# Modèle de réseaux multiplexes pour l'étude de l'influence sur Twitter

**Lobna Azaza, Éric Leclercq, Marinette Savonnet**

*Laboratoire d'Informatique de Bourgogne - EA 7534*
*Univ. Bourgogne Franche-Comté*
*9, Avenue Alain Savary*
*F-21078 Dijon - France*
*prénom.nom@u-bourgogne.fr*

RÉSUMÉ. *Les réseaux issus de données réelles tels que Twitter ou Facebook sont appelés réseaux complexes ou graphes de terrain. La plupart des travaux sur ces réseaux ont été abordés sous l'angle de la théorie ou de l'algorithmique des graphes mais très peu sous l'angle de la modélisation des données. Dans ce contexte, nous introduisons les notions fondamentales sur lesquelles est basée notre modélisation des réseaux sociaux en prenant comme exemple Twitter. Notre contribution consiste, dans un premier temps, à déterminer le modèle le plus adapté, c'est-à-dire un réseau multiplexe hétérogène, et à le spécialiser pour spécifier les relations engendrées par les interactions entre les utilisateurs. Dans un second temps, nous exploitons ce réseau multiplexe afin de mesurer l'influence des utilisateurs en incluant les différentes couches (relations). Pour ce faire nous utilisons une extension pour les réseaux multiplexes de l'algorithme PageRank. Notre contribution sur ce point consiste à représenter les données extraites de Twitter dans le modèle de réseau multiplexe que nous avons défini et à exploiter la richesse des relations traduisant l'influence et enfin d'étudier les paramètres qui modifient le comportement de l'algorithme.*

ABSTRACT. *Networks derived from real data such as Twitter and Facebook, etc. are called complex networks or real-world graphs. Most of the work on these networks has been approached from the perspective of graph theory or algorithmic, but very little from the perspective of data modeling. In this context, we introduce the fundamental notions on which our social network modeling is based, using Twitter as an example. Our contribution consists, first, in determining the most appropriate model, i.e., an heterogeneous multiplex network, and specializing it to specify the relationships generated by the interactions between users. In a second step, we operate this multiplex network to measure the influence of users by including the different layers. To do this, we use an extension for the multiplex networks of the PageRank algorithm. Our contribution on this point consists in representing the data extracted from Twitter in the multiplex network model we have defined and exploiting the richness of the relationships that reflect the influence and finally studying the parameters that modify the algorithm's behavior.*

MOTS-CLÉS : *Réseau multiplexe hétérogène, PageRank, Twitter, Influence*

KEYWORDS: *Multiplex Networks, PageRank, Twitter, Influence*

## 1. Introduction

Dans de nombreux contextes applicatifs (Internet, biologie, Réseaux sociaux), de grands graphes ne montrant pas de structure forte apparente comme des graphes réguliers, des arbres, sont appelés réseaux complexes ou graphes de terrain, par opposition aux graphes explicitement construits par un modèle ou une théorie.

La plupart des travaux sur les réseaux complexes ont été abordés sous l'angle de la théorie ou de l'algorithmique des graphes. En revanche, il existe très peu de travaux de modélisation de tels réseaux et les aspects liés à la sémantique des liens sont peu développés. Notre objectif est de proposer un modèle pour traduire la richesse des données afin de contextualiser les résultats produits par les algorithmes et ainsi aider leur interprétation. Nous prendrons comme exemple de réseau complexe des données réelles issues de *Twitter*. Si le principe de fonctionnement de *Twitter* est extrêmement simple, *tweeter* un texte court, l'utilisation par chacun peut être très variable. *Twitter* présente une grande hétérogénéité sémantique que les modèles de données et les algorithmes doivent prendre en compte. En effet, les liens sont créés par les opérateurs (retweet RT, mentionne @, hashtag #, etc.) qui, selon le domaine et les intentions de l'utilisateur, induisent des sémantiques différentes. L'utilisation de @utilisateur en début de *tweet* sert à interpeller ou répondre à un utilisateur alors que l'utilisation de @media en fin de *tweet* sert à accroître sa visibilité. Les relations n-aire sont aussi nécessaires pour décrire l'usage combiné des différents opérateurs, par exemple l'utilisation conjointe de RT et @ dans le même *tweet* ou pour détailler les objets contenus dans un *tweet*.

Dans la section 2, nous montrons que la modélisation classique des réseaux complexes sous forme de graphe simple ou de multi-graphe est insuffisante. Puis, en section 3, nous proposons et formalisons une modélisation sous la forme d'un réseau multiplexe hétérogène. Dans la section 4, nous détaillons l'algorithme PageRank puis une de ses extensions aux réseaux multiplexes. Dans la section 5, nous appliquons cet algorithme sur un jeu de données issu du projet pluridisciplinaire TEE'2014 – étude de la communication politique sur *Twitter* durant les élections européennes de 2014 – afin de déterminer l'influence des candidats, puis nous étudions les paramètres qui modifient le comportement de l'algorithme.

## 2. Modélisation des données des réseaux sociaux sous forme de graphe

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

Pour représenter les réseaux complexes tels que *Twitter* tout en ayant la possibilité de visualiser leurs différentes entités, plusieurs modèles de graphe ont été proposés allant du graphe simple à des formes plus évoluées.

