Évaluation et Prédiction de la Centralité de Groupes de Recherche dans un Réseau de Collaborations Scientifiques

Aurélien Bossard*, Mario Cataldi*
Myriam Lamolle*, Chan Le Duc*

*Université Paris 8 - LIASD-EA4383
IUT de Montreuil
{a.bossard,m.cataldi,m.lamolle,c.leduc}@iut.univ-paris8.fr,

Résumé. De nos jours, il y a un fort intérêt pour de nouvelles méthodes d'évaluation des groupes de recherche afin de quantifier l'impact de leur travail sur toute la communauté scientifique et de tenter de prédire leurs performances dans le futur. Dans ce contexte, nous proposons une nouvelle approche hybride qui mesure la centralité d'un groupe de chercheurs publiants. Cette mesure profite de l'expressivité et de la capacité d'inférence apportées par une modélisation ontologique des groupes et des thématiques inférées, et d'une modélisation en graphe qui permet d'explorer les interactions entre ces différents groupes au fil du temps. Ce modèle permet également de détecter les groupes capables de collaborer avec d'autres tout en maintenant un haut niveau de production, et d'identifier ceux qui sont plus déterminants sur les thématiques déduites, afin de développer des collaborations de recherche plus fructueuses.

1 Introduction

De nos jours, une des méthodes les plus populaires pour évaluer le travail d'un chercheur ou d'un groupe de recherche est de considérer ses publications scientifiques et d'évaluer leurs incidences sur les communautés scientifiques nationales et internationales. Cependant, la façon d'évaluer cet impact est encore débattue ; la plupart des méthodes existantes insistent sur l'analyse quantitative du nombre d'articles produits et essaient d'estimer, avec des approches très différentes, le nombre et la qualité des éventuelles citations. Ces mesures représentent bien évidemment des outils précieux pour l'analyse du monde académique mais elles ne considèrent pas les relations scientifiques et leurs impacts.

Sachant cela, et que bon nombre de ces mesures sont également utilisées pour la distribution des financements et/ou recrutements, il est crucial d'analyser les relations scientifiques et les collaborations entre groupes de recherche afin d'estimer la capacité d'un groupe à travailler et à produire des résultats *sans* les autres groupes qui ont aidé son travail jusqu'à une date donnée. De plus, l'étude des collaborations scientifiques peut permettre d'analyser la centralité de chaque groupe dans le réseau national et international de collaborations de recherche.

Une collaboration peut être en effet définie comme un processus à double sens où les organisations partagent leurs idées et leur expérience pour produire des résultats scientifiques

en commun. Les collaborations sont nécessaires en raison de la difficulté évidente pour les groupes de recherche de mener individuellement plusieurs recherches impactantes. Pour cela, l'un des aspects clé d'un groupe qui vise une certaine réussite est le développement d'un réseau étendu et actif de collaborateurs qui peut l'aider à apporter de nouvelles solutions et à proposer, en continu, de nouvelles idées à la communauté de recherche. D'autre part, le système d'évaluation des groupes a besoin d'un processus inverse dont l'objectif principal est de comprendre le rôle de chaque groupe et son impact spécifique sur la communauté de recherche, dans cet environnement collaboratif.

Pour atteindre cet objectif, nous proposons un nouveau modèle hybride visant *(i)* à évaluer les collaborations scientifiques d'un groupe au fil du temps et *(ii)* à estimer sa centralité dans la communauté à partir de l'analyse précédente. Ce nouveau modèle utilise en premier lieu l'expressivité apportée par une modélisation ontologique des collaborations scientifiques et de toutes les thématiques traitées qui permet, selon les besoins, d'extraire ces informations selon des points de vue différents. Les relations entre les groupes de recherche et les thématiques sont ensuite modélisées au travers d'un graphe de collaboration [1]. Ce graphe est finalement exploité pour estimer la centralité de chaque groupe de chercheurs dans ce réseau des contributeurs de la conférence EGC.

Cette approche hybride permet d'exploiter, de façon automatique, la puissance du moteur d'inférence ontologique pour déduire, à plusieurs niveaux, différentes définitions d'un groupe (*i.e.* chercheurs avec la même affiliation et/ou personnes qui travaillent sur le même sujet de recherche) et, pour chaque définition, calculer l'importance de chaque groupe dans la communauté de la conférence EGC analysée. Par cette approche, nous pouvons détecter les groupes les plus déterminants pour le bon développement de collaboration de recherche à l'intérieur d'une communauté, et prédire ceux qui ont une haute probabilité de le devenir dans le futur.

