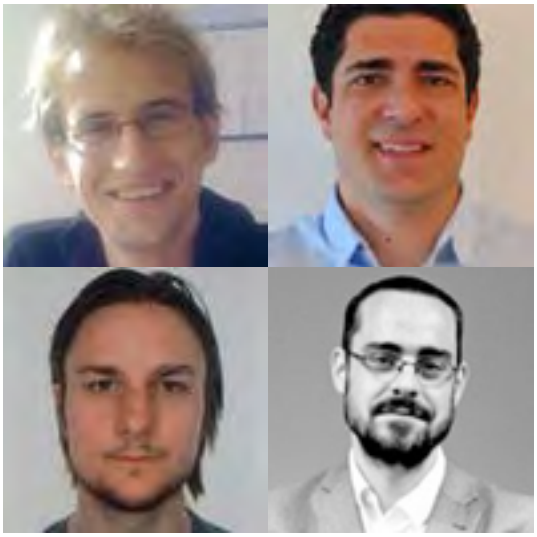
MÉTHODES

Présidentielle 2017 : l'analyse des tweets renseigne sur les recompositions politiques



Pierre LATOUCHE
Maître de Conférences en Mathématiques Appliquées, Université Paris 1

Charles BOUVEYRON
Professeur de Mathématiques Appliquées, Université Côte d'Azur

DAMIEN MARIE
Ingénieur, Société d'accélération de transfert technologique « IDFINNOV »

GUILHEM FOUETILLOU
Professeur associé, Sciences Po Paris

Voici un exemple d'utilisation de Big Data pour observer la société. C'est d'analyse politique qu'il s'agit ici. Les auteurs tirent parti des tweets émis juste avant et juste après le premier tour de l'élection présidentielle française d'avril 2017 pour regrouper les comptes twitter, en tenant compte à la fois des contenus des messages et des liens entre ces comptes. Ayant attribué des noms de partis politiques aux groupes issus de leur travail, ils peuvent proposer une analyse des forces en présence avant et après l'élection, ainsi que des transferts entre ces forces. Les détenteurs d'un compte twitter ne forment certainement pas un échantillon représentatif de toute la population électorale : et pourtant les résultats sont remarquablement proches de ceux de l'élection.

Introduction

Emmanuel Macron a été élu à la présidence de la République sur un programme dont une des priorités est la recomposition de la vie politique. La période précédant les législatives était donc sujette à de fortes interrogations quant à la réorganisation à venir des partis politiques. Afin d'apporter un éclairage sur ce point, nous avons étudié pendant les semaines qui ont précédé le second tour de l'élection présidentielle les mouvements et transferts entre les partis, avec un prisme particulier : celui du web social. En partenariat avec l'entreprise Linkfluence, nous avons analysé la recomposition des partis sur Twitter suite au premier tour.

Figure 1 : Visualisation d'un partitionnement (« clustering ») du réseau de tweets sur le premier tour de l'élection présidentielle 2017[1].

La plupart des outils permettant d'analyser ce type de données voient les tweets comme un ensemble de documents et ont pour objectif d'étudier le choix des mots, les thèmes de discussion majoritaires, et les sentiments relayés par les tweets. Cependant, les tweets sont par nature des données plus riches que de simples documents puisqu'ils caractérisent des interactions entre des individus. Par exemple, un individu A interagit avec B s'il retweete un message de B ou s'il écrit un message faisant référence à B. Un ensemble de tweets est alors vu comme un réseau « social ». Malheureusement, les outils traditionnels d'analyse de réseaux sont eux aussi limités et ne peuvent généralement gérer que des interactions binaires (interagit ou n'interagit pas) entre les individus.

Méthode et données

L'analyse des réseaux est un domaine de recherche particulièrement actif dont un des objectifs est l'extraction automatique d'informations pertinentes à partir des interactions observées entre des individus. Les méthodes ont été créées à l'origine en sciences sociales. Depuis, l'immense majorité des outils ont été proposés par des physiciens/informaticiens afin de maximiser un critère bien particulier, la modularité. Ce critère vise à identifier des groupes d'individus ayant plus de connexions entre eux qu'avec des individus d'autres groupes. C'est le principe de la communauté. Nous observons des communautés dans les réseaux sociaux vérifiant le principe de transitivité, i.e. l'ami de mon ami est mon ami. Malheureusement, les réseaux en général et sociaux en particulier sont souvent construits à partir d'autres types de groupes. Il existe par exemple des individus ayant une forte influence sur les avis/comportements des autres. On parle alors de groupes d'influenceurs et d'influencés. De la même manière, nous trouvons également régulièrement des structures inversées où il existe plus de connexions entre des individus de groupes différents qu'entre des individus d'un même groupe. La recherche en

