

Thèse présentée pour obtenir le grade de  
**Docteur de l'Université Lumière Lyon 2**

École Doctorale Informatique et Mathématiques (ED 512)

Laboratoire ERIC (EA 3083)

**Discipline : Informatique**

---

## **Diffusion de l'information dans les médias sociaux**

### **Modélisation et analyse**

---

**Par : Adrien Guille**

Présentée et soutenue publiquement le 25 novembre 2014, devant un jury composé de :

<b>Pascal Poncelet</b> , Professeur des Universités, Université Montpellier 2	Rapporteur
<b>Emmanuel Viennet</b> , Professeur des Universités, Université Paris 13	Rapporteur
<b>Vincent Labatut</b> , Maître de Conférences, Université d'Avignon et des Pays du Vaucluse	Examinateur
<b>Christine Largeron</b> , Professeur des Universités, Université Jean Monnet Saint-Etienne	Examinaterice
<b>Cécile Favre</b> , Maître de Conférences, Université Lumière Lyon 2	Co-directrice
<b>Djamel Zighed</b> , Professeur des Universités, Université Lumière Lyon 2	Directeur



# Abstract

Social media have greatly modified the way we produce, diffuse and consume information, and have become powerful information vectors. The goal of this thesis is to help in the understanding of the information diffusion phenomenon in social media by providing means of modeling and analysis.

First, we propose *MABED* (Mention-Anomaly-Based Event Detection), a statistical method for automatically detecting events that most interest social media users from the stream of messages they publish. In contrast with existing methods, it doesn't only focus on the textual content of messages but also leverages the frequency of social interactions that occur between users. *MABED* also differs from the literature in that it dynamically estimates the period of time during which each event is discussed rather than assuming a predefined fixed duration for all events. Secondly, we propose *T-BASIC* (Time-Based ASynchronous Independent Cascades), a probabilistic model based on the network structure underlying social media for predicting information diffusion, more specifically the evolution of the number of users that relay a given piece of information through time. In contrast with similar models that are also based on the network structure, the probability that a piece of information propagate from one user to another isn't fixed but depends on time. We also describe a procedure for inferring the latent parameters of that model, which we formulate as functions of observable characteristics of social media users. Thirdly, we propose *SONDY* (SOcial Network DYnamics), a free and extensible software that implements state-of-the-art methods for mining data generated by social media, *i.e.* the messages published by users and the structure of the social network that interconnects them. As opposed to existing academic tools that either focus on analyzing messages or analyzing the network, *SONDY* permits the joint analysis of these two types of data through the analysis of influence with respect to each detected event.

The experiments, conducted on data collected on Twitter, demonstrate the relevance of our proposals and shed light on some properties that give us a better understanding of the mechanisms underlying information diffusion. First, we compare

---

the performance of *MABED* against those of methods from the literature and find that taking into account the frequency of social interactions between users leads to more accurate event detection and improved robustness in presence of noisy content. We also show that *MABED* helps with the interpretation of detected events by providing clear textual descriptions and precise temporal descriptions. Secondly, we demonstrate the relevancy of the procedure we propose for estimating the pairwise diffusion probabilities on which *T-BASIC* relies. For that, we illustrate the predictive power of users' characteristics, and compare the performance of the method we propose to estimate the diffusion probabilities against those of state-of-the-art methods. We show the importance of having non-constant diffusion probabilities, which allows incorporating the variation of users' level of receptivity through time into *T-BASIC*. We also study how – and in which proportion – the social, topical and temporal characteristics of users impact information diffusion. Thirdly, we illustrate with various scenarios the usefulness of *SONDY*, both for non-experts – thanks to its advanced user interface and adapted visualizations – and for researchers – thanks to its application programming interface.

**Keywords.** Social media data mining ; Event detection and tracking ; Modeling and predicting information diffusion ; Scientific software development.

# Résumé

Les médias sociaux ont largement modifié la manière dont nous produisons, diffusions et consommons l'information et sont de fait devenus des vecteurs d'information importants. L'objectif de cette thèse est d'aider à la compréhension du phénomène de diffusion de l'information dans les médias sociaux, en fournissant des moyens d'analyse et de modélisation.

Premièrement, nous proposons *MABED*, une méthode statistique pour détecter automatiquement les événements importants qui suscitent l'intérêt des utilisateurs des médias sociaux à partir du flux de messages qu'ils publient, dont l'originalité est d'exploiter la fréquence des interactions sociales entre utilisateurs, en plus du contenu textuel des messages. Cette méthode diffère par ailleurs de celles existantes en ce qu'elle estime dynamiquement la durée de chaque événement, plutôt que de supposer une durée commune et fixée à l'avance pour tous les événements. Deuxièmement, nous proposons *T-BASIC*, un modèle probabiliste basé sur la structure de réseau sous-jacente aux médias sociaux pour prédire la diffusion de l'information, plus précisément l'évolution du volume d'utilisateurs relayant une information donnée au fil du temps. Contrairement aux modèles similaires également basés sur la structure du réseau, la probabilité qu'une information donnée se diffuse entre deux utilisateurs n'est pas constante mais dépendante du temps. Nous décrivons aussi une procédure pour l'inférence des paramètres latents du modèle, dont l'originalité est de formuler les paramètres comme des fonctions de caractéristiques observables des utilisateurs. Troisièmement, nous proposons *SONDY*, un logiciel libre et extensible implémentant des méthodes tirées de la littérature pour la fouille et l'analyse des données issues des médias sociaux. Le logiciel manipule deux types de données : les messages publiés par les utilisateurs, et la structure du réseau social interconnectant ces derniers. Contrairement aux logiciels académiques existants qui se concentrent soit sur l'analyse des messages, soit sur l'analyse du réseau, *SONDY* permet d'analyser ces deux types de données conjointement en permettant l'analyse de l'influence par rapport aux événements détectés.

---

Les expérimentations menées à l'aide de divers jeux de données collectés sur le média social Twitter démontrent la pertinence de nos propositions et mettent en lumière des propriétés qui nous aident à mieux comprendre les mécanismes régissant la diffusion de l'information. Premièrement, en comparant les performances de *MABED* avec celles de méthodes récentes tirées de la littérature, nous montrons que la prise en compte des interactions sociales entre utilisateurs conduit à une détection plus précise des évènements importants, avec une robustesse accrue en présence de contenu bruité. Nous montrons également que *MABED* facilite l'interprétation des évènements détectés en fournissant des descriptions claires et précises, tant sur le plan sémantique que temporel. Deuxièmement, nous montrons la validité de la procédure proposée pour estimer les probabilités de diffusion sur lesquelles repose le modèle *T-BASIC*, en illustrant le pouvoir prédictif des caractéristiques des utilisateurs sélectionnées et en comparant les performances de la méthode d'estimation proposée avec celles de méthodes tirées de la littérature. Nous montrons aussi l'intérêt d'avoir des probabilités non constantes, ce qui permet de prendre en compte dans *T-BASIC* la fluctuation du niveau de réceptivité des utilisateurs des médias sociaux au fil du temps. Enfin, nous montrons comment, et dans quelle mesure, les caractéristiques sociales, thématiques et temporelles des utilisateurs affectent la diffusion de l'information. Troisièmement, nous illustrons à l'aide de divers scénarios l'utilité du logiciel *SONDY*, autant pour des non-experts, grâce à son interface utilisateur avancée et des visualisations adaptées, que pour des chercheurs du domaine, grâce à son interface de programmation.

**Mots-clés :** Fouille de données issues des médias sociaux ; Détection et suivi d'évènements ; Modélisation et prévision de la diffusion de l'information ; Développement de logiciel scientifique.

# Remerciements

Je tiens en premier lieu à adresser mes remerciements à Hakim Hacid pour son aide dans la préparation de ma candidature en doctorat, ainsi que pour ses conseils sur la manière de mener mes travaux de recherche.

Je remercie ensuite mes encadrants, Cécile Favre et Djamel Zighed, pour m'avoir permis de réaliser cette thèse au sein du laboratoire ERIC dans les meilleures conditions, et pour m'avoir donné une grande liberté d'action. Je remercie également les nombreux relecteurs anonymes à travers le monde, dont les remarques constructives ont contribué à l'enrichissement des travaux présentés dans ce manuscrit de thèse.

Je tiens aussi à remercier les membres du jury, Pascal Poncelet et Emmanuel Vienet en tant que rapporteurs, ainsi que Vincent Labatut et Christine Largeron en tant qu'examinateurs, pour avoir accepté d'évaluer ces travaux.

Par ailleurs, je remercie les membres du laboratoire ERIC, et tout particulièrement Marian-Andrei Rizoiu et Mathilde Forestier pour l'ambiance de travail très agréable.

Enfin, je tiens à exprimer ma gratitude envers mes parents et mes amis.



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>19</b>
1.1	Problématiques et contributions . . . . .	21
1.1.1	Déetecter les évènements . . . . .	22
1.1.2	Modéliser et prévoir la diffusion de l'information . . . . .	23
1.1.3	Identifier les utilisateurs influents . . . . .	24
1.2	Organisation du manuscrit de thèse . . . . .	25
<b>2</b>	<b>Médias sociaux et diffusion de l'information</b>	<b>29</b>
2.1	Les médias sociaux . . . . .	30
2.1.1	Comparaison avec les médias traditionnels . . . . .	32
2.1.2	Le média social type : Twitter . . . . .	33
2.2	Diffusion de l'information . . . . .	36
2.3	Vue d'ensemble de la recherche sur la diffusion de l'information dans les médias sociaux . . . . .	40
<b>3</b>	<b>Déetecter les évènements</b>	<b>43</b>
3.1	Introduction . . . . .	44
3.2	État de l'art . . . . .	48
3.2.1	Pondération statistique des termes . . . . .	49
3.2.2	Modélisation probabiliste des thématiques latentes . . . . .	52
3.2.3	Classification non supervisée de termes . . . . .	54
3.2.4	Synthèse de l'état de l'art . . . . .	56
3.3	Méthode proposée . . . . .	58
3.3.1	Formulation du problème . . . . .	58
3.3.2	Vue d'ensemble de la méthode proposée . . . . .	59
3.3.3	Détection des évènements à partir de l'anomalie dans la fréquence de création de mentions . . . . .	61
3.3.4	Sélection des mots décrivant les évènements . . . . .	64

3.3.5	Génération de la liste des évènements . . . . .	66
3.3.6	Algorithme général . . . . .	68
3.4	Expérimentations . . . . .	70
3.4.1	Protocole expérimental . . . . .	70
3.4.2	Évaluation quantitative . . . . .	73
3.4.3	Évaluation qualitative . . . . .	76
3.5	Implémentation et visualisations . . . . .	79
3.6	Discussion . . . . .	82
3.6.1	Résumé des travaux présentés . . . . .	83
3.6.2	Perspectives de travail . . . . .	83
<b>4</b>	<b>Modéliser et prévoir la diffusion de l'information</b>	<b>89</b>
4.1	Introduction . . . . .	90
4.2	État de l'art . . . . .	92
4.2.1	Modélisation n'exploitant pas la structure du réseau . . . . .	93
4.2.2	Modélisation basée sur la structure du réseau . . . . .	96
4.2.3	Synthèse de l'état de l'art . . . . .	100
4.3	Méthode proposée . . . . .	102
4.3.1	Formulation du problème . . . . .	102
4.3.2	Vue d'ensemble de la méthode proposée . . . . .	103
4.3.3	Description du modèle . . . . .	105
4.3.4	Espace de représentation . . . . .	107
4.3.5	Estimation des paramètres du modèle . . . . .	109
4.4	Expérimentations . . . . .	117
4.4.1	Protocole expérimental . . . . .	117
4.4.2	Évaluation de la procédure d'estimation des probabilités de diffusion . . . . .	120
4.4.3	Évaluation du modèle <i>T-BASIC</i> . . . . .	122
4.4.4	Analyse des facteurs impactant la diffusion de l'information . . . . .	127
4.5	Discussion . . . . .	131
4.5.1	Résumé des travaux présentés . . . . .	131
4.5.2	Perspectives de travail . . . . .	132

<b>5 Un logiciel libre pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux</b>	<b>135</b>
5.1 Introduction . . . . .	136
5.2 État de l'art . . . . .	139
5.2.1 Détection d'évènements et analyse de l'influence dans les médias sociaux . . . . .	139
5.2.2 Logiciels pour la fouille et l'analyse de données issues des médias sociaux . . . . .	144
5.2.3 Synthèse de l'état de l'art . . . . .	147
5.3 Logiciel proposé . . . . .	148
5.3.1 But du logiciel, publics visés et architecture générale . . . . .	149
5.3.2 Service de manipulation des données . . . . .	152
5.3.3 Service de détection d'évènements . . . . .	154
5.3.4 Service d'analyse du réseau social . . . . .	157
5.3.5 Service d'import d'algorithmes et API . . . . .	160
5.4 Exemples de scénarios d'utilisation . . . . .	161
5.4.1 Utilisation par un non-expert . . . . .	161
5.4.2 Utilisation par un chercheur du domaine . . . . .	165
5.5 Discussion . . . . .	170
<b>6 Conclusion</b>	<b>173</b>
6.1 Résumé de la thèse . . . . .	173
6.2 Perspectives de travail . . . . .	175
<b>Bibliographie</b>	<b>177</b>
<b>Annexes</b>	<b>187</b>
Liste des publications . . . . .	188
Revue internationale . . . . .	188
Conférence internationale et atelier international . . . . .	188
Conférence nationale . . . . .	188



# Table des figures

1.1	Logos de certains des médias sociaux les plus populaires. . . . .	20
1.2	Déroulement typique du processus de fouille de données. . . . .	21
1.3	Structuration des travaux de thèse. De haut en bas : le phénomène étudié, les problématiques de recherche et les contributions apportées. . . . .	26
2.1	Un média social fictif, utilisé par 5 personnes ayant publié 8 messages. . . . .	31
2.2	Le flux de messages généré par les utilisateurs du média social fictif représenté par la figure 2.1. . . . .	32
2.3	Chiffres clés résumant l'activité générée par Twitter en 2012. . . . .	33
2.4	Interfaces du média social Twitter. . . . .	35
2.5	Un média social fictif représenté par un graphe, dans lequel les nœuds colorés en gris foncé (u1, u2, u4) ont publié chacun un message (m1, m2, m4) à propos d'une même thématique. . . . .	39
2.6	Taxonomie structurant les principales pistes de recherche explorées dans les travaux portant sur la diffusion de l'information dans les médias sociaux. . . . .	41
3.1	De gauche à droite, les « trends » détectés par Twitter le 8 juillet 2014 pour la France, les États-Unis et le monde entier. . . . .	45
3.2	Dynamique temporelle des thématiques saillantes correspondant aux discussions engendrées par 7 événements dans un média social fictif. .	46
3.3	Un flux de messages continu et le même flux de messages dont l'axe temporel a été discréétisé. . . . .	49
3.4	Fréquence des mots pour un flux de messages fictif. . . . .	51
3.5	Représentation en plaques du modèle LDA. . . . .	53
3.6	Déroulement de la méthode proposée, <i>MABED</i> . . . . .	60
3.7	Identification d'un événement lié au mot « kadhafi ». L'aire algébrique sous la fonction d'anomalie correspond aux zones grises. . . . .	63

3.8	Un graphe des événements stockant deux événements. Les mots principaux sont représentés par les noeuds blancs avec des bordures noires.	67
3.9	Identification de la redondance entre deux événements fictifs $e_0$ et $e_1$ .	68
3.10	Résultat de la fusion entre les événements fictifs $e_0$ et $e_1$ .	70
3.11	Précision, F-mesure et <i>DERate</i> de <i>MABED</i> sur le corpus $\mathcal{C}_{en}$ pour différentes valeurs de $\sigma$ .	76
3.12	Anomalie mesurée pour les mots « hood », « fort » et « shooting » du 5 au 7 novembre à minuit (CST).	79
3.13	Distribution de la durée des événements détectés par <i>MABED</i> .	80
3.14	Capture d'écran de l'interface centrée-temps. Elle se décompose en deux parties : la partie inférieure permet la navigation à travers le temps tandis que la partie supérieure donne des détails à propos des événements.	81
3.15	Capture d'écran de l'interface centrée-impact. Chaque événement est associé à une couleur.	81
3.16	Capture d'écran de l'interface centrée-thématique. Les noeuds gris correspondent aux mots principaux. Leur diamètre est proportionnel à l'impact des événements.	82
3.17	Distribution du poids des catégories des événements détectés par <i>MABED</i> dans les corpus $\mathcal{C}_{en}(c_0)$ , $\mathcal{C}_{en}(c_1)$ , $\mathcal{C}_{en}$ et $\mathcal{C}_{en}$ (aléatoire)	87
4.1	Représentations graphiques des modèles épidémiologiques <i>SI</i> et <i>SIR</i> .	94
4.2	Allure typique des courbes de diffusion obtenues avec le modèle <i>SIR</i> .	95
4.3	Illustration du fonctionnement du <i>Linear Influence Model</i> . Le volume de messages publiés au fil du temps est obtenu en sommant les fonctions d'influence des utilisateurs initialement actifs : u1, u2 et u3.	96
4.4	Un processus de diffusion modélisé selon le <i>Independent Cascades Model (IC)</i> . À gauche : un extrait du réseau servant de support à la diffusion, annoté avec les probabilités de diffusion pour chaque arc visible. À droite, le processus de diffusion initié par les deux noeuds colorés en gris foncé.	98
4.5	Prévision de la diffusion d'une information à l'aide du modèle <i>T-BASIC</i> .	104

4.6 Illustration du processus de construction du jeu de données d'entraînement. La structure représentée correspond au graphe $G$ , un arc ( $ux \rightarrow uy$ ) signifie donc que l'utilisateur $ux$ est exposé aux messages publiés par $uy$ . . . . .	112
4.7 Représentation du classifieur linéaire construit à partir de la fonction $f$ . . . . .	116
4.8 Visualisation sous forme de graphe des instances de $D$ ayant pour modalité $y_i = 1$ extraites à partir d'un même évènement. . . . .	118
4.9 Séries temporelles réelles et prédictives représentant l'évolution du volume d'utilisateurs influencés, pour trois processus de diffusion représentatifs. Les points symbolisés par des carrés correspondent aux séries temporelles réelles, tandis que ceux symbolisés par des cercles correspondent aux séries temporelles prédictives. . . . .	124
4.10 Rapports de cotes pour différents attributs, mesurés par rapport aux utilisateurs $ux$ et $uy$ . La direction des barres traduit la direction de la relation entre chaque attribut et la probabilité de diffusion : vers la gauche, l'effet est négatif, vers la droite, l'effet est positif. . . . .	128
4.11 Distribution des valeurs de $Re(ux)$ en fonction de la modalité prise par $y_i$ (1 : diffusion, 0 : non-diffusion). . . . .	130
5.1 Comportement typique de l'indicateur <i>MACD</i> . . . . .	140
5.2 Décomposition en trois enveloppes d'un réseau comportant 11 membres.	143
5.3 Interfaces des logiciels <i>SAP Social Media Analytics</i> (a) et <i>BrandWatch Analytics</i> (b). . . . .	145
5.4 Interface utilisateur du logiciel Gephi pour la fouille de graphe. . . . .	146
5.5 Interfaces du logiciel <i>SONDY</i> : manipulation des données (a), détection et visualisation des évènements (b), analyse et visualisation du réseau social (c) et import de nouveaux algorithmes (d). . . . .	150
5.6 Positionnement des services du logiciel <i>SONDY</i> dans le processus typique de fouille de données. . . . .	151
5.7 Architecture du logiciel <i>SONDY</i> . . . . .	151

---

5.8 La principale fenêtre correspond à l'interface du service de manipulation des données. Les deux petites fenêtres (b') montrent des extraits de fichiers CSV (à gauche, les messages, à droite le réseau social) pouvant être importés par SONDY. . . . .	153
5.9 La principale fenêtre correspond au cœur de l'interface du service de détection d'évènements. Les deux autres fenêtres correspondent respectivement, de haut en bas, à la fenêtre pour l'exploration des messages (d') et à la frise chronologique des évènements détectés (f') . . . . .	155
5.10 L'interface principale du service d'analyse du réseau permet de naviguer dans le réseau social coloré et de consulter la distribution des rangs identifiés par les algorithmes. La seconde fenêtre (d') permet de naviguer parmi les messages publiés par les utilisateurs membres du réseau. . . . .	158
5.11 Quatre des étapes d'une séquence d'activation capturées à partir de l'interface du service d'analyse du réseau social. . . . .	159
5.12 Exploration des évènements en lien avec Google : (a) sélection des données à analyser, (b) détection des évènements et (c) frise chronologique.	162
5.13 Exploration des évènements en lien avec Google : (a,b) courbes de fréquence des évènements et (c,d) messages liés. . . . .	163
5.14 Identification d'utilisateurs influents à propos de l'évènement le plus marquant concernant Google. . . . .	164
5.15 Fenêtre de log retracant les opérations effectuées. . . . .	165
5.16 Détection d'évènements avec différents algorithmes : <i>Peaky Topics</i> , <i>Trending Score</i> , <i>Persistent Conversations</i> et <i>EDCoW</i> . . . . .	166
5.17 Résultats obtenus par la méthode <i>Trending Score</i> pour différentes préparations d'un même jeu de données. L'intervalle temporel est défini en jours, le début de la période couverte par le jeu de données étant associé au jour 0. . . . .	168
5.18 Analyse de l'influence au sein d'un réseau social à l'aide de différentes méthodes. . . . .	169

# Liste des tableaux

3.1	Matrice de comparaison des méthodes existantes pour la détection d'évènements. . . . .	56
3.2	Matrice de comparaison des méthodes existantes pour la détection d'évènements. . . . .	56
3.3	Liste des notations utilisées dans le chapitre 3. . . . .	58
3.4	Statistiques sur les corpus (@ : proportion de tweets qui contiennent des mentions, $RT$ : proportion de retweets). . . . .	71
3.5	Performances des cinq méthodes sur les deux corpus. . . . .	75
3.6	Liste des 20 évènements ayant eu le plus fort impact sur les utilisateurs, détectés par <i>MABED</i> à partir du corpus $\mathcal{C}_{en}$ . Les mots principaux sont en gras et le poids de chaque mot lié est donné entre parenthèses. Les intervalles temporels sont exprimés en temps UTC. . . . .	77
4.1	Matrice de comparaison des modèles existants pour la prévision de la diffusion de l'information dans les médias sociaux. . . . .	101
4.2	Liste des notations utilisées dans le chapitre 4. . . . .	103
4.3	Instanciation possible d'un vecteur $v_{ux,uy}^t$ . . . . .	110
4.4	Performances des six classifieurs sur le jeu de données $D_{test}$ . . . . .	122
4.5	Erreur mesurée pour les trois méthodes. . . . .	125
4.6	Effet des caractéristiques des utilisateurs sur la probabilité de diffusion : la couleur orange traduit un effet positif tandis que la couleur grise traduit un effet négatif, l'intensité de la couleur traduisant l'importance de l'effet mesurée selon le log-odds-ratio. . . . .	130
5.1	Matrice synthétisant les fonctionnalités des logiciels développés dans le milieu académique pour les tâches de détection d'évènements et l'analyse de l'influence. . . . .	147



# Chapitre 1

## Introduction

Le phénomène de diffusion est observé et étudié depuis longtemps dans de nombreux domaines de la science : propagation des maladies (*Kermack et McKendrick*, 1927; *Hethcote*, 2000) ou des virus informatiques (*Serazzi et Zanero*, 2004), diffusion des innovations technologiques (*Rogers*, 1995), déplacements humains (*Brockmann et al.*, 2006), etc. En particulier, le phénomène de diffusion de l'information – que l'on définit comme l'action de propager des éléments d'information auprès d'un public – suscite depuis plusieurs années un intérêt accru au sein de la communauté scientifique.

En effet, les moyens de télécommunication modernes, notamment Internet, ont largement modifié la manière dont nous produisons, diffusons et consommons l'information. Les médias sociaux – tels que Twitter, Facebook ou Google+, dont les logos sont repris au centre de la figure 1.1 – représentent une part importante du paysage actuel du Web, l'une des applications majeures d'Internet. Ils sont utilisés par des centaines de millions de personnes à travers le monde pour se connecter avec d'autres utilisateurs et communiquer. Ces personnes créent et partagent librement de l'information liée à divers types d'évènements – allant d'évènements personnels banals à des évènements importants et/ou globaux – en publiant ou retransmettant des messages, et ce en temps réel. Les réseaux sociaux sous-jacents servent de supports à des phénomènes de diffusion d'information à grande échelle. *Bakshy et al.* (2012) observent d'ailleurs que ces réseaux favorisent la propagation d'idées nouvelles et de points de vue différents. Leur efficacité en tant que vecteurs d'information ainsi que leur capacité à influencer la société ont été étudiées et soulignées à plusieurs reprises, par exemple lors des élections présidentielles américaines de 2008 (*Hughes et Palen*, 2009), ou encore durant le « Printemps Arabe » initié fin 2010 (*Howard et Duffy*,



FIGURE 1.1 – Logos de certains des médias sociaux les plus populaires.

2011).

Les utilisateurs des médias sociaux étant à la fois producteurs et consommateurs d'information, l'augmentation continue de leur nombre s'accompagne d'une augmentation continue du volume de messages publiés. Le constat fait il y a plus de 40 ans par *Simon* (1971) est ainsi plus que jamais d'actualité : « *A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it* ». Autrement dit, le principal facteur limitant la diffusion de l'information aujourd'hui n'est plus la disponibilité de l'information, mais la disponibilité de l'attention de ses potentiels récepteurs. Face à cette situation de surcharge informationnelle, parfois appelée infobésité, il devient nécessaire de développer des outils favorisant une distribution et une consommation efficace de l'information véhiculée par les médias sociaux.

L'informatique, en tant que science du traitement des informations, est une des nombreuses branches de la science – telles que la sociologie, la psychologie cognitive ou encore les sciences de la communication et de l'information – qui cherchent à apporter des solutions à ce problème. Devant l'abondance des données produites par les

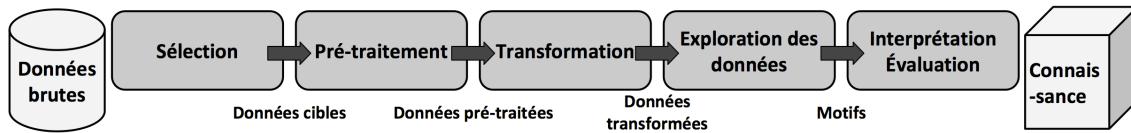


FIGURE 1.2 – Déroulement typique du processus de fouille de données.

médias sociaux, un domaine de l'informatique s'y intéresse particulièrement : celui de la fouille de données. La fouille de données – aussi appelée exploration de données ou « *data mining* » en anglais – consiste à collecter des données, puis à les pré-traiter et à les transformer dans le but d'y appliquer des algorithmes pour les analyser. L'objectif de ces algorithmes est d'identifier des motifs latents, dont l'interprétation et la visualisation permettent d'extraire des connaissances utiles à la résolution des problèmes liés aux données analysées. La figure 1.2 décrit les principales étapes du processus de fouille de données – des données brutes jusqu'à la connaissance.

La fouille des données générées par les médias sociaux peut, parce qu'elle nous aide à mieux comprendre la dynamique humaine (Aggarwal, 2011), nous aider à développer des outils favorisant une distribution et une consommation plus efficace de l'information. Certains travaux menés dans le domaine s'intéressent par exemple à l'identification de groupes d'utilisateurs similaires en analysant la structure du réseau social les interconnectant (Combe et al., 2013; Stattner et Collard, 2014). Connaissant les communautés d'utilisateurs, il est par exemple possible de leur recommander des éléments d'information susceptibles de les intéresser. D'autres travaux s'intéressent à la prédiction de liens entre utilisateurs (Backstrom et Leskovec, 2011) voire la prédiction de la désinscription d'utilisateurs (Ngonmang et al., 2014), ainsi que de nombreuses autres problématiques (Aggarwal, 2011; Can et al., 2014).

## 1.1 Problématiques et contributions

Dans cette thèse, nous nous intéressons à trois problématiques qui découlent du phénomène de diffusion de l'information dans les médias sociaux (Guille, 2013; Guille et al., 2013c) : (i) détecter les événements importants qui suscitent l'intérêt des utilisateurs, (ii) modéliser et prévoir la diffusion de l'information, et (iii) identifier des

utilisateurs influençant la diffusion de l'information. Les contributions apportées par nos travaux sont les suivantes :

- *MABED* : une méthode statistique pour la détection et le suivi des évènements dans les médias sociaux.
- *T-BASIC* : un modèle probabiliste pour prévoir la diffusion de l'information dans les médias sociaux.
- *SONDY* : un logiciel implémentant des méthodes de la littérature pour la détection d'évènements et l'identification d'utilisateurs influents.

Ces contributions, que nous positionnons et décrivons brièvement dans la suite de cette section, sont formulées de manière générique par rapport aux médias sociaux, et évaluées par rapport à un média social en particulier, à savoir Twitter.

### 1.1.1 Déetecter les évènements

**Problématique.** Les utilisateurs des médias sociaux partagent, discutent et retransmettent de l'information à propos d'évènements divers – allant d'évènements personnels et/ou banals à des évènements importants et/ou globaux – en temps réel. Le volume sans-cesse croissant de messages publiés sur les médias sociaux fait de ces derniers des sources d'information à la fois riches et réactives, ce avec quoi les médias traditionnels peuvent difficilement rivaliser. Toutefois, l'augmentation de ce volume engendre un phénomène de surcharge informationnelle et il est de plus en plus difficile d'identifier des éléments d'information pertinents liés à des évènements importants. Par « évènement important » nous entendons ici un évènement réel et susceptible d'être couvert par les médias traditionnels. Cela nous amène à formuler la question suivante : *comment peut-on exploiter les médias sociaux pour détecter automatiquement les évènements importants ?* Répondre à cette question permettrait notamment d'analyser les évènements, ou les types d'évènements, qui suscitent le plus l'intérêt des utilisateurs des médias sociaux – ce qui serait utile dans le cadre de la veille d'information, du journalisme de données, du marketing, etc.

**Contribution.** Détecter automatiquement les évènements importants à partir des médias sociaux est une tâche complexe, puisque les messages se rapportant à ces évènements sont noyés dans un grand volume de messages sans rapport (*i.e.* du bruit). Nous proposons *MABED* (*Mention-Anomaly-Based Event Detection*, (*Guille et*

(*Favre, 2014a,b*)), une méthode statistique pour détecter automatiquement les événements importants qui suscitent l'intérêt des utilisateurs des médias sociaux à partir du flux de messages qu'ils publient, dont l'originalité est d'exploiter la fréquence des interactions sociales entre utilisateurs, en plus du contenu textuel des messages. La méthode *MABED* diffère par ailleurs des méthodes existantes en ce qu'elle estime dynamiquement la durée de chaque évènement, plutôt que de supposer une durée commune et fixée à l'avance pour tous les évènements. Les expérimentations menées montrent la pertinence de la méthode proposée. Notamment, en comparant les performances de *MABED* avec celles de méthodes récentes tirées de la littérature, nous montrons que la prise en compte des interactions sociales entre utilisateurs conduit à une détection plus précise des évènements importants, avec une robustesse accrue en présence de contenu bruité. Nous montrons également que *MABED* facilite l'interprétation des évènements détectés en fournissant des descriptions claires et précises, tant sur le plan sémantique que temporel.

### 1.1.2 Modéliser et prévoir la diffusion de l'information

**Problématique.** Au-delà de la détection *a posteriori* des évènements ayant suscité l'intérêt des utilisateurs d'un média social, il est également utile, dans certains cas, de pouvoir anticiper la réaction des utilisateurs à un évènement spécifique, *i.e.* anticiper la diffusion de l'information liée à cet évènement. La prédiction du phénomène de diffusion de l'information dans les médias sociaux est une tâche qui suscite un fort intérêt de la part des chercheurs en fouille de données, cependant, la manière dont cette tâche est abordée dans la littérature fait que nous en savons encore peu à propos des facteurs qui sous-tendent le processus de diffusion au sein des médias sociaux. Cela nous amène donc à formuler les questions suivantes. D'une part, *quels facteurs influent sur la diffusion de l'information dans les médias sociaux ?* D'autre part, *comment prévoir la diffusion de l'information à partir de ces facteurs ?* Répondre à ces questions nous permettrait de mieux comprendre le phénomène de diffusion dans les médias sociaux et nous permettrait de mieux l'anticiper – ce qui serait utile dans le cadre du marketing viral, de la communication de crise, *etc.*

**Contribution.** Modéliser et prévoir le phénomène de diffusion de l'information à travers les médias sociaux est une tâche ardue en raison de l'intrication entre les dy-

namiques humaines et les structures sociales de grande ampleur. Nous proposons *T-BASIC* (*Time-Based ASynchronous Independent Cascades*, (Guille et Hacid, 2012; Guille et al., 2012)), un modèle probabiliste basé sur la structure de réseau sous-jacente aux médias sociaux pour prévoir la diffusion de l'information, plus précisément l'évolution du volume d'utilisateurs relayant une information donnée au fil du temps. Contrairement aux modèles similaires et également basés sur la structure du réseau, la probabilité qu'une information donnée se diffuse entre deux utilisateurs connectés n'est pas constante mais dépendante du temps. Nous décrivons aussi une procédure pour l'inférence des paramètres latents du modèle (probabilité de diffusion et délai de transmission pour chaque lien du réseau), dont l'originalité est de formuler les paramètres comme des fonctions de caractéristiques observables des utilisateurs. Les expérimentations menées montrent la validité de la procédure d'estimation des paramètres, ainsi que l'intérêt d'avoir des probabilités dépendantes du temps, ce qui permet de prendre en compte dans *T-BASIC* la fluctuation du niveau de réceptivité des utilisateurs des médias sociaux au fil du temps. Par ailleurs, nous montrons comment, et dans quelle mesure, les caractéristiques sociales, thématiques et temporelles des utilisateurs affectent la diffusion de l'information.

### 1.1.3 Identifier les utilisateurs influents

**Problématique.** De nombreux acteurs de la société (e.g. les entreprises, les services gouvernementaux, les journalistes) cherchent à exploiter et analyser les médias sociaux à des fins diverses (e.g. analyser la réaction des consommateurs à propos de certains produits et les promouvoir, détecter des informations et utilisateurs dangereux, détecter des événements importants et interroger les utilisateurs). Généralement, la démarche qu'ils cherchent à mettre en œuvre consiste à détecter les événements animant les discussions des utilisateurs, puis à identifier les utilisateurs influents par rapport à ces événements, afin de prendre des décisions et éventuellement agir. Pour que cette démarche soit efficace, elle doit reposer sur des méthodes de détection d'événements et d'analyse de l'influence adaptées au contexte des médias sociaux. Néanmoins, nous constatons que les chercheurs qui développent de telles méthodes ne partagent pas systématiquement leurs implémentations. Cela nous amène à formuler les deux questions suivantes. D'une part, *comment permettre à des non-*

*experts d'analyser efficacement des données collectées sur les médias sociaux ? D'autre part, comment favoriser le partage et la réutilisation des implémentations des méthodes nécessaires à cette analyse ?* Les réponses à ces questions bénéficieraient autant aux non-experts ayant besoin d'analyser les données dont ils disposent, qu'aux chercheurs, en leur permettant d'une part de partager leurs nouvelles méthodes et d'autre part en leur permettant de réutiliser les méthodes existantes.

**Contribution.** Nous proposons *SONDY* (*SOcial Network DYnamics*, (Guille et al., 2013a,b)), un logiciel libre et extensible qui implémente des méthodes tirées de la littérature pour la fouille et l'analyse des données issues des médias sociaux. Le logiciel manipule deux types de données : les messages publiés par les utilisateurs, et la structure du réseau social interconnectant ces derniers. Contrairement aux logiciels académiques existants qui se concentrent soit sur l'analyse des messages, soit sur l'analyse du réseau, *SONDY* permet d'analyser ces deux types de données conjointement en permettant l'analyse de l'influence par rapport aux évènements détectés. Utilisé comme logiciel autonome, *SONDY* offre une interface utilisateur avancée, accessible aux non-experts, et des visualisations adaptées. Utilisé comme bibliothèque, il permet d'intégrer facilement les méthodes implémentées dans d'autres programmes, par exemple pour automatiser la comparaison de leurs performances. Les expérimentations menées illustrent, à l'aide de scénarios concrets, l'intérêt de *SONDY* pour les deux publics qu'il vise.

## 1.2 Organisation du manuscrit de thèse

La figure 1.3 présente l'organisation des travaux de thèse par rapport au phénomène étudié. La suite de ce manuscrit est organisée comme suit. Le chapitre 2 aborde des notions générales à propos des médias sociaux, de la diffusion de l'information et de la recherche menée sur ces sujets dans le domaine de la fouille de données. Nous consacrons ensuite un chapitre à chacune des problématiques évoquées précédemment. À chaque fois, nous développons l'état de l'art avant de présenter notre contribution. Le chapitre 3 présente une première contribution, qui est une méthode pour la détection des évènements suscitant l'intérêt des utilisateurs des médias sociaux, *MABED*. Le chapitre 4 est consacré à la modélisation du phénomène de diffusion de l'information dans les médias sociaux et nous y présentons une seconde contribution,

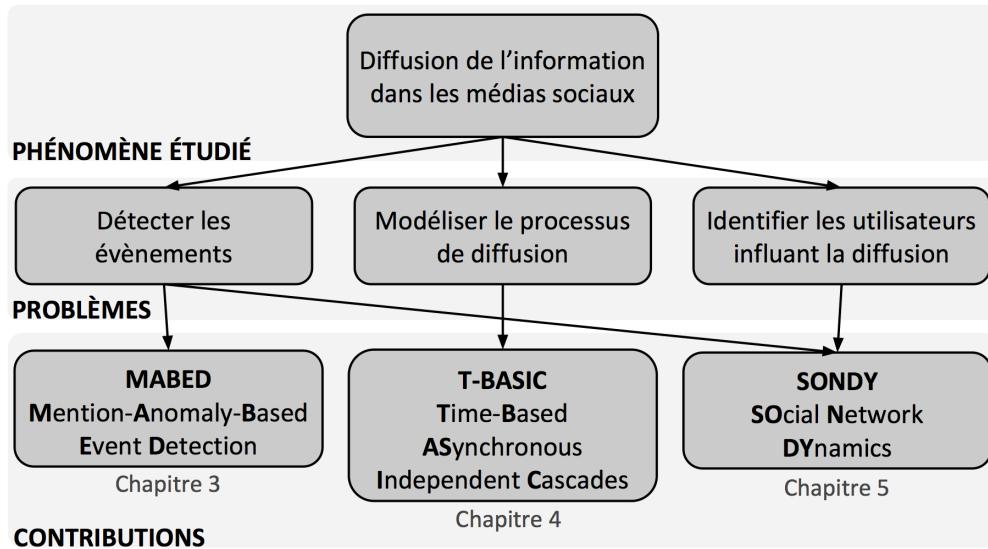


FIGURE 1.3 – Structuration des travaux de thèse. De haut en bas : le phénomène étudié, les problématiques de recherche et les contributions apportées.

qui consiste en un modèle prédictif que nous nommons *T-BASIC*. Dans le chapitre 5 nous présentons le logiciel libre que nous développons pour la fouille des données issues des médias sociaux, *SONDY*, qui permet notamment d'identifier les membres jouant des rôles importants dans la diffusion de l'information dans les médias sociaux. Enfin, le chapitre 6 conclut ce manuscrit de thèse.

Le présent manuscrit de thèse couvre du contenu tiré des publications internationales suivantes :

1. A. Guille et C. Favre. Mention-anomaly-based Event Detection and Tracking in Twitter. Full paper (18% acceptance rate).  
*ASONAM '14 – Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, 2014.
2. A. Guille, C. Favre, H. Hacid et D. Zighed. SONDY : an Open Source Platform for Social Dynamics Mining and Analysis. Demonstration paper (35% acceptance rate).  
*SIGMOD '13 – Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2013.

3. A. Guille. Information Diffusion in Online Social Networks. Full paper (46% acceptance rate).  
*SIGMOD/PODS Ph.D. Symposium '13 – Proceedings of the SIGMOD/PODS Ph.D. symposium*, 2013.
4. A. Guille, H. Hacid, C. Favre and D. Zighed. Information Diffusion in Online Social Networks : A Survey. Journal article (impact factor 2013 : 0.955).  
*ACM SIGMOD Record* – Volume 42, Number 2, 2013.
5. A. Guille et H. Hacid. A Predictive Model for the Temporal Dynamics of Information Diffusion in Online Social Networks. Full paper (57% acceptance rate).  
*WWW '12 Companion – Proceedings of the International Conference Companion on World Wide Web : Workshop on Mining Social Network Dynamics*, 2012.

Ces publications internationales ont accumulé, d'après Google Scholar<sup>1</sup>, plus de 100 citations au moment de la rédaction de ce manuscrit. Les travaux décrits dans ces publications internationales ont également fait l'objet de publications, présentations et démonstrations à la conférence francophone EGC (Extraction et Gestion des Connaissances) lors des éditions 2012, 2013 et 2014. Une liste exhaustive des publications liées à cette thèse est donnée en annexes (page 188).

---

1. Profil Google Scholar consultable à l'adresse : [http://scholar.google.fr/citations?user=mM\\_oO18AAAAJ](http://scholar.google.fr/citations?user=mM_oO18AAAAJ)



# Chapitre 2

## Médias sociaux et diffusion de l'information

### Sommaire

---

2.1	Les médias sociaux . . . . .	30
2.1.1	Comparaison avec les médias traditionnels . . . . .	32
2.1.2	Le média social type : Twitter . . . . .	33
2.2	Diffusion de l'information . . . . .	36
2.3	Vue d'ensemble de la recherche sur la diffusion de l'information dans les médias sociaux . . . . .	40

---

Dans la première partie de ce chapitre nous détaillons la notion de média social, nous définissons la structure commune aux médias sociaux puis nous précisons leurs spécificités – notamment par rapport aux médias traditionnels. Dans la seconde partie de ce chapitre, nous abordons les notions et théories de base à propos de la diffusion de l'information. Pour cela, nous nous appuyons sur des recherches issues du domaine de l'analyse des médias sociaux, mais également des travaux menés en sociologie et en économie. Enfin, la troisième partie de ce chapitre donne un bref tour d'horizon des travaux de recherche récents à propos de la modélisation et l'analyse de la diffusion de l'information dans les médias sociaux.

## 2.1 Les médias sociaux

Dans cette thèse, nous définissons un média social comme un service en ligne qui permet essentiellement deux choses à ses utilisateurs :

- Premièrement, ceux-ci créent une page de profil sur laquelle ils peuvent publier des messages.
- Deuxièmement, ils se connectent à d'autres utilisateurs afin de suivre leurs publications.

Cette définition générale est similaire à celle proposée par *Boyd et Ellison* (2007) et couvre un grand nombre de services web qui, tout en respectant cette définition, présentent chacun des spécificités. Les différents médias sociaux se distinguent d'une part en fonction de la visibilité et de l'accessibilité des pages de profil de leurs utilisateurs. Par exemple, 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. Les profils créés sur Facebook sont au contraire privés sauf si son créateur en décide autrement. Les différents médias sociaux se distinguent d'autre part selon la manière dont leurs utilisateurs se connectent entre eux. 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 (« following » en anglais) et permet à la personne ayant initié la connexion de recevoir automatiquement les messages publiés par l'utilisateur ciblé. D'autres médias sociaux – comme Facebook – se basent sur un mode de connexion bilatéral, ce qui signifie que les deux utilisateurs doivent autoriser

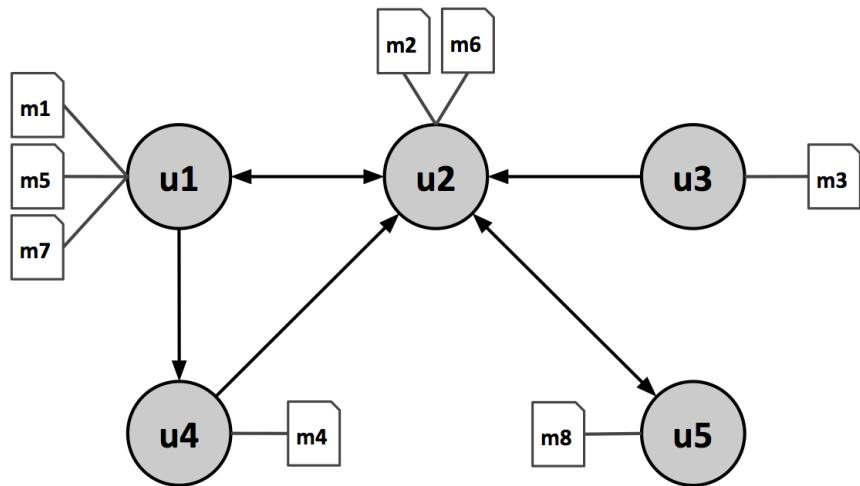


FIGURE 2.1 – Un média social fictif, utilisé par 5 personnes ayant publié 8 messages.

la création du lien. L'information circule alors dans les deux sens et ce lien est appelé « lien d'amitié ». Le terme d'amitié est employé par de nombreux médias sociaux pour désigner les connexions entre utilisateurs. Néanmoins, comme l'observe Boyd (2006), 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.

Formellement, un média social est représenté par un graphe étiqueté, où les nœuds correspondent aux utilisateurs du service et où les liens représentent les connexions entre utilisateurs. Ce graphe peut être orienté ou non, selon les spécificités du média social considéré (*i.e.* selon que le mode de connexion soit unilatéral ou bilatéral). Les sommets sont étiquetés avec les messages publiés par l'utilisateur correspondant. Un message est décrit par (i) son auteur, (ii) son contenu et (iii) sa date de publication. La figure 2.1 présente un média social fictif reposant sur un mode de connexion unilatéral utilisé par cinq personnes ( $u_1, \dots, u_5$ ) ayant publié 8 messages ( $m_1, \dots, m_8$ ). Sur cette figure, un arc ( $u_x \rightarrow u_y$ ) signifie que l'utilisateur  $u_x$  s'est connecté à l'utilisateur  $u_y$  et reçoit donc automatiquement les messages de ce dernier. Cette représentation révèle par exemple que l'utilisateur  $u_1$  est exposé aux messages publiés par  $u_2$  et  $u_4$ , tandis que personne ne reçoit les messages publiés par l'utilisateur  $u_3$ .

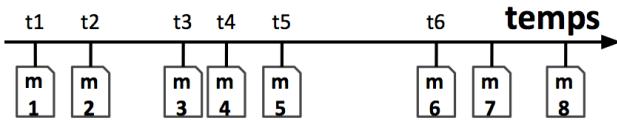


FIGURE 2.2 – Le flux de messages généré par les utilisateurs du média social fictif représenté par la figure 2.1.

Lorsque l'on s'intéresse à la publication de messages non plus au niveau individuel mais au niveau d'un média social dans son ensemble, on observe un flux continu de messages. La figure 2.2 représente le flux de messages généré par les cinq utilisateurs du média social fictif décrit par la figure 2.1. Chaque utilisateur d'un média social est exposé à une part plus ou moins importante du flux total, en fonction des connexions qu'il a établies avec les autres utilisateurs.

### 2.1.1 Comparaison avec les médias traditionnels

Les médias sociaux diffèrent des médias traditionnels en de nombreux points, que nous listons ci-après.

- **Expressivité.** En contraste avec les articles publiés par les médias traditionnels, les messages publiés par les utilisateurs des médias sociaux sont courts, voire parfois très courts. La longueur des messages est parfois limitée par le service.
- **Volume.** Là où les médias traditionnels reposent sur un petit nombre de contributeurs, chaque média social compte un grand nombre d'utilisateurs, chacun d'entre eux publant plus ou moins régulièrement des messages.
- **Hétérogénéité.** Les messages publiés abordent des sujets divers et variés, allant d'évènements banals de la vie quotidienne à des évènements importants et/ou globaux. Qui plus est, contrairement aux médias traditionnels, les messages ne sont pas catégorisés, ni structurés.
- **Rapidité.** Enfin, la grande force des médias sociaux par rapport aux médias traditionnels est l'immédiateté de la publication. Il n'y a en effet (en principe) aucun filtrage sur le contenu publié. Cette rapidité est amplifiée par le fait que pour la plupart des médias sociaux, une part importante des utilisateurs accèdent au service depuis un terminal mobile – l'accès mobile permettant de

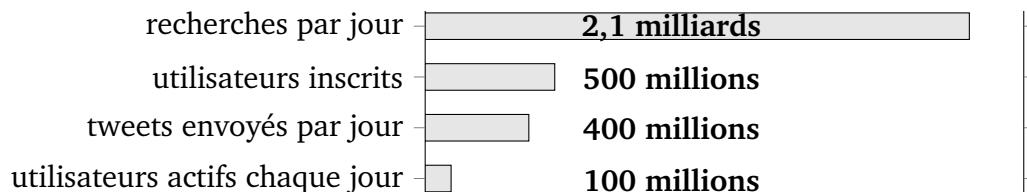


FIGURE 2.3 – Chiffres clés résumant l'activité générée par Twitter en 2012.

publier des messages à tout moment.

Nous illustrons ces particularités dans la sous-section suivante.

### 2.1.2 Le média social type : Twitter

Bien que les contributions apportées par ces travaux de thèse – que ce soit sous forme d'algorithme, de modèle ou de logiciel – soient applicables à la plupart des médias sociaux, nous avons choisi de mener nos expérimentations sur un média social en particulier : Twitter. Deux raisons principales motivent ce choix. Premièrement, l'engouement pour Twitter est un phénomène global qui pousse chaque jour des internautes du monde entier à s'inscrire puis prendre part aux discussions. Par conséquent, Twitter occupe une place sans cesse plus importante dans notre environnement médiatique. Il est de fait devenu un outil de communication prisé de beaucoup de journalistes, acteurs de la vie politique ou encore entreprises. L'étude menée par *Hughes et Palen* (2009) révèle par exemple le rôle important de Twitter au sein de la stratégie de communication adoptée par Barack Obama durant la campagne présidentielle de 2008 aux États-Unis. L'étude conduite par *SimplyMeasured* (2014) montre quant à elle que les petites comme les grandes entreprises intègrent Twitter dans leurs plans de communication. Notamment, elle indique que les 100 plus grandes compagnies selon le classement InterBrand<sup>1</sup> ont publié en 2013 en moyenne chacune 12 tweets par jour. Deuxièmement, Twitter – contrairement à la majorité des médias sociaux – permet d'accéder gratuitement à une part importante de ses données, ce qui pousse beaucoup de chercheurs à l'étudier.

1. Liste consultable à l'adresse : <http://www.interbrand.com/en/best-global-brands/2013/Best-Global-Brands-2013.aspx>

**Spécificités.** Twitter<sup>2</sup> est l'un des médias sociaux les plus populaires, lancé en juillet 2006 aux États-Unis. Ses créateurs le définissent ainsi : « Twitter offre à chacun l'opportunité de créer et de partager instantanément des idées et des informations, sans aucune barrière<sup>3</sup> ». Les utilisateurs inscrits publient des messages limités à 140 caractères, appelés « tweets », et se connectent entre eux de manière unilatérale selon le principe d'abonnement (*i.e. following*). Le réseau formé par ces connexions est appelé le graphe des abonnements. Comme l'indique la figure 2.3, Twitter comptait en 2012 environ 500 millions d'utilisateurs à travers le monde, qui ont publié en moyenne 400 millions de tweets chaque jour. Le réseau d'abonnements entre ces utilisateurs était alors formé de plus de 20 milliards de connexions (Myers *et al.*, 2014). Les utilisateurs ont la possibilité de rédiger et de publier des tweets en temps-réel, notamment grâce aux terminaux mobiles. En 2013, 75% des accès à Twitter se sont faits à partir de terminaux mobiles (Techcrunch, 2013). Chaque tweet apparaît sur la page de profil de son auteur et est instantanément transmis à ses abonnés, qui le reçoivent dans leur « timeline ». La timeline consiste en l'empilement en ordre chronologique inverse des tweets publiés par les utilisateurs suivis (*i.e. followees*). La figure 2.4.a montre la page de profil d'un utilisateur de Twitter, tandis que la figure 2.4.b montre la timeline de cet utilisateur.

Les messages publiés par les utilisateurs de Twitter abordent des thématiques diverses et variées, qu'elles soient d'ordre public – c'est-à-dire en lien avec des éléments d'information susceptibles d'intéresser un large public – ou bien d'ordre personnel. Les tweets appartenant à ce second groupe de thématiques représentent plus de la moitié des tweets publiés (PearAnalytics, 2009; Zheng *et Han*, 2013).