2 État de l'art

Les indicateurs bibliométriques sont de plus en plus utilisés pour évaluer les publications des groupes de recherche fondés sur leur liste de publications scientifiques. Le simple nombre d'articles publiés et les citations reçues peuvent capturer respectivement la quantité et l'impact de l'ensemble des œuvres d'un auteur.

Cependant, ces méthodes n'aident pas à comprendre la contribution réelle d'un groupe de recherche dans un réseau collaboratif. Par exemple, Slone (1996) souligne en premier lieu le problème de *co-auteurs immérités*. D'autre part, l'analyse simple de la position d'un chercheur dans la liste des co-auteurs ne suffit pas (Imperial et Rodríguez-Navarro, 2007). En effet, l'ordre des auteurs généralise un concept qui suit des règles inconnues.

Un autre point intéressant de l'analyse est donné par le cas suivant : un auteur d'un seul article très cité et *a contrario* d'un ensemble de travaux très peu cités. En considérant ce problème, les approches qui se fondent sur les citations et le nombre d'articles publiés sont inefficaces car elles doivent se fonder sur un seuil prédéfini généralement difficile à choisir et à affiner. Un exemple de ce fait est le nombre d'*articles importants* (Hirsch, 2005), défini comme le nombre d'articles ayant plus d'un certain nombre de citations. (Hirsch, 2005), (et plus tard (Hirsch, 2007)) a introduit l'indice h bien connu, une mesure pour l'évaluation des CV de

1. dont une définition plus précise sera donnée par ailleurs

chercheurs, des impacts des revues et des communautés de recherche. Selon cette mesure, un auteur (ou un centre de recherche) a un indice h s'il a publié h papiers cités au moins h fois chacun. Comme il a été démontré dans Sidiropoulos et al. (2007), il mesure indirectement à la fois la qualité et la quantité.

Selon un autre point de vue, Ausloos (2013) a proposé que l'évaluation scientifique d'un chercheur ne soit pas fondée sur le nombre de citations de ses articles, mais sur le nombre de chercheurs avec lesquels il a été en mesure de collaborer afin de produire des publications scientifiques. Van den Besselaar et al. (2012) ont travaillé avec cette même motivation et ont proposé un indicateur d'indépendance composé par plusieurs dimensions différentes. Dans Pepe et Rodriguez (2010) et Rodriguez et Pepe (2008), les auteurs cherchent à détecter les affiliations universitaires, la position et le pays d'origine des réseaux socio-académiques (Taramasco et al., 2010) en mettant l'accent sur l'évolution des groupes de recherche. Yan et Ding (2009) utilisent quatre mesures de centralité au sein d'un réseau de collaborations restreintes, montrant qu'ils sont significativement corrélés avec le nombre de citations. Dans Radicchi et al. (2009), les auteurs aspirent à découvrir la diffusion de crédits scientifiques dans la communauté de recherche en se fondant sur leur réseau de citations.

Dans cet article, nous avons encore étendu ce travail afin d'analyser les collaborations, au fil du temps, de chaque groupe de recherche (avec plusieurs définitions possibles) tout en prenant en compte la façon dont chaque collaboration évolue dans différents intervalles de temps.

3 Étude de centralité d'un groupe sur le graphe de collaboration selon une thématique de recherche

Dans cette section, nous analysons le problème de l'évaluation des profils de recherche dans le but de proposer une nouvelle métrique pour estimer l'impact, dans le temps, d'un groupe de recherchesur toute la communauté de recherche. Pour ce faire, nous proposons une nouvelle méthode pour estimer la centralité d'un groupe par rapport à une thématique à travers l'évaluation de sa capacité à maintenir la même production scientifique dans un environnement collaboratif différent sur la thématique évaluée.

Dans ce contexte, nous définissons un *groupe de recherche* comme un ensemble d'auteurs d'articles scientifiques sous des affiliations communes et/ou liés par des thématiques de recherche. Une *collaboration* est donc définie comme l'activité commune de deux groupes (ou plus) de chercheurs visant à atteindre un but commun, à savoir la production de nouveaux articles scientifiques.

Nous introduisons en premier lieu les méthodes pour détecter, de façon automatique, les groupes de recherche avant d'exposer notre approche de calcul de centralité.