### 2.1. *Multi-graphe*

Dans (Basaille *et al.*, 2016), nous avons proposé une modélisation sous forme d'un **multi-graphe orienté avec labels** implémentée dans Neo4j. La figure 1 présente cette implémentation où les entités (*tweets*, utilisateurs, hashtags, etc.) sont les nœuds et les relations sont décrites par les liens. Les liens représentés par un pointillé sont déduits, les liens représentés par un trait plein sont des liens directs.

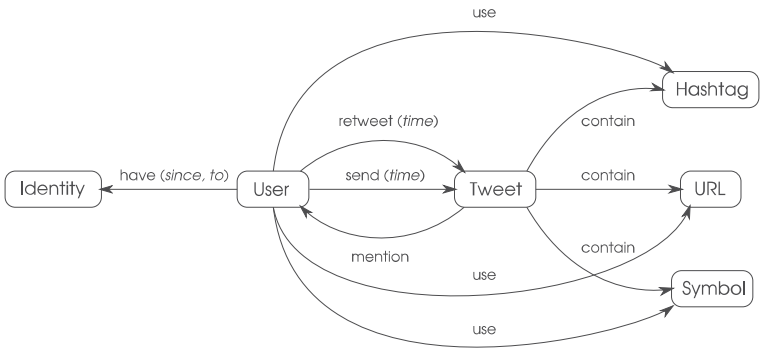


*Figure 1. Représentation multi-graphe orienté avec labels de Twitter*

La modélisation traditionnelle des réseaux complexes sous forme de graphe simple ou multi-graphe n'est pas suffisante. Par exemple, la relation *Mention* peut s'appliquer entre un *tweet* et un utilisateur mais aussi entre deux utilisateurs lorsque l'on agrège le nombre de mentions d'un utilisateur fait par l'utilisateur source ($Mention_U$) ou la relation $Mention_{ctx}$ qui agrège les relations *Écrire* et *Mention* (voir le tableau 1). Les relations n-aire sont aussi nécessaires pour décrire l'usage des différents opérateurs, par exemple l'utilisation conjointe de RT et @ dans le même *tweet* ou pour détailler les objets contenus dans un *tweet*, par exemple le *tweet* $t_{1234}$ peut contenir une URL et deux hashtags. Un nouveau cadre plus général est alors fourni par les hypergraphes.

### 2.2. *Hypergraphe*

Un hypergraphe $H$ est un couple $(V, E)$ où $V$ est un ensemble de nœuds et $E$ est un ensemble d'hyperliens. Chaque hyperlien $e \in E$ peut contenir arbitrairement plusieurs nœuds, $E$ est donc défini comme un sous-ensemble de $2^V$. Cette modélisation permet de combiner des informations sur les utilisateurs et le contenu.

*Twitter* peut être modélisé comme un hypergraphe orienté dont $V$ est l'ensemble des nœuds qui représentent les utilisateurs $V_U$, les *tweets* $V_T$ et les objets $V_O$ tels que les hashtags, les URLs et les symboles (emoji) [1] et un ensemble de relations $R$ entre ces nœuds. Ces deux ensembles sont définis comme suit : $V = \{V_U \cup V_T \cup V_O\}$ et

1. Les intersections des ensembles pris deux à deux sont vides, c'est-à-dire $V_U \cap V_T = \emptyset$ et $V_U \cap V_O = \emptyset$ et $V_T \cap V_O = \emptyset$.

$R = \{R_i, i = 1, \ldots, m\}$ avec $R_i : V^p \rightarrow \mathbb{N}$. Le tableau 1 présente quelques relations possibles dans *Twitter*.

*Tableau 1. Exemple de relations possibles dans Twitter*

| **Relation** | **Signature** |
|---|---|
| *Écrire* | *utilisateur* $\times$ *tweet* $\rightarrow \mathbb{B}$ |
| *Suivre* | *utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{B}$ |
| *Retweet*$_T$ | *tweet* $\times$ *tweet* $\rightarrow \mathbb{B}$ |
| *Retweet*$_U$ | *utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{N}$ |
| *Retweet*$_{ctx}$ | *utilisateur* $\times$ *tweet* $\times$ *utilisateur* $\times$ *tweet* $\rightarrow \mathbb{B}$ |
| *Mention* | *tweet* $\times$ *utilisateur* $\rightarrow \mathbb{B}$ |
| *Mention*$_U$ | *utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{N}$ |
| *Mention*$_{ctx}$ | *utilisateur* $\times$ *tweet* $\times$ (*utilisateur*)$^n \rightarrow \mathbb{B}$ |
| *Réponse* | *tweet* $\times$ *tweet* $\rightarrow \mathbb{B}$ |
| *Réponse*$_U$ | *utilisateur* $\times$ *utilisateur* $\rightarrow \mathbb{N}$ |
| *Réponse*$_{ctx}$ | *utilisateur* $\times$ *tweet* $\times$ *utilisateur* $\times$ *tweet* $\rightarrow \mathbb{B}$ |
| *Contenir* | *tweet* $\times$ *objet* $\rightarrow \mathbb{B}$ |
| *Contenir*$_U$ | *utilisateur* $\times$ *objet* $\rightarrow \mathbb{N}$ |
| *Contenir*$_{ctx}$ | *utilisateur* $\times$ *tweet* $\times$ (*objet*)$^n \rightarrow \mathbb{B}$ |

Prenons l'exemple d'un réseau *Twitter* comprenant les nœuds et relations décrites ci-dessous : $u_1$ à $u_4$ sont les utilisateurs, $t_1$à $t_5$ sont les tweets et $h_1$, $h_2$ représentent deux hashtags qui sont des objets particuliers.