**Accessibilité des données.** Twitter, contrairement à la majorité des médias sociaux permet la collecte de données. Au moment de la rédaction du manuscrit de thèse, les principaux points d'accès proposés gratuitement par Twitter sont les suivants<sup>4</sup> :

- **Streaming.** Ce point d'accès permet de collecter en temps réel des tweets correspondant à une requête portant sur son contenu et certaines métadonnées (*e.g.* langue de l'interface de l'auteur, localisation). L'accès gratuit et anonyme à cette source de données limite le volume de tweets reçus à 1% du volume

---

2. Twitter est accessible sans inscription à l'adresse : <http://twitter.com>.

3. Plus d'informations sont disponibles à l'adresse : <https://about.twitter.com/fr>.

4. Nous décrivons les fonctionnalités de l'API Twitter 1.1 : <https://dev.twitter.com/docs/api/1.1>.

The screenshot shows a Twitter profile for a user named Adrien Guille (@adrienguille). The top navigation bar includes Home, Notifications, Discover, Me, and a search bar. Below the navigation is a large profile picture of a man standing by a bridge over a river. The profile summary shows 6 tweets, 1 photo/video, 17 accounts followed, and 23 followers. A 'More' button and an 'Edit profile' link are also present.

**Annotations:**

- Centre de notifications : messages dans lesquels l'utilisateur est mentionné**: Points to the 'Notifications' tab in the top bar.
- Réseau social : abonnements et abonnés**: Points to the 'Following' and 'Followers' counts at the bottom of the profile summary.
- Messages publiés par l'utilisateur**: Points to the 'Tweets' section of the timeline.

(a) Profil d'un utilisateur.

The screenshot shows the Twitter timeline for the same user, Adrien Guille (@adrienguille). The top navigation bar is identical to the profile page. The main area displays a feed of tweets from various users, with the first tweet by Adrien Guille visible. The profile summary at the top left shows the same statistics as the profile page.

**Annotations:**

- Composition d'un message contenant une mention**: Points to the first tweet in the timeline, which contains a mention of @NoelGallagher.
- Timeline : flux des messages publiés par les abonnements**: Points to the general flow of tweets in the timeline.

(b) Timeline d'un utilisateur.

FIGURE 2.4 – Interfaces du média social Twitter.

total de tweets publiés. Autrement dit, les requêtes couvrant moins de 1% du volume de tweets publiés à chaque instant sont censées renvoyer l'intégralité des tweets correspondants. Au-delà, un échantillonnage non-aléatoire est réalisé par Twitter comme le révèle l'étude menée par *Morstatter et al.* (2013).

- **Search.** Ce point d'accès permet d'accéder aux tweets historisés par Twitter, selon une requête ou pour un utilisateur spécifique. Ce point d'accès limite les résultats aux 3600 derniers tweets publiés par chaque utilisateur. L'accès anonyme et gratuit à cette API est limité à 180 requêtes par tranche de 15 minutes.
- **Followers.** Ce point d'accès permet de récupérer la liste des utilisateurs abonnés à un utilisateur spécifique. L'accès gratuit et anonyme à cette API limite la collecte à 1000 listes d'abonnés par tranche de 24 heures.

## 2.2 Diffusion de l'information

Ayant décrit les médias sociaux, identifié et illustré leurs spécificités, nous nous intéressons maintenant au phénomène de diffusion de l'information auquel ils servent de support.

Chaque message publié sur un média social peut être décrit par une thématique (*Makkonen et al.*, 2004), selon l'un des formalismes donnés dans la définition 2.1.

**Définition 2.1** (Thématique). Une thématique est un ensemble sémantiquement cohérent de termes portant sur un sujet particulier. En pratique, on trouve trois principales interprétations de cette définition : (i) un terme seul, *e.g.* « Obama », (ii) un ensemble de termes, *e.g.* {« Obama », « visite », « Chine »}, ou bien (iii) un ensemble pondéré de termes, *e.g.* {(« Obama », 1), (« visite », 0.75), (« Chine », 0.85)}.

Le flux de messages produit par un média social peut être vu comme une séquence de décisions – chaque décision portant sur la publication ou non d'un message à propos de la thématique – où les utilisateurs consultant leur flux à un moment donné observent les décisions précédemment prises par leurs voisins (*i.e.* dans le cas de Twitter, les utilisateurs auxquels ils sont abonnés). Les utilisateurs peuvent donc être influencés par les décisions des autres, de par les thématiques à propos desquelles ils

publient des messages. Cet effet, connu sous le nom d'influence sociale (*Anagnosopoulos et al.*, 2008), est défini ainsi :

**Définition 2.2** (Influence sociale). L'influence sociale – aussi appelée imitation – est un phénomène social que les utilisateurs des médias sociaux peuvent subir et exercer, traduisant le fait que les actions d'un utilisateur peuvent induire ses connexions à se comporter d'une manière similaire. L'influence se manifeste parfois explicitement dans les médias sociaux, par exemple sur Twitter, lorsqu'un utilisateur « re-tweete » un message – c'est-à-dire lorsqu'il recopie un message en créditant l'auteur original.

Qui plus est, comme le montrent entre autres *Crandall et al.* (2008) et *Aiello et al.* (2012), les facteurs de similarité et d'homophilie renforcent cet effet dans les médias sociaux. D'une part, la similarité entre les utilisateurs connectés dans les médias sociaux a tendance à augmenter au fil du temps du fait de l'influence sociale. D'autre part, les utilisateurs ont tendance à se connecter à d'autres utilisateurs qui leurs sont déjà similaires, selon le principe d'homophilie, ce qui a tendance à amplifier l'effet d'influence sociale.

La diffusion de l'information dans les médias sociaux dépend grandement du phénomène d'influence sociale, comme le rapportent notamment *Mochalova et Nanopoulos* (2013). On distingue deux manières de concevoir le phénomène de diffusion, selon que l'influence sociale soit le facteur central moteur de la diffusion, ou un des éléments contribuant au processus. Dans le premier cas on parle de cascade d'information, définie par la définition 2.3, tandis que dans le second cas, on parle de comportement de foule, défini par la définition 2.4.

**Définition 2.3** (Cascade d'information). Une cascade d'information se produit lorsque les utilisateurs d'un média social publient à propos d'une même thématique seulement par inférence à partir des décisions prises par leurs voisins, quel que soit leur propre signal informationnel. Autrement dit, on considère un monde fermé où l'information ne peut être acquise qu'au sein du média social.

**Définition 2.4** (Comportement de foule). Un comportement de foule s'observe lorsque les utilisateurs d'un média social publient successivement à propos d'une même thématique, en se basant à la fois sur leur propre signal informationnel et sur les décisions prises par leurs voisins. Autrement dit, on considère un monde ouvert où l'information peut être acquise au sein du média social mais également en dehors.

Une expérience de psychologie sociale menée à la fin des années 60 par *Milgram et al.* (1969) (connu notamment pour ses travaux sur le phénomène du « petit monde » (*Milgram*, 1967)) illustre de manière parlante les concepts d'influence sociale et de cascade informationnelle. Cette expérience consiste à placer des groupes d'une à quinze personnes au coin d'une rue. Les membres du groupe se contentent de scruter un ciel dégagé, où il ne se passe rien. Les résultats montrent que plus le groupe de personnes est grand, plus les passants s'arrêtent et observent également le ciel (45% des passants s'arrêtent pour observer le ciel avec le groupe de 15 personnes). Comme le notent *Easley et Kleinberg* (2010), une façon d'interpréter ces résultats est de considérer que les passants n'avaient au départ pas de raison de s'arrêter pour observer le ciel (*i.e.* leur signal informationnel propre leur indique qu'il n'y a rien de particulier dans le ciel), mais que le nombre grandissant de personnes observant le ciel a influencé les futurs passants qui ont rationnellement décidé qu'il devait y avoir une raison de regarder le ciel et se sont arrêtés, donnant naissance à une cascade d'information. D'une manière semblable, les utilisateurs des médias sociaux, influencés par leur voisinage, peuvent être amenés à publier des messages à propos de thématiques pour lesquelles ils n'ont aucun signal informationnel propre. C'est le cas par exemple lorsque des utilisateurs publient des messages et relaient de l'information à propos d'évènements auxquels ils n'ont pas assisté.

*Banerjee* (1992) illustre un autre cas de cascade informationnelle. Admettons qu'une personne doive choisir un restaurant pour dîner dans une ville qu'elle ne connaît pas, et qu'après avoir effectué ses propres recherches elle ait décidé d'aller à un restaurant *A*. Néanmoins, en arrivant sur place elle s'aperçoit que le restaurant *A* est vide tandis que le restaurant *B* juste à côté est presque complet. En supposant que les clients aient des goûts similaires aux siens et qu'ils connaissent les endroits où dîner, il peut sembler rationnel à cette personne de se joindre aux clients du restaurant *B* et ignorer ses propres connaissances. Dans cet exemple, les premiers clients ayant choisi le restaurant *B* ont influencé les clients suivants en transmettant une information à propos de leurs connaissances concernant les endroits où dîner. La cascade d'information se développe alors séquentiellement, lorsque les clients suivants ignorent leurs propres connaissances et se décident sur la base d'inférences faites à partir des décisions des clients précédents. Ce genre de cascades informationnelles se produit également à travers les médias sociaux et contribuent à orienter les discus-

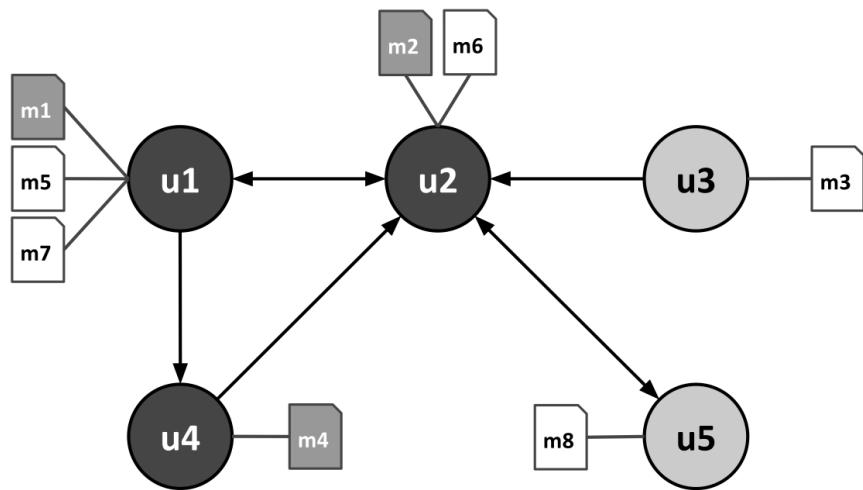


FIGURE 2.5 – Un média social fictif représenté par un graphe, dans lequel les nœuds colorés en gris foncé ( $u_1$ ,  $u_2$ ,  $u_4$ ) ont publié chacun un message ( $m_1$ ,  $m_2$ ,  $m_4$ ) à propos d'une même thématique.

sions vers certaines thématiques plutôt que d'autres.

La manière la plus simple de décrire la diffusion d'une information dans un média social consiste à considérer qu'un utilisateur est soit « actif », c'est-à-dire qu'il a publié un message à propos de la thématique décrivant l'information considérée, ou bien « inactif ». Ainsi, le processus de propagation peut être vu comme une séquence d'activation, au sens de la définition 2.5.

**Définition 2.5 (Séquence d'activation).** Une séquence d'activation est un ensemble ordonné de couples (utilisateur, instant d'activation) capturant l'ordre selon lequel les utilisateurs d'un média social sont devenus actifs à propos d'une thématique.

La figure 2.5 reprend le média social fictif déjà illustré par la figure 2.1. On observe que les nœuds  $u_1$ ,  $u_2$  et  $u_4$  ont chacun publié un message à propos d'une thématique donnée (messages grisés). On sait aussi d'après la figure 2.2 (page 32) représentant le flux des messages publiés par les utilisateurs de ce média social que les messages ont été publiés dans l'ordre  $\langle m_1, m_2, m_4 \rangle$ . On en déduit donc la séquence d'activation suivante :  $\langle (u_1, t_1), (u_2, t_2), (u_4, t_4) \rangle$ .

## 2.3 Vue d'ensemble de la recherche sur la diffusion de l'information dans les médias sociaux

Après avoir détaillé des notions générales ayant trait aux médias sociaux et à la diffusion de l'information, nous dressons ci-après un bref tour d'horizon de la recherche menée à propos de ce phénomène. Pour structurer et synthétiser les travaux décrits dans la littérature, nous construisons la taxonomie présentée par la figure 2.6 (page 41), dont le second niveau reprend les trois problématiques à la base de cette thèse.

Ces travaux couvrent plusieurs champs du domaine de la fouille de données ; par exemple, les méthodes existantes pour la détection d'événements dans les médias sociaux se concentrent sur la détection de thématiques saillantes à l'aide de techniques de fouille de textes : pondération statistiques des termes (*e.g.* avec une variante de  $tf \cdot idf$ ), modélisation des thématiques latentes (*e.g.* avec une variante de l'allocation de Dirichlet latente), ou encore classification non supervisée de termes (*e.g.* avec une variante de la méthode des  $k$  plus proches voisins).

Concernant la prévision de la diffusion dans les médias sociaux, nous catégorisons les modèles existants en deux familles, selon que la structure du réseau social soit prise en compte ou non. Dans les deux cas, les paramètres latents de ces modèles sont estimés en résolvant des problèmes d'optimisation à partir des données (*e.g.* maximisation de la vraisemblance).

En ce qui concerne l'analyse de l'influence dans les médias sociaux, les méthodes existantes exploitent la structure des réseaux sociaux sous-jacents avec des techniques de fouille de graphes (*e.g.* en développant une variante de la décomposition du réseau en  $k$ -enveloppes, ou en modélisant une marche aléatoire sur le réseau), que ce soit pour identifier des utilisateurs ayant une influence positive ou négative sur le phénomène de diffusion de l'information.

Dans les trois chapitres suivants, chacun consacré à une problématique, nous développerons un état de l'art décrivant plus en détail ces travaux.

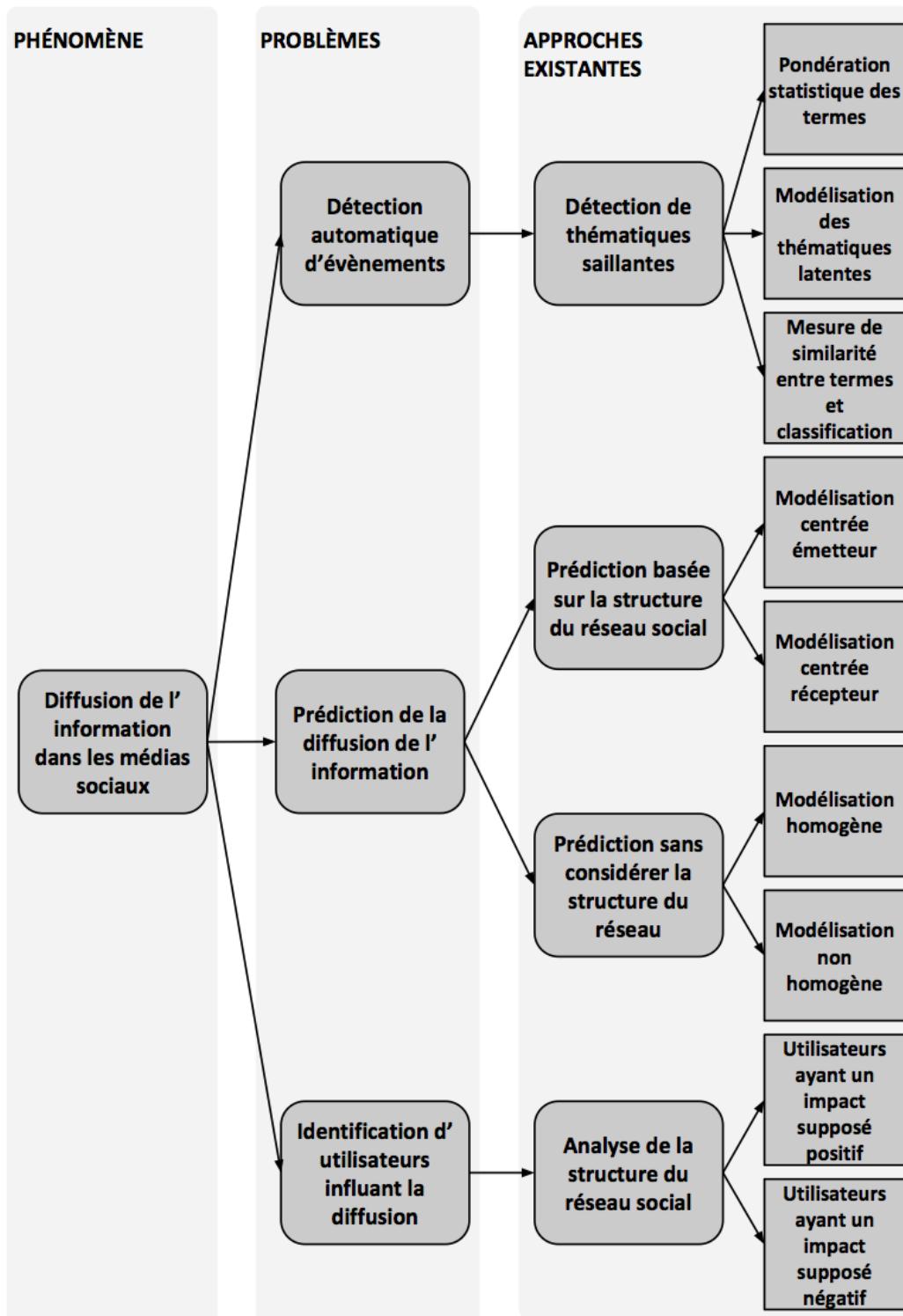


FIGURE 2.6 – Taxonomie structurant les principales pistes de recherche explorées dans les travaux portant sur la diffusion de l'information dans les médias sociaux.



# Chapitre 3

## Déetecter les évènements

### Sommaire

---

<b>3.1 Introduction . . . . .</b>	<b>44</b>
<b>3.2 État de l'art . . . . .</b>	<b>48</b>
3.2.1 Pondération statistique des termes . . . . .	49
3.2.2 Modélisation probabiliste des thématiques latentes . . . . .	52
3.2.3 Classification non supervisée de termes . . . . .	54
3.2.4 Synthèse de l'état de l'art . . . . .	56
<b>3.3 Méthode proposée . . . . .</b>	<b>58</b>
3.3.1 Formulation du problème . . . . .	58
3.3.2 Vue d'ensemble de la méthode proposée . . . . .	59
3.3.3 Détection des évènements à partir de l'anomalie dans la fréquence de création de mentions . . . . .	61
3.3.4 Sélection des mots décrivant les évènements . . . . .	64
3.3.5 Génération de la liste des évènements . . . . .	66
3.3.6 Algorithme général . . . . .	68
<b>3.4 Expérimentations . . . . .</b>	<b>70</b>
3.4.1 Protocole expérimental . . . . .	70
3.4.2 Évaluation quantitative . . . . .	73
3.4.3 Évaluation qualitative . . . . .	76
<b>3.5 Implémentation et visualisations . . . . .</b>	<b>79</b>
<b>3.6 Discussion . . . . .</b>	<b>82</b>
3.6.1 Résumé des travaux présentés . . . . .	83
3.6.2 Perspectives de travail . . . . .	83

---

Dans le chapitre précédent nous avons présenté les concepts et théories de base en rapport avec les médias sociaux et la diffusion de l'information, ainsi qu'une vue d'ensemble des recherches menées dans le domaine. Dans le présent chapitre, nous abordons la première des contributions de cette thèse, qui porte sur la détection des événements suscitant l'intérêt des utilisateurs des médias sociaux.

## 3.1 Introduction

Comme cela a déjà été discuté précédemment, les utilisateurs des médias sociaux publient des messages courts sur des sujets très variés – allant d'événements banals de la vie quotidienne à des événements importants et/ou globaux – en temps réel. Le nombre sans cesse croissant d'utilisateurs à travers le monde fait des médias sociaux des sources d'information précieuses. Dans le même temps, cela donne naissance à un phénomène de surcharge informationnelle et il devient de plus en plus difficile d'identifier des éléments d'information pertinents liés à des événements importants. Par « événement important » nous entendons ici un événement réel et susceptible d'être couvert par les médias traditionnels. Ces faits nous amènent donc à la question suivante : *Comment peut-on utiliser les médias sociaux pour détecter automatiquement les événements ?* Répondre à cette question aiderait à analyser les événements, ou les types d'événements, qui intéressent le plus les utilisateurs des médias sociaux. Cela est également important pour l'analyse journalistique des événements, retracer leur déroulement, etc. Cependant, la plupart des médias sociaux ne fournissent aucune assistance pour cette tâche. Même si Twitter fait exception et propose une liste de « trends », celle-ci n'est que peu utile puisqu'elle n'énumère que des mots-clés sans information temporelle ni information à propos du niveau d'attention que lui accordent les utilisateurs. La figure 3.1 reprend à titre d'exemple les « trends » détectés par Twitter le 8 juillet 2014, pour la France, les États-Unis et le monde entier. On constate notamment que les « trends » concernant la France sont difficilement interprétables, avec des expressions isolées telles que « Joyeux Anniversaire » ou « WAllah ».

Détecter automatiquement les événements importants à partir des médias sociaux est une tâche difficile puisque les messages rapportant ou discutant les événements intéressants sont noyés par un grand volume de messages sans rapport, i.e. du bruit. Les médias sociaux délivrent un flux de messages continu, ce qui permet d'étudier com-

## Déetecter les évènements

---

France Trends · Change	United States Trends · Change	Worldwide Trends · Change
#ProblemesDeMecs	#WeLoveYouConnor	#KeepCalmAndVoteJokowi
#BREALL	#BeforeMeloDecides	#RailBudget
#TeamInsomnique	#5sostribe	#WeLoveYouConnor
Xavier Bertrand	#VoteGRich	#KüçükkenHiçUnutmam
Touquet	Dequan	#CanimisiyanSeyNeBiliyormusun
#vacances	Daquan	£11m for Ross McCormack
#OntTaTejUnePommeDeTerre	#cantsleep	Happy Thanksgiving
Joyeux Anniversaire	Netflix	Fulham
WAallah	George Lopez	Dequan
Londres	FaceTime	Main Chick

FIGURE 3.1 – De gauche à droite, les « trends » détectés par Twitter le 8 juillet 2014 pour la France, les États-Unis et le monde entier.

ment les thématiques apparaissent et disparaissent au cours du temps (*Leskovec et al.*, 2009). En particulier, les méthodes pour la détection d'évènements se concentrent sur les thématiques « saillantes » qui sont supposées signaler des évènements (*Kleinberg*, 2002). Au niveau individuel, une thématique saillante se caractérise par une apparition soudaine liée à sa montée en popularité, laquelle ne persiste que sur une durée relativement courte. Au niveau global, la dynamique temporelle des thématiques saillantes est une succession de pics d'attention. La figure 3.2 illustre ces propriétés en représentant l'évolution de la fréquence de 7 thématiques saillantes au sein d'un média social fictif. Comme nous le verrons dans l'état de l'art qui sera détaillé par la suite, les méthodes de détection d'évènements existantes mettent en œuvre différentes approches : pondération statistique des termes (*Shamma et al.*, 2011; *Benharnodus et Kalita*, 2013), modélisations probabilistes des thématiques latentes (*Lau et al.*, 2012; *Yuheng et al.*, 2012), ou encore clustering (*Weng et Lee*, 2011; *Li et al.*, 2012; *Parikh et Karlapalem*, 2013).

Malgré la richesse de la recherche portant sur cette problématique, les travaux issus de la littérature se concentrent sur le contenu textuel des messages échangés et négligent l'aspect social. Or, les utilisateurs insèrent souvent du contenu non-textuel dans leurs messages. En particulier, les utilisateurs ont la possibilité d'insérer (*i.e.* de mentionner) dans leurs messages les pseudonymes d'autres utilisateurs. Cette pratique – appelée « mentioning » – se retrouve sur la plupart des médias sociaux, notamment Twitter, Facebook et Google+ qui respectent tous trois une même syntaxe, à savoir « @pseudonyme ». Ces mentions sont en fait des liens dynamiques créés soit

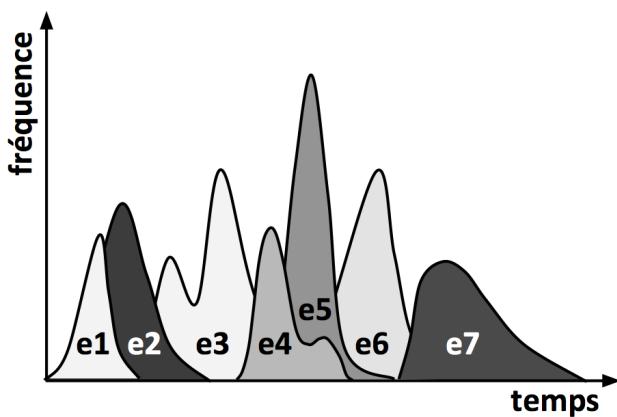


FIGURE 3.2 – Dynamique temporelle des thématiques saillantes correspondant aux discussions engendrées par 7 évènements dans un média social fictif.

intentionnellement pour engager la discussion avec des utilisateurs spécifiques, soit automatiquement lors d'une réponse à un message, ou bien encore sur Twitter, lors d'un re-tweet. On considère ce type particulier de lien comme dynamique puisqu'il est lié à une période temporelle spécifique, *i.e.* la durée de vie du message, et à une thématique spécifique, *i.e.* celle abordée par le message.

**Proposition et positionnement.** Nous traitons le problème de la détection d'évènements dans les médias sociaux en proposant une nouvelle méthode statistique, fondée sur l'analyse de l'anomalie dans la fréquence de création de mentions : *MABED* (*Mention-Anomaly-Based Event Detection*). Cette méthode produit une liste d'évènements, chaque évènement étant décrit par (i) un mot principal et un ensemble pondéré de mots liés, (ii) une période de temps et (iii) sa magnitude d'impact sur les utilisateurs, laquelle est proportionnelle à l'ampleur de la réaction suscitée auprès des utilisateurs.

La méthode *MABED* diffère des méthodes existantes pour plusieurs raisons. Tout d'abord, elle intègre l'aspect social des flux de messages en analysant les mentions pour capturer au mieux les évènements qui intéressent les utilisateurs. Ensuite, elle introduit par rapport à l'existant une structuration dans la description textuelle de chaque évènement, à travers la distinction faite entre mots principaux et mots liés, pour faciliter leur lecture. Par ailleurs, *MABED* estime dynamiquement la durée de

chaque évènement – là où les méthodes existantes supposent généralement une durée commune à tous les évènements – ce qui permet également d'estimer plus précisément leur magnitude d'impact sur les utilisateurs. Enfin, la méthode que nous proposons repose strictement sur des mesures statistiques à partir d'un flux de messages, ce qui la rend indépendante de la langue utilisée et aisément utilisable, contrairement à certaines méthodes de la littérature qui font appel à de l'information externe (e.g. Wikipédia, médias traditionnels).

**Résultats.** Nous menons une évaluation quantitative et qualitative de la méthode proposée sur un corpus francophone et un corpus anglophone, de plusieurs millions de tweets chacun. Nous montrons que *MABED* est capable d'extraire une vision retrospective claire et précise des évènements discutés dans chaque corpus, avec des temps de calcul courts. L'efficacité de la méthode est amplifiée par le fait qu'elle traite des tweets bruts, c'est-à-dire qu'elle ne requiert pas de pré-traitements coûteux, tels que la lemmatisation ou l'identification d'entités nommées. Pour étudier la précision et le rappel, nous demandons à des annotateurs humains de juger si les évènements détectés sont compréhensibles et significatifs. Nous démontrons empiriquement la pertinence de l'approche basée sur l'anomalie dans la fréquence de création de mentions en montrant que *MABED* obtient de meilleurs résultats qu'une variante qui ignore la présence ou non de mentions dans les tweets. Nous montrons également que *MABED* améliore l'état de l'art en comparant ses performances avec celles de méthodes récentes tirées de la littérature.

**Application.** *MABED* est utilisé depuis décembre 2013 pour analyser en continu les discussions en rapport avec François Hollande sur Twitter. À partir de tweets collectés en temps réel via l'*API streaming* de Twitter, *MABED* identifie et suit les évènements au plus fort impact et aide à analyser la façon dont les opinions se forment puis se propagent sur Twitter<sup>1</sup>. L'implémentation de la méthode est publique<sup>2</sup> et est également intégrée dans l'outil *SONDY*, un logiciel libre et gratuit pour la fouille des données issues des médias sociaux que nous décrivons dans le chapitre 5 de ce manuscrit.

Ce chapitre est organisé de la manière suivante. Dans la section 3.2, nous présentons une synthèse de l'état de l'art puis dans la section 3.3 nous décrivons formellement la méthode proposée. Ensuite, nous détaillons les résultats de l'évaluation

---

1. <http://mediamining.univ-lyon2.fr/people/guille/twitterstream.php>  
2. <http://mediamining.univ-lyon2.fr/people/guille/mabed.php>

quantitative et qualitative que nous avons menée dans la section 3.4, puis nous présentons le prototype implémentant la méthode proposée dans la section suivante. Enfin, nous concluons ce chapitre et discutons des perspectives dans la section 3.6.

## 3.2 État de l'art

Les méthodes pour la détection d'évènements à partir des médias sociaux reposent sur de nombreux travaux portant sur la détection de thématiques, de motifs saillants et d'évènements à partir de flux textuels, un évènement se définissant comme une thématique propre à une certaine période de temps. Dans une étude fondatrice, *Kleinberg* (2002) s'intéresse à la détection de motifs saillants à partir d'un flux d'e-mails. Supposant que tous les e-mails reçus traitent de la même thématique, il propose de modéliser les motifs saillants à l'aide de chaînes de Markov cachées. Pour traiter un flux de documents textuels abordant diverses thématiques, *AlSumait et al.* (2008) développent *OLDA*, une modélisation probabiliste des thématiques latentes qui permet de construire des matrices caractérisant l'évolution de la distribution des thématiques au cours du temps, à partir desquelles les thématiques saillantes peuvent être identifiées. *Fung et al.* (2005) proposent d'identifier les mots saillants en étudiant l'allure de la distribution de leur fréquence, puis de les regrouper en analysant leurs cooccurrences pour former des descriptions d'évènements. Ils nomment leur approche *feature-pivot clustering*, par opposition à l'approche classique *document-pivot clustering* qui vise à former des groupes de documents similaires pour ensuite extraire une description d'évènement pour chaque groupe.

Cependant, comme nous l'avons déjà évoqué dans le chapitre 2, les flux de messages produits par les médias sociaux présentent plusieurs spécificités, du point de vue de l'expressivité, du volume, etc. Cela limite l'efficacité des méthodes traditionnelles, tant au niveau de la précision qu'en matière de passage à l'échelle et il est par conséquent nécessaire de développer de nouvelles méthodes mieux adaptées (*Bouillot et al.*, 2012). Dans les sous-sections suivantes nous décrivons des méthodes représentatives tirées de la littérature, que nous regroupons en trois familles en fonction du type d'approche qu'elles mettent en œuvre, à savoir : (i) pondération statistique des termes, (ii) modélisation probabiliste des thématiques latentes et (iii) classification non supervisée de termes.

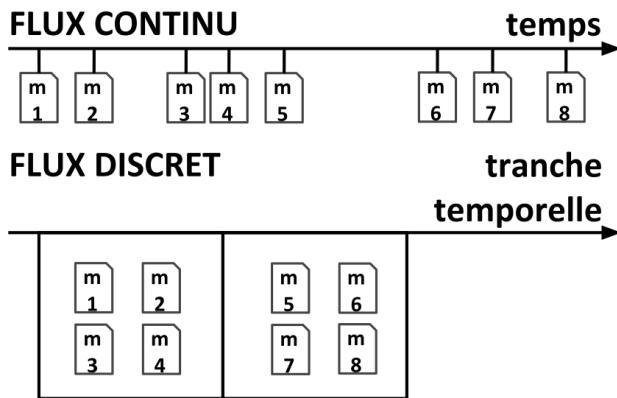


FIGURE 3.3 – Un flux de messages continu et le même flux de messages dont l’axe temporel a été discrétisé.

**Entrée, pré-traitement et sortie.** Ces méthodes reçoivent en entrée un corpus  $\mathcal{C}$  de messages, dont le vocabulaire est noté  $V$ . Elles requièrent toutes le même pré-traitement du corpus, qui consiste à partitionner les messages en  $n$  tranches temporelles de durées égales afin de discrétiser l’axe temporel, notamment pour permettre le calcul de fréquences. Ce pré-traitement est illustré sur la figure 3.3. On y voit comment un flux continu de messages est transformé en une séquence de collections de messages. En sortie, elles produisent une liste d’évènements ordonnés selon leur significativité.

### 3.2.1 Pondération statistique des termes

La première famille de méthodes pour la détection d’évènements à partir des médias sociaux regroupe des méthodes qui définissent des métriques permettant de scorer les termes rencontrés dans les messages, de telle sorte que les termes liés à des évènements obtiennent les scores les plus élevés. Pour faciliter l’identification des évènements, les termes sont classés selon ces scores. Nous décrivons ici deux méthodes issues de cette famille.

*Shamma et al. (2011)* proposent la méthode nommée *Peaky Topics* qui repose sur le calcul d’une mesure de fréquence normalisée  $ntf_{t,i}$  pour chaque mot  $t \in V$  en chaque tranche temporelle  $i \in [1; n]$ . La fréquence est normalisée par le nombre total

d'occurrences du mot  $t$  dans le flux de messages, afin que les mots fréquents en une tranche temporelle et rares dans le reste du flux aient une valeur de  $ntf$  élevée en cette tranche temporelle et faible pour les autres. La fréquence normalisée est définie ainsi :

$$ntf_{t,i} = \frac{tf_{t,i}}{cf_t}$$

où  $tf_{t,i}$  est la fréquence du mot  $t$  à la  $i^{\text{ème}}$  tranche temporelle et  $cf_t$  est le nombre total d'occurrences du mot  $t$  dans le corpus  $\mathcal{C}$ . Pour établir le classement des mots liés à des évènements, les auteurs définissent ensuite un score pour chaque mot  $t$  :  $peakiness_t = \max(ntf_{t,i})$ . Chaque entrée du classement est décrite par un mot, son score et la tranche temporelle maximisant la métrique  $ntf$ . La figure 3.4 se base sur un flux de messages fictif dont le vocabulaire  $V$  contient 4 mots,  $V = \{t1, t2, t3, t4\}$ , partitionné en 4 tranches temporelles. On observe sur cette figure que la fréquence des mots  $t1$  et  $t2$ , en gris foncé et gris clair, est distribuée de manière quasiment uniforme dans le temps. Par conséquent, ces deux mots ont un score  $peakiness$  proche de  $\frac{1}{4}$ . Au contraire, les mots  $t3$  et  $t4$  n'apparaissent chacun que dans une tranche temporelle, leur score  $peakiness$  est donc égal à 1. Cet exemple met en évidence une des limitations de la méthode *Peaky Topics*. En effet, la normalisation proposée ne tient compte que de la variabilité de la fréquence du mot considéré à travers le temps. Par conséquent, les mots dont la fréquence est uniformément distribuée – qu'ils soient toujours très fréquents ou toujours très rares – auront un score  $peakiness$  proche de  $\frac{1}{n}$  (où  $n$  est le nombre de tranches temporelles). De la même façon, un mot qui n'apparaît que dans une tranche temporelle, quelle que soit sa fréquence absolue dans cette tranche, aura un score égal à 1. L'autre limitation est liée au fait qu'un seul mot peut ne pas être suffisant pour décrire un évènement complexe, à cause de la possible ambiguïté et du manque de contexte.

Pour pallier à ces limitations, *Benhardus et Kalita* (2013) suggèrent de s'intéresser aux N-grammes de mots (*i.e.* séquences de  $N$  mots consécutifs) et plus particulièrement les bigrammes et trigrammes. Pour chaque N-gramme et tranche temporelle, ils proposent de calculer un score, nommé *trending score* (aussi noté *TS*), équivalent à une fréquence normalisée. La normalisation est effectuée par rapport au nombre total d'occurrences du N-gramme dans le corpus de messages et aussi – contrairement à la méthode *Peaky Topics* – par rapport à la fréquence des autres N-grammes dans la

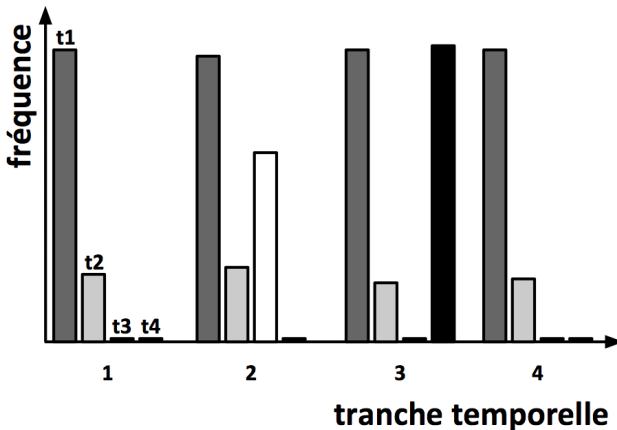


FIGURE 3.4 – Fréquence des mots pour un flux de messages fictif.

même tranche temporelle. Le *trending score* est défini de la façon suivante :

$$TS_{t,i} = \frac{ntf_{t,i}}{atf_{t,i}}$$

$$\text{où } ntf_{t,i} = \frac{tf_{t,i}}{\sum_{k \in V} tf_{k,i}}$$

ce qui normalise la fréquence du mot  $t$  par rapport à celle des autres mots du vocabulaire, et :

$$atf_{t,i} = \sum_{j, j \neq i} \frac{ntf_{t,j}}{n - 1}$$

ce qui normalise la fréquence du mot  $t$  par rapport à sa fréquence dans les autres tranches temporelles. Néanmoins, même si l'utilisation de N-grammes peut améliorer la sémantique des résultats, ils ne permettent pas de capturer les relations entre des mots trop éloignés dans les messages et sont d'autant plus sensibles au bruit que  $N$  est grand.

### 3.2.2 Modélisation probabiliste des thématiques latentes

Dans le but d'augmenter la sémantique des évènements détectés, plusieurs méthodes basées sur la modélisation probabiliste des thématiques latentes ont été développées. Ces méthodes décrivent chaque évènement à l'aide d'un ensemble pondéré de mots.

*Lau et al.* (2012) développent *On-line LDA*, une variante en ligne du modèle *LDA* (*i.e.* Latent Dirichlet Allocation (*Blei et al.*, 2003)) et une technique pour mesurer l'évolution des thématiques et ainsi détecter les évènements. *LDA* est un modèle génératif probabiliste qui apprend un ensemble de thématiques latentes à partir d'une collection de documents, chaque document étant considéré comme un sac de mots. Un document est caractérisé par une distribution sur un nombre fixé ( $K$ ) de thématiques et une thématique est une distribution de probabilités sur le vocabulaire  $V$  des mots employés dans les documents. La figure 3.5 résume le processus génératif à la base du modèle *LDA* à l'aide de la notation en plaques. Ainsi, les noeuds gris clairs représentent les variables latentes, tandis que le noeud gris foncé représente la seule variable observée,  $w$ , à savoir les mots dans les documents. Sur cette figure,  $K$  est le nombre de thématiques,  $W$  le nombre de mots dans un message et  $M$  le nombre total de messages dans la collection. Les variables  $\phi$  et  $\theta$  représentent respectivement la distribution des mots en fonction des thématiques et la distribution des thématiques en fonction des documents, tandis que  $z$  donne l'affectation des thématiques pour les mots de chaque document. Les variables  $\alpha$  et  $\beta$  sont les paramètres de concentration des priors de Dirichlet du modèle génératif qui influencent  $\phi$  et  $\theta$ . La particularité d'*On-line LDA* est que le modèle génératif est entraîné – à l'aide de l'échantillonnage de Gibbs (*Casella et George*, 1992), une méthode du type Monte-Carlo – en chaque tranche temporelle. Afin de maintenir la comparabilité une à une entre les thématiques d'une tranche temporelle à une autre, *Lau et al.* (2012) introduisent dans ce modèle un nouveau paramètre  $c \in [0; 1]$  qui quantifie l'influence des paramètres  $\alpha$  et  $\beta$  appris en  $i$  sur les paramètres  $\alpha'$  et  $\beta'$  appris en  $i + 1$ . En choisissant une valeur de  $c$  différente de 0, il est possible d'étudier l'évolution de la distribution des mots par thématique (*i.e.*  $\phi$ ) d'une tranche temporelle à une autre. Ainsi, les auteurs suggèrent de mesurer le score d'évolution de la thématique  $T$  à la tranche  $i$  comme  $e(T, i) = \text{JSD}(\phi_T^i, \phi_T^{i-1})$ , où  $\text{JSD}$  est la mesure de divergence de

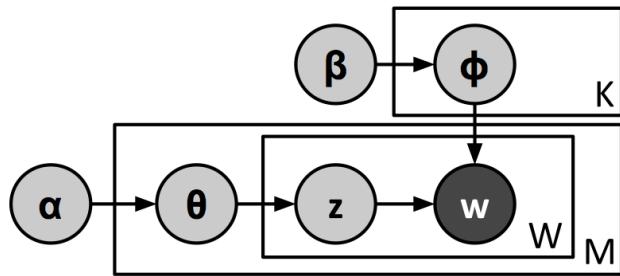


FIGURE 3.5 – Représentation en plaques du modèle LDA.

Jensen-Shannon (*Lin, 1991*). L'ensemble des évènements détectés correspond à l'ensemble des couples  $(T, i)$ , ordonné selon le score d'évolution  $e$  avec un seuil manuel sur la valeur minimale de  $e$ .

*Yuheng et al. (2012)*, constatant que le modèle *LDA* s'adapte mal aux documents courts, comme c'est le cas des messages publiés dans les médias sociaux, développent *ET-LDA* (*i.e. Event and Tweets LDA*). Leur approche présente deux spécificités. Premièrement, les auteurs proposent d'enrichir les messages publiés sur les médias sociaux. Chaque message est utilisé comme requête sur un moteur de recherche traditionnel et est ensuite enrichi par l'adjonction des mots les plus fréquents dans les résultats de la recherche. Deuxièmement, *ET-LDA* modélise conjointement les thématiques dans le corpus de messages enrichis et un corpus d'articles tirés des médias traditionnels afin de favoriser la distinction entre les thématiques d'arrière plan et les thématiques liées à des évènements. Néanmoins il n'est pas clair en quoi cette approche permet d'améliorer les résultats. En effet, les évènements détectés correspondent uniquement à ceux traités dans le corpus d'articles utilisé, et même si les messages issus des médias sociaux peuvent potentiellement apporter des informations supplémentaires, leur impact est minoré par le fait que ces messages ont été altérés par l'information apportée par un moteur de recherche traditionnel.

L'étude réalisée par *Aiello et al. (2013)* montre que les méthodes de détection d'évènements reposant sur la modélisation des thématiques latentes souffrent de plusieurs limitations. Notamment, du fait du coût important du processus d'apprentissage du modèle génératif, ces méthodes passent difficilement à l'échelle. Il apparaît aussi que ce type de méthode est particulièrement inefficace pour traiter des flux de

messages dans lesquels de nombreux événements distincts sont discutés.

### 3.2.3 Classification non supervisée de termes

Les méthodes basées sur la classification non supervisée de termes visent à offrir un coût computationnel moindre que celui des méthodes de la famille précédente tout en identifiant des descriptions des événements sémantiquement riches, sous la forme d'ensembles de termes.

*Weng et Lee (2011)* proposent une approche reposant sur la technique des ondelettes, nommée *EDCoW* (*i.e. Event Detection with Clustering of Wavelet-based signals*). Dans un premier temps, un signal capturant la saillance de chaque mot  $t \in V$  en fonction du temps est construit. L'idée consiste à calculer la transformée en ondelettes discrète de la série temporelle représentant la fréquence d'un mot, par morceaux, et de créer un signal dont chaque point correspond à la différence entre les mesures d'entropie normalisée de Shannon de la distribution des énergies des ondelettes pour deux morceaux successifs. Ce signal permet ainsi de capturer les irrégularités dans la fréquence de chaque mot. Les mots apparaissant de manière régulière, c'est-à-dire ceux ne présentant pas de motif saillant dans leur fréquence, sont filtrés sur la base de la mesure d'autocorrélation de leur signal et de l'indice *MAD* (*i.e. Median Absolute Deviation (David et al., 1983)*). Dans un second temps, *EDCoW* construit une matrice de similarité  $X$  entre les mots restants, où  $X_{t_1 t_2}$  est égal à la mesure de corrélation croisée (sans décalage temporel, ce qui revient alors au coefficient de corrélation de Pearson) entre les signaux correspondant à  $t_1$  et  $t_2$  pour une fenêtre de taille fixée (qui détermine donc la durée de tous les événements détectés). Afin de rendre chaque matrice creuse, l'indice *MAD* est de nouveau mesuré et toutes les cellules dont la valeur est inférieure à la valeur de l'indice sont ramenées à 0. Dans un troisième temps, les événements sont identifiés en formant des groupes de mots fortement corrélés à partir de chaque matrice  $X$  – que l'on considère comme la matrice d'adjacence d'un graphe non-orienté et pondéré – avec l'algorithme spectral d'optimisation de la modularité de *Newman (2006)*. Enfin, les événements détectés sont classés selon un score  $\epsilon$ , avec un seuil sur la valeur minimale, qui est proportionnel à la somme des poids du sous-graphe lié à chaque événement. Le principal problème de cette méthode est que la similarité entre deux mots n'est mesurée que du point de vue temporel, ce

qui a pour effet de regrouper ensemble les mots liés à des évènements se produisant simultanément.

*Li et al.* (2012) proposent la méthode *TwEvent* qui vise à regrouper non pas des mots mais des N-grammes de mots. Les N-grammes candidats sont d'abord identifiés et sélectionnés sur la base d'informations statistiques fournies par *Microsoft Web N-Gram service*<sup>3</sup> ainsi que leur fréquence d'apparition sur Wikipedia en tant que labels de liens pointant vers d'autres articles. Les auteurs définissent ensuite de manière paramétrique la probabilité qu'un N-gramme soit saillant en une tranche temporelle donnée, en fonction d'une mesure de fréquence normalisée. Les N-grammes sont ensuite regroupés selon une stratégie du type « *k* plus proches voisins » à l'aide de l'algorithme décrit par *Jarvis et Patrick* (1973), pour une fenêtre de taille fixe – de façon similaire à *EDCoW*. Mais contrairement à *EDCoW*, l'estimation de la similarité entre deux N-grammes  $t_1$  et  $t_2$  n'est pas seulement basée sur leurs motifs d'apparition dans le temps durant cette fenêtre. La similarité prend également en compte le contenu lié à chaque N-gramme durant cette fenêtre. Cette mesure de similarité de contenu est établie en sélectionnant les deux ensembles de messages correspondant respectivement aux messages contenant  $t_1$  et aux messages contenant  $t_2$  dans la fenêtre temporelle, puis en mesurant la similarité cosinus (*Singhal*, 2001) de ces deux ensembles de N-grammes pondérés à l'aide de  $tf \cdot idf$  (*Salton et McGill*, 1986). Les groupes de N-grammes de mots formés sont alors évalués selon le *newsworthiness score* défini par les auteurs, qui est proportionnel à la fréquence d'apparition des N-grammes comme liens sur Wikipédia, filtrés selon un seuil manuel puis classés.

Constatant que l'extraction des N-grammes par *TwEvent* est à la fois coûteuse et grandement influencée par Microsoft Web N-Gram et Wikipédia, *Parikh et Karlapalem* (2013) développent la méthode *ET*. Cette méthode ne considère que les bigrammes, et plus particulièrement les bigrammes saillants qui sont détectés à l'aide d'une mesure de fréquence normalisée par rapport au temps. Les bigrammes saillants sont regroupés par classification ascendante hiérarchique selon la similarité entre bigrammes, mesurée à partir de la similarité entre leur fréquence dans le temps et les bigrammes avec lesquels ils apparaissent souvent.

Les méthodes à base de clustering présentent certains inconvénients. Notamment, elles ont tendance à produire des clusters de grande taille. *Valkanas et Gunopoulos*

---

3. Beta privée accessible sur invitation : <http://web-ngram.research.microsoft.com>.

TABLE 3.1 – Matrice de comparaison des méthodes existantes pour la détection d'évènements.

	Description textuelle	Description temporelle	Prise en compte de l'aspect social
<i>Peaky topics</i>	un mot	une tranche temporelle	non
<i>Trending score</i>	un n-gramme	une tranche temporelle	non
<i>On-line LDA</i>	un ensemble pondéré de mots	une tranche temporelle	non
<i>ET-LDA</i>	un ensemble pondéré de mots	indéfinie	non
<i>EDCoW</i>	un ensemble de mots	un nombre fixe de tranches temporelles	non
<i>TwEvent</i>	un ensemble de n-grammes	un nombre fixe de tranches temporelles	non
<i>ET</i>	un ensemble de bigrammes	un nombre fixe de tranches temporelles	non

TABLE 3.2 – Matrice de comparaison des méthodes existantes pour la détection d'évènements.

(2013) notent à ce propos que les méthodes à base de clustering, en regroupant de manière « aggressive » certains termes, incorporent du bruit dans les descriptions des évènements.

### 3.2.4 Synthèse de l'état de l'art

La matrice présentée par la table 3.2 synthétise cet état de l'art en comparant les méthodes selon trois critères : (i) la manière dont les évènements sont décrits, (ii) la manière dont la durée des évènements est déterminée et (iii) la prise en compte de l'aspect social du flux de messages.

**Description textuelle des évènements.** Nous constatons qu'il n'existe pas de consensus sur la manière de décrire un évènement. Néanmoins, il semble que la manière la plus fine de décrire un évènement soit un ensemble pondéré de mots, comme le font les méthodes basées sur la modélisation probabiliste des thématiques latentes, qui décrivent chaque évènement par une distribution de probabilités sur un

ensemble de mots. L'avantage de la pondération est qu'elle permet, si cela est nécessaire, de contrôler la taille des descriptions en sélectionnant un sous-ensemble de mots en fonction de leur poids. Par contraste, les méthodes basées sur la classification non-supervisée de termes – telles que *EDCoW* qui adopte une approche basée sur la modularité ou *TwEvent* qui applique les  $k$  plus proches voisins – ont tendance à produire des descriptions très longues et donc difficiles à interpréter.

**Description temporelle des évènements.** Il apparaît que les méthodes existantes se basent sur une hypothèse commune, à savoir que tous les évènements détectés ont une même durée. Certaines méthodes considèrent que la durée de chaque évènement est égale à la durée d'une tranche temporelle – c'est par exemple le cas des méthodes de pondération statistique des termes que nous avons décrites précédemment – ou bien, égale à la durée d'une séquence de longueur fixe de tranches temporelles – ce que font les trois méthodes basées sur la classification non-supervisée de termes. Dans le premier cas, les auteurs partitionnent généralement les messages par tranches de 24 heures, tandis que dans le second cas, la longueur des séquences est souvent fixée à 24 heures, mais la durée des tranches temporelles peut être ajustée afin de définir le niveau de détail voulu pour la mesure de similarité temporelle. Néanmoins tous les évènements ne suscitent pas le même intérêt auprès des utilisateurs des médias sociaux, et par conséquent, la durée pendant laquelle ceux-ci continuent d'en discuter varie de quelques heures à plusieurs jours. Or, en fixant la durée de tous les évènements, cette information est perdue et il paraît crucial d'estimer dynamiquement la durée de chaque évènement.

**Aspect social.** Toutes les méthodes décrites dans cet état de l'art se concentrent sur la fréquence des mots ou des N-grammes de mots. Or, les médias sociaux ont pour vocation de favoriser les interactions sociales entre leurs utilisateurs, lesquelles sont notamment symbolisées au sein des messages publiés via des marqueurs spécifiques, tels que les mentions. Par conséquent, il semble important de ne pas se concentrer uniquement sur l'aspect textuel des messages, mais également sur l'aspect social des flux de messages produits par les médias sociaux.

Dans la section suivante, nous décrivons la méthode que nous proposons pour pallier à ces limitations.

TABLE 3.3 – Liste des notations utilisées dans le chapitre 3.

Notation	Définition
$N$	Nombre total de messages dans le corpus
$N^i$	Nombre de messages à la $i^{\text{ème}}$ tranche temporelle
$N_t^i$	Nombre de messages à la $i^{\text{ème}}$ tranche temporelle contenant le mot $t$
$N_{@t}$	Nombre de messages dans le corpus contenant le mot $t$ et au moins une mention
$N_{@t}^i$	Nombre de message contenant le mot $t$ et au moins une mention à la $i^{\text{ème}}$ tranche temporelle

## 3.3 Méthode proposée

### 3.3.1 Formulation du problème

**Entrée.** Soit un corpus de messages  $\mathcal{C}$ . Nous discrétons l'axe temporel en partitionnant les messages en  $n$  tranches temporelles de même durée. Soit  $V$  le vocabulaire des mots employés dans l'ensemble des messages et  $V_@$  le vocabulaire des mots employés dans les messages contenant au moins une mention. La table 3.3 donne les notations utilisées dans le reste de ce chapitre.

