Défi EGC 2016
Vues Conceptuelles des Collaborations aux Conférences EGC depuis 2004: Une modélisation descriptive

Erick Stattner

Université des Antilles
erick.stattner@univ-ag.fr,
http://www.erickstattner.com

Résumé. Dans ce travail, nous analysons les données concernant les articles publiés à la conférence EGC. Notre objectif est d'identifier et de comprendre les tendances en matière de collaborations. Pour ce faire, nous adoptons une modélisation descriptive, à travers une approche réseau qui consiste à générer tout d'abord le réseau de collaborations des auteurs à partir des données. Nous enrichissons ensuite les noeuds de ce réseau d'une dizaine d'attributs individuels extraits à partir des données. Enfin, nous recherchons des vues conceptuelles, une approche récente de clustering de liens, qui permet de synthétiser des réseaux en mettant en évidence les ensembles d'attributs retrouvés fréquemment liés dans le réseau. Les résultats obtenus montrent les tendances existantes dans les comportements de collaborations. Dans ce papier, nous présentons ces tendances et montrons comment elles évoluent selon différents seuils d'extraction.

1 Introduction

L'étude des réseaux est devenue un des axes de recherche très actif du 21e siècle, qu'on retrouve dans la littérature comme étant la *Science des réseaux* (Barabasi et Crandall, 2003). En effet, grâce à l'amélioration des capacités de stockage et de calculs et l'hétérogénéité des données qui sont actuellement extraites de systèmes en ligne, de plus en plus de travaux se sont intéressés à des approches qui combinent plusieurs sources d'informations et qui redéfinissent les schémas traditionnels de connaissance. C'est par exemple le cas des approches de clustering de réseaux. Alors que les méthodes traditionnelles de clustering dédiées aux réseaux ont exploité uniquement la structure du réseau pour extraire des communautés (Fortunato, 2010), les approches récentes se sont, elles, intéressées à la fois à la structure du réseau et aux attributs des noeuds pour identifier de nouveaux types de groupes (Zhou et al., 2009).

Ainsi dans ce travail, nous adoptons une approche réseau, et plus particulièrement une approche de clustering, dans le but d'analyser les données qui concernent les articles publiés à la conférence internationale francophone *Extraction et Gestion des Connaissances (EGC)* de 2004 à 2015. Notre objectif est de mettre en évidence les tendances en matière de collaboration en recherchant et en identifiant des régularités cachées dans la co-écriture des articles.

Pour ce faire, nous adoptons une modélisation descriptive qui s'appuie sur une méthodologie à trois niveaux. (i) Nous générons, après pré-traitement des données, le réseau de collaborations des auteurs à partir des articles publiés. (ii) Les noeuds du réseau ainsi obtenu sont ensuite enrichis d'une dizaine d'attributs individuels qui viennent caractériser chaque auteur et qui sont générés à partir de la fréquence

des publications et de certaines propriétés des articles et des collaborateurs. (iii) Nous analysons ce réseau enrichi d'attributs pour extraire des *vues conceptuelles*, une approche récente de clustering de liens, qui permet de synthétiser des réseaux en mettant en évidence les ensembles d'attributs retrouvés fréquemment liés. Les résultats obtenus montrent qu'il existe des tendances au sein des comportements de collaborations. Dans cet article, nous présentons ces tendances et montrons comment elles évoluent selon différents seuils d'extraction.

L'article est organisé comme suit. La Section 2 présente le jeu de données utilisé ainsi que les traitements préalables effectués. La Section 3 détaille le réseau de collaboration obtenu et les attributs identifiés pour caractériser les noeuds. La Section 4 décrit les tendances identifiées après extraction des vues conceptuelles. Enfin, la Section 5 conclut l'article et présente des pistes d'orientation.

2 Jeu de données : préparation et exploration

Le jeu de données fourni représente les 1041 articles publiés à la conférence EGC sur la période s'étalant de 2004 à 2015. Il se compose d'un unique fichier texte dans lequel chaque ligne décrit un article, selon 6 attributs : (1) l'année de publication, (2) le titre de l'article, (3) le résumé (4) la liste des auteurs, (5) une url pointant vers une version pdf de la première page de l'article et (6) une url pointant vers l'article complet. Dans ce travail, nous avons choisi d'écarter les attributs (5) et (6) puisque nous n'en avons pas besoin pour l'étude que nous menons ici.