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

Ce réseau est modélisé par l'hypergraphe donné en figure 2 avec :
$V = \{u_1, u_2, u_3, u_4, t_1, t_2, t_3, t_4, t_5, h_1, h_2\}$, $R = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7\}$

Cette modélisation ne permet pas de représenter l'orientation des hyperliens mais l'hypergraphe peut être représenté par sa matrice d'incidence $A_{ir}$ à coefficients entiers appartenant à l'ensemble $\{0, +1, -1\}$ telle que chaque colonne correspond à un hyperlien $r$, et chaque ligne à un nœud $i$ de l'hypergraphe $H$. Si un hyperlien $r$ arrive au nœud $i$, alors $A_{ir} = 1$, si un hyperlien $r$ sort du nœud $i$, alors $A_{ir} = -1$, $A_{ir} = 0$ sinon (voir la figure 3).

Bien que les hypergraphes offrent un moyen de modéliser *Twitter* selon plusieurs points de vue sur une même relation, cette modélisation reste limitée puisque les hyperliens n'ont pas de sémantique explicite et que certaines informations sont perdues. Par exemple, les mentions peuvent être utilisées plusieurs fois dans un même *tweet*,

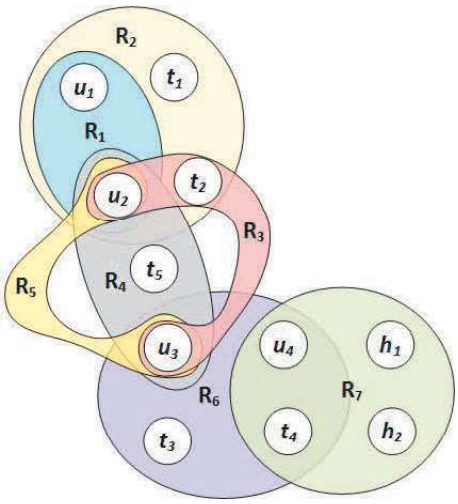


*Figure 2. Exemple d'une modélisation de Twitter sous la forme d'un hypergraphe*

ainsi l'arité [2] n'est pas fixée. Nous pouvons les modéliser par un ensemble de relations *Mention* ou $Mention_{ctx}$ mais l'ordre dans lequel les utilisateurs sont mentionnés est perdu. De plus, dans des réseaux complexes tels que *Twitter*, nous pouvons atteindre des milliers d'interactions, avec une telle modélisation, nous serons confrontés à plusieurs types d'hyperliens, ce qui peut compliquer la lisibilité des interactions. Par ailleurs, comparé au graphe simple, peu d'algorithmes ont été développés pour les hypergraphes même si dans (Estrada, Rodriguez-Velazquez, 2005) des mesures de centralité et le coefficient de clustering ont été étendus aux hypergraphes.

## 3. Modélisation des réseaux sociaux sous forme de réseaux multi-couches

Récemment, plusieurs travaux de recherche ont proposé d'améliorer et de généraliser les approches existantes pour prendre en compte des réseaux incluant plusieurs sous-systèmes et différents niveaux de connectivité entre les systèmes. Bien que la terminologie varie largement, en fonction de chaque cas spécifique, de tels réseaux peuvent généralement être appelés réseaux multi-couches.

### *3.1. Typologie des réseaux multi-couches*

Les notations utilisées pour des graphes simples sont étendues pour permettre de représenter les structures qui ont des couches en plus des nœuds et des liens. Dans

2. L'arité d'une fonction est le nombre d'arguments. Par exemple, dans un *tweet* on peut utiliser n mentions n n'étant pas connu.

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

*Figure 3. Matrice d'incidence de l'exemple donné en figure 2*

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

– $Y$ est l'ensemble des couches, $Y = \{\alpha, \alpha \in \{1, 2, \ldots, M\}\}$ et $M$ indique le nombre total de couches ($M = |Y|$) ;

– $\overrightarrow{G}$ une liste ordonnée de réseaux (graphes) caractérisant les interactions à l'intérieur de chaque couche $\alpha = 1, 2, \ldots, M$ soit $\overrightarrow{G} = (G_1, G_2, \ldots, G_\alpha, \ldots, G_M)$ pour lesquels $G_\alpha = (V_\alpha, E_\alpha)$ et $N_\alpha = |V_\alpha|$ ;

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

Un nœud peut exister dans plusieurs couches et il apparaît dans au moins une couche. Ainsi, le n-uplet nœud-couches $(i, \alpha_1, \cdots, \alpha_M)$ indique si nœud $i$ existe dans une couche $\alpha_j, j \in \{1, \ldots, M\}$.

Il existe plusieurs catégories de réseaux multi-couches, comme les réseaux multi-relationnels, hétérogènes, temporels, multiplexes. Les termes les désignant ont évolué au fur et à mesure de la publication des travaux de recherche. On peut remarquer que l'article de Kivelä (Kivelä *et al.*, 2014) permet une unification de la terminologie. Dans la suite de cette section, nous nous intéressons aux types réseaux multi-couches permettant de modéliser le réseau *Twitter*.

