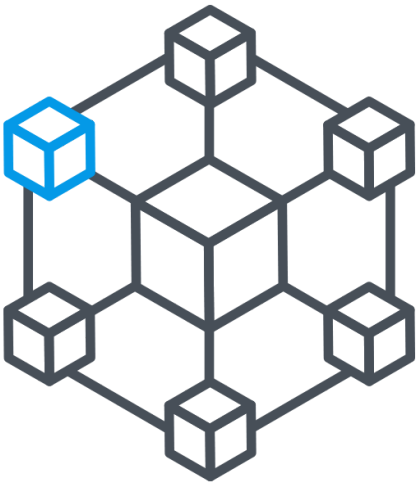
Analyse graphique des travaux de recherche autour de la blockchain



Enseignant : Rémy Cazabet

EISO Master Épistémologie et Ingénierie de la Science Ouverte

Sommaire

1) Introduction

a) Origine des données (HAL)

Dans ce rapport nous avons réalisé une analyse réseau liée au domaine de la blockchain. Le terme blockchain désigne une nouvelle technologie qui s'inscrit comme un mode de stockage et de transmission de données sous forme de blocs liés les uns aux autres et protégés contre toute modification. Les données que nous souhaitons utiliser se trouvent sur HAL, archive ouverte contenant des articles scientifiques en ligne en libre accès. On y trouve des métadonnées, comme les titres, auteurs, affiliations, résumés mais pas l'article en lui-même. HAL permet à n'importe qui de faire des requêtes, et renvoie le résultat : c'est donc à partir de ce paramètre possible que nous avons trouvé notre sujet puis établi une démarche et employé une méthodologie.

b) Sujet et démarche

Ce domaine nous a particulièrement intéressé dans la mesure où celui-ci est extrêmement jeune et nouveau. Nous avons donc décidé de l'analyser pour essayer de comprendre comment se structurait la littérature existante : Y a-t-il des disciplines plus orientées que d'autres sur ces questions ? Existe-il des réseaux d'auteurs très prolifiques dans le domaine ? Les recherches sur le domaine ont-elles connu une certaine direction en fonction des années, des zones géographiques, ou encore des domaines scientifiques ? Cette étude réseau va nous permettre d'étudier ces questions et, pourquoi pas, tenter d'y apporter une réponse. Nous commencerons donc à établir une requête avec les informations pré-sélectionnées, établir des réseaux afin de répondre à nos problématiques tout en les analysant. Enfin, nous conclurons notre étude.

c) Méthodologie

Concernant la méthodologie employée, nous avons requêté à l'aide de l'outil Google Collab, l'API de l'archive ouverte HAL (https://hal.archives-ouvertes.fr/). La requête concernait l'ensemble des publications contenant le mot "blockchain" de 2016 jusqu'en 2021.

- Informations exportées

Pour tenter de mener à bien ces objectifs, nous avons donc défini les informations dont nous aurions besoin à l'export :

— Titre
— Auteurs
— Structures
— Dates de publication
— Domaines
— Mots-clefs
— Types de documents

Ces métadonnées nous permettent d'obtenir plusieurs résultats afin de pouvoir exécuter d'autres requêtes à partir de celles-ci.

- Requête finale

La requête finale utilise l'API de HAL et se présente donc sous la forme suivante :

```
q = "title_t:blockchain"
format = "csv"
n_rows = "900"
cols =
"title_s,authFullName_s,keyword_s,publicationDateY_i,structName_s,
level0_domain_s, domainAllCode_s,docType_s"
response = requests.get("http://api.archives-ouvertes.fr/search/?q=" +
q + "&wt=" + format + "&rows=" + n_rows + "&fl=" + cols)
#response =
requests.get("http://api.archives-ouvertes.fr/search/?q=title_t:sorc*
OR keyword_t:blockchain OR
abstract_t:blockchain&wt=csv&rows=900&fl=title_s,authFullName_s,keyword
_s,publicationDateY_i,structName_s, level0_domain_s,
domainAllCode_s,docType_s")
as_csv = pd.read_csv(io.StringIO(response.text),sep=",")
as_csv.to_csv("results-blockchain.csv")
as_csv
```

