Tweet, Retweet et Follower : que recommander et à qui ?

Quentin Grossetti, Cédric Du Mouza, Nicolas Travers

Tweet, Retweet et Follower : que recommander et à qui ?

Quentin Grossetti[1,2], Cédric du Mouza[2] et Nicolas Travers[2]

***Abstract*— Twitter véhicule une quantité incroyable de messages - les *tweets* - sur un réseau gigantesque d'utilisateurs. Appréhender la dimension de ce réseau et le comportement des utilisateurs pour en extraire du sens est une tâche complexe, et les métriques sont nombreuses. Cet article propose l'étude d'un jeu réel représentatif de Twitter composé d'un réseau entre utilisateurs et de leurs messages. Cette étude se focalise sur les liens entre les utilisateurs, le comportement des retweets et la similarité entre les utilisateurs. Les résultats amènent à proposer une approche de recommandations pour les réseaux sociaux de micro-blogging.**

I. Introduction

Depuis plus de dix ans, les réseaux sociaux - et plus spécifiquement les plateformes de microblogging - ont connu une émergence extraordinaire avec l'arrivée du Web 2.0. Les plateformes comme `Twitter`, `Pinterest`, `Instagram`, `Weibo` ou `Tumblr` ont chacune un public ciblé ainsi que des fonctionnalités uniques, cependant elles ont en commun un cœur de fonctionnement qui est aujourd'hui bien établi dans la culture internet. On distingue ainsi la possibilité de suivre un utilisateur sans que cette relation soit réciproque, la possibilité de publier du contenu et enfin la possibilité de partager le contenu d'un autre utilisateur. Depuis leur création, le nombre d'utilisateurs ne cesse de croître de façon exponentielle et la quantité d'informations qui transitent sur ces réseaux est énorme, `Twitter` culminant par exemple à plus de 500 millions de messages publiés par jour en 2017.

Trouver une information intéressante devient un enjeu considérable, ce pourquoi `Twitter` a intégré à sa plateforme un système de recommandation de façon à mettre en avant de manière personnalisée des messages publiés [13]. On dénombre plusieurs leviers pour façonner les systèmes de recommandation, on peut par exemple décrire les objets à recommander afin de les proposer aux gens qui aiment ce type de contenu [9]. Cette approche - content-based - est hélas limitée sur les plateformes de microblogging, du fait de la difficulté à décrire efficacement des messages aussi courts (140 caractères pour `Twitter`, une image pour instagram). Un second levier, le filtrage collaboratif, consiste à capturer les similarités entre utilisateurs avec l'intuition que des utilisateurs qui ont aimé les mêmes choses dans le passé aimeront les mêmes choses dans le futur. Cette approche est actuellement la plus répandue car elle fournit des recommandations très pertinentes [12]. Parmi les modèles de filtrage collaboratif, les décompositions matricielles sont aujourd'hui légions dans la communauté [6] ou a base de confiance sociale [10]. Cependant ces techniques, qui réussissent grâce à une série d'optimisations [11], donnant un très bon compromis entre pertinence des recommandations et temps de calcul, sont difficiles à appliquer à l'échelle de `Twitter` par exemple. En effet, la taille de la matrice des intéractions entre utilisateurs et messages est gigantesque (environ 1 000 milliards de messages publiés depuis la naissance de `Twitter`) et augmente chaque jour ce qui rend délicat la tâche de calculer les scores de recommandation pour des messages récents.

Afin d'explorer une approche différente de l'exploitation de ce graphe d'utilisateurs et des tweets qui transitent sur le réseau, nous proposons une étude approfondie d'un jeu de données sur `Twitter`. En effet, au vu du problème d'échelle, que ce soit en quantité d'information ou rapidité de traitement de celles-ci, il est nécessaire d'extraire les informations clés permettant de comprendre ce qui a une forte influence sur le réseau et la vie d'un tweet sur le réseau.

Cet article présente un jeu de données provenant de `Twitter`, comprenant aussi bien un large réseau d'utilisateurs comprenant *followers* et *followee*, ainsi que les tweets et retweets associés à ces mêmes utilisateurs. Afin de définir la recommandation d'un tweet, il faut comprendre les liens entre les utilisateurs, les chemins dans le réseau et la distance entre ces utilisateurs, leur similarité ou dissimilarité (aussi bien en followers/followees que tweets/retweets), mais également s'intéresser au comportement d'un tweet sur ce même réseau avec sa durée de vie. Il faut également s'intéresser à l'homophilie dans ce réseau, le fait que les utilisateurs liés entre eux ont un comportement similaire au niveau de leur diffusion d'information (retweet). Cette étude nous permettra d'amener une discussion sur la conception d'une système de recommandation adapté au contexte de `Twitter`.