3.1 Détection des groupes et de leurs thématiques par ontologie

Afin de déterminer les groupes et leurs thématiques de recherche, nous représentons par des ontologies les liens sémantiques explicites entre des concepts (simples ou complexes). Ces liens sont exprimés en un langage d'ontologie (OWL [2]) avec une sémantique formelle et

2. `http://www.w3.org/TR/owl2-overview/`

non ambiguë. Les liens sémantiques *implicites* ou *inférés* entre les concepts sont déduits des ontologies à l'aide d'un raisonneur (HermiT [3] pour cette étude) via une interface munie du langage de requête DL [4].

L'avantage de l'approche ontologique est qu'une fois les connaisances bien modélisées dans une ontologie, les utilisateurs peuvent l'interroger via une API (*e.g.* par des services Web avec OWLAPI [5]) en utilisant des requêtes pertinentes par rapport à leurs demandes. En d'autres termes, les efforts de programmation pour l'implémentation d'une nouvelle mesure provenant d'une interprétation des données existantes sont réduits aux efforts de composition de "bonnes" requêtes.

3.1.1 Ontologie DataOnto

Dans cette étude, nous cherchons à modéliser les liens sémantiques entre les auteurs des articles, leurs affiliations (les intitulés, les adresses, etc.), les métadonnées sur les articles, les années de publication et les mots-clés extraits des résumés des articles [6]. Dans un premier temps, ces liens sémantiques sont traduits en axiomes dans une ontologie OWL, appelée DataOnto. Ces liens sémantiques proviennent d'une base de données qui modélise les données de la conférence EGC. Plus précisément, DataOnto exprime les classes, les propriétés, les relations de subsumptions suivantes :

1. Classes :
   - `Author` : ensemble des chercheurs ayant écrit au moins un article,
   - `Team` : ensemble des groupes de chercheurs ayant écrit au moins un article ensemble,
   - `Paper` : ensemble des articles de la conférence EGC,
   - `ResearchArea` : ensemble des domaines de recherche constitué à partir des mots-clé extraits des résumés,
   - `Affiliation` : ensemble des affiliations des auteurs,
   - `Country`, `City`.
2. Propriétés : les relations binaires entre concepts sont principalement
   - `isMemberOf(Author, Team)` qui permet d'inférer les membres d'un groupe,
   - `write(Author, Paper)` qui spécifie les auteurs d'un article,
   - `relatedTo(Paper, ResearchArea)` qui rattache un article à différents domaines de recherche,
   - `workOn(Author, ResearchArea)` qui détermine les thématiques de recherche d'un auteur,
   - `isCoauthorWith(Author, Author)` qui permet de trouver tous les co-auteurs d'un chercheur,
   - `isLivingIn(Author, City)` qui situe géographiquement un auteur,

L'ontologie DataOnto peut fournir les informations quantitatives à l'aide de requêtes DL et d'un raisonneur comme HermiT Motik et al. (2009). Par exemple, l'utilisateur peut interroger l'ontologie pour connaître le nombre de publications en commun entre deux chercheurs.

3. `http://hermit-reasoner.com/`
4. Description Logic
5. cf. `http://owlapi.sourceforge.net/`
6. l'article complet au format pdf n'étant pas toujours fournie par RNTI

En sus des liens sémantiques provenant directement de la base de données, l'ontologie DataOnto est enrichie par les relations entre les mots-clés et les mots apparaissant dans les titres et les résumés des articles. Cet enrichissement a pour but de faciliter la découverte des collaborations existantes ou potentielles entre les chercheurs et les groupes.

3.1.2 Ontologie ExtendedOnto

Afin d'enrichir les liens sémantiques exprimés dans DataOnto et les résultats quantitatifs issus des requêtes DL, il est nécessaire d'avoir une version étendue de celle-ci, appelée ExtendedOnto, qui modélise dans la mesure du possible les notions de collaboration sur un ensemble de thématiques de recherche sémantiquement liées telles que introduites dans la section 3.1.1. Pour ce faire, nous introduisons :