2) Premiers résultats de notre analyse

a) Critique

Nous avons été confrontés à quelques problèmes lors de notre requête sur HAL. En effet, comme mentionné dans le domaine de la blockchain, le sujet est assez nouveau, ce qui ne nous a donné que peu de résultats. De plus, ces résultats ne représentent pas la période totale d'étude du domaine : il est certain qu'il existe des travaux avant 2016 ou d'autres plus importants voire précurseurs non présents dans HAL. Faire cette requête sur HAL est aussi risqué car on sait que le domaine de la blockchain n'est pas aussi étudié en France qu'au États-Unis ou en Chine par exemple. Nous nous attendions donc à avoir une représentativité assez faible de l'état de l'art général du sujet. Nous avons aussi observé, à travers les métadonnées, des biais de résultats, notamment concernant la date de publication où nous observions plusieurs dates ou encore des chiffres sans correspondance (203 par exemple). De plus, lors de notre requête, pour observer les types de travaux autour de la blockchain nous avons eu comme résultat "undefined" à 5%, ce qui interfère notre interprétation du résultat.

b) Premier graphique

Le résultat de notre première requête API de HAL nous a donné un total de 532 documents différents dans le domaine de la blockchain. La première figure représente l'évolution du nombre de publications présentes sur HAL pour notre requête par année de publication, avec la quantité de documents correspondant en fonction des années de publication. Ici, une barre du graphique représente 1 année.

Nous pouvons voir qu'entre 2016 et 2021, le nombre de publications a augmenté de façon exponentielle. Cela s'explique du fait que ce sujet est de plus en plus un objet de recherche, étant nouveau, difficile à comprendre et à fort potentiel pour l'avenir. Le vocabulaire dans le domaine de la blockchain est de plus en plus étudié, parlé et discuté. Cela peut aussi s'expliquer par l'intérêt grandissant des personnes pour le bitcoin, en tant que monnaie virtuelle et actif financier. L'intérêt pour ce champ de recherche a suscité la production de plus en plus de travaux, en parallèle à la politique de HAL d'augmenter le nombre de dépôt depuis ces dernières années.

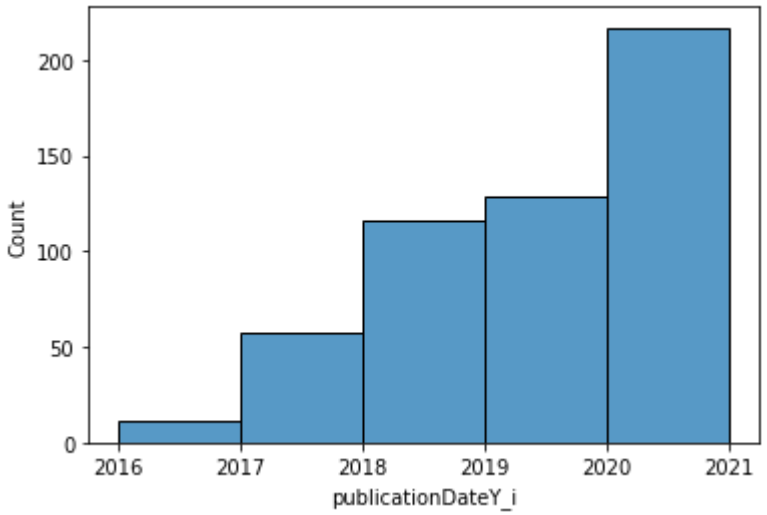


Figure 1: Évolution du nombre de publications présentes sur HAL pour notre requête par année de publication. Une barre du graphique représente 1 année.

La figure 2 représente la répartition des types de documents présents dans notre corpus. Les communications dans des congrès sont majoritaires, représentant la moitié de nos travaux à partir de nos résultats (49%), ensuite, ce sont les articles et les chapitres d'ouvrages. Il est intéressant d'observer ce type de résultat et de voir que ce sujet est énormément discuté dans des congrès et non sur d'autres supports. Cela suit la logique évoquée précédemment, du fait de sa nouveauté et qu'il accompagne souvent d'autres disciplines (que nous verrons dans un réseau ultérieurement). Cependant, il devient de plus en plus un sujet à part entière et unique lors de congrès, conférences ou séminaires.

Ce graphique nous montre ainsi que notre corpus n'est pas totalement hétérogène. Si l'on regarde le pourcentage de communications dans des congrès sur HAL qui est de 25,1%, on peut déjà remarquer que le domaine de la blockchain a trouvé son type de document lorsqu'on souhaite l'étudier.