L'article s'organise de la manière suivante, tout d'abord nous présenterons la conception de notre jeu de données et sa composition brute dans la section II. Puis, nous présenterons le comportement des tweets/retweets dans la section III, ce qui nous permettra d'étudier dans la section IV le comportement des utilisateurs dans le réseau `Twitter`. Une discussion sur les résultats obtenus et une proposition de solution est présentée en section V, et une conclusion en section VI.

II. Jeu de données sur Twitter

Nous présentons ici les caractéristiques principales de notre jeu de données issu de `Twitter`, sur lequel reposent les différentes analyses de nos travaux, qui permettent de

*This work was not supported by any organization

[1]`quentin.grossetti@upmc.fr`
[2]Laboratoire CEDRIC, Cnam - `prenom.nom@cnam.fr`

Nombre total de nœuds	2 182 867
Nombre total d'arcs	325 451 980
Nombre total de tweet	3 001 502 711
Nombre moyen de followers	69,4
Nombre moyen de followees	57,8
Nombre maximum de followers	185 401
Nombre maximum de followees	348 595
Diamètre du réseau	15
Taille moyenne du plus court chemin	3,7
Liens reciproques	49 819 171

TABLE I

CARACTÉRISTIQUES PRINCIPALES DU JEU TWITTER

mieux appréhender les différents mécanismes d'une plateforme de microblogging.

Ce jeu a été constitué en plusieurs étapes. Dans un premier temps nous avons extrait une composante connexe du réseau de Twitter rendu disponible en 2009 par Kwak and al. [7]. Pour chacun des nœuds issus de cette composante, nous avons collecté les arcs entrants, les arcs sortants ainsi que les tweets associés aux comptes utilisateurs afin d'être à jour car le réseau social réel a changé entre 2009 et aujourd'hui.

Pour ce faire, nous avons utilisé l'API[1] de Twitter pour récupérer les données des followers et des followee de chacun des nœuds de ce réseau. Du fait des limites de l'API de Twitter sur le nombre de tweets récupérés pour chaque utilisateur, nous ne pouvions récupérer que les 3 200 derniers tweets associés à chaque compte. De fait, les comptes sont bornés à cette limite, quelle que soit la fréquence de publication de ceux-ci.

Les caractéristiques de notre jeu de données sont présentées dans le tableau I. Nous allons présenter par la suite les résultats de notre analyse qui peuvent avoir un impact sur la définition d'un modèle de recommandation.

Avec 2 millions d'utilisateurs et 3 milliards de tweets, nous avons une moyenne de 1 375 tweets par utilisateur. La composante connexe de notre réseau représente 325 millions d'arcs, avec en moyenne 69 arcs entrants par utilisateur, le maximum étant de 185. Le diamètre de ce réseau s'étend jusqu'à une distance de 15 entre deux utilisateurs avec une moyenne de parcours de 3,7. Par ailleurs, sur les 325 millions de liens unidirectionnels (followers ou followee), près de 50 millions d'entre-eux sont des liens réciproques, ce qui veut dire que seulement 15% des utilisateurs se suivent mutuellement.

Nous allons commencer notre analyse par une étude sur les tweets et retweets effectués par les utilisateurs dans ce réseau, puis nous nous intéresserons à la topologie du réseau Twitter et la caractérisation de l'homophilie de ses utilisateurs.

III. COMPORTEMENT DES RETWEETS

Afin de mieux comprendre les effets des tweets dans l'environnement Twitter, nous allons nous focaliser sur les retweets qui témoignent explicitement de l'intérêt d'un utilisateur pour un message. Comprendre les intérêts et les engagements des utilisateurs est une étape clé pour construire efficacement un système de recommandation.