**Sortie.** L'objectif consiste à produire une liste  $L$ , telle que  $|L| = k$ , contenant les  $k$  événements distincts ayant eu les plus grandes magnitudes d'impact sur le comportement des utilisateurs, tant en terme de publication de messages qu'en terme d'interaction entre eux. Nous définissons un événement comme une thématique saillante, dont la magnitude d'impact est caractérisée par un score. Les définitions 3.1 et 3.2 ci-après, définissent respectivement les concepts de thématique saillante et d'événement.

**Définition 3.1** (Thématique saillante). Étant donné un intervalle  $I$ , une thématique  $T$  est considérée saillante si elle a suscité un niveau d'attention particulièrement important durant cet intervalle en comparaison avec le reste de la période d'observation. La thématique  $T$  est décrite par un terme principal  $t$  et un ensemble pondéré  $S$  de mots. Les poids varient entre 0 et 1, un poids proche de 1 signifiant que le mot est central à la thématique durant la période saillante, tandis qu'un poids proche de 0 signifie que ce mot est moins spécifique à cette période.

**Définition 3.2** (Évènement). Un évènement  $e$  est caractérisé par une thématique saillante  $TS = [T, I]$  et une valeur  $Mag > 0$  reflétant la magnitude d'impact de l'évènement sur les utilisateurs.

### 3.3.2 Vue d'ensemble de la méthode proposée

La méthode proposée, *MABED* (*Mention-Anomaly-Based Event Detection*), suit un processus en deux phases et s'appuie sur trois composants : (i) la détection des évènements à partir de la mesure d'anomalie dans la fréquence de création de mentions, (ii) la sélection des mots décrivant le mieux chaque évènement et (iii) la génération de la liste des  $k$  évènements ayant eu le plus d'impact. Le déroulement de la méthode, illustrée sur la figure 3.6, est brièvement décrit ci-après.

1. La fréquence de création de mention associée à chaque mot  $t \in V_{@}$  est analysée avec le premier composant. Il en ressort une liste d'évènements partiellement définis, dans le sens où l'ensemble  $S$  de mots décrivant chaque évènement est vide. Autrement dit, chaque évènement est caractérisé par un mot principal, un intervalle temporel et sa magnitude d'impact. Cette liste des évènements est triée par ordre décroissant selon leur magnitude d'impact.
2. La liste est parcourue en partant de l'évènement ayant eu le plus fort impact. Pour chaque évènement, le second composant sélectionne l'ensemble  $S$  des mots le décrivant le mieux. La sélection se base sur la cooccurrence et la dynamique temporelle des mots employés durant l'intervalle  $I$ . Chaque évènement traité par ce composant est ensuite passé au troisième composant, qui est chargé de sauvegarder les descriptions des évènements tout en gérant les évènements redondants. Enfin, lorsque  $k$  évènements *distincts* (*i.e.* en n'incluant pas les éventuels évènements redondants) ont été traités, le troisième composant fusionne les éventuels évènements redondants et renvoie la liste  $L$  contenant les  $k$  évènements ayant eu les plus grandes magnitudes d'impact sur les utilisateurs. Chaque évènement dans cette liste est alors décrit par un terme principal, un ensemble pondéré de mots liés, un intervalle temporel et sa magnitude d'impact.

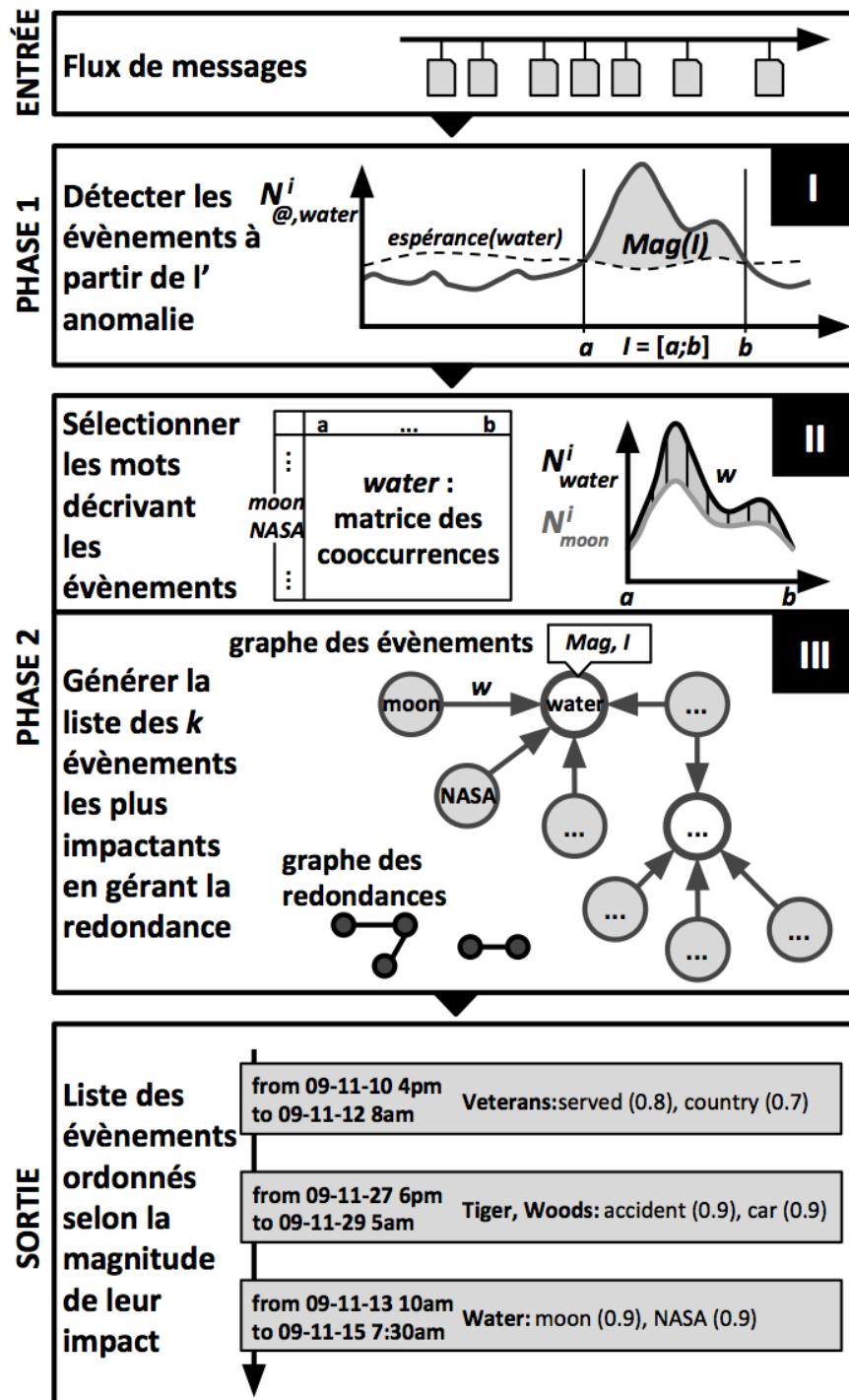


FIGURE 3.6 – Déroulement de la méthode proposée, MABED.

### 3.3.3 Détection des évènements à partir de l'anomalie dans la fréquence de création de mentions

L'objectif de ce composant est d'identifier précisément quand les évènements ont été rapportés et discutés par les utilisateurs, ainsi que d'estimer la magnitude de leur impact sur ces mêmes utilisateurs. Il se base sur l'identification de pics à partir d'une mesure d'anomalie dans la fréquence de création de mentions pour chaque mot du vocabulaire  $V_{@}$ . Les méthodes existantes supposent habituellement une durée fixée et commune à tous les évènements, ce qui n'est pas le cas de *MABED*. Dans la suite de cette section, nous décrivons comment calculer l'anomalie liée à un mot en une tranche temporelle, puis nous montrons comment mesurer la magnitude d'impact d'un mot pour une séquence contiguë de longueur quelconque de tranches temporelles. Enfin, nous expliquons comment identifier les intervalles qui maximisent la magnitude d'impact pour tous les mots de  $V_{@}$ .

**Calcul de l'anomalie en un point.** Avant de formuler la mesure d'anomalie, nous définissons le nombre espéré de messages contenant le mot  $t$  et au moins une mention pour chaque tranche temporelle  $i \in [1; n]$ , en supposant que ce mot ne soit lié à aucun évènement. Pour cela, nous supposons que le nombre de tels messages à la  $i^{\text{ème}}$  tranche temporelle,  $N_{@t}^i$ , suit un modèle génératif probabiliste. Ainsi il est possible de calculer la probabilité  $P(N_{@t}^i)$  d'observer  $N_{@t}^i$ . Pour un corpus suffisamment grand, il semble raisonnable de modéliser ce type de probabilité avec une loi binomiale (*Fung et al.*, 2005). Par conséquent nous pouvons écrire :

$$P(N_{@t}^i) = \binom{N^i}{N_{@t}^i} p_{@t}^{N_{@t}^i} (1 - p_{@t})^{N^i - N_{@t}^i}$$

où  $p_{@t}$  est la probabilité qu'un message contienne le mot  $t$  et au moins une mention, quelle que soit la tranche temporelle. Comme le nombre de messages  $N^i$  est grand dans le contexte des médias sociaux, nous pouvons raisonnablement supposer que  $P(N_{@t}^i)$  peut être approximée par une loi normale, c'est-à-dire :

$$P(N_{@t}^i) \sim \mathcal{N}(N^i p_{@t}, N^i p_{@t} (1 - p_{@t}))$$

Il en découle que la quantité espérée de messages contenant le mot  $t$  et au moins

une mention à la  $i^{\text{ème}}$  tranche temporelle est :

$$E[t|i] = N^i p_{@t}, \text{ où } p_{@t} = N_{@t}/N$$

Enfin, nous définissons l'anomalie dans la fréquence de création de mentions liée au mot  $t$  à la  $i^{\text{ème}}$  tranche temporelle comme suit :

$$\text{anomalie}(t, i) = N_{@t}^i - E[t|i]$$

Avec cette formulation, l'anomalie est positive uniquement lorsque la fréquence observée de création de mentions est strictement supérieure à l'espérance. Les mots liés à des événements et spécifiques à une période temporelle particulière auront tendance à avoir des valeurs d'anomalie positives élevées durant cette période. Au contraire, les mots récurrents (*i.e.* triviaux) qui ne sont pas liés à un événement auront des valeurs d'anomalie qui divergeront peu par rapport à l'espérance. Par ailleurs, contrairement à des approches plus sophistiquées comme par exemple la modélisation des fréquences à l'aide de mixtures gaussiennes, cette formulation passe facilement à l'échelle et s'adapte donc facilement à la taille du vocabulaire.

**Calcul de la magnitude d'impact.** La magnitude d'impact,  $Mag$ , d'un événement associé à l'intervalle  $I = [a; b]$  et au mot principal  $t$  est donnée par la formule ci-dessous. Elle correspond à l'aire algébrique sous la fonction d'anomalie sur l'intervalle  $[a; b]$ .

$$\begin{aligned} Mag(t, I) &= \int_a^b \text{anomalie}(t, i) \, di \\ &= \sum_{i=a}^b \text{anomalie}(t, i) \end{aligned}$$

L'aire algébrique est obtenue en intégrant la fonction discrète d'anomalie, ce qui revient dans ce cas à une somme.

**Identification des événements.** Pour chaque mot  $t \in V_@$ , nous cherchons à iden-

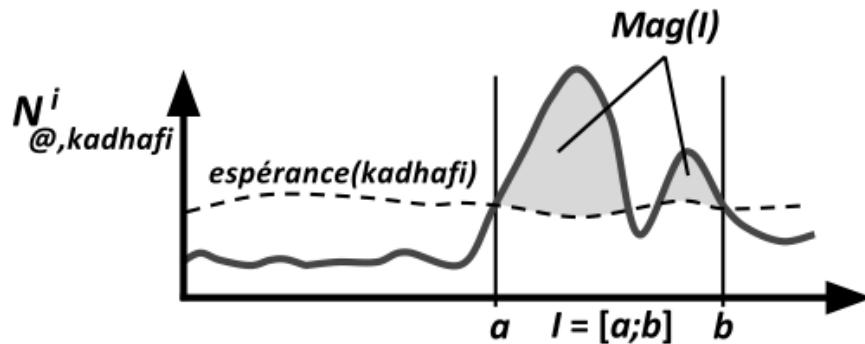


FIGURE 3.7 – Identification d'un évènement lié au mot « kadhafi ». L'aire algébrique sous la fonction d'anomalie correspond aux zones grisées.

tifier l'intervalle qui maximise la magnitude d'impact, c'est-à-dire :

$$I = \underset{I}{\operatorname{argmax}} \ Mag(t, I)$$

Or, nous avons montré précédemment que la magnitude d'impact d'un évènement décrit par le mot principal  $t$  et l'intervalle  $I = [a; b]$  correspond à la somme de l'anomalie sur cet intervalle. Par conséquent, cela revient à résoudre un problème du type « Sous-séquence contiguë de somme maximale » (SSCSM), un type de problème courant en fouille de flots de données (Lappas et al., 2009), qui trouve également des applications dans divers domaines tels que la bio-informatique (Fan et al., 2003) ou la fouille de règles d'associations (Fukuda et al., 1996). En d'autres termes, pour un mot  $t$ , nous cherchons à identifier l'intervalle  $I = [a; b]$  tel que :

$$Mag(t, I) = \max \left\{ \sum_{i=a}^b \text{anomalie}(t, i) \mid 1 \leq a \leq b \leq n \right\}$$

Cette formulation permet à l'anomalie d'être négative en certains points de l'intervalle, si et seulement si cela permet d'étendre l'intervalle tout en augmentant la magnitude. C'est une propriété intéressante, puisque cela permet d'éviter la fragmentation de longs évènements s'étendant sur plusieurs jours et dont l'anomalie associée devient négative par exemple la nuit, du fait du faible niveau d'activité nocturne sur

le média social étudié. Une autre propriété intéressante de cette formulation est qu'un mot donné ne peut être considéré comme le mot principal que d'un seul évènement. Cela augmente la lisibilité des résultats pour la raison suivante. Plus le nombre d'évènements pouvant être décrits par un même mot est grand, moins ce mot est spécifique à chaque évènement. Par conséquent, ce mot devrait plutôt être considéré comme un mot lié que comme un mot principal. La figure 3.7 montre comment un évènement lié au mot « kadhafi » est détecté. On observe que l'anomalie négative au milieu de l'intervalle identifié est compensée par l'anomalie positive mesurée sur la deuxième partie de l'intervalle.

Nous résolvons ce problème à l'aide de l'algorithme en temps linéaire décrit par Bentley (1984). Finalement, chaque évènement détecté suivant ce processus est décrit par : (i) un mot principal  $t$ , (ii) une période de temps  $I$  et (iii) la magnitude de son impact sur le comportement des utilisateurs,  $Mag(t, I)$ .

### 3.3.4 Sélection des mots décrivant les évènements

Partant du constat que les méthodes à base de clustering tendent à produire des descriptions longues et bruitées, nous adoptons une approche différente, que nous décrivons par la suite, avec pour objectif de produire des descriptions sémantiquement plus claires.

Dans le but de limiter la surcharge informationnelle, nous choisissons de limiter le nombre de mots utilisés pour décrire un évènement. Cette limite est un paramètre fixé manuellement noté  $p$ . Ce choix se justifie par la brièveté des messages publiés sur les médias sociaux. En effet, comme ces messages ne contiennent que peu de mots, il ne semble pas rationnel qu'un évènement soit décrit par un grand nombre de mots (Weng et Lee, 2011).

**Identification de mots candidats.** L'ensemble des mots candidats pour décrire un évènement est l'ensemble des mots avec les  $p$  plus fortes cooccurrences avec le mot  $t$  durant la période de temps  $I$ . Les mots les plus pertinents sont sélectionnés parmi les candidats selon la similarité entre leur dynamique temporelle et celle du mot  $t$  durant l'intervalle  $I$ . Pour ce faire, nous calculons un poids  $w_q$  pour chaque mot candidat  $t'_q$ . Nous proposons d'estimer ce poids à partir des séries temporelles  $N_t^i$  et  $N_{t'_q}^i$  et du coefficient de corrélation proposé par Erdem et al. (2012). Ce coefficient – initialement

conçu pour analyser des données boursières, réputées non-stationnaires – possède deux propriétés intéressantes pour notre application : (i) il est non-paramétrique et (ii) il ne requiert pas d'hypothèse de stationnarité contrairement, par exemple, au coefficient de Pearson. Ce coefficient prend en compte le décalage temporel afin de capturer au mieux la direction de la co-variation des deux séries temporelles au fil du temps. Par souci de concision, nous ne donnons ici que la formule permettant d'approximer ce coefficient, étant donné les mots  $t$ ,  $t'_q$  et l'intervalle temporel  $I = [a; b]$  :

$$\rho_{Ot,t'_q} = \frac{\sum_{i=a+1}^b A_{t,t'_q}}{(b-a-1)A_t A_{t'_q}},$$

où  $A_{t,t'_q} = (N_t^i - N_t^{i-1})(N_{t'_q}^i - N_{t'_q}^{i-1})$

$$A_t^2 = \frac{\sum_{i=a+1}^b (N_t^i - N_t^{i-1})^2}{b-a-1}$$

$$A_{t'_q}^2 = \frac{\sum_{i=a+1}^b (N_{t'_q}^i - N_{t'_q}^{i-1})^2}{b-a-1}$$

Cela correspond quasiment à l'auto-corrélation du premier ordre des séries temporelles  $N_t^i$  et  $N_{t'_q}^i$ . Erdem *et al.* (2012) fournissent la preuve que  $\rho_O$  satisfait la condition  $|\rho_O| \leq 1$  en utilisant l'inégalité de Cauchy-Schwartz. Enfin, nous définissons le poids du mot  $t'_q$  comme une fonction affine de  $\rho_O$  afin de se conformer à notre définition de thématique saillante, *i.e.*  $0 \leq w_q \leq 1$  :

$$w_q = \frac{\rho_{Ot,t'_q} + 1}{2}$$

Parce que la dynamique temporelle des mots toujours très fréquents est moins impactée par un évènement particulier, cette formulation – d'une certaine manière comme  $tf \cdot idf$  – diminue le poids des mots généralement fréquents dans le flux de messages et augmente le poids des mots qui le sont moins, *i.e.* les mots plus spécifiques.

**Sélection des mots les plus pertinents.** L'ensemble final des mots retenus pour décrire un évènement est l'ensemble  $S$ , tel que  $\forall t'_q \in S, w_q > \theta$ . Les paramètres  $p$  et  $\theta$

permettent à l'utilisateur de la méthode *MABED* d'ajuster la quantité et la granularité de l'information dont il a besoin.

### 3.3.5 Génération de la liste des événements

Chaque fois qu'un événement a été traité par le second composant, il est passé au troisième composant. Ce composant est chargé de sauvegarder la description des événements détectés tout en limitant la redondance (*i.e.* la duplication d'événements). Pour cela, il utilise deux structures de graphes : (i) le graphe des événements et (ii) le graphe des redondances. Le premier est un graphe orienté, pondéré et étiqueté qui modélise les descriptions des événements. La représentation d'un événement  $e$  dans ce graphe est comme suit. Un nœud représente le mot principal  $t$  et est étiqueté avec l'intervalle  $I$  et le score  $Mag$ . Chaque mot lié  $t'_q$  est représenté par un nœud et possède un arc dirigé vers le mot principal, dont le poids est égal à  $w_q$ . La figure 3.8 montre un graphe des événements stockant deux événements. Nous pouvons voir que l'un d'eux a pour mot principal « kadhafi » et qu'il y a quatre mots liés, à savoir « financé », « campagne », « 2007 », « sarkozy ». La seconde structure est un simple graphe non-orienté utilisé pour modéliser les éventuelles redondances entre les événements détectés, où un événement est représenté par son mot principal.

Soit  $e_1$  l'événement traité par ce composant. Tout d'abord, il vérifie s'il est redondant avec un événement déjà sauvegardé dans le graphe des événements ou non. Si ce n'est pas le cas, sa description est introduite dans le graphe des événements et le nombre d'événements distincts détectés est incrémenté d'un. Dans le cas contraire, en admettant que  $e_1$  est redondant avec l'événement  $e_0$  déjà présent dans le graphe, une arête est ajoutée dans le graphe des redondances entre les nœuds  $t_1$  et  $t_0$  (et le nombre d'événements détectés reste inchangé). Lorsque le nombre d'événements distincts détectés atteint  $k$ , le composant fusionne les événements redondants et retourne la liste contenant les  $k$  événements aux plus fortes magnitudes d'impact. Par la suite, nous décrivons comment les événements redondants sont identifiés et comment ils sont fusionnés.

**Identification des événements redondants.** L'événement  $e_1$  est considéré comme redondant avec l'événement  $e_0$  déjà représenté dans le graphe des événements si (i) les mots principaux  $t_1$  et  $t_0$  seraient mutuellement connectés et (ii) si le coefficient de

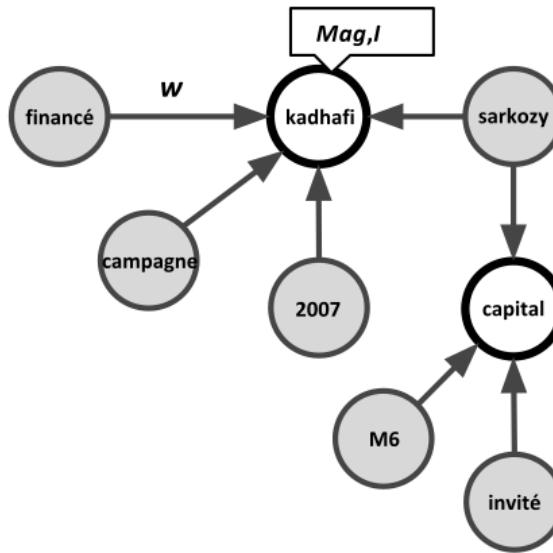
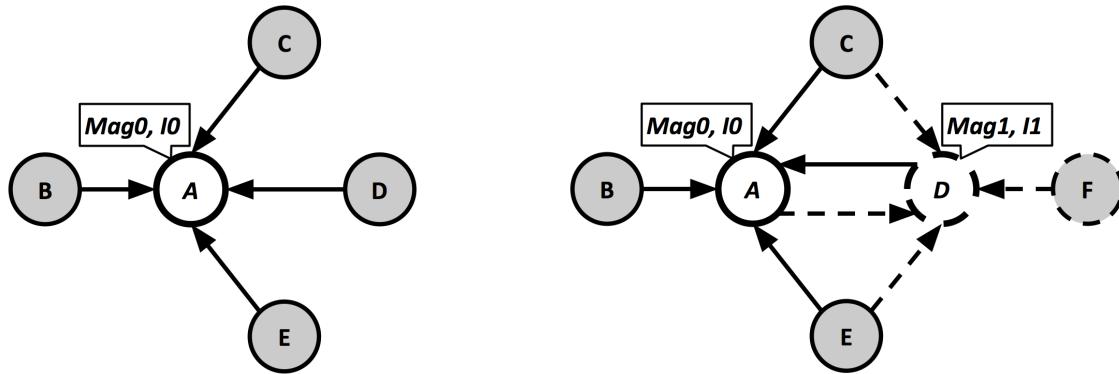


FIGURE 3.8 – Un graphe des évènements stockant deux évènements. Les mots principaux sont représentés par les noeuds blancs avec des bordures noires.

recouvrement des deux périodes de temps  $I_1$  et  $I_0$  dépasse un seuil fixe. Le coefficient de recouvrement est défini comme  $r(I_0, I_1) = \frac{|I_1 \cap I_0|}{\min(I_1, I_0)}$  et le seuil est noté  $\sigma$ ,  $\sigma \in ]0; 1]$ . Dans ce cas, la description de l'évènement  $e_1$  est sauvegardée en marge et une relation est ajoutée entre  $t_1$  et  $t_0$  dans le graphe des redondances. La partie gauche de la figure 3.9 montre la représentation graphique de l'évènement fictif  $e_0$ , dont le mot principal est noté  $A$ . Sur la partie droite de cette même figure, on observe la structure du graphe des évènements si l'on insérait un évènement fictif  $e_1$ , dont les éléments sont tracés en pointillés et dont le mot principal est noté  $D$ . Il apparaît que le mot  $D$  est un mot lié à l'évènement  $e_0$  et que  $A$  est un mot lié à l'évènement  $e_1$ . Par conséquent, si la condition  $r(I_0, I_1) > \sigma$  est vérifiée, l'évènement  $e_1$  est jugé redondant avec  $e_0$  et n'est pas inséré dans le graphe des évènements.

**Fusion des évènements redondants.** Identifier quels évènements redondants doivent être fusionnés ensemble revient à identifier les composantes connexes dans le graphe des redondances. Cela se fait en temps linéaire à l'aide de l'algorithme décrit par *Hopcroft et Tarjan* (1973). Dans chaque composante connexe se trouve exactement un noeud correspondant à un évènement représenté dans le graphe des


 FIGURE 3.9 – Identification de la redondance entre deux événements fictifs  $e_0$  et  $e_1$ .

événements. Sa magnitude d'impact (qui est nécessairement supérieure à celles des événements redondants, du fait du tri réalisé à la fin de la première phase) et son intervalle temporel restent inchangés, par contre, sa description textuelle est enrichie selon le principe suivant. La thématique décrivant cet événement est mise à jour selon les informations supplémentaires apportées par les descriptions des événements redondants. Le mot principal devient l'agrégation des mots principaux des événements redondants. Les mots liés décrivant l'événement mis à jour sont les  $p$  mots parmi tous les mots liés des événements redondants avec les  $p$  plus grands poids. En reprenant l'exemple utilisé à la figure 3.9, on fusionnerait les événements  $e_0$  et  $e_1$  comme sur la figure 3.10. Le terme principal de l'événement  $e_0$  devient l'agrégation des mots A et D, ce dernier étant le mot principal de l'événement  $e_1$ . Aussi, en supposant que  $p > 3$ , F est ajouté comme mot lié à l'événement  $e_0$ .

### 3.3.6 Algorithme général

Pour conclure cette section, nous donnons l'enchaînement des étapes que nous venons de décrire avec l'algorithme 1 (page 69).

**Algorithme 1 :** Déroulement général de la méthode MABED.

---

**Données :** Un corpus  $\mathcal{C}$  de messages partitionné en  $n$  tranches temporelles, le vocabulaire correspondant  $V_{@}$

**Paramètres :** nombre d'évènements  $k$ , limite de mots liés  $p$ , poids minimal des mots liés  $\theta \in [0; 1]$ , seuil de recouvrement temporel pour la fusion  $\sigma \in ]0; 1]$

**Résultat :** Une liste ordonnée  $L$  contenant les  $k$  évènements au plus fort impact  
/\* Phase 1 \*/

Initialiser la pile  $P$  utilisée pour stocker les évènements lors de la phase 1;

**pour chaque mot**  $t \in V_{@}$  **faire**

    Identifier l'intervalle  $I = [a; b]$  tel que :

$Mag(t, I) = \max\{\sum_{i=a}^b \text{anomalie}(t, i) | 1 \leq i \leq n\};$

    Ajouter l'évènement  $e = [t, \emptyset, I, Mag(t, I)]$  à la pile  $P$ ;

**fin**

Trier la pile d'évènements  $P$  par ordre décroissant de magnitude d'impact;

/\* Phase 2 \*/

Initialiser le graphe des évènements  $G_E$  et le graphe des redondances  $G_R$ ;

Initialiser la variable *compteur* à 0;

**tant que** *compteur*  $< k$  et  $|P| > 0$  **faire**

    Dépiler l'évènement  $e$  au sommet de la pile  $P$ ;

    Sélectionner les mots liés à l'évènement  $e$ , avec la limite  $p$  et le seuil  $\theta$ ;

**si**  $e$  est redondant avec un évènement  $e'$  présent dans  $G_E$  pour le seuil  $\sigma$  **alors**

        Ajouter une relation entre  $e$  et l'évènement  $e'$  dans  $G_R$ ;

        Sauvegarder la description de  $e$  en marge;

**sinon**

        Insérer la description de l'évènement  $e$  dans le graphe  $G_E$ ;

        Incrémenter la variable *compteur*;

**fin**

**fin**

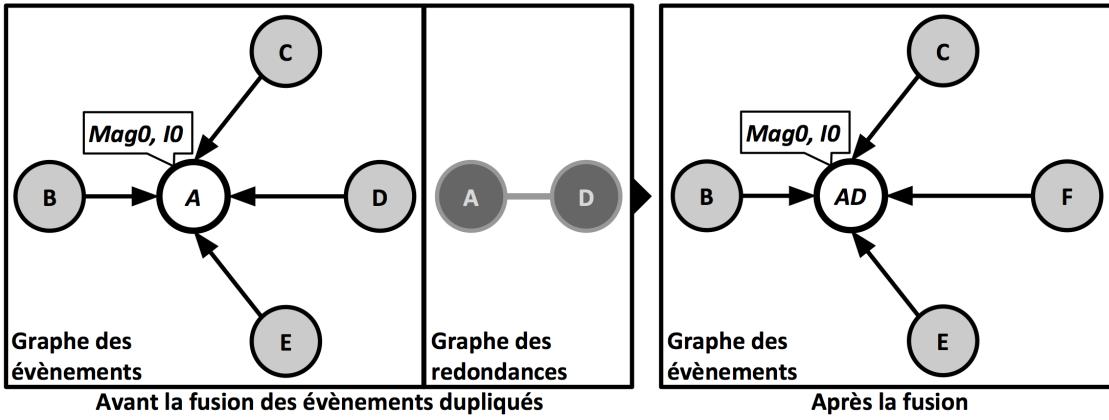
Identifier les évènements à fusionner à partir de  $G_R$  puis mettre à jour  $G_E$ ;

Transformer le graphe  $G_E$  en la liste d'évènements  $L$ ;

Trier la liste  $L$  par ordre décroissant de magnitude d'impact;

**retourner**  $L$ ;

---

FIGURE 3.10 – Résultat de la fusion entre les évènements fictifs  $e_0$  et  $e_1$ .

## 3.4 Expérimentations

Dans cette section, nous présentons la synthèse des résultats obtenus lors de l'étude expérimentale que nous avons menée à l'aide de données issues de Twitter pour évaluer *MABED*. Tout d'abord, à travers une évaluation quantitative, nous démontrons la pertinence de l'approche basée sur la mesure de l'anomalie dans la fréquence de création de mentions, et nous mesurons les performances de *MABED* par rapport à plusieurs méthodes de la littérature. Pour évaluer la précision et le rappel, nous avons demandé à des annotateurs humains de juger si les évènements détectés sont compréhensibles et significatifs. Lors de l'évaluation qualitative, nous montrons que les descriptions d'évènements détectés par *MABED* sont plus précises, temporellement et sémantiquement, que celles extraites par les méthodes existantes, ce qui favorise une compréhension aisée des résultats.

### 3.4.1 Protocole expérimental

**Corpus.** Étant donné que les corpus utilisés par les auteurs des méthodes existantes ne sont pas accessibles, nous basons nos expérimentations sur deux corpus différents. Le premier corpus – noté  $\mathcal{C}_{en}$  – contient 1 437 126 tweets rédigés en anglais, collectés avec une stratégie centrée-utilisateur par Yang et Leskovec (2011). Ils

TABLE 3.4 – Statistiques sur les corpus (@ : proportion de tweets qui contiennent des mentions, *RT* : proportion de retweets).

Corpus	# tweets	# auteurs	@	RT
$\mathcal{C}_{en}$	1 437 126	52 494	0,54	0,17
$\mathcal{C}_{fr}$	2 086 136	150 209	0,68	0,43

correspondent à l'intégralité des tweets publiés durant le mois de novembre 2009 par 52 494 utilisateurs américains de Twitter. Ce corpus contient beaucoup de bruit. Selon l'étude menée par *PearAnalytics* (2009), la proportion de tweets sans rapport avec aucun évènement pourrait atteindre 50%. Le second corpus – noté  $\mathcal{C}_{fr}$  – contient 2 086 136 de tweets rédigés en français collectés avec une stratégie centrée-mots-clés en mars 2012, durant la campagne pour l'élection présidentielle. Nous avons obtenu ces tweets via l'*API streaming* de Twitter, en utilisant les noms des principaux candidats comme mots-clés. Ce corpus cible donc les thématiques politiques en rapport avec la France. Les mots triviaux sont retirés des messages à l'aide de listes de mots vides francophone et anglophone. Les dates de publication des messages sont au format UTC. La table 3.4 donne des détails supplémentaires à propos de chaque corpus.

**Méthodes comparées.** Nous considérons d'une part une variante de la méthode *MABED* –  $\alpha$ -*MABED* – qui ignore la présence ou l'absence de mentions dans les messages. Cela signifie que le premier composant détecte les évènements et estime la magnitude de leur impact à partir des valeurs de  $N_t^i$  au lieu de  $N_{@t}^i$ . Nous considérons d'autre part deux méthodes récentes tirées de la littérature, que nous avons décrites dans la section consacrée à l'état de l'art (3.2), à savoir : *TS* et *ET*. *TS* est une méthode de pondération statistique des N-grammes développée par *Benhardus et Kalita* (2013). Nous appliquons cette méthode aux bigrammes (*TS2*) et aux trigrammes (*TS3*). *ET* (*Parikh et Karlapalem*, 2013) crée des clusters de bigrammes en réalisant une classification ascendante hiérarchique basée sur la similarité temporelle et de contenu entre bigrammes. Faute d'avoir pu obtenir l'implémentation de ces méthodes auprès des auteurs, nous les avons réimplémentées en Java. Nous avons également implémenté la méthode *EDCoW* (*Weng et Lee*, 2011), néanmoins nous avons décidé de ne pas intégrer les résultats obtenus dans la suite de cette section, ceux-ci étant inexploitables – chaque évènement étant décrit par un grand nombre de mots sans rapport entre eux.

Il faut noter que *Valkanas et Gunopoulos* (2013) ont également réimplémenté cette méthode et ont obtenu des résultats médiocres en l'appliquant à des données extraites de Twitter, puisque la précision qu'ils mesurent est de 1/588. Nous excluons également la comparaison avec les méthodes basées sur la modélisation des thématiques latentes, du fait des temps de calcul prohibitifs qui les rendent pratiquement inutilisables. À titre d'exemple, l'implémentation parallélisée d'*On-line LDA* fournie par *Lau et al.* (2012)<sup>4</sup> ne parvient à traiter que la moitié des messages contenus dans le corpus  $\mathcal{C}_{en}$  (autrement dit les messages publiés durant les 15 premiers jours du mois de novembre 2009, environ 700 000) après un mois de calcul monopolisant totalement les ressources de la machine utilisée.

**Choix des paramètres.** Pour *MABED* et  $\alpha$ -*MABED* nous partitionnons les deux corpus en tranches temporelles de 30 minutes, ce qui permet une bonne précision temporelle tout en maintenant un nombre de tweets suffisant en chaque tranche temporelle. Les paramètres  $p$  et  $\theta$  – respectivement le nombre maximum de mots décrivant chaque évènement et le poids minimum des mots liés – permettent aux utilisateurs de *MABED* de définir le niveau de détail requis. Étant donné que les messages publiés sur Twitter contiennent en moyenne 10,7 mots par phrase d'après l'étude menée par *Oxford* (2009), nous fixons  $p = 10$ . Pour les besoins de cette évaluation, nous fixons  $\theta = 0,7$  de telle sorte que seuls les mots spécifiques à chaque évènement soient présentés aux annotateurs. Il y a un paramètre qui peut affecter les performances de *MABED*, à savoir  $\sigma$ , le seuil de recouvrement temporel des évènements à fusionner. Par la suite, nous présentons les résultats pour  $\sigma = 0,5$  (nous examinons l'impact de  $\sigma$  dans la section 3.4.2).

Pour *ET* et *TS*, puisque ces méthodes supposent une durée fixée commune à tous les évènements – qui correspond à la durée d'une tranche temporelle – nous partitionnons les corpus par tranches de 24 heures, comme cela est typiquement fait dans la littérature. *ET* a deux paramètres, pour lesquels nous utilisons les valeurs optimales indiquées par les auteurs.

**Métriques d'évaluation.** En l'absence de vérité terrain, nous avons demandé à deux annotateurs humains<sup>5</sup> de juger si les évènements détectés sont compréhensibles et significatifs, en attribuant à chaque évènement une note de 0 (*i.e.* non significatif)

---

4. L'implémentation est téléchargeable à l'adresse suivante : [http://ww2.cs.mu.oz.au/~tim/etc/online\\_lda.zip](http://ww2.cs.mu.oz.au/~tim/etc/online_lda.zip).

5. Les annotateurs sont des étudiants de master et ne sont pas autrement impliqués dans ce projet.

ou 1 (*i.e.* significatif). Un évènement est considéré comme étant significatif s'il pourrait être repris par les médias traditionnels (*e.g.* via la publication d'un article à son sujet). Globalement, un évènement détecté est significatif si les deux annotateurs lui ont attribué la note de 1. Les annotateurs indiquent également si un évènement est redondant avec un évènement qu'ils ont précédemment annoté dans la même liste. Étant donné que chaque corpus couvre une période d'un mois et que l'annotation des évènements détectés est une tâche chronophage, nous limitons l'évaluation aux 40 évènements les plus impactants détectés par chaque méthode (*i.e.*  $k = 40$ ) dans chaque corpus. Nous mesurons la précision comme la fraction d'évènements détectés notés 1 par les deux annotateurs, c'est-à-dire :

$$P = \frac{\text{nombre d'évènements détectés significatifs}}{k}$$

Nous mesurons le rappel comme la fraction d'évènements significatifs distincts parmi tous les évènements détectés, de la même façon que *Li et al.* (2012) :

$$R = \frac{\text{nombre d'évènements détectés significatifs} - \text{nombre d'évènements redondants}}{k}$$

Nous mesurons également le *DERate*, qui correspond à la proportion d'évènements redondants parmi tous les évènements significatifs détectés (*Li et al.*, 2012), autrement dit :

$$DERate = \frac{\text{nombre d'évènements redondants}}{\text{nombre d'évènements détectés significatifs}}$$

### 3.4.2 Évaluation quantitative

Par la suite, nous examinons les performances des différentes méthodes sur la base des notes données par les annotateurs. L'accord inter-annotateurs mesuré selon le Kappa de Cohen vaut  $\kappa = 0,76$ , ce qui dénote un fort accord. La table 3.5 (page 75) reporte la précision, la F-mesure définie comme la moyenne harmonique entre précision et rappel, le *DERate* et le temps de calcul (temps moyen pour trois exécutions) de chaque méthode pour chaque corpus.

**Performances de *MABED* par rapport à celles des méthodes de référence.** La

table 3.5 (page 75) indique que *MABED* obtient les meilleures performances concernant la précision et la F-mesure, avec une précision de 0,775 pour une F-mesure de 0,682 sur le corpus  $\mathcal{C}_{en}$ , et une précision de 0,825 pour une F-mesure de 0,825 sur le corpus  $\mathcal{C}_{fr}$ . Par rapport à  $\alpha$ -*MABED*, nous observons un gain relatif moyen concernant la F-mesure de 17,2%. Cela vérifie de manière empirique notre principale hypothèse, dans le cadre de Twitter, à savoir que la prise en compte de la fréquence de création de mentions liées aux événements conduit à une détection plus précise des événements significatifs. Globalement, nous observons que *MABED* surpassé les méthodes de référence avec une marge plus importante du point de vue de la F-mesure sur  $\mathcal{C}_{en}$  que sur  $\mathcal{C}_{fr}$ . Or,  $\mathcal{C}_{en}$  contient beaucoup plus de bruit que  $\mathcal{C}_{fr}$ . Cela suggère que prendre en compte le comportement des utilisateurs des médias sociaux en matière de création de mentions permet une détection plus robuste des événements à partir d'un flux de tweets bruité. Le *DERate* révèle que *MABED* n'a dédoublé aucun événement significatif parmi ceux détectés dans  $\mathcal{C}_{fr}$ , mais que – en dépit de la gestion explicite de la redondance par le troisième composant – 6 ( $DERate = 0,193$ ) des 31 ( $P = 0,775$ ) événements significatifs détectés dans  $\mathcal{C}_{en}$  sont redondants. Ce *DERate* reste toutefois inférieur à celui mesuré pour les méthodes *TS2* ou *TS3*, et *MABED* obtient néanmoins le meilleur rappel sur ce corpus.

**Explication de la performance de *MABED*.** Il apparaît que les événements significatifs détectés par les méthodes de référence sont un sous-ensemble de ceux détectés par *MABED*. L'analyse plus approfondie des résultats d' $\alpha$ -*MABED*, *TS2* et *TS3* révèle que la plupart des événements jugés non-significatifs sont aisément assimilables à du spam. La non-détection de ces événements non-significatifs par *MABED* suggère que la prise en compte des mentions limite la sensibilité au spam, ce qui expliquerait en partie l'amélioration plus importante de la F-mesure de *MABED* sur  $\mathcal{C}_{en}$  que  $\mathcal{C}_{fr}$  par rapport aux méthodes de référence. En ce qui concerne *ET*, nous remarquons que la longueur moyenne des descriptions des événements est de 17,25 bigrammes (i.e. plus de 30 mots). Nous constatons que les descriptions des événements détectés par cette méthode à base de classification non supervisée sont bruitées. Les descriptions des événements non-significatifs sont essentiellement des ensembles de bigrammes sans rapport les uns avec les autres, dont il est impossible d'extraire le moindre sens. Comme le font remarquer *Valkanas et Gunopoulos* (2013), cela s'explique par une stratégie de regroupement des termes trop aggressive.

TABLE 3.5 – Performances des cinq méthodes sur les deux corpus.

Méthode	Corpus : $\mathcal{C}_{en}$			
	Précision	F-mesure	DERate	Temps de calcul
<i>MABED</i>	0,775	0,682	0,193	96s
$\alpha$ - <i>MABED</i>	0,625	0,571	0,160	126s
<i>ET</i>	0,575	0,575	0	3480s
<i>TS2</i>	0,600	0,514	0,250	80s
<i>TS3</i>	0,375	0,281	0,4	82s

Méthode	Corpus : $\mathcal{C}_{fr}$			
	Précision	F-mesure	DERate	Temps de calcul
<i>MABED</i>	0,825	0,825	0	88s
$\alpha$ - <i>MABED</i>	0,725	0,712	0,025	113s
<i>ET</i>	0,700	0,674	0,071	4620s
<i>TS2</i>	0,725	0,671	0,138	69s
<i>TS3</i>	0,700	0,616	0,214	74s

**Efficacité.** En ce qui concerne l’efficacité, il apparaît que les temps de calcul de *MABED* sont du même ordre que ceux de *TS*, tandis que les temps de calcul d’*ET* sont notoirement plus longs. Nous remarquons également que *MABED* est plus rapide que  $\alpha$ -*MABED*. La principale raison à cela est que  $|V_{@}| \leq |V|$ , ce qui accélère la première phase du traitement. Il est à noter que les temps indiqués dans la table 3.5 n’incluent pas le temps nécessaire à la préparation des données, c’est-à-dire préparer les vocabulaires et indexer les messages pour extraire les fréquences des termes.

**Impact de  $\sigma$  sur *MABED*.** Lors de la construction de la liste d’évènements par *MABED*, le seuil de recouvrement  $\sigma$  contrôle la sensibilité aux évènements redondants. La figure 3.11 donne la précision, la F-mesure et le *DERate* obtenus par *MABED* sur le corpus  $\mathcal{C}_{en}$  pour des valeurs de  $\sigma$  allant de 0,2 à 1 par pas de 0,1. On peut observer que la valeur de  $\sigma$  agit principalement sur le *DERate*. Plus spécifiquement, le *DERate* augmente avec  $\sigma$ , puisque de moins en moins d’évènements redondants sont fusionnés. Pour  $\sigma = 1$ , la précision augmente à 0,825 du fait de la proportion importante d’évènements significatifs redondants. Globalement, il apparaît que la meilleure F-mesure est atteinte pour des valeurs de  $\sigma$  allant de 0,2 à 0,5. Malgré cela, même en fixant  $\sigma = 1$ , *MABED* atteint une F-mesure de 0,582, ce qui est supérieur à la

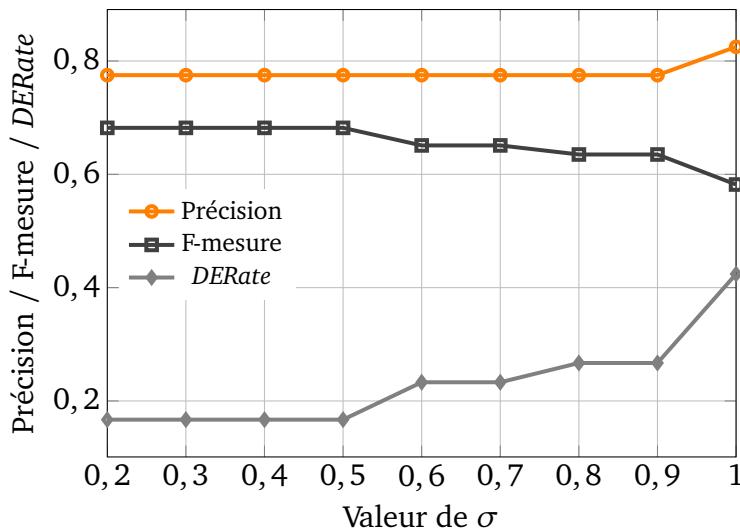


FIGURE 3.11 – Précision, F-mesure et *DERate* de *MABED* sur le corpus  $\mathcal{C}_{en}$  pour différentes valeurs de  $\sigma$ .

F-mesure obtenue avec les méthodes comparées.

### 3.4.3 Évaluation qualitative

Maintenant, nous analysons d'un point de vue qualitatif les résultats de *MABED* et montrons en quoi la méthode fournit des informations pertinentes à propos des événements détectés. La table 3.6 (page 77) liste les 20 événements avec les plus fortes magnitudes d'impact sur les utilisateurs dans  $\mathcal{C}_{en}$ . Nous basant sur cette liste, nous faisons plusieurs observations ayant trait à trois aspects : lisibilité, précision temporelle et redondance.

**Lisibilité.** La mise en avant des mots principaux facilite la lecture des descriptions des événements, d'autant plus qu'ils correspondent souvent à des entités nommées, *e.g.* Fort Hood (événement #6), Chrome (événement #7), Tiger Woods (événement #8), Obama (événement #13). La distinction entre mots principaux et mots liés favorise la compréhension rapide des événements en mettant en lumière les personnages/produits/lieux clés, ce que les méthodes existantes ne permettent pas directement. Qui plus est, la pondération des mots liés et la limitation de leur nombre améliore la lisibilité des résultats, notamment en comparaison avec les méthodes à

TABLE 3.6 – Liste des 20 évènements ayant eu le plus fort impact sur les utilisateurs, détectés par *MABED* à partir du corpus  $\mathcal{C}_{en}$ . Les mots principaux sont en gras et le poids de chaque mot lié est donné entre parenthèses. Les intervalles temporels sont exprimés en temps UTC.

#	Intervalle	Thématique
1	du 25 09h30 au 28 06h30	<b>thanksgiving, turkey</b> : hope (0.72), happy (0.71) <i>Les twittos célèbrent Thanksgiving</i>
2	du 25 09h30 au 27 09h00	<b>thankful</b> : happy (0.77), thanksgiving (0.71) <i>Lié à l'évènement # 1</i>
3	du 10 16h00 au 12 08h00	<b>veterans</b> : served (0.80), country (0.78), military (0.73), happy (0.72) <i>Commémoration du 11 novembre, « Veterans Day »</i>
4	du 26 13h00 au 28 10h30	<b>black</b> : friday (0.95), amazon (0.75) <i>Les twittos discutent des offres proposées par Amazon la veille du « Black Friday »</i>
5	du 07 13h30 au 09 04h30	<b>her, bill, health, house, vote</b> : reform (0.92), passed (0.91), passes (0.88) <i>La Chambre des représentants des États-Unis adopte la réforme de santé</i>
6	du 05 19h30 au 08 09h00	<b>hood, fort</b> : ft (0.92), shooting (0.83), news (0.78), army (0.75), forthood (0.73) <i>Une fusillade a lieu dans l'enceinte de la base militaire américaine de Fort Hood</i>
7	du 19 04h30 au 21 02h30	<b>chrome</b> : os (0.95), google (0.87), desktop (0.71) <i>Google rend public le code source de Chrome OS pour PC</i>
8	du 27 18h00 au 29 05h00	<b>tiger, woods</b> : accident (0.91), car (0.88), crash (0.88), injured (0.80) <i>Tiger Woods est victime d'un accident de la route</i>
9	du 28 22h30 au 30 23h30	<b>tweetie, 2.1, app</b> : retweets (0.93), store (0.90), native (0.89), geotagging (0.88) <i>L'application Tweetie sort sur l'apple store et inclut de nouvelles fonctions, e.g. retweets</i>
10	du 29 17h00 au 30 23h30	<b>monday, cyber</b> : deals (0.84), pro (0.75) <i>Les twittos partagent les offres commerciales high-tech du « Cyber Monday »</i>
11	du 10 01h00 au 12 03h00	<b>linkedin</b> : synced (0.86), updates (0.84), status (0.83), twitter (0.71) <i>Linkedin permet à ses utilisateurs de synchroniser leurs statuts avec Twitter</i>
12	du 04 17h00 au 06 05h30	<b>yankees, series</b> : win (0.84), won (0.84), fans (0.78), phillies (0.73), york (0.72) <i>Les Yankees, l'équipe de baseball de New York remporte la World Series face aux Phillies</i>
13	du 15 09h00 au 17 23h30	<b>obama</b> : chinese (0.75), barack (0.72), twitter (0.72), china (0.70) <i>Lors d'une visite en Chine, Barack Obama admet n'avoir jamais utilisé Twitter</i>
14	du 25 10h00 au 26 10h00	<b>holiday</b> : shopping (0.72) <i>Les twittos réagissent par rapport au « Black Friday », un jour férié dédié au shopping</i>
15	du 19 21h30 au 21 16h00	<b>oprah, end</b> : talk (0.81), show (0.79), 2011 (0.73), winfrey (0.71) <i>Oprah Winfrey annonce la fin de son talk-show en septembre 2011</i>
16	du 07 11h30 au 09 05h00	<b>healthcare, reform</b> : house (0.91), bill (0.88), passes (0.83), vote (0.83) <i>Lié à l'évènement #5</i>
17	du 11 03h30 au 13 08h30	<b>facebook</b> : app (0.74), twitter (0.73) <i>Pas d'évènement correspondant</i>
18	du 18 14h00 au 21 03h00	<b>whats</b> : happening (0.76), twitter (0.73) <i>Twitter demande maintenant « What's happening ? » et plus « What are you doing ? »</i>
19	du 20 10h00 au 22 00h00	<b>cern</b> : lhc (0.86), beam (0.79) <i>Les faisceaux de particules circulent à nouveau dans l'accélérateur LHC du CERN</i>
20	du 26 08h00 au 26 15h00	<b>icom</b> : lisbon (0.99), roundtable (0.98), national (0.88) <i>Tenue de la table ronde de l'ICOM à propos des marchés financiers portugais</i>

base de clustering qui ont tendance à extraire des descriptions très longues.

**Précision temporelle.** *MABED* estime dynamiquement la période de temps durant laquelle chaque évènement est discuté. Cela améliore la précision temporelle par rapport aux méthodes existantes qui fixe typiquement la durée des évènements à une journée. Nous illustrons en quoi cela améliore la qualité des résultats avec l'exemple suivant. Le 6<sup>ème</sup> évènement de la table 3.6 correspond aux utilisateurs de Twitter signalant la fusillade qui a eu lieu à la base militaire américaine de *Fort Hood* le 5 Novembre 2009 entre 13h34 et 13h44 CST<sup>6</sup> (*i.e.* 19h34 et 19h44 UTC). Le pic d'activité engendré par cet évènement est détecté par *MABED* dès la tranche temporelle couvrant la période 19h30-20h00 UTC du 5 novembre. *MABED* donne la description suivante :

(i) du 05/11 19h30 au 08/11 9h00 (UTC) ; (ii) hood, fort ; (iii) ft (0,92), shooting (0,83), news (0,78), army (0,75), forthood (0,73).