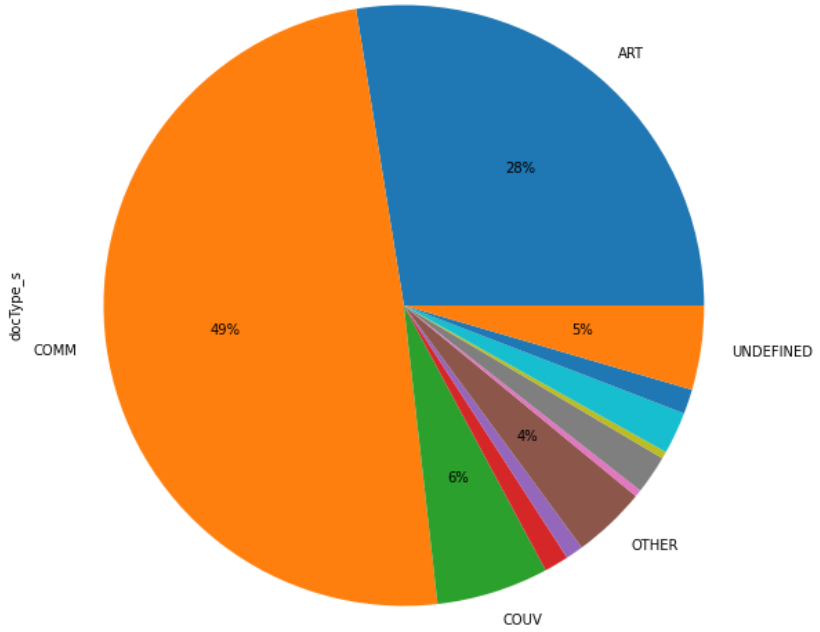


Figure 2 –
Représentation graphique de la répartition des types de documents. COMM correspond à des communication dans des congrès, ART à des articles de revues, COUV à des chapitres d'ouvrages, OTHER à d'autres types de livrables et UNDEFINED à non définis

La figure 3 représente la répartition des différentes disciplines associées aux travaux de notre requête : 74 disciplines sont représentées avec certaines plus ou moins lisibles. Ici, 26 ont au moins une occurrence de 3. Les disciplines les plus représentées sont logiquement l'informatique et les sciences humaines et sociales. Nous observons aussi des interdisciplines : en informatique avec info-dc (direct connect), info-ni, info-et, info-cr, info-ai et les SHS en droit et en gestion. Nous pouvons déjà avoir des éléments de réponse à notre problématique, en observant que 38% des travaux viennent du domaine de l'informatique et 11% en viennent des Sciences humaines et sociales..

Nous nous attendions à ce que la blockchain se retrouve dans le domaine de l'informatique, mais pas forcément dans cette proportion en SHS. Ces résultats montrent ainsi que la blockchain peut également avoir un intérêt dans le domaine des sciences humaines et sociales, dans la mesure où celle-ci tente de répondre à certains problèmes sociaux et sociétaux très complexes.

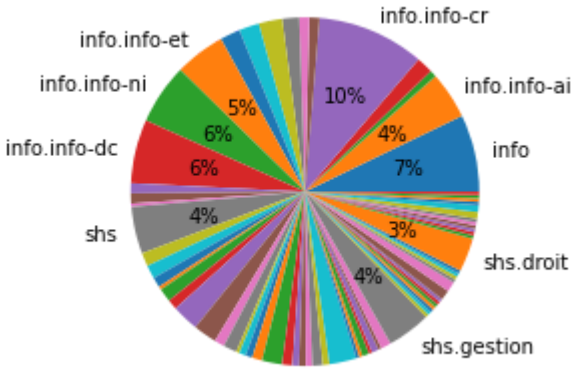


Figure 3 - Répartition des disciplines pour les travaux de notre requête. Plusieurs disciplines peuvent être indiquées par document. Par souci de lisibilité, nous n'affichons pas les proportions ni les labels pour les disciplines représentées à moins de 3 %. SHS signifie « Science de l'homme et société ».

3) Réseau de mots clés

a) Hypothèses

Il était important pour nous de commencer par l'étude de réseaux de mots clés, étant donné la richesse du sujet. Certains d'entre nous n'ont aucune connaissance dans le domaine, il était donc important d'observer le réseau autour de la blockchain et de comprendre le lien entre les différents mots-clés liés aux travaux obtenus par notre requête. L'idée est de définir le contexte autour de la blockchain et quels sujets sont les plus abordés. Nous nous faisions déjà l'idée que la blockchain était discutée autour de la sécurité, la vie privée et traçabilité, étant donné que ce système est totalement transparent, disposant d'un haut niveau de sécurité et fonctionnant sans organe central de contrôle. Cependant, nous voulions aussi voir si cela allait être abordé différemment avec des sujets totalement étrangers pour nous.