1. Chaque point (=nœud) représente un point d'origine d'un tweet ; le « cluster » auquel ce point appartient, identifié par une couleur, correspond à la proximité par rapport à un des candidats, identifiée sur son identifiant et/ ou son contenu. Les courbes connectant les points (=arêtes) correspondent aux échanges (réponses aux tweets). La couleur et l'épaisseur d'une arête correspondent à l'intensité (multiplicité) des échanges sur un point de discussion et sa direction (intra- ou inter- cluster). Les couleurs des arêtes ont une signification différente des couleurs des nœuds et reflètent les sujets abordés.

Mathématiques, et en particulier en Statistique, a fourni ces quinze dernières années plusieurs solutions permettant de pallier les limites des outils existants. Ces approches permettent en particulier d'identifier des individus organisés en communautés, mais également en d'autres types d'organisations sociales. La recherche française en Statistique a largement contribué aux avancées théoriques et méthodologiques dans ce domaine.

Dans le cadre d'un projet de collaboration entre les laboratoires de Mathématiques des universités Paris 1 Panthéon-Sorbonne et Paris Descartes, nous avons proposé un nouveau modèle statistique, dénommé STBM (Stochastic Topic Block Model)[2], et une méthode d'estimation associée permettant de réaliser une analyse conjointe d'un réseau et d'un ensemble de textes. Le réseau social à analyser n'est alors plus vu comme un objet binaire. Un individu A interagit avec un individu B sur un texte donné. A peut par exemple envoyer plusieurs e-mails à B. Dans ce cas, l'interaction de A vers B est caractérisée par cet ensemble d'e-mails. Pour des données de type tweet, une interaction de A vers B rassemble tous les tweets écrits par A faisant directement ou indirectement (retweet) référence à B. L'analyse de ce réseau social permet alors d'identifier des groupes d'individus en fonction de à qui ils s'adressent et de quoi ils parlent. La méthode détermine les thèmes de discussion propres aux échanges entre les groupes. Elle permet ainsi de dire : le groupe G1 identifié discute beaucoup avec le groupe G2, sur le sujet S1 identifié.

Adapté aux réseaux de taille modérée à grande (de quelques centaines à plusieurs centaines de milliers d'individus), STBM peut ainsi analyser des échanges de textes, que ce soient des e-mails, des contenus scientifiques, des tweets, etc. D'un point de vue plus technique, le modèle au cœur de l'algorithme STBM est une généralisation de deux modèles statistiques reconnus : le SBM[3] (Stochastic Block Model) qui permet de modéliser la structure d'un réseau par partitionnement (« clustering ») et le LDA[4] (Latent Dirichlet Allocation) qui permet d'analyser les thèmes abordés dans des textes. La dépendance entre les deux modèles est faite au niveau des groupes des individus : les paramètres gérant la partie du modèle liée au texte dépendent des groupes des émetteurs et récepteurs des communications. L'inférence de ce modèle statistique repose sur un algorithme CVEM (Classification Variational Expectation-Maximization) qui optimise séquentiellement la vraisemblance des parties réseaux et textes.
STBM est ainsi capable d'étudier conjointement le contenu des échanges et les interactions entre des individus ou des groupes d'individus. A titre d'exemple, STBM a été appliqué à l'analyse du réseau des e-mails de l'entreprise Enron[5], qui a connu une faillite très médiatique au début des années 2000, et à l'analyse de réseaux de co-publications scientifiques. Notons que notre plateforme Linkage.fr permet à chacun de faire traiter par STBM ses propres données de réseaux (e-mails, PubMed, Arxiv, Twitter, ...). L'exemple suivant[6] permet d'illustrer l'approche.
A partir de tous les tweets des français liés à la politique, nous nous sommes concentrés sur deux périodes : 17-18 avril et 24-25 avril 2017, c'est à dire quelques jours avant, et juste après le premier tour. Les tweets liés à l'élection présidentielle ont été extraits et formatés par Linkfluence. L'ensemble de données fourni par Linkfluence s'appuie sur Radarly, le logiciel propriétaire développé par l'entreprise permettant de suivre en temps réel la quasi totalité du web social au niveau mondial. Dans ce cas précis, la totalité des mentions des 5 candidats principaux ont été captées sur le réseau social Twitter. Ainsi, environ 5 millions de verbatims ont été extraits pour l'analyse. La méthodologie statistique a été appliquée sur les réseaux