À la lecture de cette description, nous pouvons clairement comprendre (i) qu'il s'est passé quelque chose aux alentours de 19h30 UTC le 5 novembre, (ii) à la base de Fort Hood et (iii) qu'il s'agit d'une fusillade (*i.e.* shooting). Par contraste,  $\alpha$ -*MABED* ne détecte cet évènement que le 7 novembre, lorsque la couverture médiatique était la plus importante.

**Redondance.** Certains évènements sont associés à plusieurs mots principaux. C'est le cas des évènements #1 (Thanksgiving, turkey), 5 (HCR, bill, health, house, vote), 6 (Hood, Fort) et 8 (Tiger, Woods) entre autres. Ceci est dû aux fusions opérées par le troisième composant de la méthode *MABED* pour éliminer les duplicita. La redondance est également limitée grâce à l'estimation dynamique de la période durant laquelle chaque évènement est discuté. Nous illustrons cela toujours à l'aide du même exemple, à savoir l'évènement #6. La figure 3.12 (page 79) montre l'évolution de l'anomalie mesurée pour les mots « hood », « fort » et « shooting » entre le 5 et le 7 novembre 2009. Nous pouvons voir que l'anomalie est proche de 0 durant la nuit, donnant une allure à « deux pics » à la courbe. Néanmoins, *MABED* détecte un unique évènement discuté pendant plusieurs jours, plutôt que de reporter deux évènements distincts, durant chacun une journée. L'importance de l'estimation dynamique de la durée durant laquelle chaque évènement est discuté est renforcée par la figure 3.13 (page 80), qui montre la distribution de la durée des évènements détectés dans

---

6. [http://en.wikipedia.org/wiki/Fort\\_Hood\\_shooting](http://en.wikipedia.org/wiki/Fort_Hood_shooting)

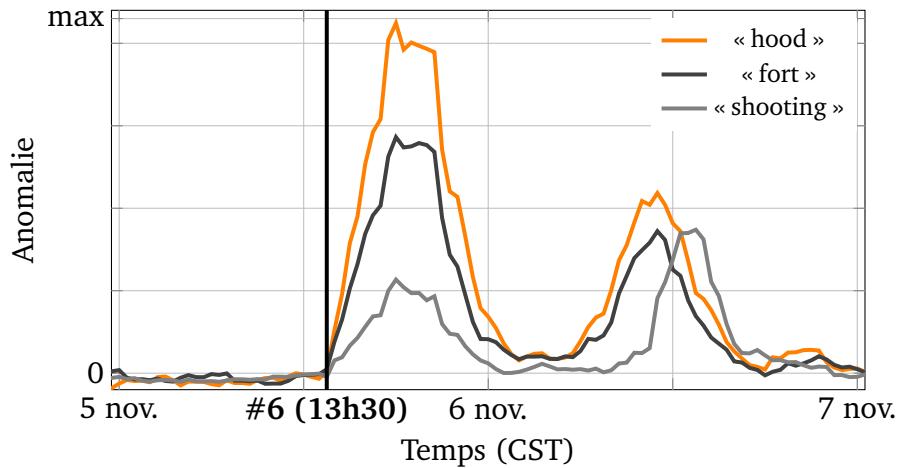


FIGURE 3.12 – Anomalie mesurée pour les mots « hood », « fort » et « shooting » du 5 au 7 novembre à minuit (CST).

chaque corpus. Nous observons que certains évènements sont discutés pendant moins de 12 heures (ce qui est le cas uniquement pour 3 évènements détectés à partir de  $\mathcal{C}_{en}$ ), tandis que d'autres sont discutés pendant plus de 60 heures. Qui plus est, nous notons que les durées des évènements détectés à partir de  $\mathcal{C}_{fr}$  sont normalement distribuées – au sens du test de *Shapiro et Wilk* (1965), avec  $W = 0.92$ , une  $p$ -valeur de 0.54 et un seuil de signification à 0.05 – et ont tendance à être plus longues que celles des évènements détectés à partir de  $\mathcal{C}_{en}$ . Autrement-dit, les évènements en lien avec la campagne électorale ont tendance à animer les discussions plus longtemps. Cela coïncide avec les résultats de l'étude empirique menée par *Romero et al.* (2011), qui indiquent que les thématiques controversées et notamment les thématiques politiques persistent plus longtemps sur Twitter que les autres.

### 3.5 Implémentation et visualisations

L'implémentation de la méthode *MABED* est disponible sur internet<sup>7</sup> et est également incluse dans le logiciel *SONDY*, qui sera détaillé ultérieurement dans ce manuscrit. Pour permettre une exploration efficace des évènements détectés, l'implé-

7. <http://mediamining.univ-lyon2.fr/people/guille/mabed.php>

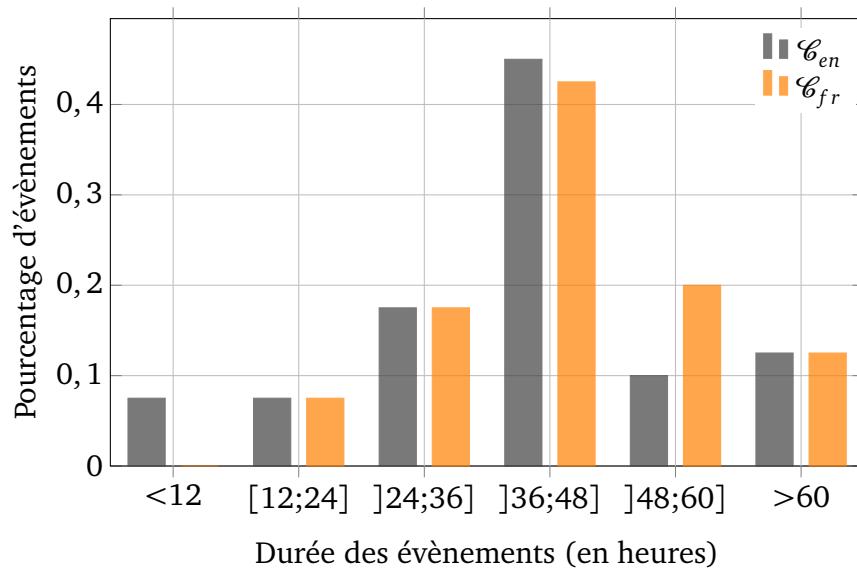


FIGURE 3.13 – Distribution de la durée des évènements détectés par *MABED*.

mentation de *MABED* propose trois visualisations que nous décrivons ci-après. Nous proposons une interface pour chacune des dimensions caractérisant les évènements : temps, impact et thématique.

**Temps.** La figure 3.14 montre une capture d’écran de l’interface centrée-temps, qui est une frise chronologique. Le ruban dans la partie inférieure permet la navigation à travers le temps tandis que la partie supérieure donne des détails à propos de l’évènement sélectionné. La description de chaque évènement est enrichie d’une image et d’une URL. Ces deux éléments correspondent à la première image et à la première URL renvoyées par le moteur de recherche Bing, en utilisant les descriptions des évènements extraites à partir des messages par *MABED* comme requêtes. La frise chronologique reprise par la figure 3.14 a été générée à partir des évènements détectés dans  $\mathcal{C}_{fr}$ .

**Impact.** La figure 3.15 montre une capture d’écran de l’interface centrée-impact. Elle consiste en un graphique interactif qui permet de visualiser et comparer l’évolution dans le temps de l’impact des évènements sur les utilisateurs du média social étudié. Ce graphique a été généré à partir des cinq évènements au plus fort impact détectés durant la journée du 30 mai 2014 par la version en ligne de *MABED* qui analyse en permanence les tweets francophones mentionnant François Hollande.

## Détecter les évènements

---



FIGURE 3.14 – Capture d’écran de l’interface centrée-temps. Elle se décompose en deux parties : la partie inférieure permet la navigation à travers le temps tandis que la partie supérieure donne des détails à propos des évènements.

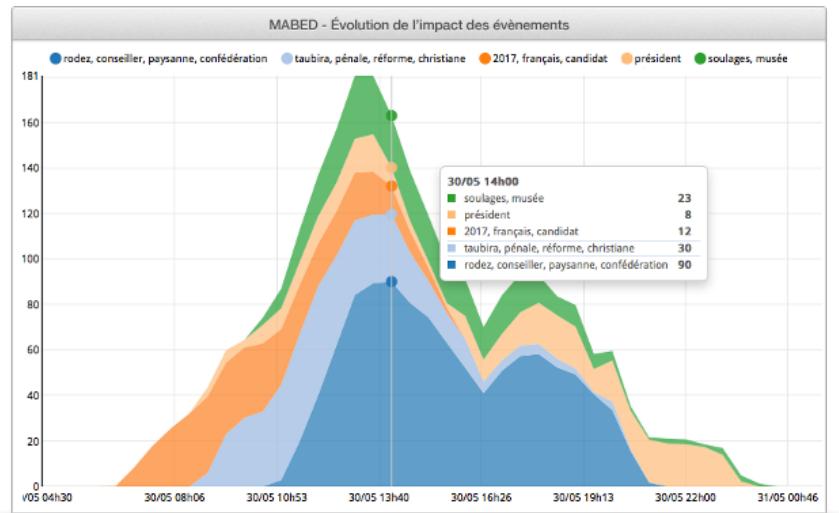


FIGURE 3.15 – Capture d’écran de l’interface centrée-impact. Chaque évènement est associé à une couleur.

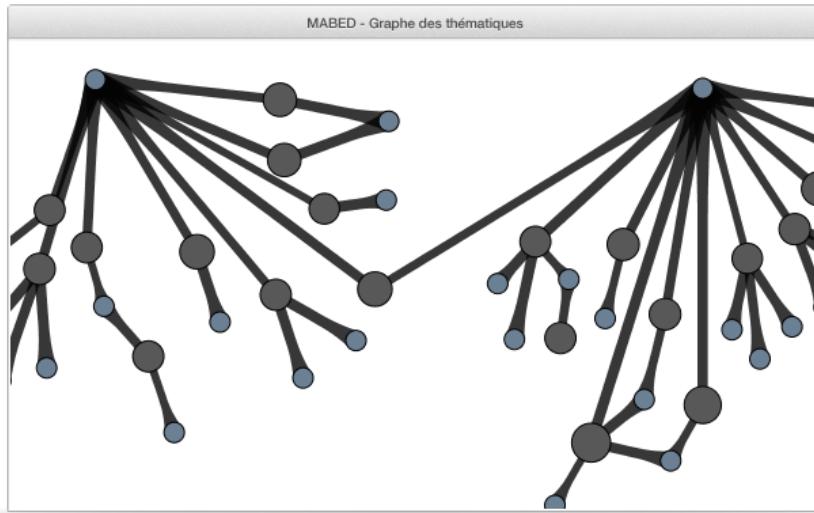


FIGURE 3.16 – Capture d’écran de l’interface centrée-thématique. Les nœuds gris correspondent aux mots principaux. Leur diamètre est proportionnel à l’impact des évènements.

**Thématique.** La figure 3.16 montre une capture d’écran de l’interface centrée-thématique. Elle permet de naviguer dans le graphe des évènements construit par le troisième composant de *MABED*, ce qui facilite l’identification des évènements proches. Pour simplifier la lecture du graphe, les labels des nœuds ne s’affichent que sur demande. La figure 3.16 montre un extrait du graphe construit à partir du corpus  $\mathcal{C}_{fr}$ . Les nœuds bleutés en haut à gauche et à droite de la fenêtre correspondent respectivement à François Hollande et Nicolas Sarkozy, *i.e.* les deux principaux candidats à la présidence en 2012.

## 3.6 Discussion

Pour conclure ce chapitre, nous résumons dans un premier temps les travaux que nous venons de présenter. Ensuite, nous présentons la principale piste de recherche que nous comptons explorer dans nos futurs travaux.

### 3.6.1 Résumé des travaux présentés

Dans ce chapitre nous avons décrit *MABED*, une nouvelle méthode efficace reposant sur la mesure de l'anomalie dans la fréquence de création de mentions pour détecter les évènements dans les médias sociaux. Contrairement aux méthodes de la littérature, *MABED* prend en compte l'aspect social des flux de messages à travers les mentions que les utilisateurs insèrent dans leurs messages pour engager la discussion avec d'autres. Par ailleurs, notre approche diffère des travaux existants en ce qu'elle estime dynamiquement la période de temps durant laquelle chaque évènement est discuté, plutôt que de supposer une durée fixée et commune à tous les évènements détectés. Les expérimentations que nous avons menées sur deux jeux de données collectés sur Twitter ont démontré la pertinence de notre approche. Quantitativement parlant, *MABED* a obtenu de meilleurs résultats que  $\alpha$ -*MABED* – une variante qui ignore la présence ou l'absence de mentions dans les messages – pour tous les tests que nous avons menés, et surpassé également deux méthodes récentes tirées de la littérature. Nous avons ainsi pu valider empiriquement – sur Twitter – notre hypothèse de départ, à savoir que la prise en compte des mentions conduit à une détection plus précise des évènements significatifs. Les résultats obtenus suggèrent également que la prise en compte des mentions accroît la robustesse de la détection en présence de données très bruitées. Qualitativement parlant, nous avons montré que la distinction entre mots principaux et mots liés augmente la lisibilité des descriptions des évènements. Nous avons également mis en évidence l'intérêt des informations temporelles fournies par *MABED*. D'une part, elles permettent de savoir précisément quand les évènements se sont produits. D'autre part, l'estimation dynamique de la période temporelle durant laquelle chaque évènement est discuté par les utilisateurs permet de limiter la fragmentation et la duplication des évènements détectés.

**Impact.** Ces travaux ont notamment fait l'objet d'un article long, présenté à la conférence internationale *IEEE/ACM ASONAM* en 2014.

### 3.6.2 Perspectives de travail

Partant de l'intuition que les communautés d'utilisateurs au sens social – c'est-à-dire les communautés identifiables à partir de la structure du réseau social que forment les utilisateurs d'un média social – sont similaires aux communautés d'utili-

sateurs au sens thématique – c'est-à-dire les communautés identifiables à partir des évènements à propos desquels les utilisateurs réagissent, une perspective de travail consisterait à explorer la complémentarité entre la tâche de détection d'évènements à partir des messages et la tâche de détection de communautés d'utilisateurs à partir de la structure du réseau social. Par exemple, il serait intéressant d'exploiter la détection d'évènements pour évaluer la pertinence des communautés détectées à partir de la structure du réseau social, voire même pour améliorer la pertinence des communautés détectées.

Pour justifier l'intérêt de cette piste de recherche, nous présentons dans la suite de cette section quelques résultats préliminaires. Ces résultats se basent d'une part sur les évènements détectés à partir du corpus de messages  $\mathcal{C}_{en}$ , et d'autre part, sur le réseau social formé par les liens d'abonnement entre les auteurs des messages. Le corpus  $\mathcal{C}_{en}$  – déjà utilisé dans la section précédente – contient 1 437 126 tweets publiés en novembre 2009 par 52 494 utilisateurs. Le réseau social interconnectant ces utilisateurs comprend 5 793 961 liens d'abonnements, lesquels ont été collectés fin 2009 par Kwak *et al.* (2010).

**Identification des communautés à partir de la structure du réseau social.** Pour ce faire, nous utilisons la méthode de Louvain (Blondel *et al.*, 2008), couramment utilisée pour détecter les communautés dans les grands graphes et en particulier les réseaux sociaux. Cette méthode repose sur une heuristique d'optimisation de l'indice de modularité. Cet indice défini par Newman (2006) mesure la différence entre la densité d'une instance de graphe et la densité d'un graphe aléatoire possédant la même distribution des degrés. La méthode de Louvain identifie deux communautés :  $c_0$ , qui contient 25 625 membres et  $c_1$ , qui contient 26 869 membres.

**Détection des évènements à partir des messages.** Le partitionnement des utilisateurs en deux communautés nous amène à former deux sous-corpus,  $\mathcal{C}_{en}(c_0)$ , qui contient 479 899 messages, et  $\mathcal{C}_{en}(c_1)$ , qui contient 932 699 messages. Ces corpus correspondent respectivement aux messages publiés par les membres des communautés  $c_0$  et  $c_1$ . Nous utilisons la méthode MABED pour détecter dans chaque corpus les 10 évènements ayant eu le plus fort impact sur les membres de chaque communauté, ce qui conduit à l'élaboration de deux listes d'évènements,  $L_0$  et  $L_1$ , qui correspondent respectivement aux évènements détectés à partir de  $\mathcal{C}_{en}(c_0)$  et de  $\mathcal{C}_{en}(c_1)$ .

**Analyse des évènements détectés par communauté.** Nous associons tout d'abord

manuellement chaque évènement détecté à une des catégories proposées par *McMinn et al.* (2013) pour classifier les évènements sur Twitter, à savoir :

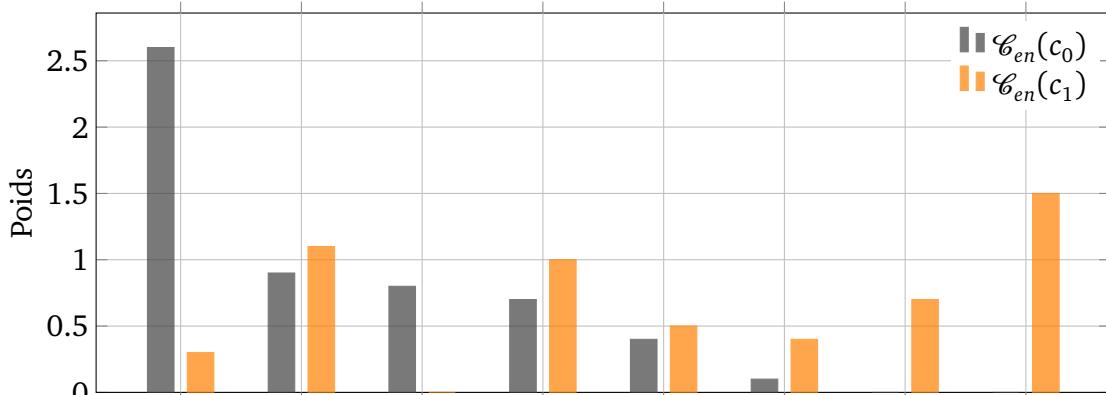
- Science et technologie ;
- Justice, politique et scandales ;
- Business et économie ;
- Art, culture et divertissement ;
- Catastrophes et accidents ;
- Sport ;
- Conflits armés et attaques ;
- Divers.

Ensuite, du fait de la différence importante entre les volumes de messages contenus dans  $\mathcal{C}_{en}(c_0)$  et  $\mathcal{C}_{en}(c_1)$ , nous proposons non pas de comparer directement les deux listes d'évènements sur la base de la magnitude d'impact des évènements par catégorie, mais plutôt en fonction de la position des évènements dans chaque liste. À cette fin, nous définissons le poids d'un évènement comme une fonction affine de sa position dans la liste :  $1 - ((\text{position} - 1) \times 0,1)$ . Ainsi le premier évènement a un poids de 1, tandis que le dernier évènement de la liste a un poids de 0,1. Enfin, pour chaque catégorie, nous sommes les poids des évènements associés.

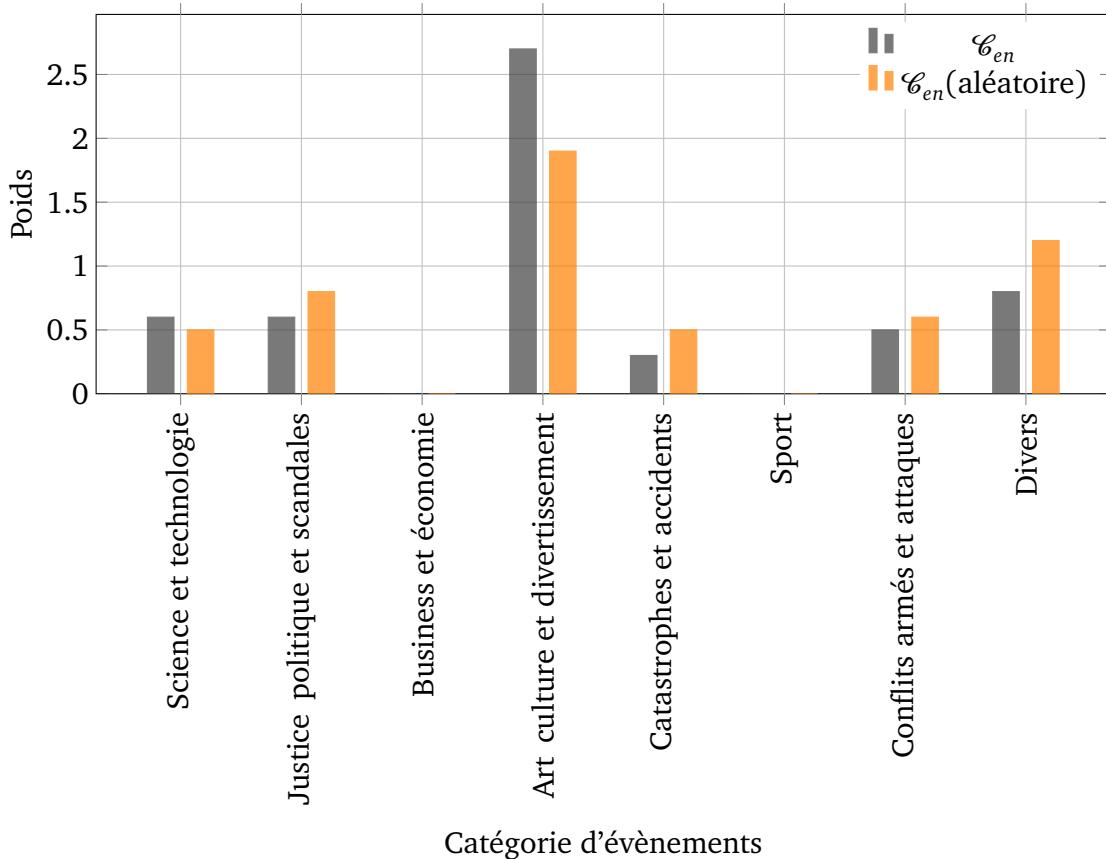
La figure 3.17.a (page 87) donne la distribution des poids des catégories des évènements détectés à partir des messages par *MABED* en fonction des communautés détectées avec la méthode de Louvain à partir de la structure du réseau social. Elle révèle que les deux distributions sont très différentes. On observe par exemple que la catégorie « Divers » a le poids le plus élevé pour la communauté  $c_1$ , tandis qu'aucun des évènements détectés pour la communauté  $c_0$  n'appartient à celle-ci. À l'inverse, la catégorie « Science et technologie » a le poids le plus important pour  $c_0$ , tandis qu'elle est la seconde catégorie au poids le plus faible pour  $c_1$ , après « Business et économie », pour laquelle seule  $c_0$  a un poids non nul. Pour renforcer ce constat, nous présentons sur la figure 3.17.b la distribution des poids des catégories des 10 évènements au plus fort impact détectés à partir des messages par *MABED* dans le corpus  $\mathcal{C}_{en}$  et le corpus  $\mathcal{C}_{en}(\text{aléatoire})$ . Ce dernier contient 725 806 tweets correspondant à tous les messages publiés par 26 000 auteurs sélectionnés aléatoirement parmi les auteurs des messages dans le corpus  $\mathcal{C}_{en}$ . Contrairement au cas précédent, on observe ici deux distributions similaires. Notamment, la catégorie au plus fort poids dans les

deux corpus est « Art, culture et divertissement », tandis que les catégories « Business et économie » et « Sport » ont des poids nuls pour les deux corpus.

**Lien entre communautés au sens social et communautés au sens thématique.** La corrélation entre les distributions des poids obtenus à partir de  $\mathcal{C}_{en}(c_0)$  et  $\mathcal{C}_{en}(c_1)$  mesurée selon le coefficient de corrélation linéaire de Bravais-Pearson (*Wilcox et Muska, 2001*) vaut  $-0,36$ , tandis que le même coefficient mesuré pour les distributions obtenues à partir de  $\mathcal{C}_{en}$  et  $\mathcal{C}_{en}(\text{aléatoire})$  vaut  $0,93$ . Cela signifie, au sens de ce coefficient, que les centres d'intérêt des deux communautés détectées selon la méthode de Louvain sont différents, voire opposés, tandis que les types d'événements détectés à partir d'un échantillon aléatoire sont très semblables à ceux détectés pour l'ensemble de la population. Dans ce cas, les événements détectés confirment la pertinence des communautés identifiées par la méthode de Louvain. Nous pouvons également imaginer la situation inverse, où les événements détectés seraient semblables pour deux ou plusieurs communautés, ce qui pourrait permettre de les fusionner ensemble. D'une part, développer une mesure de qualité du partitionnement d'un réseau social exploitant la détection d'événements permettrait d'évaluer et de comparer d'une nouvelle manière les méthodes de détection de communautés dans les réseaux sociaux. D'autre part, cela pourrait permettre d'améliorer les résultats obtenus avec des méthodes strictement basées sur la structure du réseau. Par exemple, il serait possible d'intégrer cette mesure dans la méthode de détection de communautés proposée par *Newman (2006)*, qui procède à des dichotomies récursives du réseau en optimisant la modularité, en vérifiant que les deux communautés résultant d'une dichotomie ne soient pas trop semblables du point de vue des événements qui suscitent leur intérêt. Il serait également possible d'utiliser cette métrique pour déterminer le paramètre optimal de résolution de la méthode de Louvain.



(a) Évènements détectés dans les corpus  $\mathcal{C}_{en}(c_0)$  et  $\mathcal{C}_{en}(c_1)$ .



(b) Évènements détectés dans les corpus  $\mathcal{C}_{en}$  et  $\mathcal{C}_{en}(\text{aléatoire})$ .

FIGURE 3.17 – Distribution du poids des catégories des évènements détectés par *MABED* dans les corpus  $\mathcal{C}_{en}(c_0)$ ,  $\mathcal{C}_{en}(c_1)$ ,  $\mathcal{C}_{en}$  et  $\mathcal{C}_{en}(\text{aléatoire})$



# Chapitre 4

## Modéliser et prévoir la diffusion de l'information

### Sommaire

---

4.1	Introduction	90
4.2	État de l'art	92
4.2.1	Modélisation n'exploitant pas la structure du réseau	93
4.2.2	Modélisation basée sur la structure du réseau	96
4.2.3	Synthèse de l'état de l'art	100
4.3	Méthode proposée	102
4.3.1	Formulation du problème	102
4.3.2	Vue d'ensemble de la méthode proposée	103
4.3.3	Description du modèle	105
4.3.4	Espace de représentation	107
4.3.5	Estimation des paramètres du modèle	109
4.4	Expérimentations	117
4.4.1	Protocole expérimental	117
4.4.2	Évaluation de la procédure d'estimation des probabilités de diffusion	120
4.4.3	Évaluation du modèle <i>T-BASIC</i>	122
4.4.4	Analyse des facteurs impactant la diffusion de l'information	127
4.5	Discussion	131
4.5.1	Résumé des travaux présentés	131
4.5.2	Perspectives de travail	132

---

Dans le chapitre précédent, nous avons présenté une nouvelle méthode pour détecter les évènements faisant réagir les utilisateurs des médias sociaux, de manière rétrospective. Dans ce chapitre, nous décrivons la deuxième contribution de cette thèse, qui porte sur la modélisation et la prévision de la diffusion de l'information au sein des médias sociaux, dans le but de comprendre et anticiper la réaction des utilisateurs par rapport aux évènements.

## 4.1 Introduction

La méthode présentée dans le chapitre précédent permet de détecter *a posteriori* les évènements ayant suscité l'intérêt des utilisateurs d'un média social. Néanmoins, dans certains cas, il est utile de pouvoir anticiper la réaction que les utilisateurs auront par rapport à un évènement, par exemple dans le but d'anticiper l'efficacité d'une campagne marketing virale. Dans d'autres cas, il peut être utile de prévoir l'évolution d'un phénomène de diffusion en cours, par exemple dans le but de combattre la propagation d'informations erronées. La modélisation et la prévision du phénomène de diffusion de l'information dans les médias sociaux sont des tâches qui suscitent un fort intérêt de la part des chercheurs en fouille de données. De nombreuses méthodes pour prévoir la diffusion de l'information ont été développées, qu'elles soient inspirées de modèles épidémiologiques (Leskovec *et al.*, 2007; Wang *et al.*, 2012) ou bien qu'elles soient des variantes de modèles initialement développés en marketing (Galuba *et al.*, 2010; Saito *et al.*, 2010a; Motoda, 2011). Cependant, nous en savons encore peu à propos des facteurs qui gouvernent le processus de diffusion au sein des médias sociaux, tant au niveau social, temporel que thématique. Cela nous amène donc à formuler les questions suivantes. D'abord, *quels facteurs influent sur la diffusion de l'information dans les médias sociaux ?* Puis, *comment prévoir la diffusion de l'information à partir de ces facteurs ?*

Nous sommes plus particulièrement intéressés par la prévision du volume d'utilisateurs relayant une information spécifique au sein d'un média social fondé sur un réseau social explicite, décrivant quels utilisateurs sont exposés aux messages publiés par quels utilisateurs. Ce réseau correspond par exemple, dans le cas de Twitter, au graphe des abonnements.

**Proposition et positionnement.** Nous proposons un nouveau modèle probabi-

liste pour la diffusion de l'information dans les médias sociaux, ainsi qu'une procédure pour estimer ses paramètres : *T-BASIC* (*Time-Based ASynchronous Independent Cascades*). Ce modèle repose sur la structure de réseau sous-jacente aux médias sociaux et deux paramètres pour chaque lien du réseau : la probabilité de diffusion de l'information, et le délai de transmission entre les deux utilisateurs.

Contrairement aux modèles similaires également basés sur la structure de réseau, qui reposent sur des probabilités de diffusion constantes pour chaque lien du réseau, *T-BASIC* repose sur des probabilités dépendantes du temps, ce qui permet d'intégrer la fluctuation du niveau de réceptivité des utilisateurs en fonction du temps dans la modélisation. La procédure que nous décrivons pour estimer les paramètres de *T-BASIC* diffère aussi des approches existantes. Plutôt que d'estimer directement tous les paramètres latents du modèle à partir de séquences d'activation, nous proposons d'exprimer les paramètres en fonction de caractéristiques observables des utilisateurs. Cela nous permet de réduire le coût de la phase d'estimation, puisqu'il ne s'agit plus d'estimer directement tous les paramètres du modèle – dont le nombre est proportionnel au nombre de liens dans le réseau étudié – mais d'estimer les paramètres de deux fonctions, l'une modélisant la probabilité de diffusion et l'autre le délai de transmission selon les caractéristiques des utilisateurs. Nous décrivons donc d'abord un espace de représentation des utilisateurs, composés d'attributs sociaux, thématiques et temporels, puis nous montrons comment inférer les paramètres de ces fonctions à partir des caractéristiques mesurées et de séquences d'activation. Enfin, cette approche nous permet d'analyser l'effet des facteurs sociaux, thématiques et temporels sur le processus de diffusion de l'information dans les médias sociaux.

**Résultats.** Nous menons une évaluation en deux temps par rapport à des méthodes tirées de la littérature, avec des données collectées sur le média social Twitter. D'abord, nous démontrons la validité de la démarche proposée pour estimer les paramètres de *T-BASIC*, puis nous évaluons les capacités prédictives du modèle *T-BASIC* en comparant les séries temporelles réelles et prédites modélisant la diffusion de différents éléments d'information. Par ailleurs, nous menons une analyse des facteurs influençant le phénomène de propagation de l'information à travers l'étude des paramètres de la fonction modélisant la probabilité de diffusion entre deux utilisateurs. Cette analyse révèle que les caractéristiques des utilisateurs, tant sur le plan social, que sur le plan thématique ou temporel, ont des effets importants sur la probabilité

de diffusion, dont la direction varie selon qu'elles soient mesurés par rapports aux utilisateurs exerçant l'influence, ou ceux la subissant. Par ailleurs, cette analyse indique que, pour une paire d'utilisateurs connectés ( $ux \rightarrow uy$ ), ce sont les caractéristiques de l'utilisateur  $ux$  qui influent le plus sur la probabilité de diffusion, c'est-à-dire l'utilisateur subissant l'influence de  $uy$ , plus que celles de l'utilisateur  $uy$ , qui relaie déjà l'information.

**Application.** L'implémentation du modèle *T-BASIC* est publique et son code source en Java est distribué librement<sup>1</sup>.

Ce chapitre est organisé de la manière suivante. Dans la section 4.2 nous présentons une synthèse de l'état de l'art, puis dans la section 4.3, nous décrivons formellement la méthode proposée. Ensuite, nous présentons les expérimentations que nous avons menées dans la section 4.4. Enfin, nous concluons ce chapitre et discutons des perspectives dans la section 4.5.

## 4.2 État de l'art

Les modèles pour la prévision de la diffusion de l'information dans les médias 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 de 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 deux 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 épidémiologiques 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.

---

1. <http://mediamining.univ-lyon2.fr/people/guille/tbasic.php>

#### 4.2.1 Modélisation n'exploitant pas la structure du réseau

**Modèles classiques.** Nous décrivons ici des modèles développés en épidémiologie, dits modèles compartimentaux, conçus pour modéliser la diffusion d'une maladie au sein d'une population constante de  $N$  individus. Ils supposent d'une part que les contacts entre les  $N$  individus se font aléatoirement, et d'autre part que les membres se trouvent dans des états particuliers (dus à la diffusion), ce qui permet de les « compartimenter ». On parle de modélisation sous forme de « mélange homogène » puisque ces modèles considèrent d'une part que les individus d'un même compartiment sont connectés selon une structure régulière avec les individus des autres compartiments, et d'autre part que les individus changent de compartiments de façon homogène. Les modèles compartimentaux caractérisent le processus de diffusion à travers l'évolution de la taille de chaque compartiment dans le temps, modélisée à l'aide d'équations différentielles. Ils se concentrent donc par nature sur l'aspect temporel de la diffusion.

*Kermack et McKendrick (1927)* décrivent le modèle le plus simple, *SI*, qui considère deux états : « *Susceptible* » (*S*) et « *Infected* » (*I*). Les membres du réseau dans l'état *Susceptible* peuvent contracter la maladie au contact des membres dans l'état *Infected*. La seule transition possible, comme l'illustre la figure 4.1.a, se fait donc depuis l'état *Susceptible* vers l'état *Infected*. Ce modèle suppose que tout individu dans le compartiment *S* a une probabilité constante  $\beta$  d'être infecté par un individu appartenant au compartiment *I*. Soit  $S$  la taille du compartiment contenant les individus dans l'état *S* et  $I$  la taille du compartiment regroupant les individus dans l'état *I*. On peut alors exprimer les taux de changement de la taille des deux compartiments de la façon suivante :

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dI}{dt} = \frac{\beta SI}{N}$$

Comme on a  $N = S(t) + I(t)$ , on peut ré-écrire la seconde équation comme suit :

$$\frac{dI}{dt} = \beta I(1 - \frac{I}{N})$$

Il apparaît que la taille du compartiment *I* croît selon un modèle de Verhulst, ce

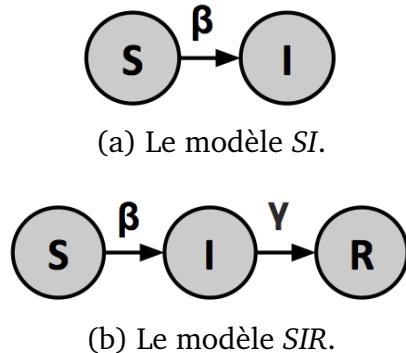


FIGURE 4.1 – Représentations graphiques des modèles épidémiologiques *SI* et *SIR*.

qui signifie que  $I(t)$  suit une fonction logistique (*Verhulst*, 1845), jusqu'à ce que tous les individus soient dans l'état I (*Hethcote*, 2000).

*Kermack et McKendrick* (1927) proposent d'enrichir ce modèle en ajoutant une nouvelle transition depuis I vers R, un nouvel état nommé « Removed » en épidémiologie ou « Refractory » dans les travaux en rapport avec la diffusion de l'information. La représentation graphique du modèle *SIR* est donnée par la figure 4.1.b. Les individus appartenant au compartiment I rejoignent le compartiment R avec une probabilité constante  $\gamma$ . Une fois dans cet état, les individus ne peuvent ni contracter, ni transmettre la maladie. On modélise l'évolution de la diffusion comme suit :

$$\frac{dS(t)}{dt} = -\frac{\beta S(t)I(t)}{N}$$

$$\frac{dI(t)}{dt} = \frac{\beta SI}{N} - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$

L'ajout du compartiment R modifie notablement le comportement du modèle, puisque la quantité d'individus dans le compartiment I n'évolue plus nécessairement de façon monotone,  $I(t)$  étant une fonction croissante dans le cas de *SI*. En effet, comme le montre la figure 4.2, la courbe représentant l'évolution de la taille du compartiment I a une allure en cloche lorsque l'on choisit des valeurs de  $\beta$  et  $\gamma$  supérieures

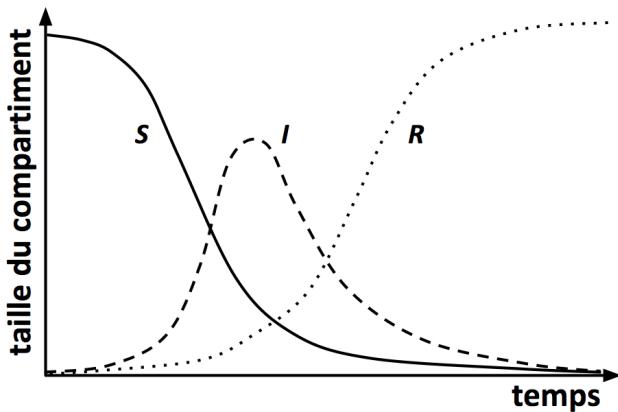


FIGURE 4.2 – Allure typique des courbes de diffusion obtenues avec le modèle *SIR*.

à 0 pour le modèle *SIR*.

**Modèles pour la diffusion de l'information dans les médias sociaux.** Ces modèles compartimentaux, ou des variantes, sont couramment utilisés pour modéliser puis prévoir la diffusion de l'information dans les médias sociaux (Saito *et al.*, 2011; Cheng *et al.*, 2013), leurs paramètres étant estimés à l'aide de traces de diffusion passées. Leskovec *et al.* (2007) proposent par exemple de modéliser la diffusion de l'information parmi les utilisateurs des médias sociaux à l'aide du modèle épidémiologique *SIS*. Celui-ci est semblable au modèle *SIR* – *I* signifiant dans ce cas que l'utilisateur a reçu l'information et participe à sa diffusion – sauf que les utilisateurs atteignant l'état *R* retournent immédiatement dans l'état *S*. Cependant, cette modélisation implique que l'on suppose que tous les individus soient aussi influents les uns que les autres – ce degré d'influence commun à tous les individus étant proportionnel à la valeur du paramètre  $\beta$ , ce qui n'est pas une hypothèse satisfaisante dans le cadre des médias sociaux. Qui plus est, les auteurs ne proposent pas de méthode pour estimer la valeur optimale de  $\beta$  mais la déterminent manuellement par essais successifs.

Face à ce problème, Yang *et Leskovec* (2010) proposent le *Linear Influence Model* (*LIM*) qui se fonde sur l'estimation de l'influence de chacun des utilisateurs pour prévoir l'évolution du volume d'utilisateurs relayant l'information. *LIM* associe à chaque utilisateur un paramètre, à savoir une fonction d'influence  $I$ , telle que la fonction  $I_{ux}$  associée à l'utilisateur  $ux$  exprime le volume d'utilisateurs qu'il influence au fil du

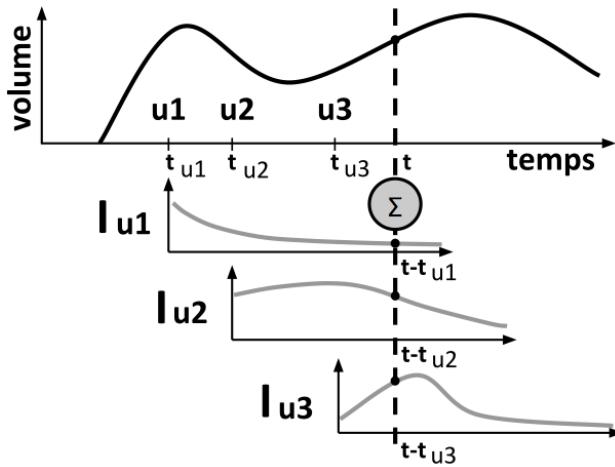


FIGURE 4.3 – Illustration du fonctionnement du *Linear Influence Model*. Le volume de messages publiés au fil du temps est obtenu en sommant les fonctions d'influence des utilisateurs initialement actifs :  $u_1$ ,  $u_2$  et  $u_3$ .

temps après avoir relayé une information. Ainsi, comme nous l'illustrons à l'aide de la figure 4.3, partant d'un ensemble d'utilisateurs initialement informés et connaissant le moment où ils publient un message à propos de cette information, *LIM* prévoit l'évolution du volume d'utilisateurs relayant l'information en sommant les fonctions d'influence des utilisateurs initiaux. Pour pouvoir estimer les fonctions d'influence, les auteurs décrivent une formulation non paramétrique de *LIM*, où les fonctions sont représentées par des vecteurs de longueur fixe. Pour estimer les fonctions d'influence des utilisateurs à l'aide de traces de diffusion passées (*i.e.* l'évolution du volume d'utilisateurs relayant l'information et le moment où chaque utilisateur a relayé l'information), ils formulent un problème d'optimisation de type « moindres carrés non négatifs » qui peut être résolu à l'aide de la méthode décrite par *Coleman et Li* (1996).

#### 4.2.2 Modélisation basée sur la structure du réseau

**Modèles classiques.** Ces travaux, menés initialement dans le domaine du marketing, modélisent le processus de diffusion au sein d'une population constante de  $N$  individus interconnectés par un réseau statique décrit par un graphe orienté  $G$ . 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, selon que la modélisation soit :

- centrée sur les récepteurs – on parle alors de « modèle de seuil », tel que le modèle de seuil linéaire développé par (*Granovetter, 1978*) ;
- ou bien centrée sur les émetteurs – on parle alors de « modèle de cascade », tel que le modèle des cascades indépendantes proposé par (*Goldenberg et al., 2001*).

Dans les deux cas, on considère que chaque membre du réseau peut être soit inactif, soit actif, un membre actif étant un individu ayant reçu l'information et participant à sa propagation. 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 monotone (*i.e.* les membres actifs ne peuvent pas redevenir inactifs) et synchrone. Par conséquent, et contrairement aux modèles compartimentaux, ces modèles se concentrent sur l'aspect structurel de la diffusion.

Les modèles de seuil sont fondés sur le principe selon lequel le passage de l'état inactif à l'état actif d'un membre du réseau dépend de l'influence exercée par ses voisins actifs dans le réseau. Chaque arc ( $ux \rightarrow uy$ ) du graphe  $G$  est associé à un paramètre  $w_{xy}$  quantifiant le degré d'influence qu'exerce le membre  $uy$  sur  $ux$ , et chaque nœud  $ux$  du graphe est associé à un seuil d'influence  $\theta_{ux}$ . Par ailleurs, à chaque nœud  $ux$  du graphe  $G$  (dont l'ensemble des voisins sortants est noté  $\Gamma_{ux}^{\rightarrow}$ ) est associée une fonction croissante  $g_{ux}$  définie sur  $\mathcal{P}(\Gamma_{ux}^{\rightarrow})$ ,  $\mathcal{P}(\Gamma_{ux}^{\rightarrow})$  représentant l'ensemble des parties de l'ensemble  $\Gamma_{ux}$ , tel que  $g_{ux}(\emptyset) = 0$ . Étant donné un ensemble  $S$  de membres du réseau initialement actifs, le processus de diffusion se déroule itérativement, comme suit. À une itération  $t$ , on évalue pour chaque nœud inactif  $ux$  la fonction  $g_{ux}$  pour l'ensemble de ses voisins actifs. Si sa valeur excède le seuil  $\theta_{ux}$ , alors le nœud  $ux$  devient actif à l'itération  $t + 1$ . Le processus s'achève lorsqu'aucune nouvelle activation n'est possible. *Granovetter* (1978) propose le *Linear Threshold Model (LT)*, qui, comme son nom l'indique, définit la fonction  $g$  comme une somme linéaire des degrés d'influence de tous les voisins actifs  $uy$ , c'est-à-dire :  $g_{ux} = \sum w_{uy}$ .

Les modèles de cascade requièrent, quant à eux, que l'on définisse pour chaque arc ( $ux \rightarrow uy$ ) la probabilité  $p_{ux,uy}$  que le membre  $ux$  influence son voisin inactif  $uy$  de sorte que celui-ci passe dans l'état actif. Étant donné un ensemble  $S$  de membres du

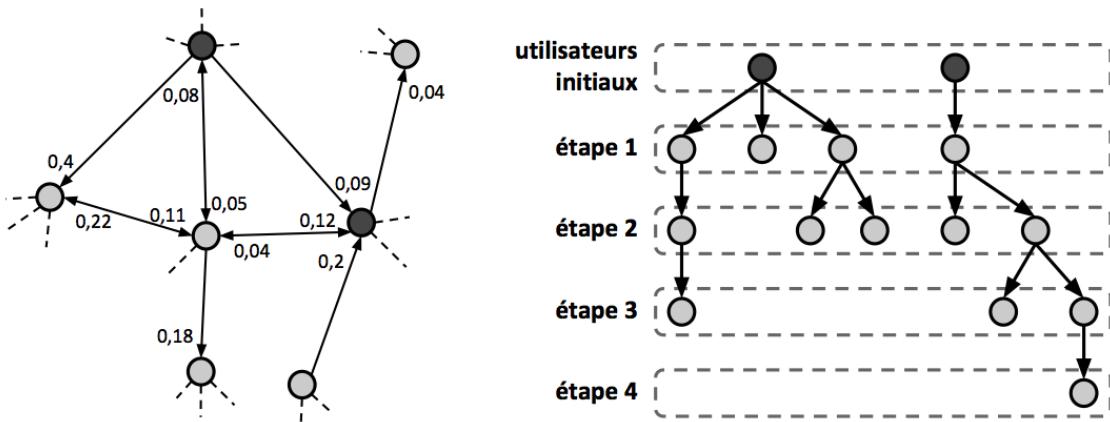


FIGURE 4.4 – Un processus de diffusion modélisé selon le *Independent Cascades Model* (*IC*). À gauche : un extrait du réseau servant de support à la diffusion, annoté avec les probabilités de diffusion pour chaque arc visible. À droite, le processus de diffusion initié par les deux nœuds colorés en gris foncé.

réseau initialement actifs, le processus de diffusion se déroule itérativement, comme suit. À une itération  $t$ , chaque nœud  $ux$  devenu actif à cette itération influence avec une probabilité  $p_{ux,uy}$  chacun de ses voisins  $uy$ , qui deviennent actifs le cas échéant à l'itération  $t+1$ . Le processus s'achève lorsqu'aucune nouvelle activation n'est possible. Goldenberg *et al.* (2001) proposent le *Independent Cascades Model* (*IC*), qui suppose que chaque nœud  $ux$  nouvellement actif influence indépendamment des autres chacun de ses voisins inactifs.

La manière dont ces modèles sont définis, ainsi que le processus itératif sur lequel ils reposent, induisent des changements d'états synchrones. Comme le montre la figure 4.4, qui illustre un processus de diffusion modélisé selon le modèle *IC*, la diffusion ne se déroule pas le long d'un axe temporel concret, mais le long d'un axe temporel simple qui correspond à une série d'étapes. Or, dans de nombreuses situations, les processus de diffusion observés sont asynchrones. Par conséquent, ces modèles ne peuvent pas reproduire les motifs temporels liés à des processus de diffusion réels asynchrones, comme c'est le cas des processus de diffusion d'information à travers les médias sociaux. Pour cette raison, Saito *et al.* (2010a) développent les modèles *AsLT* et *AsIC* qui étendent respectivement les modèles *LT* et *IC*, en associant à chaque arc

( $ux \rightarrow uy$ ) du graphe  $G$  un paramètre  $r_{ux,uy}$  quantifiant le délai d'activation. Ainsi, dans le cas du modèle *AsIC* par exemple, un nœud  $ux$  influencé par  $uy$  à un instant  $t$  est activé en  $t + r_{ux,uy}$ . Par conséquent, les changements d'état sont asynchrones et le processus de diffusion modélisé le long d'un axe temporel concret.

**Modèles pour la diffusion de l'information dans les médias sociaux.** *Saito et al.* (2009) proposent d'estimer les paramètres du modèle *IC*, à savoir les probabilités de diffusion pour chaque lien du réseau, à partir de traces de diffusion passées dans un média social, plus exactement des séquences d'activation des nœuds engendrées par la diffusion de diverses informations. Pour ce faire, ils formulent la vraisemblance qu'un ensemble donné de paramètres (*i.e.* l'ensemble des probabilités de diffusion pour les liens du réseau) ait généré ces séquences d'activation, et décrivent une méthode itérative de type *EM* (*i.e.* Espérance-Maximisation (*Do et Batzoglou, 2008*)) afin d'identifier l'ensemble de paramètres maximisant la vraisemblance. *Galuba et al.* (2010) suggèrent d'employer le modèle *LT* pour modéliser la diffusion d'une information (les auteurs considèrent uniquement la diffusion d'URL) à travers un média social. Chaque lien du réseau est donc associé à un degré d'influence et chaque nœud du réseau est associé à un seuil. Ce seuil est une fonction de deux variables latentes : (i) la viralité de l'information et (ii) la probabilité que cet utilisateur relaie n'importe quelle information. Pour définir les valeurs optimales des degrés d'influence et des deux variables latentes caractérisant chaque seuil, les auteurs formulent un problème d'optimisation à partir de séquences d'activation, qu'ils résolvent de manière itérative selon la méthode du gradient (*Snyman, 2005*).

L'inconvénient des modèles *IC* et *LT* est que ceux-ci supposent des changements d'états synchrones, ce qui fait qu'il n'est pas possible de les utiliser pour prévoir à long terme l'évolution du volume d'utilisateurs relayant l'information à travers un média social. Par exemple, *Galuba et al.* (2010) n'utilisent en pratique le modèle qu'ils proposent que pour anticiper une étape dans le processus de diffusion. Autrement dit, ils évaluent leur modèle pour la tâche qui consiste à, ayant observé le début d'un processus de diffusion, prédire quels utilisateurs parmi le voisinage direct de ceux déjà actifs vont s'activer. Comme nous l'avons indiqué précédemment, des variantes asynchrones – *AsIC* et *AsLT* – permettant de modéliser la diffusion le long d'un axe temporel concret ont été développées. Similairement aux travaux cités précédemment, *Motoda* (2011) décrit une méthode de type *EM* pour estimer les paramètres de ces variantes à partir

de séquences d'activation.

Il est également intéressant de mentionner le *Diffusive Logistic Model (DLM)* récemment proposé par *Wang et al.* (2012) qui – à l'instar des modèles épidémiologiques que nous avons décrits dans la précédente section – modélise le processus de diffusion à l'aide d'équations différentielles. Qui plus est, comme le modèle *SI* (*Kermack et McKendrick*, 1927) décrit précédemment, il modélise l'évolution du volume d'utilisateurs influencés selon une équation logistique, à la différence près que *DLM* ne caractérise pas la diffusion uniquement dans le temps, mais également dans l'espace. En effet – en supposant que le processus de diffusion est initié par un unique utilisateur – le réseau peut alors être décrit de manière superficielle en terme de distance entre l'utilisateur initial et les autres utilisateurs. Ainsi, *DLM* décrit l'évolution du volume d'utilisateurs influencés dans le temps en fonction de leur distance (notée  $x$ , *i.e.* le nombre de sauts) par rapport à la source selon l'équation suivante :

$$\frac{dI(t, x)}{dt} = rI(x, t)\left(1 - \frac{I(x, t)}{K}\right)$$

où  $r$  et  $K$  sont des paramètres caractérisant respectivement la vitesse à laquelle l'information se propage parmi les utilisateurs situés à une même distance de la source, et la densité maximale d'utilisateurs pouvant être influencés à une distance donnée.

### 4.2.3 Synthèse de l'état de l'art

La matrice présentée par la table 4.1 synthétise cet état de l'art en comparant les méthodes existantes selon quatre critères : (i) la prise en compte de la structure du réseau, (ii) la prise en compte du temps, (iii) la prise en compte de la spécificité de l'information qui se diffuse et (iv) la manière dont l'influence entre les utilisateurs est mesurée.

**Réseau.** Nous remarquons que seuls les modèles de seuil ou de cascade exploitent totalement la structure du réseau, ce qui leur permet de modéliser précisément l'influence entre chaque paire d'utilisateurs connectés. Prendre en compte la structure du réseau est important, comme le démontre l'étude menée par *Katona et al.* (2011). Les auteurs observent principalement deux effets liés au réseau : l'effet de degré, selon lequel un utilisateur connecté à beaucoup d'utilisateurs relayant une information a plus

TABLE 4.1 – Matrice de comparaison des modèles existants pour la prévision de la diffusion de l'information dans les médias sociaux.

