Analyse et visualisation de l'intrication d'arêtes dans les réseaux multiplex

Benjamin Renoust

Graduate School of Mathematics and Computer Science (EDMI)
University of Bordeaux

Analysis and Visualisation of Edge Entanglement in Multiplex Networks

by Benjamin Renoust

A thesis submitted in fulfillment of the requirements for the degree of Philosophy Doctor in Computer Science

defended on December 18, 2013 before the committee:

President:	Frédérique Segond	Viseo
Reviewers:	Fabien Gandon	INRIA Sophia Antipolis
	Georges Grinstein	University of Massachusetts Lowell
Examiners:	Alberto Cottica	EdgeRyders
	Fleur Mougin	University of Bordeaux
Advisors:	Guy Melançon	University of Bordeaux
	Marie-Luce Viaud	National Institute for Audiovisual

Thesis number: 4985

ABSTRACT

When it comes to comprehension of complex phenomena, humans need to understand what interactions lie within them. These interactions are often captured with complex networks. However, the interaction pluralism is often shallowed by traditional network models. We propose a new way to look at these phenomena through the lens of *multiplex* networks, in which *catalysts* are drivers of the interaction through *substrates*.

To study the *entanglement* of a multiplex network is to study how edges intertwine, in other words, how catalysts interact. Our entanglement analysis results in a full set of new objects which completes traditional network approaches: the *entanglement homogeneity and intensity* of the multiplex network, and the *catalyst interaction network*, with for each catalyst, an *entanglement index*.

These objects are very suitable for embedment in a visual analytics framework, to enable *comprehension* of a complex structure. We thus propose a visual setting with coordinated multiple views. We take advantage of mental mapping and visual linking to present simultaneous information of a multiplex network at three different levels of abstraction. We complete brushing and linking with a *leapfrog* interaction that mimics the back-and-forth process involved in users' comprehension.

The method is validated and enriched through multiple applications including assessing group cohesion in document collections, and identification of particular associations in social networks.

Keywords: *Networks; Graphs; Complex Networks; Social Networks; Multiplex Networks; Document Networks; Document Analysis; Analysis; Network Analysis; Visual Analytics; Sense-making; Visualization; Interaction; Design; Multiple Coordinated Views; Mental Mapping; Brushing and Linking;*

RÉSUMÉ

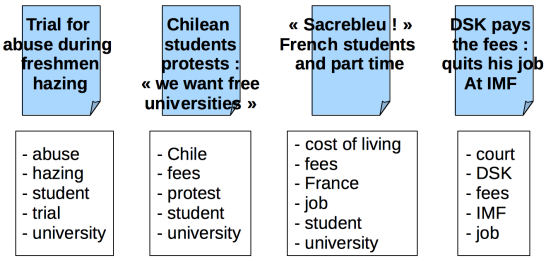
Les travaux de cette thèse montrent comment, depuis l'analyse de corpus de documents de l'Institut National de l'Audiovisuel (INA), nous avons mis en place une méthodologie d'analyse d'intrication dans les réseaux multiplexes. Cette analyse enrichie par le biais d'interactions et de visualisations pertinentes est ensuite appliquée à des domaines très variés, allant de l'analyse des réseaux sociaux jusqu'aux réseaux financiers. La première partie de ce document traite essentiellement de la nature de nos données et de l'analyse des documents sous la perspective des réseaux multiplexes. La seconde partie de ce manuscrit argumente sur la pertinence et l'implantation de nos outils analytiques dans un système de visualisation, notamment au travers d'exemples et applications. La dernière partie de notre document met finalement en jeu l'analyse visuelle de l'intrication d'un réseau multiplexe dans de nombreux domaines d'application.

Modèles et données

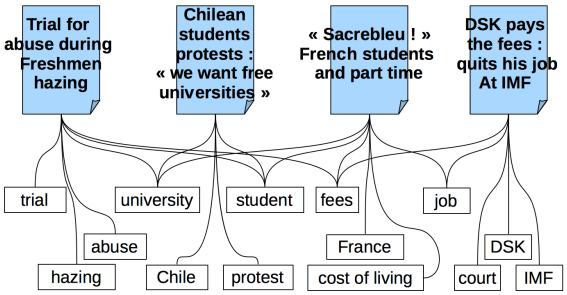
Notre premier chapitre justifie et détaille les applications des modèles de graphe, et tout particulièrement celui du graphe multiplexe. Cette partie pose aussi les motivations de notre recherche, c'est-à-dire le cadre de l'analyse de documents. En premier lieu, nous présentons les groupes de documents de l'INA enrichis d'annotations sémantiques. Nous discutons ensuite les enjeux et approches traditionnelles de l'analyse de documents (modèles vectoriels, analyse latente...). Puis, nous introduisons successivement les modèles de réseaux et la problématique des réseaux complexes. Ces modèles correspondent parfaitement à la représentation des relations dans les données réelles. Nous avançons aussi les approches classiques de l'analyse de réseau qui permettent d'appréhender les réseaux complexes.

Nous démontrons par la suite comment le modèle de réseau multiplexe s'applique très bien aux réseaux complexes, ainsi que les différentes stratégies pour l'aborder. Nous terminons ce chapitre en appliquant le modèle de réseau multiplexe à un réseau de documents liés par la sémantique. Malgré l'intérêt de ces sujets, la méthodologie utilisée pour former des groupes de documents, ainsi que celle qui autorise la formation de liens entre documents, ne seront pas abordées. En effet, d'une part un groupe de documents nous est *donné*

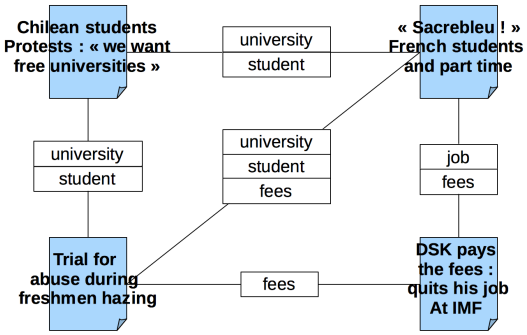
(manuellement, par requête, ou par regroupement), et d'autre part le lien sémantique qui associe deux documents se matérialise par des mots clefs communs. Nous formerons ainsi un réseau de *"substrats"*, nos documents, reliés entre eux par des *"catalyseurs"*, leurs mots clefs communs.



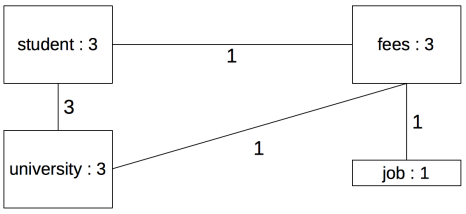
(a). Association de documents et de termes.



(b). Construction d'un réseau bipartis documents-termes.



(c). Réseau multiplexe de documents.



(d). Réseau d'interaction des termes

Figure 1: Les différentes étapes dans l'analyse de documents sous l'angle des réseaux multiplexes telles que nous les proposons.

Analyse de l'intrication

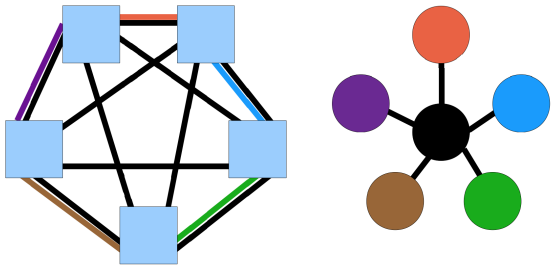
Ce chapitre présente notre première contribution, dans l'analyse des réseaux. Inspiré par les travaux de Ronald Burt (Burt and Schøtt, 1985), nous y amenons le concept d'*ambiguïté* dans les réseaux sociaux. Nous généralisons ce concept à la notion d'*intrication*, étendue à n'importe quel réseau multiplexe. A partir de solides fondations en algèbre linéaire, nous avons détaillé la méthodologie qui permet de déterminer l'*indice d'intrication* d'un catalyseur dans un réseau multiplexe. Cet indice nous permet de dériver deux autres mesures, l'*intensité d'intrication* et l'*homogénéité d'intrication* au niveau du groupe de substrat formant le réseau multiplexe. La détermination de cette mesure amène dans la foulée un autre objet qui se révélera central dans notre analyse, le *réseau d'interaction des catalyseurs* (Figure 1). Par la suite, nous indiquons les précautions à prendre lors de l'analyse de l'intrication. Nous détaillons aussi les nuances que cette analyse est capable de prendre en compte (Figure 2), ainsi que les limitations de nos différentes mesures.

Nous ouvrons enfin de nombreuses perspectives, correspondant à des travaux presque matures. Celles-ci incluent une généralisation de l'analyse de l'intrication pour les réseaux multiplexes pondérés. Nous avons aussi étudié une mesure de l'homogénéité et de l'intensité d'intrication au niveau d'un substrat au sein de son voisinage. Au delà du modèle pondéré, nos perspectives s'avancent sur d'autres aspects propres au réseaux de terrain: le cas du réseau orienté, la question de l'aspect multi-échelle, et bien sûr la dynamique dans ces systèmes complexes.

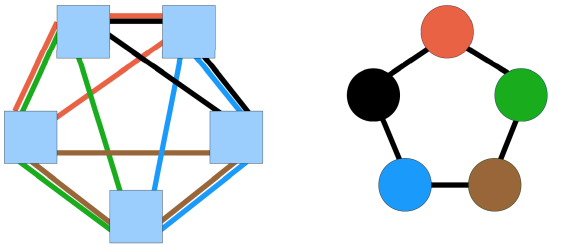
Un autre sujet pique notre curiosité, celui de l'analyse réciproque, où le réseau à l'origine du modèle multiplexe est un réseau bipartis. Chaque parti du réseau original peut en effet servir à la fois de substrats et de catalyseurs. Une évolution éventuelle vers les réseaux n-partis nous intéresseraient fortement. Tout particulièrement si l'on tient compte de l'ensemble des relations entre chaque parties du réseaux. Par exemple, ces relations peuvent être hiérarchiques.

D'autres perspectives nous intéressent, comme la mesure de la participation d'un substrat à l'intrication du groupe dans son ensemble, au-delà du concept d'intensité et d'homogénéité d'intrication "locale au substrat". Des résultats préliminaires se sont montrés intéressants. Ce type de mesure nous permettrait, par exemple, de proposer de nouvelles alternatives pour le partitionnement de réseaux multiplexes. Ceci nous pousse en conséquence, à chercher des améliorations et optimisations algorithmiques pour effectuer des calculs sur de grandes masses de données. D'autres pistes peuvent nous amener par exemple à considérer des classes d'équivalence de catalyseurs comme le suggérait déjà les premiers travaux de Ronald Burt.

En outre, l'algèbre linéaire pourrait nous offrir parmi ses nombreux outils, la possibilité de mieux comprendre la relation entre la topologie du réseau d'interaction des catalyseurs et les mesures



A multiplex graph, with a star-shape interaction pattern



A multiplex graph, with an interaction pattern shaped as a cycle

Figure 2: Notre analyse d'intrication permet de capturer ces différentes nuances. A gauche, les réseaux multiplexes de substrats, avec pour catalyseurs les arêtes de couleurs. A droite, les réseaux d'interaction des catalyseurs. Les poids que l'on pourrait associer aux deux réseaux sont identiques, 2 pour chaque arête externe et 1 sur chaque arête interne. Seule la topologie du réseau d'interaction des catalyseurs permet réellement de les différencier.

d'intrications. En ce sens, nous aimerions étudier le comportement de nos différentes mesures sur des réseaux aléatoires générés par différents modèles. Ainsi une large étude comparative entre nos mesures et celles proposées par la communauté scientifique sur les réseaux multiplexes serait sans aucun doute une nouvelle contribution.

Conceptions et outils pour l'analyse visuelle

Ce nouveau chapitre introduit le concept d'analyse visuelle. Nous démontrons combien sont idéales les représentations visuelles de réseaux afin de soutenir une analyse de haut niveau. Nous commençons dans ce chapitre par nous intéresser au processus de *compréhension*. C'est-à-dire, les mécanismes (en psychologie cognitive) qui permettent à l'homme de saisir une information complexe, et ce tout particulièrement au travers de supports visuels (Figure 3). Ceci nous permet de mettre en oeuvre une conception particulière d'un système d'analyse visuelle pour tirer au maximum profit de la cognition humaine et répondre aux multiples enjeux du processus de *compréhension*.

C'est tout particulièrement dans cet objectif que nous avons présenté le modèle de conception *imbriqué* de Tamara Munzner (Munzner, 2009). Ce modèle permet de découper la conception en quatre niveaux d'abstraction afin d'éviter tout piège dans l'articulation

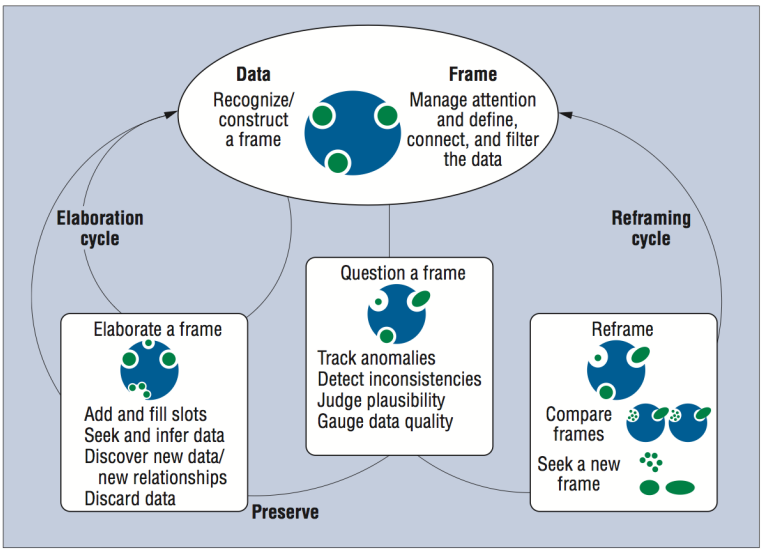


Figure 3: Le modèle macro-cognitif de la compréhension (Klein et al., 2006b).

de ces niveaux ce qui affecterait terriblement l'efficacité du système. Cela nous amène directement à questionner les différentes tâches qu'un système d'analyse visuel doit effectuer. Nous avons détaillé dans ce même but la typologie de Matthew Brehmer et de Tamara (Brehmer and Munzner, 2013). Cette typologie permet d'expliciter chaque manipulation au niveau du système, afin d'éviter un maximum de pièges dus aux subtilités de l'interaction homme-machine.

Nous avons présenté par la suite les mécanismes de perception humaine dont nous pouvons tirer avantage (Figure 4), et la grande variétés d'outils que la visualisation d'information met à notre disposition. Tout particulièrement, nous avons introduit les techniques de visualisation de données multivariées et de réseaux complexes. Nous terminons ce chapitre par la présentation des outils qui nous semblent adaptés pour notre analyse des réseaux multiplexes: la visualisation interactive de réseaux. Elle se présente sous la forme de vues multiples et liées, dont l'interaction se fera par la technique de *brush-*

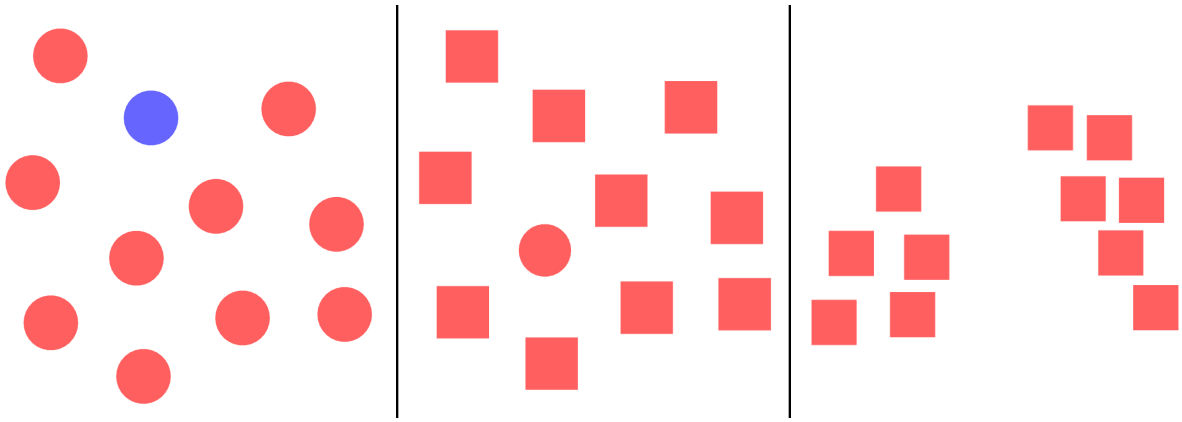


Figure 4: Trois exemples de regroupements pré-attentifs: *couleur, forme*, et *densité*.

ing and linking. Il s'agit d'une sélection manuelle que nous pouvons "promener" au travers d'une vue et dont la correspondance dans les autres vue s'affiche de manière instantanée.

Comprendre visuellement les réseaux multiplexes

Ce chapitre rentre directement dans le détail de nos choix et de l'implémentation de notre système d'analyse visuelle de l'intrication des réseaux multiplexes. Nous commençons avant tout par présenter les tâches au plus haut niveau, qui soutiennent la *compréhension* de la structure de l'interaction dans un réseau multiplexe. Nous avons ensuite, suivant le modèle imbriqué de Tamara, détaillé nos choix au niveau de l'abstraction et de la manipulation des données. C'est ici que nous mettons en oeuvre notre analyse d'intrication qui permet la *compréhension*. Après quoi nous sommes descendus au niveau de l'encodage visuel et interactions. Il s'agit de tous les mécanismes que nous avons mis oeuvre pour immerger l'utilisateur dans notre représentation du réseau multiplexe.

Nous avons tout particulièrement détaillé un algorithme de dessin de graphe bipartis qui permet d'*harmoniser* visuellement deux représentations séparées du même graphe (Figure 5). Cet algorithme associe le dessin d'une partie A au dessin d'une partie B. Pour cela il choisit les éléments les plus "pertinents" de B qu'il va fixer, autour desquels les éléments de A vont s'organiser à partir d'un algorithme de force. Les deux parties A et B pouvant ensuite être représentées séparément, auront une correspondance "cartographique" (les éléments de A et de B qui sont liés partageront ainsi la même zone dans leur dessins respectifs).

Nous validons la pertinence de nos choix d'implémentation en les incluant dans la typologie de tâches multi-niveaux présentée au chapitre précédent. Nous discutons ensuite de l'usage particulier que nous faisons de la correspondance visuelle. Elle permet d'associer mentalement nos différentes vues en *un seul tout* (Figure 6). Nous mettons aussi en avant notre technique d'*"articulation interactive"* (Figure 7). Celle-ci autorise simplement de passer d'une vue à l'autre sans perdre nos repères visuels. Elle permet de très facilement faire acheminer un raisonnement sur la donnée nécessitant de multiples allers et retours entre substrats et catalyseurs.

Nous discutons ensuite des perspectives ouvertes par nos techniques: de l'inclusion des avancées que nous acheminons depuis l'analyse, et de l'intégration dans notre système de nouvelles informations externes et multivariées. Nous concluons ce chapitre avec l'ouverture qu'offre notre système d'analyse visuelle. En séparant l'analyse d'intrication de la visualisation d'un réseau multiplexe notre système présente des avantages pour l'analyse de données multivariées de manière générale.

Dans la perspective de l'intégration de mesures au niveau des

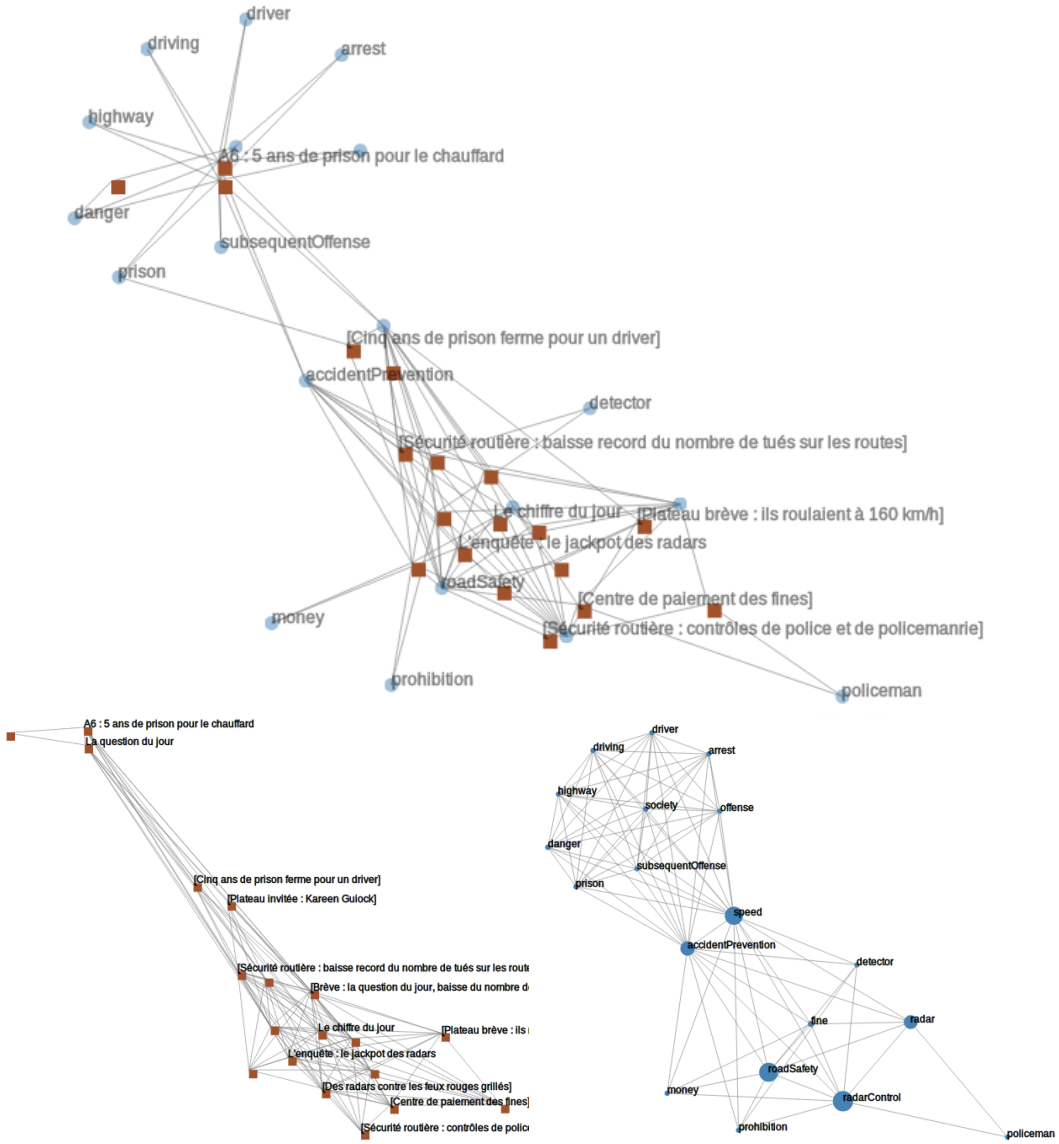


Figure 5: Illustration de notre algorithme de dessin, il permet à partir d'un graphe bipartis de placer les substrats en fonction de leurs catalyseurs spécifiques.

substrats, nous aimerions mettre en avant les zones du réseau de substrats ou nous observons la plus forte intrication. Un peu comme le font les cartes thermiques (*heat maps*). Naturellement, il est toujours possible d'associer de nouvelles mesures aux variables visuelles. Le défi est de le faire de manière à ce que la représentation soit à la fois pertinente et assez simple de sorte à n'induire aucune surcharge cognitive, qui rendrait l'utilisation de notre système visuel pénible voire confuse. Nous prévoyons aussi de travailler sur les représentations d'arêtes, un point que nous avons jusqu'alors écarté. Un encodage visuel sur les arêtes, ou bien le *"groupage d'arêtes"* (ou *edge bundling*) semblent être d'intéressantes pistes. Il y a là-aussi un équilibre à trouver entre l'information que l'on voudrait transmettre et son impact sur la lisibilité de la représentation. En ce sens, une adaptation aux réseaux multiplexes des *power graphs* (simplification d'un graphe

par motifs et méta-nœuds) pourrait bien être une voie à suivre.

Dans l'objectif d'extraire l'essentiel de notre approche visuelle de l'analyse de l'intrication, nous devons nous orienter vers l'intégration visuelle de l'aspect multivarié des données. Pour cela, il est important de se tourner vers des techniques de représentations traditionnelles (comme les graphiques) qui ont prouvé leur efficacité, et d'y intégrer notre articulation interactive. Ainsi nous devrons faire face aux problématiques de la dimensionnalité, et éventuellement de la représentation dynamique des données.

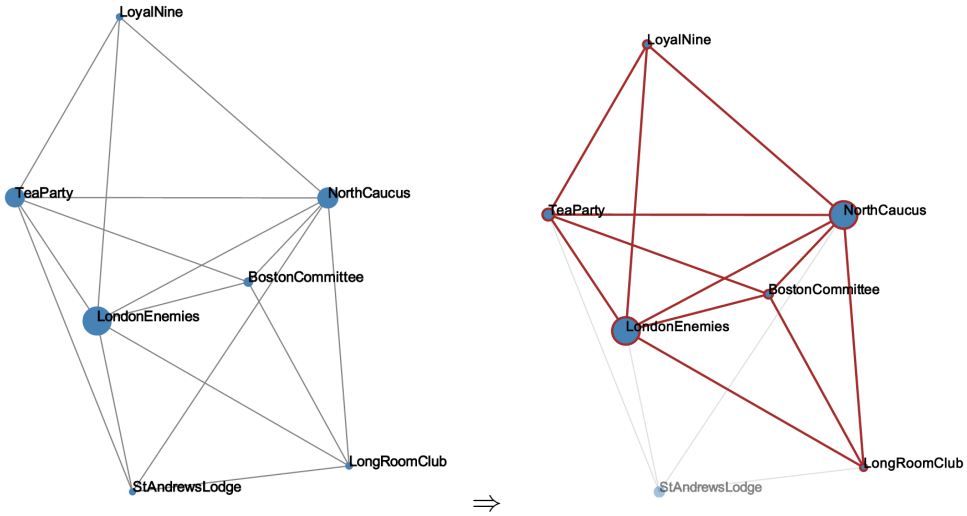
L'algorithme de dessins *harmonisés* de réseau biparti offre aussi une pleine liberté dans l'étude de ses paramètres. Une optimisation algorithmique pour étudier l'influence de ces paramètres en temps réel serait un avantage certain. L'algorithme s'appuie pour le moment sur un critère arbitraire, séparant les catalyseurs en deux groupes. Le choix de ce critère est resté libre et son implémentation est des plus empirique. Une étude sur l'influence de ce critère et un travail sur une possible continuité entre ces deux groupes (plutôt qu'une séparation binaire) permettrait de généraliser notre algorithme.

En dernier lieu, l'évaluation de notre travail est un point que nous aimerions compléter. Le chapitre suivant présente en effet une étude utilisateur, mais réalisée à petite échelle, avec des experts des documents INA. Pour réellement valider l'approche de nos choix de conception et l'efficacité de notre implémentation, une étude utilisateur à grande échelle est nécessaire. Ce serait de surcroît un apport innovateur à la visualisation des graphes multiplexes.

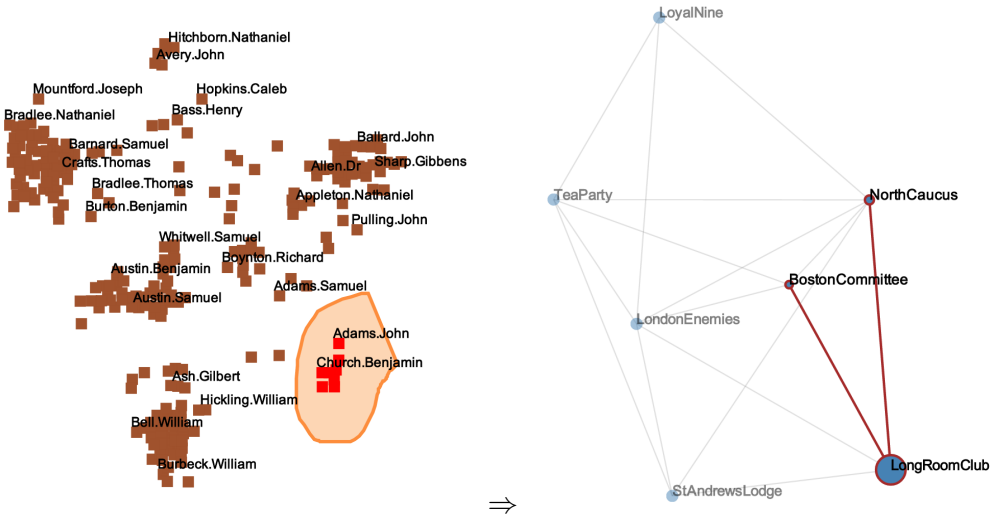
Contextes d'application et exemples

Ce dernier chapitre rend compte de l'utilité de notre approche en la confrontant à des cas d'applications et données réelles. Tout naturellement, nous sommes revenus sur les données qui ont motivé notre travail: les documents. Nous présentons ainsi des exemples détaillés d'analyse de réseaux de documents illustrant nos méthodes. Nous avons aussi pris le temps de prendre du recul sur la nature des groupes que nous avons observés sous l'éclairage de l'analyse de l'intrication. Nous avons présenté les résultats d'une évaluation utilisateur de petite envergure avec des experts de l'INA. Bien que non qualifiante pour l'évaluation de notre système, celle-ci nous a permis de confirmer la pertinence de notre approche et de tirer les leçons nécessaires à une future évaluation de plus grande envergure. Dans le cas de l'application aux documents, nous avons observé que la forme du réseau d'interaction des catalyseurs semble avoir un pouvoir explicatif (Figure 8). Il y a une relation certaine entre cette forme et les différents sujets qu'aborde un groupe de documents. Nous voudrions exploiter ce réseau afin de bien identifier ces différents sujets.

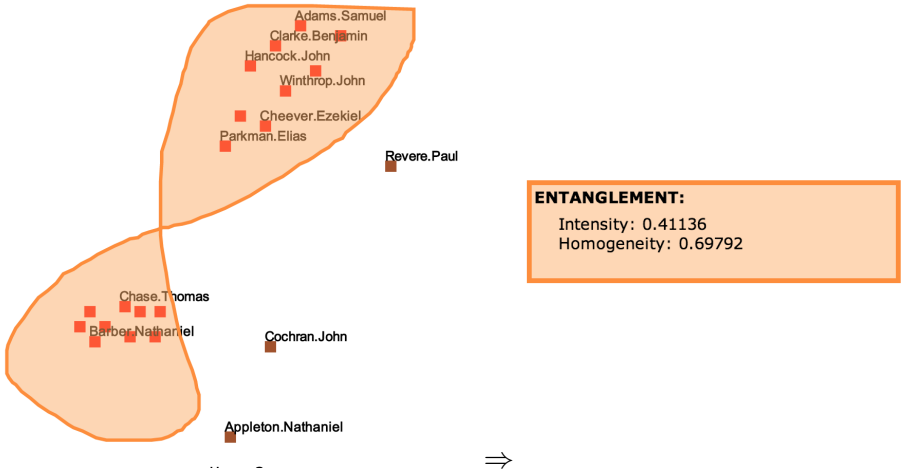
Toutes ces discussions nous permettent d'envisager des amélio-



Préservation de la carte mentale au travers d'une sélection d'un même type de données (ici les catalyseurs)



Préservation de la carte mentale entre deux types de données différents



Préservation de la correspondance des couleurs entre une sélection et les mesures de groupe

Figure 6: Nos différentes techniques permettent d'associer différents éléments de la visualisation de manière pré-attentive, dans l'articulation de différentes vues.

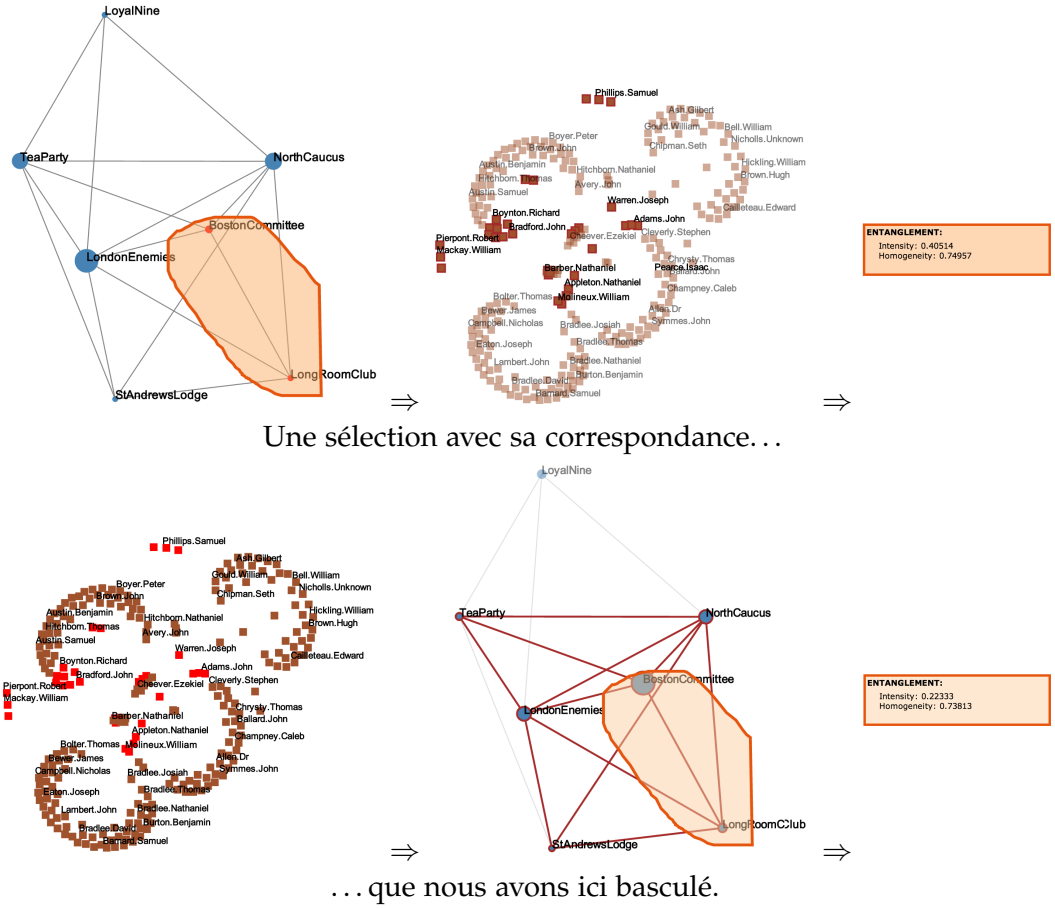


Figure 7: Bascule d'une sélection (en haut), dont la destination devient une nouvelle source (en bas). On remarque la mise à jour de l'encodage visuel.

rations de notre système dans le cadre de l'application aux groupes de documents. Entre autres, la dimension du temps qui offrira à notre système son plein avantage dans l'étude des évènements médiatiques. Nous sommes ensuite passé à l'application dans le domaine des réseaux sociaux, et nous avons détaillé deux exemples sur des jeux de données "classiques" à la communauté d'analyse de réseaux sociaux et de visualisation: celui de l'interaction entre acteurs et réalisateurs du corpus IMDB, et celui du réseau de coauteurs proposé par le Challenge InfoVis 2004. Nous nous sommes aussi ouverts à bien d'autres réseaux, qui motivent de potentielles perspectives. Grâce à eux, nous avons commencé la généralisation de notre méthodologie au cas pondéré. La comparaison de notre méthodologie sur des réseaux sociaux, dits *d'affiliation*, dont l'analyse a été nourrie par la littérature, fait partie de nos perspectives.

Enfin, illustrant l'aspect générique de notre approche, nous avons proposé un champs plus large d'applications. Celles-ci incluent les réseaux financiers de la Banque Mondiale, ou encore les interactions protéines/ontologies dans le cadre de la bioinformatique. Chaque nouveau domaine où nous appliquons notre méthodologie est une source de richesse, pour laquelle les perspectives sont nombreuses.

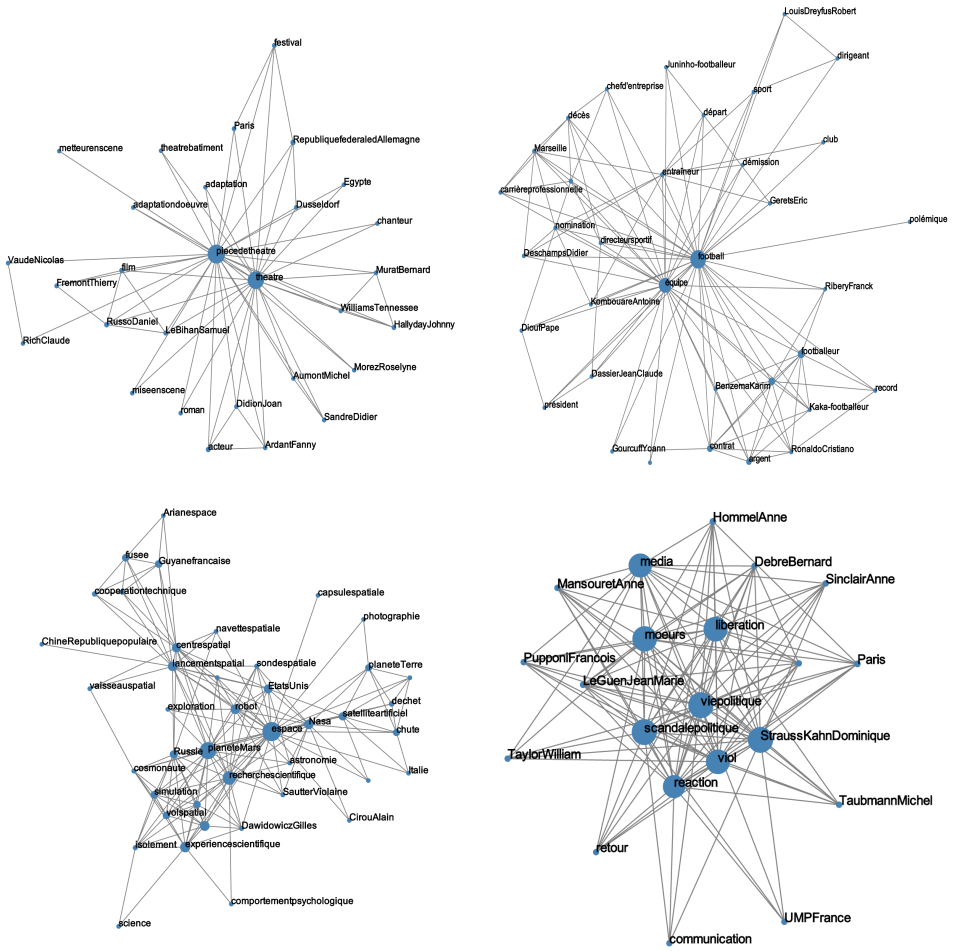


Figure 8: Différents types de réseaux d'interaction de termes dans les groupes de documents de l'INA. Les réseaux qui sont les plus thématiques semblent former une topologie "en étoile" (en haut à gauche, le théâtre) alors que les plus évènementiels sont beaucoup plus denses (en bas à droite, l'affaire DSK). Entre deux, nous avons des évènements dans une thématique (en haut à droite, football) ou bien l'association d'évènements autour d'une thématique similaire (en bas à gauche, recherche astronomique/lancement de fusée).

Elles nous poussent à dépasser les limites de notre méthodologie. C'est pour cela que nos ouvertures du côté applicatif se tournent vers tous les domaines possibles.

L'expertise de spécialistes d'un domaine permet de nous orienter dans la justesse de notre approche. L'hétérogénéité des domaines applicatifs rend possible un recul qui motive nos avancées tant du côté analytique que visuel. Si l'on prend l'exemple de la Banque Mondiale, les objectifs de compréhension ne passent pas seulement par l'*exploration*, mais aussi par l'*explication*. Nous devrons nous intéresser par exemple aux stratégie statistiques qui s'y associent, comme par exemple l'*autocorrellation* dans un réseau multiplexe. Si l'on s'intéresse aux mécanismes que la bioinformatique aspire à comprendre, la complexité se retrouve à un niveau encore supérieur. Les produits d'une interaction eux-mêmes intègrent un système encore plus complexe. Nous y observons des réseaux multiplexes au travers de réseaux multiplexes, tous hétérogènes.

Conclusion

Ce manuscrit présente ainsi une nouvelle méthodologie pour l'analyse des réseaux multiplexes. Nous l'amenons tout du long, depuis ses fondations en analyse, au travers de son intégration en visualisation, jusqu'à son application concrète. Notre approche, l'intrication dans les réseaux multiplexes, a été motivée par la problématique très concrète de l'INA dans l'analyse de ses groupes de documents. Elle nous a menés non seulement à exploiter le réseau multiplexe, un objet qui a jusque là peu intéressé la communauté d'analyse des réseaux, mais encore a en offrir une analyse originale qui semble se distinguer des approches classiques de réseaux complexes. Un des intérêts de cette approche est de fournir en une seule série de calculs un indice local associé à un nouveau réseau et deux mesures de groupe.

En utilisant ces objets issus de notre analyse, nous avons proposé un système d'analyse visuel complet pour l'étude d'un réseau multiplexe, à son tour peu exploité par la communauté en visualisation de réseau. La conception de ce système de visualisation a été pensée de manière à éviter les multiples pièges inhérents à la problématique de la visualisation analytique. Ainsi notre approche met en avant, et valide, trois techniques particulières ayant trait à la coordination de vues multiples d'un même réseau multiplexe. Un algorithme relie deux vues hétérogènes dans leurs données en fonction de la position des éléments dans leur dessin. La préservation mentale des références graphiques à tous points de vues, permet d'associer très facilement l'information au plus haut niveau d'abstraction avec celle au plus bas niveau. Cette association permet d'élaborer sans effort des comparaisons qui auraient été compliquées autrement. Enfin, une interaction particulière, autorise un va-et-vient entre les différents niveaux d'abstraction, qui mime le raisonnement par étape de l'utilisateur sur de telles données.

Le travail que nous avons réalisé durant cette thèse a été clairement motivé par notre curiosité et notre soif de compréhension. L'élégance de la représentation par le réseau est fascinante. C'est un objet abstrait dont l'analyse se base sur de solides fondations théoriques, et qui pourtant se comprend aisément une fois dessiné. L'exemple de la "carte mentale" (*mindmap*) en est une preuve frappante. Elle est utilisée depuis le milieu de l'éducation jusqu'au monde des affaires, sans que personne ne la remette en question. Les réseaux multiplexes sont bien sûr des objets autrement complexes. Notre défi était de donner du sens à la représentation de ce qui semble être chaotique de prime abord afin d'en extraire un réelle compréhension. Relever ce défi a été l'un des aspects les plus intéressants de cette thèse. Un autre aspect est l'incroyable potentiel d'application de ces réseaux multiplexes. C'est justement là une chance de faire de la science appliquée, de pouvoir s'adresser à un public aussi hétérogène que varié. C'est une sorte de voyage où à chaque application l'on peut confronter ses idées et "changer la forme de sa pensée" si l'on veut citer Alphonse de

Lamartine. Ces remises en question demandent une adaptation particulière, notamment aux langages et aux concepts qui peuvent nous être étrangers. Une bonne communication est essentielle pour atteindre un but commun. Mais l'effort que cela comporte est toujours récompensé par l'enrichissement de nos modèles et de nos méthodes, et le travail de cette thèse en amène une démonstration.

ACKNOWLEDGMENTS

First of all, I would like to thank Marie-Luce Viaud and Guy Melançon for their trust in hiring me from across the planet. Not only they have been intellectually, and technically, supporting, and challenging me during the thesis, but also they are rare models of combined sharp spirit, kindness, and open-mindedness, making the time of this thesis unforgettable.

I also want to acknowledge INA, the ANRT, INRIA Bordeaux Sud-Ouest, and the European FP7 program (project Emergence by Design – MD), the University of Bordeaux and LaBRI for financing my thesis and hosting me in the best conditions possible. Special thanks to the whole Tulip software[1] team, and to both INA and LaBRI engineering and administration members without whom most of the work could not have been done

I would then thank all the members of my jury. Fabien Gandon, from INRIA Sophia-Antipolis at Valbone (France), and Georges Grinstein, from the University of Massachussets Lowell at Lowell (Ma, USA), for reviewing my manuscript and thoughtful remarks that have enhanced the quality of this body of work. I would also thank Alberto Cottica from the Council of Europe, and now EdgeRyders at Brussels (Belgium), Fleur Mougin from the ISPED and the University of Bordeaux (France), and Frédérique Segond from Viseo at Paris (France), for accepting to be jury of my defense.

I want to thank all my collaborators at INA, especially Nicolas Hervé, Laurent Joyeux, Pierre Letessier, Agnès Saulnier, Jérôme Thièvre, and Olivier Buisson, but also all the research department at INA and the INA DLWeb team who have been a great help and support during this thesis. I want to thank also my collaborators at LaBRI, all the new friends who arrived during this thesis, and all the other, particularly Daniel Archambault, David Auber, Romain Bourqui, Maylis Delest, Maurin Nadal, Bruno Pinaud, François Queyroi, Sébastien Rufiange, Arnaud Sallabarry, Patrick Mary and Alice Rivière, for the many scientific and technical advises without which I could not have achieved this thesis.

I wish to thank all the people I have been sharing passionate work. These collaborations led to very productive open minding insights and applications among other work (here in alphabetical order): James Abello from the DIMACS center at Rutgers University (New Brunswick, NJ, USA); Milica R. Begovic and Giulio Quaggiotto from the UNDP at Podgorica (Montenegro) and Bratislava (Slovak

1. http://tulip.labri.fr/

Republic); Nozha Boujemaa from INRIA Saclay at Palaiseau (France); Kenneth Chomitz, Alex McKenzie, and Prasanna Lal Das from the WorldBank at Washington DC (DC, USA); Alberto Cottica from the Council of Europe, and now EdgeRyders at Brussels (Belgium); Michael McGuffin from the ETS at Montreal (Qc, Canada); Tamara Munzner from University of British Columbia at Vancouver (BC, Canada); and Christophe Viau from Datapad at San Francisco (Ca, USA). I also want to thank Helen Purchase from the University of Glasgow (UK) and Patricia Thebault from the University of Bordeaux (France), for sharing with me some of their light while I was writing this thesis. I would also thank all the partners of the ANR-10-CORD-000 OT-Media and FP7 MD projects, too numerous to be exhaustively listed here. Many thanks to Fabien Colombo and Quentin Gardrat both from the University of Toulouse Le Mirail (France) for winning our many debates[2], in which we always bring back together Philosophy and Science. I would like to thank Alexander Pak from Google Zurich for his priceless advises and help in spreading my work. I also want to thank Rolf Reinhardt, from Pearson et Munich (Germany) and the Learning Agency Network at Brussels (Belgium), for his support and kind invitation to attend TEDx Brussels.

I want to especially thank Mike and Elizabeth Stamp, Guillaume Le Louëdec, and Barbara Crisp for their amazing job in revising this thesis for the enjoyment of our dear readers.

I would like to thank all my friends, my family, 나의 한국에 있는 가족에게, my parents, my sister, and my girlfriend, who have never stopped believing in me, and supported me even during the toughest times.

Last, but not the least, I will thank the many professors and their teams who have guided my path and supported my choices and applications, leading to Research, and now the Ph.D. degree. It includes, by chronological order, David Laiymani from the University of Franche-Comté at Belfort (France), David Whitley from the CMD at the University of Portsmouth (UK), Jean-Christophe Bailli from Gostai now Aldebaran Robotics at Paris (France), Yassine Ruichek from the UTBM at Belfort (France), and Jong C. Park from the NLP⋆CL lab at KAIST in Daejeon (South Korea).

I dedicate this work to all the great thinkers, who I have and will meet and read, who cultivate in me this sense of *curiosity*.

2. http://www.ted.com/talks/daniel_h_cohen_for_argument_s_sake.html

PREFACE

My thesis took place between the National Institute for Audiovisual (INA[3], Paris), financing my thesis under the CIFRE scholarship, and the Bordeaux Laboratory of Informatics (LaBRI[4], Bordeaux), with INRIA[5] Bordeaux Sud-Ouest also financing my work. INA aims to archive every audiovisual broadcast since the creation of radio and TV in France. My mission there was to analyse and visualize news excerpts from the point of view of graphs and integration in the OTMedia project[6] (which aims to analyse and visualize news from very diverse sources).

3. www.ina.fr
4. www.labri.fr
5. www.inria.fr

6. www.otmedia.fr
ANR-10-CORD-000

INA needs to exploit and understand documents to deliver expert consultancy and document packages. A network's perspective can support INA's experts, and it was already made a weapon of choice for INA's document exploration strategy. We focused in this thesis on the semantics that tie these documents.

The document groups we were to observe were collected at the end of a long process. The documents are rich structures, with meta information, annotated, manually or automatically, heterogenous, and sometimes incomplete. The source of the information could be also heterogeneous and varied, and we gathered documents in many types of groups, depending on the usage they were destined to. Tackling such information, heterogeneous and of varying quality, is itself a challenge. We were aiming at finding structures in a whole, that appeared chaotic at a first glance. Thus we focused on understanding the semantics that binds documents in a group.

We have been working with most of the traditional tools for documents analysis from text, to first tackle the networks of documents, and drew them with different algorithms. The networks looked dense and it was difficult to understand anything out of local densities of document similarities. We explored different outcomes through tuning the different stages of our algorithm or filtering the sets and the links. At this stage, it appeared that we needed a different way to see things.

Everything seemed tightly linked to the quality of the semantic extraction, a task that was being undertaken by some of our colleagues. From all those multiple attempts to make sense of these networks, we started to rely more and more on solely semantic annotations. We figured out it would be interesting to tackle documents from their perspective. It is only then we have looked at the documents on a complex network perspective, with overlapping layers of semantics:

our little *"eureka"* moment. From this perspective, we could not only address our issues with document groups, but finally open ourselves to a very wide range of applications.

Another very interesting challenge I faced throughout this thesis was the collaboration with people from various fields, different to information technologies. I have been working with journalists and documentalists at INA, but also many other people during the last part of my thesis. The FP7[7] project Emergence by Design[8] studies *innovation*, and stability in an *"innovation society"*. This rather broad subject interests many disciplines and gave me the opportunity to closely collaborate with social scientists, economists, and policy makers. It taught me many important principles for a successful collaboration. Indeed, when we are very specialized in a particular field, it takes careful listening, and pedagogic explanations to understand properly each other's vision on a same issue. This is especially true in analysis and visualization. It is critical to understand *the right questions* we are addressing, not only the general purpose, but also the objects we are all manipulating.

7. http://cordis.europa.eu/fp7

8. http://emergencebydesign.org/

This situation is similar to international communication: we may all talk about the same objects, but we have different point of views such as definitions overlapping, do not match exactly. It is important for data scientists to rationalize the boundaries of these definitions in order to enable data processing (and answer the *right* question). It is also equally important to understand and communicate the boundaries of the processing outcomes. Indeed, many tools we use involve approximations and distortions, and we need to communicate well these effects to avoid over-interpretation – of which the most well known is certainly the difference between correlation and causation. This is especially true with visual analytics where the visual support is a baseline for higher purpose analysis: visualizations *do* lie and play tricks on our minds. Fortunately open-mindedness, good design, and good communication are solutions to avoid any problem.

This leads to a fundamental question for our analytical work: *What are we showing and how is it understood?*. This question is even more relevant with the intricate nature of our objects of analysis (complex networks). Understanding the structure of information led us to use metrics, associated with interactive visualization to display and explain the inner structure of the objects we are observing. Comprehension was the key to enable higher purpose analysis, and the network model seemed a perfect object to support it.

The work we achieved lifted us from document group cohesion analysis to lead into the more general *multiplex network* analysis. We combined analytical methods with interactive visual interfaces to bring understanding of complex structures. This document presents the insights and results from our research together with state-of-the-art approaches, and aims to bring you new tools in hand to tackle the fascinating world of complexity.

CONTENTS

Appendices

INTRODUCTION

1

> *"As long as the centuries continue to unfold, the number of books will grow continually, and one can predict that a time will come when it will be almost as difficult to learn anything from books as from the direct study of the whole universe. It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes."*
> Denis Diderot, *"Encyclopédie"* in (Diderot et al., 1751-72) *Vol. 5 (1755), pp. 635–648A*

Science started thousands of years ago, describing the world's phenomena with natural phenomena, and understanding the techniques needed to explain and reproduce them: the first tools in the development of Science were *empirical* (Gray and Szalay, 2007). Centuries later, we started to describe these natural phenomena building models with laws. We postulated theories we could refute or confirm: the development of Science accelerated adding *theoretical* capabilities to its toolbox (Gray and Szalay, 2007). In the last decades, models of the same natural phenomena became more and more complex, and the combinatorics of theories expanded greatly so we needed computational power to simulate the complex phenomena: the development of Science accelerated even faster becoming *computational* (Gray and Szalay, 2007). Since the beginning of the *XXIst* century, computational capabilities have extended the granularity and precision of simulations, with the production of abstract information, and sensed information that is expanding exponentially to the point at which the information is growing faster than our capability to analyse it[1]. Science is now adding to its set of tools Gray's fourth paradigm of Science: *data exploration*, an exploration that aims to unify experiments, theories, and simulations (Gray and Szalay, 2007; Hey et al., 2009).

1. The production of information now exceeds Moore's law: http://www.emc.com/about/news/press/2011/20110628-01.htm

A challenge in this new Science paradigm is the management of complex, diverse and numerous data. This is the so-called *Big Data* challenge. Behind this buzzword hides a need for constant rationalization of phenomenon through descriptive, explanatory and predictive data, and the need to efficiently process and analyse this data. Capabilities to generate this information have never been so efficient, yet never so obvious is our inability to handle numerous, diverse and complex information: this information is made of a large quantity of data, often coming from heterogeneous sources with complex data structures that are also heterogeneous and multidimensional. With

our growing need to analyse and understand what hides behind the data, the relationships between the phenomena we observe get more and more complex, and the relationships we intend to find in the information we collect does not get any simpler. Fortunately, the tools we are developing are increasingly efficient.

In 2011, Gartner – a well-known IT consulting company – made popular the first *three Vs* of Big Data[2]:

- *Volume*: with the exponential growth of data production, volume becomes an issue for storage and analysis,
- *Velocity*: lots of data is now streamed and velocity involves the ability to generate and broadcast data as well as real-time analysis, and
- *Variety*: data is stored and sourced from many different locations, and the variety of information implies complex and heterogeneous data structures.

Later enriched by IBM with another *V*:

- *Veracity*[3], multiple reasons make the quality of data very uncertain and one challenge is to assert trust in this information and ensure a right interpretation.

These are the main challenges we are facing dealing with Big Data[4]. Although the work in this thesis has not been directly driven by the Big Data challenges of scalability (*Volume* or *Velocity*), it still falls within concerns of *Variety* and *Veracity*. This thesis aims to examine *Variety* and *Veracity* through novel measures embedding them in a Visual Analytics framework. The challenges that motivated this thesis, exploration and utilization of document collections, forced us to deal with *Variety* of information: heterogeneous multi-dimensional data involving semantics. Tackling the uncertainty and ambiguity involved in semantics pushed us to ensure *comprehension* and assert *Veracity*.

For ages human beings have been supporting their cognitive reasoning with visual representations. One of the earliest example is a bone from an eagle's wing, with notches carved in it by men during paleolithic times, representing the moon phases, or some sort of marking of time[5] (Marshack, 1991) – discussing the intertwined role of visual representations in abstract thinking (eventually catalyzing the rise of human beings and civilizations) is the subject of other books (Gattis, 2003). Indeed humans are hard-wired to interpret the world from visual information[6]. Human perception allows very fast processing (Healey and Enns, 2012). Preattentive processing is able to identify in parallel multiple features of a scene (Treisman, 1985), and it is done by the brain *before* any attention is focused on a scene, *before* any conscious cognitive process – and our favourite magicians know very well how to trick our preattentive processing (Olson et al.,

2. Big Data's 3 Vs were popularized in http://www.gartner.com/newsroom/id/1731916, but the concepts were first introduced by Laney (2001)

3. See in TechAmerica Foundation's report for IBM, *"Demystifying Big Data"*: http://public.dhe.ibm.com/common/ssi/ecm/en/iml14336usen/IML14336USEN.PDF

4. More recently the phylogenetic tree of Big Data's Vs were completed with exotic species from all around the globe: Value, Validity, Visibility, *etc*.Arguing around those is beyond the scope of this manuscript.

5. Here may be an early "barchart" counting 28 marks that can be admired at the Aquitaine Museum in Bordeaux.

6. Recent studies suggests about 20% of the brain is involved in vision of which *"semantic representation is analogous to retinotopic representation"* (Huth et al., 2012).

2012). It becomes a very powerful tool when combined with graphical semantics, as depicted in Jacques Bertin's[7] most well known legacy: Bertin (1967, p. 2) Sémiologie Graphique:

"Graphic representation constitutes one of the basic sign-systems conceived by the human mind for the purposes of storing, understanding, and communicating essential information".

It is especially with the goal of designing Information Visualization optimized for human perception that Collin Ware wrote his excellent book (Ware, 2000). With the knowledge of the Information Visualization tools, we designed techniques to make sense of complex information modelled with complex networks. Visual Analytics supports analytical reasoning with interactive visualization (Thomas and Cook, 2005), a difference that turns passive viewers into active users.

Mind maps have been used to represent the structure of our knowledge and ideas for centuries[8]. The visual representation of a mind map is analogous to that of a network. Therefore network representations seem then ideal objects to visually support comprehension of a complex information. The power of this representation is not limited to visualization, and networks empower analysis with the tools of graph theory (West et al., 2001). Although the theory has been formalized by Euler (1735), network analysis' popularity has only been rising since the 1950's, and already have found itself successful in a tremendous number of applications (Wolfe, 2010), and ultimately drives the *Internet* – International Network – and the recent evolutions in *social media*[9].

This body of work can be placed at the intersection of analytics, network science, information visualization, and interaction[10]. We contributed to the field by proposing the study of a new concept of complex networks, the *entanglement* of multiplex networks, with new measures, and their use for visual analytics with innovative interaction and representations. Inspired by past work on social science by Ronald Burt (Burt and Schøtt, 1985), we questioned the structure of complex networks of news documents through the entanglement of multiple edges. Deriving from an *entanglement index*, we proposed to measure also the *entanglement homogeneity* and *intensity* of a multiplex network. We developed frameworks to visualize a complex network and the structure of its inner entanglement. It is interactive and query-able through brushing and linking across multiple views, supporting comprehension and sense-making of a complex network.

Chapter 2 will describe the context of our research, the data we had to process and how it is modelled as a *multiplex network*. In Chapter 3, we introduce the concept of *entanglement* and how we can measure it. Chapter 4 presents the *embedding* of entanglement measures in a Visual Analytics framework. Next, in Chapter 6, we propose applications of this work, and conclude the thesis in Chapter 7, opening

7. Jacques Bertin (1918-2010) was a French cartographer and passed away during the time I made my thesis, he left us an immense legacy, founding the graphical semiology, which has applications not limited to cartography. He can be considered as one of the founding father of modern Information Visualization and his work will still continue to guide our graphical designs for years.

8. The first historical use of mind maps has been alleged to Porphyry of Tyros (233-310 AD), in his *Isagoge - Introduction to Categories* in which he discusses the work of Aristotle. Here is a VIIth Century copy of a mind map in the manuscript, by Syrian Athanase of Balad (686 AD) (image with the courtesy of Bibliothèque Nationale de France).

9. see for example Google's search or LinkedIn's recommendation algorithms

10. *"Every science overlaps with others: they are two continuous branches off a single trunk."*, Denis Diderot, *"Encyclopédie"* in (Diderot et al., 1751-72)

new perspectives of research on tangled complex networks.

Part of our work was submitted and published, and the content of these publications formed a basis for our different Chapters:

- In Chapter 2, for the general setting of our methodology: (Viaud et al., 2010; Renoust et al., 2011a);
- in Chapter 3, as basis for the analysis of entanglement in complex networks: (Renoust et al., 2011b, 2013f,d,c,e);
- in Chapter 5, inspiring the design of a visual analytics framework: (Renoust et al., 2011a, 2013f,c);
- and in Chapter 6, for the applicative aspects of our work: (Viaud et al., 2010; Renoust et al., 2011a,b, 2013f,d; Melançon et al., 2012; Renoust et al., 2013e,d,c,a; Renoust and Begovic, 2013; Renoust et al., 2013b).

DATA AND MODELS

2

This chapter presents the ground knowledge for both document analysis and multiplex homophily networks. After introducing some necessary definitions, towards understanding of the world of networks, we present the data that motivates our research, INA's document corpora. Then we cover classical approaches to document analysis in the research community. We surveyed the literature in usage of complex network models using classical approaches to handle them, showing that document corpora are good candidates for such models. However, traditional network approaches are not sufficient for a complete document analysis, and thus we upgraded our approach to multiplex complex networks. Finally we introduce our modelling choice that marks the baseline of all our work.

2.1 Foreword on complex networks and documents

We approach document analysis from the complex network point of view, and propose, in Chapters 3 and 4, measures and frameworks that can apply to the most complex networks. Before doing any analysis, we introduce complex networks, and the characteristics of our document corpora that justify this analogy.

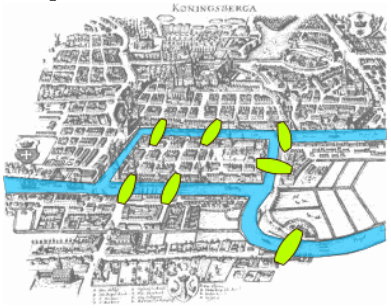
2.1.1 A few network definitions

A ***network*** is the representation of a mathematical model called a ***graph***[1] which is defined by graph theory as a set of objects (***nodes*** often called ***vertices***) from which some pairs are connected by ***links*** (also called ***edges***).

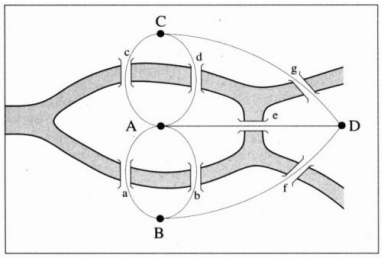
A ***graph*** is defined as $G = (V, E, f)$, V being the set of nodes, E the set of edges, and f a *surjective* application $f : E \mapsto V \times V$ that associates two nodes. $V(G)$ and $E(G)$ designate the nodes and respectively the edges of the graph G. The ***order*** of a graph represents the number of nodes $|V|$, and the ***size*** of a graph its number of edges $|E|$. For the sake of simplicity, we will not mention the application f unless necessary, and assume edges as pairs of nodes.

By definition, an ***edge*** $e = (u, v)$ connects the nodes u and v. Edges can be directed ($(u, v) \neq (v, u)$), weighted ($(u, v) = w_{(u,v)}$), and/or *loops* ($e = (u, u)$). An edge $e(u, v)$ designates an edge e between nodes u and v.

1. Graph theory was formally introduced by Leonhard Euler in 1735, The seven bridges of Kőnigsberg (Euler, 1735). The problem was to find a route across the city of Kőnigsberg, that would cross *each* bridge once and *only once*. With the foundations of graph theory, Euler has proven this problem to have *no* solution.



The Kőnigsberg bridges, (from Wikimedia commons.)



the network pattern of connections across the bridges... (from B.W. Carroll)

...somehow looks like a pretzel ! (from M. Taylor)

A graph $G = (V, E, f)$ is ***not directed*** when f is *symmetric*, $(u, v) = (v, u)\ ,\ \forall v,\ u \in V$.

A graph $G = (V, E, f)$ is ***simple*** when f is *bijective*, a simple graph admits no loop nor multiple edges between a same pair of nodes. A *multiple graph* is a graph that is not simple and presents multiple edges between a pair of nodes.

A graph $G = (V, E, f)$ is ***weighted*** when f is a function that associates a weight to every edge.

A graph $G = (V, E)$ can be represented by an ***adjacency matrix***[2]. The adjacency matrix M of a common graph is a square matrix of dimension $|V| \times |V|$ for which each line/column represents a node and one entry of the matrix $m_{u,v} \neq 0$ represents an *existing* edge between nodes u and v. The diagonal entries are all equal to 0 when G presents no loop. $m_{u,v} = \alpha$, $\alpha \in \mathbb{R}$ when G is a weighted graph, M is a boolean matrix otherwise ($\alpha \in \{0, 1\}$ as in Figure 2.1). M is *not* symmetric if G is a directed graph.

2. The adjacency matrix and the Laplacian – a popular normalization of the adjacency matrix– are very powerful tools especially for spectral analysis of graphs, out of the scope of this document but worth mentioning (Gkantsidis et al., 2003; McGraw and Menzinger, 2008).

A ***subgraph*** $G' = (V', E')$ of graph $G = (V, E)$ is a graph such as $V' \subseteq V$ and $E' \subseteq E$.

An ***induced subgraph*** $G' = (V', E')$ is a subgraph of $G = (V, E)$ induced by a set of nodes such as $V' \subseteq V$ and $E' = \{(u, v) \in E, u \in V', v \in V'\}$.

A ***complete graph*** $G = (V, E)$ is a graph for which every node is connected to all other such as $|E| = \frac{1}{2}|N|(|N| - 1)$ in an undirected graph or $|E| = |N|^2$ in a directed graph.

$$M = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

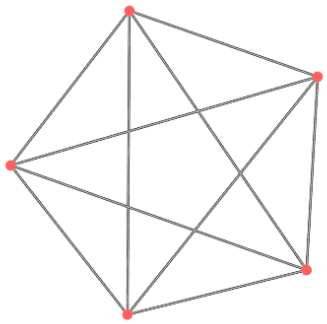


Figure 2.1: A *5-clique* adjacency matrix and its corresponding node-link diagram. We can count 5 nodes and $\frac{5\times4}{2} = 10$ edges

A ***k-clique*** V' is a set of nodes in $G = (V, E)$ such as the induced subgraph $G' = (V', E')$ forms a complete graph with order k (such as $|E'| = \frac{k(k-1)}{2}$ in an undirected graph, Figure 2.1)

A ***path*** in $G = (V, E)$ is a sequence of nodes and edges $(v_1, e_1, v_2, e_2, \ldots, v_i, e_i, \ldots, e_{n-1}, v_n)$ such as $e_i = (v_i, v_{i+1})$, $e \in E$, $\forall i \in [1..n-1]$ and $v_i \in V$, $\forall i \in [1..n]$ with $\forall i$, $v_i \neq v_{i+1}$ and $e_i \neq e_{i+1}$. A path is *simple* if it does not go through any node more than once.

The ***length*** of a path is measured by the number of edges in its sequence, or the sum of its edge weights in a weighted graph.

The ***distance*** $\delta_G(u, v)$ between nodes u and v in a graph G is the

length of the shortest path between u and v in G. When there is no path between u and v (they are not connected) the distance $\delta_G(u,v)$ is *infinite* or is undefined.

The ***eccentricity*** $\epsilon(u)$ of a node u in a graph $G = (V,E)$ is the maximum distance of a node to any other node in the graph, $\epsilon(u) = max_v \delta_G(v,u)$, $v \in V$, $v \neq u$.

The ***radius*** of a graph $G = (V,E)$ is the minimum eccentricity among its nodes $r_G = min_u \epsilon(u)$, $u \in V$.

The ***diameter*** of a graph $G = (V,E)$ equals the maximum distance between any pair of nodes in G (the lengthiest of the shortest paths in G, also the maximum eccentricity $d_G = max_u \epsilon(u)$, $u \in V$).

A graph $G = (V,E)$ is ***connected*** if a path exists between every pair of nodes in the graph $\delta_G(u,v) \neq \infty$, $\forall u \in V$, $\forall v \in V$. When a graph is not connected, it is composed of a union of connected subgraphs called ***connected components***.

The ***neighbourhood*** of a node u in $G = (V,E)$ is the set of nodes $N_G(u) = v \in V, \exists e(u,v) \in E$.

The ***degree*** of a node u in G is then $d_G(u) = |N_G(u)|$. In a directed graph, ***in-degree*** $d_G^-(u)$ refers only to $|N_G^-(u)| = |\{v \in V\}|$, $\exists e = (v,u) \in E$ and ***out-degree*** $d_G^+(u)$ to $|N_G^+(u)| = |\{v \in V\}|$, $\exists e = (u,v) \in E$.

