fournit

une méthode de la division de la valeur totale du jeu entre les différents acteurs. Il existe de nombreux concepts de solutions comme le noyau, la valeur de "Shapley", le nucléole, etc. Le concept de "Shapley", que l'algorithme $SPIN$ utilise, est développé par Shapley Lloyd S [45]. Dans l'algorithme 3, une file est créée et est nommée par $RankList$, qui trie l'ensemble des sommets par ordre décroissant en fonction de leur valeur "Shapley".

Après le tri des sommets, R. Narayanam et Y. Narahari proposent l'algorithme 4 qui construit l'ensemble *topknodes[index]*. Les $k - top$ sont choisis dans l'algorithme de *SPIN* en se basant sur la valeur du Shapley [45]. Dans cette approche, les auteurs voient le problème comme un jeu coopératif et ils choisissent des jetons gagnants.

Dans tous ces travaux, l'approximation n'est pas garantie. En 2011, Goyal Amit [46] proposent *SIMPATH*, un algorithme efficace pour la détection des semences dans le problème de maximisation de l'influence, basée sur le modèle LT. Cette approche prend un ensemble de sommets et calcule la somme des propagations de chaque sommet. A. Goyal *et al.* en proposant Simple Path (SIMPATH), ils utilisent une approche de sélection dynamique et une approche gloutonne pour garantir l'approximation. Ils utilisent l'optimisation de *CELF* proposée en 2007 par Leskovec Jure *et al.* [47] dont ils (Goyal Amit *et al.*) ont proposé une amélioration en 2011 [48]. Ils déterminent les semences de la même manière que *lazy forward* dans *CELF* qu'on developpera dans le prochaine paragraphe. Dans cette même extraction de sous-graphe, en 2012 Wang Chi *et al.* [49] proposent Maximum influence paths (MPI) basé sur le modèle de propagation IC. L'idée principale de cette heuristique est d'utiliser les structures d'arborescence de chaque sommet pour une approximation de la propagation de l'influence. D'abord, ils déterminent le plus court chemin entre deux sommets en utilisant l'algorithme de *Dijkstra*[3]. Ce chemin est appelé la région d'influence pour chaque sommet y appartenant. Ils ignorent les MPI qui ont

3. Il a proposé un algorithme qui donne tous les plus courts chemins, si c'est possible, entre un sommet u et les autres du graphe. Son algorithme prend en paramètre un graphe dont le poids de chaque arêtes est positif et un sommet u de départ. Sa complexité est $O(n^2)$

$S_i(\pi_j)$ représente l'ensemble des sommets après la permutation π_j
$MC[i]$ représente la contribution marginale de i
$\phi(i)$ représente la valeur Shapley du sommet i
Soit π_i le $j - ime$ permutation de Ω
Soit R le nombre de répétitions

pour $i = 1$ *à* n **faire**
  $MC[i] \longleftarrow 0$;
fin
pour $j = 1$ *à* j **faire**
  pour $i = 1$ *à* n **faire**
    $temp[i] \longleftarrow 0$;
  fin
  pour $r = 1$ *à* R **faire**
    Assigner des seuils aléatoires aux sommets;
    pour $i = 1$ *à* n **faire**
      $temp[i] \longleftarrow temp[i] + v(S_i(\pi_i \bigcup \{i\}) - v(S_i(\pi_i))$;
    fin
  fin
  pour $i = 1$ *à* n **faire**
    $MC[i] \longleftarrow temp[i]/R$;
  fin
fin
pour $i = 1$ *à* n **faire**
  Calculer $\phi(i) = MC[i]/t$;
fin

Algorithme 3 : *Construction de RankList (SPIN)*

```
1  Initialisation index ⟵ 1 et status[1..n] à 0;
2  pour j = 1 à n faire
3  |   flag ⟵ 0;
4  |   pour i=0 à index faire
5  |   |   si topknodes[i] = RankList[j] Ou topknodes[i] est adjacent à RankList[j]
   |   |   alors
6  |   |   |   flag ⟵ 1;
7  |   |   |   break;
8  |   |   fin
9  |   fin
10 |   si flag=0 alors
11 |   |   si index=k alors
12 |   |   |   goto 28
13 |   |   fin
14 |   |   topknodes[index] ⟵ RankList[j];
15 |   |   status[j] ⟵ 1;
16 |   |   index ⟵ index + 1;
17 |   fin
18 fin
19 pour j=0 à n faire
20 |   si status[j] ≠ 1 alors
21 |   |   si index = k alors
22 |   |   |   goto 28
23 |   |   fin
24 |   |   topknodes[index] ⟵ RankList[j];
25 |   |   index ⟵ index + 1;
26 |   fin
27 fin
28 Déclarer les sommets qui sont dans topknodes[] comme les k-top
```

Algorithme 4 : *Choix des $k-Top$ sommets (SPIN)*

une probabilité plus petite qu'un seuil d'influence θ défini. Ensuite, la réunion des forme l'arborescence (). Pour chaque chemin $P =< u = p_1, p_2, ..., p_m = v >$, ils définissent la probabilité de propagation $pp(P)$ donnée par l'équation 3.2.

$$pp(P) = \Pi_{i=1}^{m-1} pp(p_i, p_{i+1}) \tag{3.2}$$

Intuitivement, la probabilité pour que le sommet u active le sommet v dans le chemin P est $pp(P)$. Cette dernière est la probabilité d'activer tous les sommets appartenant à P. Si cette probabilité est supérieure au seuil d'influence alors ce chemin sera dans l'arborescence sinon il sera ignoré. Ils proposent d'utiliser pour estimer l'influence d'un sommet à un autre. Soit $\Omega(G, u, v)$ tous les chemins du sommet u vers le sommet v. Le *MIP* entre u et v est donné par l'équation 3.3

$$MIP_G(u, v) = argmax_p\{pp(P)|P \in \Omega(G, u, v) \tag{3.3}$$

3.1.3 Sélection gloutonne

Sur cette même lancée, nous avons plusieurs heuristiques qui sélectionnent dynamiquement les semences en utlisant uniquement un algorithme glouton. Cette approche est proposée par Kempe David *et al.* [11] en 2003. Son approche nécessite un modèle de propagation ψ connu. Elle est développée dans le paragraphe 2.4. Les approches algorithmiques et de centralités sont plus rapides que les approches gloutonnes qui n'ont pas une bonne complexité. Les approches utilisant l'algorithme glouton sont très gourmandes en terme de complexité. C'est comme si on demandait un enfant de choisir *k-utilisateurs* les plus influents, il les choisit un à un. Mais les approches gloutonnes garantissent l'approximation. Toujours dans le but de diminuer la complexité, en 2007 Leskovec Jure *et al.* [47] proposent un algorithme appelé Cost-Effective Lazy Forward (CELF). Ce dernier joue sur les propriétés de la sous-modularité de la fonction d'influence $\sigma()$ pour diminuer la complexité temporelle de Greedy hill climbing

proposé par Kempe David *et al.* [11]. L'algorithme $CELF$ peut réduire le nombre de simulations *Monte Carlo* à 700 unité de temps appelé par greedy simple. Le principe est que le gain marginal d'un sommet dans l'itération actuelle ne peut pas être plus que celui dans les itérations précédentes, donc le nombre d'appels d'estimation de propagation peut être de taille très grande.

La simulation *Monte Carlo* est une technique mathématique informatisée qui permet de tenir compte du risque dans l'analyse quantitative et la prise de décision. La simulation *Monte Carlo* procède de l'analyse du risque par élaboration de modèles de résultats possibles, en substituant une plage de valeurs, une distribution probabiliste, à tout facteur porteur d'incertitude. Elle calcule et recalcule ensuite ces résultats selon, à chaque fois, un ensemble distinct de valeurs aléatoires des fonctions de probabilités. Suivant le nombre d'incertitudes et les plages spécifiées pour les représenter, une simulation $Monte\ Carlo$ peut impliquer, pour être complétée, des milliers ou même des dizaines de milliers de calculs et recalculs. La simulation produit des distributions de valeurs d'issues possibles.

En 2011 Goyal Ami *et al.* [48] proposent l'algorithme *Cost-Effective Lazy Forward* (CELF++). Ils optimisent d'avantage l'algorithme $CELF$ en exploitant la sous-modulairité de la fonction σ et ils montrent qu'il est plus rapide que $CELF$ [47] de 30% à 55%. Toujours dans le but de réduire la complexité de la fonction d'influence σ, en 2013 Zhou Chuan *et al.* [50], en 2014, proposent une approche nommée Upper Bound basé sur Lazy Forward (UBLF) qui se base sur l'algorithme glouton sous le modèle de propagation LT.

Dans une itération, un sommet ne peut avoir un gain marginal plus petit que dans l'itération précédente. Contrairement à *CELF* et *CELF++*, *UBLF* utilise la limite supérieure *(**U**pper **B**ound)* apparaissant dans le théorème 3.1 pour classer tous les sommets dans l'étape d'initialisation, qui finalement réduit le nombre total d'estimations de la propagation.

Théorème 3.1 *La borne supérieure de propagation $\sigma_L(S)$ est :*

$$\sigma_L(S) = \sum_{t=0}^{N-S} \Pi_0^S.W^t.1 \tag{3.4}$$

où, $W = w_{ij}$ est le poids de ij dans la matrice.

3.2 Synthèse de l'état de l'art

La détection des diffuseurs initiaux dans le problème de maximisation de l'influence est une tâche très difficile. Plusieurs travaux ont été effectués. Dans ce domaine comme, il est presque impossible sous un modèle de propagation ψ donné, d'avoir une combinaison de $k-$*utilisateurs* qui va influencer tous les utilisateurs du réseau social. Dans la sélection statique, chaque utilisateur a une mesure qui représente son degré d'influence. Dans les mesures de centralité proposées, certaines s'adressent à la portée de la propagation de l'influence d'un sommet dans une seule direction comme $k-$*path edge*, d'autres au nombre de voisinages comme centralité de degré dans les RSM et centralité de *multi degré* dans les RSMC. Dans ces réseaux, les mêmes centralités sont utilisées en prenant en compte qu'un utilisateur peut avoir des représentants dans les différentes couches. Or l'information à propager peut avoir une portée très grande dans une direction et pas dans une autre direction. Dans la selection dynamique, l'atout principal est le fait de tenir en compte les semences déjà sélectionnées en modifiant la mesure des autres utilisateurs. En d'autres mots, si une semence est sélectionnée, la centralité de chacun de ses voisin est modifiée. Dans les différentes approches, seules celles utilisant l'algorithme glouton garantissent une bonne approximation mais elles sont difficiles de mettre en œuvre avec les fonctions sous-modulaires. Donc cela nécessite une bonne compréhension des mathématiques. Aussi, les approches gloutonnes sont très gourmandes en complexité. Or dans les résolutions des problèmes informatiques, le temps est un paramètre très important.

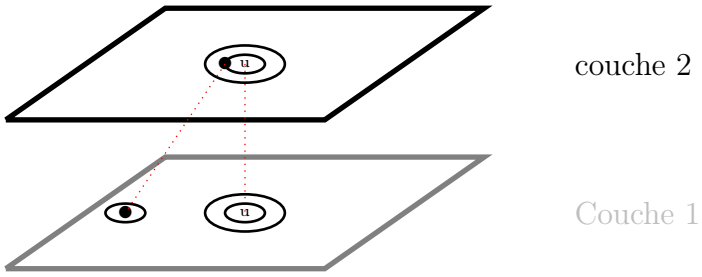

FIGURE 3.1: *Expansion à partir de u dans les réseaux sociaux multicouches*

3.3 Contributions

Nos contributions peuvent être regroupées en deux parties. D'abord, nous intéressons à l'information qui se propage dans toutes les directions. Un utilisateur émet une information, on évalue son influence sur ses voisins N^1, ensuite sur les voisins N^2 jusqu'au voisins N^ℓ. C'est comme dans une expansion, plus les voisins qui sont dans le rayon de l'expansion sont influents plus l'individu v est important dans le réseau social. Dans la figure 3.1, le sommet u diffuse une information, pour le niveau de voisinage fixé à 9, son importance est mesuré selon le nombre d'individus susceptibles d'être influencé. Pour considérer cette expansion dense, nous avons proposé une centralité dans [51], qui prend en compte les voisins $N(\ell, v)$ et pour chacun, son degré d'influence vers ses voisins N^1. Nous notons que cette mesure de centralité se base sous le modèle IC et nous l'appelons *centralité de degré de diffusion ℓ-ième* et nous la notons par C_{dd}^{ℓ}. Cependant, cette mesure n'est pas efficace dans les RSMC. Nous avons redéfini la mesure de centralité C_{dd}^{ℓ} dans [8]. Nous avons utilisé le même principe que C_{dd}^{ℓ} mais au lieu de considérer cette expansion dans une seule couche, nous la considérons dans toutes les couches. Elle sera amplifiée d'une couche à une autre via les sommets qui ont plusieurs représentants comme le montre la figure 3.2. Tous les représentants d'un individus sont déterminés en considérant leur classe d'équivalence [8]. Nous notons cette mesure par C_{dd}^{MLN} et nous l'appelons centralité de *degré de multi-diffusion*. Enfin, après avoir étudié les deux familles de modèles de

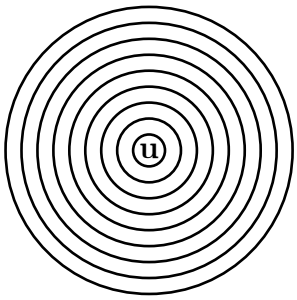

FIGURE 3.2: *Expansion à partir de u dans les réseaux sociaux monoplex*

propagation de base, à savoir les modèles cascade indépendant (IC) et les modèles seuil linéaire (LT). Dans ces modèles, à une itération donnée, on traite seulement les utilisateurs non influencés. On parle de rétroaction le fait de traiter un utilisateur déjà influencé. Nous avons proposé une approche qui fait un pré-traitement pour empêcher les rétroactions vers les utilisateurs semences. Dans un premier temps, nous avons proposé un algorithme d'extraction, appelé SGC, qui prend un graphe non orienté connexe. Cet algorithme est en deux versions, SGC_{v1} qui construit les descendants aléatoirement, proposé dans [52], et SGC_{v2} qui les construit en se basant sur le nombre de leurs voisins dans le graphe initial, proposé dans [53]. Ensuite, nous donnons une généralisation de chacun des deux versions. Elle prend tous les types de graphes et elle est appelée algorithme SG. Après l'extraction du graphe couvrant de maximisation, nous utilisons des heuristiques existantes pour déterminer les semences dans le graphe partiel. Nous avons montré que ce pré-traitement améliore les heuristiques existantes.

Dans cette section, nous avons présenté les approches proposées dans la détection des diffuseurs initiaux en se basant sur des paramètres qu'on juge importants. Dans la suite, nous les détaillerons.

3.4 Heuristique degré de diffusion $\ell-$ième

Ici, nous proposons dans [51], une heuristique nommée Centralité degré de diffusion ℓ-ième et elle est notée par C_{dd}^{ℓ}. Cette heuristique est basée sur le modèle de propagation Cascade Indépendant et utilise la contribution des voisins de niveaux ℓ et de leur probabilité de diffusion. D'abord, nous parlerons de niveaux de voisinage. Ensuite, nous donnons les limites des heuristiques de référence. Enfin, nous proposons une heuristique qui prend en compte ces manquements en proposant un algorithme.

3.4.1 Niveau d'un sommet par rapport à un autre

Dans un graphe, un sommet peut avoir des voisins directs, alors on peut parler de voisinage N^1. Dans les réseaux sociaux, on parle tout simplement d'amis. Cette appellation peut être trompeuse car la nature des relations peut être par exemple familiale, professionnelle, ... La mesure de centralité de degré, proposée par Kempe David *et al* dans [11], se base sur la notion de voisinage de niveau 1. Cette heuristique considère plus qu'une personne a des voisins plus qu'elle est importante dans la diffusion de l'influence. Nous considérons à titre d'illustration le graphe social de la figure 3.3. Il est composé de 11 utilisateurs reliés par des relations de même nature. Ici, le sommet numéro 3 entouré en rouge représentant un utilisateur avec trois voisins N^1 entourés en vert. Prenons v un sommet d'un graphe, nous notons par N^1(v) l'ensemble des voisins de niveau 1 du sommet v. Dans le graphe de la figure 3.3, nous avons $N^1(3)$=$\{2,5,4\}$. Si nous prenons le graphe de la figure 3.4, précédemment, nous avons donné les voisins de niveau 1. L'ensemble N^2(v) est considéré de tous les voisins de chaque sommet de N^1(v) qui ne sont pas bien sûr dans l'ensemble N^1(v). Comme illustration, nous avons en bleu, dans la figure 3.4, tous les voisins de niveaux 2 du sommet 3. Nous avons l'ensemble $N^2(3)$ =$\{1, 7, 6, 8\}$.

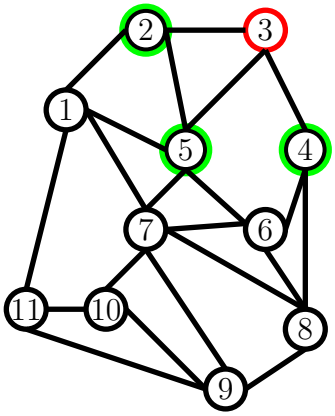

FIGURE 3.3: *Voisinage de niveau 1 du sommet 3*

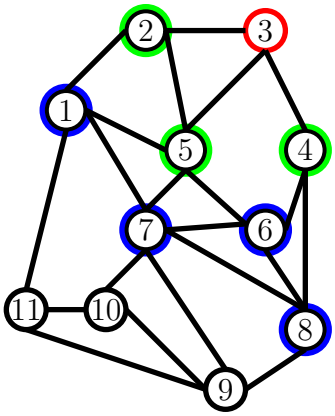

FIGURE 3.4: *Voisinage de niveau 2 du sommet 3*

Suivant ce même principe, nous pouvons avoir tous les voisins $N^{\ell}(v)$. L'ensemble des voisins de niveau 1 jusqu'au niveau ℓ est noté par $N(\ell, v)$ et est défini par la réunion de tous les voisins de niveau 1 à ℓ. Il est donné par l'équation 3.5.

$$N(\ell, v) = N^1(v) \bigcup N^2(v) \bigcup \cdots \bigcup N^{\ell}(v) \tag{3.5}$$

3.4.2 Approche du degré de diffusion ℓ-ième

L'information à diffuser peut être de nature diverse. Les individus peuvent être sensibles à une information α et non à une information β. Donc, diffuser une information peut dépendre de la pression sociale (l'influence) et aussi de la nature de l'information à diffuser. Dans notre modèle, pour prendre en compte la nature de l'information et de la pression sociale, nous intégrons la probabilité qui représente le taux d'influence d'un utilisaeur vers ses voisins N^1. Dans le modèle IC, un utilisateur nouvellement influencé, va essayer, à son tour, d'influencer tous ses voisins N^1. Alors, si un utilisateur détient l'information, il est important que ce dernier ait plusieurs voisins influents. Jusqu'ici, les heuristiques existantes ne prennent pas en compte des voisins $N(\ell, v)$ et de leur degré d'influence. Dans notre heuristique, nous allons determiner le nombre de voisins N^1 jusqu'à N^ℓ et pour chacun d'eux, nous utilisons son taux d'influence.

3.4.3 La centralité de degré de diffusion ℓ-ième

3.4.3.1 Modèle mathématique

La centralité d'un sommet est un concept fondamental dans l'analyse des réseaux sociaux (ARS). Elle donne une indication sur comment un sommet est connecté dans le réseau social. Les mesures de centralité proposées peuvent être efficaces dans une application et non dans une autre. Dans la maximisation de l'influence, nous proposons une heuristique qui se base sur la centralité qui donne le degré d'influence d'un sommet dans un réseau social. Nous la nommons *degré de diffusion $\ell-$ième* et elle est notée par C_{dd}^{ℓ}. Dans cette centralité, nous utilisons celle proposée par Freeman Linton *et al.* dans [54] appelée centralité de degré et elle est notée par C_d.

