Exploration, navigation et visualisation des réseaux multi-couches à travers les sciences humaines et sociales

Antoine Laumond

LaBRI

université de BORDEAUX

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par **LAUMOND Antoine**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Exploration, navigation et visualisation des réseaux multi-couches à travers les sciences humaines et sociales

---

Date de soutenance : 30/06/2020

Devant la commission d'examen composée de :

Guy Melançon	Pr, LaBRI, Univ. Bordeaux	Co-directeur
Bruno Pinaud	MCF HDR, LaBRI, Univ. Bordeaux ..	Co-directeur
Xavier Blanc	Pr, LaBRI, Univ. Bordeaux	Président du jury
Florence Sedes	Pr, IRIT, Toulouse	Rapporteuse
Cyril de Runz	MCF HDR, LIFAT, IUT de Blois	Rapporteur
Christine Largeron .	Pr, Hubert Curien, Univ. Jean Monnet	Examinatrice
Mohammad Ghoniem	Ch.R, LIST, Luxembourg	Examinateur

Laboratoire d'accueil Laboratoire Bordelais de Recherche en Informatique (LaBRI, UMR 5800)

Keywords multilayer, network, human and social sciences, exploration, navigation, visualisation, digital humanities

Mots-clés multi-couche, réseau, sciences humaines et sociales, exploration, navigation, visualisation, humanités numériques

Abstract Nowadays networks are becoming more and more common subjects of study in many fields : from online social networks to criminal networks and linked collections of documents. Among these networks, a particular type of network, called "multi-layer networks", are composed of several sets of elements of distinct types ("the layers"). These objects are commonly encountered in the humanities and social sciences research fields but can be difficult to exploit because of their semantic complexity. To this end, we present M-QuBEEE, a method to explore multi-layer networks by successive and evolutive subnetwork extractions. Specifically adapted to the methodology of experts in the human and social sciences, M-QuBEEE determines the pertinence of each element of the network in order to propose a partial view relevant to the users. Users can then continue to interact iteratively on these sub-networks to improve their views or explore new directions.

Résumé De nos jours les réseaux deviennent des sujets d'étude de plus en plus répandus dans de nombreux domaines : des réseaux sociaux en ligne jusqu'aux réseaux criminels en passant par des corpus de documents liés. Parmi ces réseaux existent ce que l'on appelle des "réseaux multi-couches", des réseaux se décomposant en plusieurs ensembles d'éléments de types différents ("les couches"). Ces objets sont communément rencontrés en sciences humaines et sociales mais peuvent être difficiles à exploiter à cause de leur complexité sémantique. Nous proposons alors M-QuBEEE, une méthode d'exploration par extractions successives et évolutives de sous-réseaux. Spécialement adaptée à la méthodologie des experts des sciences humaines et sociales, M-QuBEEE détermine l'intérêt de chaque élément du réseau pour l'utilisateur afin de lui proposer une vue partielle pertinente. Celui-ci peut ensuite continuer à interagir itérativement sur ces sous-réseaux pour améliorer ses vues ou explorer de nouvelles pistes.

Remerciements

Malgré les conventions, je vais remercier en premier ma compagne Marie qui a sans aucun doute rendu tout cela possible par son soutien indéfectible et absolu depuis non seulement le début de ma thèse mais depuis même bien avant le tout début de mes études dans le supérieur ! Son aide est inquantifiable.

Je remercie évidemment Bruno Pinaud et Guy Melançon, mes directeurs de thèse, pour qui j'ai été successivement stagiaire, ingénieur, doctorant et ATER et qui m'ont sûrement apporté plus que tous les enseignants rencontrés dans mon chemin universitaire et scolaire réunis. De même, je remercie aussi Mohammad Ghoniem, mon co-encadrant, qui m'a accueilli chaleureusement et accompagné pour chaque mission au Luxembourg.

Je remercie également mon jury, qui a pu se rendre disponible malgré la pandémie afin de se pencher sur mes travaux et de rendre possible ma soutenance : Xavier Blanc (président du jury), Florence Sèdes et Cyril de Runz (rapporteur·e·s) ainsi que Christine Largeron (examinatrice).

Cette thèse a été financée par le projet BLIZAAR (ANR/FNR) et a donc pris place au sein d'une coopération entre plusieurs équipes. Je remercie les historiens du CVCE et particulièrement Marten During qui, grâce à nos différents échanges, a aidé à dessiner précisément les contours de ces travaux. Je remercie également les membres du LIST et de l'EISTI dont Maria Malek, Sébastien Ruffiange, Stefan Bornhofen, Simone Zorzan, Fintan McGee et particulièrement Mickael Stefas (qui m'a fait connaître entre autre les meilleurs restaurants d'Esch-sur-Alzette !).

Je remercie Bénédicte Lavaud-Legendre et Mathieu Noucher pour m'avoir offert l'opportunité de travailler sur mes premiers projets de recherche qui, à terme, m'ont permis d'aiguiller mes travaux postérieurs de thèse.

Je remercie le LaBRI, mon laboratoire d'accueil, et les membres de l'équipe Bkb (ex-MaBioVIS) pour m'avoir offert un cadre de travail serein. Je remercie encore plus particulièrement mes collègues doctorants et ingénieurs : Joris, Claire, Aaron, Gaëlle, Fred, Antoine, Alexandre, Arnaud ainsi que le duo Norbert et Jason qui m'ont traîné jusqu'au Grand Canyon d'Arizona.

Je remercie spécialement Kubilaï, mon chien, qui a donné son nom à ma méthode (M-Qube^3 $\rightarrow$ M-Kubi) et mes amis pour m'avoir permis de descendre en pression les fois où c'était nécessaire (Damien, Seb, Math, Guillaume, Bruno, Tom, Jordan, Romain, Greg). Enfin, je remercie ma famille qui m'a toujours offert un sanctuaire où se replier en cas de pépin : que cela soit mes parents qui ont tout fait pour que cela soit possible ou mes deux soeurs qui m'ont montré l'exemple depuis toujours.

Table des matières

Table des figures

Chapitre 1

Introduction

Sommaire

Ces dernières années, les avancées technologiques et informatiques se sont démocratisées jusqu'à concerner presque tous les pans de l'activité humaine. Ce qui était auparavant l'apanage d'un domaine particulier est aujourd'hui ré-utilisé naturellement dans un tout autre contexte : la première automobile "moderne", destinée à transporter les canons français [51], est aujourd'hui utilisée par le monde entier ; les machines à calculer et les ordinateurs d'hier, jadis réservés au monde scientifique, sont dans les poches de chacun ; Internet, initialement destiné à la défense américaine [60], est finalement utilisé par tout le monde. Il est aujourd'hui difficile d'isoler un savoir à un domaine restreint car chaque domaine de l'activité humaine bénéficie à se connecter avec les autres. C'est aussi le cas des différents domaines de la science et, parmi eux, la science des données fait figure d'élève modèle plus qu'aucune autre.

Presque par définition, la science des données (consistant à traiter, analyser et visualiser des jeux de données) se lie aux autres sciences. Tout domaine scientifique, quel qu'il soit, nécessite d'exploiter de l'information. La science des données intervient alors pour aider à la manipuler, l'analyser et la visualiser. C'est notamment le cas dans le cadre de ces travaux proposant de joindre science des données et sciences humaines et sociales.

1.1 Contexte : Sociologie et données numériques

Ces travaux de thèse s'inscrivent dans la lignée d'une succession de partenariats avec différents experts de différents champs des sciences humaines et sociales. Cet ensemble d'expériences, malgré des contextes et des objectifs variés, est lié par une problématique commune qui s'est dessinée à travers les scénarios que nous avons rencontrés. Qu'il s'agisse de la présentation d'une visualisation brute d'un jeu de données géographiques, de la conception de la base de données d'un réseau criminel ou du développement d'une méthode d'exploration de données historiques, chacun de ces jeux de données avait pour particularité d'être modélisable sous la forme d'une structure particulière appelée réseau multi-couche (Chapitre 2).

Dans ces objets, les types ou attributs des sommets ou des liens déterminent ce que l'on appelle des couches. Riches sémantiquement, les méthodes couramment utilisées en visualisation des données ne permettent pas aux experts de tirer entièrement partie de ces objets et des informations qu'ils contiennent. Le plus souvent des compromis sont faits et on analyse alors le réseau sans mettre l'accent sur son caractère multi-couche. Par exemple, la visualisation se concentre sur les interactions entre seulement deux types de données, génère des vues en agrégeant une partie des informations pour mettre en valeur des interactions pré-déterminées ou mène directement un processus de visualisation standard comme s'il s'agissait d'un graphe simple.

Si ces procédés sont entièrement fonctionnels dans bien des domaines, ce n'est pas nécessairement le cas pour les sciences humaines et sociales, domaine se focalisant tout particulièrement sur les individus [9]. Certaines études peuvent d'ailleurs grandement se complexifier lorsqu'elles sont réalisées à l'échelle d'un groupe [74]. Cela impacte évidemment les méthodologies utilisées et induit d'adapter les méthodes traditionnellement utilisées en science des données, ce que nous verrons dans la suite. Par ailleurs, les graphes multi-couches ont souvent dans la pratique une proximité marquée avec le concept du "Big Data" dans les sciences humaines et sociales. Le "Big Data" n'a pas de définition faisant consensus dans la communauté scientifique mais une définition régulièrement admise est celle des trois "v" : volume, variété et vélocité [43]. Si la notion de volume de données est souvent celle qui retient l'attention (le qualificatif de "Big" aidant), les notions de variété (traiter des données hétérogènes i.e. de types différents) et de vélocité (la contrainte d'effectuer des traitements en temps acceptables) ne sont pas moins importantes. La richesse sémantique des graphes multi-couches fait directement écho à la variété et au volume et le besoin d'explorer de tels objets contraint à considérer la question de la vé-

locité. La numérisation de notre société et l'océan de données numériques qui en découle sont les causes directes d'une intersection croissante entre le "Big Data" et les sciences humaines et sociales. Cette évolution du domaine social ne laisse d'ailleurs pas indifférent ses différents spécialistes qui assistent et constatent les changements déjà présents [10,11,40] et essayent d'en deviner les conséquences à long termes [15, 53] parfois même à un niveau directement politique voir philosophique [5] (ici en citant même André Malraux :"Le XXIème siècle sera mystique ou ne sera pas."). À l'optimisme d'une partie des experts répond alors un scepticisme voir un rejet de l'autre camp, ce dont parle notamment Etienne Ollion à travers sa conférence au collège de France de 2014 (https://www.college-de-france.fr/site/pierre-michel-menger/symposium-2014-06-02-16h30.htm) ainsi que plusieurs de ses publications [63, 64].

Dans la liste des défauts imputés, il cite notamment une qualité des données souvent insuffisante, un problème d'alignement pouvant générer des biais entre l'objectif initial lors de la création des données et l'objectif final lors de l'analyse, un positivisme latent autour du "Big Data" ou le problème concernant la vie privée des usagers dont les données ont été récoltées. A cette liste déjà conséquente, il reproche au "Big Data" de générer auprès de ses enthousiastes la volonté d'"aller chercher l'information là où elle est disponible plutôt que d'avoir l'inventivité d'aller chercher des moyens de trouver une réponse à sa question" en ré-utilisant la métaphore de l'ivrogne qui cherche ses clés non pas à l'endroit où il les a perdues mais sous la lumière du réverbère. Il dénonce ainsi une possibilité de fascination empirique qui pourrait induire en erreur les experts.

Ollion énonce aussi un certain nombre de promesses [63] qui selon lui n'ont pas abouti : d'un point de vue empirique l'augmentation du volume de données disponible n'est pas corrélable avec leur utilité au niveau de la recherche, d'un point de vue méthodologique la quantité disponible ne permet pas nécessairement d'études intégrales ("on dispose de plus de données mais pas de toutes") et d'un point de vue théorique les attentes de découvertes de nouvelles lois sociales sont restées "lettre morte".

Après ce constat amère, Ollion déclare néanmoins que ce débat dans les sciences sociales n'est pas celui du "Big Data" mais celui de la multiplication des données numériques. Provoquant un lourd changement des sciences humaines et sociales de manière interne (en mettant à disposition une profusion de données riches) et externe (en influant des corps de métiers qui ne seraient sinon pas entrés en contact avec les sciences sociales), la multiplication des données induit une nécessité de développer de nouvelles compétences dont de nouveaux traitements efficaces des

données et une compréhension accrue de leur production.

C'est ce que nous proposons à travers les travaux réalisés dans cette thèse en proposant une méthode spécialisée pour les sciences sociales afin de visualiser et d'explorer les larges champs de données que peuvent représenter les réseaux multi-couches. Si E. Ollion insiste sur le manquement aux "promesses" du "Big Data", il s'agit essentiellement d'attentes placées par les experts sur ce domaine et qui n'ont pas trouvé satisfaction depuis. C'est le signe supplémentaire qu'il reste en effet de nombreux travaux à réaliser en science humaine et sociale pour apprivoiser ces jeux de données. Empiriquement, tous les cas que nous avons rencontrés exploitant de larges jeux de données n'étaient pas un choix mais une nécessité. La question n'est donc pas de juger ici la légitimité du "Big Data" mais de trouver un moyen de répondre à nos objectifs sans être freiné ou bloqué par celui-ci. En outre, les sciences humaines et sociales contraignent à infuser leur méthodologie propre aux méthodes traditionnellement utilisées en science des données. C'est donc à la frontière de deux domaines que ces travaux ont été réalisés. Dans la suite, nous détaillons d'une part les contributions scientifiques de la thèse puis le plan de ce manuscrit.

1.2 Contributions

En partenariat avec plusieurs experts des sciences humaines et sociales, nous avons conceptualisé et implémenté une méthode répondant au leitmotiv constaté entre leurs différentes problématiques : un besoin d'explorer de larges réseaux dont l'observation et l'analyse doivent s'effectuer au niveau de l'individu et de ses relations plutôt qu'à l'échelle de la population ou du graphe.

Nos contributions à travers ces travaux sont les suivantes :

- Un mécanisme d'exploration incrémental, `M-QuBE`^{3}, permettant d'améliorer graduellement la pertinence des visualisations en fonction d'un score d'intérêt déterminé par les actions de l'utilisateur.
- Une méthode itérative, **eScore**, permettant de calculer l'intérêt de l'utilisateur pour les différents sommets d'un réseau à travers un mécanisme de sélections successives spécialisé pour les graphes multi-couches.
- La plateforme BLIZAAR, la combinaison des deux précédentes contributions, implémentée et validée par des retours utilisateurs d'experts des données.

Dans les chapitres suivants, nous justifions en outre les choix effectués lors de la conception de nos méthodes à la lumière des caractéristiques et méthodes propres

aux sciences humaines et sociales. La structure de la thèse est définie autour de ces contributions selon le plan détaillé ci-dessous.

1.3 Structure du manuscrit

Cette thèse est structurée ainsi :

- **Chapitre 2, Problématique et données** : Dans ce chapitre, nous détaillons les problématiques, les objectifs et les données des experts en sciences humaines et sociales qui ont mené au développement de nos méthodes. Qu'il s'agisse des travaux menés avec nos collègues géographes, juristes ou historiens, les données exploitées étaient toujours modélisables sous-forme de graphes multi-couches. Si nos anciens projets nous ont fait prendre conscience de la nécessité d'adaptation issue du caractère multi-couche, notre partenariat actuel avec les historiens du CVCE (Centre Virtuel de Connaissance sur l'Europe) nous a permis de mettre au point une méthode spécialisée pour répondre à cette problématique.
- **Chapitre 3, Visualisation et sciences humaines et sociales** : Dans ce chapitre, nous réalisons dans un premier temps un tour d'horizon de la visualisation puis de ses applications dans le domaine des sciences humaines et sociales. A la lumière des spécificités de ce domaine et ce qu'elles induisent sur la visualisation, nous explicitons alors les choix de conception effectués pour le développement de notre méthode, qui est décrite en détail dans le chapitre suivant.
- **Chapitre 4, Exploration et Navigation** : Dans ce chapitre, nous décrivons `M-QuBE`^{3} (**M**ultilayer network : **Qu**erying **B**ig networks by **E**volutive **E**xtraction and **E**xploration), la méthode mise au point pour explorer et naviguer dans les réseaux multi-couches issus des sciences humaines et sociales en fonction d'une estimation de l'intérêt. Cette méthode d'exploration fonctionne par vues partielles itérativement générées où chaque nouvelle vue se base sur les précédentes pour gagner en pertinence grâce aux interactions de l'utilisateur. Dans le chapitre suivant, nous proposons une métrique d'estimation d'intérêt dédiée spécialement aux graphes multi-couches et exploitant le caractère itératif particulier de `M-QuBE`^{3}.
- **Chapitre 5, Estimation d'intérêt** : Dans ce chapitre, nous expliquons le fonctionnement d'eScore, un calcul d'intérêt adapté aux réseaux multi-couches et aux sciences humaines et sociales. `M-QuBE`^{3} étant itératif, le calcul

de l'intérêt a la possibilité d'évoluer en fonction de ces itérations afin d'améliorer graduellement la pertinence des vues partielles générées. En combinant M-QuBE^3 et eScore, nous obtenons alors un processus d'exploration complet, itératif et interactif, dépendant des choix et des actions de l'utilisateur dont l'implémentation et l'évaluation sont détaillées dans le chapitre suivant.

- **Chapitre 6, Implémentation et Validation** : Dans ce chapitre, nous décrivons l'implémentation de nos méthodes à travers le développement d'une plateforme en ligne ainsi que le processus de validation effectué auprès des différents experts des données. Nous détaillons le retour utilisateur obtenu et discutons des améliorations et fonctionnalités supplémentaires qui pourraient être efficaces pour mettre en valeur nos méthodes.
- **Chapitre 7, Conclusion et perspectives** : Enfin, ce manuscrit se conclut sur la synthèse de nos travaux, les démarches à effectuer pour compléter la procédure d'évaluation ainsi que deux axes d'amélioration potentiels : accroître la confiance de l'utilisateur et faciliter la configuration et la prise en main de nos méthodes.

Chapitre 2

Objectifs et données

Les travaux de ce manuscrit ont été réalisés dans le cadre du projet BLIZAAR, un projet collaboratif international issu d'un partenariat entre la France et Luxembourg financé par l'ANR et le FNR (https://blizaar.list.lu/doku.php). Regroupant plusieurs acteurs provenant des domaines de l'informatique, des humanités numériques (voir section suivante) et de la biologie, il a pour objectif l'analyse et la compréhension des ensembles de données modélisables par des réseaux multi-couches ainsi que la création de nouvelles représentations interactives pour les réseaux multi-couches à travers la combinaison ou la création de nouvelles méthodes de visualisation. Les travaux réalisés dans ce manuscrit ont été effectués en étroite collaboration avec des experts des données à travers des discussions régulières afin de comprendre leur domaine, leurs problématiques et les solutions à mettre en oeuvre pour y répondre.

Dans ce chapitre, nous détaillons le domaine et l'activité de nos experts des données afin de pouvoir expliquer leurs objectifs (Section 2.1). Nous décrivons ensuite les données des experts et leur modélisation (Section 2.2) ainsi que les différentes similitudes rencontrées avec nos expériences passées (Section 2.3) afin de conclure sur les questions relatives aux contraintes et spécificités qui trouvent réponses dans les chapitres suivants de ce manuscrit (Section 2.4).

2.1 Contexte et objectifs

Les travaux de ce manuscrit ont été réalisés conjointement avec les historiens du C2DH (Centre for Contemporary and Digital History), oeuvrant dans le domaine de recherche des humanités numériques (Digital Humanities). La défini-

tion de ce domaine, à l'instar du "Big Data", peine à faire consensus parmi ses membres. Un historien, Fred Gibbs, écrit même en introduction d'un de ses travaux : "S'il y a deux choses dont le milieu universitaire n'a pas besoin, ce sont un autre livre sur Darwin et un autre post de blog sur la définition des humanités numériques." [32]. On peut cependant le définir d'une manière générale par "l'application du « savoir-faire des technologies de l'information [et de l'informatique/infosciences] aux questions de sciences humaines et sociales»" (wikipédia : https://fr.wikipedia.org/wiki/Humanités_numériques). L'un des rôles majeurs attribués aux humanités numériques est celui de la préservation et la valorisation du patrimoine culturel numérique : il consiste en la collection, la préservation, l'analyse et la diffusion au plus grand nombre des oeuvres numériques fabriquées par l'homme au cours du temps et en lien avec notre patrimoine culturel et historique. C'est ce que proposent les historiens avec le CVCE (Centre Virtuel de la Connaissance sur l'Europe, https://www.cvce.eu/), une infrastructure de recherche exploitant et mettant à disposition des milliers de documents de tout genre : image (photos, caricatures, affiches, etc.), vidéo (films d'archive, séminaires, interviews, etc.), son (discours, témoignages, interviews, etc.) et texte (articles, rapports, lettres, etc.) dont l'objectif est de pouvoir retracer le processus de la construction européenne. Ces données sont presque intégralement consultables sur le site du CVCE.

À partir de ce vaste corpus de documents, les historiens du CVCE composent des "ePublications", des articles en ligne sur des thématiques ou périodes historiques précises comprenant d'une part un texte descriptif ou analytique du sujet et d'autre part un ensemble de documents utilisés comme sources et/ou comme compléments d'information (https://www.cvce.eu/epublications). La conception de ces ePublications nécessite cependant de trouver les sources nécessaires. Ceci va donc être l'objectif phare des experts des données : **comment naviguer dans un vaste corpus de documents hétérogènes afin de constituer une ensemble bibliographique satisfaisant ?** Le terme satisfaisant est important car à cet objectif s'ajoutent des contraintes afin de considérer pertinent l'ensemble bibliographique : les documents utilisés doivent couvrir les éléments majeurs reconnus dans le domaine mais les documents plus marginaux (i.e. moins connus/utilisés) sont aussi valorisés et peuvent être considérés comme très intéressants. En plus de cela, une seconde contrainte est l'homogénéité de type dans l'ensemble bibliographique : la sélection des documents doit couvrir de manière équitable, dans la mesure du possible, les différents types de document existants i.e. avoir une représentation non déséquilibrée des différents types dans l'ensemble bibliographique.

Outre la conception d'ePublications, un autre objectif assez similaire est d'offrir à des utilisateurs tiers (visiteurs du site du CVCE par exemple) la possibilité de naviguer simplement et intuitivement dans le corpus afin de prendre connaissance des différents thèmes et documents majeurs qui s'y trouvent avec, encore une fois, la nécessité de pouvoir présenter des documents marginaux ou en marge si certaines pistes ou thèmes sont approfondis.

Notre but va donc être de proposer une méthode permettant de répondre simultanément aux différents objectifs des experts en offrant la possibilité de naviguer dans le corpus et de l'explorer tout en facilitant aux experts le respect des contraintes s'imposant à la conception de la base bibliographique. Dans la section suivante, nous décrivons nos données et le système complexe qu'elles forment, modélisable par un réseau multi-couche, puis les différentes expériences rencontrées relatives à ces objets dans d'autres domaines des sciences humaines et sociales.

2.2 Données et modélisation

Comme vu précédemment, l'essentiel des données utilisées par les experts correspond à un vaste corpus de documents de types variés. En plus de cela, des fiches de méta-données renseignent diverses informations dont la nature varie en fonction des types de documents (le nom des auteurs pour un écrit ou une illustration, des informations géographiques pour un lieu ou un rapport de meeting, des informations temporelles, etc.). Ces informations ne sont pas renseignées systématiquement ainsi une majorité de documents ne sont par exemple pas datés.

A partir de ces informations, une base de données a été générée référençant non seulement les ePublications et les documents du corpus mais aussi les différentes "entités" liées à ce corpus : personnes, lieux, institutions, organisations ou groupes sociaux. Pour cela, ces "entités" sont extraites des documents en utilisant essentiellement un calcul de co-occurence des mots basé sur le coefficient de Jaccard [37] au sein des documents ainsi qu'à partir des méta-données disponibles. DBpedia (https://wiki.dbpedia.org/) est aussi utilisé afin d'enrichir certaines entités lorsqu'un parallèle est possible entre l'entité et DBpedia. En plus de cela, les relations entre les différents documents et entités sont conservées : un lien est présent entre une entité et un document dans lequel elle apparaît, entre deux entités apparaissant au sein du même document, entre une ePublications et les documents sur lesquels elle s'appuie, entre deux documents partageant un même document, etc. (voir Fig. 2.1). Il en résulte un réseau composé de 51798 sommets (documents, entités et ePublications) et 1 074 643 liens.

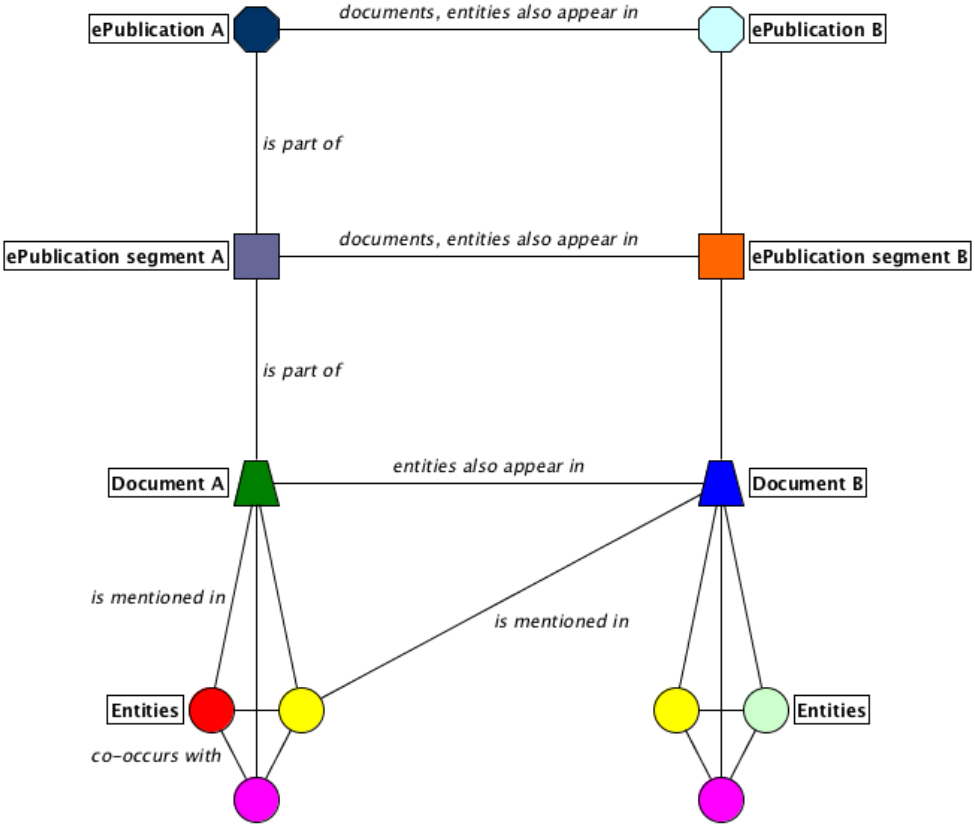


FIGURE *2.1* – ***Structure des données*** *: Les ePublications, documents et entités sont reliés entre eux. La sémantique des liens varie en fonction des types des deux éléments reliés (illustration extraite du site du projet :* `https://blizaar.list.lu/doku.php?id=BLIZAAR` *).*

Ces données forment alors un système complexe (Fig. 2.2) modélisable comme un ensemble de sous-systèmes (ou couches) inter-connectés : un réseau multi-couche. Un réseau multi-couche est défini par un ensemble de sommets (ici les documents, entités et ePublications) et un ensemble de liens (les relations entre les documents et entités) mais propose en plus un ensemble de couches i.e. des ensembles de sommets ou liens définis par les différents types possibles (l'ensemble de sommets représentant des personnes va définir la couche personne, l'ensemble des sommets représentant des localisations va définir la couche localisation, etc.) [41, 56]. Ces couches rendent complexe une analyse claire de la topologie du réseau tant la distribution peut varier en fonction des couches ou des liens inter-

FIGURE *2.2* – ***Vue générale des données*** *: Représentation noeud-lien des données de nos collègues historiens provenant de la base de données du CVCE (Centre Virtuel de Connaissance sur l'Europe). Réseau complexe de 51798 sommets et 1074643 arêtes, sa distribution est essentiellement multipolaire et hiérarchique. Un tel ensemble de données, à cause de la surcharge visuelle, de la superposition des sommets et de la complexité sémantique, impose le développement de méthodes spécialisées pour pouvoir mener une analyse efficace et pertinente. La conception et l'implémentation de ces méthodes sont développées dans les chapitres suivants.*

couches. Dans le cas présent, malgré une construction hiérarchique des données, la distribution est essentiellement multi-polaire avec des points centraux (institutions

majeures, personnalités majeures de l'Europe, etc.) inter-connectés par des sommets intermédiaires appartenant à des couches diverses. Les réseaux multi-couches sont donc de véritables mille-feuilles sémantiques nécessitant alors des méthodes propres pour être exploités pleinement.

Enfin, il est à noter que les historiens du CVCE sont déjà familiarisés avec la notion de réseaux, ce qui a permis une étroite collaboration dans la conception des méthodes présentées dans la suite. Des travaux antérieurs au projet BLIZAAR ont permis à nos collègues historiens de créer un outil, Histograph [24], permettant de filtrer et visualiser leurs données en utilisant des vues noeuds-liens. La plupart des visualisations de réseaux actuelles en humanité numérique sont publiées dans "la tradition de l'ère d'imprimerie : sous forme d'images statiques expliquées par du texte et une légende". Il a été constaté que ces visualisations peuvent être "traîtresses" car "exigent que les personnes comprennent l'objectif visé, la conceptualisation des données et les biais potentiels". Histograph a donc été développé dans l'objectif de répondre à cette observation. Permettant des visualisations proposant animations, filtres et informations contextuelles, il permet de faciliter la compréhension générale et le développement d'analyses plus efficacement qu'avec une "image statique". Dans la conclusion de ces travaux, les historiens constatent la présence d'"opportunités pour les projets d'analyse de réseau". Le projet BLIZAAR s'inscrit en continuité de cette dynamique : proposer cette fois une méthode permettant l'exploration des données, la gestion du caractère multi-couche du réseau et la prise en compte des objectifs et contraintes des experts lors de la navigation.

2.3 Travaux préliminaires

Ce n'est pas la première fois que nous exploitons des réseaux multi-couches dans le cadre des sciences humaines et sociales. Nos expériences passées nous ont amenés à travailler avec différents spécialistes (géographes, géomaticiens, juristes et sociologues) à travers deux projets, GEOBS (2015-2017) et TETRUM (2016), ayant chacun eu un impact sur la conception de la méthode `M-QuBE`^{3}. C'est notamment ces travaux qui nous ont amenés progressivement à considérer ces données comme des réseaux multi-couches et a comprendre les spécificités de ce modèle.