[1]https://dev.twitter.com/rest/public

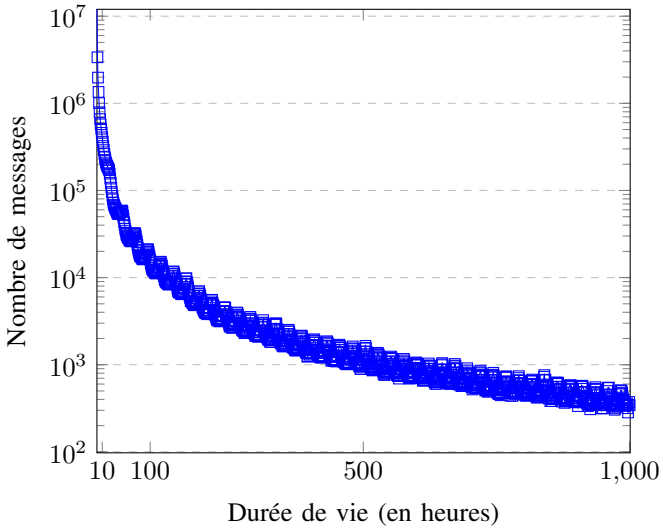

Fig. 1. Durée de vie d'un message

A. Durée de vie d'un tweet

Un message peut être partagé par n'importe quel utilisateur de façon à intégrer celui-ci à ses statuts et à le mettre en avant pour ses followers. Dans ce contexte, on peut considérer que le temps écoulé entre la date de publication d'un message (tweet de t_1) et la dernière fois que celui-ci a été partagé (dernier retweet de t_1) peut être perçu comme sa durée de vie. De manière assez évidente seuls les messages ayant été partagés au moins une fois sont considérés dans cette analyse (*i.e.,* durée de vie > 0). Les résultats sont visibles dans la Figure 1.

On peut remarquer que la grande majorité des messages ont une durée de vie inférieure à une heure (40%), pour 90% elle est inférieure à 3 jours, et il est extrêmement rare pour un message d'être partagé au-delà de cette limite.

Nos résultats montrent une diminution de la durée de vie des messages comparés aux résultats de Kwak [7]. En 2009, cette étude montrait que des messages pouvaient être partagés jusqu'à un mois après la parution d'un message. Notre hypothèse est qu'en 2009, beaucoup moins de contenus étaient publiés sur Twitter et qu'il était donc encore possible d'accéder facilement à des messages anciens. De plus, les utilisateurs sont plus à l'aise avec l'utilisation de Twitter qu'en 2009, les comportements des utilisateurs sont davantage normalisés.

Ce qui nous paraît particulièrement intéressant, et d'autant plus lors de la conception d'un système de recommandation par exemple, c'est le caractère prépondérant de la fraîcheur d'un message pour son partage. En d'autres termes, un message perd de l'intérêt de manière exponentielle au fil du temps. Il devient alors critique de pouvoir recommander des tweets en temps réel en se fondant sur une fenêtre très courte des activités du réseau.