Elle est définie par l'équation 3.6.

$$C_d(v) = \sum_{(u \in V - v)} \sigma(v,u),\ ou\ \sigma(v,u) = 1\ si\ \{u,v\} \in E\ et\ 0\ sinon \tag{3.6}$$

Dans la centralité C_{dd}^{ℓ}, calculer la centralité d'un sommet v revient d'abord à calculer sa propre contribution dans la diffusion de l'influence. Ensuite celle de ses voisins N^1 jusqu'aux voisins N^{ℓ}

↬ **La contribution de l'utilisateur** v

Comme nous l'avions dit précédemment, ce ne sont pas toutes les informations qu'un utilisateur, dans un réseau social, accepte de diffuser et les degrés d'influence des utilisateurs peuvent être différents. Et s'il accepte la diffusion, combien de voisins de niveau 1 va-t-il influencer. Un utilisateur peut être influencé comme il peut ne pas être influencé, alors nous intégrons une probabilité de diffusion dans l'heuristique définie par Freeman Linton [54]. Dans l'equation 3.7, nous redéfinissons alors la centralité de degré (C_d) en intégrant la probabilité d'influence de l'utilisateur sur ses voisins.

$$C'_{dd}(v) = \lambda_v * C_d(v) \tag{3.7}$$

Dans cette centralité, en plus du nombre de voisins N^1, nous prenons en compte la probabilité λ de chaque sommet qui donne son taux d'influence. Alors, ce n'est pas seulement le nombre de voisins qui fait qu'un sommet influent. Dans la figure 3.5, nous avons un graphe social dont les utilisateurs sont numérotés de 1 à 6. Si on utilise l'algorithme de Freeman Linton C [54] qui prend les sommets qui ont plus de voisins comme les plus influents. Ces sommets seront donc 2 et 5. Mais, ils peuvent ne pas être influents sur leurs voisins N^1. Alors, leur apport dans la diffusion peut être plus petit que les autres qui ont un taux d'influence plus élevé. Ils ne seront pas considérés comme les plus influents dans l'heuristique centralité de degré de diffusion ℓ-ième. Dans cette dernière, ce n'est pas le nombre de voisins qui est important mais

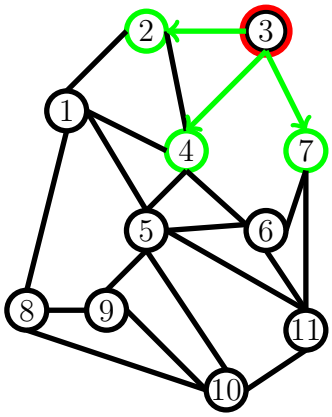

FIGURE 3.6: *Contribution du sommet* 3

le degré d'influence. Une fois que le sommet a une bonne probabilité de diffusion, l'heuristique cherche alors combien de voisins le sommet va essayer d'influencer, si ces sommets voisins vont-ils diffuser l'information en tenant compte de leur taux d'influence. Cette heuristique se focalise sur l'apport du sommet v et de ces voisins N^1 jusqu'aux voisins N^ℓ. Dans la figure 3.6, on veut calculer $C^\ell_{dd}(3)$. On va évaluer son apport dans la diffusion. Il a comme N^1 les sommets 2, 4 et 7. On va évaluer la probabilité d'activation de ses trois sommets. On parle d'apport du sommet lui même.

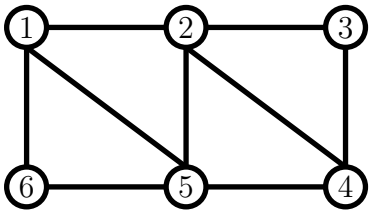

FIGURE 3.5: *Nombre de voisins vs plus influent*

Après avoir donné l'apport du sommet lui même, la centralité C^ℓ_{dd} prend aussi en compte l'apport donné par ses voisins de $N(\ell, v)$.

↬ **La contribution des** $N(\ell, v)$

Dans l'équation 3.7, nous avons la contribution de sommet v dans la diffusion de l'influence. Dans la suite, nous évaluons l'apport de ses voisins de N^1 jusqu'aux voisins N^ℓ. Si v influence ses voisins N^1, ces derniers à leur tour vont tenter d'influencer eux aussi leurs voisins de N^1 (qui sont des voisins N^2 du sommet de départ v). Dans la centralité C_{dd}^ℓ, l'apport des voisins de N^1 aussi est important. Dans C_{dd}^ℓ, un sommet est influent si ses voisins sont influents. Le processus de voisinage continue jusqu'aux voisins N^ℓ. Donc, l'importance d'un sommet est donnée par la portée qu'il donne à l'influence. Comme le montre la figure 3.1, plus les voisins qui sont dans le rayon de l'expansion sont influents plus le sommet v est important. Nous avons dans l'équation 3.8 la contribution de $N^1(v)$. Elle sera généralisée dans l'equation 3.9 pour tous les voisins de niveau 1 de v. Ainsi nous définissons la contribution des voisins de niveau 1 dans le processus de diffusion de l'influence. Nous avons dans l'équation 3.8 la contribution d'un voisin de niveau 1 de v. Elle sera généralisée dans l'equation 3.9 pour tous les voisins de niveau 1 de v. Ainsi nous définissons la contribution des voisins de niveau 1 dans le processus de propagation. Dans l'équation 3.9, l'ensemble des voisins de niveau 1 de v est défini soit par N soit par N^1.

$$\lambda_u * C_d(u) \tag{3.8}$$

où $u \in V$ et $u \in N(v)$.

$$\sum_{u \in V, u \in N^1(v)} \lambda_u * C_d(u) \tag{3.9}$$

où N^1(v) est l'ensemble de voisins de niveau 1 du sommet v.

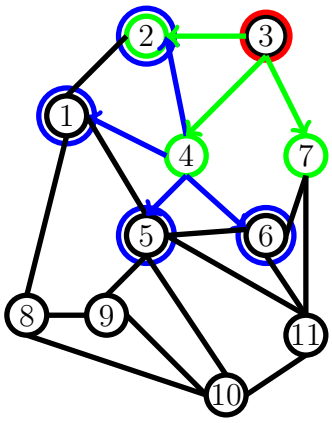

FIGURE 3.7: *Contribution d'un voisin de niveau 1*

Dans la figure 3.7, pour évaluer la centralité degré de diffusion *ℓ-ième* du sommet 3 ($C_{dd}^{\ell}(3)$), après avoir évalué sa propre contribution dans la figure 3.6, nous évaluons les contributions des voisins de niveau 1. Le sommet 4 qui est un élément de $N^1(3)$, va essayer d'influencer les sommets 2, 1, 5 et 6. On le fait pour les sommets de $N^1(3)$. Ainsi nous avons la contribution des sommets de $N^1(3)$ dans la diffusion de l'information.

Une fois que les voisins N^1 sont évalués, le processus continue vers les voisins de N^2. Ils vont participer aussi à la diffusion. En se basant sur l'équation 3.9, nous déterminons la contribution des voisins de niveau 2 (la contribution des amis de mes amis) dans l'equation 3.10. Cette équation est tout simplement la sommation des contributions de tous les sommets de niveau 2 par rapport à v.

$$\sum_{u \in V, u \in N^2(v)} \lambda_u * C_d(u) \tag{3.10}$$

où N^2(v) est l'ensemble de voisins de niveau 2 du sommet v.

Le processus de contribution continue aux voisins N^3, N^4, jusqu'aux voisins N^{ℓ} initié par le sommet de début. Dans l'équation 3.11, nous donnons la contribution

des sommets de niveau ℓ par rapport à v dans la diffusion.

$$\sum_{u \in V, u \in N^{\ell}(v)} \lambda_u * C_d(u) \tag{3.11}$$

où N^{ℓ}(v) est l'ensemble des voisins de niveau ℓ.

La contribution de tous les ℓ^{ime} voisins de v (niveau 1 jusqu'au niveau ℓ) sera la somme des contributions de tous les niveaux. Nous définissons alors , l'ensemble de sommets de niveau 1 jusqu'au niveau ℓ. Elle (la contribution de tous les ℓ^{ime} voisins sera définie dans l'équation 3.12

$$\sum_{u \in V, u \in N(v,\ell)} \lambda_u * C_d(u) \tag{3.12}$$

Où $N(\ell, v)$(v) est l'ensemble de tous les voisins de niveau entre 1 et ℓ.

Une fois la contribution du sommet v et ses voisins $N(\ell, v)$, qui est définie dans l'équation 3.5, dans la diffusion de l'influence, la centralité degré de diffusion ℓ-ième sera définie par leur somme. Dans l'équation 3.13, nous avons une mesure qui est la centralité degré de diffusion ℓ-ième.

$$C_{dd}^{\ell}(v) = \lambda_v * C_d(v) + \sum_{u \in V, u \in N(v,\ell)} \lambda_u * C_d(u) \tag{3.13}$$

En guise d'application, nous pouvons utiliser cette mesure de centralité comme une heuristique dans la maximisation de la diffusion de l'influence. Elle consiste à calculer pour chaque sommet du réseau social sa centralité de degré de diffusion ℓ-ième et les k-top i.e. les k utilisateurs qui ont la plus grande mesure, sont considérés comme les semences que l'on notera S_k^*.

3.4.3.2 Modèle algorithmique

Dans cette section, nous proposons un modèle algorithmique pour calculer la C_{dd}^{ℓ} donnée par 3.13. L'algorithme prend en entrée un graphe représentant un réseau social, un sommet représentant un utilisateur et les probabilités qui sont les taux d'influence. Il donne en sortie la mesure de centralité degré de diffusion ℓ qui sera utilisée comme une heuristique dans la maximisation de l'influence. Cet algorithme

```
1  Les variables et sous programmes utilisés dans algorithme 5
2  currentLevel : le niveau courant qui va de 0 jusqu'à ℓ
3  Level^v(u) : Niveau da u par rapport à v
4  F^v : la file d'attente de v
5  enfiler(F,v) et défiler(F) : enfilé et défilé la file F
6  λ_v : La probabilité pour que v diffuse l'information
   Données : G (le graphe du réseau social), ℓ, v
   Résultat : C_dd^ℓ(v)
7  Initialisation : F^v ⟵ φ, marquer(v) ; currentLevel ⟵ 0;
8  C_dd^ℓ(v) ⟵ λ_v * C_d(v) ; Level^v(v) ⟵ 0;
9  v1 ⟵ v;
10 tant que (currentLevel ≤ ℓ) faire
11 |  pour each (u ∈ N(v1)) ET (non marquer(u)) faire
12 |  |  Level^v(u) ⟵ Level^v(v1) + 1
13 |  |  enfiler(F^v, u)
14 |  |  marquer(u)
15 |  fin
16 |  si (v1 ≠ v) alors
17 |  |  C_dd^ℓ(v) ⟵ C_dd^ℓ(v) + λ_v1 * C_d(v1);
18 |  fin
19 |  v1 ⟵ défiler(F^v) ;
20 |  currentLevel ⟵ Level^v(v1) ;
21 fin
22 return C_dd^ℓ(v);
```

Algorithme 5 : *Algorithme* C_{dd}^{ℓ}

utilise une file F^v. Un sommet est considéré comme un maillon qui a plusieurs informations. Le niveau de tous les voisins N^1(v) est mis à 1 par rapport à v. Le niveau sera stocké dans la variable $Level^v(u)$. L'idée générale de l'algorithme 5 peut être divisé en deux étapes : calculer l'apport d'un sommet et enfiler tous ces voisins

dans la file. De la ligne 7 à la ligne 9, l'algorithme passe à la phase d'initialisation. C^{ℓ}_{dd}(v) est initialisé par l'apport de sommet v, il sera marqué comme à visiter, son niveau par rapport à lui même sera mis à zéro et il initialise une file à vide. Une variable *currentLevel* est initialisée à zéro. Elle sera utilisée pour connaître le niveau de sommet par rapport à v. De la ligne 10 à la ligne 21, l'algorithme 5 prend le sommet tête de la file, le marque à visiter et il enfile tous ses voisins en mettant leur niveau, niveau courant ou le niveau du sommet courant incrémenté d'une unité. A la ligne 17, l'approt du sommet courant sera ajouté à C^{ℓ}_{dd}(v). A la ligne 19, il défile la file et si le niveau du sommet nouvellement défile est plus grand que le paramètre ℓ l'algorithme retourne la variable C^{ℓ}_{dd}(v). Sinon, une autre itération sera effectuée. De la ligne 11 à la ligne 15, l'algorithme récupère tous les voisins du sommet courant non visités et il les met dans la file F^v, leur marque à visiter et leur niveau au niveau courant incrémenté d'une unité.

Ainsi, une nouvelle mesure de centralité qui prend en compte le nombre de voisins $N(\ell, v)$ et leur probabilité de diffuser l'information, est mise en place. Les $k - top$ sont considérés comme les sommets, dans la maximisation de l'influence, les plus influents (l'ensemble S^*_k). Nous avons proposé un modèle mathématique et un modèle algorithmique. Pour les performances de cette approche, des simulations sont effectuées dans la sections 4.3.1. Des résultats ont montré que les approches basées dans les RSM ne sont pas efficaces dans les RSMC. Dans la suite, nous proposons une heuristique appelée *degré multi-diffusion* et est notée par C^{MLN}_{dd} dans les RSMC qui est une re-définissions de la centralité degré de diffusion ℓ-ième . Cet algorithme est efficace et a une complexité $O(mn)$ où n le nombre d'utilisateurs et m le nombre de liens entre eux.

3.5 Heuristique Degré multi-diffusion (C_{dd}^{MLN})

Les travaux menés par Zhao Dawei *et al.* [39] montrent que les résultats notés dans les RSM ne sont pas efficaces dans les RSMC. Plusieurs travaux faits dans les RSM sont en train d'être adaptés dans les RSMC, bien vrai que la plupart de ces travaux restent théoriques. Dans cette partie, nous proposons une nouvelle heuristique qui prend en compte la probabilité d'influence d'un utilisateur et l'apport de ses voisins dans la propagation de l'influence dans toutes les couches. Dans la suite, nous développerons un modèle mathématique et un un modèle algorithmique. Les travaux menés par Zhao Dawei *et al.* [39] montrent que les résultats notés dans les RSM ne sont pas efficaces dans les RSMC. Plusieurs travaux faits dans les RSM sont en train d'être adaptés dans les RSMC, bien vrai que la plupart de ces travaux restent théoriques. Dans cette partie, nous proposons une nouvelle heuristique qui prend en compte la probabilité d'influence d'un utilisateur et l'apport de ses voisins dans la propagation de l'influence dans toutes les couches. Dans la suite, nous poserons un modèle mathématique et un un modèle algorithmique.

3.5.1 Approche du degré de multi-diffusion

Dans les RSMC, plusieurs travaux effectués dans les RSM sont en train d'être redéfinis dans les RSMC. Par exemple la mesure de centralité degré est redéfinie par Bródka Piotr *et al.* [36] et par Magnani Matteo *et al.* [9]. Dans [55], Berlingerio Michele *et al.* proposent une redéfinition de plusieurs mesures utilisées dans les RSM tels le voisinage, la centralité par proximité, etc. Nous avons proposé une extension de la mesure C_{dd}^{ℓ} pour RSMC qui peut être une agrégation de plusieurs réseaux sociaux monoplex, un réseau social qui a plusieurs natures de liens tels que l'amitié, la famille, professionnel, ... Les travaux existants dans ces RSMC ne prennent pas en compte la probabilité de diffuser l'influence et l'apport de voisins dans la propagation de l'influence. Alors, nous avons proposé une nouvelle heuristique appelée *degré*

de Mutli-diffusion que nous notons C_{dd}^{MLN}. Dans cette heuristique, nous avons considéré le degré d'influence de l'utilisateur en introduisant une probabilité de diffusion dans chaque couche. Cet apport peut être différent d'une couche à une autre. Nous considérons également l'apport des voisins de niveau 1 dans toutes les couches. Dans la figure 3.2, nous avons un RSMC comprenant deux couches. Nous prenons l'utilisateur u comme une semence. Il va propager l'information dans la couche 2. Dans la couche 1, l'information sera propagée via son représentant. Dans la couche 2, on tient en compte de l'apport des voisins de v niveau 1 qui vont à leur tour propager l'information. Mais un voisin v peut avoir des représentants dans d'autres couches. Alors chaque représentant va essayer de diffuser l'information dans dans toutes les couches où il est représenté. Nous proposons une redéfinition de cette heuristique dans les RSMC. Dans cette partie, nous allons proposer premièrement un modèle mathématique en donnant la contribution de l'utilisateur dans la même couche dans le processus de diffusion et sa contribution dans toutes les couches. Deuxièmement, nous allons proposer un modèle algorithmique qui prend en entrée un graphe social multicouche et un sommet v_k^i et fournit en sortie la C_{dd}^{MLN}.

3.5.2 Modèle mathématique

Nous proposons une nouvelle mesure de centralité que l'on peut appliquer dans les RSMC. Les K utilisateurs qui ont la plus grande mesure seront considérés comme les plus influents. Cette mesure de centralité se base d'abord sur les voisins de niveau 1 et 2 dans toutes les couches et de leur taux d'influence. Nous appelons cette mesure de centralité de *degré de multi-diffusion*[4] et nous la notons par C_{dd}^{MLN}. Pour la détermination de cette mesure de centralité, nous définissons $\lambda_{v_k^i}$, la probabilité pour que l'utilisateur v_k^i qui est le $i-$ième utilisateur de la couche k, diffuse de l'influence. Dans les RSM, une heuristique qui donne les semences qui vont maximiser l'influence a été développée dans [51]. Cette heuristique prend la contribution de l'utilisateur

4. Multi-Diffusion Degree centrality

dans toutes les couches et celle de ses voisins de niveaux N^1 et N^2 dans toutes les couches.

3.5.2.1 La contribution de l'utilisateur

Ici, nous évaluons la contribution de l'utilisateur dans sa couche et dans toutes les autres couches. Prenons un sommet v_k^i qui est l'utilisateur i dans la couche k du graphe représentant le RSMC. Nous définissons la probabilité $\lambda_{v_k^i}$ de diffuser l'information dans la couche k de l'utilisateur v_k^i. Nous définissons cette contribution dans l'équation 3.14. L'utilisateur v_k^i peut diffuser l'information à tous ses voisins de la même couche avec la probabilité $\lambda_{v_k^i}$ de vouloir diffuser cette information.

$$\lambda_{v_k^i} * C_d^k(v_k^i) \tag{3.14}$$

où $C_d^k(v_k^i)$ est la centralité de degré (le nombre de voisins) de v_k^i dans la couche k. Mais le sommet v_k^i peut avoir des représentants dans les autres couches. Sa contribution ne se limite pas seulement dans sa couche mais aussi dans la diffusion de l'information dans les autres couches. Dans chaque couche, nous déterminons sa contribution dans cette diffusion. Dans la section 1.1.2, nous avons représenté les RSMC via les graphes. Pour extraire les mêmes utilisateurs dans le réseau social multicouche, nous définissons des matrices de mappage. Pour construire une matrice de mappage entre deux couches différentes, nous avons utilisé une relation d'équivalence $\Re$ afin de determiner les mêmes utilisateurs dans les différentes couches. Dans la contribution de l'utilisateur v_k^i, au lieu de le considérer lui seul, nous allons considérer sa classe d'équivalence $class(v_k^i)$. Pour chaque représentant, nous déterminons sa contribution dans sa couche. La contribution d'un utilisateur est la somme de chacun des représentants dans leur couche. Mais la probabilité de diffuser l'information peut différer d'une couche à l'autre. Alors dans chaque couche, nous utilisons une probabilité différentes $\lambda_{v_k^i}$. Cette probabilité sera appliquée uniquement dans la

couche du représentant. L'équation 3.15 donne la la contribution de l'utilisateur v_k^i dans toutes les couches.

$$\sum_{v_{k'}^{i'} \in class(v_k^i)} \lambda_{v_{k'}^{i'}} * C_d^{k'}(v_{k'}^{i'}) \tag{3.15}$$

Une fois avoir déterminé la contribution de l'utilisateur dans toutes les couches, nous déterminons la contribution des voisins de niveau 1 de chaque représentant dans toutes les couches.