Dans le jeu de données fourni, nous observons que certains enregistrements sont incomplets, c'est-à-dire des enregistrements pour lesquels au moins un des 4 premiers attributs sus-mentionnés n'est pas renseigné. 185 enregistrements ont ainsi été identifiés comme incomplets ; nous avons choisi de les écarter de l'étude. Le jeu de données initial a donc été réduit aux 856 articles pour lesquels les 4 premiers attributs sont tous renseignés.

Dans une première étape, nous avons mené une analyse exploratoire pour comprendre comment ont évolué les publications aux conférences EGC depuis 2004. La Figure 1 montre (a) le nombre de papiers publiés par année, (b) le nombre d'auteurs distincts par année, (c) la distribution du nombre d'auteurs par papier, (d) le min, la moyenne et le max des auteurs sur un papier par année, (e) le nombre d'articles écrits par un seul auteur, et (f) le nombre moyen d'auteurs par article.

Nous pouvons tout d'abord observer qu'il existe de fortes fluctuations dans le nombre d'articles publiés chaque année (cf. Figure 1(a)). En effet, si nous notons une augmentation du nombre de papiers de 2004 à 2008, celui-ci connait une baisse à partir de 2009.

Le nombre total d'auteurs connait également quelques variations sur ces mêmes périodes (cf. Figure 1(b)). Deux tendances peuvent être distinguées : une phase de croissance de 2005 à 2009, puis une décroissance de 2011 à 2015. Ici également, l'année 2010 semble être une année qui marque un ralentissement dans l'activité de publication, puisque nous notons une baisse assez prononcée du nombre d'auteurs. Ce ralentissement est toutefois assez intéressant, puisqu'il coïncide avec le début de la crise au sein de la zone euro.

En ce qui concerne les collaborations sur les papiers, nous pouvons observer que la majorité des papiers publiés a requis la participation de plusieurs auteurs (cf. Figure 1(c)). En effet, 20% des articles sont co-écrits par deux auteurs, alors que sur 60% des articles, nous retrouvons au moins trois auteurs. Seul 10% des papiers sont, eux, écrits par un unique auteur.

Pour aller plus loin sur les collaborations, nous montrons sur la Figure 1(d) comment évoluent, chaque année, le min, la moyenne et le max des auteurs sur un papier. Nous pouvons observer un pic en matière de collaboration en 2010. En effet, on retrouve un article pour lequel une dizaine d'auteurs a collaboré. A l'exception de l'année 2010, le nombre d'auteurs maximal sur un papier est compris entre 5 et 7 quelle que soit l'année considérée.

En ce qui concerne plus particulièrement les articles publiés par un seul auteur (cf. Figure 1(e)), nous observons des différences selon les années. Pour les années 2004 et 2005, on dénombre en moyenne

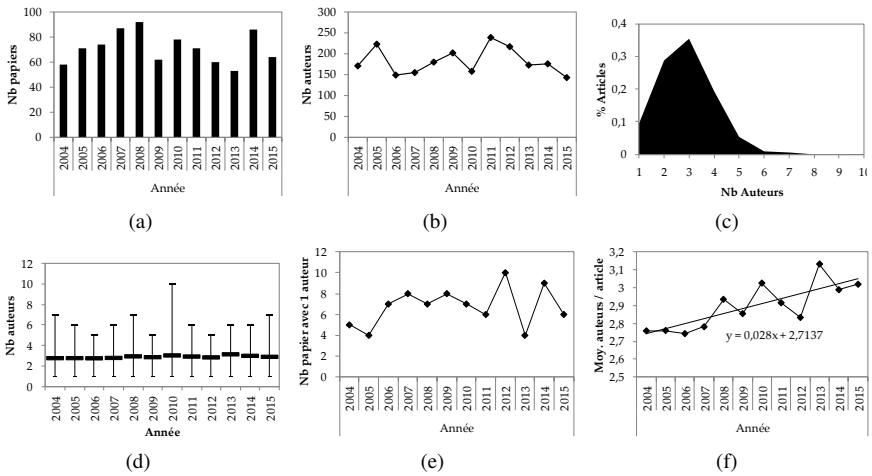

FIG. 1 – *Analyse exploratoire du jeu de données après suppression des enregistrements incomplets*

4.5 articles ne possédant qu'un seul auteur. En revanche sur la période allant de 2006 à 2015, nous observons une augmentation du nombre d'articles écrits seul ; 7.2 articles en moyenne. Enfin, si nous nous intéressons plus spécifiquement au nombre d'auteurs moyen par article (cf. Figure 1(f)), il croît de façon plus ou moins linéaire avec les années.