A ***bipartite*** graph $G = (V,E)$ (Figure 2.4, left) is a graph with a partition $(V_1, V_2), V_1 \cap V_2 = \emptyset, V_1 \cup V_2 = V$ such as $\forall e(u,v) \in E$, $u \in V_1$, $v \in V_2$. A bipartite graph noted $G = (V_1 + V_2, E)$, often represents networks of relations between two different types of nodes. The matrix representation A of a bipartite graph is thus an $|V_1| \times |V_2|$ rectangle matrix where $a_{ij} \neq 0$ represents a relationship $e(u_i, v_i)$. In our model, we refer to each partition of a bipartite graph as ***substrates*** and ***catalysts***. ***Substrates*** are entities of interest among which we wish to observe interactions of ***catalysts***.

A ***cycle*** in a graph $G = (V,E)$ is a path starting from and ending at node v_1 in a graph such as $(v_1, e_1, \ldots, v_1)$.

An ***acyclic graph*** is a graph that contains no cycle.

A ***tree*** $T = (V,E)$ is an acyclic connected graph. In a tree, there is *exactly one* path between two nodes. An edge in a tree is called a ***branch*** and every node u with degree $d_T(u) = 1$ is called a ***leaf***.

A ***rooted tree*** $T = (V,E)$ is a tree with a root node r inducing a directed path to every other node. In a rooted tree, with the exception of the ***root*** r, every node u has an exact *in-degree* $d_T^+(u) = 1$ and the associated node is called its ***parent*** (Figure 2.2). Any nodes preceding u in the directed path $(r, e, \ldots, u)$ can also be referred by extension as *one* parent of node u.

A ***random graph*** $G = (V,E)$ is a graph generated by a stochastic process, for which we suppose that all the edges $e \in E$ have been randomly selected from all possible subsets of $V \times V$. Different models are available in the literature, of which the Erdős-Rényi (Erdős and Rényi, 1959) is the most well known (Figure 2.3). It supposes a fixed probability p for an edge creation between every pair of nodes

Figure 2.2: A rooted tree. The *root* node of the tree (in *green*) only has children nodes and gives a natural direction to every edge (or *branch*). A bottom node, of degree 1, is called a *leaf* (in *red*) and only has one *parent* node.

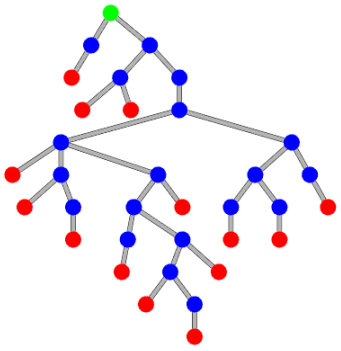
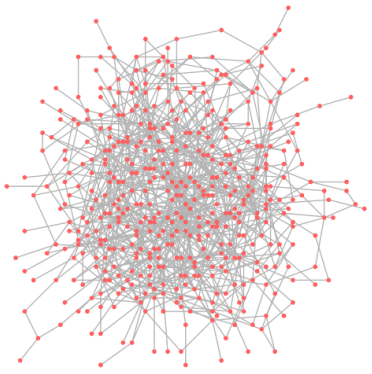


Figure 2.3: An Erdős-Rényi random graph with 477 nodes and p=0.56



in $V \times V$ so that the number of edges follows a binomial distribution $|E| \sim \left(\binom{n}{2}, p\right)$. Other models have been surveyed in (Goldenberg et al., 2010).

A ***multiplex graph*** (also called a *multilayer* and *layered graph*, Figure 2.4, center.) is a multiple graph $G = (V, E) = (V, \sum_i E_i)$ that differentiates layers of edges $E_1, \dots E_n$ such as $\bigcap_i^n E_i = \emptyset$ and $\bigcup_i^n E_i = E$. In our model, nodes represent *substrates* and layers of edges *catalysts*.

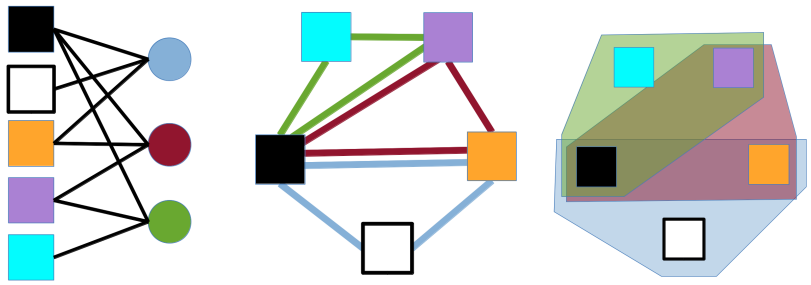


Figure 2.4: Representation examples of a bipartite graph (left), of a multiplex graph (center), and of a hyper-graph (right). These models are very closely related, and coloured nodes or edges have correspondence in every other model (respective to the colour of entities).

A ***hyper-graph*** (Figure 2.4, right) $H = (X, E)$ is a generalization of a graph of which X designates a set of nodes and E is a set of non-empty set of nodes called *hyper-edges*[3]. In our model, nodes represent *substrates* and hyper-edges *catalysts*.

3. This thesis will not cover the theory of hyper-graphs, but only use them representated as bipartite graphs. Interested readers are invited to deepen their knowledge by reading (Berge and Minieka, 1973) and (Berge, 1984).

The ***clustering coefficient*** of a node u in a graph $G = (V, E)$ is the ratio between the number of edges $|E_u|$ within its neighbourhood $N_g(u)$ and the number of edges if they were all connected together $\frac{1}{2}(|N_g(u)| + 1)(|N_g(u)|)$ so:

$$C_u = \frac{2|E_u|}{(|N_g(u)| + 1)|N_g(u)|}$$

This computes the number of triangles existing in a node's neighbourhood over the number of all possible triangles. The clustering coefficient of a graph is the average over its nodes $C_G = \frac{1}{|V|} \sum_{v \in V} C_v$.

The ***betweenness centrality*** of a node u (or of an edge e) in the graph $G = (V, E)$ is the number of shortest paths between every pair or nodes $(v, w) \in V$, $v \neq u$, $w \neq u$ that goes through u (respectively e) (Freeman, 1977).

The ***density*** of a graph corresponds to the size of a graph $G = (V, E)$ in comparison with the size of a complete graph of the same order:

$$\rho_G = \frac{2|E|}{|V|(|E| - 1)}$$

The ***drawing of a graph*** (Figures 2.1 or 2.2) is a graphical representation of the graph in a displayable 2D or 3D space. A node-link diagram is a representation for which every node is represented by a symbol and has a coordinate in its display space, and every link is

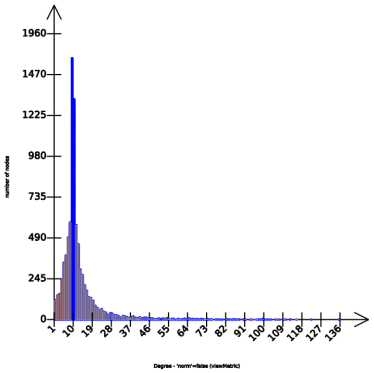
represented by a line connecting its adjacent nodes. Other visual representations such as a matrix view are also possible (Ghoniem et al., 2005).

2.1.2 *Complex networks*

The study of complex systems is the study of *"how relationships between parts give rise to the collective behaviors of a system and how the system interacts and forms relationships with its environment"* (from wikipedia - complex systems). Network science is embedded in the study of complex systems as it involves the study of "network representations of physical, biological, and social phenomena leading to predictive models of these phenomena" (Committee on Network Science for Future Army Applications, 2005).

Complex networks have been studied for a long time, and their characterization and definition is still the subject of publications (Kim and Wilhelm, 2008; Costa et al., 2007). We can however admit that a complex network is a network which presents characteristics we cannot find in simple networks nor in random networks, but that we very often find in real-world networks (Watts and Strogatz, 1998). More empirically, the literature has uncovered (Albert and Barabási, 2002) and still uncovers (Pagani and Aiello, 2013) complex network features in many – if not all – real world networks. Of such characteristics the most famous are (Albert and Barabási, 2002) a small world appearance – the diameter of the network is rather limited in comparison with one of a random graph with the same number of nodes and edges; a scale free degree distribution – some observe a long tail in distribution of nodes degree (Figure 2.5), often referenced as *power law*; or a high clustering coefficient – the coefficient is much higher for real world networks than for random networks. A complex network presents at least some of these features – which are not limited to the ones depicted previously (Claussen and Wilhelm, 2008). In a real-world graph, we can also observe many disconnected components with a lot of variation in their size (probably induced by some inner scale-free feature as described above). One or more largest components may present most of the features of complex networks.

Figure 2.5: Here is a histogram view of nodes degree of a similarity network of 10,000 news documents collected at INA, with about 1.5M edges (and spread among nearly 300 clusters), notice the long tail in the distribution. This graph presents a clustering coefficient of 0.697 versus 0.031 for a random graph of the same size and order.



Strogatz in his work (Strogatz, 2001), defines six different ways for networks to display complexity:

- structural complexity (edges are tangled)
- network evolution (the network evolves over time)
- connection diversity (weights/directions/signs of edges)
- dynamical complexity (node states can vary with time)
- node diversity (different types of nodes)
- meta-complication (a combination of the preceding complications)

Networks made from news documents appear to display characteristics of complex networks, such as a high clustering coefficient with a long tail degree distribution (see Figure 2.5). Before explaining how we constructed our networks of news documents, we can see that the document network displays all the previous complicating factors. Such network is very tangled; news have an existence over time as they are published over real world events; semantic proximity between news can be weighted; citations can be directed; and news can be typed – *e.g.* with its editor or other meta information. In our context, one news excerpt, once broadcasted, is considered as one point in time. It has been published once, then archived and remains unchanged. Later in the manuscript, we will use another dimension of connection diversity: the different types of edges; and add one more dimension of complication in the structural complexity: the edge entanglement.

2.1.3 *Documents*

Before examining documents from a network analysis point of view, we will learn a little more about those documents. Firstly we will formally define a news excerpt as a *document* d indexed by *terms* t_i given a vocabulary $\mathcal{T}_\mathcal{H}$. Terms are unique and chosen to be unambiguous. A document can be then defined as

$$d = \{t_1, \ldots, t_n\}, t_i \in \mathcal{T}, t_i \neq t_j \forall i \forall j$$

It is worth noting that such a representation already induces a distortion since two different documents in the real-world can have the same representation of index terms. However, our focus is mainly directed to the study of the semantics of a document group. In addition, we also used two other meta information as filters, the document's publication date and its diffusion channel. We also use the document's *title* convenient for overview.

The vocabulary defining our terms is based on a *thesaurus*[4] maintained by INA. Discussion of the thesaurus itself is beyond the scope of this thesis – as we had but limited access and did very little research on it[5] – however mentioning its structure has significant implications in the results we are showing later in the manuscript.

This thesaurus can be defined as a rooted tree $\mathcal{T}_\mathcal{H} = (V, E)$ on which each node ($v \in V$) is either a term or a category linked to its parent *term* or *category* (and rooted in the category "thesaurus"). *Terms*, *categories*, and relationships are developed and constantly updated by INA's specialists, and only *terms* included in the vocabulary $\mathcal{T}$ allow for document indexing ($\mathcal{T} \subseteq \mathcal{T}_\mathcal{H}$). The *categories* exist only for understanding of thesaurus, but not indexing. Figure 2.6 shows an extract of the thesaurus of common nouns rooted at *Science*. Under the root node *thesaurus* of the tree, there are two main categories: *common nouns* and *places* – there is also *people*'s information but that is kept in a different database out of our reach.

4. INA's *thesaurus* is an organized and controlled list of terms representing less ambiguous possible concepts for documentation and indexation purpose, it is proprietary and well kept since it is used at the base for INA's services and retrieval tasks.

5. As an agent of INA, I had the opportunity to show to my collaborators who were in charge of the thesaurus, how visualizations can be helpful to get insights and understanding of their process. We also published, with colleagues at LaBRI, a new method for such visualization (image below), based on space-filling curves (Auber et al., 2013), but this will not be discussed in this manuscript.

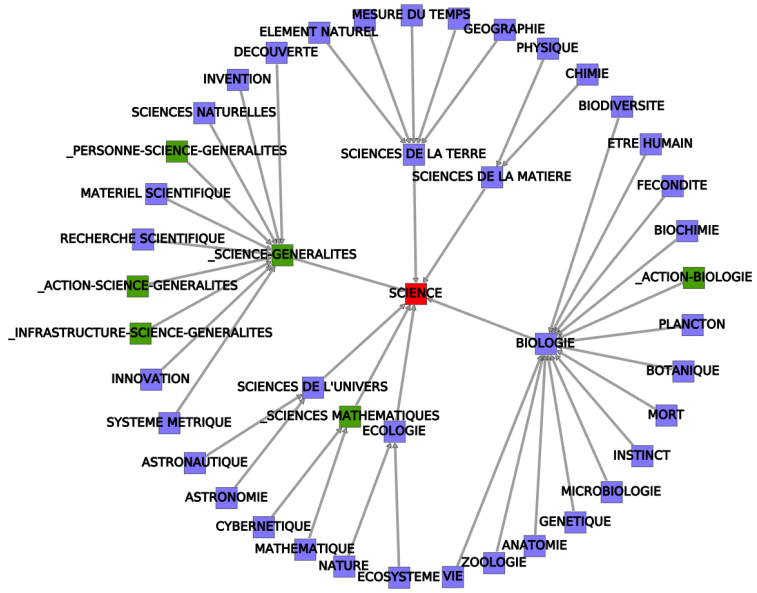


Figure 2.6: An extract of the common nouns thesaurus rooted at the term *Science* in red), with two levels of depth. Categories which are in green, such as (translated in English) *Science-General-Concepts* are not suitable for indexing, contrarily to the other terms (in blue) such as *Biology*.

The document groups we studied came from very different sources. They were the result of a query on an index search engine (such as Apache Lucene[6]), and were filtered results of a sub-corpus (such as results of a query divided by channels of publication, or by week), and also the results of different clustering methods (image-based clustering (Joly et al., 2012) or document clustering (Thièvre, 2006; Viaud et al., 2010)). No matter which grouping method we used, we were constantly surprised by the constant uneven quality of the resulting groups. There always seemed to be a trade-off between what the machine understood of a query, or a group, and the intentions of the person posing the question. Thus in efforts to devise techniques to enhance current grouping technologies, we have used different groups as a base for analysis, motivating our study of group semantics, questioning them and turning them into objects of study.

In our model of analysis, we will refer to documents as *substrates* and to terms as *catalysts*.

Now, a few facts and figures about one typical example of a news document corpus (part of the OTMedia capture efforts[7]):

- 9,802 documents[8],
- 7109 different terms index the corpus,
- there are between 8 and 9 index terms per document on average (see Figure 2.7 top left for distribution),
- terms occur on average 11 times in the corpus, ranging from 1 to 1140 times[9] – (see Figure 2.7 top right for distribution),

6. Apache Lucene available at http://lucene.apache.org/

7. Further description of this project at http://www.otmedia.fr, (Hervé et al., 2013)

8. Captured between 2011-09-01 and 2011-11-29, and spread among nine different news channels:

Channel	News subjects
TF1	1678
France 2	1674
France 3	1328
M6	1181
Canal +	969
LCI	964
ARTE	841
BFM TV	697
France 24	470

9. The term that occurs most often *economical crisis* !

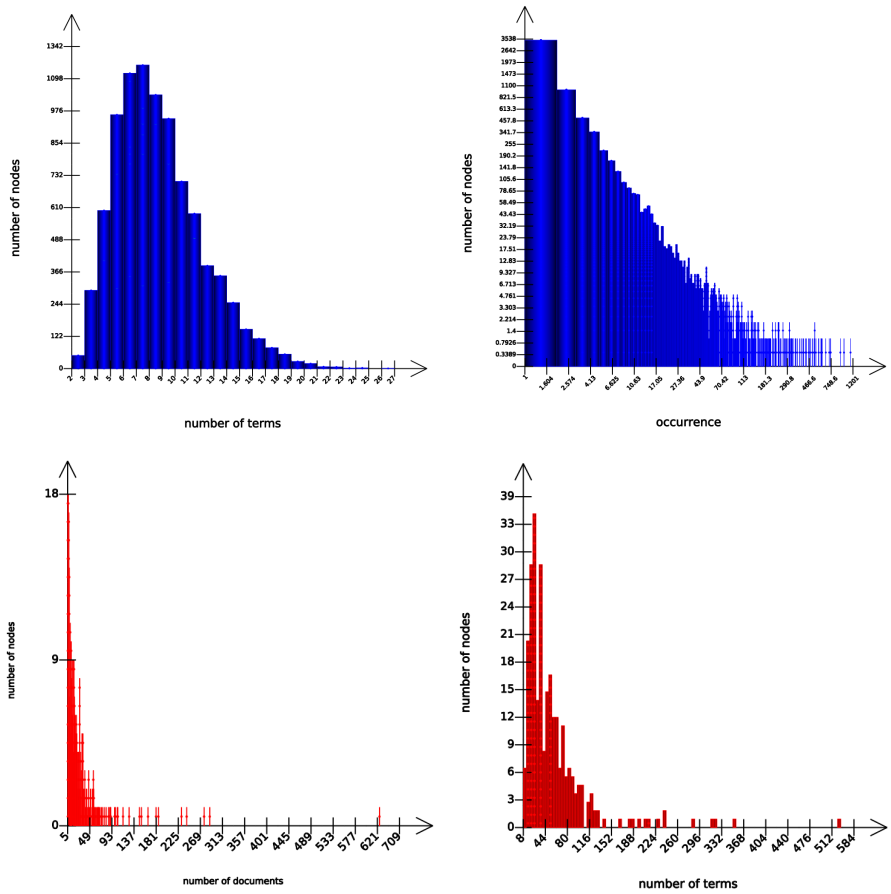


Figure 2.7: Top left, the overall distribution of terms per document, ranges from 2 to 27, with an average of 8.28. Top right, the overall (log)distribution of term occurrences in the corpus (x and y on log scale) shows that most terms occur on a really small number of documents while a few dominate the news by 3 orders of magnitude. Bottom left shows the distribution of the number of documents per cluster, 33.80 on average. Bottom right, the distribution of terms per cluster is 60.93 on average. It is interesting to note the imbalance in the clusters with respect to the number of documents and terms, suggesting, perhaps, some inequality between the different groups.

- documents spread over 289 preprocessed clusters (Viaud et al., 2010), and
- about 34 documents per cluster on average, with 61 unique terms indexing each (see Figure 2.7 bottom diagrams for distributions).

2.2 A few notes on document analysis

With such document collection, classical analysis includes several areas of expertise including data mining, knowledge discovery, information retrieval, and most particularly Topic Detection and Tracking (TDT). TDT aims to deploy technology to monitor broadcasts, and to alert us when something new or interesting happens (Allan, 2002).

With news, TDT typically consists in detecting and following events. There are five main challenges in research for TDT (Allan, 2002):

- *Story segmentation*, cuts a stream into different meaningful elements of independent stories. This part is already processed in our case since we have documents representing identified news excerpts.
- *Topic detection*, builds groups of documents containing stories about the same topic. This work involves mainly clustering techniques[10], bearing in mind that a story can typically be assigned to more than one cluster.
- *Topic tracking*, selects and tracks clusters, given a set of example stories (sort of queries). This is in fact a way of filtering the corpus.
- *First story detection*, detects a story that concerns no known topic and forms a new group. This supposes to maintain a constant knowledge of known topics, and continuously update the document grouping.
- *Link detection*, establish links when two (or more) stories are connected. This involves topic proximity, and networks.

10. More on data clustering and recent advances can be found in (Jain, 2010)

Our work has not focused on the temporal aspects of the information, but on the semantic aspects of our document groups. Because language inevitably carries ambiguities, automatic or supervised classification processes are not entirely satisfying (Tomoharu et al., 2008; Chuang et al., 2012). Even when indexed from thesauri, documents may be grouped according to descriptors of varying levels of generality, inducing fuzziness in groups. Content based grouping may still contain noise and as a consequence, users might need to perform additional tasks to validate the relevance of the document grouping and/or selection. This involves doubting and inspecting the proposed groupings, going backwards and forwards over the different levels of representation of the data. It is certainly true that meaningful and valuable analysis requires constant alertness, in order to avoid conclusions based on noisy and/or fuzzy geometric proximities – matters that can affect even the most basic tasks.

We believe our contributions offer tools that can assist many of the TDT aspects that contribute to sound analysis. The semantic entanglement of documents, which we shall define in Chapter 3, can help judge the quality of detected topics; offer a pattern for as in, for example tracking stories; display new semantic relations that did not appear beforehand; or relate groups of documents with similar semantic intertwining.

To support those challenges, many tools are available, and most of them provide more or less accurate distance measure between documents and/or documents relevancy ranking given a query. This

enables clustering of document corpus, seeking particular features in documents, or exploring a document's neighbourhood, covering most Information Retrieval tasks (Ingwersen, 1996).

2.2.1 *From the Vector Space Model...*

Given the description of documents we made previously, one natural model can be applied to delve more deeply into such document collection: Salton's Vector Space Model (VSM) (Salton et al., 1975)[11]. This model builds on top of the Bags-of-Words model (Harris, 1954) from which a document's description can be represented as an unordered collection of words. In the VSM any document and any query can be represented via vectors, for which each dimension is represented by a word or a concept, and, for comparison purpose, these vectors need to be aligned:

$$\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{n,j})$$
$$\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

Many operations can be started from this model such as feature-based indexing for clustering purposes or multi-dimensional scaling for visualization. We often measure the cosine similarity between documents, or between documents and a query to retrieve for example the most relevant documents given a query. The cosine similarity between a document and a query is expressed as:

$$cos\theta = \frac{\vec{d}.\vec{q}}{\|\vec{d}\|\|\vec{q}\|}$$

All vector entries are positive or null. A cosine value $cos\theta = 0$ means that the document and the query vectors are orthogonal, so there is no common vocabulary between both members. A cosine $cos\theta = 1$ means that both document and query make a perfect match with the exact same use of vocabulary. For example, the distance between documents allows us to rank relevant documents of a query, and select the *k-nearest neighbors*, which is useful for document classification (Cover and Hart, 1967).

The terms of a document are very often weighted, and the Term Frequency - Inverse Document Frequency (TF-IDF) is very often associated to the VSM as proposed by Salton et al. (1975). From a vector representation of document x, defined by $\vec{d_x} = (w_{1,x}, w_{2,x}, \ldots, w_{n,x})$, the weight $w_{i,x}$ of a term i in the vector is

$$w_{i,x} = tf_{i,d_x}.idf_i$$

where tf_{i,d_x} is the frequency of term i in document d_x (of length N_{d_x}) and idf_i is the inverse frequency of documents in the document corpus D that contain the term i:

$$tf_{i,d_x} = \frac{n_i}{N_{d_x}}$$

11. The history of the Vector Space Model is an interesting case of citation bias in research. The citation of Salton et al. (1975) is actually a misunderstanding of Salton's work, that has been repeated over years of citations (Dubin, 2004). Since we only introduce the concept we will stick to this most-commonly cited article.

$$idf_i = log \frac{|D|}{|\{d_y \in D | t \in d_y\}|}$$

Despite being a classic in Information Retrieval, the VSM presents limitations such as issues with dimensionality on documents containing large number of terms; the probably wrong assumption that terms are statistically independent from one another; or again the loss of term ordering. Even so, the model still remains a very flexible framework to begin analysis of document corpora, and many models have successfully extended the VSM. For example, the Generalized Vector Space Model will avoid the independence assumption (Wong et al., 1985), and Latent Semantic Analysis (LSA) offers much as we shall now see.

2.2.2 *..and beyond: studying the document-term space*

LSA or LSI (Latent Semantic Indexing) (Deerwester et al., 1990) is an indexing technique for information retrieval that studies the latent structure of relationships between terms and documents in raw texts. It assumes that concepts are composed of words that tend to have similar meanings which are used in the same context. The methodology is based on a singular value decomposition[12] of a (weighted)[13] $n \times m$ term-document occurrence matrix (n documents in the corpus and m different terms).

Only the k-principal components are kept for the output, reducing the complexity of the space and keeping the most represented words as representative of the corpus. The resulting output would be a document-term matrix of much lower size – documents are described in a low-dimensional space – from which distances between documents are more relevant, easing clustering or querying methods. This analysis can also be transposed to study the relationships from the terms point of view, with a term-document matrix.

The probabilistic model that lies behind LSA assumes that terms are normally distributed, even when a Poisson distribution is observed. To cope with this issue, pLSA (or probabilistic Latent Semantic Analysis) (Thomas, 1999), is an extended model that focuses on the probability of co-occurrence as a mixture of independent multinomial distributions:

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

where w is a word in a document d and belongs to a concept (latent class) c. This probabilistic model can fit some real data by finding its parameters with for example an Expectation-Maximization algorithm[14]. Outputs of pLSA are similar to those of LSA.

Latent Dirichlet Allocation (or LDA) (Blei et al., 2003) further extends pLSA, by supposing that each document is a mixture of a small number of topics, and each word has a probability of belonging to each topic in the document. LDA differs to pLSA by assuming a

12. A singular value decomposition is a process of linear algebra, enabling the factorization of a rectangle matrix of real or complex entries. The factorization decomposes the matrix to an orthonormal base. For the sake of understanding, we can interpret in our case, that the resulting orthonormal space describes the structure of the document-term space, enabling the extraction of terms that are most descriptive of each document in this space.

13. TF-IDF is one of many types of possible weights.

14. The EM method is an iterative algorithm that finds the maximum likelihood estimates of parameters of statistical models, given unobserved latent variables. The description of the algorithm is rather long and does not bring anything relevant to this manuscript, but the curious reader is invited to consult (Alpaydin, 2004; Witten and Frank, 2005) for an introduction to machine learning or (McLachlan and Krishnan, 2007) for expertise.

Dirichlet prior the topic distribution (where pLSA assumes an uniform distribution). In other words, LDA worries about the distribution of topics which are distributions of words among documents. One inconvenience of such technique is the need to set, in advance, an approximate number of topics. Extensions of LDA offer hierarchical topic modelling (Griffiths et al., 2004), then offers a non-parametric solution to LDA, based on a hierarchical Dirichlet process (Teh et al., 2006). LDA outputs[15] a number of topics with associated probabilities for terms that might belong to such topic. It is then quite straightforward to regroup documents around common topics, given their associated terms.

15. The description of an algorithm solving LDA requires strong foundations in statistics (involving Gibbs sampling, details in (McLachlan and Krishnan, 2007)) and will not be detailed here, however curious readers may refer to the original paper (Blei et al., 2003) for advice on implementation.)

There are also plenty of interesting studies such as looking for semantic correlations between documents, or for the construction of term ontologies to enhance mining, or even in natural language processing to efficiently identify salient meaningful words. Though fascinating, these challenges have been covered by many books (Baeza-Yates et al., 1999; Maimon and Rokach, 2005; Manning et al., 2008) and are beyond the scope of this thesis.

2.3 A document corpus as a graph

Now that we have seen there might be a complex network structure behind a document corpus, and different (basic) ways and strategies to mine such a corpus, we can examine how graph models may be applied to such a group of documents.

2.3.1 A simple graph of documents

A simple graph of documents uses nodes to represent documents, and links to represent relations between those documents. With the example of the web, links between documents are natural hypertext links (Henzinger, 2001). Similarly, scientific papers give rise to a citation network (Hummon and Dereian, 1989). In the case of co-authorship (Newman, 2004), documents are papers, and links between documents are shared when documents are published by the same authors, or in another case, when documents bear the same keywords. Semantic proximity sometimes enables connections across documents (Viaud et al., 2010). Such proximity can be built upon a version of the vector space model, where distances could range from the length of the set intersection (keyword co-occurrence) (Sallaberry et al., 2010) to more complex distances built on top of natural language processing techniques (such as a distance of paths within a thesaurus (Hossain and Angryk, 2007)).

Keeping in mind that our usage needs focus on the structure and quality of semantic groups, the graphs built upon INA's documents are naturally based on semantic proximity. To do so, we have mainly considered TF-IDF (for our largest collections of documents indexed

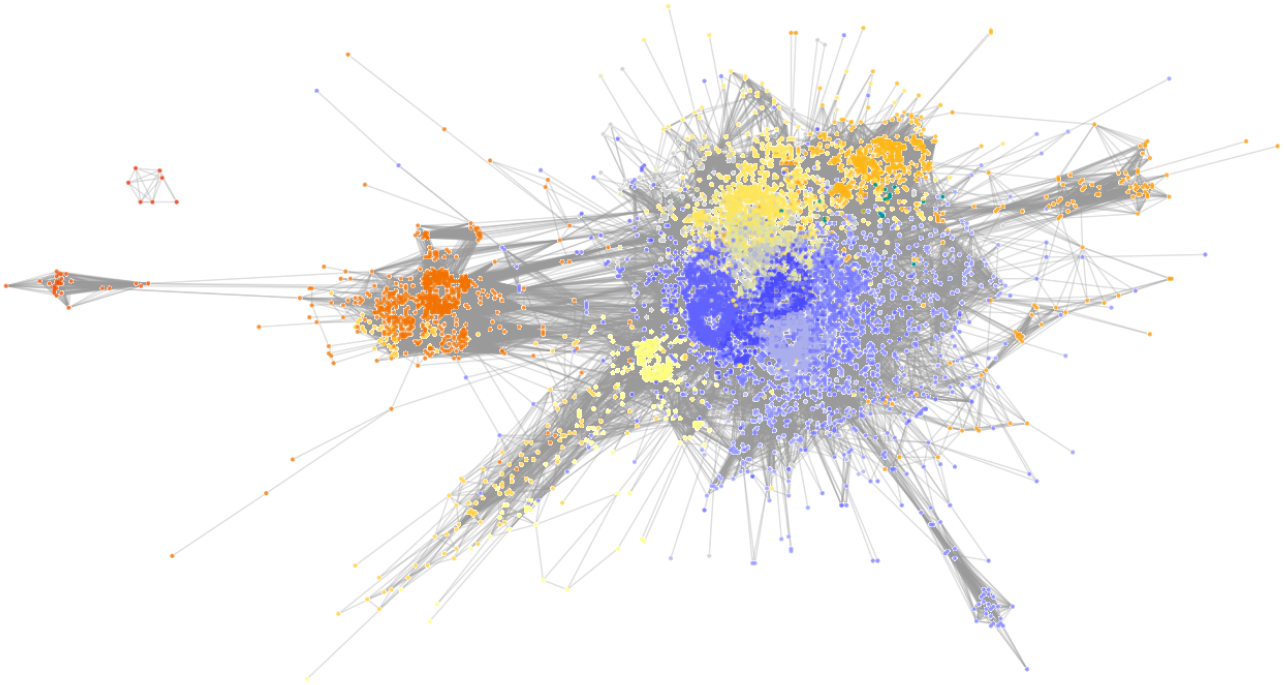


Figure 2.8: A semantic proximity network made with 10,000 news excerpts connected by keywords co-occurrence: set intersection $|A \cap B| \geq 2$. The colours display the different communities computed by the modularity-based "Louvain" algorithm (Blondel et al., 2008) and the layout is FM^3 (Hachul and Jünger, 2005).

from Lucene[16]), Jaccard's distance[17], and keyword set intersection (Figure 2.8).

16. Lucene offers a very convenient interface to compute TF-IDF distances

2.3.2 *Tasks for graph manipulations*

We use the graph model to represent relationships, but what are we looking for when we face a network? Social science has long observed global phenomena as consequences of a social association (for example (White et al., 1981; Kogut, 2000)), and network autocorrelation models (Dow et al., 1984; Leenders, 2002) are weapons of choice in such cases. But when looking deeply into networks, exploring structures and questioning networks, what are the tasks we need to achieve? *"Looking into"* is a visual metaphor and, fortunately, the field of *graph visualization* has already pondered such tasks (Lee et al., 2006) offering us a quite complete taxonomy, separating *overview*, from *topology-based* tasks, and *attribute-based* tasks.

17. Jaccard's distance (Jaccard, 1901) measures the ratio between intersection and union of sets:

$$J_\delta(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The **overview tasks** are part of an exploratory task which aims to estimate, speedily, an idea of the local features of a graph, as in, for example, estimating the size of a network, its diameter or community structure. Overview estimation allows immediate focus on interesting features corresponding to local areas. Other overview tasks include finding patterns and outliers, as in, for example, redundant *k-cliques*, or oddly connected nodes.

The **topology-based tasks** are focused only on the pattern of con-

nection between nodes:

- ***Adjacency (direct connection)****: Finding the set of nodes adjacent to a node*. Here we question the number of adjacent nodes to another node (*e.g.* in disease spread (Shirley and Rushton, 2005), how many people can catch a disease from patient A), or which node has direct access to certain information from a specific node (*e.g.* again in disease spread, who can catch a disease from patient A?). We also question patterns of adjacency, a node that is connected only to another node (*e.g.* in a social network, A is completely dependent of B to access the *Club*'s network), or nodes that have the maximum number of adjacent nodes (*e.g.* people most-likely to receive and transmit information).

- ***Accessibility*** *(direct or indirect connection): Finding the set of nodes accessible from a node*. With directed connection, we can question the number of nodes accessible from a particular node, or the number of nodes that access that node (*e.g.* to study route traffic).
  Finding the set of nodes accessible from a node where the distance is less than or equal to n (*e.g.* looking for distant family members).

- ***Common Connection****: Given nodes, finding a set of nodes that are connected to all of them* (*e.g.* common friends of two people).

- ***Connectivity****: Finding the shortest path between two nodes*. This is very useful for routing information, as in a car's GPS, or for a researcher to know one's *Erdős number*[18].
  Identifying clusters, which helps us understand community structures in graphs, has become its own area within graph processing (Fortunato, 2010). It is useful for classification.
  Identifying connected components. There is no possible path between nodes of different connected components and two connected components may be treated as two separate cases.
  Finding bridges. A bridge is an edge such as when it is removed, the graph composes two connected components. It has strategic importance for someone who wants to isolate an area.
  Finding articulation points. Articulation points, or nodes that are bridging communities (*e.g.* people bridging *structural holes*[19]) are at advantageous places (Burt, 2004) in their network.

Attribute-Based Tasks in a graph are more classic, similar to usual search or retrieval tasks with the exception that they apply here to graph objects (nodes and links).

- ***On nodes****: Finding the nodes having a specific attribute value*, and identify a specific entity or group of entities within the network.
  Reviewing the set of nodes. This allows us to check measures, or features of a selected group of nodes (for example questioning the degree of a selected set of nodes).

18. Paul Erdős was a very influential mathematician who published at least 1,525 papers during his life. The *Erdős number* is a humorous tribute to his work. It represents the "collaborative distance" with Paul Erdős, *i.e.* the co-authorship distance of an author with him (Goffman, 1969). Direct collaborators of Paul Erdős have a number of 1, collaborators of collaborators a number of 2 and so on...

19. Ronald S. Burt, sociologist, did a thorough study of social capital in social networks, more specifically on the concept of *structural holes*: In social structures, when there is no redundant information across two separate groups, there is a *structural hole* between them (Burt, 1992).

- ***On links**: Given a node, finding the nodes connected only by certain types of links.* Links can also bear attributes like nodes, and one can find a specific link, or a set of links, for example the biggest migration flows.

- ***Browsing Tasks**: Following a given path.* For example find the first common ancestor between two persons.

- ***Revisit**: Returning to a previously visited node.* This may apply more specifically to exploration, and in the case of web browsing, we often come back from a query answer looking for alternate answers to the same query.

2.3.3 Measures in network analysis

We should now see the common measures used for graph exploration and mining. We will present only *endogenous* measures, derived from the network's topology – as opposed to *exogenous* measures that are calculated from the external attributes of the raw data. We can find measurements at a global level (*i.e.* for a whole graph) and a local level (attributed to nodes and/or edges). Note that most of the following measurements are presented from their original definitions but can also be adapted to weighted graphs.

Among global measures, we have already introduced ***order*** $|N|$ and ***size*** $|E|$, giving us information on the ***density*** of a graph. Areas of higher densities are often at the centre of communities, and areas with very low density might be source of structural holes. The ***clustering coefficient***, ***diameter*** and ***radius*** of a graph offer information on the small world characteristic of the graph we are dealing with (Watts and Strogatz, 1998).

Another type of measures derives from the community structures of the graph, and this is used to express the quality of a clustering within the network. One basic notion is the *coverage* that only computes the fraction of edges in a graph that fall into communities. The most popular is Newman's ***modularity*** Q (Newman and Girvan, 2004; Newman, 2006). The modularity globally measures the fraction of edges, in a graph, that falls into groups, minus the expected fraction of edges that would fall into the same groups if edges were distributed at random. In a graph $G = (V, E)$, given a partition $C = (C_1, \ldots, C_n)$ of its nodes, the modularity of the partition C is defined by

$$Q(G, C) = \sum_{i=1}^{n} \left(\frac{e_{ii}}{|E|} - \left(\frac{d_i}{2|E|} \right)^2 \right)$$

with $d_i = 2e_{ii} + \sum_{j \neq i} e_{ij}$ the sum of node degrees in the group C_i, e_{ii} being the fraction of edges with both ends in the group C_i. This measure is, however, much questioned in certain literature (Fortunato and Barthelemy, 2007; Good et al., 2010), and improvements of Q have been proposed (Reichardt and Bornholdt, 2006; Arenas et al., 2008),

along with other quality measures (Mancoridis et al., 1998; Rosvall and Bergstrom, 2008; Queyroi et al., 2013).

This leads us to the numerous graph clustering and community detection methods[20], some based on quality measures (Blondel et al., 2008), on spectral analysis (Ng et al., 2002) or random walks (Pons and Latapy, 2005), on graph cuts (Ding et al., 2001), or even on visual layouts (Noack, 2003; van Ham and van Wijk, 2004).

Local measures of graphs allows also to study the position of a node or an edge in the network.

Jaccard's index can also be applied to edges $e(u, v)$ by comparing its ends' neighbourhood:

$$W_J(u, v) = \frac{|N_G(u) \cap N_G(v)|}{|N_G(u) \cup N_G(v)|}.$$

Edges with a higher index are better anchored in their neighbourhood. Note that many other closely related metrics can be found at (Melançon and Sallaberry, 2008).

There are multiple ***centrality*** measures (Freeman, 1977, 1979; Borgatti, 2005) and here is a non exhaustive list of the most influential centrality indices[21]:

- The ***degree centrality*** of a node u in a graph $G = (V, E)$ defined as
$$C_d(u) = \frac{d_G(u)}{|V| - 1}$$
draws attention onto nodes of highest degree, which are those that have the greatest interaction in a network.

- The ***betweenness centrality*** of a node, previously defined in Section 2.1.1, highlights the nodes that host the highest flow of information, and these nodes may have great importance for the network (for example, a node connecting two communities). Freeman (1979) proposes a theoretical maximum centrality of $\frac{1}{2}|V|^2 - 3|V| + 2$ which can be used as a normalization factor.

- The ***closeness centrality*** of a node u in a graph $G = (V, E)$ defined as
$$C_c(u) = \frac{|V] - 1}{\sum_{v \in V} \delta(u, v)}$$
and this denotes how easily a node interacts with all others

- The ***Eigen centrality*** (Bonacich, 1972), refers to the principal eigenvector v of the adjacency matrix M of a graph $G = (V, E)$ defined as $\lambda v = Mv$. This measures the influence of a node in the network, since an influent node would be connected to other influent nodes. (Borgatti et al., 1990) offer to normalize each entry of the vector by $\sqrt{\frac{1}{2}}$ as this would be the maximum value given by the raw vector.

20. For more readings on graph clustering, we invite the reader to refer to (Schaeffer, 2007; Leskovec et al., 2010; Plantié and Crampes, 2013).

21. An interesting study of correlations of different centrality measures can be found in (Valente et al., 2008).

The last centrality measures naturally lead us to the ***Hubs and Authorities*** measure (Kleinberg, 1999) of directed graphs. *Hub* refers to nodes with many out-links conveying much information and *authority* to nodes with many in-links, thus centralizing information. The hub score $W_H(i)$ and the authority score $W_A(i)$ of a edge i in a graph $G = (V, E)$ are calculated iteratively:

$$W_H(i) = \sum_{(i,j)\in E} W_A(j)$$
$$W_A(i) = \sum_{(j,i)\in E} W_H(j)$$

PageRank[22] (Page et al., 1999) also aims to identify influential nodes in a directed network. The PageRank of a node u is measured as:

$$W_{PR}(u) = (1-q) + q \sum_{(u,v)\in V} \frac{W_{PR}(v)}{d_G^+(v)}$$

with q denoting the *damping factor*[23]. PageRank's model is based on random walks throughout the network, and the damping factor reflects the probability of the walk stoping at any step.

Other interesting measures include the ***coreness*** (maximum connectedness of a node), and the ***strength*** (counting triangular patterns in a node's neighbourhood (Granovetter, 1973)). We could not find a complete survey in the literature, but many other metrics can be found, as in (Botafogo et al., 1992) or in (Wasserman and Faust, 1994).

Efforts have also been made by the research community to study robustness and fragility of graphs (the number of nodes/edges that need to be removed in order to disconnect a connected graph) with techniques ranging from percolation theory (Callaway et al., 2000) to graph topology (Dekker and Colbert, 2004).

22. PageRank is the algorithm implemented in the first prototype of a search engine named Google that launched the famous *firm of Mountain View*.

23. The damping factor is one parameter of their search engine kept secret by Google.

2.4 Multiplex networks

Given the many tools we have for network analysis, we can now focus on complex networks through the lens of multiplex networks.

2.4.1 Different types of ties

In the previous sections, we have read many words referring to different animals in the zoo of graphs: the *subgraph*, the *clique*, the *tree* that represents the thesaurus, the many adjectives such as *directed*, *weighted*, *simple*, *multilayered*, *random*, *bipartite*, or even *complex* graphs, *etc*. These describe different structures and patterns of connections between entities, however we must now question what creates a connection between two entities.

Stephen Borgatti in (Borgatti et al., 2009) has proposed a typology of the different ties studied in social network analysis (here almost directly transposed from Figure 3 in (Borgatti et al., 2009)), in which nodes represent agents:

- ***Similarities*** (or ***homophily***). Similarities can be of *location*, agents share the same spatial and temporal space; *membership*, agents share the same circles, attended the same events *etc.*; *attribute*, agents can be defined with different attributes such as gender, age, education, *etc.*
- ***Social Relations***. Many possible social relations, *kinship* family tie between agents; *other role* such as friendship, professor/student, co-worker *etc.*; *affective* ties could be love, hate and so on; *cognitive* relations are reflections of an agent to another *e.g.* knowing, knowing about, seeing as happy.
- ***Interactions*** of any sort: has talked to, had sex with, collaborated with, helped, harmed *etc.*
- ***Flows*** such as information, beliefs, personnel, resources, migrations *etc.*

In a complex network, many different types of ties are often laid out together. If we take for example a network constructed from someone's Facebook friendships, with all the possible metadata, we can differentiate social relationships (family, friends, coworkers *etc.*), from attribute similarity (location, age, gender, social groups), from interactions (pokes), and flow (of information, through the exchange of posts, conversation). We often study such graphs only from the perspective of one interaction, yet this might hide other inner types of relations: you do not usually interact with pure strangers, you usually meet for lunch with business partners, colleagues, or friends. When many types of relations are captured together, the structure of the resulting network is completely blurred by overlapping structures of the subnetworks induced by each of the different types of ties. As a result it can become difficult to spot real community structures among nodes (Nick et al., 2013).

The consequences of most social relations are often materialized by interaction, after all aren't regular and frequent calls between two people a footprint of a strong social tie between them? Capturing the network structure from the remanence of these interactions can be easier than obtaining, expressly, the types of relations in the network – of course, this depends on the type of entities we observe. For example, Latapy et al. (2008a) not only observe the complex network of the internet through dissection of multiple answers in *traceroute*-like queries, but also differentiate ego-centred views of the internet among different machines. Nick et al. (2013) look for a core structure of dorm communities (*membership* relations) behind a global Facebook friendship network.

We propose, in this document, to study networks from the perspective of *homophily*. Homophily is often embodied in bipartite networks, where entities of a given type $\mathcal{A}$ (authors; movie actors) connect through entities of a different type $\mathcal{B}$ (papers; directors). Referring to the work of Manski (1993), we take the notion of a *group* to

be the central paradigm in an analysis of homophily networks. Numerous authors have indeed confronted homophily in many social behaviors or phenomenon (influence, contagion, information diffusion, *e.g.*) (Aral et al., 2009; Shalizi and Thomas, 2011; Bakshy et al., 2012), questionning Manski's *group effect* as the driving force behind the observed phenomenon: individuals in the same group tend to display similar behaviors because of similarities in their characteristics or environment.

2.4.2 *Bipartite structures everywhere*

There are many ways to model complex networks, with perhaps one specific for each type of real-world network[24]. However Guillaume and Latapy (2005) have suggested that *any* complex network could be represented as a bipartite graph: whenever one faces a complex network, it is possible to assign each node to one maximum clique[25] it belongs to.

Many real-world networks have a ***natural bipartite*** structure (Breiger, 1974), (Figure 2.4, left), also called *affiliation networks, dual networks* or *two-mode networks* (Borgatti and Everett, 1997). It can be derived from a natural *n-partite* structure from entities' attributes (Lambiotte and Ausloos, 2006; Zhou et al., 2007b), for example we can consider triplets $\{person, gender, country\}$ of which nodes would represent the different people, gender and locations. The literature shows numerous fields of application: authors are connected to papers in co-authorship networks (Newman, 2001a,b, 2004; Goldstein et al., 2005); actors are connected to movies in which they have participated (Watts and Strogatz, 1998; Newman et al., 2001); board members are connected to companies they lead (Battiston and Catanzaro, 2004; Robins and Alexander, 2004; Conyon and Muldoon, 2004); sexual interactions between male and female partners (Lind et al., 2005; Rocha et al., 2010); sports teams and members (Bonacich, 1972; Onody and de Castro, 2004); student registering to courses (Holme et al., 2007); consumers and producers in financial networks (Caldarelli et al., 2004; Dahui et al., 2006); recommendation networks connecting users to products (Perugini et al., 2004); people's attendance at theatre performances (Agneessens et al., 2004); participation of politicians to political events, or people to protests (Faust et al., 2002; Boudourides and Botetzagias, 2007); in peer-to-peer exchanges, peers are connected to the files they share or search (Voulgaris et al., 2004; Guillaume et al., 2004); internet networks, with computers connected to their local networks (Tarissan et al., 2013); or documents and words associations (i Cancho and Solé, 2001; Dhillon, 2001). From this last type of application, we can easily see how we may tackle our news documents from a complex network perspective.

Another inherent bipartite structure derives from the definition of ***hyper-graphs*** (Berge and Minieka, 1973; Zykov, 1974; McPherson, 1982) $H = (X, E)$ when hyper-edges E are sets of sets of nodes. This

24. *"Complex network"* has become quite a buzzword and the literature is full of application-oriented articles referring to network analysis of real-world phenomena. As for 2013, a quick search results show applications in meteorology (Boers et al., 2013), wireless networks (Leonel et al., 2013), bioinformatics (Dolinski et al., 2013), transportations (Zanin and Lillo, 2013) and so on

25. A maximum clique of a node u in $G = (V, E)$ is a maximum induced subgraph from its neighbourhood, which forms also a complete graph.

means that each hyper-edges $e \in E$ can be assigned to a set of nodes $V' \subseteq X$, giving rise to the bipartite model $G = (X + E, F)$ with edge $f(n, e) \in F$ if and only if $e \in E$ and $n \in X, n \in e$ (Figure 2.4, right). The bipartite graph model of a hyper-graph is also called an incident graph (Levi, 1942). Hypergraphs are also used to model numerous abstract problems (Gallo et al., 1993) such as transportation systems – a transportation line representing an hyper-edges (Nguyen and Pallottino, 1988, 1989) – or databases – minimal covers for functional dependencies as hyper-edges – (Maier, 1980).

A last category concerns the ***multiplex networks***, (multiple networks, multilayer networks, or layered networks) which consider complex networks as part of larger systems (Kurant and Thiran, 2006) (Figure 2.4, center). The different layers of such complex network may look different but they might also be strongly dependent on one another. As an example, the IP topology of the internet relies on a physical layer of wiring (Kurant and Thiran, 2005) and other computer networks (Cetinkaya and Knightly, 2004). Among other applications, multiplex networks have been studied as suitable models for economical systems (Snyder and Kick, 1979; Nemeth and Smith, 1985), biological systems (Sharan et al., 2005), or social networks (Dow, 2007; McPherson et al., 2001; Gould, 1991; Burt and Schøtt, 1985). Although multiplex networks $G = (V, \sum_i E_i)$ are by definition tied to one type of entity, we can model them with bipartite graphs such as $G' = (V + E, F)$ with E designating the different families of edges E_i and F representing an edge $f'_G(v, e_i)$ when a node $v \in V(G)$ is connected to another node $v' \in V(G)$ through an edge $e(v, v') \in E_i(G)$. Note however that this representation of G' induces a loss of specific information on individual paths $(v, v') \in E_i(G)$ by compressing them through one application $f : (v_1, \ldots, v_n) \mapsto F$, and re-projection, as described below, might result in the formation of undesirable edges[26] (see Figure 2.9).

26. This would lead to a "similarity"-network, in which nodes are connected together when they share a similar adjacent family of edges.

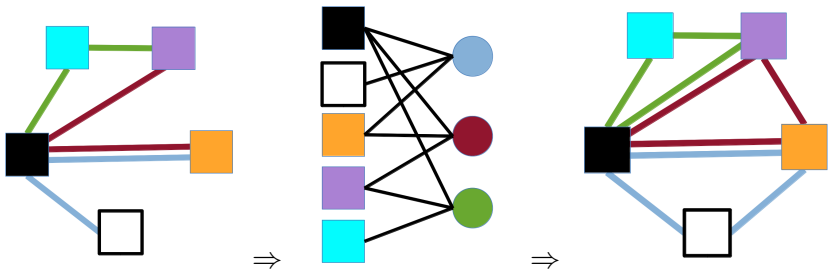


Figure 2.9: Successive transformations and side effects: we start from a multiplex graph (on the left), then convert it into a bipartite graph (middle) with the right (round shape) entities corresponding to adjacent edges in the first graph, and finally project the bipartite graph onto a multiplex graph (right) where new edges appear. Note that the right multiplex graph could be considered as an *edge-similarity* multiplex graph.

Inversely, multiplex networks are also obtained by projection of a bipartite structure (Newman, 2003; Zhou et al., 2007b; Jackson, 2010;

Neal, 2013). From a bipartite graph $G = (V_1 + V_2, E)$ we define two projections[27] into two multilayer graphs $G_1 = (V_1, F_1)$ respectively $G_2 = (V_2, F_2)$, with:

$$F_1 = \bigcup_{i \in V_2} F_{1,i},\ f = (v, v_i),\ v \in V_1,\ v_i \in V_2,\ \delta_G(v, v_i) = 1,\ f \in F_{1,i}$$

$$F_2 = \bigcup_{i \in V_1} F_{2,i},\ f = (v, v_i),\ v \in V_2,\ v_i \in V_1,\ \delta_G(v, v_i) = 1,\ f \in F_{2,i}$$

27. *N-partite* (multi-mode) graphs might also be projected by pairs of modes, and it would be really interesting to study more complex projections (a nice research agenda that may involve knowledge in advanced geometry and n-dimensional space projections).

Projecting a bipartite graph inducing relationships between entities of a same type is a common strategy (Zhou et al., 2007b; Neal, 2013) that, however, has the disadvantage of containing lots of cliques [28] (Guillaume and Latapy, 2005). The advantage of this model is to focus observations of relationships between entities of one partition of the nodes (Robins and Alexander, 2004; Borgatti and Everett, 1997; Neal, 2013). However, limited to these inter-entities relationships, the projection results in a loss of information as single associations of nodes across the bipartite graph would not create any new relationship in a projection(Robins and Alexander, 2004; Latapy et al., 2008b; Wang et al., 2009).

28. The node-clique association can also be modelled with bipartite graphs (Everett and Borgatti, 1998).

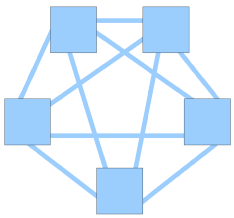
In this thesis we propose the study of entanglement in multiplex graphs. Since all the models presented can result in bipartite graph representations, and bipartite graph representations can be projected into a multiplex graph, the domain covered by our proposed methodology is quite large.

2.4.3 *Nuances in the multiplex model*

The complexity of multiplex networks can be approached in many ways. A weighted model of "flat" projections cannot capture them all, and weights have to be well chosen (Neal, 2013). In such graph, we will refer to *substrate* as entities of one type (document, actors, authors *etc.*) and to *catalysts* as entities of another type which are the entities interacting over the edges of a *substrate* network (keywords, movies, papers *etc.*)[29]. We will present four very different multiplex graphs that result exactly in the same 5-clique projection (Figure 2.11).

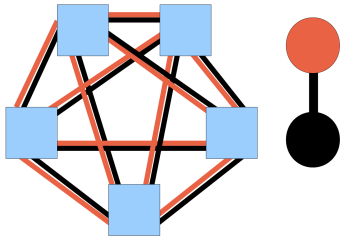
29. *Catalysts* can be seen as the elements that induce interactions across *substrates*.

Figure 2.11: A five node clique resulting in the "flat" projection of examples in Figures 2.10 and 2.12.

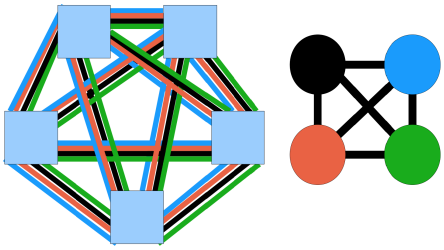


Consider the 5-nodes topology of the substrate graphs in Figure 2.10 – the network of square nodes on the left. Obviously a weighted model can capture the difference of dimensions in both networks, the one on top displaying 2 different types of catalysts, the one on bottom showing 4 different types of catalysts. In both cases, the catalyst interaction network corresponds to a clique. This weight is sometimes called the *multiplexity* (Podolny and Baron, 1997).

This weighted model has limitations since it does not capture the different types of interactions across substrate edges. Now, Figures 2.12 underline the "nuance" we wish to bring to the analysis of homophily networks. We have considered substrates with catalyst interaction in two different manners as shown in the Figure. Substrates in the square node graphs (left) are linked by an edge whenever they

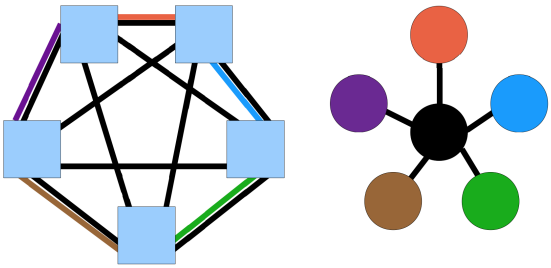


A multiplex clique with two layers on each edge

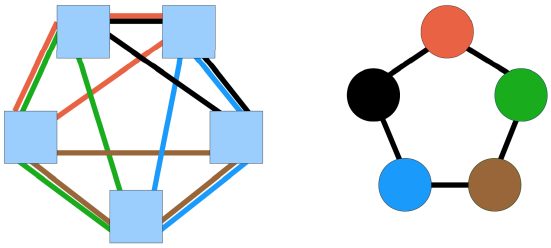


A multiplex clique with four layers on each edge

Figure 2.10: As a first "nuance" in multiplex graphs, we examine the way catalysts interact on substrate edges. In both figures, the square node graph (left) links substrates (documents, authors, movie directors, *e.g.*) whenever they are linked to the same catalysts (keywords, movie actors, *e.g.*). Catalysts appear as colours on induced links. The round node graph (right) describes how catalysts interact, that is when they co-occur as colours on an edge. The catalysts interaction network clearly distinguishes the two situations, whereas the projected single-type network shows identical topologies (Figure 2.11). Notice, however, that a simple edge weighting would differentiate both situations – with a weight of 2 on each edge of the graph on top, and a weight of 4 on each edge of the bottom graph.



A multiplex graph, with a star-shape interaction pattern



A multiplex graph, with an interaction pattern shaped as a cycle

Figure 2.12: Another example underlining the "nuance" we emphasize, looking at how *catalysts* interact. In both figures, the square node graph (left) link *substrates* projects to an identical single-type network (Figure 2.11). The weights one could assign to this projected single-type network are identical in both cases – we can assign 2 to the "outer" edges and 1 to the "inner edges". Only the topology of the *catalysts* interaction graph enables us to differentiate both cases.

share an attribute. Observe that in both situations the pairwise "distance" between substrates is the same – which would likely be assigned as a weight – because any two substrates share exactly two catalysts (or one), ending in identical topologies. As a consequence, based on pairwise distance, these two groups are somehow equivalent.

Now, consider the circle node graphs (right) describing how catalysts interact *within the whole group* of actors. Clearly, *all* substrates having the *black* catalyst give this catalyst a central position. If there were a reason explaining why these substrates form a group, it would certainly rely on the *black* catalyst, the other catalysts being somehow accessory. The second situation is much more balanced (although catalysts do not mix as intensely as they could). This small example points to situations where the analysis may be misleading, in situation where we only inspect single-type networks.

2.4.4 *Analysis of multiplex graphs*

The tools we shall now present, for the analysis of homophily networks, come from both the bipartite graph analysis and multiplex graph analysis research communities. It is generally admitted that, due to the loss of information during projection as previously discussed, bipartite graph analysis is preferred to projection analysis. However, such projections are mostly made on a "flattened" version of the multiplex graph (considering it only a simple graph) (Robins and Alexander, 2004; Borgatti and Everett, 1997; Podolny and Baron, 1997; Newman et al., 2001; Robins and Alexander, 2004; Zhou et al., 2007b; Borgatti, 2012).The analysis of multiplex graphs mostly extends existing concepts of "flat" graphs – as opposed to multiplex graphs. Despite that, some measures do not need to be updated from their regular graph version to their multiplex version, as one can consider traditional network analysis approaches on a weighted "flat" projection of a multiplex graph whilst considering as weight the *multiplexity* (Podolny and Baron, 1997) – number of parallel edges (in $G = (V, \sum_{t=1}^{n} E_t), w(e') = |\{e_t\}|, e_t \in E_t)$.

In a bipartite graph $G = (V_1, V_2, E)$, there are two ***orders*** to consider, one for each partition of nodes $|V_1|$ and $|V_2|$, although the definition of ***size*** remains the same. In a multiplex graph[30] $G = (V, \sum_{t=1}^{n} E_t)$, the notion of *order* remains the same, but there are multiple notions of *size*: one for each layer of edges, one for the whole graph, and one for its "flattened" version. The number of different layers of edges n is also an interesting metric.

30. A recent work (De Domenico et al., 2013) has proposed a very sound and complete mathematical background for multiplex networks problems formulated with tensors.

In a bipartite graph $G = (V_1, V_2, E)$, the ***clustering coefficient*** of a node u considers squares, not triangles. It is defined by (Lind et al., 2005), improved by (Zhang et al., 2008), as follows:

$$C_C(u) = \frac{q_{uvw}}{(|V_1| - \eta_{uvw}) + (|V_2| + \eta_{uvw}) + q_{uvw}}$$

where v and w are pair of neighbors of the node u, q_{uvw} is the number of squares that include the three nodes u, v, w. $\eta_{uvw} = 1 + q_{uww} + \Theta_{vw}$ with $\Theta_{vw} = 1$ if v and w are connected and 0 otherwise.
Clustering coefficients in a multiplex graph can be computed of each level of edges separately.

Many clustering or community detection procedures for bipartite graphs (Dhillon, 2001; Zha et al., 2001; Barber, 2007; Gaume et al., 2013; Crampes and Plantié, 2013) and multiplex graphs (Klein and König-Ries, 2002; Mucha et al., 2010; Gómez et al., 2013) are available. There are also many updates of the modularity quality measures in bipartite (Barber, 2007; Crampes and Plantié, 2013), and multiplex networks (De Domenico et al., 2013)[31].

31. Formulations of bipartite and multiplex modularity would be too long and not clearly relevant to this document, but we invite interested readers to refer to the articles (Barber, 2007; De Domenico et al., 2013) for further readings.

The measure of edges with Jaccard's index is straightforward for analysis of a bipartite graph $G = (V_1, V_2, E)$ since to a node $u \in V_1$ can be associated a set of *nodes* $\{v\} \in V_2, e(u, v) \in E$ (and inversely). We need a slight update for multiplex graph analysis $G = (V, \sum_{t=1}^{n} E_t)$. To a node u is associated a set of *edges types* $\{t\}$, $e(u, v) \in E_i$. Note that the intersection size of both sets, corresponding to both end-nodes of an edge, *might* differ from the actual number of edge types across this pair of nodes, given the input data.

Density in an undirected bipartite graph $G = (V_1, V_2, E)$ is defined in (Borgatti and Everett, 1997; Borgatti et al., 2009) as the fraction of edges over all possible inter-partite edges:

$$D_G = \frac{|E|}{|V_1||V_2|}$$

We can extend this definition of density to an undirected multiplex graph $G = (V, \sum_{i=1}^{n} E_i)$ as the fraction of observed edges over all possible edges among the different layers:

$$D_G = \frac{2|\sum_{i=1}^{n} E_i|}{n|V|\,(|V| - 1)}$$

Centrality measures have also been extended to bipartite graphs (Faust, 1997), as proposed here in (Borgatti and Everett, 1997):

- The ***degree centrality*** of a node u in a bipartite graph $G = (V_1 + V_2, E)$ defined as
$$C_d(u) = \frac{d_G(u)}{|V_X|}$$
where V_X designates V_2 if $n \in V_1$ and respectively V_1 if $n \in V_2$. In a multiplex graph $G = (V, \sum_{t=1}^{n} E_t)$, the degree centrality could easily be extended to:
$$C_d(u) = \frac{d_G(u)}{n\,(|V| - 1)}$$

- The ***betweenness centrality*** in a bipartite graph $G = (V_1 + V_2, E)$ can be computed as usual and normalized for a node u as follows:

$$\begin{aligned} C_{BC}(u) =& 2\left(|V_0| - 1\right)\left(|V_X| - 1\right), && if\ |V_0| > |V_X| \\ C_{BC}(u) =& \frac{1}{2}|V_X|\left(|V_X| - 1\right) \\ &+ \frac{1}{2}|V_X|\left(|V_0| - 1\right)\left(|V_0| - 2\right) \\ &+ \left(|V_0| - 1\right)\left(|V_X| - 1\right), \\ & && if\ |V_0| \leq |V_X| \end{aligned}$$

  where V_X and V_0 designate opposed sets of nodes, $V_X = V_2, V_0 = V_1$ if $u \in V_1$ and respectively $V_X = V_1, V_0 = V_2$ if $u \in V_2$.

- The ***closeness centrality*** – average distance to all other nodes – a node u in a bipartite graph $G = (V_1 + V_2, E)$ is defined as

$$C_c(u) = \frac{|V_X| + 2|V_0| - 2}{\sum\limits_{v \in V} \delta(u, v)}$$

  where V_X and V_0 designate opposed sets of nodes, $V_X = V_2$, $V_0 = V_1$ if $u \in V_1$ and respectively $V_X = V_1$, $V_0 = V_2$ if $u \in V_2$.

- The ***Eigen centrality*** of a bipartite graph $G = (V_1 + V_2, E)$, represented by the rectangle matrix $M_{|V_0|,|V_1|}$, can be obtained in a similar fashion to the Eigen centrality of a regular graph, from the first factor of the singular value decomposition of the matrix M (or principal component of the product MM'). However, two Eigen centralities need to be computed, one for each partition of the graph. Additionally, Borgatti and Everett (1997) proposes a normalization factor $\sqrt{\frac{1}{2|V_0|}}$ where V_0 designates the partition to which the node belongs.
  Eigen centrality can also be applied to multiplex graphs when represented by a weighted adjacency matrix. Additionally, Bonacich (1972) propose a factorial analysis of centralities within each layer of a multiplex graph.

The research community also tackled 2-cliques (or *bicliques*) searches (Peeters, 2003) which can be seen as cohesive subgroups in bipartite graphs (Borgatti, 2012). A biclique in a bipartite graph $G = (V_1 + V_2, E)$ is a subgraph $G' = (V_1' + V_2', E')$ such that $\forall u \in V_1'$, $\forall v \in V_2'$, $\exists e(u, v) \in E'$.

Among other interesting approaches, *bipartite network autocorrelation* has also been studied (Fujimoto et al., 2011; Bavaud, 2013), and PageRank has been applied to a bipartite graph of matrix M with a slight modification of the product MM^T(Bauckhage, 2008), and a recent generalization to multiplex graphs has been proposed in (Halu

et al., 2013), defined for a node u, in a layer i, of a multiplex graph $G = (V, \sum_{i=1}^{n} E_i)$:

$$W_{i+1}(u) = \alpha \sum_{v \in N_G(u)} x_u^{\beta} \frac{X_v}{G_v} + (1-\alpha) \frac{|V| x_u^{\gamma}}{\sum_{v \in V} x_v^{\gamma}}$$

with parameters $\alpha > 0$, $\beta \geq 0$ and $\gamma \geq 0$, $G_v = \sum_{w \in N_G(v)} x_w^{\beta}$, $N_G(w)$ denoting the neighbourhood of node w in G, and x_w the PageRank score of node w in the layer $i+1$.

Other measures specifically address the analysis of layer complexity in bipartite and multiplex networks.

Redundancy (Latapy et al., 2008b) is a measure of the neighbourhood overlap of a node u in a bipartite graph $G = (V_1 + V_2, E)$:

$$rc(u) = \frac{2|\{\{v,w\} \in N_G(u), \exists u' \neq u, e(u',v) \in E \text{ and } e(u',w) \in E\}|}{|N_G(u)|\,(|N_G(u)|-1)}$$

where $N_G(u)$ is the neighbourhood of u. This measures the fraction of pairs of nodes in the neighbourhood of u, that are linked to a node other than v, which would be connected together in a projection even if u is not in the network.

Network exposure (Fujimoto et al., 2011) of an actor measures the influence of events in an actor–events bipartite model projected on an actor–actor network (of matrix $M_{N \times N}$):

$$E = \frac{\sum_{j=1}^{n} C_{i,j} Y_j}{\sum_{j=1}^{n} C_{i,j}}, for\ i,j \in [1,N], i \neq j$$

where C_{ij} is a weighted adjacency matrix of the actor-actor network (with the number of attended events in the diagonal), and Y_j is a vector of attendant events.

Ambiguity (Burt and Schøtt, 1985) is a measure of links overlap in multiplex networks that inspired our work and will be thoroughly described in Section 3.3.1.

2.5 Our general setting

This section takes a closer look at the general framework we use, starting from homophily networks to multiplex graphs.

Bearing in mind that we are looking for semantic cohesion in groups of documents, we know this will be easier to achieve with smaller groups. Inspecting a group, in an effort to understand why and how cohesion is embodied in that group requires validation based on user knowledge, and makes more sense when conducted on small scale groups, gathering hundreds of nodes at most.

Our motivations came from simple questions that come to mind when inspecting a group, such as "*How can we assess that group of substrates really forms a cluster*?" "*How can we make sure all substrates of a cluster really belong to it*?" "*Should we suspect that group contains marginal (outlier) substrates*?".

The central ingredient we use to answer these questions is a set of metrics that capture the intensity and homogeneity of interaction between attributes in a group of actors, as we will see in Chapter 3. These metrics can be viewed as an aid to assess the internal cohesion of a group.

2.5.1 Multiplex graph and interaction network

Here is the description of the multiplex graph model we have chosen to represent documents and by extension any *homophily* network.

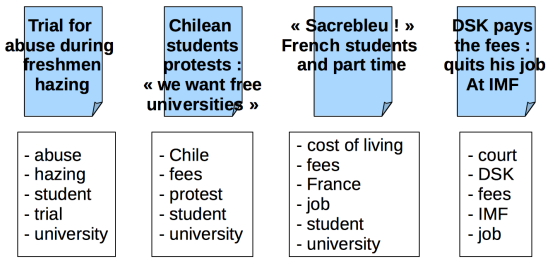
Our starting point is a set of substrates $\mathcal{A}$ with associated catalysts $\mathcal{B}$, as shown in Figure 2.13 (a). Many techniques use a vector of catalysts to compute distances between substrates and infer *semantically close* groups of substrates (as presented in Section 2.2).