GEOBS

Ces dernières années, le développement des moyens de diffusion de l'information numérique, l'amélioration des techniques de géolocalisation et l'utilisation accrue

des informations environnementales par les politiques publiques ont résulté en une augmentation considérable des flux d'informations géographiques. Pour contrôler ces flux, de nombreux investissements ont été débloqués ces dernières années par les autorités publiques afin de créer des structures spécialisées : les Infrastructures de Données Géographiques (Code de l'environnement - Article L127-1) ou "IDG".

Le projet région Aquitaine GEOBS (http://geobs.cnrs.fr/) avait pour objectif d'analyser d'une part les flux d'informations transitant à travers et entre les IDG (une IDG peut partager ou moissonner les données d'autres IDG) et d'autre part les usages et moyens mis en oeuvre autour de ces plateformes. De manière simplifiée, ces IDG se présentent sous la forme d'un site internet à partir duquel il est possible d'accéder aux différentes informations et études géographiques qui y sont stockées. Ces données sont toutes accompagnées d'une fiche de méta-données renseignant la zone géographique concernée par la donnée (l'emprise), les thèmes de l'étude, les auteurs, etc. Nos travaux [62] ont consisté à utiliser ces méta-données afin de pouvoir modéliser et exploiter différents graphes permettant ainsi à nos collègues géographes d'analyser la qualité et la circulation des informations intra et inter IDG. Plusieurs approches ont donc été réalisées afin de répondre aux questions de nos experts notamment :

- Une analyse de la couverture thématique basée sur des calculs de similarité entre les mots clés et descriptifs – Est-ce que les thèmes sont équitablement répartis entre les données ? Est-ce qu'il existe des communautés thématiques majeures ? (Fig. 2.3)
- Une analyse de la gouvernance des données basée sur les différentes informations relatives aux acteurs ayant généré les données – Quels sont les acteurs phares dans le milieu ? Y a t-il des groupes d'acteurs en concurrence ou coopération ?
- Une analyse de la couverture spatiale utilisant les informations de géolocalisation – Y a t-il une homogénéité des emprises géographiques dans le territoire ? Quel est le degré de superposition des emprises des différentes études ? (Fig. 2.4)

Tous les travaux réalisés, y compris les exemples précédents, ont un point commun : chaque analyse est basée sur une métrique spécifique utilisant des attributs différents du même jeu de données. Autrement dit, chaque nouvelle analyse a nécessité d'ajuster la manière de calculer un score pour s'adapter à de nouvelles informations sémantiques issues du même jeu de données. Chaque métrique y détermine un score qui est comparé à une valeur seuil définie par l'utilisateur. Ce

seuil permet ainsi de filtrer les données afin de ne conserver que celles considérées représentatives ou intéressantes pour l'utilisateur. Par exemple, pour le graphe de similarité (Fig. 2.3), un lien n'est affiché qu'à partir d'un pourcentage de ressemblance des thèmes traités. Un seuil maximal ne va alors afficher dans le graphe que les arêtes entre des études géographiques identiques ou redondantes alors qu'un seuil nul va générer un graphe avec l'ensemble des liens (dont la sur-abondance n'est ni représentative de l'objectif ni exploitable). Pour le graphe de couverture spatiale (Fig. 2.4), le schéma est identique mais l'analyse étant centrée sur la superposition des surfaces couvertes par les études géographiques, le seuil est défini en fonction de l'intersection spatiale entre deux études. Ainsi, deux études sont liées si elles ont une surface en commun supérieure à la valeur fixée par le seuil. Le projet GEOBS a continué à posteriori du début du projet BLIZAAR et a généré plusieurs autres publications, notamment sur la communication et l'usage utilisateur des IDG [28].

Si au moment de ces travaux, nous n'avions pas encore de focus sur les graphes multi-couches, le cadre est pourtant comparable : les différents liens entre les sommets définissent des couches (similarité sémantique, superposition spatiale, gouvernance pour les trois exemples ci-dessus) à partir desquelles il est nécessaire de faire ressortir ce qui est intéressant pour l'utilisateur, comme pour les données du CVCE. De ces travaux, nous avons donc tiré deux enseignements ré-exploités lors de la conception de `M-QuBE`[3] : la nécessité de différencier le traitement pour l'adapter à chaque "couche" sémantique d'un même réseau / jeu de données (liens thématiques, liens d'appartenance, liens spatiaux...) et la nécessité de restreindre la visualisation à ce qui est le plus pertinent pour l'utilisateur, en évaluant et comparant les éléments traités en fonction des objectifs définis.

TETRUM

Les réseaux de traite des humains ne sont pas nouveaux mais ont subi un changement dû aux nouvelles techniques de communication et de partage des informations. Avec le développement d'internet, c'est toutes les pratiques et stratégies criminelles qui ont évolué. Un projet interdisciplinaire (PEPS/IdEx) comprenant juristes, sociologues et informaticiens a donc été mis en place afin d'analyser et comprendre les nouvelles formes, usages et modes opératoires de ces réseaux criminels [47] (et a été mentionné dans un article du Figaro : `https://www.labri.fr/images/uploads/Art%20Figaro-Trafic%20%EAtre%20humains`).

Contrairement aux projets décrits précédemment, les données initiales dont

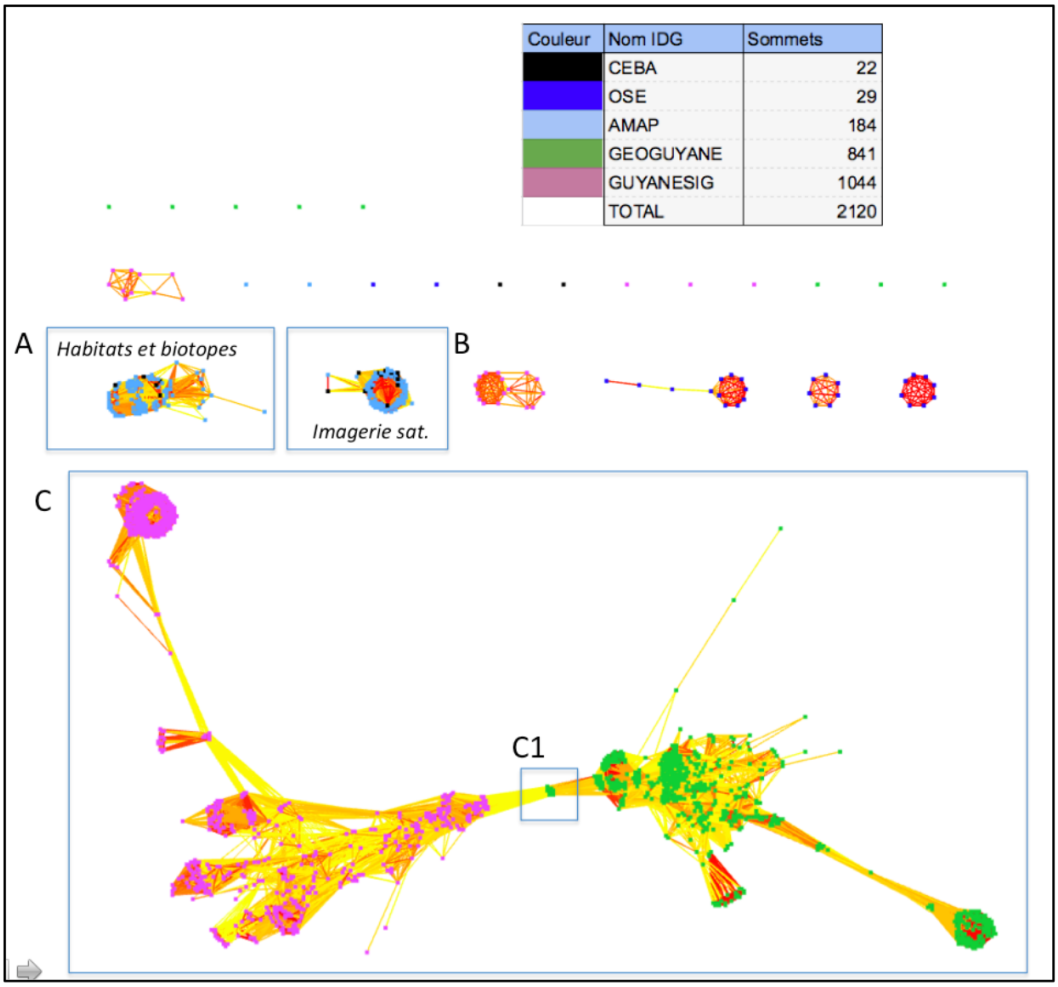


FIGURE *2.3* – ***Graphe de similarité*** *: Ce graphe est construit à partir d'un calcul de similarité entre les différentes données de cinq IDG en utilisant les mots-clés, thèmes et descriptifs contenus dans les méta-données. Les sommets représentent les études géographiques et une arête entre deux sommets indique qu'ils ont un score de similarité sémantique supérieur au seuil défini par les experts (arête jaune : similarité minimale, arête rouge : similarité maximale). Les différentes communautés représentent alors des groupements thématiques attribuables aux différentes IDG. Il est aussi possible de voir les thématiques en commun entre deux IDG, permettant ainsi de connaître l'intersection de leurs couvertures sémantiques. Image provenant de [62].*

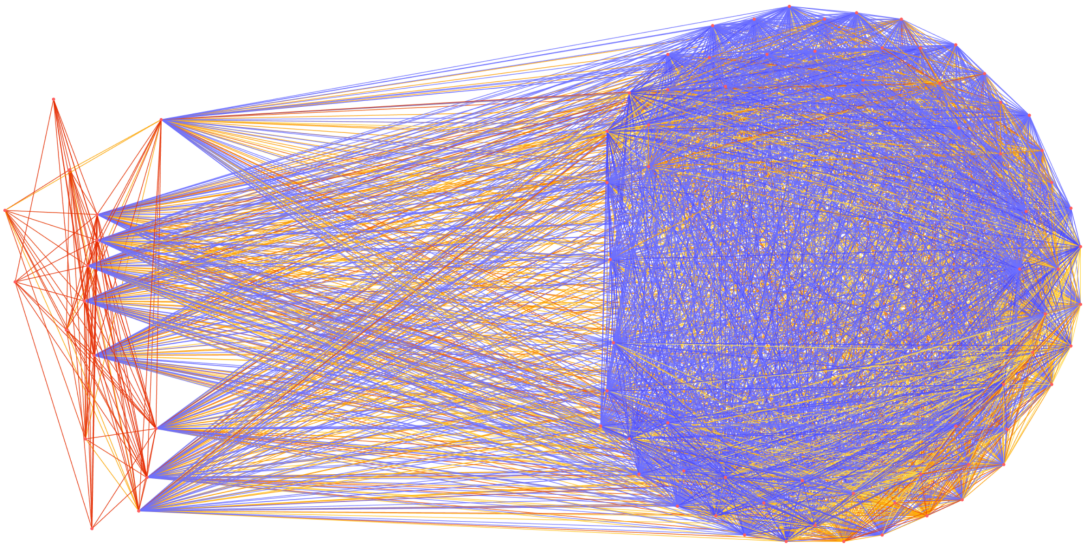


FIGURE *2.4* – ***Graphe d'emprises géographiques*** *: Extrait du graphe d'emprise de PIGMA (l'IDG Aquitaine). Les sommets représentent les études géographiques et une arête entre deux sommets indique que les emprises spatiales des deux études se superposent et que leur surface d'intersection est supérieure à un seuil fixé par les experts. Plus la surface d'intersection est élevée plus l'arête tend du bleu vers le rouge. Dans ce graphe, on peut ainsi voir que l'essentiel des études ne se recouvrent que légèrement (arêtes bleues), signe d'une couverture spatiale homogène, exceptée une communauté d'études hautement superposées (arêtes rouges).*

nous avons disposé sont entièrement physiques : un corpus composé de plusieurs dossiers judiciaires d'environ 25000 pages chacun, traitant d'affaires répertoriées par la police et relatives à ces réseaux. De nombreux types de documents y sont consignés : témoignages, écoutes téléphoniques, rapports d'interrogatoire, liste de numéros de téléphones suspects, rapports de police...

Une étape essentielle va donc être de numériser ces données afin de pouvoir les visualiser et les exploiter efficacement. Cette étape est difficilement automatisable notamment à cause de l'hétérogénéité des documents et de la mauvaise qualité d'impression des feuilles excluant de scanner automatiquement les dossiers. En plus de cela, beaucoup de document nécessitent une analyse humaine afin de s'assurer de la pertinence voir de la véracité des informations. Un interview, un témoignage ou un interrogatoire peut apporter des informations en contradiction avec d'autres, non fiables ou même volontairement fausses (exemple d'une personne ne donnant

pas son vrai nom, numéro de téléphone, etc.). Ces éléments requièrent alors une intervention humaine mais, avec des dizaines de milliers de page à étudier, il est nécessaire d'avoir une aide informatique afin de stocker et interroger efficacement ces données.

Pour ce faire, un modèle abstrait de données a été réalisé et utilisé au sein d'une plateforme en ligne permettant de faciliter la consultation et la saisie des informations. Cette plateforme a été réalisée simultanément avec l'exploration des dossiers par nos collègues des sciences humaines et sociales, le modèle abstrait a donc évolué au fur et à mesure des découvertes et a été mis à jour régulièrement en même temps que la plateforme était implémentée.

Pour cette raison, nous avions commencé le projet avec une base de données relationnelle, solution classique pour des données basées sur des relations. Les données permettent de définir un réseau multi-couche (Fig. 2.5) où chaque couche est définie par les types de relations : lien financier, lien sexuel, lien de sang, lien de réseau, lien de connaissance, lien de soutien et lien juju (une cérémonie religieuse incitant une personne à se prostituer pour rembourser une dette sous peine de "mauvais sort" [46]).

Cependant, cette solution s'est vite révélée problématique au niveau de la conception des requêtes et de leurs performances. Les requêtes extrêmement complexes, en raison du grand nombre de jointures dû aux nombreux types de liens et entités, ainsi que la nécessité de changer ou faire évoluer régulièrement le modèle de données ne conviennent pas à la rigidité du modèle relationnel [45].

C'est pourquoi nous avons, dans le cadre du CVCE, stocké et utilisé ces données à travers, d'une part, une base de donnée graphe (une base de données spécialement conçue pour l'exploitation des réseaux) et, d'autre part, Tulip [3], une infrastructure logicielle spécialisée dans les réseaux afin de bénéficier d'une souplesse dans la conception du modèle (et ainsi permettre son évolution) ainsi qu'une performance accrue pour toute requête nécessaire à l'analyse ou la visualisation. Plus de détails sur le rôle et l'utilisation de ces objets sont disponibles à la section 6.1.

2.4 Synthèse

A travers les objectifs et les données de nos collègues historiens du CVCE, nous proposons une méthode permettant à la fois une navigation simple et intuitive dans leur corpus de documents ainsi qu'un respect facilité des contraintes inhérentes à la conception des ePublications.

Pour ce faire, les données dont nous disposons prennent la forme d'une interconnexion complexe de sous-systèmes : un réseau multi-couche. Cet objet nécessite cependant des techniques et méthodes spécialisées pour pouvoir bénéficier pleinement de sa sémantique riche, que nos expériences passées ont fait émerger notamment à travers **une différenciation des traitements des couches**, **un focus accru sur les éléments intéressants pour l'utilisateur** et **des choix techniques optimisés pour ce type de réseau**.

Cependant, concevoir une méthode de navigation et de visualisation pour des sciences informatiques ou physiques n'est pas le même exercice que pour les sciences humaines et sociales. Les méthodologies sont susceptibles de différer, tant par les types de visualisation nécessaires que la manière d'utiliser ces visualisations.

Dans le chapitre suivant, nous faisons donc un tour d'horizon des spécificités des sciences humaines et sociales afin de déterminer les éléments et contraintes nécessitant d'être pris en compte pour modeler notre méthode au plus près des besoins et méthodologies de ce domaine.

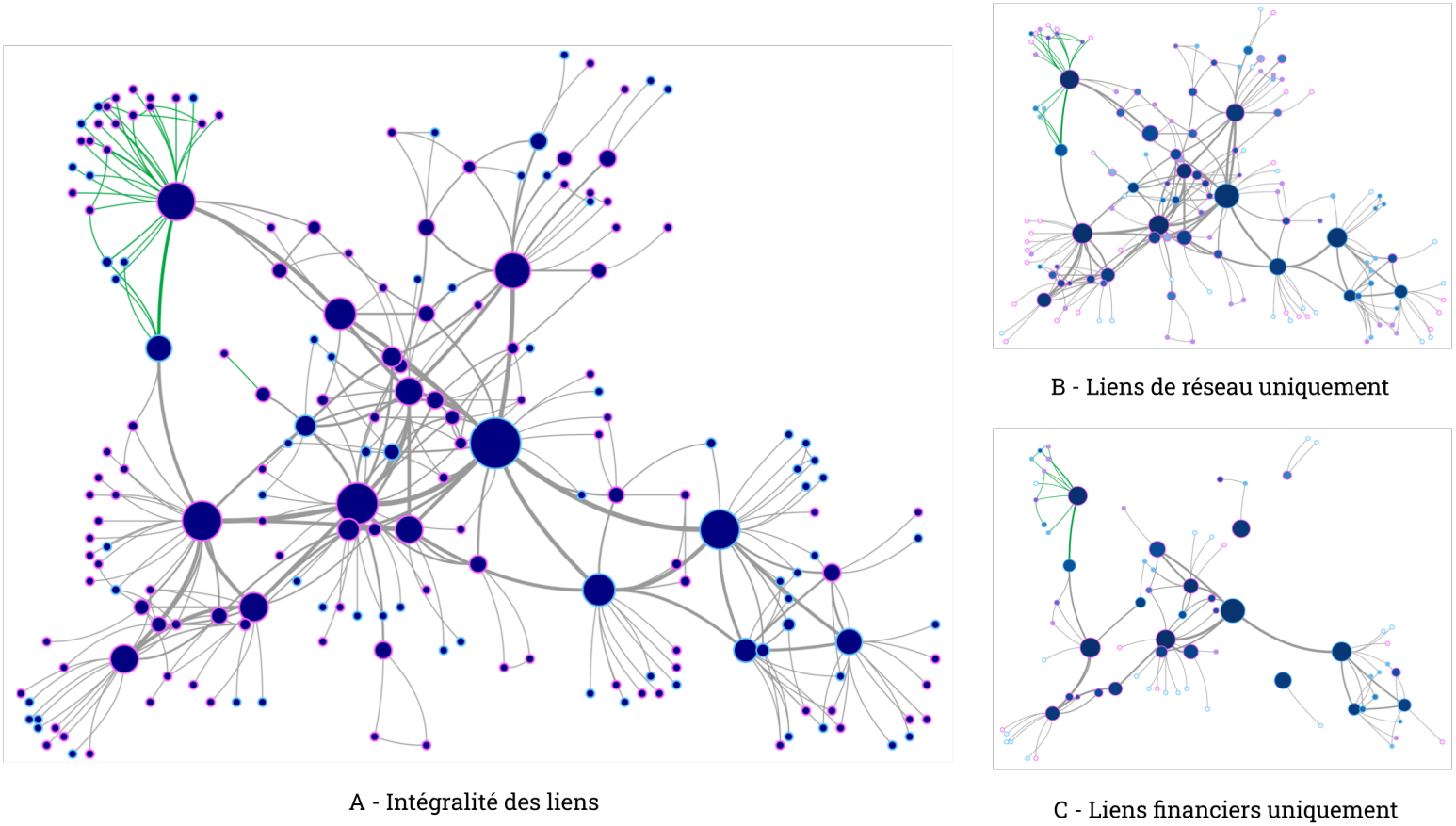


FIGURE *2.5* – ***Réseau TETRUM*** *: Voici un extrait du réseau TETRUM. Chaque sommet représente une personne dont la couleur du contour indique le sexe. Le réseau A correspond au réseau complet. Les réseaux B et C représentent respectivement uniquement les liens de réseau (témoigne d'une action au sein du réseau criminel entre deux personnes) et les liens financiers (témoigne d'un échange de nature financière entre deux personnes) ainsi que les sommets reliés par ces liens. Les liens verts correspondent à la même sélection de sommets dans les trois réseaux présentés.*

Chapitre 3

Visualisation et sciences humaines et sociales

Sommaire

La navigation et l'exploration de réseaux est à la croisée des chemins de plusieurs domaines. Il est évidemment nécessaire de s'immerger dans l'analyse structurelle et sémantique des réseaux afin de comprendre les objets que nous utilisons et s'adapter en conséquence (c'est ce qui a été réalisé dans le chapitre 2) mais il est tout autant nécessaire de se focaliser sur la visualisation générale en elle même.

Si jusqu'ici nous avons toujours traité d'éléments relatifs aux graphes et aux réseaux, la visualisation couvre cependant un spectre d'objets bien plus large [67]. Plus exactement : là où il y a des données, il y a de la visualisation.

Ce domaine n'est pas nouveau. Déjà dans les années 1850, Charles Joseph Minard a produit plusieurs représentations graphiques de données et peut être considéré comme un pionnier de la visualisation analytique. Dans plusieurs de ses travaux (notamment ses "cartes figuratives et approximatives" : "des quantités de viandes de boucherie envoyées sur pied par les départements et consommateurs à Paris", "des tonnages des Grand Ports et des principales Rivières d'Europe", "représentant pour l'année 1858 les émigrants du globe", etc. - voir Annexe A), C.J Minard essaye déjà d'exprimer visuellement les interconnexions entre de multiples

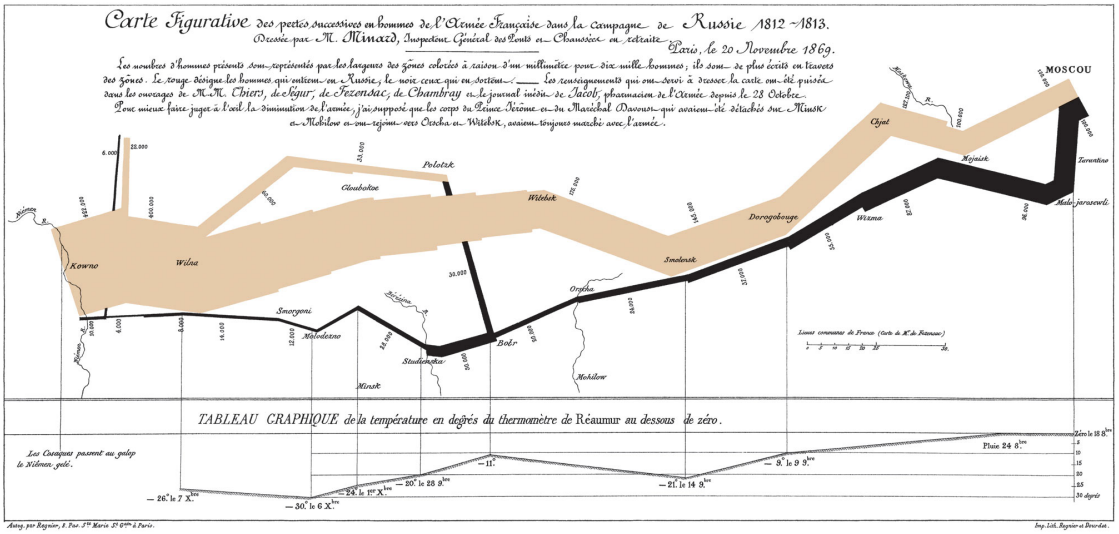


FIGURE *3.1* – ***Début de la visualisation*** *: Ceci est le travail de Charles Joseph Minard, un des pionniers de la représentation graphique des données. Cette carte réalisée en 1879 est son travail le plus célèbre et représente le flux des soldats français lors de la campagne de Russie en 1812–1813 sous le règne de Napoléon. La largeur du trait orange représente le nombre d'individus encore en vie et le trait noir le nombre de pertes humaines. Ces éléments peuvent ensuite être analysé par rapport à l'emplacement géographique, la température et les divers renforts indiqués par les traits transversaux.*

types de données. Son travail le plus célèbre, la "Carte figurative des pertes successives en hommes de l'Armée française dans la campagne de Russie en 1812–1813" (Figure 3.1, https://fr.wikipedia.org/wiki/Carte_figurative), détaille simultanément de nombreuses informations différentes : la chronologie de l'évènement, la température, l'itinéraire des soldats, la position géographique, le nombre de soldats à travers le temps ainsi que les renforts et pertes humaines.

Ces travaux sont précurseurs d'une des problématiques de la visualisation moderne : comment présenter un ensemble de données de différents types simultanément et efficacement. S'il est trivial de recouper deux types de données afin d'analyser les relations qu'ils entretiennent, le problème se complexifie lorsque de nouvelles dimensions sont ajoutées aux données étudiées. Cette problématique, présente depuis longtemps tant en science des données (avec par exemple les questionnements relatifs à l'implémentation de ces visualisations plus complexes [79]) qu'en sciences humaines (avec par exemple les travaux menés sur des données à la fois spatiales

et temporelles [34] avec les 'Space-time cubes", revisités de nombreuses années plus tard [42] et qui forment encore aujourd'hui la base de la "time-geography"), prend une importance capitale avec l'étude des graphes multi-couches.

Les réseaux multi-couches (voir Chapitre 2) font partis de ces objets complexes, aux multiples dimensions, qui nécessitent des méthodes spécifiques pour exprimer leur plein potentiel. S'il est toujours possible d'occulter des caractéristiques d'un réseau multi-couche pour le réduire à un objet facilement visualisable (souvent un graphe simple), beaucoup de domaines, dont les sciences humaines, ont un intérêt à conserver l'entièreté des informations. De ce fait, des méthodes spécifiques doivent être déployées. C'est ce que nous proposons de faire avec `M-QuBE`^{3}, une méthode d'exploration spécialisée pour les réseaux multi-couches dans le cadre des sciences humaines et sociales conçue afin de répondre aux besoins de nos collègues historiens et s'appuyant sur les expériences accumulées auprès de collègues juristes, sociologues et géographes.

Dans ce chapitre, nous faisons un tour d'horizon du domaine de la visualisation (section 3.1) puis de quelques unes de ses applications dans le domaine des sciences humaines et sociales. A travers les liens avec la science des données (section 3.2), les réseaux multi-couches (section 3.3) et les méthodes exploratoires (section 3.4), nous explicitons les spécificités des sciences humaines et sociales afin de justifier les choix qui ont été faits dans la conception et le développement de `M-QuBE`^{3} décrit dans les chapitres suivants.

3.1 Visualisation : méthodes et mantras

Comme dit précédemment, la visualisation n'est pas un concept récent. Ce domaine s'est construit graduellement en suivant des préceptes dont certains ont été hissés au rang de mantra. Intervenant à deux niveaux, ces règles sont liées d'une part à la conception de la visualisation et d'autre part à l'application même de la visualisation.

Le premier et plus connu des mantras est celui de Shneiderman [70] qui définit une caractérisation et l'ordre des interactions nécessaires : **"Overview first, zoom and filter, then details-on-demand"** (on établit d'abord une vue d'ensemble, ensuite on zoom et on filtre puis on détaille à la demande). Par ce mantra, on détermine une méthodologie à adopter pour maximiser l'efficacité de la visualisation. Ceci impacte lourdement la phase de conception car nécessite des interactions et des visualisations spécifiques. En effet, une processus de visualisation n'est pas nécessairement un processus fixe. Cela peut à la fois désigner un processus à une

seule étape, où la visualisation va être calculée puis présentée comme conclusion à l'utilisateur, ou suivre une méthodologie en plusieurs étapes, suivant par exemple le mantra précédemment cité, permettant de rentrer dans les données, d'y naviguer et d'en faire ressortir ce qui est intéressant pour un utilisateur donné [78]. Ce second scénario correspond à la notion d'exploration, étroitement corrélée à la visualisation tant l'exploration de données ne peut s'effectuer sans visualisation et que la visualisation ne se justifie parfois que par l'exploration.

Le domaine de l'exploration de données devient progressivement indispensable à l'heure où nos usages quotidiens d'internet, des objets connectés ou des réseaux sociaux génèrent des quantités phénoménales de données. Comme dit précédemment, exploration et visualisation sont liées et suivent des règles communes. Cependant, le domaine d'application qui nous intéresse dans le cadre de notre projet, i.e. celui des sciences humaines et sociales, a des spécificités qui influent sur la méthodologie habituelle.

En effet, le mantra de Shneirderman est habituellement amplement suffisant pour mener à terme une visualisation efficace sur les données : à partir d'une vue globale, l'utilisateur peut cibler les points d'intérêt ou, en cas de difficulté, filtrer cette vue pour les trouver plus aisément. Puis, on concentre l'analyse sur ces points. Cependant, si les données sont trop volumineuses ou connectées, une vue d'ensemble naïve posera des problèmes de lisibilité à cause d'un encombrement visuel trop important [30, 31] et ne répondra donc pas à son objectif. Un solution communément utilisée est alors d'agréger les données afin d'offrir une vue d'ensemble simplifiée mais significative. L'agrégation nécessite alors de pouvoir grouper des informations pour conserver et mettre en avant essentiellement ce qui est porteur de sens et d'intérêt pour l'utilisateur. Voici quelques exemples représentatifs nécessitant l'agrégation des données :

- imMens [50] est un système de visualisation dont les méthodes utilisées sont implémentées et optimisées pour fonctionner sur les navigateurs internet. Le credo autour d'imMens est de proposer des visualisations dont la limite ne serait pas le volume de données à visualiser mais la résolution de l'écran utilisé pour la visualisation. Un tel objectif induit forcément de faire des concessions sur l'affichage en déterminant précisément ce qui sera montré à l'utilisateur pour éviter une surcharge visuelle qui détériorerait la lisibilité. De plus, l'interactivité et l'exécution en temps réel de la visualisation ajoutent une couche de difficulté supplémentaire et nécessitent encore davantage à agréger les données visualisées pour conserver des performances stables. Il y a donc nécessité de représenter comme des éléments individuels des groupes

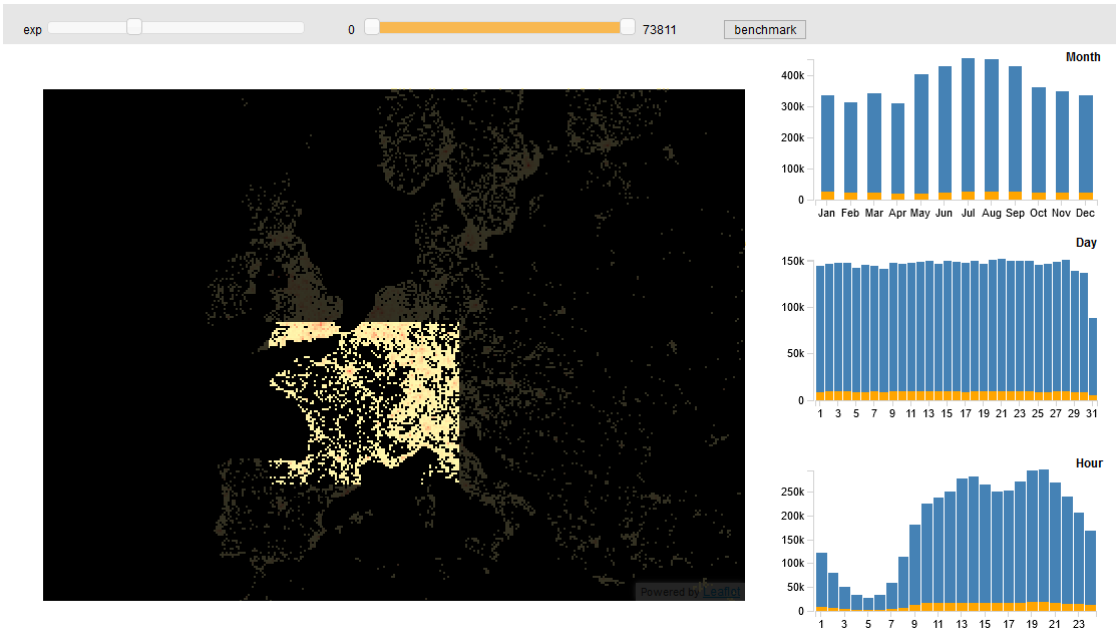


FIGURE *3.2* – ***imMens*** *: Ceci est une capture d'écran provenant d'une version de démonstration d'imMens. La vue de gauche est une carte de chaleur représentant le nombre de personnes enregistrées pour des voyages en avion sur une période d'un an. Les trois autres visualisations montrent ces même enregistrements en fonction des mois, des jours et des heures. En sélectionnant une zone comprenant la France sur la carte (points jaunes), on peut observer l'ensemble des éléments correspondant dans les trois autres visualisations. Ainsi sélectionner une zone de quelques pixels sur la carte peut sélectionner plusieurs milliers d'éléments dans les autres vues.*

d'éléments ou des ensembles d'informations : sélectionner un point sur imMens correspond alors à sélectionner un ensemble potentiellement très grand de valeurs liées à ce point (voir Figure 3.2).