	Prise en compte du réseau	Aspects pris en compte pour la prévision : temps	information	influence
modèles compartimentaux	-	✓	-	influence uniforme à travers le réseau
<i>LIM</i>	-	✓	-	influence globale de chaque utilisateur
<i>IC-EM</i>	✓	-	-	influence pour chaque lien du réseau
<i>LT-Gradient</i>	✓	-	✓	influence pour chaque lien du réseau
<i>AsIC-EM</i> <i>AsLT-EM</i>	✓	✓	-	influence pour chaque lien du réseau
<i>LDM</i>	partielle	✓	-	influence globale de l'utilisateur source

de chances de relayer l'information à son tour, et l'effet de clustering, selon lequel la densité de connexions au sein d'un groupe d'utilisateurs relayant une information a un effet positif important sur la probabilité que les utilisateurs connectés à ce groupe relaient à leur tour l'information. Cependant, seules les variantes asynchrones de ces modèles permettent d'intégrer le temps dans la prédiction.

**Thématicité.** Nous remarquons par ailleurs que seule l'adaptation du modèle *LT* proposée par *Galuba et al.* (2010) intègre une propriété de l'information, à savoir sa viralité, dans le processus d'estimation des paramètres et de prévision. Or, comme le montre l'étude menée sur Twitter par *Romero et al.* (2011) l'information se diffuse différemment selon les thématiques abordées. Par conséquent, il paraît crucial d'intégrer l'aspect thématique dans la prévision de la diffusion de l'information.

**Estimation des paramètres.** Il ressort de cet état de l'art que les modèles existants sont paramétrés par un grand nombre de variables latentes non observables. Dans le cas des modèles de seuil, ce sont les degrés d'influence de tous les liens du réseau et les seuils d'influence de tous les noeuds, tandis que dans le cas des modèles de cascade de base, ce sont les probabilités de diffusion de tous les liens du réseau. Les approches existantes estiment ces paramètres en formulant différents problèmes d'optimisation qui exploitent un seul type de variable observable, à savoir des séquences d'activation, ce qui pose deux problèmes. D'une part, plus le réseau étudié est grand et/ou dense,

plus la quantité de données nécessaires à l'estimation des paramètres du modèle est grande afin qu'elle soit satisfaisante, ce qui augmente d'autant plus le coût de l'estimation des paramètres. D'autre part, en estimant directement les paramètres à partir de séquences d'activation, on ne met pas en avant de facteurs observables permettant d'interpréter clairement les mécanismes régissant la diffusion.

Dans la section suivante, nous décrivons la méthode que nous proposons pour pallier aux limitations des méthodes existantes.

## 4.3 Méthode proposée

### 4.3.1 Formulation du problème

**Entrée.** Soit un ensemble  $U$  d'utilisateurs d'un média social et un corpus  $\mathcal{C}$  de messages, reflétant l'activité de ces utilisateurs pendant une certaine période de temps dans le passé. Le graphe orienté  $G = (U, E)$  modélise le réseau social interconnectant ces utilisateurs, où  $E (\subset U \times U)$  est l'ensemble des liens connectant les utilisateurs, de telle sorte qu'un lien  $(ux \rightarrow uy)$  signifie que l'utilisateur  $ux$  est connecté à l'utilisateur  $uy$  et est exposé au contenu publié par ce dernier. Soit une thématique  $T$  décrite par un mot principal et un ensemble pondéré de mots liés. Soit un ensemble  $S \subset U$  d'utilisateurs étant les premiers à relayer l'information décrite par la thématique  $T$ . La table 4.2 donne les notations utilisées dans le reste de ce chapitre.

**Sortie.** En supposant que l'information se propage par cascade d'information, l'objectif consiste à prévoir la série temporelle  $\hat{s}_T$  modélisant l'évolution du volume d'utilisateurs influencés à propos de la thématique  $T$ , laquelle se propage à travers le graphe  $G$ , en partant des utilisateurs de l'ensemble  $S$ . Supposer que la diffusion d'une thématique  $T$  est due à une cascade d'information, revient à considérer un monde fermé, où – hormis les utilisateurs de l'ensemble  $S$  – les utilisateurs ne sont influencés que par leurs voisins dans le graphe  $G$ . Autrement dit, nous ne considérons pas de sources d'influence externes.

TABLE 4.2 – Liste des notations utilisées dans le chapitre 4.

Notation	Définition
$U$	Ensemble des utilisateurs
$ux$	Un utilisateur appartenant à l'ensemble $U$
$G = (U, E)$	Réseau social interconnectant les utilisateurs
$E$	Ensemble des liens orientés du réseau social
$\Gamma_{ux}^{\rightarrow}$	Ensemble des voisins sortants de $ux$ dans le graphe $G$
$\Gamma_{ux}^{\leftarrow}$	Ensemble des voisins entrants de $ux$ dans le graphe $G$
$\mathcal{C}$	Corpus de messages publiés par les utilisateurs de l'ensemble $U$
$\mathcal{C}_{ux}$	Corpus des messages publiés par l'utilisateur $ux$ ( $\mathcal{C}_{ux} \subset \mathcal{C}$ )
$\mathcal{C}_{ux}^@$	Corpus des messages publiés par l'utilisateur $ux$ contenant au moins une mention ( $\mathcal{C}_{ux}^@ \subseteq \mathcal{C}_{ux}$ )
$V_{ux}$	Vocabulaire des mots employés dans le corpus $\mathcal{C}_{ux}$
$M_{ux}^@$	Ensemble des utilisateurs mentionnés dans le corpus $\mathcal{C}_{ux}^@$
$N_{ux}^@$	Nombre de messages dans le corpus $\mathcal{C}$ mentionnant l'utilisateur $ux$
$T$	Une thématique décrite par un terme principal et un ensemble de mots liés
$t$	Variable modélisant le temps

### 4.3.2 Vue d'ensemble de la méthode proposée

La méthode que nous proposons, dont le déroulement est schématisé à la figure 4.5, comporte deux volets. Tout d'abord, nous proposons un modèle probabiliste *T-BASIC* (*Time-Based ASynchronous Independent Cascades*) basé sur la structure de réseau sous-jacente aux médias sociaux, et qui est une extension du modèle *AsIC* proposé par (Saito et al., 2010a). Comme *AsIC*, il repose sur deux paramètres pour chaque lien du réseau, la probabilité de diffusion entre deux utilisateurs et le délai de transmission, et modélise le processus de diffusion comme des cascades indépendantes asynchrones. Néanmoins, dans le cas de *T-BASIC*, les probabilités ne sont pas constantes mais dépendante du temps, ce qui permet d'intégrer la fluctuation du niveau d'attention des utilisateurs au fil du temps.

Ensuite, nous proposons une procédure pour estimer les paramètres du modèle *T-BASIC*. Plutôt que d'estimer directement tous les paramètres latents du modèle, dont le nombre est égal à deux fois le nombre de liens dans le réseau étudié, nous proposons de les exprimer comme des fonctions de caractéristiques observables des

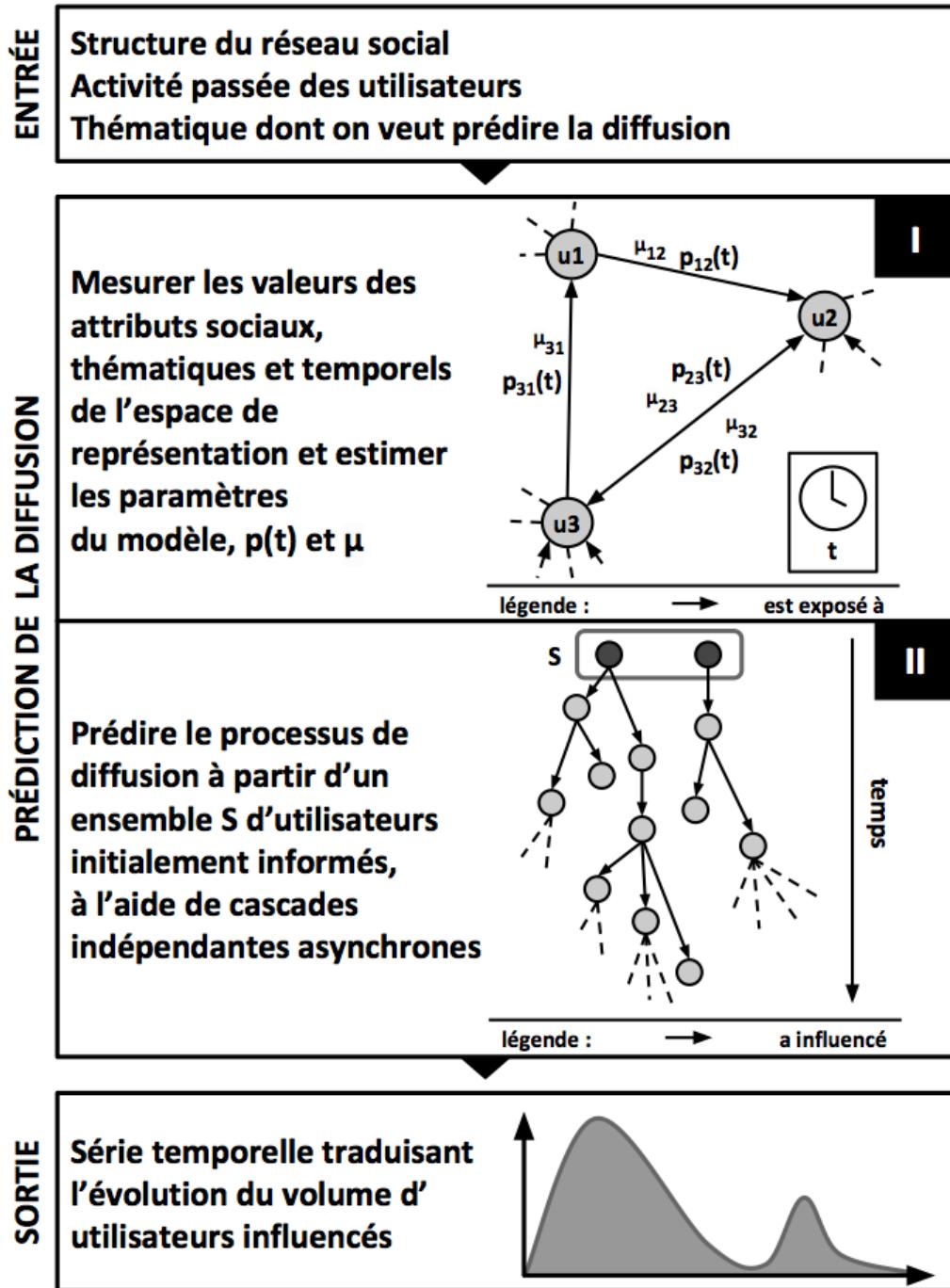


FIGURE 4.5 – Prévision de la diffusion d'une information à l'aide du modèle *T-BASIC*.

utilisateurs, selon trois aspects : social, thématique et temporel. Ainsi, le nombre de paramètres à estimer n'est plus lié à la taille du réseau, mais au nombre de caractéristiques mesurées pour chaque paire d'utilisateurs, ce qui diminue grandement le coût de la phase d'estimation. L'autre avantage de cette démarche est qu'elle permet aisément d'analyser l'effet des caractéristiques des utilisateurs sur le phénomène de diffusion de l'information.

### 4.3.3 Description du modèle

Le modèle *T-BASIC* est une extension du modèle *AsIC* (Saito *et al.*, 2010a), lui-même une extension du modèle *IC* (Goldenberg *et al.*, 2001). *T-BASIC* modélise le processus de diffusion le long d'un axe temporel concret, sous la forme de cascades indépendantes, à travers le graphe orienté  $G = (U, E)$ . Il y a pour chaque lien  $(ux \rightarrow uy) \in E$  deux paramètres :

- Une fonction réelle  $p_{ux,uy}(t)$  à valeur dans  $[0; 1]$  donnant la probabilité que l'utilisateur  $uy$  influence  $ux$  à un instant  $t$ .
- Une valeur réelle  $\mu_{ux,uy}$ , telle que  $\mu_{ux,uy} > 0$ , correspondant au délai après lequel l'utilisateur  $ux$  publie un message à propos de la thématique, une fois qu'il a été influencé avec succès par l'utilisateur  $uy$ .

Le processus de diffusion, décrit par l'algorithme 2 (page 106), est initié à partir d'un ensemble  $S$  d'utilisateurs activés. Chaque utilisateur  $uy$  nouvellement activé à un instant  $t$  peut influencer chacun de ses voisins entrants  $ux$  inactifs avec une probabilité  $p_{ux,uy}(t)$ . Nous considérons que l'utilisateur  $ux$  est influencé avec succès par  $uy$  si la probabilité  $p_{ux,uy}(t)$  est strictement supérieure à une valeur aléatoire réelle comprise entre 0 et 1, tirée aléatoirement selon une loi uniforme. Les voisins qu' $uy$  a influencés deviennent actifs à leur tour à l'instant  $t + \mu_{ux,uy}$ . La condition d'arrêt est la même que pour *IC* et *AsIC*, c'est-à-dire lorsqu'aucune nouvelle activation n'est possible.

---

**Algorithme 2 :** Prévision de la diffusion avec le modèle *T-BASIC*.

---

**Données :** Un ensemble  $U$  d'utilisateurs, le graphe  $G = (U, E)$  décrivant le réseau social

**Paramètres :** Une thématique  $T$ , les probabilités de diffusion  $p_{ux,uy}(t)$ , les délais de transmission  $\mu_{ux,uy}$ , un ensemble  $S$  d'utilisateurs initiant la diffusion

**Résultat :** Une série temporelle  $s$  caractérisant la diffusion

Initialiser l'horloge  $t$ ;

Initialiser la table  $H$  utilisée pour mémoriser les couples (utilisateur, instant d'activation);

**pour** chaque utilisateur  $u \in S$  **faire**

| Ajouter un couple  $c = (u, t)$  à la table  $H$ ;

**fin**

**tant que** il existe un couple  $c = (uy, ty) \in H | ty \geq t$  **faire**

| Initialiser la variable *compteur* à 0;

| **pour** chaque utilisateur  $uy \in H$  nouvellement activé à l'instant  $t$ , i.e.  $ty = t$  **faire**

| Incrémenter la variable *compteur*;

| **pour** chaque utilisateur inactif  $ux \in \Gamma_{uy}^{\leftarrow}$  n'apparaissant pas dans  $H$  **faire**

| Tirer une valeur aléatoire,  $v$ , selon une loi uniforme;

| **si**  $p_{ux,uy}(t) > v$  **alors**

| | /\* L'utilisateur  $uy$  a influencé  $ux$  \*/

| | Ajouter une entrée  $(ux, t + \mu_{ux,uy})$  à la table  $H$ ;

| **fin**

| **fin**

| **fin**

|  $s(t) \leftarrow \text{compteur};$

| Incrémenter l'horloge  $t$ ;

**fin**

**retourner**  $s$ ;

---

#### 4.3.4 Espace de représentation

L'estimation des paramètres du modèle *T-BASIC* repose sur les attributs numériques que nous définissons ci-après. Ils caractérisent chaque paire d'utilisateurs ( $ux,uy$ ) connectés dans le réseau social par un lien ( $ux \rightarrow uy$ ) et couvrent trois aspects : social, thématique et temporel. Ces attributs sont mesurés à partir du corpus de messages  $\mathcal{C}$  qui reflète le comportement des utilisateurs dans le passé.

**Aspect social.** Ces attributs servent à quantifier les interactions sociales ayant lieu entre les utilisateurs. Les métriques utilisées pour les calculer et que nous décrivons ci-après sont essentiellement basées sur les interactions explicites, lesquelles sont matérialisées par les mentions présentes dans les messages. L'importance des mentions dans le cadre de la détection d'évènements dans les médias sociaux a été soulignée dans le chapitre précédent, et leur pouvoir prédictif dans le cadre de la diffusion de l'information est également mis en avant dans l'étude menée par *Yang et Counts* (2010).

- *Activité* (Ac) : cet attribut caractérise le degré d'activité de chaque utilisateur. Il correspond au nombre moyen de messages publiés par heure, borné à 1. La valeur de cet attribut est calculée ainsi :

$$Ac(ux) = \begin{cases} \frac{|\mathcal{C}_{ux}|}{\epsilon} & \text{si } |\mathcal{C}_{ux}| < \epsilon, \\ 1 & \text{sinon.} \end{cases} \quad Ac(uy) = \begin{cases} \frac{|\mathcal{C}_{uy}|}{\epsilon} & \text{si } |\mathcal{C}_{uy}| < \epsilon, \\ 1 & \text{sinon,} \end{cases}$$

où  $\epsilon$  exprime la durée en heures de la période couverte par le corpus  $\mathcal{C}$ .

- *Homogénéité sociale* (Hs) : cet attribut porte sur la paire d'utilisateurs ( $ux,uy$ ). Son but est de quantifier la similarité entre les ensembles d'utilisateurs avec lesquels chacun des deux utilisateurs connectés interagissent,  $M_{ux}$  et  $M_{uy}$ . Plus cette similarité est grande, plus il est probable que les deux utilisateurs aient des centres d'intérêt proches. La valeur de cet attribut est calculée selon la métrique de Jaccard, qui correspond à la cardinalité de l'intersection des deux ensembles, normalisée par la cardinalité de leur union :

$$Hs(ux,uy) = \frac{|M_{ux} \cap M_{uy}|}{|M_{ux} \cup M_{uy}|}$$

- *Rôle* (Ro) : cet attribut vise à caractériser le rôle joué par chaque utilisateur dans le processus de diffusion de l'information et, pour cela, nous proposons de mesurer la proportion de messages publiés contenant au moins une mention. Une proportion élevée indique que l'utilisateur a tendance à jouer un rôle actif dans le processus de propagation de l'information, en ciblant activement d'autres utilisateurs. Au contraire, une proportion faible indique un rôle plus passif par rapport à la diffusion. Cette proportion est calculée comme suit :

$$Ro(ux) = \begin{cases} \frac{|\mathcal{C}_{ux}^@|}{|\mathcal{C}_{ux}|} & \text{si } |\mathcal{C}_{ux}| > 0, \\ 0 & \text{sinon.} \end{cases} \quad Ro(uy) = \begin{cases} \frac{|\mathcal{C}_{uy}^@|}{|\mathcal{C}_{uy}|} & \text{si } |\mathcal{C}_{uy}| > 0, \\ 0 & \text{sinon.} \end{cases}$$

- *Mention* (Me) : cet attribut binaire a pour but de signaler une interaction explicite, *i.e.* mention, entre deux utilisateurs connectés ( $ux \rightarrow uy$ ). Autrement dit, si  $ux$  mentionne  $uy$  dans l'un des messages contenus dans  $\mathcal{C}_{ux}$ ,  $Me(ux, uy)$  vaut 1 et 0 autrement. L'attribut est calculé – symétriquement – pour chaque utilisateur :

$$Me(ux, uy) = \begin{cases} 1 & \text{si } uy \in M_{ux}, \\ 0 & \text{sinon.} \end{cases} \quad Me(uy, ux) = \begin{cases} 1 & \text{si } ux \in M_{uy}, \\ 0 & \text{sinon.} \end{cases}$$

- *Taux de mention* (Tm) : cet attribut est similaire au « *mention rate* » décrit par *Yang et Counts* (2010). Sa valeur est proportionnelle à  $N_{ux}^@$ , la quantité de messages mentionnant le pseudonyme de chaque utilisateur. C'est un indicateur fiable de la popularité de chaque utilisateur et également un indicateur pertinent pour prédire la diffusion comme le montrent *Yang et Counts* (2010). Il est calculé comme suit :

$$Tm(ux) = \frac{N_{ux}^@}{|\mathcal{C}|} \quad Tm(uy) = \frac{N_{uy}^@}{|\mathcal{C}|}$$

**Aspect thématique.** Nous considérons également un attribut binaire (Th) capturant le lien entre chaque utilisateur et la thématique dont on cherche à prévoir la diffusion. Cet attribut signale si le mot principal,  $m$ , de la thématique  $T$  a déjà été employé dans l'ensemble des messages publiés par chaque utilisateur, c'est-à-dire :

$$\text{Th}(ux, T) = \begin{cases} 1 & \text{si } m \in V_{ux}, \\ 0 & \text{sinon.} \end{cases} \quad \text{Th}(uy, T) = \begin{cases} 1 & \text{si } m \in V_{uy}, \\ 0 & \text{sinon.} \end{cases}$$

**Aspect temporel.** Enfin, nous considérons aussi l'aspect temporel, et ce dans le but de prendre en compte la fluctuation de la réceptivité de chaque utilisateur au cours de la journée. Nous modélisons la réceptivité (Re) d'un utilisateur au cours de la journée de manière non paramétrique, à l'aide d'un vecteur non négatif de longueur fixe noté  $re$ . Chaque composante du vecteur représente le niveau d'attention durant une période de la journée. Nous proposons ici de partitionner une journée en 6 périodes de 4 heures chacune et de définir la valeur de chaque composante comme la proportion de messages que chaque utilisateur publie durant chaque période (*i.e.* [0h; 4h[, [4h; 8h[, etc.). Autrement dit, la valeur de la composante de  $re_{ux}^i$  donne la probabilité qu'un message choisi aléatoirement dans  $\mathcal{C}_{ux}$  ait été publié durant la  $i^{\text{ème}}$  période d'une journée. À un instant  $t$ , tel que l'heure de la journée correspondante soit  $h$ , on obtient donc la valeur de cet attribut ainsi :

$$\text{Re}(ux, t) = re_{ux}^{t'}, \quad \text{Re}(uy, t) = re_{uy}^{t'},$$

où l'on détermine la composante du vecteur correspondant à l'instant  $t$  ainsi :

$$t' = \lfloor \frac{h}{4} \rfloor,$$

c'est à dire le quotient de la division entière de l'heure  $h$  par 4.

**Espace de représentation complet.** Pour chaque lien du réseau ( $ux \rightarrow uy$ ), nous définissons à un instant  $t$  le vecteur  $v_{ux,uy}^t$ , pour une thématique  $T$  donnée, dont les 13 composantes correspondent aux valeurs des 13 mesures – toutes définies sur  $[0; 1]$  – que nous venons de décrire. Pour illustrer cela, la table 4.3 donne une instanciation possible du vecteur  $v_{ux,uy}^t$ .

#### 4.3.5 Estimation des paramètres du modèle

Plutôt que d'estimer directement tous les paramètres du modèle *T-BASIC*, *i.e.* les probabilités de diffusion en fonction du temps et les délais de transmission pour tous les liens du réseau, nous proposons de les exprimer en fonction des attributs de l'es-

TABLE 4.3 – Instanciation possible d'un vecteur  $v_{ux,uy}^t$ .

	Social					Thématique	Temps
	Ac	Hs	Ro	Me	Tm	Th( $T$ )	Re( $t$ )
$ux$	0,78	0,12	0,7	0	0,65	1	0,5
$uy$	0,22		0,23	1	0,10	1	0,33

pace de représentation que nous venons de décrire. Ainsi, nous définissons la probabilité de diffusion  $p_{ux,uy}(t)$  comme une fonction paramétrique  $f(v_{ux,uy}^t)$  des attributs des utilisateurs  $ux$  et  $uy$ . Nous définissons le délai de transmission  $\mu_{ux,uy}$  comme une fonction  $g(\text{Ac}(ux))$  du degré d'activité de l'utilisateur  $ux$ . Dans la suite de cette section, nous montrons comment estimer les paramètres de ces fonctions à partir d'un jeu de données adapté.

**Construction du jeu de données pour l'estimation des paramètres de  $f$ .** Le jeu de données servant à estimer les paramètres de la fonction  $f$  est un ensemble  $D$  de  $n$  instances décrites par un vecteur  $v_i = \{v_{i1}, v_{i2}, \dots, v_{i13}\}$  et une variable qualitative  $y_i$  à deux modalités, « diffusion » ou « non-diffusion », notée respectivement 1 et 0 pour simplifier les écritures. Nous avons donc  $D = (v_i, y_i)_{i=1}^n, y_i \in \{1, 0\}$ , où la modalité  $y_i = 1$  signifie que le vecteur  $v_i$  correspond aux attributs d'une paire d'utilisateurs  $(ux, uy)$  mesurés par rapport à un instant  $t$  et une thématique  $T$ , tel que l'utilisateur  $uy$  a influencé  $ux$  à l'instant  $t$  à propos de la thématique  $T$ . Au contraire, la modalité  $y_i = 0$  signifie que l'information ne s'est pas diffusée entre les deux utilisateurs décrits par le vecteur  $v_i$ . La construction de ce jeu de données se base sur les deux éléments suivants :

- Un corpus  $\mathcal{C}$  contenant l'intégralité des messages publiés par les utilisateurs de l'ensemble  $U$  durant une période de temps donnée  $P_{\mathcal{C}}$ . Les valeurs des attributs composant les vecteurs  $v_i$  sont mesurées à partir de ce corpus.
- Un ensemble  $A$  de séquences d'activation liées à la diffusion d'éléments d'information à travers le réseau social modélisé par le graphe  $G$  durant une période de temps donnée  $P_A$ . Une séquence d'activation  $a$  est liée à une thématique  $T$  décrite par un terme principal et un ensemble pondéré de mots liés, ainsi qu'un intervalle temporel  $I$ , c'est-à-dire une thématique saillante, telle que nous l'avons définie dans le chapitre 3. Une séquence d'activation est une

séquence de couples (utilisateur, instant d'activation), qui indique quels utilisateurs de l'ensemble  $U$  ont relayé l'information et à quel moment. Par exemple, le couple  $(ux, tx)$  indique que l'utilisateur  $ux$  a publié un message à l'instant  $tx$ .

Les périodes de temps  $P_{\mathcal{C}}$  et  $P_A$  sont choisies telles qu'elles soient disjointes et que  $P_{\mathcal{C}}$  soit antérieure à  $P_A$ , de sorte que les attributs soient mesurés sur une période précédent la période durant laquelle les séquences d'activation sont extraites.

L'objectif consiste, pour toutes les séquences d'activation, à identifier les paires d'utilisateurs  $(ux, uy)$  pour lesquelles – étant donnée l'hypothèse de monde fermé – il apparaît que l'utilisateur  $ux$  a été influencé par  $uy$  à propos de la thématique  $T$ .

- Premièrement, il existe dans le graphe  $G$  un arc  $(ux \rightarrow uy)$ . Autrement dit,  $ux$  est exposé aux messages publiés par l'utilisateur  $uy$ .
- Deuxièmement, les utilisateurs  $ux$  et  $uy$  apparaissent dans une même séquence d'activation, liée à une thématique  $T$  ;
- Troisième condition, l'utilisateur  $ux$  a relayé l'information décrite par  $T$  après  $uy$ , c'est-à-dire que  $tx > ty - tx$  et  $ty$  désignant les moments où respectivement  $ux$  et  $uy$  ont relayé l'information ;
- Enfin, parmi les utilisateurs appartenant à  $\Gamma_{ux}^{\rightarrow}$  (l'ensemble des voisins sortants de  $ux$ ) seul  $uy$  satisfait la troisième condition.

Pour chaque paire  $(ux, uy)$  vérifiant ces quatre conditions, nous construisons une instance positive :  $(v_{ux,uy}^{ty}, 1)$ . Les valeurs des attributs  $\text{Re}(ux)$  et  $\text{Re}(uy)$  – qui mesurent la réceptivité des utilisateurs – sont mesurées en  $ty$ , c'est-à-dire au moment de la journée où  $uy$  a relayé l'information. Les attributs  $\text{Th}(ux)$  et  $\text{Th}(uy)$  sont quant à eux évalués pour le terme principal de la thématique  $T$ . Par ailleurs, pour chaque paire  $(ux, uy)$  identifiée à partir de la séquence d'activation liée à la thématique  $T$ , nous choisissons un utilisateur  $uz$ , tel que :

- Premièrement, l'utilisateur  $uz$  n'a pas relayé l'information décrite par  $T$ , c'est-à-dire qu'il n'apparaît pas dans la séquence d'activation ;
- Deuxièmement, l'utilisateur  $uz$  est exposé aux messages publiés par  $uy$ , c'est-à-dire qu'il appartient à l'ensemble  $\Gamma_{uy}^{\leftarrow}$ , l'ensemble des voisins entrants de  $uy$  dans  $G$ .

Ainsi, pour un utilisateur  $uz$  satisfaisant ces conditions, nous construisons une instance négative :  $(v_{uz,uy}^{ty}, 0)$ . Au final, nous obtenons un jeu de données équilibré à

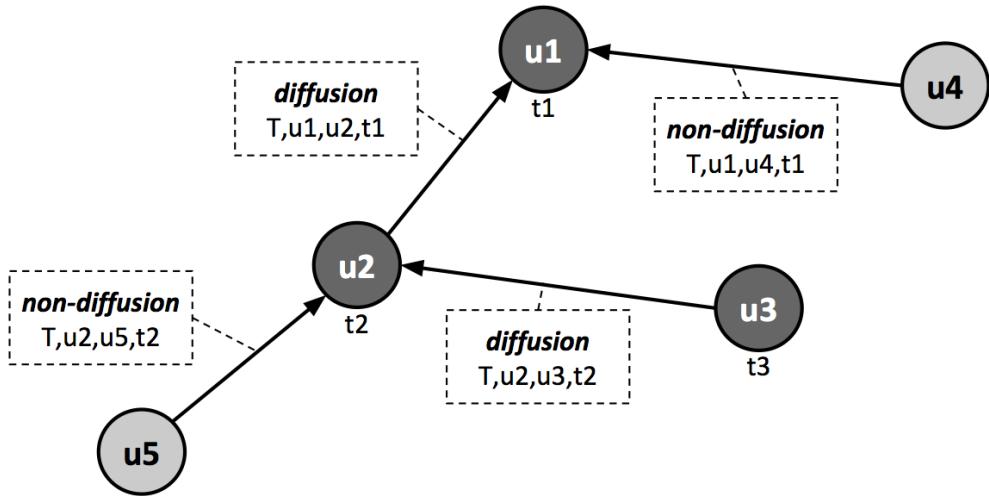


FIGURE 4.6 – Illustration du processus de construction du jeu de données d’entraînement. La structure représentée correspond au graphe  $G$ , un arc  $(ux \rightarrow uy)$  signifie donc que l’utilisateur  $ux$  est exposé aux messages publiés par  $uy$ .

partir duquel nous pouvons estimer la fonction  $f$ . La figure 4.6 illustre ce processus : les nœuds  $u_1$ ,  $u_2$  et  $u_3$  ont relayé l’information décrite par une thématique  $T$  respectivement en  $t_1$ ,  $t_2$  et  $t_3$ , tandis que les nœuds  $u_4$  et  $u_5$  n’ont pas relayé l’information.

**Estimation des paramètres de  $f$ .** Nous définissons la probabilité  $p_{ux,uy}(t)$  – la probabilité que l’utilisateur  $uy$  influence  $ux$  à un instant  $t$  à propos d’une thématique donnée – comme une fonction du vecteur d’attributs les décrivant, c’est-à-dire :  $p_{ux,uy}(t) = f(v_{ux,uy}^t)$ . Comme nous souhaitons que cette fonction soit interprétable et nous permettent d’analyser l’impact des différents attributs sur la probabilité de diffusion, nous proposons d’exprimer  $f$  comme une fonction monotone à valeurs dans  $[0; 1]$  d’une combinaison linéaire des composantes du vecteur d’attributs  $v$  de la forme  $w_0 + \sum_{j=1}^{13} w_j v_j$ . Pour simplifier l’écriture de cette combinaison linéaire, nous modifions le vecteur  $v$  de sorte à avoir  $v_0 = 1$ , ce qui nous permet de la ré-écrire comme le produit scalaire  $w \cdot v$ . La fonction  $f$  étant monotone, l’analyse du vecteur de coefficients  $w$  permet de quantifier l’impact, négatif ou positif, des différents attributs sur la probabilité de diffusion de l’information entre deux utilisateurs. Plusieurs formes paramétriques pour la fonction  $f$  sont envisageables. Nous choisissons ici d’utiliser la

fonction sigmoïde et définissons donc  $f$  de la manière suivante :

$$f(v) = \frac{\exp(w \cdot v)}{1 + \exp(w \cdot v)}.$$

Nous proposons d'estimer le vecteur de coefficients  $w$  à l'aide du jeu de données  $D = (v_i, y_i)_{i=1}^n$  par maximisation de la vraisemblance. Étant donnée une instance de ce jeu de données, nous avons la relation suivante :

$$f(v_i) = P(Y = 1|v_i)$$

Par ailleurs, la probabilité qu'un vecteur  $v_i$  soit associé à la modalité  $y_i \in \{0; 1\}$  s'écrit comme suit :

$$P(Y = y_i|v_i) = \begin{cases} P(Y = 1|v_i) & \text{si } y_i = 1, \\ 1 - P(Y = 1|v_i) & \text{si } y_i = 0. \end{cases}$$

Bénéficiant du fait que  $y_i \in \{0; 1\}$ , nous pouvons écrire cette probabilité d'une façon plus compacte :

$$P(Y = y_i|v_i) = P(Y = 1|v_i)^{y_i} (1 - P(Y = 1|v_i))^{1-y_i}$$

Ainsi, en supposant l'indépendance entre les instances du jeu de données  $D$ , nous mesurons la vraisemblance du vecteur  $w$  pour le jeu de données  $D$  (*i.e.*  $P(D|w)$ ) comme suit :

$$\begin{aligned} L(D, w) &= \prod_{i=1}^n P(Y = y_i|v_i) \\ &= \prod_{i=1}^n P(Y = 1|v_i)^{y_i} (1 - P(Y = 1|v_i))^{1-y_i} \\ &= \prod_{i=1}^n \left( \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right)^{y_i} \left( 1 - \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right)^{1-y_i} \end{aligned}$$

Estimer le vecteur  $w$  s'apparente donc au problème d'optimisation qui consiste à

maximiser la vraisemblance du jeu de données  $D$  :

$$\hat{w} = \operatorname{argmax}_w L(D, w)$$

ce qui se fait en dérivant  $L(D, w)$  par rapport à  $w$ . Or, le logarithme naturel étant une fonction strictement croissante, maximiser  $L(D, w)$  équivaut par conséquent à maximiser la log-vraisemblance,  $\ell(D, w)$  – dont la dérivation est plus simple. Nous avons donc le problème d'optimisation suivant, équivalent au précédent :

$$\hat{w} = \operatorname{argmax}_w \ell(D, w)$$

Nous exprimons la log-vraisemblance comme suit :

$$\begin{aligned}\ell(D, w) &= \ln \left( \prod_{i=1}^n \left( \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right)^{y_i} \left( 1 - \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right)^{1-y_i} \right) \\ &= \sum_{i=1}^n \left( y_i \ln \left( \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right) + (1 - y_i) \ln \left( 1 - \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right)^{1-y_i} \right).\end{aligned}$$

En exploitant le fait que :

$$P(Y = 0 | v_i) = 1 - P(Y = 1 | v_i) = 1 - \frac{\exp w \cdot v_i}{1 + \exp w \cdot v_i} = \frac{1}{1 + \exp(w \cdot v_i)}$$

on obtient alors :

$$\begin{aligned}\ell(D, w) &= \sum_{i=1}^n (y_i(w \cdot v_i) - y_i \ln(1 + \exp(w \cdot v_i)) - (1 - y_i) \ln(1 + \exp(w \cdot v_i))) \\ &= \sum_{i=1}^n (y_i(w \cdot v_i) - \ln(1 + \exp(w \cdot v_i)))\end{aligned}$$

Enfin, nous exprimons la dérivée partielle de  $\ell(D, w)$  par rapport à  $w$  comme suit :

$$\begin{aligned}\frac{\partial \ell(D, w)}{\partial w_k} &= \frac{\partial}{\partial w_k} \sum_{i=1}^n (y_i(w \cdot v_i) - \ln(1 + \exp(w \cdot v_i))) \\ &= \sum_{i=1}^n \left( y_i v_{ik} - \frac{1}{1 + \exp(w \cdot v_i)} v_{ik} \exp(w \cdot v_i) \right) \\ &= \sum_{i=1}^n v_{ik} \left( y_i - \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right)\end{aligned}$$

Or, il apparaît que l'équation :

$$\sum_{i=1}^n v_{ik} \left( y_i - \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right) = 0$$

n'est pas linéaire en  $w$ , ce qui ne permet pas de trouver une solution analytique. Néanmoins, il est possible d'appliquer la méthode numérique itérative de Newton-Raphson telle qu'elle est décrite par *McCullagh et Nelder* (1989) pour résoudre ce système d'équations non-linéaires. Cela nécessite, en plus du vecteur gradient défini par les dérivées partielles premières, de définir la matrice Hessienne qui est constituée des dérivées partielles secondes :

$$\begin{aligned}\frac{\partial^2 \ell(D, w)}{\partial w_k \partial w_l} &= \sum_{i=1}^n v_{ik} \frac{\partial}{\partial w_l} \left( y_i - \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right) \\ &= - \sum_{i=1}^n v_{ik} \frac{\partial}{\partial w_l} \left( \frac{\exp(w \cdot v_i)}{1 + \exp(w \cdot v_i)} \right) \\ &= - \sum_{i=1}^n v_{ik} v_{il} \frac{\exp(w \cdot v_i)}{(1 + \exp(w \cdot v_i))^2}\end{aligned}$$

La méthode de Newton-Raphson repose sur la relation suivante entre les solutions à l'itération  $p$  et  $p + 1$ , l'inverse de la matrice Hessienne étant la matrice de variance-covariance des coefficients (*McCullagh et Nelder*, 1989) :

$$w^{p+1} = w^p - \left( \frac{\partial^2 \ell(D, w)}{\partial w \partial w'} \right)^{-1} \times \frac{\partial \ell(D, w)}{\partial w}$$

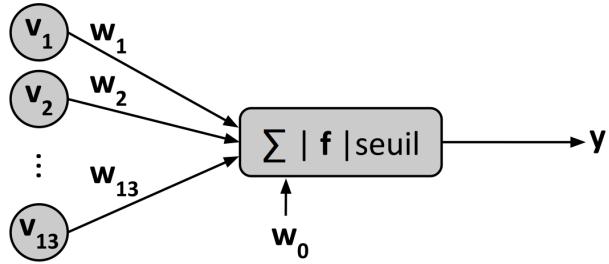


FIGURE 4.7 – Représentation du classifieur linéaire construit à partir de la fonction  $f$ .

Si l'on se donne un seuil (e.g. 0,5), cette démarche ayant pour but l'estimation des paramètres de la fonction  $f$  à partir du jeu de données  $D$  conduit à la construction d'un classifieur binaire, équivalent à un perceptron simple utilisant comme fonction d'activation la fonction sigmoïde ou un classifieur par régression logistique. Ainsi, lorsque  $f(v)$  est supérieure au seuil, le classifieur prédit la modalité 1, et 0 sinon. La figure 4.7 schématise le classifieur construit, avec à gauche les entrées et à droite la sortie.

**Construction du jeu de données et estimation des paramètres de  $g$ .** Nous définissons le délai de transmission  $\mu_{ux,uy}$ , exprimé en heures, comme une fonction du degré d'activité de l'utilisateur  $ux$ ,  $Ac(ux)$ , c'est-à-dire :  $\mu_{ux,uy} = g(Ac(ux))$ . Plus particulièrement, nous définissons  $g$  de la façon suivante, de sorte que le délai de transmission minimum soit d'une heure :

$$g(Ac(ux)) = \begin{cases} w_0 + w_1 Ac(ux) & \text{si } w_0 + w_1 Ac(ux) > 1, \\ 1 & \text{sinon.} \end{cases}$$

Pour estimer les paramètres  $w_0$  et  $w_1$ , nous construisons un jeu de données dérivé de  $D$ , noté  $D'$ . Ce jeu de données est constitué de  $m = n/2$  instances décrites par deux variables  $(x_i, z_i)$ , selon le principe suivant :

- Pour chaque instance  $(v_i, y_i)$  de  $D$  décrite par un vecteur  $v_i = v_{ux,uy}$  et telle que  $y_i = 1$ , nous créons l'instance suivante  $(Ac(ux), \delta_{ux,uy})$ , où  $\delta_{ux,uy}$  est le délai de transmission entre  $uy$  et  $ux$  exprimé en heures. Autrement dit, pour chaque paire d'utilisateurs  $(ux, uy)$  identifiée lors de la construction du jeu de données  $D$  et entre lesquels l'information s'est diffusée (cf. figure 4.6), nous

créons une instance de  $D'$  décrite par le délai de transmission observé, et le degré d'activité de l'utilisateur  $ux$  qui a été influencé par  $uy$ .

Nous proposons d'estimer les paramètres  $w = \{w_0, w_1\}$  selon la méthode des moindres carrés (*Cornillon et Matzner-Løber, 2007*), *i.e.* en minimisant la somme des carrés des différences entre les délais de transmission observés et les délais prédits, le critère des moindres carrés étant le suivant :

$$\widehat{w} = \underset{w_0, w_1}{\operatorname{argmin}} \sum_{i=1}^m (z_i - w_0 - w_1 x_i)^2$$

## 4.4 Expérimentations

Dans cette section, nous présentons la synthèse des résultats obtenus lors de l'étude expérimentale que nous avons menée avec des données issues de Twitter. Premièrement, nous examinons la validité de l'approche que nous proposons pour estimer les paramètres de *T-BASIC*, plus précisément la procédure proposée pour estimer les paramètres de la fonction  $f$ . Ensuite, nous évaluons la capacité du modèle *T-BASIC* à prévoir la dynamique temporelle du processus de diffusion, puis pour conclure cette section, nous analysons l'effet des caractéristiques des utilisateurs sur la diffusion de l'information.

### 4.4.1 Protocole expérimental

**Corpus.** Les résultats expérimentaux que nous présentons dans cette section reposent sur trois corpus de messages collectés par *Yang et Leskovec (2011)*. Chacun représente l'intégralité des tweets écrits en anglais, publiés par 52 494 utilisateurs américains durant une période d'un mois :

- Le corpus  $\mathcal{C}_1$  : 1 399 840 tweets publiés du 01/10/2009 au 31/11/2009 ;
- $\mathcal{C}_2$  : 1 437 126 tweets publiés du 01/11/2009 au 30/11/2009 ;
- $\mathcal{C}_3$  : 1 141 740 tweets publiés du 01/12/2009 au 31/12/2009.

**Structure du réseau.** Le réseau d'abonnements qui interconnectent ces 52 494 utilisateurs forme un graphe orienté comprenant 5 793 961 liens, collecté fin 2009 par *Kwak et al. (2010)*. Le graphe a un diamètre de 8, tandis que la longueur moyenne du chemin entre deux nœuds est de 2,55.

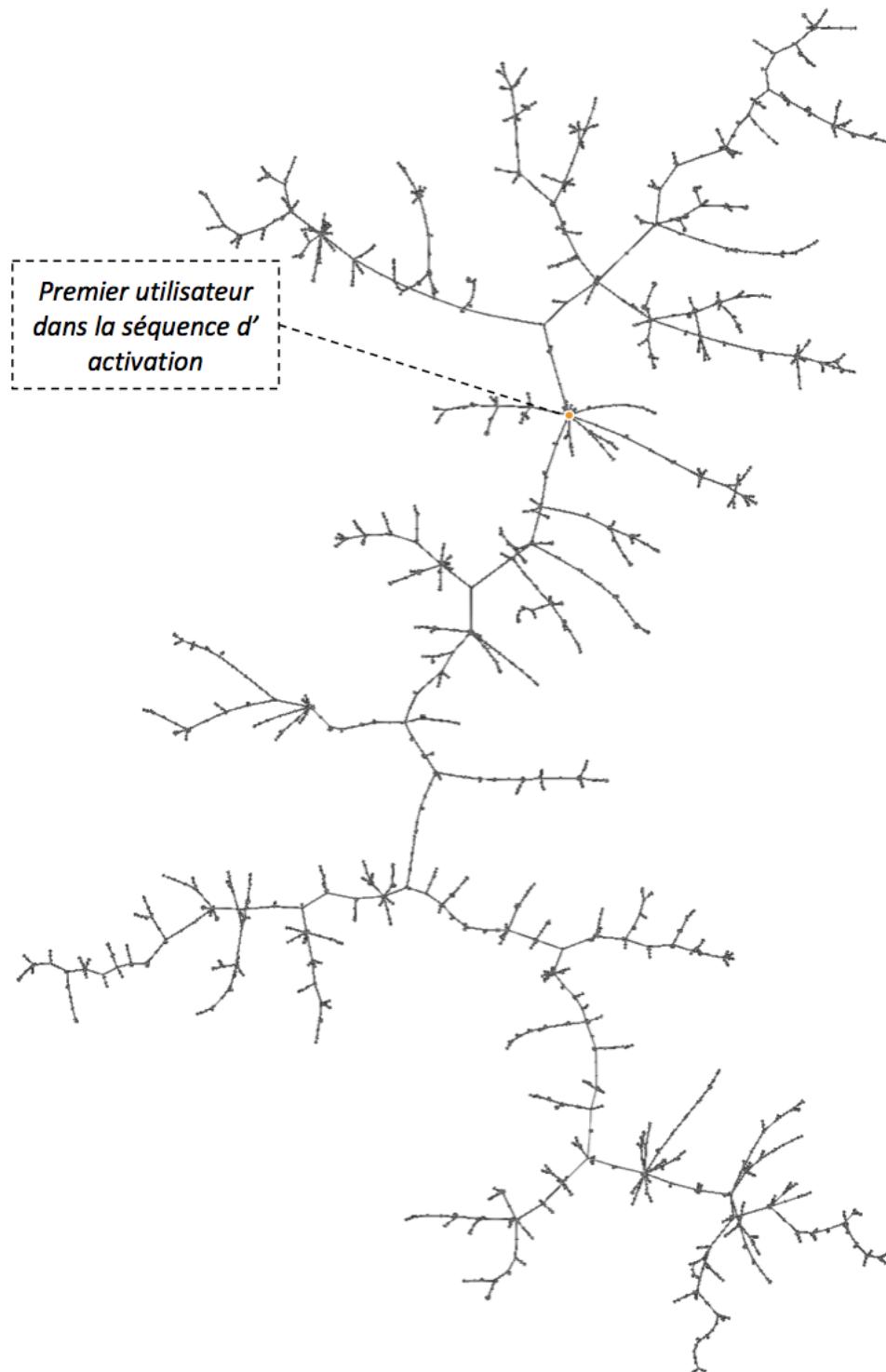


FIGURE 4.8 – Visualisation sous forme de graphe des instances de  $D$  ayant pour modalité  $y_i = 1$  extraites à partir d'un même évènement.

**Construction des jeux de données d'apprentissage.** Les instances des jeux de données d'apprentissage  $D$  et  $D'$  sont décrites par les attributs des utilisateurs mesurés à l'aide du corpus  $\mathcal{C}_1$ . Nous utilisons la méthode *MABED* – que nous avons décrite dans le chapitre 3 – pour détecter des événements dans le corpus  $\mathcal{C}_2$ , à partir desquels nous extrayons les séquences d'activation nécessaires à la construction des jeux de données d'apprentissages. Plus spécifiquement, nous appliquons la méthode *MABED* avec comme paramètres  $k = 30$ ,  $p = 3$ ,  $\theta = 0,75$  et  $\sigma = 0,5$ , ce qui génère une liste  $L$  de 30 événements décrits chacun par au plus 3 mots liés. Pour chaque événement  $e \in L$ , décrit par un terme principal, un ensemble de mots liés et un intervalle temporel, nous extrayons une séquence d'activation en identifiant les auteurs des messages publiés durant cet intervalle et contenant le terme principal et les mots liés à l'événement. Ceci nous permet, en ordonnant les auteurs des messages selon les dates de publication, de construire la séquence d'activation liée à cet événement. Pour chacune des séquences d'activation ainsi extraite, nous appliquons le processus que nous avons décrit dans la section précédente afin de générer les instances de  $D$  et  $D'$ , avec néanmoins une contrainte supplémentaire. Afin que les instances identifiées soient significatives, nous ne considérons que les utilisateurs ayant publié au moins 6 messages dans chacun des corpus  $\mathcal{C}_1$  – le corpus à partir duquel les attributs caractérisant les utilisateurs sont mesurés – et  $\mathcal{C}_2$  – le corpus à partir duquel les séquences d'activation sont extraites. Nous obtenons ainsi le jeu de données  $D$  équilibré comportant 28980 instances et le jeu de données  $D'$  comportant 14490 instances. La figure 4.8 montre par exemple le dessin d'un graphe où les liens correspondent aux instances de  $D$ ,  $(v_i, y_i)|y_i = 1$ , extraites à partir d'un des événements détectés dans  $\mathcal{C}_2$ . Cette visualisation permet de voir comment l'information s'est diffusée spatialement à partir du premier utilisateur dans la séquence d'activation, coloré en orange sur ce dessin.

**Construction des jeux de données de test.** Afin d'évaluer la procédure proposée pour estimer les probabilités de diffusion sur lesquelles repose *T-BASIC*, nous construisons un jeu de données de test noté  $D_{\text{test}}$  et comportant 24722 instances, selon le même principe que celui employé pour construire  $D$ , sauf que les attributs des utilisateurs sont mesurés à partir de  $\mathcal{C}_2$  et que les 30 séquences d'activation sont extraites à partir de  $\mathcal{C}_3$ . Afin d'évaluer la capacité de *T-BASIC* à prévoir la diffusion de l'information, plus précisément sa capacité à prévoir l'évolution du nombre d'utilisateurs

relayant une information donnée au fil du temps, nous construisons un ensemble  $\mathcal{S}$  de 30 séries temporelles de référence. Chaque série temporelle  $s_T \in \mathcal{S}$  est construite à partir de la séquence d'activation liée de l'un des évènements, et est donc liée à une thématique  $T$  décrite par un terme principal et un ensemble de mots, et un intervalle temporel  $I$ . Les séries temporelles ont une résolution de 4 heures, c'est-à-dire que  $s_T(0)$  donne le nombre d'utilisateurs activés durant les 4 premières heures de l'intervalle  $I$ ,  $s_T(1)$  donne le nombre d'utilisateurs activés entre 4 et 8 heures après le début de l'évènement, etc.

#### 4.4.2 Évaluation de la procédure d'estimation des probabilités de diffusion

Dans cette section nous nous intéressons plus particulièrement à l'efficacité de la procédure que nous proposons pour estimer les probabilités de diffusion – la probabilité  $p_{ux,uy}(t)$  étant donnée par la fonction  $f(v_{ux,uy}^t)$ , dont les paramètres sont estimés à partir du jeu de données  $D$ . Or, comme nous l'avons déjà mentionné, en se dotant d'un seuil  $\theta$ , cette procédure s'apparente à la construction d'un classifieur linéaire binaire. Par conséquent, nous pouvons évaluer la qualité de cette modélisation en évaluant ses performances pour la tâche qui consiste à classifier les instances du jeu de données  $D_{\text{test}}$ . Pour une instance  $v_i$  de  $D_{\text{test}}$ , la modalité prédictive est alors donnée par la fonction  $h(v_i)$  suivante :

$$h(v_i) = \begin{cases} 1 & \text{si } f(v_i) > \theta, \\ 0 & \text{sinon.} \end{cases}$$

**Méthodes comparées.** Nous considérons deux classificateurs couramment utilisés en apprentissage automatique, que nous entraînons avec le jeu de données  $D$  : le séparateur à vaste marge proposé par *Cortes et Vapnik* (1995) et le classifieur bayésien naïf, décrit par *John et Langley* (1995). Un séparateur à vaste marge (*SVM*) est un classifieur binaire qui cherche à identifier l'hyperplan séparant les instances avec la plus vaste marge. Plus spécifiquement, nous considérons le *SVM* à marge souple capable de traiter les cas non linéairement séparables, qui cherche donc à maximiser la marge tout en minimisant les erreurs de classification. Un classifieur bayésien naïf est un classifieur simple qui repose sur l'hypothèse selon laquelle les attributs  $v_{ij}$  sont deux à deux indépendants conditionnellement à la valeur  $y_i$ .