The data can be usefully modelled as a bipartite substrate-catalyst network $G^*_{\mathcal{AB}} = (\mathcal{A} + \mathcal{B}, E)$, with edges $a - t$, whenever substrates a has associated catalysts t (see Figure 2.13 (b)). Two other networks are derived from the substrate-catalyst network, namely a substrate multiplex network G (and its "flattened" projection $G_{\mathcal{A}}$ the substrate interaction graph) and a catalyst interaction network $G_{\mathcal{C}}$. The multiplex network is usually built from the substrate-catalyst network by projecting paths $a - t - a'$ (linking substrates $a, a' \in \mathcal{A}$ through catalysts $t \in \mathcal{B}$) onto an edge $a - a'$, directly linking substrates. We also need to consider the catalyst t as a layer for the edge $a - a'$.

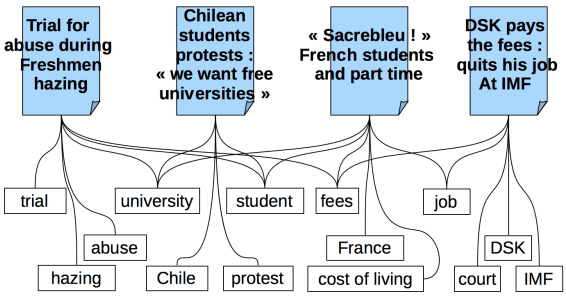
This defines a multiplex network of substrates $G = \mathcal{A}, \sum_{t \in \mathcal{C}} E_t$, with $\mathcal{C}$ the collection of all edges $a - a'$ bearing the catalyst t. This multiplex network is the ground base for analysis. It derives from the bipartite document-term network, focusing on documents as substrates. For the sake of simplicity we can consider that edges in $G_{\mathcal{A}}$ are thus labelled by subsets of catalysts (all catalysts $t, t', \ldots$ collected from triples $a - t - a'$, $a - t' - a'$, $\ldots$).

We filtered out some of the edges. Since we are focusing on document group cohesion and on terms co-occurrence, loops were discarded and we took into consideration only links across documents that contains catalyst interaction: term **co-**occurrence. We therefore only kept edges $a - a'$ that were inferred from at least two different terms t, t'. The resulting network is shown in Figure 2.13 (c).

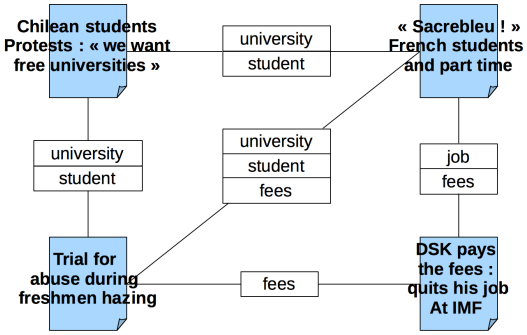
Note that some use cases involves data models that are starting directly from a multiplex network perspective, deriving from a superposition of layers of ties. For example, in Section 6.2.2, we consider a co-authorship network with *interaction*-type of ties – coauthorship – under the perspective of homophily – publication keywords. Recon-



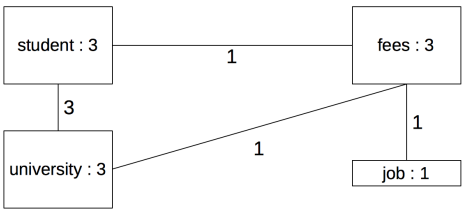
(a). Substrates (documents) $\mathcal{A}$ with catalysts (indexing terms) $\mathcal{C}$



(b). Bipartite graph $G^*_{\mathcal{AB}} = (\mathcal{A} + \mathcal{C}, E)$



(c). Multiplex graph $G = (\mathcal{A}, \sum_{t \in \mathcal{C}} E_t)$, $\mathcal{C} \subseteq \mathcal{B}$



(d). Catalysts (terms) interaction network $G_{\mathcal{C}} = (\mathcal{C}, E_{\mathcal{C}})$

Figure 2.13: The initial data (a) is formed of substrates (documents) having associated catalysts (keywords indexing documents). This situation is modelled by a bipartite graph linking substrates to catalysts (b) (documents concerning the same topics). We then consider the projected multiplex network with catalysts as different edges (c) and derive the resulting *catalyst interaction network* (d).

structing the bipartite network author–keywords and re-projecting it into our multiplex model would result in a very different network of relationships which reveals similarities between authors[32]. It is also possible to consider edges bearing only one attribute, however when facing documents with a large vocabulary, these edges bring a great amount of noise[33].

The catalyst interaction network is a central artefact in our methodology and will be further presented in Section 3.2.2.

32. The study of similarities and differences between an interaction-based network and its corresponding attribute similarity-based network would be very interesting work to tackle.

33. Within the whole corpus of documents presented in 2.1.3, this triples the amount of edges: from 1,479,362 to 4,644,837.

2.6 Conclusion

We have exposed in this chapter the main issues related to document analysis from the *multiplex network*'s perspective, in which *substrates* correspond to documents and *catalysts* to terms. We have introduced the data which has motivated the research and some traditional challenges and approaches related to this data. We have formally presented the graph model and its related concepts, the use and analysis of a graph, and how a document graph fits as a complex network. We have then advanced to the multiplex network model, a model that can be applied to many real-world issues. We have detailed the nuances and considerations proper to multiplex networks, and extended the common graph analysis methods this model. We have finally proposed the general setting that enables the analysis of our document collections under the light of multiplex graphs, offering all its potential applications of network analysis and visualization. This general setting is of course included to most of our publications related to document analysis, however, the premise of this setting has been introduced in (Viaud et al., 2010; Renoust et al., 2011a) and (Renoust et al., 2011b).

ENTANGLEMENT ANALYSIS

3

In this chapter we will use our methodology to tackle edge complexity through measures of entanglement in a multiplex network. We will firstly present the origins of entanglement analysis from studies in social sciences, introducing the *entanglement index* as a measure of intertwined layers of edges. After this we shall extend the concept to global entanglement measures of a multiplex network: *intensity* and *homogeneity*. We will then do a more thorough study of those measures, introducing the central role of the *catalyst interaction network* as a tool for investigating the interactions in the multiplex network. We will introduce the limitations of entanglement analysis and its interpretation, and conclude by presenting work in progress, along with further ideas about these measures.

3.1 From Burt's ambiguity to the entanglement index

In the previous chapter we modelled a corpus of documents as a multiplex graph. Social network analysis offers us powerful tools when it comes to understanding the properties of a graph from its topology, but most approaches to multiplex networks are confined to the study of the topology of a projected network. Our aim is to delve further into the source of complexity represented by edge intertwining. We believe the way edges mix in such networks can help to both assess and explain cohesion among a community of nodes.

We addressed this issue by looking at how well *catalysts* mix within a group. This is accomplished using the *entanglement index* computed for each catalyst, measuring how intensely and how homogeneously a catalyst co-occurs with all other catalysts in a group of substrates. Edge entanglement, at the group level, is then computed from the individual catalyst entanglement indices. For instance, optimal cohesion is reached whenever catalysts have equal entanglement indices. This is the case when all substrates have the exact same associated catalysts, and that all catalysts equally co-occur within substrates.

3.1.1 Ambiguity in social networks analysis

Social networks analysis is a discipline that studies social relationships using graph models and network theory, both offering very

useful methodologies in the study of human systems and social behaviour. Often, in sociology, groups of people and their relational ties are the baseline for observation before modelling. When collected by hand, through interviews with people, the data is sometimes "messy", and certain facts are sometimes described quite differently, although they could refer to the same concept. Given a group of people, people can claim many different types of relationships: *co-worker, classmate, lover, friend, sports-mate, employee, boss* and so on. Some of them are quite ambiguous: when we consider *employee*, *superior*, *co-worker* we can clearly see that these relationships are very similar, and what means *employee* to one, might mean *boss* or *co-worker* to another.

This ambiguity between relationships brings noise (meaning interference) into the analysis of social networks, especially when this ambiguity occurs across different domains (*e.g.* professional and private). Burt and Schøtt (1985) tackled this issue by looking at a way to solve the ambiguity, or at least measure it, in order to understand how relationships coincide, so that they could assess the network for higher level analysis. Addressing this matter, Burt and Schøtt firstly wondered about the distance between these relationships: is it possible to create classes of equivalent relationships? The recursive nature of ambiguity also caught their attention: an ambiguous relationship is even more ambiguous when its related people also share other ambiguous relationships.

In terms of graphs, people are defined as nodes, and relationships as edges. This gives us a setting of multiple edges – of different types – across the same pair of nodes: it is precisely here where ambiguity occurs. Studying ambiguity Burt and Schøtt wished to find patterns of coinciding relationships over a given network, in other words, socio-metric questions on ambiguous relationships.

3.1.2 Entanglement of a graph

We have seen in Section 2.4.2 that multiplex graphs can be derived from any bipartite structures such as the documents from INA. We were really interested in working around the notion of ambiguity. Indeed, relationships can be intuitively transposed to terms between documents. Semantics can be ambiguous in themsevles, but the documents from INA are annotated with a thesaurus that is designed to avoid semantic ambiguity. Ambiguity (as defined by Burt and Schøtt) between terms measures the entanglement of those terms: coincidence between relationships becomes co-occurrence between terms.

Terms that are not ambiguous by their definition (let's say *Europe* and *elections*) can be measured as ambiguous in a given context (such as *the European elections*). Hence the more terms that are ambiguous between documents, the more numerous the documents concerning the same subject matter. Hence we can directly transpose the notion of ambiguity to entanglement of terms within document proximity.

Entanglement is not limited to documents, but actually to any multiplex network (representing *substrate* interactions through *catalysts*).

The interaction between catalysts across substrates' edges measures the entanglement. This can be confusing: we do not study co-occurrence of catalysts within substrates, but in the case of documents, within pairs of substrates. This subtle difference firstly reinforces the co-occurrence relationship – it has to happen between at least a pair of substrates. Counting the edge occurrences allows freedom in building the substrate graph, as we will see in Section 3.4.3, and as illustrated by the different cases of Chapter 6.

3.1.3 Other "entanglement" measures from the literature

Etymologically, the term *"entanglement"* comes from the verb *"tangle"*, and designates the state of some intertwined elements that are mixed together (such as *tangled hair*). It is precisely from this analogy, with hair, that we decided to name our analysis, the *entanglement analysis*. However, *entanglement* has already been used in previous literature to describe complex but unrelated concepts.

In physics, *entanglement* corresponds to a physical phenomenon that occurs at the quantum level. If we want to briefly describe this phenomenon, it refers to the interdependence of the quantum state of particles of a same quantum system[1]. Graphs are widely used in physics since they are useful models of complex interactions. Euler himself was also a physicist, though he certainly had no knowledge of what quantum entanglement could be. It is not surprising to find that studies of entanglement in quantum systems are modelled with graphs (Hein et al., 2004; Lu et al., 2007). In such systems, graph states are multi-particle entanglement states, where nodes represent the quantum spin of particles, and edges a certain type of interaction. Even if the appellation *"multi-party entanglement"* feels close, other interests are at stake here, for *quantum entanglement* is measured differently from a single-mode graph with specific characteristics.

Entanglement has also been proposed as a measure in directed single-mode graphs (then extended to undirected graphs) in mathematics, especially *fixed point theory*[2] (Berwanger and Grädel, 2005; Belkhir and Santocanale, 2007). It has interesting applications in game theory[3], as when modelling the game of *thieves and policemen*. A thief stands at a position $u \in V$ in a graph $G = (V, E)$ and policemen are outside the graph; at every turn, the policemen can enter the graph, keep their position, move to an adjacent node, which could be the thief's position; at every turn, the thief knows the position of the policemen, and the thief has to move to an adjacent free position if the policeman has moved the thief's position. The policemen win when there are no positions left for the thief; the *entanglement* k is then the minimal number of policeman that are needed to catch the thief. In this case, the *entanglement* refers again to a very different concept from the *multiplex entanglement*.

1. Our introduction to quantum physics will not go further, but original mentions of quantum entanglement can be found in (Einstein et al., 1935; Schrödinger, 1935).

2. Fixed point theory is a branch of mathematics studying... fixed points in function applications, and their properties (Dugundji and Granas, 1982).

3. Game theory (founded by John von Neumann (Neumann, 1928)) can be quickly defined as the mathematical study of strategic decision making. The many applications of game theory (economics, political science, biology, logic, or psychology) are quite popular, such as *the prisoner's dilemma*, and were even depicted in Ron Howard's 2001 Hollywood movie *"A beautiful mind"* illustrating the life of the mathematician John Nash.

3.1.4 Computing the entanglement index

Given a multiplex graph of $G = (\mathcal{A}, \sum_{t \in \mathcal{C}} E_t)$, where nodes are substrate and edges catalysts (social relationships, terms...). Catalysts can be of type t *or* $t', \in \mathcal{C}$. $\mathcal{C}$ is the set of all existing catalysts, allowing multiple edges across nodes in the graph. $G_{\mathcal{A}} = (V_{\mathcal{A}}, E_{\mathcal{A}})$ is then the "flattened" projection of G, with no multiple edges.

In order to measure interaction within G, we are interested in counting how often catalysts co-occur between substrates in the considered group. We define $n_{t,t}$ to count the number of edges in $E_{\mathcal{A}}$ carrying the catalyst t, and $n_{t,t'}$ the number of edges carrying both catalysts t and t'. In other words, $n_{t,t}$ counts the number of occurrences of t in $E_{\mathcal{A}}$ and $n_{t,t'}$ counts the number of co-occurrence of t and t' among $E_{\mathcal{A}}$.

The interaction matrix $N_{\mathcal{C}}$ which collects all these $n_{t,t'}$, is nonnegative and symmetric. Note this matrix can be seen as an adjacency matrix giving rise to the catalyst interaction graph $G_{\mathcal{C}}$, a central artifact for our visual analytics approach detailed in Section 3.2.2. Note that when the matrix $N_{\mathcal{C}}$ is considered reducible (and displays a block structure) the induced graph $G_{\mathcal{C}}$ is not connected; conversely, when $G_{\mathcal{C}}$ is connected, the matrix $N_{\mathcal{C}}$ is *irreducible*.

Now that we have defined the interactions, we are interested in observing how often they occur throughout our graph. We can easily compute the occurrence frequency of a catalyst among the edges as

$$c_{t,t} = \frac{n_{t,t}}{|E_{\mathcal{A}}|} \tag{3.1}$$

i.e. the proportion of edges carrying catalyst t among all $|E_{\mathcal{A}}|$ edges in $G_{\mathcal{A}} = (\mathcal{A}, E_{\mathcal{A}})$. Conditional frequency of co-occurrence of catalysts also interests us, and we define it as the following ratio:

$$c_{t,t'} = n_{t,t'} / n_{t,t} \tag{3.2}$$

reflecting the proportions of edges *given* catalyst t, that are also bearing catalyst t'. Differences between $c_{t,t'}$ and $c_{t',t}$ reflect the reciprocal nature of the interaction between t and t'

The ratios $c_{t,t'}$ can now be used to fill an interaction frequency matrix $C_{\mathcal{C}}$, which is also filled with nonnegative values but is *not* symmetric. Note that $C_{\mathcal{C}}$ though having a probabilistic interpretation of its entries, is *not* a stochastic matrix[4].

4. Stochastic matrices, are a class of non-negative square matrices with entries that represent probabilities, and sum to 1 in rows. Stochastic matrices support transitions in a Markov chain (Hastings, 1970), a common model for memory-less random processes.

The entanglement index for each attribute measures how much a catalyst t contributes to the overall cohesion of a substrate group, and we need to measure this, with λ denoting the maximum value among entanglement indices λ_t of attributes $t \in \mathcal{C}$. In other words, the entanglement index of attribute t is a fraction of λ, namely $\lambda_t = \gamma_t \cdot \lambda$.

The entanglement value of a catalyst t is reinforced through interactions with other highly entangled catalysts. With a probabilistic

interpretation of the matrix entries $c_{t,t'}$ in mind, we are able to postulate the following equation which defines the values γ_t.

$$\gamma_t \cdot \lambda = \sum_{t' \in T} c_{t',t} \gamma_{t'} \tag{3.3}$$

The first clause of the Perron-Frobenius theorem stipulates (from Ding and Zhou (2009) Chapter 2 Theorem 2.6):

> *"Let A a nonnegative irreducible square matrix,*
>
> 1. *The spectral radius $r(A)$ of A is a positive eigenvalue with a positive eigenvector x and a positive left eigenvector y*
> 2. *The eigenvalue $r(A)$ is geometrically simple*
> 3. *Any nonnegative eigenvector of A is a positive scaler multiple of x corresponding to $r(A)$*
> 4. *The maximal eigenvalue $r(A)$ is algebraically simple"*

The corollary this would be (from Ding and Zhou (2009, Chapter 2 Corollary 2.6)):

> *"Let A be a nonnegative square matrix.*
> *There exists a unique vector x such that $Ax = r(A)x$, $x > 0$,*
> *and $e^T x = 1$*
> *and there exists a unique vector y such that $y^T A = r(A)y^T$,*
> *$y > 0$, and $y^T e = 1$*
> *with $r(A)$ the spectral radius of A, and e such as $AA^{-1} = e$."*

For a such nonnegative matrix, there exists therefore one unique positive maximal eigenvalue which happens to be its spectral radius, associated to a nonnegative eigenvector.

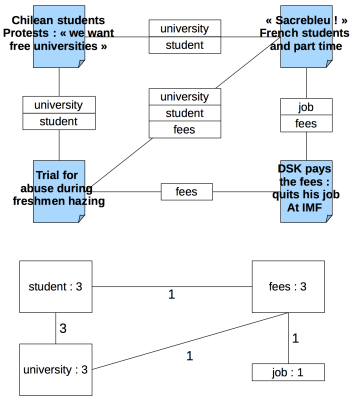
The Perron-Frobenius theory holds for irreducible matrices, that is when the graph $G_{\mathcal{C}}$ is connected. Hence, the connected components in $G_{\mathcal{C}} = (\mathcal{C}, E_{\mathcal{C}})$ (or irreducible blocks in $C_{\mathcal{C}}$) must be inspected independently. We hereafter assume that $C_{\mathcal{C}}$ is irreducible.

Eq. (3.3) gives rise to the matrix equation $\gamma \cdot \lambda = C'_{\mathcal{C}} \cdot \gamma$. From the preceding Corollary 2.6, the solution of this equation will be the *right* eigenvector of the transposed matrix $C'_{\mathcal{C}}$, associated to the its spectral radius, i.e. its maximum eigenvalue. The solution $\gamma = (\gamma_t)_{t \in T}$ collects values for all catalysts t.

The maximum entanglement index thus equals the unique maximal eigenvalue λ of matrix $C'_{\mathcal{C}}$.

The actual entanglement index values λ_t are of lesser interest, we are actually interested in the relative γ_t values. We call γ_t, the *entanglement index* of a catalyst t.

5. Network diagrams (c) and (d) from Figure 2.13):



Considering the example in Figure 2.13 (replicated here[5] for convenience). A multiplex graph is obtained by inducing edges between documents labelled with terms (Figure 2.13 (c)). The resulting attribute interaction network directly links the terms is shown in Figure 2.13 (d). The matrices $N_{\mathcal{C}}$ and $C_{\mathcal{C}}$ (built over the catalysts *student*, *fees*, *university* and *job*) then read:

$$N_{\mathcal{C}} = \begin{bmatrix} 3 & 3 & 1 & 0 \\ 3 & 3 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$C_{\mathcal{C}} = \begin{bmatrix} \frac{3}{4} & 1 & \frac{1}{2} & 0 \\ 1 & \frac{3}{4} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{2} & 1 \\ 0 & 0 & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

Hence the entanglement indices for these catalysts are:

$$\gamma = \begin{bmatrix} 0.66, & 0.66, & 0.35, & 0.10 \end{bmatrix}$$

Notice that two indices are equal, and correspond to catalysts *student* and *university*.

3.1.5 Complexity

The construction of the interaction matrix $N_{|\mathcal{C}|\times|\mathcal{C}|}$ is straightforward and can be achieved in $O(|E_{\mathcal{A}}||\mathcal{C}|)$ (in the worst case scenario, every catalyst appears on every substrate edge).

The construction of the conditional frequency matrix $C_{|\mathcal{C}|\times|\mathcal{C}|}$ can then be done in $O(\frac{1}{2}|\mathcal{C}|(|\mathcal{C}|-1))$ (since $N_{\mathcal{C}}$ is symmetric).

Entanglement indices, and hereafter *entanglement intensity* and *homogeneity*, are then directly computed from the Eigen analysis of C. Although eigenvalues are computed in $O(|\mathcal{C}|^3)$ (Demmel et al., 2007), further work might prove that approximations may be expected to converge faster as it was seen in PageRank (Page et al., 1999).

The whole calculation should hence take an overall complexity of $O(|\mathcal{C}|(|E_{\mathcal{A}}| + \frac{|\mathcal{C}|-1}{2} + |\mathcal{C}|^2))$.

3.2 Measuring the entanglement of a graph

Now we have a tool to measure entanglement at a catalyst level, but what happens at a group level? We will introduce two group measures of the entanglement of a group, briefly followed by the term interaction network as an analytical tool to study interactions between catalysts.

3.2.1 *Entanglement intensity and homogeneity*

This section introduces *entanglement intensity* and *entanglement homogeneity* as group network measures. The focus here is on interactions among catalysts, and aims to reveal how cohesive the group of substrates is considering this set of catalysts.

It is straightforward to interpret the entanglement index of a catalyst t as its participation in a hypothetical group entanglement, and to infer that in a case within even participation of all catalysts to this group entanglement, each of the catalyst would have an equal index. Hence the eigenvector that gives us the entanglement indices also offers us some information about the balance of catalyst interaction.

The associated eigenvalue is also of interest and, as we will shortly see, it can easily be normalized. It also has a geometrical interpretation as a scale factor in the transformation from the space described by the matrix $C_{\mathcal{C}}$ to the "eigenspace of entanglement". In this "eigenspace of entanglement", it is the principal eigenvector that describes how each catalyst contributes relative to the entanglement. The principal eigenvalue tells us how much of entanglement space is occupied by our catalysts. This is possible thanks to the probabilistic interpretation of the matrix $C_{\mathcal{C}}$ since every entry of the matrix admits a maximal (and minimal) limit, and since linear algebra gives us theoretical boundaries for the maximal eigenvalue, which depends on the matrix' entries (Ding and Zhou, 2009). We will now present intuitions and theoretical foundations that allow us to define the *entanglement homogeneity* $\mathcal{H}_{\mathcal{G}}$, and *the entanglement intensity* $\mathcal{I}_{\mathcal{G}}$.

Given our definition of entanglement, a catalyst may rarely co-occur on the graph's edges and still be very tangled if it constantly mixes with other tangled catalysts. This is especially relevant for a subgroup of catalysts that always co-occur together and only together, thus giving the highest entanglement values. Interestingly this is the pattern we are looking for when questioning cohesion in a group of documents (substrates). As we assume that an optimally cohesive group of documents will display exactly the same set of terms (catalysts). Hence, those catalysts, taken together, should display the highest entanglement indices relative to their corresponding subset of documents. As we are only measuring the *relative* entanglement index, an optimal case displays equal entanglement index for every term.

More formally, in such a situation, $N_{\mathcal{C}}$ and all entries coincide with $n_{i,j} = k, \forall\,(i,j) \in \mathcal{C}$. Hence $C_{\mathcal{C}}$ displays $c_{i,j} = 1$ by construction, and entries of the principal eigenvector coincide as well: $\gamma_i = \frac{1}{\sqrt{|\mathcal{C}|}}$ after normalization.

With this intuition in mind, we shall now question the observed graph in comparison to an optimal one with those values.

Having calculated the spread of the catalysts' interactions, or the intensity of their interactions, we will compare the spread of our catalysts to that of an even case. We then question how close our catalysts'

interaction is to that of a full interaction?

With this intuition in mind, we shall now question the observed graph in comparison to an optimal one with those values. We propose then to answer the questions *How even is the catalyst interaction? How intense is the catalyst interaction?* by comparison to our optimal case. And the shape of our γ vector and λ value are good indicators.

We define now the *entanglement homogeneity*

$$\mathcal{H}_{\mathcal{G}} = |\langle \gamma, 1 \rangle| \tag{3.4}$$

Since we have stated that in the optimal case, all the entanglement indices γ_t coincide (*i.e.* the catalysts offer all an equal share of the overall entanglement), the homogeneity $\mathcal{H}_{\mathcal{G}}$ will be measured to 1

The cosine brings a natural interpretation of the measure, but other measures, including Shannon's entropy (Shannon, 1948) and Guimera's participation coefficient (Guimera et al., 2005), offer interesting alternatives to cosine similarity.

An even spread among catalysts does not explain the inner structure of intertwining edges, moreover many equivalent situations can occur as discussed in the next Section, and the measure of intensity can help us differentiate between those situations.

Bearing that we are looking for the optimal case, ($C_{\mathcal{C}}$ filled with 1), we have yet to exploit the principal eigenvalue. Luckily, another corollary of Perron-Frobenius' theorem of non-negative matrices (Ding and Zhou, 2009, Chapter 2 Corollary 2.4) gives us very convenient boundaries for the maximal eigenvalue of such a matrix.

"Let $A_{n\times n}$ be a nonnegative square matrix, then:

$$\min_{1\leq i\leq n} \sum_{j=1}^{n} a_{i,j} \leq \lambda \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} a_{i,j}$$

and

$$\min_{1\leq j\leq n} \sum_{i=1}^{n} a_{i,j} \leq \lambda \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} a_{i,j}$$

"

with $a_{i,j}$ an entry of a nonnegative matrix A. In plain English this means that λ is bounded between the minimum and maximum sum of the matrix columns (or lines). In our case, thanks to the probabilistic definition of our matrix, $a_{i,j} = c_{i,j} \leq 1$, minimum and maximum would then be equal to $1 * |\mathcal{C}|$, the number of catalysts in the best case scenario. This ratio thus provides a measure for entanglement *intensity* among all substrates *with respect to catalysts in $\mathcal{C}$*, allowing comparison between graphs with different numbers of catalysts.

We define *entanglement intensity*

$$\mathcal{I}_{\mathcal{G}} = \frac{\lambda}{|\mathcal{C}|} \in [0, 1] \tag{3.5}$$

The entanglement intensity $\mathcal{I}_{\mathcal{G}}$ is then bound to the maximum eigenvalue of our interaction matrix $C_{\mathcal{C}}$, and captures the amount of

interaction in our graph. Normalized, intensity $\mathcal{I}_\mathcal{G}$ gets closer to 1 as the amount of interaction in the graph gets closer to its maximum possible.

Recalling the example in Section 3.1.4, we measure $\mathcal{I}_\mathcal{G} = 0.505$ and $\mathcal{H}_\mathcal{G} = 0.885$.

3.2.2 The catalyst interaction graph

The matrices $N_\mathcal{C}$ and $C_\mathcal{C}$ appear very close to common adjacency matrices describing graphs. Let us see how we can use this artifact.

We have seen in Section 2.4.2 that behind the model of a complex system we can set up a bipartite graph. The measures we have previously introduced in this section consider a projection from this bipartite graph on one side: substrates. However in a substrate-catalyst bipartite graph $G^*_{\mathcal{AB}} = (\mathcal{A} + \mathcal{B}, E)$, this projection has a corresponding projection on the other side: the catalysts related to substrates $G_\mathcal{B}$. The entities taken into account by matrices $N_\mathcal{C}$ and $C_\mathcal{C}$ actually correspond to a subset of this corresponding projection ($G_\mathcal{C} \subset G_\mathcal{B}$). It is a graph with *almost* the same set of nodes and edges as the projection onto catalysts. Only catalysts involved in multiple content relationships in $G_\mathcal{A}$ are taken into account.

Each matrix $N_\mathcal{C}$ and $C_\mathcal{C}$ describe a different graph, with different relationships between the same set of catalysts. $N_\mathcal{C}$ presents *undirected* relationships between co-occurring catalysts, with edge weights that are a co-occurrence quantity relative to the whole set of edges $E(G_\mathcal{A})$ in the substrate graph – and loops correspond similarly to occurrences.

$C_\mathcal{C}$ describes a graph with *directed* edges between catalysts, however each edge is doubly directed with different nonnegative weights. Those weights can be interpreted as a conditional probability of two catalysts co-occurring *given* one of them. Loops are weighted with the frequency of occurrence of a catalyst onto $E\left(G_\mathcal{A}\right)$.

One could argue that the methodology we have presented for computing entanglement indices is very similar to that of PageRank, Eigen centrality, a random walk, or any slightly equivalent approach. It is indeed a recursive definition that solves similarly to a Markov chain with a spectral analysis of a square matrix, to which corresponds a transition graph. The catalyst interaction graph would then correspond to this transition graph, with *possibilities* of transition (a *probability* of transition would imply that all entries of a line or column would sum to one). However constructing this graph is not trivial. Even if the two adjacency matrices $N_\mathcal{C}$ and $C_\mathcal{C}$ describe different graphs, the undirected simple graph that covers them (*i.e.* a graph that covers all nodes and presents no loop or multiple edges) is exactly the same, and itself represents an interesting object of study. We call this graph the *catalyst interaction graph*.

The inspection of a group of substrates and associated catalysts raises several questions. For instance, it might be important to know

whether catalysts homogeneously map to *all* substrates in the group. Conversely, a *misleading transitivity effect* might be suspected. Indeed, we may have catalysts t, t' co-occurring between substrates a and a', and catalysts t', t'' co-occurring between substrates a' and a'', may lead one to believe that catalysts t, t', t'' co-occur simultaneously between all three substrates a, a', a''.

The basic shape of this graph, its topology, is a very good indicator, explaining everything that ties substrates together. Studies of connected components, degree, centrality, neighbourhood, and communities, lead to the understanding of the structure that ties substrates together. Each of these measure has an interpretation that depends on the nature of the catalysts and the network of substrates. (For example, if catalysts are terms of returned documents by a query, the most central catalysts most often corresponds to the terms of the request). The graph is a very tangible object, offering diagrammatic interpretation and support to understanding, but above all, it is a central artifact in our visual analytics approach (as we will present in Chapter 5, and especially in the example of INA's documents in Section 6.1.6).

3.3 Behaviour of the measures

Now that we know how to compute those measure, we shall examine their application.

3.3.1 Comparison of the entanglement index with other known measures

All that we obtain from our definition of the entanglement index is a *relative* value for each catalyst. The actual entanglement values are of lesser interest since we are mainly interested in comparing catalysts with one another, knowing which ones are dominant in tangling with others. Given this information, we might suspect the most (co-)occurring catalysts to be the most tangled, and this would be a logical observation since a catalyst needs to appear (occur) often to have a better chance of tangling (co-occurring) with others. But this notion of entanglement is much finer than that, and 3.1 shows how different it can be. The entanglement index gives a sense on how a catalyst mixes with other catalysts. In that sense, a group of rarely occurring catalysts which always co-occur with one another will show high entanglement indices (lower part of the catalysts shown in Figure 3.1).

We can also compare the entanglement index with other network metrics commonly used, in a catalyst interaction network. We also compared other measures taken from the same example represented in Figure 3.1 (catalyst occurrence, degree, betweenness centrality and PageRank, see in Figure 3.2).

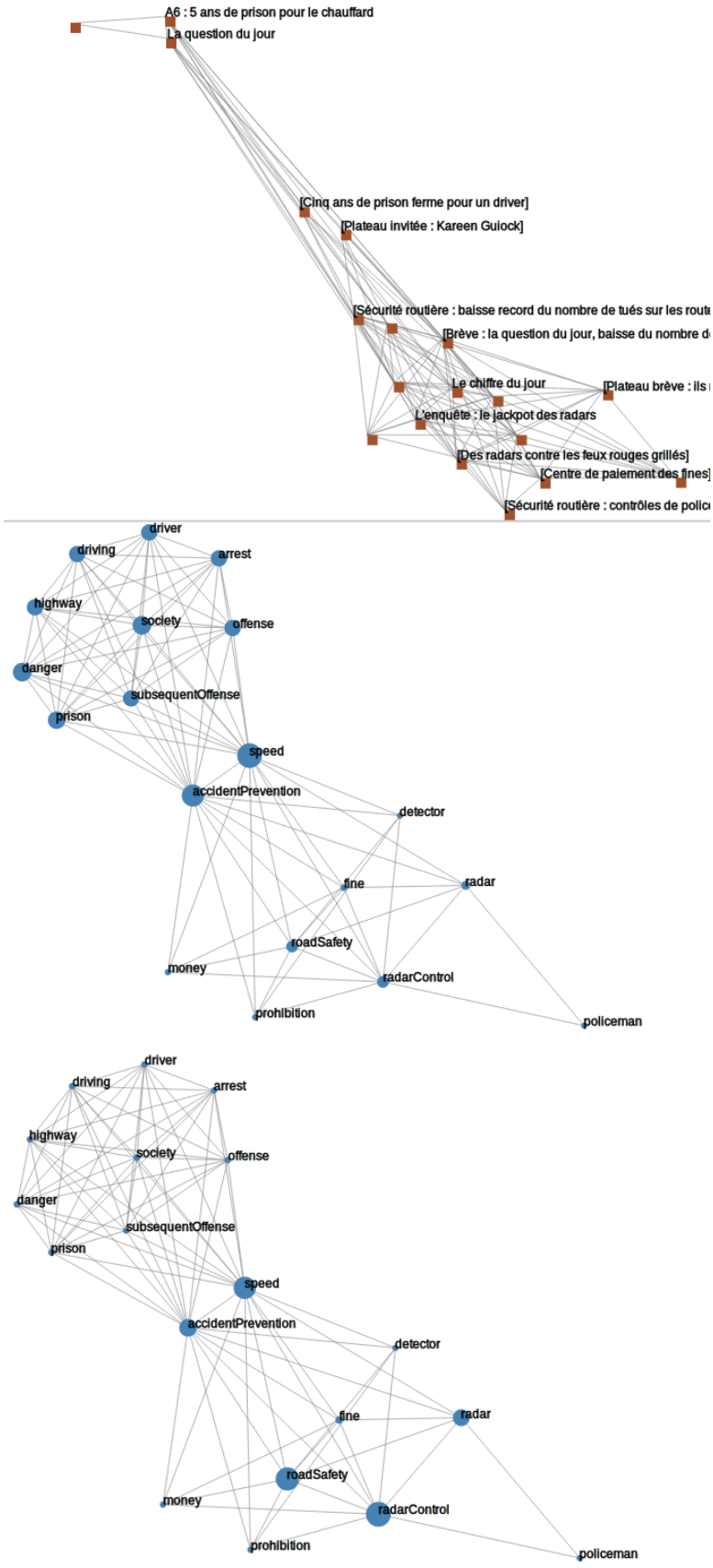


Figure 3.1: *Road safety* documents (top), the catalyst interaction network, with node size mapped to entanglement index (middle) and number of co-occurrences (bottom). The difference in size clearly shows that these two statistics act complementarily. 19 terms are co-occurring across 18 documents.

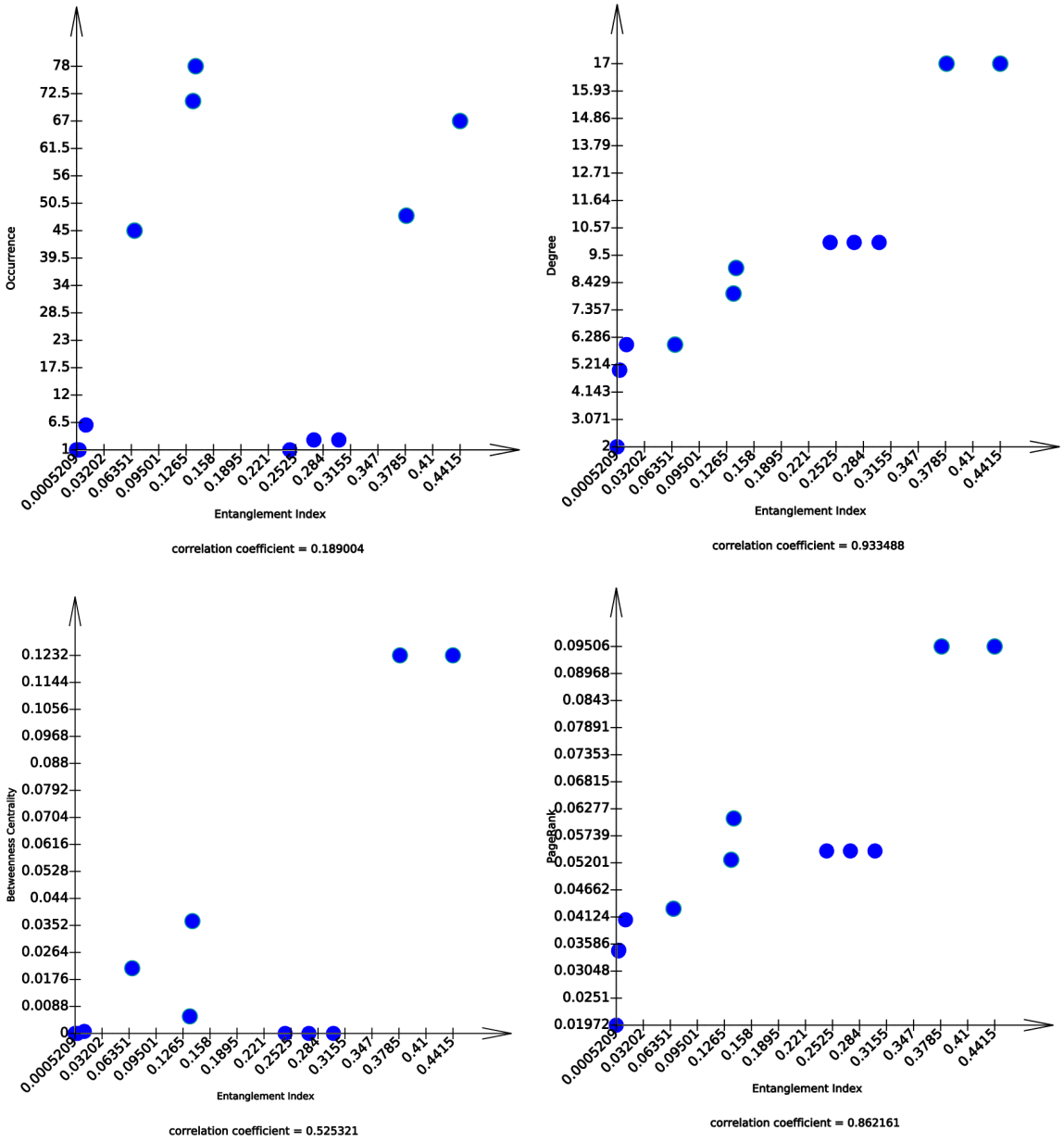


Figure 3.2: Comparison from the example network of Figure 3.1 between the entanglement index with catalyst occurrence on substrate edges, degree in the catalyst interaction graph, betweenness centrality, and PageRank. Although we can see some level of correlation between these measure, the entanglement index captures a fine difference. The case of application in a document–term network certainly induces a bias, and further study on random graphs would confirm such difference.

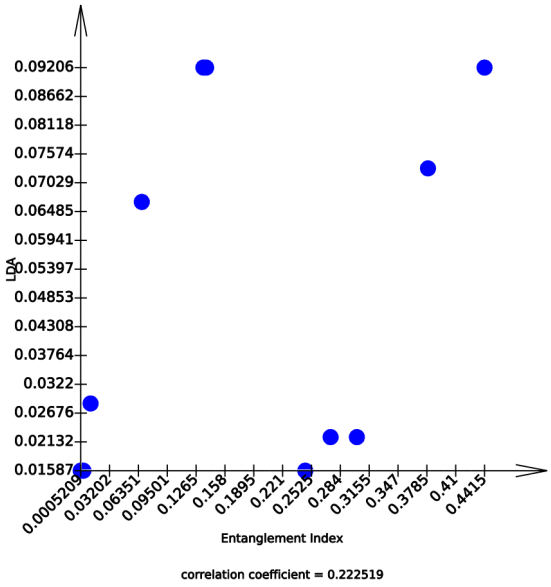


Figure 3.3: Comparison from the example network of Figure 3.1 between the entanglement index and the probability of each term to belong to one topic of LDA. This example clearly shows a difference between the different measures.

Even with a small number of observations, Figure 3.1 shows some level of correlation between the entanglement index and measures in the catalyst interaction graph. However we are not only proposing the entanglement index, but a whole methodology starting from a multiplex view of a complex network.

We might also strongly suspect some bias induced by the nature of the data itself[6]. Since terms are chosen from a thesaurus that can be modelled with a tree, high-level general terms tend to occur more often than low-level specific terms, resulting in higher degree and betweenness centrality for the terms in their interaction graph.

6. We need to first study correlations of the entanglement index on random graphs, and then on other application contexts, to confirm the bias.

However, LDA also offers us other challenges, in terms of choosing the right number of topics, even though we have new methodologies to find the natural number of topics, they rely on big datasets, for which a small number of topics and detailed analysis are irrelevant since the "document" population would be too small for reliable statistics. However, the shape of the catalyst interaction network offers an idea of the number of topics one can observe.

If we keep following the same example, asking LDA for only 1 topic as results show in Figure 3.3, the two measures do not return the same results – but the comparison itself feels quite inappropriate. However, from the shape of the catalyst interaction network (as shown in Figure 3.1), we can suspect 2 different topics are at play in this example. We observe the topic distribution in Table 3.1 asking for two topics.

From this example, we can see that the top 5 terms in both topics are well matched with the different communities we can observe from the catalyst interaction network in Figure 3.1. The following terms are however quite mixed among both communities with equivalent

Topic 1		*Topic 2*	
Term	**P(t)**	**Term**	**P(t)**
vitesse	0.056	securiteroutiere	0.121
preventiondesaccidents	0.056	controleradar	0.121
danger	0.043	vitesse	0.101
societe	0.043	radar	0.101
prison	0.043	preventiondesaccidents	0.072
conduite	0.031	amende	0.034
autoroute	0.031	detecteur	0.024
recidivejustice	0.031	gendarme	0.024
infraction	0.031	impot	0.014
automobiliste	0.031	valdoise	0.014
arrestation	0.031	interdiction	0.014
argent	0.031	fillonfrancois	0.014
securiteroutiere	0.031	sondagedopinion	0.014
circulationroutiere	0.031	contravention	0.014
controleradar	0.031	codedelaroute	0.014
alencon	0.019	policier	0.014
prevision	0.019	photographie	0.014
policedelaroute	0.019	paiement	0.014
gendarmerie	0.019	manchedepartement	0.014
vacances	0.019	volinfraction	0.014

Table 3.1: Two topics identified with LDA and their respective distribution of terms.

probabilities. This might be due to the limited amount of redundant information (19 terms co-occurring across 18 documents).

We can think of our methodology as a process that goes right after an application of LDA-based clustering, and ranking of entanglement indices among communities formed in the catalyst-interaction network are consistent with the highest topic probabilities returned by LDA. Nonetheless with the lowest probable catalysts in small topics, LDA can offer some level of noise, but then a threshold is needed to determine whether a catalyst is relevant or not.

Finally, the way we set edges between substrates, in combinations with the ratio and variety of catalysts, also has a certain influence on the differences between those metrics. When substrate proximity is based on catalyst co-occurrence, it is natural to find a level of correlation with all occurrence-based measures. However this index becomes especially relevant as computed on-the-fly, in the midst of study of homophily relationships in a network observed from another type of relationships, such as the co-authoring network presented in Section 6.2.2.

Section 3.5.2 offers further investigation into comparisons of the entanglement intensity and homogeneity.

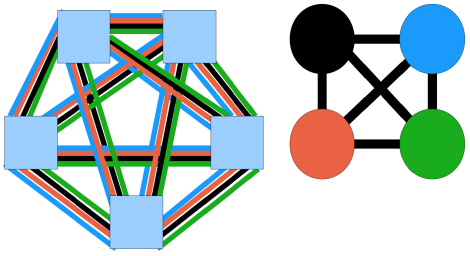
3.3.2 *Different topologies, different measures*

"Group" measures, intensity and homogeneity, capture different aspects of the network at a group level.

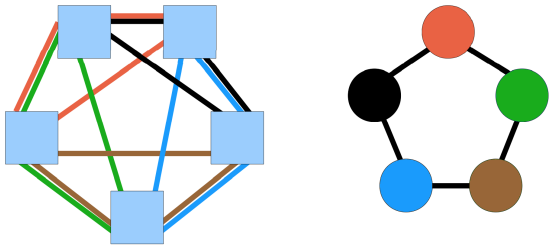
Now, more practically, we need to consider what homogeneity and intensity are measuring in a multiple network. Let us consider a multiple network of substrates. The substrate network connectedness has no influence on the intensity and homogeneity measures. Intensity and homogeneity consider only the edges of the network, no matter which substrate they bind. The edges can support any possible combinations of catalysts. The number of catalysts to be considered is another influential factor in the measure of entanglement. Given these catalysts and substrate edges, homogeneity and intensity are two measures of the possible combination of catalysts over the edges.

Intensity gives a sense of how much catalyst interaction occurs over the edges, and how much more might occur. We can see it as a degree of freedom, the lower the intensity, the more space there is, and the greater the likelihood of more catalysts on the edges. When intensity is at maximum, there is no space on the edges for any additional catalysts, as the edges are already saturated with catalysts.

Note that entanglement only takes into account interaction over substrate edges. Hence the topology of the substrate interaction network has no theoretical influence on the entanglement measure. However, observations on real-world networks suggest these two parameters are not totally independent.



$$\gamma = \begin{bmatrix} 0.5, & 0.5, & 0.5, & 0.5 \end{bmatrix}, \mathcal{H}_{\mathcal{G}} = 1, \mathcal{I}_{\mathcal{G}} = 1$$



$$\gamma = \begin{bmatrix} 0.45, & 0.45, & 0.45, & 0.45, & 0.45 \end{bmatrix}, \mathcal{H}_{\mathcal{G}} = 1, \mathcal{I}_{\mathcal{G}} = 0.193$$

Figure 3.4: These two patterns of interaction show maximal homogeneity with different intensity. Both intensity values and the shape of the catalyst interaction graphs allow to understand and distinguish those cases.

Homogeneity does not focus on substrate edges, but on catalysts interactions. Homogeneity exists if catalyst interaction is even from

a one catalyst type to any other. If every catalyst shares exactly *the same number* of interactions with *the same number* of different catalysts as every other catalyst of the substrate edges, wherever this interaction takes place, it does not matter, as then it will be optimal. If interactions are randomly distributed on the substrate edges among a random number of catalysts, the homogeneity will most probably be measured as low. As a consequence, several different topologies are identifiable as optimally homogeneous, with different measures of intensity, as Figure 3.4 shows.

Although homogeneity and intensity do not vary independently, we can plot those values onto a 2D space, where homogeneity is plotted along the x-axis and intensity is plotted along the y-axis (Figure 3.5). Any term interaction network would then be plotted as a 2D point $(x, y) = (\frac{\langle 1_T, \gamma \rangle}{\|1_T\| \|\gamma\|}, \frac{\lambda}{|F|})$ in the plane that could be divided into rough areas of which extremes correspond to typical network profiles.

Figure 3.5 displays different values of homogeneity and intensity measures in real and artificial cases. There is an obvious dependency between intensity and homogeneity: high intensity cannot be achieved without some amount of homogeneity, explaining the empty top left part of the plane. Indeed, in the case of high intensity, it is less likely to find many different possibilities of catalyst organizations among edges, and thus our chances of observing optimal homogeneity are very high.

We can however observe typical profiles in this space. Here we study this space in general terms of graphs, and we will recall these

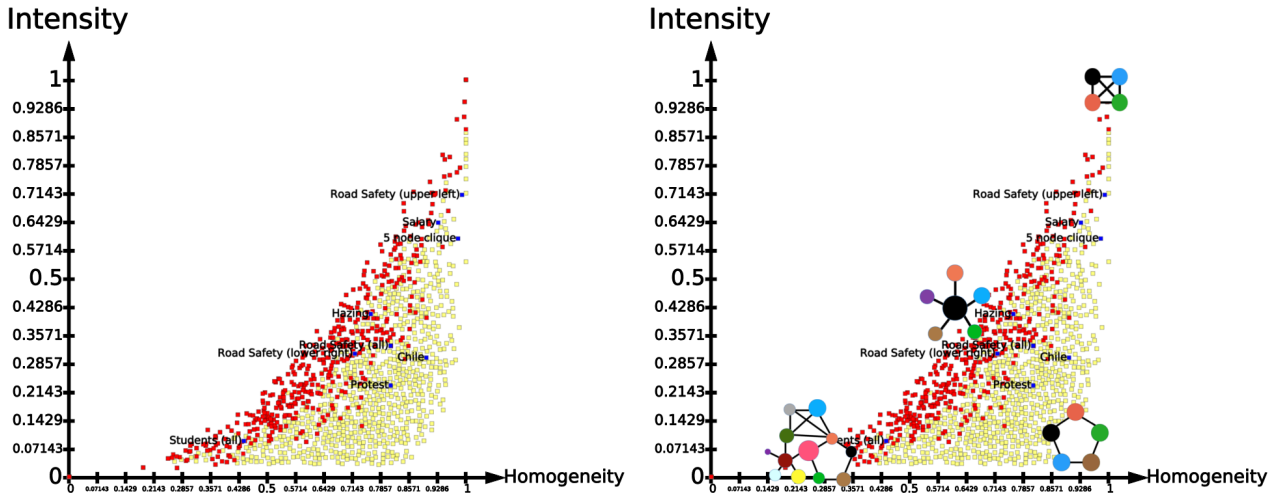


Figure 3.5: Entanglement profiles can roughly be categorized by combining intensity and homogeneity.
Left: Measures on INA's document groups are in red. Randomly generated multiplex graphs are in yellow. Examples of Chapter 6 are in blue with labels.
Right: Different regions can be characterized with typical catalyst interaction shapes. The top right area tends to correspond to the optimal case, a balanced clique. On the bottom right area, we can find balanced cycles. The bottom left area is an unbalanced mess where finding patterns can be difficult. Along the intensity/homogeneity limit, hierarchies across catalysts appear.

profiles in Section 6.1.4 and try to interpret them in terms of document groups.

- The equally weighted catalyst clique was presented as the archetype of an optimal interaction network located at the top rightmost position $(x, y) = (1, 1)$ in the plot. The top-right area thus collects these relatively dense and evenly interacting networks.
- The lower-left case gathers networks with low intensity and low homogeneity. This is a rather common case, usually gathering a lot of substrates and catalysts with loose interactions. This is a situation where many catalysts could appear as satellites of more central catalysts. A typical catalyst network would have a low density (few edges) and a random link structure, leading to a sparse $N_{\mathcal{C}}$ matrix with ϵ entries.
- The lowest-right area corresponds to relatively high homogeneity and lower intensity: catalysts almost all interact with one another, but not as much as the substrate graph $G_{\mathcal{A}}$ theoretically allows. $N_{\mathcal{C}}$ matrices are non-sparse, and they have large diagonal but rather low off-diagonal entries.
- The area further left is tricky (particularly the central part, at the limit). This case occurs when catalysts are nested. This situation translates into consecutive inclusion of catalysts among substrate edges (*i.e.*, edges in $G_{\mathcal{A}}$ associated with catalyst t include all edges associated with catalyst t' plus some other edges).

3.4 The difficulties in measuring entanglement

Entanglement is a complex concept and some precautions need to be taken into account when computing it.

3.4.1 Granularity and dimensionality curse

Entanglement measures are built around two types of information: the available substrate edge space, and the number of different catalysts. For a given number of catalysts, available substrate edge space defines only the level of granularity at which you can observe catalyst interaction. The more edge space there is, the more we can measure catalyst interaction at a finer level. On the contrary, very small substrate networks, with only one or two edges, return very often near-optimal values, thus they are not very relevant for most analysis.

The number of catalysts itself reflects the level of complexity between interactions. With a low intensity, the more catalysts you have (given a large enough edge space), the more difficult homogeneity will be to interpret. One may indeed find a very low entanglement index for most of the catalysts with one or a few catalysts having a high

entanglement index. Such pattern, with balanced low interaction, will lead to a cosine measure (given the definition of homogeneity) that is close to one. The pattern of catalyst interaction in such cases is indeed quite homogeneous, but homogeneity will not, of itself, detect that one or few catalysts are dragging all the interactions. The dimensionality curse can occur in homogeneity measurement giving rise to high values that need some care in their interpretation.

A possible solution is to observe such catalysts entanglement indices with an histogram, or more directly through the shape of the catalyst interaction network.

3.4.2 *What to do when faced with the multiple connected components of catalysts?*

With multiple connected components, the interaction matrix N can be organized in blocks. In each block, catalysts interact only together but *not* with catalysts from other blocks – given the original substrate network. Therefore, we can identify in the original substrate network a sub-network (from a subset of edges) that concerns only the catalysts of the block. Given this sub-network we can compute proper entanglement measures. The entanglement study of the original network will then be a collection of entanglement intensity and homogeneity for each block.

Thus we need to make a choice if we intend to offer but one measure of homogeneity and intensity. It is arbitrary depending on the application context. With INA documents as substrates, we usually return entanglement values of the the biggest catalyst connected component which corresponds to the largest substrate sub-network, because other blocks only concern a very few catalysts on a very few substrate edges (this is a side effect due to our method of selection of substrate networks). When all the blocks are of reasonable size (reasonable needs to be defined by the application context), we can return averaged values for intensity and homogeneity. Note however that we will always get one unique individual entanglement index for each catalyst.

3.4.3 *Which edges should be taken into account?*

A basic workflow as presented in Section 2.5.1 suggests a projection of the bipartite graph onto a multiplex substrate interaction graph. Entanglement then focuses on multiple relationships between substrates (hence only the substrates that share *at least* two catalysts). This is fine when working with a limited set of catalysts but in the case of a larger set of substrates and/or catalysts, a different strategy should be observed. The rule we use to build proximity between documents, by projecting only multiple content relationships, may be considered as equivalent to a condition on a distance function f, where $f(d_i,\ d_j) = |T_i \cap T_j|$, $f(d_i, d_j) > 1$, with $T_i,\ T_j$ representing the

set of terms of documents d_i, d_j. Other rules could also be considered, based on measures such as intersection with another network, or even manual selection of edges. The computation of entanglement relies *only* on edges, which makes it a powerful tool when studying any kind of multiplex graph.

Another question often arises as to which substrate edge should be taken into account for measuring entanglement. This question mostly matters when building the conditional frequency matrix. Thus, two cases need to be identified:

- The first case is the one where the substrates edges are *not* the result of a homophily (similarity) network, but whatever other kind of network that clearly states multiple edges (such as a physical network of interconnected cities). In this case, all substrate edges should be taken into account, even when only one catalyst appears on it. The conditions to fill the conditional frequency matrix will still be respected (nonnegative values between 0 and 1), but more edge space will be taken into account, certainly leading to lower intensity measurement as if we had only considered the "interacting" edges (*i.e.* those displaying at least 2 catalysts).
- The second case appears when edges are the result of similarity between substrates. In this case, only the interaction pattern of catalysts is the focus of attention in the entanglement study. We therefore need to consider only the interacting catalysts, and would discard all other edges since they would bring noise into the analysis.

3.5 Perspectives

We have shown, in this chapter, how we extended a theory taken from a social analysis approach to any multiplex network analysis. We have detailed computation, usage and limitations of a local measure, the entanglement index, and two group measures, the entanglement homogeneity and intensity. Extended work is in preparation and still remains unfinished at the time of writing this thesis.

3.5.1 Extension to more general graph models

In real-world networks, relationships across entities are not always equal, and we often need to apply weights to edges of the graph model we use. These weights could, for example, represent a geographical distance, or an information flow, between two entities. Although the integration of a weighted model in our analysis is not yet complete, we will present now a promising lead.

Let us consider an undirected weighted substrate interaction graph of $G_{\mathcal{A}} = (V_{\mathcal{A}}, E_{\mathcal{A}})$. We have a map $w : E_{\mathcal{A}} \to \mathbb{R}^+$ (where $\mathbb{R}^+$ denotes the set of reals $r \geq 0$), hence denoting the weight of an edge e as w_e.

We extend these notations and write $w(F) = \sum_{e \in F} w_e$ for any subset $F \subset E_{\mathcal{A}}$.

We moreover assume edges carry catalysts, that is we additionally have a map $\tau : E_{\mathcal{A}} \to 2^{\mathcal{C}}$. The value $\tau(e) \subset \mathcal{C}$ is the set of all catalysts $t \in \mathcal{C}$ associated with edge $e \in E_{\mathcal{A}}$. An edge e bears catalyst t whenever $t \in \tau(e)$. Conversely, $\tau^{-1}(t) \subset E_{\mathcal{A}}$ is the set of edges bearing catalyst t.

We define:

$$n_t = w(\tau^{-1}(t)) = \sum_{e \in \tau^{-1}(t)} w_e \tag{3.6}$$

$$n_{t,t'} = w\big(\tau^{-1}(t) \cap \tau^{-1}(t')\big) \tag{3.7}$$

$$c_{t,t'} = \frac{n_{t',t}}{n_{t'}} \tag{3.8}$$

Since we want to preserve here the probabilistic interpretation of the c_t and $c_{t,t'}$ values. Hence, we need to further define:

$$c_t = \frac{n_t}{w(E)} \tag{3.9}$$

- Equation (3.9) may be interpreted as the probability that an edge bears catalyst t.
- Equation (3.8) may be interpreted as the conditional probability that an edge carries t *given it already bears* t'.

With this generalization, we fall back on the ordinary situation if we consider equal weights $w_e = 1$ for all edges $e \in E$. As a consequence, we may still define the entanglement index through the recursive definition: "catalyst entanglement is reinforced through interactions with other tangled catalysts".

Now, catalysts may not be equal (some may weigh more than others), and the interaction through *a same* catalyst across two different pairs of substrate may weigh differently, and generalization of our model to these different cases is still in progress.

The case of a directed multiplex graph came also to our mind, and we have not yet been able to give a proper sense to the opposition of directions in terms of catalyst interaction. However, it is still possible to consider opposite edges of a directed graph as different edges ($e(u, v)$ and $e(v, u)$) on which occur independently catalyst interaction.

The consideration of dynamical complexity, as another dimension that represents time, is another very interesting topic in graph analysis (of which nodes and edges have existence over time). We have not yet started to consider such approaches, but we suspect, in the case of news events, that the dimension of catalysts grows with time as more index term are brought into play. The evolution of the shape of the catalyst interaction graph might be a good indicator of how much

an event can be so illustrated, that it becomes a general theme, and related sub-events could appear from that theme. Such a dynamical study would help us tackle the TDT challenges of *First story detection* and *Link detection* (as presented in Section 2.2).

3.5.2 *Entanglement intensity and homogeneity of a substrate node*

We have briefly started to explore the possibility of applying the entanglement homogeneity and intensity at a substrate level. The intuition came from the following question: *how cohesive is a substrate node relatively to its neighbourhood?* We can straightforwardly apply the entanglement analysis to a subgraph "ego-centred" on each node u with its neighbourhood in $G_{\mathcal{A}}$. Hence we have two measures, of entanglement intensity $\mathcal{I}_{\mathcal{N}_{G_{\mathcal{A}}}(u)}$ and one of homogeneity $\mathcal{H}_{\mathcal{N}_{G_{\mathcal{A}}}(u)}$.

A few facts need, however, to be taken into consideration for such measure. First of all, does the "ego-centred" subgraph needs to include the edges between neighbours of u – in other words, should we compute the entanglement measures on the induced subgraph of $N_{G_{\mathcal{A}}}(u)$? or should we consider only the edges adjacent to u? This difference (illustrated in Figure 3.6) might be quite philosophical, if we question the cohesion of a node's connections, in which even the second case might be a better choice. However, if we question the legitimacy of a node within its neighbourhood, then the first choice might be a better fit – *e.g.* this might be adequate for a greedy clustering approach.

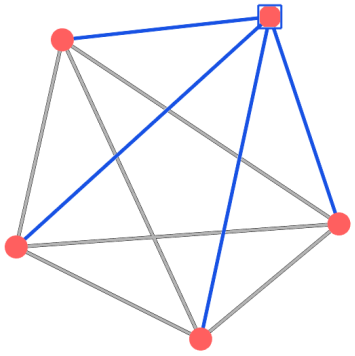
Moreover another subtle difference might direct our choice. When taking the induced subgraph of a neighbourhood into account, every node that is part of a clique will display the exact same values of entanglement. This is not true when considering only a node's adjacent edges, enabling us to discriminate nodes belonging to a clique. Figure 3.7 suggests there are many nodes of interest with maximal intensity and homogeneity whatever the type of neighbouring subgraph chosen.

Additionally, substrate nodes can bridge different communities of catalysts. Indeed we have seen previously that blocks in the interaction matrices correspond to disconnected components. Whatever the type of neighbourhood subgraph we consider, there might be different connected components of a catalyst at play within the subgraph[7]. Surely, such node brings noise in the network, but it does not mean that it does not belong to one or all of the communities. What to do in such case? We merely raise the question here and leave it open.

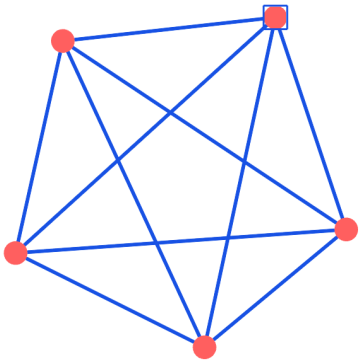
We also started to study the distribution and correlation[8] between the entanglement and other common measures (detailed plots in Appendix A), and build Table 3.2 of correlations[9].

Substrate entanglement intensity and homogeneity do not seem correlated (or anti-correlated) to any kind of known regular network measures. Of course, further investigation is needed before drawing any conclusion. Except for *multiplexity*, we have only made compar-

Figure 3.6: Should we take into account all edges adjacent to a node...



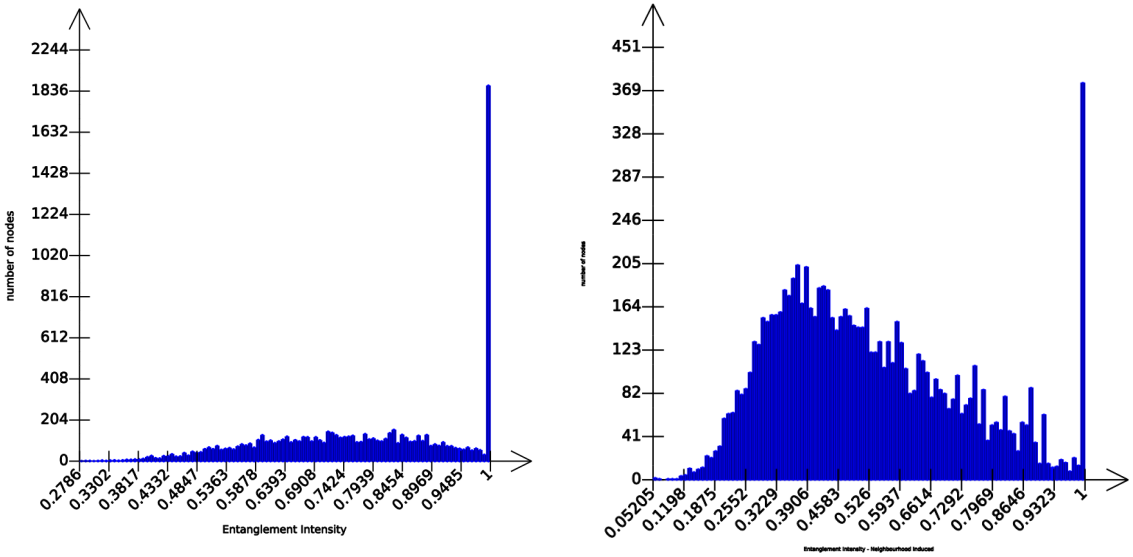
... or the whole induced neighbourhood?



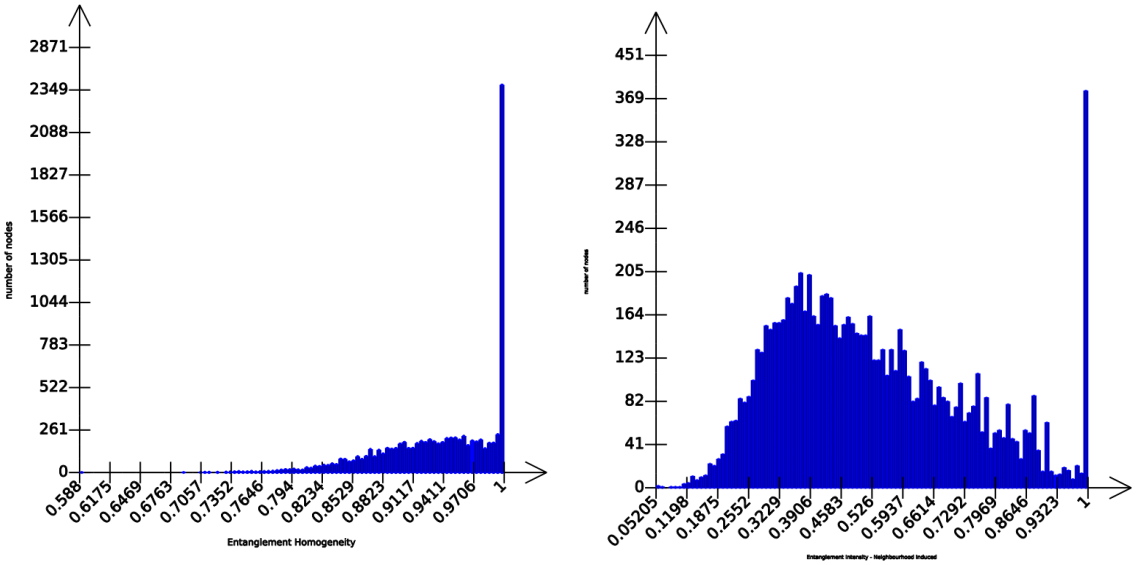
7. We have found 36 substrate nodes in our corpus – with both methods for selecting the "ego-centred subgraph" – that connects at least 2 connected components of catalysts, and these nodes do not appear to be peripheral in their own clusters, and often belong to cliques.

8. We use Pearson's correlation coefficient (Pearson, 1901), defined as $corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ with cov(X,Y) the covariance of both variables X and Y and σ_X the standard deviation of the variable X, although other correlation coefficient comparing ranks such as Spearman's might also be appropriate (Spearman, 1904; Kendall, 1948).

9. This led us to question the correlation of all other measures together in a document network, raw results are presented in Appendix B.