- Si imMens propose des visualisations sur des données non inter-connectées, il existe aussi de nombreux exemples pour la visualisation de graphes et réseaux. C'est par exemple le cas de JASPER [73] qui se propose de représenter un réseau entier sous forme de mosaïque pixelisée (voir Figure 3.3) dont les carreaux (les zones d'une même couleur) représentent les communautés. Les pixels de ces zones représentent les différents éléments constituant les communautés qu'elles définissent. Une telle visualisation fait le choix d'orienter la

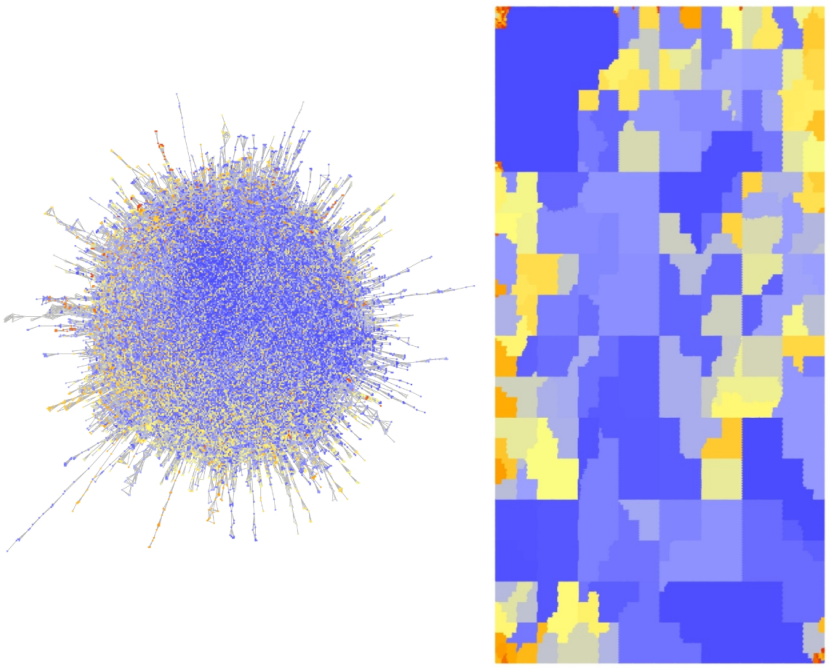


FIGURE *3.3* – ***JASPER*** *: Cette méthode propose une représentation orientée pixel permettant de mettre en avant les communautés et leur disposition au sein du réseau en agrégeant les données représentées. La visualisation de gauche correspond à la représentation noeud-lien des données (utilisant l'algorithme de dessin* FM^3 *[33] pour déterminer la position des noeuds). Celle de droite représente les même données mais utilisant JASPER. Les zones de couleurs représentent les différentes communautés et les pixels les composants sont les différents noeuds du réseau. Image provenant de [73].*

visualisation pour se concentrer essentiellement sur les communautés, quitte à ne pas représenter les liens entre les différents sommets. Ces liens sont néanmoins pris en compte afin de conserver une proximité entre les sommets voisins dans la visualisation. Cela permet ainsi d'avoir aisément un aperçu représentatif des relations inter-communautaires, de l'étendue de ces communautés ou de l'évolution du réseau lors d'une étude de propagation.

- Un autre exemple similaire sont les travaux autour de la navigation multi-échelle [4,12]. Le concept est de transformer un graphe dont la taille et la complexité ne permettent pas une lecture simple à un objet plus réduit. Ce faisant, des structures visuelles autrement imperceptibles apparaissent, rendant

le graphe plus facilement analysable. De nombreuses méthodes existent [49] passant par exemple par le groupement d'arêtes (représentation d'un ensemble d'arêtes du graphe en une seule arête) ou la création de méta-noeud (représentation d'un sous-réseau par un unique sommet). Ceci nécessite néanmoins de pouvoir hiérarchiser les données afin de déterminer ce qui est agrégable, ce qui n'est pas toujours possible.

Si le mantra de Shneiderman est le plus couramment utilisé, il n'est pas nécessairement adapté à tous les scénarios. Celui-ci est essentiellement centré sur la visualisation. Elle est le point de départ du processus et l'entièreté de l'analyse en dépend. Une approche davantage basée sur l'analyse visuelle a été proposée par Keim *et al.* dans un nouveau mantra inspiré de celui de Shneiderman : **"Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand"** [39] (d'abord on analyse, ensuite on montre ce qui est important, on zoom, on filtre, on analyse plus en profondeur puis on détaille à la demande). Celui-ci ne s'abstrait pas de la nécessité d'utiliser un vue globale simplifiée. Les deux premières étapes : l'analyse et la mise en valeur de ce qui est intéressant nécessite une agrégation des données pour procéder aux phases postérieures de la visualisation (avec le zoom, le second filtrage et l'éventuelle extension de ce qui est montré). Cela peut être contraignant voir rédhibitoire si une vue globale est difficile ou semble impertinente dans le scénario (voir section suivante).

Chaque nouveau scénario apporte son lot de variations et de contraintes susceptibles d'imposer un ajustement des mantras existants. On peut néanmoins classer les différents travaux vus précédemment en deux catégories :

- **Top-Down :** Comme son nom l'indique, les visualisations de type *top-down* sont celles qui vont commencer par établir une vue générale (*top*) avant d'offrir un moyen de se pencher à posteriori sur les détails des données (*down*). Toutes les visualisations citées en exemple dans cette section sont de type *top-down* car découlant du mantra de Shneiderman ou Keim. Elles sont contraintes de passer par de l'agrégation, du filtrage ou de s'appliquer à des jeux de données qui sont pas trop volumineux ou complexes.
- **Bottom-Up :** Inversement, l'approche *bottom-up* propose de partir au plus près des données (*bottom*) puis, à partir de ces données, de pouvoir analyser ou se faire un aperçu de l'ensemble (*up*). Cette approche est privilégiée lorsqu'une vue générale n'est pas possible à cause des contraintes liées au scénario ou au domaine de l'étude. C'est par exemple le cas pour la visualisation d'ontologie [23] qui n'a pour le moment pas moyen de mener une

étude sans se focaliser en premier lieu sur des éléments précis et ne peut donc pas se baser essentiellement sur une vue globale. Une illustration de cette approche sont les travaux de S.van den Elzen et J.J. van Wijk [75]. Mêlant simultanément une vue centrée sur les détails et une vue globale des données au sein d'une même visualisation, ils proposent par une analyse et une sélection d'éléments précis de générer une vue agrégée en fonction des éléments intéressants pour l'utilisateur. Cette approche impose néanmoins aussi des contraintes : il est nécessaire de pouvoir naviguer dans le détail des données efficacement, un mécanisme de filtrage et de sélection est donc absolument nécessaire. Plus qu'en science des données, cette approche devient incontournable pour certains domaines dont les sciences humaines et sociales, ce que nous verrons dans la section suivante.

Nous avons présenté les approches classiques de la visualisation. Cependant, la structure multi-couche de nos données (cf. Chapitre 2) et les contraintes inhérentes aux sciences humaines et sociales nous obligent à diverger des méthodes traditionnelles. Dans la partie suivante, nous détaillons ces spécificités et ce qu'elles induisent afin d'expliquer, d'une part, les challenges relevés par notre méthode et, d'autre part, les particularités de notre méthode.

3.2 Science des données et sciences humaines et sociales

La science des données et les sciences humaines et sociales ont des manières distinctes d'initier et de mener leurs analyses respectives. Par leurs différences d'objectifs et de données, les méthodologies sont inévitablement différentes. Ces particularités et cette divergence par rapport à la science des données sont notamment développées dans les travaux de Borgatti et al. [9]. L'analyse et l'exploitation des réseaux sont ancrées dans un large panel de domaines des sciences humaines, de l'Histoire à l'économie en passant par la psychologie. La différence fondamentale avec la science des données est que l'attention des experts va se concentrer non pas sur une analyse du réseau lui même mais sur "comment des individus autonomes peuvent se combiner pour créer des sociétés durables et fonctionnelles". L'essentiel de l'attention est alors placé dans une analyse rapprochée des éléments du réseau. Borgatti cite en exemple les travaux de Moreno [59] qui fait l'analogie entre ces réseaux et des systèmes physiques. Les personnes et autres éléments constitutifs du réseau analysé deviennent alors des "atomes sociaux" soumis aux lois de la

“gravitation sociale” (lois impactant l’apparition des liens et donc la structure du réseau), faisant ainsi écho au terme de “physique sociale” du philosophe français Auguste Comte [18] presque cent ans plus tôt. Pour accentuer encore davantage l’importance capitale mise sur les individus, il cite aussi l’ouvrage de Durkheim [25] où les sociétés humaines sont comparées à des organismes biologiques constitués de l’interrelation de leurs éléments. On pourrait citer aussi la notion d’“ego network” [2, 55], au nom explicite, où l’analyse s’effectue à partir d’un individu (*ego*) pour analyser les différentes couches de son entourage (*alter*), toujours afin de centrer l’analyse sur l’individu et son entourage. On observe alors deux niveaux d’analyse, une niveau inter-individuel, se concentrant sur les individus, et un niveau inter-organisationnel, se concentrant sur les communautés d’individus. Si ces deux perspectives se voient accorder une importance différente en fonction du courant dans lequel se situent les experts (individualisme ou holisme), elles ne sont pas nécessairement distinctes et il est possible de les prendre en compte conjointement afin d’établir une étude plus globale [48]. La notion de “Noyau-Périphérie” [8], où l’on considère un groupe indivisible d’individus (*core*) connectés à des individus externes (*periphery*), est déjà positionnable de manière intermédiaire entre le niveau inter-individuel et le niveau inter-organisationnel. Cette notion est même élargie à un modèle continu, où il existe potentiellement plusieurs classes d’individus semi-périphériques, amenuisant encore plus les différences entre ces deux niveaux.

Parmi tous ces modèles, le point commun est l’importance apportée à ce qui constitue le réseau plutôt que le réseau lui même. C’est notamment ce que dit Borgatti lorsqu’il développe les perceptions mutuelles entre la science des données et les sciences sociales [9]. Chaque domaine est susceptible de considérer l’autre comme uniquement descriptif. Les scientifiques peuvent reprocher aux sociologues de ne pas opposer leurs mesures des propriétés du réseau à des modèles théoriques là où des sociologues peuvent considérer les travaux des scientifiques comme horriblement simplistes et génériques (“*alarmingly simplistic and coarse-grained*”). Borgatti indique par exemple que des modèles d’étude comme ceux basés sur l’utilisation de graphes aléatoires paraissent extrêmement naïfs pour les sociologues, s’apparentant pour eux à “comparer un gratte-ciel à une distribution aléatoire de la même quantité de matériaux”. Borgatti explique cette différence en sciences sociales et humaines par un intérêt supérieur pour le sommet individuel (représentant tant un individu qu’un collectif d’individus) que pour le réseau lui même en tant qu’ensemble.

Il s’agit donc essentiellement d’un paradigme différent auquel notre méthode va devoir s’adapter pour être en accord avec la méthodologie et les objectifs généraux

des sciences humaines et sociales. En plus d'opter pour une approche majoritairement centrée sur les sommets, il est aussi nécessaire de capturer et d'exploiter le caractère multi-couche des réseaux utilisés en sciences humaines et sociales. C'est ce que nous proposons dans la section suivante.

3.3 Réseaux multi-couches et sciences humaines et sociales

En sciences humaines et sociales, l'étude d'un réseau n'est pas tant l'analyse du réseau lui même (topologie) que directement l'analyse de la sémantique qu'il véhicule. Or, comme dit précédemment (voir Chapitre 2), un réseau sémantiquement riche est facilement modélisable sous forme de réseau multi-couche.

Avant même que le concept de multi-couche se démocratise, ces objets étaient déjà traités inconsciemment même sans avoir spécifiquement de méthodologies affirmées centrées sur les réseaux multi-couches. Nos expériences, tant avec les géographes de GEOBS que les juristes et sociologues de TETRUM (voir Chapitre 2), reflètent d'ailleurs cette situation. Les géographes ont par exemple souvent cherché à générer des graphes bipartis en utilisant deux types parmi l'ensemble des métadonnées dont on disposait [62]. La situation est alors similaire au fait d'établir un filtrage sur le réseau multi-couche correspondant à l'ensemble des données pour ne conserver et analyser que les interactions de deux de ses couches. Pour TETRUM, la conception et le développement du modèle de données a mené progressivement à considérer notre réseau comme multi-couche [45]. Si les sommets du réseau criminel représentent tous des personnes, le caractère multi-couche du réseau vient cette fois du large spectre de types de lien possibles entre ces personnes (liens financiers, liens de prostitutions, liens de sang, etc.). On parle alors de réseau multiplexe, une sous-catégorie de réseau multi-couche où les couches sont définies non pas par le type des sommets mais par le type des liens du réseau [41]. D'autres exemples de travaux sur les réseaux criminels [13, 68] comportent aussi cette caractéristique : même sans expliciter l'aspect multi-couche du réseau, l'analyse est menée en ayant une pleine connaissance de ses couches voir en concentrant l'analyse sur celles-ci.

Pour autant, il est courant de voir en sciences humaines et sociales des réseaux exploités comme des graphes simples. C'est le cas par exemple des travaux précédemment cités de Borgatti [8] sur la notion de "Noyau-Périphérie". Dans cet exemple, les sommets dit noyaux (groupes sémantiquement indivisibles d'éléments), ceux appartenant à la périphérie ainsi que ceux appartenant aux différents

niveaux intermédiaires (pour le modèle continu) sont le résultat d'une analyse structurelle du réseau ayant pour but de faire ressortir des classes d'éléments sémantiquement significatifs. Ces classes sont néanmoins estimées indépendamment des types d'origine des sommets alors qu'ils pourraient pourtant avoir une importance élevée lors d'une analyse sémantique du réseau. D'une manière plus générale, les études basées sur du partitionnement [35,58] et autres analyses structurelles [54] ne différencient pas nécessairement les types de sommets et exploitent donc le réseau multi-couche comme sil était mono-couche.

Il existe néanmoins des cas d'utilisation du caractère multiplexe des réseaux. C'est par exemple le cas des travaux de Perer et Shneiderman [65]. Ceux-ci y proposent un système de visualisation permettant de trier et filtrer les sommets en fonction de métriques (degrés, différents types de centralité, barycentres, etc.) afin d'obtenir une liste ordonnée de sommets. Ceux-ci sont ensuite affichés à travers une vue noeud-lien personnalisable où l'utilisateur peut décider d'afficher uniquement certaines couches (donc uniquement les sommets ayant un certain type de lien) ainsi que des sommets ayant un certain classement ou score via les métriques. Le caractère multi-couche du réseau est ici utilisé mais comme filtre visuel et non pour établir la pertinence des sommets à afficher.

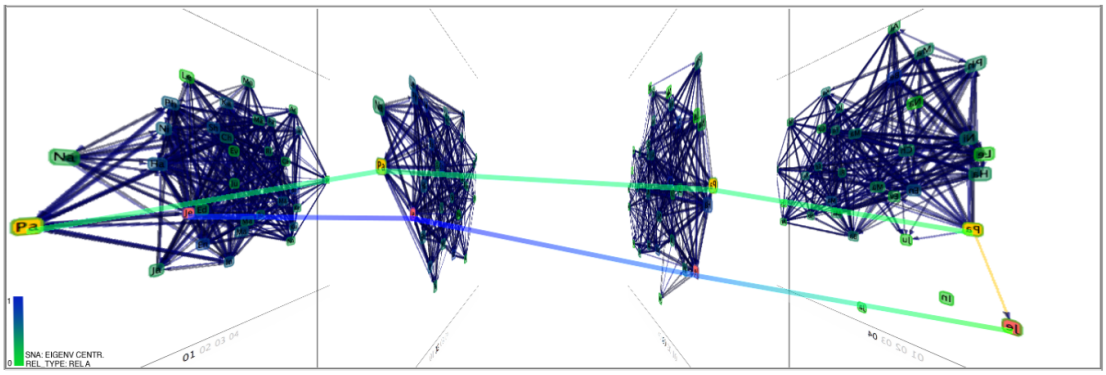


FIGURE *3.4* – ***Visualisation en 2.5D*** *: Ceci est une visualisation en 2,5D de quatre ensembles de sommets d'un réseau dynamique. Chacun représente une couche temporelle différente. Les liens entre les différentes couches permettent de suivre l'évolution des sommets à travers les différentes périodes de temps. Image provenant de [26].*

Un autre exemple ne référençant pas directement l'aspect multi-couche mais pouvant néanmoins l'utiliser est l'approche de Federico et al. [26] sur les réseaux

dynamiques. Dans ces travaux, une visualisation en 2.5D (Figure 3.4) représente plusieurs ensembles de sommets correspondant chacun à des périodes de temps. Les liens entre ces ensembles de sommets donnent alors une indication de l'évolution temporelle des sommets en suivant leurs présences dans les différents ensembles. On peut concevoir ces ensembles comme autant de couches dont le type est déterminé par leurs dates. Ce faisant, on peut établir une visualisation des couches et des liens inter-couches d'un réseau. Une adaptation aux graphes multi-couches de la visualisation en 2,5D a d'ailleurs été réalisée par un partenaire du projet BLIZAAR dont un aperçu est accessible sur le site du projet (https://blizaar.list.lu/doku.php?id=eisti_tool).

Comme nous venons de voir, plusieurs exemples exploitant plus ou moins étroitement le coté multi-couche des réseaux sociaux existent. Cependant, notre méthode a besoin de pouvoir utiliser directement cette particularité pour impacter les visualisations. La solution adoptée doit donc d'une part pouvoir être spécialisée dans l'analyse de réseau et d'autre part prendre en compte tous les types (et donc couches) des sommets pour définir une visualisation qui sera pertinente pour l'utilisateur.

Enfin, il reste à déterminer comment explorer ces réseaux. Notre méthode doit donc utiliser les particularités des sciences humaines et sociales afin de définir une manière de naviguer et d'interagir permettant une exploration efficace. C'est ce que nous détaillons dans la partie suivante.

3.4 Exploration et sciences humaines et sociales

L'exploration est le but initial de nos travaux. Nos collègues ayant de vastes jeux de données, il est nécessaire de mettre au point un mécanisme permettant de naviguer efficacement et sans se perdre dans les données. S'il existe déjà des outils et méthodes pour visualiser des réseaux multi-couches [21,22,56], il y a néanmoins peu de méthodes spécialisées à la fois pour l'exploration et les sciences humaines et sociales comme l'attestent Perer et Shneiderman en conclusion de leurs travaux précédemment cités [65]. Pour répondre à ce problème et mettre au point une méthode adaptée aux besoins des experts des données, il est nécessaire de concevoir notre méthode en prenant en compte tant les spécificités du domaine que leur méthodologie. Nous avons vu précédemment qu'une attention toute particulière est placée à l'échelle de l'individu ou du groupe d'individus plutôt que sur le réseau lui même. L'exploration, comme l'analyse, se doit donc de suivre ce schéma est de pouvoir être menée directement à partir des éléments du réseau et de leurs

détails. Ceci fait écho à une approche que nous avons vue dans une des parties précédentes (Section 3.1), le *bottom-up*.

Les approches *bottom-up* permettent d'initier analyses et explorations au plus près des données afin de pouvoir à posteriori élargir le spectre d'analyse à des communautés, sous-réseau voir réseau tout entier. Les travaux de S.van den Elzen et J.J. van Wijk [75], précédemment cités, s'ancrent dans cette dynamique en permettant la génération d'une vue agrégée de l'ensemble des données en fonction des sélections effectuées à l'échelle des individus. Ainsi, en prenant connaissance des éléments et de leurs interconnexions à l'échelle atomique, l'utilisateur va forger par ses interactions une vue plus globale permettant de comprendre la structure du réseau à l'échelle globale. Un autre exemple plus amplement lié aux sciences humaines et sociales est l'étude menée par Ghanie et al [29]. Le cadre de ces travaux est très similaire au notre : une coopération avec des experts en sciences humaines et sociales où les données sont modélisables sous forme de réseau multicouche. Une nécessité de simplification visuelle est également nécessaire afin de pouvoir créer des visualisations utiles et exploitables par les experts des données. Parmi les différentes alternatives présentées afin de réduire la complexité visuelle, se trouve notamment le "diviser pour mieux régner" (*divide and conquer*) : subdiviser le problème ou les données à visualiser en sous-ensembles de tailles réduites jusqu'à obtenir un résultat compréhensible et analysable. Ce concept synergise particulièrement bien avec les sciences humaines et sociales en répondant d'une part au problème de complexité visuelle et en permettant d'autre part d'axer la visualisation sur une échelle proche des individus et groupes d'individus.

La méthode utilisée, PNLBs ("*Parallel Node-Link bands*"), est d'ailleurs similaire dans son concept aux travaux menés par notre collègue du projet BLIZAAR (dont un aperçu est disponible sur le wiki officiel : `https://blizaar.list.lu/doku.php?id=list_tool`). Le réseau est présenté à l'expert sous la forme de listes de sommets inter-connectées entre elles à la manière de coordonnées parallèles où les listes représentent les différentes couches du réseau. Encore une fois, la visualisation est essentiellement menée à l'échelle des sommets et permet de fait de montrer une différence au niveau des paradigmes habituellement utilisés. C'est une des particularités de l'approche *bottom-up*, les mantras habituels tel que le "Overview first, zoom and filter, then details-on-demand" de Shneiderman [70] ou le "Analyze first, Show the Important, Zoom, filter and analyze further, Details on demand" de Keim [39] ne sont pas adaptés. Plus que de simples différences de méthodes, le *bottom-up* induit un changement au niveau même du paradigme utilisé.

Un certain nombre de travaux ont d'ailleurs proposé leurs propres versions

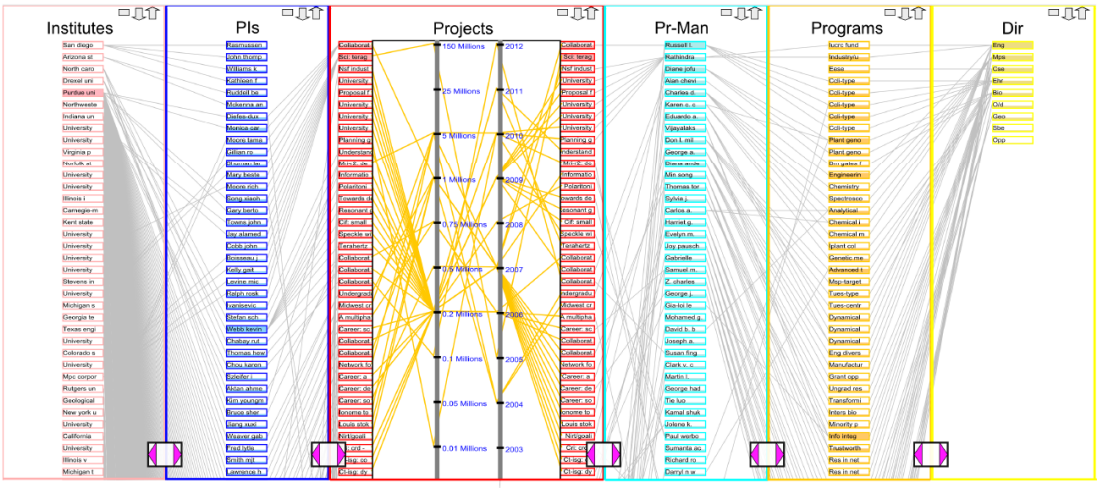


FIGURE *3.5* – ***PNLBs : visualisation par listes***. *Cette visualisation présente les couches du réseau sous forme de listes inter-connectées, mettant ainsi l'accent sur les connexions inter-couches. Image provenant de [29].*

des mantras de visualisations ré-adaptés au *bottom-up* comme le "Details-first, show context, overview last" de Luciani et al [52] (détailler en premier, montrer le contexte et finir par une vue générale) ou le "Search, show context, expand on demand" (chercher, montrer du contexte, étendre à la demande) de Van Ham et Perer [76] dont nous parlerons plus en détail dans le chapitre 5. Nos travaux s'inscrivent dans ce dernier par la proximité avec les habitudes des experts : à partir d'un point de départ connu (ex : une personnalité politique, un évènement historique, une communauté sociale donnée, etc.), on enrichit sémantiquement le contexte (on met en valeur les éléments environnants, les connexions à des groupes significatifs, etc.) puis on étend l'analyse si nécessaire (à des pistes connexes, transversales ou simplement en enrichissant sémantiquement encore davantage le contexte actuel).

Si le *bottom-up* est parfaitement en accord avec le fondement des sciences humaines et sociales, il reste néanmoins un aspect de leur méthodologie qui doit être impérativement pris en compte. L'exploration en sciences humaines et sociales prend souvent la forme d'un tâtonnement : on part d'un élément défini sans savoir précisément à l'avance vers où notre piste va nous mener. Dans certains cas, une piste peut mener vers d'autres pistes et le processus d'exploration est alors à recommencer à partir d'un nouveau référentiel pour explorer ces nouvelles pistes, et ainsi de suite... La méthode d'exploration doit donc prendre en compte une dé-

marche itérative où l'on doit pouvoir essayer et ré-essayer de visualiser et explorer différentes pistes voir même prendre en compte l'aspect arborescent d'une telle procédure, chaque piste pouvant mener à d'autres pistes.

3.5 Vers la méthode M-QuBE^3

A notre connaissance, une méthode proposant à la fois une **échelle centrée autour de l'individu**, **une gestion du caractère multi-couche des réseaux utilisés en sciences humaines et sociales** et **une gestion de l'aspect itératif et arborescent de la méthode de travail du domaine**, n'a pas encore été développée. Pour répondre à ce besoin, nous proposons notre méthode, `M-QuBE`^{3}, conçue pour l'exploration et la visualisation de réseaux multi-couches en sciences humaines et sociales.

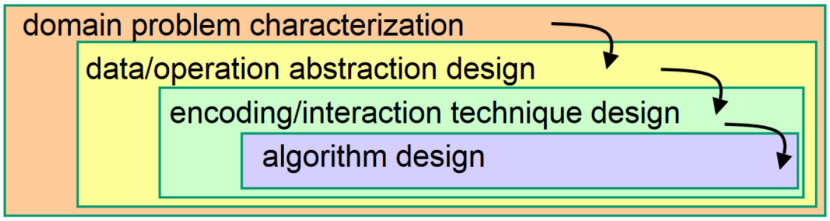


FIGURE *3.6* – ***Nested model*** *: Ce modèle se compose de quatre couches imbriquées décrivant les différentes étapes nécessaires à la conception d'une visualisation performante. L'objectif est de mener une réflexion "en entonnoir" où l'on commence par formaliser le problème puis analyser de plus en plus en détail le contexte et les besoins jusqu'au design final de l'algorithme à utiliser. Image provenant de [61].*

La conception de `M-QuBE`^{3} suit le "nested model" de Munzner (voir Figure 3.6). Ce modèle propose d'expliciter les différentes étapes nécessaires pour produire une visualisation en accord aux besoins de l'utilisateur tout en évitant les erreurs majeures pouvant y survenir (incompréhension des besoins ou du domaine, présentation des mauvaises informations et de la mauvaise manière, algorithme inefficace). La première étape est de prendre connaissance des spécificités des problèmes et des données de l'utilisateur. Cette formalisation s'établit notamment vis à vis des objectifs personnels de l'utilisateur mais aussi du domaine dans lequel il évolue. Cette étape a été réalisée directement avec nos collègues historiens afin d'avoir l'idée la plus précise possible du contexte et des challenges auxquels il est nécessaire de répondre. Dans une seconde et troisième étape, on définit une abstraction

sur les types de données et les opérations qui vont être utilisées pour répondre à ces objectifs avant d'établir la charte graphique, les techniques et les interactions qui vont donner corps aux opérations souhaitées. Ces étapes ont été réalisées en travaillant directement à partir des données (voir chapitre 2) déjà utilisées par le CVCE (https://www.cvce.eu/) et toujours en étroite collaboration avec un expert des données afin de s'assurer de répondre efficacement aux particularités du domaine. Enfin, les algorithmes qui générant la visualisation sont conçus, implémentés et validés.

Ces étapes sont capitales au bon déroulement de toute visualisation : s'appuyant directement sur les données et les contraintes du domaine, elles permettent de conceptualiser et développer la visualisation au plus près des besoins de l'utilisateur. La méthode M-QuBE^3 répond ainsi à l'échelle centrée sur les individus par une approche *bottom-up* et *divide and conquer* où l'aspect itératif et arborescent de la méthode de travail des experts est exploitée par M-QuBE^3 via une boucle d'interaction elle même itérative dont les traces arborescentes sont conservées et exploitables (voir Chapitre 4). Le caractère multi-couche est quant à lui utilisé à travers eScore, une mesure d'intérêt prenant en considération les différentes couches du réseau, permettant ainsi à M-QuBE^3 de pouvoir répondre à l'entièreté des besoins du domaine (voir Chapitre 5). Enfin, M-QuBE^3 et eScore sont évaluées par des experts des données les découvrant pour la première fois afin de juger leur efficacité à exploiter et explorer les réseaux multi-couches en sciences humaines et sociales (voir Chapitre 6).