3.5.2.2 La contribution des voisins

Précédemment, nous avons déterminé la contribution de l'utilisateur dans toutes les couches. Il reste à déterminer la contribution de chaque représentant de la classe de l'utilisateur dans toutes les couches. Chaque représentant de $class(v_k^i)$ a des voisins dans la couche où il est défini. Ses voisins aussi peuvent avoir des représentants dans les autres couches. Soit $v_k^{i'}$, un voisin de v_k^i dans la même couche. Dans l'équation 3.16, nous avons la contribution des voisins de l'utilisateur v_k^i dans la même couche qui n'est rien d'autre que la somme des participations de chaque voisin.

$$\sum_{v_k^{i'} \in N^k(v_k^i)} \lambda_{v_k^{i'}} * C_d^k(v_k^{i'}) \tag{3.16}$$

$N^k(v_k^i)$ désigne l'ensemble des voisins de l'utilisateur v_k^i dans la même couche. Toutefois l'utilisateur v_k^i a des représentants dans les autres couches. Au lieu de travailler avec seulement l'utilisateur concerné, nous travaillerons avec sa classe d'équivalence. Pour chaque représentant de $class(v_k^i)$, nous considérons la contribution de ses voisins dans la même couche. La contribution totale sera alors la somme des contributions de tous les voisins de chaque représentant dans $class(v_k^i)$. Cette somme est définie dans l'équation 3.17. Soit $N(class(v_k^i))$ l'ensemble de tous les voisins de v_k^i dans toutes les couches. Pour sa construction, nous prenons chaque représentant

de la classe de v_k^i et on ajoute ses voisins dans sa couche.

$$\sum_{v_{k'}^j \in N(class(v_k^i))} \lambda_{v_{k'}^{i'}} * C_d^{k'}(v_{k'}^{i'}) \tag{3.17}$$

$v_{k'}^j$ qui est un voisin d'un représentant de la classe $class(v_k^i)$ peut avoir lui aussi des représentants dans les autres couches. Or dans l'équation 3.17, les représentants des voisins ne sont pas pris en compte. Dans l'équation 3.18, nous avons la contribution d'un voisin dans toutes les couches. En d'autres mots, nous avons la contribution de chaque représentant d'un voisin $v_{k'}^j$.

$$\sum_{v_{k''}^l \in class(v_{k'}^j)} \lambda_{v_{k''}^l} * C_d^{k''}(v_{k''}^l) \tag{3.18}$$

L'équation 3.17, donne la contribution de chaque voisin d'un représentant de $class(v_k^i)$ dans la même couche que ce dernier. Mais nous savons qu'un voisin peut avoir un représentant dans les autres couches. Dans l'équation 3.18, nous avons la contribution des voisins dans toutes les couches. Pour avoir donc la contribution de tous les voisins dans toutes les couches, nous considérons pour chaque voisin sa classe d'équivalence. Au lieu de considérer $v_{k'}^j$ seulement, nous considérons sa classe d'équivalence. Cette contribution est donnée par l'équation 3.19.

$$\sum_{v_{k'}^j \in N(class(v_k^i))} \left(\sum_{v_{k''}^l \in class(v_{k'}^j)} \lambda_{v_{k''}^l} * C_d^{k''}(v_{k''}^l) \right) \tag{3.19}$$

Mais l'équation 3.19 présente des redondances. Supposons le scénario suivant : en considérant un utilisateur α_1 dans une couche L_1. On veut évaluer sa mesure de centralite degré multi-diffusion. Cet utilisateur a un représentant α_2 dans la couche L_2. Soit aussi un utilisateur β_1 dans la couche L_1 et son représentant β_2 dans la couche L_2. Supposons qu'il y a un lien $\alpha_1\beta_1$ et un lien $\alpha_2\beta_2$ alors l'équation 3.19 va calculer la contribution du voisin β_1 deux fois. Il sera évalué en tant que voisin de α_1

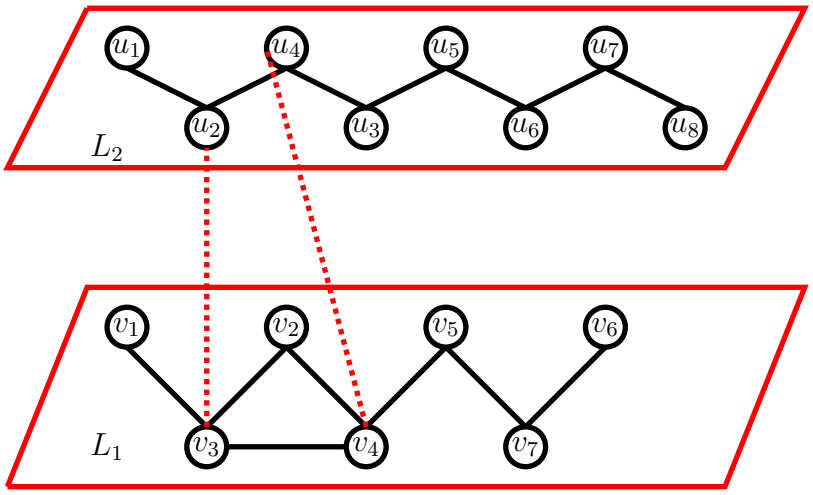

FIGURE 3.8: *Redondances dans l'équation 3.19*

et sa classe d'équivalence sera comptée. Il sera aussi compté comme un voisin de α_2 et la classe d'équivalence de β_2 sera aussi évaluée. Mais comme β_1 et β_2 représentent la même personne alors nous évaluons la même classe deux fois. En guise d'application, nous avons la figure 3.8 qui représente un réseau social de deux couches. Nous avons $class(v3) = \{v3, u2\}$ et $class(v4) = \{v4, u4\}$. $v3$ et $v4$ sont des voisins dans la couche L_1. La contribution de chaque représentant de $class(v4) = \{v4, u4\}$ est calculée. Nous avons aussi $u2$ et $u4$ des voisins dans la couche L_2. $class(u4) = \{v4, u4\}$ qui est le même ensemble que la $class(v4) = \{v4, u4\}$, sera elle aussi évaluée. Donc, nous avons le même ensemble évalué deux fois ce qui donne des redondances dans l'équation 3.19. Pour enlever ces derniers, nous considérons la réunion de toutes les classes d'équivalence afin de les évaluer une seule fois. Dans l'équation 3.20, nous avons la contribution de tous les voisins de chaque représentant de V_k^i dans toutes les couches.

$$\sum_{\substack{v_{k''}^l \in \bigcup class(v_{k'}^j) \\ v_{k'}^j \in N(class(V_k^i)}} \lambda_{v_{k''}^l} * C_d^{k''}(v_{k''}^l) \tag{3.20}$$

3.5.2.3 L'heuristique degré de multi-diffusion

La mesure de centralité de degré multi-diffusion est la somme des contributions de chaque représentant de la classe de l'utilisateur dont on veut calculer la centralité (équation 3.15) et celle des voisins de chaque représentant dans toutes les couches (équation 3.20). Elle est définie dans l'équation 3.21.

$$C_{dd}^{MLN}(v_k^i) = \sum_{v_{k'}^{i'} \in class(v_k^i)} \lambda_{v_{k'}^{i'}} * C_d^{k'}(v_{k'}^{i'}) + \sum_{\substack{v_{k''}^{l} \in \bigcup class(v_{k'}^{j}) \\ v_{k'}^{j} \in N(class(V_k^i)}} \lambda_{v_{k''}^{l}} * C_d^{k''}(v_{k''}^{l}) \qquad (3.21)$$

Dans cette partie, nous avons proposé un modèle mathématique pour déterminer la centralité degré multi-diffusion. Dans la suite, nous proposerons un modèle algorithmique qui s'appuie sur le modèle mathématique.

3.5.3 Modèle Algorithmique

Dans cette partie, nous proposons un modèle algorithmique de l'équation 3.21. Dans cet algorithme, nous utilisons les classes d'équivalence de chaque sommet donné par les matrices de mappagge d'un graphe social multicouche. L'algorithme peut être divisé en trois étapes.

Premièrement, de la ligne 1 à la ligne 4, on détermine l'apport du sommet v_k^i dont on veut calculer la centralité. On évalue la classe du sommet v_k^i et pour chaque sommet de la classe, on fait une sommation partielle. Dans cette sommation, on prend la probabilité du représentant qu'on multiplie par le nombre de voisins dans la couche où ce dernier est représenté.

Deuxièmement, de la ligne 5 à la ligne 13, on calcule la réunion de toutes les classes d'équivalence de chaque voisin d'un représentant de la classe de v_k^i. L'idéé de calculer cette réunion permet d'éviter les redondances notées précédemment. Si une classe est déjà dans l'ensemble des classe de chaque voisin de v_k^i alors elle sera ignorée. C'est

pourquoi dans le modèle mathématique, nous avons utilisé la réunion des classes des voisins. L'idée est de prendre chaque élément de la classe de v_k^i et on calcule sa classe d'équivalence et on le met dans l'ensemble $N(classe(v_k^i))$ s'il n'y est pas.

Troisièmement, de la ligne 14 à la ligne 18, on determine l'apport des voisins dans toutes les couches de chaque élément de $classe(v_k^i)$. L'algorithme a une complexité de $O(mn)$ où n le nombre d'utilisateur du RSMC et m le nombre de liens entre les utilisateurs. Dans cette section, nous avons proposé une heuristique qui donne les

Données : Un graphe multicouche $L = (L_1, L_2, \cdots, L_\eta, MM)$;
v_k^i;
Résultat : $C_{dd}^{MLN}(v_k^i)$;

1 Calculer $classe(v_k^i)$;
2 $C_{dd}^{MLN}(v_k^i) \longleftarrow \phi$;
3 **pour** *tout* $v_{k'}^j \in classe(v_k^i)$ **faire**
4 $\quad C_{dd}^{MLN}(v_k^i) \longleftarrow C_{dd}^{MLN}(v_k^i) + \lambda_{v_{k'}^j} * C_d^{k'}(v_{k'}^j)$;
5 **fin**
6 $N(classe(v_k^i)) \longleftarrow \phi$;
7 **pour** *tout* $v_{k'}^j \in classe(v_k^i)$ **faire**
8 $\quad$ **pour** *tout* $v_{k'}^{j'} \in N(v_{k'}^j)$ **faire**
9 $\quad\quad$ Calculer $classe(v_{k'}^{j'})$;
10 $\quad\quad$ **si** $classe(v_{k'}^{j'}) \not\subset N(classe(v_k^i))$ **alors**
11 $\quad\quad\quad N(classe(v_k^i)) \longleftarrow N(classe(v_k^i)) \bigcup classe(v_{k'}^{j'})$;
12 $\quad\quad$ **fin**
13 $\quad$ **fin**
14 **fin**
15 **pour** *tout* $classe(v_{k''}^l) \subset N(classe(v_k^i))$ **faire**
16 $\quad$ **pour** *tout* $v_{k'''}^h \in classe(v_{k''}^l)$ **faire**
17 $\quad\quad C_{dd}^{MLN}(v_k^i) \longleftarrow C_{dd}^{MLN}(v_k^i) + \lambda_{v_{k'''}^h} * C_d^{k'''}(v_{k'''}^h)$;
18 $\quad$ **fin**
19 **fin**
20 retourner $C_{dd}^{MLN}(v_k^i)$;

Algorithme 6 : *Algorithme* C_{dd}^{MLN}

utilisateurs les plus influents dans un RSMC. Dans cette heuristique, nous avons mis l'importance sur la probabilité de diffuser l'information et sur l'apport des voisins N^1. Nous avons proposé un modèle mathématique suivi d'un modèle algortihmique.

Pour montrer la performance de cette heuristique, des simulations sont effectuées dans la section 4.3.2. Dans la suite, nous développerons une autre approche qui prévient les rétroactions vers les sommets de S_k^* sous les modèles de propagations IC et LT.

3.6 Graphe couvrant de maximisation

Dans la maximisation de l'influence, l'information peut revenir vers un utilisateur déjà influencé sous les modèles de diffusion de base et leurs dérivés. On parle de rétroaction. Dans cette section, nous développons une approche qui extrait un graphe couvrant particulier, qu'on notera par SG, pour éviter les rétroactions. Dans cette extraction, nous utilisons deux mesures de centralité, degré et proximité. D'abord, nous expliquerons les deux mesures de centralité. Ensuite, nous parlerons de la rétroaction vers un sommet sous les modèles IC et LT. Enfin, nous proposerons des versions d'algorithmes d'extraction de SG en fonction de la nature du graphe social d'entrée.

3.6.1 Mesure de centralité par proximité et degré

Ici, nous déterminons un graphe couvrant particulier qui va donner de meilleures semences que le graphe social initial. Le défi est de bien choisir le sommet qui va commencer la construction de ce graphe couvrant. Dans les mesures de centralité, nous avons la centralité par proximité C_c proposée par Freeman Linton C [54] qui donne la position du sommet dans le graphe. Le sommet qui a la plus petite mesure de centralité par proximité est celui qui est plus proche de tous les autres sommets. Si nous commençons la construction à partir de ce sommet, nous aurons un graphe couvrant équilibré par rapport au sommet de début. Pour visiter tous les sommets, nous faisons moins d'itérations. Cette mesure de centralité est donnée par l'équation 3.22 et pour la calculer, nous suivons les étapes suivantes :

– Calculer la plus petite distance (Distance géodésique [5]) entre le sommet u et les autres sommets du graphe.
– Calculer la somme de tous les distances géodésiques.

$$C_c(u) = \sum_{v \in V, v \neq u} d(u, v) \tag{3.22}$$

où $d(u, v)$ représente la distance géodesique entre les sommets u et v. Dans certains de nos algorithmes, les descendants ne sont pas choisis aléatoirement. Nous utilisons le voisinage de chaque sommet dans le graphe initial. Le nombre de voisins est donné par la centralité degré et il est donné par l'équation 3.23

$$C_d(u) = \sum_{v \in V} \lambda(u, v) \tag{3.23}$$

où $\lambda(u, v) = 1$ si $\{u, v\} \in E$ et 0 sinon.

3.6.2 La rétroaction (Information feedback)

Dans les réseaux sociaux, nous avons souvent plusieurs connexions. Les graphes représentant de ces réseaux sont souvent connexes. Les modèles de propagation dans la diffusion de l'information dans les RSM ou RSMC, peuvent être divisés en deux familles : la famille des modèles linéaires avec seuil et la famille des modèles cascades. Dans les modèles de propagation proposées, un utilisateur u peut diffuser une information δ vers ses voisins. Le processus de diffusion continue et à un certain moment, l'information δ peut revenir vers u. Les modèles de diffusion de base vont l'ignorer. Parce que, un utilisateur influencé le reste à jamais. En d'autres mots, si un utilisateurs influencé ne peut pas nier d'avoir diffusé l'information. Donc, les liens qui permettent la rétroaction, ne sont pas importants dans la sélection des

5. geodesic distance

semences. Alors, nous avons proposé d'empêcher la rétroaction au moment de leur détermination. Pour cela, nous avons proposé d'extraire un graphe couvrant particulier. Dans sa construction, le choix de l'utilisateur de départ est très important. Avec cette approche, nous suivons la méthodologie suivante, représentée aussi dans la figure 3.9 pour la détermination des semences.

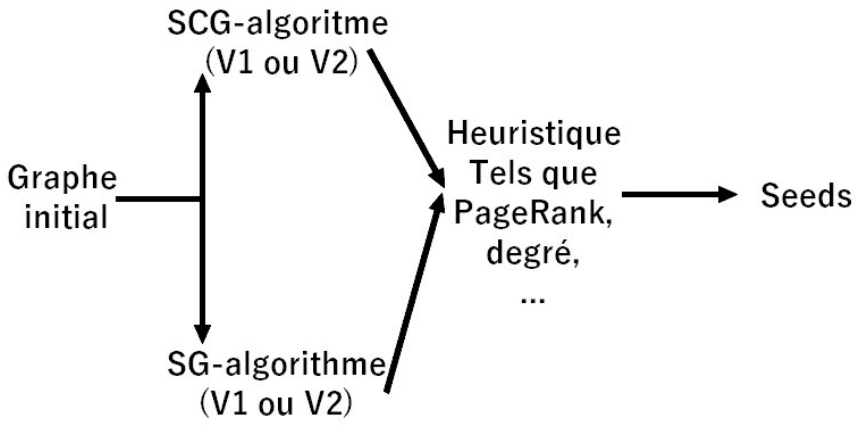

FIGURE 3.9: *La méthodologie de selection de semences en prévenant les rétroactions*

↬ Premièrement, nous allons extraire le graphe couvrant particulier SG (Spanning graph). Pour cette extraction, nous avons proposé deux types algorithmes selon le type du graphe social d'entrée. S'il est connexe à partir du sommet qui au centre du graphe, nous utilisons une des versions des algorithmes SCG. Sinon, nous utilisons une des versions des algorithmes SG comme le montre la figure 3.9.

↬ Deuxièmement, nous utilisons une heuristique telles que C_{dd}^{ℓ}, $k-shell$, ... dans le graphe couvrant particulier donné par un des algorithmes, pour determiner les semences.

3.6.3 Approche du graphe couvrant de maximisation

Dans la maximisation de l'influence dans les réseaux sociaux, le modèle de diffusion est très important. Tous les travaux que nous avons menés se basent sur les modèles IC et LT. Dans la détection des S_k^*, les travaux ne prennent pas en compte la rétro-

action vers les sommets qui pourraient être des semences. Dans la propagation de l'information, un utilisateur qui diffuse et influence un voisin ne peut plus nier avoir fait la diffusion. Alors un utilisateur est soit influencé (actif) et il le reste à jamais, soit non influencé (inactif). Dans les modèles de diffusion de base, à une itération donnée, on essaye d'influencer seulement les utilisateurs qui le ne sont pas encore. Pour prévenir cette rétroaction vers les utilisateurs semences, nous avons proposé un pré-traitement du graphe social initial avant de les détecter. Ce pré-traitement permet d'extraire un graphe couvrant particulier le plus équilibré possible par rapport à l'utilisateur qui est plus proche de tous des autres. Nous avons proposé un algorithme d'extraction [53] qui se basant sur le modèle de propagation IC. Dans cette approche, la construction des voisins d'un utilisateur se faisait aléatoirement. Une amélioration de ce dernier est proposée dans [8]. Dans l'extension, les voisins sont choisis en fonction du nombre de leurs voisins et on a montré par des simulations que la prévention est efficace dans le modèle LT. D'abord, on développera deux versions d'algorithmes d'extraction de graphes couvrants appelées *Spanning Connected Graph* et on note SCG. Ces deux versions prennent en entrée un graphe connexe et donne en sortie un arbre couvrant. Enfin, on donnera une généralisation de la première version d'algorithme SCG appelé GG qui prend en entrée tout les types de graphes sociaux et donne en sortie un arbre couvrant ou une forêt couvrante selon la connexité du graphe social.