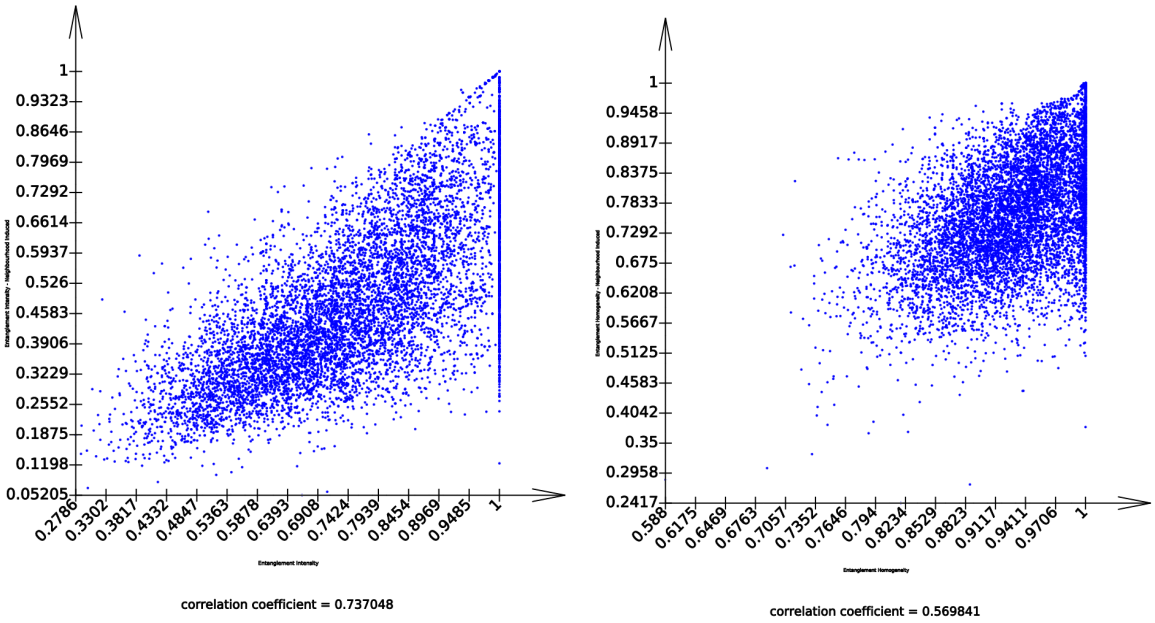


Distribution of entanglement intensity $\mathcal{I}_{\mathcal{N}_{G_{\mathcal{A}}}(u)}$

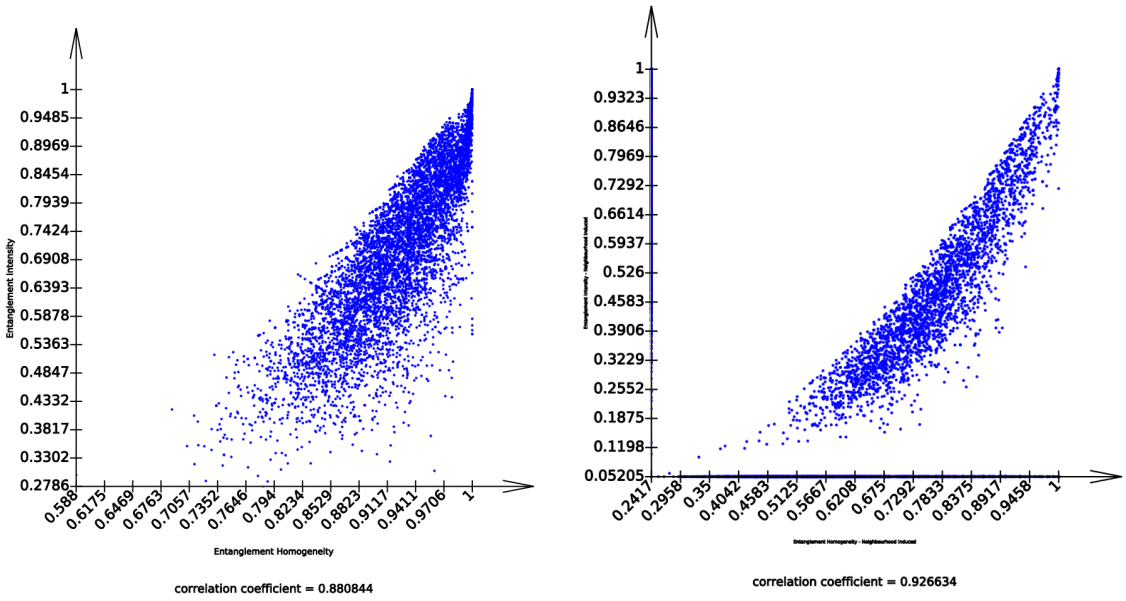


Distribution of entanglement homogeneity $\mathcal{H}_{\mathcal{N}_{G_{\mathcal{A}}}(u)}$

Figure 3.7: Distribution of entanglement measures among the 10,000 news document spread in 300 clusters (from Section 2.1.3. Adjacent "ego-centred subgraphs" are on the left, and induced neighbourhoods on the right. Notice the high number of maximal or near maximal values, suggesting areas of interest.



Entanglement intensity and homogeneity of both methods compared together



Entanglement intensity compared to homogeneity in both methods

Figure 3.8: Comparison of measures of intensity (top left) and homogeneity (top right) between both types of "ego-centred subgraphs", display significant differences. The comparison of the intensity $\times$ homogeneity space from the adjacent neighbouring edges only (bottom left) and the induced neighbourhood (bottom right) display similar shapes (reminding us of the documents group positions (in red) in Figure 3.5).

	Ent. intensity	*Ent. homogeneity*
Betweenness centrality	-0.130	-0.104
Closeness centrality	-0.083	-0.030
Clustering coefficient	0.403	0.295
Degree	-0.108	-0.060
Eccentricity	0.032	0.010
***K*-cores**	0.152	0.104
Multiplexity	-0.095	-0.039
Catalysts	-0.117	-0.078
PageRank	-0.266	-0.170
Strength	0.321	0.251

Table 3.2: Comparison of Pearson's correlation coefficient between the entanglement intensity and homogeneity and nine other measures.

isons with single-mode network measures, but not with *redundancy*, *network exposure* or other multiplex measures as presented in Section 2.4.4. These would be necessary to assess the originality of the entanglement measures. Mathematical properties of the substrate entanglement measures need to be studied in order to be included in any other algorithm. Correspondence with any measures should also be studied in random graphs. However, we feel quite strongly that we have a solid lead here, on new perspectives for multiplex networks analysis.

3.5.3 Other analysis

There are several leads for measures we have yet to explore. From Burt and Schøtt (1985), we find a proposed distance function which, applied between catalysts, would allow us to create equivalence classes of catalysts, and this might be useful when analysing a large number of catalysts. The Perron Frobenius theory is rich in tools and we have only exploited a small proportion of them when it comes to entanglement analysis. For instance, we have yet to explore what information might be obtained from the other eigenvalues, or from the spread we can observe in entanglement intensity values (since the maximum eigenvalue of a nonnegative matrix admits maximum and minimum limits).

We have also started to explore leads into measures at the substrate level. It could be useful to support visual analytics (see next Chapter) and to explore substrates which offer more, or less, interaction between catalysts. One effect we have tried to report is the catalysts contribution to entanglement at the substrate level, as the two bind together. We are able to propose a naive way to measure the contribution of a substrate s in entanglement, as $d_s = \sum_e \sum_t g_t / e$.

Another way would lead us to analyse entanglement in both projections by swapping between catalysts and substrates.

A last goal would be to study profiles of entanglement in differ-

ent random graph models (Skvoretz and Faust, 1999; Guillaume and Latapy, 2005; Wang et al., 2009), with many real world application graphs, and compare entanglement measures with other metrics if applicable (Latapy et al., 2008b; Fujimoto et al., 2011).

We also have not tackled the algorithmic optimization yet, and we expect to find indexation-based approaches to compute the entanglement measures with the lowest complexity possible.

3.6 Conclusion

We have presented here the *entanglement analysis* as a new methodology to tackle multiplex networks. We have introduced our motivations from early work in Social Network Analysis. We have extended the model to *any* multiplex network (and, with careful interpretation, applicable by extension to most complex networks and n-partite graph). We have defined the entanglement in a graph of which study brings the *catalyst entanglement index*, *group entanglement intensity* and *homogeneity*, and the *catalyst interaction graph*. These objects express three levels of granularity in a multiplex network, with substrates corresponding to the lowest entity level. Catalysts make an intermediate level that creates the ties and brings groups of substrates together. Entanglement intensity and homogeneity clearly point at the top level, considering the complex network as a whole. It is one sure advantage of our methodology to be able to address these different levels *at once*. We have studied the behaviours of these measures, and detailed the many considerations one should have before applying the model. We have finally opened perspectives among them, an extension to a weighted model, and another extension to *local entanglement homogeneity* and *intensity*.

The work presented in this Chapter has been first introduced in (Renoust et al., 2011b). It has since then been extended and presented in (Renoust et al., 2013f,d,c) and soon in (Renoust et al., 2013e).

DESIGNS AND TOOLS FOR VISUAL ANALYTICS

4

Sound and powerful mathematical tools are very important in any successful analysis of complex information. However, the results of such analysis can be quite difficult to understand. Indeed not all users of an analytical system have the knowledge, or the time, to understand the mechanisms that are at play behind such systems. Thus the communication of all results could prove a challenge that is best supported by many tasks of visualization[1]. However, Visual Analytics goes further that mere communication of analytical results as it supports reasoning, as described by Thomas and Cook (2005, Page 4), Visual Analytics is "*the science of analytical reasoning facilitated by interactive visual interfaces*".

1. "The greatest benefits of data visualization are found in its ability to bring important findings to light and to help us think productively and swiftly, leading to the poignant experience of: ***Aha, I see!***" (Few, 2006).

The design of such frameworks is an extensive work that spans from considerations of human factors, teaching us why a system must interact with its users in a specific way, to the identification of which relevant tasks should be addressed. The design also has to include techniques for technical implementation, and Nanard and Nanard (1995) gave us a perfect example of such design.

In this chapter, we present the cognitive elements that explain how, exactly, visual analytics support reasoning. These explanations form the base line of good design and sound validations of a visual analytics framework. We will then consider how human beings might perceive the visual information in order to identify the tools of information visualization and interaction that will most accurately aid reasoning in our context, especially in the case of complex networks.

4.1 Visualizing and analysing for understanding

Searching for an information is actually even more than just finding: we search, we find, we *learn*, and naturally *explore* the neighborhood of our recent findings (Marchionini, 2006). The typical way in which scientists process data is reported in Springmeyer et al. (1992). The scientific process first investigates the data, then integrates any newly discovered insights. Four main actions were identified, and these seemed to be applied consistently across all scientific research domains: *"Interaction with representations (the exploration phase), Applying mathematics (to further investigate or confirm insights), Manoeuvring (movement within and among programs, to organize the data, and*

choose and set up representations), and Expressing ideas (keeping record of all the data transformations, observations before sharing and concluding on insights) . . . " We now can see how visual analytics contribute to this process.

4.1.1 Added-value of visual analytics

Visual interactions are thus aimed at facilitating analysis of complex problems in order to detect known features (the expected) and discover new insights (the unexpected) (Wong and Thomas, 2004). A framework supporting visual analysis should also support the discovery of the unexpected, a process that involves *serendipity*. Serendipity, as in André et al. (2009), is indeed the whole process of accidental discovery. In this case, we are interested in finding unexpected new information while one is only observing current information. It is made by the reasoning process that leaps from one *understanding* to another until a new insight is formed. *Understanding* analytics and visualization is therefore an important key in discovery, greatly enhanced by an environment rich in strong analytical tools and relevant visualizations. Network visualizations thus improve visual discovery, encouraging users to question the relationships that they are observing (Tamassia, 2013, Chapter 22). Other user centric advocates, such as Dörk et al. (2011)'s *flaneur*, propose exploration as the guiding principle in information seeking. Visual exploration is also a perfect way to deal with heterogeneous and noisy data as it is intuitive and requires no understanding of the complex mathematical and algorithmic procedures (Keim, 2002).

Making sense of visualization is a complicated process (Friel et al., 2001) that involves multiple factors, including understanding of conventions in visualization design, manipulations of the information, and general abstraction of the acquired knowledge. With regard to manipulation, a good interaction design should offer users the possibility of adapting their own perspective on the information, and is therefore an essential component that assists reasoning (Pike et al., 2009).

No matter how good the visual tools that we manipulate, they will still involve many steps of interaction between humans and machines[2], a possible source of failure in the system's assistance of reasoning. The goal of successful design (Norman, 1988) is then to reduce the distance between the user's intention of action and the system's allowable action (*the Gulf of Execution*), as well as the distance between the system's representation, and the user's perception and interpretation, which can also differ from the user's expectations and intentions (*the Gulf of Evaluation*).

Hutchins et al. (1985) offer more details on these different distances. An interface should be well matched in respect of the level of description of a task (and at the output of the action), and the level at which the user thinks of this task (or expects the output) by provid-

2. *The seven stages of action* by Donald Norman (1988), decomposes human action, allowing us to identify possible source of failures in design. They can be listed as:
"*1. Forming the goal*
2. Forming the intention
3. Specifying an action
4. Executing the action
5. Perceiving the state of the world
6. Interpreting the state of the world
7. Evaluating the outcome"

ing abstraction structures (*e.g.* instant values or progressions). The *semantic distance* reflects how much of such a structure is provided by the system and how much by the user. The *articulatory distance* concerns the shape that takes the description of a task (or its output). For example, using a mouse to select a screen area reduces the *articulatory distance* in comparison to a box selection coordinates. Wisely minimizing such distances naturally engages users and reduces the cognitive load[3] of the system's interaction.

The different gaps in perception, especially *rationale gap* underlined by Amar and Stasko (2004), are critical issues that any visualization must address. This rationale gap necessarily emerges between the actual relationship (at the data level) and the perceived relationship between elements in the visualization. Coping with this gap involves "*being able to explain confidence in the relationship, as well as its usefulness*". This analytic gap moreover may suffer from any distortion that graphical representations inevitably carry, often inherited from data transformation or geometrical projections (Kaski et al., 2003; Lespinats and Aupetit, 2009, 2011).

Wehrend and Lewis (1990) have already defined the *effectiveness* of a visualization as its ability to faithfully translate similarities so that items close to one another can be inferred to be similar[4]. Thus it is essential for a visualization to limit such an effect so that users can effectively trust a given representation as a base for reasoning, otherwise counter-effects would be a nuisance more than an aid (Klein, 2007)[5]. Many solutions are investigated to ensure trustworthiness, from design to techniques displaying uncertainties (Pang et al., 1997; Kaski et al., 2003; Skeels et al., 2008; Chuang et al., 2012).

4.1.2 Cognitive elements

Visual Analytics support analysts in their goals, helping them to make sense of complex bodies of data. *Sense-making* is itself a well studied process, and is defined in Russell et al. (1993) *"as the process of searching for a representation and encoding data in that representation to answer task-specific questions"*, completely coinciding with the goals of Visual Analytics. Visual representations can be very flexible and efficient representations reducing the cognitive/external resource cost of the different operations required in the sense-making process (as explained in Section 4.3.1.

Klein et al. (2006a) have pointed out five psychological concepts that do not quite define sense-making itself but do shed light on the possible components of sense-making: ***creativity*** such as the ability to build solutions to puzzles, ***curiosity*** which motivates exploratory behaviours, ***comprehension*** which tackles the understanding of individual stimuli from a complex situation, ***mental modeling*** which is a memory representation of salient aspects of a situation expressed in concepts and knowledge, and ***situation awareness*** which comes after *mental modeling* and refers to the state of which principles, knowl-

3. The cognitive load is roughly the amount of the brain's cognitive processing that is needed to achieve a task. The interested reader can find more on *cognitive load* theory in (Sweller et al., 1998; Paas et al., 2003; Sweller et al., 2011)

4. *Efficiency* and *effectiveness* are ISO recommended criteria (for Standardization, 1998), very often used in usability studies (Frøkjær et al., 2000)

5. Many funny examples of bad designs can be found in (Flanders and Willis, 1998), and (ill designed) websites such as www.webpagesthatsuck.com, www.baddesigns.com, wtfviz.net, or cartastrophe.wordpress.com.

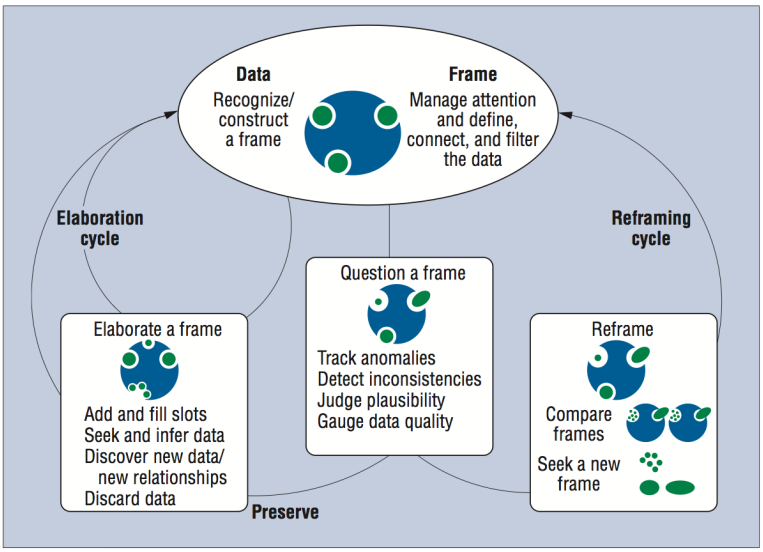


Figure 4.1: The macro cognitive model of sense making by Klein et al. (2006b).

edge and inferences can be drawn from the initial situation/data. We can see that *mental modeling* can be well supported by visualization tools, and *creativity* and *curiosity* are the necessary degrees of freedom which the visual analytics framework should cultivate, and not hinder. Overall *comprehension* would be the goal of the visual analytics framework, and *situation awareness* the goal of the user.

Although these five concepts involve different aspects related to sense-making, they are not sufficient to define it. Klein et al. (2006b) propose a very high level macro-cognitive model of sense-making based on data and frames (Figure 4.1). This model insists on how a *frame*, which defines a perspective, viewpoint or framework on a situation, and how *data* may be interdependent. The frame shapes the relevant data (we often select information from the intuition of an explanatory lead), yet that data limits the range of possible frames. This cyclic and iterative refinement process appears in most of the frameworks that we will soon be demonstrating. Our challenge, in visual analytics, is to offer users constant opportunities to reformulate their perspective, rendering them increasingly relevant, eventually pointing out conclusions, the need for more data, or the need for new rerepresentations[6], or new frames.

6. "Rerepresentation *re-construes parts of compared situations in order to improve a match. It is an important process in analogical reasoning and learning.*" (Yan et al., 2003)

The whole sense-making process is well described in (Pirolli and Card, 2005), and Figure 4.2 is an adaptation from Pirolli and Card in (Thomas and Cook, 2005). It is worth describing the sense-making loop in order to see where Visual Analytics can leverage difficulties. The first thing to notice is, at any step, the process is recursive and iterative. We often iteratively refine analysis at different levels of granularity. The process can also be bottom-up – from data to conclusions – or top-down – confirming hypotheses or theories with data. We often "jump" between both approaches depending on findings and opportunities.

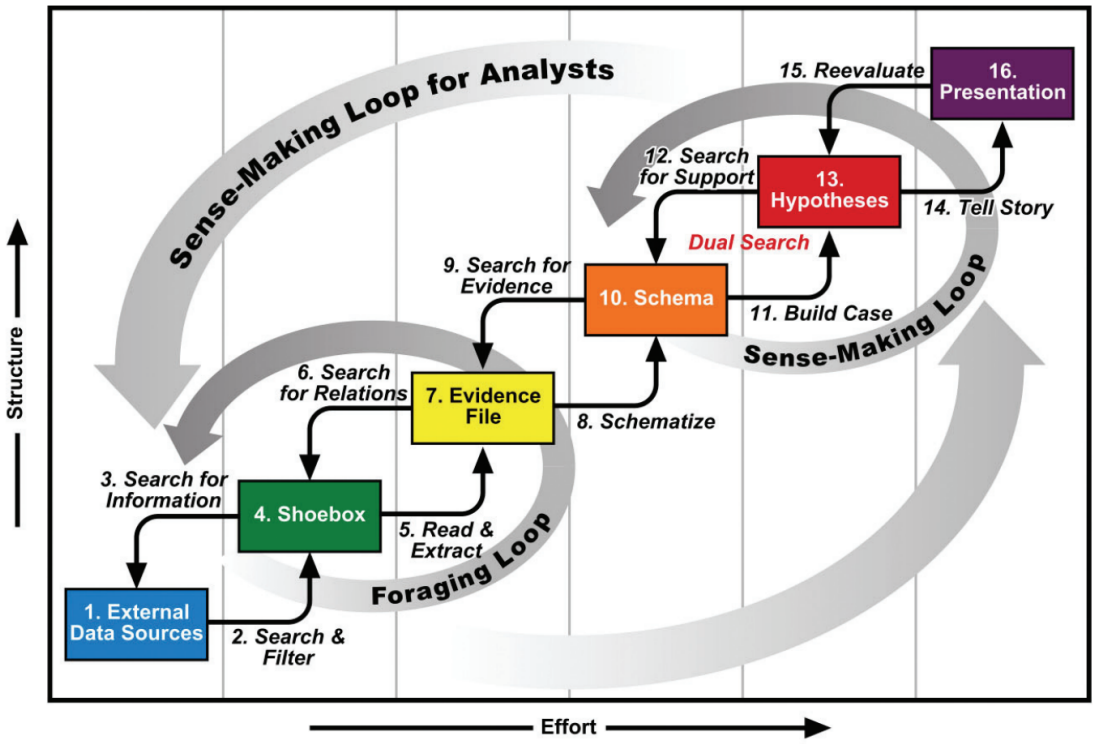


Figure 4.2: The sense-making loop for analytics by Thomas and Cook (2005), adapted from (Pirolli and Card, 2005).

The first three steps, *external data sources*, *shoebox* and *evidence file* compose the *foraging loop*. This is about selecting the relevant data for analysis, possibly translating them in a structured way facilitating each transition. Visualization is at play when the (re)representation step enters: *schematize*. *Searching for evidence* can be made easier with visually organized information – for example, by discarding outliers in the data. Depending on the quality of the relevant data, the process can go as deep as searching and filtering from the original sources. Once the information is well selected and schematized, we enter the *sense-making loop* composed of *schema*, *hypotheses*, and finally *presentation*. One can indeed, from cases formulate hypotheses enabling explanation of the data, although the explanation might not be completely satisfactory so the process needs to step backwards. Visualization also supports illustration to facilitate communication of the results. Moreover, interaction in visualization eases back-and-forth processes in these final steps. Notice too, that a precise *dual* search occurs here. Visual analytics can play a particular role in the articulation of both loops, the transition from concrete evidence to abstract representations. The first loop consists in manipulating very concrete information – low level, and the second in manipulating pure abstractions of the information.

From (Card et al., 1999), information visualization enhances analysts' cognitive abilities in six main ways which, combined with data analysis and interaction, can be applied to analytic reasoning and

support the sense-making process (Thomas and Cook, 2005).

- *"increasing the memory and processing resources available to the users"*: encoding of information and making it comparable, *e.g.* the length of a set placed on a node size
- *"reducing the search for information"*: showing order where none obviously appeared, *e.g.* proximity between elements
- *"using visual representations to enhance the detection of patterns"*: *e.g.* a force directed network diagram displaying tight communities
- *"enabling perceptual inference operations"*: *e.g.* correlations of two variables when plotted
- *"using perceptual attention mechanisms for monitoring"*: *e.g.* animation, colour coding, or blinking red indicators for potentially threatening or dangerous situations
- *"encoding information in a manipulable medium"*: *e.g.* allowing interaction through many means such as selections or scales

We can clearly see that each of the listed points can enhance most, if not all steps in the sense-making loop.

At each step, the sense making process places heavy demands on cognitive processes, and thus depends on the liberation of the analysts' cognitive and perceptual capabilities. When liberated, they can focus on the analytical problem and augment their potential discoveries. Visual Analytics aims precisely at supporting these goals, and Keim et al. (2008) rewrites and adapts the sense-making process in a formal Visual Analytics process in terms of datasets, operations and data representations.

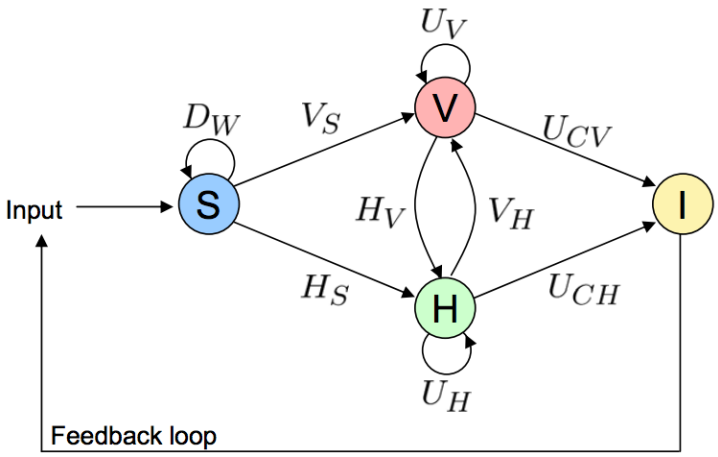


Figure 4.3: The formal Visual Analytics process by Keim et al. (2008).

In Figure 4.3 we observe, again, the natural iterative structure of the Petri network that is described, with the recursion emphasizing that Visual Analytics is an iterative process. S describes the data as input, to which some pre-processing D_W is often applied. This part mostly covers the *foraging loop* in the sense-making loop for analytics. A branching is then made on both visualization V and hypotheses formulation H. These can be expressed directly from the data (V_S, H_S) or from one another (V_H and H_V). It is interesting to note that this part not only covers the sense-making loop, but also frees both schematization/hypothesis formulations from the order proposed by Pirolli and Card (2005). This expresses more elegantly why the sense-making loop needs a "dual search" artifact: instead of seeing opportunistic top-down/bottom-up approaches, this part is done not only in parallel (one can equally formulate hypotheses on the data *before* or *after* visualizing it), but many iterations from one another help the analysis to converge to conclusions (or insights, I in the schema).

These steps are supported by human intelligence (user interaction, U_V and U_H) which also drives the process to insights I (from visualization U_{CV} or from hypothesis U_{CH}). These exact insights I can then feed back into the process (*feedback loop*) to refine information until the final insights and conclusions do not require further iteration. The user interaction part corresponds in the sense-making loop to the transitions between each state. (Keim et al., 2008) elegantly succeeded in formalizing the Visual Analytics process, expressing it in terms of application functions of the data, and especially identifying key roles of user's intelligence.

4.2 Design and validation

Now that we know where visual analytics operate in the analysis process, there are many ways for visualization to support analysis. However designing visualization is a delicate process that covers many layers of abstraction, each of which contains potential threats that could lead to bad designs.

4.2.1 Framework design

Munzner (2009) provides us with some guidelines when designing an efficient system for problem solving, and it is driven by the four points upon which visualization systems usually fail. Although this approach seems directed to problem solving, sense-making is obviously one important part at any step of problem solving.

- *"**Wrong problem**": the system does not address its target users problem.* This is the *domain problem characterization* level, and at this level we need to identify which "concrete" problem our system needs to answer.

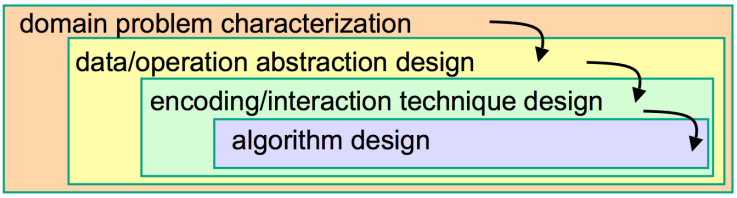


Figure 4.4: The four-part *nested model* to design a visualization system (Munzner, 2009).

- *"**Wrong abstraction**": the system does not display what is interesting users in their problems*. This is the *data abstraction/operation abstraction* level, and at this level we need to identify which data abstraction and data operations are relevant to solving our problem.
- *"**Wrong encoding, wrong interaction technique**": the system displays abstraction in an inefficient way, complicating the user's workflow*. This is the *visual encoding and interaction technique* level, at this level, we need to design the right visual encoding and the right interactions, leveraging the cognitive load for users in solving their problem.
- *"**Wrong algorithm**": the system is too slow*. This is the *algorithm* level, at this level, we need to design algorithms that are efficient so that computation will not be a bottleneck in the decision making process for problem solving.

Meyer et al. (2012) refine the nested model with a block and guidelines model, offering a lower level of granularity within each level of the nested model. Although it lacks a taxonomy for blocks and guidelines, it highlights the fact that implementing a system is not as linear as one might understand from the nested model, but interaction patterns can be observed within and between the different levels of abstraction. Moreover, this paper reflects the tremendous benefits of describing clear tasks at distinct levels of abstractions.

From this nested model, we can suspect that integration of the entanglement tools are involved at the *algorithm level*, reading and interacting with data, driven by the entanglement measures, and intervening at the *visual encoding and interaction technique* level. A multiplex graph would occur at the *data abstraction/operation abstraction* level, and this multiplex graph would offer sense-making in problems identified by the *domain problem characterization* level. However, we will discuss the subject in the next Chapter, Section 5.1.

4.2.2 Analytic tasks

The literature is rich in taxonomies of tasks that support reasoning, whether it's in the field of human-computer interaction, cartography, information retrieval, visualization, or visual analytics. They can be

high level or low level, domain specific or generic. Meyer et al. (2012) highlights the gap in the literature between low and high level tasks classifications, that strongly motivated the excellent work of Brehmer and Munzner (2013), which describes a typology for visualization tasks at *any* level of abstraction. We will detail this key contribution in the next Section, but will now go through a few other articles identifying tasks that are relevant to visual analytics, and especially tasks that are useful for complex networks exploration. The list is extensive but we believe it will help readers understand this necessary aspect of applied analytics[7].

Amar et al. (2005) provides a set of domain-free low level analytic tasks from an extended set of user behaviour observations. These tasks are centred on numerical operation and data selection, upon which users achieve reasoning.

- *Retrieve Value*, find attributes of a data (sub)set.
- *Filter*, find a data (sub)set satisfying some conditions on attributes.
- *Compute Derived Value*, compute a numeric representation of a data (su)bset.
- *Find Extremum*, find a data subset with extreme attribute value in the observed data (sub)set.
- *Sort*, rank a (sub)set of data given an attribute.
- *Determine Range*, find the span of an attribute value given a data (sub)set.
- *Characterize Distribution*, characterize the distribution of an attribute given a data (sub)set.
- *Find Anomalies*, given criteria and relationships, find the anomalies in a data (sub)set.
- *Cluster*, group (sub)sets of data of similar attributes.
- *Correlate*, determine relationships between attributes among a data (sub)set.

We might notice that most of these tasks can be expressed in terms of operations and combinations of the first three tasks. For example, *Find extremum* can be seen as *Retrieve Value* of attribute *a*, *Compute Derived Value* max and min of *a*, *Filter* data entries that satisfy value of *a* equals extremum. An efficient Visual Analytics framework should easily support such tasks, at least, in our case for observations of the different entanglement measures.

Amar and Stasko (2004) refer to *analytic gaps* as obstacles between visualization and higher level tasks – which more or less depict the gap mentioned in (Meyer et al., 2012) and Norman's gulfs (Norman,

7. Of course, one cannot avoid citing the most famous InfoVis paper that gave rise to the most famous visual information seeking mantra *"Overview first, zoom and filter, then details on demand"* (Shneiderman, 1996) which provides per-data type a description of the low-level tasks of *Overview, Zoom, Filter, Details-on-Demand* (in addition *Relate, History, Extract*, but these are not in the mantra).

1988), on two levels, the *Rational gap* and the *WorldView gap*, and proposes visualization tasks to avoid falling into these gaps. The *Rational gap* refers to the difference between a perceived relationship through the visual system, and the ability of such system to explain reasons for this apparent relationship. To bridge this, Amar and Stasko propose the following tasks: *Expose uncertainty* in measures and outcome, for example by showing population, average and standard deviation of a statistical summary. *Concretize relationships* by explicitly displaying them, for example linking and brushing among derived data highlighting the source data. *Formulate cause and effect* possibly allowing live editing to investigate effects from causes. The *Worldview gap* is the gap between what is displayed from what needs to be displayed (including the limitations of a display). The authors thus propose the following tasks: *determination of domain parameters*, *multivariate explanations*, and *confirm hypotheses*.

Buja et al. (1996) describe some multivariate data tasks for interactive visualization. Although they might be part of the visual analytics culture now, and feel quite basic, these higher level tasks still remain a good reference: *Finding structural patterns*: focusing individual views in order to figure out some pattern or overview (choosing data subsets/attributes/visual parameters); *posing queries*: linking multiple views in order to highlight relationships (search for an element in all views, brush among axes and highlights in other views); *making comparisons*: arranging many views in a comparable way (with similar layouts, or a scatter plot matrix for example). Some other works are also focused on interaction tasks and design: Dix and Ellis (1998) looked for simple interaction tasks to augment existing static visualizations (highlight and focus, accessing extra information, overview and context, keep representation and change parameters, and keep data change representation).

Chi and Riedl (1998) describes an operator-based framework for visualization (low-level) design and manipulation tasks. Without going through all the stages and detailed taxonomy of their model, a very interesting aspect of this contribution is the separation of data manipulations and visualization manipulations. It derives from earlier work (Chuah and Roth, 1996) that described the semantics of interaction in visualization systems – which also distinguished operations from graphical, and more arguably data and set operations. This makes a clear distinction as to whether a set of tasks influence the visualization or the data itself. Operator (of interaction) are then tasks or manipulations of such views, or data, and can easily be assembled in a pipeline.

Gotz and Zhou (2009) identified a whole catalogue of 21 different low-level tasks from quantitative observations of usage of Visual Analytics systems. Three categories have been identified. *Exploration actions* separate visual exploration tasks and data exploration tasks, *Insights actions* manipulate the visual insights or knowledge insights taken from exploration, and *Meta actions* operates at the application

level on user's activity. More recently, Heer and Shneiderman (2012) provide three high-level categories of tasks and 12 different tasks: *data and view specification* (visualize, filter, sort, and derive); *view manipulation* (select, navigate, coordinate, and organize); and *analysis process and provenance* (record, annotate, share, and guide). The categories correspond to high-level tasks, and are critical in enabling iterative visual analysis. These tasks are relevant for visualization creation, interactive querying, multi-view coordination, history, and collaboration.

Close to this spirit, Kandel et al. (2012) studied the analysis process directly *from* analysts. They have identified five challenges (high-level tasks) that analysts face in their everyday job. Their results are well illustrated with concrete examples: *"Discover data necessary to complete an analysis tasks (*e.g. *finding a data set online, locating the database tables in a MySQL database, or asking a colleague for a spreadsheet). Wrangle data into a desired format (e.g parsing fields from log files, integrating data from multiple sources into a single file or extracting entities from documents). Profile data to verify its quality and its suitability for the analysis tasks (*e.g. *inspecting data for outliers or errors and examining the distributions of values within fields). Model data for summarization or prediction. (*e.g. *computing summary statistics, running regression models, or performing clustering and classification). Report procedures and insights to consumers of the analysis."*

Ward and Yang (2004), instead of focusing on operators/interaction tasks, focus on interaction spaces, which at a highly refined level of analysis are surprisingly variate: they can be the screen space (pixel level), the data-value space (data entries), the data-structure space (data structure that represents the data), the data graphical attributes space (graphical attributes representing data characteristics), the graphical object space (graphical objects that represents the data), and the visualization structure space (parameters structuring the visualization).

The list is very long[8], but all the tasks are described at different levels and are very relevant for analytics. This underlines how difficult it is to identify not only the task itself but also the level of granularity to which it applies. Moreover, ignoring the knowledge of this taxonomy zoo makes it difficult to make the statement about positioning, contributions, and validation of work in visual analytics.

8. Further reading: (Wehrend and Lewis, 1990), (Raskin, 2000), (Crystal and Ellington, 2004), (Valiati et al., 2006), (Yi et al., 2007), (Liu and Stasko, 2010) or (Roth, 2012).

4.2.3 Validation

Validation and evaluation of work involving human interactions, such as visual analytics, is always a delicate process and the subject remains a challenge (Plaisant, 2004; Scholtz, 2006; van Wijk, 2013). Quantitative criteria of information retrieval – *e.g.* precision, recall, f-measure (Powers, 2011) – do not quite apply. Comparison of aesthetic criteria, such as in (Purchase, 2002), are possible for very specific cases of application. Case studies in realistic settings (Kang et al.,

2009), most often demonstrating proof of feasibility, have the disadvantage of being extremely time consuming and not really reproducible (Plaisant, 2004). Usability evaluations (Sutcliffe et al., 2000) only provide feedback, which may offer good advice on enhance a design. User evaluations through controlled experiments[9] can quickly become difficult to put together without any bias (Nielsen, 1994; Saraiya et al., 2006); they also require time and resources, especially when a large number of participants is necessary for statistical validations.

9. All the best advice for user experiments can be found in (Purchase, 2012).

A tasks oriented design and validation offers an alternative approach. A clear description of tasks, based on strong cognitive theory, would enrich the validation of a framework, designed to answer specific goals. It may not be sufficient, but it can enrich additional validation processes, such as usability evaluation, or even prepare the ground tasks for an extended user experiment.

Brehmer's multilevel typology of visualization tasks (Brehmer and Munzner, 2013) is based on a large survey of taxonomies beyond the one we examined in Section 4.2.2. This typology presents numerous advantages. The first advantage is that typology completes the definition of tasks and characterizes them around three questions: *why?*, *how?* and *what?* as in Figure 4.5. *Why?* describes the motivations, the goals of a task, eventually narrowing from high- to low-level of abstraction. *How?* describes the methods, the visualization/interaction techniques, this aspect was often the most researched and considered in the taxonomies described in the previous Section 4.2.2. *What?* rationalizes a task by specifying its inputs and eventual outputs. This definition is purely abstract and enables the translation of any type of interesting task in the *why/how/what* framework, making it clear and almost ready for implementation, especially when we want to integrate a task in Munzner (2009)'s nested model. Thanks to the *what?* aspect of a task, this typology allows the linking of several tasks (which eases the integration for the block and guidelines extension of the nested model (Meyer et al., 2012)). Though useful for design, this work developed during our writing of this thesis, and thus we will only be using this typology in the description and evaluation of our own framework (see Section 5.2).

4.3 Information visualization

10. *Data visualization* is a very broad field issuing from *computer graphics*, and roughly two main disciplines compose it; *scientific visualization* focuses on the visualization of concrete phenomena often involving spatial 3D information (such as meteorological phenomena, fluid simulation, or the human body); and *information visualization* which focuses on abstract representation of data designed for understanding and analysis.

Now that we know the different frameworks in which we should embed visualization, to support analytical process, we can take another look at the Information Visualization field[10] and identify the different tools it offers us to visualize such information. The benefits of Information visualization are multiple (Fekete et al., 2008), and we can find numerous examples where visual indications clearly assist performance in achieving the tasks for which they were designed (Norman, 1993).

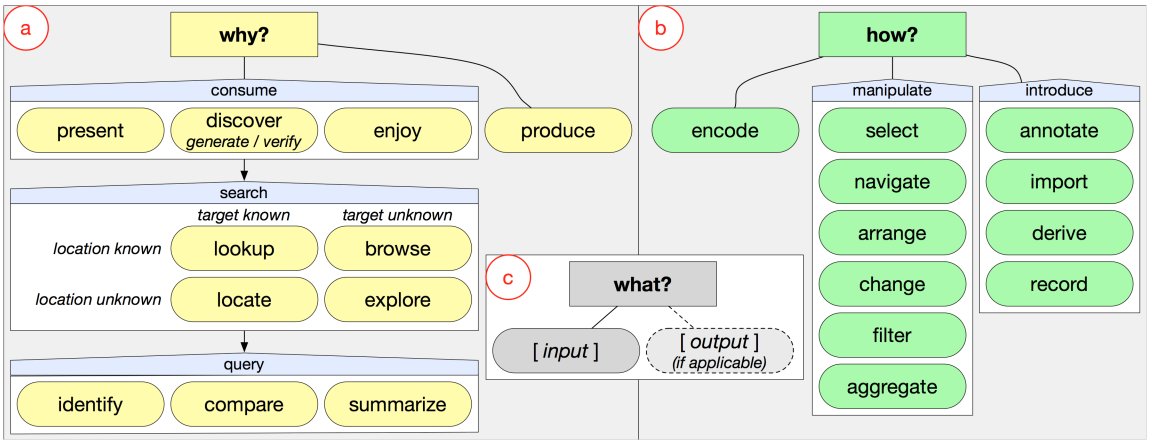


Figure 4.5: This visualization task typology (Brehmer and Munzner, 2013), places a task into context, answering the questions *why?* *how?* and *what?*. Freedom in describing inputs and outputs (through *what?*) allows a multilevel description of tasks answered by a system.

Visualization helps cognitive support (Card et al., 1999) by increasing memory and resource processing, by reducing search and recognition of information and patterns, and by encoding the information visually in order to utilize perceptual attention and perceptual inference as tools.

4.3.1 *Visual perception*

The mechanisms allowing utilization of the human perceptual capabilities have been demonstrated at a low-level with the *pre-attentive processing* theory (Treisman, 1985), and at a higher level by the *Gestalt* theory (Koffka, 1935).

Pre-attentive processing theory describes different features that are unconsciously captured by the brain even before any other attention or cognition process – with retinal rods and cones responding under 200ms (Goldstein, 2008, Chapter 3)[11]. A non exhaustive list of features (Healey et al., 1993; Ware, 2000) that are recognized at glance by pre-attentive processing (see Figure 4.6) and include *hue*, *orientation*, *position*, *direction*, *shape*, *size*, *density*, *etc.* Features that are not only recognized individually but also by groups (If and Blake, 1990). Changes of such features are also instantly captured (Ware, 2000, Chapter 6)[12]. Pre-attentive processing then becomes a very effective tool in directing someone's attention onto elements of interest. We can use features and groups of features to enforce associations of similar elements, thus allowing users to avoid many of the cognitive tasks that include grouping and associating elements.

Gestalt theory[13] (Koffka, 1935) can be applied to studies of visual perception as a whole system in which six principles bring us understanding of an image (Ware, 2000, Chapter 6):

- ***Proximity:*** *perception groups things that are close together;*
- ***Similarity:*** *perception groups things that share similar attributes;*

11. We invite our curious reader to learn more on the mechanisms of perception by reading the very complete work of Goldstein (2008)

12. Changes in features are not captured with the same priority and our brain can play some tricks with our vision, the experiment of Suchow and Alvarez (2011) is a striking example of such trick.

13. From the German *Gestalt* that translates as *shape* in English, Gestalt theory is a trend in psychology which aims at understanding the laws behind human's ability "*to acquire and maintain stable percepts in a noisy world*" (from wikipedia – Gestalt theory). One principle of Gestalt theory is that the brain is capable of understanding an entire system more than just a collection of isolated feature: "*the whole is greater than the sum of the parts*" (Koffka, 1935).

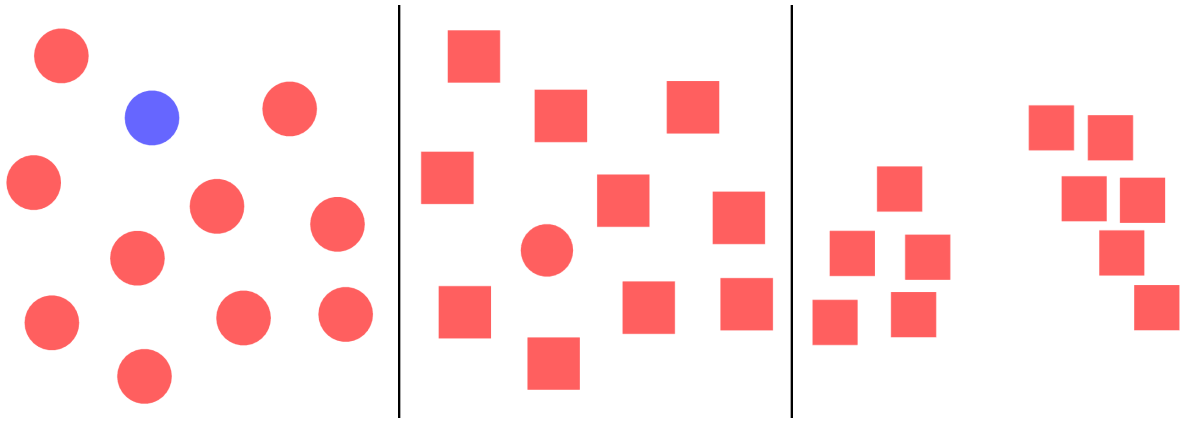


Figure 4.6: Three examples of different features pre-attentively perceived: *colour*, *shape*, and *density*.

- ***Continuity:*** *perception groups things that are connected or continuous;*
- ***Symmetry:*** *perception groups things that are symmetrically arranged;*
- ***Closure:*** *perception sees closed contours as a whole object;*
- ***Relative size:*** *perception sees smaller components of a pattern as objects, and larger ones as background.*

We can intuitively see how pre-attentive processing features can support understanding of an image.

4.3.2 *Visual encodings*

Using perceptual mechanisms to enhance visualization is a well studied field (Healey et al., 1993; Pickett et al., 1995). The field of cartography, and especially Bertin (1967) brought us a many parameters – called *visual variables* – that we can work with to *encode* information. These visual variables correspond to attributes of visual objects (see Figure 4.7) and can be listed as follows: *shape*, *size*, *value* (or *intensity* such as brightness), *orientation*, *colour*, and *grain* (which corresponds to a density of a texture). According to Bertin (1967) these visual variables can be associated with perceptual properties that are *selection*, *quantity*, *order* and *association*. *Selection* allows us to isolate groups of entities sharing the same value. *Order* allows us to naturally compare values one with another. *Quantity* encodes numerical attributes through mapping of quantities[14]. *Quantity* is more specific than *order*, and it allows us to estimate distance between two values. *Association* allows to group elements together, and it is possible to use a visual variable for *association* when the modification of their value does not affect the visibility of the object it represents.

Two types of information visualization are of particularly interest[15]. The first one is obviously network visualization since we are focusing on networks, where we are processing data with various attributes and the second one involves multivariate data visualization.

14. While mapping quantitative variables, scales that are not well chosen can induce severe distortions (Tufte and Graves-Morris, 1983), the choice of these scales is beyond the scope of this thesis but recommendations can be found in (Herman et al., 2000a; Harrower and Brewer, 2003; Bertini et al., 2007).

15. The field of information visualization is a zoo full of wild and strange animals and we invite our reader to take *"a tour through the visualization zoo"* (Heer et al., 2010). Most of the statistical diagrams we use have been credited to William Playfair (1801), however more accurate historical details can be found in (Spence, 2005). Additionally, historical use of visualizations that catalyzed Science can be found in (Freeland and Corones, 1999).

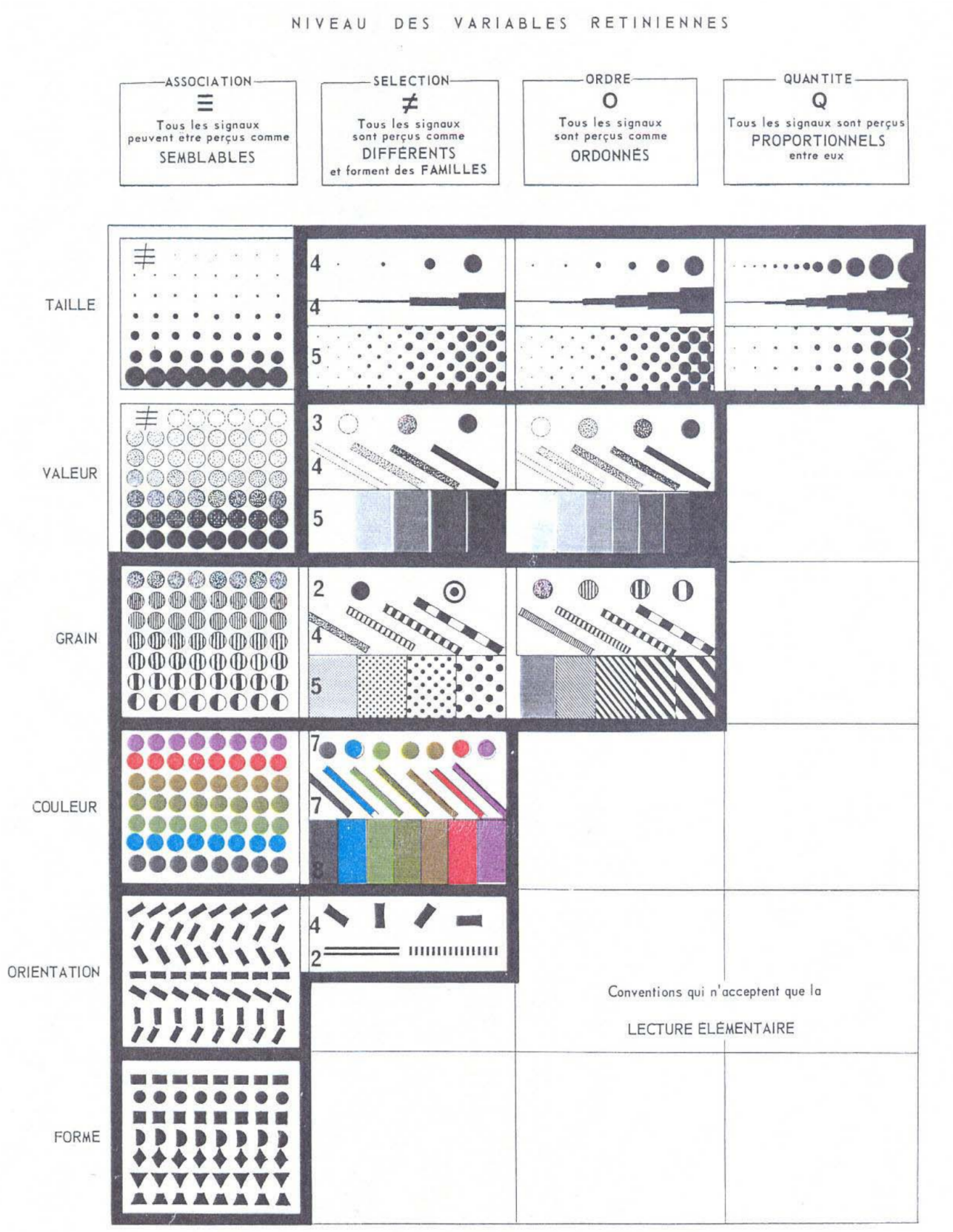
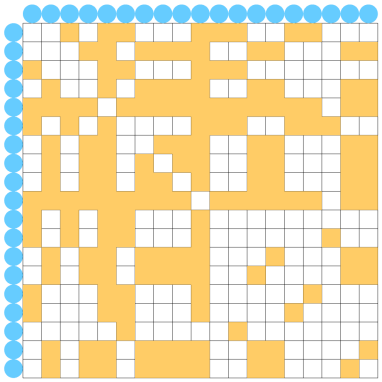


Figure 4.7: The different visual variables and their association to retinal attributes from (Bertin, 1967).

4.4 Visual analytics of complex networks

Figure 4.8: A basic matrix representation of the network example of Figure 2.13.



We can apply Big Data's challenges to network visualization. *Volume* obviously deals with large networks representations. *Velocity* might concern dynamic networks. *Variety* refers to attributes of nodes and edges, and in our case, layer complexity. *Veracity* addresses the trustworthiness of a network's visual representation. This section presents how visualization takes care of network representation and complex data structures, with interactions that support the analysis of such visualization.

4.4.1 Networks visualization

As we have seen before in Chapter 2, networks representations[16] are based on graph models. Graphs focus on the interactions between entities and different types of representations have been studied since (Bertin, 1967) (Figure 4.9). Two types of representations are widely used for general graphs (as opposed to specific graphs such as trees): *the matrix view* and *the node-link diagram*[17] (Herman et al., 2000b).

Matrix views (Figure 4.8) rely on the representation of a graph's adjacency matrix as a graphical object. Instead of showing numbers in rows and columns, they can encode relationships with colours or intensity. A matrix can be well reordered to show grouping and identification of bridges (Bertin, 1967). Although it does not show the issue of edge crossing that one can encounter in a node-link diagram of a large graph (Ghoniem et al., 2004), it is very difficult to follow a path in such matrix without the support of interaction. Moreover, we focused our exploration on node-link diagram approaches because of their relevancy in dealing with networks of moderate size and low density (Ghoniem et al., 2005).

In the literature, we noticed that node-link diagrams were supporting social network analysis as early as (Moreno, 1934) (Figure 4.10). It intuitively depicts relationships with nodes represented as glyphs, and edges as lines (or curves (Riche et al., 2012)). This type of representation benefits from previous *Graph drawing* research, offering efficient algorithms and aesthetic representations of networks (Tamassia, 2013). Although 3D representations of such a graph exist, we are currently only considering diagrams in 2D space[18]. Representation of the nodes with glyphs allows the encoding of node's information in all possible visual variables. The positioning of nodes and links in the representation space is called a graph layout. A correct layout is the base for node-link diagram understanding (Misue et al., 1995). Many layout algorithms are available (Herman et al., 2000b; Tamassia, 2013; McGuffin, 2012) of which force-directed layouts are the most popular for general graphs.

A force-directed layout (also known as spring-embedding, or energy-based layout) algorithm aims to draw a graph based only on its structural information (Tamassia, 2013, Chapter 12). The principle is

16. Network representation differs from *graph drawing*, which focuses on the mathematical properties of graph (such as *planarity*) to tackle specific drawing challenges corresponding to quality measures (such as minimizing the edge crossing).

17. A comparison and a hybridization of both representations can be found in (Henry et al., 2007).

18. There is considerable controversy in the field as to which technique –2D or 3D visualization– works best. This mostly depends on the data and the tasks, as 3D information visualizations are very difficult to set up correctly, often leading to cognitive overload and misunderstandings in interpretation (Sebrechts et al., 1999; Herman et al., 2000b).

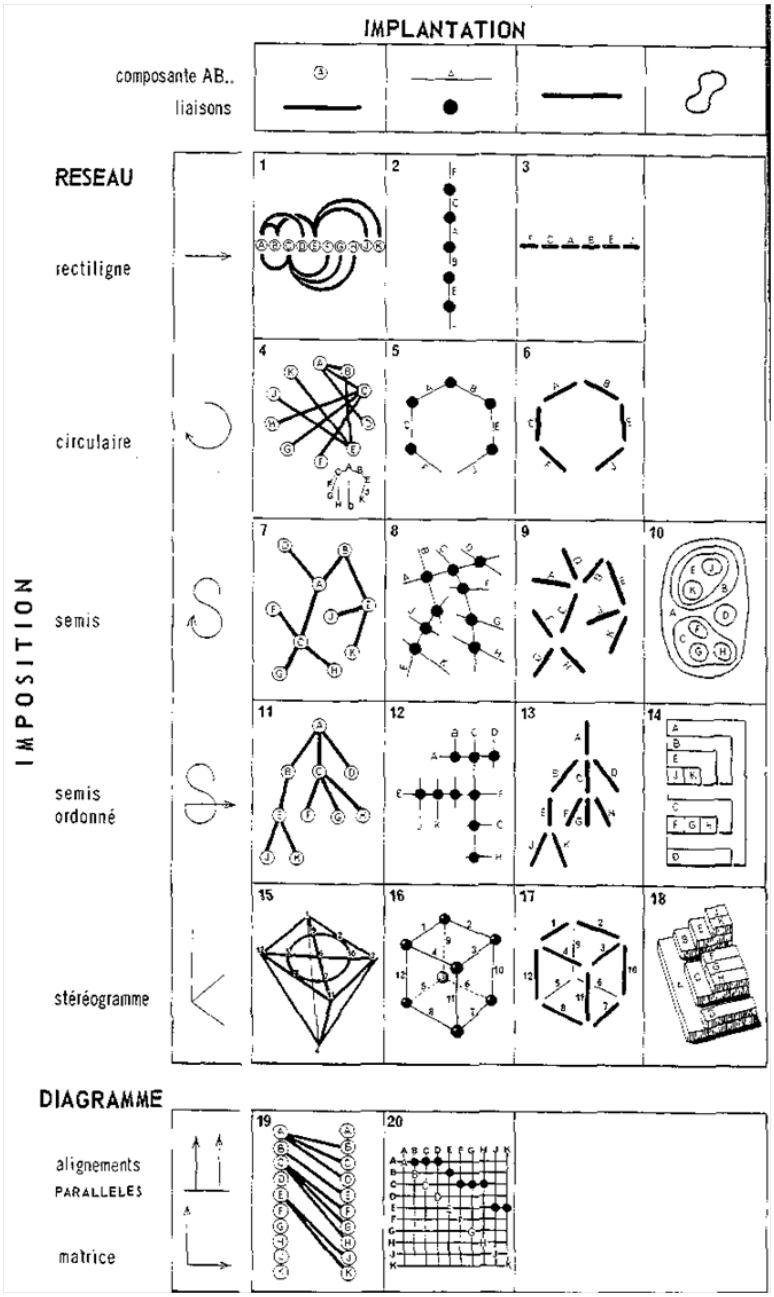


Figure 4.9: The network representation taxonomy from (Bertin, 1967). It not only covers most challenges in representation addressed in (Herman et al., 2000b), but also matrix views and bipartite representations.

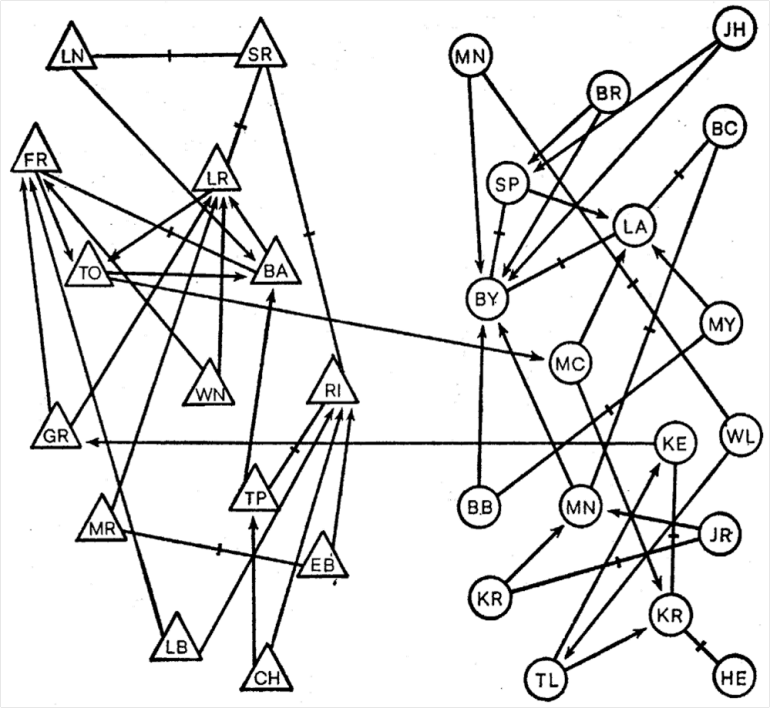


Figure 4.10: One of the oldest node-link diagrams applied to social networks analysis found in the literature, called *socio-gram*, depicting an attraction relationship network between 4th grade students (Moreno, 1934).

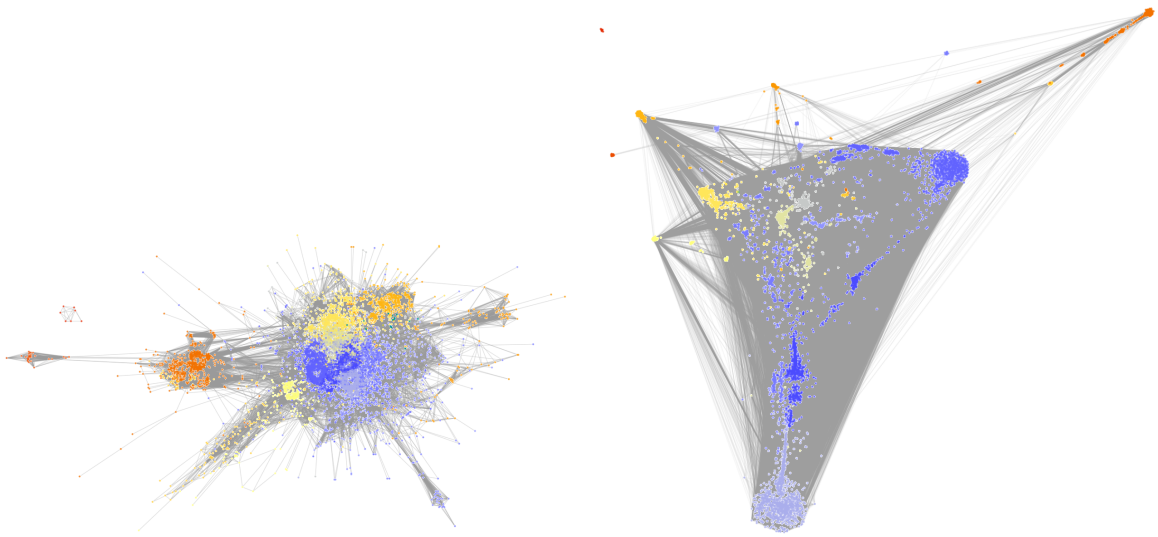


Figure 4.11: Two representations of the document graph of Section 2.3.1, drawn with two different layout algorithms: (left) Hachul and Jünger (2005)'s FM^3, and (right) Noack (2006)'s edge LinLog. Colours correspond to communities as detected by Louvain's method (Blondel et al., 2008).Notice that LinLog's output separates more areas of high density, however computing the layout takes a lot more time.

simple, it embeds forces in the nodes and links of a graph and iteratively moves each node in the layout until it reaches a stable state. An easy analogy would be to represent edges with springs stretched between nodes, the nodes could be initially placed anywhere and once they are freed, the springs' force moves the nodes until equilibrium is reached. Of course oscillations can occur but all algorithms ensure convergence in one way or another. Basic forces in a graph $G = (V, E)$ as defined by Fruchterman and Reingold (1991), define two types of forces, attractive forces $f_a(u, v), u, v \in V$ between nodes that are connected and repulsive forces $f_r(u, v)$ between other nodes:

$$f_{total}(u) = \sum_{v \in N_G(v)} f_a(u, v) + \sum_{v \notin N_G(v)} f_r(u, v)$$
$$f_a(u, v) = \frac{d(u, v)^2}{k}$$
$$f_r(u, v) = \frac{k^2}{d(u, v)}$$

with $d(u, v)$ a distance vector from node u to node v and k the radius around nodes. Three algorithms attracted our attention: *GEM* (Frick et al., 1995) for the aesthetic quality of its output on small graphs, FM^3 (Hachul and Jünger, 2005) for its performances with large networks, and EdgeLinLog (Noack, 2006) for its ability to visually discriminate communities (see Figure 4.11).

4.4.2 *Visualizing complex data structures*

As we have seen in previous chapters, the type of information we process and our methodology always make us consider entities of different types, called *substrates* and *catalysts*. Our goal here is thus to observe how *catalyst* entities influence the interaction between *substrate* entities and it makes sense to focus our interest on bipartite and multivariate network[19] visual representations. We can additionally note significant efforts in hyper-graphs visualizations involving aesthetic Euler diagrams (Verroust and Viaud, 2004; Simonetto and Auber, 2008) and metro map designs (Wolff, 2007).

19. A multivariate network is a network whose nodes present multi-dimensional attributes.

The research community has not yet focused its efforts on bipartite and multivariate networks, and many of the contributions we find are very much oriented towards applications.

Graph drawing approaches to bipartite graph visualization aim at displaying, directly, the bipartite graph as a whole, with significant efforts on minimalizing the edge crossings between each part (Eades and Wormald, 1994; Makinen and Sieranta, 1994; Shahrokhi et al., 2001), of which radial representations would seem to be the current trend (de Klerk et al., 2012; Dumas et al., 2012b,a). Other efforts focus on the drawing of Galois lattices[20] to represent the bipartite network (Duquenne, 1999; Berry et al., 2011), which quickly become very complex. Usui et al. (2007) tackle bipartite network visualization through a very aesthetic spherical embedding. Misue (2006); Ito et al. (2009,

20. A Galois lattice is a model of partially ordered sets often used in Formal Concept Analysis – further reading in (Ganter et al., 1997).

2010) use an interesting method to explore a bipartite graph within an anchored circular (spherical and multi-circular) layout (Figure 4.13, left).

Bipartite graphs have recently been used in the design of a website traffic analysis system (Didimo et al., 2011) and Spanurattana and Murata (2011) try to summarize bipartite graphs with a visualization of three computed measures, but it excludes any possible exploration of relationships. An optional but common strategy consists in projecting the graph inducing relationships between entities of a same type (Zhou et al., 2007a; Douglas et al., 2005; Borgatti, 2012), and even across both projections of a bipartite (Schulz et al., 2008; Bhavnani et al., 2012), with the obvious disadvantage of inducing a lot of edge cluttering[21] (Figure 4.13, right).

21. The *edge cluttering* happens with node-links diagrams that are drawn with so much edge crossings that it becomes impossible to read paths in a static image.

Associations of separated views of different parts of a network can be coordinated *on-the-fly* with a query, as in (Shneiderman and Aris, 2006), through display of edges across one part to the other. The visualization of multi-graphs tends quickly to overstep the domain of hierarchical graph visualization, as in (Dogrusoz and Genç, 2006). Close to that spirit, Di Giacomo et al. (2007) used a bipartite graph to display and cluster web queries results, supported by a tree of snippets clusters based on semantics, while maintaining a proximity graph representation. (Giacomo et al., 2010) provides a nice way to interconnect multiple graphs through bundled edges[22], but following the edges path tend to be quite difficult (Figure 4.13, centre).

22. *Edge bundling* is a technique designed to overcome the cluttering effect by agglomerating edges, with very aesthetic outcomes. Nice examples can be found in (Holten, 2006; Lambert et al., 2010)

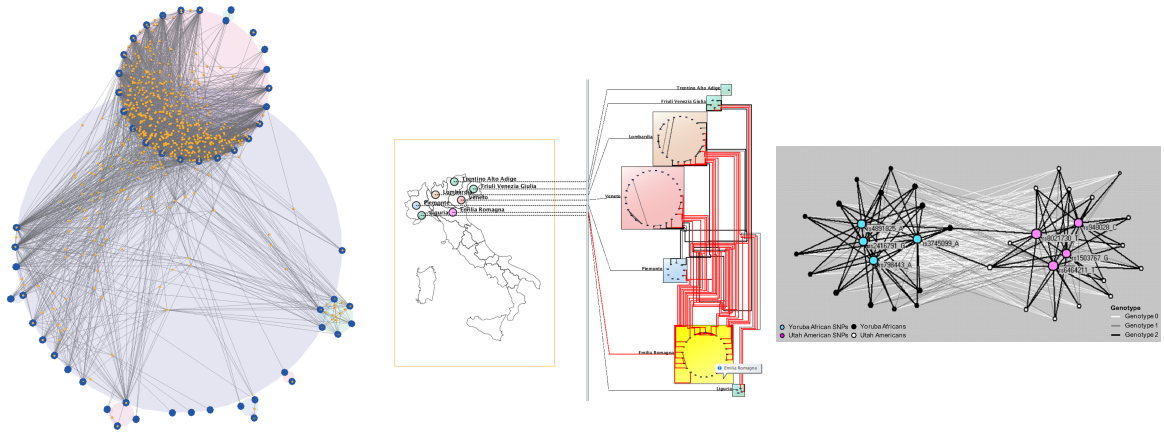


Figure 4.12: Three examples of complex networks representations: anchored radial bipartite layout from (Ito et al., 2010), bundled one-to-many matching from (Giacomo et al., 2010), and dual projections from (Bhavnani et al., 2012).

Multivariate network visualization joins multivariate data visualization with networks that are enriched with many attributes. These approaches benefit from multi-dimensionnal data analysis and embed outputs into network visualization, encoding analysis results into classical node-link diagram representations (De Leeuw and Michailidis, 2000). It very often combines different classical visualizations, with some ad-hoc novel visualizations, in a *perspective*, that is a col-

lection of simultaneous views linked one way or another, to support discovery (Eick, 2000; Stolte et al., 2002; Auber, 2004; Dunne et al., 2012; Wong et al., 2012; Shamir and Stolpnik, 2012). Such combinations, allow for a profuse number of possible applications, and we will present only illustrative examples here[23].

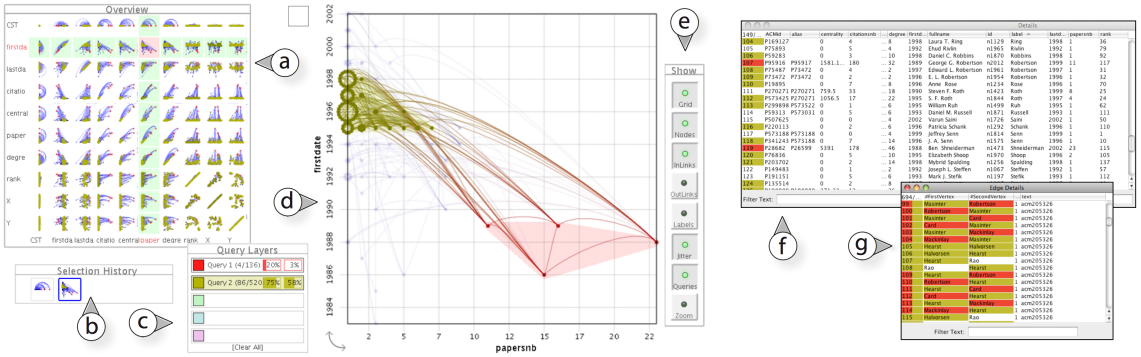
Didimo et al. (2011) designed a system to support web site traffic analysis, placing in parallel two different linkages of a same graph enriched by analysis. (Jusufi et al., 2013) presents a constrained graph layout with regard to the different attributes each node can bear. (Elmqvist et al., 2008; Bezerianos et al., 2010) support multivariate visualization with matrices of scatter plots – a matrix of which each element is a scatter plot with for axes the combination of two attributes (see Figure 4.15). Other authors designed systems that combine a broad set of diverse, creative and interactive visualizations with a core suite of document analysis, comparison, and summarization techniques, see for instance (Chen, 2006; Oelke et al., 2008; Görg et al., 2013; Cao et al., 2010; Paulovich et al., 2012; Cui et al., 2012; Gorg et al., 2013).

Additionally, we can mention two remarkable contributions of multivariate data analysis and visualizations that enable comprehensive displays of points in n-dimensional space. The multidimensional scaling (Borg, 2005), is a field of research dedicated to projections of n-dimensional space, and can be applied to visualization in result of viewable layouts of data points where distance in the viewable space reflects distance in the n-dimensional space; and self-organizing maps (Kohonen, 2001) which are representations of artificial neural networks, arranging – with training – elements based on their proximity in their original descriptive space.