Nous avons dû trier le jeu de données en gardant seulement les mots clés qui apparaissent au moins 3 fois. Nous avions de base réduit à 1 fois mais nous obtenions des mots clés identiques à d'autres (pluriel-singulier), et qui auraient pu être regroupés.

b) Calculs

Nous avons calculé la densité et le diamètre afin de voir s' il y a beaucoup de liens, si les voisins du nœud principal (blockchain) sont connectés entre eux et si notre réseau est dense.
Diamètre = 2, la valeur maximale des distances entre les nœuds est faible.
Densité = 0,057, il y a très peu de liens entre les nœuds. En effet ils sont tous reliés à la blockchain mais pas entre eux, à part pour certains. Il y a également certains nœuds isolés dont un très dense.
Coefficient de clustering = 0,6, confirme que les voisins du noeud principal ne sont pas connectés entre eux

c) Observations

Le réseau présenté dans la figure 4 nous permet de confirmer les précédents calculs et de répondre à un élément de notre problématique.

On observe d'une part que les mots clés ont quasiment tous la même centralité autour du mot clé "Blockchain". Le nœud central de notre requête est celui qui représente la betweenness la plus élevée. De plus, on remarque qu'on ne peut pas parler de blockchain sans "bitcoin", "Ethereum", "cryptocurrency" ou les autres mots clés en lien avec lui. La blockchain est un élément central autour de la décentralisation, de la crypto-monnaie et de la vie privée. Nous observons également que ce mot clé est le seul à avoir le plus de connexion avec les autres nœuds, voire le seul. On peut voir que seulement les mots clés "security" ou "Ethereum" sont connectés entre eux, ce qui explique le coefficient de clustering du réseau très faible.

Nous avons par surprise observé plusieurs mots clés isolés dont "Blockchain Technology" en doublon, ainsi que le mot clé "NAN", signifiant "Nano Token". Pour affiner un peu plus notre travail nous aurions pu trouver une équation qui réunit les mots clés ayant le même sens comme le doublon "Blockchain technology" où la seule lettre "T" est différente. Nous aurions pu l'associer directement avec le mot clé central Blockchain.