3.6.4 Algorithme SCG

Dans cette partie, nous développons deux versions de l'algorithme SCG qui utilisent en entrée un graphe connexe et donnent en sortie un graphe couvrant particulier qu'on appellera *graphe couvrant de maximisation de l'influence*. La première version (algorithme 7) utilise la mesure de centralité par proximité pour déterminer l'utilisateur qui débute la construction du graphe couvrant. Il sera noté par BeginNode. Cet utilisateur est donné par l'équation 3.24. Une fois le premier sommet déterminé,

l'algorithme détermine les voisins N^1, N^2, ..., aléatoirement. Par exemple, supposons la construction de u et v qui sont des voisins N^{i-1}(BeginNode). On suppose aussi que w est un élément de N^i(BeginNode) et dans le graphe social initial, w est en lien avec u et v. Alors, dans le graphe couvrant particulier, w sera soit voisin de u, soit v. L'algorithme 7 choisit un de ces deux utilisateurs (u et v) aléatoirement comme voisin de w dans le graphe couvrant. Ce choix aléatoire sera corrigé dans l'algorithme 8 proposé dans [53]. Le voisin du sommet w sera choisi en fonction du nombre de voisins de u et de v dans le graphe social.

$$BeginNode = \arg \min_{v, v \in V} \quad C_c(v) \tag{3.24}$$

où $C_c(v)$ représente la centralité par proximité de u.

3.6.4.1 Algorithme SCG_{v1}

Nous développons ici la première version de l'algorithme SCG que l'on notera par SCG_{v1}. Le but est de prévenir la rétroaction vers les utilisateurs semences avant leur détermination. Cet algorithme extrait un graphe couvrant particulier (qui est ici, un arbre couvrant) à partir d'un graphe social connexe. Le principe de fonctionnement peut être décrit en deux étapes, le choix du premier sommet (*BeginNode*) et le choix de N^1(*BeginNode*), N^2(*BeginNode*), etc.

↬ Le choix du premier sommet est très important dans la construction du graphe couvrant de maximisation. Dans [54], [56], [37], [57], [58], plusieurs mesures de centralité sont définies. Parmi eux, nous avons la mesure de centralité par proximité qui donne la position du sommet par rapport aux autres sommets. Le sommet qui a la plus petite mesure de centralité par proximité est le sommet central du réseau. Nous le notons par *BeginNode* et il est donné par l'équation 3.24.

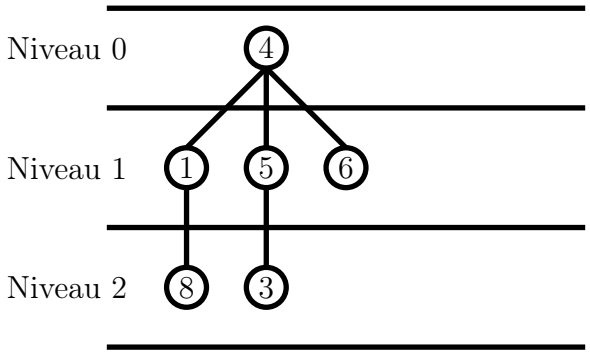

FIGURE 3.10: *Un arbre avec le niveau des sommets à partir de 4*

↬ Après le choix de ce sommet, nous déterminerons ses voisins N^1, N^2, ... Nous prenons zéro comme le niveau du premier sommet (*BeginNode*) de l'arbre couvrant. Dans la figure 3.10, nous savons le sommet 4 comme *BeginNode* et son niveau est zéro. Pour la construction de $N^1(BeginNode)$, nous cherchons tous les voisins de *BeginNode* dans le graphe social de départ qui ne sont encore exploités comme ses voisins dans l'arbre couvrant. Pour les voisins $N^2(BeginNode)$, nous prenons aléatoirement les voisins de chaque utilisateur de $N^1(BeginNode)$ dans le graphe social initial qui ne sont pas encore exploités, comme ses voisins dans le graphe couvrant de maximisation. Toujours dans le même exemple, les sommets, 1, 5 et 6 sont les voisins de niveau 1 ($N^1(BeginNode)$), ils seront construits après le choix du sommet de départ. Ensuite, l'algorithme détermine $N^2(BeginNode)$, $N^3(BeginNode)$, etc. d'une manière aléatoire.

La complexité de cet algorithme est $O(mn)$. Dans le pire des cas, nous avons un graphe social complet. C'est à dire, un réseau social donc tout le monde est en lien avec tout le monde. Alors, *BeginNode* sera en lien avec les $n-1$ utilisateurs du réseau social. L'algorithme fera m itérations et tous les utilisateurs seront exploités. Pour chaque voisins de *BeginNode*, ses $n-1$ voisins seront exploités. Ce qui donne une complexité $O(mn)$ pour la construction du graphe couvrant de maximisation.

L'algorithme 7 calcule la centralité par proximité C_c de tous les sommets aux ligne 6 à 8. A la ligne 9, il détermine le sommet qui débute la construction du graphe

Importantes variables utilisées dans l'algorithme 7
SG : un graphe couvrant
E_{SG} : Ensemble des arêtes de SG
V_{SG} : Ensemble des sommets de SG
$Level^v(u)$: Le niveau du sommet u par rapport au sommet v

Données : Un graphe connexe $G(V, E)$
Résultat : un graphe couvrant $SG(V_{SG}, E_{SG})$

pour *all node* $v \in V$ **faire**
  Calculer $C_c(v)$;
fin
$BeginNode \longleftarrow \arg\min_{v, v \in V} C_c(v)$;
$V_{SG} \longleftarrow \{BeginNode\}, E_{SG} \longleftarrow \phi, level \longleftarrow 0$;
$Level^{BeginNode}(BeginNode) \longleftarrow level$;
tant que $(\mid V_{SG} \mid \neq \mid V \mid)$ **faire**
  pour *all* $u \in V_{SG}$ *et* $Level^{BeginNode}(u) = level$ **faire**
    pour *all sommets* $z \in N(u)$ *et* $z \notin V_{SG}$ **faire**
      $E_{SG} \longleftarrow E_{SG} \bigcup \{u, z\}$;
      $V_{SG} \longleftarrow V_{SG} \bigcup \{z\}$;
      $Level^{BeginNode}(z) \longleftarrow level + 1$;
    fin
  fin
  $level \longleftarrow level + 1$;
fin
retourner SG;

Algorithme 7 : *Algorithme* SCG_{v1}

couvrant de sortie stocké dans la variable BeginNode. A la ligne 13, nous considérons tous les sommets u de même niveau par rapport à BeginNode. De la ligne 14 à la ligne 18, nous considérons tous les voisins de z de chaque u qui ne sont pas encore exploités. Ces voisins vont devenir les descendants de u dans le graphe couvrant. Leur niveau sera mis à $level + 1$, c'est à dire le niveau de u incrémenté d'une unité. Nous supposons le scénario suivant : soient u et v deux sommets de niveau *currentlevel* et w sommet qui n'est encore visité. Supposons aussi que w est un voisin de u et de v dans le graphe initial. L'algorithme 7 choisit aléatoirement entre le sommet u et le sommet v le voisin de w. Cet algorithme à été proposé dans [52]. Dans une extension [53], le choix des voisins n'est plus aléatoire, nous nous basons sur le voisinage de chaque utilisateur. Nous avons montré aussi que le graphe couvrant de maximisation est efficace sous les modèles de diffusion LT. L'algorithme SCG_{v1} a une complexité polynomiale $O(mn)$, où m est le nombre d'arêtes et n le nombre de sommets.

Choix voisins N^{α}(*BeginNode*)

Nous considérons le réseau social représenté par le graphe social de la figure 3.11. L'algorithme 7 le prend comme donnée d'entrée. Il calcule la centralité par proximité C_c de tous les sommets. Il choisit le sommet qui a la plus petite C_c. Dans cet exemple, c'est le sommet 1. La construction du graphe couvrant particulier va débuter à partir de ce sommet. Son niveau sera mis à zéro et il sera ajouté dans l'ensemble des sommets de SG. A la deuxième itération, le niveau courant vaut 0. Les descendants de ce sommet seront construits. Avec ce sommet, nous n'avons pas de problème de choix par ce qu'il est unique. Tous ses voisins dans le graphe initial seront repris et nous aurons l'ensemble *children(1)*={2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 18, 20, 22, 32} comme les sommets de niveau 1. A la deuxième itération, l'algorithme passe à la construction des descendants de l'ensemble *children(1)*. Algorithme 7 commence aléatoirement la construction des voisins de children(1). Dans la figure 3.11, le sommet 34 est voisin des sommets 32 et 9 dans le graphe initial. il sera voisin

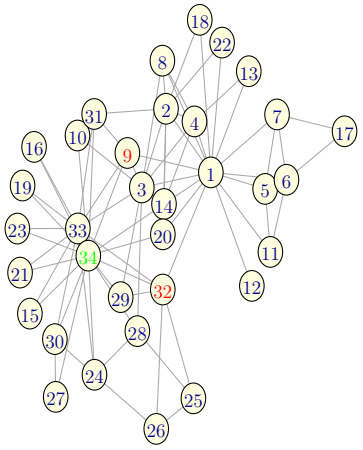

FIGURE 3.11: *Un graphe social connexe*

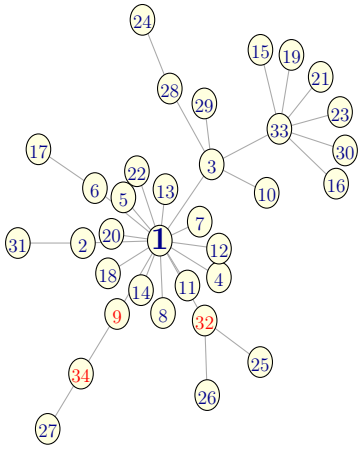

FIGURE 3.12: *Le graphe couvrant donné par l'algorithme 7*

soit du sommet 32 soit du sommet 9 dans le graphe couvrant. Ce choix est aléatoire. Dans la figure 3.12, dans le graphe couvrant donné par l'algorithme 7, on voit que le sommet 9 a été choisi comme voisin du sommet 32.

Processus de construction

Dans ce paragraphe, les figures 3.15, 3.16, 3.17 et 3.18 montrent le processus de construction du graphe couvrant particulier par l'algorithme 7. Dans la figure 3.13**Dolphins** [6], nous avons un graphe social qui représente le réseau social *Dolphins*. Nous le prenons comme graphe d'entrée de l'algorithme 7. Dans la figure 3.14, nous avons la sélection du sommet de départ qui est exactement le sommet 37. La construction du graphe couvrant particulier va débuter à partir de ce sommet et son niveau par rapport à lui est zéro. A la première itération, nous allons construire les voisins $N^1(BeginNode$ qui sont représentés dans la figure 3.15. Le même principe continue pour les voisins $N^2(BeginNode)$ comme le montre la figure 3.16, les voisins $N^3(BeginNode)$ comme on le voit dans la figure 3.17, les voisins $N^4(BeginNode)$,

6. https ://networkdata.ics.uci.edu/data.php ?id=6, visité en 2015

comme dans la figure 3.18.

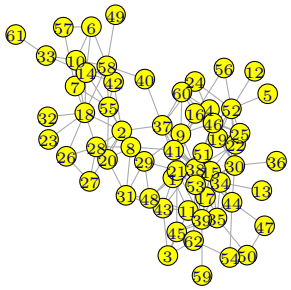

FIGURE 3.13: *Le graphe du réseau Dolphins*

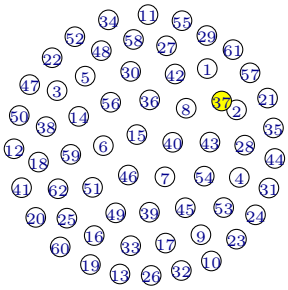

FIGURE 3.14: *Étape 1 de l'algorithme 7*

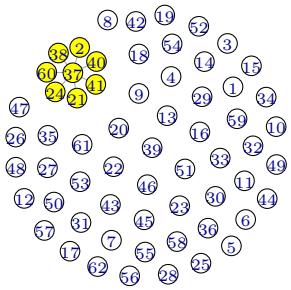

FIGURE 3.15: *Étape 2 de l'algorithme 7*

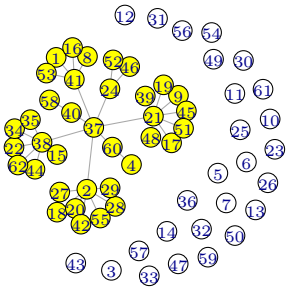

FIGURE 3.16: *Étape 3 de l'algorithme 7*

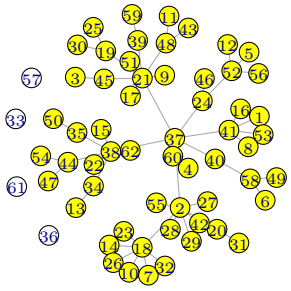

FIGURE 3.17: Étape 4 de l'algorithme 7

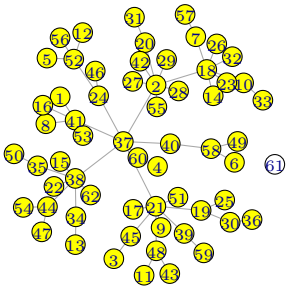

FIGURE 3.18: *Étape 5 de l'algorithme 7*

3.6.4.2 Algorithme SCG_{v2}

Nous développons ici la deuxième version de l'algorithme SCG que l'on notera par SCG_{v1}. Le but est d'empêcher la rétroaction vers les utilisateurs semences avant leur détermination. Cet algorithme extrait un graphe couvrant particulier (qui est ici, un arbre couvrant) à partir d'un graphe social connexe. Le principe de fonctionne est quasiment le même que la première version. Il peut être décrit en deux étapes, le choix du premier sommet (*BeginNode*) et le choix de N^1(*BeginNode*), N^2(*BeginNode*), etc. Le choix du premier sommet se fait de la même manière que celui dans la première version. Après son choix, nous déterminerons ses voisins N^1, N^2, etc. Nous prenons zéro comme le niveau du premier sommet (*BeginNode*) de l'arbre couvrant. Pour la construction de N^1(*BeginNode*), nous cherchons tous les voisins de *BeginNode* dans le graphe social de départ qui ne sont encore exploités comme ses voisins dans l'arbre couvrant. Dans la construction de N^1(*BeginNode*), le choix de voisinage ne pose pas problème parce que nous avons un seul *BeginNode*. Pour les voisins N^2(*BeginNode*), nous allons les construire en fonction de leur nombre de voisins dans le graphe social initial. En d'autres mots, nous commençons la construction des voisins du sommet vérifiant l'équation 3.25. Nous allons trier tous les sommets de N^1(*BeginNode*) en fonction de leur nombre de voisins dans le graphe social initial. Nous construisons d'abord les voisins du sommet qui a le plus de voisins dans le graphe social initial.

La complexité de cet algorithme, pour les mêmes raisons que la première version, est $O(mn)$ si on ne tient pas compte du tri. En tenant en compte de calculer la complexité sera le maximum entre $O(mn)$ et la complexité de l'algorithme de tri utilisé pour l'ordre de construction de voisins de même niveau.

$$u = \arg \max_{v,v \in V \ etLevel^{BeginNode}(v)=level} C_d(v) \tag{3.25}$$

Dans l'algorithme 8, le choix des voisins des utilisateurs d'un niveau donné 'est

Importantes variables utilisées dans l'algorithme 8
SG : Graphe couvrant de sorti
E_{SG} : Ensemble des arêtes du graphe couvrant SG
V_{SG} : Ensemble des sommets du graphe couvrant SG
$Level^v(u)$: Niveau du sommet u par rapport au sommet v
$SortNodeSet^v(1)$: L'ensemble des sommets de niveau α par rapport au sommet v
Données : Un graphe connexe G(V,E)
Résultat : Graphe couvrant SG(V_{SG},E_{SG})

pour *all sommet* $v \in V$ **faire**
  Calculer $C_c(v)$
fin
$BeginNode \longleftarrow \arg\min_{v,v \in V} C_c(v)$;
$V_{SG} \longleftarrow \{BeginNode\}, E_{SG} \longleftarrow \phi, level \longleftarrow 0$;
$Level^{BeginNode}(BeginNode) \longleftarrow level$;
tant que $(\mid V_{SG} \mid \neq \mid V \mid)$ **faire**
  $SortNodeSet^{BeginNode}(level) \longleftarrow$ Tous les sommets de niveau $level$;
  Trier $SortNodeSet^{BeginNode}(level)$ en fonction de leur nombre de voisins dans G;
  pour *all* $u \in SortLevelNodeSet$ **faire**
    pour *all sommets* $z \in N(u)$ *et* $z \notin V_{SG}$ **faire**
      $E_{SG} \longleftarrow E_{SG} \bigcup \{u, z\}$;
      $V_{SG} \longleftarrow V_{SG} \bigcup \{z\}$;
      $Level^{BeginNode}(z) \longleftarrow level + 1$;
    fin
  fin
  $level \longleftarrow level + 1$;
fin
retourner SG ;

Algorithme 8 : *Algorithme* SCG_{v2}

plus aléatoire. Il utilise leur voisinage dans le graphe social initial. L'algorithme 8 prend en entrée un graphe connexe et donne en sortie un graphe couvrant comme dans dans la premier version. Dans cette deuxième version, nous perdons moins d'informations que le graphe couvrant donné par la première version. Par exemple, la mesure de centralité d'un sommet dans le graphe couvrant est presque la même dans le graphe initial, à la différence des arêtes supprimées qui permettent la rétroaction. Le principe ne change pas. De la ligne 7 à 9, l'algorithme 8 calcule la mesure de centralité par proximité de tous les sommets du graphe initial. A la ligne 10, il determine le sommet qui a la plus petite mesure de centralité par proximité (BeginNode) et il débute la construction par ce sommet. Son niveau par rapport à lui est mis à zéro. De la ligne 17 à 21, Pour tout sommet u dont son niveau est le niveau courant, l'algorithme 8 prend tous ses voisins z dans le graphe initial qui ne sont pas encore exploités et les considère comme les descendants du sommet u. L'arête uz sera mis dans l'ensemble E_{SG}, le sommet z dans l'ensemble V_{SG} et le niveau de z est mis à niveau de u incrémenté d'une unité. À la ligne 14, il crée un ensemble appelé $SortNodeSet^{BeginNode}(level)$ qui a tous les sommets de niveau *level* par rapport à BeginNode. A la ligne suivante, il passe au tri de cet ensemble par rapport aux voisins de chaque sommet de l'ensemble dans le graphe G. Le premier sommet de $SortNodeSet^{BeginNode}(level)$ sera le sommet qui a plus de voisins dans le graphe initial, le dernier le sommet qui a le moins de voisin dans le graphe initial. Avec ce tri, il corrige l'inconvéniant noté dans la première version. Soit le scénario suivant, soit un sommet w non exploité, soient deux sommets u et v dans le graphe couvrant et de même niveau. Si le sommet u à plus de voisins alors w sera considéré comme son voisin dans le graphe couvrant sinon, il sera considéré comme le voisin de l'autre sommet. Donc, il construit les voisins du sommet vérifiant l'equation 3.25.