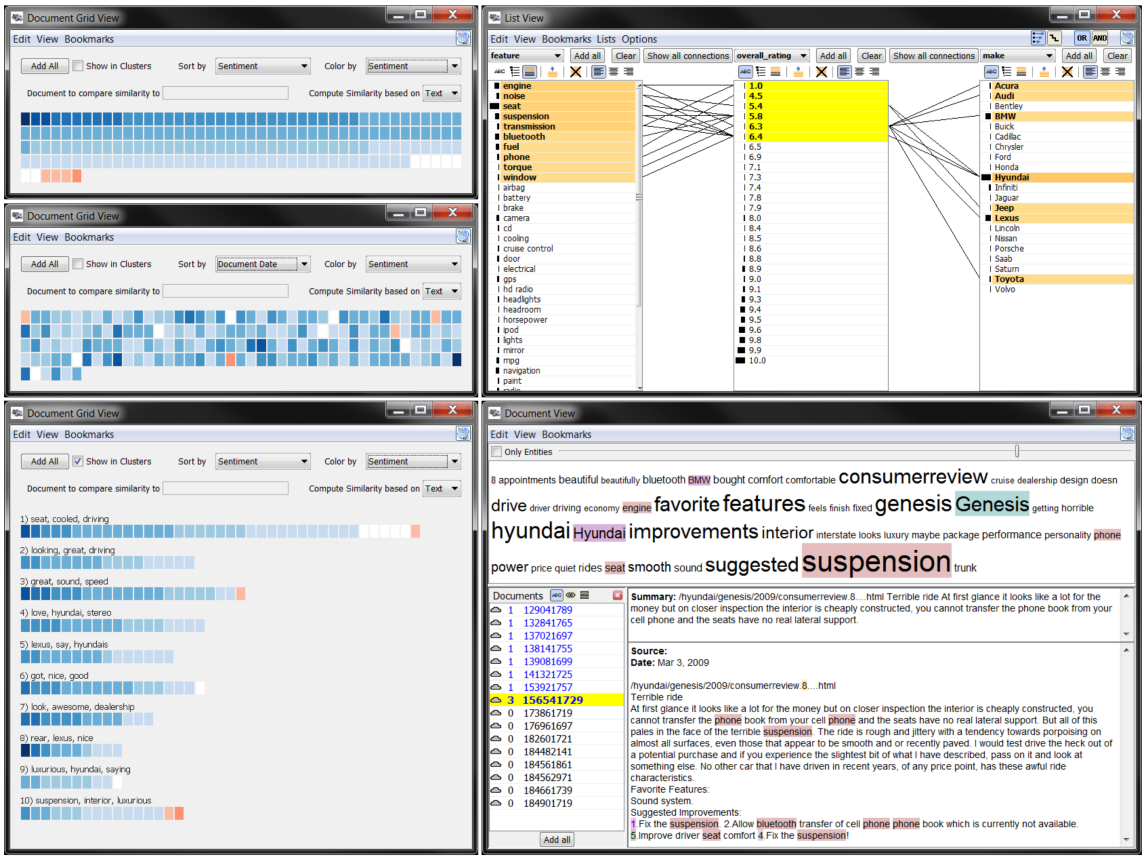
23. The literature is really broad on this subject and our curious reader, interested in visual data analysis, may refer to the many contribution in the field with examples of (Wong and Bergeron, 1997; Grinstein et al., 2001; Fayyad et al., 2002; Keim, 2002).

4.4.3 Enriching visualization with interaction

Many interactions are available for node-link diagrams and multiple-views representations, but, in this section we will only consider works related to common human-computer interface involving a mouse and a keyboard. Many basic interaction techniques are necessary for network exploration. Such interaction are the necessary building blocks for humans to refine their analysis, and Herman et al. (2000b) present four basic classes of interactions for navigation. *Zooming and panning* are the most classical operations when exploring a 2D space (Perlin and Fox, 1993), such as a road map on the internet, our favourite photo editor, or a graph layout (Van Wijk and Nuij, 2003), such operation sets up the context of user's observation. *Magnifying lenses* and other space-deformation techniques are good support features which encourage users to focus on a particular area of a graph whilst keeping the user aware of the overall context (Bier et al., 1993; Furnas, 1986; Tominski et al., 2006) (see Figure 4.14). *Dynamic modifications of the layout* would be another way to focus on specific areas whilst preserving a sense of the context. As an example (Wong et al., 2003)



(a). GraphDice from (Bezerianos et al., 2010)



(b). JigSaw from (Gorg et al., 2013)

Figure 4.13: Two examples of multiple view of multivariate data. (a), from (Bezerianos et al., 2010), is focused on graph representation, and presents a scatter plot matrix. (b), from (Gorg et al., 2013) is focused on documents analysis.

rearrange edges depending on the focus, to limit the cluttering effect, and (Yee et al., 2001) completely rearranges the layout when focused on a node. *Incremental exploration* techniques allow us to navigate a graph step by step, such as the neighbourhood exploration proposed in (Plaisant et al., 2002), that allows us to jump from one node to another; and the *bring and go* technique (Moscovich et al., 2009) (see Figure 4.14) which combines the advantages of magnifying lenses, dynamic layout updates, and incremental navigation.

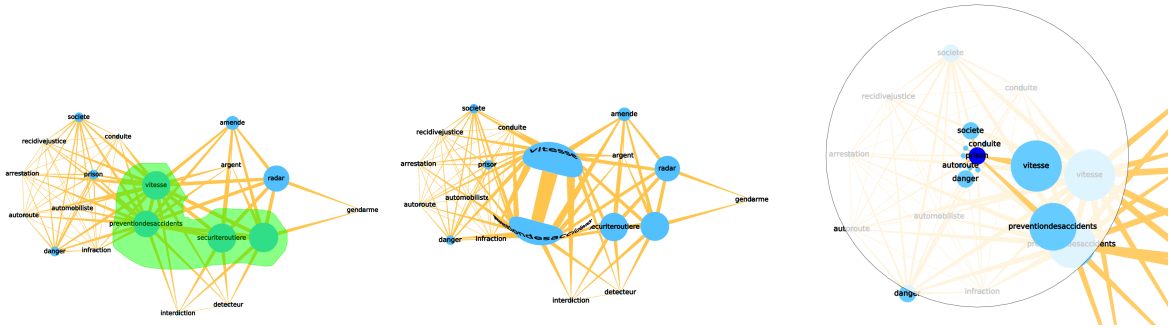


Figure 4.14: Three examples of interactions (implementations from Tulip (Auber et al., 2012)).
(Left), a lasso selection. (Centre), a fisheye view (from (Furnas, 1986)).
(Right), a neighbourhood highlighting, *bring-and-go* (from (Moscovich et al., 2009)).

Another category of interactions allows the selection of particular features of a graph[24]. Manual selection/search on nodes/edges features, search of shortest paths, or of a neighbourhood are also very classical methods (Wills, 1996) (see Figure 4.14). However more advanced selections have improved these approaches (McGuffin and Jurisica, 2009), including an update of the classic lasso selection of elements in a 2D view. Brushing techniques extend selection techniques as early as in (Becker and Cleveland, 1987), where the brush defines a selection that can be dragged within the view space. This tool is the weapon of choice for multivariate visualizations (Becker and Cleveland, 1987; Martin and Ward, 1995), very helpful, for example, in linking parallel-coordinates views (Hauser et al., 2002; Kosara, 2011).

24. Many advanced interactions can also be found in (McGuffin and Jurisica, 2009; Shamir and Stolpnik, 2012).

Linking of multiple views of a complex dataset naturally follows brushing (see Figure 4.15), as in (Chen et al., 2007, Chapter 9) which is dedicated to the topic. Visual views, such as an entry in a scatter plot matrix, are often referred to *facets* (Becker and Cleveland, 1987; Elmqvist et al., 2008). Linking needs mechanisms that insure correspondence(s) between facets to propagate selections (Chen et al., 2007, Chapter 9). Linking also necessitates a visual representation of such correspondences, for which coloured highlighting is a very common procedure (Isenberg and Fisher, 2009; Elmqvist et al., 2008; Kosara, 2011). Linking not only applies to multiple parameters of the same visualization data as in a scatter plot matrix, it also applies on heterogeneous visualizations and heterogeneous data (Wang Baldonado et al., 2000; Konyha et al., 2006; Shrinivasan and van Wijk, 2008;

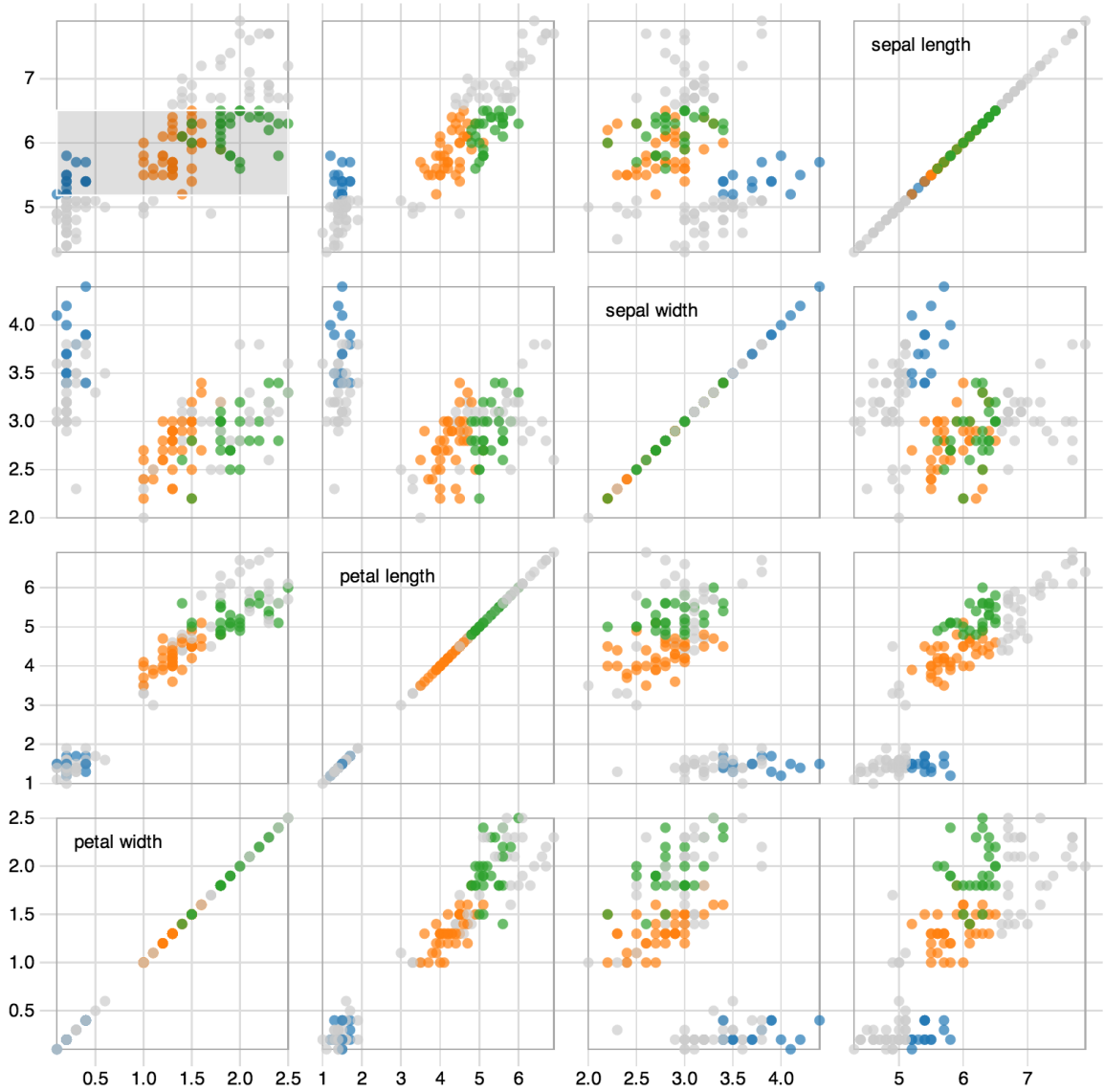


Figure 4.15: Scatter-plot matrix brushing and linking, with d3.js (Bostock et al., 2011). The scatter-plot matrix highlights the selection made in the top-left scatter plot onto every other scatter plots. Brushing allows us to interactively move the selection with on-the-fly updates of highlights in the other scatter plots.

Dork et al., 2012; Chuang et al., 2012; Emerson et al., 2013). This, of course, also applies to network representations (Munzner et al., 2003; Bezerianos et al., 2010).

4.5 Conclusion

In this Chapter, we have presented the many challenges and tasks for complex information understanding, and how Visual Analytics can support it. The realization of a visual analytics framework is a long process centred on the human-machine interaction. The human factor is not a numerical value we can easily control, and its complexity leaves space for many design failures. We have then studied

the rules and the design that would lead a visual analytics framework to success, or at least avoid many possible failures. With a well studied design, the human nature that is part of the interaction can carefully be utilized in order to increase the efficiency of the framework. In that spirit, we have presented the mechanisms of human perception that can be used in the encoding of Information Visualization. We have finally introduced different propositions made by the research community to tackle complex visualization of networks and multivariate data structures and the interactions that bridges the visualization/machine level and the cognition/human level.

It is worth mentioning here that we have opted for a perspective joining multiple views of the multiplex network we are observing. Indeed, interactions with multiple views are valuable and important as they support dynamic comparative analysis, but perhaps more importantly is the human-in-the-loop who is mentally integrating the multiple views and gaining connections, and often realizing insight from those multiple views.

We need to present the substrates, the catalysts, and group-level informations. We have then chosen to base our representations on the very common node-link diagrams. However, due to the complexity of the objects we want to visualize (heterogeneous, different levels), we are pushed to choose very carefully our visual encoding. We unify these views through mental linking and mapping, supported by interaction. Here also, we need to be cautious of the techniques we will use, otherwise we may very easily have our users lost in the system's complexity. That is why the next Chapter will present our choice to assemble very simple techniques, yet empowering complex structures understanding.

VISUALLY COMPREHENDING MULTIPLEX NETWORKS

5

Our work has been motivated by very high-level yet concrete and simple questions. We were initially driven by issues concerning INA's document collections and the groupings of such when presented to journalists and documentalists. The effects of clustering inevitably carry noise, sometimes making the groups difficult to use for higher purpose: *"How come these documents are grouped together?" "How do I know a group of documents really forms a cluster?" "How can I be sure all documents of a cluster really belong to it?" "Should I suspect that the group contains marginal documents?"*.

These questions are driven by curiosity and the desire to make sense of complex information. We are, however, dealing with human information, thus the relevancy of a group of documents is highly subjective, depending on the focus of the analyst concerned. When tackling the different groupings, whatever structure we were enforcing in the data, the problem remains tight and complex. We thus attempted to cope with this issue by exposing the relationships, understanding why the data remained tight, without disintegrating into trivial data subsets, and thus possibly losing a major portion of the information.

The complex network perspective on document analytics exposed us to a much broader field of applications cases which involve any domain facing complex networks. *Understanding* is key to many high-level analytical tasks, such as Amar and Stasko (2004)'s *rationale gap* – "*being able to explain confidence in the relationship, as well as its usefulness*".

We presented, in Chapter 3, analytical tools that can help us unveil the inner structure of complex networks. The objects that result in the entanglement analysis correspond to three levels of granularity of a multiplex network. Substrates correspond to the lowest entity level. Catalysts make an intermediate level that brings groups of substrates together. Entanglement intensity and homogeneity clearly point at the top level, considering the complex network as a whole.

We propose in this chapter, to bring these tools into play, within a visual analytics framework to support *understanding* of complex systems, and in turn, to answer to many domain-dependent tasks. Our framework takes its roots in very simple, yet well adjusted, visualizations and interactions, and even includes a fitted layout algorithm.

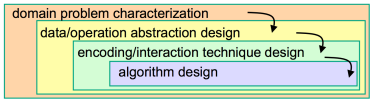
We followed Munzner (2009)'s guidelines for design and Brehmer and Munzner (2013) for validation. We conclude this Chaptter by identifying possible weaknesses in our framework and propose solutions, thus suggesting new leads for future research.

5.1 Tasks and framework design

The highest level task we want to support is the *comprehension* of complex information. As seen in Section 4.1 humans *reason* to *understand* and, since reasoning is a recursive process, *comprehension* of lower level issues is needed to *reason* at higher levels. We are not, at this time, proposing to understand *faster* complex structures, nor to understand *large quantities* of complex information, but we are proposing a *better understanding*, in the sense of an in-depth comprehension of the structure that ties complex information. To do so, we shall lean on our analysis of complex structures using simple representations and techniques.

In this section, we will follow Munzner (2009)'s guidelines[1] for design and validation: *Problem characterization*, *Data operation/manipulation*, *Visual encoding/interaction* and *Algorithm*.

1. Munzner (2009)'s nested model here as a reminder.



5.1.1 Problem characterization

We shall partially answer the *comprehension* task by exposing the complexity of relationships in information that can be modelled as a multiplex graph, and more extensively as a bipartite graph. The model we apply is the one of complex networks, where we observe relationships in a group of entities of type $\mathcal{A}$ (substrates), induced by factors of type $\mathcal{B}$. Each different factor of a relationship can be also seen as an entity of type $\mathcal{C}$ (a catalyst). The relationships between these factors give rise to a network of substrates. The construction of complex structures is described in Section 2.5.1 and takes the shape of two network $G_{\mathcal{A}}$ and $G_{\mathcal{C}}$. With the help of these networks, we can reformulate part of the question from ***Why*** *do we observe* ***this*** *complex structure?* into the two questions *How do substrates interact together in* this *group?* and ***How*** *do catalysts interact together in* this *group?* which imply a third question ***What*** *are the bridges between substrates and catalysts?*.

To answer these very high-level questions we propose using Buja et al. (1996)'s higher level tasks. As pointed out by Amar and Stasko (2004)'s *rationale gap*, to understand the structure of relationships in a complex network, is to expose these relationships and their structure. Buja's *"finding structural patterns"* is possible with the exposition of these relationships, which are, in our case, threefold (relationships across substrates, across catalysts, and across both). We have also seen in Section 3.2.1 that we observe an extremum case in which all catalysts interact through all substrates in a group. Another obvious extremum case occurs when a group of substrates does not show any

interaction. Knowing these extrema, *"making comparisons"*, as proposed by Buja, is possible at any time. *"Posing queries"* would be translated in such a way as to question the structural characteristics of a substrate group in comparison to extremum cases. Finer comparisons can also be established by studying the structural patterns of two different queries.

This leads us to three foci of attention that depend on their domain of application when applying such exploration to real-world complex structures:

- *How close is this group to the idea we have of it (the expected)?*: in this case we are looking for a comparison to a known case, such as an extremum case. For example, we are looking for areas of substrates with a structure close to that which we believe to be optimal.
- *What actually makes this group complex?*: we are looking for the shape of the interaction between catalysts, as an explanation of this situation.
- *Is there anything unexpected, and why?*: we are looking here for outliers, and understanding of why such outlier might appear.

Each different focus is taken *prior* to the inspection of the group of substrates, however, making sense of a group often requires us to jump between these three questions.

Many high-level domain-dependent tasks can find answers starting with such a focus. We see applications of the *understanding* of complex networks in many domains to which they apply. Document networks for journalism and social sciences, social networks, financial networks for policy makers and economists, protein-gene ontology networks in bioinformatics, peer-to-peer networks in computer science, *etc.* Every different domain of application has, of course, a different interpretation of the notion of interaction, patterns of interaction and maximum cohesion. Here is a list of several high-level tasks we have confronted, or which have been suggested to us by many different domain experts (the list is, of course, not exhaustive):

- *Clustering:* When facing the results of a clustering technique of substrates, in which we postulate that catalysts are the reason for a cluster's existence, we can measure if the cluster is optimal and understand *why* some clusters are not as expected; and perhaps see the side-effects of the clustering technique.
- *Semantic cohesion:* In the case of a group of documents with concepts, we can ask what concepts are at the heart of a group of semantic documents, if a group is semantically cohesive.
- *Query:* In case of the result of a query with words, semantics can be ambiguous, which can be observed through the relationships between catalysts (*e.g.* "apple" the company, or the fruit).

- *News events:* Similarly, with news documents associated to keywords, the relationships between catalysts highlights the different facets that narrate the event.
- *Co-authors:* In the case of co-authors and topics, co-authors in the same domain make separate communities who publish on the same topics. Specific interactions can be identified when co-publication *does not* imply cohesion of topics.
- *Finances:* In a network of projects and suppliers from the World-Bank, less cohesive groups of suppliers across projects might mean more equity in the project procurement process.
- *Affiliation networks:* In a network of people affiliated to groups, the organization of groups across people allow to identify specific actors in the groups. Cohesion of groups, such as companies, actors, members of boards, may indicate a possible assimilation of the groups.
- *Peer-to-peer:* In a p2p network, community structures depend on the exchange of files, large communities can gather around a small number of files, and a large number of files can be shared across a limited community.
- ...

5.1.2 Data abstraction and manipulation

Our domain-problem characterization, as proposed by Tamara Munzner (2009), is therefore very abstract (which is quite opposed to Tamara's recommendations), but corresponds in its application to a wide range of concrete questions. We can still propose, as a domain-problem, the *better comprehension of complex networks* – better in terms of depth of understanding, which does not take dynamical aspects nor concerns very large structures.

The framework we propose is dedicated to this specific data abstraction of complex networks as modelled with multiplex networks. The operation abstraction that can lift comprehension corresponds to Buja's high level-tasks.

"Finding structural patterns" is possible by exposing the multiplex network's structure, which is represented by both the substrate interaction network, and the catalyst interaction network. The entanglement measures we propose in Chapter 3 enrich the representation with entanglement indices that explain the role of each catalyst in the network, and the entanglement intensity and homogeneity are also indicators of the complexity of the multiplex network as a whole.

"Posing queries" is one operation that results in a subgraph of the complex network. It can be done in two ways that are *not* equivalent.

- The first type of query is focused on substrates; it allows the *selection* of a subgraph of the substrate-interaction network; finds

its correspondence in the multiplex network; and results in a corresponding catalyst-interaction network enriched by the analysis of entanglement. The correspondence in the multiplex network is built on top of the edges of the selected substrate interaction subgraph[2].

- The second type of query is focused on catalysts; it allows the *selection* of a subgraph of catalysts; finds its correspondence in the multiplex network; and similarly results in a new analysis. Keeping in mind that catalysts correspond to edge types in the multiplex network, the subgraph that is associated with a selection of catalysts in not only a subset of substrate nodes that are connected through catalyst-typed edges, but also a subset of edges that excludes any other catalyst than those selected. The resulting analysis will be then achieved on a multiplex graph with as many layers of edges as the number of catalysts selected[3].

2. For the sake of simplicity, we consider querying only substrate *nodes*, and their induced subgraph. Future works will consider querying substrates nodes *and* edges.

3. Notice that we could, here, take specific links across catalysts into account, and this will be proposed in future works. However, since a catalyst corresponds to a set of substrate edges, we have found it convenient to propose, by default, the ***Or*** operator (as opposed to ***And***). The ***Or*** operator will consider edges across substrates in the multiplex networks if they interact through *any* catalysts (as opposed to *all* catalyst).

The *selection* here is rather abstract, and can be done in many ways, including manual selection. The resulting association, of catalysts from substrates, or of substrates from catalysts, can also be taken as selection. The association of substrates and catalysts is not commutative, so the selection is not symmetric as described in Property 5.1. This enables a recursive process, in which we can *leapfrog* from catalyst to substrates, and vice-versa, until we reach conclusions:

$$
\begin{aligned}
&\textit{Knowing that } \mathcal{P}_{\mathcal{A}} : G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t\right) \mapsto G_{\mathcal{A}}\left(\mathcal{A}, F\right)\\
&\textit{and } \mathcal{P}_{\mathcal{C}} : G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t\right) \mapsto G_{\mathcal{C}}\left(\mathcal{C}, H\right)\\
&R_{\mathcal{A}} : G'_{\mathcal{A}} \subseteq G_{\mathcal{A}} \mapsto G' \subseteq G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t\right)\\
&R_{\mathcal{C}} : G'_{\mathcal{C}} \subseteq G_{\mathcal{C}} \mapsto G' \subseteq G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t\right)\\
&\textit{with } \mathcal{P}_{\mathcal{C}}\left(R_{\mathcal{A}}\left(G'_{\mathcal{A}}\right)\right) = G''_{\mathcal{C}} \textit{ and } \mathcal{P}_{\mathcal{A}}\left(R_{\mathcal{C}}\left(G''_{\mathcal{C}}\right)\right) = G''_{\mathcal{A}}\\
&\textit{then } G'_{\mathcal{A}} \neq G''_{\mathcal{A}}\\
&\textit{also } \mathcal{P}_{\mathcal{A}}\left(R_{\mathcal{C}}\left(G'_{\mathcal{C}}\right)\right) = G''_{\mathcal{A}} \textit{ and } \mathcal{P}_{\mathcal{C}}\left(R_{\mathcal{A}}\left(G''_{\mathcal{A}}\right)\right) = G''_{\mathcal{C}}\\
&\textit{then } G'_{\mathcal{C}} \neq G''_{\mathcal{C}}
\end{aligned}
\tag{5.1}
$$

With $G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_{\mathcal{C}}\right)$ the multiplex graph. $\mathcal{A}$ designates substrates and $\mathcal{C}$ catalysts. G'_X designates a subgraph of G_X. $\mathcal{P}_X$ is the application of the multiplex graph that maps to a substrate/catalyst graph. R_X is the application of a catalyst/substrate subgraph that matches a multiplex graph.

"Making comparisons" in a multiplex network is possible in very many ways.

- At the *entity* level[4], catalysts can be compared to one another with their entanglement index (Section 3.1.4), additionally the network structure allows comparison of entities through their topological characteristics (position in the network, degree, centrality, *etc.* more information on such characteristics in Section 2.3.3).

- At the *query* level, a query corresponds to a sub-network of the complex network $G(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t)$ we wish to analyse. A multiplex sub-network $G'(\mathcal{A}', \sum_{t'\in\mathcal{C}'} E'_t)$ can be compared to its parent network since it ensures the following properties:

$$\begin{aligned} & G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t\right) \mapsto G_{\mathcal{A}}\left(\mathcal{A}, F\right) \text{ and } G_{\mathcal{C}}\left(\mathcal{C}, H\right) \\ & \forall G'\left(\mathcal{A}', \sum_{t'\in\mathcal{C}'} E'_t\right), \exists G'_{\mathcal{A}}\left(\mathcal{A}', F'\right) \text{ and } G'_{\mathcal{C}}\left(\mathcal{C}', H'\right) \\ & \text{such as } \mathcal{A}' \subseteq \mathcal{A}, F' \subseteq F \text{ and } \mathcal{C}' \subseteq \mathcal{C}, H' \subseteq H \end{aligned} \tag{5.2}$$

  Therefore all the structural properties specific to our multiplex network and its entanglement analysis can be compared to those of its sub-networks.

- Queries can also be compared with one another through the same structural properties and entanglement analysis as depicted above.[5]

5.1.3 *Visual encoding*

We will now describe the visual encoding choices corresponding to the three data manipulation tasks.

We have stated, that, in order to *"find structural patterns"*, we need to expose the substrate interaction network, the catalyst interaction network, and to enrich the representation with the different entanglement measures. We propose to do so by offering a *perspective*[6] of three different views (see Figure 5.3).

A first view – the *group view* – presents the measure at the level of a group, entanglement homogeneity and entanglement intensity. This view corresponds essentially to a coloured rectangle area, displaying the numerical values of both entanglement intensity and entanglement homogeneity. To avoid the cognitive load of reading numerical values, we enriched this view with colours. The two numerical values are mapped onto the same colour scale, a 9-class sequential *Yellow-Orange-Brown* scale[7], as recommended in (Harrower and Brewer, 2003)), so we have a different colour for each value, ranging from low hue to high hue, for low values to high values. The colour of the intensity is placed at the *background* of the rectangular area, and the colour of the homogeneity is placed at the *border* of the area, made thick to be clearly seen. The colours are additionally made slightly

4. Not yet integrated into our framework, the entanglement measures at the level of the individual substrate (in Section 3.5.2 and 3.5.3), will also allow comparison between substrates.

5. Although we have no higher-level analytical tasks that have considered this, the analysis of a different multiplex networks of different nature might well be similarly possible.

6. A *perspective* presents simultaneous views of the same information, as defined in Section 4.4.2

7. This scale and more can be found at: http://colorbrewer2.org/index.php?type=sequential&scheme=YlOrBr&n=9

transparent so that they match exactly to those of the lasso, which we will explain a little later.

Two separate views now present each other, the substrate interaction network, and the catalyst interaction network. As mentioned earlier, the nature of the data we are dealing with, and the indices we bring into play, have led us to consider node-link diagrams as the core ingredient for our visualization. However, we are aware that the simultaneous display of two node-link diagrams undoubtedly increases the cognitive load required to manipulate and exploit the visualization[8].

8. *"As with any graphic, networks are used in order to discover pertinent groups, or to inform others of the groups and structures discovered. It is a good means of displaying structures, however, it ceases to be a means of discovery when the elements are numerous. The figure rapidly becomes complex, illegible and untransformable."* Jacques Bertin (1981), p. 129: About why draw a network?

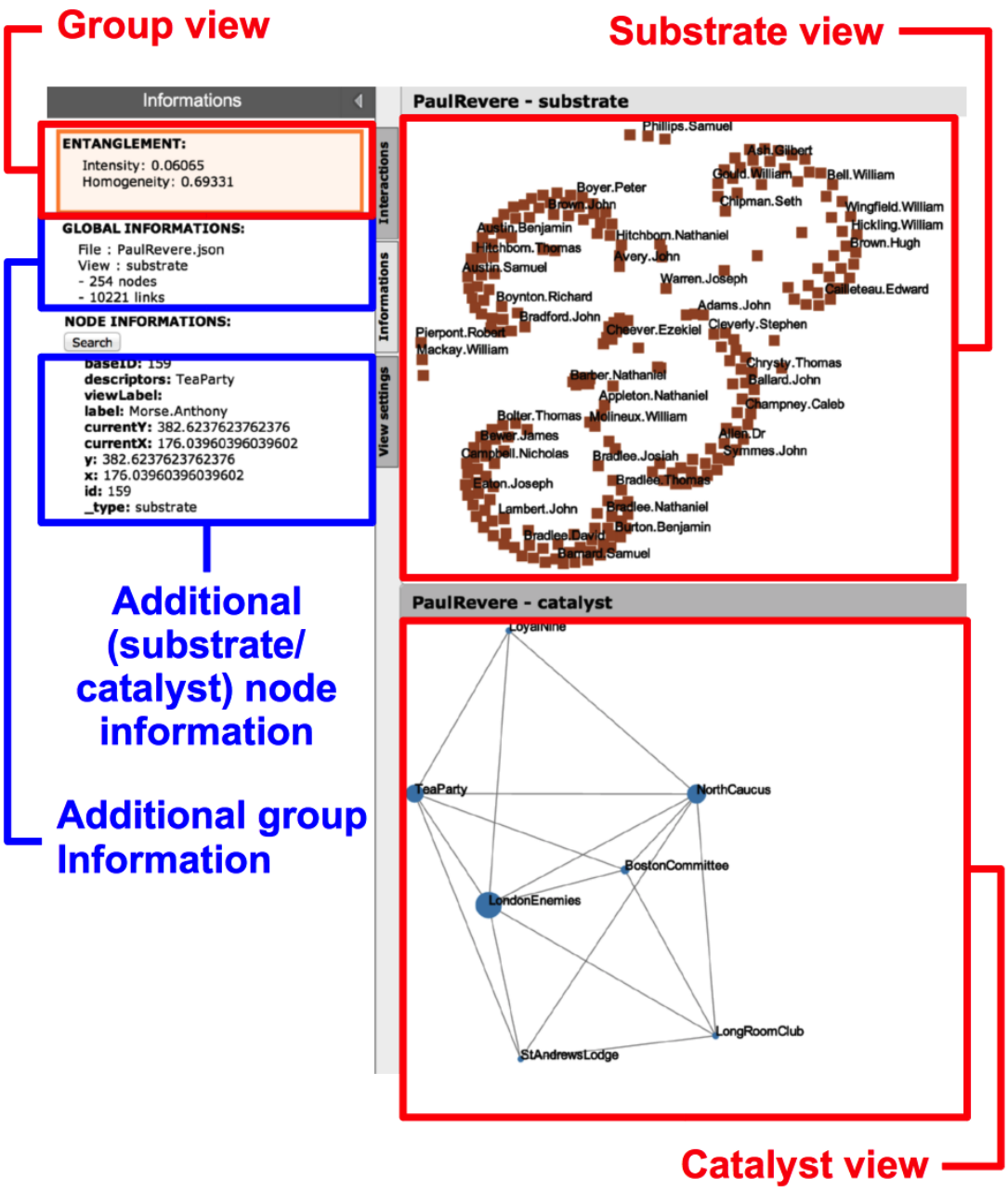


Figure 5.1: The general setting of our framework

As a consequence, we had to design a strategy to display adequate layouts. The layouts of the substrate interaction network $\mathcal{L}_{G_\mathcal{A}}$ and the catalyst interaction network $\mathcal{L}_{G_\mathcal{C}}$ were harmonized so that the position of substrate nodes in $\mathcal{L}_{G_\mathcal{A}}$ would match the catalysts nodes in $\mathcal{L}_{G_\mathcal{C}}$ (Figure 5.6). Our early prototype included a layout for a bipartite representation $G_{\mathcal{A},\mathcal{C}}\left(\mathcal{A}+\mathcal{C}, D\right)$ of the multiplex network $G\left(\mathcal{A}, \sum_{t\in\mathcal{C}} E_t\right)$. In this bipartite representation, a substrate node $u \in G_{\mathcal{A},\mathcal{C}}(\mathcal{A})$ is connected to a catalyst node $t \in G_{\mathcal{A},\mathcal{C}}(\mathcal{C})$ if the same substrate node $u \in G(\mathcal{A})$ is connected to at least one other substrate node $v \in G(\mathcal{A})$ through an edge of catalyst t such as $e(u,v) \in G\left(E_t\right)$. Incidentally, we designed a layout for $G_{\mathcal{A},\mathcal{C}}$ where catalysts would be located closer to substrates they connect (the algorithm is described in Section 5.1.3). The bipartite representation with this layout was soon abandoned because it is rather clumsy and difficult to read. Nonetheless, the layouts for $G_\mathcal{A}$ and $G_\mathcal{C}$ are restriction of this layout for $G_{\mathcal{A},\mathcal{C}}$. Figure 5.6 on top, shows an example; the layouts for $G_\mathcal{A}$ and $G_\mathcal{C}$ obtained from it appear in Figure 5.6 on bottom. The strength of this layout is it preserves a mental map between substrates and catalysts, in which substrates are placed in the same relative area as that of the catalysts to which they relate.

This layout does not try to minimize edge crossings and thus tends to display a high level of edge cluttering, especially when the density of edges gets high. This led us to consider avoiding displaying edges across substrates. This is especially true when the topology of the substrate interaction network does not present any obvious explanatory features such as communities (the substrate interaction is nearly a clique, or a big *hairy ball*). However in some application cases, such as the co-authorship network in Section 5.1.5, keeping edges is important. Particularly in this case, a force-directed layout that displays the community structure is more efficient at explaining the graph. Displaying the edges in $\mathcal{L}_{G_\mathcal{C}}$ is important since the entanglement analysis focuses on the interaction between catalysts[9]. By default, layouts are *harmonized*, and edges are displayed only if they are low in number (we fixed an arbitrary threshold of 1,000 edges). However users are also offered the possibility of manually updating the layout; to switch any layout to a force-directed layout, a multidimensional scaling, and a radial layout; and to display or hide edges (see Figure 5.2). This falls within the *"Reconfigure"* category of interactions as discussed by Yi et al. (2007) by offering users alternative views of the same data structure.

9. However when a lot of interaction in observed across a large number of catalysts, we face the traditional challenges of network visualizations.

Additionally, substrates and catalysts were assigned different shapes and colours (contrasts) so they could easily be distinguished (Figure 5.3). We have also mapped the entanglement index of each catalyst onto its size so dominant catalysts will pop-up in the visualization (in accordance with Buja's *Making comparisons* tasks).

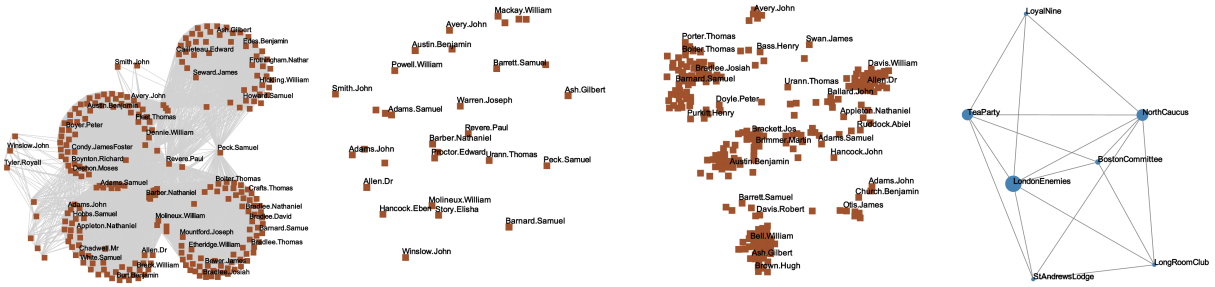


Figure 5.2: Three different layouts for the *Paul Revere* dataset. From left to right, force-directed (with links), multidimensional scaling, harmonized layouts, and the force-directed representation of catalysts. Notice the catalysts drawn with blue circles, and substrates are drawn with brown squares.

5.1.4 *Interactions*

Interaction techniques support navigation and exploration within our complex network. Basic interaction techniques include the aforementioned changes of layout, in addition to node size and colour mapping to topological metrics (*e.g.* degree and centrality).

The – mouse driven – layout navigation is amongst the most classic of methods, allowing zooming and panning such as in a map. We use a *semantic zoom*, as proposed in (van Ham and van Wijk, 2004), such that the size of nodes remains constant at any level of zoom. This enables us to emphasize the *perceived* grouping effect of a layout algorithm at high-level zoom.

Selection is a fundamental interaction, which incidentally forms another interaction category of Yi et al.'s taxonomy, and enables Buja's *"Posing queries"* as discussed in Section 5.1.2, which, in turn, enables *"Making comparisons"* as further discussed in Section 5.3.1.

In our case, having multiple views, *selection* enables link highlighting across views by matching elements in both layouts. It is indeed through selection that users implicitly point at a group asking to be informed about its internal structure and cohesion. The cognitive load of selection is relieved when the layouts match, and also when *selection* is linked through both views (Figure 6.2) as we will explain

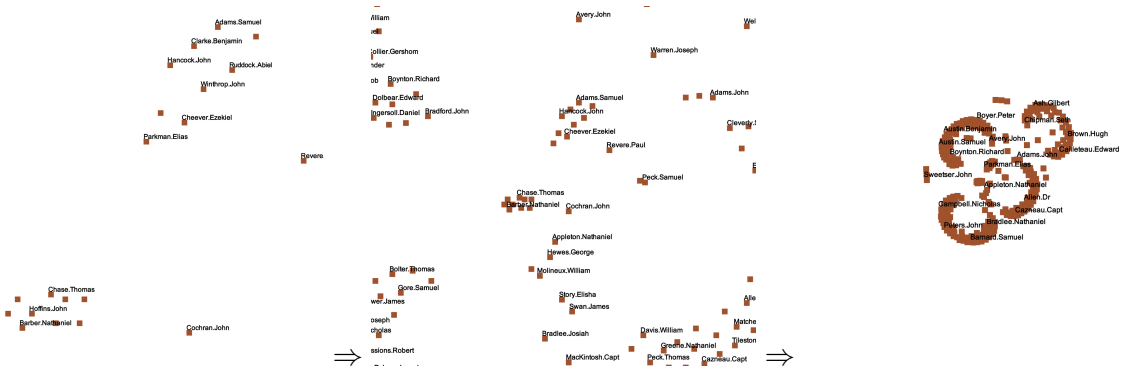


Figure 5.3: Three different levels with "semantic" zooming, for closer on the left to further on the right. The clusters are readable at closer points of view, and aggregate when zooming out.

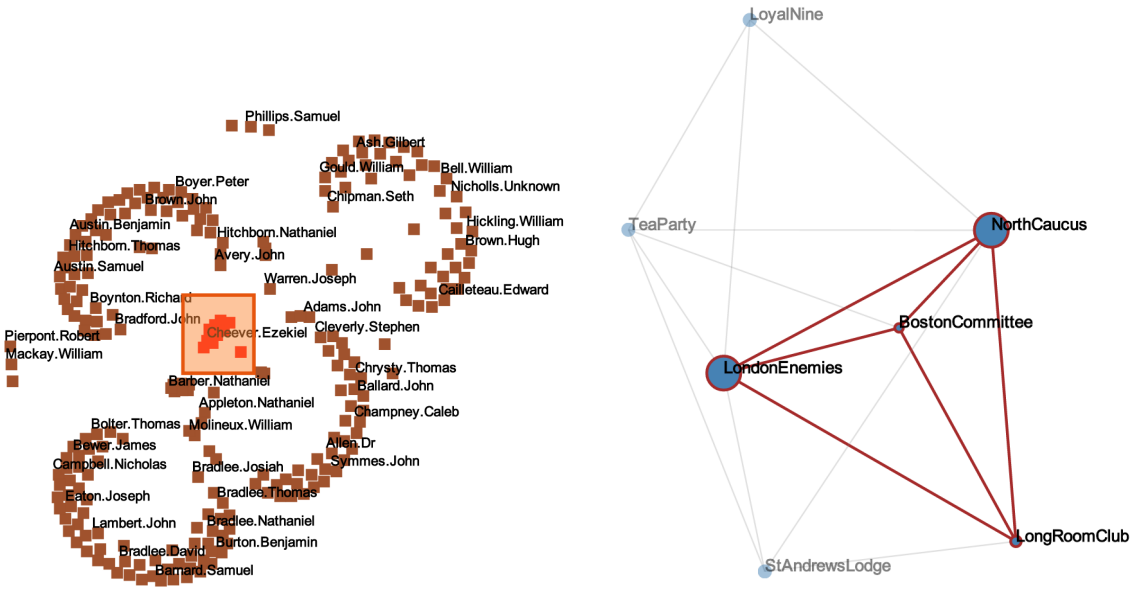


Figure 5.4: A linking operation has been triggered. Selected substrate nodes are highlighted and the lasso updated (on the left). Catalyst nodes (on the right) that correspond to the substrate selection (on the left) are highlighted when the other nodes are dimmed. Notice also the change of size of the selected catalyst nodes (as compared to Figure 5.2, right).

below.

The selection in one view is naturally highlighted (nodes are turned red, Figure 5.4, left), and its corresponding subgraph in the other view is highlighted on-the-fly (Figure 5.4, right). This second highlighting technique not only contours the selection in a red colour, but also dims – without hiding – the other graph elements that do not correspond to the selection. It is upon the selected elements that the entanglement computation is triggered on-the-fly. The size of the concerned catalyst nodes is updated according to their entanglement index in the selection, so is the group view with the new intensity and homogeneity.

Because geometric proximity inevitably results in distortions due to the chosen layout algorithm, we implemented flexible selection mechanisms. McGuffin and Jurisica (2009)'s lasso is used to grasp all elements users want to question. This lasso mode mixes the traditional rectangle-shape, with which everyone is accustomed, with a complex lasso shape, depending on how the user draws their lasso (Figure 5.5).

The lasso thus defines a closed region capturing elements of interest. The closed region and its contour are then assigned colours that reflect the entanglement intensity and homogeneity – the colour encoding is exactly the same as that we used in the rectangle of the group view. Interactive brushing is achieved by moving the lasso around, and users may estimate whether or not the selected group shares a similar structure to its overall neighbourhood.

The selection can be extended/restrained by moving a new lasso with the usual *control/shift*-selection operation. This allows users to

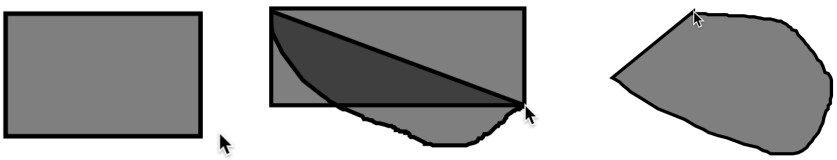


Figure 5.5: A rectangle-lasso selection (middle). The ratio between the starting point / ending point distance, and the total drawn distance allows us to choose between a rectangle (left) and a free shape (right) selection.

detect whether an element has a critical impact on the overall structure of a group, either by adding new elements or discarding already selected ones. These rather classical interactions enter the "Filter" category according to Yi *et al.*, defining the perimeter on which the entanglement computation is to be performed. Incidentally, selection can be also made as a result of a search query on the substrates or catalysts attributes.

Users, through selection, can easily identify subgroups of interest, but they want to identify the substrate outliers to expose the uncertainty of the group of substrates. They may further discard them to reinforce the group's structural cohesion. Users may filter out a set of substrates from a given selection, enabling them to find outliers, and understand which catalysts brought them into play. As a complement, a user can also restrain the study of a substrate group to only its selection, discarding all unselected substrates.

The *leapfrog* selection (see Figure 5.10) allows us to flip the selection over: from any given selection in one view, its associated elements in the other view can be used to trigger a new selection. At the moment this is implemented by an external button, but since it's a frequent manipulation, we are hoping to shorten the interaction access by, for example, associating it to the mouse *middle-click*.

5.1.5 Algorithm

We now tackle Munzner (2009)'s recommendations for algorithm design. The visual encoding and interactions described above rely mostly on two algorithms. The first algorithm, which analyses the entanglement, actually does many computations at once: from a multiplex network, it determines a catalyst-interaction network, and computes entanglement indices, as well as the group's homogeneity and intensity. The second algorithm computes the initial layout given both catalyst interaction network, and the substrate interaction network.

We must be able to compute the entanglement analysis *on-the-fly*. The bottleneck here is not to find a corresponding sub-network of the multiplex network from a selection but the analysis itself. The complexity of the entanglement analysis, as indicated in Section 3.1.5, limits its capacity to run on the fly due to the complexity of the selected multiplex sub-network. Unsurprisingly, the nature of the data

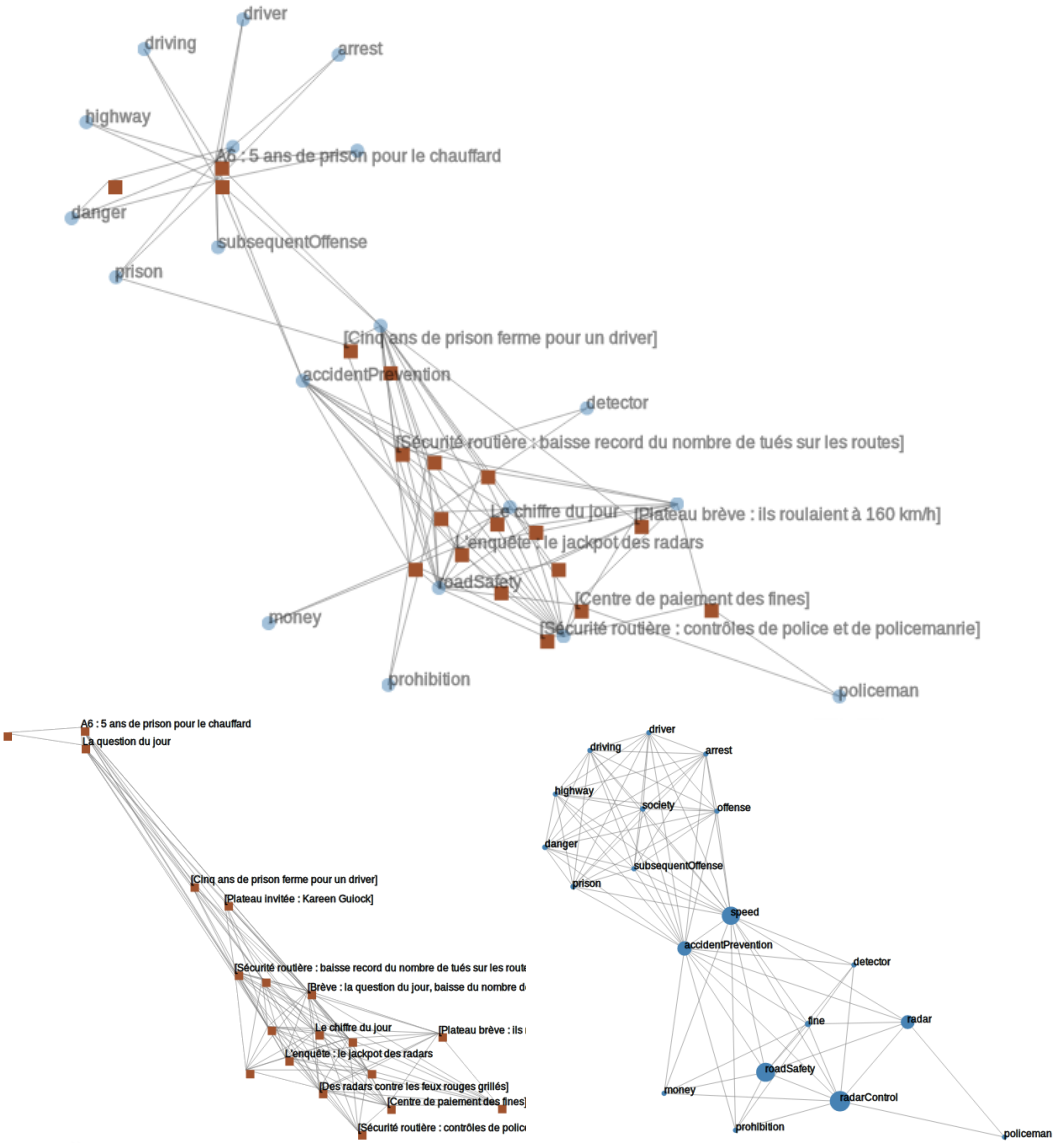


Figure 5.6: The *road safety* bipartite layout $\mathcal{L}_{G'_{\mathcal{A},\mathcal{C}}}$ (top) and the layout for substrates $\mathcal{L}_{G_{\mathcal{A}}}$ (bottom left) and catalysts $\mathcal{L}_{G_{\mathcal{C}}}$ (bottom right) once our algorithm has been applied.

will greatly influence on the speed of the analysis, as that depends on the number of substrate edges and the number of catalyst. Fortunately, we can easily handle a few hundred substrates, with a few dozen catalysts near real-time. The analysis on-the-fly makes sense for small data sets, for, as we argued in the beginning of Section 5.1, our goal is not to be *bigger*, or *faster*, but to offer *better* understanding.

We now detail the strategy we use to coordinate the layouts of the substrate interaction network and the catalyst interaction network (Algorithm 1). Because the substrate interaction network $G_{\mathcal{C}}$ is a central object in our analysis, we first compute a layout $\mathcal{L}_{G_{\mathcal{C}}}$ for this graph. We use a force-directed layout so that interacting catalysts are more likely to be positioned close to one another. Users may optionally edit this layout by moving nodes around, or applying another

layout method.

The layout for the substrate-catalyst bipartite network $\mathcal{L}_{G'_{\mathcal{A},\mathcal{C}}}$ is obtained by aggregating substrates to catalysts, starting from the previously computed layout. That is, catalysts are assigned the same positions they have in $\mathcal{L}_{G_\mathcal{C}}$. From early observations, we found that users give a special focus to substrates related to little occurring catalysts, since they are bringing "details" into the general structure of a substrate group. Following this intuition, catalysts in $G_\mathcal{C}$ are considered as being *general* or *specific*, according to a given rule. In our current setting, we define as *general* catalysts that interact over at least half of the overall substrate edges, and *specific*, catalysts occurring in up to half of the overall substrate edges. The user may use his or her *specific/general* distinction, setting the threshold between general and specific over any measure of the catalysts (*e.g.* degree or centrality) This defines an application $f : u \in \mathcal{A} \mapsto \{v\} \in \mathcal{C}$ that associates substrate nodes and catalyst nodes.

The bipartite graph $G_{\mathcal{A},\mathcal{C}}$ represents all associations of a substrate node of $u \in \mathcal{A}$ *at least to one catalyst* edge of type $t \in \mathcal{C}$. For substrates nodes of $G_{\mathcal{A},\mathcal{C}}$ that are connected to at least one *specific* catalysts, we *only* keep the edges connecting these nodes to *specific* catalysts. If a substrate node is *not* connected to *any specific* catalyst, we keep all its connections to *general* catalysts. The resulting graph $G'_{\mathcal{A},\mathcal{C}}$ is indeed bipartite and covered by the original bipartite graph $G'_{\mathcal{A},\mathcal{C}}(E) \subseteq G_{\mathcal{A},\mathcal{C}}(E)$.

We run Noack (2006)'s layout *anchoring* the catalyst nodes to their initial position – *i.e.* catalyst nodes will not move, but they keep their influence on the network. Hence substrate nodes cluster around their specific terms if any, see Figure 5.6 as an example. The final layout $\mathcal{L}_{G_\mathcal{A}}$ is obtained by simply "forgetting" catalyst nodes and inserting the initial edges between substrates. The mental map between $\mathcal{L}_{G_\mathcal{C}}$ and $\mathcal{L}_{G_\mathcal{A}}$ is therefore preserved. Users can easily locate substrates related to specific catalysts, since they are located in the same areas of both graph representations. An example is provided in Figure 6.2 showing (top) $\mathcal{L}_{G_\mathcal{C}}$, and (bottom) $\mathcal{L}_{G_\mathcal{A}}$. Substrates in the lower part of $\mathcal{L}_{G_\mathcal{A}}$ are related to the lower catalysts of $\mathcal{L}_{G_\mathcal{C}}$, as the highlighting suggests.

5.2 Design validation

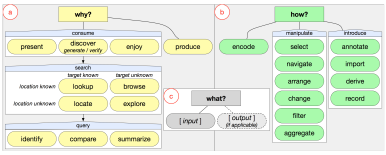
We now present a multi-level description of our tasks and implementations with the typology we discussed in 4.2.2, as proposed by Brehmer and Munzner (2013). This description separates the highest level of abstraction – *i.e.* the analytical sense-making process – from the low-level implementation – *i.e.* visual encoding and interactions. The tasks are described in terms of *why?*, *how?* and *what?* using the categories Brehmer and Munzner proposed. However some categories group a series of subcategories, and when our description

Data: $G_{\mathcal{A}} = (V_{\mathcal{A}}, E_{\mathcal{A}})$, $G_{\mathcal{C}} = (V_{\mathcal{C}}, E_{\mathcal{C}})$, $f : u \in V_{\mathcal{A}} \mapsto \{v\} \in V_{\mathcal{C}}$
Result: $\mathcal{L}_{G_{\mathcal{A}}}$
Compute or set up a layout $\mathcal{L}_{G_{\mathcal{C}}}$
$G'_{\mathcal{A},\mathcal{C}} = (V'_{\mathcal{A},\mathcal{C}}, E'_{\mathcal{A},\mathcal{C}})$
for $u \in V_{\mathcal{C}}$ **do**
  $V'_{\mathcal{A},\mathcal{C}} \leftarrow u$
  $\mathcal{L}_{G'}(u) \leftarrow \mathcal{L}_{G_{\mathcal{C}}}(u)$
end
for $u \in V_{\mathcal{A}}$ **do**
  $V'_{\mathcal{A},\mathcal{C}} \leftarrow u$
  for $v \in f(u)$ **do**
    $E'_{\mathcal{A},\mathcal{C}} \leftarrow (u, v)$
  end
end
while *force-directed layout is not finished* **do**
  if *node* $u \in V_{\mathcal{C}}$ **then**
    Do not update $\mathcal{L}_{G'_{\mathcal{A},\mathcal{C}}}(u)$
  else
    Update $\mathcal{L}_{G'_{\mathcal{A},\mathcal{C}}}(u)$
  end
end
for $u \in V_{\mathcal{A}}$ **do**
  $\mathcal{L}_{G_{\mathcal{A}}}(u) \leftarrow \mathcal{L}_{G'_{\mathcal{A},\mathcal{C}}}(u)$
end

Algorithm 1: Harmonizing both layouts: The application f allows the construction, from the substrate graph $G_{\mathcal{A}}$ and the catalyst graph $G_{\mathcal{C}}$, of a bipartite graph $G'_{\mathcal{A},\mathcal{C}}$. The layout of the catalyst graph $\mathcal{L}_{G_{\mathcal{C}}}$ allows the computation of each substrate node's position $\mathcal{L}_{G'_{\mathcal{A},\mathcal{C}}}(u)$ in the bipartite graph. These positions are then assigned back to form the layout of substrates $\mathcal{L}_{G_{\mathcal{A}}}$.

refers specifically to the parent category, it also implies all the subcategories. We will refer to the terms use in this typology, pictured in Figure 4.5[10], and further details on all the different categories can be found in Brehmer and Munzner (2013).

10. Reproduced here for convenience:



5.2.1 *Making sense of a multiplex network*

At the highest level, multiplex network analysis tasks can be described as in Table 5.1. In this case, the task *Consume* includes all three tasks *Present*, *Discover* and *Enjoy*. *Produce* corresponds to the creation of new data structures and visualizations; *Discover* corresponds to "the generation and verification of hypotheses" and includes the whole *Search* and *Query* subcategories. *Encode* refers to visual encoding; *Derive* refers to the construction of persistent information derived from the input information; *Manipulate* here includes the whole set of subcategories *Select*, *Navigate*, *Arrange*, *Change*, *Filter*, and *Aggregate*.

This level is purely abstract with networks and measures combin-

Task	*Why?*	*How?*	*What?*
Entanglement analysis of a multiplex network	*Consume* + *Produce*	*Encode* + *Derive*	*In*: multiplex network *Out*: substrate interaction network + catalyst interaction network + catalyst measures + group measures
Focus on substrate interaction	*Discover*	*Manipulate*	*In*: substrate network *Out*: multiplex sub-network
Focus on catalyst interaction	*Discover*	*Manipulate*	*In*: catalyst network *Out*: multiplex sub-network
Combine all three levels of granularity	*Produce*	*Derive*	*In*: substrate interaction network + catalyst interaction network + group measures *Out*: Insights

Table 5.1: The description of the multiplex network analysis, according to Brehmer and Munzner (2013)'s typology.

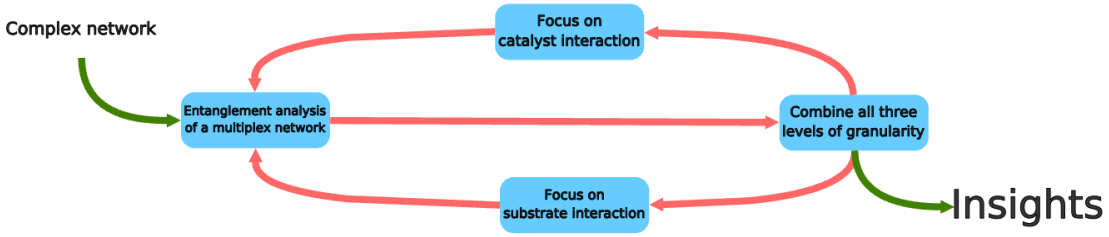


Figure 5.7: The high-level sense-making process of a multiplex network with entanglement analysis, with inputs and outputs as described in Table 5.1. The analysis starts from the complex multiplex network, and iteratively refines to insights.

ing data abstraction/manipulation and visual encoding/interaction. These tasks form a sequence, pictured in Figure 5.7, that re-inforces the recursive structure of the sense-making processes we have seen in Section 4.1. *"Focus on substrate interaction"* and *"Focus on catalyst interaction"* include the whole analysis of the representation of the network, eventually ending in the *selection* of a sub-network that matches a subset of the original multiplex network. This particular description emphasizes the importance of *leapfrogging* from one structure to another as we will further discuss in Section 5.3.2. Here, *entanglement analysis* is seen as a combination of multiple foci. Although *insights* could be found in both steps *"Focus substrate interaction"* and *"Focus catalyst interaction"* depending on the domain dependent task we want to answer, we emphasize the role of the *entanglement analysis* as the main source when seeking *insights*.

5.2.2 *Visual encoding and interactions*

Table 5.2 describes the different tasks of visualization and interaction. The task of *Visualizing a multiplex network* is high-level and encompasses all the visual encodings necessary to generate the whole

perspective of three views (substrate interaction, catalyst interaction, and group views). It also regroups the two interactions (zooming and panning) necessary for a basic exploration of *one* view. A second set represents advanced interaction techniques that allow selection brushing and linking through views. We describe the selection mechanisms first, and then the linking procedure. The linking procedure includes a few analytical steps that are included here. Although analytically speaking, the steps are exactly the same as the first analysis of the multiplex network, we make a distinction here, because implementing the visualization process is slightly different: we do not need all the new networks information, and we want to emphasize *comparison* by highlighting the results of a query. The last set represents filters that allow one to restrain the exploration, from the original multiplex network, to only a subset of this network.

As above, we used categorial typologies when all their subcategories apply. *Query* refers to *Identify*, *Compare* and *Summarize*. *Select* applies when the system allows differentiation of selected and unselected elements. *Filter* reduces the data observed; *Aggregate* relates to the production and aggregation of information at a different granularity level – note, however, that we are manipulating, simultaneously, three levels of granularity in this framework (substrates, catalysts, and group). *Navigate* naturally refers to updates of a user's viewpoint; *Change* applies to modifications of visual encoding. *Arrange* corresponds to updates in the arrangement of visual elements of the display. The interconnection of these tasks, rather complicated is presented in Figure 5.8.

5.3 Discussions and perspectives

We are here discussing further the choices we have made to relieve our users' cognition, with visual cues that use mental mapping and associations, and how *leapfrogging* is an interaction that drives reasoning to insights. This opens new ways to handle the challenge of multivariate multiplex network visualization.

5.3.1 Mental mapping and linking

It is worth pointing out the different linking mechanisms we used to relieve the cognitive load of such complex visualization, enabling Buja's *Making comparisons*.

The first mechanism is the layout discussed in Section 5.1.5, which preserves a common layout mapping of both layouts. The semantics of this mapping depends, of course, on the selection criteria of the algorithm, but the criteria we propose allow us to position substrates relatively to the catalysts that make them *specific*. This is of course a weak approximation that shades many characteristics of an individual substrate. However, by definition, the structure of the catalyst interaction graph indicates their interactions, and the hypothesis that

Task	*Why?*	*How?*	*What?*
Visualization of a multiplex network	*Consume + Produce*	*Encode + Derive*	*In*: an entanglement analysis *Out*: a *perspective* of three views
Draw/change a layout	*Produce*	*Encode + Arrange + Derive*	*In*: network matrix *Out*: node(-link) layout in the view space
Compute and encode a measure	*Produce*	*Encode + Change + Derive*	*In*: (multivariate) network data + layout *Out*: visual mapping onto node glyphs
Harmonize layouts	*Discover*	*Arrange*	*In*: the multiplex network and catalyst layout *Out*: a new catalyst layout
Pann a view	*Consume*	*Navigate*	*In*: a viewpoint on a view *Out*: a new viewpoint on this view
Semantic zoom in a layout	*Consume*	*Navigate + Arrange*	*In*: a scale factor on a layout *Out*: a new scale factor on a layout
Brushing and linking a multiplex network	*Produce*	*Encode + Derive*	*In*: a selection in a view *Out*: the whole updated *perspective*
Query (Lasso / Filter / *Leapfrog*)	*Query*	*Change + Select + Filter*	*In*: a view and layout *Out*: a selection of nodes in a view + remove all highlights + highlight this selection
Brush a view	*Discover*	*Arrange + Change*	*In*: a lasso selection *Out*: a new lasso selection
Link a selection to the multiplex network	*Produce*	*Encode*	*In*: a selection of nodes in a view *Out*: a corresponding multiplex sub-network
Compute entanglement measures	*Produce*	*Aggregate*	*In*: a multiplex network *Out*: an entanglement analysis
Highlight a sub-network analysis	*Discover*	*Encode + Select + Change*	*In*: an entanglement analysis *Out*: highlight of *corresponding* sub-network in substrate/catalyst views + mapping onto catalyst node + mapping onto lasso + update of the *group view*
Filtering a multiplex network	*Produce*	*Encode + Derive*	*In*: a multiplex network *Out*: a multiplex sub-network
Discard a selection of substrates	*Query*	*Filter*	*In*: a selection of substrates *Out*: a multiplex sub-network
Restrain to a selection of substrates	*Query*	*Filter*	*In*: a selection of substrates *Out*: a multiplex sub-network

Table 5.2: The description of our visualization and interactions, according to Brehmer and Munzner (2013)'s typology. Note that the data abstractions (as input or output of tasks) bridge this visualization task description to the higher level tasks described in Table 5.1.

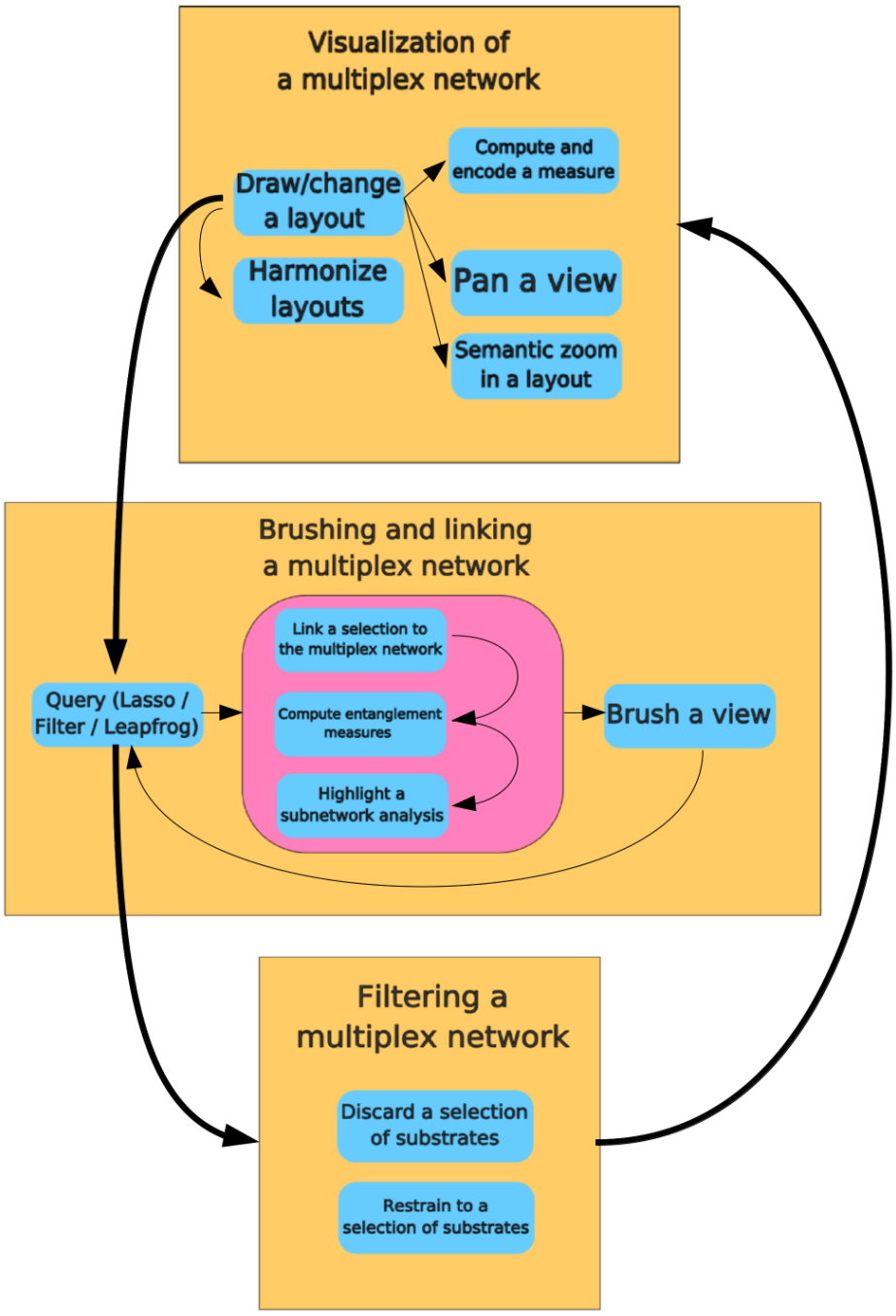


Figure 5.8: The interconnection of our visualization and interaction tasks, as described from their inputs and outputs in Table 5.2. Blue areas correspond to tasks. Orange areas represent high-level visualization and interactions, the pink area emphasizes a necessary analytical process.

a *specific* catalyst neighbourhood might indicate other more *general* catalysts seems reasonable. Hence, positioning substrates based on the suggestions of our layout algorithm *is* still approximate, but not as weak as one can imagine, see Figure 5.9, middle.

