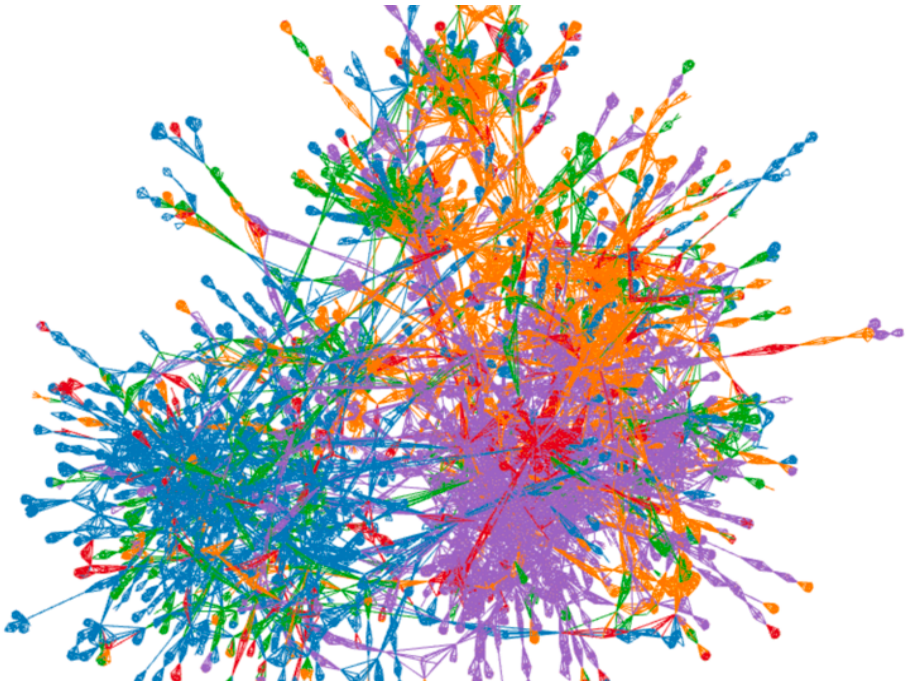


Linkage : Un outil d'intelligence artificielle pour l'analyse de réseaux

LINKAGE : QUI PARLE A QUI ET DE QUOI ?

LINKAGE EN 2 MIN ...

Le projet Linkage

CONTEXTE

L'intérêt croissant pour l'analyse des réseaux s'explique d'une part par la forte présence de ce type de données dans le monde numérique d'aujourd'hui et, d'autre part, par les progrès récents dans la modélisation et le traitement de ces données. Les méthodes de clustering permettent en particulier de découvrir une structure en groupes cachés dans le réseau. Dans ce cadre, les méthodes statistiques d'intelligence artificielle (IA) présentent l'avantage d'offrir une segmentation fine des données dont l'interprétation est facilitée par le modèle statistique sous-jacent. Malgré les nombreux développements dans ce domaine, l'analyse conjointe des réseaux et des textes associés n'a reçu qu'une attention très limitée, alors même que la plupart des réseaux sociaux sont aujourd'hui associés à du texte (emails, Facebook, Twitter, ...). La méthodologie implémentée dans le logiciel Linkage permet de pallier ce manque et autorise l'analyse conjointe de réseaux et de textes.

LA METHODOLOGIE

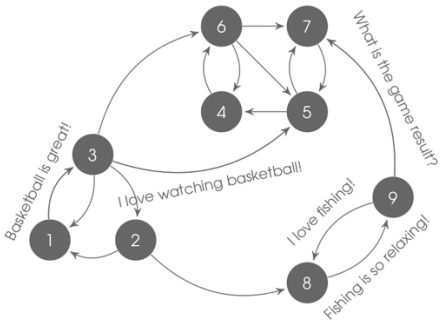
Le modèle statistique qu'implémente Linkage permet de segmenter les nœuds (individus) d'un réseau avec arcs ou arêtes textuels, tout en identifiant les thèmes de discussions (*topics*) utilisés. Les réseaux peuvent être dirigés (emails, tweets, ...) ou non dirigés (publications scientifiques, posts facebook, ...). Linkage requiert uniquement la donnée d'un ensemble d'échanges de textes entre des individus, ou plus généralement entre des entités. Par exemple, on peut considérer les échanges de textes entre des individus d'un réseau social, ou les échanges d'emails entre les employés d'une entreprise, ou encore les co-publications de brevets ou publications scientifiques. Linkage est également capable de déterminer automatiquement le nombre de groupes d'individus et le nombre de *topics* utilisés. De ce fait, le *process* des données est totalement automatisé et ne requiert aucun paramétrage expert. Les sorties du logiciel sont les classifications des individus dans les groupes (assortis de probabilités) ainsi qu'une description des *topics* à l'aide des mots les plus représentatifs.