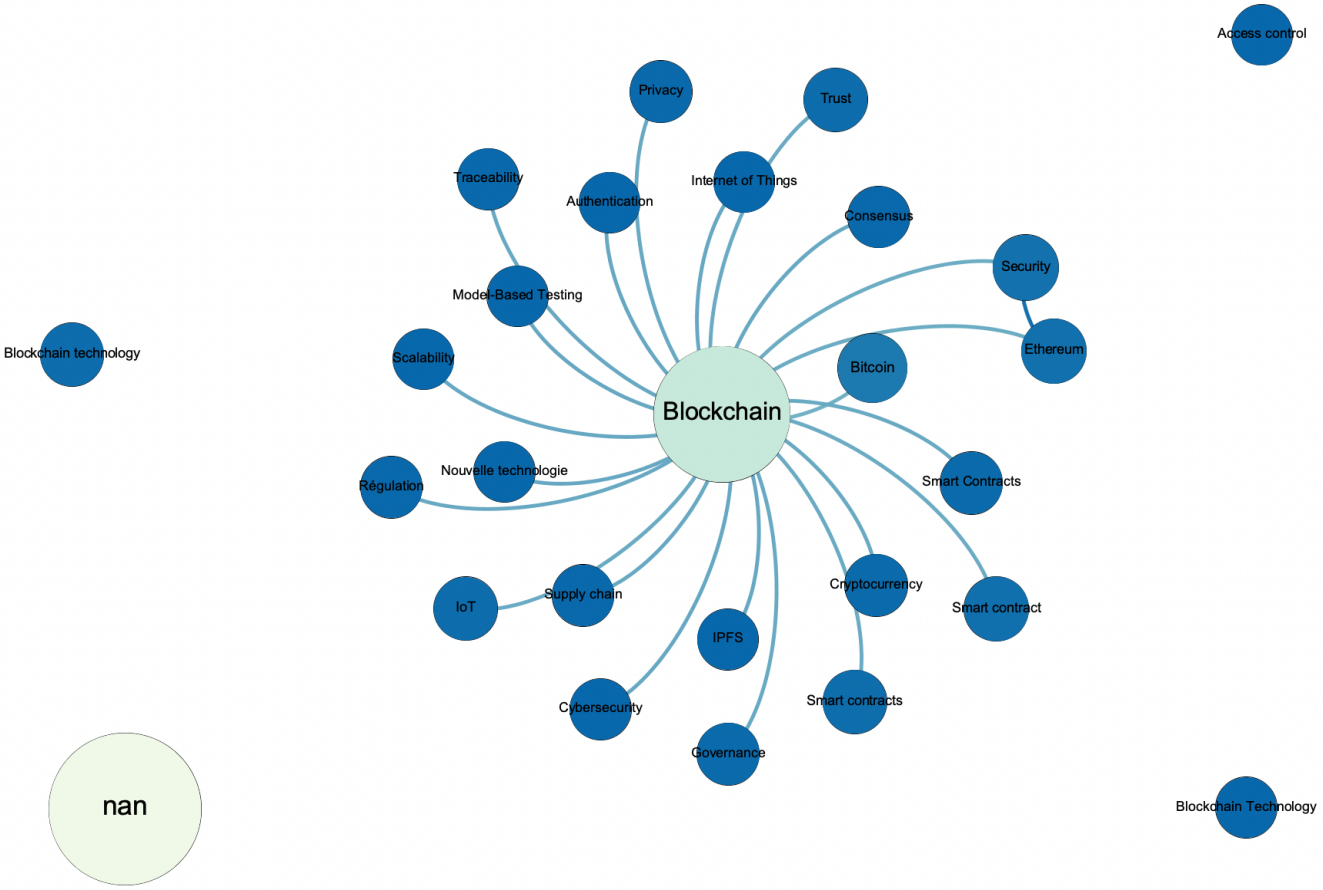


Figure 4 – Réseau des mots clefs représentant la closeness et la betweenness des différents nœuds.

4) Réseau d'auteurs

a) Hypothèses

Pour notre réseau d'auteurs, nous voulions créer un réseau et voir lesquels revenaient le plus dans les travaux dans le domaine de la blockchain. En effet nous ne savons pas quelles personnes sont spécialistes du domaine et comment elles collaborent. Ainsi, nous voulions obtenir un réseau de co-auteurs structuré et simple à analyser pour discuter sur les auteurs, et voir même le genre représenté. Il serait intéressant de voir s' il y a autant de femmes qui travaillent dans ce domaine que d'hommes par exemple.

Nous avons donc créé un réseau dont les nœuds sont les auteurs, pour voir s'ils apparaissent ensemble dans le même article. Les auteurs avec le plus de connexion et une betweenness plus forte représenteront les auteurs les plus productifs ayant déposé sur HAL. La closeness dans ce réseau montre les différentes "communautés" d'auteurs présentes autour de la blockchain

b) Calculs

Comme pour le précédent réseau, nous avons calculé le diamètre, la densité et le coefficient de clustering pour déjà permettre de nous éclaircir et de formuler des premiers éléments d'analyse.
Diamètre = 8, la valeur maximale des distances entre les nœuds est forte.
Densité = 0,004, il y a très peu de liens entre les nœuds.
Coefficient de clustering = 0,9 , les voisins des nœuds principaux sont connectés entre eux. Ces résultats concernent le premier réseau d'auteurs de co-occurrence.

c) Observations

Dans la figure 5, on remarque que le graphique obtenu avec Gephi est constitué de plusieurs îlots regroupant différents chercheurs. Un îlot principal se détache des îlots secondaires, on peut donc se représenter les collaborations entre chercheurs autour de la blockchain. Pour ce qui est de l' interprétation : la palette représente les occurrences des différents nœuds, la co-occurrence ici est la présence simultanée de deux ou de plusieurs auteurs ou dans le même article. Les auteurs avec le plus d'occurrences auront alors un nœud plus large, de façon décroissante. On voit que malgré certains gros îlots, la disparité domine assez. Le domaine étant

encore émergent cela pourrait expliquer le fait que les chercheurs ne collaborent pas encore nécessairement à grande échelle. Certains auteurs ne sont même pas connectés aux autres (en rouge).