- une propriété `isSameTeam(Author, Author)` modélisant la notion de groupe comprenant les co-auteurs, et un axiome
$$\texttt{isCoauthorOf} \sqsubseteq \texttt{isSameTeam}$$
imposant que les auteurs qui sont co-auteurs doivent appartenir à un groupe. Grâce au chaînage
$$\texttt{write(Author, Paper)} \circ \texttt{relatedTo(Paper, ResearchArea)}$$
un raisonneur peut calculer les liens réels de la propriété `workOn(Author, ResearchArea)`, et donc trouver toutes les thématiques abordées par les auteurs,
- une propriété `isPotentialCoauthor(Author, Author)` exprimant le lien entre des auteurs qui travaillent sur les mêmes thématiques,
- une propriété *transitive* `isPotentialSameTeam(Author, Author)` modélisant la notion de groupe *potentiel*. Un tel groupe comprend non seulement les co-auteurs qui ont écrit au moins un article en commun mais aussi ceux qui pourraient potentiellement collaborer pour produire une publication sur un ensemble de thématiques liées sémantiquement. Les contraintes sémantiques présentées sont exprimées par les deux axiomes suivants :
$$\texttt{isPotentialCoauthor} \sqsubseteq \texttt{isPotentialSameTeam}$$
$$\texttt{workOn} \circ \texttt{workOn}^{-} \sqsubseteq \texttt{isPotentialCoauthor}$$
La collaboration potentielle entre deux auteurs est déduite de la transitivité de la propriété `isPotentialSameTeam` et du premier axiome. Le deuxième axiome, quant à lui, permet à un raisonneur de détecter toutes les thématiques de recherche sur lesquelles les auteurs concernés pourraient collaborer.

Afin de déterminer s'il y a une collaboration potentielle entre deux chercheurs quelconques c_i et c_j sur une thématique w (considérés comme des individus des ontologies ci-dessus), nous composons les requêtes DL (avec OWLAPI) suivantes :

$$\text{isPotentialCoauthor}\ (c_i, c_j),\ \text{isPotentialSameTeam}\ (c_i, c_j),\ \text{workOn}\ (c_i, \text{w})$$

Les résultats obtenus par l'intermédiaire d'un raisonneur servent à former les groupes dont les membres ont collaboré ou pourront potentiellement collaborer sur un ensemble de thématiques de recherche liées sémantiquement. La flexibilité du requêtage permet de répondre à différentes questions portant sur les éléments sémantiques modélisés dans les ontologies. À titre d'exemple, une nouvelle détection des groupes de chercheurs en fonction des informations géographiques (*eg. Paris*) et une thématique précise peut se faire par les requêtes supplémentaires suivantes :

isLivingIn (c_i, Paris), workOn (c_i, Ontology)

L'ontologie ExtendedOnto est accessible sur le lien

https ://github.com/ontoEGC/DefiEGC

3.2 Formalisation du graphe de collaboration

En considérant un groupe de recherche G_i, formé par deux chercheurs c_i et c_j, obtenu grâce à l'approche ontologique, nous formalisons son ensemble de publications scientifiques de recherche $P^t_{G_i}$ (appelés également papiers scientifiques dans cet article) antérieures à l'instant t, comme [7] :

$$P^t_{G_i} = P^t_{c_i} \cap P^t_{c_j} = \{p^t_{(c_i,c_j),1}, p^t_{(c_i,c_j),2}, \cdots, p^t_{(c_i,c_j),m}\}, \quad (1)$$

où $p^t_{(c_i,c_j),k}$ est le k-ème article de recherche co-écrit par c_i et c_j au moment t et m est le nombre total d'articles co-écrits par les chercheurs au moment t. Il est alors possible de quantifier leur productivité au moment t comme $|P^t_{c_i,c_j}|$. Notons que cette approche est applicable à tout groupe de recherche avec tout type de cardinalité.

De plus, soit un groupe de recherche $G_i = c_1, c_2, \cdots c_n$ formé par n chercheurs, nous formalisons les collaborations scientifiques de ce groupe comme

$$Net^t_{G_i} = \{G^t_1, G^t_2, \cdots, G^t_h\} \quad (2)$$

où h est le nombre total des groupes de recherche, au moment t, ayant eu au moins une collaboration scientifique avec e_i (*i.e.* ils ont co-écrit au moins un article ensemble). De même, considérant deux groupes G_i et G_j, nous formalisons leur coopération scientifique comme $Net^t_{G_i,G_j}$, au temps t, comme l'ensemble des groupes ayant collaboré au moins une fois avec les deux groupes ensemble ($G_k \in Net^t_{G_i,G_j} \rightarrow \exists p^t_x \in P^t_{G_i,G_j}$, et p^t_x est également co-écrit par G_k).

Dans les sections suivantes, nous allons utiliser ces formalisations pour présenter notre modèle d'évaluation des groupes.