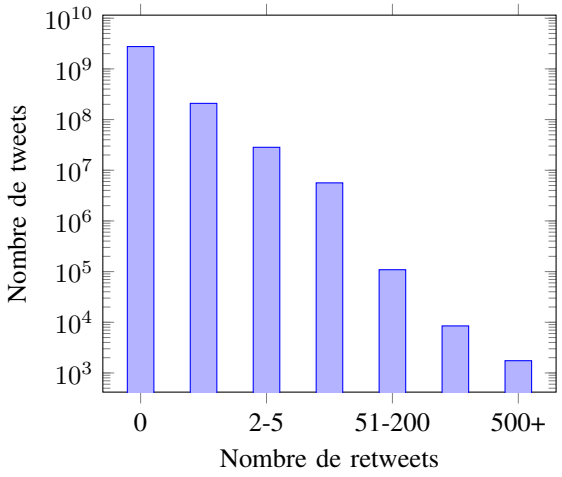

Fig. 2. Répartition du nombre de retweets par tweet

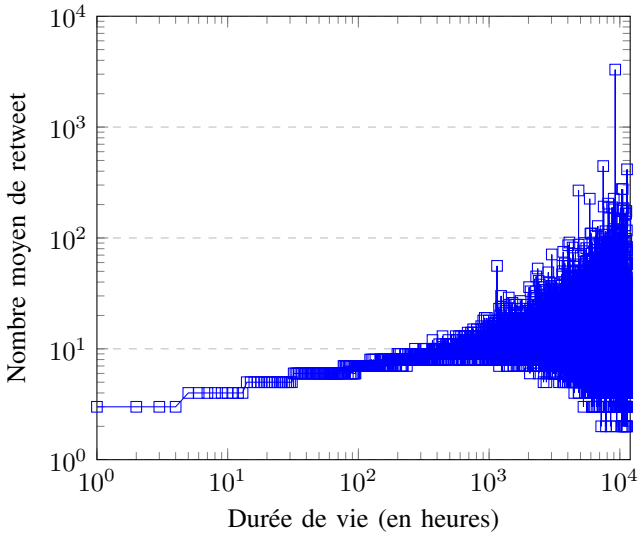

Fig. 3. Corrélation entre popularité et durée de vie

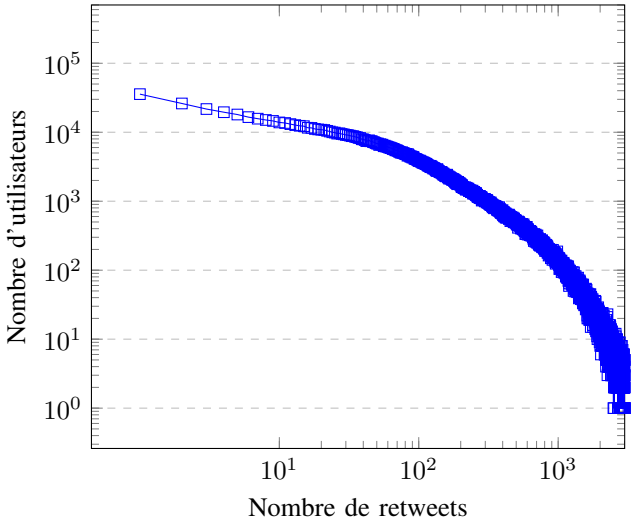

Fig. 4. Répartition des retweets par utilisateurs

B. *Popularité des tweets*

Dans cette partie, nous nous concentrons sur la répartition de la popularité des tweets avec le nombre de retweets par message. Les résultats sont représentés dans la Figure 2. La grande majorité des messages ne sont jamais retweetés ($\approx$ 90%), voire très peu, avec à peine 2-3 retweets ($\approx$ 2%). Ces résultats sont très cohérents par rapport à l'étude de Kwak [7] qui trouvait une très grande partie de messages jamais retweetés ou retweetés une seule fois.

De fait, dès qu'un tweet devient important (*i.e.,* plus de 2-3 retweets), un système de recommandation doit accorder un poids non négligeable à ce type de message. Toutefois, le réseau en lui-même favorise la propagation de ce type de message (effet réseau social). À contrario, être capable de recommander un tweet récent avec peu de retweets (moins de 5) est un réel défi pour rendre des messages populaires, correspondant au seuil critique de lancement d'un message sur le réseau.

C. *Popularité d'un message et durée de vie*

Afin de déterminer si un message populaire est toujours synonyme d'une durée de vie plus longue, nous avons représenté la popularité moyenne des messages en fonction de leur durée de vie dans la Figure 3. Les résultats sont conformes à ce que l'on pouvait attendre, plus un message est partagé, plus celui-ci se répand dans le réseau, ce qui allonge sa durée de vie. Cependant, on constate qu'au-delà de la limite de 1 000 heures, la corrélation s'effondre, ce qui signifie que des vieux messages non populaires peuvent refaire surface bien longtemps après leur parution.

D. *Retweets utilisateurs*

Pour terminer sur le comportement des retweets, nous allons maintenant nous intéresser à leur répartition parmi les utilisateurs. Les résultats sont présentés dans la Figure 4, nous pouvons constater sans grande surprise que très peu d'utilisateurs font un grand nombre de retweets. La grande majorité des utilisateurs effectuent entre 10 et 100 retweets. Du point de vue du système de recommandation, le problème se pose pour les utilisateurs ayant très peu de retweets, voire zéro, qui représentent une part importante des utilisateurs (un utilisateur sur 4 n'a jamais retweeté). Les méthodes qui se basent sur du filtrage collaboratif peuvent difficilement fournir des recommandations pour ces utilisateurs du fait de leur inactivité.

E. *Conclusion*

Nos résultats montrent que les utilisateurs sont très friands de messages récents et que peu de messages sont partagés tous les jours. Au vu du nombre d'utilisateurs et du nombre de messages, il n'est pas raisonnable de prendre en considération l'ensemble des données.