Choix voisins N^{α}(*BeginNode*) pour l'algorithme 8

Soit un réseau social représenté par le graphe connexe de la figure 3.11. L'algorithme 8 donne en sortie le graphe couvrant de la figure 3.19 en prenant comme entrée

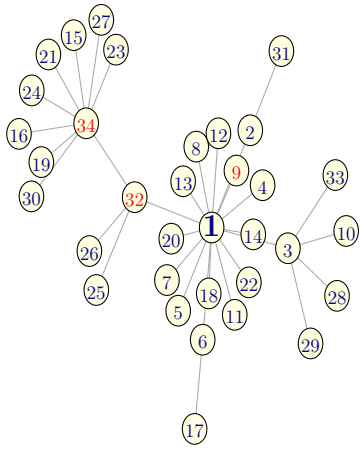

FIGURE 3.19: *Graphe couvrant donné par algorithme 8*

le graphe de la figure 3.11. A la ligne 7 de l'algorithme 8, l'ensemble $SortNodeSet^1(1)$ constitue simplement les descendants du sommet 1. Il donne $children(1)$ ={2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 18, 20, 22, 32}. Cet ensemble sera trié à la ligne 8 et il donne le même ensemble mais ordonné en fonction de leur nombre de voisins dans G. On aura $children(1)'$=[3, 2, 4, 32, 9, 14, 8, 6, 7, 11, 20, 5, 18, 13, 22, 12]. Le sommet 3 a plus de voisins que les autres. D'abord, l'algorithme 8 construit les descendants de 3 et il termine avec les descendants de 12 qui ont moins de voisins. Le sommet 34 est voisin de 9 et de 32 qui sont de même niveau. Le sommet 9 sera considéré comme un descendant du sommet 32 qui a plus de voisins que le sommet 9.

3.6.5 Algorithme *SG*

Dans cette partie, nous proposons une généralisation de l'algorithme 7 et nous l'appellons *SG*. Pour une généralisation de l'algorithme 8, on intègre seulement le tri des sommets de même niveau avant de construire leurs voisins. Sa complexité est $O(mn)$. Cet algorithme peut prendre en entrée tous les types de graphes sociaux (connexe ou non connexe, orienté ou non orienté). La connexité par rapport à un sommet dans les graphe orienté est définie dans le premier chapitre.

L'idée de l'algorithme 9 n'a pas beaucoup de différences par rapport aux deux versions précédentes. On n'a pas de contraintes sur le type de graphe d'entrée. Le graphe couvrant de maximisation sera une forêt couvrante si le graphe d'entrée est orienté et non connexe par rapport au sommet qui a la plus petite mesure de centralité par proximité sinon il sera un arbre couvrant comme le cas des deux premières versions. Dans cet algorithme, si le graphe social d'entrée n'est pas connexe, nous travaillons avec ses composantes connexes. On applique le principe de l'algorithme *SCG* dans chaque composante connexe. On commence la construction par le sommet u qui a la plus petite mesure de centralité par proximité. Du u, nous allons ajouter tous les autres sommets qui lui sont accessibles à partir lui. L'ensemble forme le premier arbre couvrant. S'il reste des sommets qui ne sont pas encore exploité dans le graphe initial G(V,E), on cherche à nouveau le sommet qui a la plus petite mesure de centralité par proximité et qui n'est pas exploité. Un nouvel arbre couvrant débute sa construction à partir de ce sommet. Ainsi, le processus continue jusqu'à exploiter tous les sommets du graphes G(V,E). A la fin du processus, tous ces arbres couvrants constituent notre graphe couvrant de maximisation qui est aussi appelé une forêt couvrante de maximisation.

Importantes variables utilisées dans l'algorithme 9

i Une composante connexe i

$E_{SG(i)}$ Ensemble des arêtes de la i^{ime} composante connexe

$V_{SG(i)}$ Ensemble des sommets de la i^{ime} composante connexe

$Level^v(u)$ Le niveau de u par rapport au sommet v

Données : Un graphe quelconque G(V,E)

Résultat : Une forêt couvrante SG(V_{SG},E_{SG})

pour *all sommets* $v \in V$ **faire**
    calculer $C_c(v)$;
fin
$E_{SG} \longleftarrow \phi$,$V_{SG} \longleftarrow \phi$,i$\longleftarrow 1$;
tant que $(\mid V_{SG} \mid \neq \mid V \mid)$ **faire**
    $BeginNode(i) \longleftarrow \arg \min_{v,v \in V \text{ et } v \notin V_{SG}} C_c(v)$;
    $V_{SG(i)} \longleftarrow \{BeginNode(i)\}, E_{SG(i)} \longleftarrow \phi, level \longleftarrow 0$;
    $Level^{BeginNode(i)}(BeginNode(i)) \longleftarrow level$;
    $CC(i) \longleftarrow true$;
    tant que *CC(i)* **faire**
        $leafNode \longleftarrow$ all sommets de niveau $level$;
        si $leafNode$ *non vide* **alors**
            pour *all* $u \in leafNode$ **faire**
                pour *all sommet* $z \in N(u)$ *et* $z \notin V_{SG}$ **faire**
                    $E_{SG(i)} \longleftarrow E_{SG(i)} \bigcup \{u, z\}$;
                    $V_{SG(i)} \longleftarrow V_{SG(i)} \bigcup \{z\}$;
                    $Level^{BeginNode(i)}(z) \longleftarrow level + 1$;
                fin
            fin
            $level \longleftarrow level + 1$;
        sinon
            $CC(i) \longleftarrow false$;
        fin
    fin
    $V_{SG(i)} \longleftarrow V_{SG} \bigcup V_{SG(i)}$;
    $E_{SG(i)} \longleftarrow E_{SG} \bigcup E_{SG(i)}$;
    $i \longleftarrow i + 1$;
fin
retourner SG;

Algorithme 9 : *Algorithme SG*

De la ligne 7 à la ligne 8, l'algorithme 9 calcule la centralité par proximité de tous les sommets dans le graphe social d'entrée. De la ligne 1 à la ligne 33, il determine la forêt couvrante. A la ligne 11, il determine le sommet qui a la plus petite mesure de centralité par proximité. Ce sommet sera appelé *BeginNode(i)* et le niveau par rapport à lui même sera mis à zéro. Il est le sommet de début de la $i-ième$ composante connexe. Tous les sommets accessibles à partir de *BeginNode(i)* sont dans la composante connexe i. Ensuite, il passe à la construction de la $i+1-ième$ en determinant un autre sommet de début pour une autre composante connexe. A la ligne 16, l'algorithme 9 determine l'ensemble de tous les sommets dont leur niveau est le niveau courant (*level*). Cet ensemble est appelé *Leafnode*. Pour les sommets u dans *Leafnode*, leurs voisins dans le graphe G non visités vont devenir des descendants de u dans le graphe couvrant SG. Si plusieurs sommets de même niveau ont les mêmes voisins non visités dans le graphe G, l'algorithme 9 choisit aléatoirement leurs voisins dans le graphe SG. Pour une généralisation de la deuxième version (algorithme 2), dans une composante connexe donnée, les sommets de même niveau seront triés et la construction de leur voisins débute à partir du sommet qui a le plus de voisins dans le graphe social de début.

Après avoir construit tous les voisins $N^{\alpha}(BeginNode(i))$, on passe au niveau suivant à la ligne 25 de l'agorithme 9. Un autre ensemble de *leafnode* est créé. Si il n'est pas vide, les voisins $N^{\alpha+1}(BeginNode(i))$ sinon il passe à la création de la composante connexe $i+1$ en cherchant le sommet de début BeginNode(i+1). Le processus continue jusqu'à exploiter tous les sommets du graphe de début.

Soit un réseau social unilatéral représenté par le graphe social orienté de la figure 3.20. Nous l'utilisons comme le graphe d'entrée de l'algorithme 9 qui n'a pas de contraintes sur le type du graphe. Le graphe de sortie est une forêt couvrante. Dans l'exemple, on voit des sommets seuls qui forment une composante connexe. Si on regarde la figure 3.20, ces sommets ont souvent un seul lien. Soit le lien est sortant soit il est entrant. Dans la maximisation de l'influence, ces sommets n'ont

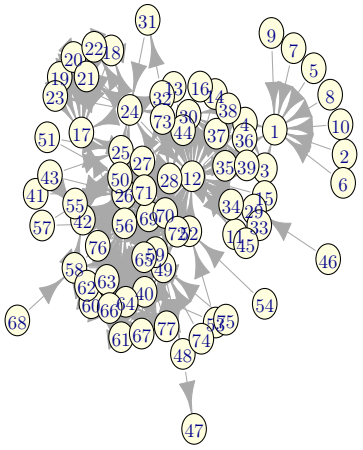

FIGURE 3.20: *Un graphe orienté non connexe par rapport au sommet Begin-Node*

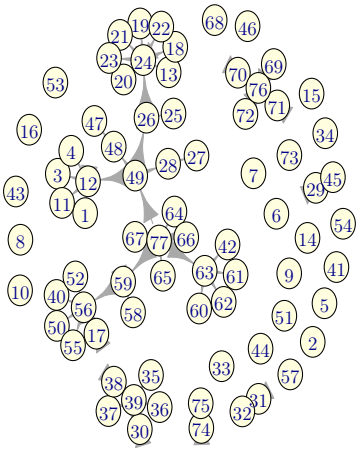

FIGURE 3.21: *Le graphe couvrant donné par algorithm 9*

pas beaucoup d'importance.

Conclusion

Ce chapitre s'articule sur une présentation de quelques heuristiques existantes et sur les contributions effectuées dans la détections des semences dans la maximisation de l'infuence. Nous avons présenter deux mesures de centralités. L'une concerne les réseaux sociaux monoplex et l'autre les réseaux sociaux multicouches. Nous avons aussi proposé une approche qui empêche les rétroactions vers les utilisateurs semences sous les modèles IC et LT et leur dérivés. Dans cette approche, on a proposé trois algorithmes, deux versions de *SCG* pour un graphe social connexe et une généralisation pour la première version. Pour la généralisation de la deuxième version, il suffit de tenir compte de l'ordre de construction des voisins de même niveau. Pour les performances de nos approches, des simulations sont effectuées dans le chapitre suivant.

Chapitre 4

Validation

Dans le chapitre précédent, nous avons présenté des approches pour la détection des semences dans le problème de maximisation de l'influence. Nous avons proposé deux mesures de centralités C_{dd}^{ℓ} et C_{dd}^{MLN} respectivement dans les réseaux sociaux monoplex et multicouches. Nous avons aussi développé une approche qui est d'extraire un graphe couvrant de maximisation avant de détecter les semences à partir des heuristiques existantes. Dans ce chapitre, nous montrerons par des simulations que nos approches donnent de meilleurs semences que les références. Il peut être divisé en deux parties. D'abord, nous présenterons l'outil de simulation. Enfin, nous montrerons par des simulations que nos approches sont plus performantes que les existantes en expliquant les jeux de données utilisés et en calculant la propagation de l'influence de nos modèles et ceux des références les plus connues

4.1 Outils

4.1.1 Quelques outils

Dans l'analyse des réseaux sociaux (ARS), il existe plusieurs outils de simulation. On peut citer des logiciels open-source ou non, commerciaux ou non et même dans

certains langages de programmation, tels que *Python*, le langage *C*, etc on peut également citer soit des classes pour les langages tels que *python*, soit des bibliothèques tels que le langage. Dans ces dernières, plusieurs fonctions de manipulation et de visualisation des graphes ont été implémentées. Dans cette section, nous allons voir quelques outils et leurs caractéristiques. Dans [59] vous pouvez trouver d'amples informations sur les logiciels d'analyse et de visualisation des réseaux sociaux.

↬ **Commetrix**

Commetrix est un outil d'analyse et de visualisation des réseaux sociaux dynamiques. Il prend comme données d'entrée des fichiers qui sont crées par lui même. Dans ce logiciel, on peut exporter les tables d'extension $^*.CSV$. Il est développé en Java et il est supporté par plusieurs systèmes d'exploitation. Il est commercial mais il existe une version d'essai pour quelques jours.

↬ **CoSBiLab Graph**

CoSBiLab Graph est un logiciel de visualisation, d'analyse et de manipulation des réseaux sociaux. Il prend en données d'entrée plusieurs formats tels que $^*.dot$, $^*.txt$, $\cdots$. Il prend comme données de sortie les mêmes extensions et $^*.png$ pour la visualisation. Il fonctionne uniquement sous Windows et requiert $.NET3.5$ pour un bon fonctionnement. Les sommets peuvent être agrégés et arrangés dans l'espace.

↬ **Cytoscape**

Il est un logiciel de visualisation, d'analyse et d'intégration des réseaux sociaux. A l'origine, il était développé pour la recherche en bio-informatique. Plusieurs plugins sont implémentés pour étendre ses fonctionnalités. Il est supporté par tous les systèmes qui ont une machine virtuelle java.

↬ **DEX**

Il est un logiciel d'analyse des réseaux sociaux développé en java et en C++. Il peut supporter des millions de sommets et d'arêtes. Il prend comme données d'entrée des fichiers $^*.CSV$ et $^*.jdbc$. Il prend aussi comme données de sorties de fichiers $^*.CSV$. Il existe une version sous Windows et une version sous Linux. Ce logiciel est

commercial.

↬ **igraph**

Igraph est une bibliothèque développée *C* pour l'analyse et la visualisation des graphes de grande taille. Dans *Igraph* plusieurs fonctions de manipulations des graphes sociaux sont implémentés. Il est open source et il existe des versions pour les systèmes d'exploitations Windows, Linux et Mac OS. Il prend en données entrées $^*.txt$ (liste d'arêtes), $^*.graphml$, $^*.gml$, ... Les données de sorties ont les mêmes extensions que les données d'entrées.

↬ **Pajek**

Il est un outil d'analyse et de visualisation des réseaux de taille très grande. il existe des versions pour les systèmes d'exploitations Windows, Linux et Mac OS. Il n'est commercial et il n'est pas aussi open-source.

↬ **R**

R est un logiciel de simulation open-source et libre pour les statistiques. On peut l'utiliser dans plusieurs domaines. Il suffit d'avoir la bibliothèque ou de développer une bibliothèques et l'intégrer. Dans l'analyse des réseaux sociaux, *R* utilise la bibliothèque *igraph* qui a presque toutes les fonctionnalités dans l'analyse et la visualisation de graphes. *R* peut lire presque toutes les extensions et peut générer aussi plusieurs extensions.

4.1.2 Choix

Notre choix se porte sur le logiciel *R* qui est open-source et il existe dans plusieurs systèmes d'exploitations. Avec le logiciel R, on peut faire des simulations dans plusieurs domaines tels que data mining, la recherche opérationnelle, etc. Si on travaille dans un domaine, il suffit de trouver la classe pour ce dernier. Dans la figure 4.1, on a l'interface d'accueil de R dans Windows. Dans la plateforme CRAN [1], on a plusieurs bibliothèques. Nous avons utilisé la bibliothèque *Igraph* [2] qui contient

1. https ://cran.r-project.org/

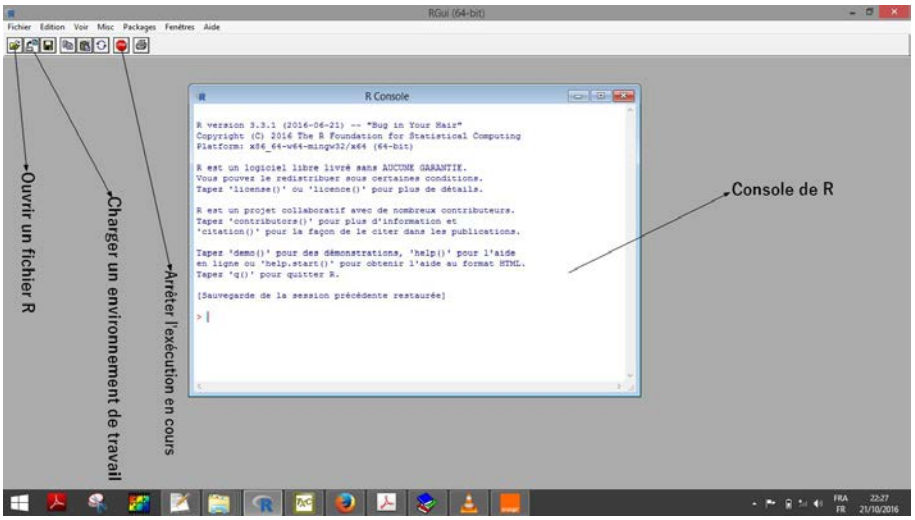

FIGURE 4.1: *Interface d'acueil de R*

plusieurs fonctions d'analyse et de visualisation des graphes. Cette classe existe dans plusieurs outils de simulations comme dans R, avec *Python*, avec le langage *C*. Dans la dernière version de la bibliothèques *Igraph*, nous notons d'importants changements dont les majeurs sont surtout la facilitation de mémoriser les noms des fonctions. Mais les anciens noms des fonctions fonctionnent toujours. On installe la bibliothèque *Igraph* de la même façon que les autres dans *R*. Dans la console, on lance la commande $install.packages(''Nom\_de\_la\_classe'')$ avec bien sûr une connexion internet. Les dossiers seront téléchargés dans un serveur. Après le lancement de la commande $install.packages(''igraph'')$, on sélectionne le serveur comme le montre la figure 4.2. Une fois effectuer l'installation, on passe au chargement de la bibliothèque *Igraph* pour pouvoir utiliser ses fonctions. On lance la commande $library(Nom\_de\_la\_classe)$. Ici, il faut noter que les guillemets présents dans l'installation, ne seront pas mis dans le chargement de la bibliothèque. Alors, pour charger la bibliothèque *Igraph*, on lance la commande $library(''igraph'')$. Maintenant, toutes les fonctions de la bibliothèque sont utilisables. Avec *Igraph*, nous avons plusieurs manières de lire un graphe. Dans nos simulations, nous avons utilisé les

2. http ://igraph.org/2015/06/24/igraph-1.0.0-r.html

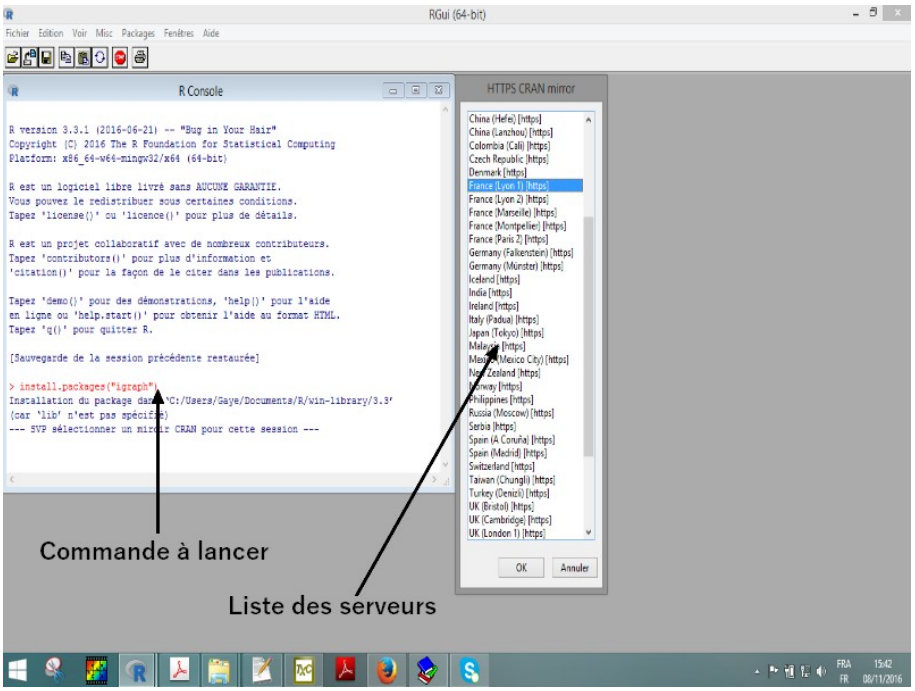

FIGURE 4.2: *Choix du serveur de téléchargement dans R*

graphes de formats *gml* [3]. Les fichiers de codage $*.gml$ représentent les données sous un vecteur. La force de R est aussi la représentation vectorielle sous forme de listes. Dans la section 4.4 de l'annexe, on a un graphe codé sous l'extension *gml*. Dans ce graphe, on définit d'abord l'orientation. On va donner une valeur de zéro à "directed" si le graphe n'est orienté et 1 sinon 0. Ensuite, on liste les sommets par les mots clés : *node* suivi de l'*id* et du *label*. Enfin, on liste les arêtes du graphe par le mot clé *edge*, on donne le sommet source par le mot clé *source* de celui de la destination par le mot clé *target*. Dans le cas des graphes orientés, si on intervertit la source et la destination, ça ne change pas le graphe. Maintenant, il reste à charger le graphe pour éventuellement d'analyse ou la visualisation. Pour le chargement, on exécute la fonction *read.graph* qui prend en argument le chemin du graphe et l'extension d'ouverture. Si le chargement s'est bien effectué, on peut calculer par exemple les mesures de centralités, visualiser le graphe, donner le diamètre du graphe, etc. Dans

3. **G**eography **M**arkup **L**anguage

TABLEAU 4.1: *Quelques fonctions de la bibliothèque Igraph*

Fonction	Arguments	description
plot	G(V,E)	représentation sagittale de G
betweenness	G(V,E) G(V,E),node	centralité intermédiaire de tous les sommets de G centralité intermédiaire du sommet node
degree	G(V,E) G(V,E), node	centralité degré de tous les sommets du graphe centralité degré du sommet node
closeness	G(V,E) G(V,E),node	centralité par proximité de tous les sommets de G centralité par proximité du sommet node
Rank	G(V,E) G(V,E),node	PageRank de tous les sommets de G centralité Page Rank du sommet node
read.graph	url, format	lire un graphe
length	ensemble	la taille de l'ensemble
write.graph	G(V,E), url, format	écrire G(V,E) sous le format de sortie donné dans url
runif	n,bornes inf et sup	Génère uniformément n nombre compris entre inf et sup

le tableau 4.1, nous avons quelques fonctions qui sont dans *igraph*.