Our second mechanism is very straightforward, the colour encoding between the lasso and the group view is exactly the same. However the lasso needs a level of transparency otherwise it is simply impossible to draw, verify and brush a selection. Fortunately, Harrower and Brewer (2003)'s colour scales do not mind transparency, and for a perfect mapping, nothing stops us applying the exact same transparency factor to the rectangle of the group view, see Figure 5.9, bottom.

Finally, our third mental mapping is possible thanks to the properties of the multiplex network and of any selection as presented in Property 5.2 (Figure 5.9, top). Since any subgraph $G'_{\mathcal{A}}$ has a corresponding subgraph $G'_{\mathcal{C}}$ and vice-versa, the layout and properties of $G'_{\mathcal{C}}$, and respectively $G'_{\mathcal{A}}$, naturally benefit from the layout of their parent set. Although such subgraphs could show a totally different topology (such as a clique or communities) we believe the preservation of the mental map, established at the whole group level, is more beneficial than re-computing a new layout. Apart from the obvious computing gain, the preservation of the layout allows *comparison* between different selections. It is true that more cognitive resources will then be needed to figure out the topological pattern of the resulting layout, however a lot more cognitive resources would be necessary to compare completely different layouts[11].

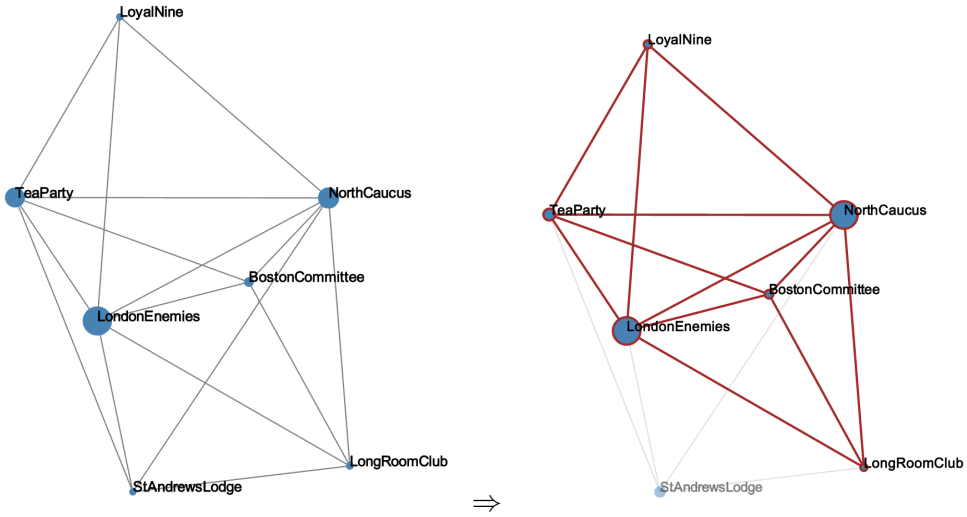
We see our mental mapping and linking mechanisms as key elements that bridge our three representations of levels of granularity: the substrates, the catalysts, and the group.

11. This a well known issue in dynamic graph visualization, to which mental map preservation has been proven important (Purchase et al., 2007).

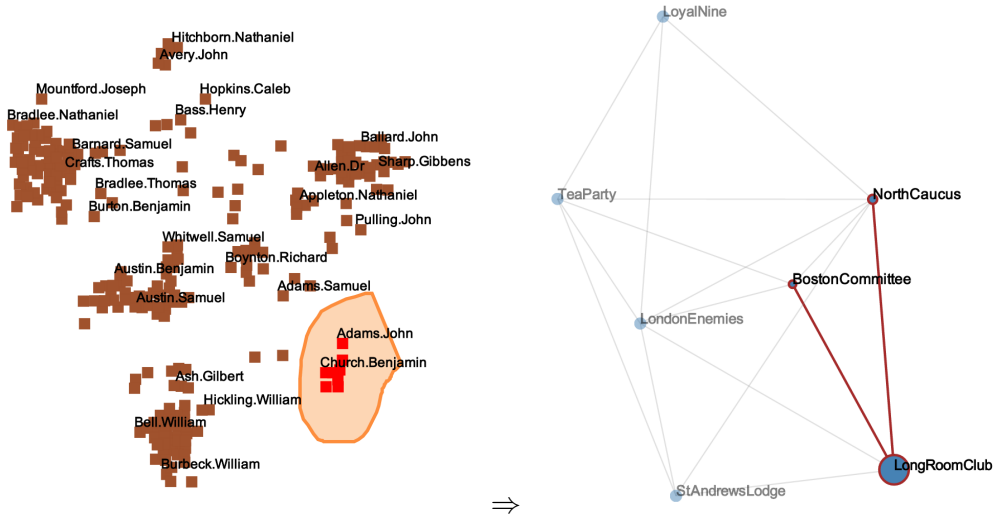
5.3.2 *Leapfrogging*

We call *leapfrogging* the repeated exploration process that makes our focus literally leapfrog from one point of view on the substrates representation, to its corresponding catalyst representation. *Leapfrogging* is not something that occurs only at the level of visualization (as in Figure 5.10), for it also occurs in our minds when we study a complex network. It was originally achieved by repeatedly selecting entities from one view, then switching to another, and we particularly noticed that this process was being repeated when users started their exploration with a selection of catalysts. The case usually started with one catalyst that the user selected, moving on to the corresponding substrate nodes, to finally study which catalysts were in play among the selected nodes

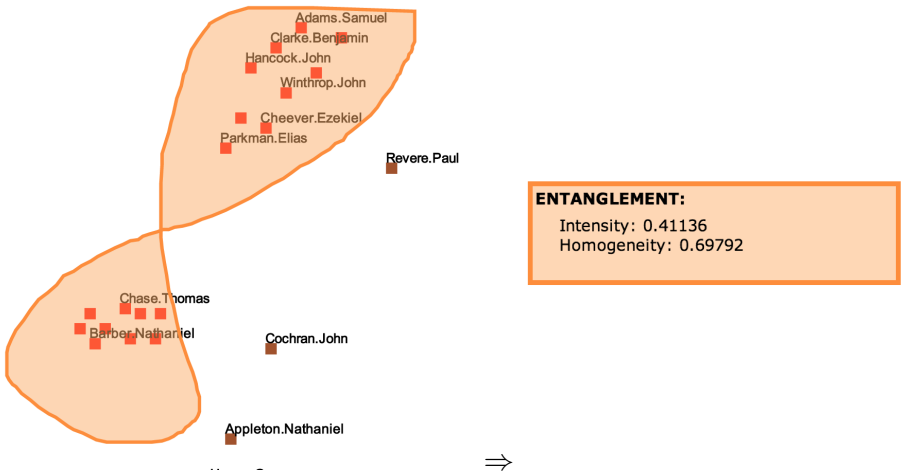
One very simple example was the one with news documents, starting a query from a specific term, to check its entanglement with all the other terms involved across the corresponding documents. If the entanglement measures were optimal, it would mean the whole



Mental linking of the sub-networks within the same map (here on the catalysts)



Mental map preservation between catalysts and substrates



colour mapping preservation between selection brush and group measures

Figure 5.9: Three techniques that preserve similar encoding allowing pre-attentive processing to “associate” elements of the visualization.

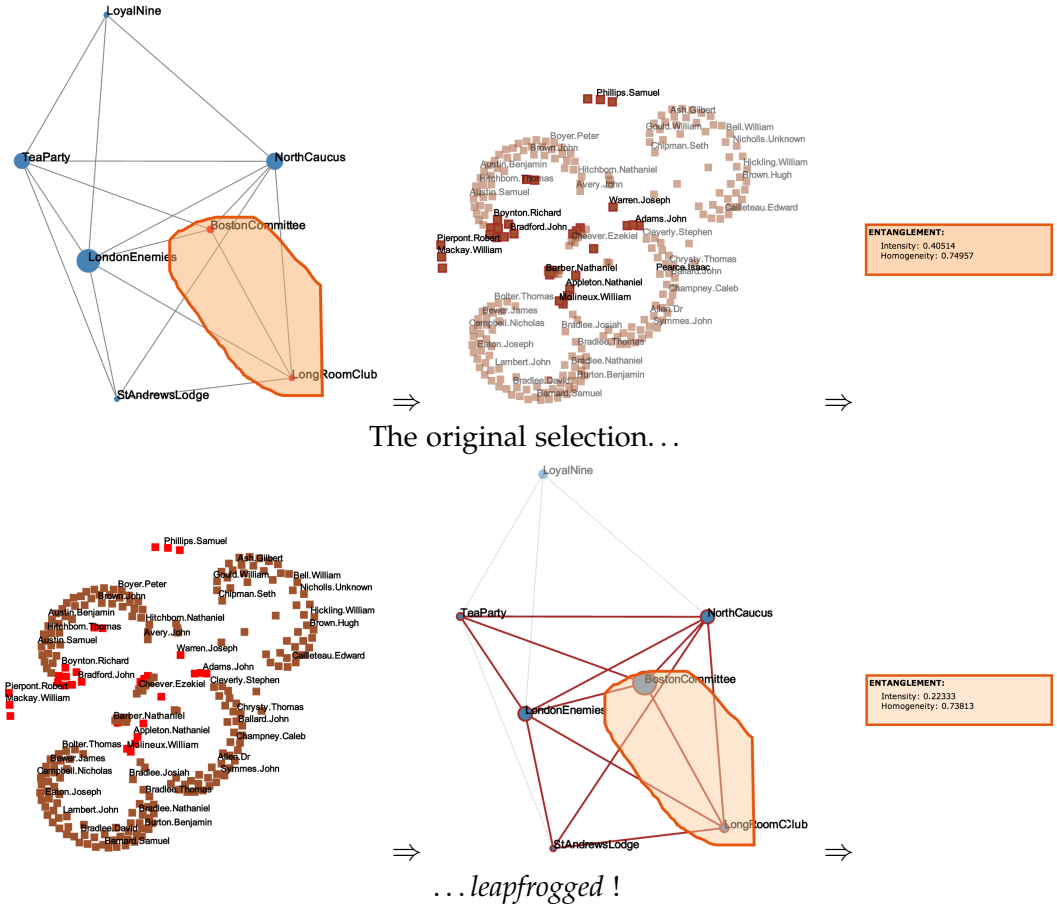


Figure 5.10: *Leapfrogging* from the results of a selection (top), to a new query (bottom). Notice all the visual changes and mapping updates.

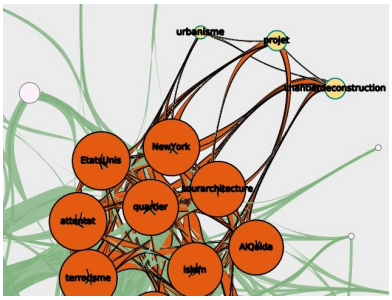
vocabulary formed a cohesive tie across the documents the queried term annotates. On the contrary, with low entanglement measures, we could identify dominant terms that annotated the concerned documents.

The importance of the process was so striking that in our first prototype we implemented the leapfrog immediately on catalyst selection[12]. It induced two entanglement analyses for one selection of catalysts, and they were visually mapped on top of each other with the same colour coding, but with different catalyst node highlights to distinguish both cases (we offered contrast with a different border colour). The problem was not computational, but cognitive. It was extremely difficult for users who were not familiar with the concept of entanglement analysis to cope with the additional complex representations, as there were too many visual variables corresponding to rather complicated processes.

We eventually abandoned this double process and replaced it with a straightforward *leapfrog* selection query, which involves the user in the process. Although we have not formally evaluated it yet, this second solution seems much appreciated by all users.

Our non-commutative mapping, combined with the properties of

12. Here is an example of catalysts leapfrogging in our first prototype, the original selection (here in yellow) shows little entanglement intensity, but the leapfrogged selection (in orange) shows higher intensity.



our substrates and catalysts networks (in Property 5.2), guaranties convergence to *every possible* corresponding node in the selection when we use the catalyst *or* operator for mapping, and to *not any* corresponding node in the selection when we use the *and* operator. This is especially useful when we query larger groups. Manual updates of the selection (for example, discarding an outlier from the selection) and flexible changes of the catalyst's correspondence operator, enable a smooth back-and-forth selection process in the visualization.

Leapfrogging appears to be not only a convenient means of interaction, but, perhaps, it is also the key concept that leads to *comprehension* of phenomena in a multiplex network.

5.3.3 *Multivariate information*

We have barely mentioned, in this Chapter, the multivariate nature of the information, choosing, instead, to focus on the manipulation of complex networks. However, real-world data often comes with much external information which can be attributed to nodes and edges in the multiplex network. Of course we have not ignored this information, and we can find two very convenient uses of it.

The first use naturally follows from the entanglement analysis we are presenting in this manuscript. Any external attribute is one more reason to create new families of ties in a multiplex network. We can use these families, focusing on substrates, and creating many coordinated catalysts views. It might induce projections that could prove an interesting analysis challenge (Figure 5.11). It would also induce another visualization challenge, that of linking all these views, a small multiple visualization, such as a "catalyst–matrix" in the spirit of "NodeTrix" (Henry et al., 2007), with multiple asymmetric coordinations, which would also be a challenge for someone to tackle.

The second use would be simpler. We could consider the multivariate attributes as possible criteria for visual mapping in node-link diagrams, and search and filter. We would also hope to coordinate with our framework of classical visualizations, such as histograms or scatterplots when applicable. This is driven by the reality, on

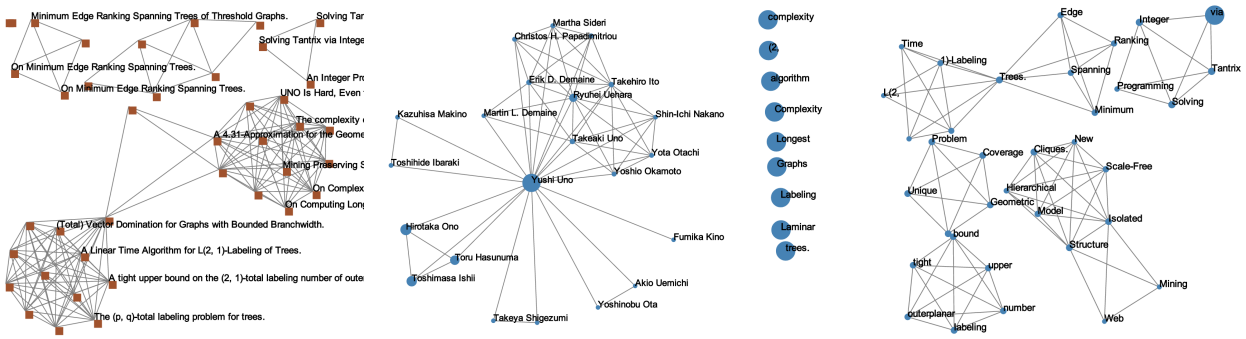
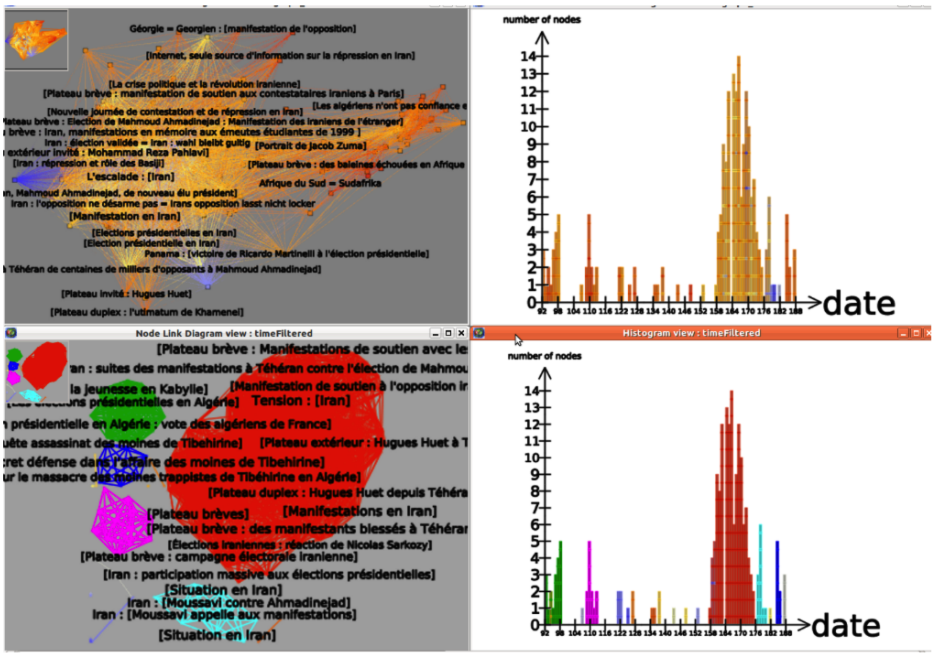


Figure 5.11: Two different types of catalysts over this co-authorship network. Publications are substrates (left), authors are catalysts (middle), and also title vocabulary (right).



Segmenting the substrate network over time.

Linked multiple views, with selection linking.

Figure 5.12: Using external information. On top, an early prototype to distinguish different slices of time concerning a same news cluster. Below, current development of linked highlighting, also across traditional views such as the scatterplot on the left, substrate nodes are mapped onto external measures as well.

the ground, of real-world applications, where such traditional views make excellent tools for analysis – for example nothing is more intuitive than a well set scatterplot to give us an idea of the comparison of distributions between two variables (Figure 5.12). Using external attributes as filters also leads to other interesting tasks – such as comparing how two different broadcast channels approach the same news event. Even better, we began to inspect the notion of time in just such a multiplex network, enabling comparisons between slices of time. This particular research is in its infancy (Figure 5.12), and further work will definitely include these approaches.

5.3.4 *Layout, visual encoding and interactions*

We now examine the structure and the nature of the original data questions, the relevancy of node-link diagrams, and appropriate layouts. We have seen, in the case of a social network, that a force directed layout might be an adapted strategy for exploration. If, for example, we were to observe substrates as countries, a geographical mapping might be more appropriate. Here we encounter the wider challenge of graph visualization, how do we choose the right visual metaphor. Switching between layouts as we propose – or even views such as a matrix view – is one way to cope with this issue, another solution would be to add up linked visualization, but the cognitive load is undoubtedly high in both cases. Nonetheless the harmonized layout we propose has the merit of focusing on the structural aspects of the original multiplex network.

Much further work is expected to extend this harmonized layout. We would hope to make the initial set up of the layout more flexible, such as in the metric and the conditions under which we distinguish the two families of catalysts (*general* and *specific*). Assigning weights to edges in order to drive the layout computation would also be a good idea, especially when the layout can be computed on-the-fly. The possibility of reversing the roles between substrates and catalysts, choosing the best harmonization, depends on the multiplex network topological information (number/density of substrates/catalysts).

As discussed in Section 3.5.2 and 3.5.3, we are also looking for a way to enrich the entanglement analysis on the substrates side, diligently mapping the measures onto substrate nodes, without overloading cognition of our visual analytics framework. Figure 5.13 and 5.13 are already showing promising results. Along the way, we would also like to enhance the rendering of edges in node-link diagrams, with edge bundling and, perhaps, metric mapping onto edges (Figure 5.13). *Power graphs* (Royer et al., 2008) adapted for multiplex networks might be an interesting alternative.

The interaction of our visual analytics framework does not yet allow modification of individual elements. "Dirty" information is very common in real-world application and the correction of individual "misspelled" information could be supported by our framework, with automatic adjustement of all views. The framework could then support then verification and validation of any correction.

We have offered in Section 5.1.4 to filter raw data and eliminate any substrate information we do not want to analyse, and we should plan to do as much on the catalysts side. Finally, we intend to accelerate our analysis algorithms and index structures to make the interaction flow more efficiently when facing large structures.

We provide here a validation of our design based on tasks typology, and we will provide in the next Chapter application cases and a small scale user experiment. However none of these techniques can replace a full scale user study in evaluating our visual frame-

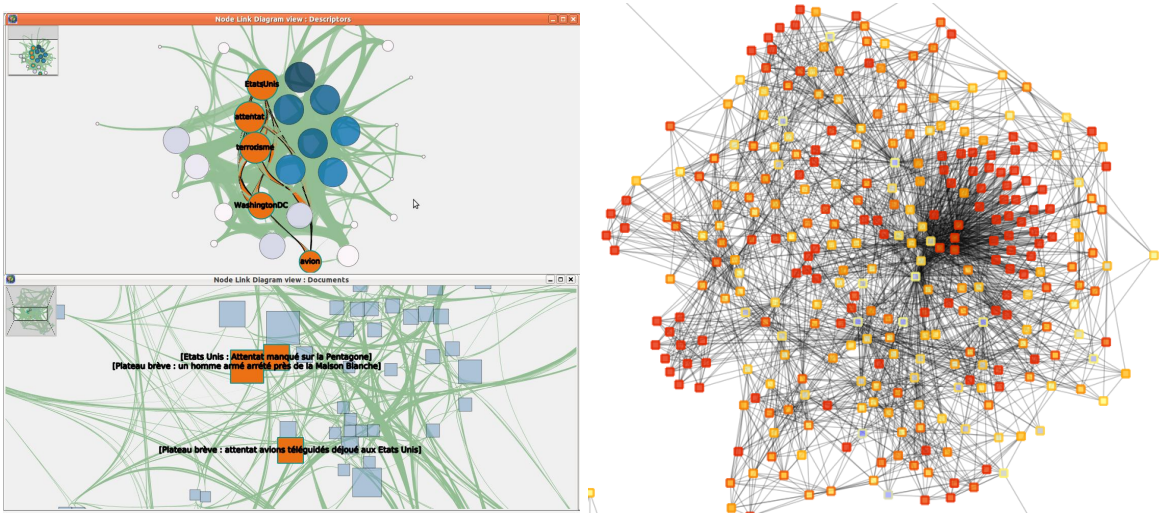


Figure 5.13: New ways for visual mapping. On the left, some prototypal tests for edge size mapping and bundling, as well as size mapping on substrate nodes (lower part). Here we tried to inversely map substrate node size with a custom measure discussed in Section 3.5.3 so irrelevant substrates would pop-out the screen and be quickly removed. Early feedback has confirmed the usefulness of this approach. On the right, our first trial for colour mapping the local entanglement homogeneity and intensity onto substrate nodes. Some areas seem to pop out better than others.

work. The design, implementation and study of such evaluation is a thorough process, that we have initiated with our small scale user experiment, and we also plan to complete it soon.

5.4 Conclusion

In this Chapter we have presented the framework we designed with regards to the recommendations presented in Chapter 4. This framework is designed to handle exploration of multiplex networks for *comprehension*, with coordinated multiple views, and a fitted layout algorithm. Another key contribution is the well suited coordination and mental mapping which preserve interactions, easing leapfrogging between the entity-level of a multiplex network (substrates) and its concept-level (catalysts), whilst preserving the group structure (entanglement homogeneity and intensity mappings). We have validated the design with a task oriented typology, which can cope with the multivariate information, a must, when tackling a great variety of field data.

Despite our many efforts, we have yet failed to publish most of work on the visualization side.The main difficulty we faced was to correctly stress out our contributions in visual analysis, and we fell into the trap of bringing upfront the entanglement study, that is purely analytical. The validation of our work was also a difficulty we managed to overcome recently. The exercise of writing this thesis, enlightened by wise advice from Tamara Munzner, has however clarified our discourse. We have nonetheless introduced parts of our contributions in early works, less visually oriented publications, as in (Renoust et al., 2011a, 2013f) and (Renoust et al., 2013c). The enthusi-

asm of users and domain experts who experienced our framework, as explained in the next chapter, strongly encourages us into pursuing our efforts for publishing this work.

APPLICATION CONTEXT AND EXAMPLES

6

We have started this manuscript by mentioning the nature of the information that motivated our research, and modelled it with multiplex networks. We have presented a framework for analysis and interactive visualization of such network. In this Chapter, we propose to illustrate our work into application. We are presenting the study of some application examples which describes how our framework can be used when applied to more concrete data in a varied range of fields. It feels natural to provide an important space for the study of news document groups in varied cases. We expose also the results of a small scale user experiment we have conducted early in this particular context of document networks. We finally extend to other fields of applications.

6.1 INA's news documents

This section discusses the entanglement analysis embedded in a visual analytics framework, employed in our main application context (INA's documents) to explore a document collection and assess the cohesion of document groups. It is intended to demonstrate the manner in which users explore the document collection and reason about the document content in order to understand the structure of the semantic space. The examples are designed to highlight different aspects of the exploration while emphasizing how the different actions lead users to these discoveries.

The general setting that we consider is one in which users have queried documents using a few keywords. Although documents (substrates) are queried with a clear focus in mind, they introduce semantic entities (catalysts) that were not targeted by the query. Semantic similarities and ambiguities typically mix within the document space, thereby challenging the interpretation of the visual display. Documents then usually undergo a grouping process based on semantic proximity, (such as discussed in Sections 2.2 and 2.3.3). The data we used here is presented in Section 2.1.3 and modelled with a multiplex network as suggested in Section 2.5.1, outputting groups of varying sizes and homogeneity.

At this point, the knowledge that users have gained roughly is "*These documents* concern *these terms*". This is where we enter the

scene. What does "*these terms*" really mean? Do all documents address all terms? Do documents more or less split among terms? What terms make the split explicit? In other words, users must be able to elucidate to what extent, and possibly how and why, the group of documents forms a relatively homogeneous semantic unit.

When the documents are news excerpts, for example, they may well form a group because they all describe the same event (and were published around the same time). Contrarily, cohesion within a document group may follow from general index terms, *i.e.*, documents that all concern the same general subject matter. The examples that we describe illustrate the variety of situations that users may encounter.

The central ingredient of this inspection process is the terms that are used in index documents. A close examination of how terms link to one another *and* how they tie documents together is crucial. Considering the manner in which subject matters are revealed through index terms helps build knowledge and reasoning about the document group.

The interface may treat collections of a few hundred documents; however, we have chosen small collections to illustrate the different aspects of how our visual analytic system performs the different tasks that users face when analysing a set of documents.

6.1.1 Road Safety

This first example illustrates the manner in which the linked views, together with visual cues, help users assess the local homogeneity and identify how lower level terms interact within a group of documents.

Our example is a small group that contains approximately 20 documents concerning *road safety*. A visualization is shown in Figures 3.1 and 6.1. The term-interaction network (top) already describes the "shape" of the semantic space, and we may suspect documents to split between two spheres formed by the upper left and the bottom right cliques. The upper clique gathers terms such as *arrest*, *danger*, *driver*, *offense* and *prison*, for example, and concerns a few documents. The lower clique concerns the majority of documents, which are gathered around terms such as *fine*, *money*, *prohibition* and *radar*, etc. The terms *prevention* and *speed* serve as bridge nodes in this network and belong to both cliques. Recall that as much as possible, the positions of documents mirror those of their specific index terms.

When the upper-left clique is selected, the background color of the selector shape indicates that the group is close to optimally homogeneous. The view linking highlights four documents in the document view. Almost all of these documents are indexed with *all* of the selected terms, which is exactly why the clique is optimal. These documents more precisely relate to a specific news excerpt that concerns a driver who was charged and jailed. All documents tell the same

story and end up being indexed by the same terms (*arrest*, *offense*, *prison*, etc.).

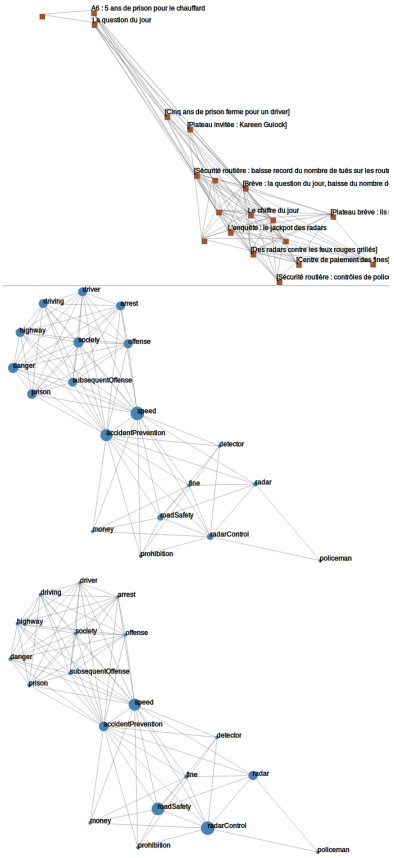
The lower clique gathers terms that spread over the majority of documents with a much lower homogeneity. These documents do not tightly link together, in contrast to the news excerpts of the upper clique. The documents discuss the current road safety policy developed by the government in recent years, and some documents focus on the implementation of automated radar. Terms are loosely connected, and some terms only index a few documents among the whole group.

Figure 3.1[1] shows the *road safety* term-interaction network with node size mapped to number of co-occurrences (top) and entanglement index (bottom). The difference in size clearly indicates that these two statistics act complementarily; the bottom clique contains prominent terms, but those terms do not occur homogeneously over the documents.

The visual cues clearly indicate that the two subjects are embodied through distinct subsets of documents and in opposite manners. There is one specialized subject matter, i.e., a story about a severe road violation, that is equally and exhaustively covered by a small subset of documents. In contrast, there is a larger subset of documents that each contribute to a different aspect of a general subject matter, i.e., the ongoing debate about the state's road safety policy.

Looking at the upper-left clique, we can clearly see that terms' occurrences are low, but their cohesion is high. Indeed those terms tend to co-occurr on every document linked together. There is a large variety of terms but only 4 documents correspond to those terms. As a consequence, occurrences of those terms are low, but the group of documents related to those terms are highly cohesive: intensity is measure at 0.72, and homogeneity at 0.99.

1. Reproduced here for convenience:



The *road safety* documents (top) and the term-interaction network, with node size mapped to entanglement index (middle) and number of co-occurences (bottom). The difference in size clearly indicates that these two statistics act complementarily.

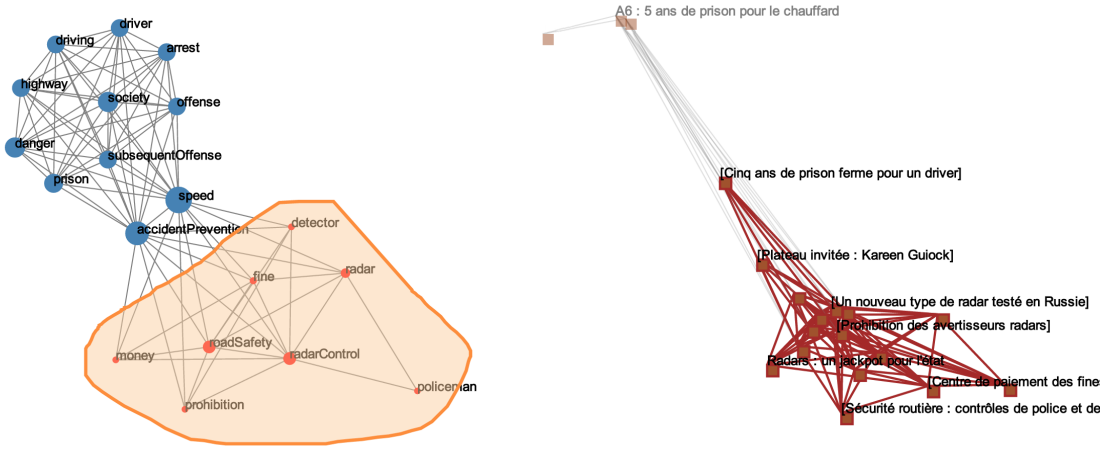


Figure 6.1: The selection of the terms down part of the *Road Safety* highlights the corresponding documents and shows low entanglement measures in the lasso.

The terms of the lowest right part of the graph – do not form a clique – correspond to 19 documents (Figure 6.1), with intensity and homogeneity of 0.29 and 0.63. With the subgroup of documents at the lowest part of the subset, intensity and homogeneity are measured at 0.50 and 0.82: *radar control, radar* and *road safety* present the highest cohesion indices as they are contextual to this subgroup of documents, whereas *speed* and *accident prevention* are incidental to this context (as well as *policeman*). From the rest of this subset, *speed* and *accident prevention* take a similar level of cohesion indices than *road safety* and *radar control*, all the other terms present a very low index, and the overall intensity and homogeneity are set to 0.46 and 0.74.

This first example reveals the benefits of interactive exploration of this document space relative to a topic model. First, algorithms are capable of guessing the number of natural topics – and so the number of natural subject matters – that exist in a document space, whereas the shape of the interaction network readily reveals the number of such topics. Second, a topic model solely indicates that there are two topics and does not yield any additional knowledge regarding the manner in which the topics are embodied in the document space.

6.1.2 Higher education

This application example illustrates how brushing helps to discern among sub-groups of documents. The term-interaction network and the document network can both be queried in a back-and-forth process to focus the exploration on marginal (outlier) documents.

This example gathers a group of 36 documents that are related to higher education. A low global homogeneity (presented in a side panel of the visualization), together with a much looser overall topology, suggests the presence of different stories in the document group (Figure 6.2).

Nodes with higher entanglement indices (mapped to node size) in the term-interaction network are primarily general terms: *students*, *universities* and *higher education*. There is one more specific and unexpected term that occurs: *hazing*.

By utilizing the linking between the two views, going from terms to documents and back to terms is an efficient strategy for understanding the structure of the semantic space and discovering the highest-impact terms. Selecting *hazing* in the term network view yields a dozen documents that form a tightly connected subgroup with much greater local homogeneity. We may then switch to the document view and brush these highlighted documents. Performing the selection in document view highlights all co-occuring terms in the documents that were initially associated with *hazing*. Thus, we now have a view regarding a larger semantic space, which hints to the user that there may be more than a single story that concerns *hazing*.

Indeed, using a brush with a smaller diameter allows the user to locally inspect the document subgroup (Figure 6.3). When moving

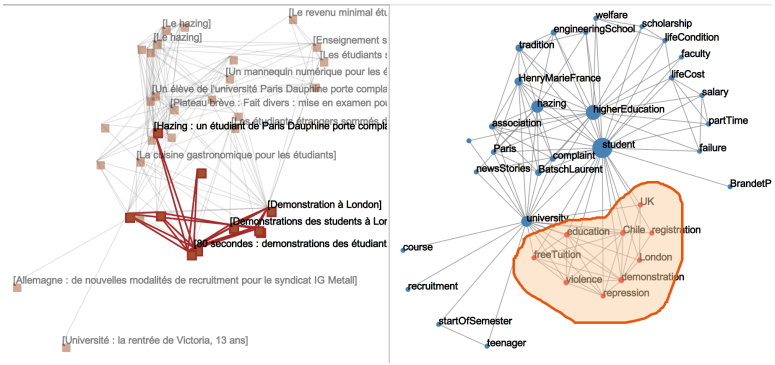


Figure 6.2: A document set (left view) related to higher education. The topology of the term-interaction network (right view) is intricate but roughly splits over three dense neighborhoods/topics: hazing (center); foreign student demonstrations (bottom); and students' living conditions (top). A linked selection from a set of terms (lasso drawn on bottom view) reveals documents (highlighted in the top view) related to students and hazing. Notice the correspondence between the selected documents' positions and their terms' positions (top-left area of both views).

the brush around, changes in the brush's background color indicate that the subgroup is not uniformely homogenous. Observe that because the layouts are harmonized, brushing documents in the leftmost (rightmost) part of the subgroup have a higher chance of highlighting terms in the leftmost (rightmost) part of the term network. Incidentally, a few brush moves allow one to easily identify two poles that formed around *tradition* and *complaint*.

Again, we are able to split the documents. A first, more focused, story is about a *hazing* event that resulted in a *complaint* filed by a student; this story is confirmed by the presence of nodes such as *Laurent Batsch* (dean of the university at which the abuses occurred), *association* and *Marie France Henry* (leader of an anti-hazing association). A second story addresses *hazing* as a societal issue in higher education that is considered a *tradition* at *engineering schools*.

We have thus identified a semantic subspace defined around the term *hazing*. Putting this subspace aside, we are left with two dense neighborhoods that can be easily grasped using a lasso. One clearly relates to *demonstrations* that occurred in *Chile*, and were echoed in Europe (in France and in *London*, *UK*). The second neighborhood relates to the more general issue of student living conditions and gathers terms such as *life conditions*, *life cost*, *social security* and *scholarship*.

Now, what if we discard the documents that are concerned with each of these three main subject matters? By selecting each of these term neighborhoods, we may update the document space by momentarily deleting them. Selection of the remaining documents highlights the most general terms in the term network: *student*, *higher education* and *universities*. Although these documents belong to the same semantic space, they only do so through very general subject matters and do not relate to the three specific stories we previously identified.

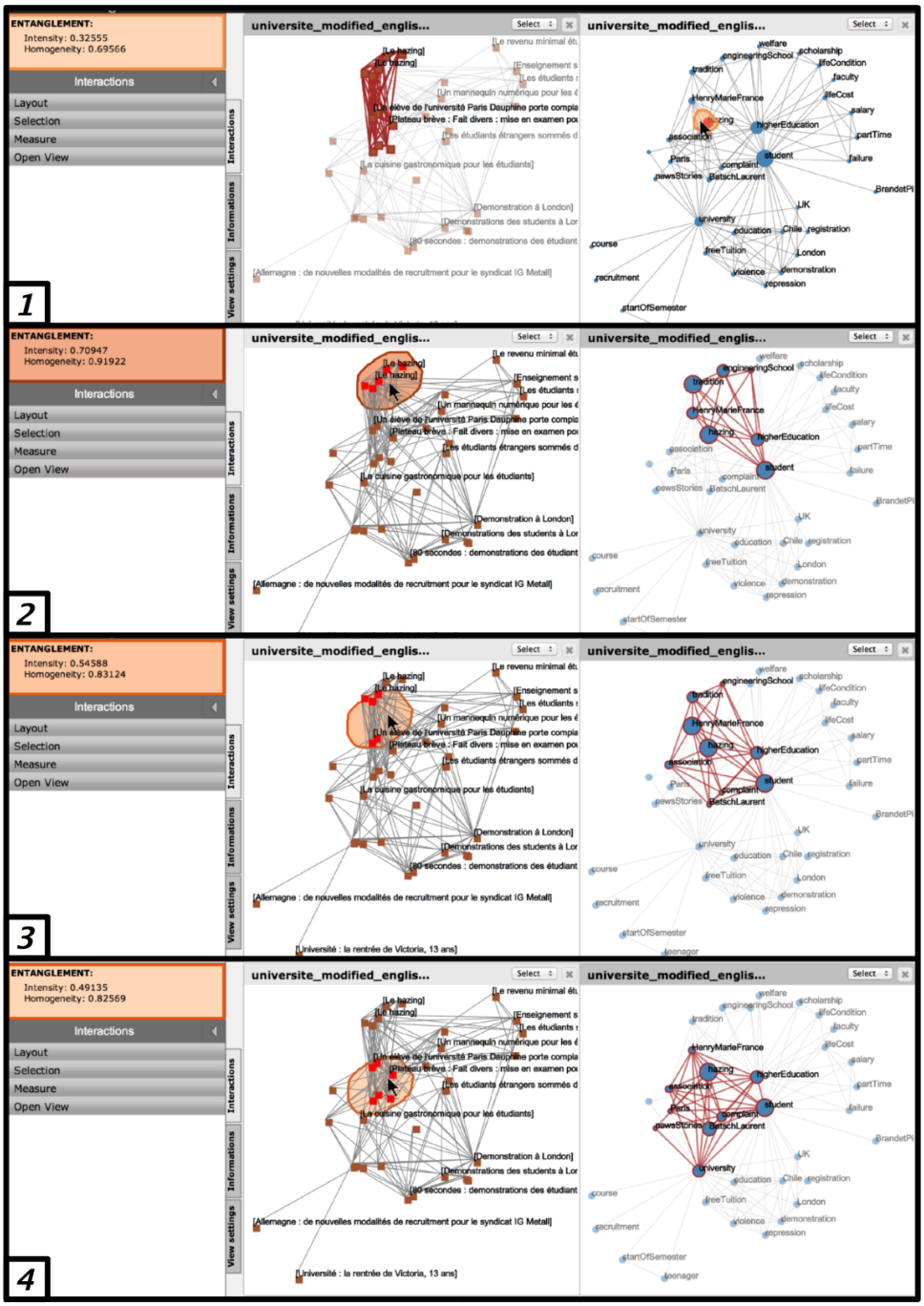


Figure 6.3: A step-by-step exploration of the document space corresponding to *hazing* (step 1). Brushing around (step 2-4) reveals two subgroups of documents concerned by *hazing*, a first one around *tradition* (step 2), and a second one around *complaint* (step 4).

6.1.3 *Swine flu*

This application case illustrates how leapfrogging between documents and term selection enables users in handling much bigger sets of documents.

The example gathers 226 documents about the swine flu outbreak, gathered around 93 terms (Figure 6.5). Note that 180 terms originally

indexed those documents, but only 93 of them interacts through pairs of documents, reducing the complexity the original term space. The group displays very low intensity of 0.05 and low homogeneity of 0.26.

In this example, the node-link diagram display of the documents brings nothing but edge cluttering, so edges are not displayed. The user can then choose between different layout of documents, multi-dimensionnal scaling (from Jaccard's distance), force layout of the document network, or harmonized layout as shown in Figure 6.4.

The vocabulary we observe is very wide, centered around *flu*, *public health*, *epidemic* and *contamination*. Following right after are *swine*, *Mexico*, *medical treatment*, *surveillance*, *prevention*, *sick etc.*, and lots of detailed vocabulary surrounding these terms.

The topology of the term interaction graph does not suggest a lot of obvious subjects since everything seems tied to the flu. However, we can find subgroups of peripheral terms such as *tabacco*, *smoking*, *smoker*, *cigarrette* and *school*, *closing*, *Creteil*, *child*. These groups can be outliers gathered because of one particular term – which is the case of the first group around *tabacco*, gathered by *prevention* – or specific terms related to one specific event – such as the closing of a school in Creteil.

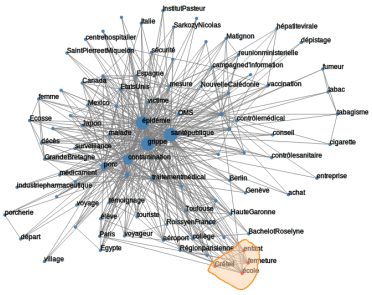
With any meaningful layout of the documents' view, brushing remains a powerful strategy to locate subsets of interests. Any selection of documents from the user can be specified as an input for a new bipartite analysis given those documents. This turns out to be very handy when facing large groups of documents.

Users now can query the graph depending on their angle on the news. For example, if we are interested by the schools that were closed during the outbreak, we can select the second group of terms previously listed. A group of related document is then returned through linked selection (Figure 6.5). The user can now ask to leapfrog through the linked selection to the corresponding documents and observe the vocabulary and the entanglement.

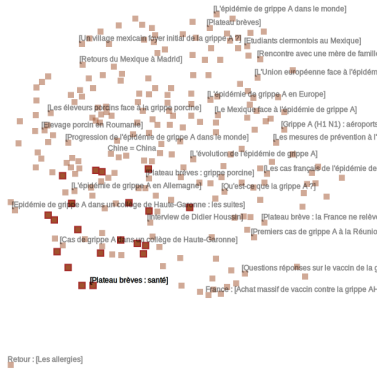
The global measures show a much higher intensity of 0.25 and homogeneity of 0.57, surrounding 23 documents and 18 terms, and we can use this set of documents as an input for an analysis. This gives users the freedom to explore this subset as described in the previous application cases, perhaps removing some outliers to do so, and constructing their own corpus of documents to work with.

The same operations are extended to documents. Let's say our user is interested in an outlier document such as *The come back of allergies* in order to detect and discard similar outliers, we can leapfrog through the selection in this manner as well. In this case, it will return all documents centered on *prevention*.

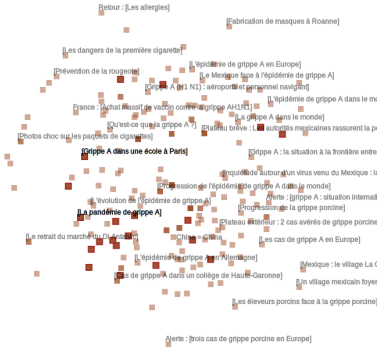
Figure 6.4: Three different layouts linked with the same selection of *swine flu* documents.



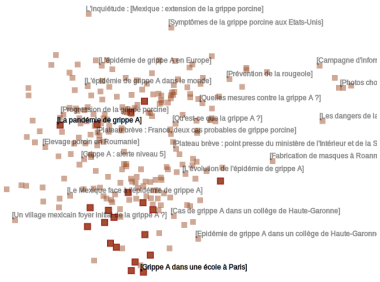
Terms with selection



Documents with a force layout



Documents with a MDS layout



Documents with an harmonized layout

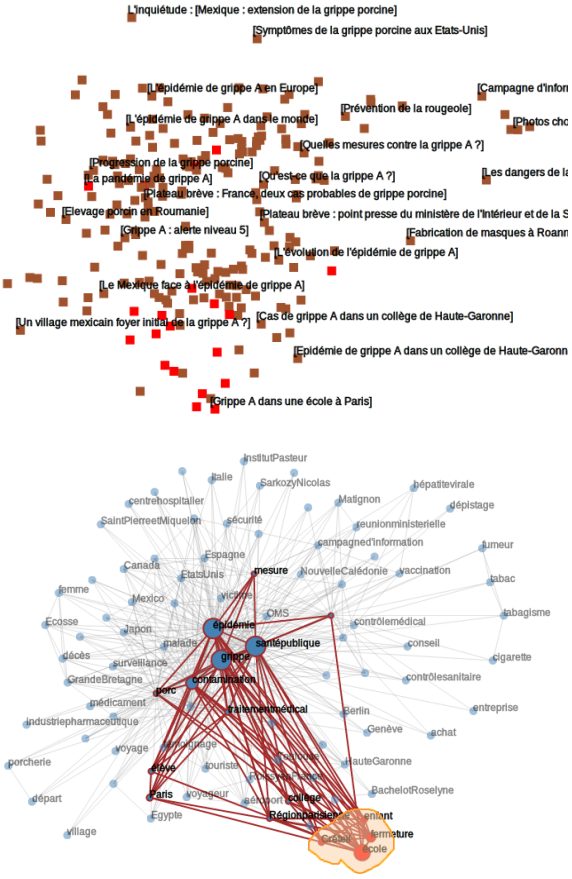


Figure 6.5: A leapfrogged selection in the *swine flu* group. The selection is leapfrogged from the same selection as in 6.4 : *school*, *closing*, *Creteil*, *child*, thereby displaying all the vocabulary with its entanglement measures (term size and brush color) related to the corresponding subset of documents.

6.1.4 *Structure of the (Intensity × Homogeneity) space*

We have put forward how the interaction networks express the structure of the documents, and how the group entanglement measurement relates to a document group's cohesion. Now that we have seen this whole frame in application, we can recall the study of the (Intensity × Homogeneity) space we started in Section 3.3.2, and complete it with our application cases. We bring again here Figure 6.6 for convenience, and we can comment the structure of this space further.

In Figure 6.6, the red area corresponds to our entire corpus of INA documents, the blue nodes refer to the previous examples, and the rest corresponds to the random networks. Keeping in mind the dependency between intensity and homogeneity, and the dependency on documents to vocabulary, we can attempt to explain why the document networks are so close to the leftmost border of the scatter plot:

- There is nothing to question on the top rightmost position in the plot, since it corresponds to the most cohesive groups of documents. The *Road Safety* upper sub-network considered in the first application case (Figure 3.1, Section 6.1.1). Higher inten-

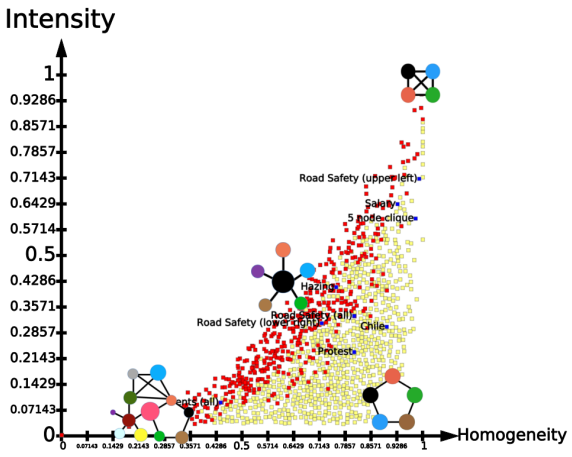


Figure 6.6: From Figure 3.5. Entanglement profiles can roughly be categorized by combining intensity and homogeneity. Measures on INA's document groups are in red. Randomly generated multiplex graphs are in yellow. Cases examples are in blue with labels. Typical catalyst interaction networks can be associated on homogeneity/intensity regions.

sity and homogeneity are much easier to achieve with smaller document and term subsets.

- The lower-left case gathers networks with low intensity and low homogeneity. A term set covering a wider semantic scope inevitably induces a network falling in this category. The *Swine flue* (Figure 6.5), and more arguably, the starting interaction network in the *Higher education* application case, with an intensity of 0.09 and a homogeneity of 0.44, fall within this category.

- The upper-right area corresponds to relatively high homogeneity and lower intensity. Some variety of vocabulary without centralized concept, or separated concepts could lead to such a case. Example networks in this category are the *Chile* sub-network, with intensity 0.30 and homogeneity 0.90, and the *Road safety* sub-network (Figure 6.1) is also part of this category.

- The leftmost area may be the most common cases for documents. This case occurs when terms are expressing similar concepts at different generality levels. We pointed out the *Hazing* sub-network as a prototype of this phenomenon, and maybe the lowest part of *Road safety*.

The leftmost border means correlation between entanglement homogeneity and intensity. Bigger clusters mean more vocabulary, the number of terms per document is averaged between 8 and 9 terms per document (Section 2.1.3), and the distribution of terms in the corpus follows a power-law distribution (most certainly a Zipfian distribution[2]). The interaction here corresponds to co-occurrence, and since it strongly depends on occurrence (a term needs to occur before

2. A Zipf's law (Zipf, 1935) models among other phenomena the distribution of words occurrence in a language. And our terms make no exception, even though we use a controlled vocabulary.

co-occurring), we might suspect to observe such power law in the distribution of entanglement. Hence, due to this power distribution, for a given intensity, we often observe a very low homogeneity (minimal for *this* intensity).

Even if we still need more work to confirm all these hypotheses, we believe that, in definitive, the distribution of clusters in this space could very well be distinctive of the data and the clustering technique applied to it.

6.1.5 Validation through a small scale user experiment

We aimed to confirm if the entanglement relates to the semantic cohesion of a document group (*i.e.* if a group of documents makes sense as a whole), and how it can help the user explore and analyse a document group. The semantic cohesion, however, is a notion that is often qualitatively evaluated. We thus felt the need to confront the entanglement index with expert users to assess its relevancy.

We have conducted an experiment with four expert documentalists (2 senior, 2 junior). Although informal and involving very few users, the experiment was designed according to a strict protocol. Four different document sets were used; smaller samples contained between 20 and 40 documents, and larger samples contained between 60 and 80 documents. Each sample contained documents related to events covered in the news (the documents themselves corresponded to TV news excerpts). Samples showed clear contrasts (in terms of entanglement and term interaction network profile).

Users had access to documents in three different settings. The baseline situation consisted in a list of documents (text) listing titles, content and index terms. In a second situation, users had access to a node-link display of the document interaction network, with document details and classical graph statistics (degree, centrality). The third situation offered users our prototype as illustrated in the above examples.

Users had a short training period to ensure that they could use the interface and understood the tasks they were asked to perform. They were then given random combinations of document samples and display to avoid potential biases (learning or fatigue). The experiment ended with a questionnaire and face-to-face interview. The experiment took 2.5 hours per user on average. Users were asked several questions:

- Evaluate the overall cohesion of a document sample (how well do documents fit together).
- Eliminate "noisy" documents to reinforce cohesion within the remaining documents (off-topic documents).
- Users were given a story line and were asked to find documents

they felt were more relevant to the story. Conversely, they were asked to point at the documents that did not relate to the story.

- Tell the story covered by the document sample (which should more or less correspond to a news event from which documents were extracted).

- Rate the confidence they had in their overall analysis (e.g., discarding the *right* documents, recovering the *right* story, *etc.*).

One goal of the experiment was to get feedback on the ability of the linked view to sort out documents according to the previous questions, and to evaluate the utility of the entanglement in successfully performing these tasks. Users were asked to rate several aspects of the prototype (usefulness of the linked views over alternative scenarii, usefulness of the entanglement index mapped on node size, readability of the layouts, *etc.*). The ratings were collected to guide future prototyping iterations, not to draw any statistical evidence.

A second goal of the experiment was to evaluate how much the overall prototype contributed to build trust in their analysis. Face-to-face interviews brought up interesting observations. The fact that the layouts of both networks matched was noted as useful. Documents linked to marginal terms could easily be identified and discarded as "outliers".

We firstly aimed at comparing times acquired while they were processing tasks, however curiosity drove users much further into digging the collections of documents, and time measurement would have proven our system wrong whereas the quality of their insights, and the enjoyment of manipulating the system were obvious.

Users enjoyed the informative readability of the term network representation and agreed that it helped them to differentiate between terms, as well as to build a story for the whole document sample. We also gained confidence that the network shape supported users' intuition to identify salient characteristics (e.g., central term subsets, outliers, *etc.*).

The confidence level of users clearly improved when using the term interaction network to explore a document sample. Finally, users felt that the support gained from the interaction network became more critical as network and document collection size increased.

Here is a small extract of some user's comments (translated in English), supporting the use of the catalyst-interaction network:

- *I did not know much about graph visualization, but even if the system is complex, I like that it is intuitive to learn.*

- *The catalyst view makes it easy to isolate the difference concepts concerning the group of documents.*

- *It is like a word-cloud but better because it gives a sense of a hierarchy between terms.*

- *I like that we can associate the documents from the terms.*
- *The catalyst view summarizes well what happens in the documents.*

Obviously, this small scale experiment does not qualify as a formal user experiment from which statistical evidence can be drawn. We nevertheless felt the use to perform it soon enough in the design and development process to validate our design choices.

6.1.6 *Discussion*

There is a surprising power in the shape of the catalyst interaction network, especially in the case of documents. It feels almost as the shape of the term interaction network could narrate the story about a specific group of documents. Dense areas, almost communities, seem to refer to independent stories, and star-shaped network seem to be very thematic. This is additionally interesting when one looks at a news event through the lens of different broadcasts (Figure 6.8). As

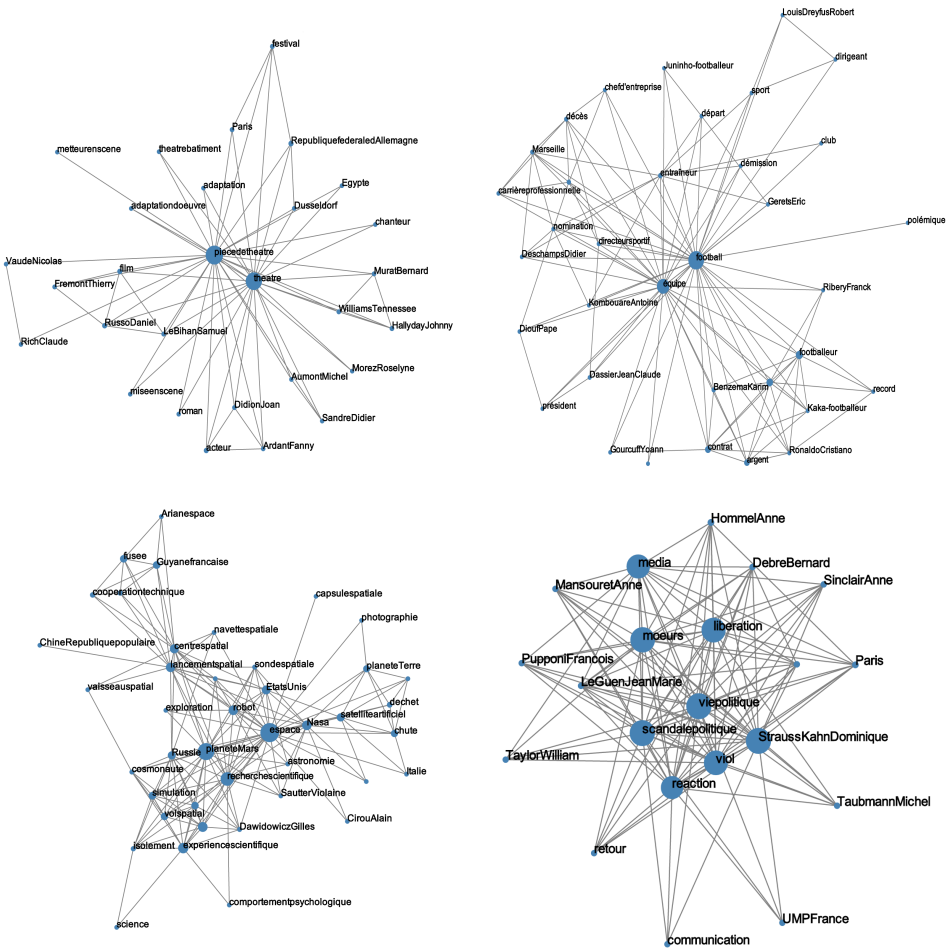


Figure 6.7: Different shapes of term interaction networks. The most *theme-based* groups of documents seem to display a star shape (top left, *theatre plays*). The most *event-based* seem to display a more dense shape (bottom right, *the DSK case*). Many cases are in between: (top right, *European football*) events within a theme, or (bottom left, *space research/rocket launch*) events regroup and form a theme.

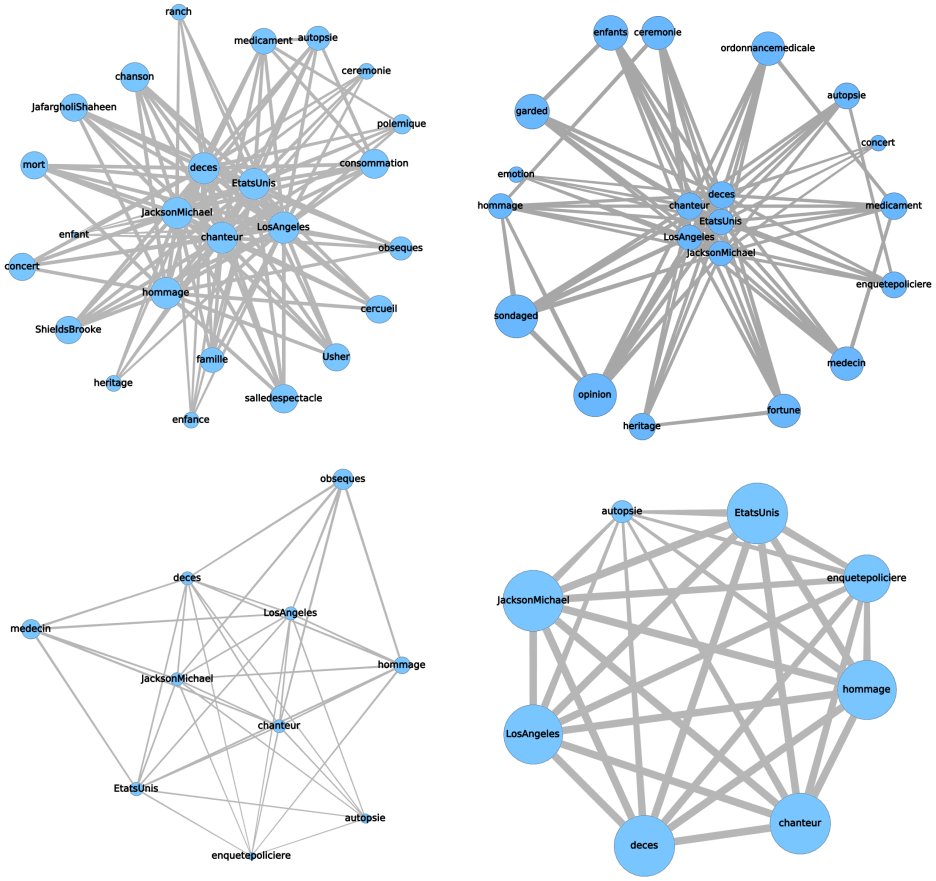


Figure 6.8: *The death of Michael Jackson* narrated by four different news broadcasts, TF1 (top left), France 2 (top right), France 3 (bottom left), and M6 (bottom right).

a picture is worth a thousand words, the examples of Figure 6.7 give an idea of the different species we can find in this corpus. It includes groups made from event, or themes, or fusion of events, events in themes, and so on.

There is definitively a lot to learn from this perspective, and inference from the typological features of the term interaction network may characterize the type of a group of documents.

We still can do much more investigating of the document collection, and many tools need to be embedded in the framework in order to even deepen the analysis. Time is a dimension we have barely approached, and the data is rich, channels, named entities identification, geographical links, social network of personalities and more. We have not yet delved into the multimedia aspect of the documents such as links across locutors and conversations, illustrative pictures, and the sources of the pictures and information. It requires a lot of process, and fortunately all these issues are addressed by the OTMedia project and soon we will be in a position to try our analysis taking these multiple factors into account (Hervé et al., 2013).

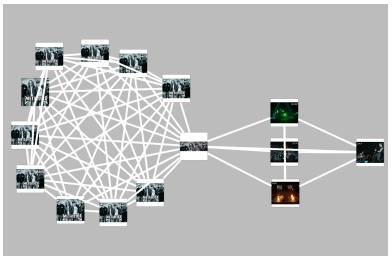
As a starter, we also informally applied our analysis on image-copy matching retrieval (in which we have linked images to parts

that are shared between images, as in (Letessier et al., 2011)) and the *precision*[3]) of the system was very high so we mostly had optimally cohesive groups with little exceptions[4]. We had only time for a very quick test on-the-side, but further work is intended. As an example, we would like to measure the effect in the group cohesion of parametrizing the retrieval engine for better recall or better precision.

Further than INA's documents, we would like to explore the relationships of the catalyst-interaction network with already existing ontologies or semantic graphs we could extract from a lexical database such as Wordnet[5]. We would also like to further investigate the graphs that the W3C Resource Description Framework (RDF)[6] model describes, in which case we have already natural multiple relationships to explore.

3. *Precision* and *recall* are two quality measures used to evaluate an Information Retrieval system. Roughly speaking, *precision* addresses the non-relevant items that are retrieved, and *recall*, the relevant item that are not retrieved.

4. Here is a sample of an image group in which every image shares the logo of the rock band *Metallica*



5. http://wordnet.princeton.edu/

6. http://www.w3.org/RDF/

6.2 Social networks

The application cases we describe in this section aim at showing how the entanglement indices, and the group entanglement of networks help users explore social networks and reason about the homophily content. By navigating the network and getting feedback about these indices, users can question the structure of the space that binds entities together. The examples are designed to highlight different aspects of the exploration, each time underlining how the indices contribute to better understanding the group structure of the homophily network.

Roughly speaking, the knowledge users gain after applying a grouping procedure (clustering, community detection) is that "*a group of agents*" share "*a list of attributes*". This is where the entanglement index enters the scene. What does "*a list of attributes*" really mean? Do all agents share all attributes? Do actors more or less split between attributes? What particular attribute(s) make(s) the split explicit? In other words, users must be able to elucidate to what extent, and possibly how/why, the group of agents form a more or less cohesive unit.

6.2.1 IMDB

This application case is built from the Internet Movie DataBase, a largely used dataset[7]. Auber et al. (2003) had visualized a small world subset of the IMDB co-acting graph. Starting from the main actors in this subset, we have additionally extracted the corresponding movie directors to form a bipartite network where directors connect to movie actors they have directed. Applying our methodology we compute (i) a movie director network (substrates), where two directors connect when the set of actors (catalysts) they have directed share at least two actors, together with (ii) the corresponding actor interaction network. The data may thus be used to find homogenous

7. www.imdb.com

subgroups of movie directors, those whose artistic signature rely on similar movie casts.

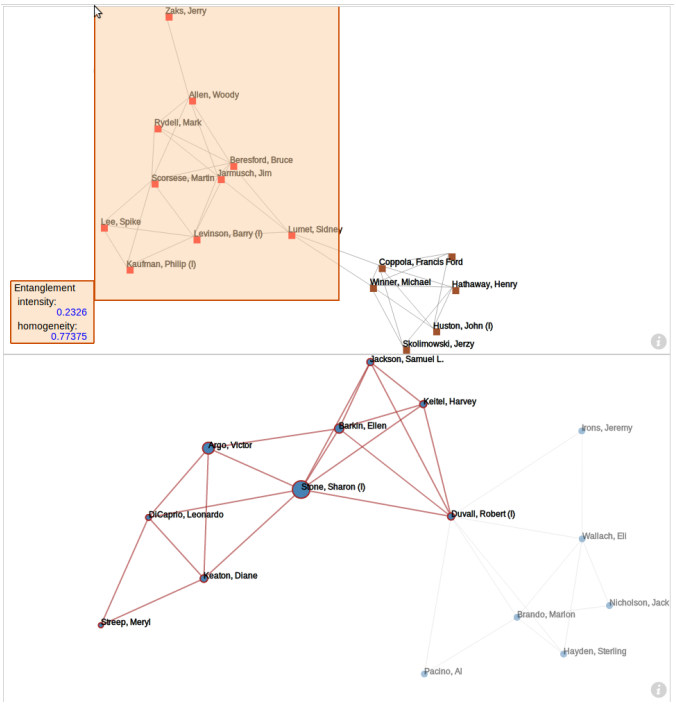


Figure 6.9: IMDB - *directors* appear on top; the *actor* interaction network is displayed in the bottom panel. Selecting a group of directors highlights the corresponding actors, with node size mapped to their entanglement index. This group of directors shows low homogeneity and intensity. We can clearly see that the distribution of actors is unbalanced, partly because *Sharon Stone* plays by far a central role in the interactions between directors – the directors all have, at some point, directed her.

This first example gathers 15 actors and 16 directors (see Figure 6.9). A low intensity and medium homogeneity, together with a loosely connected actor interaction network topology suggests that actors and directors roughly split into two communities. The director network has medium homogeneity that corresponds to a quite balanced distribution of actors among them. Intensity is not optimal: the directors did not individually direct *each* of these actors although, as a group, they did direct *all* of these actors. The low values of the network indices readily indicate the need to dig further into the network and try to "nuance" the structure of this group. Roughly speaking, low intensity follows from the fact that most directors have directed only a small number of actors relative to the whole set.

As can be seen from Figure 6.9 (bottom), the two communities of actors are connected through Robert Duvall, and the two communities of directors are connected through Sidney Lurnet. Apart from Robert Duvall, the bottom right community of actors is formed around Marlon Brando, Al Pacino, Jeremy Irons, Jack Nicholson, *etc.* The top left community of actors is formed around Sharon Stone, Harvey Keitel, Samuel Lee Jackson, Leonardo DiCaprio, Meryl Streep, *etc.* Clearly, there is a generation gap between those two communi-

ties of actors with Robert Duval filling the gap – just as Sidney Lurnet does it in the director network.

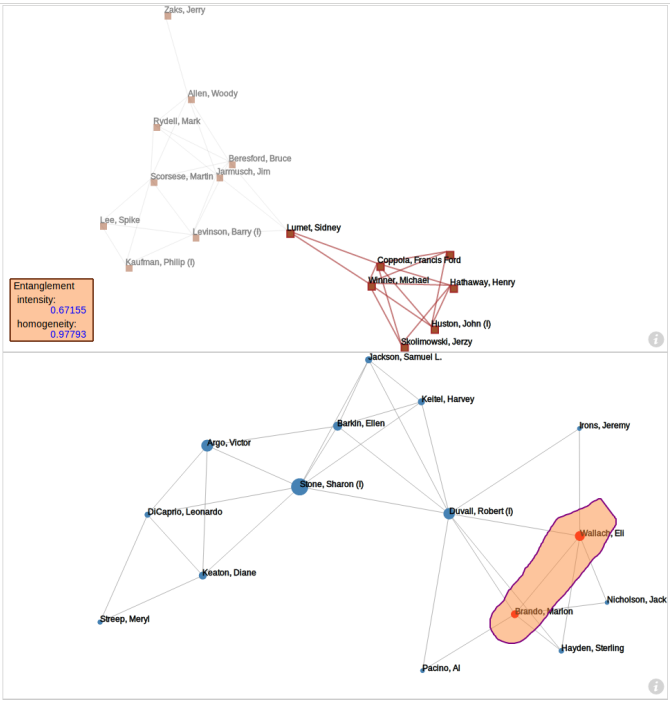


Figure 6.10: The selected subgroup of actors show high intensity and homogeneity: many of the corresponding directors directed each of them. Interesting, these actors refer mostly to directors from a same separated community.