2. C. Bouveyron, P. Latouche and R. Zreik, *The Stochastic Topic Block Model for the Clustering of Networks with Textual Edges, Statistics and Computing*, 2017 (https://doi.org/10.1007/s11222-016-9713-7).
3. K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 96(455):1077–1087, 2001.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.
5. Voir https://linkage.fr/blog/Enron-Scandal pour une analyse détaillée.
6. L'analyse et ses graphiques dynamiques sont accessibles sur le site linkage.fr, permettant de compléter l'information statique présentée dans cet article. L'accès sur le site Linkage.fr est libre, moyennant l'ouverture totalement libre et gratuite d'un compte. Rechercher : "French Twitter Politics Discussion Groups before the 2nd round of the presidential elections of 2017" sous l'onglet "Jobs".

ainsi constitués et a identifié cinq thèmes de discussion et dix groupes d'individus, sur les deux périodes (c'est-à-dire avant et après le premier tour de l'élection).

Résultats de l'analyse

Pour la 1ère période (17-18 avril), quatre des thèmes trouvés correspondent aux tweets des français à propos des principaux candidats. Cependant, il est particulièrement intéressant de constater que le cinquième thème rassemble uniquement les tweets critiquant le système politique en général. Ce thème, au cœur de la campagne, est relayé par tous les partis politiques. Un examen des comptes présents dans chacun des groupes identifiés par la méthode nous a également permis d'étiqueter chaque groupe vis-à-vis de sa tendance politique. Un groupe dont les identifiants mentionnent explicitement un parti ou un candidat donné de son parti est étiqueté du nom de ce parti. Contrairement à tous les partis, le parti socialiste se retrouve isolé et n'interagit pas ou peu avec le groupe central en gris sur la Figure 2, rassemblant les comptes Twitter des candidats et des principaux médias.

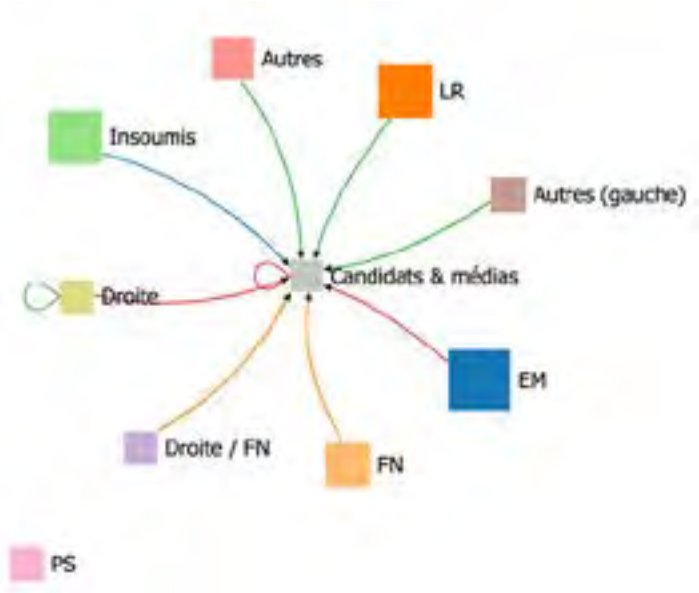


Figure 2 : représentation agrégée de la Twittosphère politique française des 17 et 18 avril[7].

De manière surprenante, les poids des partis que nous avons identifiés se sont avérés proches du vote des Français (Figure 3). 24,1% des comptes analysés ont ainsi été classés dans le groupe EM. Pour rappel, Emmanuel Macron a obtenu 24.01% des voix.

7. Chaque carré (=nœud) caractérise un ensemble de tweets regroupé sur leur proximité avec un parti ou un candidat de ce parti à partir de leur identifiant et/ ou de leur contenu. La taille de chaque nœud est proportionnelle au nombre d'éléments qu'il contient. Sa couleur a été attribuée arbitrairement pour permettre de les distinguer. Le positionnement de ces nœuds, figé sur cette vision statique, n'a pas de signification particulière ici. Les flèches indiquent les directions des tweets en terme de contenu émanant d'un groupe destiné à un ou des éléments d'un ou plusieurs groupes ou vers lui même (cf. boucles au niveau des carrés intitulés « Droite » ou « Candidats & Médias »). La couleur des flèches indique les thèmes majoritaires de discussion. Le choix automatisé de cette couleur par le logiciel est indépendant du choix des couleurs pour les nœuds, et répond à la légende suivante : thème Insoumis (bleu), thème FN (orange), thème Critique du système (vert), thème EM (rouge).