3 Vers une analyse réseau

La première étape de notre approche a consisté à générer le réseau sous-jacent de collaborations des auteurs à partir des articles publiés. Dans ce réseau, chaque noeud correspond à un auteur et deux noeuds sont connectés si les auteurs associés ont co-écrit au moins un article. Les auteurs d'un même article forment donc un sous-graphe complet dans le réseau de collaborations. Pour étudier cette structure, nous avons dans un premier temps étudié ces principales propriétés topologiques (cf. Figure 2).

Nous pouvons ainsi observer que le réseau obtenu est composé de 1397 noeuds et 2366 liens. La densité de liens y est donc assez faible. Le degré minimum est de 0, ce qui signifie que certains auteurs n'ont eu aucune collaboration. Le degré moyen est de $3,3$, c'est-à-dire qu'en moyenne un auteur collabore avec 3 auteurs différents, alors que le degré maximum est, lui, de 40. La distribution du degré ne suit pas une loi de puissance, mais apporte tout de même des informations utiles. En effet, 95% des auteurs ont collaboré avec moins de 10 auteurs, quand 80% ont collaboré avec moins 5 auteurs.

Le réseau n'est pas connexe, c'est-à-dire qu'il y a des noeuds pour lesquels il n'existe pas de chemin permettant de les relier. Cela laisse à penser qu'il existerait des petites communautés d'auteurs qui publient entre eux. Les résultats obtenus sur les composantes connexes tendent à le confirmer. En effet, 214 composantes ont été identifiées. En moyenne, une composante est composée de $6,5$ noeuds. Si il existe une composante géante qui regroupe $42,5\%$ des noeuds, la plupart sont de petites composantes. Par exemple, 93% des composantes regroupent moins de 10 noeuds et 80% regroupent moins de 5 noeuds.

Enfin, nous observons que le coefficient de clustering est relativement élevé. Cela traduit le fait que les noeuds connectés à un auteur tendent à être connectés eux mêmes, formant ainsi des "*triangles*" dans la structure. Ce résultat est particulièrement intéressant, puisque nous observions précédemment que la

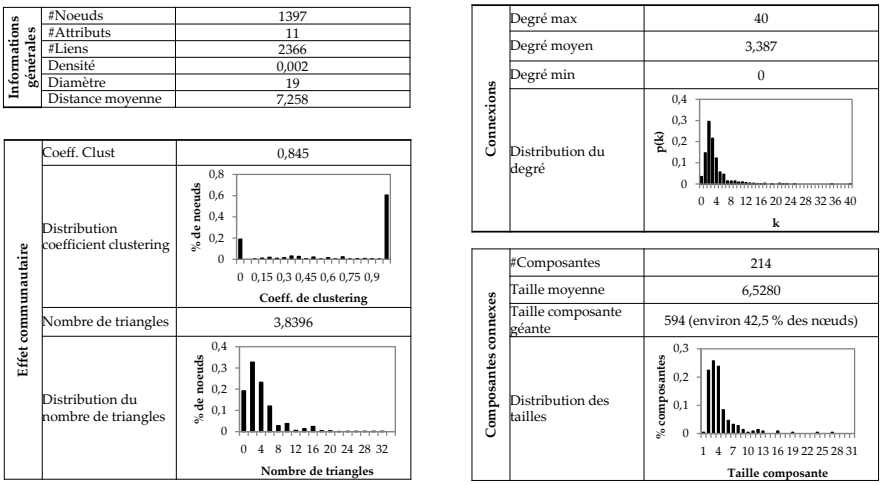

FIG. 2 – *Principales propriétés du réseau de collaborations*

plupart des noeuds sont répartis dans de petites composantes. On peut ainsi supposer que la densité de connexion est forte au sein d'une composante.

Dans une deuxième étape, nous avons ensuite analysé le jeu de données dans le but d'enrichir les noeuds d'un ensemble d'attributs individuels liés à la fois au comportement de publication et à la structure du réseau de collaborations ci-dessus présentée. Chaque noeud est ainsi caractérisé par 11 attributs individuels : (1) label, (2) nombre d'articles, (3) nombre moyen de co-auteurs par papier, (4) année où il a le plus d'articles, (5) taille moyenne des résumés, (6) nombre moyen d'articles par an, (7) position moyenne dans la liste d'auteurs, (8) nombre de fois qu'il apparait premier auteur, (9) nombre de collaborateurs, (10) coefficient de clustering et (11) nombre de triangles formés dans le réseau.