3.3 Centralité d'un groupe dans le graphe de collaboration

Étudions maintenant chaque collaboration scientifique d'un groupe de recherche afin d'estimer sa centralité dans la communauté de recherche EGC.

Pour ce faire, nous définissons un graphe de collaboration parmi les groupes de recherche, C^t, comme un graphe non orienté qui exprime la centralité de chaque groupe, au temps t, dans toute la communauté scientifique. Plus formellement, nous définissons la communauté scientifique au moment t, CS^t, comme

$$CS^t = (V^t, A^t, w), \quad (3)$$

– $V^t = (G^t_1, G^t_2, \cdots, G^t_n)$ étant l'ensemble complet des n groupes scientifiques de recherche au moment t (groupes qui ont publié au moins un article à l'instant t) ;

7. Dans cette étude, l'unité temporelle considérée est l'année.

– A^t étant l'ensemble des arêtes non orientées, où chaque $a_{i,j} \in A$ représente une collaboration entre G_i et G_j (où $G_i, G_j \in V$) motivée par la présence à l'instant t d'au moins un article co-écrit par les deux groupes ;
– w étant la fonction de centralité au moment t du groupe G_i à l'intérieur du graphe CS^t.

Grâce à cette mesure de centralité w, nous essayons de compter le nombre de fois qu'un nœud agit comme un point de passage du plus court chemin entre deux autres nœuds. Les nœuds qui ont une forte probabilité d'apparition sur les plus courts chemins entre deux nœuds choisis au hasard ont alors un haut degré de centralité.

Ainsi, la centralité d'un nœud $v \in V^t$ (représentant un groupe de recherche) est calculée selon la méthode suivante :

1. Pour chaque couple de nœuds u et z (où u et $z \in V^t$), nous calculons les plus courts chemins les reliant.
2. Pour chaque couple de nœuds u et z (où u et $z \in V^t$), nous déterminons la proportion de plus courts chemins passant par v.
3. Nous sommons cette fraction sur tous les couples possibles de nœuds du graphe considéré.

De façon plus concise, la centralité peut être représentée par :

$$w_v = \sum_{u \neq v \neq z \in V} \frac{\sigma_{uz}(v)}{\sigma_{uz}} \tag{4}$$

où σ_{uz} est le nombre total de plus courts chemins du nœud u au nœud z et $\sigma_{uz}(v)$ est le nombre de tels chemins qui passent par v. Nous normalisons donc la valeur de centralité en divisant le résultat par le nombre de couples de nœuds ne comprenant pas v, nombre qui est égal à $(N-1)(N-2)/2$ dans un graphe non orienté.

Pour mieux comprendre l'approche de centralité proposée, considérons un graphe étoile non orienté : le nœud central sera contenu dans tout plus court chemin ayant une centralité égale à $(N-1)(N-2)/2$ (qui, après normalisation, sera donc égale à 1). *A contrario*, les nœuds "feuilles", dans un graphe étoile, ne sont contenus dans aucun plus court chemin ; ils auront une valeur de centralité égale à 0.

Notons que cette approche nécessite le calcul des valeurs de centralité pour tous les nœuds du graphe. Cela requiert donc de calculer les plus courts chemins entre toutes les paires de nœuds d'un graphe, soit une complexité de $O(N^2 \log N + NM)$ pour un graphe à M arêtes et N nœuds.

3.4 Trajectoire de centralité : évolution et prédiction de la centralité d'un groupe dans le temps

Dans la section précédente, nous avons introduit une nouvelle façon d'estimer la centralité d'un groupe, à un moment précis, dans le réseau scientifique en fonction de la productivité de chaque collaboration du groupe au sein de ce réseau. Ces valeurs peuvent maintenant être utilisées pour définir la courbe de centralité d'un groupe G_i comme

$$\overrightarrow{w_{G_i}} = \{w^t_{G_i}, w^{t+1}_{G_i}, \cdots, w^{t+n}_{G_i}\}, \tag{5}$$

où t est l'année de la première publication de G_i et n exprime la différence arithmétique entre la dernière et la première année de publication du groupe.