Il faut donc réduire l'espace pour pouvoir en extraire du sens et cibler les recommandations. À partir de ce constat, une piste envisageable de réduction de cet espace est de mettre l'accent sur des messages plutôt récents et partagés

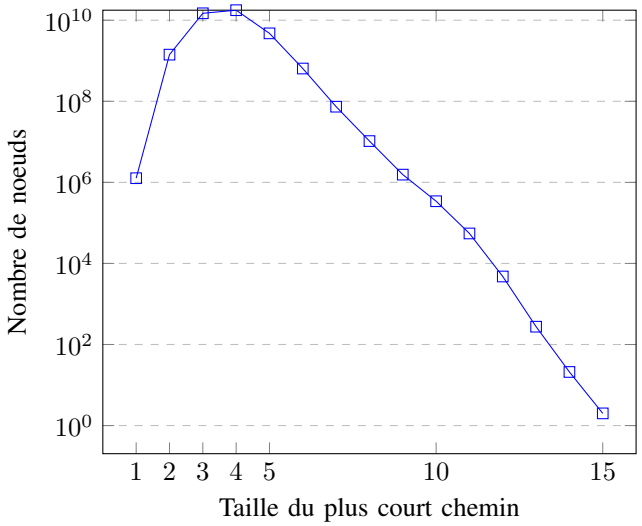

Fig. 5. Distribution des plus courts chemins dans Twitter

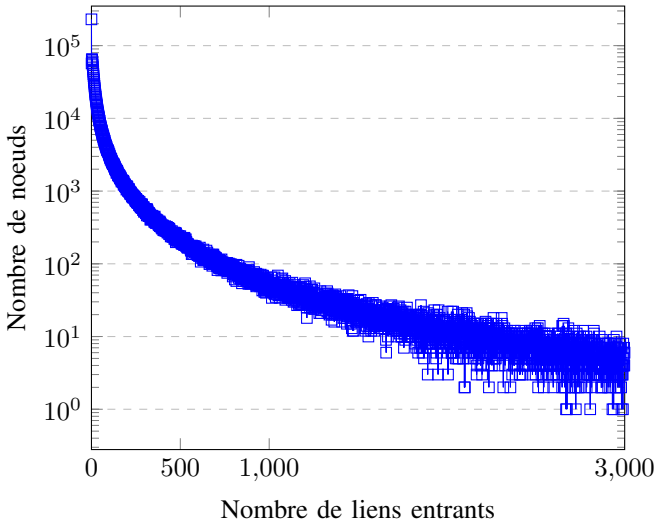

Fig. 6. Répartition des liens entrants

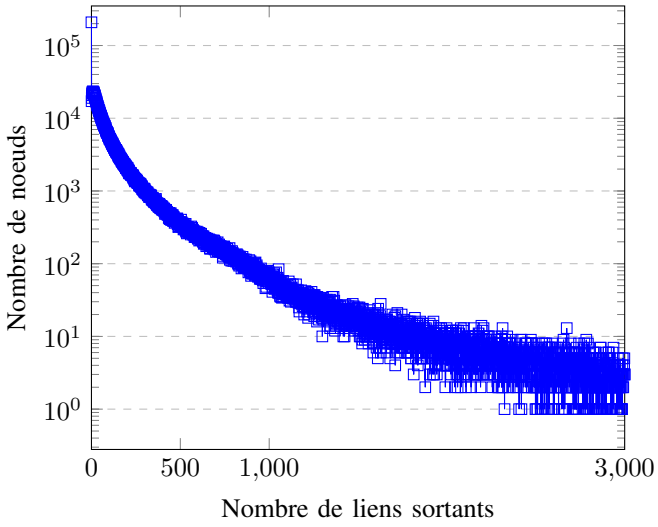

Fig. 7. Répartition des liens sortants

(retweet) au moins une fois. Cela permettrait ainsi de faire un premier tri lors d'un processus de recommandation. De plus, la qualité du système de recommandation doit également prendre en compte le fait de proposer une information à l'utilisateur *avant* que celui-ci ne la retweet lui-même. Ainsi, il faut pouvoir faire "sauter" le message dans le réseau, c'est à dire ne pas se concentrer sur le voisinage direct. Mais pour cela, il nous faut étudier le comportement des retweets sur ce réseau `Twitter`.

IV. Analyse du réseau Twitter

Après avoir étudié le comportement des tweets et des retweets, nous allons maintenant nous intéresser à la topologie du réseau d'utilisateurs, puis aux similarités entre ces utilisateurs afin de pouvoir mieux appréhender l'influence d'un tweet sur ce réseau.