Enfin, avant de mener la phase de clustering sur le réseau enrichi des attributs, nous avons utilisé l'outil Weka (Hall et al., 2009) pour discrétiser tous les attributs (à l'exception du label) sur 5 classes ayant des tailles plus ou moins équivalentes.

4 Vues conceptuelles des collaborations

L'extraction de vues conceptuelles est une nouvelle approche de clustering qui combine à la fois la structure et réseau et les attributs des noeuds pour extraire des liens retrouvés fréquemment entre des ensembles d'attributs. Contrairement aux approches traditionnelles qui se donnent pour objectif de partitionner les noeuds du réseau, une vue conceptuelle met, elle, en lumière des groupes de liens.

Des groupes de noeuds sont d'abord crées sur la base du partage d'attributs communs. La fréquence des liens entre ces groupes est ensuite évaluée. Lorsque cette fréquence dépasse un seuil de support β, on parle de ***lien conceptuel fréquent (FCL)***, dans le sens où il s'agit d'un lien qui lit deux ensembles d'attributs, c'est-à-dire deux *concepts* dans le domaine de l'analyse de concepts formels (Ganter et al., 2005). L'ensemble des FCL est utilisé pour former une nouvelle structure de réseau appelée ***vue conceptuelle***, qui synthétise toute la connaissance extraite du réseau initial. Dans ce réseau, chaque noeud correspond à un ensemble d'attributs (on parle aussi de ***Meta-Noeud***) et un lien correspond à un FCL.

Dans ce travail, nous utilisons l'algorithme MFCLMin (Stattner et Collard, 2012) qui effectue une recherche ascendante des liens conceptuels, c'est-à-dire que les groupes de tailles t, sont utilisés pour construire ceux de taille $t+1$. Notez également que l'algorithme MFCLMin ne conserve que les groupes

maximaux, c'est-à-dire ceux qui ne sont pas inclus dans d'autres. On parle alors de liens conceptuels fréquents maximaux et nous le notons *mfcl*.

Ainsi, nous avons donc recherché les vues conceptuelles du réseau de collaborations pour comprendre s'il existait des tendances en matière de collaboration. La Figure 3 montre les vues obtenues en utilisant différents seuils de support. Par simplicité, nous notons chaque noeud avec $(k_1, k_2, ...k_{11})$ ou k_i correspond à la valeur de l'attribut (i). Lorsque $k_i = *$, cela signifie que que l'attribut peut prendre n'importe quelle valeur.

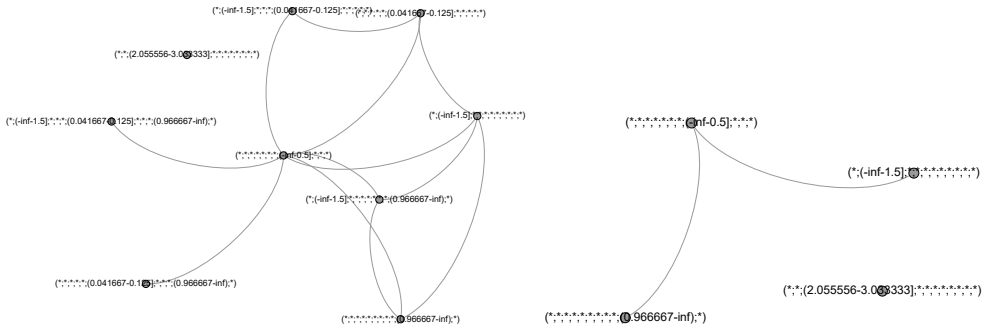

FIG. 3 – *Vues conceptuelles du réseau de collaborations : (a)* $\beta = 0.24$ *et (b)* $\beta = 0.30$

Par exemple, si on s'intéresse aux résultats obtenus avec le seuil $\beta = 0.24$, nous pouvons observer que 24% des collaborations ont eu lieu entre des auteurs dont le nombre moyen d'articles par an est compris entre 0.04 et 0.125 et des auteurs qui ne sont pas premier auteur sur les articles auxquels ils participent. Si on restreint un peu plus les groupes extraits en augmentant le seuil de support $\beta = 0.30$, nous observons par exemple que 30% des collaborations ont eu lieu entre des auteurs aux propriétés similaires, c'est-à-dire qui ont entre 2 et 3 co-auteurs.