PLATEFORME SAAS

Une plateforme bridée de démonstration, accessible à l'adresse https://linkage.fr, permet à tout un chacun d'expérimenter l'usage de Linkage sur ses propres données ou sur des jeux de données publiques. La création d'un compte sur la plateforme est gratuite mais le volume des données qui peut être traité sur la plateforme est limité. Les meilleures librairies de calculs et les protocoles les plus récents de sécurité ont été utilisés pour garantir un traitement optimal des données. Sur la plateforme, il est possible d'uploader ses propres données dans un format simple (.csv), de récupérer son réseau de mail grâce à l'API Gmail ou de faire des requêtes sur des données publiques (Twitter, PubMed, Arxiv, ...). De nombreuses sorties graphiques permettent d'exploiter facilement les résultats de l'analyse. De plus, il est possible de télécharger les résultats bruts pour les exploiter dans un logiciel de visualisation, tel que Gephi par exemple.

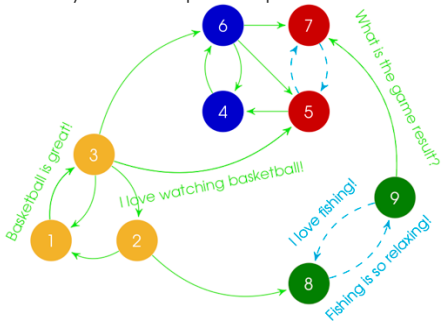
ANALYSE DE RESEAUX DE COMMUNICATION

La technologie Linkage permet d'analyser des réseaux de communications dont les acteurs (personnes, comptes Twitter, ...) échangent des contenus composés de textes. A titre d'illustration, considérons le réseau (simulé) d'emails entre 9 personnes, présenté dans la figure ci-dessous.



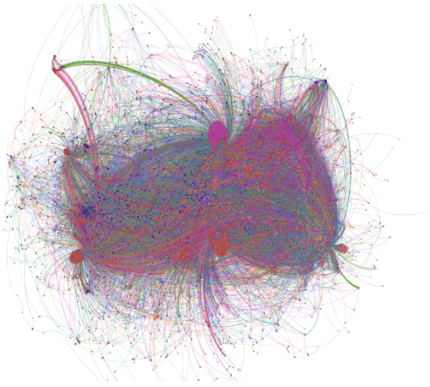
SEGMENTATION DES INDIVIDUS ET DES TEXTES

Sur un réseau de petite taille comme celui ci-dessus, un analyste sera capable de déterminer qu'il y a 4 communautés d'individus (couleurs des nœuds) et que ceux-ci échangent sur 2 sujets (couleurs des arcs). Linkage a bien sûr cette capacité également, comme l'illustre le résultat ci-dessous, mais sera capable de faire de même sur de très gros volumes de données, là où l'analyste n'en n'a plus la capacité !



UNIVERSITÉ PARIS DESCARTES — LABORATOIRE MAP5 MATHÉMATIQUES APPLIQUÉS — cnrs — IDF INNOV De l'invention à l'innovation — SAMM — UNIVERSITÉ PARIS 1 PANTHÉON SORBONNE

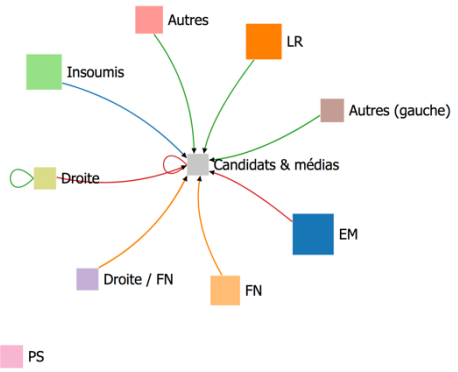
Présidentielle 2017



ANALYSE DES DONNEES TWITTER

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 notre partenaire Linkfluence. L'ensemble de données fourni couvre la totalité des mentions des 5 candidats principaux sur le réseau social Twitter. Ainsi, environ 5 millions de verbatims ont été extraits pour l'analyse. La méthodologie statistique implémentée par Linkage a ensuite été appliquée sur les réseaux ainsi constitués et a identifié cinq thèmes de discussions et dix groupes d'individus, dans les deux cas. La figure ci-dessus propose une visualisation du clustering du réseau de tweets sur le premier tour de l'élection présidentielle 2017.