**Choix des paramètres.** Nous considérons quatre configurations pour le *SVM* : (i) un *SVM* utilisant un noyau linéaire (*SVM-l*), un *SVM* utilisant un noyau gaussien (*SVM-g*), deux *SVM* utilisant pour l'un un noyau polynomial (*SVM-p2*) de degré 2 et l'autre un noyau polynomial de degré 3 (*SVM-p3*). Le paramètre élémentaire de tout *SVM* est la valeur  $\gamma$  qui définit l'influence accordée à chaque instance du jeu de données d'entraînement. Un *SVM* à marge souple requiert aussi la définition d'un paramètre  $C > 0$  caractérisant le compromis entre maximisation de la marge et minimisation des erreurs de classification. Pour chacune de ces quatre configurations, nous appliquons la démarche « grid-search », c'est-à-dire une recherche exhaustive, en utilisant la validation croisée sur  $D$ , pour des valeurs comprises entre  $10^{-4}$  et  $10^4$  afin d'identifier le couple  $(\gamma, C)$  donnant les meilleurs résultats. Enfin, nous fixons  $\theta = \frac{1}{2}$ , ce qui signifie que le classifieur  $h$  basé sur la fonction  $f$  prédit la modalité la plus probable.

**Métriques d'évaluation.** Puisque nous sommes intéressés par la capacité à prévoir la diffusion de l'information entre des paires d'utilisateurs connectés, nous mesurons donc la précision par rapport à la modalité  $y = 1$ , comme le rapport entre le nombre d'instances de  $D_{\text{test}}$  correctement associées par le classifieur à la modalité 1 et le nombre total d'instances associées par le classifieur à la modalité 1, c'est-à-dire :

$$P = \frac{\text{nombre d'instances correctement associées par le classifieur à la modalité } y = 1}{\text{nombre d'instances associées par le classifieur à la modalité } y = 1}$$

De même, nous définissons le rappel par rapport à la modalité  $y = 1$  comme le rapport entre le nombre d'instances correctement associées par le classifieur à la modalité 1 et le nombre total d'instances associées à la modalité 1 dans  $D_{\text{test}}$ , c'est-à-dire :

$$R = \frac{\text{nombre d'instances correctement associées par le classifieur à la modalité } y = 1}{\text{nombre d'instances réellement associées à la modalité } y = 1}$$

Enfin, nous combinons précision et rappel en calculant la F-mesure, définie comme la moyenne harmonique de ces deux métriques :

$$F = \frac{2PR}{P + R}$$

La table 4.4 reporte la précision, le rappel et la F-mesure obtenus par chaque

TABLE 4.4 – Performances des six classifieurs sur le jeu de données  $D_{\text{test}}$ .

Métrique	$(h \mid f, \theta)$	SVM-l	SVM-g	SVM-p2	SVM-p3	Bayésien naïf
Précision	0,700	0,712	0,712	0,721	0,688	0,708
Rappel	0,822	0,799	0,798	0,755	0,692	0,777
F-mesure	0,756	0,753	0,753	0,738	0,690	0,741

classifieur sur le jeu de données  $D_{\text{test}}$ .

**Validité de la procédure d'estimation des paramètres de la fonction  $f$ .** La lecture de la table 4.4 révèle que tous les classifieurs obtiennent des performances satisfaisantes sur le jeu de données équilibré  $D_{\text{test}}$ , la F-mesure la plus faible (0,690) étant obtenue par le SVM à noyau polynomial de degré 3. Nous observons que la meilleure précision est obtenue par le SVM à noyau polynomial de degré 2, tandis que le meilleur rappel et la meilleure F-mesure sont atteints avec le classifieur  $h$  basé sur le couple  $(f, \theta)$ . La consistance des résultats obtenus en classification montre le pouvoir prédictif des attributs retenus pour constituer l'espace de représentation des utilisateurs, *i.e.* les vecteurs à partir desquels les probabilités de diffusion du modèle *T-BASIC* sont estimées. Globalement, nous observons que le rappel est supérieur à la précision. Ceci suggère l'existence d'un motif spécifique lié aux caractéristiques des utilisateurs et qui se retrouve très souvent lorsque l'information se diffuse d'un utilisateur vers un autre. Néanmoins, le fait que la précision soit inférieure au rappel indique que la présence de ce motif n'induit pas systématiquement la diffusion de l'information.

#### 4.4.3 Évaluation du modèle *T-BASIC*

Dans cette section, nous évaluons la capacité de *T-BASIC* à prévoir la diffusion de l'information en nous basant sur l'ensemble  $\mathcal{S}$  contenant 30 séries temporelles de référence, extraites à partir du corpus  $\mathcal{C}_3$ . Chaque série temporelle  $s_T \in \mathcal{S}$  décrit l'évolution réelle du volume d'utilisateurs relayant une thématique  $T$  décrite par un mot principal et un ensemble de mots liés, durant un intervalle temporel  $I$ . Il s'agit donc de prévoir la série temporelle  $\hat{s}_T$  représentant l'évolution du volume d'utilisateurs influencés à propos de la thématique  $T$  durant l'intervalle temporel  $I$ .

**Méthodes comparées.** Nous considérons la même méthode de référence que celle

utilisée par *Yang et Leskovec* (2010) pour évaluer le *Linear Influence Model*, à savoir le *One-time-lag Predictor* (*OTL*). Cette méthode simple consiste à prédire  $\hat{s}_T(t)$  comme étant égal à  $s_T(t - 1)$ . Nous considérons d'autre part une variante de *T-BASIC*,  $\alpha T$ -*BASIC*, telle que les probabilités de diffusion soient constantes, *i.e.* indépendantes du temps. Pour  $\alpha T$ -*BASIC* nous définissons donc  $p_{ux,uy}$  comme la moyenne de  $p_{ux,uy}(t')$ , mesurée en chacune des 6 tranches temporelles considérées pour l'attribut Re, c'est-à-dire :

$$p_{ux,uy} = \sum_{t'} \frac{p_{ux,uy}(t')}{6}, \text{ où } t' \in \{0\text{h}, 4\text{h}, 8\text{h}, 12\text{h}, 16\text{h}, 20\text{h}\}$$

**Métriques d'évaluation.** Comme *Yang et Leskovec* (2010), nous mesurons la différence entre le volume réel d'utilisateurs influencés à propos d'une thématique  $T$  et le volume prédit,  $\hat{s}_T(t)$ , à un instant  $t$  :

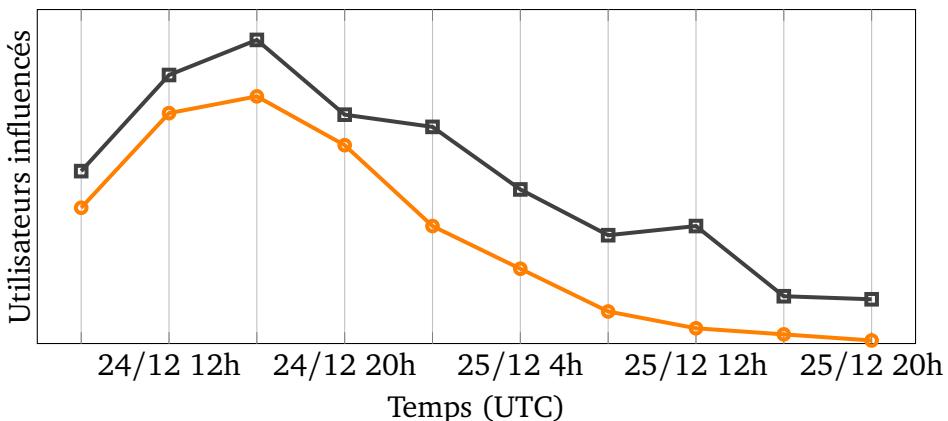
$$E_T(t) = s_T(t) - \hat{s}_T(t)$$

à partir de laquelle nous définissons l'erreur relative globale par rapport au volume :

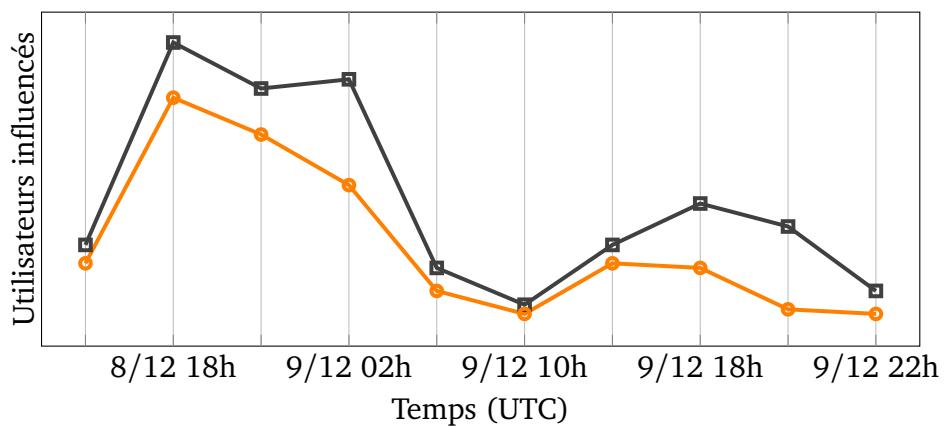
$$E_{\text{volume}} = \frac{\sqrt{\sum_{T,t} E_T(t)^2}}{\sqrt{\sum_{T,t} s_T(t)^2}}$$

**Choix des paramètres.** Les paramètres des fonctions  $f$  et  $g$  sont estimés à l'aide des jeux de données  $D$  et  $D'$  (pour rappel, les attributs sont mesurés à partir de  $\mathcal{C}_1$  et les séquences d'activation sont extraites à partir de  $\mathcal{C}_2$ ). Pour la prédiction, les attributs des utilisateurs sont à nouveau mesurés à partir du corpus  $\mathcal{C}_2$ . La prédiction est réalisée avec un pas d'une heure, autrement dit, *T-BASIC* et  $\alpha T$ -*BASIC* prédisent le nombre d'utilisateurs nouvellement activés par heure. Pour chacune des thématiques  $T$  dont on cherche à prévoir la diffusion, l'ensemble  $S$  d'utilisateurs initiant le processus de diffusion est défini comme l'ensemble réel des utilisateurs activés durant la première heure suivant le début de l'évènement décrit par  $T$  (*i.e.* la prévision débute après 1 heure d'observation). Afin de lisser les prédictions et d'obtenir une résolution identique à celle des séries temporelles de référence, 4 heures, le volume prédit par *T-BASIC* et  $\alpha T$ -*BASIC* est cumulé par tranches successives de 4 heures.

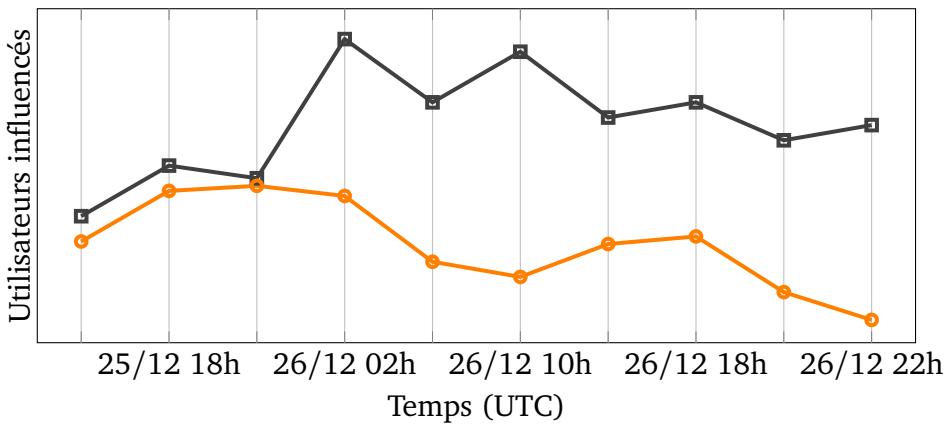
**Analyse des performances de *T-BASIC*.** La table 4.5 donne l'erreur mesurée par rapport au volume pour chacune des trois méthodes. Nous comparons les perfor-



(a) Exemple pour le premier cas.



(b) Exemple pour le deuxième cas.



(c) Exemple pour le troisième cas.

FIGURE 4.9 – Séries temporelles réelles et prédictives représentant l'évolution du volume d'utilisateurs influencés, pour trois processus de diffusion représentatifs. Les points symbolisés par des carrés correspondent aux séries temporelles réelles, tandis que ceux symbolisés par des cercles correspondent aux séries temporelles prédictives.

TABLE 4.5 – Erreur mesurée pour les trois méthodes.

Métrique	<i>One-time-lag Predictor</i>	<i>T-BASIC</i>	$\alpha T\text{-BASIC}$
$E_{\text{volume}}$	0,522	0,412	0,447

mances de *T-BASIC* et  $\alpha T\text{-BASIC}$  par rapport à celles du *One-time-lag Predictor* en mesurant le taux de réduction de l'erreur par rapport au volume,  $r$ , de la façon suivante :

$$r_{T\text{-BASIC}} = \frac{E_{\text{volume}}(T\text{-BASIC}) - E_{\text{volume}}(\text{OTL})}{E_{\text{volume}}(\text{OTL})}$$

$$r_{\alpha T\text{-BASIC}} = \frac{E_{\text{volume}}(\alpha T\text{-BASIC}) - E_{\text{volume}}(\text{OTL})}{E_{\text{volume}}(\text{OTL})}$$

Le modèle *T-BASIC* réduit l'erreur relative globale sur le volume de 21,2% par rapport au *One-time-lag Predictor*, tandis que la variante  $\alpha T\text{-BASIC}$  ne réduit l'erreur relative globale sur le volume que de 14,3%. Ceci suggère que la dynamique globale du phénomène de diffusion ne dépend pas uniquement de la topologie du réseau social à travers lequel l'information se propage, mais dépend également de la fluctuation du niveau de réceptivité de chaque utilisateur. L'analyse approfondie des résultats indique que les performances de *T-BASIC* semblent dépendre du type d'évènement lié au phénomène de diffusion à prévoir. En examinant par quels liens du réseau l'information s'est propagée et les probabilités de diffusion associées, nous identifions trois cas. *T-BASIC* obtient les meilleurs résultats dans deux de ces cas :

- Le premier cas correspond à des évènements rapidement relayés par une part importante des utilisateurs. La diffusion étant initiée par un ensemble  $S$  contenant beaucoup d'utilisateurs répartis à travers le réseau, l'information se propage rapidement en passant même par des liens associés à des probabilités de diffusion faibles, grâce à l'effet de degré déjà observé par (Katona *et al.*, 2011). L'allure des séries temporelles réelles et prédites présente dans ce cas un seul pic d'activation important dès le début de la diffusion. La figure 4.9.a illustre l'un de ces cas, basé sur un évènement détecté le 24 décembre 2009, décrit par le mot principal « xmas », et les mots liés « merry » et « hope ». Cet évènement est créé par les utilisateurs qui souhaitent un joyeux noël à leurs abonnés.

- Le second cas correspond à des évènements relayés par une faible part des utilisateurs. La diffusion est alors initiée par un ensemble  $S$  contenant peu d'utilisateurs et l'information se propage principalement via des liens associés à des probabilités de diffusion élevées. L'allure des séries temporelles réelles et prédictes présente dans ce cas des pics d'activation successifs de moins en moins importants, dus à la fluctuation du niveau d'attention des utilisateurs et aux délais de transmission. La figure 4.9.b illustre l'un de ces cas, basé sur un évènement détecté le 8 décembre 2009, décrit par le mot principal « chrome », et les mots liés « google », « mac » et « beta ». Cet évènement fait référence à la sortie en version beta du navigateur Chrome développé par Google pour le système d'exploitation Mac OS.

Dans un troisième cas, nous constatons que *T-BASIC* obtient de moins bonnes performances car les séries temporelles réelles présentent une allure irrégulière. Dans ce cas, il semble que des sources d'influence externes suscitent de nombreuses activation à travers le réseau tout au long de l'évènement. Puisque *T-BASIC* modélise la diffusion en se basant uniquement sur l'influence entre utilisateurs, l'allure de la série temporelle prédictive ne correspond donc pas à celle observée. La figure 4.9.c illustre l'un de ces cas, basé sur un évènement détecté le 25 décembre 2009, décrit par le mot principal « flight », et les mots liés « northwest » et « passenger ». Cet évènement fait référence à l'attentat terroriste manqué d'un passager d'un vol de la compagnie aérienne Northwest le jour de noël. L'article Wikipédia<sup>2</sup> consacré à cette attaque manquée témoigne de l'intense activité médiatique engendrée par cet évènement, puisqu'il fait référence à plus de 170 articles de presse, principalement publiés entre le 25 décembre et le 31 décembre 2009.

Ayant montré la validité de la procédure d'estimation des paramètres de *T-BASIC* puis la capacité du modèle à prévoir la diffusion de l'information, nous analysons, dans la section suivante, comment et dans quelle mesure les caractéristiques des utilisateurs affectent le processus de diffusion de l'information.

---

2. Lien vers l'article : [http://en.wikipedia.org/wiki/Northwest\\_Airlines\\_Flight\\_253](http://en.wikipedia.org/wiki/Northwest_Airlines_Flight_253)

#### 4.4.4 Analyse des facteurs impactant la diffusion de l'information

Dans cette section, nous étudions les paramètres de la fonction  $f(v_{ux,uy}^t)$  qui modélise la probabilité que l'utilisateur  $uy$  influence  $ux$  à un instant  $t$  à propos d'une thématique donnée  $T$ ,  $p_{ux,uy}(t)$ . Afin de quantifier l'effet des facteurs sociaux, thématiques et temporels, nous mesurons les rapports de cotes – « odds-ratio » selon la terminologie anglophone (*Mosteller, 1968*) – définis ici comme le rapport de la cote de l'évènement  $y = 1$  étant donné un vecteur  $v_1$  avec celle de l'évènement  $y = 1$  étant donné un vecteur  $v_2$ .

Tout d'abord, la cote de l'évènement  $y = 1$  sachant  $v$ , *i.e.* la diffusion d'une information entre une paire d'utilisateurs décrits par le vecteur  $v$ , est mesurée ainsi :

$$\text{odds}(v) = \frac{P(Y = 1|v)}{1 - P(Y = 1|v)} = \frac{f(v)}{1 - f(v)}$$

Le rapport de cotes pour deux vecteurs d'attributs  $v_1$  et  $v_2$  est alors mesuré comme suit :

$$\text{odds-ratio}(v_1, v_2) = \frac{\text{odds}(v_1)}{\text{odds}(v_2)}$$

Cette mesure indique donc combien de fois il y a « plus de chances » d'avoir  $y = 1$  au lieu d'avoir  $y = 0$  lorsque l'on a  $v_1$  au lieu de  $v_2$ . Afin d'examiner l'impact de chaque attribut indépendamment des autres, nous mesurons les rapports de cotes pour des couples de vecteurs différant en une seule composante. Si nous nous intéressons par exemple à l'effet de l'attribut binaire  $\text{Th}(ux)$  sur la probabilité  $p_{ux,uy}(t)$  – la probabilité que l'utilisateur  $uy$  influence  $ux$  à un instant  $t$  à propos d'une thématique  $T$ , nous notons :

$$\frac{\text{Th}(ux) = 1}{\text{Th}(ux) = 0} = \text{odds-ratio}(v_1, v_2)$$

où les valeurs des composantes des vecteurs  $v_1$  et  $v_2$  sont les mêmes exceptées celles correspondant à  $\text{Th}(ux)$ , qui vaut 1 pour  $v_1$  et 0 pour  $v_2$ . Par définition, un rapport de cotes est toujours supérieur ou égal à 0, *i.e.*  $\text{odds-ratio}(v_1, v_2) \in [0; +\infty[$ , et s'interprète dans le cas présent de cette façon :

- Lorsque  $\text{odds-ratio}(v_1, v_2)$  vaut 1, l'attribut qui varie entre  $v_1$  et  $v_2$  n'a aucun effet sur la probabilité de diffusion ;
- Lorsque le rapport est supérieur à 1, la valeur de l'attribut considéré a un effet

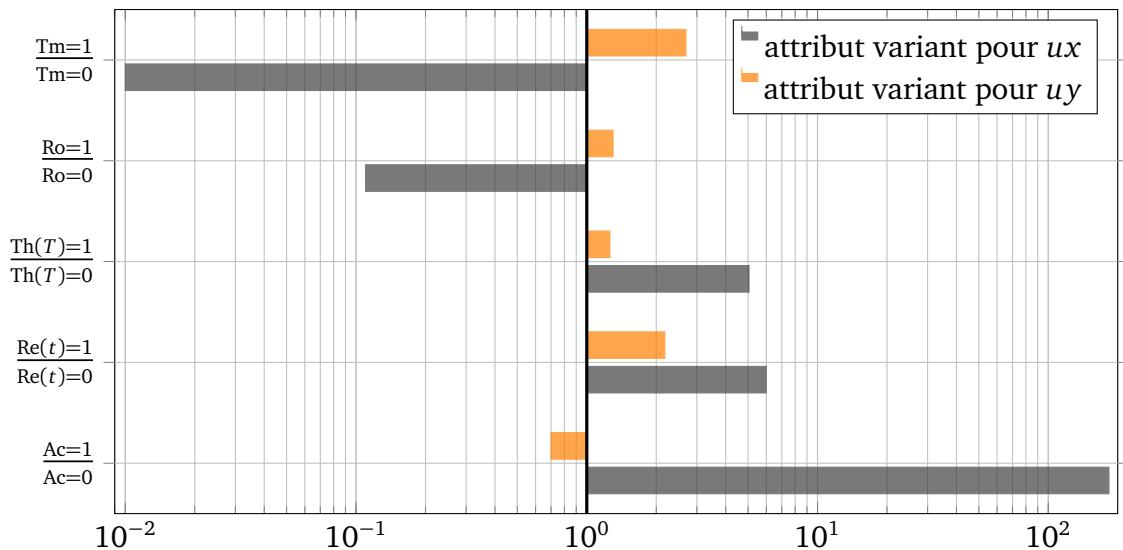


FIGURE 4.10 – Rapports de cotes pour différents attributs, mesurés par rapport aux utilisateurs  $ux$  et  $uy$ . La direction des barres traduit la direction de la relation entre chaque attribut et la probabilité de diffusion : vers la gauche, l'effet est négatif, vers la droite, l'effet est positif.

positif sur la probabilité de diffusion ;

- Au contraire, lorsque  $\text{odds-ratio}(\nu_1, \nu_2)$  est inférieur à 1, l'attribut étudié a un effet négatif sur la probabilité de diffusion.

La figure 4.10 donne les rapports de cotes en fonction de  $p_{ux,uy}(t)$  pour différents attributs, mesurés par rapport à  $ux$  et  $uy$ . Comme les attributs sont soit numériques à valeurs dans  $[0; 1]$  ou binaires avec pour valeur 1 ou 0, nous mesurons les rapports de cotes pour un attribut valant soit 1 soit 0. Bénéficiant du fait que le logarithme naturel est monotone et qu'il s'annule en 1, les rapports de cotes sont présentés selon une échelle logarithme, ce qui facilite la lecture du diagramme en faisant ressortir clairement la direction de la relation entre chaque attribut et la probabilité de diffusion.

**Impact des attributs sociaux.** Il apparaît logiquement que le facteur ayant la plus forte influence positive sur la probabilité de diffusion est le degré d'activité de l'utilisateur  $ux$ . Plus cet utilisateur est actif, *i.e.* a posté beaucoup de messages dans le passé, plus la probabilité de diffusion est importante. Au contraire, nous constatons qu'un important degré d'activité pour l'utilisateur  $uy$  a plutôt tendance à avoir un

effet négatif sur  $p_{ux,uy}(t)$ . Cela peut potentiellement s'expliquer par la dilution de l'influence de l'utilisateur proportionnellement au volume de messages qu'il produit. En effet, en publant beaucoup de messages l'utilisateur contribue à diminuer le temps durant lequel ses voisins seront exposés à chacun de ses messages. Par conséquent, même si une activité importante peut contribuer à l'influence globale d'un utilisateur, cela a tendance à diminuer son influence localement liée à chaque message. Par ailleurs, nous constatons que les taux de mention des deux utilisateurs  $ux$  et  $uy$  – que l'on peut voir comme un indicateur de popularité ou d'autorité – ont un effet important mais opposé. Nous constatons que plus un utilisateur est populaire ou jouit d'une autorité importante, moins il est influençable alors que dans le même temps cela renforce l'influence qu'il exerce sur ses voisins. Par ailleurs, la figure 4.10 révèle que l'attribut  $Ro(ux)$  qui caractérise le rôle de l'utilisateur  $ux$  (qui lorsqu'il est proche de 1 indique que  $ux$  participe de manière active au processus de diffusion en dirigeant l'information vers les utilisateurs qu'il cible directement) a un effet négatif important sur la probabilité de diffusion. Cela signifie qu'un utilisateur passif est plus susceptible d'être influencé qu'un utilisateur plus actif par rapport au processus de diffusion. Enfin, nous constatons que la valeur de l'attribut  $H(ux, uy)$  – qui mesure la similarité entre les deux ensembles d'utilisateurs avec lesquels  $ux$  et  $uy$  interagissent – a un effet positif sur la probabilité de diffusion, ce qui suggère que l'attribut  $H$  évalue d'une façon pertinente la similarité entre les centres d'intérêt des deux utilisateurs.

**Impact des attributs thématiques.** Nous observons que l'attribut thématique  $Th(ux, T)$  – qui vaut 1 si  $ux$  a publié par le passé un message contenant le mot principal de la thématique  $T$  et 0 sinon – a un effet positif sur la probabilité de diffusion, puisque nous mesurons un rapport de cotes entre  $Th(ux, T) = 1$  et  $Th(ux, T) = 0$  de 5,05, tandis que pour  $Th(uy, T) = 1$  et  $Th(uy, T) = 0$  le rapport de cotes mesuré n'est que de 1,26. Cela indique qu'un utilisateur ayant déjà abordé une thématique proche de  $T$  par le passé est plus susceptible d'être influencé à son propos, peu importe que l'utilisateur qui exerce l'influence soit familier avec ou non.

**Impact des attributs temporels.** Enfin, nous constatons que la transmission d'information de l'utilisateur  $uy$  vers  $ux$  est d'autant plus probable que la valeur de l'attribut  $Re(t)$  est élevée, le rapport de cotes mesuré entre  $Re(ux, t) = 0$  et  $Re(ux, t) = 1$  valant 5,99. Autrement dit, il est plus probable que l'utilisateur  $ux$  soit influencé par un message publié à un instant  $t$  s'il est habituellement actif durant la période de la

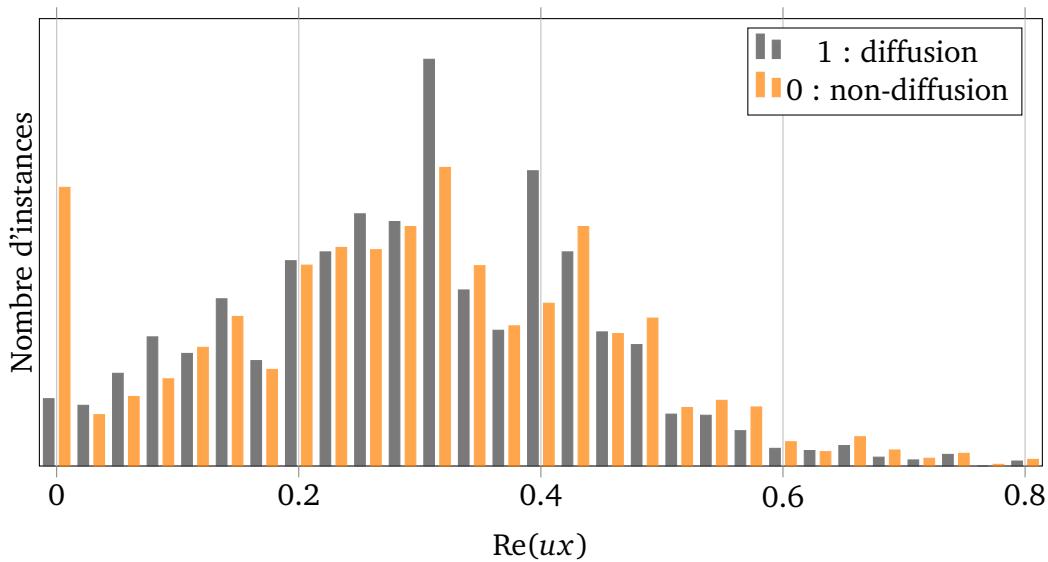


FIGURE 4.11 – Distribution des valeurs de  $\text{Re}(ux)$  en fonction de la modalité prise par  $y_i$  (1 : diffusion, 0 : non-diffusion).

journée associé à  $t$ . Pour souligner l'effet de la fluctuation du niveau de réceptivité sur le phénomène de diffusion de l'information, nous montrons avec la figure 4.11 la distribution des valeurs de l'attribut  $\text{Re}(ux)$  dans  $D$  en fonction de la modalité prise par  $y_i$ , 0 ou 1. Bien que les deux distributions aient des allures similaires pour  $\text{Re}(ux) > 0$ , il apparaît un clair déséquilibre pour  $\text{Re}(ux) = 0$  puisqu'il y a environ 4,1 fois plus d'instances pour lesquelles  $y_i$  vaut 0 que 1.

Globalement, l'étude des facteurs impactant la diffusion de l'information nous apprend deux choses. Premièrement, les trois aspects considérés sont importants,

TABLE 4.6 – Effet des caractéristiques des utilisateurs sur la probabilité de diffusion : la couleur orange traduit un effet positif tandis que la couleur grise traduit un effet négatif, l'intensité de la couleur traduisant l'importance de l'effet mesurée selon le log-odds-ratio.

	Social					Thématique	Temps
	Ac	Hs	Ro	Me	Tm	Th( $T$ )	Re( $t$ )
$ux$		Orange	Grey		Dark Grey		
$uy$		Light Orange		Light Orange	Light Orange		

puisque l'activité sociale, la familiarité avec la thématique à diffuser ainsi que la fluctuation du niveau d'attention agissent significativement sur la probabilité de diffusion. Deuxièmement, nous observons que la probabilité de diffusion est plus fortement liée aux caractéristiques de l'utilisateur subissant l'influence que de celles de l'utilisateur qui l'exerce. Pour illustrer ce constat, la table 4.6 présente un tableau structurant les caractéristiques d'une paire quelconque d'utilisateurs, et dont les cellules sont colorées selon leur impact sur la probabilité de diffusion. La couleur orange symbolise un effet positif et le gris un effet négatif, plus ou moins important selon l'intensité. La première ligne de cette table liste l'effet des attributs de l'utilisateur subissant l'influence et ressort nettement par rapport à la deuxième ligne qui se rapporte à l'utilisateur exerçant l'influence.

## 4.5 Discussion

Pour conclure ce chapitre, nous résumons dans un premier temps les travaux que nous venons de présenter. Ensuite, nous discutons des pistes de recherche ouvertes par ces travaux.

### 4.5.1 Résumé des travaux présentés

Dans ce chapitre, nous avons décrit *T-BASIC*, une modélisation probabiliste de la diffusion de l'information dans les médias sociaux basée sur la structure de réseau interconnectant les utilisateurs, ainsi qu'une procédure pour estimer les paramètres latents du modèle (*i.e.* la probabilité de diffusion et le délai de transmission pour chaque lien). Ce modèle se distingue des approches existantes basées sur la structure du réseau, de par le fait que la probabilité de diffusion en chaque lien n'est pas constante mais dépendante du temps. La procédure d'estimation des paramètres diffère également des procédures existantes, de par le fait qu'elle ne vise pas à estimer directement tous les paramètres latents à partir de séquences d'activation. Nous proposons plutôt d'exprimer la probabilité de diffusion en chaque lien du réseau comme une fonction  $f$  de caractéristiques observables des utilisateurs, et le délai de transmission comme une fonction  $g$  de leurs caractéristiques. Ainsi, estimer les paramètres du modèle *T-BASIC* consiste à estimer les paramètres définissant les fonctions  $f$  et  $g$ .

à partir de séquences d'activation et de mesures des caractéristiques des utilisateurs, ce qui a pour effet de diminuer le coût d'estimation global. Les expérimentations que nous avons menées avec des données collectées sur Twitter ont démontré la pertinence de notre proposition. Dans un premier temps, nous avons examiné la validité de la procédure d'estimation des paramètres et avons montré son efficacité. Dans un second temps, nous avons évalué les capacités prédictives du modèle *T-BASIC*. Il apparaît notamment que ses performances sont supérieures à celles d' $\alpha T\text{-BASIC}$ , une variante où les probabilités de diffusion en chaque lien du réseau sont indépendantes du temps. Nous avons pu ainsi confirmer l'effet de la fluctuation du niveau de réceptivité des utilisateurs au fil du temps sur le phénomène de diffusion de l'information. Enfin, l'analyse des paramètres de la fonction  $f$  nous a permis d'étudier l'impact des caractéristiques des utilisateurs sur la probabilité de diffusion. Nous avons montré que leurs caractéristiques, tant sur le plan social, que sur le plan thématique ou temporel, affectent le phénomène de diffusion de l'information. Par ailleurs, nous avons observé que la probabilité de diffusion est plus fortement liée aux caractéristiques de l'utilisateur influencé, qu'à celles de l'utilisateur exerçant l'influence.

**Publication.** Ces travaux ont notamment fait l'objet d'un article long présenté à l'atelier international *MSND* organisé conjointement avec la conférence internationale *WWW* en 2012.

#### 4.5.2 Perspectives de travail

Les travaux entamés à propos de la modélisation et la prévision de la diffusion de l'information ouvrent diverses perspectives de travail.

Notamment, nous identifions diverses pistes pour l'amélioration de *T-BASIC*. Dans les travaux présentés, nous supposons implicitement que la structure du réseau servant de support au processus de diffusion reste constante dans le temps. Néanmoins, il semblerait pertinent de considérer ce réseau comme dynamique. En effet, en se propagant, l'information peut susciter la création de nouveaux liens à travers le réseau, et de nouveaux utilisateurs peuvent également intégrer le réseau, ce qui peut influer sur la dynamique de diffusion. Nous supposons également qu'une thématique donnée se diffuse indépendamment des autres thématiques relayées en parallèle par les utilisateurs. Toutefois, comme le montre l'étude menée par *Myers et Leskovec (2012)*,

les différents processus de diffusion simultanés peuvent interagir entre eux et entrer en compétition, ce qui affecte les probabilités de diffusion entre les utilisateurs. Plus précisément, ils observent que la diffusion d'une thématique très populaire peut favoriser la propagation de thématiques proches moins populaires, tandis qu'elle limite encore plus la diffusion des thématiques très différentes et peu populaires. Enfin, une autre piste pour améliorer *T-BASIC* consisterait à relâcher l'hypothèse de monde fermé, puisque dans certains cas, des sources d'influence externes au réseau peuvent jouer un rôle prépondérant dans la diffusion de l'information. *Myers et al.* (2012) s'intéressent d'ailleurs à cette question et mènent une étude sur Twitter, qui révèle qu'en moyenne 71% des utilisateurs sont activés du fait de l'influence interne au réseau, les 29% d'activations restantes pouvant être attribuées à une influence externe.

Par ailleurs, nous identifions une piste de recherche intéressante concernant le problème de maximisation de la diffusion de l'information. En effet, en analysant l'impact des caractéristiques des utilisateurs sur les probabilités de diffusion, nous observons que ce sont les caractéristiques des utilisateurs subissant l'influence qui affectent le plus le phénomène de diffusion. Partant de ce constat, il serait possible de traiter le problème de maximisation de la diffusion en recherchant les utilisateurs les plus susceptibles d'être influencés, tandis que les méthodes existantes (*Kempe*, 2003; *Even-Dar et Shapira*, 2007; *Saito et al.*, 2010b) recherchent les utilisateurs le plus susceptible d'exercer une influence importante.



# Chapitre 5

## Un logiciel libre pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux

### Sommaire

---

<b>5.1 Introduction</b> . . . . .	136
<b>5.2 État de l'art</b> . . . . .	139
5.2.1 Détection d'évènements et analyse de l'influence dans les médias sociaux . . . . .	139
5.2.2 Logiciels pour la fouille et l'analyse de données issues des médias sociaux . . . . .	144
5.2.3 Synthèse de l'état de l'art . . . . .	147
<b>5.3 Logiciel proposé</b> . . . . .	148
5.3.1 But du logiciel, publics visés et architecture générale . . . . .	149
5.3.2 Service de manipulation des données . . . . .	152
5.3.3 Service de détection d'évènements . . . . .	154
5.3.4 Service d'analyse du réseau social . . . . .	157
5.3.5 Service d'import d'algorithmes et API . . . . .	160
<b>5.4 Exemples de scénarios d'utilisation</b> . . . . .	161
5.4.1 Utilisation par un non-expert . . . . .	161
5.4.2 Utilisation par un chercheur du domaine . . . . .	165
<b>5.5 Discussion</b> . . . . .	170

---

Dans ce chapitre, nous présentons la troisième contribution apportée par ces travaux de thèse. Il s'agit d'un logiciel libre que nous développons pour la fouille et l'analyse des données issues des médias sociaux, autant destiné aux chercheurs du domaine qu'à des non-experts.

## 5.1 Introduction

Les précédents chapitres ont permis de montrer la richesse de la recherche menée à propos des médias sociaux et de l'analyse du phénomène de diffusion de l'information. L'abondance de la recherche sur cette question traduit directement l'intérêt que portent de nombreux acteurs de la société aux médias sociaux et à leur analyse. Parmi ces acteurs, on retrouve par exemple les entreprises qui cherchent à analyser les réactions des consommateurs à propos de leurs produits et à promouvoir ces derniers. Parmi ces acteurs se trouvent également les services de renseignement, qui cherchent à identifier des discussions suspectes et des personnes potentiellement dangereuses, mais aussi les services météorologiques, qui cherchent à détecter des phénomènes naturels dangereux et difficilement prévisibles, à partir d'éléments relayés par le grand public. Les journalistes cherchent également à tirer parti des médias sociaux pour détecter les événements marquants, identifier des personnes à interroger à leur propos, et démarrer leurs investigations. Quel que soit le cas d'application, le processus d'analyse sous-jacent reste le même : partant d'un sujet d'étude, *i.e.* des données collectées à partir d'un média social, on cherche dans un premier temps à identifier les principaux événements animant les discussions, puis dans un second temps on cherche à identifier des personnes influentes par rapport à ces événements afin de prendre des décisions et éventuellement agir. Dans le cas d'une entreprise, le sujet d'étude peut être un ensemble de messages contenant le nom de la marque associée, et les événements peuvent correspondre aux réactions suscitées par la sortie de nouveaux produits de la marque par exemple. En identifiant les personnes influentes par rapport aux divers produits, l'entreprise peut alors mettre en place des campagnes de promotion virales, ciblées et plus efficaces. Dans le cas des services de renseignement, le sujet d'étude peut être un ensemble de messages contenant des mots-clés spécifiques, ou bien un ensemble d'utilisateurs suspects ainsi que les messages qu'ils ont publiés. Les événements peuvent correspondre à des menaces ou crises susceptibles d'affecter

un état, et les personnes influentes au sein des sous-réseaux liés à ces évènements peuvent alors correspondre à des criminels ou des leaders d'opinion dangereux. Enfin, dans le cas d'une analyse journalistique des médias sociaux, le sujet d'étude peut être l'ensemble des messages publiés sur un média social, ou bien un sous-ensemble plus spécifique, e.g. lié à la politique, le sport ou la technologie. Les évènements détectés peuvent alors faire l'objet d'une investigation journalistique, laquelle peut être menée en commençant par interroger les personnes influentes par rapport à chaque évènement, *etc.*

Étant donné le grand volume de données produites par les médias sociaux, pour que leur analyse soit efficace et utile, celle-ci doit reposer sur des techniques de fouille de données adaptées. De nombreuses méthodes ont été récemment proposées pour détecter les évènements (*Shamma et al., 2011; Weng et Lee, 2011; Lau et al., 2012; Yuheng et al., 2012; Li et al., 2012; Parikh et Karlapalem, 2013; Benhardus et Kalita, 2013; Guille et Favre, 2014a*) et analyser l'influence (*Page et al., 1998; Kitsak et al., 2010; Brown et Feng, 2011; Dugué et Perez, 2014*) à partir de données produites par les médias sociaux. Néanmoins, les chercheurs développant ces méthodes ne partagent pas systématiquement leurs implémentations et lorsque c'est le cas, elles sont programmées dans des langages différents, requièrent des formatages de données différents, *etc.* Ceci constitue un problème fondamental dans la mesure où – en particulier dans le domaine de la recherche en informatique – la reproductibilité est essentielle pour quantifier les progrès réalisés, mais aussi parce que cela ne favorise pas la réutilisation de ces méthodes par des non-experts (e.g. journalistes, enquêteurs, analystes médias) pour analyser les données dont ils disposent. Nous constatons par ailleurs que les méthodes pour l'analyse de l'influence au sein des médias sociaux, via l'analyse de la structure des réseaux sociaux sous-jacents, sont souvent utilisées pour étudier de grandes populations d'utilisateurs sans considération pour les thématiques à propos desquelles l'influence s'exerce (*Bi et al., 2014*). Les utilisateurs des médias sociaux réagissant notamment par rapport aux évènements, une manière de raffiner l'analyse de l'influence consiste justement à la mesurer par rapport aux évènements. Face à ces constatations, nous sommes donc amenés à formuler les questions suivantes. D'abord, *comment favoriser le partage et la réutilisation par les chercheurs des implémentations des méthodes développées pour la détection d'évènements et l'analyse de l'influence ?* Ensuite, *comment permettre aux non-experts d'explorer facilement*

*les données dont ils disposent ?* Bien qu'il existe des logiciels libres développés dans le milieu académique pour analyser l'influence dans les réseaux (Auber, 2004; Bastian et al., 2009), il n'existe pas encore de logiciel libre dédié à la détection d'évènements. Il existe par ailleurs des logiciels commerciaux – dont les plus connus sont *SAP Social Media Analytics*<sup>1</sup>, *NetBase*<sup>2</sup> et *BrandWatch Analytics*<sup>3</sup> – permettant une certaine analyse des évènements et de l'influence dans les médias sociaux. Néanmoins, ils n'apportent pas de solutions tangibles puisque ce sont des logiciels propriétaires qui ne dévoilent pas les algorithmes sur lesquels ils reposent, ce qui est problématique lorsqu'il s'agit d'interpréter les résultats qu'ils produisent et évaluer leur validité.

**Proposition et positionnement.** Nous proposons *SONDY* (*SOcial Network DYnamics*), qui est – à notre connaissance – le premier outil libre et extensible pour la détection d'évènements et l'analyse de l'influence à partir de données collectées sur les médias sociaux. Ces données consiste d'une part en un corpus de messages et d'autre part en la structure du réseau social interconnectant leurs auteurs. *SONDY* est un logiciel écrit en langage Java, qui inclut des outils de visualisation et implémente plusieurs méthodes de la littérature pour la détection d'évènements et l'analyse de l'influence, ainsi que des fonctionnalités avancées de préparation des données.

Contrairement aux logiciels académiques existants qui se concentrent soit sur l'analyse des messages soit sur l'analyse du réseau, *SONDY* permet d'analyser ces deux types de données conjointement, en permettant l'analyse de l'influence par rapport aux évènements. L'application est conçue de telle sorte qu'il soit simple d'y ajouter de nouveaux algorithmes en utilisant l'interface de programmation qu'elle fournit. Elle peut par ailleurs être utilisée comme bibliothèque au sein de tout programme, ce qui permet d'automatiser son fonctionnement. Son interface graphique simplifie son utilisation par des non-experts qui peuvent alors explorer les données dont ils disposent à l'aide de méthodes et visualisations adaptées, sans avoir de connaissances poussées en informatique.

**Scénarios d'utilisation.** Nous illustrons, à l'aide d'exemples et de captures d'écran, des scénarios d'utilisation montrant comment *SONDY* peut être utilisé, que ce soit par un non-expert (tel qu'un journaliste recherchant des informations et des personnes à interroger à propos d'évènements spécifiques, ou bien un analyste média souhai-

---

1. SAP Social Media Analytics : <http://www.news-sap.com/power-social-media-analytics/>
2. NetBase : <http://www.netbase.com>
3. BrandWatch Analytics : <http://www.brandwatch.com/brandwatch-analytics/>

tant analyser l'image d'une marque sur un média social) ou par un chercheur du domaine (par exemple pour comparer les types d'évènements détectés par différentes méthodes).

Ce chapitre est organisé de la manière suivante. Dans la section 5.2, nous synthétisons l'état de l'art, d'une part en matière de méthodes pour la détection d'évènements et pour l'analyse de l'influence dans les médias sociaux, et d'autre part en matière de solutions logicielles. Dans la section 5.3 nous présentons le logiciel *SONDY* et les fonctionnalités qu'il offre. Dans la section 5.4 nous illustrons le potentiel du logiciel à travers plusieurs scénarios d'utilisation. Enfin, nous concluons ce chapitre avec la discussion présentée à la section 5.5.

## 5.2 État de l'art

Dans cette section, nous présentons l'état de l'art à propos, d'une part, des méthodes pour la détection d'évènements et l'identification d'utilisateurs influents dans les médias sociaux, et d'autre part à propos des solutions logicielles pour ces tâches.

### 5.2.1 Détection d'évènements et analyse de l'influence dans les médias sociaux

**Détection d'évènements.** Un état de l'art détaillé à propos des méthodes pour la détection d'évènements dans les médias sociaux est présenté dans la section 3.2 du chapitre 3 de ce manuscrit de thèse. Pour résumer, nous pouvons dire que les méthodes existantes mettent en œuvre différentes approches se concentrant sur le contenu textuel des messages pour détecter les évènements : pondération statistique des termes (*Shamma et al.*, 2011; *Benhardus et Kalita*, 2013), modélisations probabilistes des thématiques latentes (*Lau et al.*, 2012; *Yuheng et al.*, 2012), ou encore clustering (*Weng et Lee*, 2011; *Li et al.*, 2012; *Parikh et Karlapalem*, 2013). Comme nous l'avons montré dans le chapitre 3, il est aussi possible d'exploiter le contenu hyper-textuel des messages pour améliorer la détection automatique des évènements (*Guille et Favre*, 2014a).

En complément de cet état de l'art, il est également intéressant de citer les travaux de *Rong et Qing* (2012) sur l'analyse visuelle des évènements. Ils proposent de

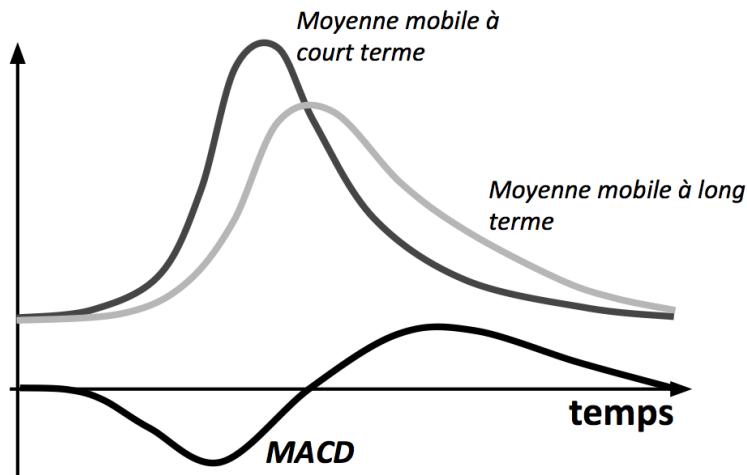


FIGURE 5.1 – Comportement typique de l'indicateur *MACD*.

visualiser leur dynamique à l'aide de courbes *MACD* (*Moving Average Convergence Divergence*) calculées à partir de la fréquence des termes liés aux événements. Ils développent à cette fin une version de l'indicateur *MACD* adaptée à cette tâche. Cet indicateur, originellement développé pour analyser la dynamique des cours de bourse par *Appel* (2005), met en valeur les tendances d'une série temporelle en mesurant la différence entre une moyenne exponentielle mobile à long terme et une moyenne exponentielle mobile à plus court terme. Ainsi, les changements de signe de l'indicateur traduisent des inversions de tendance. Constatant que la fréquence des termes peut être volatile, les auteurs proposent d'employer des moyennes mobiles simples et non pas exponentielles, lesquelles sont susceptibles d'amplifier cette volatilité du fait du poids important qu'elles donnent aux tranches temporelles récentes. La figure 5.1 illustre l'oscillation de l'indicateur *MACD* en fonction de deux moyennes mobiles. L'oscillation est facilement interprétable visuellement, suivant deux règles : (i) quand la valeur de l'indicateur *MACD* devient positive, l'intérêt des utilisateurs pour l'événement grandit et (ii) quand la valeur de l'indicateur *MACD* devient négative, l'intérêt pour l'événement s'atténue.

**Analyse de l'influence.** Il existe de nombreuses méthodes pour l'analyse de l'influence, *i.e.* autorité, des utilisateurs des médias sociaux. Celle-ci exploite la structure du réseau social interconnectant les utilisateurs, partant du postulat qu'un lien

$(ux \rightarrow uy)$  est assimilable à un vote de l'utilisateur  $ux$  en faveur de l'autorité de l'utilisateur  $uy$ .

*Page et al.* (1998) montrent que le degré entrant – dans le cas d'un réseau social, le degré entrant d'un utilisateur caractérise le nombre d'utilisateurs exposés aux messages qu'il publie – n'est pas une mesure suffisante pour déterminer la distribution de l'influence au sein d'un réseau et propose une mesure nommée *PageRank*. En plus du degré entrant, elle considère deux autres facteurs. D'abord, un lien (*i.e.* vote) provenant d'un utilisateur influent est plus significatif qu'un lien provenant d'un utilisateur peu influent. Ensuite, plus le degré sortant d'un utilisateur est grand, moins les liens sortants correspondants sont significatifs. La mesure d'autorité relative  $\mu_{ux}$  d'un utilisateur  $ux$  combine ces trois facteurs et est formulée ainsi :

$$\mu_{ux} = \sum_{(uy \rightarrow ux)} \frac{\mu_{uy}}{d_s(uy)}$$

Exprimer la mesure d'autorité pour tous les membres du réseau revient à écrire un système d'équations linéaires, que l'on peut exprimer sous forme matricielle :

$$\mu^T A = \mu^T$$

où le vecteur noté  $\mu$  représente les mesures d'autorité relatives pour tous les membres, et  $A$  est une matrice carrée telle que  $A_{xy}$  vaut  $1/d_{ux}$ , *i.e.* l'inverse du degré sortant d' $ux$ , s'il existe un lien ( $ux \rightarrow uy$ ) et 0 sinon. Or, par définition,  $\mu^T A = \mu^T$  équivaut à  $(A - I)\mu^T = 0$  et le vecteur  $\mu$  correspond alors à un vecteur propre de la matrice  $A$ . Dans le cas des grands réseaux où calculer la solution exacte est trop coûteux, la méthode de la puissance itérée est couramment employée (*Wills*, 2006) pour estimer la valeur propre de plus grand module et le vecteur propre associé. Par ailleurs, la matrice  $A$  étant stochastique puisque ses colonnes somment à 1, elle peut être interprétée comme la matrice des transitions d'une marche aléatoire. Dans ce cas, la mesure *PageRank* d'un utilisateur modélise la probabilité qu'une personne navigant aléatoirement à travers la structure du réseau social visite le profil de cet utilisateur.

La convergence de la méthode de la puissance itérée pouvant être lente dans certains cas, *Kitsak et al.* (2010) suggèrent d'utiliser la décomposition en  $k$ -enveloppes, pour mesurer plus simplement l'autorité relative des membres d'un réseau social.

Cette méthode affecte à chaque membre du réseau une mesure d'autorité relative, sur la base des deux facteurs suivants : (i) l'autorité d'un utilisateur dépend de son degré entrant et (ii) un lien (*i.e.* vote) entrant est d'autant plus significatif qu'il provient d'un utilisateur influent, *i.e.* ayant lui-même un degré entrant important. Une  $k$ -enveloppe est définie par *Seidman* (1983) comme un sous-réseau connexe et maximal, dont les membres ont un degré entrant supérieur ou égal à  $k$ . Les membres du réseau appartenant à l'enveloppe  $k$  peuvent être identifiés selon la simple méthode suivante. D'abord, les membres du réseau dont le degré entrant est inférieur à  $k$  sont retirés, ainsi que leurs liens entrants et sortants. Puis, les membres dont le degré entrant est inférieur à  $k$  suite à cette opération sont également retirés ainsi que leurs liens, et ainsi de suite jusqu'à ce qu'il n'y ait plus de membres dont le degré entrant est inférieur à  $k$ . L'ensemble d'utilisateurs restants, s'il n'est pas vide, correspond alors aux utilisateurs dont le score relatif d'influence vaut  $k$ . Cette méthode est appliquée itérativement à partir de  $k = 0$  jusqu'à ce que le cœur du réseau ait été atteint, c'est à dire l'enveloppe non-vide au plus grand  $k$ . La figure 5.2 montre la décomposition d'un réseau fictif en 3 itérations (*Kitsak et al.*, 2010). Lors de la première itération, les membres de couleur sombre sont retirés, puis les membres colorés en gris sont retirés lors de l'itération suivante. Le cœur du réseau est atteint à la troisième itération et est composé des membres colorés en blanc. La décomposition complète en  $k$ -enveloppes d'un réseau peut être calculée avec une complexité temporelle en  $\mathcal{O}(|E|)$  selon l'algorithme décrit par *Batagelj et Zaversnik* (2011).