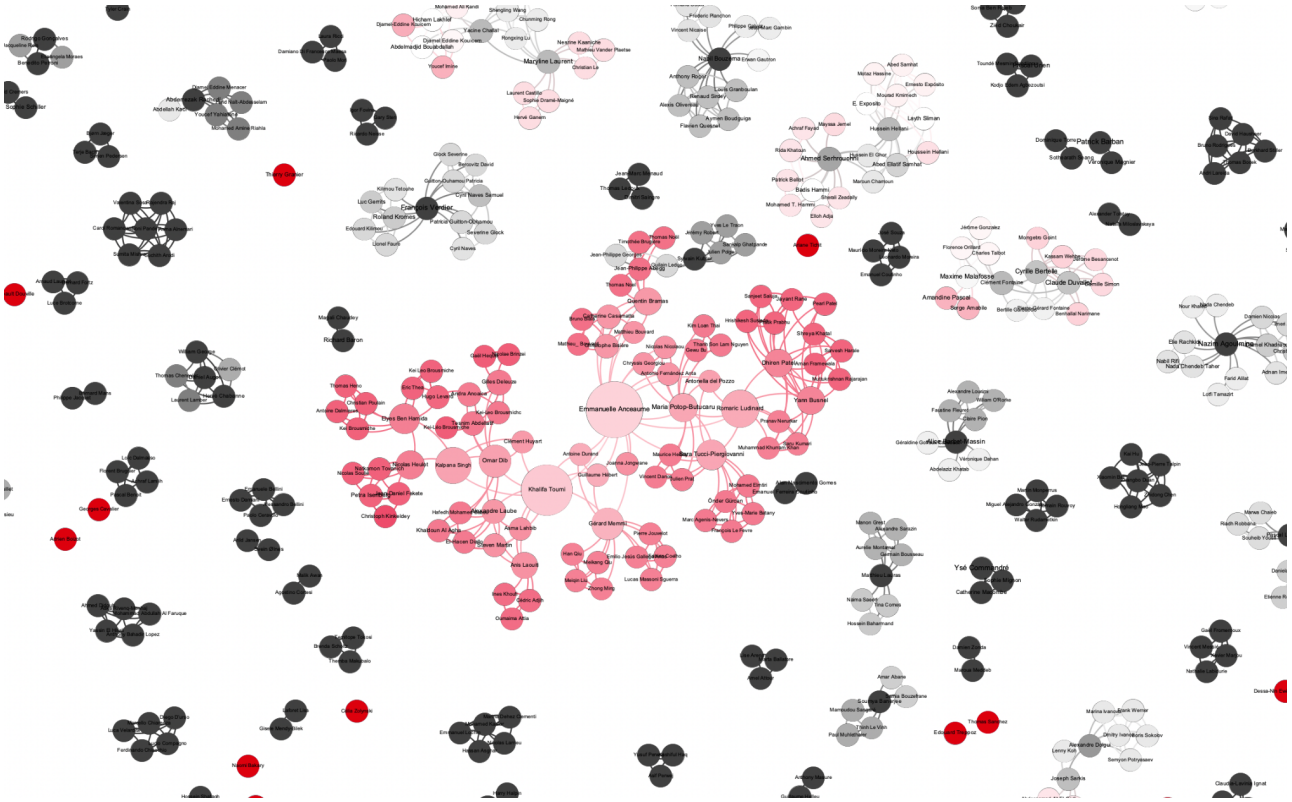


Figure 5 – Réseau des auteurs représentant la closeness et la betweenness des différents nœuds.

Si on se rapproche (figure 6), on observe un îlot principal avec plusieurs connexions. On remarque déjà des auteurs qui sont plus présents que d'autres, comme Maria Botop Butucaru, Omar Dib, Khalifa Toumi, Emmanuelle Enceaume, Elyes Ben Hamida ou encore Maryline Laurent. Ceux-ci collaborent avec de nombreux autres auteurs et créent donc leurs propres réseaux. La betweenness montre que ces auteurs agissent comme des ponts avec les autres auteurs et sont donc des nœuds plus conséquents. Plus la couleur est intense, plus la closeness est forte. On le remarque d'une part avec les différents îlots mais aussi dans l'îlot principal avec les différentes communautés, connectées aux plus grands nœuds.

En parallèle de ces observations, il est intéressant de voir que parmi les auteurs les plus prolifiques, on retrouve quasiment plus de femmes que d'hommes. Cela suggère que ce domaine n'est pas "genré", comme les mathématiques par exemple.