The "Marlon Brando" sub-community (bottom right) shows higher intensity (with homogeneity similar to the overall network). These actors appear as attributes for a subgroup of directors centred around Francis Ford Coppola, John Huston, and others (see the highlighted group located at the top of Figure 6.10). Selecting these directors clearly shows Marlon Brando, Eli Wallach and Robert Duval as influential actors. Indeed, higher intensity means they played with many other actors under the direction of these movie directors.

The community of actors located in the top left part of the panel correspond to a different group of directors (connecting to the previous group through Sidney Lurnet). It gathers Spike Lee, Jim Jarmusch, Martin Scorsese, Woody Allen and others. This community has similar intensity but higher homogeneity when compared to the overall network. This means these actors have equal influence within this group and better capture altogether the artistic signature of these directors as a group.

The upper left subgroup in the director network (see Figure 6.11) actually divides into three overlapping cliques. Two cliques reach maximal homogeneity and intensity (the exact same actors have all played under their direction). The third clique (Bruce Beresford, Jim Jarmusch, Barry Levinson, and Sidney Lurnet) – selected in the top panel of Figure 6.11 – focuses on Ellen Barklin and Sharon Stone. It has lower homogeneity and intensity indices: they don't mix that

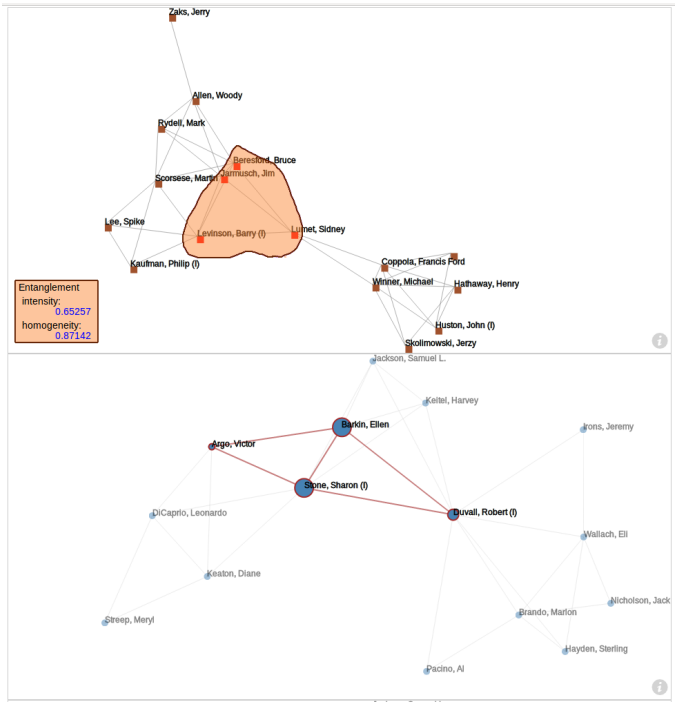


Figure 6.11: A group of directors (top) and the corresponding actors they co-directed (bottom, highlighted) with node size mapped to their entanglement index. This clique of 4 directors shows higher homogeneity and intensity than the selected group on Figure 6.9.

well with the other actors.

Not surprisingly, and even though we have not taken any time-related information into account, this study confirms the generational gap among directors and actors, from Marlon Brando to Leonardo DiCaprio, such as groups of the same generation show higher entanglement measures.

6.2.2 *InfoVis 2004 contest*

Here, we introduce data of a different nature, in which keywords, publications, and authors are linked, thereby demonstrating that the notion of entanglement applies not only to documents but also to a wider variety of entities. Co-authorship networks are an interesting case of social networks which can be seen under a different light. In these co-authorship networks, we associate people, publications, and keywords associated to the publications. This could be seen as a 3-partite network, but instead, we modelled them as multiplex networks.

We selected a subset of the InfoVis 2004 Contest dataset built from papers published in this IEEE conference proceeding over the period of 1994 to 2004 (Ke et al., 2004). The data we consider are authors indexed by keywords gathered from the papers they published. We thus compute a bipartite graph in which authors are linked to keywords. We demonstrate how the term-network/author-network

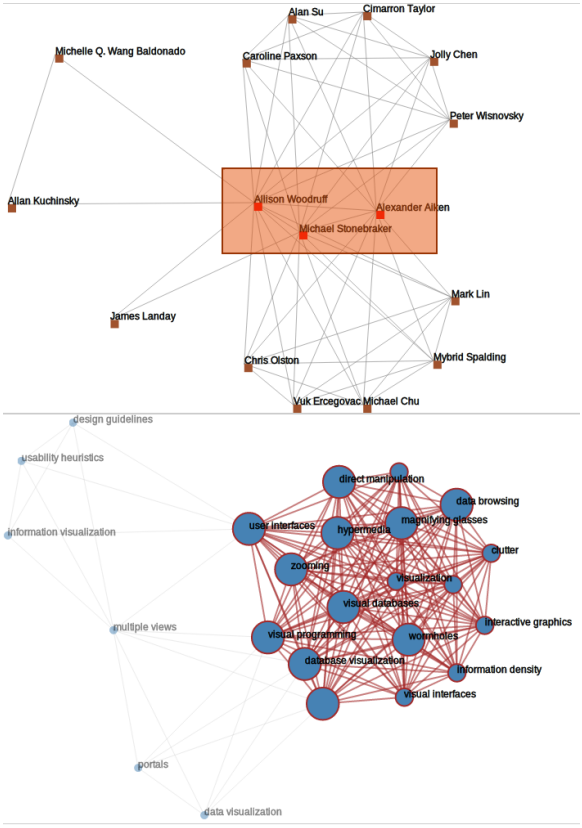


Figure 6.12: The InfoVis 2004 contest data generate a keyword-interaction network paired with an author social network. The three selected authors occupy a central position in the social network (top). Their co-publications cover a wide spectrum of research topics, as indicated by the clique of keywords in the bottom image. The entanglement intensity, although good, is not optimal: they did not pairwise co-publish on all the research topics. Indeed, we may suspect that each of them has different co-authors in the network.

paradigm helps us to easily solve two tasks of the 2004 contest:

- Where does a particular author/researcher fit within the research areas?
- What, if any, are the relationships between two or more or all researchers?

The bipartite graph yields a keyword interaction network and an author social network. The overall social network splits into several components. We will focus on the subgroup lead by Woodruff, Olston and Stonebraker (see Ke et al. (2004, leftmost part of Figure 4)).

The answer to the first question presented above is straightforward. When a single author is selected, the associated keywords are pushed to the foreground in the term network and positioned in the context of neighboring keywords. The social network displays the immediate co-authors of the selected author.

Filtering of the entire network author-by-author is useful because it provides fine-grained information regarding the network; it is, how-

ever, lengthy and tiresome and cannot reasonably be performed for larger networks.

This brings us to the second task, which requires a more elaborate exploration strategy. For this task, we take advantage of the "shape" of the social network, which exhibits the community's structure. We examine interesting sub-communities and consider keywords that are shared by authors. Conversely, we may select a subset of keywords, analyse authors who have published on these keywords and determine whether they are in the same community, for example.

The pattern of entanglement we are looking for in this example is an interesting example. We assumed with documents that cohesion was not easy to reach and we were looking for areas with higher group homogeneity and intensity. This case is the opposite, as authors have co-published they most often share the same interests on the same topics and, in order to comprehend the network, we focus our attention on areas that display the lowest group entanglement measures.

The topology of the author network (Figure 6.12, top) clearly indicates three authors as central actors (A. Woodruff, M. Stonebraker and A. Aiken) at the intersection of two different cliques. Their associated keywords cover a large part of the semantic space and form a large clique in the keyword interaction network (Figure 6.12, bottom). Variations in the entanglement indices (node size) explain the low intensity for this group and suggest that each of these three authors has a distinct set of research topics.

Selecting the authors that are part of the top clique in the social network (*Paxson*, *Wisnovsky*, *etc.*) *excepting* those central actors leaves us with a subset of authors with optimal intensity and homogeneity: they all co-published with the exact same keywords. The same is true if we select the authors that are part of the bottom clique (*except* for the central authors, i.e., *Olston*, *Spalding*, ...).

We may also select two marginal authors that lie on the left side of the social network (Baldonaldo & Kuchinsky) and observe that their semantic subspace is located outside of the "Woodruff clique" semantic space.

After inspection of these sub-communities and their associated semantic sub-space, one obvious issue comes to mind: none of these sub-communities seem to address the keywords *portals* and *data visualization*, which are located in the bottom-left region of the term-interaction network. Considering these two keywords, we readily see that they solely concern Woodruff and Olston (Figure 6.13, top). Returning to the author network and selecting Woodruff and Olston, we then see that the keywords that these two authors have in common are marginally positioned with respect to the main clique (Figure 6.13, bottom).

These manipulations demonstrate how linked views and browsing operations help decipher the nature of collaborations among au-

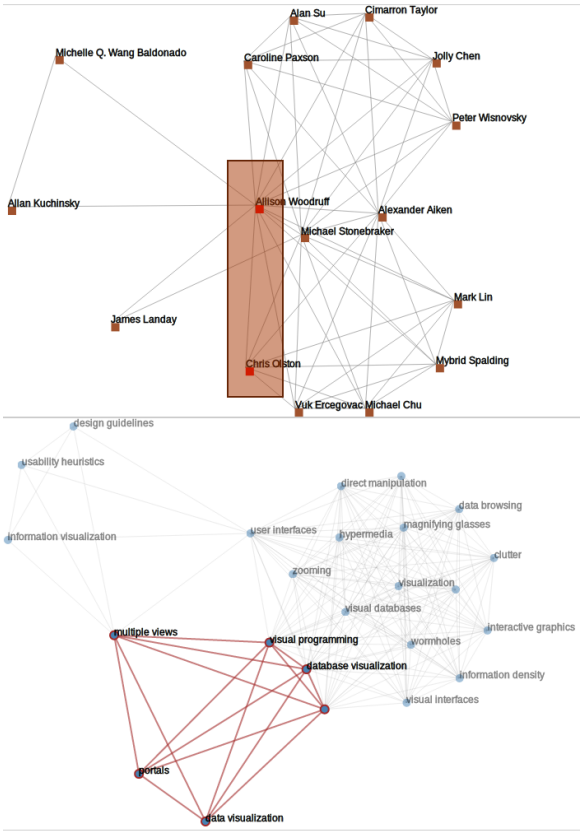


Figure 6.13: When browsing around "obvious" sub-communities of authors, the keywords portals and data visualization never arise. Direct selection in the term network brings two co-authors to the front: A. Woodruff and C. Olston (top). Selection of these authors indicates that their common research topics of interest are marginally positioned with respect to the main clique (bottom).

thors. The group entanglement of the term sub-network indicates how focused and intense these collaborations are.

Further manipulations in this network let us find more specific collaborations, between J. Landay, A. Woodruff and M. Stonebraker or between J. Chen, and A. Woodruff. The structure of all these collaborations highlights A. Woodruff as the central author who not only collaborates with the different communities, but also with very specific authors in these communities.

6.2.3 *Other social networks*

We also started to dig further into a varied types of social networks, that has brought us new insights for further investigation. Many other social networks can be similarly approached.

6.2.4 *Arnetminer*

8. www.arnetminer.org

Arnetminer[8] is a service used for academic social networks indexing and searching which provides a very convenient web API.

This system brings another aspect in a group of catalyst as it is a result of a query. The results of a query is a list of publications with

metadata. Arnetminer enables two types of queries: *keyword-based* search (Figure 6.14), as any search engine would do, returns a list of publications relevant to the query; and *author-based* search (Figure 6.15) returns a list of publications authored by the person queried.

We analyse this network under the light of these specific types of information, digging the metadata for authors and keywords. Unfortunately, the keywords, in the sense of the InfoVis 2004 dataset, are not returned by the system. Nonetheless, we put up together a quick vocabulary extraction from the publication titles.

Although we have made up this system for our pure own enjoyment and curiosity, we can get more insights on our methodology. Such a network often involves multiple connected components, emphasizing the need to treat each component as separated cases, and perhaps, build intermediate groups. Indeed, what would be the entanglement of the result of the query when we observe many separated catalyst interaction connected components?

6.2.5 *Edgeryders*

Another example that has further advanced our analysis is the Edgeryders[9] network. Edgeryders is a community of thinkers of diverse backgrounds who address innovations of all sorts with the common goal to act against Humanity's main challenges. This leads participants to have productive on-line conversations and come out with new expertise and ideas.

9. http://edgeryders.eu/

We have been approached to take a look at the structure of this network (Figure 6.16), and wanted to understand how people exchange and lead productive discussions. The data takes a shape of a long thread of a conversation, with individuals replying to each other. The conversations can split into new threads, and eventually end.

We thought entanglement analysis would be an interesting tool

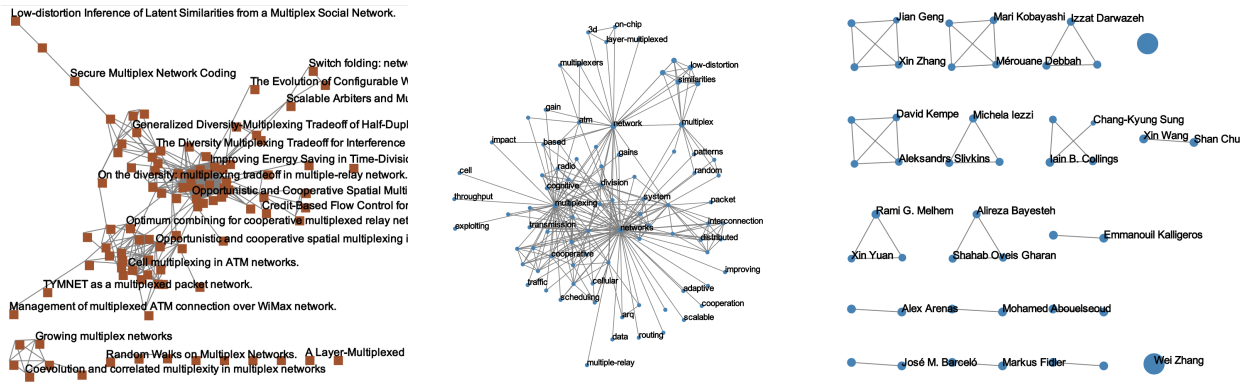


Figure 6.14: Querying "Multiplex network" on arnetminer. Publications, keywords and authors. According to a keyword-based query, studying the author/keywords is one most interesting task that teaches us more the community publishing on a particular subject: here the community appears very divided, no central author, and themes surrounding networks feel more related to computer networks and communication.

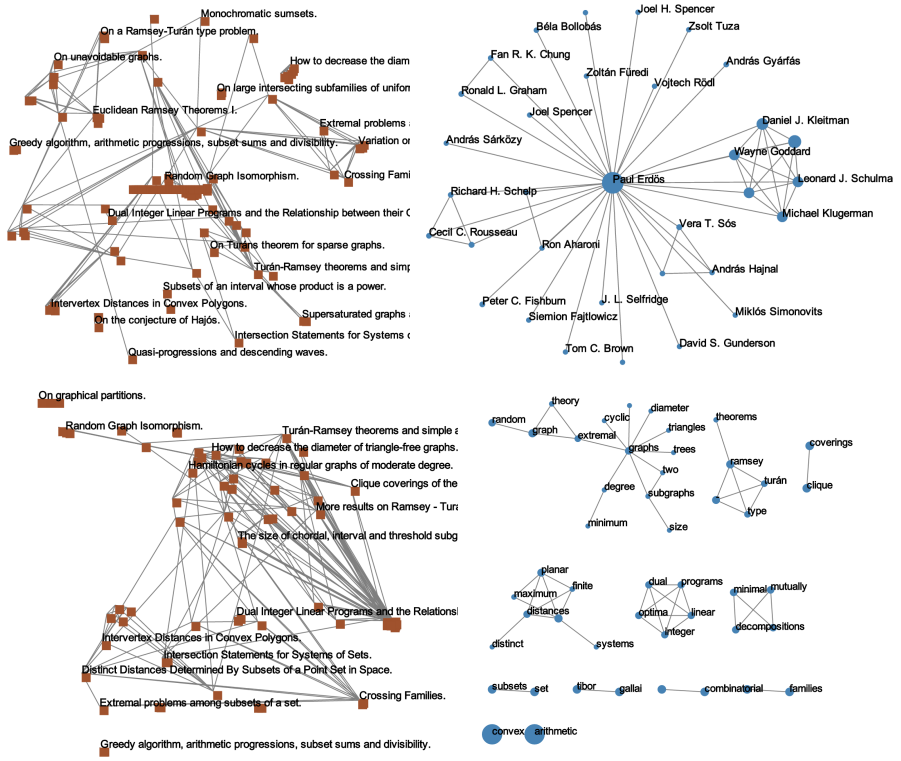


Figure 6.15: Two views on Paul Erdős as seen from Arnetminer. It is startling to see in his collaboration network (top) all the different collaborations he independently led, there is a very few transversal collaborations. In the second view, the keywords display all his interests in the publications returned by Arnetminer (a lot of them involves graph planarity).

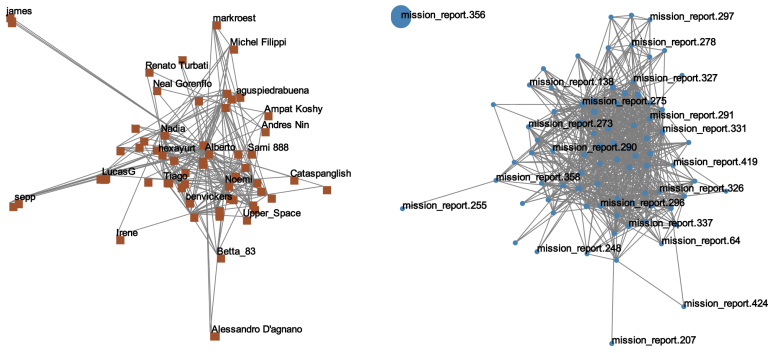


Figure 6.16: The Edgeryders community network (people are substrates, and threads are catalysts). An application case that motivated an extension of our model.

to apply here, as we could study the redundancy of conversations across people, or people across conversations. However, the results were not too satisfactory and it somehow "flattened" the reality. The model we applied was too simple, and we could not discriminate people's participations. Indeed, some answers to a thread are longer than others, and replies are also directed.

This example challenged us and led us to consider the generalized entanglement model we propose in Section 3.5.1.

6.2.6 *Two-mode networks*

A few examples can be found in the literature in social analysis of two-mode networks. Among them we tried our approach on two well-known affiliation networks: *The Women of Deep South* (Davis et al., 1941), Figure 6.18, studied for example in (Borgatti and Everett, 1997; Neal, 2013); and *Paul Revere's Ride* (Fischer, 1994), Figure 6.17, studied[10] in (Han, 2009).

These datasets affiliate people to groups or events they have participated in. The full study would be quite lengthy and applies to sociology, nonetheless, our analytical tools can highlight the structure of the two-mode association and definitely eases the comprehension of the structure that ties these people and groups together.

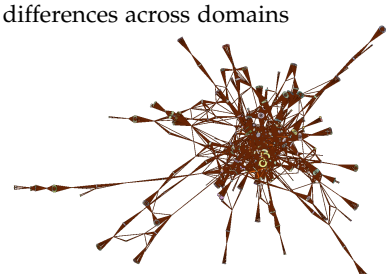
6.3 Perspectives

The fields of application are, of course, not limited to only document collections and social networks. We have seen in Section 2.4.2 that multiplex networks can be good model candidates for a very wide range of applications, and we believe our entanglement analysis also applies in any of them.

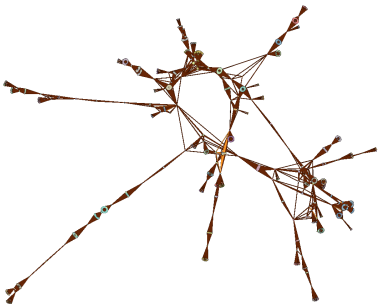
Among them we can mention applications to financial data, and we worked with the United Nations Development Program[11] (Renoust et al., 2013a), and were later appointed by the WorldBank[12]

10. Another very interesting and accessible study can be found at http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/.

11. With the UNDP (http://www.undp.org/) we got interested in how projects share common suppliers, and found interesting structural differences across domains



The *mining* domain is very much tangled, as opposed to the *education* domain



12. The WorldBank (www.worldbank.org) has put significant effort in opening their financial data (which can be found at http://data.worldbank.org/), and invited me as an external network expert during the *WorldBank DataDive* in DC, leading us to an on-going collaboration.

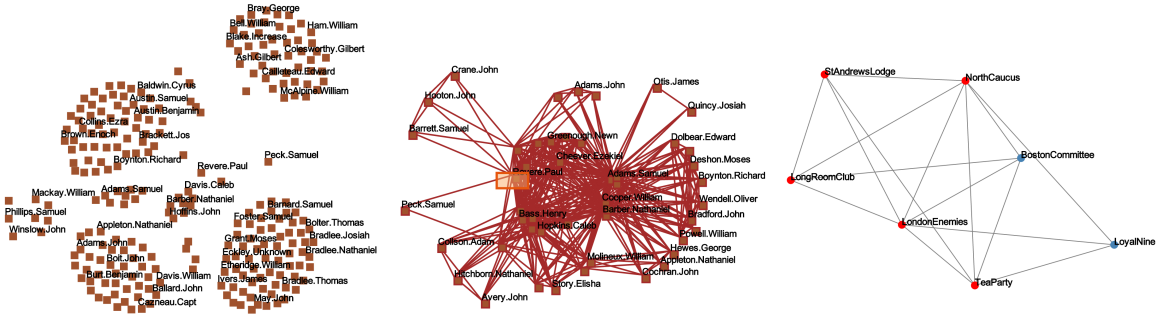


Figure 6.17: Paul Revere's Ride, with people as substrates and secret societies as catalysts. On the left the structure of communities in the dataset is striking. In the middle, a subset of people all involved in multiple societies. We have leapfrogged a selection from Paul Revere (middle) and all the societies to which Paul Revere belongs (right). A last leapfrog ends up in reaching *everybody* in the network.

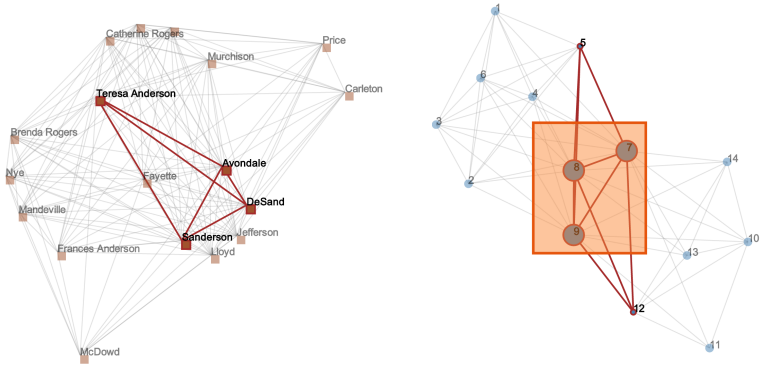


Figure 6.18: Women of the Deep South, as substrates, and events they attended as catalysts. A leapfrogged selection from the women who have attended together the exact three central events (operator *And*).

(Figure 6.19) to bring a different perspective on project procurement (Renoust et al., 2013b). This work is addressing one most interesting issue: the association of the network structure to some quality variables. We suspect the outcome of a project may be related to the many layers of ties that link the projects together, for example, geographical location, themes, sectors, expertise, social network of experts, suppliers, relationships with the teams implementing the projects, existing infrastructures, *etc.* We can only see here the many other challenges it opens: multiplex autocorrelation, combined with analysis and visualization of multivariate informations, hierarchical information, under the light of multiplex networks.

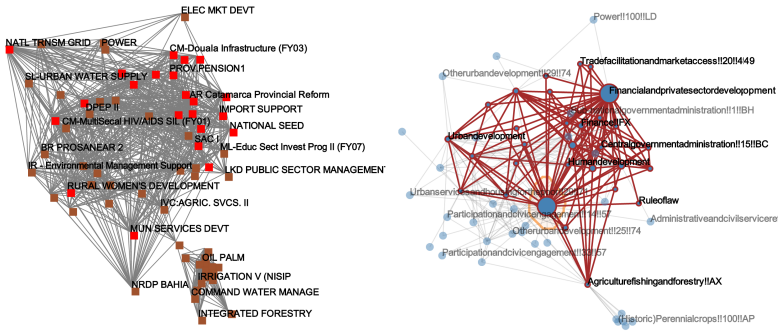


Figure 6.19: Projects and themes/sectors from the WorldBank's project API and IEG Ratings. The selected area correspond to development projects related to both public administration and private sector development.

Our laboratory at LaBRI joins also the bioinformatics field. Beyond the document/concepts associations which can be studied (very similar to our INA document networks), we are also investigating many interactions and all levels of abstraction involved in a living organism. Figure 6.20 (left) displays the different factors at play. Interaction between the functional level and the biomechanical level (such as Figure 6.20, right) is an example of study (it may answer a

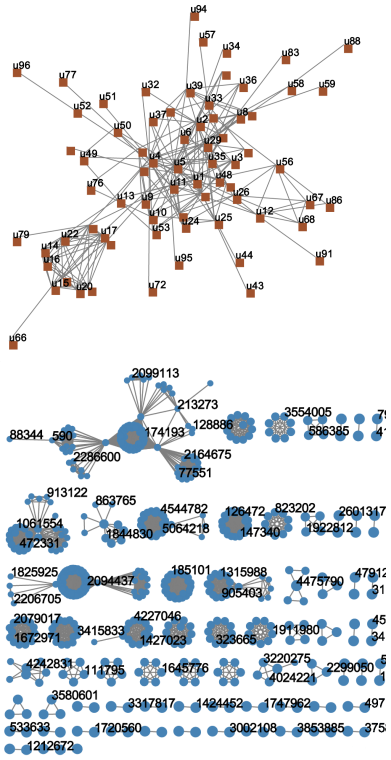
question such as: what is the role of X given Y?). Even further, interaction across every single factor, described in Figure 6.20 left, is not only possible, but also very much studied in the field. The study of entanglement is challenged here at a most interesting level. The relationships across the many families of possible catalysts could be hierarchical, or transitive, or even completely heterogeneous, with a different meaning for the entanglement associated cohesion for each pair of factors we are taking into account, such as the interaction across factors, somewhat describe another graph itself. The levels of granularity could go down to the protein sequences and up to organisms and external conditions. The combinations are possible in between and within, at every and each of these levels.

Applications simply widen our sight, opening us to new methods, and posing new challenges. Every time we apply our methodology to new fields of application, we see potential improvement either on the analytical side or the visual tools required to answer high-level issues.

Every application seems to present different patterns in their catalyst interaction structure[13], and comparing and characterizing these structures would be one great challenge.

One next step toward that direction may soon be the means to switch points of views (the substrate entities on which we observe the network) and choose catalysts on-the-fly. Another need would be to better manage the multiple catalyst components and maybe provide visual cues at a "local" cluster level.

13. Another interesting structure, a peer-to-peer network (peers are substrates, and files catalysts), with a lot more packed catalysts than substrates:



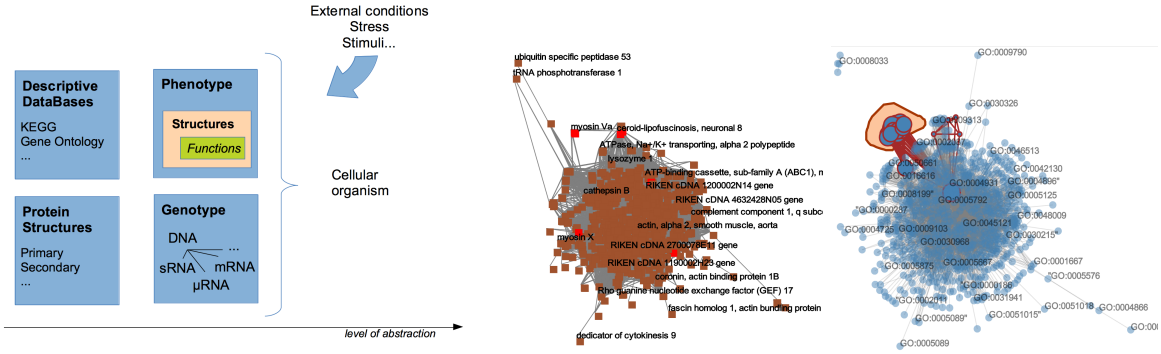


Figure 6.20: Complexity in bioinformatics: the study of living organisms shows incredible levels of complexity in which all interactions are studied.
(Left): A cellular organism reacts to external factors (e.g. temperature, environment...), and displays observable features, its phenotype, in which we can identify functions, and all such information have been structured in many databases. The organism can also be segmented according to its genotype, with interactions between DNA and many types of RNA, and some of them correspond to proteins. A very common task in bioinformatics is to understand the functions and expression of a group of protein under given conditions.
(Right): an example application. This network corresponds to an identified functional group associating *proteins* (substrates) and *functions* as described in the Gene Ontology (GO, catalysts). Here, we extend our search from a group of functions described by the GO, and question the other *functions* associated to this group through *protein* interactions, using our *leapfrog* interaction technique.

6.4 Conclusion

We have presented in this Chapter detailed examples and many applications which illustrate our methodology and confirmed the legitimacy of our approach. We firstly described the application to the field that has motivated our research: INA's documents. We have brought detailed examples of the use of our visual analytic methodology, and even displayed the results of a small scale user experiment. We have presented extended applications to social networks analysis, and went further to the point at which we extended our model. The confrontation of our choices with domains external to our motivating application is such a source of richness that it forces us to think *out of the box* and opens wider perspectives.

The applications of our work to documents has been motivations for publications since the beginning of this thesis: (Viaud et al., 2010; Renoust et al., 2011a,b, 2013f,d) and an early technical report (Melançon et al., 2012) that led us to the more mature work of (Renoust et al., 2013e). Applications to social networks have been recently presented in (Renoust et al., 2013d) and (Renoust et al., 2013c). Collaboration with policy makers for the developing world has led to the publication of (Renoust et al., 2013a) and (Renoust and Begovic, 2013), and to a further collaboration with economists (Renoust et al., 2013b).

The field of possible applications is very wide, and we are continuously seeking such collaborations with experts, and to confront and challenge our tools by applying them to new fields.

CONCLUSION

7

This Chapter summarizes and concludes this thesis. After recapping the work in its whole content, we will open perspectives for new contributions. Before concluding, we will step back and discuss more widely of the implications of the contributions of this thesis, and networks in general, under the light of anthropology and philosophy.

7.1 Summary

The first part of this thesis is purely interested in numerical analysis of complex systems under the perspective of multiplex networks.

Chapter 2 justifies and explains the application of graph models and especially the multiplex graph model for document analysis. It has presented the general setting that motivated our research: annotated document groups, and common challenges and techniques used to manipulate them. We introduced then, models of networks, complex networks, and multiplex networks, very suitable to represent relationships within real-world data. We present alongside common then specialized techniques for network, complex network and multiplex network analysis. We closed this Chapter with the multiplex network model we apply to our document groups, which shapes the base line for the rest of our work: documents are presented as *substrates* and terms indexing them as *catalysts* composing the different layers of interaction across documents.

Chapter 3 presents our first contribution in network analysis. We have brought here the *entanglement* analysis from the inspiring work of Ronald Burt and generalized the concept to *any* multiplex graph. With ground foundations in linear algebra, we have detailed the methodology which computes *catalyst entanglement* indices, and *group entanglement intensity* and *homogeneity* measures. This computation also brings along a new object of interest, the *catalyst interaction network*. After detailed discussions on how to handle an *entanglement* analysis of a multiplex network, giving behaviours and limitations of the different measures, we have closed this Chapter by opening new perspectives that include a *weighted model of entanglement*, and *local entanglement intensity* and *homogeneity* at the level of *substrates*.

The second part of this thesis discusses relevance and embedding of our analytical tool into a visual framework.

Chapter 4 reports how networks are ideal objects to be visualized and support a higher level of analysis. Here, we went back to the original process of sense-making, *i.e.* how people people process complex information, and visual information, in order to reach higher *comprehension*. We explained how visualizations and visual analytics frameworks can be designed to explicitly answer the multiple challenges and tasks of sense-making. Especially, in this part, we have introduced a nested model and a typology of tasks that supports our own framework design. We have then explained how the principles of human perception can be utilized to efficiently enforce such design. We have continued by presenting the toolbox offered by the field of information visualization and interaction, and particularly the case of multivariate data and network visualization. We have finally closed this chapter by highlighting the general settings of visual analytics that seems to us adapted to our analytical model: multiple network views that allow selection brushing and linking.

Chapter 5 brings all the details on our implementation of a visual analytics framework for multiplex networks supported by the entanglement analysis. We have started by presenting the highest-level tasks that we aim to address: *comprehension* of the structure of interactions in a multiplex network. After what we have specified our tasks into the lower level of data abstractions and manipulation which explicits how entanglement driven data manipulation supports *comprehension* in complex representations. We have then gone down to the level of visual encoding and interactions, which reports the mechanisms we put at play to enable users to visually engage with the multiplex network representation. Along the way we have presented one particular layout that contributes to the bipartite network visualization. We have finally validated the relevance of our design with a multilevel typology of tasks, and discussed two particular cases, our use of mental mapping to link visualizations, and our proposed leapfrogging interaction that allows to easily jump between the different levels of abstraction we propose in our visual framework. We have discussed potential completion of our work to more multivariate information, and inclusion in the visual framework of the perspectives opened in Chapter 3. We have finally concluded this Chapter by separating the visual framework and the entanglement analysis, and formulating what advantages makes it suit to support any multivariate visual analysis.

This last part of our document finally brings into play our visual analytics framework.

Chapter 6 demonstrates the usefulness of our approach by confronting it to real application cases. Document group analysis being the initial motivation of this thesis, we have firstly exposed detail examples with document groups. We have reported then a very short user experiment that have comforted us in our approach and design. We have also exposed general thoughts extracted from our experi-

ence manipulating document groups with our framework. We have emphasized the role of the catalyst interaction network's shape in the explanation of what compose a group of documents and discussed other potential extensions to document group explorations. We have then switched to social networks applications, with the detailed cases of relationships between movie directors and actors, and of topics in co-authorship network in scientific research. We have also opened to other social networks, some of which have motivated extensions of our analysis and visual framework. We have finally opened to a very wide range of applications that illustrates the generic aspect of our approach. We have discussed the benefits of scientific transfer between application domains, and potential perspectives for characterizing such application domain through our methodology.

7.2 Perspectives

During this manuscript, we have opened perspectives at all levels. This section aims at summarizing them.

On the analytical side, we have already proposed a generalized entanglement model for weighted multiplex networks analysis. Further extension of the model would include finer weights in directed, multi-scale, and dynamic multiplex networks. We would also like to focus on the reciprocal entanglement study of bipartite networks, where two multiplex networks can be inferred (each side could be considered as substrates). Eventually, we could extend our approach to n-partite networks, and especially taking into account the different relationships between every partial graph, with the specific case of hierarchical relationships.

Another interesting topic would be to study how substrates contribute in creating entanglement in a multiplex network. In this manner, we have proposed perspectives in the "local" entanglement homogeneity and intensity of a node within its neighbourhood, with promising early results. These could indeed drive a clustering algorithm. This consequently questions the efficiency of algorithms to compute entanglement. Optimizations would allow for analysis of larger networks. We have not yet explored all the leads from Burt and Schøtt (1985), which propose a distance function between catalysts to create equivalence classes of catalysts. It may come helpful when dealing with a large number of catalysts, thus reducing the final number of catalysts we should consider.

Linear algebra brings us strict boundaries for a maximum eigenvalue of a nonnegative matrix. We have exploited the maximum value since its interpretation was clear to us. However, the minimum boundary, and the span it defines with the maximum, may also be an important information we would like to study and take into account. More theoretical study would bring us to consider the behaviour of

entanglement in random network models, and compare them with real world graphs. A large survey with comparison of the different entanglement measures on such random graphs with known metrics will also be very interesting to conduct. It would eventually lead us to a better comprehension of what links the group entanglement measures to the topology of the catalyst interaction network.

On the visual analysis side, we are considering the integration of the various measures on the substrate side. An interesting solution might be the visualization of a heat map of tangled areas in the network. It is always possible to add up new visual cues of any sort. However, the issue is on the manner we can include those cues while keeping the visual representation simple enough, without overloading our users' cognition. We plan to enhance the rendering of edges in node-link diagrams, with edge bundling and maybe metric mapping onto edges, but we are studying ways that would not overload the readability of our diagrams. Power graphs extended to the multiplex model might as well be an interesting alternative.

The nature of the data we handle is often multivariate, and the way we combined brushing, linking, and mapping, with *leapfrogging* might also be extended to heterogeneous multivariate information. It relies on particular properties and data operations, that once formalized, can apply to any multiple heterogeneous visualizations.

We have started to study strategies to link traditional information visualization techniques with our current setting. The challenge of course increases with the number of criteria we would take into account, but the field of visualization is already full of interesting leads. Additionally, we can approach the entanglement analysis in multivariate multiplex networks with several families of catalysts by managing multiple substrate/catalyst associations. A very interesting challenge is to shrewdly link all these different perspectives of a same object.

The *harmonized* layout we have proposed relies on a bipartite model that essentially associates entities we observe to properties of interest (i.e. *specific* catalysts). Many works are also expected to extend it and we would like to make it more flexible, such as the possibility for users to choose the metric, or even specify manually what is a *specific* catalyst to their mind. We could also tune the rules on which we distinguish the two families of *general* and *specific* catalysts. We plan to further study the influence of edge weights and drive the layout computation. A good way to test all these possibilities will be to compute the layout on-the-fly and visually observe the results.

The "filter" interaction we have yet provided is only focused on removing/subset-ing substrates, we could easily extend the concept to catalyst filtering. What would be even more interesting is, *before* any filtering of the data, to preview the impact on the network. This, combined with basic "edition" interactions, could hopefully show the usefulness of our system into creating ground truth for further stud-

ies, and even in a long term detect and correct anomalies in "dirty" network data.

We have only validated our work through detailed design, tasks, and application examples. Although we led a small scale user experiment, it does not qualify as an evaluation of our work. Further perspectives will take on designing and implementing a full scale user evaluation for multiplex networks visual analysis.

On the application side, we are eager to confront and extend our methodology and visual framework to as much applications as possible. Collaboration with field experts for thorough case studies are necessary to validate our approaches on real-world applications.

The shape of the term interaction network seems able to narrate a group of INA documents. We would like to further investigate the properties of this network and maybe explain news events in their dynamics. Point-by-point comparisons with existing social analysis on two-mode affiliation networks such as the example of Paul Revere's Ride will be an interesting study to complete. The example of Edgeryders brings us further into studying quantitative social exchanges with the help of a weighted model. Another example with the WorldBank's data in which very concrete questions brings us a lot of challenges. It continues to motivate us in enhancing our approach. The need is to not only integrate new multivariate information (which could take the form of other networks), but also to use our model for a different purpose that follows exploration: explanation. Even further, the example of Bioinformatics, which very concretely face *complexity*, confirms us in the need to lift the analysis to one more level of complexity, and tackle multiple factors as families of catalysts.

Finally, we are always in search for applications, and we believe our tools to be generic enough, so that we can approach a very wide panel of new applications, in which complex networks already apply.

7.3 Discussion

Now that we have brought forward the contributions and many perspectives of our work, we can put things into perspective. We will raise our eyes beyond computer science, and attempt to catch a glimpse of the *big picture*. A striking fact came to me this last year of my thesis, when I was finally confident enough in the relevance of our tools and I came to confront formally or informally our methodology to different domains. There seem to be little limit to the applications of networks when it comes to understanding the structure of *things*. The long list of applications, presented in Sections 2.1.2 and 2.4.2, shows evidence of the broad applications of network analysis, remarkable for such a young science.

Eventually, with the revolutions in communication, access and delivery of information, networks did not limit their selves as inherent

structures of human interactions. We could go from the invention of the writing, the transportation revolutions (such as the creation of the railway network), to the revolution of the Internet, the international network. Each created network had a profound impact in the evolution of human societies. Networks gained so much success that they became the centre of attention, obvious and visible, "materialized" by the means of new "social media". These same techniques that have enabled revolutions of human societies, are also enabling the study and formalization of the same human societies.

Anthropologist Alvin Wolfe have offered a rare paper (Wolfe, 2010)[1] summarizing fifty years of watching human systems though the lens of *"the network thinking"*. Wolfe observes human systems, and his focus is on *comprehension* of such system.

1. The whole paper is eye-opening. It is a demonstration of the usefulness of network models and representations. We hope this small discussion will catch our reader's attention, and encourage its reading, of course *after* closing this manuscript.

Beyond the many applications and detailed studies he came up with, his point of view on complexity attracted our attention. He was very interested in the complexity of human systems, and confirmed with field work, that hierarchical structures appears through the dynamic aggregation of interactions between smaller structures, gaining in complexity – which is precisely the definition of complex networks. We could paraphrase the whole document that rules in favor of application of network models to study and observe complexity in all things, but this quotes sums it all: *"At each level there are different systems, in some cases billions of systems, and each of those systems is a network. And the whole is also a network, a network of networks."*

Among the many points of views he brings on the study of complexity, Wolfe discusses Herskovits' concept of *"cultural relativism"* (Herskovits, 1964), which states how things relate to each other in human cultures: *"Mass, gravity, and movement are interrelated in the one case* [general relativity in physics]; *judgment, experience, and learning, in the other* [cultural relativism] *."* Later in his paper, he points out the importance of the notion of structural and regular equivalences for understanding complex systems. What we have proposed in this manuscript joins this point of view. As he hopes research in network analysis would do, we have attempted to bring tools to study a bit further the different equivalences among which entities relate to one another. We see them in the study of catalyst interaction.

To the question, *why do network analysis?* Wolfe answers by using network analysis for *understanding "the more complex* ***wholes*** *such as multinational corporations and supranational systems. Those important entities and the problems they represent should not be left to economists and politicians!"*.

Network analysis applies anywhere we want to study relationships. In the manner of *cultural relativism*, we may suspect the network thinking to be a pure mental view. It may be proper to the human spirit, and Philosophy, the mother of all Sciences, may indeed point at the roots and uses of such complex network analysis. This

part of the discussion is highly motivated by the reading of William James (1907)[2] pragmatism, and especially his Lecture IV: *the one and the many*[3]. Wolfe pointed out the understanding of the complex *whole*, through the understanding of its *parts*. Of course, the definition of the whole and its parts is essential for formal logical reasoning, and Aristotle introduced here the first network of ideas that, put in logical relations, bring the idea of a *whole* named "Truth"[4].

Our minds seem to always separate the whole (the monistic view) to the parts (the pluralistic view). Philosophy has been divided for centuries along these axes, as if one should take over the other. The thinking process is always eager to aggregate parts for a greater understanding of a whole, and to separate a whole in parts, to grasp comprehension[5] of the whole. There might be an aesthetic canon in seeking for a whole unifying the parts, and, perhaps, we might be fooled by an evolutionary induced side-effect that pleases our lazy brains (De Neys et al., 2013). However the question remains, how does the whole relate to the parts?

James – in his Lecture, that we will quote here – argues for better comprehension of the world through pragmatism, as a union of both monism and pluralism, by leaving *"the one and the many on a par"*. We will not go to the debate that opposes philosophers disputing the superiority of a monistic approach over pluralism, but we will explain how, in his discourse, the network representation joins both views. In the end of his Lecture, James clearly stated how human beings never ceased to construct networks, giving what could be a philosophical definition of complex networks: *"The world is one just so far as its parts hang together by any definite connexion. It is many just so far as any definite connexion fails to obtain. And finally it is growing more and more unified by those systems of connexion at least which human energy keeps framing as time goes on."* This clearly joins Wolfe's empirical observation throughout his career. It also goes further, in the construction of a culture, aggregating human knowledge through conceptual connexions – which relates to the notion of cultural relativism.

The complex network representation we propose in this manuscript, as a multiplex network associating substrates and catalysts, perfectly illustrates Jame's pragmatism. It envisions at once the *whole*, the *parts* and the *Substance*[6]. James wondered how we could move from the abstract views to the concrete views of the universe. He especially questioned these different ways in various manners: such as discourse, influence, causality, genericity, or purpose. More particularly, he wondered if these ways are continuous.

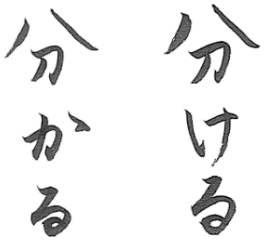
In the methodology we proposed, the whole represent the system we observe, that is composed of parts, the substrates. The *substance*, which induces continuity between the parts and the whole is precisely represented by the connexions between the parts, and corresponds to the linkage of our network. The substance is thus composed of many parts (substrates), and we have identified in catalysts the many possible ways to link the parts in the whole, ab-

2. William James directly descends through John Stuart Mill and Auguste Comte from the spiritual lineage of Claude Henri de Rouvroy, comte de Saint Simon, considered as the father of the Philosophy of networks.

3. The Lecture can be found online here: http://www.authorama.com/pragmatism-5.html

4. No wonder that Porphyry of Tyros have illustrated Aristotle with the earliest mind maps (See Chapter 1).

5. Etymology also confirms the use of both approaches: *"comprehend"* takes its roots from the latin *"com"* and *"prehendere"*: "take together".



In Japanese "comprehend" translates into *"wakaru"* (on the left), which shares the same root with (on the right) *"wakeru"*: "classify, distinguish". The root is 分く, *"waku"*: "divide".

6. The *"universal substance which alone has being in and from itself, and of which all the particulars of experience are but forms to which it gives support."*

stractions linking concrete observations. Another interesting aspect is the inherent *inseparability* of substrates and catalysts in our model. If substrates or catalysts would be put aside alone, the new whole they would compose would be without substance leading to a biased Truth, if any Truth. To each of the different catalysts corresponds a smaller network, a smaller whole. Our visual framework materialize substrates and catalysts with a similar manner, and by going back and forth between them, we finally bring *"the one and the many on a par"*.

We have presented in this discussion this importance of network models as means for comprehension of complexity. Wolfe has shown us their relevance on many – if not any – system that involve human interaction. With the pragmatism of James, networks as the perfect tools to unify views of whole and parts, supporting comprehension of a Truth, especially in the case of Wolfe's anthropology. With its root deeply embedded in human cognition, there is no wonder that visualization of networks helps this perspective.

Discussing deeper the role of visualization in the evolution of human's ways of thinking is beyond the scope of this section[7], but we can nevertheless conclude on the vision of multiplex networks as we propose it – *i.e.* under the prospective of entanglement. It not only confirms James' pragmatism, and Wolfe's network thinking in every human system, but it also shines a different light on the continuity between the whole and the parts, which have definitely a role to play in Wolfe's evolution of a supranational system.

7. However our curious reader could refer to the work of anthropologist Jack Goody, *The Domestication of the Savage Mind* (Goody, 1977).

7.4 Conclusion

This manuscript has presented a new methodology for multiplex network analysis and brought it all the way to application.

We have demonstrated its relevance in network analysis, we have contributed to complex network visualization, and we have transferred the technology to multiple application domains. Our approach has been motivated by concrete real-world issues of document analysis and document group relevance, to which we have answered through multiplex graph analysis. We have been able to generalize the approach to *any* multiplex graph model, and proposed original measures. We have exploited the outputs of our analysis to support understanding of a complex network with visualization and interaction. We have put a great deal of efforts into answering the many challenges that visual analysis of these objects impose. We have carefully identified the tasks which supports understanding, and accordingly designed visualization and interaction in our framework. It includes brushing and linking between heterogeneous multiple views, with the preservation of mental mapping, which even allowed us to contribute with a new visual layout and interaction. We have validated

our approach by confronting it to real world applications, of which some have challenged our design. By tackling these new challenged, we were able to even further improve our analysis ground, opening many perspectives.

The work done during this thesis was clearly motivated by curiosity and eagerness for comprehension. As a researcher, I have been completely fascinated by the elegance and beauty of networks in all their aspects. They are solid theoretical tools for the study of relationships and are very abstract data structure, but yet have this unique ability to be visually displayed in the same way we picture them in our mind. The clearest evidence is the use of "mind maps" in many situations, from classrooms to business meetings, and yet no one questions it and everybody is able to understand mind maps. Giving sense to the visual representations of the networks, and manipulating them in order to comprehend and extract new insights was also one most interesting aspect of this thesis. The objects we manipulate could be very complicated and we had to really exploit our visual capacities, often by using the simplest yet most efficient techniques, to make sense of our networks, even during the intermediate stages prior to analysis.

We have the chance, doing data science, to address common issues across many domains. Each time we have confronted our tools to other application domains, it was motivated by informal discussions with other domain experts. These confrontations have enriched us, very much like traveling to foreign countries enriches us[8]. It required us to adapt our vocabulary, similarly to foreign languages, and to understand what is essential in the information, abstractions, and questions our experts are facing. It requires us the ability to communicate the essence of our approach so we can reach a common goal. This synthesis work always pushed us further than our own boundaries, requestioned our models and methods, so we could reach higher wisdom. The compilation of these many wisdoms acquired during my encounters, with a wide range of thinkers and experts, is what inspired me in writing this document.

8. "There is no man more complete than the one who travelled a lot, who changed the shape of his thoughts and his life twenty times." Alphonse de Lamartine, extract from *Voyage en Orient*

Finally, since we have opened this manuscript with a quote, we will close it with another one:

> *"Don't think about why you question, simply don't stop questioning. Don't worry about what you can't answer, and don't try to explain what you can't know. Curiosity is its own reason. Aren't you in awe when you contemplate the mysteries of eternity, of life, of the marvelous structure behind reality? And this is the miracle of the human mind – to use its constructions, concepts, and formulas as tools to explain what man sees, feels and touches. Try to comprehend a little more each day. Have holy curiosity."*
> Albert Einstein during his fourth conversation with William Hermanns (in Hermanns and Einstein (1983), p 138).

AUTHORS' PUBLICATIONS

8

8.1 International Journals

- David Auber, Charles Huet, Antoine Lambert, Benjamin Renoust, Arnaud Sallaberry, and Agnès Saulnier, "GosperMap: Using a Gosper Curve for Laying out Hierarchical Data" *IEEE Transactions on Visualization and Computer Graphics*, Nov. 2013, (vol. 19 no. 11) pp. 1820-1832.

8.2 International Conferences

- Benjamin Renoust, Guy Melançon, and Marie-Luce Viaud, "Measuring group cohesion in document collections" *The 2013 IEEE/WIC /ACM International Conference on Web Intelligence (WI)*, Atlanta, USA, Nov. 2013. *Paper accepted, to be published.*
- Benjamin Renoust, Guy Melançon, and Marie-Luce Viaud, "Assessing group cohesion in homophily networks", *ACM/IEEE 2013 Advances in Social Network Analysis and Mining (ASONAM)*, Niagara Falls, Canada, Aug. 2013. *Extended version submitted by invitation to "Lecture Notes on Social Networks" (LNSN) Series, by Springer.*
- Benjamin Renoust, Guy Melançon, and Marie-Luce Viaud, "Document Corpus Analysis Based on Term Entanglement" *INSNA XXXIII Sunbelt Social Networks Conference for the International Network for Social Network Analysis*, Hamburg, Germany, May 2013.
- Marie-Luce Viaud, Benjamin Renoust, Agnès Saulnier, and Jérôme Thièvre, "Bringing Interactive Contextual Maps to Users for Information Retrieval" *2010 NEM Summit*, Barcelona, Spain, Oct. 2010

8.3 International Workshops

- Benjamin Renoust, Kenneth M. Chomitz, and Alex H. McKenzie, "Network analysis applied to project risks identification: what risk factors can explain project outcome?" *WorldBank DataDive in DC*, Washington DC, USA, Mar. 2013. *Invited researcher as a networks expert.*

8.4 Domestic Conferences and Workshops

- Benjamin Renoust, Marie-Luce Viaud, and Guy Melançon, "Mesure de cohérence dans la découverte et la visualisation d'évènements médiatiques", *Seconde conférence sur les Modèles et l'Analyse des Réseaux: Approches Mathématiques et Informatique (Marami)*, Grenoble, France, Oct. 2011.
- Benjamin Renoust, Guy Melançon, and Marie-Luce Viaud, "Mesurer l'intrication sémantique dans une collection de documents", *13e Conférence francophone sur l'Extraction et la Gestion des Connaissances / Fouille de Grands Graphes (EGCFGG13)*, Toulouse, France, Feb. 2013.
- Benjamin Renoust, Marie-Luce Viaud, and Guy Melançon, "Détection, visualisation et validation d'évènements médiatiques" *Atelier Fouille de données Complexes, Journée Fouille de Grands Graphes*, Paris, France, 2011.

8.5 Other Publications

- Milica R. Begovic, "Q&A: The day a big data scientist met a development organization" *UNDP Voices from Eurasia*, Feb. 2013 *Interview*.
- Benjamin Renoust, Milica R. Begovic, and Giulio Quaggiotto, "Big data and development organizations: What happens when you move from theory to practice?" *UNDP Voices from Eurasia,*, Jan. 2013 *Workshop report*.
- Nicolas Hervé, Marie-Luce Viaud, Jérôme Thièvre, Agnès Saulnier, Julien Champ, Pierre Letessier, Olivier Buisson, Alexis Joly, and Benjamin Renoust, "OTMedia: L'Observatoire TransMedia" *Technical Report INA, OTMedia, ANR-10-CORD-000*, Jun. 2013.
- Guy Melançon, Benjamin Renoust, and Marie-Luce Viaud, "Mesurer la cohésion sémantique dans les corpus de documents" *Technical Report RR-8075 - HAL - INRIA* , Sep. 2012.

BIBLIOGRAPHY

Agneessens, F. ; Roose, H. ; and Waege, H. . 2004. *Choices of theatre events: p* models for affiliation networks with attributes*. In Metodološki zvezki, vol. 1, no. 2, pp. 419–439. Cited on p. 23.

Albert, R. and Barabási, A.-L. . 2002. *Statistical mechanics of complex networks*. In Reviews of modern physics, vol. 74, no. 1, pp. 47–97. Cited on p. 9.

Allan, J. . 2002. *Introduction to topic detection and tracking*. In Topic detection and tracking, pp. 1–16. Springer. Cited on pp. 12 and 13.

Alpaydin, E. . 2004. *Introduction to machine learning*. MIT press. Cited on p. 15.

Amar, R. and Stasko, J. . 2004. *A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations*. In 2004 IEEE Symposium on Information Visualization, pp. 143–150. Cited on pp. 63, 69, 87, and 88.

Amar, R. ; Eagan, J. ; and Stasko, J. . 2005. *Low-level components of analytic activity in information visualization*. In Proceedings of IEEE Symposium on Information Visualization, 2005., pp. 111–117. IEEE. Cited on p. 69.

André, P. ; Schraefel, M. ; Teevan, J. ; and Dumais, S. T. . 2009. *Discovery is never by chance: designing for (un) serendipity*. In Proceedings of the seventh ACM Conference on Creativity and Cognition, pp. 305–314. ACM. Cited on p. 62.

Aral, S. ; Muchnik, L. ; and Sundararajan, A. . 2009. *Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks*. In Proceedings of the National Academy of Sciences, vol. 106, no. 51, pp. 21544–21549. Cited on p. 23.

Arenas, A. ; Fernandez, A. ; and Gomez, S. . 2008. *Analysis of the structure of complex networks at different resolution levels*. In New Journal of Physics, vol. 10, no. 5, p. 053039. Cited on p. 19.

Auber, D. ; Archambault, D. ; Bourqui, R. ; Lambert, A. ; Mathiaut, M. ; Mary, P. ; Delest, M. ; Dubois, J. ; and Melançon, G. . 2012. *The Tulip 3 Framework: A Scalable Software Library for Information Visualization Applications Based on Relational Data*. Tech. rep., INRIA Bordeaux Sud-Ouest HAL, Technical Report RR 7860. Cited on p. 83.

Auber, D. ; Chiricota, Y. ; Jourdan, F. ; and Melançon, G. . 2003. *Multiscale navigation of Small World Networks*. In Proceedings of IEEE Symposium on Information Visualisation, 2003, pp. 75–81. IEEE Computer Science Press. Cited on p. 126.

Auber, D. . 2004. *Tulip – A huge graph visualization framework*. In Graph Drawing Softwares, Mathematics and Visualization, pp. 105–126. Springer. Cited on p. 81.

Auber, D. ; Huet, C. ; Lambert, A. ; Renoust, B. ; Sallaberry, A. ; and Saulnier, A. . 2013. *GosperMap: Using a Gosper Curve for Laying out Hierarchical Data.* In IEEE transactions on visualization and computer graphics, vol. 19, no. 11, pp. 1820–1832. Cited on p. 10.

Baeza-Yates, R. ; Ribeiro-Neto, B. ; et al. 1999. *Modern information retrieval*, vol. 463. ACM press New York. Cited on p. 16.

Bakshy, E. ; Rosenn, I. ; Marlow, C. ; and Adamic, L. . 2012. *The role of social networks in information diffusion*. In Proceedings of the 21st international conference on World Wide Web, pp. 519–528. ACM. Cited on p. 23.

Barber, M. J. . 2007. *Modularity and community detection in bipartite networks*. In Physical Review E, vol. 76, no. 6, p. 066102. Cited on p. 28.

Battiston, S. and Catanzaro, M. . 2004. *Statistical properties of corporate board and director networks*. In The European Physical Journal B-Condensed Matter and Complex Systems, vol. 38, no. 2, pp. 345–352. Cited on p. 23.

Bauckhage, C. . 2008. *Image tagging using PageRank over bipartite graphs*. In Pattern Recognition, Lecture Notes in Computer Science, vol. 5096, pp. 426–435. Springer. Cited on p. 29.

Bavaud, F. . 2013. *Testing spatial autocorrelation in weighted networks: the modes permutation test*. In Journal of Geographical Systems, vol. 15, no. 3, pp. 233–247. Cited on p. 29.

Becker, R. A. and Cleveland, W. S. . 1987. *Brushing scatterplots*. In Technometrics, vol. 29, no. 2, pp. 127–142. Cited on p. 83.

Belkhir, W. and Santocanale, L. . 2007. *Undirected graphs of entanglement 2*. In FSTTCS 2007: Foundations of Software Technology and Theoretical Computer Science, Lecture Notes in Computer Science, vol. 4855, pp. 508–519. Springer. Cited on p. 37.

Berge, C. . 1984. *Hypergraphs: combinatorics of finite sets*, vol. 45. Access Online via Elsevier. Cited on p. 8.

Berge, C. and Minieka, E. . 1973. *Graphs and hypergraphs*, vol. 7. North-Holland publishing company Amsterdam. Cited on pp. 8 and 23.

Berry, A. ; Pogorelcnik, R. ; and Sigayret, A. . 2011. *Vertical Decomposition of a Lattice Using Clique Separators.* In CLA, CEUR Workshop Proceedings, vol. 959, pp. 15–29. Cited on p. 79.

Bertin, J. . 1967. *Sémiologie graphique*. Paris: Mouton, 1966, 432p. Cited on pp. 3, 74, 75, 76, and 77.

———. 1981. *Graphics and graphic information processing*. Walter de Gruyter. Cited on p. 93.

Bertini, E. ; Di Girolamo, A. ; and Santucci, G. . 2007. *See What You Know: Analyzing Data Distribution to Improve Density Map Visualization.* In Proceedings of the 9th Joint Eurographics/IEEE VGTC conference on Visualization, pp. 163–170. Eurographics Association. Cited on p. 74.

Berwanger, D. and Grädel, E. . 2005. *Entanglement–a measure for the complexity of directed graphs with applications to logic and games*. In Logic for programming, artificial intelligence, and reasoning, vol. 3452, pp. 209–223. Springer. Cited on p. 37.

Bezerianos, A. ; Chevalier, F. ; Dragicevic, P. ; Elmqvist, N. ; and Fekete, J.-D. . 2010. *Graphdice: A system for exploring multivariate social networks*. In Computer Graphics Forum, vol. 29, no. 3, pp. 863–872. Cited on pp. 81, 82, and 84.

Bhavnani, S. K. ; Bellala, G. ; Victor, S. ; Bassler, K. E. ; and Visweswaran, S. . 2012. *The role of complementary bipartite visual analytical representations in the analysis of SNPs: a case study in ancestral informative markers*. In Journal of the American Medical Informatics Association, vol. 19, no. e1, pp. e5–e12. Cited on p. 80.

Bier, E. A. ; Stone, M. C. ; Pier, K. ; Buxton, W. ; and DeRose, T. D. . 1993. *Toolglass and magic lenses: the see-through interface*. In Proceedings of the 20th annual conference on Computer graphics and interactive techniques, pp. 73–80. ACM. Cited on p. 81.

Blei, D. ; Ng, A. ; and Jordan, M. . 2003. *Latent Dirichlet allocation*. In Journal of Machine Learning Research, vol. 3, no. 5, pp. 993–1022. Cited on pp. 15 and 16.

Blondel, V. D. ; Guillaume, J.-L. ; Lambiotte, R. ; and Lefebvre, E. . 2008. *Fast unfolding of communities in large networks*. In Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, p. P10008. Cited on pp. 17, 20, and 78.

Boers, N. ; Bookhagen, B. ; Marwan, N. ; Kurths, J. ; and Marengo, J. . 2013. *Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System*. In Geophysical Research Letters, vol. 40, no. 16, p. 4386–4392. Cited on p. 23.

Bonacich, P. . 1972. *Technique for analyzing overlapping memberships*. In Sociological methodology, vol. 4, pp. 176–185. Cited on pp. 23 and 29.

Bonacich, P. . 1972. *Factoring and weighting approaches to status scores and clique identification*. In Journal of Mathematical Sociology, vol. 2, no. 1, pp. 113–120. Cited on p. 20.

Borg, I. . 2005. *Modern multidimensional scaling: Theory and applications*. Springer. Cited on p. 81.

Borgatti, S. P. . 2005. *Centrality and network flow*. In Social networks, vol. 27, no. 1, pp. 55–71. Cited on p. 20.

———. 2012. *Two-Mode Concepts in Social Network Analysis*. In Meyers, R. A. , ed., Computational Complexity - Theory, Techniques, and Applications, pp. 2912–2924. Springer. Cited on pp. 27, 29, and 80.

Borgatti, S. P. and Everett, M. G. . 1997. *Network analysis of 2-mode data*. In Social networks, vol. 19, no. 3, pp. 243–269. Cited on pp. 23, 25, 27, 28, 29, and 135.

Borgatti, S. P. ; Everett, M. G. ; and Shirey, P. R. . 1990. *LS sets, lambda sets and other cohesive subsets*. In Social Networks, vol. 12, no. 4, pp. 337–357. Cited on p. 20.

Borgatti, S. P. ; Mehra, A. ; Brass, D. J. ; and Labianca, G. . 2009. *Network Analysis in the Social Sciences*. In Science, vol. 323, no. 5916, pp. 892–895. Cited on pp. 21 and 28.

Bostock, M. ; Ogievetsky, V. ; and Heer, J. . 2011. *D^3 Data-Driven Documents*. In Visualization and Computer Graphics, IEEE Transactions on, vol. 17, no. 12, pp. 2301–2309. Cited on p. 84.

Botafogo, R. A. ; Rivlin, E. ; and Shneiderman, B. . 1992. *Structural analysis of hypertexts: identifying hierarchies and useful metrics*. In ACM Transactions on Information Systems (TOIS), vol. 10, no. 2, pp. 142–180. Cited on p. 21.

Boudourides, M. A. and Botetzagias, I. A. . 2007. *7 Networks of protest on global issues in Greece 2002–2003*. In Civil Societies and Social Movements, p. 109. Cited on p. 23.

Brehmer, M. and Munzner, T. . 2013. *A multi-level taxonomy of abstract visualization tasks*. In Proceedings of the 2013 InfoVis Conference, Transactions in Visualization and Computer Graphics, vol. 19, no. 12, pp. 2376–2385. Cited on pp. vii, 69, 72, 73, 88, 99, 100, 101, and 103.

Breiger, R. L. . 1974. *The duality of persons and groups*. In Social forces, vol. 53, no. 2, pp. 181–190. Cited on p. 23.

Buja, A. ; Cook, D. ; and Swayne, D. F. . 1996. *Interactive high-dimensional data visualization*. In Journal of Computational and Graphical Statistics, vol. 5, no. 1, pp. 78–99. Cited on pp. 70 and 88.

Burt, R. and Schøtt, T. . 1985. *Relation content in multiple networks*. In Social Science Research, vol. 14, pp. 287–308. Cited on pp. v, 3, 24, 30, 36, 58, and 141.

Burt, R. S. . 1992. *Structural holes: The social structure of competition*. Harvard University Press. Cited on p. 18.

———. 2004. *Structural holes and good ideas*. In American journal of sociology, vol. 110, no. 2, pp. 349–399. Cited on p. 18.

Caldarelli, G. ; Battiston, S. ; Garlaschelli, D. ; and Catanzaro, M. . 2004. *Emergence of complexity in financial networks*. In Complex Networks, pp. 399–423. Springer. Cited on p. 23.

Callaway, D. S. ; Newman, M. E. ; Strogatz, S. H. ; and Watts, D. J. . 2000. *Network robustness and fragility: Percolation on random graphs*. In Physical review letters, vol. 85, no. 25, pp. 5468–5471. Cited on p. 21.

i Cancho, R. F. and Solé, R. V. . 2001. *The small world of human language*. In Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 268, no. 1482, pp. 2261–2265. Cited on p. 23.

Cao, N. ; Sun, J. ; Lin, Y. ; Gotz, D. ; Liu, S. ; and Qu, H. . 2010. *FacetAtlas: Multifaceted visualization for rich text corpora*. In IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1172–1181. Cited on p. 81.

Card, S. K. ; Mackinlay, J. D. ; and Schneiderman, B. . 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann. Cited on pp. 65 and 73.

Cetinkaya, C. and Knightly, E. W. . 2004. *Opportunistic traffic scheduling over multiple network paths*. In INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, vol. 3, pp. 1928–1937. IEEE. Cited on p. 24.

Chen, C. . 2006. *CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature*. In Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 359–377. Cited on p. 81.

Chen, C.-h. ; Härdle, W. ; HaĘrdle, W. ; and Unwin, A. . 2007. *Handbook of data visualization*. Springer. Cited on p. 83.

Chi, E. H.-h. and Riedl, J. T. . 1998. *An operator interaction framework for visualization systems*. In Proceedings IEEE Symposium on Information Visualization '98, pp. 63–70. IEEE. Cited on p. 70.

Chuah, M. C. and Roth, S. F. . 1996. *On the semantics of interactive visualizations*. In Proceedings IEEE Symposium on Information Visualization '96, pp. 29–36. IEEE. Cited on p. 70.

Chuang, J. ; Ramage, D. ; Manning, C. ; and Heer, J. . 2012. *Interpretation and trust: designing model-driven visualizations for text analysis*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 443–452. ACM. Cited on pp. 13, 63, and 84.

Claussen, J. C. and Wilhelm, T. . 2008. *Network complexity: Introduction to the Session at NetSci 2008*. International Workshop and Conference on Network Science 2008. Cited on p. 9.

Conyon, M. J. and Muldoon, M. R. . 2004. *The small world network structure of boards of directors*. Tech. rep., Manchester Institute for Mathematical Sciences School of Mathematics, The University of Manchester. Cited on p. 23.

Costa, L. d. F. ; Rodrigues, F. A. ; Travieso, G. ; and Villas Boas, P. . 2007. *Characterization of complex networks: A survey of measurements*. In Advances in Physics, vol. 56, no. 1, pp. 167–242. Cited on p. 9.

Cover, T. and Hart, P. . 1967. *Nearest neighbor pattern classification*. In IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27. Cited on p. 14.

Crampes, M. and Plantié, M. . 2013. *A Unified Community Detection, Visualization and Analysis method*. In CoRR, arXiv preprint arXiv:1301.7006. Cited on p. 28.

Crystal, A. and Ellington, B. . 2004. *Task analysis and human-computer interaction: approaches, techniques, and levels of analysis.* In Proceedings of the Tenth Americas Conference on Information Systems, p. 391. Citeseer. Cited on p. 71.

Cui, W. ; Qu, H. ; Zhou, H. ; Zhang, W. ; and Skiena, S. . 2012. *Watch the story unfold with textwheel: Visualization of large-scale news streams*. In ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 2, p. 20. Cited on p. 81.

Dahui, W. ; Li, Z. ; and Zengru, D. . 2006. *Bipartite producer–consumer networks and the size distribution of firms*. In Physica A: Statistical Mechanics and its Applications, vol. 363, no. 2, pp. 359–366. Cited on p. 23.