Dans latex, la bibliothèque *tikz* permet de créer des schémas. On a pas besoin de faire des captures avec cette bibliothèque. Dans *R*, dans la bibliothèques *Igraph*, la fonction *plot* permet de visualiser les graphes chargés, le résultat d'une simulation, etc. Ce flux de la visualisation peut être redirigé dans un fichier *tikz*. Dans la section 4.5 de l'annexe, nous avons des exemples de codes *tikz* et comment générer les avec *R*

Igraph [4] est une bibliothèque qui contient plusieurs fonctions de manipulations des graphes. Il existe une version utilisable dans *R*, une version utilisable avec le langage *Python* et une version utilisable avec le langage C. Dans la dernière version, nous notons d'importants changements.

4. http ://igraph.org/2015/06/24/igraph-1.0.0-r.html

TABLEAU 4.3: *Réseaux multicouches utilisés dans nos simulations*

Réseaux sociaux	agrégation	Couche RT	Couche RP	Couche MT
Cannes2013	N=348537 M=991855	N=340349 M=496982	N=85867 M=83535	N=233735 M=411338
NY CLIMATEMARCH2014	N=102439 M=353496	N=94574 M=213754	N=7928 M=8063	N=50054 M=131679

4.2 Jeux de données

Pour montrer les performances de nos approches dans la détection des semences dans le problème de la maximisation de la diffusion de l'influence, nous avons utilisé plusieurs jeux de données détaillés dans les deux tableaux ci-dessous. Pour les RSM, les jeux de données sont dans le tableau 4.2 et ceux utilisés pour les RSMC sont dans le tableau 4.3. Dans nos simulations, nous avons utilisé le logiciel R. Ce dernier peut manipuler des graphes via la bibliothèque *igraph*. Il peut lire des fichiers $*.edge$, $*.gml$. Les sources des jeux de données sont sous forme liste d'arêtes. Nous avons utilisé le langage C pour convertir ce fichier sous l'extention $^*.gml$.

TABLEAU 4.2: *Réseaux monoplex utilisés dans nos simulations*

Réseau social	Nombre de sommets	Nombre arêtes	Heuristique
Amazon	334863	925873	C^{ℓ}_{dd}
Dolphins	62	159	C^{ℓ}_{dd}
DBLP	317080	1049866	Algorithme $SCG_{v1\ \&\ v2}$
Eron Email	36692	367662	Algorithme SG

↬ Dolphins [5]

Dolphins communauté est un réseau compilé par D. Lusseau K. Schneider [60] et composé de 62 dolphins et 159 liens entre ces dolphins. Ce réseau social peut être représenté par un graphe non orienté. Nous avons utilisé ce graphe social pour

5. https ://networkdata.ics.uci.edu/data.php ?id=6

simuler la construction du graphe couvrant de maximisation dans le chapitre 3 au paragraphe 3.6. Nous avons aussi utilisé ce réseau pour montrer la vitesse de diffusion de l'influence de l'heuristique C^{ℓ}_{dd}.

↬ **Amazon Communities** [6]

Ce réseau est un extrait du site *Amazon*. Il est basé sur les produits que les clients achètent ensemble. Chaque sommet représente un produit et si deux produits sont achetés ensemble, un lien est construit entre ces deux produits. On compte 334.863 produits étiquetés de 1 à 334.863 et 925.873 liens. Si un produit u et un produit v sont achetés ensemble, on crée une arête uv. A partir de ces données, nous avons construit un graphe non orienté. Nous l'avons utilisé pour montrer que les semences données par l'heuristique C^{ℓ}_{dd} diffusent mieux l'influence que celles données par les heuristiques de références.

↬ **DBLP** [7]

Ce réseau est la liste de documents de recherches scientifiques. On peut aussi l'appeler un réseau de co-auteurs. Dans ce réseau social, chaque auteur représente un sommet et si deux auteurs produisent ensemble un article, on crée un lien entre ces deux auteurs. Dans ce réseau, nous avons plusieurs communautés. Il y a des papiers de conférences, des journaux, $\cdots$. Chaque communauté aussi peut être vue comme une couche si on le voit comme un RSMC développé précédemment. Supposons qu'il y a un auteur α et un auteur β qui ont publié un article ensemble, on crée une arête $\alpha\beta$. Après compilation, nous avons 317.080 auteurs (sommets) et 1.049.866 de relations (arêtes). A partir de ces données, nous avons construit un graphe non orienté que nous avons utilisé pour évaluer les performances du graphe couvrant de maximisation donné par les deux versions de l'algorithme SCG.

↬ **Eron Email** [8]

6. http ://snap.stanford.edu/data/com-Amazon.html
7. http ://snap.stanford.edu/data/com-DBLP.html
8. http ://snap.stanford.edu/data/email-Enron.html

Dans ce réseau, nous avons un ensemble de Emails. Ces données ont été initialement rendues publiques et affichées sur le web, par la "Federal Energy Regulatory Commission" au cours de son enquête. Ici, les sommets représentent les e-mails. Soient deux e-mails α et β, si le premier envoie au second -email un message, un arc est créé de α vers β. Après un temps, ils récupèrent l'ensemble des e-mails et ces liens pour compiler un graphe orienté. Dans ce graphe, nous avons 36.692 e-mails (sommets) et 183.831 inter-actions (arcs). Ce réseau est utilisé pour évaluer les performances du graphe couvrant de maximisation donné par l'algorithme SG.

↬ **Cannes2013** [9] et **NY CLIMATEMARCH2014** [10]

Ces deux réseaux sont des extraits du réseau social $Twitter$. Les comptes représentent les sommets. Les liens sont divisés en trois types. L'objectif de cette division est de simuler un RSMC. Les trois couches sont notées respectivement RT,RP et MT.

Dans la couche RT, nous avons les liens ReTweet. Un utilisateur retweete (re-publie) une information d'un autre utilisateur, alors une arête de type RT est construite entre ces deux.

Dans la couche RP, nous avons les liens Reply. Un utilisateur de Twitter répond à un autre utilisateur, alors un lien de type RP est crée entre ces deux.

Dans la couche MT, nous avons les liens MenTion. Un utilisateur mentionne un autre utilisateur dans un message, alors un lien entre ces deux st créé.

9. http ://deim.urv.cat/manlio.dedomenico/data.php
10. http ://deim.urv.cat/manlio.dedomenico/data.php

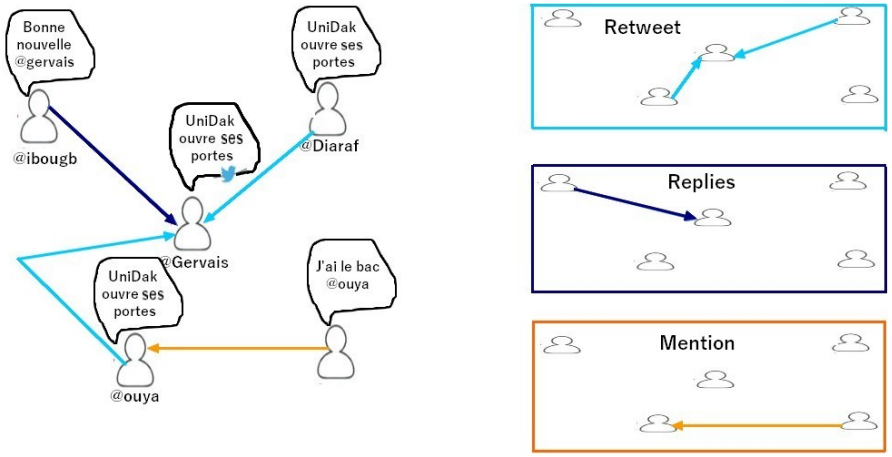

FIGURE 4.3: *Illustration des trois couches*

Dans ce réseau nous avons donc trois évènements qui composent les trois couches $\{RT, RP, MT\}$. Dans chaque couche, nous avons une seule nature de lien. Par exemple, dans la couche, la seule nature de lien est RT, dans la deuxième couche, nous avons seulement de liens de types RP et dans la dernière couche, la nature des liens est MT. Dans ce RSMC, nous pouvons avoir un tweet présent dans toutes les couches, dans deux couches ou dans une seule couche. Alors la somme des utilisateurs sera plus grande que les utilisateurs qui sont dans le réseau agrégé. Mais la somme des liens sera la même que dans le réseau agrégé. Dans le tableau 4.3, nous avons les détails de ces deux RSMC avec trois couches. Dans la figure 4.3 nous avons représenté les trois différentes natures de communication entre les tweets.

4.3 Présentation des résultats

Dans cette section, nous montrons que nos heuristiques donnent de meilleurs résultats par rapport à celles de référence. Nous travaillons principalement avec deux modèles de propagation à savoir IC et LT dont leurs implémentations avec le logiciel de simulation R sont dans la section 4.6 de l'annexe. Ici, nous présenterons d'abord les résultats de l'heuristique C_{dd}^{ℓ}, ensuite ceux de l'heuristique C_{dd}^{MLN} et enfin, ceux

du graphe couvrant de maximisation.

4.3.1 Heuristique degré diffusion $\ell-$ième (C^{ℓ}_{dd})

4.3.1.1 Heuristiques de références de paramètres de simulations

L'heuristique C^{ℓ}_{dd} est basée sur le modèle de propagation IC qui sera utilisé dans nos tests. Dans l'état de l'art sur la détection des semences, nous avons classé les heuristiques en trois approches, une approche algorithmique, une approche de centralité et une approche gloutonne. Comme la C^{ℓ}_{dd} est une approche de centralité alors, nous avons utilisé trois heuristiques basées sur la centralité et une heuristique basée sur l'algorithme qui donne aussi une métrique à chaque sommet de façon dynamique, comme des heuristiques de références qui sont :

- L'heuristique plus *haut degré*, notée par c_d, qui donne l'importance aux sommets qui ont le plus de voisins.
- L'heuristisque *intermédiaire*, notée par c_b, qui donne de l'importance aux sommets qui appartiennent à plus de chemins. Dans cette heuristique, un sommet est important s'il participe à plus de chemins.
- L'heuristique *degré discontinu*, notée par c_d *Discount*, qui se base sur la notion de voisinage. Dans cette heuristique, si un sommet v est choisi comme élément de S^*_k, pour tous ses voisins v de u, l'arête uv ne sera pas pris en compte. Le paramètre p de l'heuristique dégré discontinu (l'algortihme 2) est fixé à 0.5.
- L'heuristique *Rand* qui prend aléatoirement k sommets comme l'ensemble des semences.

Pour calculer la propagation de l'influence des semences données par ces différentes heuristiques, nous utilisons les jeux de données *Amazon* et *dolphins* dont les détails sont développés dans la section ci-après.

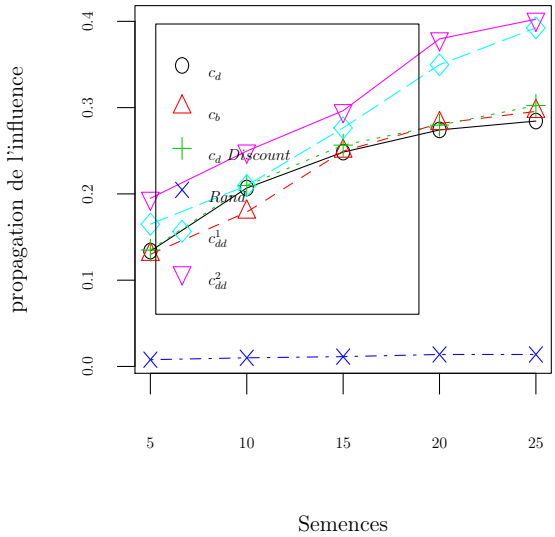

FIGURE 4.4: *Propagation de l'influence des semences données par C^{ℓ}_{dd} et les modèles références sous IC*

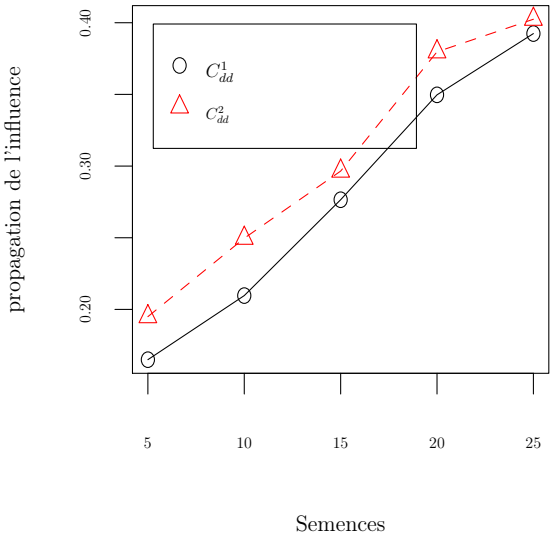

FIGURE 4.5: *Propagation de l'influence des semences données par C^{ℓ}_{dd} en fonction de ℓ sous IC*

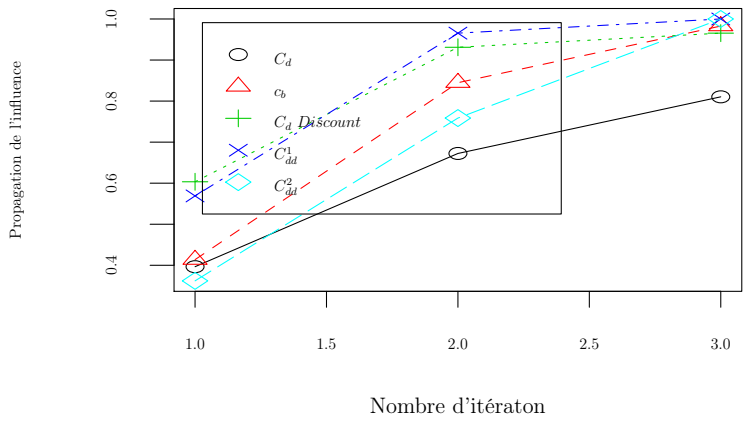

FIGURE 4.6: *Propagation de l'influence des semences données par C^{ℓ}_{dd} et les modèles références en fonction du nombre d'itération sous IC*

4.3.1.2 Résultats

Dans nos simulations, nous avons fait trois tests. D'abord, nous avons mesuré la propagation de l'influence des semences données par l'heuristique C^{ℓ}_{dd} et celles de références. Ensuite, nous avons comparé la propagation de l'influence si on change le rayon de l'heuristique C^{ℓ}_{dd}, autrement dit, si on prend en compte les voisins N^1 et N^2. Enfin, nous avons évalué la vitesse de propagation de l'influence des semences donnée par l'heuristique C^{ℓ}_{dd} et celle des références citées ci-dessus.

Dans la figure 4.4, nous comparons aux heuristiques de référence en faisant varier les semences de 5 à 25 par pas de 5 (S^*_k, $k \in [5, ..., 25]$). Pour chaque ensemble S^*_k, nous calculons sa propagation de l'influence sous le modèle de propagation IC. Le paramètre ℓ de notre heuristique est fixé à 1 pour voir la contribution de voisins N^1 et à N^2 pour voir la contribution des voisins N^2. En d'autres mots, nous calculons la propagation de l'influence des semences données par C^1_{dd} et C^2_{dd}. Dans les résultats, on voit que notre heuristique donne de meilleurs résultats que les autres heuristiques. Comme nous l'avons montré théoriquement, un utilisateur peut subir la pression sociale comme il peut ne pas en subir. Dans notre heuristique, nous avons pris en compte cela en intégrant une probabilité de propagation de l'information. Dans notre heuristique, nous prenons en compte non seulement la probabilité de propagation mais aussi le nombre de voisins de niveau ℓ. Autrement dit, plus l'expansion de rayon ℓ et de centre le sommet v a des utilisateurs influents, plus v est important dans la propagation de l'influence. Ces paramètres ne sont pas pris en compte par les autres heuristiques. Dans la figure 4.4, nous voyons que la propagation de l'influence de C^{ℓ}_{dd} est plus grande que celle des heuristiques de référence pour toutes les valeurs de k. Dans cette heuristique, nous prenons en compte dans l'influence de tous voisins de niveaux ℓ qui ne sont pas tous pris en compte pour les heuristiques références.

Dans la figure 4.5, nous mesurons les performances de notre heuristique en fonction du paramètre ℓ. Autrement dit, nous calculons la propagation de l'influence de C^{ℓ}_{dd} avec différentes valeurs de ℓ. Nous avons comparé les résultats donnés par C^1_{dd} et C^2_{dd}.

Ici, le nombre de semences varie entre 5 et 25 par pas de 5 et pour chaque ensemble de semences, nous calculons leur propagation de l'influence. Dans les résultats, nous voyons que, si on travaille avec N^1, nous aurons des résultats plus faibles que si nous travaillons avec N^2. A $k = 25$ la courbe C^1_{dd} semble dépasser la courbe C^2_{dd}, mais les grandes valeurs de k ne nous intéressent pas, parce que nous avons besoin d'une valeur de k plus petit possible comme nous l'avons posé dans la problématique.

Dans la figure 4.6, nous mesurons la vitesse de propagation de notre approche et celle des heuristiques références. Nous avons fixé le nombre de semences à quatre ($k = 4$). Le nombre d'itérations qui varie de 1 à 3 et à chaque itération, nous calculons la propagation de l'influence des semences (la nôtre et les références). Dans les résultats, pour $\ell = 1$ notre approche converge plus rapidement que les autres heuristiques références. Mais pour $\ell = 2$, nous notons un retard et c'etait prévisible. A la première itération, ce sont les semences qui vont diffuser l'information. À la deuxième itération, c'est le tour des voisins N^1 influencés par les semences alors que notre heuristique qui tardait à décoller commence à rattraper les autres et à la troisième itération, elle converge en même temps que les autres. Dans notre heuristique, à la troisième itération, c'est le tour des voisins N^2 influencés par les voisins directs des semences qui vont essayer d'influencer leurs voisins. Et dans notre heuristique nous avons pris en compte ces sommets et leur probabilité de propager l'information. Ce qui justifie que la vitesse de propagation de C^ℓ_{dd} peut être faible au début mais au fil du temps, elle évolue très rapidement.