*Brown et Feng* (2011) constatent que la distribution du nombre d'utilisateurs en fonction de l'enveloppe à laquelle ils appartiennent a généralement une allure en longue traîne (de l'anglais « long-tail distribution »). Par exemple, ils observent à partir de données collectées sur Twitter que la majorité des utilisateurs ont des valeurs de  $k$  faibles, avec un pic à  $k = 4$  et le reste des utilisateurs se répartissant dans plusieurs milliers d'enveloppes non-vides, ce qui rend difficile l'analyse des résultats. Pour remédier à ce problème, ils proposent une version modifiée de la méthode, en ce qu'elle adopte une échelle logarithme. C'est-à-dire qu'une enveloppe  $k$  rassemble les membres du réseau ayant un degré entrant supérieur ou égal à  $2^k - 1$  au lieu de  $k$ .

La majorité des méthodes existantes, telles que celles que nous venons de décrire, visent à caractériser l'influence, ou l'autorité – c'est-à-dire l'influence *a priori* positive

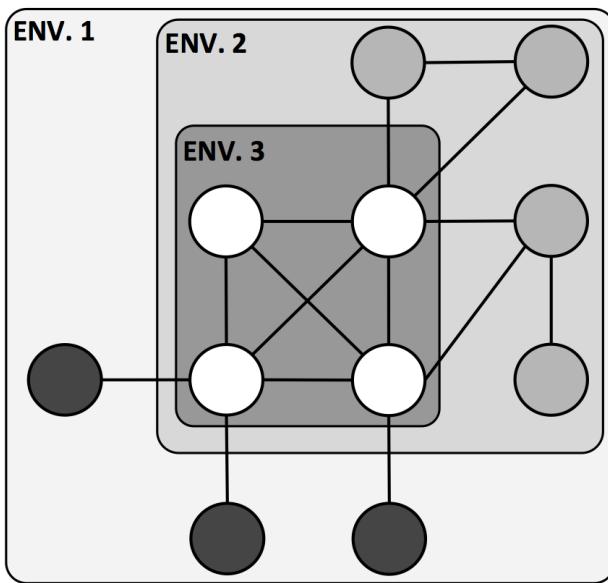


FIGURE 5.2 – Décomposition en trois enveloppes d'un réseau comportant 11 membres.

– des membres d'un média social sur la diffusion de l'information. Toutefois, *Dugué et al.* (2014) observent que certains utilisateurs des médias sociaux cherchent à maximiser leur influence de manière artificielle, selon une stratégie « capitalisme social ». Ces utilisateurs – appelés capitalistes sociaux – se connectent à un grand nombre d'utilisateurs pratiquant également cette stratégie, sans égard au contenu publié par ces derniers et dans l'espoir qu'ils établissent un lien réciproque. Ils gagnent ainsi en visibilité au sein des réseaux sociaux en augmentant leur degré, au détriment de la qualité du contenu échangé. Cette pratique remet en cause le postulat à la base des méthodes que nous avons décrites précédemment, à savoir que les liens entre utilisateurs sont assimilables à des votes pouvant être exploités comme des marqueurs d'influence. Les capitalistes sociaux ayant des degrés importants et se connectant majoritairement à d'autres capitalistes, leur influence au sens de la décomposition en  $k$ -enveloppes ou en  $\log k$ -enveloppes est importante également. La méthode *Page-Rank* semble néanmoins plus robuste face à ce genre de comportement, puisque la significativité des liens est pondérée par le degré sortant des utilisateurs à leur origine, ce qui modère l'impact des capitalistes sociaux. *Dugué et Perez* (2014) proposent

une méthode simple permettant d'identifier les utilisateurs susceptibles de pratiquer le capitalisme social. Elle repose sur la mesure du taux de recouvrement entre les voisinages entrants et sortants des utilisateurs ainsi que le rapport entre la taille de leurs voisinages entrants et sortants.

### 5.2.2 Logiciels pour la fouille et l'analyse de données issues des médias sociaux

Dans cette sous-section, nous synthétisons l'état de l'art concernant les solutions logicielles, que nous divisons en deux catégories, selon qu'elles soient développées dans l'industrie ou dans le milieu académique.

**Logiciels développés dans l'industrie.** De nombreux logiciels pour l'analyse des médias sociaux sont développés dans l'industrie, dont trois des plus populaires sont : *SAP Social Media Analytics*<sup>4</sup>, *NetBase*<sup>5</sup> et *BrandWatch Analytics*<sup>6</sup>. Ces logiciels sont orientés marketing et sont conçus pour permettre aux entreprises, à partir de données qu'elles ont ciblées, de détecter les évènements qui animent les discussions à propos de leur(s) marque(s) et identifier les utilisateurs influents à leur sujet. Ces logiciels souffrent principalement de deux limitations. Premièrement, les trois logiciels mentionnés sont payants et propriétaires, ce qui signifie que le code source n'est pas public. Cela ajouté à l'absence de communication à propos des algorithmes mis en œuvre pour traiter les données leur donne un aspect « boîte noire » qui se révèle problématique lorsqu'il s'agit d'interpréter les résultats qu'ils produisent et d'évaluer leur validité. Deuxièmement, les interfaces qu'ils proposent ne sont pas toujours adaptées aux résultats à visualiser.

À titre d'exemple et comme on peut le voir sur la figure 5.3.a, le logiciel *SAP Social Media Analytics* décrit les thématiques animant les discussions à l'aide d'un nuage de mots, lesquels sont dessinés avec plusieurs couleurs et tailles de police. Cette représentation faisant appel à beaucoup de dimensions ne permet pas d'identifier aisément les différentes thématiques. La figure 5.3.b montre quant à elle l'interface dédiée à l'identification d'utilisateurs influents proposée par le logiciel *BrandWatch Analytics*, qui consiste en une distribution et ne permet pas de visualiser la structure

4. SAP Social Media Analytics : <http://www.news-sap.com/power-social-media-analytics/>

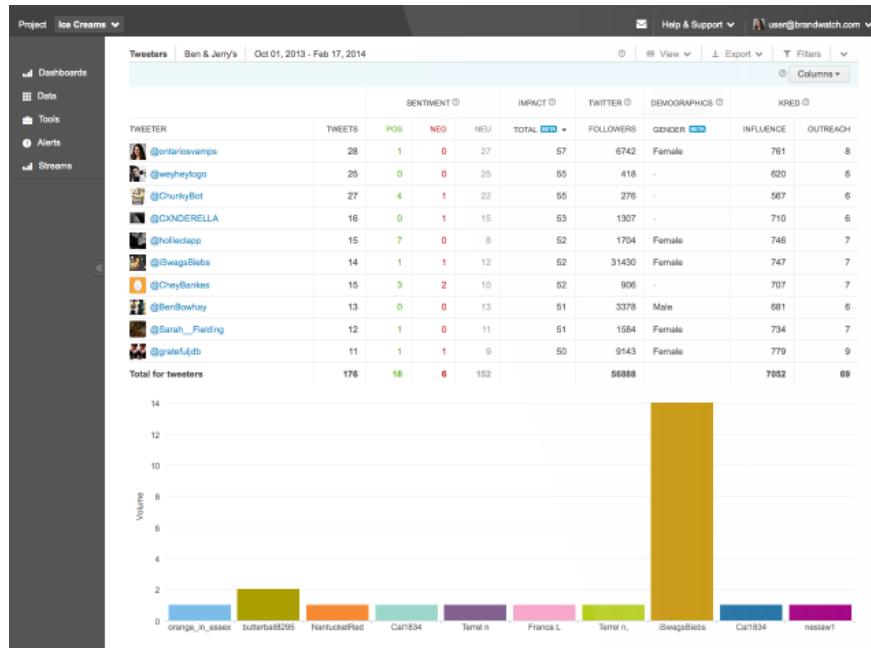
5. NetBase : <http://www.netbase.com>

6. BrandWatch Analytics : <http://www.brandwatch.com/brandwatch-analytics/>

## Un logiciel libre pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux



(a) Interface dédiée à l'analyse des messages, montrant des résultats à propos de la marque American Airlines.



(b) Interface dédiée à l'analyse de l'influence des utilisateurs, montrant des résultats pour la marque Ben & Jerry's.

FIGURE 5.3 – Interfaces des logiciels *SAP Social Media Analytics* (a) et *BrandWatch Analytics* (b).

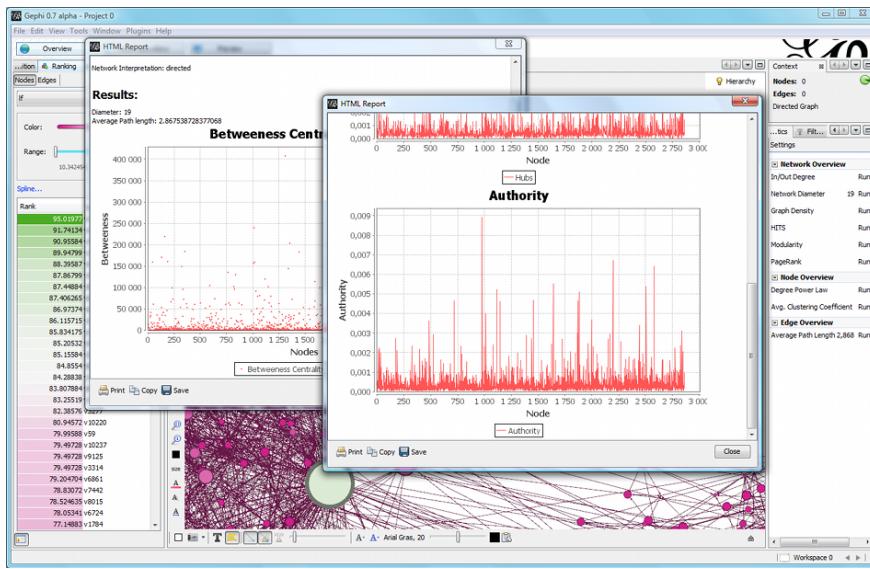


FIGURE 5.4 – Interface utilisateur du logiciel Gephi pour la fouille de graphe.

du réseau social. Or, la notion de réseau est essentielle à la notion d'influence, puisque c'est précisément parce que les utilisateurs des médias sociaux font partie d'un réseau qu'ils peuvent subir ou exercer de l'influence. Par conséquent, sa visualisation est un élément essentiel lors de l'analyse de l'influence.

**Logiciels développés dans le milieu académique.** Il n'existe à notre connaissance aucun logiciel développé dans le milieu académique permettant d'étudier à la fois les événements et l'influence à partir de données collectées sur les médias sociaux. Il existe néanmoins plusieurs prototypes non-libres pour la détection d'événements – qui implémentent chacun leur propre algorithme et ne sont pas conçus pour en intégrer d'autres – tels que *Eddi* (Bernstein *et al.*, 2010), *TwitInfo* (Marcus *et al.*, 2011) ou bien encore *KeySEE* (Lee *et al.*, 2013). Ces prototypes sont spécifiquement conçus pour Twitter et collectent directement les messages qu'ils analysent. Par ailleurs, les algorithmes que mettent en œuvre ces prototypes pour analyser les données sont peu ou pas décrits. Pour l'analyse de l'influence, il existe plusieurs logiciels libres émanant du milieu académique tels que *Gephi* écrit en Java (Bastian *et al.*, 2009), dont l'interface utilisateur est illustrée par la figure 5.4, ou encore *Tulip* (Auber, 2004) et *SNAp*<sup>7</sup>

7. Page officielle du logiciel *SNAp* : <http://snap.stanford.edu>

TABLE 5.1 – Matrice synthétisant les fonctionnalités des logiciels développés dans le milieu académique pour les tâches de détection d'évènements et l'analyse de l'influence.

		Détection d'évènements	Analyse de l'influence	
		Algorithmes	Visualisations	Algorithmes
<i>Eddi</i>	<i>ad hoc</i>	frise chronologique, nuage de mots	–	–
<i>TwitInfo</i>	<i>ad hoc</i>	frise chronologie, courbe de fréquence	–	–
<i>KeySEE</i>	<i>ad hoc</i>	frise chronologique, courbe de fréquence, nuage de mots	–	–
<i>Gephi</i>	–	–	<i>Page Rank</i> , <i>HITS</i> , intermédiairité	réseau coloré, réseau interactif, distribution de l'influence
<i>Tulip</i>	–	–	<i>Page Rank</i> , <i>k-cores</i> , intermédiairité	réseau coloré, réseau interactif, distribution de l'influence
<i>SNAP</i>	–	–	<i>Page Rank</i> , <i>HITS</i> , <i>k-cores</i> , intermédiairité	export vers Microsoft Excel

écrits en C++. La table 5.1 liste les fonctionnalités des logiciels que nous venons de citer.

### 5.2.3 Synthèse de l'état de l'art

Dans cette section, nous synthétisons brièvement l'état de l'art.

**Logiciels développés dans l'industrie.** Tout d'abord, nous constatons que les logiciels développés dans le milieu industriel sont très spécialisés et ne permettent pas d'analyser conjointement évènements et influence dans les médias sociaux. Par ailleurs, les visualisations qu'ils proposent ne sont pas toujours nécessairement adaptées, ce qui ne facilite pas l'analyse de leurs résultats, d'autant plus que nous ignorons les algorithmes mis en œuvre pour parvenir à ces résultats.

**Logiciels développés dans le milieu académique.** La table 5.1 liste les fonctionnalités des logiciels développés dans le milieu universitaire pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux. D'une part, nous constatons

que les auteurs de méthodes pour la détection d'évènements partagent rarement leurs implémentations. Par exemple, parmi les 10 méthodes couvertes par l'état de l'art (section 5.2.1), seule l'implémentation d'*On-line LDA* et de *MABED* (la méthode que nous proposons) sont partagées par les auteurs<sup>8</sup>. Qui plus est, les logiciels dédiés à la détection d'évènements se concentrent principalement sur l'aspect visualisation et implémentent des algorithmes peu ou pas décrits. De plus, ils collectent directement les données qu'ils analysent et ne permettent pas l'import de jeux de données statiques. Par conséquent, il semble important de favoriser le partage des implémentations des méthodes de détection d'évènements. D'autre part, il existe une riche offre de logiciels pour l'analyse de l'influence, qui implémentent des méthodes décrites dans la littérature et permettent d'importer manuellement les données à explorer. Or, l'influence n'étant pas universelle, il semble important de la mesurer dans un contexte précis, *e.g.* par rapport à un évènement discuté par les utilisateurs.

Face à ces constats, nous présentons dans la section suivante le logiciel libre et extensible que nous proposons pour analyser conjointement évènements et influence dans les médias sociaux.

### 5.3 Logiciel proposé

Nous proposons le logiciel *SONDY*, qui inclut des outils de visualisation et implémente des algorithmes de l'état de l'art pour la fouille et l'analyse des données issues des médias sociaux, et est doté d'une interface graphique facilitant l'accès à ses fonctionnalités. Ce logiciel est disponible librement et gratuitement sous la licence GNU GPLv3<sup>9</sup>. Il est téléchargeable à l'adresse suivante : <http://mediamining.univ-lyon2.fr/sondy>, et son code source est accessible par SVN à l'adresse qui suit : <http://mediamining.univ-lyon2.fr/websvn>. *SONDY* est développé en Java, ce qui permet de l'exécuter sur la majorité des systèmes d'exploitation, et ce avec de bonnes performances (*Taboada et al.*, 2013). La popularité de ce langage favorise par ailleurs sa réutilisation et son interopérabilité, puisque plusieurs groupes de recherche ont aussi fait le choix de ce langage, comme le « Stanford Natural Language Processing

---

8. Nous avons contacté les auteurs des autres méthodes par e-mail afin de leur demander s'ils pouvaient partager leur implémentation mais avons reçu des réponses négatives voire aucune réponse.

9. Les termes de la licence sont consultables à l'adresse : <https://www.gnu.org/licenses/quick-guide-gplv3.fr.html>.

Group<sup>10</sup> » ou le « Machine Learning Group at the University of Waikato<sup>11</sup> » entre autres.

### 5.3.1 But du logiciel, publics visés et architecture générale

**But du logiciel.** Le logiciel *SONDY* a pour but de permettre la détection d'évènements et l'identification d'utilisateurs influents à partir de données issues d'un média social. Pour cela, il offre quatre services, dotés d'interfaces graphiques adaptées :

- Le service de *manipulation* (*i.e.* import et préparation) *des données* (*cf.* figure 5.5.a) ;
- Le service de *détection et de visualisation des évènements* (*cf.* figure 5.5.b) ;
- Le service d'*analyse et de visualisation du réseau social* (*cf.* figure 5.5.c) ;
- Le service d'*import de nouveaux algorithmes* (*cf.* figure 5.5.d) ;

La figure 5.6 montre où ces services se positionnent par rapport au processus typique de fouille de données. Le service de manipulation des données se place naturellement au niveau des premières phases du processus, tandis que les trois autres services interviennent lors des dernières phases du processus qui conduisent à l'extraction de connaissances.

**Publics visés.** Le logiciel est destiné à être utilisé par des non-experts – *e.g.* journalistes, analystes médias, enquêteurs – grâce à son interface graphique claire. Il est également destiné à être utilisé par les chercheurs du domaine, puisque le logiciel est conçu pour permettre l'ajout de nouveaux algorithmes de façon simple, et peut également être utilisé comme bibliothèque au sein d'un autre programme Java.

**Architecture.** L'application traite deux types de données à la fois : les données décrivant un ensemble de messages publiés par les utilisateurs d'un média social, et les données décrivant la structure du réseau social interconnectant ces utilisateurs. La figure 5.7 décrit l'architecture logicielle de *SONDY*, *i.e.* les entrées/sorties des différents services et la manière dont ils communiquent entre eux. On observe notamment que le service de manipulation des données fait le pont entre données brutes et données manipulées par les services de détection d'évènements et d'analyse du réseau social.

---

10. Liste des logiciels développés par le Stanford NLP Group : <http://nlp.stanford.edu/software/index.shtml>

11. Liste des logiciels développés par le Waikato Machine Learning Group : <http://www.cs.waikato.ac.nz/ml/weka/index.html>

### 5.3. Logiciel proposé

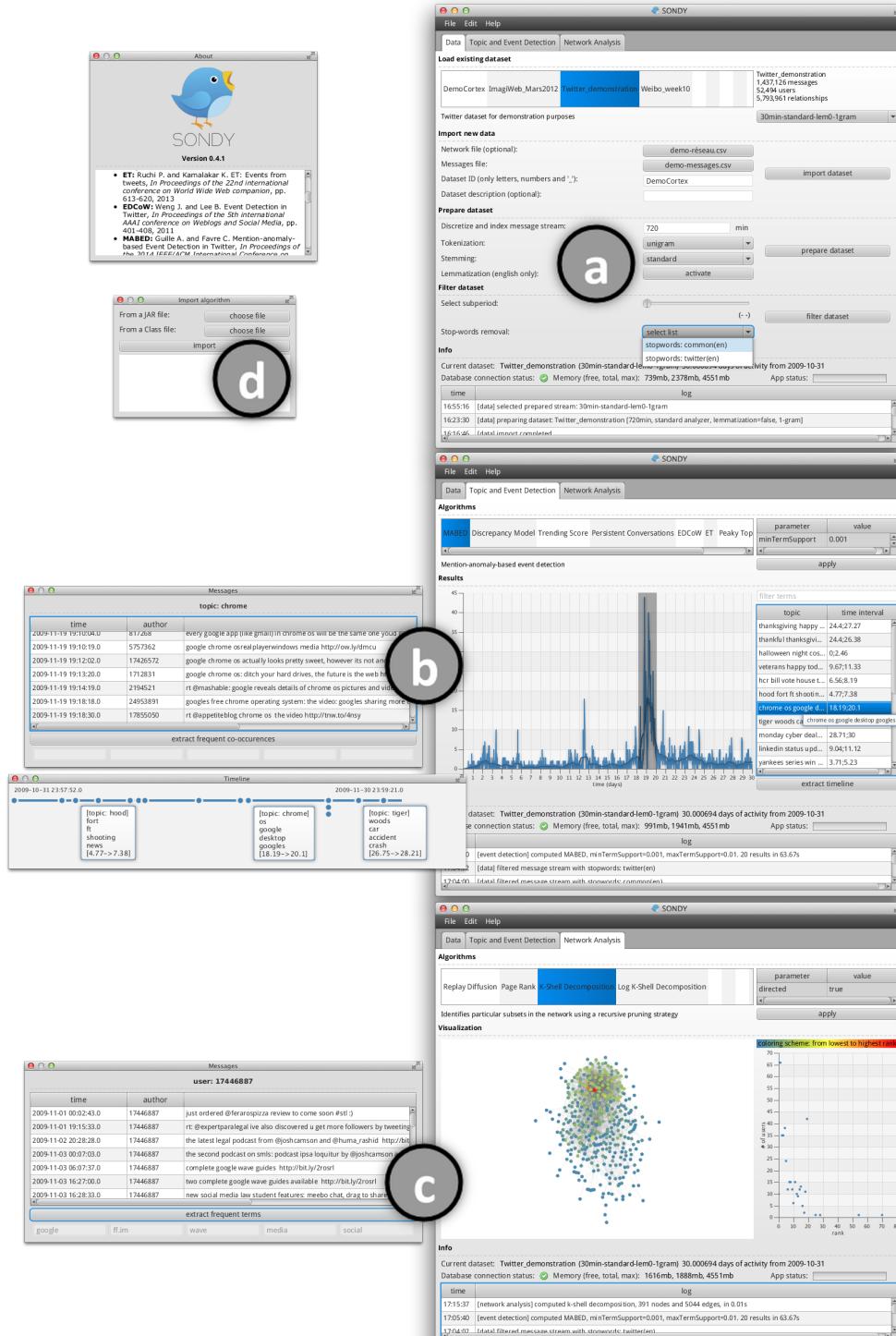


FIGURE 5.5 – Interfaces du logiciel SONDY : manipulation des données (a), détection et visualisation des évènements (b), analyse et visualisation du réseau social (c) et import de nouveaux algorithmes (d).

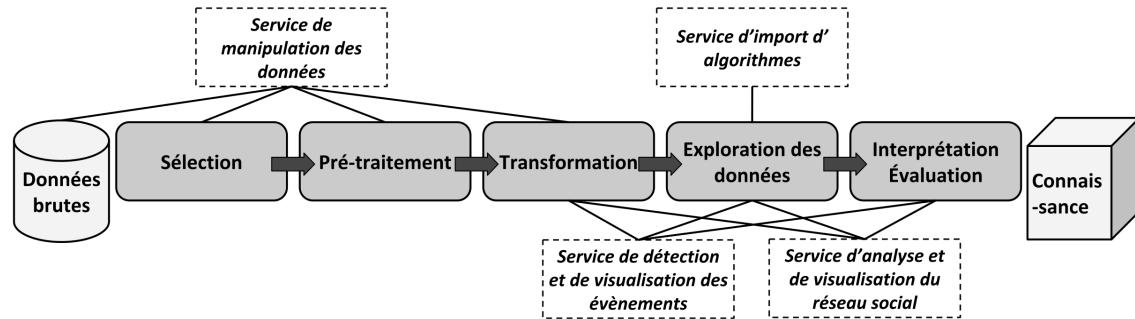


FIGURE 5.6 – Positionnement des services du logiciel SONDY dans le processus typique de fouille de données.

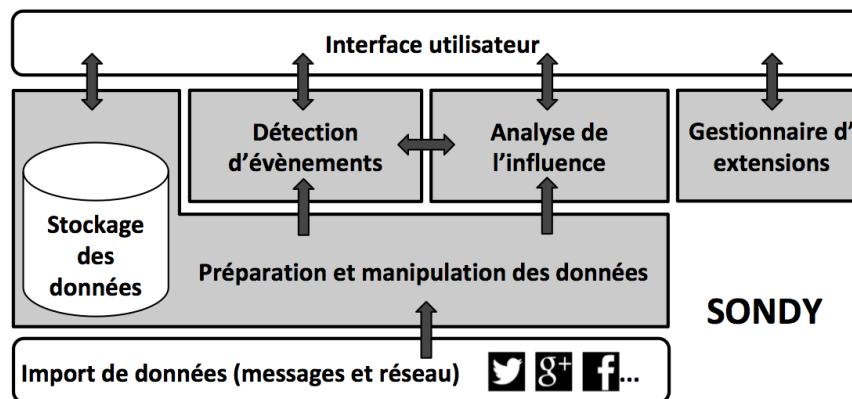


FIGURE 5.7 – Architecture du logiciel SONDY.

Dans les prochaines sections, nous décrivons en détail le rôle et le fonctionnement de chaque service.

### 5.3.2 Service de manipulation des données

Ce service gère une collection de jeux de données (*cf. figure 5.8.a*). Un jeu de données correspond à un ensemble de messages publiés sur un média social ainsi que le réseau interconnectant les auteurs.

**Import de données.** L'import d'un nouveau jeu de données est réalisé à partir de deux fichiers CSV (*cf. figure 5.8.b et b'*). L'un comporte trois colonnes – auteur, date, texte – qui permettent de représenter les messages. L'autre comporte deux colonnes qui permettent de modéliser à l'aide d'un graphe orienté le réseau interconnectant les auteurs des messages. Lors de l'import, une copie de référence du jeu de données est sauvegardée dans une base de données indexée et gérée par *SONDY*. La version actuelle du logiciel utilise un serveur de base de données libre, *MySQL*<sup>12</sup>.

**Pré-traitement des données.** Une fois un jeu de données importé, l'ensemble de messages correspondant peut être préparé afin d'optimiser son analyse par les algorithmes de détections d'évènements. Les pré-traitements implémentés dans *SONDY* sont les suivants (*cf. figure 5.8.c*) :

- *Partitionnement* : discrétise l'axe temporel en partitionnant les messages dans des tranches temporelles d'une durée égale à celle entrée en paramètre.
- *Tokenization* : définit la manière dont le contenu de chaque message est découpé – en unigrammes, bigrammes ou trigrammes.
- *Racinalisation* : supprime les préfixes et suffixes des mots pour ne conserver que leur racine – disponible pour l'anglais et le français.
- *Lemmatisation* : transforme les différentes flexions des mots en leur lemme – disponible pour l'anglais.

Lorsqu'un pré-traitement est appliqué à un jeu de données, l'ensemble de messages qui en résulte est sauvegardé et indexé à l'aide de la bibliothèque libre *Lucene*<sup>13</sup>, ceci dans le but de permettre l'analyse efficace de grands corpus de messages. L'utilisateur du logiciel peut par la suite choisir la « préparation » à utiliser à l'aide d'une liste déroulante (figure 5.8.d). Les préparations sont nommées selon le modèle : pas de

---

12. Le serveur *MySQL* est téléchargeable à l'adresse : <http://www.mysql.com>.

13. La bibliothèque *Lucene* est téléchargeable à l'adresse : <http://lucene.apache.org>.

## Un logiciel libre pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux

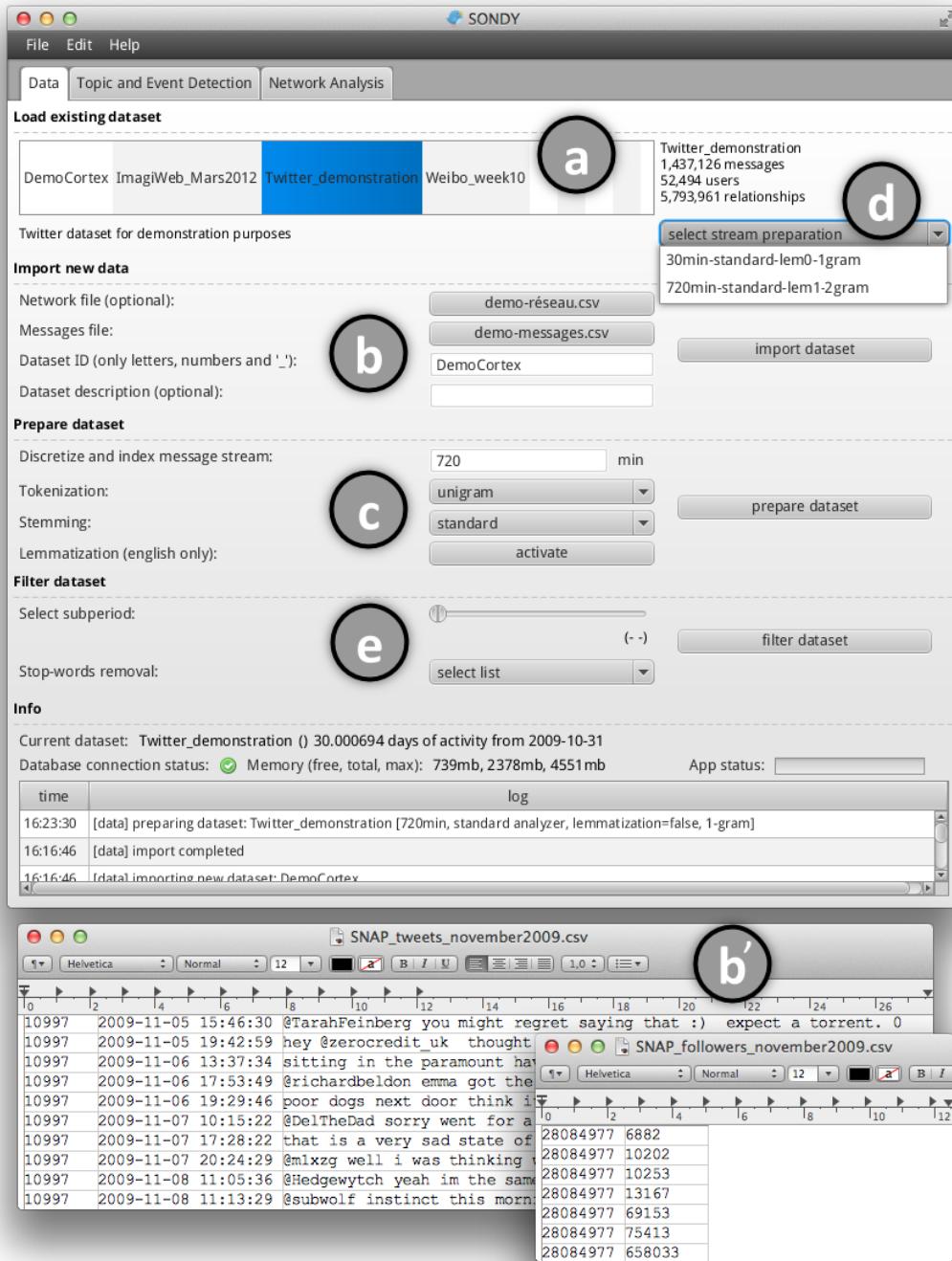


FIGURE 5.8 – La principale fenêtre correspond à l'interface du service de manipulation des données. Les deux petites fenêtres (b') montrent des extraits de fichiers CSV (à gauche, les messages, à droite le réseau social) pouvant être importés par SONDY.

discréttisation (durée en minutes) – mode de racinisation (standard, anglais, français) – lemmatisation (activée : lem1, sinon lem0) – tokenization (1gram, 2gram, 3gram).

**Filtrage des données.** Lorsqu'une préparation a été sélectionnée, il est possible de filtrer les données avant de les traiter à l'aide du service de détection d'évènements. Les filtrages disponibles sont les suivants (*cf. figure 5.8.e*) :

- *Sélection d'une période de temps* : limite la détection d'évènements à une partie du flux de message.
- *Suppression des mots vides* : retire du vocabulaire employé dans les messages les mots d'une des listes de mots vides intégrées à SONDY (anglais, chinois, français, et mot vides spécifiques à Twitter), ou une liste fournie par l'utilisateur.

### 5.3.3 Service de détection d'évènements

Ce service permet de configurer et d'appliquer des algorithmes pour la détection automatique d'évènements (*cf. figure 5.9.a*) à partir d'un jeu de données préparé et éventuellement filtré avec le service de manipulation des données, puis d'explorer les résultats.

**Algorithmes implémentés pour la détection d'évènements.** Plusieurs méthodes récentes tirées de la littérature sont implémentées, à savoir :

- *Peaky Topics* : une méthode de pondération statistique des termes pour détecter les évènements très localisés dans le temps (*Shamma et al., 2011*).
- *Persistent Conversations* : une méthode de pondération statistique des termes pour détecter les évènements suscitant l'intérêt des utilisateurs pendant une période de temps prolongée (*Shamma et al., 2011*).
- *Trending Score* : une méthode de pondération statistique des termes plus spécifiquement adaptée aux N-grammes (*Benhardus et Kalita, 2013*).
- *EDCoW* : une méthode de classification non supervisée des termes fondée sur la mesure de la similarité temporelle entre termes à l'aide de la théorie des ondelettes (*Weng et Lee, 2011*).
- *ET* : une méthode de clustering des termes par classification hiérarchique ascendante fondée sur la mesure de la similarité temporelle et sémantique entre termes (*Parikh et Karlapalem, 2013*).

## Un logiciel libre pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux

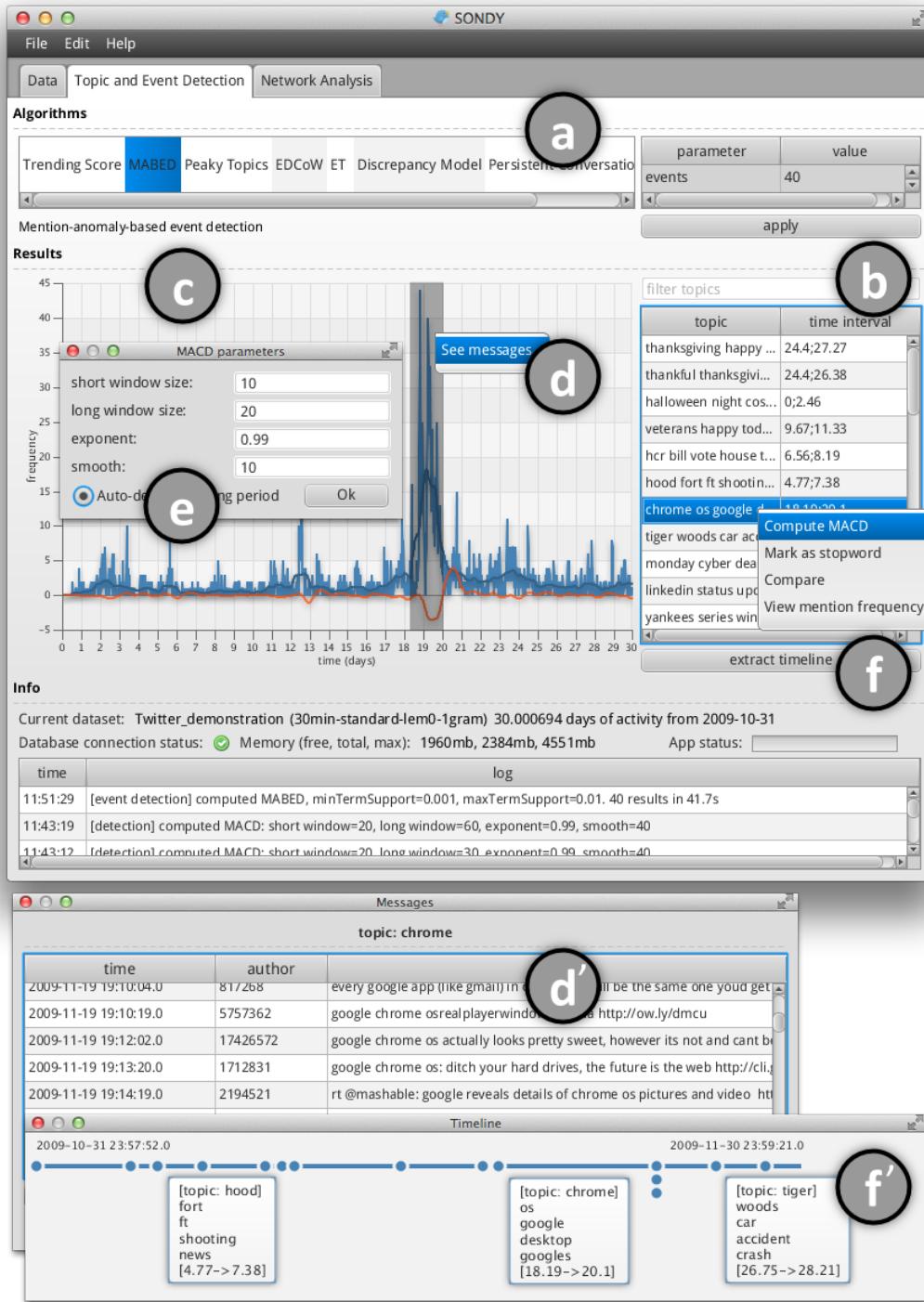


FIGURE 5.9 – La principale fenêtre correspond au cœur de l'interface du service de détection d'évènements. Les deux autres fenêtres correspondent respectivement, de haut en bas, à la fenêtre pour l'exploration des messages (d') et à la frise chronologique des évènements détectés (f').

- *MABED* : une méthode basée sur la mesure de l'anomalie dans la fréquence de création de mentions (*Guille et Favre*, 2014a).
- *Pont vers On-line LDA* : exporte le jeu de données préparé et filtré dans un format compatible avec l'implémentation en python fournie par les auteurs (*Lau et al.*, 2012).

Pour permettre une exploration efficace des évènements détectés par ces algorithmes, *SONDY* offre plusieurs visualisations. Ces visualisations, qui couvrent les trois dimensions des évènements – à savoir l'impact, la thématique et le temps – sont décrites ci-après.

**Liste des évènements détectés.** Lorsqu'un algorithme est appliqué, les évènements détectés sont listés par magnitude d'impact – au sens de l'algorithme utilisé – décroissante dans une table (cf. figure 5.9.b). Chaque évènement est décrit dans cette table par une thématique (un ou plusieurs termes selon l'algorithme utilisé) et un intervalle temporel. Le contenu de la table peut être filtré selon les thématiques, en entrant une expression régulière dans le champ de recherche au-dessus de celle-ci (cf. figure 5.9.b).

**Courbe de fréquence.** Lorsqu'un évènement est sélectionné, sa fréquence d'apparition dans les messages est donnée par le graphique à gauche de la table (cf. figure 5.9.c) et l'intervalle temporel identifié par l'algorithme est automatiquement grisé. Pour faciliter la lecture de la courbe de fréquence, *SONDY* propose de calculer l'indicateur *MACD* (*Rong et Qing*, 2012), qui fait ressortir les irrégularités de la courbe de fréquence. Cet indicateur peut également aider l'utilisateur à raffiner l'intervalle temporel associé aux évènements détectés par certains algorithmes dont la précision temporelle est faible. Pour ce faire, la zone grisée peut être librement modifiée, déplacée, etc.

**Exploration des messages liés aux évènements.** Un clic droit dans la zone grisée (cf. figure 5.9.d) permet d'extraire les messages liés à la thématique de l'évènement et publiés durant l'intervalle temporel sélectionné. La fenêtre pour l'exploration des messages (5.9.d') permet également d'extraire les termes les plus fréquents dans cet ensemble de messages, ce qui est particulièrement utile lorsque l'algorithme utilisé identifie des descriptions d'évènements sémantiquement faibles.

**Frise chronologique des évènements.** Enfin, il est possible de générer (cf. figure 5.9.f') une frise chronologique reprenant les évènements détectés. Le survol avec la

souris des points sur l'axe temporel fait apparaître la description de l'évènement correspondant.

### 5.3.4 Service d'analyse du réseau social

Ce service permet de configurer et d'appliquer des algorithmes pour l'analyse du réseau social des auteurs des messages liés à l'évènement sélectionné dans le service de détection des évènements (*i.e.* un évènement sélectionné dans la table et l'intervalle temporel correspondant à la zone grisée sur le graphique donnant la courbe de fréquence).

**Algorithmes implémentés pour l'analyse du réseau social.** Plusieurs méthodes récentes tirées de la littérature sont implémentées (cf. figure 5.10.a). Leur but est d'associer un rang – dont la signification dépend de la méthode utilisée – à chaque membre du réseau social analysé :

- *Page Rank* : estime le rang de chaque membre du réseau en fonction de son influence, mesurée comme la probabilité qu'un surfeur aléatoire visite le nœud correspondant (Page *et al.*, 1998).
- *K-shell decomposition* : partitionne de manière récursive les membres du réseau en enveloppes, l'enveloppe à laquelle chaque membre appartient correspondant à son rang (Batagelj *et Zaversnik*, 2011).
- *Log K-shell decomposition* : partitionne les membres de façon similaire à la décomposition en k-enveloppes, mais avec une échelle logarithmique pour favoriser la lecture des rangs dans les grands réseaux sociaux (Brown *et Feng*, 2011).
- *Centralité d'intermédiarité* : mesure pour chaque nœud le nombre de chemins les plus courts passant par celui-ci, depuis tous les nœuds, vers tous les autres (Freeman, 1977).
- *Social capitalist identification* : estime le rang de chaque membre du réseau en fonction de sa propension au capitalisme social, mesuré comme le taux de recouvrement entre les ensembles d'arcs entrants et sortants pour chaque membre du réseau (Dugué *et Perez*, 2014).

**Distribution des rangs des utilisateurs.** Lorsqu'un algorithme est appliqué, la distribution du nombre de membres du réseau en fonction des rangs déterminés par

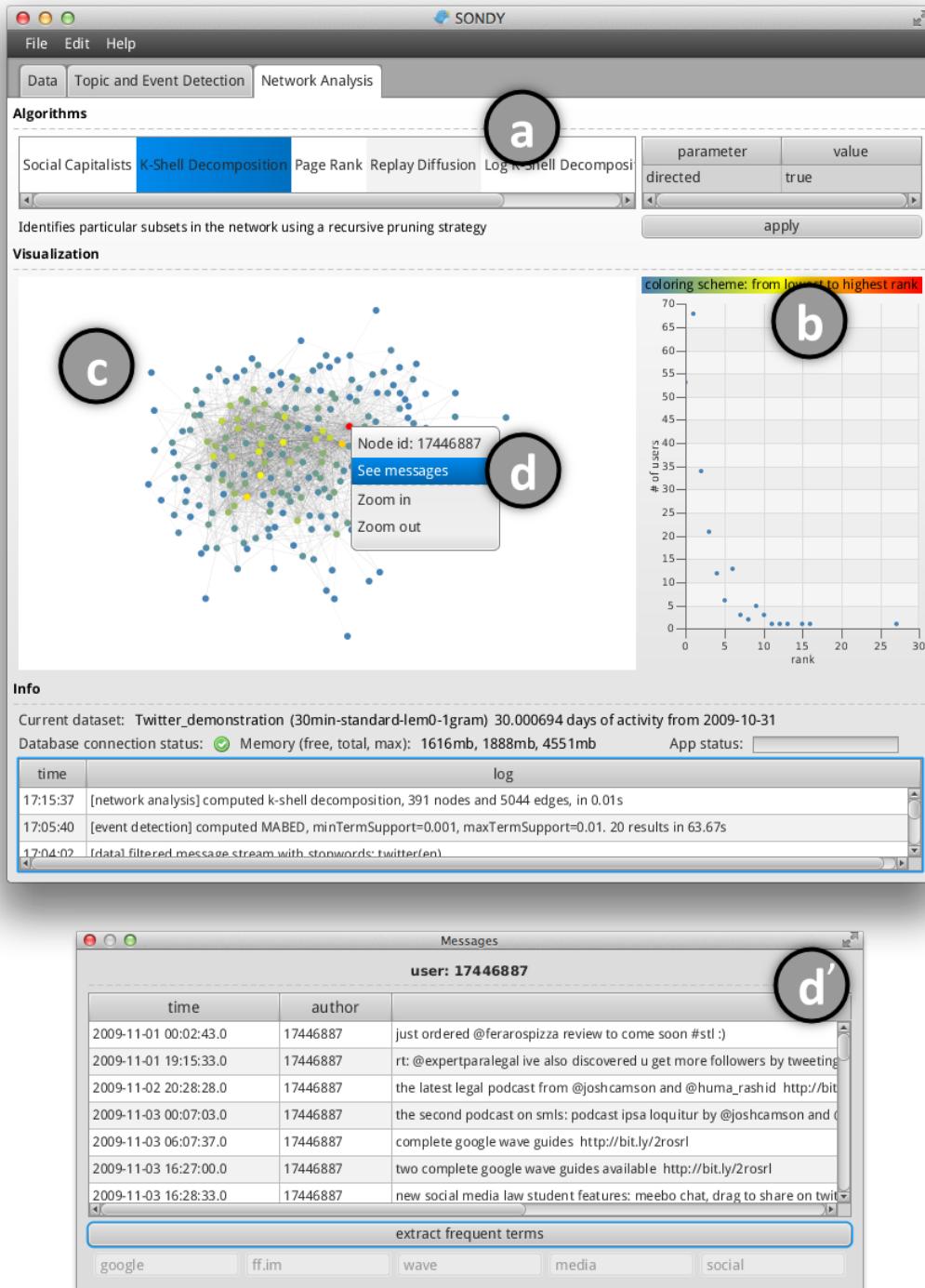


FIGURE 5.10 – L'interface principale du service d'analyse du réseau permet de naviguer dans le réseau social coloré et de consulter la distribution des rangs identifiés par les algorithmes. La seconde fenêtre (d') permet de naviguer parmi les messages publiés par les utilisateurs membres du réseau.

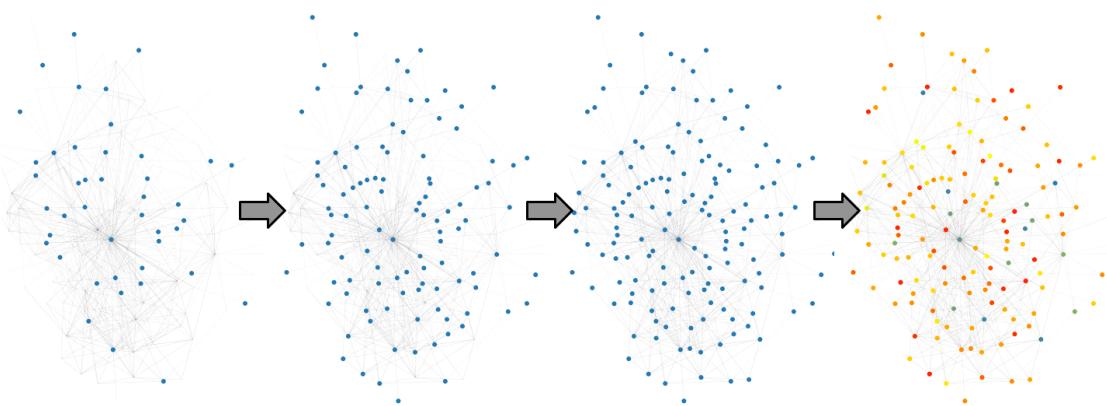


FIGURE 5.11 – Quatre des étapes d'une séquence d'activation capturées à partir de l'interface du service d'analyse du réseau social.

cet algorithme est donnée dans la partie droite de l'interface principale (cf. figure 5.10.b). Au-dessus de cette distribution est présenté un gradient allant du bleu au rouge en passant par le jaune, qui correspond à la répartition des couleurs associées aux noeuds du réseau en fonction de leur rang.

**Navigation dans le réseau social coloré.** Le réseau social analysé est dessiné et coloré selon le gradient et les rangs déterminés par l'algorithme utilisé (cf. figure 5.10.c). Les noeuds du réseau sont positionnés selon l'algorithme de type force-directed proposé par *Fruchterman et Reingold* (1991) et le rendu du graphe est obtenu grâce à la bibliothèque *GraphStream* développée par l'université du Havre (*Dutot et al.*, 2007). La visualisation est interactive et permet à la fois de déplacer le graphe dans le plan de dessin et d'ajuster la distance entre la caméra et le plan.

**Exploration des messages publiés par les utilisateurs.** Un clic droit sur un noeud du réseau social révèle l'identifiant de l'utilisateur qu'il représente et permet d'explorer tous les messages qu'il a publiés (cf. figure 5.10.d'). Cette interface permet également d'extraire les termes les plus fréquents dans les messages de l'utilisateur sélectionné, ce qui peut aider par exemple à déterminer ses centres d'intérêt.

**Visualisation de la séquence d'activation.** Parmi la liste des méthodes disponibles (cf. figure 5.10.a) se trouve une méthode permettant de rejouer la séquence d'activation liée à un évènement, en affichant les noeuds un à un. La figure 5.11 montre quelques étapes de la séquence d'activation engendrée par la diffusion d'une

information à travers le réseau des utilisateurs ayant réagi à son sujet. Une fois tous les nœuds activés, le réseau est coloré en fonction du rang de chaque utilisateur dans la séquence d'activation.

#### 5.3.5 Service d'import d'algorithmes et API

*SONDY* fournit une API qui permet le développement de nouveaux algorithmes compatibles, qui peuvent être importés dynamiquement à l'aide du service d'import d'algorithmes – grâce à un système de plug-ins – sous la forme d'une classe Java ou d'un JAR si plusieurs ressources sont nécessaires à l'exécution de l'algorithme importé.

L'API fournit un ensemble de méthodes pour la manipulation des données ainsi que des interfaces pour concevoir les classes implémentant les algorithmes de détection d'évènements ou d'analyse du réseau social.

**Fonctions pour la manipulation des données.** L'API donne accès aux fonctions de bases pour manipuler le flux de messages : extraction du vocabulaire des termes, fréquence des termes, cooccurrences entre termes, etc. Elle fournit également les fonctions de base pour manipuler la structure du réseau social : liste des nœuds et des liens, identification du voisinage entrant ou sortant d'un nœud.

**Interface pour les algorithmes de détection d'évènements.** Les nouveaux algorithmes doivent implémenter une interface de l'API. Cette interface permet la définition des paramètres de l'algorithme de sorte qu'ils s'intègrent automatiquement à l'interface utilisateur du logiciel *SONDY*. Elle définit par ailleurs une structure générique permettant de stocker les évènements détectés par tout algorithme, qui est reprise automatiquement dans l'interface utilisateur pour différents usages (e.g. listes des évènements, frise chronologique, courbe de fréquence).

**Interface pour les algorithmes d'analyse du réseau social.** Similairement, l'API fournit une interface à implémenter pour développer de nouveaux algorithmes d'analyse du réseau social, qui gère l'intégration des paramètres et des résultats. Elle permet aussi de maintenir la cohérence avec le service de détection d'évènements, en fournissant la structure du sous-graphe correspondant aux utilisateurs ayant publié des messages en lien avec l'évènement sélectionné.

## 5.4 Exemples de scénarios d'utilisation

Dans cette section, nous présentons deux scénarios d'utilisation illustrant les capacités du logiciel *SONDY*. Le premier scénario décrit comment un non-expert peut utiliser le logiciel pour analyser les données dont il dispose. Le deuxième scénario montre comment un chercheur peut utiliser *SONDY* pour réutiliser et comparer les méthodes développées dans le domaine.

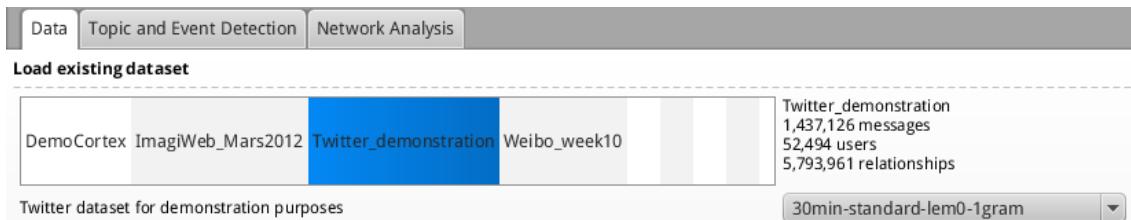
### 5.4.1 Utilisation par un non-expert

On se propose ici d'étudier le média social Twitter du point de vue d'un non-expert. Plus particulièrement, on se propose de s'intéresser à la société Google et à ses produits.