Chapitre 4

M-Qube^3, une méthode d'exploration itérative par extractions successives de vues partielles

Sommaire

Dans le chapitre précédent, nous avons explicité les spécificités des sciences humaines et sociales. A partir de cela, nous dégageons les éléments auxquels notre méthode doit apporter une solution :

- Une attention centrée essentiellement sur les individus et leurs interconnexions en réduisant l'échelle du graphe à un objet lisible et analysable par l'utilisateur sans passer par une agrégation ou une simplification des données.
- Une méthode de travail nécessitant de créer en parallèle une arborescence de pistes dans lesquelles il est nécessaire de pouvoir naviguer.

- Une gestion du caractère multi-couche des réseaux en sciences humaines et sociales afin de pouvoir différencier les traitements en fonction des types de données considérées.

Nous introduisons ainsi `M-QuBE`[3] (**M**ultilayer network : **Qu**erying **B**ig networks by **E**volutive **E**xtraction and **E**xploration), une méthode permettant la construction d'une succession arborescente de sous-réseaux pour convenir à la méthodologie des experts des données, en offrant la possibilité, à n'importe quel moment du processus, d'affiner le chemin suivi jusqu'à présent pour parfaire son analyse, de revenir à un état antérieur pour essayer d'autres chemins ou simplement de décider d'explorer de nouvelles pistes découlant d'essais antérieurs.

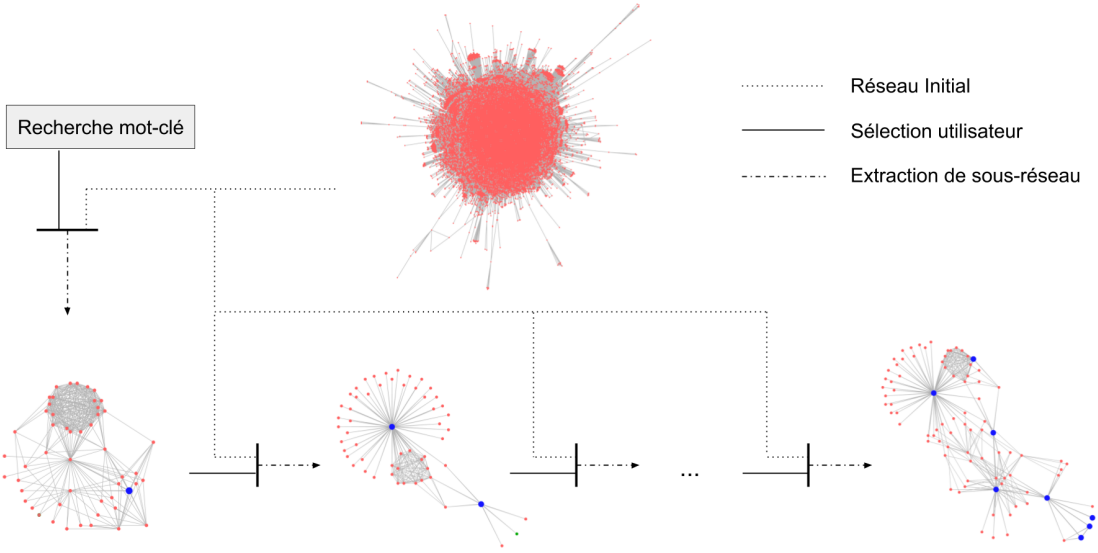


FIGURE *4.1* – ***Fonctionnement itératif du processus*** *: Création d'une série de sous-graphes. Les sommets bleus sont sélectionnés par l'utilisateur. Chaque sélection utilisateur permet la création d'un nouveau sous-graphe d'intérêt supérieur. Ce dernier est issu du graphe initial et prend en compte les informations sélectionnées dans les sous-graphes précédents.*

Le fonctionnement général de ce processus, comme illustré dans la Fig. 4.1, consiste à utiliser l'interaction de l'utilisateur (sélection de sommet ou recherche par mot-clé) pour créer, à partir du réseau initial, une succession potentiellement arborescente de sous-réseaux de plus en plus pertinents pour l'utilisateur.

Présentons une illustration concrète de cette méthode (Fig. 4.2) utilisant les données du CVCE (et la configuration décrite dans la section 6.1.3). Dans ce

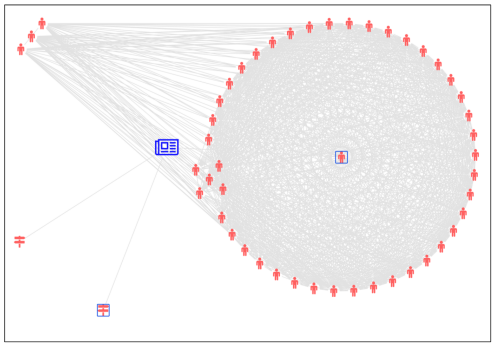
scénario, nous souhaitons évaluer l'influence du Royaume-Uni dans l'histoire de la construction européenne. Notre point de départ est la photo d'un meeting ayant eu lieu à Londres entre Margaret Thatcher et Helmut Schmidt témoignant d'une interaction entre le Royaume-Uni et un représentant de l'Union européenne. Un premier sous-réseau est généré à partir de la sélection de ce document (Fig. 4.2a). Dans ce sous-réseau, nous sélectionnons M.Thatcher et Londres pour valoriser les éléments du réseau en lien avec l'une des plus célèbres politiciennes anglaises ou un lieu phare du Royaume-Uni.

Nous obtenons un nouveau sous-réseau avec un nouveau contexte cette fois centré sur Margaret Thatcher, Londres et le document original. Nous constatons l'apparition d'un nouveau document sur l'entrée d'un nouveau pays dans l'Union européenne et de nombreuses personnalités politiques étant liées à ce document. Parmi eux figurent Jacques Delors et Pierre Werner qui sont liés à la fois à Margaret Thatcher ainsi qu'à plusieurs autres personnes en lien avec elle. Nous les sélectionnons afin d'orienter le contexte sémantique vers davantage d'informations relatives aux acteurs majeurs de l'Europe afin de pouvoir les corréler avec les informations relatives à M.Thatcher (Fig. 4.2b).

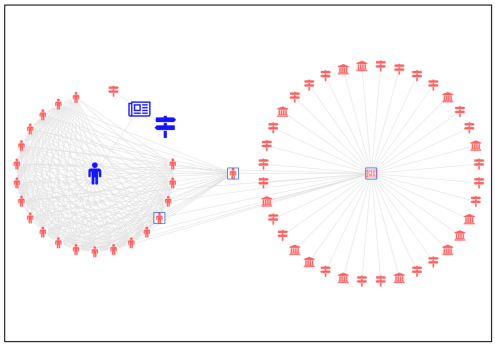
Le nouveau sous-réseau généré fournit une large gamme de nouvelles informations (Fig. 4.2c). Des documents relatifs aux trois acteurs sélectionnés apparaissent et concernent des sujets directement liés à notre objectif tels que le référendum anglais de 1975 ou des commentaires sur la vision de l'Europe de M.Thatcher. On peut également noter que ces documents estimés pertinents sont liés à des institutions européennes telles que la Commission européenne ou la Communauté économique européenne (CEE) qui figurent également dans ce nouveau sous-réseau. La sélection de ces entités va permettre d'orienter les documents et les personnalités et donc de faire évoluer à nouveau le contexte sémantique.

Nous décidons plutôt d'explorer une nouvelle voie car notre curiosité nous pousse à faire évoluer notre objectif initial. Pour ce faire, nous ne retenons de la sélection actuelle que M.Thatcher et nous sélectionnons un sommet représentant la République française, le Président français ainsi qu'un article francophone très critique envers l'Europe. Le nouveau sous-réseau généré (Fig. 4.2d) est entièrement construit à partir de la nouvelle sélection et offre ainsi un nouvel horizon de recherche en accord avec cette nouvelle voie.

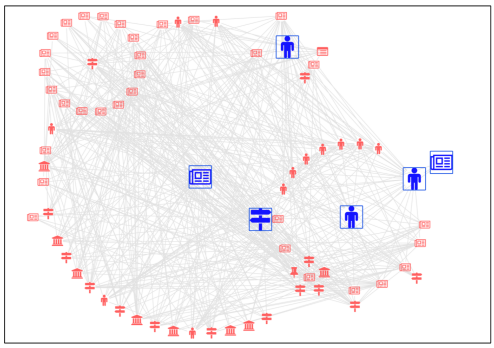
Avec `M-QuBE`^{3}, les experts sont donc en mesure d'explorer et de guider leurs explorations à travers un large réseau simplement en analysant les successions de sous-réseaux de tailles réduites. Outre le processus de génération de sous-réseaux et les interactions utilisateur qu'il utilise, il est aussi nécessaire de proposer aux



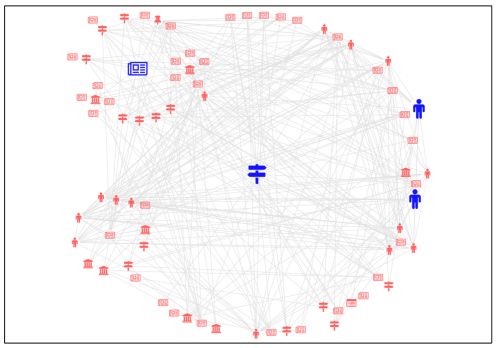
(a) Commençons notre exploration à partir d'un document sur un meeting européen à Londres (icône journal bleue). Nous sélectionnons Margaret Thatcher (icône personne encadrée en bleu) et Londres (icône pancarte encadrée en bleu) dans ce premier sous-réseau.



(b) Un second sous-réseau est généré à partir de la sélection réalisée en a) (les trois sommets agrandis bleus). Dans ce nouveau sous-réseau, nous sélectionnons un document sur l'entrée dans l'Europe d'un nouveau pays ainsi que deux acteurs majeurs de la scène européenne (Jacques Delors et Pierre Werner), tous liés à M.Thatcher et encadrés en bleu.



(c) Une nouvelle piste est explorée en ne conservant de la sélection actuelle uniquement Margaret Thatcher et en sélectionnant à la place les sommets représentant le rôle de président français, la république de France et un document traitant de la "Décadence Européenne".



(d) Ce nouveau sous-réseau présente un horizon de recherche entièrement nouveau calculé à partir de cette nouvelle sélection. Il est possible d'accéder à de nouveaux documents, nouvelles localisations, nouvelles institutions et nouvelles personnes liés directement ou indirectement à notre nouvelle sélection.

FIGURE *4.2 – Le rôle du Royaume-Uni dans le développement européen*

utilisateurs un moyen d'interagir et de naviguer dans la succession arborescente de sous-réseaux générés correspondant aux différentes pistes ayant été suivies (où "arbre de trace").

La méthode M-QuBE^3 se divise donc en deux parties : d'une part le processus de génération des sous-réseaux (Section 4.1), d'autre part la gestion de cette succession arborescente de sous-réseaux (Section 4.2).

4.1 Génération itérative de sous-réseaux

Le processus de génération de sous-réseaux de M-QuBE^3 est lui-même divisé en trois phases distinctes (Fig. 4.3). Le processus commence par une recherche par mot clé (partie A) afin de sélectionner des sommets. Ces sommets composent l'*ensemble focus* initial : une liste de sommets de référence permettant de définir une liste de sommets candidats qui seront potentiellement affichés dans le prochain sous-graphe à extraire. Un score est calculé pour ces sommets candidats en considérant les informations sémantiques des données ainsi que la structure du réseau afin d'obtenir une estimation d'intérêt pour l'utilisateur (partie B). A partir de ces scores, un classement est effectué pour déterminer une liste des sommets les plus intéressants (liste des élus) qui sont ensuite utilisés pour extraire le sous-réseau qui sera présenté à l'utilisateur (partie C).

Les experts commençant avec un détail et progressant pas à pas, le processus est structuré de manière similaire. Les étapes de sélection et d'extraction peuvent donc être répétées itérativement afin d'explorer les données plus en détail et avec une précision croissante.

L'utilisateur interagit avec le sous-graphe extrait en sélectionnant de nouveaux sommets qui lui semblent pertinents. Ces sommets vont venir enrichir l'ensemble focus et ainsi améliorer le prochain sous-réseau. Cet ensemble est conservé à travers les itérations et utilisé pour le calcul du prochain sous-réseau. Nous détaillons dans la suite les différentes étapes suivies pour chaque itération du processus.

Dans les parties suivantes, les figures représentant en détail les parties A, B et C (Fig. 4.4, 4.5 et 4.6) sont toutes trois extraites de la figure globale de M-QuBE^3 (Fig. 4.7). Les positionnements des éléments au sein des figures sont donc choisis afin de pouvoir être rassemblés dans la figure globale.

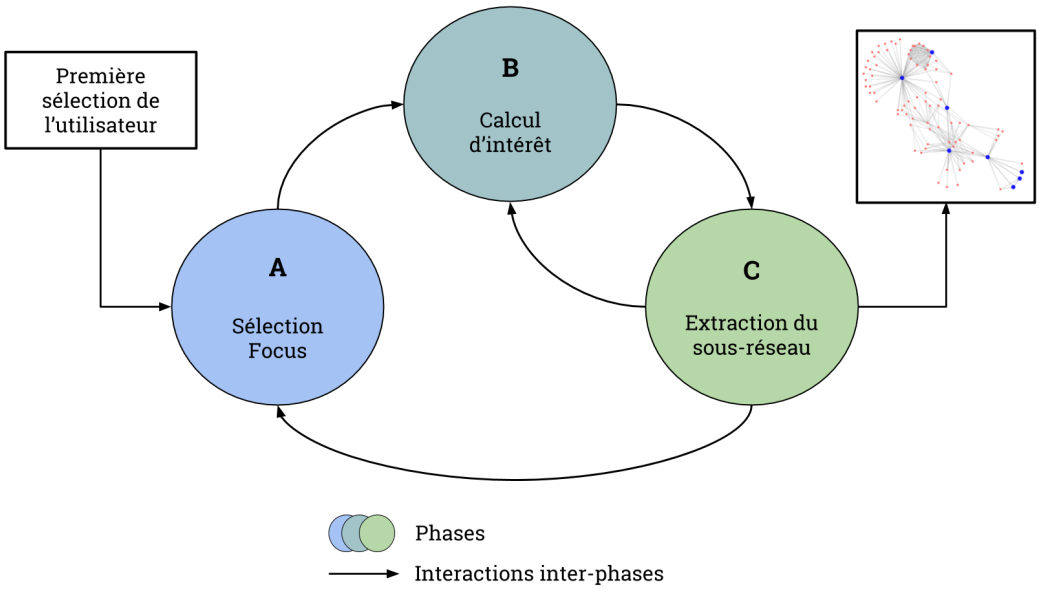


FIGURE *4.3* – ***Interactions entre les différentes phases du processus de génération de sous-réseaux de*** `M-QuBE`^{3} *: Pour commencer le processus, une première sélection est effectuée. A partir de cette sélection (A), plusieurs calculs d'intérêt vont être effectués dans le réseau (B) afin d'établir un classement entre les sommets et déterminer le plus intéressant pour l'utilisateur (C). Le sommet le plus intéressant est conservé dans un sous-réseau et, si celui-ci n'est pas de taille suffisante, un calcul est ré-itéré (B) en considérant de nouveaux éléments. Lorsque le sous-réseau est de taille satisfaisante pour l'utilisateur, celui-ci est montré et l'utilisateur peut alors ré-itérer le processus complet en établissant une nouvelle sélection (A).*

4.1.1 Sélection de l'ensemble focus

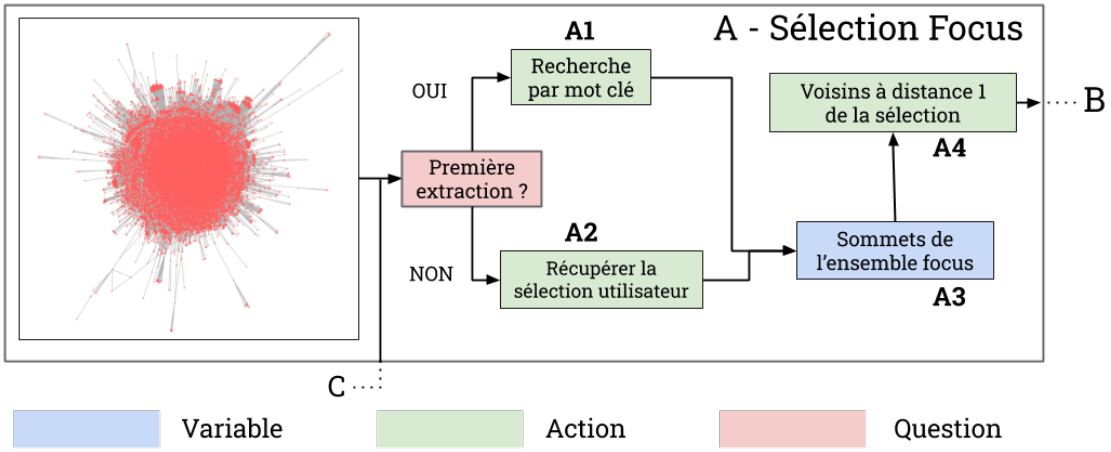


FIGURE *4.4* – ***Phase de sélection de l'ensemble focus*** *: La phase de sélection permet de définir l'ensemble focus, un ensemble de sommets de référence à partir duquel les sommets à évaluer sont déterminés. S'il s'agit du premier passage dans la phase A, la sélection s'effectue par recherche de mot-clé. Le cas échéant, la sélection sera définie par l'utilisateur en interagissant avec la vue d'un sous-réseau (par une zone de sélection ou un clic par exemple). Le voisinage des sommets focus sont ensuite transmis à la phase B.*

Comme énoncé précédemment, l'ensemble focus est l'ensemble des sommets sélectionnés par l'utilisateur. Dans un premier temps, il doit être initialisé ou être mis à jour. S'il s'agit de la première itération, l'ensemble focus (Fig. 4.4, A3) est défini par une recherche par mot clé d'un ou plusieurs sommets (A1). Ensuite, entre chaque nouvelle itération, l'ensemble focus est modifié par l'ajout ou la suppression de sommets par l'utilisateur (A2). Les voisins de chaque sommet de l'ensemble focus (A4) composent ensuite un ensemble appelé la liste de candidats (voir section suivante : Fig. 4.5, B2). Cette liste contient les sommets potentiellement montrés dans le prochain sous-réseau extrait, en fonction de l'estimation d'intérêt calculée dans le partie B. Cette liste de candidats est amenée à évoluer au cours de processus.

La prochaine phase est le calcul des différentes métriques utilisées pour déterminer le score d'intérêt final.

4.1.2 Calcul d'intérêt

La phase de calcul d'intérêt détermine un score (Fig. 4.5, B7) représentant l'intérêt de l'utilisateur pour un sommet donné n. Plus le score est élevé, plus la probabilité que le sommet n soit sélectionné et montré à l'utilisateur dans le prochain sous-réseau est grande (Fig. 4.6, C6).

Le score final est une métrique définie à partir de plusieurs scores. Premièrement, un score directement lié à la volonté utilisateur, l'*eScore* (B3), ainsi qu'un score défini en fonction de la position des sommets dans le réseau, *pScore* (B4), sont calculés. Ces deux scores sont ensuite combinés en un score pondéré, le *wScore* (B5). Le score final est finalement obtenu en prenant en compte les scores des voisins (B3' à B5') à travers un calcul de diffusion de l'intérêt (B6).

eScore (B3). L'eScore (**e**xploratory **Score**) est une métrique d'estimation de l'intérêt calculée pour tous les sommets dans la liste de candidats (B1). L'objectif est de représenter la volonté générale de l'utilisateur notamment les contraintes et les objectifs à prendre en compte lors de la recherche. Lors d'une première utilisation de `M-QuBE`^{3} sur un nouveau jeu de données ou en cas d'objectifs imprécis ou encore non-définis de la part de l'utilisateur, il est tout à fait possible d'utiliser des métriques communément utilisées en sciences humaines et sociales (comme les exemples énoncés dans les travaux de Mainas [54] appliqués aux réseaux criminels). Notre solution spécialisée d'eScore pour les réseaux multi-couches est présentée dans la section 5.2 du chapitre 5 pour adapter précisément `M-QuBE`^{3}aux réseaux issus des sciences humaines et sociales.

pScore (B4). L'utilisateur interagit avec le processus en sélectionnant des sommets à chaque itération (Partie A, Fig. 4.4). Nous supposons que les sommets proches d'un sommet sélectionné dans le graphe ont plus de chances d'être considérés comme intéressants par l'utilisateur. A cette fin, la sélection de l'utilisateur constitue ce que nous appelons une zone focale et un sommet inclus ou proche de cette zone est pondéré positivement (voir le calcul du score pondéré ci-dessous). Pour ce faire, une fonction basée sur le centroïde est utilisée pour calculer la distance moyenne entre le sommet évalué x et les sommet de l'ensemble focus, déterminant ainsi sa position dans la zone focale.

Cette fonction est définie ainsi :

$$C(x, Y) = \frac{\sum_{y \in Y} d(x, y)}{|Y|}$$

avec Y l'ensemble focus et d une fonction de distance. La fonction la plus pertinente pour d indépendamment du contexte est souvent un calcul du plus court chemin entre deux sommets. La distance euclidienne peut également être utilisée, mais les coordonnées des sommets issues de l'algorithme de dessin du graphe doivent avoir un sens, ce qui nécessite d'abord un travail sur le modèle et la représentation du réseau.

$pScore(x, Y)$ est la distance normalisée entre le sommet x et les sommets de l'ensemble Y :

$$pScore(x, Y) = 1 - \frac{C(x, Y) - c_{min}}{c_{max} - c_{min}}$$

où c_{min} et c_{max} sont respectivement la valeur minimum et maximum de C dans le réseau. Une normalisation est ici nécessaire pour pouvoir effectuer l'étape suivante en mettant à la même échelle les informations sémantiques (eScore) et les informations topologiques (pScore).

wScore (B5). Les métriques $eScore$ et $pScore$ sont combinées pour obtenir le score pondéré $wScore$ défini tel que :

$$wScore(x, Y) = (1 - w) \times eScore(x|Y) + w \times pScore(x, Y)$$

où w est une constante définie sur $[0; 1]$ en fonction de l'importance que veut donner l'utilisateur à la zone focale. $wScore$ est donc l'estimation d'intérêt d'un sommet pour l'utilisateur en tenant compte à la fois de la sémantique (eScore) et des informations structurelles du réseau (pScore).

Diffusion (B6). Le calcul de l'intérêt des sommets se termine par la phase de diffusion. Un problème possible de ce processus est le même que celui rencontré dans les travaux de van Ham et Perer [76]. La liste des sommets candidats s'étend itérativement à la manière d'un algorithme glouton : lorsqu'un sommet est sélectionné pour être montré dans le sous-réseau, ses sommets voisins sont ajoutés à la liste de candidats. Cependant, si un sommet très intéressant (un sommet avec un score élevé) est entouré de sommets avec un score faible, l'algorithme itératif

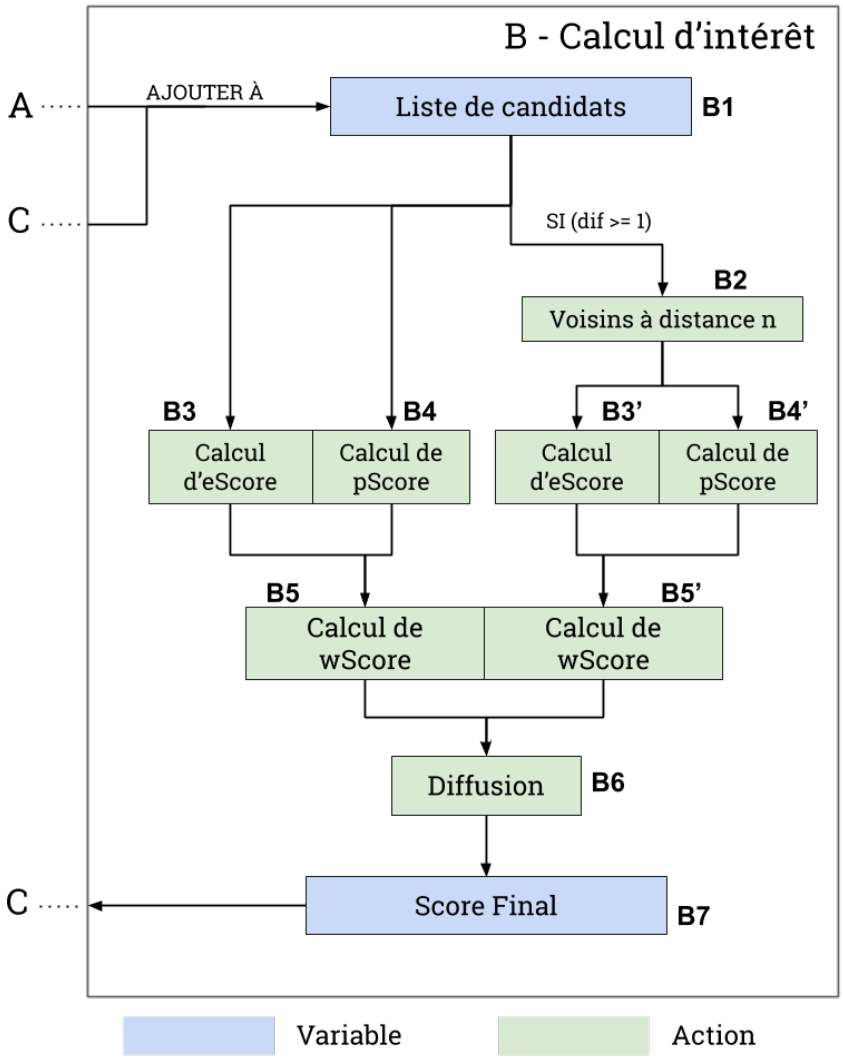


FIGURE 4.5 – ***Phase de calcul de l'intérêt*** *: Cette phase commence en initialisant une liste de candidat à partir des sommets reçus (B1). Si ceux-ci n'ont jamais été évalués, des scores basés sur des fonctions d'intérêt (B3) et sur la position des sommets dans le réseau (B4) sont calculés. A partir de celui-ci est défini un score pondéré (B5) qui sera ensuite impacté par son voisinage (B6) pour obtenir un score final (B7). La branche parallèle (B2) est utilisée pour calculer cette diffusion à partir du voisinage des candidats. Il faut alors calculer de la même manière les scores pour les voisins des candidats (B3',B4',B5') afin de finaliser la diffusion (B6). Une fois la phase B terminée, les scores finaux sont transmis à la phase C.*

peut ne jamais le sélectionner (puisque ses voisins peuvent ne jamais être sélectionnés). Afin d'éviter une situation où ces extrema locaux sont isolés, il est possible d'effectuer une diffusion du score d'intérêt.

La solution consiste pour chaque sommet intéressant à diffuser une partie de son intérêt à son voisinage. Pour ce faire, les utilisateurs sélectionnent un degré de diffusion *dif*. Plus *dif* est élevé, plus le diamètre du réseau extrait peut être large (si *dif* est nul, alors le mécanisme de diffusion n'est pas utilisé). Pour faire une diffusion de degré *dif* d'un sommet, il faut calculer le score des sommets non candidats à distance *dif* de celui-ci (B2). Ensuite, une fois le score calculé pour tous les sommets requis (B3',B4',B5'), chaque sommet gagne un pourcentage du $wScore$ du sommet le plus intéressant (i.e. le sommet avec le score le plus élevé) à distance *dif* ou moins (B6). Ce pourcentage est également déterminé par l'utilisateur. Plus il est élevé, plus les scores des sommets vont s'homogénéiser.

Ce mécanisme optionnel peut améliorer la pertinence des sous-réseaux d'intérêt. Cependant, un degré élevé de diffusion peut impacter négativement les performances du processus si le réseau est très connecté. De même, un pourcentage élevé de diffusion homogénéise les scores, ce qui diminue l'impact sur la sélection des scores d'intérêts et des classements qu'ils génèrent. Ce mécanisme doit donc être utilisé avec prudence. Par ailleurs, le fait qu'un sommet théoriquement intéressant soit isolé par son entourage peu intéressant peut ne pas déranger l'expert si celui-ci met l'accent dans son analyse sur les liens ou les communautés. Auquel cas, il n'est pas nécessaire de procéder à une étape de diffusion.

4.1.3 Extraction du sous-réseau

Cette phase commence par le calcul de la liste des sommets choisis (Fig. 4.6, C3). Cette liste contient les sommets sélectionnés dans la liste des candidats pour composer le nouveau sous-réseau. Les sommets sélectionnés par l'utilisateur sont automatiquement dans la liste des sommets choisis. La finalité du processus complet est donc de remplir la liste des sommets choisis en fonction des scores obtenus afin d'avoir un sous-réseau d'intérêt optimal pour l'utilisateur.

Un classement des sommets est effectué en utilisant les scores finaux des phases précédentes (C1) et le sommet ayant le score le plus élevé est ajouté à la liste des sommets choisis (C2). Si le nombre de sommets présent dans la liste des sommets choisis correspond au nombre souhaité par l'utilisateur (C4), on extrait le sous-réseau composé des sommets de la liste et des liens présents entre ces sommets dans le réseau initial (C6).