4.3.2 Heuristique degré multi-diffusion (C^{MLN}_{dd})

4.3.2.1 Heuristiques de références de paramètres de simulations

Dans ces simulations, nous avons utilisé l'heuristique plus haut degré appelée *multi-degré* redéfinie par Magnani Matteo *et al.* [9]. C'est une réadaptation de l'heuristique plus haut degré définie par kempe David *et al.* [11] mais dans les réseaux sociaux

multicouhes. Elle est définie dans l'équation 4.1. Si on prend un voisin, la nouvelle heuristique cherche ses voisins dans la même couche et dans les autres couches. Avec les deux jeux de données, nous calculons la propagation de l'influence des deux ensembles de semences données par notre approche et celle de référence. Le nombre d'itération est fixé à 3 pour le graphe *NYClimateMarch2014* et à 4 pour le graphe *Cannes2013*. La probabilité d'activation est choisie uniformément entre 0 et 1 par la fonction *runif* de *R*. Le modèle proposé est basé sur le modèle de propagation IC. Alors, nous avons utilisé le modèle IC adapté par Salehi Mostafa [30] dans les RSMC.

$$\delta(v) = |P_{eqIMi}(\bigcup_{i\in[1..n],(u,v)\in E_i} u)| \tag{4.1}$$

4.3.2.2 Résultats

Pour montrer les performances de notre modèle dans les RSMC, nous déterminons les $k - top$ donnés par notre heuristique et l'heuristique de référence (*multi-degré*). Nous calculons la propagation de l'influence initiée par chacun des deux ensembles de semences. Dans les figures 4.7 et 4.8, nous avons utilisé le RSMC *Cannes2013* et dans les figures 4.10 et 4.9, nous avons utilisé le RSMC *NYClimateMarch2014*.

Dans les figures 4.7 et 4.9, nous calculons la propagation de l'influence des deux ensembles de semences sous le modèle IC après un nombre d'itérations fixe. Le nombre de semences varie entre 5 et 30 par pas de 5. Pour chaque nombre de semences, notre approche donne de meilleurs résultats que l'approche de référence. En effet, comme nous l'avions dit théoriquement, notre approche prend en compte la probabilité qu'un sommet diffuse l'information et les voisins de niveau 2 et leur probabilité de diffuser. Ces deux paramètres ne sont pas pris en compte par le modèle de référence. Ce qui justifie que la propagation de l'influence de notre modèle est plus loin que le modèle de référence.

Dans les figures 4.8 et 4.10, nous calculons la vitesse de propagation de l'influence des deux modèles. Nous choisissons 30 semences pour chaque modèle. Dès la première itération, les semences données par notre modèle propage plus l'information. Le modèle de base ne prend pas en compte la probabilité de diffuser alors que le nôtre la prend en compte. A la deuxième itération, notre modèle prend en compte le nombre de voisins. Or dans la deuxième itération, c'est le tour de voisins niveau 1 activés qui vont propager l'information. Notre modéle prend en compte du nombre de voisins de niveau 2 des sommets semences et aussi leur probabilité de diffusion. Ce qui justifie que, lors de la deuxième itération, notre modèle donne de meilleurs résultats. La vitesse de propagation de l'information est plus rapide qu'avec notre modèle que celui de référence.

Dans les RSMC, les approches proposées ne prennent pas en compte l'apport des voisins de niveau 2 de chaque semence et la probabilité de diffusion dans toutes les couches qui sont des paramètres tres importants dans la diffusion de l'information. Sur le plan théorique et pratique, nous avons montré que si ces deux paramètres sont pris en compte, la propagation de l'influence est plus importante et sa vitesse aussi est rapide.

4.3.3 Graphe couvrant de maximisation

4.3.3.1 Heuristiques de référence et paramètres de simulations

Comme dans les deux modèles de propagation références, un sommet u peut diffuser une information et cette même information revient vers le sommet u mais elle sera ignorée. Pour prévoir le retour de l'information vers les sommets semences, nous utilisons trois heuristiques, basées sur la centralité des sommets, qui sont :

- *haut degré*
- *degré discontinu* et le paramètre p fixé à 0.01.
- *PageRank* qui est utilisée par le moteur de recherche *google* pour chercher les

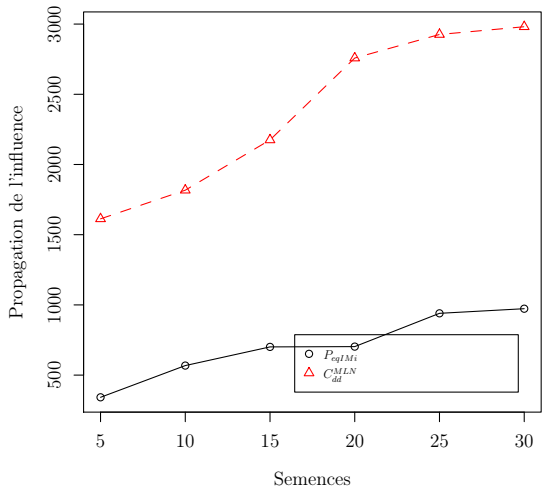

FIGURE 4.7: *Propagation C_{dd}^{MLN} vs P_{eqIMi} sous IC*

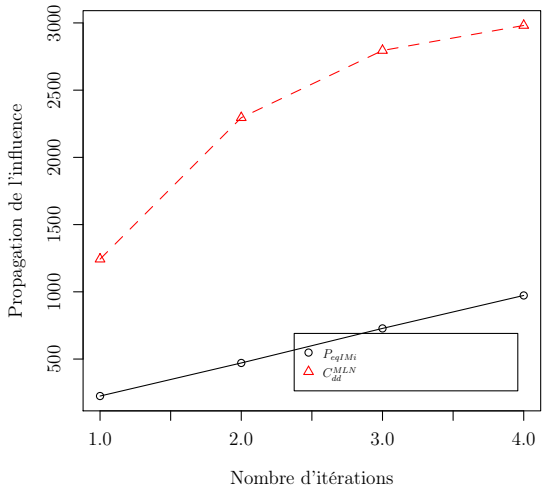

FIGURE 4.8: *Propagation C_{dd}^{MLN} vs P_{eqIMi} sous IC en fonction de l'itération*

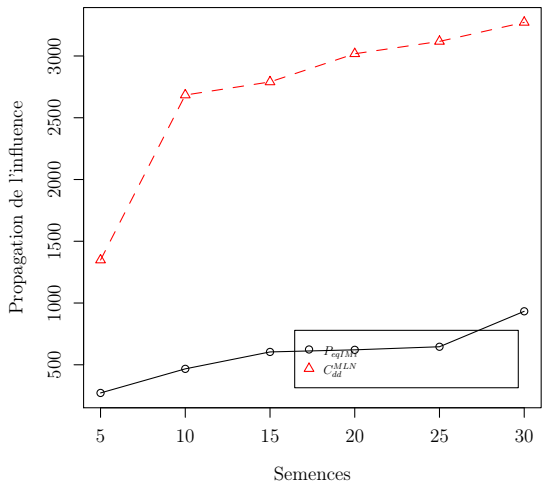

FIGURE 4.9: *Propagation C_{dd}^{MLN} vs P_{eqIMi} sous IC*

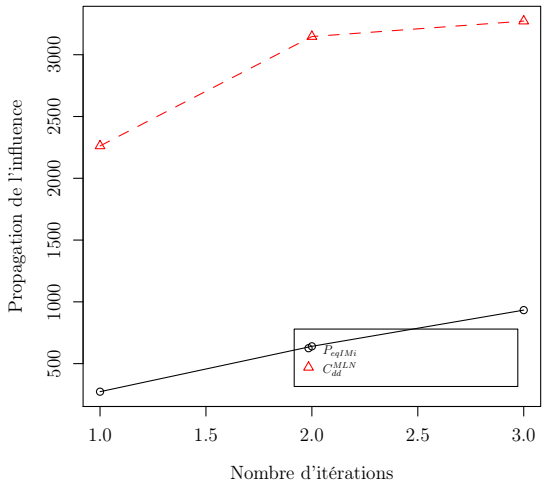

FIGURE 4.10: *Propagation C_{dd}^{MLN} vs P_{eqIMi} sous IC en fonction e l'itération*

pages webs en fonction d'une requête d'un utlisateur. Le paramètre d'amortissement [11] p est fixé à 0.8.

Pour calculer la propagation de l'influence des ces différentes heuristiques, nous utilisons les deux modèles de propagation LT et IC et les jeux de données $DBLP$ et $Enron$.

4.3.3.2 Résultats

Pour montrer la pertinence de prévenir le retour d'information vers les sommets semences avant leurs détections, nous utilisons une heuristique quelconque tels que degré, PageRank, etc et nous calculons la propagation de l'influence des semences sous un modèle de propagation ψ. Cette heuristique sera appliquée dans le graphe initial et dans le graphe couvrant de maximisation. Dans la suite, nous utilisons les ensembles de semences suivants :

- s_0^k : ensemble de k semences extrait dans le graphe initial G=(V,E),
- s_1^k : ensemble de k semences extrait dans le graphe couvrant de maximisation obtenu à partir de l'algorithme SCG_{v1},
- s_2^k : ensemble de k semences extrait dans le graphe couvrant de maximisation obtenu à partir de l'algorithme SCG_{v2},
- s_3^k : ensemble de k semences extrait dans le graphe couvrant de maximisation obtenu à partir de l'algorithme SG,

Pour les tests, nous faisons varier le nombre de semences et nous calculons pour chaque s_k la propagation de l'influence après un nombre d'itérations fixé.

Algorithme SCG

11. damping

plus haut degré (C_d)

En utilisant l'heuristique plus haut degré sous les modèle LT et IC, les figures 4.11 et 4.12, montrent que les semences (s_1^k et s_2^k) données par notre approche donnent de meilleurs résultats que s_0^k données par le graphe initial. Dans la construction du graphe couvrant, si on prend en compte du nombre de voisinage, on a encore de meilleurs résultats avec l'heuristique plus haut *degré* sous les modèles de propagations LT et IC.

degré discontinu

En utilisant l'heuristique *degré dicontinu* sous les modèle LT et IC, les figures 4.13 et 4.14, montrent que les semences données par notre approche (s_1^k et s_2^k) donnent de meilleurs résultats que celles données par le graphe initial (s_0^k). Dans la construction du graphe couvrant, si on prend en compte du nombre de voisinage, on a encore de meilleurs résultats avec l'heuristique *degré dicontinu* sous les modèles de propagations LT et IC.

Page Rank

En utilisant l'heuristique *PageRank* sous les modèle LT et IC, les figures 4.15 et 4.16, montrent que les semences données par notre approche (s_1^k et s_2^k) donnent de meilleurs résultats que celles données par le graphe initial (s_0^k). Dans la construction du graphe couvrant, si on prend en compte du nombre de voisinage, on a encore de meilleurs résultats avec l'heuristique *PageRank* sous les modèles de propagations LT et IC.

Algorithme SG

En utilisant les heuristiques plus haut degré et degré discontinu sous les modèle LT et IC, les figures 4.18 et 4.17, montrent que les semences donnés par notre approche (s_3^k) donnent de meilleurs résultats que celles données par le graphe initial (s_0^k). Dans la construction du graphe couvrant, si on prend en compte du nombre

de voisinage, on a encore de meilleurs résultats.

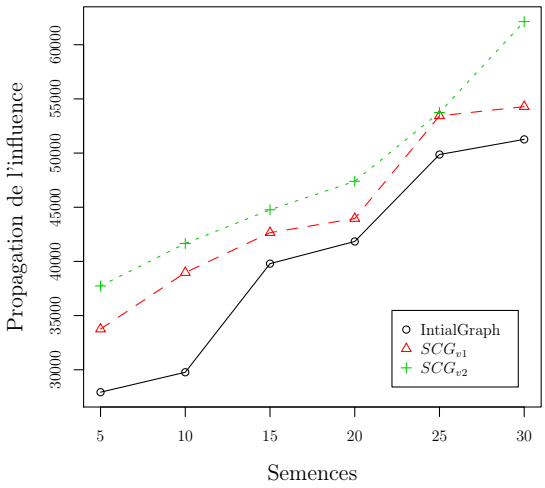

FIGURE 4.11: *Propagation* s_0^k *vs* s_1^k *vs* $s_2^k C_d$ *sous IC et* C_d

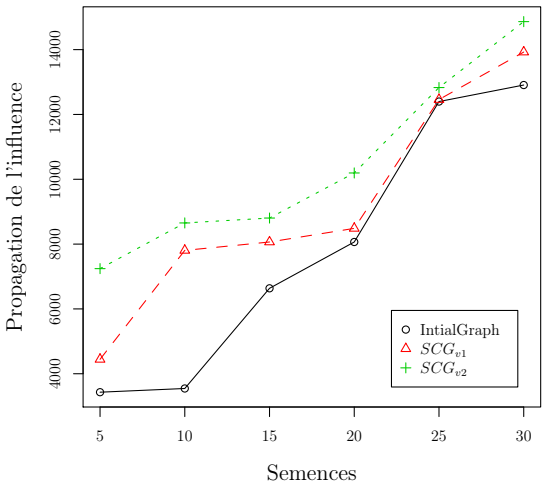

FIGURE 4.12: *Propagation* s_0^k *vs* s_1^k *vs* $s_2^k C_d$ *sous LT et* C_d

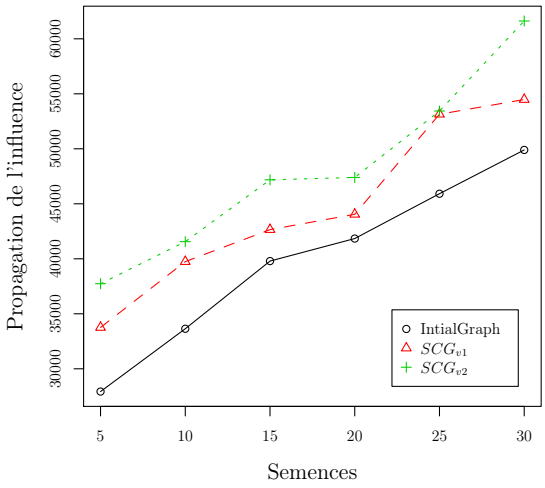

FIGURE 4.13: *Propagation* s_0^k *vs* s_1^k *vs* $s_2^k C_d$ *sous IC et* C_d *discontinu*

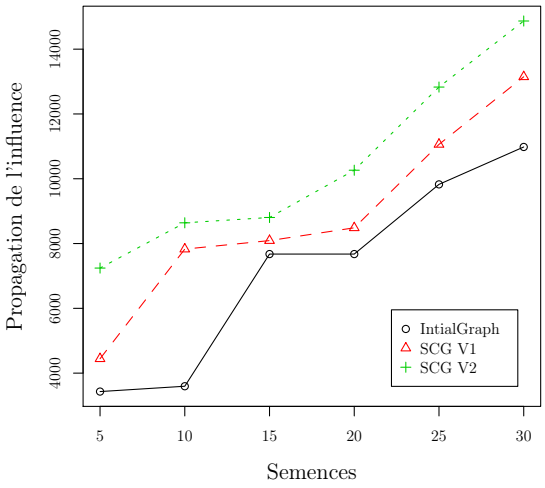

FIGURE 4.14: *Propagation* s_0^k *vs* s_1^k *vs* $s_2^k C_d$ *sous LT et* C_d *discontinu*

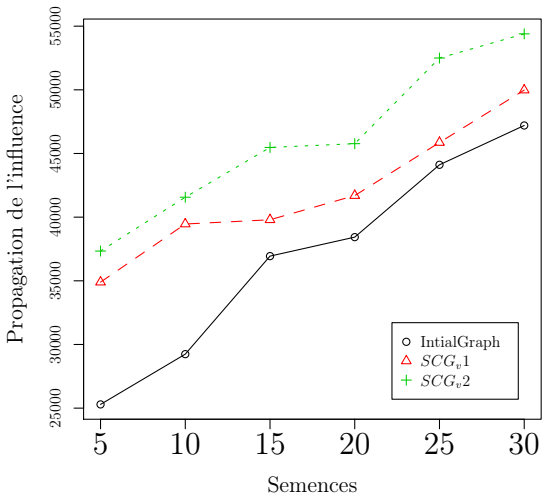

FIGURE 4.15: *Propagation s_0^k vs s_1^k vs $s_2^k C_d$ sous IC et PageRank*

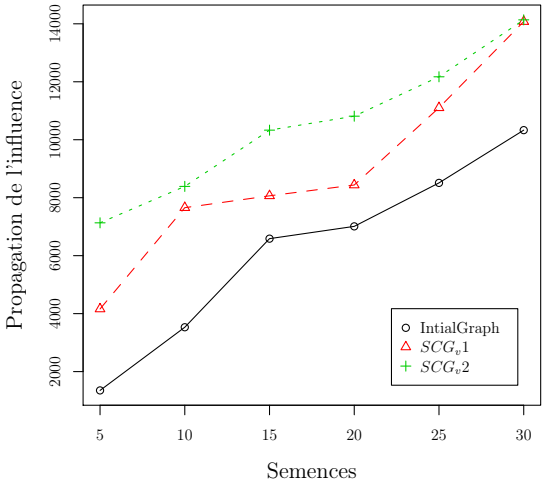

FIGURE 4.16: *Propagation s_0^k vs s_1^k vs $s_2^k C_d$ sous LT et PageRank*

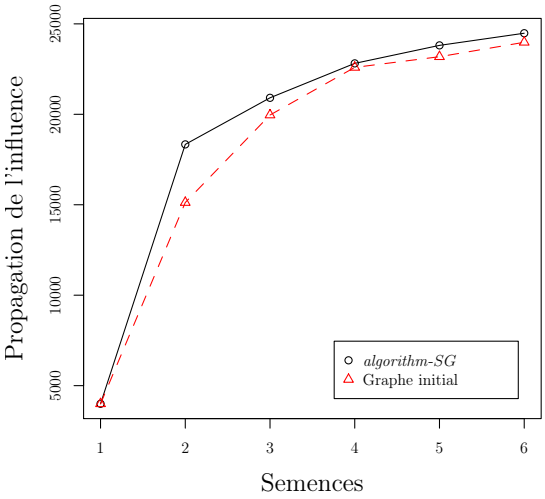

FIGURE 4.17: *Propagation s_0^k vs s_3^k C_d sous IC et C_d*

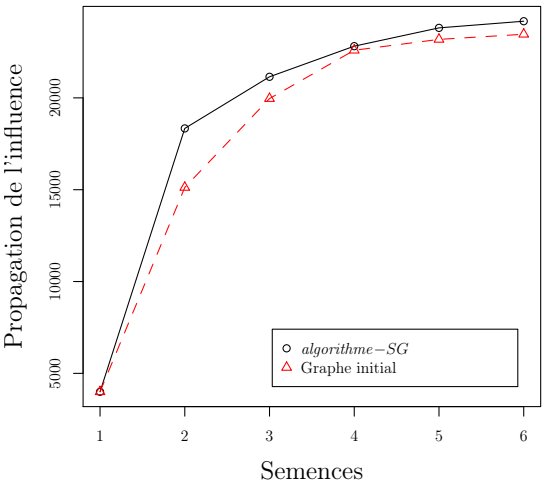

FIGURE 4.18: *Propagation s_0^k vs s_3^k C_d sous IC et C_d discontinu*

Toutes ces figures montrent que, les semences obtenues en prévenant le retour d'information donnent de meilleurs que celles obtenues sans la prévention sous les modèles

de propagations cascades et seuils. Comme nous l'avions montrer théoriquement, dans ces deux modèles, à une itération donnée, on essaye d'activer seulement les sommets inactifs. Nous avons proposé de prévenir ce retour d'information au moment de la détection les semences. Nos simulations confirment que cette élimination est pertinante dans la maximisation de la propagation de l'influence.