Dans un réseau multi-relationnel, les liens sont étiquetés avec un certain type, chaque type représente une relation spécifique entre les nœuds dans le réseau et se traduit en une couche. L'ensemble des liens $E$ est divisé en classes disjointes : $E = \bigcup_{r \in R} E_r$, où $R$ est l'ensemble de relations possibles. En conséquence, les couches sont implicites et dans chaque couche l'ensemble des nœuds est homogène, les liens également. Pour les réseaux sociaux tels que *Twitter* et afin de modéliser l'hétérogénéité des relations, (Dai *et al.*, 2012 ; Kivelä *et al.*, 2014) proposent une modélisation

sous la forme de réseau multi-relationnel. Cependant, cette modélisation n'est pas complètement satisfaisante car elle fixe une contrainte forte : l'ensemble des nœuds est homogène. Or, dans *Twitter*, les relations peuvent exister entre différents types de nœuds comme l'illustre le tableau 1.

Les réseaux hétérogènes ou réseaux multi-types possèdent des nœuds étiquetés avec un certain type qui peuvent être adjacents à des nœuds étiquetés avec le même type ou un type différent (Vazquez, 2006 ; Allard *et al.*, 2009). *Twitter* peut être vu comme un réseau hétérogène puisque les nœuds peuvent avoir plusieurs types comme utilisateurs, *tweets*, hashtags, etc. Mais cette modélisation reste insuffisante puisque l'hétérogénéité des liens n'est pas prise en compte.

Dans un réseau temporel, chaque couche correspond à un évènement, à un timestamp ou un intervalle de temps. Ainsi, les réseaux temporels permettent de représenter un ensemble d'évènements consécutifs comme une séquence ordonnée de réseaux. Dans le cas d'un ensemble d'évènements, les évènements peuvent être représentés comme des triplets $e = (i, j, t)$ où $i, j \in V$ sont des nœuds et $t \in T$ est le timestamp d'un évènement. *Twitter* peut être modélisé comme un réseau temporel puisque chaque *tweet* possède un timestamp.

Un réseau multiplexe est une spécialisation de réseau multi-couches qui impose au moins un nœud en commun entre les couches et qui permet que les nœuds n'existent pas dans toutes les couches (Kanawati, 2015) (voir la figure 4). (Bianconi, 2018) propose une modélisation issue d'une simplification de la structure des réseaux multi-couches généraux respectant les propriétés suivantes :

– les réseaux multiplexes sont des réseaux multi-couches pour lesquels il existe une correspondance un à un entre les nœuds des différentes couches (nommés nœuds réplicas). Les correspondances peuvent être omises si tous les nœuds sont présents dans toutes les couches ;

– si il existe des liens entre les différentes couches (*interlinks*), ils expriment uniquement des correspondances entre nœuds réplicas.

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

Dans la plupart des réseaux multiplexes, les nœuds des différentes couches sont en correspondance 1-1 pour indiquer qu'il s'agit de la même entité. Dans certains cas, il peut être utile de distinguer l'identité des nœuds dans les différentes couches. On a alors recours à des liens inter-couches qui peuvent porter un poids spécifiant, comme dans les réseaux de transports, le coût associé à un changement de couche. Pour cela, on adopte une notation complémentaire pour les $N$ nœuds et les $M$ couches. Chaque nœud $i$ peut avoir $M$ réplicas représentés par des couples $(i, \alpha)$ avec $i = 1, 2, \ldots, N$

et $\alpha = 1, 2, \ldots, M$, $V_\alpha = \{(i, \alpha), i \in \{1, 2, \ldots, N\}\}$. Un réseau multiplexe avec des liens inter-couches est alors modélisé par un triplet $\mathcal{M} = (Y, \overrightarrow{G}, \mathcal{G})$ avec $\overrightarrow{G} \neq \emptyset$ et $\mathcal{G} \neq \emptyset$. Chaque couche $\alpha$ est représentée par un graphe $G_\alpha = (V_\alpha, E_\alpha)$ qui lui-même peut être représenté par sa matrice d'adjacence, $A^{[\alpha]}$ de taille $N \times N$ incluant éventuellement des liens dirigés et/ou pondérés. Le réseau de couplage représentant les liens entre les couches $\mathcal{G}_{\alpha,\beta} = (V_\alpha, V_\beta, E_{\alpha,\beta})$ connecte les nœuds réplicas avec les liens définis par $E_{\alpha,\beta} = \{((i, \alpha), (i, \beta)), i \in \{1, 2, \ldots, N\}\}$ et représenté par une matrice d'adjacence $\mathscr{C}^{[\alpha,\beta]}$, nommée alors matrice de couplage de taille $N \times N$, où les éléments non diagonaux sont nuls. Afin d'exploiter un réseau multiplexe, celui-ci doit être dans un format que les algorithmes sont capables de traiter directement ou bien dans un format à partir duquel il est possible de dériver les structures de données nécessaires aux algorithmes. Dans le cadre général des réseaux multi-couches, la notion de matrice de supra-adjacence est souvent utilisée. Elle peut être obtenue à partir des différentes matrices d'adjacence intra-couche (composante $\overrightarrow{G}$) et des matrices de couplage inter-couches (composante $\mathcal{G}$). La matrice de supra-adjacence de taille $(N \times M) \times (N \times M)$ $\mathscr{S}$, est définie de la manière suivante :

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

avec $A^{[\alpha]}$ la matrice d'adjacence de la couche $\alpha$ et $\mathscr{C}^{[\alpha,\beta]}$ la matrice de couplage entre les couches $\alpha$ et $\beta$ (voir figure 5).