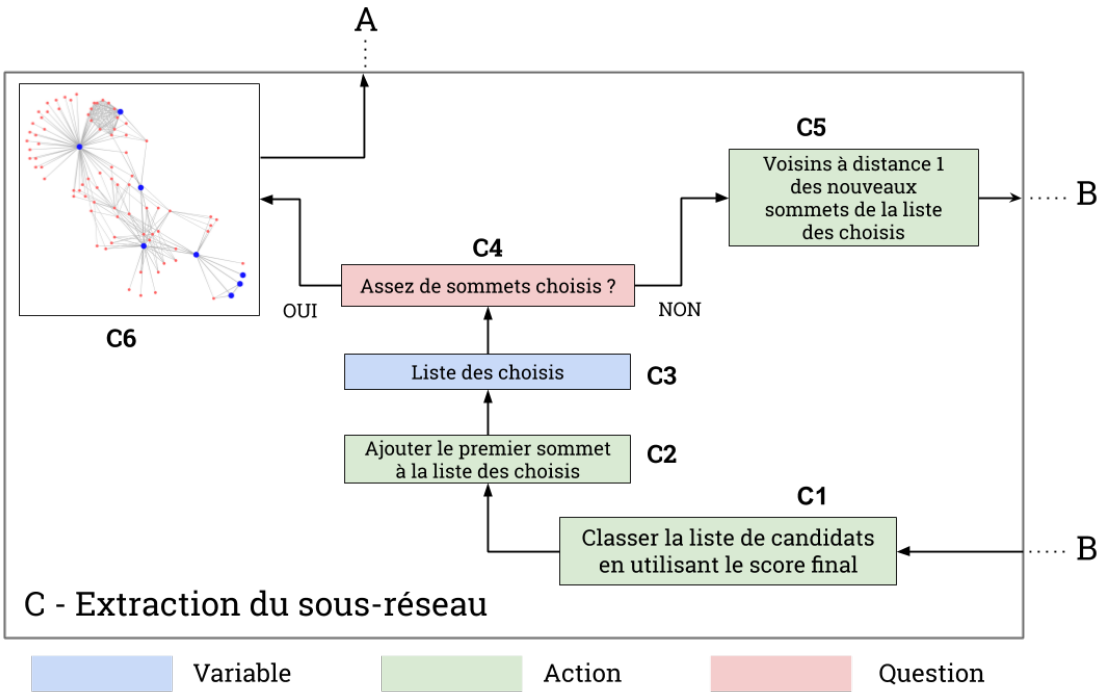


FIGURE *4.6* – ***Phase d'extraction du sous-réseau*** *: La phase C commence en effectuant un classement des sommets à partir de leurs scores finaux (C1). Le sommet en première place est ajouté à la liste des sommets choisis (C2). Cette liste (C3) correspond aux sommets qui seront présentés à l'utilisateur dans le prochain sous-réseau. La quantité de sommets présents dans la liste des choisis est évaluée (C4). Si elle ne contient pas assez de sommets par rapport au nombre souhaité par l'utilisateur, le voisinage du nouveau sommet choisi est ajouté à la liste des candidats (C5) et une nouvelle phase de calcul de l'intérêt commence (B) sur cette nouvelle liste. Si assez de sommets sont présents (C4), un sous-réseau composé des sommets de la liste des choisis et des liens entre eux est extrait et présenté à l'utilisateur (C6). A partir de ce sous-réseau, l'utilisateur pourra sélectionner de nouveaux sommets et ainsi recommencer la procédure à partir de la phase A.*

Si le nombre de sommets n'est pas suffisant, les sommets voisins du dernier sommet choisi sont ajoutés, s'ils n'y sont pas déjà (C5), à la liste des candidats. La mise à jour de la liste de candidats requiert alors que les nouveaux sommets n'ayant pas encore de scores soient traités. La procédure est alors répétée depuis B1. Les scores des sommets précédemment traités n'ont néanmoins donc besoin d'être ré-évalués car leurs scores ne subiront aucun changement.

4.1.4 Processus de génération complet

Une fois ces phases complétées (Fig. 4.7), un sous-réseau est généré. L'utilisateur peut alors sélectionner de nouveaux sommets qui lui semblent pertinents dans ce sous-réseau pour réitérer le processus depuis le début. Le processus complet recommence alors entièrement avec un nouvel ensemble focus enrichi des nouvelles sélections et dé-sélections de l'utilisateur (A2). De nouveaux scores sont calculés engendrant ainsi un nouveau sous-réseau sémantiquement plus proche des nouvelles indications de l'utilisateur. Cette procédure est réitérée jusqu'à satisfaction de l'utilisateur vis à vis des sous-réseaux obtenus.

La répétition de ces étapes permet de générer des sous-réseaux se succédant. Il est néanmoins nécessaire de rappeler que chaque sous-réseau est extrait du graphe initial et généré à partir des sélections utilisateurs provenant d'un sous-réseau père. Chaque sous-réseau ayant alors un sous-réseau père mais potentiellement plusieurs sous-réseaux fils, cette succession n'est pas linéaire et il est alors nécessaire de proposer un moyen de représenter et d'utiliser cette arborescence de sous-réseaux.

4.2 Arbre de traces

La méthodologie de travail de nos experts est éminemment arborescente : lorsqu'une piste d'analyse est suivie, il est fréquent que de nouvelles opportunités apparaissent et ouvrent la voie à de nouvelles pistes à explorer. La méthode `M-QuBE`[3] suit cette méthodologie en proposant à l'expert d'utiliser un "arbre de traces" interactif, inspiré par la notion d'*"history tree"* [14, 71].

Chaque action effectuée par l'utilisateur (sélection/dé-sélection d'un sommet, augmentation/diminution du nombre de sommets requis par sous-réseau ou changement de l'algorithme de positionnement des sommets) génère une "trace" dans l'arbre i.e. un sommet représentant le sous-réseau résultant de cette action. Ce sommet est lié à son père (le sous-réseau sur lequel on a effectué l'action) par une arête dont le type représente l'action effectuée (sélection, augmentation, positionnement).

Cet arbre ne se contente pas d'être descriptif mais propose à l'utilisateur un système de navigation entre tous les sous-réseaux obtenus. Lorsque l'utilisateur clique sur un sommet de l'arbre de trace, le sous-réseau correspondant est sélectionné et affiché. Toute nouvelle action de la part de l'utilisateur va alors générer un sous-réseau dont le père est le sous-réseau sélectionné, permettant ainsi la création d'une nouvelle branche sur l'arbre de trace (Fig. 4.8).

Ce mécanisme fait ainsi écho à la méthode de travail des experts en permettant

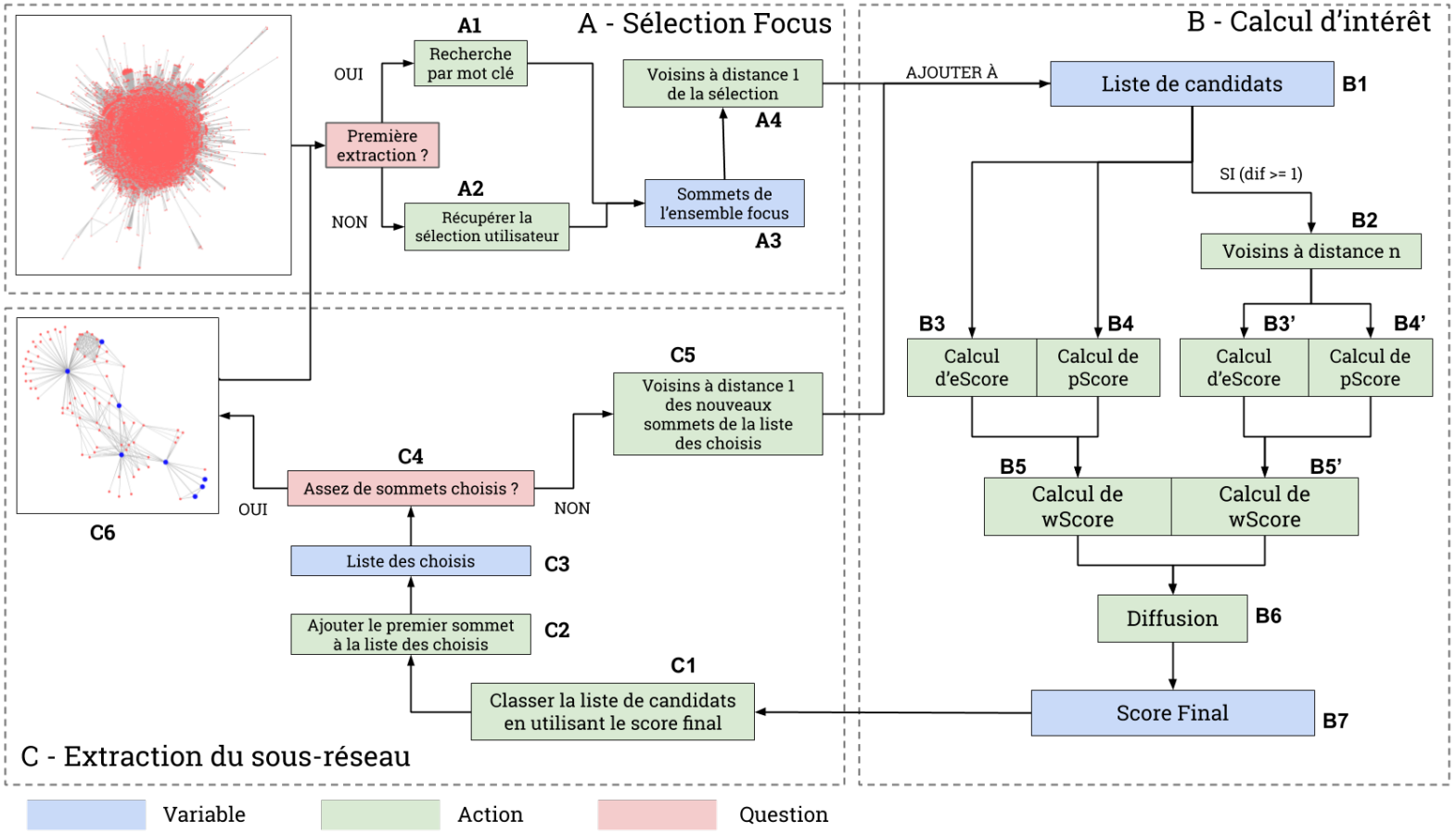


FIGURE *4.7* – ***Algorithme complet de*** `M-QuBE`^{3} *: Schéma non simplifié des interactions entre les différents phases du processus. Le processus se comporte comme un algorithme glouton, une boucle apparaît entre les phases B et C (de C5 vers B1) pour étendre l'évaluation des sommets au voisinage du sommet choisi (C2). Un seconde boucle apparaît entre la phase C et la phase A (de C6 vers A2) lorsque le processus* `M-QuBE`^{3} *est re-sollicité afin d'ajouter un nouveau sous-réseau à la succession de sous-réseaux déjà générés.*

de suivre et développer plusieurs pistes simultanément et a donc été particulièrement bien accueilli par les experts des données. Pour rendre cela possible sans pertes de performance et sans duplication de données entre les sous-réseaux, nous utilisons la structure de données Tulip [3] basée sur une hiérarchie de graphes. Ceci s'ancre dans la continuité des travaux commencés avec Porgy [66] dont le fonctionnement est aussi basé sur un graphe de graphes assuré par le modèle de données Tulip. Des informations supplémentaires quant à l'implémentation, l'emploi et les retours utilisateurs sont disponibles dans le chapitre 5.

4.3 Synthèse

La méthode M-QuBE^3 a pour objectif de répondre à plusieurs particularités des sciences humaines et sociales. Parce que les sciences humaines et sociales opèrent à une échelle proche des individus et groupes d'individus, M-QuBE^3 propose d'utiliser des vues partielles pour explorer le réseau global : en offrant en permanence des sous-réseaux facilement lisibles et analysables, chaque individu du réseau est accessible et peut servir à étendre l'exploration en offrant de nouveaux sous-réseaux construits autour de ses spécificités. Par ailleurs, tant par la création successive de sous-réseaux que les possibilités offertes par l'arbre de trace, M-QuBE^3 respecte la méthode de travail des experts des sciences humaines et sociales en proposant d'explorer différentes pistes simultanément, de revenir en arrière dans leurs recherches à tout moment et d'approfondir autant que souhaité les pistes qui semblent prometteuses.

Cependant, pour répondre aux trois points précédemment cités en introduction de ce chapitre, il reste encore à prendre en considération le caractère multi-couche inhérent aux réseaux des sciences humaines et sociales. Nous avons vu dans la section 4.1.2 que M-QuBE^3 est construit autour de différents calculs de scores dont l'eScore, un score d'intérêt orienté vers la sémantique du réseau. Dans le chapitre suivant, nous expliquons comment, par l'intermédiaire de ce score au centre du fonctionnement de M-QuBE^3, nous offrons la possibilité aux experts de ces réseaux de pouvoir orienter leurs explorations et recherches en prenant en compte l'importance sémantique des réseaux multi-couches.

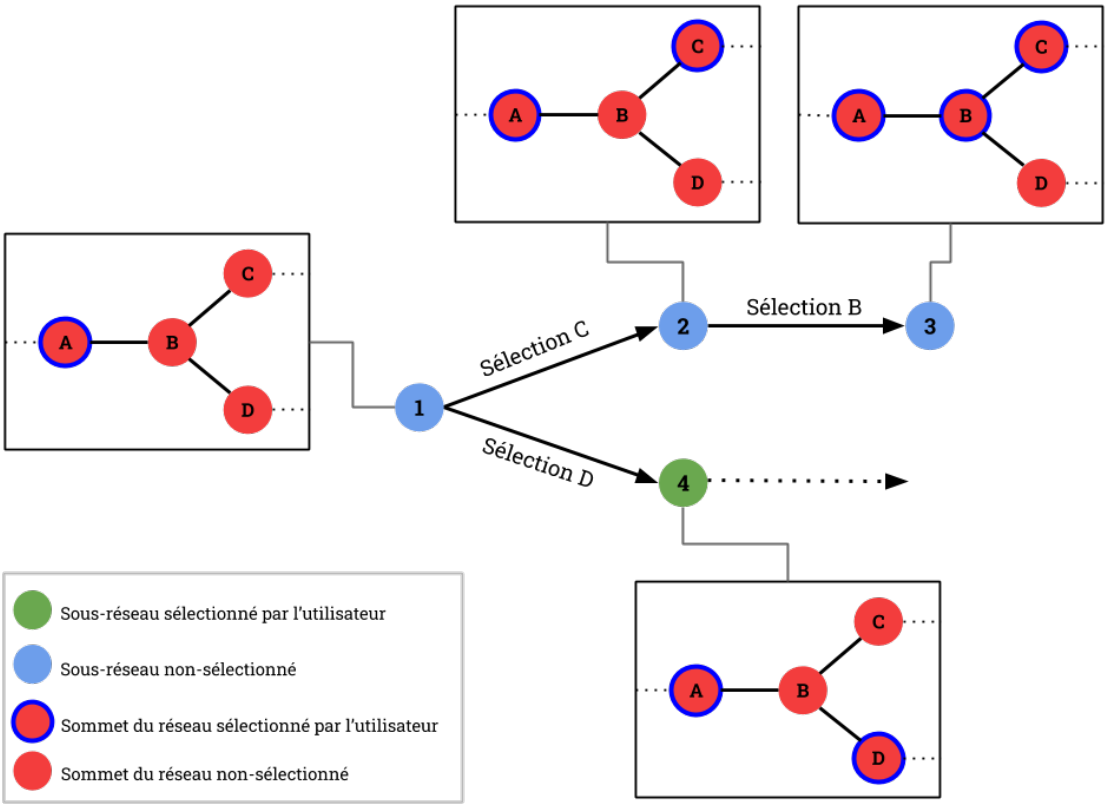


FIGURE *4.8* – ***Exemple d'arbre de traces*** *: Les sommets A,B,C,D représentent des éléments du réseau sur lequel* `M-QuBE`^{3} *est appliqué. Lorsqu'une action de l'utilisateur génère un nouveau sous-réseau, un sommet représentant ce sous-réseau est créé (les sommets 1,2,3,4). Ainsi le sommet 1 de l'arbre de trace représente le premier sous-réseau généré où seul le sommet A a été sélectionné par l'utilisateur (sommet sur-ligné en bleu). Par la suite, lorsque l'utilisateur a sélectionné le sommet C afin de générer un nouveau sous-réseau, le sommet 2 le représentant a été créé. L'arête entre son père (1) et lui indique l'action l'ayant généré. Le sommet vert représente le sous-réseau actuellement sélectionné et visualisé : une fois les réseaux 1,2,3 générés, l'utilisateur a cliqué sur le sommet 1 afin de le sélectionner et de créer une nouvelle branche en sélectionnant D au lieu de C. La sélection est maintenant sur son fils, le sommet 4, qui sera le père de la prochaine action effectuée par l'utilisateur.*

Chapitre 5

Estimation d'intérêt dans les réseaux multi-couches

Sommaire

Quantifier l'intérêt des éléments des structures de données n'est pas nouveau. De nombreux auteurs ont établi des métriques d'estimation d'intérêt à des fins de visualisation ou de navigation dans des contextes applicatifs précis. Pour les graphes multi-couches, le concept est encore plus pertinent tant cela entre en synergie avec la forte richesse sémantique de ces objets. Le concept s'est néanmoins développé progressivement avec en premier lieu des estimations sur des objets simples en ne prenant en compte que les informations directement accessibles, la topologie de la structure, avant de s'étendre à la sémantique.

Nous commençons donc par présenter les différents éléments marquants de la quantification de l'intérêt (Section 5.1) puis notre version spécialisée pour les réseaux multi-couches (Section 5.2).

5.1 Estimation de l'intérêt

L'estimation d'intérêt est aujourd'hui omniprésente dans énormément de domaines et fonctionnalités que nous utilisons chaque jour. De la recommandation de vidéos [20] aux moteurs de recherches [7, 19] en passant par l'"Eye-Tracking" [72], il est devenu capital dans la masse de données accessibles de pouvoir isoler ce qui est pertinent pour l'utilisateur. Pour répondre à ce problème, une solution fréquemment utilisée est d'établir un classement des éléments accessibles (comme les exemples précédents et ceux à venir). Il est alors nécessaire d'avoir un critère permettant de quantifier l'intérêt que l'utilisateur a de chaque élément afin de pouvoir les hiérarchiser.

Si la problématique est très actuelle (plateformes de vidéos à la demande, moteurs de recherche, streaming audio et vidéo, etc.), le concept de score d'intérêt ne l'est pas. Déjà en 1986, les travaux de Furnas [27] proposent une première approche de quantification de l'intérêt appelée $DOI_{fisheye}$ (*Degree of Interest Fisheye*) S'appliquant sur les arbres. Le $DOI_{fisheye}$ est alors utilisé afin de simplifier visuellement des fichiers textes hiérarchisés comme du code C en n'affichant que les lignes ayant un score d'intérêt suffisant. Pour ce faire, chaque ligne du texte correspond à un élément de l'arbre et la profondeur de la ligne dans la hiérarchie des blocs de code est équivalente à la profondeur dans l'arbre à partir de la racine.

Pour chaque sommet x de l'arbre (i.e. chaque ligne du code), le $DOI_{fisheye}$ est calculé ainsi :

$$DOI_{fisheye}(x|y) = API(x) - D(x, y)$$

où y est un sommet de référence appelé le sommet $focus$ représentant le point d'attention actuel de l'utilisateur (ici la ligne actuellement éditée ou étudiée), où $API(x)$ ("à priori") est une fonction qui représente l'intérêt "absolu" d'un élément de l'arborescence (indépendemment de y) pour l'utilisateur et où $D(x, y)$ est une fonction qui correspond à la distance entre les sommet x et y dans l'arbre. Dans cet exemple, API correspond à la distance du sommet x par rapport à la racine de l'arbre. La racine représentant le plus haut niveau de l'arborescence du code (le bloc de la fonction elle même), plus une ligne est intriquée dans des blocs moins son score est élevé. Ainsi, les instructions profondément enfouies dans des blocs de codes, eux même enfouis dans d'autres blocs, etc. sont jugées moins pertinentes pour l'utilisateur et seront donc moins susceptibles d'être montrées à l'utilisateur. Enfin, la fonction D a pour effet de renforcer le score final des éléments à proximité de l'élément focus y. Ainsi, plus un élément va être éloigné du focus, moins il sera

jugé intéressant. Avec ces deux fonctions, le $DOI_{fisheye}$ s'appuie sur la topologie de l'arbre afin de supposer une vue simplifiée plus intéressante pour l'utilisateur. Celui-ci fixe ensuite un seuil et toutes les lignes ne dépassant pas ce seuil sont agrégées et remplacées par un "..." dans la marge du texte visualisé final.

Si le $DOI_{fisheye}$ de Furnas concerne exclusivement les arbres, le concept a ensuite été étendu (comme pour le DOI Tree [16] interactif) ou appliqué à divers objets (ontologies [36], coordonnées parallèles [17], etc.) par d'autres auteurs. Parmi ces travaux, certains mettent l'accent sur la prise en compte de la sémantique des données en utilisant des variations du $DOI_{fisheye}$ [38,77] dont Van Ham et Perer ont ensuite proposé une généralisation appliquée aux graphes [76]. Dans cette généralisation, ils proposent une nouvelle version enrichie du DOI permettant d'extraire un sous-graphe pertinent pour un utilisateur à partir des informations topologiques et sémantiques d'un ensemble sinon difficile à analyser.

En plus de la distance et de la fonction API déterminées à partir des informations structurelles du graphe, Van Ham et Perer proposent d'utiliser une fonction UI (*User Interest*) basée sur la sémantique du graphe (mots-clés, tags, valeurs d'attributs, etc.). Ce DOI est ainsi défini :

$$DOI(x|y,z) = \alpha.API(x) + \beta.UI(x,z) + \gamma.D(x,y)$$

où UI utilise une fonction utilisateur z et où α, β et γ sont des constantes servant de levier afin de moduler l'importance des informations structurelles (API, D) ou sémantiques (UI) dans l'estimation de l'intérêt. La fonction z est une requête de l'utilisateur représentant son intérêt pour les informations sémantiques issues du sommet x. Plus z tend à être validée par les informations issues du sommet x plus le score de UI est élevé. Par exemple, il peut s'agir de la similarité d'un mot clé z avec un attribut texte du sommet x, une note par rapport à un attribut numérique, etc. L'essentiel pour une utilisation efficace du DOI est donc de faire correspondre au mieux les différentes fonctions aux besoins et objectifs des utilisateurs, ce qui peut être fait en utilisant les diverses métriques et indicateurs utilisés en sciences humaines et sociales (comme les exemples donnés dans les travaux de Mainas vus précédemment [54]).

Cependant, le DOI de Van Ham et Perer peut ne pas convenir dans le cas de graphes spécifiques à cause de sa généricité. C'est par exemple le cas des graphes dynamiques [1] dont une adaptation a été proposée afin de pouvoir bénéficier des spécificités de ces objets. De manière analogue, il est nécessaire d'adapter le calcul d'estimation de l'intérêt pour convenir aux réseaux multi-couches et à `M-QuBE`^{3} en prenant en compte à la fois les différents types de sommets (donc les couches) et

à la fois le caractère itératif de la méthode M-QuBE^3. C'est ce que nous proposons avec notre calcul d'intérêt spécialisé, l'eScore.

5.2 eScore, une métrique adaptée et appliquée aux réseaux multi-couches

L'eScore se propose comme une adaptation itérativement applicable pour les réseaux multi-couches inspirée du $DOI_{fisheye}$ de Furnas [27] et de la dimension sémantique de Van Ham et Perer [76]. Pour ce faire, un score est également calculé individuellement pour chacun des sommets du réseau mais en prenant cette fois en compte les informations sémantiques issues des couches et un ensemble de sommets focus évoluant en fonction des sélections de l'utilisateur.

Parce que M-QuBE^3 se veut interactif, eScore doit aussi se baser sur les choix et actions de l'utilisateur. Celui-ci va donc intervenir de deux manières afin d'impacter le calcul d'intérêt : dans un premier temps, et préalablement à toute procédure, à travers le choix d'un ensemble de contraintes et d'objectifs liés aux couches du réseau (les "fonctions de pilotage") et, dans un second temps, en exploitant les sommets jugés pertinents sélectionnés par l'utilisateur qui définissent l'"ensemble focus" (voir sous-section 4.1.1).

Dans M-QuBE^3, chaque nouvelle itération génère un nouveau sous-réseau à partir d'une nouvelle sélection. L'eScore va alors se comporter de la même manière en évoluant en fonction de chaque nouvelle sélection. L'utilisateur se voit ainsi proposer un nouveau sous-réseau donc le calcul a été impacté par eScore et donc la sélection d'un de ses sommets va impacter le prochain calcul d'eScore, impactant le prochain sous-réseau, etc.

A noter que la sélection de l'ensemble focus dans M-QuBE^3 s'effectue en amont du calcul de l'eScore (respectivement phase A et B dans la Fig. 4.3). Cette sélection s'effectue par recherche par mots-clés ou via une visualisation interactive. L'eScore calculant une estimation d'intérêt en fonction d'une sélection utilisateur changeante, il est alors possible d'utiliser l'eScore dans n'importe quel processus évolutif où le calcul de l'eScore peut être ré-itéré au fur et à mesure des nouveaux éléments pré-sélectionnés par l'utilisateur.

Dans la suite, nous considérons que l'eScore est utilisé conjointement avec M-QuBE^3. Dans un premier temps, nous définissions formellement l'eScore (Section 5.2.1) afin, dans un second temps, d'expliquer et définir ce que sont les fonctions de pilotage qu'il utilise (Section 5.2.2).

5.2.1 Formalisation

En s'inspirant du modèle de Kivelä *et al.* [41], notre réseau multi-couches est défini par $G(V, L, E)$ où V est l'ensemble des sommets, L l'ensemble des couches tel que $\forall l \in L, l : (0,1)^{|V|}$ et E l'ensemble des arêtes tel que $E : (V, L) \times (V, L)$. Pour chaque $v \in V$, $b_l(v) : (0,1)^{|L|}$ renvoie un vecteur binaire indiquant les couches auxquelles appartient le sommet v.

La volonté utilisateur pour chaque sommet v de V est définie par un ensemble F de fonctions. Chaque f de F s'applique à un sous-ensemble de couches $L' \subseteq L$ (aspects) et peut être définie par un vecteur binaire b_{lf} indiquant les couches sur lesquelles s'applique la fonction f tel que : $b_{lf}(f) : (0,1)^{|L|}$. Chaque fonction de F renvoie un score normalisé entre 0 et 1.

L'eScore pour un sommet x compte tenu de l'ensemble focus Y peut être défini ainsi :

$$eScore(x|Y) = \frac{\sum_{i=1}^{|F|} f_i(x, Y, L'_i, b_l(x))}{|F|}$$

où $b_l(x) \subseteq L'_i \subseteq L$.

Le rôle de chaque fonction est de guider la navigation en prenant en compte les différences sémantiques et les différences d'intérêt utilisateur entre les couches. Ces fonctions s'appellent des "fonctions de pilotage". Par exemple, soient deux couches données composées d'un ensemble de sommets représentant des fichiers vidéos dans une des couches et de fichiers audio dans l'autre. Une fonction commune au deux couches peut ne pas avoir de sens. Les différents attributs des couches peuvent contraindre à une différenciation lors de l'estimation de l'intérêt et ainsi contraindre à déterminer des méthodes différentes. Il est aussi possible que l'utilisateur ne porte pas une attention égale aux différentes couches et souhaite focaliser son attention davantage sur les vidéos que sur les pistes audio. L'eScore propose donc d'établir une attribution des fonctions de pilotage en fonction des différentes associations de couches possibles.

Dans notre exemple, il est possible d'avoir une fonction attribuée à chacune des couches audio et vidéo et une fonction commune aux deux couches. Ainsi un sommet représentant un fichier vidéo sera considéré par une fonction de pilotage spécialisée qui prendra en compte son type et ses attributs spécifiques et sera aussi considéré par la fonction de pilotage globale pour permettre une comparaison et une analyse plus générale.

Autre exemple plus formel pour un réseau à 4 couches : une fonction attribuée

à l'association de couches $b_{lf}(f) = (0, 1, 1, 0)$ va influer sur les scores de chaque sommet x dont, pour $\forall n \in \{0, 1\}$, $b_l(x) = (n, 1, n, n)$ ou $(n, n, 1, n)$ (Fig.5.1).

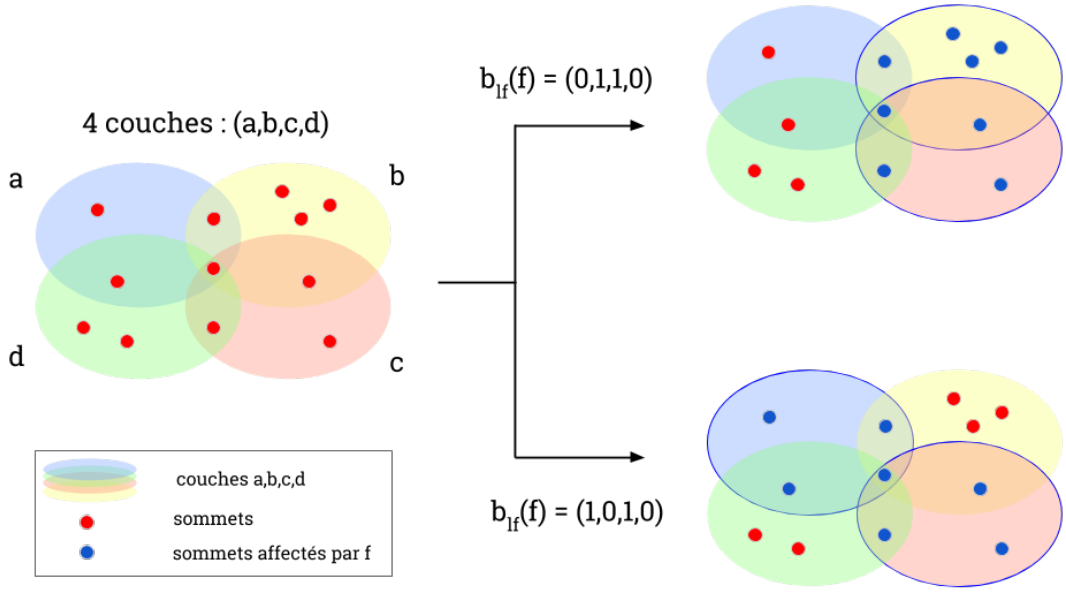


FIGURE *5.1* – ***Domaine d'application des fonctions de pilotage*** *: Le vecteur binaire lié à une fonction de pilotage représente son domaine d'application i.e les différentes couches sur lesquelles elle s'applique. Un sommet est donc utilisé par chaque fonction comprenant dans son domaine d'application la couche à laquelle il appartient.*

Actuellement, les fonctions de pilotage doivent être sélectionnées et paramétrées avec les experts des données pour s'assurer qu'elles soient en accord le plus étroitement possible avec leur volonté afin de s'assurer d'une pertinence optimale. Dans la partie suivante, nous allons proposer une catégorisation de ces fonctions et quelques exemples ayant été utilisés.

5.2.2 Fonctions de pilotage

Déterminer une mesure d'intérêt à partir d'un sommet peut découler de la sémantique des données (recherche de mots dans des champs texte, sélection manuelle de sommets par l'utilisateur, analyse des attributs des sommets), de la topologie du réseau (i.e centralité d'intermédiarité, degrés, faire partie d'une clique ou

d'une triade, etc.) ou, parfois, des deux simultanément. Les fonctions de pilotage suivent ce schéma en pouvant s'orienter à la fois vers la sémantique et la topologie.

En plus de cela, nous distinguons ces fonctions selon deux critères : l'"interaction" (dépendant de l'ensemble focus ou constant) et le "domaine d'application" (couche simple ou association de couches). Les fonctions de pilotage sont définies à travers ces deux critères en fonction de leurs rôles et objectifs afin d'obtenir une formalisation mathématique de la volonté utilisateur (Fig. 5.2).