A. Topologie générale

Pour commencer, nous allons préciser la distance qui sépare deux utilisateurs dans le réseau. Le plus court chemin qui sépare en moyenne un utilisateur de n'importe quel autre utilisateur est très proche des résultats présentés par Kwak and al. [7]. Dans son cas, ils trouvaient une distance moyenne de 4,12 là où nous trouvons 3,7 (Figure 5). L'écart entre ces deux valeurs étant très faible, nous pensons que c'est un bon point pour témoigner de la représentativité de notre jeu de données.

De plus, le diamètre, c'est à dire le chemin le plus long qui sépare deux utilisateurs, est de 15 (contre 18 dans l'article de Kwak). Les distributions des nombres de followers (Figure 6) et de followee (Figure 7) sont classiques et très proches de celles mises en avant dans d'autres travaux comme Weng and al. [14] ou Lerman and al. [8]. De fait, la plupart des utilisateurs sont reliés à moins de 500 autres nœuds avec une orientation dominante vers les liens sortants (followee).

B. Influence des Top-N

Afin de mieux comprendre l'impact des retweets sur le réseau, nous allons nous intéresser à la similarité entre les utilisateurs. En effet, si nous pouvons extraire des corrélations entre utilisateurs à travers les retweets qu'ils ont en commun, nous pouvons ainsi extraire des propriétés intéressantes du réseau. Pour ce faire, nous allons utiliser un facteur clé des systèmes à filtrage collaboratif : la liste des *Top-N*. Cette liste se base sur la similarité entre les utilisateurs et ne conserve pour chaque utilisateur que les N scores les plus élevés. Cela permet en définitive de filtrer sur les comportements les plus similaires dans le réseau. Toutefois, la question est de savoir si ces utilisateurs similaires sont proches ou non dans le réseau.

Dans notre jeu de données, nous avons extrait un sous-ensemble d'utilisateurs et calculé la liste des *top-N* scores de similarités avec tous les utilisateurs. Pour ce faire, nous avons utilisé une distance Jaccard, adaptée pour y intégrer la

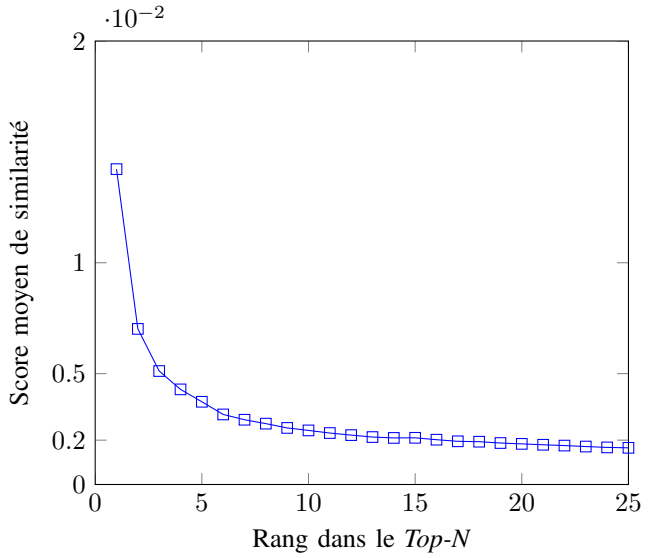

Fig. 8. Score de similarité moyenne en fonction du rang dans le *Top-N*

popularité des tweets, comme proposé par Breese et al. [2]. Cet ajustement permet de donner un poids plus fort à un retweet commun s'il est moins populaire, avec l'idée que deux personnes sont d'autant plus proches que leurs goûts communs sont rares. La distance est ainsi définie par :

$$sim(u,v) = \frac{\sum_{i \in L_u \cap L_v} \frac{1}{log(1+m(i))}}{|L_u \cup L_v|} \quad (1)$$

La fonction $sim(u,v)$ définit le score de similarité entre les utilisateurs u et v. L_u est le profil de l'utilisateur u correspondant, c'est à dire l'ensemble des messages sur lesquels il a fait un retweet (respectivement pour L_v). L'impact de la popularité d'un tweet $m(i)$ (nombre de retweet pour le message i) est compensé par la formule : $\frac{1}{log(1+m(i))}$.