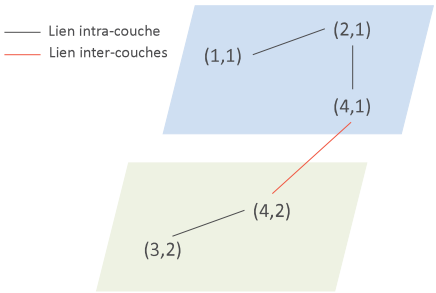


*Figure 4. Exemple de réseau multiplexe ayant 4 nœuds et 2 couches*

### *3.2. Modélisation de Twitter sous forme de réseau multiplexe hétérogène*

Un modèle adapté aux données de *Twitter* est le modèle de réseau multiplexe puisque les différentes couches peuvent être exploitées pour représenter les différents types de relations. Or, ces relations peuvent exister entre différents types de nœuds (voir le tableau 1). Comme le modèle de réseaux multiplexes ne contraint pas les nœuds à avoir le même type, ni à tous exister dans toutes les couches, ce modèle est le plus pertinent. Dans le cadre de nos expérimentations, nous ne nous intéressons pas à l'aspect temporel, en revanche, nous souhaitons présenter les différentes entités

$$
\begin{array}{c|cccc|cccc}
 & (1,1) & (2,1) & (3,1) & (4,1) & (1,2) & (2,2) & (3,2) & (4,2) \\
\hline
(1,1) & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
(2,1) & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
(3,1) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
(4,1) & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
\hline
(1,2) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
(2,2) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
(3,2) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
(4,2) & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0
\end{array}
$$

*Figure 5. Matrice de supra-adjacence de l'exemple donné en figure 4*

du réseau telles que les *tweets* et les utilisateurs par différents types de nœuds, par conséquent, nous proposons de modéliser *Twitter* sous forme d'un réseau multiplexe hétérogène.

Nous spécialisons le modèle de réseau multiplexe général proposé par Bianconi (2018) et nous proposons un modèle pour *Twitter* défini par $\mathcal{M} = (Y, \overrightarrow{G}, \mathcal{G})$ où :

– $Y$ est l'ensemble des couches correspondant aux différentes relations binaires du tableau 1 :

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

– $\overrightarrow{G}$ est la liste des graphes représentant les couches définies à partir de l'ensemble des nœuds $V = \{V_U \cup V_T \cup V_O\}$ pour lesquels $V_U$ est l'ensemble des utilisateurs, $V_T$ l'ensemble des tweets et $V_O$ l'ensemble des objets c'est-à-dire un hashtag, une URL ou un symbole (emoji par exemple). Les graphes de $\overrightarrow{G}$ sont :

- $G_{\acute{E}crire} = (V_U \cup V_T, E_{\acute{E}crire})$ est un graphe orienté avec $E_{\acute{E}crire} \subseteq V_U \times V_T$ ;

- $G_{Suivre} = (V_U, E_{Suivre})$ est un graphe orienté avec $E_{Suivre} \subseteq V_U \times V_U$ ;

- $G_{Retweet_T} = (V_T, E_{Retweet_T})$ est un graphe orienté qui représente les retweets avec $E_{Retweet_T} \subseteq V_T \times V_T$ ;

- $G_{Retweet_U} = (V_U, E_{Retweet_U})$ est un graphe orienté pondéré avec $E_{Retweet_U} \subseteq V_U \times V_U$, dans ce graphe les liens représentent le nombre de retweets d'un utilisateur donné effectués par un autre ;

- $G_{Mention} = (V_U \cup V_T, E_{Mention})$ est un graphe orienté qui représente les mentions des utilisateurs dans les *tweets*, $E_{Mention} \subseteq V_U \times V_T$ ;

- $G_{Mention_U} = (V_U, E_{Mention_U})$ est un graphe orienté pondéré avec $E_{Mention_U} \subseteq V_U \times V_U$, dans ce graphe, les liens représentent le nombre de mentions d'un utilisateur dans les *tweets* émis par un autre utilisateur ;

- $G_{R\acute{e}ponse} = (V_T, E_{R\acute{e}ponse})$ est un graphe orienté qui représente les *tweets* qui sont des réponses à d'autres *tweets*, $E_{R\acute{e}ponse} \subseteq V_T \times V_T$ ;

- $G_{Réponse_U} = (V_U, E_{Réponse_U})$ est un graphe orienté pondéré avec $E_{Réponse_U} \subseteq V_U \times V_U$, dans ce graphe, les liens représentent le nombre de réponses d'un utilisateur à un autre, tous *tweets* confondus ;

- $G_{Contenir} = (V_T \cup V_O, E_{Contenir})$ est un graphe orienté qui représente les objets contenus (hashtag, emoji, URL, images) dans les *tweets*, $E_{Contenir} \subseteq V_T \times V_O$ ;

- $G_{Contenir_U} = (V_U \cup V_O, E_{Contenir_U})$ est un graphe orienté pondéré avec $E_{Contenir_U} \subseteq V_U \times V_O$, dans ce graphe, les liens représentent le nombre de fois où un utilisateur a utilisé un objet, tous *tweets* confondus.