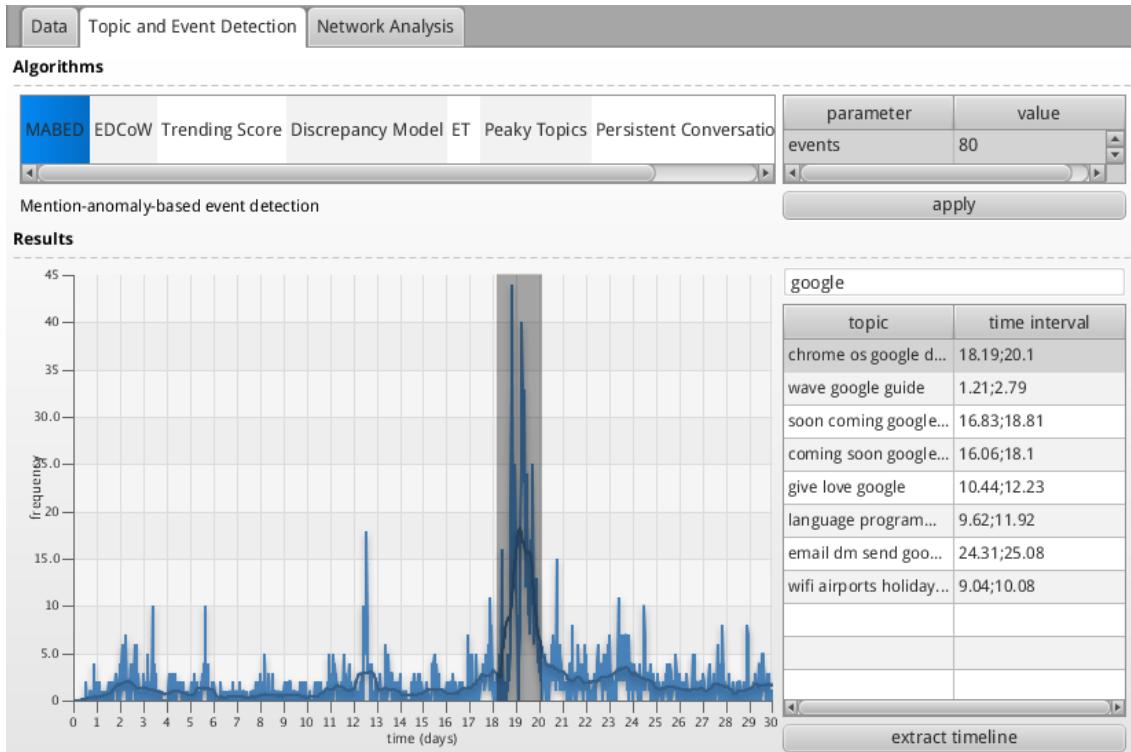
**Sélection et préparation des données.** Pour ce scénario, nous utilisons un jeu de données Twitter. Il représente l'intégralité des 1 437 126 de tweets publiés du 1 au 31 novembre 2009 par 52 494 utilisateurs nord-américains interconnectés par 5 793 961 liens d'abonnement (*cf.* figure 5.12.a). À l'aide du service de manipulation des données, nous partitionnons le flux de messages en tranches de 30 minutes et nous découpons et indexons le contenu des messages en unigrammes. Avant de passer à l'étape de détection des évènements, nous filtrons le vocabulaire du flux de messages avec deux listes de mots vides inclus dans *SONDY*, à savoir la liste des mots communs dans la langue anglaise et la liste des mots vides propres à Twitter (*e.g.* RT, via).

**Détection et analyse des évènements.** Nous utilisons d'abord la méthode *MABED* pour détecter les 80 évènements ayant eu le plus d'impact auprès de ces utilisateurs en novembre 2009, puis nous filtrons la liste obtenue par le mot clé « Google » (*cf.* figure 5.12.b). Il reste 8 évènements dont on peut visualiser la répartition temporelle en générant la frise chronologique (*cf.* figure 5.12.c). L'évènement qui a le plus fait réagir les utilisateurs est la publication par Google du code source du projet Chrome OS, qui a retenu leur attention les 19 et 20 novembre. On observe que les évènements # 6 – la sortie d'un langage de programmation, Go, conçu par Google – et # 8 – l'annonce de l'accès wifi offert dans les aéroports américains par Google – ont suscité l'intérêt des utilisateurs du 10 au 11 novembre (*cf.* figure 5.13.a et b). La lecture des messages liés à ces évènements permet d'en approfondir la compréhension (*cf.* figure

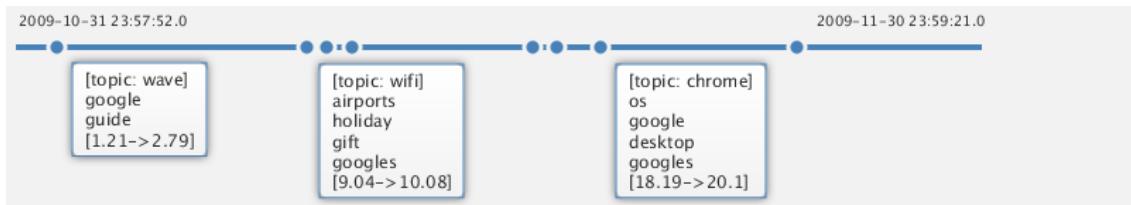
## 5.4. Exemples de scénarios d'utilisation



(a) Sélection du jeu de données.



(b) Liste des évènements concernant Google détectés avec la méthode *MABED* et fréquence de l'évènement # 1.



(c) Frise chronologique des évènements.

FIGURE 5.12 – Exploration des évènements en lien avec Google : (a) sélection des données à analyser, (b) détection des évènements et (c) frise chronologique.

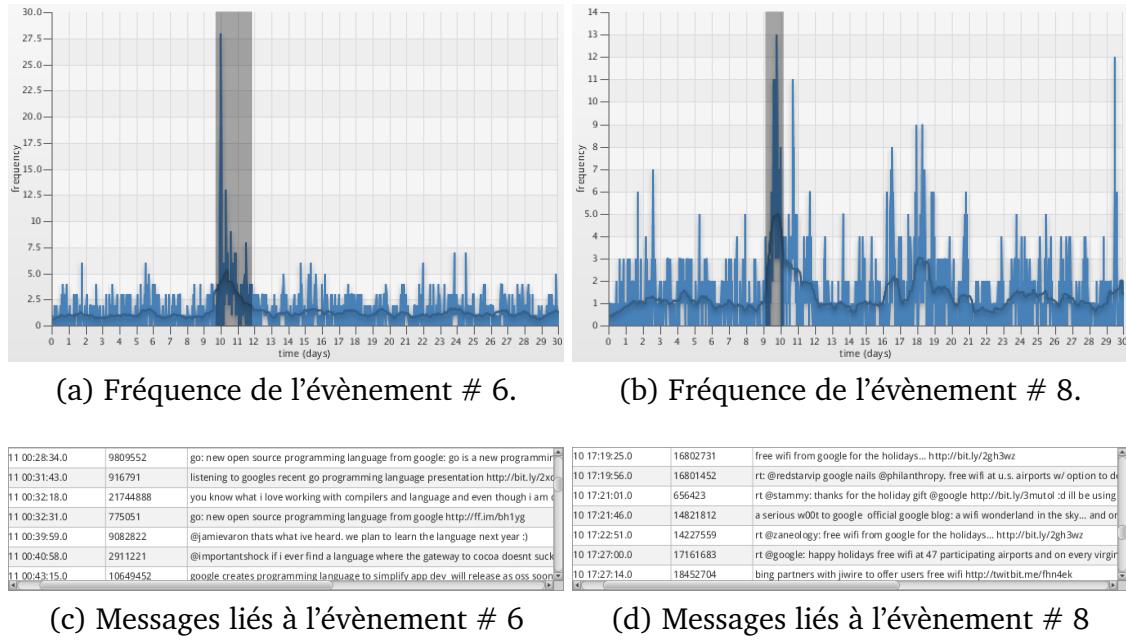
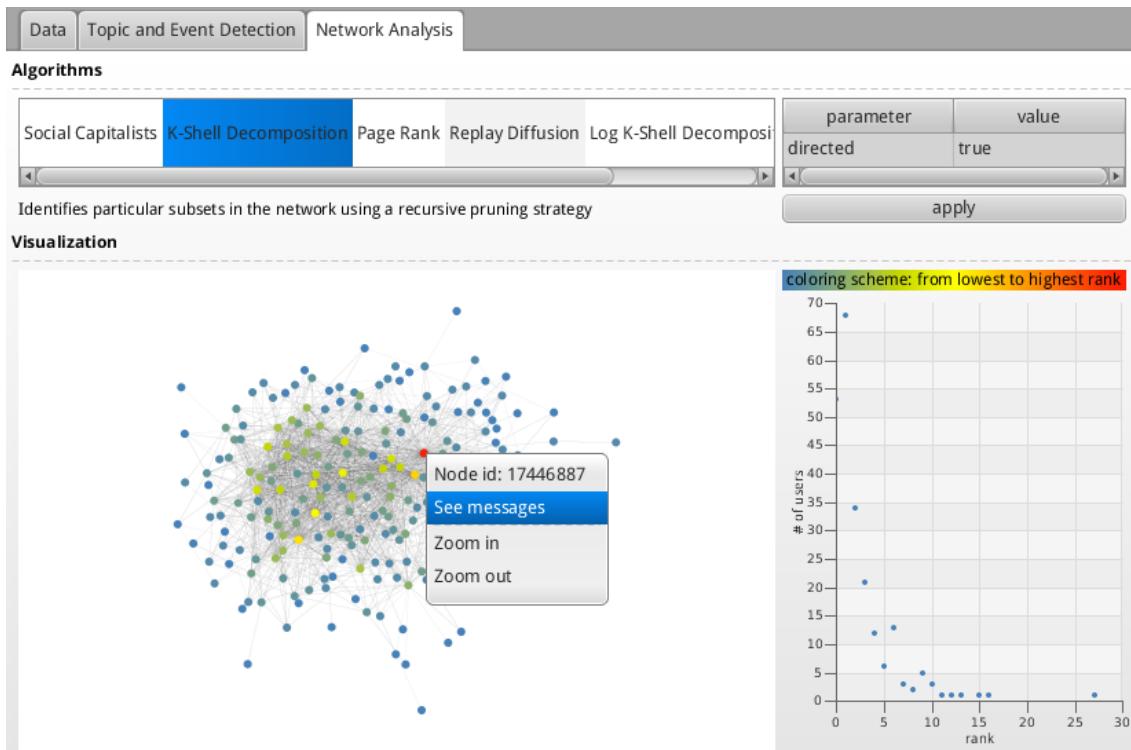


FIGURE 5.13 – Exploration des évènements en lien avec Google : (a,b) courbes de fréquence des évènements et (c,d) messages liés.

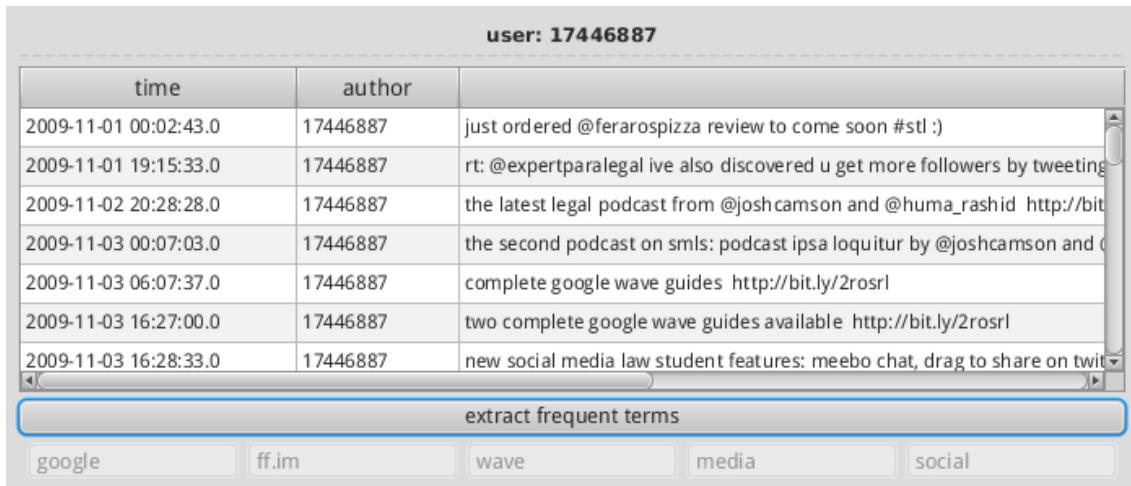
5.13.c et d). Un journaliste spécialisé préparant un dossier sur ce nouveau langage trouvera des ressources intéressantes en lisant les messages liés à l'évènement # 6 où figurent entre autres un lien vers un podcast présentant le langage, un lien vers la page officielle du projet, et diverses réactions d'utilisateurs. La lecture des messages liés à l'évènement # 8 peut par exemple permettre au département marketing de Google de se faire une idée de la réaction suscitée auprès des utilisateurs de Twitter à propos de cette annonce.

**Analyse de l'influence.** Ayant sélectionné le premier évènement de la liste (cf. figure 5.12.b), nous décidons d'étudier l'influence au sein du réseau social formé par les connexions entre les auteurs des messages liés. Nous appliquons la méthode de décomposition en  $k$ -enveloppes (cf. figure 5.14.a). Il apparaît qu'un utilisateur occupe une place centrale dans ce réseau puisque la distribution indique que tous les membres du réseau ont une valeur de  $k$  inférieure à 17 sauf celui-ci dont la valeur  $k$  est 28. La visualisation du réseau permet d'identifier aisément le sommet coloré en rouge qui lui correspond. Cela nous permet de consulter l'ensemble des messages

## 5.4. Exemples de scénarios d'utilisation



(a) Mesure et visualisation de l'influence selon la méthode de décomposition en  $k$ -enveloppes au sein du réseau des auteurs.



(b) Ensemble des messages de l'utilisateur sélectionné, i.e. le plus influent au sens de la méthode appliquée.

FIGURE 5.14 – Identification d'utilisateurs influents à propos de l'évènement le plus marquant concernant Google.

time	log
11:21:59	[event detection] computed peaky topics, minTermSupport=0.001, maxTermSupport=0.01. 1100 results in 21.09s
11:20:06	[event detection] computed EDCoW, minTermSupport=5.0E-4, maxTermSupport=0.01, delta=8, gamma=5. 154 results in 1,013.42s
11:02:39	[data filtered message stream with stopwords: twitter(en)]

FIGURE 5.15 – Fenêtre de log retraçant les opérations effectuées.

qu'il a publiés (*cf. figure 5.14.b*). L'extraction des termes les plus fréquents dans ses messages révèle que cet utilisateur s'intéresse aux produits Google. Cet utilisateur semble donc jouer un rôle important dans le processus de diffusion de l'information liée à Google et pourrait donc être interrogé prioritairement par un journaliste s'intéressant à cet évènement. Par ailleurs, le département marketing de Google pourrait par exemple établir un contact avec cet utilisateur afin qu'il bénéficie d'informations exclusives qu'il pourrait ensuite relayer à moindre coût sur Twitter.

#### 5.4.2 Utilisation par un chercheur du domaine

Dans ce scénario, nous nous intéressons à la comparaison des résultats obtenus en variant les méthodes employées et les préparations des données.

**Comparaison des méthodes de détection d'évènements.** La fenêtre de log (*cf. figure 5.15*) permet de suivre le nombre d'évènements détectés par chacune des méthodes appliquées ainsi que leur temps de calcul.

La figure 5.16 (a,b,c,d) montre la liste des évènements détectés par les méthodes *Peaky Topics*, *Trending Score*, *Persistent Conversations* et *EDCoW* à partir d'un même jeu de données. Les différentes listes révèlent que ces méthodes décrivent les évènements de diverses manières tant du point de vue sémantique (unigramme, N-gramme, ensemble d'unigrammes) que du point de vue temporel (une tranche temporelle, une séquence de tranches temporelles). Par ailleurs, les courbes de fréquences des évènements sélectionnés révèlent que ces méthodes détectent des motifs différents (plus ou moins saillants). Nous pouvons voir en quoi les différentes préparations impactent les résultats du point de vue de la redondance, de la précision sémantique ou temporelle.

La figure 5.17 montre les évènements détectés par une même méthode pour différentes préparations – du point de vue du vocabulaire et de la discréétisation temporelle

## 5.4. Exemples de scénarios d'utilisation

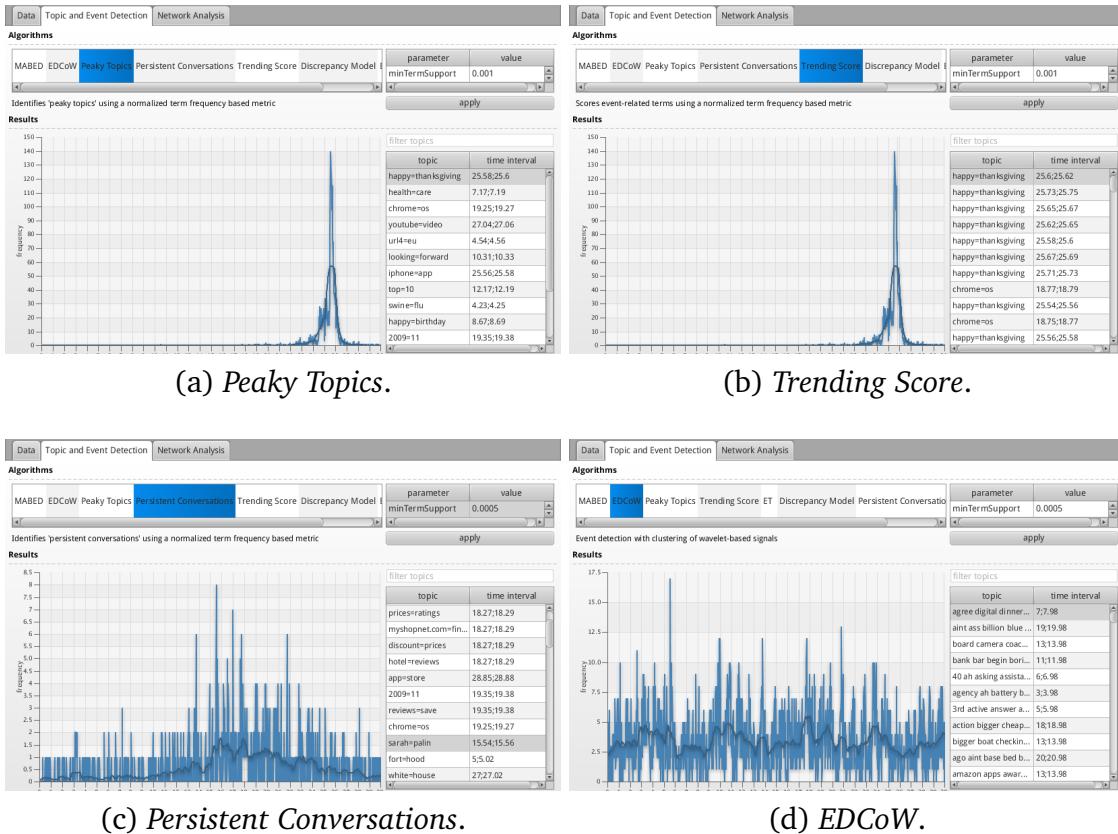


FIGURE 5.16 – Détection d'évènements avec différents algorithmes : *Peaky Topics*, *Trending Score*, *Persistent Conversations* et *EDCoW*.

– d'un même jeu de données. Cela permet d'évaluer la sensibilité de l'algorithme par rapport à la préparation des données, en analysant la variation des résultats, tant en terme de redondance que de précision sémantique ou temporelle.

**Comparaison des méthodes d'analyse de l'influence.** La figure 5.18 montre les résultats obtenus avec les méthodes *Page Rank*, *k-shell decomposition* et *log-k-shell decomposition* pour un même réseau social. Lorsque plusieurs algorithmes sont appliqués successivement au même réseau, SONDY mémorise la position des nœuds, ce qui facilite la comparaison entre les résultats produits par les différentes méthodes.

**Utilisation comme bibliothèque.** Le programme 1 ci-après (page 170) montre comment il est possible d'utiliser SONDY comme bibliothèque pour automatiser une série d'expérimentations. Dans cet exemple, un jeu de données est chargé, préparé puis une méthode de détection d'évènements est appliquée avec différents paramétrages et les résultats obtenus sont écrits sur le disque. Pour réaliser un benchmark plus complet, il serait par exemple possible d'ajouter une boucle à ce programme afin de charger différents jeux de données et appliquer différents pré-traitements, et également charger d'autres algorithmes.

topic	time interval
yankees	4.21;4.23
tiger	26.83;26.85
yankees	4.19;4.21
woods	26.83;26.85
yankees	1.19;1.21
verizon	23.25;23.27
veterans	10.58;10.6
veterans	10.62;10.65
veterans	10.56;10.58
halloween	0.04;0.06
yankees	4.23;4.25

(a) Unigrammes, 30 minutes.

topic	time interval
tiger=woods	26.83;26.85
tiger=woods	26.81;26.83
san=francisco	8.29;8.31
web=design	26.15;26.17
app=store	28.85;28.88
tiger=woods	26.85;26.88
fort=hood	4.94;4.96
happy=thanksgiving	25.6;25.62
app=store	28.9;28.92
happy=thanksgiving	25.73;25.75
cyber=monday	29.79;29.81

(b) Bigrammes, 30 minutes.

topic	time interval
happy=thanksgiving	25.62;25.83
ha=ha	30;30.21
happy=thanksgiving	25.42;25.62
cyber=monday	29.58;29.79
cyber=monday	29.79;30
fort=hood	4.79;5
tiger=woods	26.67;26.88
fort=hood	5;5.21
tiger=woods	26.88;27.08
happy=thanksgiving	25.83;26.04
chrome=os	18.75;18.96

(c) Unigrammes, 300 minutes.

topic	time interval
happy=thanksgiving	25.42;25.83
ha=ha	30;30.42
cyber=monday	29.58;30
tiger=woods	26.67;27.08
cyber=monday	29.17;29.58
happy=thanksgiving	25.83;26.25
chrome=os	19.17;19.58
fort=hood	5;5.42
chrome=os	18.75;19.17
app=store	28.75;29.17
happy=thanksgiving	25;25.42

(d) Bigrammes, 600 minutes.

FIGURE 5.17 – Résultats obtenus par la méthode *Trending Score* pour différentes préparations d'un même jeu de données. L'intervalle temporel est défini en jours, le début de la période couverte par le jeu de données étant associé au jour 0.

## Un logiciel libre pour la détection d'évènements et l'analyse de l'influence dans les médias sociaux



(a) Page Rank.



(b)  $k$ -shell decomposition



(c) log  $k$ -shell decomposition

FIGURE 5.18 – Analyse de l'influence au sein d'un réseau social à l'aide de différentes méthodes.

**Programme 1 :** Utilisation de SONDY comme bibliothèque dans un programme Java.

```

import fr.ericlab.sondy.*;
import org.apache.commons.io.FileUtils;

public class Programme {
    public static void main(String[] args) {
        AppVariables state;
        DataManipulation dataManipulation;
        // import d'un jeu de donnees
        dataManipulation.importDataset("messages.csv", "network.csv",
            "Nom", "Description optionnelle", state);
        // preparation du jeu de donnees
        dataManipulation.prepareStream(60, "English", false, state);
        // chargement de la methode MABED
        EventDetectionAlgorithm mabed = (EventDetectionAlgorithm)
            Class.forName("MABED").newInstance(state);
        for(double i = 0.2; i <= 1; i += 0.1){
            // variation du parametre sigma de la methode
            mabed.sigma = i;
            mabed.k = 40;
            mabed.theta = 0.7
            mabed.p = 10;
            mabed.apply();
            EventDetectionResults results = mabed.getResults();
            // ecriture des resultats
            FileUtils.write("chemin", results);
        }
    }
}

```

## 5.5 Discussion

Dans ce chapitre nous avons décrit SONDY, un logiciel libre et extensible pour la détection d'événements et l'analyse de l'influence à partir de données générées par les médias sociaux. Il vient combler un manque dans l'offre de logiciels développés dans le milieu académique, qui se concentrent sur l'une ou l'autre de ces tâches. Le logiciel SONDY permet, à l'aide des méthodes tirées de la littérature qu'il implémente et des outils de visualisations qu'il offre, la détection et l'exploration d'événements à partir des messages publiés par un ensemble d'utilisateurs, puis, pour chaque événement,

l'analyse de l'influence à partir de la structure du réseau social correspondant au sous-ensemble d'utilisateurs ayant réagi. Nous avons montré comment, à l'aide de plusieurs scénarios et de données réelles, un non-expert ou un chercheur du domaine peut utiliser *SONDY*. De plus, les capacités d'extensibilité du logiciel ont été démontrées par l'implémentation d'un algorithme de détection d'évènements (*EDCoW*) avec l'API de *SONDY* par des étudiants de master de l'université Lumière Lyon 2, ainsi qu'un algorithme d'analyse de l'influence (*Social capitalists identification*) par des chercheurs de l'université d'Orléans. Une des principales perspectives de travail consiste à améliorer la documentation (tant pratique que technique) du logiciel, afin de faciliter son utilisation.

**Impact.** Ces travaux ont notamment fait l'objet d'un article dans la session démonstration et d'une présentation interactive à la conférence internationale *ACM SIGMOD* en 2013. Le logiciel *SONDY* a été téléchargé plus de 700 fois entre mars 2013 et octobre 2014.



# Chapitre 6

## Conclusion

Pour conclure ce manuscrit de thèse, nous résumons tout d'abord les travaux que nous avons présentés, puis nous terminons en synthétisant les principales perspectives de recherche ouvertes par ces travaux.

### 6.1 Résumé de la thèse

Dans cette thèse, nous nous sommes intéressés à la diffusion de l'information dans les médias sociaux, et avons apporté des solutions à certaines des problématiques majeures liées à ce phénomène.

Premièrement, nous avons proposé *MABED* (Mention-Anomaly-Based Event Detection), une méthode statistique pour détecter automatiquement les événements importants qui suscitent l'intérêt des utilisateurs des médias sociaux à partir du flux de messages qu'ils publient, dont l'originalité est d'exploiter la fréquence des interactions sociales entre utilisateurs, en plus du contenu textuel des messages. La méthode *MABED* diffère par ailleurs des méthodes existantes en ce qu'elle estime dynamiquement la durée de chaque événement, plutôt que de supposer une durée commune et fixée à l'avance pour tous les événements. Deuxièmement, nous avons proposé *T-BASIC* (Time-Based ASynchronous Independent Cascades), un modèle probabiliste basé sur la structure de réseau sous-jacente aux médias sociaux pour prévoir la diffusion de l'information, plus précisément l'évolution du volume d'utilisateurs relayant une information donnée au fil du temps. Contrairement aux modèles similaires également basés sur la structure du réseau, la probabilité qu'une information donnée se diffuse entre deux utilisateurs n'est pas constante mais dépendante du temps. Nous avons décrit une procédure pour l'inférence des paramètres latents du modèle (probabilité

de diffusion et délai de transmission pour chaque lien du réseau), dont l'originalité est de formuler les paramètres comme des fonctions de caractéristiques observables des utilisateurs. Troisièmement, nous avons proposé *SONDY* (SOcial Network DYnamics), un logiciel libre implémentant des méthodes tirées de la littérature pour la fouille et l'analyse des données issues des médias sociaux. Le logiciel manipule deux types de données : les messages publiés par les utilisateurs, et la structure du réseau social interconnectant ces derniers. Contrairement aux logiciels académiques existants qui se concentrent soit sur l'analyse des messages, soit sur l'analyse du réseau, *SONDY* permet d'analyser ces deux types de données conjointement en permettant l'analyse de l'influence par rapport aux événements détectés. Utilisé comme logiciel autonome, *SONDY* offre une interface utilisateur avancée et des visualisations adaptées. Utilisé comme bibliothèque, il permet d'intégrer facilement les méthodes implémentées dans d'autres programmes.

Les expérimentations que nous avons menées à l'aide de divers jeux de données collectés sur le média social Twitter (plusieurs millions de messages publiés par des utilisateurs interconnectés par plusieurs millions de liens) ont démontré la pertinence de nos propositions et ont mis en lumière des propriétés qui nous aident à mieux comprendre les mécanismes régissant la diffusion de l'information. Premièrement, en comparant les performances de *MABED* avec celles de méthodes récentes tirées de la littérature, nous avons montré que la prise en compte des interactions sociales entre utilisateurs conduit à une détection plus précise des événements importants, avec une robustesse accrue en présence de contenu bruité. Nous avons également montré que *MABED* facilite l'interprétation des événements détectés en fournissant des descriptions claires et précises, tant sur le plan sémantique que temporel. Deuxièmement, nous avons montré la validité de la procédure proposée pour estimer les probabilités de diffusion sur lesquelles repose le modèle *T-BASIC*, en illustrant le pouvoir prédictif des caractéristiques des utilisateurs que nous avons sélectionnées et en comparant les performances de la méthode d'estimation proposée avec celles de méthodes tirées de la littérature. Nous avons également montré l'intérêt d'avoir des probabilités non constantes, ce qui permet de prendre en compte dans *T-BASIC* la fluctuation du niveau de réceptivité des utilisateurs des médias sociaux au fil du temps. Enfin, nous avons montré comment, et dans quelle mesure, les caractéristiques sociales, thématiques et temporelles des utilisateurs affectent la diffusion de l'information. Troisièmement,

nous avons illustré à l'aide de divers scénarios d'utilisation l'utilité du logiciel *SONDY*, autant pour des non-experts que pour des chercheurs du domaine. Nous avons par exemple montré comment une société peut utiliser *SONDY* pour détecter les évènements la concernant et qui font réagir les utilisateurs des médias sociaux, et ensuite identifier des personnes influentes, lesquelles pourraient potentiellement participer à des campagnes marketing virales. Nous avons aussi montré comment un chercheur du domaine peut utiliser *SONDY* comme logiciel autonome pour expérimenter des méthodes de la littérature, mais aussi comment il peut concevoir un programme utilisant l'interface de programmation de *SONDY*, afin par exemple d'automatiser une série d'expérimentations.

## 6.2 Perspectives de travail

Les travaux de recherche entamés durant cette thèse et présentés dans ce manuscrit ouvrent plusieurs perspectives de recherche intéressantes.

En conclusion du chapitre 3 nous avons présenté une perspective de travail, appuyée par des premières expérimentations prometteuses menées à l'aide de *MABED* et de la méthode de Louvain (*Blondel et al.*, 2008), reposant sur l'intuition selon laquelle les communautés d'utilisateurs au sens social – c'est-à-dire les communautés identifiables à partir de la structure du réseau social que forment les utilisateurs d'un média social – sont similaires aux communautés d'utilisateurs au sens thématique – c'est-à-dire les communautés identifiables à partir des évènements à propos desquels les utilisateurs réagissent. En particulier, il semblerait intéressant de pouvoir exploiter la détection d'évènements pour évaluer la pertinence des communautés identifiées à partir de la structure du réseau social, voire même pour améliorer la pertinence des communautés identifiées.

Concernant la modélisation et la prédiction de la diffusion de l'information, nous avons mentionné en conclusion du chapitre 4 plusieurs pistes pour améliorer nos travaux. L'une d'elles consisterait à supposer que le réseau servant de support à la propagation de l'information puisse évoluer, plutôt que le considérer statique comme c'est le cas avec *T-BASIC*. Aussi, plutôt que de modéliser la diffusion de chaque thématique indépendamment des autres, il serait intéressant de modéliser ces processus simultanément, afin de pouvoir prendre en compte les interactions entre thématiques. Enfin,

puisque dans certains cas l'influence externe à un média social peut jouer un rôle prépondérant dans le processus de diffusion, il pourrait être intéressant de relâcher l'hypothèse de monde fermé sur laquelle se fonde notre approche.

Au-delà de ces perspectives directes, nous envisageons plusieurs autres directions pour nos futurs travaux. Par exemple, il serait intéressant d'expérimenter les algorithmes et modèles que nous avons proposés avec des données collectées à partir de médias sociaux autres que Twitter – chose que nous n'avons malheureusement pu faire jusqu'à présent du fait de la difficulté à obtenir des données à partir d'autres services. Cela nous permettrait d'étudier la générnicité pratique de nos propositions, mais aussi de savoir dans quelle mesure les propriétés du phénomène de diffusion de l'information mis en avant par nos travaux sont indépendantes ou non du média social étudié. Il serait par ailleurs intéressant de considérer la problématique de la détection de sentiments et d'opinion. Cela pourrait par exemple nous permettre d'améliorer la description des évènements détectés avec *MABED* à partir des médias sociaux, et pourrait également faire l'objet d'un service supplémentaire dans le logiciel *SONDY*.

# Bibliographie

- Aggarwal, C. C. (2011), *Social Network Data Analytics*, Springer.
- Aiello, L. M., A. Barrat, R. Schifanella, C. Cattuto, B. Markines, et F. Menczer (2012), Friendship prediction and homophily in social media, *ACM Trans. Web*, 6(2), 137–170.
- Aiello, L. M., G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, Y. Kompatsiaris, et A. Jaimes (2013), Sensing trending topics in twitter, *IEEE Trans. Multimedia*, 15(6), 1–15.
- AlSumait, L., D. Barbará, et C. Domeniconi (2008), On-line lda : Adaptive topic models for mining text streams with applications to topic detection and tracking, in *ICDM '08*, pp. 3–12.
- Anagnostopoulos, A., R. Kumar, et M. Mahdian (2008), Influence and correlation in social networks, in *KDD '08*, pp. 7–15.
- Appel, G. (2005), Technical analysis power tools for active investors, *Financial Times Prentice Hall*, pp. 166–167.
- Auber, D. (2004), Tulip – a huge graph visualization framework, in *Graph Drawing Software*, pp. 105–126, Springer.
- Backstrom, L., et J. Leskovec (2011), Supervised random walks : Predicting and recommending links in social networks, in *WSDM '11*, pp. 635–644.
- Bakshy, E., I. Rosenn, C. Marlow, et L. A. Adamic (2012), The role of social networks in information diffusion, in *WWW '12*, pp. 519–528.
- Banerjee, A. V. (1992), A simple model of herd behavior, *The Quarterly Journal of Economics*, 107(3), 797–817.
- Bastian, M., S. Heyman, et M. Jacomy (2009), Gephi : An open source software for exploring and manipulating networks, in *ICWSM*, pp. 361–362.
- Batagelj, V., et M. Zaversnik (2011), Fast algorithms for determining (generalized) core groups in social networks, *Advances in Data Analysis and Classification*, 5(2), 129–145.

- Benhardus, J., et J. Kalita (2013), Streaming trend detection in twitter, *IJWBC*, 9(1), 122–139.
- Bentley, J. (1984), Programming pearls : algorithm design techniques, *CACM*, 27(9), 865–873.
- Bernstein, M. S., B. Suh, L. Hong, J. Chen, S. Kairam, et E. H. Chi (2010), Eddi : Interactive topic-based browsing of social status streams, in *UIST '10*, pp. 303–312.
- Bi, B., Y. Tian, Y. Sismanis, A. Balmin, et J. Cho (2014), Scalable topic-specific influence analysis on microblogs, in *WSDM '14*, pp. 513–522, doi :10.1145/2556195.2556229.
- Blei, D., A. Ng, et M. Jordan (2003), Latent dirichlet allocation, *JMLR*, 3, 993–1022.
- Blondel, V., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008), Fast unfolding of communities in large networks, *Journal of Statistical Mechanics : Theory and Experiment*, P10008, 1–12.
- Bouillot, F., P. Nhat Hai, N. Béchet, S. Bringay, D. Ienco, S. Matwin, P. Poncelet, M. Roche, et M. Teisseire (2012), How to extract relevant knowledge from tweets ?, in *ISIP '12*, pp. 111–120.
- Boyd, D. (2006), Friends, friendsters, and myspace top 8 : Writing community into being on social network sites, *First Monday*, 11(12).
- Boyd, D., et N. Ellison (2007), Social network sites : Definition, history, and scholarship, *Journal of Computer-Mediated Communication*, 13(1), 210–230.
- Brockmann, D., L. Hufnagel, et T. Geisel (2006), The scaling laws of human travel, *Nature*, 439(7075), 462–465.
- Brown, P., et J. Feng (2011), Measuring user influence on twitter using modified k-shell decomposition, in *ICWSM '11 Workshops*.
- Can, F., T. Özyer, et F. Polat (2014), *State of the Art Applications of Social Network Analysis*, Springer.
- Casella, G., et E. I. George (1992), Explaining the gibbs sampler, *The American Statistician*, 46(3), 167–174.
- Cheng, J.-J., Y. Liu, B. Shen, et W.-G. Yuan (2013), An epidemic model of rumor diffusion in online social networks, *The European Physical Journal B*, 86(1).

## Bibliographie

---

- Coleman, T. F., et Y. Li (1996), A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables, *SIAM J. on Optimization*, 6(4), 1040–1058.
- Combe, D., C. Largeron, E. Egyed-Zsigmond, M. Géry, et al. (2013), Totem : une méthode de détection de communautés adaptées aux réseaux d'information, in *EGC '13*, pp. 305–310.
- Cornillon, P., et E. Matzner-Løber (2007), *Régression*, Springer.
- Cortes, C., et V. Vapnik (1995), Support-vector networks, *Machine Learning*, 20(3), 273–297.
- Crandall, D., D. Cosley, D. Huttenlocher, J. Kleinberg, et S. Suri (2008), Feedback effects between similarity and social influence in online communities, in *KDD '08*, pp. 160–168.
- David, H., F. Mosteller, et J. Tukey (1983), *Understanding Robust and Exploratory Data Analysis.*, John Wiley & Sons.
- Do, C. B., et S. Batzoglou (2008), What is the expectation maximization algorithm ?, *Nature Biotechnology*, 26, 897 – 899.
- Dugué, N., V. Labatut, et A. Perez (2014), Identifying the community roles of social capitalists in the twitter network, in *ASONAM '14*, p. à paraître.
- Dugué, N., et A. Perez (2014), Social capitalists on twitter : detection, evolution and behavioral analysis, *Social Network Analysis and Mining*, 4(1), 178.
- Dutot, A., F. Guinand, D. Olivier, et Y. Pigné (2007), Graphstream : A tool for bridging the gap between complex systems and dynamic graphs, in *EPNACS '07*.
- Easley, D., et J. Kleinberg (2010), *Networks, Crowds, and Markets : Reasoning About a Highly Connected World*, Cambridge University Press.
- Erdem, O., E. Ceyhan, et Y. Varli (2012), A new correlation coefficient for bivariate time-series data, in *MAF*, pp. 58–73.
- Even-Dar, E., et A. Shapira (2007), A note on maximizing the spread of influence in social networks, in *Internet and Network Economics, Lecture Notes in Computer Science*, vol. 4858, pp. 281–286, Springer Berlin Heidelberg.
- Fan, T.-H., S. Lee, H.-I. Lu, T.-S. Tsou, T.-C. Wang, et A. Yao (2003), An optimal algorithm for maximum-sum segment and its application in bioinformatics, in *CIAA*, pp. 251–257.

- Freeman, L. (1977), A set of measures of centrality based on betweenness, *Sociometry*, 40, 35–41.
- Fruchterman, T. M. J., et E. M. Reingold (1991), Graph drawing by force-directed placement, *Software : Practice and Experience*, 21(11), 1129–1164.
- Fukuda, T., Y. Morimoto, S. Morishita, et T. Tokuyama (1996), Data mining using two-dimensional optimized association rules : scheme, algorithms, and visualization, in *SIGMOD '96*, pp. 13–23.
- Fung, G. P. C., J. X. Yu, P. S. Yu, et H. Lu (2005), Parameter free bursty events detection in text streams, in *VLDB*, pp. 181–192.
- Galuba, W., K. Aberer, D. Chakraborty, Z. Despotovic, et W. Kellerer (2010), Outtweeting the twitterers - predicting information cascades in microblogs, in *WOSN '10*, pp. 3–11.
- Goldenberg, J., B. Libai, et E. Muller (2001), Talk of the network : A complex systems look at the underlying process of word-of-mouth, *Marketing Letters*.
- Granovetter, M. (1978), Threshold models of collective behavior, *American journal of sociology*, pp. 1420–1443.
- Guille, A. (2013), Information diffusion in online social networks, in *SIGMOD/PODS Ph.D. Symposium '13*, pp. 31–36.
- Guille, A., et C. Favre (2014a), Mention-anomaly-based event detection and tracking in twitter, in *ASONAM '14*, pp. 375–382.
- Guille, A., et C. Favre (2014b), Une méthode pour la détection de thématiques populaires sur twitter, in *EGC '14*, pp. 83–88.
- Guille, A., et H. Hacid (2012), A predictive model for the temporal dynamics of information diffusion in online social networks, in *WWW '12 (companion volume)*, pp. 1145–1152.
- Guille, A., H. Hacid, et C. Favre (2012), Une approche multidimensionnelle basée sur les comportements individuels pour la prédition de la diffusion de l'information sur twitter, in *EGC '12*, pp. 239–244.
- Guille, A., C. Favre, H. Hacid, et D. Zighed (2013a), Sondy : An open source platform for social dynamics mining and analysis, in *SIGMOD '13*, pp. 1005–1008.
- Guille, A., C. Favre, et D. Zighed (2013b), Sondy : une plateforme open-source d'analyse et de fouille pour les réseaux sociaux en ligne, in *EGC '13 (Recueil des démonstrations)*, pp. 45–48.

## Bibliographie

---

- Guille, A., H. Hacid, C. Favre, et D. Zighed (2013c), Information diffusion in online social networks : A survey, *SIGMOD Record*, 42(2), 17–28.
- Hethcote, H. W. (2000), The mathematics of infectious diseases, *SIAM REVIEW*, 42(4), 599–653.
- Hopcroft, J., et R. Tarjan (1973), Algorithm 447 : efficient algorithms for graph manipulation, *CACM*, 16(6), 372–378.
- Howard, P. N., et A. Duffy (2011), Opening closed regimes, what was the role of social media during the arab spring ?, *Project on Information Technology and Political Islam*, pp. 1–30.
- Hughes, A., et L. Palen (2009), Twitter adoption and use in mass convergence and emergency events, *International Journal of Emergency Management*, 6(3), 248–260.
- Jarvis, R. A., et E. A. Patrick (1973), Clustering using a similarity measure based on shared near neighbors, *IEEE Trans. Comput.*, 22(11), 1025–1034.
- John, G. H., et P. Langley (1995), Estimating continuous distributions in bayesian classifiers, in *UAI*, pp. 338–345.
- Katona, Z., P. Zubcsek, et M. Sarvary (2011), Network effects and personal influences : The diffusion of an online social network, *Journal of Marketing Research*, 48(3), 425–443.
- Kempe, D. (2003), Maximizing the spread of influence through a social network, in *KDD '03*, pp. 137–146.
- Kermack, W. O., et A. G. McKendrick (1927), A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society A : Mathematical, Physical and Engineering Sciences*, 115(772), 700–721.
- Kitsak, M., L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, et H. Makse (2010), Identification of influential spreaders in complex networks, *Nature Physics*, 6(11), 888–893.
- Kleinberg, J. (2002), Bursty and hierarchical structure in streams, in *KDD*, pp. 91–101.
- Kwak, H., C. Lee, H. Park, et S. Moon (2010), What is Twitter, a social network or a news media ?, in *WWW '10*, pp. 591–600.
- Lappas, T., B. Arai, M. Platakis, D. Kotsakos, et D. Gunopoulos (2009), On burstiness-aware search for document sequences, in *KDD*, pp. 477–486.

- Lau, J. H., N. Collier, et T. Baldwin (2012), On-line trend analysis with topic models : #twitter trends detection topic model online, in *COLING*, pp. 1519–1534.
- Lee, P., L. V. Lakshmanan, et E. Milios (2013), Keysee : Supporting keyword search on evolving events in social streams, in *KDD '13*, pp. 1478–1481.
- Leskovec, J., M. Mcglohon, C. Faloutsos, N. Glance, et M. Hurst (2007), Cascading behavior in large blog graphs, in *SDM '07*.
- Leskovec, J., L. Backstrom, et J. Kleinberg (2009), Meme-tracking and the dynamics of the news cycle, in *KDD '09*, pp. 497–506.
- Li, C., A. Sun, et A. Datta (2012), Twevent : Segment-based event detection from tweets, in *CIKM*, pp. 155–164.
- Lin, J. (1991), Divergence measures based on the shannon entropy, *IEEE Transactions on Information Theory*, 37(1), 145–151.
- Makkonen, J., H. Ahonen-Myka, et M. Salmenkivi (2004), Simple semantics in topic detection and tracking, *Inf. Retr.*, 7(3-4), 347–368.
- Marcus, A., M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, et R. C. Miller (2011), Twitinfo : aggregating and visualizing microblogs for event exploration, in *CHI '11*, pp. 227–236.
- McCullagh, P., et J. A. Nelder (1989), *Generalized Linear Models, Second edition.*, Chapman and Hall, London.
- McMinn, A. J., Y. Moshfeghi, et J. M. Jose (2013), Building a large-scale corpus for evaluating event detection on twitter, in *CIKM '13*, pp. 409–418.
- Milgram, S. (1967), The small-world problem, *Psychology Today*, 1(1), 61 ?67.
- Milgram, S., L. Bickman, et L. Berkowitz (1969), Note on the drawing power of crowds of different size, *Journal of Personality and Social Psychology*, 13(2), 79–82.
- Mochalova, A., et A. Nanopoulos (2013), On the role of centrality in information diffusion in social networks, in *ECIS '13*, pp. 101 – 112.
- Morstatter, F., J. Pfeffer, H. Liu, et K. M. Carley (2013), Is the sample good enough ? comparing data from twitter's streaming api with twitter's firehose, in *ICWSM*, pp. 400–408.
- Mosteller, F. (1968), Association and estimation in contingency tables, *Journal of the American Statistical Association*, 63(321), 1–28.

## Bibliographie

---

- Motoda, H. (2011), Learning information diffusion models from observation and its application to behavior analysis, in *SocInfo '11*, pp. 6–11.
- Myers, S., et J. Leskovec (2012), Clash of the contagions : Cooperation and competition in information diffusion, in *ICDM '12*.
- Myers, S. A., C. Zhu, et J. Leskovec (2012), Information diffusion and external influence in networks, in *KDD '12*, pp. 33–41.
- Myers, S. A., A. Sharma, P. Gupta, et J. Lin (2014), Information network or social network ? : The structure of the twitter follow graph, in *WWW (companion volume)*, pp. 493–498.
- Newman, M. E. J. (2006), Modularity and community structure in networks, *Proceedings of the National Academy of Sciences of the USA*, 103(23), 8577–8582.
- Ngonmang, B., E. Viennet, et M. Tchuente (2014), Predicting users behaviours in distributed social networks using community analysis, in *State of the Art Applications of Social Network Analysis*, pp. 119–138.
- Oxford, U. P (2009), OUP dictionary team monitors twitterer and tweets. <http://blog.oup.com/2009/06/oxford-twitter/>.
- Page, L., S. Brin, R. Motwani, et T. Winograd (1998), The pagerank citation ranking : Bringing order to the web, in *WWW '98*, pp. 161–172.
- Parikh, R., et K. Karlapalem (2013), Et : events from tweets, in *WWW (companion volume)*, pp. 613–620.
- PearAnalytics (2009), Twitter study, *Tech. rep.*, <http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf>.
- Rogers, E. M. (1995), *Diffusion of Innovations.*, 4 ed., Free Press.
- Romero, D. M., B. Meeder, et J. Kleinberg (2011), Differences in the mechanics of information diffusion across topics : idioms, political hashtags, and complex contagion on twitter, in *WWW*, pp. 695–704.
- Rong, L., et Y. Qing (2012), Trends analysis of news topics on twitter, *International Journal of Machine Learning and Computing*, 2(3), 327–332.
- Saito, K., M. Kimura, R. Nakano, et H. Motoda (2009), Finding influential nodes in a social network from information diffusion data, in *SBP '09*, pp. 138–145.
- Saito, K., M. Kimura, K. Ohara, et H. Motoda (2010a), Selecting information diffusion models over social networks for behavioral analysis, in *PKDD '10*, pp. 180–195.

- Saito, K., M. Kimura, K. Ohara, et H. Motoda (2010b), Discovery of super-mediators of information diffusion in social networks, in *DS '10*, pp. 144–158.
- Saito, K., M. Kimura, K. Ohara, et H. Motoda (2011), Efficient discovery of influential nodes for sis models in social networks, *Knowledge and Information Systems*.
- Salton, G., et M. J. McGill (1986), *Introduction to Modern Information Retrieval*.
- Seidman, S. B. (1983), Network structure and minimum degree, *Social Networks*, 5(3), 269 – 287.
- Serazzi, G., et S. Zanero (2004), Computer virus propagation models, in *Performance Tools and Applications to Networked Systems, Lecture Notes in Computer Science*, vol. 2965, pp. 26–50.
- Shamma, D. A., L. Kennedy, et E. F. Churchill (2011), Peaks and persistence : modeling the shape of microblog conversations, in *CSCW*, pp. 355–358.
- Shapiro, S. S., et M. Wilk (1965), An analysis of variance test for normality (complete samples), *Biometrika*, 52(3 et 4), 591–611.
- Simon, H. (1971), Designing organizations for an information-rich world, *Computers, Communication, and the Public Interest*, pp. 37–72.
- SimplyMeasured (2014), How companies use twitter : Big brands vs. small brands.
- Singhal, A. (2001), Modern information retrieval : A brief overview, *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), 35–43.
- Snyman, J. A. (2005), *Practical Mathematical Optimization : An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*, Springer.
- Stattner, E., et M. Collard (2014), Matching of communities and frequent conceptual links, *Social Network Analysis and Mining*, 4(1).
- Taboada, G. L., S. Ramos, R. R. Expósito, J. Touriño, et R. Doallo (2013), Java in the high performance computing arena : Research, practice and experience, *Science of Computer Programming*, 78(5), 425–444.
- Techcrunch (2013), Mobile twitter : 164m+ access from handheld devices monthly.
- Valkanas, G., et D. Gunopulos (2013), How the live web feels about events, in *CIKM*, pp. 639–648.
- Verhulst, P.-F. (1845), Recherches mathématiques sur la loi d'accroissement de la population, *Nouveaux Mémoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, 18, 1–42.

## Bibliographie

---

- Wang, F., H. Wang, et K. Xu (2012), Diffusive logistic model towards predicting information diffusion in online social networks, in *ICDCS '12 Workshops*, pp. 133–139.
- Weng, J., et B.-S. Lee (2011), Event detection in twitter, in *ICWSM*, pp. 401–408.
- Wilcox, R. R., et J. Muska (2001), Inferences about correlations when there is heteroscedasticity, *British journal of mathematical and statistical psychology*, 54, 39–47.
- Wills, R. (2006), Google's pagerank, *The Mathematical Intelligencer*, 28(4), 6–11.
- Yang, J., et S. Counts (2010), Predicting the speed, scale, and range of information diffusion in twitter, in *ICWSM '10*, pp. 355–358.
- Yang, J., et J. Leskovec (2010), Modeling information diffusion in implicit networks, in *ICDM '10*, pp. 599–608.
- Yang, J., et J. Leskovec (2011), Patterns of temporal variation in online media, in *WSDM*, pp. 177–186.
- Yuheng, H., J. Ajita, D. S. Dorée, et W. Fei (2012), What were the tweets about? topical associations between public events and twitter feeds, in *ICWSM*, pp. 154–161.
- Zheng, L., et K. Han (2013), Multi topic distribution model for topic discovery in twitter, in *ICSC*, pp. 420–425.



# **Annexes**

## Liste des publications

### Revue internationale

1. A. Guille, H. Hacid, C. Favre and D. Zighed. Information Diffusion in Online Social Networks : A Survey.  
*ACM SIGMOD Record* – Volume 42, Number 2, pp. 17-28, 2013.

### Conférence internationale et atelier international

1. A. Guille et C. Favre. Mention-anomaly-based Event Detection and Tracking in Twitter.  
*ASONAM '14 – Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pp. 375-382, 2014.
2. A. Guille, C. Favre, H. Hacid et D. Zighed. SONDY : an Open Source Platform for Social Dynamics Mining and Analysis.  
*SIGMOD '13 – Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1005-1008, 2013.
3. A. Guille. Information Diffusion in Online Social Networks.  
*SIGMOD/PODS Ph.D. Symposium '13 – Proceedings of the SIGMOD/PODS Ph.D. symposium*, pp. 31-36, 2013.
4. A. Guille et H. Hacid. A Predictive Model for the Temporal Dynamics of Information Diffusion in Online Social Networks.  
*WWW '12 Companion – Proceedings of the International Conference Companion on World Wide Web : Workshop on Mining Social Network Dynamics*, pp. 1145-1152, 2012.

### Conférence nationale

1. A. Guille, C. Favre. Une méthode pour la détection de thématiques populaires sur Twitter.  
*EGC '14 – Actes de la 14ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances*, pp. 83-88, 2014.

## Bibliographie

---

2. A. Guille, C. Favre et D. Zighed. SONDY : une plateforme open-source d'analyse et de fouille pour les réseaux sociaux en ligne.  
*EGC '13 – Recueil des démonstrations de la 13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances*, pp. 45-48, 2013.
3. A. Guille, H. Hacid et C. Favre. Une approche multidimensionnelle basée sur les comportements individuels pour la prédiction de la diffusion de l'information sur Twitter.  
*EGC '12 – Actes de la 12ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances*, pp. 405-410, 2012.