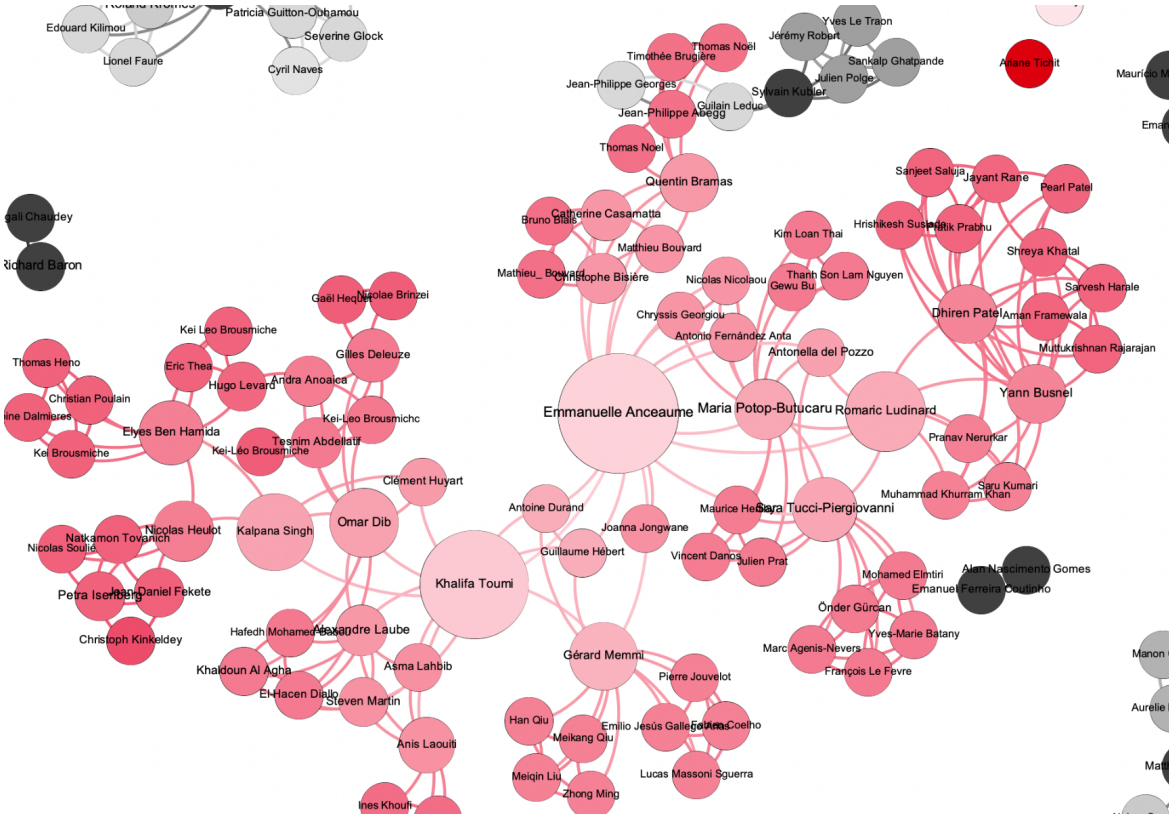


Figure 6 - Réseau de la figure 5 rapproché avec la closeness (intensité de couleur) et la betweness (taille) des différents nœuds.

d) Réseau bipartite d'auteurs avec les domaines

Après avoir réalisé le réseau d'auteurs, permettant de voir lesquels étaient les plus présents dans la blockchain sur HAL, ceux qui collaborent le plus parmi eux et les différences de genre, nous voulions ensuite observer dans quels domaines les auteurs se situaient et ainsi quelles disciplines étaient les plus présentes. Même si la figure 3 nous montre la répartition des disciplines, ainsi que leur pourcentage, il était important pour nous ici de le visualiser sur Gephi et retrouver les auteurs que nous avions identifié dans la figure 6 afin d'identifier à quelle discipline ceux-ci sont rattachés.

Nous remarquons dans la figure 7 que l'informatique est majoritaire (cf figure 3). Les SHS sont aussi bien présentes mais ne représentent pas les plus gros nœuds, plus réparties dans le réseau : le nœud "SHS" est plus petit que les interdisciplines, à l'inverse de l'informatique où on retrouve un nœud central "info" puis les autres dérivés.