Interaction	Application	Exemples
Sélection Utilisateur	Couche simple	Nombre défini de politiciens souhaité dans la sélection utilisateur (un sommet appartenant à la couche politicien est favorisé si la quantité choisie n'est pas atteinte)
	Association de couches	Homogénéité du nombre de sommets sélectionnés par l'utilisateur à travers les différentes couches reliés à des types de documents (article, vidéo, interview, etc.)
Statique	Couche simple	Connectivité d'une personne dans la couche personne (les sommets des autres couches ne sont pas considérés)
	Association de couches	Calcul de centralités sur l'intégralité du réseau

FIGURE 5.2 – ***Classification des fonctions de pilotage*** *: Chaque fonction de pilotage peut être classée selon deux critères. Elle peut être dépendante de la sélection utilisateur ou être constante/statique. Une fonction basée sur la sélection utilisateur rendra le calcul d'intérêt dépendant de l'action de l'utilisateur et peut rendre ainsi le processus utilisant le score interactif. En plus de cela, dans chacune de ses catégories, elle peut être liée à un couche simple ou à un ensemble/association de couche. Une association de couche peut comprendre l'ensemble de toutes les couches du réseau si la fonction doit s'appliquer sur tous les sommets du réseau.*

Interaction

L'interaction d'une fonction de navigation se définit par l'impact donné aux actions de l'utilisateur sur le calcul de score entre les différentes exécutions de `M-QuBE`^{3}/eScore. Une fonction peut appartenir à deux catégories : soit basée sur la sélection utilisateur (tous les sommets de l'ensemble focus) soit statique (indépendant de tout choix utilisateur).

Fonctions basées sur la sélection utilisateur Ces fonctions sont basées sur la sélection utilisateur. Parce qu'elles utilisent l'ensemble focus, le résultat peut varier d'une itération à l'autre.

Leur objectif est d'une part de renforcer l'interactivité en mettant l'utilisateur aux commandes et d'autre part de permettre à la méthode de s'adapter si des contraintes ou des objectifs doivent être respectés lors de la recherche.

Un exemple est l'homogénéité de type dans l'ensemble focus. Les utilisateurs veulent avoir un nombre équivalent de sommets des différents types possibles dans leur sélection. Il est donc nécessaire de maximiser les scores des sommets qui améliorent l'homogénéité de cette sélection s'ils doivent être sélectionnés (et inversement minimiser le score des sommets déséquilibrant davantage l'homogénéité de la sélection s'ils sont sélectionnés). Cette procédure est similaire à un calcul d'optimisation d'entropie. En uniformisant le nombre de chaque type dans la sélection, les scores des sommets deviennent alors égaux car leurs sélections seraient d'un impact équivalent sur l'homogénéité. L'utilisateur ayant ainsi un nombre maximum de choix pour sa sélection, l'entropie est alors maximisée. Un cas pratique utilisant cet exemple est présenté dans le chapitre 6.

Fonctions constantes Une fonction constante est une fonction avec aucun prérequis de la part de l'utilisateur pour calculer son score. Il est néanmoins possible de l'appliquer sur une ou plusieurs couches (voir paragraphe suivant). Ces fonctions peuvent être topologiques ou sémantiques. Par exemple, nous calculons dans notre réseau un classement basé sur les degrés de tous les sommets (et donc de la topologie du réseau) ou différents types de centralité. Pour une fonction orientée vers la sémantique, on peut utiliser par exemple un score de proximité entre un mot clé et des attributs des sommets, comme utilisé par Van Ham et Perer [76].

Parce que ces fonctions sont indépendantes du contexte, elles peuvent être calculées antérieurement à toute action utilisateur et leurs résultats peuvent être conservés entre les différentes itérations en cas de processus avec de multiples calculs de l'eScore.

Domaine d'application

Le domaine d'application d'une fonction de pilotage correspond aux couches sur laquelle elle va avoir une influence. Les fonction de pilotage peuvent ainsi soit s'appliquer sur une couche soit sur une association de couches. Il est à noter que, comme dit précédemment, des chevauchements sont possibles entre les différentes

fonctions. Ainsi, une couche donnée peut être comprise et concernée par plusieurs fonctions mono-couches et/ou plusieurs fonctions d'association de couches.

Couche simple Il est parfois nécessaire de pouvoir définir un objectif spécifique pour une catégorie de sommets du réseau. Dans notre exemple, les historiens voulaient trouver des personnes importantes liées à autant d'autres personnalités que possible dans le réseau. Nous avons donc ajouté un calcul de degré interne à la couche personne (en ne considérant que les liens entre deux sommets appartenant à la couche personne). De telles fonctions peuvent aussi être utilisées pour simplement pondérer une couche en particulier du réseau. Si les experts ne sont pas intéressés par un pan des données, il est alors possible d'attribuer uniquement par cet intermédiaire un score bas pour la couche qui est jugée non pertinente.

Association de couches Les fonctions basées sur l'association de couches permettent de mettre en valeur l'interaction entre les différentes couches du réseau ou de faire l'union de certaines couches autour d'un objectif commun. Un exemple orienté sur la topologie est de calculer la centralité dans un sous-réseau composé de sommets inclus uniquement dans une association de couches donnée. Pour la sémantique, l'exemple précédent sur l'homogénéisation de types correspond à nouveau : les types de documents (interview audio, interview vidéo, journal télévisé, article, etc.) sont en réalité une association de couches basée sur les couches interview audio, interview vidéo, etc. sur laquelle s'applique une même fonction d'homogénéisation afin d'obtenir au mieux la même quantité de documents par type dans la sélection. Par ailleurs, il est aussi possible d'instancier une fonction qui s'applique à une association correspondant à toutes les couches du réseau. Ce faisant, les fonctions topologiques classiques (centralité de proximité, centralité d'intermédiarité, degrés,etc.) peuvent être utilisées avec chaque sommet du réseau pour calculer un score.

5.3 Synthèse

Comme vu dans le chapitre précédent, `M-QuBE`^{3} est définie comme une méthode à l'**échelle des individus**, **itérative et arborescente**, en accord avec la méthodologie des experts des sciences humaines et sociales. Parmi nos objectifs initiaux, il reste alors à répondre au caractère multi-couche des réseaux utilisés.

La méthode `M-QuBE`^{3} a été construite autour d'un score d'estimation d'intérêt expressément pour apporter une réponse à cet objectif. En plus de répondre au

caractère multi-couche, il est aussi possible de faire écho aux précédents objectifs en proposant au score de prendre en compte la sélection utilisateur changeante au fil des itérations et de s'adapter en conséquence.

C'est pourquoi nous proposons eScore, une métrique d'estimation de l'intérêt pouvant être utilisée à travers des processus itératifs et spécialisée pour les réseaux multi-couches.

Permettant à la fois d'adapter son traitement à des associations de couches tout en ayant la possibilité de considérer la sélection de l'utilisateur, l'eScore permet ainsi une gestion tant des réseaux multi-couches que des cas plus génériques. Selon la catégorisation que nous avons défini, il est par exemple possible de reproduire l'API de Furnas [27] (en sélectionnant une fonction de pilotage statique appliquée sur l'association de l'ensemble des couches) et l'UI de Van Ham et Perer [76] (en sélectionnant une fonction de pilotage basée sur la sélection de l'utilisateur, ne comportant qu'un seul focus et appliquée à l'association de l'ensemble des couches).

Si eScore peut être utilisé sans `M-QuBE`^{3} et inversement, ces deux travaux ont été conçus et développés afin de fonctionner en synergie pour répondre aux problèmes de nos experts. C'est pourquoi, dans le chapitre suivant, nous détaillons l'implémentation d'une plateforme donnant corps à `M-QuBE`^{3} et eScore afin de valider nos méthodes à travers les cas réels de nos experts des données, issus des sciences humaines et sociales.

Chapitre 6

Implémentation et Validation

Sommaire

Dans les chapitres précédents, nous avons défini M-QuBE^3 (Chapitre 4), une méthode d'exploration et de navigation pour les réseaux multi-couches utilisant une estimation de l'intérêt, l'eScore (Chapitre 5). Ces travaux ont été menés en étroite collaboration avec nos collègues historiens afin de correspondre au mieux aux besoins et aux méthodes de travail de leur domaine. Il est donc important d'éprouver M-QuBE^3 et son implémentation afin de s'assurer que la méthode remplisse ses objectifs. Cette validation utilise les données sur lesquelles travaillent quotidiennement ces historiens (voir Chapitre 2). Ils ont ainsi à la fois une connaissance aiguë du contexte sémantique des données et de leur structure et sont donc à même de pouvoir juger au mieux la viabilité et l'intérêt des résultats obtenus.

Une plateforme a été développée afin de pouvoir donner corps aux différentes vues et méthodes des partenaires (sous-section 6.1) et ainsi les présenter aux experts. Si la version initiale était conçue pour être générique et facilement personnalisable (sous-section 6.1.1), la conception et le développement de M-QuBE^3 ont

néanmoins contraint à faire évoluer cette architecture (sous-section 6.1.2) afin de pouvoir proposer, après une phase de paramétrage (sous-section 6.1.3), une exploration adaptée aux besoins et à la méthodologie des experts. En plus d'estimer la pertinence des vues et des concepts proposées par M-QuBE^3, un objectif important est de savoir si les experts trouvent la méthode adaptée et efficace pour explorer un nouveau jeu de données. Des sessions de tests ont donc été effectués afin de valider M-QuBE^3 à travers différents scénarios liés à ces données européennes (section 6.2). Après avoir déterminé un protocole (sous-section 6.2.1), les experts ont été interrogés afin d'analyser les différents aspects spécifiques de M-QuBE^3(sous-section 6.2.2).

6.1 Implémentation

Le processus M-QuBE^3 a été développé dans le contexte du projet ANR BLIZAAR (voir chapitre 2). Nous avons convenu au début du projet d'une architecture commune avec nos partenaires à partir de laquelle développer nos outils (sous-section 6.1.1). Cependant des évolutions du modèle initial ont été nécessaires afin de parvenir à nos fins lors de la conception de M-QuBE^3(sous-section 6.1.2). Si l'architecture initiale est induite par les spécificités du projet et ses évolutions par celles de M-QuBE^3, nous présentons ensuite la configuration de M-QuBE^3 qui est directement liée aux volontés des experts (sous-section 6.1.3).

6.1.1 Architecture initiale

Conformément aux objectifs du projet, l'architecture technologique repose sur une plateforme web sur laquelle chaque partenaires peut ajouter ses propres outils de calculs et de visualisations (Fig. 6.1).

La plateforme s'articule autour d'un serveur central sur lequel se greffent des serveurs périphériques dédiés à la gestion et aux calculs sur le réseau issu des données européennes. Ce serveur central est alors une interface pour accéder aux données sans avoir à se soucier de leur gestion en amont. Il utilise notamment deux bases de données NoSQL ("Not only SQL") pour stocker d'une part les informations relatives à la gestion des utilisateurs et d'autre part aux données du réseau. Ce serveur est également le point d'accès de l'utilisateur qui peut ainsi bénéficier des différents services de visualisation et d'exploration offerts par la plateforme directement sur son navigateur.

Un serveur utilisant le langage javascript, NodeJS (https://nodejs.org/en/), a été sélectionné pour l'utilisation facilitée des bibliothèques utilisées pour la ges-

tion et l'affichage des données (voir ci-après). Des serveurs additionnels peuvent ensuite être greffés au serveur NodeJS afin de répondre aux besoins des technologies imposées par les visualisations et les autres fonctionnalités que l'on souhaite ajouter. Ces serveurs communiquent par le biais de fichiers JSON (un format de données inspiré des objets du langage Javascript : https://fr.wikipedia.org/wiki/JavaScript_Object_Notation) répertoriant les sommets, liens et les divers informations utiles à afficher coté client.

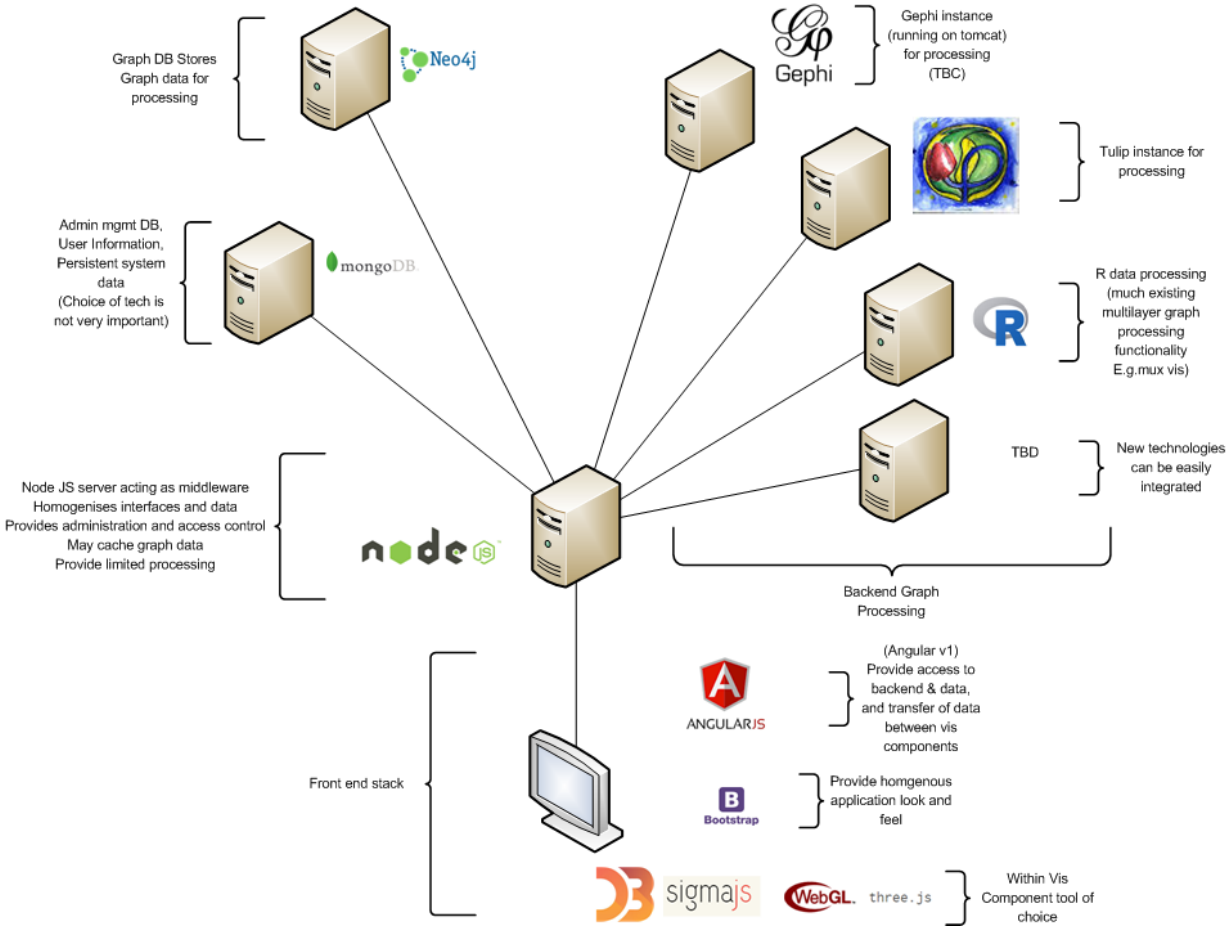


FIGURE *6.1* – ***Architecture initiale du projet BLIZAAR*** *: un serveur central nodeJS est connecté à un ensemble de bases de données stockant les données relatives aux réseaux et aux utilisateurs ainsi qu'à un ensemble de serveurs dédiés aux traitements sur les réseaux ou les données. Image provenant du site du projet : https://blizaar.list.lu/doku.php?id=project_architecture_tools_and_design_standards_document*

Du coté des bases de données, les données définissant des réseaux imposent des contraintes dans la modélisation afin de pouvoir être exploitées de manière optimale. Il est en effet difficile de stocker et établir des requêtes sur un réseau complexe (hétérogène et très connecté) et volumineux avec une base de données relationnelle comme l'ont attesté les travaux précédemment effectués lors du projet TETRUM [45] (chapitre 2). Une base de données graphe (i.e. une base de donnée où les données sont stockées nativement sous forme de sommets et de liens entre les sommets définissant ainsi un graphe) a donc été sélectionnée pour permettre une utilisation efficace du réseau. Le modèle choisi contenant les données historiques est une base Neo4J (https://neo4j.com/) avec laquelle nous avons déjà travaillé et qui est familière aux différents membres du projet BLIZAAR. Celle-ci est interrogée par le serveur central afin de créer un réseau intermédiaire, le graphe tampon ("master graph" dans la documentation anglophone du projet), correspondant à une image partielle des données créée à partir d'une requête de l'utilisateur. Ce sous-réseau, plus léger et plus simple, permet ensuite d'être exploité directement par une requête vers le serveur NodeJS afin d'être utilisé par les différentes vues sans avoir à solliciter la base Neo4j.

En plus de la base de données Neo4j, il y a aussi une base orientée document de type mongodb (https://www.mongodb.com/fr) dédiée principalement aux informations liées aux utilisateurs (sessions, mots de passe, identifiants) ou à des sauvegardes de listes de sommets pour des vues spécifiques de nos partenaires.

Les serveurs de calcul sollicités par le serveur central utilisent principalement R (https://www.r-project.org), Gephi [6] (https://gephi.org) et Tulip [3] (http://tulip.labri.fr). Leurs rôles sont utilisés en fonction des besoins des visualisations à développer et des préférences individuelles des développeurs.

Du coté client, l'aspect visuel général est assuré par Bootstrap (une bibliothèque en langage CSS et Javascript permettant de structurer les pages sur une grille afin d'avoir une résultat ordonné et régulier, https://getbootstrap.com/) et Angular(https://angular.io/), une bibliothèque permettant de modifier facilement la structures des pages ainsi que facilitant l'utilisation des données transitant depuis le serveur central. Pour l'affichage des visualisations coté client, les vues noeuds-liens des réseaux sont assurées par sigmaJS, une API de visualisation spécialisée dans les graphes, ou directement en webGL, une variante web d'OpenGL (https://www.opengl.org/, une bibliothèque de référence pour faire de la synthèse d'image) permettant notamment d'utiliser le processeur graphique de l'ordinateur via un navigateur pour des performances accrues. Les visualisations d'autres types sont essentiellement assurées par D3.js (https://d3js.org/), une

bibliothèque graphique proposant une grande liberté en terme de représentations graphiques mais limitée à des jeux de données plus réduits.

Cependant, le développement de M-QuBE^3 a demandé de dévier de cette architecture pour répondre à des besoins techniques.

6.1.2 Évolution de l'architecture

L'architecture initiale est conçue avec pour préoccupation première la généricité et l'adaptabilité. Néanmoins, des limites sont apparues lors de la conception de M-QuBE^3. Les spécificités de M-QuBE^3 ont révélé deux éléments imposant de dévier légèrement de la direction initiale (Fig.6.2).

Comme indiqué précédemment, les données utilisées par les vues transitent dans la première architecture à travers le graphe tampon (l'image partielle du réseau) contenue dans le serveur central. Ceci pose un problème à M-QuBE^3 qui nécessite d'accéder au réseau entier lors de son calcul d'intérêt. En effet, celui-ci se comporte de manière analogue à un algorithme glouton : à partir de la sélection utilisateur, les différents scores sont calculés pour le voisinage de chaque sommet de cette sélection. Après avoir établit un classement à partir de ces scores, le sommet ayant le meilleur score d'intérêt aura son voisinage ajouté à la population à évaluer. La procédure est réitérée et la zone analysée dans le réseau s'étend au fur et à mesure que des voisinages de sommets en tête de classement sont ajoutés. Avoir une image partielle du réseau est alors problématique car la zone à analyser peut facilement sortir hors de ce sous-réseau et ainsi imposer une mise à jour du graphe tampon à chaque nouveau classement pour récupérer les sommets manquants. Au final, il est alors plus efficace de pouvoir directement utiliser un réseau complet pour éviter d'alourdir significativement le temps de calcul de M-QuBE^3. En pratique, la base Neo4j est donc directement liée à un serveur Flask où un graphe Tulip est généré avant toute action de l'utilisateur à partir des données de Neo4j. Le graphe tampon n'est donc plus utilisé. Toutes les opérations requises par M-QuBE^3 sur le réseau sont ensuite opérées sur le graphe Tulip stocké et utilisé sur le serveur Flask.

Le deuxième élément est l'utilisation du graphe de trace (graphe des actions de l'utilisateur, voir chapitre 4). Ce graphe a pour particularité d'être un graphe de graphes i.e. un graphe où chaque sommet est un sous-réseau qui a été généré par M-QuBE^3(voir chapitre 4). Il est alors nécessaire de pouvoir gérer une deuxième structure de graphe parallèlement à celle issue des données et aussi de pouvoir accéder facilement aux différents sous-réseaux générés par M-QuBE^3. Le serveur nodeJS se veut initialement générique et utilisable avec différents serveurs de calculs

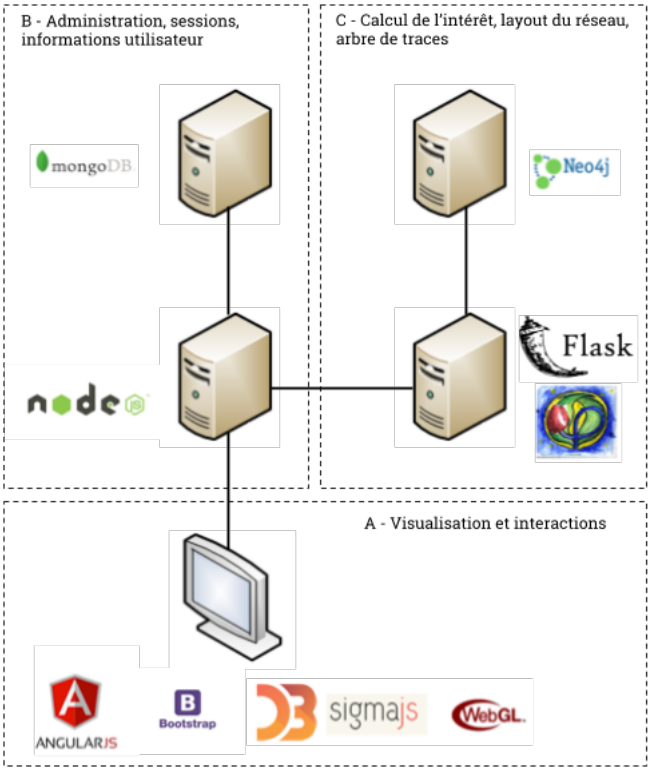


FIGURE *6.2* – ***Architecture utilisée par*** `M-QuBE`^{3} *: Le processus* `M-QuBE`^{3} *nécessite d'avoir un accès complet aux données afin de pouvoir traiter l'entièreté du réseau. Le serveur Flask utilise donc la bibliothèque Tulip sur une image du réseau complète extraite directement de Neo4j à partir de Tulip (C). Le serveur NodeJS est maintenant essentiellement dédié à la gestion des utilisateurs et aux procédure d'affichage des données reçues par Flask (B). Ces données sont ensuite transmises au coté client afin d'être visualisées (A). Lorsque l'utilisateur interagit avec la vue, les informations relatives à ses actions sont transmises à Flask par l'intermédiaire de Nodejs qui peut alors recommencer une procédure de calcul.*

(coté serveur) et différentes vues (coté client). Modifier sa structure spécifiquement pour `M-QuBE`^{3} en tordant son implémentation va à l'encontre de son objectif premier. Cela nécessite encore une fois d'avoir une autre structure graphe en dehors du serveur central et à proximité du réseau des données européennes sur lequel s'établissent les calculs de `M-QuBE`^{3}. Chaque vue partielle du graphe de trace doit en effet pouvoir être sauvegardée et lue avec un minimum d'aller-retours entre les différents serveurs, ce mécanisme est donc aussi assuré par Tulip sur le serveur

Flask en générant un deuxième graphe Tulip spécialement dédié au graphe de trace. Chaque sommet du graphe de trace référence alors les différents éléments (sommets et liens) composant un sous-réseau précédemment généré par M-QuBE^3 et permet de facilement ré-accéder à ces données pour pouvoir les afficher à la volée. Par ailleurs, Tulip a des facilités natives pour gérer ce type de structures complexes [66], permettant ainsi d'éviter de ré-implémenter un structure de donnés capable de gérer les graphes de graphes ou de dupliquer des données.

Après avoir décrit l'implémentation et le déploiement derrière M-QuBE^3, voyons maintenant son paramétrage.

6.1.3 Paramétrage

Le processus M-QuBE^3 est pensé comme un outil de navigation en deux temps (voir section 4) : en premier lieu, on initialise l'outil avec des fonctions de pilotage centrées principalement sur la topologie du réseau. Par cette première configuration, on mène une première exploration générale pour prendre connaissance de la sémantique et de la structure des données. A partir de cette compréhension accrue du réseau, il est ensuite possible de configurer plus finement l'outil afin de procéder à une seconde exploration plus précise, directement liée aux données, afin de répondre à un objectif défini ou explorer plus amplement autour d'un centre d'intérêt donné.

Cette évaluation est centrée sur l'étape de première exploration. Même si cette première étape se veut générale et non liée à un contexte sémantique précis du jeu de données, il est néanmoins nécessaire de la configurer en sélectionnant des fonctions de pilotage cohérentes pour l'ensemble des scénarios envisagés et en accord avec la volonté des utilisateurs. Le processus M-QuBE^3 a ainsi été initialisé et paramétré de la même manière pour l'ensemble des scénarios afin de pouvoir comparer les résultats obtenus.

Comme dit précédemment, le processus M-QuBE^3 essaye de coller au plus près au "Nested Model" de Munzner [61] dans son concept. Aussi, si l'on ne connaît pas précisément les sommets du réseau à l'avance, les différents types de couches (et donc les différents types de sommets) sont connus et il est alors possible d'attribuer les fonctions de pilotage en fonction des intérêts et contraintes des experts sur ces différentes couches.

Chaque couche est donc soumise à un même ensemble de fonctions de pilotage (Fig. 6.3) basé sur la topologie (fonctions A et B sur la figure) ou la contrainte utilisateur (figure C).

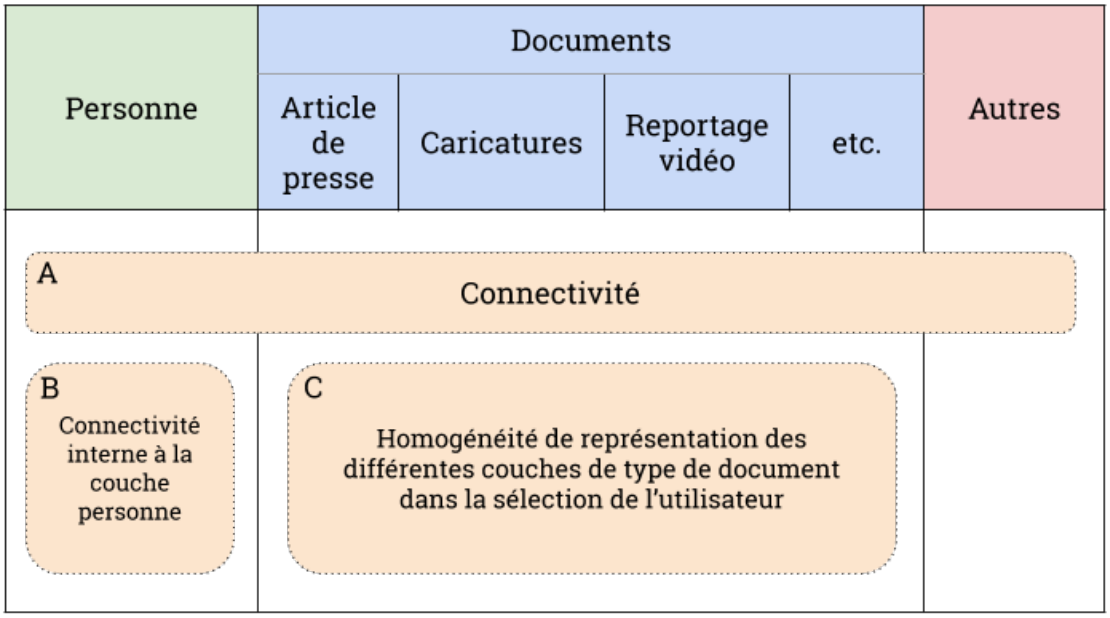


FIGURE *6.3* – ***Choix des fonctions de pilotage*** *: Les historiens ont un intérêt accru pour les éléments très connectés du réseau, une fonction de pilotage générale sur l'ensemble des couches est donc ajoutée utilisant un score basé sur le degré (A). De la même manière, les historiens veulent se concentrer sur les acteurs de la construction européenne en relation avec d'autres acteurs. Aussi, une seconde fonction de connectivité basée sur les degrés est ajoutée mais celle-ci ne considère que les sommets présents dans la couche personne lors de son calcul de score (B). Enfin, les historiens veulent pouvoir sélectionner un ensemble de documents pour créer une bibliographie mais avec pour contrainte que les différents types de document soient équitablement représentés. Une fonction est donc attribuée aux documents afin d'homogénéiser les types de document dans la sélection utilisateur en proposant les types manquants à la sélection (C).*

A : Les historiens ont un intérêt accru pour des éléments avec un voisinage important, une fonction basée sur les degrés des sommets à donc été privilégiée. D'autres types de centralité peuvent évidemment être utilisés cependant, à des fins d'évaluation, la mesure de degré est facilement compréhensible par un non spécialiste et a donc été privilégiée. Cette fonction de pilotage retourne donc pour chaque sommet un score basé sur son rang dans un classement basé sur le degrés des sommets. Un score de 1 va correspondre au premier

sommet du classement (avec le degré le plus haut) et un score de 0 va correspondre au dernier (avec le degré le plus faible). Cette fonction permet ainsi de valoriser les sommets les plus connectés lors du choix des sommets à montrer dans le prochain sous-réseau. Selon la classification des fonctions de pilotage (voir section 5.2.2), elle est une fonction statique appliquée sur une association de couches comprenant l'ensemble des couches du réseau afin de pouvoir être appliqué à tous les sommets du réseaux. Étant statique, elle n'est pas dépendante des actions de sélection de l'utilisateur et peut donc être pré-calculée à des fins d'optimisation.

B : Si la première fonction de pilotage s'applique sur l'ensemble de couches et donc l'ensemble des sommets du réseau, cette fonction se concentre uniquement sur la couche des personnes. Cette fonction est donc une fonction statique mono-couche. Les historiens veulent voir ressortir des acteurs de l'histoire hautement connectés pour pouvoir s'appuyer sur eux ou leur voisinage lors de leurs recherches suivantes. Le score est donc aussi basé sur le degré mais cette fois-ci seul le voisinage étant compris dans la couche personne est considéré. Le classement est donc restreint aux sommets de la couche personne et propose de mettre en avant les personnes étant le plus connectées à d'autres personnes ainsi faire ressortir les acteurs importants pour les experts.

C : Les historiens ont pour nécessité lors de la rédaction d'une publication d'avoir une couverture homogène des types de document utilisés pour sa documentation (article de presse, interview vidéo, caricature, etc.). Ainsi, lorsque l'utilisateur sélectionne un document, il est nécessaire par la suite de valoriser les autres types de documents parmi les sommets proposés dans le prochain sous-réseau. Cette dernière fonction propose donc cette fois une estimation de l'intérêt en fonction de la sélection de l'utilisateur appliquée à une association de couches correspondant à l'ensemble des différentes couches décrivant un type de document (articles de presse, caricatures, reportages vidéos, interview télévisés, etc.). Ce problème peut être assimilé à un calcul d'optimisation d'entropie où, à chaque itération, on essaye de maximiser le nombre de choix possibles pour sélectionner un document. En cas de sélection déséquilibrée (i.e. un type de document prédomine dans la sélection de l'utilisateur), le choix se réduit autour des types manquants et l'entropie diminue conséquemment. Inversement, en homogénéisant la sélection, tout type de document devient un choix potentiel et l'entropie est alors maximale. La fonction utilisée, basée sur l'entropie de Shannon [69], aide ainsi les experts à avoir la diversité de bibliographie exigée par leur domaine.