– $\mathcal{G}$ représente les liens inter-couches modélisés par les trois types de graphes suivants :

- correspondances d'identités entre utilisateurs : $\{\mathcal{G}_{\alpha,\beta}\}$ avec $\alpha$ et $\beta \in \{Écrire, Suivre, Retweet_U, Mention, Mention_U, Réponse_U, Contenir_U\}$ sont vingt-et-un graphes non orientés ($(7 \times 6)/2$) qui expriment les correspondances 1-1 entre utilisateurs entre des couches dont un des types de nœud inclut $V_U$, $E_{\alpha,\beta} \subseteq V_U \times V_U$ avec comme contrainte $|E_{\alpha,\beta}| \leqslant |V_U|$ ;

- correspondances d'identités entre tweets : $\{\mathcal{G}_{\alpha,\beta}\}$ avec $\alpha$ et $\beta \in \{Écrire, Retweet_T, Mention, Réponse, Contenir\}$ sont dix graphes non orientés qui traduisent les correspondances 1-1 entre *tweets* dans les couches qui incluent des nœuds de type $V_T$, $E_{\alpha,\beta} \subseteq V_T \times V_T$ avec comme contrainte $|E_{\alpha,\beta}| \leqslant |V_T|$ ;

- correspondances d'identités entre objets : $\{\mathcal{G}_{\alpha,\beta}\}$ avec $\alpha$ et $\beta \in \{Contenir, Contenir_U\}$ est un graphe non orienté qui traduit les correspondances 1-1 entre objets dans les deux couches qui incluent des nœuds de type $V_O$, $E_{\alpha,\beta} \subseteq V_O \times V_O$ avec comme contrainte $|E_{\alpha,\beta}| \leqslant |V_O|$.

## 4. Influence et algorithme PageRank étendu au réseau multiplexe

Dans (Leavitt *et al.*, 2009), l'influence dans le réseau social *Twitter* est définie comme la capacité d'un utilisateur à provoquer une action chez un autre utilisateur. Le terme « action » désigne les différentes relations possibles entre les nœuds (utilisateurs). Par conséquent, la mesure de l'influence dans *Twitter* est un problème complexe puisque *Twitter* offre plusieurs types de relations (*retweet*, *réponse*, *mention*, *suivre*), qui peuvent être combinés pour former différentes interactions.

Pour estimer l'influence d'un nœud, de nombreuses approches ont été proposées (Riquelme, González-Cantergiani, 2016 ; Al-Garadi *et al.*, 2018) : certaines basées sur un résumé statistique de *Twitter*, d'autres sur la topologie du réseau avec différentes mesures de centralités ou encore en classant les nœuds avec des algorithmes basés sur le PageRank et le HITS, ou encore sur la fusion d'information afin de tenir compte de la diversité des liens. Certaines mesures de centralité et l'algorithme de PageRank ont été transposés pour le modèle des réseaux multiplexes. Solé-Ribalta *et al.* (2014) proposent une mesure de centralité d'intermédiarité qui prend en compte la structure inhérente des réseaux multiplexes et proposent un algorithme pour la calculer de manière efficace. Spatocco *et al.* (2018) ont présenté une méthodologie géné-

rale de calcul de mesures de centralité de manière itérative, basée sur le théorème de Perron-Frobenius [3], permettant les classements sur une série de matrices représentant un réseau multiplexe. Dans la suite, nous présentons le PageRank et un algorithme de PageRank étendu au réseau multiplexe nommé Multiplex PageRank.

### *4.1. PageRank*

L'idée principale du PageRank est que « Les pages les plus importantes (des sites Web) sont susceptibles de recevoir plus de liens à partir d'autres pages » (Page *et al.*, 1999). PageRank suppose que l'importance d'une page Web est déterminée par la quantité et la qualité des pages qui s'y rattachent. Initialement, chaque nœud (c'est-à-dire, page) reçoit une valeur $PR$. Ensuite, chaque nœud distribue uniformément sa valeur $PR$ à ses voisins à travers les liens sortants. Mathématiquement, la valeur $PR$ du nœud $i$ à l'étape $t$ est :

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

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

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

### *4.2. Multiplex PageRank*

Des auteurs ont proposé une modification de l'algorithme PageRank pour l'adapter aux réseaux multiplexes. Halu *et al.* (2013) décrivent le pageRank multiplexe sur un réseau multiplexe ayant deux couches, Iacovacci et Bianconi (2016) introduisent en plus une fonction de similarité entre couches. Nous reprenons cet algorithme, et nous l'adaptons à notre formalisation du réseau multiplexe décrivant *Twitter*. Le principe

3. http://www.bibmath.net/dico/index.php?action=affiche&quoi=./p/perron-frobenius.html

de base consiste à choisir une couche principale $\alpha$ avec sa matrice d'adjacence $A^{[\alpha]}$ et à calculer le PageRank $x_i^{[\alpha]}$ , $i = 1, \ldots, N_\alpha$ pour chaque nœud présent dans cette couche par la formule habituelle soit :

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

où $d_\alpha$ est le facteur de téléportation de la couche $\alpha$, $g_j = max(1, \sum_{r=1}^{N_\alpha} A_{jr})$.

L'ensemble des valeurs constitué par chaque $x_i^{[\alpha]}$ est représenté par le vecteur $X^{[\alpha]}$ pour la couche $\alpha$. Ensuite, l'algorithme PageRank multiplexe mesure une forme de centralité pour tous les nœuds $(i, \beta)$ présents dans une autre couche $\beta$ en utilisant la matrice d'adjacence $A^{[\beta]}$ et prenant en compte les mesures effectuées sur la couche $\alpha$. Les mesures sur la couche $\beta$ sont influencées par le voisinage du nœud $(i, \beta)$ et par les nœuds réplicas dans la couche $\alpha$. On obtient ainsi une formule générale du PageRank :

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

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

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

Le second terme de l'équation (1) traduit la contribution de la centralité du nœud $(i, \alpha)$ au nœud $(i, \beta)$. En effet, même si un nœud de $\beta$ n'a pas la capacité à attirer des voisins importants, il peut quand même bénéficier de sa centralité acquise dans la couche $\alpha$. Une autre manière de formuler cette propriété est : l'importance d'un nœud dans une couche est affectée de manière bénéfique par l'importance qu'il possède dans une autre couche indépendamment de sa capacité à attirer d'autres nœuds. Pour ce faire on ajoute un facteur multiplicateur sur la partie de la formule liée au saut aléatoire qui est la valeur du PageRank de $(i, \alpha)$ dans la couche $\alpha$ divisée par la moyenne des valeurs des PageRank des nœuds de la couche $\alpha$ soit $\frac{x_i^{[\alpha]}}{\langle X^{[\alpha]} \rangle}$.