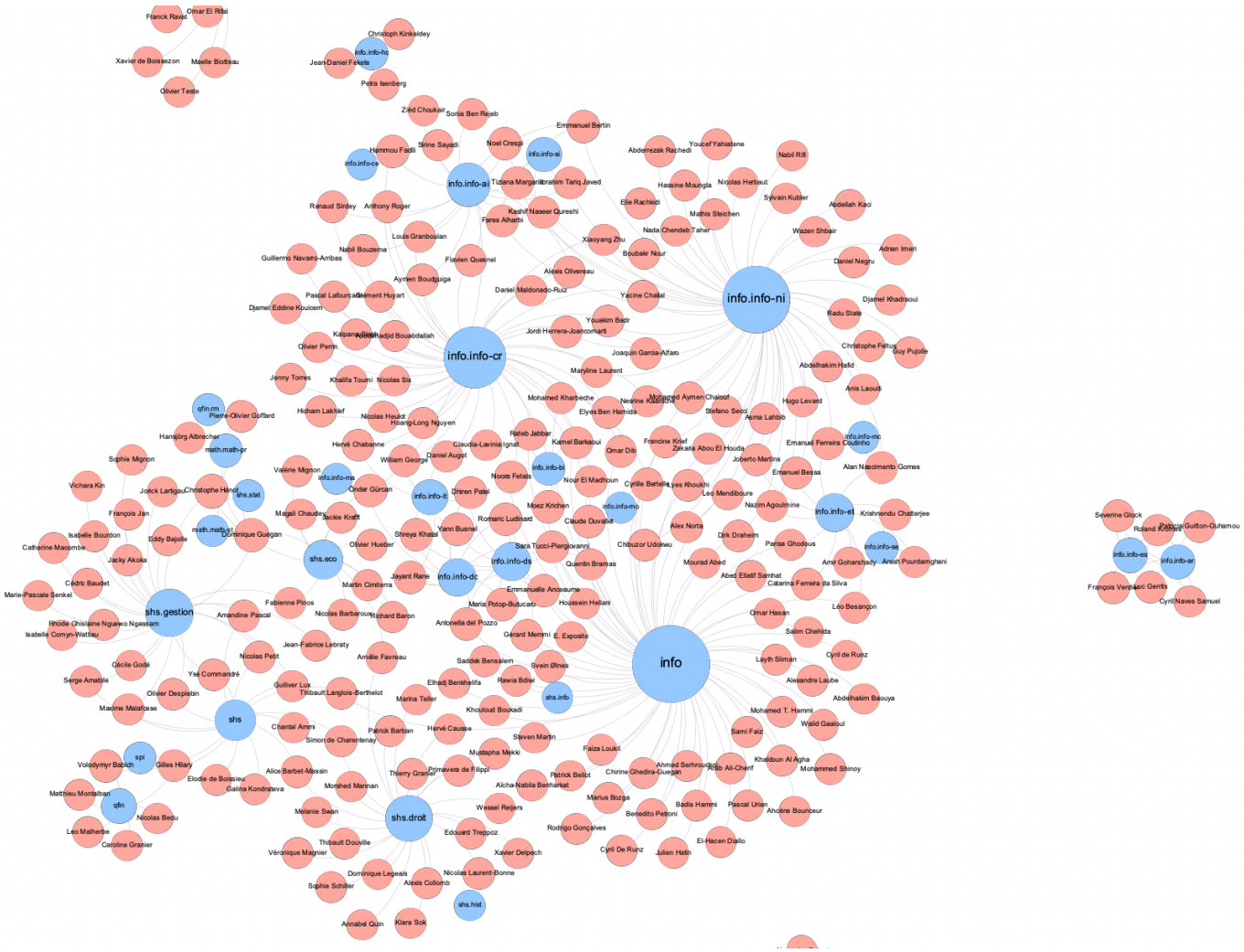


Figure 7 - Réseau bipartite auteurs/disciplines. Les couleurs permettent de repérer les deux types de nœuds (en bleu les disciplines, en rose les auteurs).

5) Conclusion

Notre travail s'est articulé autour de l'utilisation du terme Blockchain dans les titres des publications scientifiques sur l'archive ouverte HAL. Nous avons ainsi requêté et analysé les résultats avec pour objectif de mieux comprendre la répartition des différents domaines de recherches, des auteurs, ou encore l'utilisation des différents mots clés utilisés avec ce sujet. Le sujet étant encore un sujet très émergent, nous nous sommes également intéressés à la dimension temporelle de tout cela. Un gros questionnement que nous avions concernait également les domaines de recherche. Pour concevoir notre requête, nous avons récupéré tous les travaux comportant le terme "blockchain" dans le titre des publications. Nous voulions ainsi nous rendre compte de l'évolution du traitement du terme blockchain au fil des années. Nous souhaitions également nous rendre compte au mieux de l'évolution des domaines de recherche traitant de ce sujet. Nous voulions également mettre en lumière les différents auteurs qui étaient les plus prolifiques concernant ce domaine novateur et qui, pour l'heure, ne dispose pas vraiment d'expert à proprement parler.

Nous avons identifié plusieurs limites à notre analyse. Tout d'abord, il est évident que HAL est une archive parmi d'autres (française, qui plus est) et il est probable que bon nombre de publications aient échappé à notre requête. Nous avons également requêté uniquement le terme "blockchain" dans le titre des articles, ce qui a potentiellement laissé passer des publications parlant de la blockchain mais pas dans le titre. Également, nous avons eu beaucoup de mal à concevoir une requête satisfaisante. Nous voulions aller plus loin dans notre analyse mais les problèmes rencontrés ont donc été des freins dans notre analyse. Néanmoins, ce rapport nous a permis de nous familiariser avec Gephi et de comprendre les termes vu en cours. Il nous a également apporté des éléments de réponses au sujet de la blockchain, avec la connaissance de nouveaux termes, d'auteurs et de champs disciplinaires.