Conclusion

Dans ce chapitre, nous avons montré, par des tests, que nos approches sont plus performantes que celles existantes. D'abord, nous avons donné quelques outils de simulations et nous avons choisi un qui est open source et très facile à utiliser. Enfin nous avons utilisé des jeux de données et nous avons fait deux types de comparaison à savoir calculer la vitesse de propagation de l'influence et calculer la propagation de l'influence pour un ensemble de semences donné. Dans tous les tests, nos approches donnent de meilleurs résultats.

Conclusions et travaux futurs

De nos jours, l'internet est devenu indispensable dans la vie quotidienne des individus. Les réseaux sociaux avec leurs nombres d'utilisateurs qui n'arrêtent de croître exponentiellement, font partis des systèmes les plus utilisés sur internet. Avec ces réseaux, plusieurs types d'informations tels que des documents, des photos, des vidéos, etc sont partagés. L'analyse des réseaux sociaux (ARS) [12] devient un domaine de recherche très vaste. Le thème traité dans ce manuscrit est la maximisation de l'influence dans les réseaux sociaux. Elle consiste de trouver une combinaison de k-utilisateurs (S_k^*) dans les réseaux sociaux qui vont maximiser la propagation de l'influence avec un modèle de propagation ψ. Cette problématique consiste à trouver les diffuseurs initiaux et un modèle de diffusion optimal. Nos travaux se portent essentiellement sur la détection des diffuseurs initiaux. Nous avons proposé d'abord, une mesure de centralité appelée degré de diffusion ℓ-ième et elle est notée par C_{dd}^{ℓ}. Ensuite, souvent les réseaux sociaux ont plusieurs type de liens. Chaque nature de liens peut être vue comme un réseau social monoplex. Dans une agrégation de plusieurs réseaux sociaux, les mêmes utilisateurs peuvent être connus. Alors un système qui a plusieurs de types liens peut être vu comme un réseau multicouche. Des travaux ont montré que les modèles dans les réseaux monoplex ne sont pas efficaces. Nous avons proposé une mesure de centralité appelée degré multi diffusion et elle est notée par C_{dd}^{MLN}. Enfin, après avoir étudié le fonctionnement des modèles de propagations de bases à savoir LT et IC, nous avons proposé de faire un pré-traitement du

12. Social network analysis (SNA)

graphe initial en éliminant les retours d'information vers les sommets semences. Le traitement donne un graphe couvrant qu'on a appellé graphe couvrant de maximisation. Ce manuscrit est composé de quatre chapitres. Nous avons commencé dans le premier chapitre par donner les définitions et terminologie dans la maximisation de l'influence. Dans le second chapitre, nous avons fait un état de l'art dans la diffusion de l'information dans les réseaux sociaux en expliquant le principe de fonctionnement de ces derniers. Dans le troisième chapitre, nous avons détaillé nos modèles proposés en faisant un états de l'art sur la détection des semences dans les réseaux sociaux monoplex et multicouches. Enfin, dans le quatrième chapitre, nous avons montré avec des simulations que nos approches sont plus performantes que les existantes.

Dans les réseaux sociaux (monoplex et multicouches), il reste plusieurs travaux à faire. Dans le futur, nous comptons :

Utiliser de vrais réseaux sociaux afin voir qualitativement, l'apport de nos heuristiques dans la propagation de l'influence.

Dans la détection des diffuseurs initiaux, les modèles, que nous avons proposé, utilisent une probabilité pour qu'un utilisateur diffuse l'information. Nous pensons utiliser faire une corrélation entre l'information à diffuser et les activités de chaque utilisateur. Cette corrélation nous permet d'avoir l'importance de chaque utilisateur pour une information donnée.

Nous pensons aussi à utiliser l'apprentissage approfondi [13], qui est une nouvelle approche de fouille de données, dans la détection de liens, des communautés et aussi des semences.

Dans les réseaux multicouches, la plupart des travaux restent théoriques. Et les modèles proposés dans les réseaux monoplex ne sont pas efficaces dans les réseaux multicouches. Alors, tous les résultats dans les réseaux monoplex peuvent être redéfinis pour les réseaux mutlicouches.

13. deep learning

CONCLUSIONS ET TRAVAUX FUTURS

Dans tous nos travaux, nous avons traité le problème de la maximisation de l'influence dans les réseaux sociaux. Il est important que cette influence soit positive. La détection de la nature de l'influence avant la maximisation est très importante. Par exemple, si l'influence est une rumeur, on parle d'influence négative, plutôt il faut minimiser sa propagation. Dans le futur, nous comptons travailler dans la détection et la minimisation de l'influence.

Annexe

4.4 Exemple de graphes avec le codage GML

Le logiciel *R* avec la bibliothÃ¨que *igraph* ont plusieurs formats de représenter les graphes. Dans nos travaux, nous avons utilisé la représentation *.gml* qui est une liste de listes. D'abord, il faut commencer par renseigner l'orientation du graphe par le mot clé : "directed" qui prendra la valeur 0 ou 1. Ensuite, dans la deuxième sous, nous passons par lister les sommets en utilisant le mot clé : "node" qui est caractérisé par son id, son label et poids si le graphe est pondéré. Enfin, nous terminons par lister les arêtes ou les arcs selon l'orientation. Pour cela, nous utilisons le mot clé : "edge" en donnant la source et la destination. Ci-après, nous avons un exemple de graphe de format *.gml*. graph

```
[
    directed 0
    node
    [
        id 1
        label "Beescratch"
    ]
    node
    [
```

```
            id 2
            label "Bumper"
        ]
    edge
    [
            source 1
            target 2
        ]
]
```

4.5 Code tikz

Dans le logiciel *R*, nous avons une classe nommée *tikzdevice* qui permet de récupérer les résultats sous le format compatible avec latex. *Tikz* est un package de latex permettant de tracer des figures avec une excellente qualité. Ci-après, nous avons deux codes qui représentent successivement, la figure 2.11 et et le résultat donné par la méthode *matplot* de *R*.

4.5.1 Graphe de la figure 2.11

```
\begin{tikzpicture}[x=15pt,y=15pt]
\draw [line  width=2pt, draw=red](2,6) circle (0.5)node[center]{4};
\draw [line  width=2pt, draw=red](8,6) circle (0.5)node]center]{10};
\draw [line  width=2pt](0,3) circle (0.5)node[center]{1};
\draw [line  width=2pt](2,3) circle (0.5)node[center]{5};
\draw [line  width=2pt](4,3) circle (0.5)node[center]{6};
\draw [line  width=2pt](7,3) circle (0.5)node[center]{7};
\draw [line  width=2pt](9,3) circle (0.5)node[center]{11};
```

\draw [line width=2pt](2,0) circle (0.5)node[center]{3} ;

\draw [line width=2pt](0,0) circle (0.5)node[center]{8} ;

\draw (-3,6) node[center]{Itération 0} ;

\draw (-3,3) node[center]{Itération 1} ;

\draw (-3,0) node[center]{Itération 2} ;

\draw (-3,-3) node[center]{Itération 3} ;

\draw [line width=2pt,->](2,5.5) – (0,3.5) ;

\draw [line width=2pt,->](2,5.5) – (2,3.5) ;

\draw [line width=2pt,->](2,5.5) – (4,3.5) ;

\draw [line width=2pt,->](0,2.5) – (0,0.5) ;

\draw [line width=2pt,->](2,2.5) – (2,0.5) ;

\draw [line width=2pt,->](8,5.5) – (7,3.5) ;

\draw [line width=2pt,->](8,5.5) – (9,3.5) ;

\draw [line width=2pt](-2,7) – (10,7) ;

\draw [line width=2pt](-2,4.5) – (10,4.5) ;

\draw [line width=2pt](-2,1.5) – (10,1.5) ;

\draw [line width=2pt](-2,-1.5) – (10,-1.5) ;

\end{tikzpicture}

4.5.2 Le flux de la méthode plot de la classe *Igraph*

Dans code, on passe au chargement du package, ensuite on met en place le flux avec la méthode *tikz*, enfin on utilise *matplot* de *R*. Tout le fux sera mis dans le fichier *degree.tex*. library(tikzDevice)

tikz(file = "C ://Users/Mr GAYE/Desktop/these/trav/article3/resul/degree/degree.tex", width = 5, height =5)

SpanningTree_degree<-

scan("C ://Users/Mr GAYE/Desktop/these/trav/article3/resul/degree/S_tree_degree.txt")

```
degree<-
scan("C ://Users/Mr GAYE/Desktop/these/trav/article3/resul/degree/degree.txt")
gen<-seq(5, 30, by=5)
matplot(gen,
cbind(SpanningTree_degree,degree),
type="o",pch=1 :2,
col=1 :2,
xlab="Nb Node",
ylab="Influence propagation",
sub="Spanning Tree VS Degree")
legend("bottomright", inset=.05, legend=c("Spanning TreeDegree","Degree"),
pch=1 :2, col=1 :2
)
```

4.6 Code R des modèles de diffusions IC et LT

Dans cette partie de l'annexe, nous avons donné les codes des modèles de propagation IC et LT. Le nombre de balayage est fixé à 4. Nous donnons è à la fin la propagation de l'influence de l'ensemble sk qui représente ici les semences.

4.6.1 Code R de IC

```
tailledebut <- length(sk)
Nbactiver <-0
nbBalayage <-4
l<-0
sktest <- sk
i<-0
s<-sktest
```

```
tete <- 1
while (i!=nbBalayage){
    tailleSK <- length(sktest)
    for(j in tete :tailleSK){
        neig <- neighbors(dolphins,sktest[j])
        tailleneg<-length(neig)
        for(z in 1 :tailleneg)
            if( !(neig[z] %in% s))
                if(prob[sktest[j]]>=prob[neig[z]]){
                    Nbactiver <- Nbactiver+1
                    s <- cbind(s,neig[z])
                }
    } print("iteration")
    print(1+i)
    print("nb actif")
    print(Nbactiver)
    tete <- length(sktest)
    sktest <- s
    i <- i + 1
    print("————————————————————————————————")
}
```

4.6.2 Code R de LT

```
seuil<-0.04
Nbactiver <-0
nbBalayage <- 4
```

```
it <- 0
while (it !=nbBalayage){
        for(i in 1 :length(V(dolphins))){
                if( !(i %in% sk)){
                        neigh<-neighbors(dolphins,i)
                        somseuil<-0
                        for(j in 1 :length(neigh))
                                if(neigh[j] %in% sk) somseuil<-somseuil+prob[neigh[j]]
                            if(somseuil>=seuil){
                                sk <- cbind(sk,neigh[j])
                                Nbactiver <- Nbactiver+1
                            }
                    }
            } print("iteration")
        print(1+it)
        print("nb actif")
        print(Nbactiver)
        it <- it + 1
        print("————————————————————————————————")
}
```

Liste des publications

Conférences internationales

1. ***I. Gaye***, G. Mendy, S. Ouya, and D. Seck. *New Centrality Measure in Social Networks Based on Independent Cascade (IC) Model.* In Proceedings of Future

Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on IEEE, 2015.

2. ***I. Gaye***, G. Mendy, S. Ouya, and D. Seck. *Spanning graph for maximizing the influence spread in Social Networks.* In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (Paris), 2015.
3. ***I. Gaye***, G. Mendy, S. Ouya, I. Diop and D. Seck. *Multi-Diffusion Degree centrality measure to maximize the influence spread in the multilayer social networks.* in proceedings of AFRICOMM 2016.

Chapitre de livres

1. ***I. GAYE***, G. MENDY, S. OUYA, D. SECK, *An Approach to Maximize the Influence Spread in the Social Networks.* In Springer International Publishing AG 2017, Lecture Notes in Social Networks, DOI 10.1007/978-3-319-53420-6_9

Bibliographie

[1] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A : mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[2] Giuseppe Serazzi and Stefano Zanero. Computer virus propagation models. In *Performance Tools and Applications to Networked Systems*, pages 26–50. Springer, 2004.

[3] Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[4] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075) :462–465, 2006.

[5] Wenjun Wang and W Nick Street. A novel algorithm for community detection and influence ranking in social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 555–560. IEEE, 2014.

[6] Chao Li, Jun Luo, Joshua Zhexue Huang, and Jianping Fan. Multi-layer network for influence propagation over microblog. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 60–72. Springer, 2012.

[7] Eriola Kruja, Joe Marks, Ann Blair, and Richard Waters. A short note on

the history of graph drawing. In *International Symposium on Graph Drawing*, pages 272–286. Springer, 2001.

[8] Ibrahima Gaye, Gervais Mendy, Samuel Ouya, and Diaraf Seck. Multi-Diffusion Degree centrality measure to maximize the influence spread in the multilayer social networks. *To appear in the proceedings AFRICOMM 2016*, 2016.

[9] Matteo Magnani and Luca Rossi. The ml-model for multi-layer social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 5–12. IEEE, 2011.

[10] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15. ACM, 2008.

[11] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[12] Nicole B Ellison et al. Social network sites : Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1) :210–230, 2007.

[13] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad. Opening closed regimes : what was the role of social media during the arab spring ? 2011.

[14] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3-4) :248–260, 2009.

[15] BOYD Danah Michele. Friends, freindsters, and myspace top 8 : Writing community into being on social network sites, 2006.

[16] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM, 2008.

[17] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6) :1268–1282, 2013.

[18] Alessio Cardillo, Massimiliano Zanin, Jesús Gómez-Gardenes, Miguel Romance, Alejandro J García del Amo, and Stefano Boccaletti. Modeling the multilayer nature of the european air transport network : Resilience and passengers re-scheduling under random failures. *The European Physical Journal Special Topics*, 215(1) :23–33, 2013.

[19] Matteo Magnani, Barbora Micenkova, and Luca Rossi. Combinatorial analysis of multiple networks. *arXiv preprint arXiv :1303.4986*, 2013.

[20] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31) :13636–13641, 2010.

[21] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3) :203–271, 2014.

[22] Simon R Broadbent and John M Hammersley. Percolation processes : I. crystals and mazes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 53, pages 629–641. Cambridge University Press, 1957.

[23] David Easley and Jon Kleinberg. *Networks, crowds, and markets : Reasoning about a highly connected world*. Cambridge University Press, 2010.

[24] M. E. J. Newman. Networks : An introduction. *Epidemics on Networks*, 1, chapitre 17(50) :627â676, 2010.

[25] Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6) :1420–1443, 1978.

[26] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network : A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3) :211–223, 2001.

[27] Masahiro Kimura and Kazumi Saito. Tractable models for information diffusion in social networks. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 259–271. Springer, 2006.

[28] Cédric Lagnier and Gaussier Eric. Etude de la maximisation de l'influence dans les réseaux sociaux. In *4ième conférence sur les modèles et l'analyse des réseaux : Approches Mathématiques et informatiques*, page 32, 2013.

[29] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11) :888–893, 2010.

[30] Mostafa Salehi, Rajesh Sharma, Moreno Marzolla, Matteo Magnani, Payam Siyari, and Danilo Montesi. Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering*, 2(2) :65–83, 2015.

[31] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[32] Gerard Cornuejols, Marshall L Fisher, and George L Nemhauser. Exceptional paperâlocation of bank accounts to optimize float : An analytic study of exact and approximate algorithms. *Management science*, 23(8) :789–810, 1977.

[33] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsâi. *Mathematical Programming*, 14(1) :265–294, 1978.

[34] George L Nemhauser and Laurence A Wolsey. Integer and combinatorial optimization. interscience series in discrete mathematics and optimization. *ed : John Wiley & Sons*, 1988.

[35] Ding-Zhu Du, Ronald L Graham, Panos M Pardalos, Peng-Jun Wan, Weili Wu, and Wenbo Zhao. Analysis of greedy approximations with nonsubmodular potential functions. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 167–175. Society for Industrial and Applied Mathematics, 2008.

[36] Piotr Bródka, Krzysztof Skibicki, Przemysław Kazienko, and Katarzyna Musiał. A degree centrality in multi-layered social network. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pages 237–242. IEEE, 2011.

[37] Suman Kundu, CA Murthy, and Sankar K Pal. A new centrality measure for influence maximization in social networks. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 242–247. Springer, 2011.

[38] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara, and Angela Ricciardello. A novel measure of edge centrality in social networks. *Knowledge-based systems*, 30 :136–150, 2012.

[39] Dawei Zhao, Lixiang Li, Shudong Li, Yujia Huo, and Yixian Yang. Identifying influential spreaders in interconnected networks. *Physica Scripta*, 89(1) :015203, 2013.

[40] Jordan Viard and Matthieu Latapy. Identifying roles in an ip network with

temporal and structural density. In *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*, pages 801–806. IEEE, 2014.

[41] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1) :107–117, 1998.

[42] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[43] Ramasuri Narayanam and Yadati Narahari. A shapley value-based approach to discover influential nodes in social networks. *IEEE Transactions on Automation Science and Engineering*, 8(1) :130–147, 2011.

[44] B Myerson Roger. Game theory : analysis of conflict, 1991.

[45] Lloyd S Shapley. A value for n-person games. *The Shapley value*, pages 31–40, 1988.

[46] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Simpath : An efficient algorithm for influence maximization under the linear threshold model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 211–220. IEEE, 2011.

[47] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[48] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++ : optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48. ACM, 2011.

[49] Chi Wang, Wei Chen, and Yajun Wang. Scalable influence maximization for independent cascade model in large-scale social networks. *Data Mining and Knowledge Discovery*, 25(3) :545, 2012.

[50] Chuan Zhou, Peng Zhang, Jing Guo, and Li Guo. An upper bound based greedy algorithm for mining top-k influential nodes in social networks. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 421–422. ACM, 2014.

[51] Ibrahima Gaye, Gervais Mendy, Samuel Ouya, and Diaraf Seck. New centrality measure in social networks based on independent cascade (ic) model. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 675–680. IEEE, 2015.

[52] Ibrahima Gaye, Gervais Mendy, Samuel Ouya, and Diaraf Seck. Spanning graph for maximizing the influence spread in social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1389–1394. ACM, 2015.

[53] Ibrahima Gaye, Gervais Mendy, Samuel Ouya, and Diaraf Seck. *An approach to maximize the influence spread in the social networks.* in Springer International Publishing AG 2017, R. Missaoui et al. (eds.), Trends in Social Network Analysis, Lecture Notes in Social Networks, DOI 10.1007/978-3-319-53420-6_9, 2017.

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

[55] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Multidimensional networks : foundations of structural analysis. *World Wide Web*, 16(5-6) :567–593, 2013.

[56] Enrico Bozzo and Massimo Franceschet. Resistance distance, closeness, and betweenness. *Social Networks*, 35(3) :460–469, 2013.

[57] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks : Generalizing degree and shortest paths. *Social networks*, 32(3) :245–251, 2010.

[58] Martin G Everett and Stephen P Borgatti. Induced, endogenous and exogenous centrality. *Social Networks*, 32(4) :339–344, 2010.

[59] Social network analysis : Theory and applications. *http ://code.pediapress.com/*, pages 12 – 42, 2011.

[60] O. J. Boisseau P. Haase E. Slooten D. Lusseau, K. Schneider and S. M. Dawson. Behavioral Ecology and Sociobiology. *Bulletin of advanced technology research*, 54, 2003.