En outre, nous cherchons également à obtenir un système d'évaluation pour résumer, en une seule valeur, la centralité globale d'un groupe dans l'intervalle de temps considéré. Ainsi, compte-tenu de l'ensemble complet des valeurs de centralité, nous calculons la *trajectoire* de centralité d'un groupe en calculant l'écart, au cours du temps, de chaque valeur de centralité de valeur optimale 1 (soit un score de centralité de 1). Par cette approche, nous cherchons donc à évaluer la centralité totale du groupe au cours des années de publications analysées. Plus formellement, étant donné un groupe G_i, nous définissons sa trajectoire de centralité comme

$$\overrightarrow{t_{G_i}} = \{em^t_{G_i}, em^{t+1}_{G_i}, \cdots, em^{t+n}_{G_i}\}, \tag{6}$$

où $em^t_{G_i}$ est calculé selon la méthode suivante :

$$em^t_{G_i} = \sqrt{\frac{\sum_{a_k \in Net_{G_i}} (C^t_{G_i})^2}{|Net_{G_i}|}}. \tag{7}$$

Avec cette formule nous calculons l'écart moyen des valeurs de centralité précédemment calculées par rapport à 0. Plus $em^t_{G_i}$ est élevé, plus le travail de G_i est central pour la collaboration avec les autres groupes de la communauté analysée. Par cette formule, le système tente de détecter des anomalies dans les modèles de centralité de collaboration à l'égard de certaines valeurs attendues. On attend, en fait, qu'un groupe, tout au long de son activité, augmente son réseau de collaboration et, par conséquent, devienne plus central dans le réseau de collaboration de recherche.

3.5 Progression de centralité dans le temps et prédiction de valeur future

Une fois obtenue la courbe de centralité d'un groupe G_i, $\overrightarrow{c_{G_i}}$, il est possible de prédire la valeur future de centralité de G_i dans le graphe de collaboration (avec l'hypothèse de maintenir les mêmes nœuds du graphe) en utilisant un modèle de régression linéaire.

Le modèle prend donc en compte la distribution à deux variables définie par (i) la courbe de centralité d'un groupe et (ii) la séquence des années durant lesquelles le groupe a publié dans la conférence considérée. Avec cette entrée, nous construisons la droite de régression linéaire (Figure 1) par la méthode des moindres carrés ordinaires. Cette droite permet d'estimer la valeur de centralité d'un groupe dans le futur.

4 Expérimentations

Dans cet article, nous avons présenté un modèle hybride qui permet de concentrer l'évaluation d'un groupe de recherche (avec plusieurs définitions ontologiques possibles) sur l'analyse de ses collaborations scientifiques. Pour nos expérimentations, nous avons considéré l'ensemble des articles extraits de la base de données bibliographiques EGC [8] qui contient des informations sur tous les auteurs de la conférence EGC et leurs articles scientifiques.

8. http ://editions-rnti.fr/files/RNTI_articles_export.txt.zip

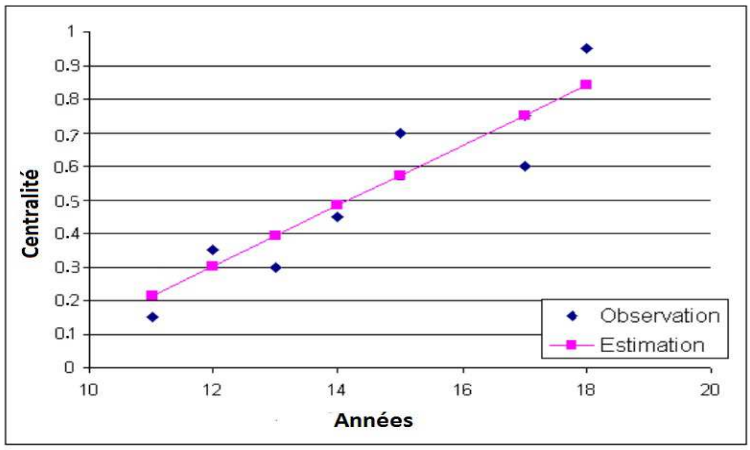

FIG. 1 – *Exemple de droite de régression linéaire d'une courbe de centralité*

Dans cette section, nous illustrons les résultats de plusieurs études menées pour évaluer l'importance des groupes de recherche à travers leur centralité dans le graphe des collaborations académiques, au fil du temps, dans la conférence considérée. Pour ces expérimentations, nous avons déduit les groupes en considérant l'affiliation commune (*i.e.*, deux auteurs appartiennent au même groupe s'ils ont signé un papier avec la même affiliation).