Ces fonctions de pilotage, bien que liées aux données, demeurent relativement génériques dans le sens où elles ne sont pas spécifiques à un scénario défini. En outre, la contrainte liée à l'homogénéité des types de documents et l'importance des personnes très connectées dans un réseau sont des problématiques constatées avec l'intégralité de nos partenaires en sciences humaines et sociales. Ces fonctions détermineront la manière dont le score d'intérêt des sommets est estimé pour les quatre scénarios.

Dans la suite, nous allons détailler la méthodologie et le déroulement de l'évaluation puis, dans un second temps, les résultats obtenus.

6.2 Validation

Lors de ces sessions de validation, un même protocole a été appliqué pour l'ensemble des partenaires afin de minimiser l'apparition de biais et pouvoir compiler et comparer les résultats obtenus. Dans un premier temps, nous allons donc expliciter la méthode employée pour effectuer les sessions de validation puis nous détaillerons les résultats obtenus selon les différentes caractéristiques de M-QuBE^3 analysés.

6.2.1 Méthodologie

Le processus de validation a été effectué en quatre sessions d'une heure, chacune avec un expert d'un domaine spécifique lié à la construction de l'Europe. Ces domaines sont très différents (développement du Benelux, émergence des banques européennes, impact des acteurs européens majeurs comme Jacques Delors ou Robert Schuman) et permettent de couvrir un large spectre de thématiques (géopolitique, économique, social ou des combinaisons de ces thématiques). Ces experts ont chacun un ensemble de scénarios liés à leurs domaines spécifiques et sont familiers des données utilisées par le processus M-QuBE^3.

Une session de validation a une durée d'une heure. Elle se compose d'une présentation générale de M-QuBE^3 à l'expert en expliquant l'objectif de la méthode, son fonctionnement global (sans rentrer dans les détails techniques sauf requête de l'utilisateur) et sa manipulation. Cela est suivi d'un exemple théorique sur les données (en l'occurrence sur la guerre froide avec pour point de départ le sommet représentant Joseph Staline) afin d'illustrer concrètement le déroulement d'une utilisation de M-QuBE^3.

Vient ensuite une phase où, à deux devant un ordinateur portable, l'expert est libre d'explorer le jeu de données en suivant les scénarios qu'il avait préalablement préparés (histoire d'une institution, d'une personnalité européenne ou d'une zone géographique spécifique par exemple). Pour éviter d'être freiné par des soucis de manipulation, l'expert peut décider s'il souhaite ou non manipuler l'outil lui-même. Le cas échéant, j'ai manipulé l'outil selon ses indications.

Cette phase prend majoritairement la forme d'une discussion. Les différents points ou choix effectués dans le design de `M-QuBE`^{3} sont présentés au fur et à mesure des explorations à l'expert afin de pouvoir bénéficier de son opinion, de ses analyses et d'éventuelles suggestions. Des notes sont prises en parallèle pour conserver des informations ou suggestions jugées significatives. Ces différents points sont le concept général d'évoluer dans un jeu de données par vues partielles et la méta-navigation offerte par le graphe de trace (navigation dans les vues utilisées pour naviguer dans les données), les vues proposées et les interactions qui y sont liées et, enfin, les sommets et les scores utilisés pour les obtenir.

A l'issue du processus de validation, un questionnaire est remis à l'expert (voir Annexe B). Celui-ci comprend plusieurs groupes de questions dont l'objectif est d'évaluer les différentes parties de chaque point discuté. Ces questions demandent soit de donner une note de 1 (qualité basse) à 7 (qualité haute) sur les différents points, soit proposent des réponses plus ouvertes afin que l'expert puisse donner plus amplement son avis et soumettre des idées de fonctionnalités additionnelles. Chaque question ouverte a été présentée comme facultative, les experts ont donc été libres de répondre ou non aux différentes questions s'ils n'avaient pas de suggestions ou remarques particulières à l'issue du processus de validation.

Dans la partie suivante, nous traitons des différents retours que nous avons obtenus pour chaque aspect du processus `M-QuBE`^{3}.

6.2.2 Résultats

Les résultats suivant sont issus des retours obtenus par les différentes discussions avec les experts ainsi que par les questionnaires qu'ils ont complétés. Toutes les remarques et suggestions, orales comme écrites, ont été prises en compte afin de présenter une opinion globale de la méthode. Ces résultats s'articulent en plusieurs aspects analysés séparément et présentés ci-après.

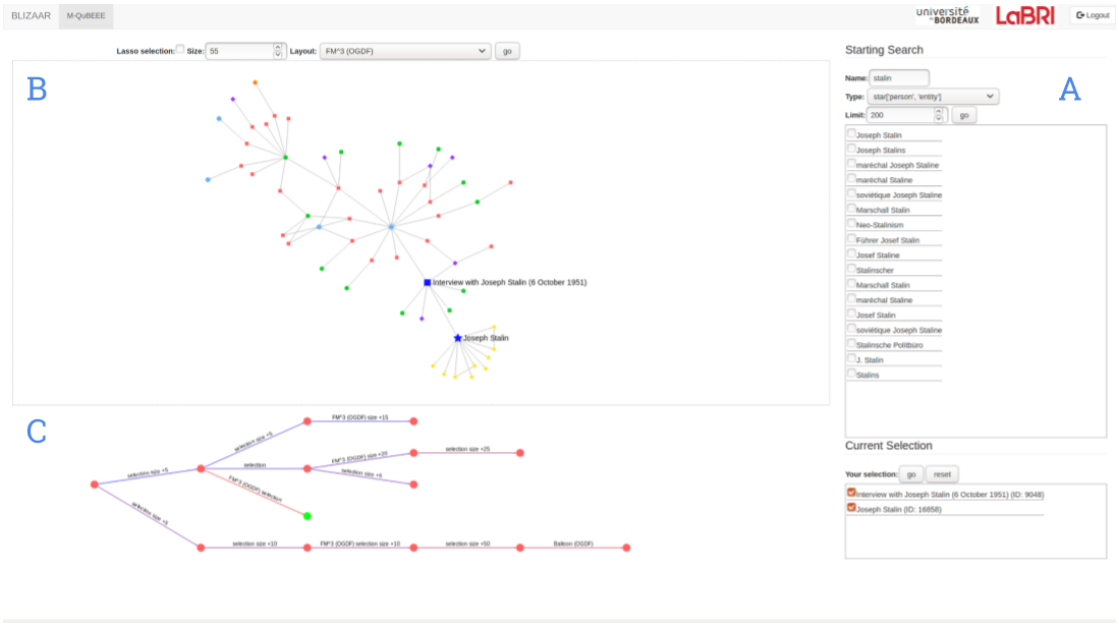


FIGURE *6.4* – ***Capture d'écran de*** `M-QuBE`^{3} *: Voici la version utilisée par les experts lors de la session de validation. Le panneau latéral (A) permet de sélectionner un ou plusieurs points de départ par une recherche via mot-clé dont la liste des sommets proposés peut être filtrée par type de sommet. Le canvas (B) affiche le sous-réseau actuellement sélectionné. Lorsque l'utilisateur sélectionne un ou plusieurs sommets depuis cette vue, un nouveau sous-réseau est créé et remplace l'actuel. En plus de cela, lorsqu'un sous-réseau est généré, un sommet le représentant est ajouté à l'arbre de trace (C). Celui-ci est ensuite lié au sous-réseau à partir duquel il a été généré. En cliquant sur un sommet de l'arbre de trace, l'utilisateur affiche dans le canvas le sous-réseau correspondant et peut alors générer de nouveaux sous-réseaux à partir de celui-ci.*

Concept et arbre de traces

Le concept général de navigation dans les données à travers des vues partielles a été jugé facilement compréhensible par trois experts sur quatre (présenté comme "naviguer dans la mer et regarder par le hublot du bathyscaphe") et a été favorablement reçu. Le dernier expert a néanmoins manifesté une gène pour l'aspect "boîte noire" à savoir ne pas comprendre pourquoi certains éléments sont favorisés aux détriments des autres et surtout ne pas avoir connaissance des éléments environnants non affichés.

Proposer un paramétrage personnalisé des fonctions de pilotage directement dans un panneau de l'interface pourrait permettre, en mettant l'utilisateur encore davantage aux commandes, de régler l'aspect de la compréhension et de la confiance envers les scores d'intérêt. Une possibilité de pré-configuration ou d'aide à la configuration semble néanmoins nécessaire pour les experts néophytes en analyse de réseaux. Pour les éléments environnants, une fonctionnalité permettant de rajouter un nombre défini de voisins d'un sommet en cliquant dessus permet d'y répondre partiellement. Néanmoins, le concept même du processus est de réduire le réseau à des sous-ensembles de sommets analysables. Initialement, `M-QuBE`[3] est conçu pour s'appliquer à de grands volumes de données hautement connectées et hétérogènes. Dans ce contexte, il est alors souvent difficile ou impossible d'afficher des portions complètes du réseau sans nuire à la lisibilité [31]. Fournir des informations générales sur le réseau (nombre de sommets du voisinage, distribution des types de sommets du voisinage, ect.) peut permettre de mitiger l'impression de "boîte noire" mais il semble délicat de faire disparaître entièrement cette impression sans nuire à l'efficacité du concept de base.

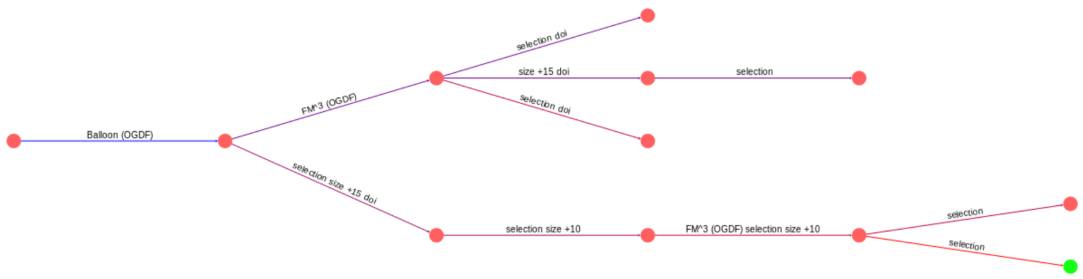


FIGURE *6.5* – ***Arbre de traces*** *: Ceci est l'arbre de trace permettant de naviguer entre les différents sous-réseaux (sommets rouges). Le sommet sur-ligné en vert est le sous-réseau actuellement affiché. En plus de la sélection utilisateur, changer le dessin du sous-réseau et augmenter ou diminuer le nombre de sommets génèrent également un nouveau sous-réseau. Chaque opération ayant permis la création d'un sous-réseaux est spécifiée sur le lien entre le sous-réseau généré et le sous-réseau sur lequel l'opération a été effectuée.*

Dans un deuxième temps, l'arbre de traces (Figure 6.4, C et Figure 6.5) correspondant aux différentes actions et recherches des historiens a été reçu avec beaucoup d'enthousiasme. Son fonctionnement correspond en effet à la manière que les experts ont d'établir et initier des recherches et synergise donc organiquement

avec leur méthode de travail. Actuellement, l'arbre de traces permet de revenir d'un simple clic à un sous-réseau précédemment généré pour l'utiliser comme point de départ pour de nouvelles branches. Les différents dessins générés par l'historien sont sauvegardés et accessibles dans l'arbre de traces. Les experts ont soumis l'idée de plusieurs fonctionnalités additionnelles. Premièrement, inscrire des commentaires à la manière de Post-It® sur les sommets et les branches serait utile pour se repérer facilement avec cette méthode de travail non linéaire. Cette problématique n'est pas spécifique à nos collègues historiens et avait déjà été constatée avec nos collègues juristes et sociologues du projet TETRUM [47] (voir chapitre 2). Deuxièmement, afficher des informations comme le pourcentage des types de sommet présents ou l'état de la sélection sur les sous-réseaux produits pourrait aider à la comparaison et l'analyse des pistes. Ces améliorations ergonomiques permettraient aux experts de ne pas se perdre dans leurs pistes de recherche tout en leur offrant des possibilités analogues à leur méthode de travail habituelle.

Le point suivant concerne l'aspect visuel et les interactions au sein du processus.

Vue et interacteurs

Si le processus utilise un réseau pour effectuer ses calculs et estimations sur les données de manière interne, il est tout à fait possible d'avoir recours à d'autres types de visualisation que des vues noeuds-liens pour présenter les vues partielles à l'utilisateur. Les seules contraintes sont d'avoir un mécanisme de sélection et une visualisation convenant à un ensemble restreint de données. La vue noeuds-liens a été ici privilégiée (Figure 6.4, B) pour sa simplicité d'utilisation et sa mise en avant des liens entre les sommets. Dans les expérience précédentes issues des projets GEOBS et TETRUM (voir chapitre 2), nous avons remarqué que proposer un graphe comme objet d'analyse n'est pas forcément intuitif pour quelqu'un qui n'est pas familier avec la théorie des graphes. Avec BLIZAAR, il a été étonnant de constater que la vue noeuds-liens a été uniformément bien accueillie par les experts des données. Cela peut supposément être attribué à la taille réduite des réseaux produits. Par soucis de clarté, les sommets ont en outre une forme et une couleur déterminant leurs types, permettant d'avoir intuitivement une idée des couches observées et des interactions entre sommets. Les réseaux obtenus sont donc facilement lisibles et analysables bien que des problèmes de chevauchement de textes aient été observés. Le nombre de sommets des sous-réseaux est néanmoins paramétrable et peut être augmenté ou diminué graduellement pour un même sous-réseau. La valeur par défaut est de 50 sommets mais l'intégralité des experts ont

préféré établir des sous-réseaux plus grands (de 75 à 200 sommets). Si un réseau petit est facilement lisible, l'intérêt des utilisateurs est basé sur les interactions et les liens et nécessite donc suffisamment de sommets pour voir des communautés et liens entre les communautés se dessiner.

En plus de l'aspect visuel se pose aussi la question de l'interactivité afin d'assurer une utilisation fluide et facile de `M-QuBE`^{3}. Pour commencer le processus, l'utilisateur sélectionne le(s) sommet(s) de départ en utilisant une recherche par mot-clé via un panneau latéral dédié (Figure 6.4, A). Un filtrage par couche de la liste des sommets permet en outre de faciliter le sélection. Dans la vue proposée, l'outil de sélection est basé sur un interacteur de type "lasso" (Figure 6.6) demandant de tracer une zone autour des sommets à sélectionner. Pour avoir des informations complémentaires sur les sommets présents dans le sous-réseau ou accéder aux documents intégraux associés, il suffit de cliquer sur un sommet pour avoir accès à un bulle contenant les informations souhaitées et un lien vers le document complet (Figure 6.7). En outre, survoler un sommet avec le curseur met en surbrillance son voisinage afin de mettre en valeur les interactions avec les autres sommets et ainsi faciliter la lecture du graphe. Ces fonctionnalités basiques ont été jugées suffisantes et satisfaisantes et n'ont pas posé de difficultés aux experts des données.

Le dernier point concerne le score d'intérêt et le choix des sommets à exposer.

Score et sommets

La question de la pertinence du choix des sommets montrés est l'élément le plus difficile à évaluer. Étant donné que la première exécution de `M-QuBE`^{3} se veut générique, elle doit donner un rapide aperçu du contexte immédiat autour de la sélection de l'utilisateur et simultanément faire le pont vers d'autres thèmes ou éléments intéressants au sein du réseau (i.e. proposer des chemins vers d'autres sommets ou communautés de sommets). Cependant, ce comportement est à pondérer par la notion de distance utilisée dans l'estimation de score (voir section 5).

En effet, lorsqu'un élément est sélectionné, son entourage à proximité est valorisé (i.e. son score devient plus élevé). Lorsque plusieurs sommets sont sélectionnés, les sommets les plus proches de chacun des sommets de la sélection vont être valorisés. Ces sommets déterminent une "zone focale" : une zone délimitée par les sommets sélectionnés par l'utilisateur et composée des sommets aux scores de distance les plus élevés. Une sélection de plusieurs sommets autour du même thème ou dans une même communauté va donc augmenter les chances de proposer des

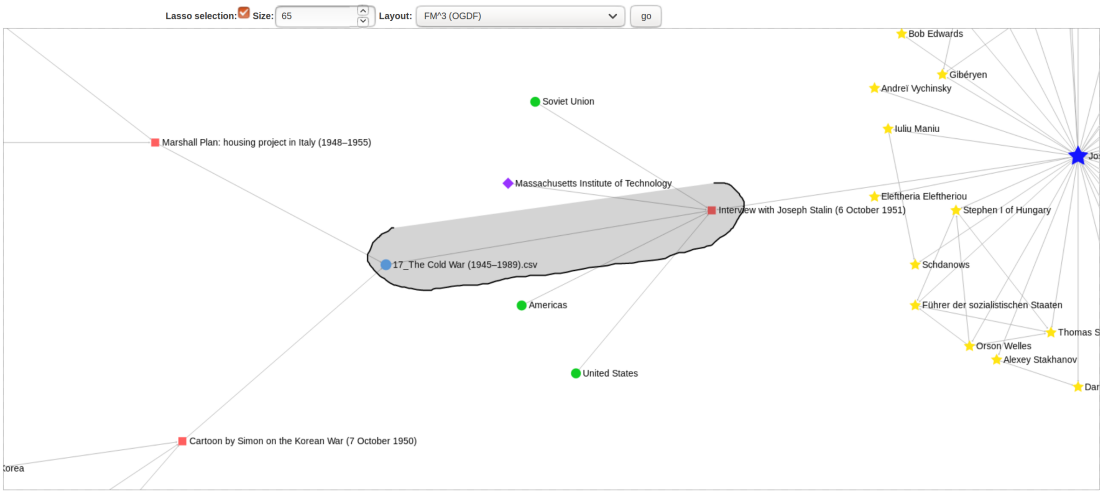


FIGURE *6.6* – ***Sélecteur Lasso*** *: Ceci est l'intéracteur permettant de sélectionner des sommets dans les différents sous-réseaux. Tout les sommets compris dans la zone délimitée par le mouvement du curseur sont ajoutés à la sélection actuelle et génèrent un nouveau sous-réseau basé sur cette nouvelle sélection.*

sommets autour de ce thème ou cette communauté en déplaçant la zone focale vers cette zone d'intérêt du réseau.

Si ce mécanisme permet d'amplifier le score d'intérêt dans une zone du réseau, cela ne se substitue pas au deuxième passage de `M-QuBE`^{3}. C'est en effet ce qui a été observé lors des différents scénarios des experts. Plusieurs sommets présentés dans les sous-graphes ont été jugés étonnants ou surprenants. S'il y a des sommets attendus et directement liés aux sujets, certains ont un lien très indirect avec la sélection et y sont reliés par des chemins de taille élevée rendant difficile d'établir un lien sémantique avec les sommets de la sélection. Ce comportement est néanmoins attendu : parce que la version évaluée est paramétrée sans favoriser un contexte sémantique précis, `M-QuBE`^{3} tend à faire ressortir des éléments centraux du réseau sans que ceux-ci ne soient directement liés à la sélection.

C'est le rôle et l'objectif du deuxième passage qui va mettre l'accent sur un contexte précis et permettre de faire une recherche localisée sur le réseau. Un des experts a trouvé cette méthode en deux temps perturbante. Mixer ces deux approches en permettant notamment d'ajouter à la volée de nouvelles fonctions de pilotage permettrait de créer dans l'arbre de trace des branches spécialisées qui se substituerait au deuxième passage de `M-QuBE`^{3}. Ce faisant, la comparaison entre les

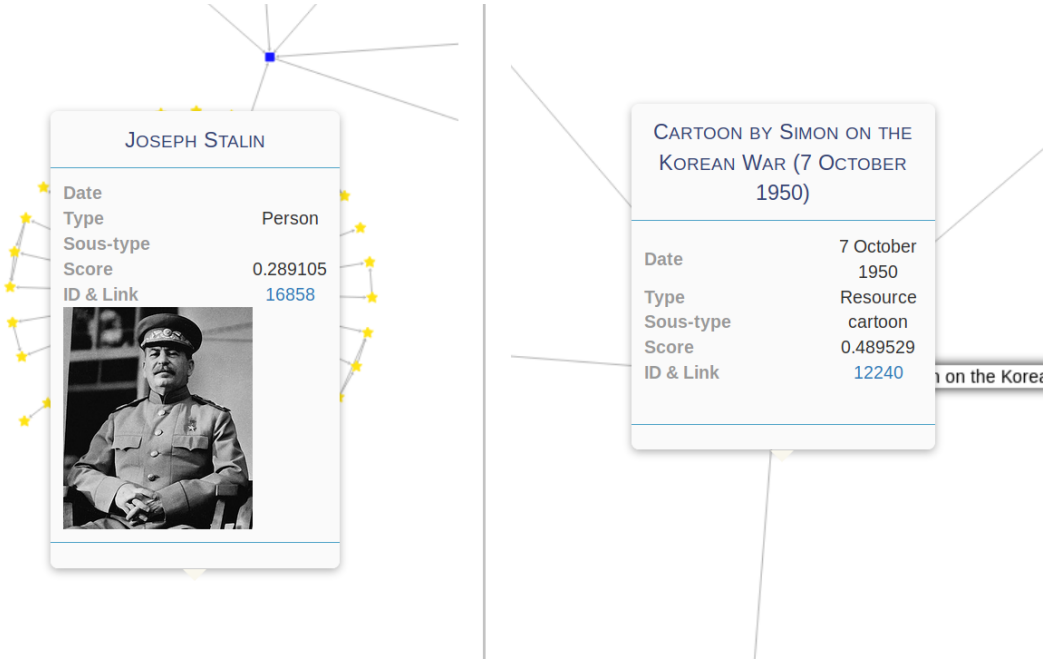


FIGURE *6.7* – ***Bulle informative*** *: Une bulle d'information apparaît lorsque l'on clique sur un sommet. Celle-ci renseigne l'utilisateur sur le score d'intérêt et le type du sommet. Certaines informations facultatives peuvent aussi être disponibles comme l'éventuel sous-type du sommet, une image le représentant et/ou un lien vers la page correspondant à l'entité représentée par le sommet sur le site du CVCE pour obtenir davantage d'informations.*

différentes branches serait néanmoins biaisée car l'évaluation des scores ne s'effectuerait plus nécessairement de la même manière. Il faudrait alors prévenir l'expert de n'établir de comparaisons qu'au sein d'une branche dont la spécialisation (i.e. les fonctions de pilotage) est identique.

6.3 Synthèse

Une plateforme ayant évoluée au fil du projet BLIZAAR a été mise en place pour accueillir et proposer le processus `M-QuBE`^{3} décrit au chapitre 4.

Cette procédure d'évaluation avait deux objectifs : prendre connaissance des ressentis des experts vis à vis des sous-réseaux obtenus et estimer l'efficacité de l'approche pour explorer des données inconnues. Les experts ayant pour certains

une connaissance très large des données, il a pu être estimé si les sous-réseaux permettaient de mettre en lumière les éléments importants amenant à une compréhension générale du réseau. Le processus M-QuBE^3 a ainsi été unanimement considéré comme intéressant pour explorer un nouveau jeu de données par les experts.

Cependant, le concept d'établir une analyse complète en deux temps à travers deux exécutions a semblé être perturbante pour les experts. La gène majeure a souvent été de ne pas pouvoir se concentrer sur un domaine ou un thème du réseau lors de l'exploration, ce qui est habituellement effectué lors du second passage du processus. Établir un pont entre ces deux passages semble important afin de passer organiquement du premier passage au second passage. Cela permettrait d'optimiser l'expérience utilisateur et fluidifier l'expérience.

Il manque en outre un certain nombre de fonctionnalités d'amélioration de la qualité de vie notamment la possibilité de pouvoir marquer des informations sur les branches d'explorations, les sous-réseaux ou les sommets. Le concept général est donc considéré efficace et fonctionnel mais il reste des améliorations en terme d'interface et d'aide aux utilisateurs pour optimiser la procédure et la rendre plus accessible.

Chapitre 7

Conclusion et perspectives

L'ensemble des travaux présentés dans ce manuscrit, y compris les travaux préliminaires (Section 2.3), ont été réalisés avec des experts issus des sciences humaines et sociales. A travers ces collaborations s'est dégagé un scénario commun. A chaque fois leurs données se sont résumées en sous systèmes inter-connectés traduisibles en réseaux multi-couches. Les experts ont voulu visualiser et explorer ces données en accordant une attention particulière aux individus et en utilisant leurs métriques propres (Chapitre 3).

Nous avons donc développé `M-QuBE`^{3} (Chapitre 4), un mécanisme d'exploration incrémental permettant d'améliorer graduellement la pertinence des visualisations en fonction d'un score d'intérêt déterminé par les actions de l'utilisateur. Afin d'être au plus près des besoins des experts, ce dernier utilise l'**eScore** (Chapitre 5), une méthode itérative permettant de calculer l'intérêt de l'utilisateur pour les différents sommets d'un réseau à travers un mécanisme de sélections successives spécialisé pour les graphes multi-couches.

Ces deux travaux complémentaires ont fait l'objet d'une publication commune [44] à Electronic Imaging (Visualization and Data Analysis 2019) et ont été implémentés à travers la plateforme BLIZAAR qui a été validée par les retours d'experts des données (Chapitre 6). D'un point de vue de la visualisation, les solutions pour visualiser les graphes multi-couches [21,56] peuvent bénéficier du duo `M-QuBE`^{3} et eScore, ceux-ci proposant en même temps une application adaptée à ces visualisations ainsi qu'une procédure d'exploration permettant de placer l'utilisateur aux commandes de sa navigation, un besoin précédemment identifié par McGee et al. [57].

Dans la suite, nous indiquons les démarches restantes à effectuer pour **compléter la procédure d'évaluation** puis proposons deux axes potentiels d'amé-

lioration : **accroître la confiance** de l'utilisateur et **faciliter la configuration et la prise en main** de nos méthodes.

Évaluation Un des objectifs du CVCE est celui de la valorisation des connaissances : permettre à des utilisateurs tiers (comme les visiteurs de leur site) de se familiariser avec les données, de les comprendre et d'y naviguer. Notre processus de validation était axé sur l'objectif d'exploration dédié aux experts du CVCE et n'a pas été effectué à plus grande échelle avec des personnes n'ayant potentiellement aucune connaissance des données et/ou de leur contexte. En plus de cela, même si `M-QuBE`[3] et eScore ont été construits dans le contexte d'une coopération avec le CVCE, ils se veulent génériques car applicables dans d'autres contextes des sciences humaines et sociales. Si l'on a eu des données et expériences de trois domaines différents (géographique, socio-légal et historique), le processus de validation n'a été effectué qu'avec des experts issus du CVCE, certes avec des objectifs et parcours différents, mais néanmoins dans un domaine d'application général restreint. Un dernier élément relatif au processus de validation est le fait de comparer nos méthodes à celles existantes. Comme dit précédemment, il y a peu de méthodes consacrées à l'exploration en sciences humaines et sociales [65] et conséquemment encore moins si l'on considère en plus le caractère multi-couche des réseaux. La comparaison pourrait néanmoins s'effectuer non pas avec des méthodes identiques mais avec des méthodes partiellement correspondantes i.e proposant certaines fonctionnalités similaires. Évidemment, de tels comparatifs apportent des biais à prendre en compte mais permettraient néanmoins d'analyser individuellement et sous un autre angle les différents aspects de la validation.

Confiance et contexte Un des retours qui est souvent revenus dans nos expériences est le problème de confiance : "Comment puis-je être certain que les éléments qui me sont présentés sont les éléments les plus intéressants ? Est-ce qu'il n'y a pas des trésors cachés en arrière plan qui ne me sont pas présentés ?". Cette crainte n'est pas triviale à faire disparaître mais des travaux sur le contexte des vues partielles pourraient aider à atténuer ce sentiment. En hybridant des vues centrées sur les individus, comme actuellement, avec des représentations permettant de les situer dans le réseau global, l'utilisateur pourrait peut être plus facilement se situer sémantiquement et ainsi se guider plus efficacement dans les données. De la même manière, proposer des informations annexes supplémentaires (nombre total d'éléments du réseau concernant une thématique lié à l'élément observé, position de cet élément dans un classement basé un fonction de pilotage précise, etc.) pourrait

faciliter le contextualisation. Enfin, enrichir l'arbre de trace avec des informations sur les sous-réseaux et leurs sélections pourrait permettre aux experts d'établir des comparaisons entre les sous-réseaux et ainsi seconder les prises de décisions quant aux pistes à développer ou à créer.

Configuration et usages Un autre aspect est celui du paramétrage des fonctions de pilotage. Actuellement la configuration de ces fonctions se fait après de longs entretiens avec les experts des données afin de cerner leurs besoins, les formaliser et les ré-injecter dans nos méthodes. Plusieurs souhaits ou configurations communes se sont dégagés (homogénéité, similarité, etc.), il serait donc possible de proposer aux experts une manière de configurer eux-même les rails de leurs explorations à travers des listes d'options pré-remplies. La difficulté serait alors de rendre cela intuitif et compréhensible pour un public potentiellement non scientifique. La difficulté est encore accrue lorsque l'on considère le fait de faire utiliser M-QuBE^3/eScore a des personnes n'ayant pas spécialement d'objectifs de recherche. Une piste imaginable pour répondre à ces points serait d'essayer d'analyser les usages des spécialistes pour proposer des pré-paramétrages. Il serait même intéressant d'imaginer faire intervenir de l'apprentissage dans le paramétrage voir dans l'estimation des scores en conservant les informations relatives à l'utilisation des autres utilisateurs. Ainsi, on pourrait questionner les utilisateurs en fin d'exploration afin, en cas de satisfaction, d'orienter les prochaines exécutions de la méthode vers les éléments ou pistes ayant été jugés intéressants. Cela pose néanmoins le problème de miser sur la coopération des utilisateurs et un échantillon suffisant de données utilisateur récoltées. Une alternative serait alors de s'appuyer sur l'utilisation d'informations externes : utiliser une base de données annexe ou un équivalent de DBpedia en parallèle dont la fréquence de consultation des pages/entités serait accessible et permettrait de valoriser les éléments les plus consultés (ou, dans le cas de M-QuBE^3 où l'utilisateur a déjà effectué une sélection, valoriser les éléments consultés par les individus ayant déjà consultés d'autres éléments appartenant à la sélection). Pour l'exemple des historiens, le site du CVCE permettrait d'obtenir ces informations notamment lorsque quelqu'un accède à une ePublication ou une autre ressource en ligne.