Ainsi, les effets de l'interaction entre les différentes couches du réseau sur la centralité de type PageRank sont directement pris en compte dans la marche aléatoire. En particulier, en fonction de l'intensité de l'interaction entre les couches, Iacovacci

et Bianconi (2016) définissent différentes versions du PageRank multiplexe : additif, multiplicatif et combiné.

$$x_i^{[\beta]} = d_\beta \sum_j^{N_\beta} (x_i^{[\alpha]})^b A_{ji}^{[\beta]} \frac{x_j^{[\beta]}}{G_j} + (1 - d_\beta) \frac{1}{N_\beta} \left( \frac{x_i^{[\alpha]}}{\langle X^{[\alpha]} \rangle} \right)^a \tag{2}$$

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

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

– PageRank multiplexe additif $(a = 1, b = 0)$, dans ce cas, $G_j = max(1, \sum_{r=1}^{N_\beta} A_{jr}^{[\beta]})$. Le premier terme pondère la centralité des nœuds de la couche $\beta$ en fonction de la centralité de leurs voisins et le deuxième terme permet à chaque nœud de la couche $\beta$ de tirer un avantage supplémentaire du fait qu'il est central dans la couche $\alpha$, quelle que soit la pertinence des nœuds qui le désignent dans la couche $\beta$.

– PageRank multiplexe multiplicatif $(a = 0, b = 1)$, l'effet de la couche $\alpha$ se produit sur le voisinage des nœuds de la couche $\beta$. Ici, tous les avantages d'être central dans la couche $\alpha$ dépendent des connexions que le nœud reçoit des nœuds centraux de la couche $\beta$.

– PageRank multiplexe combiné $(a = 1, b = 1)$ ajoute les deux effets précédents.

## 5. Application à l'étude de l'influence sur Twitter

Bianconi fournit une fonction Matlab pour calculer le PageRank multiplexe [4]. Nous avons utilisé cette fonction avec les données du projet TEE'2014 dont l'objectif global est d'observer et d'analyser la communication des politiques sur *Twitter* durant les élections européennes de mai 2014. Nous évaluons l'influence des candidats *via* leurs scores PageRank dans un réseau multiplexe où chacune des trois relations *Retweet*, *Mention* et *Réponse* représente une couche. Afin de construire les couches, nous considérons les candidats et les utilisateurs ayant interagi avec eux selon ces trois relations. Les couches sont représentées par leur matrice d'adjacence, le facteur de téléportation utilisé a été fixé à $0,85$. Le tableau 2 résume les données utilisées.

Dans le cadre du PageRank multiplexe, la couche *Retweet* est choisie pour être la couche principale puisqu'elle est plus significative en terme d'influence comme l'a indiqué Cha *et al.* (2010). C'est à partir de cette couche que les autres calculs sont établis dans les trois cas : PageRank multiplexe multiplicatif, additif et combiné. La figure 6

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

*Tableau 2. Paramètres des données du corpus TEE'2014*

| | | | |
|---|---|---|---|
| **Nombre de comptes** | 73 264 | **Nombre de candidats** | 616 |
| **Nombre de relations** | 614 191 | **Nombre de *retweets*** | 145 633 |
| **Nombre de *mentions*** | 434 207 | **Nombre de *réponses*** | 34 351 |

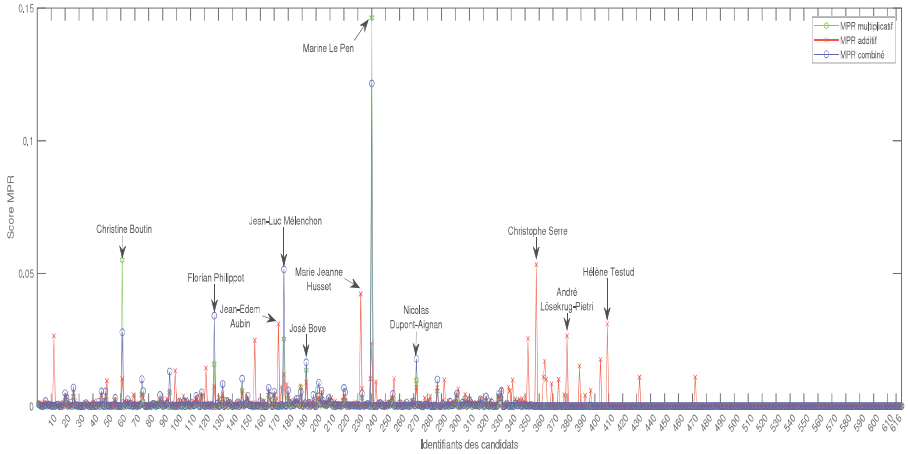


*Figure 6. PageRank multiplexe multiplicatif, additif et combiné des candidats français*

synthétise les scores PageRank pour les trois versions du PageRank multiplexe. À des fins de comparaison, nous avons aussi calculé le PageRank simple selon chacune des relations choisies (voir la figure 7). Nous observons que sur les dix premiers candidats six candidats sont en communs.