Plus généralement, si ces structures viennent résumer l'ensemble des collaborations ayant eu lieu à EGC depuis 2004, elles nous apprennent également beaucoup sur les tendances enfouies dans ces collaborations inter-auteurs. En effet, une catégorie d'auteurs semble particulièrement centrale en matière de co-écriture d'articles ; il s'agit des individus du type (* ;* ;* ;* ;* ;* ;(-inf-0.5] ;* ;* ;*), c'est-à-dire les auteurs qui n'apparaissant premier auteur sur aucun de leur papier. Plus précisément, lorsque $\beta = 0.12$ ils interviennent dans 37 liens conceptuels, et avec $\beta = 0.24$, ils sont impliqués dans 6 liens conceptuels. Ainsi, ils semblent entretenir des liens forts avec toutes les autres catégories d'auteurs.

Une autre catégorie d'auteurs présente également des comportements intéressants ; il s'agit des auteurs ayant entre 2 et 3 publications $(*; (2.055556 - 3.033333]; *; *; *; *; *; *; *; *)$. En effet quel que soit le seuil utilisé, nous retrouvons toujours des liens très forts au sein de cette même catégorie d'auteurs. Nous pouvons décrire ce comportement comme étant "*les auteurs qui ont entre 2 et 3 articles, publient avec des auteurs qui ont eux mêmes entre 2 et 3 articles*". Ce résultat est particulièrement intéressant puisque nous avons observé que le réseau de collaboration étudié était composé de centaines de petites composantes connexes. Ainsi, au sein même de composantes éparses, nous identifions ici un schéma de collaboration qui se retrouve partout.

Ainsi, cette approche de clustering est particulièrement intéressante, Nous observons en particulier que nous retrouvons des groupes avec des seuils de support relativement élevés. Dans ce travail, des groupes sont identifiés jusqu'à un seuil de support de 30%. Cela démontre que des tendances fortes sont présentes, des motifs réguliers sous-jacents qui, bien que nous n'en soyons pas nécessairement conscients,

se mettent en place au sein des collaborations entre auteurs et structurent les comportements de co-écriture d'articles scientifiques.

5 Conclusion et perspectives

Dans ce travail, nous avons analysé les données qui concernent les articles publiés à la conférence internationale francophone EGC de 2004 à 2015. Notre objectif était de mettre en évidence les tendances en matière de collaboration en recherchant en particulier des régularités cachées dans les comportements de co-écriture d'articles scientifiques. Ainsi, les résultats obtenus dans ce travail ont permis d'identifier des tendances fortes au sein des collaborations inter-auteurs. Cela démontre qu'il existe des motifs réguliers sous-jacents qui, bien que nous n'en soyons pas conscients, se mettent en place lors des collaborations et structurent les comportements de co-écriture d'articles scientifiques.

Il serait intéressant d'aller plus loin en élargissant le jeu de données avec les informations qui peuvent, aujourd'hui, être extraites de sites en ligne tels que DBLP ou ResearchGate. Cela permettrait de confirmer les tendances observées et peut être d'en identifier de nouvelles. Il serait également intéressant de générer de nouveaux réseaux, dans lesquels les noeuds représentent un mot et le lien l'appartenance à un même article. La recherche de communautés sur de tels réseaux apporterait des éléments de compréhension supplémentaires sur les publications.

Références

Barabasi, A. et R. Crandall (2003). Linked : The new science of networks. *American journal of Physics 71*, 409.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*, 75–174.

Ganter, B., G. Stumme, et R. Wille (2005). Formal concept analysis, foundations and applications. *Lecture Notes in Computer Science 3626*.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Stattner, E. et M. Collard (2012). Social-based conceptual links : Conceptual analysis applied to social networks. *International Conference on Advances in Social Networks Analysis and Mining*.

Zhou, Y., H. Cheng, et J. Yu (2009). Graph clustering based on structural/attribute similarities. *VLDB Endowment 2*(1), 718–729.

Summary

In this article, we carried out an analysis work on the data of the articles published to the EGC conferences from 2004 to 2015. Our goal is to identify and understand the collaboration tendencies. For this purpose, we adopt a descriptive modeling, through a network approach that consist first in generating the collaboration network of authors from data provided. Then, we enrich the nodes of this network with ten individual attributes extracted from the data. Finally, we search for conceptual views, a recent link clustering approach, which allows to summarize networks by highlighting the sets of attributes found frequently linked in the network. The results show the existing trends in the behavior of collaborations. In this paper, we present these trends and show how they evolve according to different extraction levels.