La figure 8 montre l'évolution des score moyen de similarité au fûr et à mesure du classement dans le *Top-N*. Les scores de similarités sont très faibles, du fait de notre pondération par la popularité des messages et de la faible probabilité que deux utilisateurs puissent aimer les mêmes messages dans l'ensemble des tweets. Ces scores, mêmes faibles restent cependant significatifs. Ainsi, nous pouvons constater que le *top-5* reflète les utilisateurs les plus similaires à l'utilisateur u considéré. Ces 5 utilisateurs sont donc les utilisateurs qui produisent les informations qui potentiellement peuvent intéresser le plus l'utilisateur u.

C. Homophilie

Le phénomène d'`homophilie` appliqué aux réseaux sociaux est souvent décrit comme la tendance des utilisateurs à se connecter ensemble lorsqu'ils ont des traits en commun. Ce phénomène a déjà été largement étudié concernant les dimensions démographiques (Age, Sexe, Orientation politique) par exemple par Colleoni [3] ou Zamal [15]. Cette dernière étude montre qu'il est possible de prédire les informations démographiques d'un utilisateur à partir de son voisinage dans le réseau.

Distance	Nb d'utilisateurs	%	Similarité moyenne
1	3 229	2,652	0,0085
2	32 668	26,86	0,0014
3	81 645	67,132	0,0009
4	3 820	3,14	0,0010
5	43	0,03	0,0014
6	1	09	0,0008
Impossible	216	0,18	0,0017

TABLE II
RÉPARTITION DES SCORES DE SIMILARITÉ SELON LA DISTANCE AVEC L'UTILISATEUR

Des études sur l'homophilie basée sur des topics ont également été réalisées par Lerman [5] et Bhattacharya et al. [1]. Elles étudient à quel point les utilisateurs ayant des centres d'intérêts (topics) similaires ont tendance à se regrouper.

Nous cherchons ici à mettre en lumière ce phénomène au travers des méthodes de filtrage collaboratif. Autrement dit, nous examinons le lien entre score de similarité et distance dans le réseau.

Pour des raisons de complexité et de passage à l'échelle, nous avons focalisé notre étude sur un sous ensemble de 500 utilisateurs choisis aléatoirement dans le jeu de données. Pour ces utilisateurs, nous avons calculé la distance qui les sépare de chacun des utilisateurs ayant un score de similarité non nul. La distance est mesurée en conservant l'aspect orienté du réseau initial de `Twitter`. Remarquons que pour chaque utilisateur, il est possible de trouver en moyenne 240 utilisateurs ayant un score de similarité non nul avec lui.

Le tableau II présente les résultats obtenus. Nous pouvons constater que seulement 2% des paires d'utilisateurs ayant un score de similarité non nul sont directement connectés dans le réseau. Cependant, ces paires représentent un fort score de similarité par rapport aux autres. Si on s'éloigne un peu, à la distance 2, c'est à dire les followee des followee, on constate un score plus élevé que la moyenne. La majorité (68%) des paires entre utilisateurs similaires se situe à degrée 3, ce qui étant donné la topologie de notre réseau donnée en Figure 5, correspond à presque l'ensemble des utilisateurs. On peut donc faire une distinction entre d'un côté une *homophilie forte* qui décrit la tendance des utilisateurs à se connecter directement et à avoir des scores de similarité assez élevés et une *homophilie faible* qui décrit des utilisateurs ayant un score de similarité assez bon à distance 2. Globalement, la similarité moyenne entre utilisateurs décroit progressivement lorsque l'on s'éloigne dans le réseau. Ce qui nous permet de conclure que si les utilisateurs étant à distance 1 restent les plus similaires, ils sont très peu nombreux. Il apparaît alors pertinent de prendre également en compte les utilisateurs qui se trouvent à distance 2.

Le tableau III se focalise cette fois sur le rang dans le *Top-N* et la distance moyenne associée. Pour chacun des 500 utilisateurs de l'expérience précédente, nous avons conservé les 5 utilisateurs les plus proches afin de se focaliser sur la répartition des distances pour ces utilisateurs. Ainsi, le top-1 des utilisateurs les plus similaires n'est alors que dans 56%

		Réparatition des distances (en %)			
Rang	Distance Moyenne	1	2	3	4
1	1,55	57,03	31,53	10,64	0,8
2	1,68	49,60	33,13	16,87	0,4
3	1,8	42,45	36,02	20,72	0,8
4	1,86	38,71	38,71	20,56	2,02
5	1,98	31,44	40,16	27,59	0,81

TABLE III

DISTANCE DANS LE RÉSEAU EN FONCTION DU RANG DANS LE *Top-N*

des cas à une distance de 1. D'ailleurs, quelque soit le rang dans le Top-N, l'utilisateur similaire est au mieux à 50% dans son voisinage direct. Les résultats de cette analyse tendent à faire penser que les systèmes de recommandation basés sur le filtrage collaboratif doivent se focaliser en grande partie sur le voisinage indirect (distance 2, voire 3). Toutefois, il sera alors nécessaire de bien choisir ce voisinage indirect, car la taille de ce réseau explose à une distance de 3.