Les résultats obtenus ont été appréciés par nos collègues en Sciences Humaines et Sociales (SHS) participant au projet. Les PageRank multiplexe multiplicatif et combiné sont capables de distinguer correctement les personnes qui ont effectivement une influence importante dans le réseau *Twitter* et dans la vie réelle. Seuls les résultats du PageRank multiplexe additif sont très différents des autres résultats, faisant apparaître des "petits" candidats dont le cercle d'influence est plus restreint par rapport aux autres utilisateurs trouvés par les autres variations du PageRank multiplexe. Ces résultats mettant en avant de tels candidats sont certainement les plus intéressants pour nos collègues de SHS car ils présentent des singularités. Notre rôle d'informaticien n'est pas d'analyser les résulats mais de garantir leur qualité (nettoyage des données, utilisation correcte des algorithmes, etc.).

Cependant, bien que les PageRanks multiplexes permettent d'obtenir un bon classement d'utilisateurs, le score obtenu pour chaque utilisateur pris individuellement n'indique pas son influence et il est inexploitable si on ne le compare pas avec le score des autres utilisateurs.

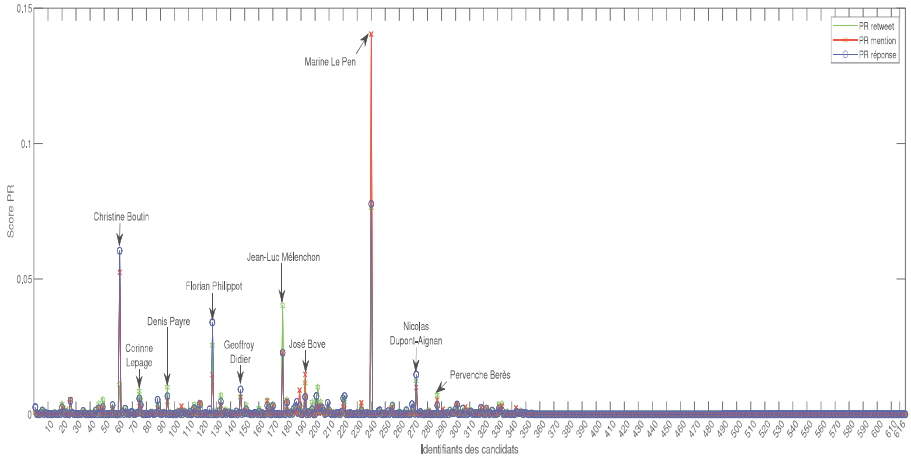


*Figure 7. PageRank des candidats français selon les relations Retweet, Mention et Réponse*

## 6. Conclusion

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

## Bibliographie

Al-Garadi M. A., Varathan K. D., Ravana S. D., Ahmed E., Mujtaba G., Khan M. U. S. *et al.* (2018, janvier). Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Comput. Surv.*, vol. 51, n° 1, p. 16:1–16:37. Consulté sur http://doi.acm.org/10.1145/3155897

Allard A., Noël P.-A., Dubé L. J., Pourbohloul B. (2009). Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics. *Physical Review E*, vol. 79, n° 3, p. 036113.

Basaille I., Kirgizov S., Leclercq E., Savonnet M., Cullot N. (2016). Towards a twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets. In *Tenth IEEE international conference on research challenges in information science, (RCIS)*, p. 1–10.

Bianconi G. (2018). *Multilayer networks: Structure and function*. Oxford University Press.

Cha M., Haddadi H., Benevenuto F., Gummadi P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, vol. 10, n° 30, p. 10-17.

Dai B. T., Chua F. C. T., Lim E.-P. (2012). Structural Analysis in Multi-Relational Social Networks. In *Proceedings of the siam international conference on data mining*, p. 451-462.

Estrada E., Rodriguez-Velazquez J. A. (2005, 06). Complex Networks as Hypergraphs.

Halu A., Mondragón R. J., Panzarasa P., Bianconi G. (2013). Multiplex pagerank. *PloS one*, vol. 8, n° 10, p. e78293.

Iacovacci J., Bianconi G. (2016). Extracting information from multiplex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 26, n° 6, p. 065306.

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

Kivelä M., Arenas A., Barthelemy M., Gleeson J. P., Moreno Y., Porter M. A. (2014). Multilayer networks. *Journal of complex networks*, vol. 2, n° 3, p. 203–271.

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

Page L., Brin S., Motwani R., Winograd T. (1999). The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th international world wide web conference*, p. 161–172.

Riquelme F., González-Cantergiani P. (2016). Measuring user influence on twitter: A survey. *Information Processing & Management*, vol. 52, n° 5, p. 949–975.

Solé-Ribalta A., De Domenico M., Gómez S., Arenas A. (2014). Centrality Rankings in Multiplex Networks. In *Proceedings of the 2014 acm conference on web science*, p. 149–155. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2615569.2615687

Spatocco C., D'Andrea A., Domeniconi C., Stilo G. (2018). A New Framework for Centrality Measures in Multiplex Networks. *CoRR*, vol. abs/1801.08026. Consulté sur http://arxiv.org/abs/1801.08026

Vazquez A. (2006). Spreading dynamics on heterogeneous populations: multitype network approach. *Physical Review E*, vol. 74, n° 6, p. 066114.