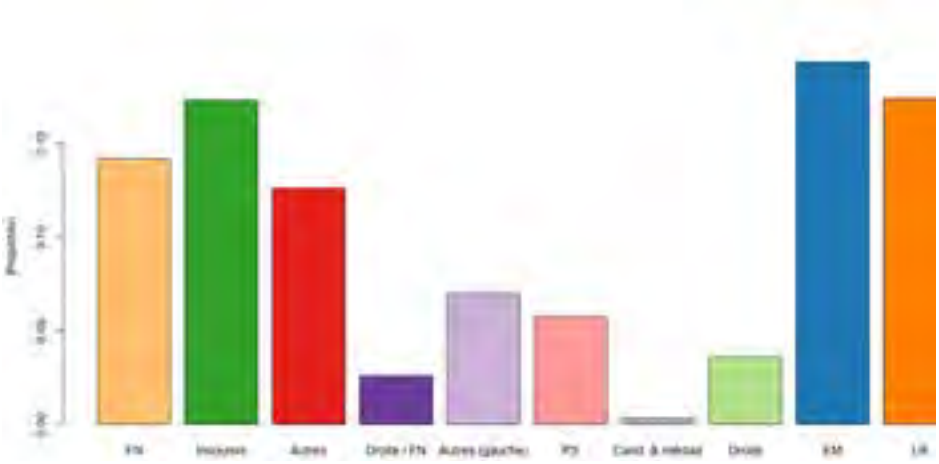


Figure 3 : poids des partis politiques sur Twitter les 17 et 18 avril[8].

Nous avons réalisé une analyse similaire sur la période 24-25 avril 2017, entre les deux tours de l'élection présidentielle, afin notamment d'observer la recomposition du paysage politique sur le réseau Twitter après les résultats du 1er tour (Figure 4). Deux thèmes sont associés à EM. Un est uniquement dédié à EM alors qu'un autre rassemble des discussions mentionnant à la fois EM et les Insoumis. Un thème correspond au FN et nous retrouvons deux thèmes de critique dont un de rejet du système politique.

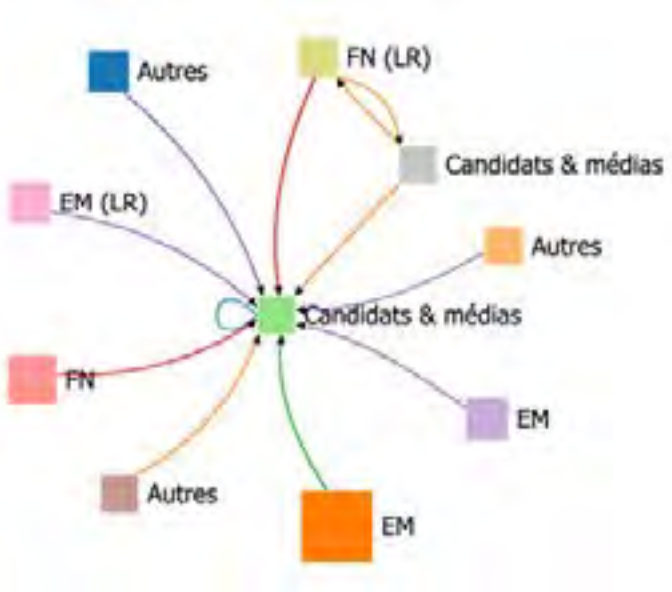


Figure 4 : représentation agrégée de la Twittosphère politique française des 24 et 25 avril.[9]

Comme pour le premier tour, nous avons pu identifier le poids des partis sur Twitter: 66% pour EM et 34% pour le FN. A la vue des résultats du 2nd tour, cette estimation du poids des partis sur le web social est bien sûr troublante. Il est néanmoins important de garder à l'esprit que le web social ne peut pas être directement utilisé aujourd'hui comme source pour le sondage car une grande partie de la population française n'est pas présente sur ces réseaux.