4.1 Extraction des affiliations et des thématiques

Les affiliations sont regroupées selon une distance de Levenshtein dont les opérations de transformation sont pondérées de manière *ad-hoc* selon le type de caractère considéré (*i.e.* la transformation d'un chiffre en un autre chiffre reçoit un poids maximum afin d'éviter le rapprochement de Paris 8 avec Paris 6). Nous considérons les thématiques comme des termes composés, que nous extrayons depuis l'ensemble des articles de la base bibliographique EGC en utilisant l'extracteur terminologique YaTea (Aubin et Hamon, 2006) afin d'extraire des thématiques candidates. Celles-ci sont ensuite validées manuellement.

4.2 Centralité des groupes dans le graphe de collaboration

Pour cette expérimentation, nous avons d'abord calculé, pour chaque groupe, la centralité de son nœud à l'intérieur du graphe de collaboration. Le résultat est présenté dans la Table 1. Nous espérons que la grande majorité des lecteurs sera d'accord avec le fait que l'approche hybride présentée dans ce papier est capable de retrouver, parmi tous les groupes détectés, ceux qui comptent le plus sur la scène scientifique de la conférence EGC. Les groupes comme celui du LIRMM, Paris-Sud ou Lyon-II représentent incontestablement des centres de recherche réputés dans le monde académique national et international. Veuillez noter que la valeur de centralité n'est pas corrélée simplement avec le nombre de publications. Par exemple, un groupe comme *Télécom ParisTech* est aussi central dans le monde collaboratif exprimé par EGC que

Équipe	Valeur de centralité	Nombre de papiers
LIRMM	438.3970	56
Université Paris-Sud	258.1097	34
Université Lumière Lyon-II	246.0740	50
Université Panthéon-Sorbonne	224.3695	46
Faculté des Sciences de Tunis	217.8984	21
Université de Lyon	178.8268	31
Orange Labs	165.4391	54
Télécom ParisTech	156.4373	16
UPMC	154.0979	20
Université Claude-Bernard Lyon 1	152.6416	18

TAB. 1 – *Classement des groupes selon leurs valeurs de centralité*

Orange Labs, même s'il semble avoir seulement un tiers de papiers publiés (16 vs 54). De même, la *Faculté des Sciences de Tunis* est plus centrale, dans la communauté EGC, que le groupe d'*Orange Labs*, même en présence d'un nombre de papiers significativement plus réduit. Ce résultat est principalement dû au fait que notre approche ne prend pas seulement en compte la pure capacité quantitative de publications mais essaie de mesurer si les autres groupes de recherche augmentent leur nombre de publications quand ils collaborent avec un groupe considéré. En ce sens-là, la capacité d'un groupe à publier avec plusieurs groupes différents tout en maintenant sa capacité d'acceptation chez EGC démontre que son travail ne dépend pas d'une collaboration particulière.

Remarquons aussi qu'en utilisant cette approche, nous explorons la manière dont chaque groupe interagit avec les autres, sur différentes thématiques, afin de détecter ceux qui ont été capables de collaborer, au fil du temps, avec d'autres groupes tout en maintenant le même rythme de production. En outre, cette approche nous permet de détecter les groupes qui paraissent les plus déterminants pour le bon développement de collaborations de recherche à l'intérieur de la communauté EGC.

4.3 Prédiction de centralité des groupes de recherche dans le futur

Dans cette section, pour un groupe G_i, nous utilisons sa courbe des valeurs de centralité dans l'intervalle de temps considéré pour estimer une prédiction de performances dans un futur proche. Pour ce faire, comme expliqué dans la section 3.5, nous calculons sa droite de régression linéaire à partir de la distribution à deux variables définie par la courbe de centralité d'un groupe et la séquence des années de publication de la conférence considérée. Nous calculons ensuite le coefficient angulaire de cette droite. Cette valeur nous permet d'estimer la croissance, en terme de centralité, du groupe de recherche. Plus la valeur du coefficient sera élevée, plus forte sera la probabilité d'avoir une haute valeur de centralité dans le futur.

Les tableaux 2 et 3 montrent les résultats à partir de deux distributions différentes. Cette expérimentation a pour but de retrouver les groupes les plus "émergents" dans la communauté de recherche. Pour ce faire, nous avons testé notre approche sur (i) les 5 dernières années et (ii) sur les 3 dernières. Les résultats montrent, encore une fois, que notre approche hybride est capable d'identifier des groupes importants au niveau national et international et de les identifier comme centraux même si le nombre de papiers publiés n'est pas forcément le plus élevé.