L'ADN de nos méthodes reste cependant de garder l'utilisateur au centre de la navigation, qu'il ait les manettes en main afin de choisir ses propres directions. Les travaux futurs devront donc trouver un équilibre afin d'accroître l'indépendance de l'utilisateur mais sans pour autant le priver de son pouvoir de décision.

Publications sur les travaux de ce manuscrit

1. *Conférence internationale avec comité de lecture*

   M-QuBE 3 : Querying Big Multilayer Graph by Evolutive Extraction and Exploration. Antoine Laumond, Guy Melançon, Bruno Pinaud, Mohammad Ghoniem. IS&T International Symposium on Electronic Imaging 2019 : Visualization and Data Analysis 2019 proceedings, Jan 2019, San Franscico, United States. (DOI : 10.2352/ISSN.2470-1173.2019.1.VDA-686).

   (https://hal.archives-ouvertes.fr/hal-02016317)

2. *Autre (poster en conférence nationale)*

   eDOI : Exploration par Degrés d'Intérêt, une Exploration Visuelle des Réseaux Multicouches. Antoine Laumond, Guy Melançon, Bruno Pinaud. EGC2018, Jan 2018, Paris, France.

   Poster : (https://hal.archives-ouvertes.fr/hal-02016075),

   Papier descriptif : (https://hal.archives-ouvertes.fr/hal-02178761)

3. *Poster en congrès international avec comité de lecture*

   Exploratory Degree of Interest : a visual interest-based exploration of multilayer networks. Antoine Laumond, Guy Melançon, Bruno Pinaud. VIS 2017, Poster session, Oct 2017, Phoenix, United States.

   Poster : (https://hal.archives-ouvertes.fr/hal-02016056),

   Papier descriptif : (https://hal.archives-ouvertes.fr/hal-01675682)

4. *Journal national avec comité de lecture*

   Analyse de réseaux criminels de traite des êtres humains. Bénédicte Lavaud-Legendre, Cécile Plessard, Antoine Laumond, Guy Melançon, Bruno Pinaud. Journal of Interdisciplinary Methodologies and Issues in Science, Journal of Interdisciplinary Methodologies and Issues in Science, 2017, Graphes et systèmes sociaux, 2, (DOI : 10.18713/JIMIS-300617-2-5).

   (https://hal.archives-ouvertes.fr/hal-01380339v6)

5. *Autre (atelier)*

   Exploration visuelle de graphes multi-couches basée sur un degré d'intérêt. Antoine Laumond, Guy Melançon, Bruno Pinaud. EGC2017, Jan 2017, Grenoble, France.

   (https://hal.archives-ouvertes.fr/hal-01462002)

6. *Autre (workshop)*

   Réseaux multiplexes à travers les sciences sociales : une adaptation et des règles dictées par les données. Antoine Laumond, Bruno Pinaud, Guy Melançon. 7ème conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques (MARAMI 2016), EISTI, Oct 2016, Cergy-Pontoise, France.

   (https://hal.archives-ouvertes.fr/hal-01382598v3)

7. *Conférence nationale avec comité de lecture*

   Un cadre d'analyse des Infrastructures de Données Géographiques pour interroger la mise en réseaux des acteurs et des outils. Matthieu Noucher, Françoise Gourmelon, Antoine Laumond, Guy Melançon, Bruno Pinaud, et al.. SAGEO : Spatial Analysis & Geomatic, Dec 2016, Nice, France.

   (https://hal.archives-ouvertes.fr/halshs-01414234)

Bibliographie

[1] J. Abello, S. Hadlak, H. Schumann, and H. J. Schulz. A modular degree-of-interest specification for the visual analysis of large dynamic networks. *IEEE Trans. on Visualization and Computer Graphics*, 20(3) :337–350, March 2014.

[2] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Analysis of ego network structure in online social networks. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 31–40. IEEE, 2012.

[3] David Auber, Romain Bourqui, Maylis Delest, Antoine Lambert, Patrick Mary, Guy Melançon, Bruno Pinaud, Benjamin Renoust, and Jason Vallet. TULIP 4. Research report, LaBRI - Laboratoire Bordelais de Recherche en Informatique, September 2016.

[4] David Auber, Yves Chiricota, Fabien Jourdan, and Guy Melançon. Multiscale visualization of small world networks. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, pages 75–81. IEEE, 2003.

[5] Gilles Babinet. *Big Data, penser l'homme et le monde autrement.* Le passeur, 2016.

[6] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi : an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*, 2009.

[7] Jöran Beel and Bela Gipp. Google scholar's ranking algorithm : an introductory overview. In *Int. Conf. on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241, 2009.

[8] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social networks*, 21(4) :375–395, 2000.

[9] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *science*, 323(5916) :892–895, 2009.

[10] Dominique Boullier. Les sciences sociales face aux traces du big data. *Revue française de science politique*, 65(5) :805–828, 2015.

[11] Dominique Boullier. Vie et mort des sciences sociales avec le big data. *Socio. La nouvelle revue des sciences sociales*, 4 :19–37, 05 2015.

[12] François Boutin. *Network multi-scaling filtering, clustering and visualisation techniques from a focus*. Theses, Université Montpellier II - Sciences et Techniques du Languedoc, November 2005.

[13] David A. Bright. *Using Social Network Analysis to Design Crime Prevention Strategies : A Case Study of Methamphetamine Manufacture and Trafficking*, pages 143–164. Springer International Publishing, Cham, 2017.

[14] Ken Brodlie, Andrew Poon, Helen Wright, Lesley Brankin, Greg Banecki, and Alan Gay. Grasparc-a problem solving environment integrating computation and visualization. In *Proceedings Visualization'93*, pages 102–109. IEEE, 1993.

[15] Roger Burrows and Mike Savage. After the crisis ? big data and the methodological challenges of empirical sociology. *Big data & society*, 1(1) :2053951714540280, 2014.

[16] Stuart K Card and David Nation. Degree-of-interest trees : A component of an attention-reactive user interface. In *AVI '02*, pages 231–245. ACM, 2002.

[17] Abhijit Chatterjee, PP Das, and Soumendu Bhattacharya. Visualization in linear programming using parallel coordinates. *Pattern Recognition*, 26(11) :1725–1736, 1993.

[18] Auguste Comte. *Social physics : from the positive philosophy of Auguste Comte*. Calvin Blanchard, 1856.

[19] Jingyu Cui, Fang Wen, and Xiaoou Tang. Real time google and live image search re-ranking. In *Proc. of the 16th ACM Int. Conf. on Multimedia*, pages 729–732. ACM, 2008.

[20] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proc. of the* 4^{th} *ACM Conf. on Recommender systems*, pages 293–296. ACM, 2010.

[21] Manlio De Domenico, Mason A. Porter, and Alex Arenas. Muxviz : a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks*, 3(2) :159–176, 2015.

[22] Mark E Dickison, Matteo Magnani, and Luca Rossi. *Multilayer social networks*. Cambridge University Press, 2016.

[23] Marek Dudáš, Steffen Lohmann, Vojtěch Svátek, and Dmitry Pavlov. Ontology visualization methods and tools : a survey of the state of the art. *The Knowledge Engineering Review*, 33, 2018.

[24] Marten Düring, Lars Wieneke, and Vincenzo Croce. Interactive networks for digital cultural heritage collections-scoping the future of histograph. In *International Conference on Web Engineering*, pages 613–616. Springer, 2015.

[25] Durkheim. *Le Suicide : Étude de sociologie*. F. Alcan, 1897.

[26] Paolo Federico, Wolfgang Aigner, Silvia Miksch, Florian Windhager, and Lukas Zenk. A visual analytics approach to dynamic social networks. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, New York, NY, USA, 2011. Association for Computing Machinery.

[27] G. W. Furnas. Generalized fisheye views. In *SIGCHI Conf. '86*, pages 16–23. ACM, 1986.

[28] Jade Georis-Creuseveau, Christophe Claramunt, Françoise Gourmelon, Bruno Pinaud, and Laurence David. A diachronic perspective on the use of french spatial data infrastructures. *Journal of Geographic Information System*, 10(04) :344, 2018.

[29] Sohaib Ghani, Bum Chul Kwon, Seungyoon Lee, Ji Soo Yi, and Niklas Elmqvist. Visual analytics for multimodal social network analysis : A design study with social scientists. *IEEE transactions on visualization and computer graphics*, 19(12) :2032–2041, 2013.

[30] Mohammad Ghoniem, J-D Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 17–24. Ieee, 2004.

[31] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. On the readability of graphs using node-link and matrix-based representations : a controlled experiment and statistical analysis. *Information Visualization*, 4(2) :114–135, 2005.

[32] Fred Gibbs. Digital humanities definitions by type. In *Defining Digital Humanities*, pages 305–314. Routledge, 2016.

[33] Stefan Hachul and Michael Jünger. Drawing large graphs with a potential-field-based multilevel algorithm. In *International Symposium on Graph Drawing*, pages 285–295. Springer, 2004.

[34] Torsten Hägerstrand. What about people in regional science ? *Papers in regional science*, 24(1) :6–21, 1970.

[35] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 170(2) :301–354, 2007.

[36] Peter Hüsken and Jürgen Ziegler. Degree-of-interest visualization for ontology exploration. *Human-Computer Interaction–INTERACT 2007*, pages 116–119, 2007.

[37] Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37 :241–72, 01 1901.

[38] Paul Janecek and Pearl Pu. Opportunistic search with semantic fisheye views. In *International Conference on Web Information Systems Engineering*, pages 668–680. Springer, 2004.

[39] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics : Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[40] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1) :2053951714528481, 2014.

[41] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3) :203–271, 07 2014.

[42] M.J. Kraak. The space-time cube revisited from a geovisualization perspective. *Proc 21st Int Cartogr Conf*, 07 2008.

[43] Douglas Laney. 3D data management : Controlling data volume, velocity, and variety. Technical report, META Group, February 2001.

[44] Antoine Laumond, Guy Melançon, Bruno Pinaud, and Mohammad Ghoniem. M-QuBE 3 : Querying Big Multilayer Graph by Evolutive Extraction and Exploration. In *IS&T International Symposium on Electronic Imaging 2019 : Visualization and Data Analysis 2019 proceedings*, San Franscico, United States, January 2019.

[45] Antoine Laumond, Bruno Pinaud, and Guy Melançon. Réseaux multiplexes à travers les sciences sociales : une adaptation et des règles dictées par les données. In *7ème conférence sur les modèles et l'analyse des réseaux : Approches*

mathématiques et informatiques (MARAMI 2016), Cergy-Pontoise, France, October 2016. EISTI.

[46] Bénédicte Lavaud-Legendre. Autonomie et protection des personnes vulnérables : le cas des femmes nigérianes se prostituant en France. Technical report, CNRS/COMPTRASEC, January 2012.

[47] Bénédicte Lavaud-Legendre, Cécile Plessard, Antoine Laumond, Guy Melançon, and Bruno Pinaud. Analyse de réseaux criminels de traite des êtres humains. *Journal of Interdisciplinary Methodologies and Issues in Science*, 2, 2017.

[48] Emmanuel Lazega and Tom AB Snijders. *Multilevel network analysis for the social sciences : Theory, methods and applications*, volume 12. Springer, 2015.

[49] Antoine Lhuillier and Christophe Hurter. Bundling, graph simplification trough visual aggregation : existing techniques and challenges. In *Proceedings of the 27th Conference on l'Interaction Homme-Machine*, page 10. ACM, 2015.

[50] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. immens : Real-time visual querying of big data. *Computer Graphics Forum*, 32(3pt4) :421–430, 2013.

[51] Jean-Louis Loubet. *« FARDIER DE CUGNOT »*. Encyclopædia Universalis SA, 2019.

[52] Timothy Luciani, Andrew Burks, Cassiano Sugiyama, Jonathan Komperda, and G Elisabeta Marai. Details-first, show context, overview last : supporting exploration of viscous fingers in large-scale ensemble simulations. *IEEE transactions on visualization and computer graphics*, 25(1) :1–11, 2018.

[53] Anders Koed Madsen, Mikkel Flyverbom, Martin Hilbert, and Evelyn Ruppert. Big data : Issues for an international political sociology of data practices. *International Political Sociology*, 10(3) :275–296, 2016.

[54] Efstathios D. Mainas. The analysis of criminal and terrorist organisations as social network structures : A quasi-experimental study. *International Journal of Police Science & Management*, 14(3) :264–282, 2012.

[55] Stephen P. Borgatti Martin Everett. Ego network betweenness. *Social Networks*, 27(1) :31 – 38, 2005.

[56] F. McGee, M. Ghoniem, G. Melançon, B. Otjacques, and B. Pinaud. The state of the art in multilayer network visualization. *Computer Graphics Forum*, 38(6) :125–149, 2019.

[57] Fintan McGee, Marten During, and Mohammad Ghoniem. Towards visual analytics of multilayer graphs for digital cultural heritage. *Towards Visual Analytics of Multilayer Graphs for Digital Cultural Heritage*, 2016.

[58] Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E Tarjan. Clustering social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 56–67. Springer, 2007.

[59] Jacob Levy Moreno. Who shall survive ? : A new approach to the problem of human interrelations. *Nervous and mental disease monograph series, no 58.*, 1934.

[60] Pierre Mounier-Kuhn. *« DÉVELOPPEMENT DU RÉSEAU INTERNET »*. Encyclopædia Universalis SA, 2019.

[61] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6) :921–928, 2009.

[62] Matthieu Noucher, Françoise Gourmelon, Antoine Laumond, Guy Melançon, Bruno Pinaud, Adeline Maulpoix, Julie Pierson, Olivier Pissoat, and Mathias Rouan. Un cadre d'analyse des Infrastructures de Données Géographiques pour interroger la mise en réseaux des acteurs et des outils. In *SAGEO : Spatial Analysis & Geomatic*, Nice, France, December 2016.

[63] Étienne Ollion. L'abondance et ses revers. big data, open data et recherches sur les questions sociales. *Informations sociales*, 191(5) :70–79, 2015.

[64] Étienne Ollion and Julien Boelaert. Au-delà des big data. les sciences sociales et la multiplication des données numériques. *Sociologie*, 3(6), 2015.

[65] Adam Perer and Ben Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE transactions on visualization and computer graphics*, 12(5) :693–700, 2006.

[66] B. Pinaud, G. Melançon, and J. Dubois. Porgy : A visual graph rewriting environment for complex systems. *Computer Graphics Forum*, 31(3pt4) :1265–1274, 2012.

[67] Frits H Post, Gregory Nielson, and Georges-Pierre Bonneau. *Data visualization : The state of the art*, volume 713. Springer Science & Business Media, 2002.

[68] Quentin Rossy, Sylvain Ioset, Damien Dessimoz, and Olivier Ribaux. Integrating forensic information in a crime intelligence database. *Forensic science international*, 230(1-3) :137–146, 2013.

[69] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1) :3–55, 2001.

[70] Ben Shneiderman. The eyes have it : A task by data type taxonomy for information visualizations. In *IEEE Symp. on Vis. Lang.*, pages 336–343, 1996.

[71] Yedendra Babu Shrinivasan and Jarke J. van Wijk. Supporting the analytical reasoning process in information visualization. In *SIGCHI Conf. '08*, pages 1237–1246. ACM, 2008.

[72] Nelson Silva, Lin Shao, Tobias Schreck, Eva Eggeling, and Dieter W Fellner. Visual exploration of hierarchical data using degree-of-interest controlled by eye-tracking. In *FMT*, pages 82–89, 2016.

[73] Jason Vallet, Guy Melançon, and Bruno Pinaud. Jasper : Just a new space-filling and pixel-oriented layout for large graph overview. *Electronic Imaging*, 2016(1) :1–10, 2016.

[74] Luc Van Campenhoudt, Jean-Michel Chaumont, and Abraham Franssen. La méthode d'analyse en groupe. *Paris, Dunod*, 2005.

[75] Stef Van den Elzen and Jarke J Van Wijk. Multivariate network exploration and presentation : From detail to overview via selections and aggregations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12) :2310–2319, 2014.

[76] Frank Van Ham and Adam Perer. "search, show context, expand on demand" : Supporting large graph exploration with degree-of-interest. *IEEE Trans. on Vis. and Comp. Graph.*, 15(6) :953–960, 2009.

[77] Frank van Ham and Jarke J. van Wijk. Interactive visualization of small world graphs. In *IEEE Symp. Infor. Vis.*, pages 199–206, 2004.

[78] Jarke J Van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86. IEEE, 2005.

[79] JPRB Walton. Now you see it-interactive visualisation of large datasets. *WIT Transactions on Information and Communication Technologies*, 3, 1970.

Annexe A

Cartes figuratives et approximatives

Charles Joseph Minard (1781-1870) est l'un des précurseurs de la visualisation moderne. Il a notamment créé durant sa vie cinquante-et-une cartes dont certaines (nommées "Cartes figuratives et approximatives") proposent notamment une représentation de multiples types de données au sein d'une même représentation.

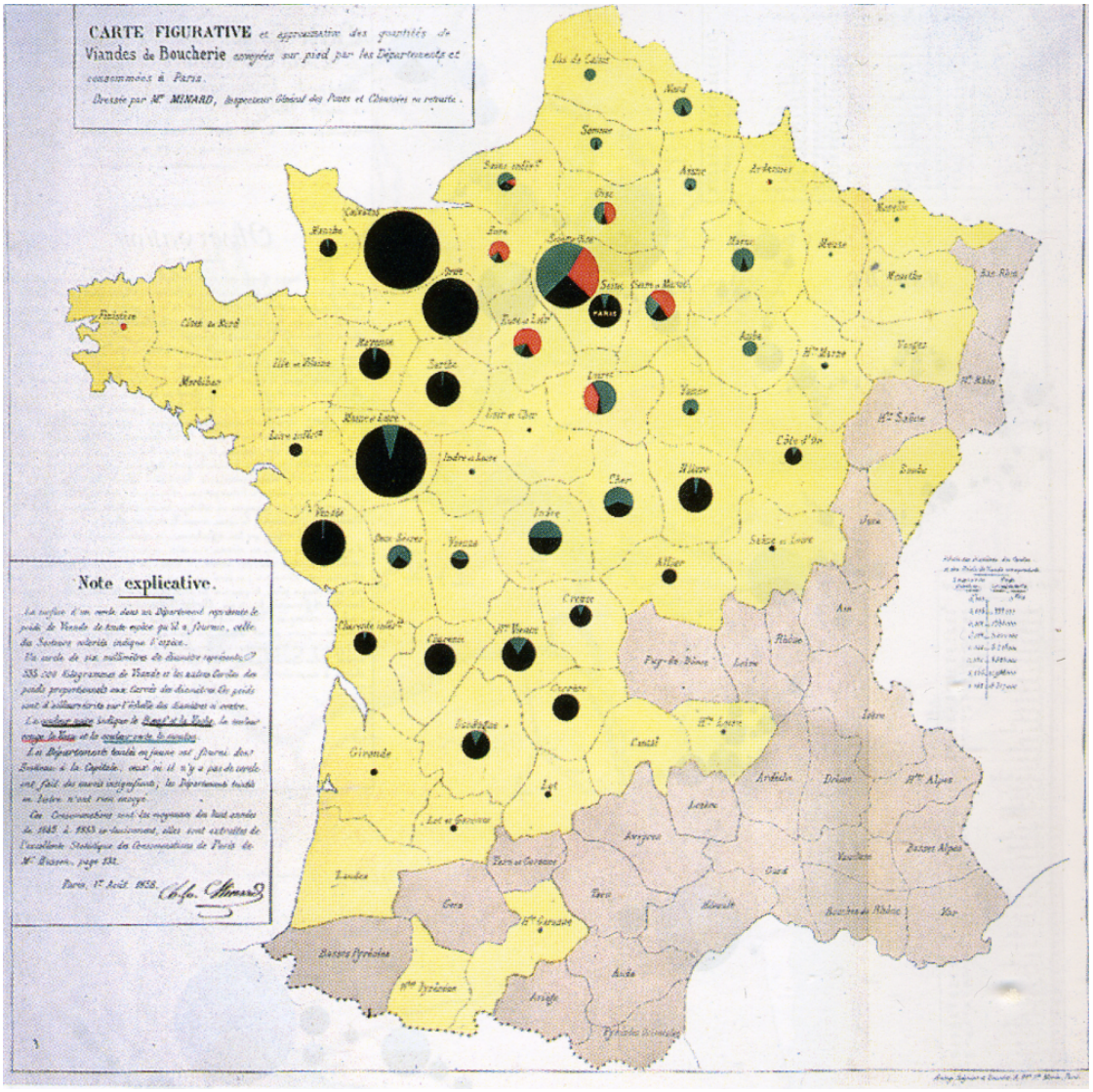


FIGURE *A.1 – Carte figurative et approximative des quantités de viandes de boucherie envoyées sur pied par les départements et consommateurs à Paris*

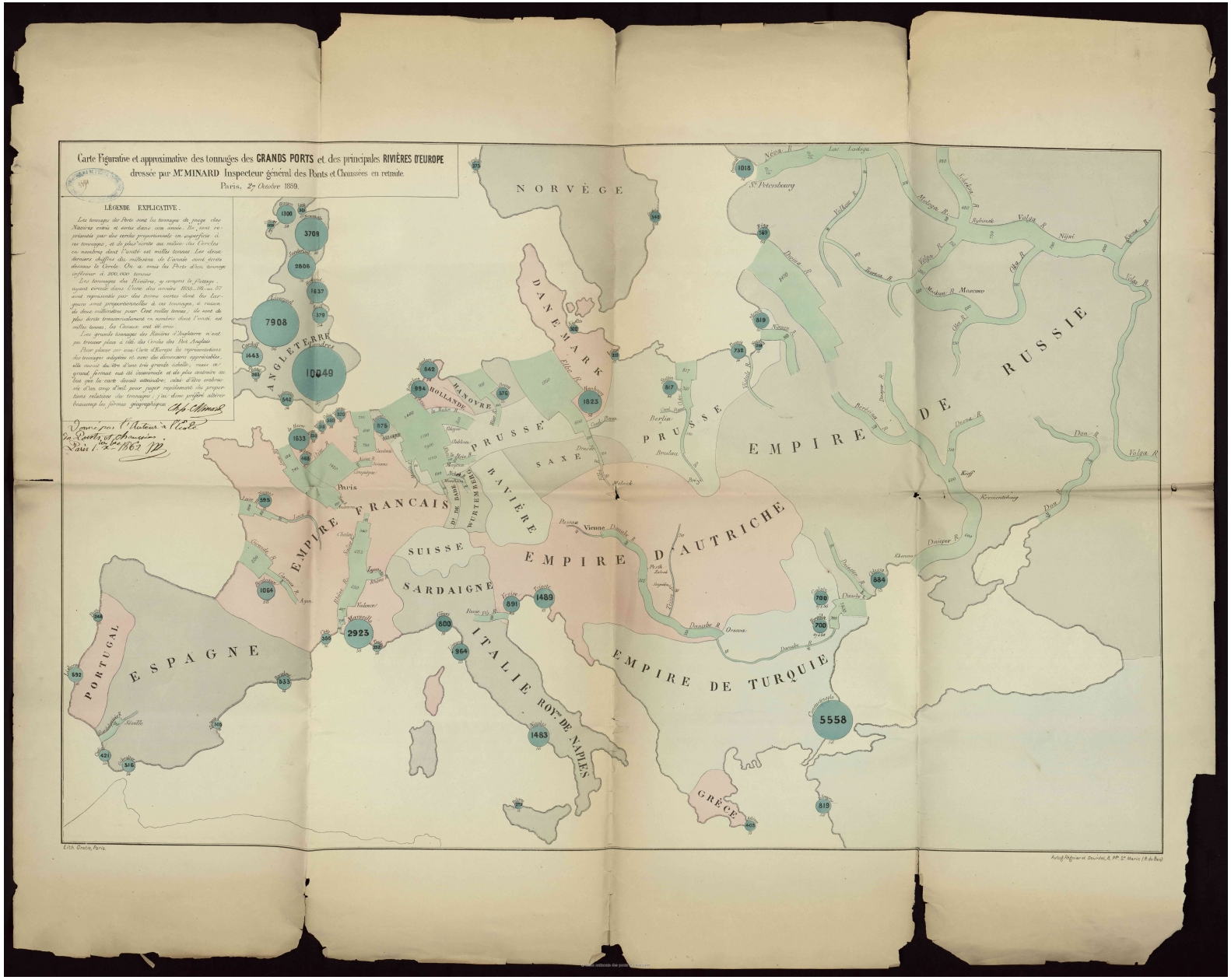


FIGURE *A.2* – *Carte figurative et approximative des tonnages des Grands Ports et des principales Rivières d'Europe*

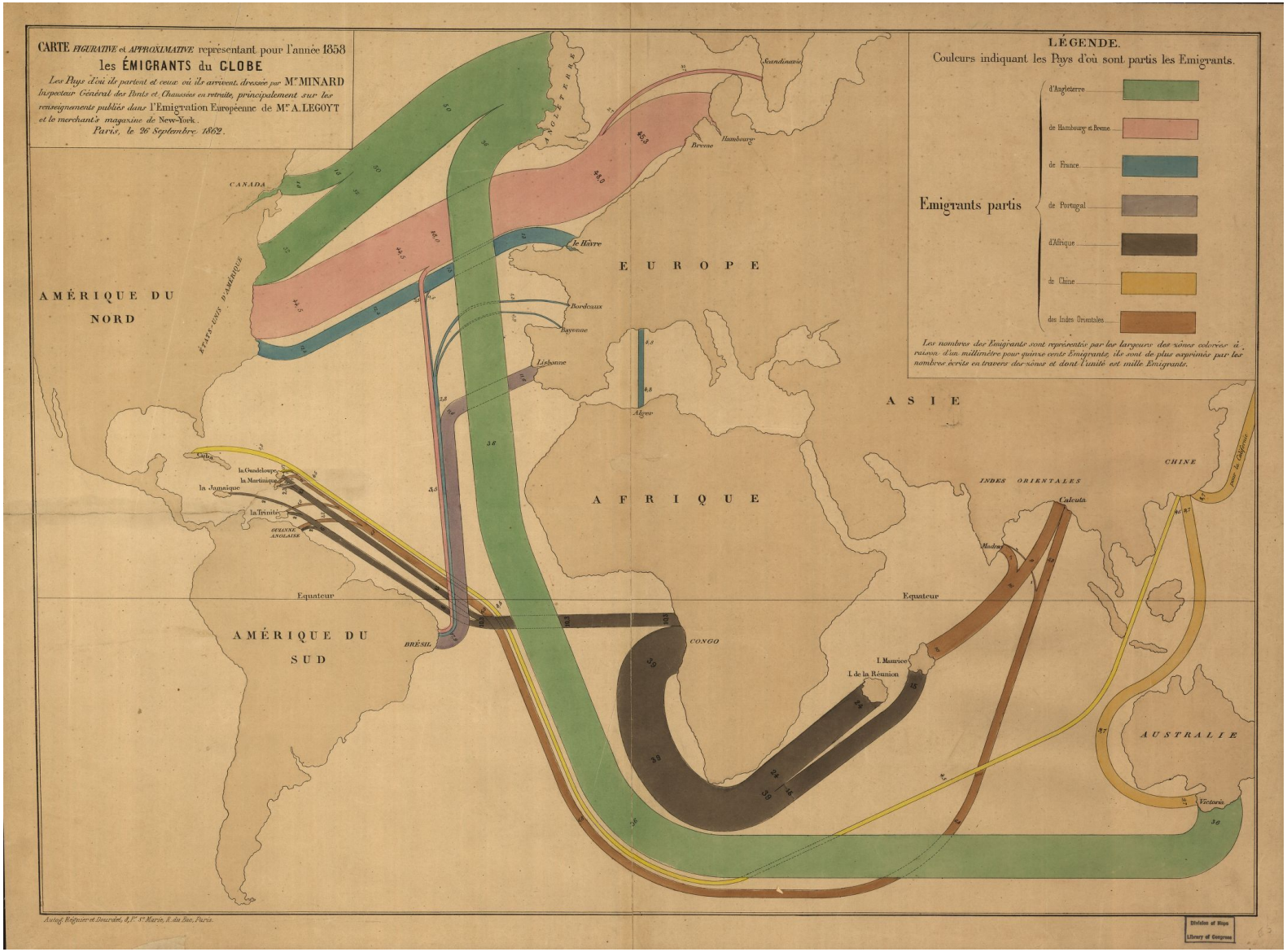


FIGURE *A.3 – Carte figurative et approximative représentant pour l'année 1858 les émigrants du globe*

Annexe B

Questionnaire de validation

Ceci est le questionnaire remis aux experts des données lors de l'évaluation. Les questions s'articulent autour de trois thématiques permettant d'évaluer les différents aspects de M-QuBE^3 : le concept général (dont l'utilisation de l'arbre de trace), la vue et les interactions qui en découlent et enfin le choix de sommets (à travers le score d'intérêt). Le détail des retours des utilisateurs est indiqué dans le chapitre 6).

M-QuBEEE : Querying Big Multilayer Network by Evolutive Extraction and Exploration

1. **Le fonctionnement du processus est-il facilement compréhensible ?**

   Difficile 1 2 3 4 5 6 7 Facile

2. **Est-il simple d'interagir avec l'application ?**

   Peu accessible 1 2 3 4 5 6 7 Facilement utilisable

3. **La vue noeud-lien vous semble t-elle satisfaisante ?**

   Pas du tout 1 2 3 4 5 6 7 Entièrement

4. **La taille des vues est actuellement autour de 50 sommets, quelle est selon vous la taille idéale ?**

   - ❏ Beaucoup moins (**-** 51% ou davantage)
   - ❏ Moins (**-** 1 à 50%)
   - ❏ Correct
   - ❏ Plus (**+** 1 à 50%)
   - ❏ Beaucoup plus (**+** 51% ou davantage)

5. **A combien estimez vous l'importance du graphe de trace ?**

   Dispensable 1 2 3 4 5 6 7 Très important

6. **Comment estimez-vous la facilité d'utilisation du graphe de trace ?**

   Difficile 1 2 3 4 5 6 7 Facile

7. **Quelle fonctionnalité serait-il intéressant de rajouter à ce graphe de trace ?**

8. **Combien de sommets attendus sont accessibles dans les sous-réseaux ?**

Peu	1	2	3	4	5	6	7	Beaucoup

9. **Comment définiriez-vous la qualité des sommets inattendus ?**

Peu pertinents	1	2	3	4	5	6	7	Pertinents

10. **Comment définiriez-vous de manière générale les sommets obtenus ?**

Peu pertinents	1	2	3	4	5	6	7	Pertinents

11. **Quelle est votre appréciation générale du processus d'exploration ?**

Pénible	1	2	3	4	5	6	7	Satisfaisant

12. **Ce processus vous semble-il intéressant pour découvrir un nouveau jeu de données ?**

Peu pratique	1	2	3	4	5	6	7	Pratique

Observations diverses :

13. **Interactivité :**

14. **Vue :**

15. **Autre :**