V. DISCUSSION

Au vu des différents résultats obtenus dans notre étude du jeu de données sur `Twitter`, nous pouvons voir qu'il est nécessaire de prendre en compte i) la fraicheur du message (quelques heures), ii) sa popularité (nombre de retweets) tout en gardant à l'esprit son partage iii) par des utilisateurs n'ayant pas forcément une énorme activité de retweets (quelques dizaines minimum), iv) sur un réseau où les utilisateurs ont tendance à suivre davantage de comptes qu'à être eux même actifs, v) avec des comportements de moins en moins similaires au fûr et à mesure de l'éloignement dans le réseau.

Une idée serait alors de proposer un méta-graphe d'utilisateurs qui se focaliserait sur les similarités entre les utilisateurs, comme présenté précédemment avec le*Top-N*. Ce méta-graphe serait alors à même de rapprocher les utilisateurs par comportements similaires, pour pouvoir faire propager des retweets commençant à être populaires plus rapidement. Ce méta-graphe pourrait également intégrer une pondération des arcs en intégrant le score de similarité, ainsi que la popularité des retweets dans ce nouvel entourage proche. La finalité serait de calculer une probabilité de recommandation influencée à la fois par la similarité entre les utilisateurs et par la popularité du message.

Le fait de proposer un méta-graphe permet également de réduire le maillage de celui-ci (*Top-N* par nœud) et ainsi de focaliser la propagation sur un sous-ensemble significatif plutôt que la quasi totalité du réseau.

VI. CONCLUSION

La conception d'un système de recommandation repose essentiellement sur notre bonne compréhension du comportement des utilisateurs. Nous avons, tout au long de cet article, détaillés les phénomènes clés qui devraient permettre de concevoir efficacement un système de recommandation pour Twitter, et plus largement pour n'importe quelle plateforme de microblogging. Nous avons donc proposé différentes analyses du comportement des utilisateurs et décrit la manière dont des utilisateurs similiaires ont tendances à se grouper. Nous pensons que les systèmes de filtrage collaboratifs ont encore de beaux jours devant eux et que nos résultats devraient aider à les adapter aux plateformes de microblogging. Nous prévoyons de concevoir un système reposant sur les principaux résultats de cette étude, qui s'appuierait sur un méta-graphe intégrant la similarité entre les utilisateurs. Il devrait permettre d'aller plus loin dans l'utilisation des intéractions entre utilisateurs que le système de recommandation par marche aléatoire actuellement en cours chez `Twitter` [4].

REFERENCES

[1] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi. Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 357–360, New York, NY, USA, 2014. ACM.

[2] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 43–52, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[3] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[4] Z. Huang, H. Chen, and D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):116–142, Jan. 2004.

[5] J. hyung Kang and K. Lerman. Using lists to measure homophily on twitter. In *AAAI workshop on Intelligent Techniques for Web Personalization and Recommendation*, July 2012.

[6] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. volume 42, pages 30–37, Los Alamitos, CA, USA, Aug. 2009. IEEE Computer Society Press.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[8] K. Lerman, R. Ghosh, and T. Surachawala. Social contagion: An empirical study of information spread on digg and twitter follower graphs. *CoRR*, abs/1202.3162, 2012.

[9] P. Lops, M. De Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[10] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 17–24, New York, NY, USA, 2007. ACM.

[11] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Dictionary learning for massive matrix factorization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1737–1746. JMLR.org, 2016.

[12] A. Said and A. Bellogín. Comparative recommender system evaluation: Benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 129–136, New York, NY, USA, 2014. ACM.

[13] A. Sharma, J. Jiang, P. Bommannavar, B. Larson, and J. Lin. Graphjet: Real-time content recommendations at twitter. *Proc. VLDB Endow.*, 9(13):1281–1292, Sept. 2016.

[14] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[15] F. A. Zamal, W. Liu, and D. Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*. The AAAI Press, 2012.