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.

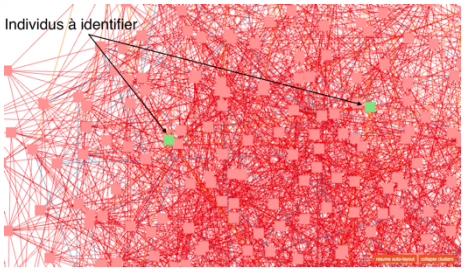


Il est intéressant de noter que les poids des partis que nous avons identifiés se sont avérés proches du vote des français : 24.1% des comptes analysés ont ainsi été classés dans le groupe EM. Pour rappel, Emmanuel Macron a obtenu 24.01% des voix lors du 1er tour.

Détection de signaux faibles

SIMULATION D'UN SIGNAL FAIBLE

Linkage implémente un algorithme statistique d'intelligence artificielle unique pour l'analyse des réseaux de communications, et ses applications en matière de détection de signaux faibles sont multiples. En comparaison des autres outils disponibles, issus des secteurs publics ou privés, français ou étrangers, Linkage a la plus grande capacité de détection de signaux faibles. A titre illustratif, nous considérons ici un réseau composé de trois groupes d'individus où deux individus d'intérêt, correspondant à un quatrième groupe, sont à identifier. Ces deux individus communiquent globalement comme n'importe quel individu du réseau et abordent les mêmes sujets dans leurs échanges. Dans le détail en revanche, ils utilisent certains mots plus que les autres. Pour complexifier la situation, ces mots (clés) ne sont pas supposés connus à l'avance. Aucun des autres outils testés d'analyse n'est capable de détecter les deux individus en question. Les approches d'analyse de textes ne voient en effet rien car les signaux (variations en terme de vocabulaire) sont trop faibles. Les méthodes se concentrant sur l'étude du réseau uniquement, et pas sur le contenu, non pas non plus la capacité de retrouver les deux individus.



RESULTATS

Linkage étant spécialisé dans l'analyse conjointe de réseaux et de textes, le groupe d'intérêt est facilement identifié comme illustré ci-dessus. Ce jeu de données est disponible sur la plateforme et librement téléchargeable afin que les utilisateurs puissent tester leurs propres outils et tenter de détecter les deux individus en question.

CAPACITE DE TRAITEMENT

Linkage est capable d'analyser des réseaux ayant des centaines de milliers ou des millions d'individus et peut traiter des volumes d'échanges particulièrement important allant du Mo au Po. A titre illustratif, le logiciel peut notamment analyser toutes les communications échangées entre 25 000 individus en moins de 2 minutes sur un ordinateur personnel standard. Son coût algorithmique est linéaire en le nombre d'individus.

CONTACT

LES AUTEURS

Charles BOUVEYRON est Professeur des Universités en Mathématiques Appliquées à l'Université Côte d'Azur, à Nice. Il est également titulaire de la Chaire d'Excellence INRIA en Science des Données. Email : charles.bouveyron@math.cnrs.fr

Pierre LATOUCHE est Professeur des Universités en Mathématiques Appliquées à l'Université Paris Descartes et à l'Ecole Polytechnique. Email : pierre.latouche@math.cnrs.fr

LES PARTENAIRES

Le projet Linkage a été mené dans le cadre d'une collaboration entre les laboratoires MAP5 (Université Paris Descartes & CNRS) et SAMM (Université Paris 1 Panthéon-Sorbonne).

COMMERCIALISATION

La commercialisation de Linkage est assurée par la SATT IdfInnov (http://idfinnov.com). Pour toutes informations concernant l'acquisition de licences, contacter la SATT. Email : bd-3s@idfinnov.com

VIDEO DE PRESENTATION

Une vidéo de présentation du logiciel est accessible à l'adresse suivante : https://youtu.be/wUNJip1jH_U

PLATEFORME

La plateforme de démonstration est accessible à l'adresse : https://linkage.fr