8. Les couleurs correspondent ici aux couleurs des carrés (nœuds) de la figure précédente.
9. Chaque nœud caractérise un groupe et sa taille est proportionnelle au nombre d'individus qu'il contient. Des couleurs leur sont attribuées sans que ce choix ni la position du nœud dans la figure n'aient d'autre but que de les individualiser. Les couleurs des flèches indiquent également les thèmes majoritaires des discussions : thème FN (rouge), thème EM-Insoumis (vert), thème EM (bleu), thème Critique du système (orange), thème Critique générale(violet). Les couleurs des flèches sont arbitraires, sans correspondance avec les couleurs des nœuds.

Fait unique, notre étude nous a permis de suivre les changements de comportement des comptes entre les deux tours. En utilisant les résultats des analyses sur les deux périodes, il nous a ainsi été possible d'estimer la recomposition du paysage politique à l'issue du 1er tour. La figure 5 permet de visualiser cette recomposition.

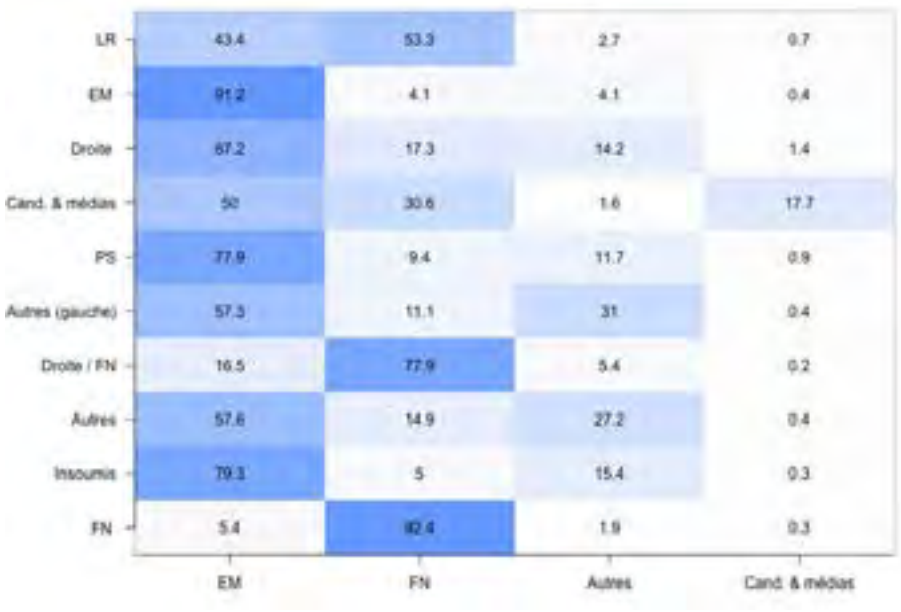


Figure 5 : Report des voix estimé par la méthode après le premier tour.
Les groupes identifiés pour les 17 et 18 avril sont en ligne.
Les groupes identifiés pour les 24 et 25 avril sont en colonne[10].

Nous avons communiqué ces résultats avant le second tour[11]. Il nous paraissait important de montrer que, sur le web social, les Insoumis semblaient finalement se tourner vers EM. Nous voulions également témoigner de la fracture que nous avions observée à droite. Une part importante des comptes actifs et proches de François Fillon a été classée FN suite au premier tour. Le reste des comptes de droite et issus de LR sont allés majoritairement vers EM.

Conclusion

Cette étude a permis de mettre en œuvre la méthodologie statistique STBM d'analyse de réseaux avec arêtes textuelles à une problématique importante qu'est l'étude d'une élection présidentielle sur le web social. Cette étude a été rendue possible par l'implémentation de la méthode STBM sur la plateforme Linkage.fr et la collaboration avec l'entreprise Linkfluence qui a capté et pré-traité les données Twitter. Outre la description synthétique de l'événement sur le web social, le résultat le plus important de cette étude est certainement la quantification de la recomposition des partis politiques entre les deux tours. En effet, la comparaison des résultats de « clustering » des deux périodes nous a permis d'estimer cette recomposition et ainsi de valider ou invalider certaines hypothèses émises par les analystes politiques.

10. Les chiffres indiquent, pour chaque colonne, les reports de voix pour le 2ème tour à l'intérieur de chaque groupe (ligne) identifié au premier tour. Le total de chaque ligne fait donc 100.
11. Tweet du vendredi 5 mai : https://twitter.com/latouche_pierre/status/860471570220929024