5 Conclusions

Le problème de l'évaluation des résultats des groupes de recherche et de leurs chercheurs a été largement étudié. Or, peu de travaux ont tenté de prendre en compte l'étude des collaborations entre les différents groupes. Nous faisons l'hypothèse que cette étude permet de mieux rendre compte de l'importance d'un groupe de recherche au sein d'une communauté. Nous avons donc proposé un modèle temporel hybride qui déduit les groupes de recherche selon plusieurs sémantiques différentes à travers la définition de relations ontologiques spécifiques, et modélise dans un graphe les collaborations entre les groupes. Ce graphe est alors utilisé pour estimer la centralité dans le temps d'un groupe par rapport à toute la communauté de recherche EGC.

Références

Aubin, S. et T. Hamon (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, Number 4139 in LNAI, pp. 380–387. Springer.

Équipe	Valeur de coefficient angulaire
Télécom Bretagne	1.1480
Orange Labs	1.1314
INRIA Rennes - Bretagne Atlantique	1.0995
UPMC	1.0934
INRIA Sophia Antipolis Méditerranée	1.0934
Télécom ParisTech	1.0636
Agrocampus-Ouest	1.0539
Université Claude-Bernard Lyon 1	1.0441
Université Panthéon-Sorbonne	1.0387
Université de Rennes 1	1.0376

TAB. 2 – *Prédiction de classement des groupes sur la base des 5 dernières années de publication EGC.*

Équipe	Coefficient angulaire
École centrale Paris	1.0349
Université François-Rabelais de Tours	1.0344
Université Lille II	1.0176
Université de Caen Basse-Normandie	1.0126
Université Montpellier 2	1.0041
SUPELEC	1.0004
Université de Cergy-Pontoise	1.0004
École Nationale d'Ingénieurs de Tunis	1.0004
Faculté des Sciences Économiques et de Gestion	1.0004
Université de Skikda	1.0004

TAB. 3 – *Prédiction des performances des groupes sur les 3 dernières années de publication EGC.*

Ausloos, M. (2013). A scientometrics law about co-authors and their ranking : the co-author core. *Scientometrics 95*(3), 895–909.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America 102*(46), 16569.

Hirsch, J. (2007). Does the h index have predictive power ? *Proceedings of the National Academy of Sciences 104*(49), 19193.

Imperial, J. et A. Rodríguez-Navarro (2007). Usefulness of Hirsch's h-index to evaluate scientific research in Spain. *Scientometrics 71*(2), 271–282.

Motik, B., R. Shearer, et I. Horrocks (2009). Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research 36*, 165–228.

Pepe, A. et M. Rodriguez (2010). An in-depth longitudinal analysis of mixing patterns in a small scientific collaboration network. *Scientometrics 85*(3).

Radicchi, F., B. Markines, et A. Vespignani (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E 80*(5), 056103.

Rodriguez, M. et A. Pepe (2008). On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics 2*(3), 195–201.

Sidiropoulos, A., D. Katsaros, et Y. Manolopoulos (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics 72*(2), 253–280.

Slone, R. M. (1996). Coauthors' contributions to major papers published in the ajr : frequency of undeserved coauthorship. *AJR. American journal of roentgenology 167*(3), 571–579.

Taramasco, C., J. Cointet, et C. Roth (2010). Academic team formation as evolving hypergraphs. *Scientometrics 85*(3), 721–740.

Van den Besselaar, P., U. Sandström, et I. Van der Weijden (2012). The independence indicator : Towards bibliometric quality indicators at the individual level.

Yan, E. et Y. Ding (2009). Applying centrality measures to impact analysis : A coauthorship network analysis. *Journal of the American Society for Information Science and Technology 60*(10), 2107–2118.

Summary

In the last decades, there is an emerging interest in finding novel methods to automatically (or semi-automatically) evaluate the work of research groups (or single researchers) to quantify their impact and the quality of their work on the surrounding research community. Such interest is due to important needs related to promotions, allocation of funds, and employment in general. In light of this, instead of focusing on a pure quantitative/qualitative evaluation of research outcomes (set of publications), we propose a novel hybrid temporal model for estimating the impact of a research group by focusing on their centrality within the scientific network formed by the contributors of a scientific conference. With this approach we aim at analyzing how each scientific group interacts with the others and detect those who result more crucial for the publishing high quality papers in the considered community. We the implemented and evaluated our model with a set of experiments on real case scenarios.