Davis, A. ; Gardner, B. B. ; and Gardner, M. R. . 1941. *Deep South: A social anthropological study of caste and class*. U of South Carolina Press. Cited on p. 135.

De Domenico, M. ; Solè-Ribalta, A. ; Cozzo, E. ; Kivelä, M. ; Moreno, Y. ; Porter, M. A. ; Gòmez, S. ; and Arenas, A. . 2013. *Mathematical Formulation of Multi-Layer Networks*. In arXiv preprint arXiv:1307.4977 physics.soc-ph. Cited on pp. 27 and 28.

De Leeuw, J. and Michailidis, G. . 2000. *Graph layout techniques and multidimensional data analysis*. In IMS Lecture Notes - Monograph Series, pp. 219–248. Cited on p. 80.

De Neys, W. ; Rossi, S. ; and Houdé, O. . 2013. *Bats, balls, and substitution sensitivity: cognitive misers are no happy fools*. In Psychonomic bulletin & review, vol. 20, no. 2, pp. 269–273. Cited on p. 145.

Deerwester, S. ; Dumais, S. ; Furnas, G. W. ; Landauer, T. K. ; and Harshman, R. . 1990. *Indexing by Latent Semantic Analysis*. In Journal of the Society for Information Science, vol. 41, no. 6, pp. 391–407. Cited on p. 15.

Dekker, A. H. and Colbert, B. D. . 2004. *Network robustness and graph topology*. In Proceedings of the 27th Australasian conference on Computer science, vol. 26, pp. 359–368. Cited on p. 21.

Demmel, J. ; Dumitriu, I. ; and Holtz, O. . 2007. *Fast linear algebra is stable*. In Numerische Mathematik, vol. 108, no. 1, pp. 59–91. Cited on p. 40.

Dhillon, I. S. . 2001. *Co-clustering documents and words using bipartite spectral graph partitioning*. In Proceedings of the seventh ACM SIGKDD International conference on Knowledge Discovery and Data mining, pp. 269–274. ACM. Cited on pp. 23 and 28.

Di Giacomo, E. ; Didimo, W. ; Grilli, L. ; and Liotta, G. . 2007. *Graph Visualization Techniques for Web Clustering Engines*. In IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 2, pp. 294–304. Cited on p. 80.

Diderot, D. ; d'Alembert, J. L. R. ; et al. 1751-72. *Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers*. André le Breton, Michel-Antoine David, Laurent Durand, and Antoine-Claude Briasson. Cited on pp. 1 and 3.

Didimo, W. ; Liotta, G. ; and Romeo, S. A. . 2011. *A Graph Drawing Application to Web Site Traffic Analysis*. In Journal of Graph Algorithms and Applications, vol. 15, no. 2, pp. 229–251. Cited on pp. 80 and 81.

Ding, C. H. ; He, X. ; Zha, H. ; Gu, M. ; and Simon, H. D. . 2001. *A min-max cut algorithm for graph partitioning and data clustering*. In Proceedings IEEE International Conference on Data Mining, ICDM 2001, pp. 107–114. IEEE. Cited on p. 20.

Ding, J. and Zhou, A. . 2009. *Nonnegative Matrices, Positive Operators and Applications*. World Scientific, Singapore. Cited on pp. 39, 41, and 42.

Dix, A. and Ellis, G. . 1998. *Starting simple: adding value to static visualisation through simple interaction*. In Proceedings of the working

conference on Advanced Visual Interfaces, AVI '98, pp. 124–134. ACM. Cited on p. 70.

Dogrusoz, U. and Genç, B. . 2006. *A multi-graph approach to complexity management in interactive graph visualization*. In Computers & Graphics, vol. 30, no. 1, pp. 86–97. Cited on p. 80.

Dolinski, K. ; Chatr-Aryamontri, A. ; and Tyers, M. . 2013. *Systematic curation of protein and genetic interaction data for computable biology*. In BioMedCentral (BMC) Biology, vol. 11, no. 1, p. 43. Cited on p. 23.

Dörk, M. ; Carpendale, S. ; and Williamson, C. . 2011. *The information flaneur: a fresh look at information seeking*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1215–1224. ACM. Cited on p. 62.

Dork, M. ; Henry Riche, N. ; Ramos, G. ; and Dumais, S. . 2012. *PivotPaths: Strolling through Faceted Information Spaces*. In IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pp. 2709–2718. Cited on p. 84.

Douglas, S. M. ; Montelione, G. T. ; and Gerstein, M. . 2005. *PubNet: a flexible system for visualizing literature derived networks*. In Genome biology, vol. 6, no. 9, p. R80. Cited on p. 80.

Dow, M. M. . 2007. *Galton's Problem as Multiple Network Autocorrelation Effects Cultural Trait Transmission and Ecological Constraint*. In Cross-Cultural Research, vol. 41, no. 4, pp. 336–363. Cited on p. 24.

Dow, M. M. ; Burton, M. L. ; White, D. R. ; and Reitz, K. P. . 1984. *Galton's problem as network autocorrelation*. In American Ethnologist, vol. 11, no. 4, pp. 754–770. Cited on p. 17.

Dubin, D. . 2004. *The most influential paper Gerard Salton never wrote*. In Library trends, vol. 52, no. 4, pp. 748–764. Cited on p. 14.

Dugundji, J. and Granas, A. . 1982. *Fixed point theory*, vol. 1. PWN-Polish Scientific Publishers. Cited on p. 37.

Dumas, M. ; McGuffin, M. J. ; Robert, J.-M. ; and Willig, M.-C. . 2012a. *Optimizing a radial layout of bipartite graphs for a tool visualizing security alerts*. In Graph Drawing, Lecture Notes in Computer Science, vol. 7034, pp. 203–214. Springer. Cited on p. 79.

Dumas, M. ; Robert, J.-M. ; and McGuffin, M. J. . 2012b. *Alertwheel: radial bipartite graph visualization applied to intrusion detection system alerts*. In Network, IEEE, vol. 26, no. 6, pp. 12–18. Cited on p. 79.

Dunne, C. ; Henry Riche, N. ; Lee, B. ; Metoyer, R. ; and Robertson, G. . 2012. *GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1663–1672. ACM. Cited on p. 81.

Duquenne, V. . 1999. *Latticial structures in data analysis*. In Theoretical Computer Science, vol. 217, no. 2, pp. 407–436. Cited on p. 79.

Eades, P. and Wormald, N. . 1994. *Edge crossings in drawings of bipartite graphs*. In Algorithmica, vol. 11, no. 4, pp. 379–403. Cited on p. 79.

Eick, S. G. . 2000. *Visual discovery and analysis*. In IEEE Transactions on Visualization and Computer Graphics, vol. 6, no. 1, pp. 44–58. Cited on p. 81.

Einstein, A. ; Podolsky, B. ; and Rosen, N. . 1935. *Can quantum-mechanical description of physical reality be considered complete?* In Physical review, vol. 47, no. 10, p. 777. Cited on p. 37.

Elmqvist, N. ; Dragicevic, P. ; and Fekete, J.-D. . 2008. *Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation*. In IEEE Transactions on Visualization and Computer Graphics, vol. 14, no. 6, pp. 1539–1148. Cited on pp. 81 and 83.

Emerson, J. W. ; Green, W. A. ; Schloerke, B. ; Crowley, J. ; Cook, D. ; Hofmann, H. ; and Wickham, H. . 2013. *The Generalized Pairs Plot*. In Journal of Computational and Graphical Statistics, vol. 22, no. 1, pp. 79–91. Cited on p. 84.

Erdős, P. and Rényi, A. . 1959. *On random graphs*. In Publicationes Mathematicae Debrecen, vol. 6, pp. 290–297. Cited on p. 7.

Euler, L. . 1735. *The seven bridges of Königsberg*. Wm. Benton (1956). Cited on pp. 3 and 5.

Everett, M. G. and Borgatti, S. P. . 1998. *Analyzing clique overlap*. In Connections, vol. 21, no. 1, pp. 49–61. Cited on p. 25.

Faust, K. . 1997. *Centrality in affiliation networks*. In Social networks, vol. 19, no. 2, pp. 157–191. Cited on p. 28.

Faust, K. ; Willert, K. E. ; Rowlee, D. D. ; and Skvoretz, J. . 2002. *Scaling and statistical models for affiliation networks: patterns of participation among Soviet politicians during the Brezhnev era*. In Social Networks, vol. 24, no. 3, pp. 231–259. Cited on p. 23.

Fayyad, U. M. ; Wierse, A. ; and Grinstein, G. G. . 2002. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann. Cited on p. 81.

Fekete, J.-D. ; Van Wijk, J. J. ; Stasko, J. T. ; and North, C. . 2008. *The value of information visualization*. In Information Visualization, pp. 1–18. Springer. Cited on p. 72.

Few, S. . 2006. *The Surest Path to Visual Discovery*. Perceptual Edge. Cited on p. 61.

Fischer, D. H. . 1994. *Paul Revere's Ride.* Oxford University Press. Cited on p. 135.

Flanders, V. and Willis, M. . 1998. *Web pages that suck: Learn good design by looking at bad design*. SYBEX Inc. Cited on p. 63.

Fortunato, S. . 2010. *Community detection in graphs*. In Physics Reports, vol. 486, no. 3, pp. 75–174. Cited on p. 18.

Fortunato, S. and Barthelemy, M. . 2007. *Resolution limit in community detection*. In Proceedings of the National Academy of Sciences, vol. 104, no. 1, pp. 36–41. Cited on p. 19.

Freeland, G. and Corones, A. . 1999. *1543 and All that: Image and Word, Change and Continuity in the Proto-scientific Revolution*. Kluwer Academic Pub. Cited on p. 74.

Freeman, L. C. . 1977. *A set of measures of centrality based on betweenness*. In Sociometry, vol. 40, no. 1, pp. 35–41. Cited on pp. 8 and 20.

———. 1979. *Centrality in social networks conceptual clarification*. In Social networks, vol. 1, no. 3, pp. 215–239. Cited on p. 20.

Frick, A. ; Ludwig, A. ; and Mehldau, H. . 1995. *A fast adaptive layout algorithm for undirected graphs (extended abstract and system demonstration)*. In Graph Drawing, pp. 388–403. Springer. Cited on p. 79.

Friel, S. N. ; Curcio, F. R. ; and Bright, G. W. . 2001. *Making sense of graphs: Critical factors influencing comprehension and instructional implications*. In Journal for Research in mathematics Education, pp. 124–158. Cited on p. 62.

Frøkjær, E. ; Hertzum, M. ; and Hornbæk, K. . 2000. *Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?* In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 345–352. ACM. Cited on p. 63.

Fruchterman, T. M. and Reingold, E. M. . 1991. *Graph drawing by force-directed placement*. In Software: Practice and experience, vol. 21, no. 11, pp. 1129–1164. Cited on p. 79.

Fujimoto, K. ; Chou, C.-P. ; and Valente, T. W. . 2011. *The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter*. In Social networks, vol. 33, no. 3, pp. 231–243. Cited on pp. 29, 30, and 59.

Furnas, G. W. . 1986. *Generalized fisheye views*, vol. 17. ACM. Cited on pp. 81 and 83.

Committee on Network Science for Future Army Applications, N. R. C. . 2005. *Network Science*. The National Academies Press. isbn 9780309100267. URL http://www.nap.edu/openbook.php?record_id=11516. Cited on p. 9.

Gallo, G. ; Longo, G. ; Pallottino, S. ; and Nguyen, S. . 1993. *Directed hypergraphs and applications*. In Discrete applied mathematics, vol. 42, no. 2, pp. 177–201. Cited on p. 24.

Ganter, B. ; Wille, R. ; and Franzke, C. . 1997. *Formal concept analysis: mathematical foundations*. Springer-Verlag New York, Inc. Cited on p. 79.

Gattis, M. . 2003. *Spatial schemas and abstract thought*. The MIT Press. Cited on p. 2.

Gaume, B. ; Navarro, E. ; and Prade, H. . 2013. *Clustering bipartite graphs in terms of approximate formal concepts and sub-contexts*. In International Journal of Computational Intelligence Systems, vol. 6, no. 6, pp. 1125–1142. Cited on p. 28.

Ghoniem, M. ; Fekete, J.-D. ; and Castagliola, P. . 2004. *A comparison of the readability of graphs using node-link and matrix-based representations*. In 2004 IEEE Symposium on Information Visualization, pp. 17–24. IEEE. Cited on p. 76.

———. 2005. *On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis*. In Information Visualization, vol. 4, no. 2, pp. 114–135. Cited on pp. 9 and 76.

Giacomo, E. D. ; Didimo, W. ; Liotta, G. ; and Palladino, P. . 2010. *Visual Analysis of One-To-Many Matched Graphs*. In Journal of Graph Algorithms and Applications, vol. 14, no. 1, pp. 97–119. Cited on p. 80.

Gkantsidis, C. ; Mihail, M. ; and Zegura, E. . 2003. *Spectral analysis of Internet topologies*. In INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies, vol. 1, pp. 364–374. IEEE. Cited on p. 6.

Goffman, C. . 1969. *And what is your Erdös number?* In The American Mathematical Monthly, vol. 76, no. 7, pp. 791–791. Cited on p. 18.

Goldenberg, A. ; Zheng, A. X. ; Fienberg, S. E. ; and Airoldi, E. M. . 2010. *A survey of statistical network models*. In Foundations and Trends® in Machine Learning, vol. 2, no. 2, pp. 129–233. Cited on p. 8.

Goldstein, E. B. . 2008. *The Blackwell Handbook of Sensation and Perception*. Wiley. com. Cited on p. 73.

Goldstein, M. L. ; Morris, S. A. ; and Yen, G. G. . 2005. *Group-based Yule model for bipartite author-paper networks*. In Physical Review E, vol. 71, no. 2, p. 026108. Cited on p. 23.

Gómez, S. ; Díaz-Guilera, A. ; Gómez-Gardeñes, J. ; Perez-Vicente, C. J. ; Moreno, Y. ; and Arenas, A. . 2013. *Diffusion dynamics on multiplex networks*. In Physical review letters, vol. 110, no. 2, p. 028701. Cited on p. 28.

Good, B. H. ; de Montjoye, Y.-A. ; and Clauset, A. . 2010. *Performance of modularity maximization in practical contexts*. In Physical Review E, vol. 81, no. 4, p. 046106. Cited on p. 19.

Goody, J. . 1977. *The domestication of the savage mind*. Cambridge University Press. Cited on p. 146.

Görg, C. ; Kihm, J. ; Choo, J. ; Liu, Z. ; Muthiah, S. ; Park, H. ; and Stasko, J. . 2013. *Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw*. In IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 10, pp. 1646–1663. Cited on p. 81.

Gorg, C. ; Liu, Z. ; Kihm, J. ; Choo, J. ; Park, H. ; and Stasko, J. . 2013. *Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw*. In IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 10, pp. 1646–1663. Cited on pp. 81 and 82.

Gotz, D. and Zhou, M. X. . 2009. *Characterizing users' visual analytic activity for insight provenance*. In IEEE Symposium on Visual Analytics Science and Technology '08., vol. 8, no. 1, pp. 123–130. Cited on p. 70.

Gould, R. V. . 1991. *Multiple networks and mobilization in the Paris commune, 1871*. In American Sociological Review, vol. 56, no. 6, pp. 716–729. Cited on p. 24.

Granovetter, M. S. . 1973. *The strength of weak ties*. In American journal of sociology, pp. 1360–1380. Cited on p. 21.

Gray, J. and Szalay, E. . 2007. *eScience–A transformed scientific method*. Presentation made to the NRC-CSTB. Cited on p. 1.

Griffiths, T. ; Jordan, M. ; Tenenbaum, J. ; and Blei, D. M. . 2004. *Hierarchical topic models and the nested Chinese restaurant process*. In Advances in neural information processing systems, vol. 16, pp. 106–114. Cited on p. 16.

Grinstein, G. ; Trutschl, M. ; and Cvek, U. . 2001. *High-dimensional visualizations*. In Proceedings of Workshop on Visual Data Mining, ACM Conference on Knowledge Discovery and Data Mining, pp. 1–14. Cited on p. 81.

Guillaume, J.-L. and Latapy, M. . 2005. *Bipartite Graphs as Models of Complex Networks*, vol. 3405 of *Lecture Notes in Computer Science*, pp. 127–139. Springer. Cited on pp. 23, 25, and 59.

Guillaume, J.-L. ; Latapy, M. ; and Le-Blond, S. . 2004. *Statistical analysis of a p2p query graph based on degrees and their time-evolution*. In Distributed Computing - IWDC 2004, Lecture Notes in Computer Science, vol. 3326, pp. 126–137. Springer. Cited on p. 23.

Guimera, R. ; Mossa, S. ; Turtschi, A. ; and Amaral, L. A. N. . 2005. *The worldwide air transportation network: anomalous centrality, community structure, and cities global roles*. In Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 22, pp. 7794–7799. Cited on p. 42.

Hachul, S. and Jünger, M. . 2005. *Drawing large graphs with a potential-field-based multilevel algorithm*. In Graph Drawing, Lecture Notes in Computer Science, vol. 3383, pp. 285–295. Springer. Cited on pp. 17, 78, and 79.

Halu, A. ; Mondragon, R. J. ; Pansaraza, P. ; and Bianconi, G. . 2013. *Multiplex PageRank*. In arXiv preprint arXiv:1306.3576 Physics and Society. Cited on p. 29.

van Ham, F. and van Wijk, J. J. . 2004. *Interactive visualization of small world graphs*. In Proceeding of 2004 IEEE Symposium on Information Visualization, pp. 199–206. IEEE. Cited on pp. 20 and 95.

Han, S.-K. . 2009. *The Other Ride of Paul Revere: The Brokerage Role in the Making of the American Revolution*. In Mobilization: An International Quarterly, vol. 14, no. 2, pp. 143–162. Cited on p. 135.

Harris, Z. S. . 1954. *Distributional structure*. In Word, vol. 10, pp. 146–162. Cited on p. 14.

Harrower, M. and Brewer, C. A. . 2003. *Colorbrewer. org: an online tool for selecting colour schemes for maps*. In The Cartographic Journal, vol. 40, no. 1, pp. 27–37. Cited on pp. 74, 92, and 105.

Hastings, W. K. . 1970. *Monte Carlo sampling methods using Markov chains and their applications*. In Biometrika, vol. 57, no. 1, pp. 97–109. Cited on p. 38.

Hauser, H. ; Ledermann, F. ; and Doleisch, H. . 2002. *Angular brushing of extended parallel coordinates*. In IEEE Symposium on Information Visualization, 2002., pp. 127–130. IEEE. Cited on p. 83.

Healey, C. G. and Enns, J. T. . 2012. *Attention and Visual Memory in Visualization and Computer Graphics*. In IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 7, pp. 1170–1188. Cited on p. 2.

Healey, C. G. ; Booth, K. S. ; and Enns, J. T. . 1993. *Harnessing preattentive processes for multivariate data visualization*. In Proceedings of Graphics Interface, pp. 107–107. Canadian Information Processing Society. Cited on pp. 73 and 74.

Heer, J. ; Bostock, M. ; and Ogievetsky, V. . 2010. *A tour through the visualization zoo.* In Commun. ACM, vol. 53, no. 6, pp. 59–67. Cited on p. 74.

Heer, J. and Shneiderman, B. . 2012. *Interactive dynamics for visual analysis*. In Commun. ACM, vol. 55, no. 4, p. 30. Cited on p. 71.

Hein, M. ; Eisert, J. ; and Briegel, H. J. . 2004. *Multiparty entanglement in graph states*. In Physical Review A, vol. 69, no. 6, p. 062311. Cited on p. 37.

Henry, N. ; Fekete, J.-D. ; and McGuffin, M. J. . 2007. *NodeTrix: a hybrid visualization of social networks*. In IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp. 1302–1309. Cited on pp. 76 and 108.

Henzinger, M. R. . 2001. *Hyperlink analysis for the web*. In Internet Computing, IEEE, vol. 5, no. 1, pp. 45–50. Cited on p. 16.

Herman, I. ; Marshall, M. S. ; and Melançon, G. . 2000a. *Density functions for visual attributes and effective partitioning in graph visualization*. In Proceedings of the 2000 IEEE Symposium on Information Visualization, pp. 49–56. IEEE. Cited on p. 74.

Herman, I. ; Melançon, G. ; and Marshall, M. S. . 2000b. *Graph visualization and navigation in information visualization: A survey*. In IEEE Transactions on Visualization and Computer Graphics, vol. 6, no. 1, pp. 24–43. Cited on pp. 76, 77, and 81.

Hermanns, W. and Einstein, A. . 1983. *Einstein and the Poet: In Search of the Cosmic Man*. Branden Books. Cited on p. 147.

Herskovits, M. J. . 1964. *Cultural dynamics*. Alfred A. Knopf. Cited on p. 144.

Hervé, N. ; Viaud, M.-L. ; Thièvre, J. ; Saulnier, A. ; Champ, J. ; Letessier, P. ; Buisson, O. ; Joly, A. ; and Renoust, B. . 2013. *OTMedia: L'Observatoire TransMedia*. Tech. rep., INA, OTMedia, ANR-10-CORD-000. Cited on pp. 11 and 125.

Hey, A. J. ; Tansley, S. ; Tolle, K. M. ; et al. 2009. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research Redmond, WA. Cited on p. 1.

Holme, P. ; Min Park, S. ; Kim, B. J. ; and Edling, C. R. . 2007. *Korean university life in a network perspective: Dynamics of a large affiliation network*. In Physica A: Statistical Mechanics and its Applications, vol. 373, pp. 821–830. Cited on p. 23.

Holten, D. . 2006. *Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data*. In IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 5, pp. 741–748. Cited on p. 80.

Hossain, M. S. and Angryk, R. A. . 2007. *Gdclust: A graph-based document clustering technique*. In Seventh IEEE International Conference on Data Mining Workshops, ICDM Workshops 2007, pp. 417–422. IEEE. Cited on p. 16.

Hummon, N. P. and Dereian, P. . 1989. *Connectivity in a citation network: The development of DNA theory*. In Social Networks, vol. 11, no. 1, pp. 39–63. Cited on p. 16.

Hutchins, E. L. ; Hollan, J. D. ; and Norman, D. A. . 1985. *Direct manipulation interfaces*. In Human–Computer Interaction, vol. 1, no. 4, pp. 311–338. Cited on p. 62.

Huth, A. G. ; Nishimoto, S. ; Vu, A. T. ; and Gallant, J. L. . 2012. *A continuous semantic space describes the representation of thousands of object and action categories across the human brain*. In Neuron, vol. 76, no. 6, pp. 1210–1224. Cited on p. 2.

If, M. B. and Blake, R. . 1990. *Preattentive vision and perceptual groups*. In Perception, vol. 19, pp. 515–522. Cited on p. 73.

Ingwersen, P. . 1996. *Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory*. In Journal of documentation, vol. 52, no. 1, pp. 3–50. Cited on p. 14.

Isenberg, P. and Fisher, D. . 2009. *Collaborative Brushing and Linking for Co-located Visual Analytics of Document Collections*. In Computer Graphics Forum, vol. 28, no. 3, pp. 1031–1038. Cited on p. 83.

Ito, T. ; Misue, K. ; and Tanaka, J. . 2009. *Sphere Anchored Map: A Visualization Technique for Bipartite Graphs in 3D Human-Computer Interaction. Novel Interaction Methods and Techniques*, vol. 5611 of *Lecture Notes in Computer Science*, pp. 811–820. Springer Berlin / Heidelberg. Cited on p. 79.

———. 2010. *Drawing Clustered Bipartite Graphs in Multi-circular Style*. In Proceedings of the 2010 14th International Conference Information Visualisation, IV '10, pp. 23–28. IEEE Computer Society. Cited on p. 80.

Jaccard, P. . 1901. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz. Cited on p. 17.

Jackson, M. O. . 2010. *Social and Economic Networks*. Princeton University Press. Cited on p. 24.

Jain, A. K. . 2010. *Data clustering: 50 years beyond K-means*. In Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666. Cited on p. 13.

James, W. . 1907. *Pragmatism: A New Name for Some Old Ways of Thinking*. Courier Dover Publications. Cited on p. 145.

Joly, A. ; Champ, J. ; Letessier, P. ; Hervé, N. ; Buisson, O. ; and Viaud, M.-L. . 2012. *Visual-based transmedia events detection*. In Proceedings of the 20th ACM international conference on Multimedia, pp. 1351–1352. ACM. Cited on p. 11.

Jusufi, I. ; Kerren, A. ; and Zimmer, B. . 2013. *Multivariate Network Exploration with JauntyNets*. In 17th International Conference on Information Visualisation (IV '13), pp. 19–27. Cited on p. 81.

Kandel, S. ; Paepcke, A. ; Hellerstein, J. M. ; and Heer, J. . 2012. *Enterprise data analysis and visualization: An interview study*. In IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pp. 2917–2926. Cited on p. 71.

Kang, Y.-a. ; Gorg, C. ; and Stasko, J. . 2009. *Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study*. In Proceedings of 2009 IEEE Symposium on Visual Analytics Science and Technology., pp. 139–146. IEEE. Cited on p. 71.

Kaski, S. ; Nikkila, J. ; Oja, M. ; Venna, J. ; Toronen, P. ; and Castren, E. . 2003. *Trustworthiness and metrics in visualizing similarity of gene expression*. In BMC Bioinformatics, vol. 4, no. 1, p. 48. Cited on p. 63.

Ke, W. ; Börner, K. ; and Viswanath, L. . 2004. *Major Information Visualization Authors, Papers and Topics in the ACM Library*. In IEEE Symposium on Information Visualization, 2004 (InfoVis) Contest. IEEE Computer Society. Cited on pp. 129 and 130.

Keim, D. A. . 2002. *Information visualization and visual data mining*. In IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 1–8. Cited on pp. 62 and 81.

Keim, D. A. ; Mansmann, F. ; Schneidewind, J. ; Thomas, J. ; and Ziegler, H. . 2008. *Visual analytics: Scope and challenges*. Springer. Cited on pp. 66 and 67.

Kendall, M. G. . 1948. *Rank correlation methods.* New York, Hafner Pub. Co. Cited on p. 55.

Kim, J. and Wilhelm, T. . 2008. *What is a complex graph?* In Physica A: Statistical Mechanics and its Applications, vol. 387, no. 11, pp. 2637–2652. Cited on p. 9.

Klein, G. . 2007. *The power of intuition: How to use your gut feelings to make better decisions at work*. Random House Digital, Inc. Cited on p. 63.

Klein, G. ; Moon, B. ; and Hoffman, R. R. . 2006a. *Making sense of sensemaking 1: Alternative perspectives*. In Intelligent Systems, IEEE, vol. 21, no. 4, pp. 70–73. Cited on p. 63.

———. 2006b. *Making sense of sensemaking 2: A macrocognitive model*. In Intelligent Systems, IEEE, vol. 21, no. 5, pp. 88–92. Cited on pp. vii and 64.

Klein, M. and König-Ries, B. . 2002. *Multi-layer clusters in ad-hoc networks – an approach to service discovery*. In Proceedings Of The First International Workshop On Peer-To-Peer Computing, pp. 187–201. Springer. Cited on p. 28.

Kleinberg, J. M. . 1999. *Authoritative sources in a hyperlinked environment*. In Journal of the ACM (JACM), vol. 46, no. 5, pp. 604–632. Cited on p. 21.

de Klerk, E. ; Pasechnik, D. V. ; and Salazar, G. . 2012. *Book drawings of complete bipartite graphs*. In CORD Conference Proceedings, pp. 1–22. Cited on p. 79.

Koffka, K. . 1935. *Principles of Gestalt psychology*. Harcourt, Brace New York. Cited on p. 73.

Kogut, B. . 2000. *The network as knowledge: generative rules and the emergence of structure*. In Strategic management journal, vol. 21, no. 3, pp. 405–425. Cited on p. 17.

Kohonen, T. . 2001. *Self-organizing maps*, vol. 30. Springer. Cited on p. 81.

Konyha, Z. ; Matkovic, K. ; Gracanin, D. ; Jelovic, M. ; and Hauser, H. . 2006. *Interactive visual analysis of families of function graphs*. In IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 6, pp. 1373–1385. Cited on p. 83.

Kosara, R. . 2011. *Indirect multi-touch interaction for brushing in parallel coordinates*. In Proceedings of Society of Photo-Optical Instrumentation Engineer, vol. 7868, p. 09. the International Society for Optical Engineering. Cited on p. 83.

Kurant, M. and Thiran, P. . 2005. *On survivable routing of mesh topologies in IP-over-WDM networks*. In Proceedings IEEE INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, pp. 1106–1116. IEEE. Cited on p. 24.

———. 2006. *Layered complex networks*. In Physical review letters, vol. 96, no. 13, p. 138701. Cited on p. 24.

Lambert, A. ; Bourqui, R. ; and Auber, D. . 2010. *3D edge bundling for geographical data visualization*. In 14th International Conference Information Visualisation (IV), pp. 329–335. IEEE. Cited on p. 80.

Lambiotte, R. and Ausloos, M. . 2006. *Collaborative tagging as a tripartite network*. In Proceedings of the 6th international conference on

Computational Science - Volume Part III, pp. 1114–1117. Springer-Verlag. Cited on p. 23.

Laney, D. . 2001. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep., Application Delivery Strategies, published by META Group Inc. Cited on p. 2.

Latapy, M. ; Magnien, C. ; and Ouédraogo, F. . 2008a. *A Radar for the Internet*. In Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on, pp. 901–908. IEEE. Cited on p. 22.

Latapy, M. ; Magnien, C. ; and Vecchio, N. D. . 2008b. *Basic notions for the analysis of large two-mode networks*. In Social Networks, vol. 30, no. 1, pp. 31–48. Cited on pp. 25, 30, and 59.

Lee, B. ; Plaisant, C. ; Parr, C. S. ; Fekete, J.-D. ; and Henry, N. . 2006. *Task taxonomy for graph visualization*. In Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization, pp. 1–5. ACM. Cited on p. 17.

Leenders, R. T. A. . 2002. *Modeling social influence through network autocorrelation: constructing the weight matrix*. In Social Networks, vol. 24, no. 1, pp. 21–47. Cited on p. 17.

Leonel, A. ; Ribeiro, C. H. ; and Brust, M. R. . 2013. *Weak Ties in Complex Wireless Communication Networks*. In Complex Networks, Studies in Computational Intelligence, vol. 424, pp. 49–56. Springer. Cited on p. 23.

Leskovec, J. ; Lang, K. J. ; and Mahoney, M. . 2010. *Empirical comparison of algorithms for network community detection*. In Proceedings of the 19th international conference on World wide web, pp. 631–640. ACM. Cited on p. 20.

Lespinats, S. and Aupetit, M. . 2009. *False neighbourhoods and tears are the main mapping defaults. How to avoid it? How to exhibit remaining ones?* In Quality issues, measures of interestingness and evaluation of data mining models, QIMIE/PAKDD 2009, pp. 55–65. Cited on p. 63.

———. 2011. *CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings*. In Computer Graphics Forum, vol. 30, no. 1, pp. 113–125. Cited on p. 63.

Letessier, P. ; Buisson, O. ; and Joly, A. . 2011. *Consistent visual words mining with adaptive sampling*. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 49. ACM. Cited on p. 126.

Levi, F. W. . 1942. *Finite geometrical systems*. University of Calcutta. Cited on p. 24.

Lind, P. G. ; Gonzalez, M. C. ; and Herrmann, H. J. . 2005. *Cycles and clustering in bipartite networks*. In Physical review E, vol. 72, no. 5, p. 056127. Cited on pp. 23 and 27.

Liu, Z. and Stasko, J. T. . 2010. *Mental models, visual reasoning and interaction in information visualization: A top-down perspective*. In IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 999–1008. Cited on p. 71.

Lu, C.-Y. ; Zhou, X.-Q. ; Gühne, O. ; Gao, W.-B. ; Zhang, J. ; Yuan, Z.-S. ; Goebel, A. ; Yang, T. ; and Pan, J.-W. . 2007. *Experimental entanglement of six photons in graph states*. In Nature Physics, vol. 3, no. 2, pp. 91–95. Cited on p. 37.

Maier, D. . 1980. *Minimum covers in relational database model*. In Journal of the ACM (JACM), vol. 27, no. 4, pp. 664–674. Cited on p. 24.

Maimon, O. Z. and Rokach, L. . 2005. *Data mining and knowledge discovery handbook*. Springer. Cited on p. 16.

Makinen, E. and Sieranta, M. . 1994. *Genetic algorithms for drawing bipartite graphs*. In International Journal of Computer Mathematics, vol. 53, no. 3-4, pp. 157–166. Cited on p. 79.

Mancoridis, S. ; Mitchell, B. S. ; Rorres, C. ; Chen, Y. ; and Gansner, E. R. . 1998. *Using automatic clustering to produce high-level system organizations of source code*. In IWPC'98 Proceedings of the 6th International Workshop on Program Comprehension, pp. 45–52. IEEE. Cited on p. 20.

Manning, C. D. ; Raghavan, P. ; and Schütze, H. . 2008. *Introduction to information retrieval*, vol. 1. Cambridge University Press Cambridge. Cited on p. 16.

Manski, C. F. . 1993. *Identification of Endogenous Social Effects: The Reflection Problem*. In The Review of Economic Studies, vol. 60, no. 3, pp. 531–542. Cited on p. 22.

Marchionini, G. . 2006. *Exploratory search: from finding to understanding*. In Communications of the ACM, vol. 49, no. 4, pp. 41–46. Cited on p. 61.

Marshack, A. . 1991. *The Tai Plaque and calendrical notation in the Upper Palaeolithic*. In Cambridge Archaeological Journal, vol. 1, no. 1, pp. 25–61. Cited on p. 2.

Martin, A. R. and Ward, M. O. . 1995. *High dimensional brushing for interactive exploration of multivariate data*. In Proceedings of the 6th Conference on Visualization '95, p. 271. IEEE Computer Society. Cited on p. 83.

McGraw, P. N. and Menzinger, M. . 2008. *Laplacian spectra as a diagnostic tool for network structure and dynamics*. In Physical Review E, vol. 77, no. 3, p. 031102. Cited on p. 6.

McGuffin, M. J. . 2012. *Simple algorithms for network visualization: A tutorial*. In Tsinghua Science and Technology, vol. 17, no. 4, pp. 383–398. Cited on p. 76.

McGuffin, M. J. and Jurisica, I. . 2009. *Interaction Techniques for Selecting and Manipulating Subgraphs in Network Visualizations*. In IEEE Transactions on Visualization and Computer Graphics, vol. 15, no. 6, pp. 937–944. Cited on pp. 83 and 96.

McLachlan, G. and Krishnan, T. . 2007. *The EM algorithm and extensions*, vol. 382. John Wiley & Sons. Cited on pp. 15 and 16.

McPherson, J. M. . 1982. *Hypernetwork sampling: Duality and differentiation among voluntary organizations*. In Social Networks, vol. 3, no. 4, pp. 225–249. Cited on p. 23.

McPherson, M. ; Smith-Lovin, L. ; and Cook, J. M. . 2001. *Birds of a feather: Homophily in social networks*. In Annual review of sociology, vol. 27, pp. 415–444. Cited on p. 24.

Melançon, G. ; Renoust, B. ; and Viaud, M.-L. . 2012. *Mesurer la cohésion sémantique dans les corpus de documents*. Tech. rep., INRIA Bordeaux Sud-Oues, HAL, Technical Report RR 8075. Cited on pp. 4 and 138.

Melançon, G. and Sallaberry, A. . 2008. *Edge Metrics for Visual Graph Analytics: A Comparative Study*. In 12th International Conference on Information Visualisation, pp. 610–615. Cited on p. 20.

Meyer, M. ; Sedlmair, M. ; and Munzner, T. . 2012. *The four-level nested model revisited: blocks and guidelines*. In Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization. ACM. Cited on pp. 68, 69, and 72.

Misue, K. . 2006. *Drawing bipartite graphs as anchored maps*. In Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation, vol. 60, pp. 169–177. Australian Computer Society, Inc. Cited on p. 79.

Misue, K. ; Eades, P. ; Lai, W. ; and Sugiyama, K. . 1995. *Layout adjustment and the mental map*. In Journal of visual languages and computing, vol. 6, no. 2, pp. 183–210. Cited on p. 76.

Moreno, J. L. . 1934. *Who shall survive?: A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Co. Cited on pp. 76 and 78.

Moscovich, T. ; Chevalier, F. ; Henry, N. ; Pietriga, E. ; and Fekete, J.-D. . 2009. *Topology-aware navigation in large networks*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2319–2328. ACM. Cited on p. 83.

Mucha, P. J. ; Richardson, T. ; Macon, K. ; Porter, M. A. ; and Onnela, J.-P. . 2010. *Community structure in time-dependent, multiscale, and multiplex networks*. In Science, vol. 328, no. 5980, pp. 876–878. Cited on p. 28.

Munzner, T. . 2009. *A Nested Model for Visualization Design and Validation*. In IEEE Transactions on Visualization and Computer Graphics, vol. 15, no. 6, pp. 921–928. Cited on pp. vi, 67, 68, 72, 88, 90, and 97.

Munzner, T. ; Guimbretière, F. ; Tasiran, S. ; Zhang, L. ; and Zhou, Y. . 2003. *TreeJuxtaposer: scalable tree comparison using Focus+ Context with guaranteed visibility*. In ACM Transactions on Graphics (TOG), vol. 22, no. 3, pp. 453–462. Cited on p. 84.

Nanard, J. and Nanard, M. . 1995. *Hypertext design environments and the hypertext design process*. In Communications of the ACM, vol. 38, no. 8, pp. 49–56. Cited on p. 61.

Neal, Z. . 2013. *Identifying statistically significant edges in one-mode projections*. In Social Network Analysis and Mining, pp. 1–10. Cited on pp. 25 and 135.

Nemeth, R. J. and Smith, D. A. . 1985. *International trade and world-system structure: a multiple network analysis*. In Review (Fernand Braudel Center), vol. 8, no. 4, pp. 517–560. Cited on p. 24.

Neumann, J. v. . 1928. *Zur theorie der gesellschaftsspiele*. In Mathematische Annalen, vol. 100, no. 1, pp. 295–320. Cited on p. 37.

Newman, M. E. J. . 2003. *The structure and function of complex networks*. In SIAM Review, vol. 45, pp. 167–256. Cited on p. 24.

———. 2001a. *Scientific collaboration networks. I. Network construction and fundamental results*. In Physical review E, vol. 64, no. 1, p. 016131. Cited on p. 23.

———. 2001b. *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality*. In Physical review E, vol. 64, no. 1, p. 016132. Cited on p. 23.

———. 2004. *Coauthorship networks and patterns of scientific collaboration*. In Proceedings of the National Academy of Sciences, vol. 101, pp. 5200–5205. Cited on pp. 16 and 23.

Newman, M. E. . 2006. *Modularity and community structure in networks*. In Proceedings of the National Academy of Sciences, vol. 103, no. 23, pp. 8577–8582. Cited on p. 19.

Newman, M. E. and Girvan, M. . 2004. *Finding and evaluating community structure in networks*. In Physical review E, vol. 69, no. 2, p. 026113. Cited on p. 19.

Newman, M. E. ; Strogatz, S. H. ; and Watts, D. J. . 2001. *Random graphs with arbitrary degree distributions and their applications*. In Physical Review E, vol. 64, no. 2, p. 026118. Cited on pp. 23 and 27.

Ng, A. Y. ; Jordan, M. I. ; and Weiss, Y. . 2002. *On spectral clustering: Analysis and an algorithm*. In Advances in neural information processing systems, vol. 2, pp. 849–856. Cited on p. 20.

Nguyen, S. and Pallottino, S. . 1988. *Equilibrium traffic assignment for large scale transit networks*. In European journal of operational research, vol. 37, no. 2, pp. 176–186. Cited on p. 24.

———. 1989. *Hyperpaths and shortest hyperpaths*. In Combinatorial Optimization, Lecture Notes in Mathematics, vol. 1403, pp. 258–271. Springer. Cited on p. 24.

Nick, B. ; Lee, C. ; Cunningham, P. ; and Brandes, U. . 2013. *Simmelian Backbones: Amplifying Hidden Homophily in Facebook Networks*. In 2013 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 525–532. IEEE. Cited on p. 22.

Nielsen, J. . 1994. *Usability inspection methods*. In Conference Companion on Human Factors in Computing Systems, CHI '94, pp. 413–414. ACM. Cited on p. 72.

Noack, A. . 2006. *Energy-based clustering of graphs with nonuniform degrees*. In Graph Drawing, Lecture Notes in Computer Science, vol. 3843, pp. 309–320. Springer. Cited on pp. 78, 79, and 99.

———. 2003. *An Energy Model for Visual Graph Clustering*. In 11th International Symposium on Graph Drawing (GD 2003), vol. 2912 of *Lecture Notes in Computer Science*, pp. 425–436. Cited on p. 20.

Norman, D. A. . 1988. *The design of everyday things*. Basic books. Cited on pp. 62 and 69.

———. 1993. *Things that make us smart: Defending human attributes in the age of the machine*. Basic Books. Cited on p. 72.

Oelke, D. ; Bak, P. ; Keim, D. ; Last, M. ; and Danon, G. . 2008. *Visual evaluation of text features fo document summarization and analysis*. In IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST'08, pp. 75–82. IEEE. Cited on p. 81.

Olson, J. A. ; Amlani, A. A. ; and Rensink, R. A. . 2012. *Perceptual and cognitive characteristics of common playing cards*. In Perception, vol. 41, no. 3, p. 268. Cited on p. 2.

Onody, R. N. and de Castro, P. A. . 2004. *Complex network study of Brazilian soccer players*. In Physical Review E, vol. 70, no. 3, p. 037103. Cited on p. 23.

Paas, F. ; Renkl, A. ; and Sweller, J. . 2003. *Cognitive load theory and instructional design: Recent developments*. In Educational psychologist, vol. 38, no. 1, pp. 1–4. Cited on p. 63.

Pagani, G. A. and Aiello, M. . 2013. *The power grid as a complex network: a survey*. In Physica A: Statistical Mechanics and its Applications, vol. 392, no. 11, pp. 2688–2700. Cited on p. 9.

Page, L. ; Brin, S. ; Motwani, R. ; and Winograd, T. . 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab. URL http://ilpubs.stanford.edu:8090/422/. Cited on pp. 21 and 40.

Pang, A. T. ; Wittenbrink, C. M. ; and Lodha, S. K. . 1997. *Approaches to uncertainty visualization*. In The Visual Computer, vol. 13, no. 8, pp. 370–390. Cited on p. 63.

Paulovich, F. V. ; Toledo, F. ; Telles, G. P. ; Minghim, R. ; and Nonato, L. G. . 2012. *Semantic wordification of document collections*. In Computer Graphics Forum, vol. 31, no. 3pt3, pp. 1145–1153. Cited on p. 81.

Pearson, K. . 1901. *LIII. On lines and planes of closest fit to systems of points in space*. In The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572. Cited on p. 55.

Peeters, R. . 2003. *The maximum edge biclique problem is NP-complete*. In Discrete Applied Mathematics, vol. 131, no. 3, pp. 651–654. Cited on p. 29.

Perlin, K. and Fox, D. . 1993. *Pad: an alternative approach to the computer interface*. In Proceedings of the 20th annual conference on Computer Graphics and Interactive Techniques, pp. 57–64. ACM. Cited on p. 81.

Perugini, S. ; Goncalves, M. A. ; and Fox, E. A. . 2004. *Recommender Systems Research: A Connection-Centric Survey*. In Journal of Intelligent Information Systems, vol. 23, no. 2, pp. 107–143. Cited on p. 23.

Pickett, R. M. ; Grinstein, G. ; Levkowitz, H. ; and Smith, S. . 1995. *Harnessing preattentive perceptual processes in visualization*. In Perceptual Issues in Visualization, IFIP Series on Computer Graphics, pp. 33–45. Springer. Cited on p. 74.

Pike, W. A. ; Stasko, J. ; Chang, R. ; and O'Connell, T. A. . 2009. *The science of interaction*. In Information Visualization, vol. 8, no. 4, pp. 263–274. Cited on p. 62.

Pirolli, P. and Card, S. . 2005. *The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis*. In Proceedings of International Conference on Intelligence Analysis, vol. 5, pp. 2–4. Cited on pp. 64, 65, and 67.

Plaisant, C. . 2004. *The challenge of information visualization evaluation*. In Proceedings of the working conference on Advanced visual interfaces, pp. 109–116. ACM. Cited on pp. 71 and 72.

Plaisant, C. ; Grosjean, J. ; and Bederson, B. B. . 2002. *Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation*. In INFOVIS 2002. IEEE Symposium on Information Visualization, pp. 57–64. IEEE. Cited on p. 83.

Plantié, M. and Crampes, M. . 2013. *Survey on Social Community Detection*. In Ramzan, N. ; Zwol, R. ; Lee, J.-S. ; Clüver, K. ; and Hua, X.-S. , eds., Social Media Retrieval, Computer Communications and Networks, pp. 65–85. Springer. Cited on p. 20.

Playfair, W. . 1801. *The statistical breviary*. Wallis. Cited on p. 74.

Podolny, J. M. and Baron, J. N. . 1997. *Resources and relationships: Social networks and mobility in the workplace*. In American sociological review, vol. 62, no. 5, pp. 673–693. Cited on pp. 25 and 27.

Pons, P. and Latapy, M. . 2005. *Computing communities in large networks using random walks*. In Computer and Information Sciences-ISCIS 2005, vol. 3733, pp. 284–293. Springer. Cited on p. 20.

Powers, D. . 2011. *Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation*. In Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63. Cited on p. 71.

Purchase, H. C. . 2002. *Metrics for graph drawing aesthetics*. In Journal of Visual Languages & Computing, vol. 13, no. 5, pp. 501–516. Cited on p. 71.

———. 2012. *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press. Cited on p. 72.

Purchase, H. C. ; Hoggan, E. ; and Görg, C. . 2007. *How important is the "mental map"?–an empirical investigation of a dynamic graph layout algorithm*. In Graph drawing, Lecture Notes in Computer Science, vol. 4372, pp. 184–195. Springer. Cited on p. 105.

Queyroi, F. ; Delest, M. ; Fédou, J.-M. ; and Melançon, G. . 2013. *Assessing the quality of multilevel graph clustering*. In Data Mining and Knowledge Discovery, pp. 1–22. ISSN 1384-5810. Cited on p. 20.

Raskin, J. . 2000. *The humane interface: new directions for designing interactive systems*. Addison-Wesley Professional. Cited on p. 71.

Reichardt, J. and Bornholdt, S. . 2006. *Statistical mechanics of community detection*. In Physical Review E, vol. 74, no. 1, p. 016110. Cited on p. 19.

Renoust, B. and Begovic, M. R. . 2013. *Q&A: The day a big data scientist met a development organization*. UNDP Voices from Eurasia: http://europeandcis.undp.org/blog/2013/02/22/qa-the-day-a-big-data-scientist-met-a-development-organization/. Cited on pp. 4 and 138.

Renoust, B. ; Begovic, M. R. ; and Quaggiotto, G. . 2013a. *Big data and development organizations: What happens when you move from theory to practice?* UNDP Voices from Eurasia: http://europeandcis.undp.org/blog/2013/01/31/big-data-and-development-organizations-what-happens-when-you-move-from-theory-to-practice/. Cited on pp. 4, 135, and 138.

Renoust, B. ; Chomitz, K. M. ; and McKenzie, A. H. . 2013b. *Network analysis applied to project risks identification: what risk factors can explain project outcome?* WorldBank DataDive in DC: http://fr.slideshare.net/renoust/network-analysis-applied-to-project-risks-identification. Cited on pp. 4, 136, and 138.

Renoust, B. ; Melançon, G. ; and Viaud, M.-L. . 2013c. *Assessing group cohesion in homophily networks*. In Advances in Social Network Analysis and Mining (ASONAM) 2013, pp. 149–155. Niagara Falls, Canada, ACM/IEEE. Cited on pp. 4, 59, 111, and 138.

———. 2013d. *Document Corpus Analysis Based on Term Entanglement*. INSNA XXXIII Sunbelt Social Networks Conference for the International Network for Social Network Analysis, Hamburg, Germany. Cited on pp. 4, 59, and 138.

———. 2013e. *Measuring group cohesion in document collections*. In The 2013 IEEE/WIC/ACM International Conference on Web Intelligence (WI) 2013, p. TBA. Atlanta, USA, IEEE/WIC/ACM. Cited on pp. 4, 59, and 138.

———. 2013f. *Mesurer l'intrication sémantique dans une collection de documents*. In 13e Conférence francophone sur l'Extraction et la Gestion des Connaissances / Fouille de Grands Graphes (EGCFGG13), pp. 18–29. Toulouse, France. Cited on pp. 4, 59, 111, and 138.

Renoust, B. ; Viaud, M.-L. ; and Melançon, G. . 2011a. *Détection, visualisation et validation d'évènements médiatiques*. Atelier Fouille de données Complexes, Journée Fouille de Grands Graphes, CNAM Paris, France. Cited on pp. 4, 33, 111, and 138.

———. 2011b. *Mesure de cohérence dans la découverte et la visualisation d'évènements médiatiques*. In Seconde conférence sur les Modèles et l'Analyse des Réseaux: Approches Mathématiques et Informatique (Marami). Grenoble, France. Cited on pp. 4, 33, 59, and 138.

Riche, N. H. ; Dwyer, T. ; Lee, B. ; and Carpendale, S. . 2012. *Exploring the design space of interactive link curvature in network diagrams*. In Proceedings of the International Working Conference on Advanced Visual Interfaces, pp. 506–513. ACM. Cited on p. 76.

Robins, G. and Alexander, M. . 2004. *Small worlds among interlocking directors: Network structure and distance in bipartite graphs*. In Computational & Mathematical Organization Theory, vol. 10, no. 1, pp. 69–94. Cited on pp. 23, 25, and 27.

Rocha, L. E. ; Liljeros, F. ; and Holme, P. . 2010. *Information dynamics shape the sexual networks of Internet-mediated prostitution*. In Proceedings of the National Academy of Sciences, vol. 107, no. 13, pp. 5706–5711. Cited on p. 23.

Rosvall, M. and Bergstrom, C. T. . 2008. *Maps of random walks on complex networks reveal community structure*. In Proceedings of the National Academy of Sciences, vol. 105, no. 4, pp. 1118–1123. Cited on p. 20.

Roth, R. E. . 2012. *Cartographic interaction primitives: Framework and synthesis*. In The Cartographic Journal, vol. 49, no. 4, pp. 376–395. Cited on p. 71.

Royer, L. ; Reimann, M. ; Andreopoulos, B. ; and Schroeder, M. . 2008. *Unraveling Protein Networks with Power Graph Analysis*. In PLoS Computational Biology, vol. 4, no. 7. Cited on p. 110.

Russell, D. M. ; Stefik, M. J. ; Pirolli, P. ; and Card, S. K. . 1993. *The cost structure of sensemaking*. In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pp. 269–276. ACM. Cited on p. 63.

Sallaberry, A. ; Zaidi, F. ; Pich, C. ; and Melançon, G. . 2010. *Interactive visualization and navigation of web search results revealing community structures and bridges*. In Proceedings of Graphics Interface 2010, pp. 105–112. Canadian Information Processing Society. Cited on p. 16.

Salton, G. ; Wong, A. ; and Yang, C.-S. . 1975. *A vector space model for automatic indexing*. In Communications of the ACM, vol. 18, no. 11, pp. 613–620. Cited on p. 14.

Saraiya, P. ; North, C. ; Lam, V. ; and Duca, K. A. . 2006. *An insight-based longitudinal study of visual analytics*. In IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 6, pp. 1511–1522. Cited on p. 72.

Schaeffer, S. E. . 2007. *Survey: Graph clustering*. In Comput. Sci. Rev., vol. 1, no. 1, pp. 27–64. ISSN 1574-0137. Cited on p. 20.

Scholtz, J. . 2006. *Beyond usability: Evaluation aspects of visual analytic environments*. In Proceedings of 2006 IEEE Symposium On Visual Analytics Science And Technology,, pp. 145–150. IEEE. Cited on p. 71.

Schrödinger, E. . 1935. *Discussion of Probability Relations between Separated Systems*. In Proceedings of the Cambridge Philosophical Society, vol. 31, p. 555. Cited on p. 37.

Schulz, H.-J. ; John, M. ; Unger, A. ; and Schumann, H. . 2008. *Visual analysis of bipartite biological networks*. In Proceedings of the First Eurographics conference on Visual Computing for Biomedicine, EG VCBM'08, pp. 135–142. Eurographics Association. Cited on p. 80.

Sebrechts, M. M. ; Cugini, J. V. ; Laskowski, S. J. ; Vasilakis, J. ; and Miller, M. S. . 1999. *Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces*. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 3–10. ACM. Cited on p. 76.

Shahrokhi, F. ; Sỳkora, O. ; Székely, L. A. ; and Vrto, I. . 2001. *On bipartite drawings and the linear arrangement problem*. In SIAM Journal on Computing, vol. 30, no. 6, pp. 1773–1789. Cited on p. 79.

Shalizi, C. R. and Thomas, A. C. . 2011. *Homophily and Contagion Are Generically Confounded in Observational Social Network Studies*. In Sociological Methods & Research, vol. 40, no. 2, pp. 211–239. Cited on p. 23.

Shamir, A. and Stolpnik, A. . 2012. *Interactive visual queries for multivariate graphs exploration*. In Computers & Graphics, vol. 36, no. 4, pp. 257–264. Cited on pp. 81 and 83.

Shannon, C. E. . 1948. *A Mathematical Theory of Communication*. In The Bell System Technical Journal, vol. 27, pp. 379–423, 623–656. Cited on p. 42.

Sharan, R. ; Suthram, S. ; Kelley, R. M. ; Kuhn, T. ; McCuine, S. ; Uetz, P. ; Sittler, T. ; Karp, R. M. ; and Ideker, T. . 2005. *Conserved patterns of protein interaction in multiple species*. In Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 6, pp. 1974–1979. Cited on p. 24.

Shirley, M. D. and Rushton, S. P. . 2005. *The impacts of network topology on disease spread*. In Ecological Complexity, vol. 2, no. 3, pp. 287–299. Cited on p. 18.

Shneiderman, B. . 1996. *The eyes have it: A task by data type taxonomy for information visualizations*. In Proceedings, 1996 IEEE Symposium on Visual Languages, pp. 336–343. IEEE. Cited on p. 69.

Shneiderman, B. and Aris, A. . 2006. *Network visualization by semantic substrates*. In Visualization and Computer Graphics, IEEE Transactions on, vol. 12, no. 5, pp. 733–740. Cited on p. 80.

Shrinivasan, Y. B. and van Wijk, J. J. . 2008. *Supporting the analytical reasoning process in information visualization*. In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1237–1246. ACM. Cited on p. 83.

Simonetto, P. and Auber, D. . 2008. *Visualise undrawable Euler diagrams*. In Proceedings of the 12th International Conference on Information Visualisation, IV'08., pp. 594–599. IEEE. Cited on p. 79.

Skeels, M. ; Lee, B. ; Smith, G. ; and Robertson, G. . 2008. *Revealing uncertainty for information visualization*. In Proceedings of the working conference on Advanced visual interfaces, AVI '08, pp. 376–379. ACM. Cited on p. 63.

Skvoretz, J. and Faust, K. . 1999. *Logit models for affiliation networks*. In Sociological Methodology, vol. 29, no. 1, pp. 253–280. Cited on p. 59.

Snyder, D. and Kick, E. L. . 1979. *Structural position in the world system and economic growth, 1955-1970: A multiple-network analysis of transnational interactions*. In American Journal of Sociology, pp. 1096–1126. Cited on p. 24.

Spanurattana, S. and Murata, T. . 2011. *Visual Analysis of Bipartite Networks*. In IEEE 11th International Conference on Data Mining Workshops (ICDMW '11), pp. 833–838. Cited on p. 80.

Spearman, C. . 1904. *The proof and measurement of association between two things*. In The American journal of psychology, vol. 15, no. 1, pp. 72–101. Cited on p. 55.

Spence, I. . 2005. *No humble pie: The origins and usage of a statistical chart*. In Journal of Educational and Behavioral Statistics, vol. 30, no. 4, pp. 353–368. Cited on p. 74.

Springmeyer, R. R. ; Blattner, M. M. ; and Max, N. L. . 1992. *A characterization of the scientific data analysis process*. In Proceedings of the 3rd conference on Visualization'92, pp. 235–242. IEEE Computer Society Press. Cited on p. 61.

for Standardization, I. O. . 1998. *ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability*. ISO. Cited on p. 63.

Stolte, C. ; Tang, D. ; and Hanrahan, P. . 2002. *Polaris: A system for query, analysis, and visualization of multidimensional relational databases*. In IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pp. 52–65. Cited on p. 81.

Strogatz, S. H. . 2001. *Exploring complex networks*. In Nature, vol. 410, no. 6825, pp. 268–276. Cited on p. 9.

Suchow, J. W. and Alvarez, G. A. . 2011. *Motion silences awareness of visual change*. In Current Biology, vol. 21, no. 2, pp. 140–143. Cited on p. 73.

Sutcliffe, A. ; Ennis, M. ; and Hu, J. . 2000. *Evaluating the effectiveness of visual user interfaces for information retrieval*. In International Journal of Human-Computer Studies, vol. 53, no. 5, pp. 741 – 763. ISSN 1071-5819. Cited on p. 72.

Sweller, J. ; Ayres, P. ; and Kalyuga, S. . 2011. *Cognitive load theory*, vol. 1. Springer. Cited on p. 63.

Sweller, J. ; Van Merrienboer, J. J. ; and Paas, F. G. . 1998. *Cognitive architecture and instructional design*. In Educational psychology review, vol. 10, no. 3, pp. 251–296. Cited on p. 63.

Tamassia, R. . 2013. *Handbook of graph drawing and visualization*. CRC Press. Cited on pp. 62 and 76.

Tarissan, F. ; Quoitin, B. ; MéRindol, P. ; Donnet, B. ; Pansiot, J.-J. ; and Latapy, M. . 2013. *Towards a bipartite graph modeling of the internet topology*. In Comput. Netw., vol. 57, no. 11, pp. 2331–2347. Cited on p. 23.

Teh, Y. W. ; Jordan, M. I. ; Beal, M. J. ; and Blei, D. M. . 2006. *Hierarchical dirichlet processes*. In Journal of the american statistical association, vol. 101, no. 476, pp. 1566–1581. Cited on p. 16.

Thièvre, J. . 2006. *Cartographies pour la recherche et l'exploration de données documentaires*. Ph.D. thesis, Université de Montpellier. Cited on p. 11.

Thomas, H. . 1999. *Probabilistic latent semantic indexing*. In 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM, Berkeley, California, United States. Cited on p. 15.

Thomas, J. J. and Cook, K. A. . 2005. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press. Cited on pp. 3, 61, 64, 65, and 66.

Tominski, C. ; Abello, J. ; van Ham, F. ; and Schumann, H. . 2006. *Fisheye tree views and lenses for graph visualization*. In Tenth International Conference on Information Visualization. IV 2006, pp. 17–24. IEEE. Cited on p. 81.

Tomoharu, I. ; Takeshi, Y. ; and Naonori, U. . 2008. *Probabilistic latent semantic visualization: topic model for visualizing documents*. In 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 363–371. ACM. Cited on p. 13.

Treisman, A. . 1985. *Preattentive processing in vision*. In Computer vision, graphics, and image processing, vol. 31, no. 2, pp. 156–177. Cited on pp. 2 and 73.

Tufte, E. R. and Graves-Morris, P. . 1983. *The visual display of quantitative information*, vol. 2. Graphics press Cheshire, CT. Cited on p. 74.

Usui, S. ; Naud, A. ; Ueda, N. ; and Taniguchi, T. . 2007. *3D-SE Viewer: A text mining tool based on bipartite graph visualization*. In International Joint Conference on Neural Networks, 2007 (IJCNN), pp. 1103–1108. IEEE. Cited on p. 79.

Valente, T. W. ; Coronges, K. ; Lakon, C. ; and Costenbader, E. . 2008. *How correlated are network centrality measures?* In Connections (Toronto, Ont.), vol. 28, no. 1, p. 16. Cited on p. 20.

Valiati, E. R. ; Pimenta, M. S. ; and Freitas, C. M. . 2006. *A taxonomy of tasks for guiding the evaluation of multidimensional visualizations*. In Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization, pp. 1–6. ACM. Cited on p. 71.

Van Wijk, J. J. and Nuij, W. A. . 2003. *Smooth and efficient zooming and panning*. In Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on, pp. 15–23. IEEE. Cited on p. 81.

Verroust, A. and Viaud, M.-L. . 2004. *Ensuring the drawability of extended Euler diagrams for up to 8 sets*. In Proc. Diagrammatic Representation and Inference, 2004 LNAI 2980, pp. 128–141. Springer. Cited on p. 79.

Viaud, M.-L. ; Renoust, B. ; Saulnier, A. ; and Thièvre, J. . 2010. *Bringing Interactive Contextual Maps to Users for Information Retrieval*. In NEM Summit. Barcelona, Spain. Cited on pp. 4, 11, 12, 16, 33, and 138.

Voulgaris, S. ; Kermarrec, A.-M. ; and Massoulié, L. . 2004. *Exploiting semantic proximity in peer-to-peer content searching*. In Proceedings 10th IEEE International Workshop on Future Trends of Distributed Computing Systems, FTDCS 2004., pp. 238–243. IEEE. Cited on p. 23.

Wang, P. ; Sharpe, K. ; Robins, G. L. ; and Pattison, P. E. . 2009. *Exponential random graph (p) models for affiliation networks*. In Social Networks, vol. 31, no. 1, pp. 12–25. Cited on pp. 25 and 59.

Wang Baldonado, M. Q. ; Woodruff, A. ; and Kuchinsky, A. . 2000. *Guidelines for using multiple views in information visualization*. In Proceedings of the working conference on Advanced visual interfaces, pp. 110–119. ACM. Cited on p. 83.

Ward, M. and Yang, J. . 2004. *Interaction spaces in data and information visualization*. In Proceedings of the Sixth Joint Eurographics - IEEE TCVG conference on Visualization, VISSYM'04, pp. 137–146. Eurographics Association. Cited on p. 71.

Ware, C. . 2000. *Information visualization: perception for design*. Elsevier. Cited on pp. 3 and 73.

Wasserman, S. and Faust, K. . 1994. *Social network analysis: Methods and applications*, vol. 8. Cambridge university press. Cited on p. 21.

Watts, D. J. and Strogatz, S. H. . 1998. *Collective dynamics of 'small-world' networks*. In Nature, vol. 393, no. 6684, pp. 440–442. Cited on pp. 9, 19, and 23.

Wehrend, S. and Lewis, C. . 1990. *A problem-oriented classification of visualization techniques*. In Proceedings of the IEEE 1st Conference on Visualization '90, pp. 139–143. IEEE Computer Society Press. Cited on pp. 63 and 71.

West, D. B. et al. 2001. *Introduction to graph theory*, vol. 2. Prentice hall Englewood Cliffs. Cited on p. 3.

White, D. R. ; Burton, M. L. ; and Dow, M. M. . 1981. *Sexual division of labor in African agriculture: a network autocorrelation analysis*. In American Anthropologist, vol. 83, no. 4, pp. 824–849. Cited on p. 17.

van Wijk, J. J. . 2013. *Evaluation: A Challenge for Visual Analytics*. In Computer, vol. 46, no. 7, pp. 56–60. ISSN 0018-9162. Cited on p. 71.

Wills, G. J. . 1996. *Selection: 524,288 ways to say "this is interesting"*. In Information Visualization'96, Proceedings IEEE Symposium on, pp. 54–60. IEEE. Cited on p. 83.

Witten, I. H. and Frank, E. . 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann. Cited on p. 15.

Wolfe, A. W. . 2010. *Anthropologist view of social network analysis and data mining*. In Social Network Analysis and Mining, vol. 1, no. 1, pp. 3–19. Cited on pp. 3 and 144.

Wolff, A. . 2007. *Drawing subway maps: A survey*. In Informatik-Forschung und Entwicklung, vol. 22, no. 1, pp. 23–44. Cited on p. 79.

Wong, N. ; Carpendale, S. ; and Greenberg, S. . 2003. *Edgelens: An interactive method for managing edge congestion in graphs*. In 2003 IEEE Symposium on Information Visualization, pp. 51–58. IEEE. Cited on p. 81.

Wong, P. C. and Bergeron, R. D. . 1997. *30 Years of Multidimensional Multivariate Visualization*. In Scientific Visualization, Overviews, Methodologies, and Techniques, pp. 3–33. IEEE Computer Society. Cited on p. 81.

Wong, P. C. ; Foote, H. ; Mackey, P. ; Chin, G. ; Huang, Z. ; and Thomas, J. J. . 2012. *A Space-Filling Visualization Technique for Multivariate Small-World Graphs*. In IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 5, pp. 797–809. Cited on p. 81.

Wong, P. C. and Thomas, J. . 2004. *Visual Analytics*. In IEEE Computer Graphics and Applications, vol. 24, no. 5, pp. 20–21. Cited on p. 62.

Wong, S. M. ; Ziarko, W. ; and Wong, P. C. . 1985. *Generalized vector spaces model in information retrieval*. In Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 18–25. ACM. Cited on p. 15.

Yan, J. ; Forbus, K. D. ; and Gentner, D. . 2003. *A theory of rerepresentation in analogical matching*. Tech. rep., DTIC Document. Cited on p. 64.

Yee, K.-P. ; Fisher, D. ; Dhamija, R. ; and Hearst, M. . 2001. *Animated exploration of dynamic graphs with radial layout*. In Proceedings of 2001 IEEE Symposium on Information Visualization, pp. 43–. IEEE Computer Society. Cited on p. 83.

Yi, J. S. ; ah Kang, Y. ; Stasko, J. T. ; and Jacko, J. A. . 2007. *Toward a deeper understanding of the role of interaction in information visualization*. In IEEE Transactions on Visualization and Computer Graphics, vol. 13, no. 6, pp. 1224–1231. Cited on pp. 71 and 94.

Zanin, M. and Lillo, F. . 2013. *Modelling the air transport with complex networks: A short review*. In The European Physical Journal Special Topics, vol. 215, no. 1, pp. 5–21. Cited on p. 23.

Zha, H. ; He, X. ; Ding, C. ; Simon, H. ; and Gu, M. . 2001. *Bipartite graph partitioning and data clustering*. In Proceedings of the tenth international conference on Information and knowledge management, pp. 25–32. ACM. Cited on p. 28.

Zhang, P. ; Wang, J. ; Li, X. ; Li, M. ; Di, Z. ; and Fan, Y. . 2008. *Clustering coefficient and community structure of bipartite networks*. In Physica A: Statistical Mechanics and its Applications, vol. 387, no. 27, pp. 6869–6875. Cited on p. 27.

Zhou, T. ; Ren, J. ; Medo, M. ; and Zhang, Y. . 2007a. *Bipartite network projection and personal recommendation*. In Physical Review E, vol. 76, no. 4, p. 046115. Cited on p. 80.

Zhou, T. ; Ren, J. ; Medo, M. ; and Zhang, Y.-C. . 2007b. *Bipartite network projection and personal recommendation*. In Physical Review E, vol. 76, no. 4, p. 046115. Cited on pp. 23, 24, 25, and 27.

Zipf, G. K. . 1935. *The psycho-biology of language.* Houghton, Mifflin. Cited on p. 121.

Zykov, A. A. . 1974. *Hypergraphs*. In Russian Mathematical Surveys, vol. 29, no. 6, pp. 89–156. Cited on p. 23.

APPENDICES

CORRELATION OF THE SUBSTRATE ENTANGLEMENT INTENSITY AND HOMOGENEITY WITH OTHER MEASURES

This section shows raw experimental results of the study of correlation of the entanglement intensity and homogeneity of a substrate node with its adjacent edges with other known measures. The dataset is composed of the 10,000 news excerpts presented in Section 2.1.3. The following table is a duplicate of Table 3.2 that summarizes the correlation between the entanglement intensity and homogeneity with other measures.

	Entanglement intensity	*Entanglement homogeneity*
Betweenness centrality	-0.130	-0.104
Closeness centrality	-0.083	-0.030
Clustering coefficient	0.403	0.295
Degree	-0.108	-0.060
Eccentricity	0.032	0.010
***K*-cores**	0.152	0.104
Multiplexity	-0.095	-0.039
Catalysts	-0.117	-0.078
PageRank	-0.266	-0.170
Strength	0.321	0.251

Table A.1: Comparison of Pearson's correlation coefficient between the entanglement intensity and homogeneity and nine other measures.

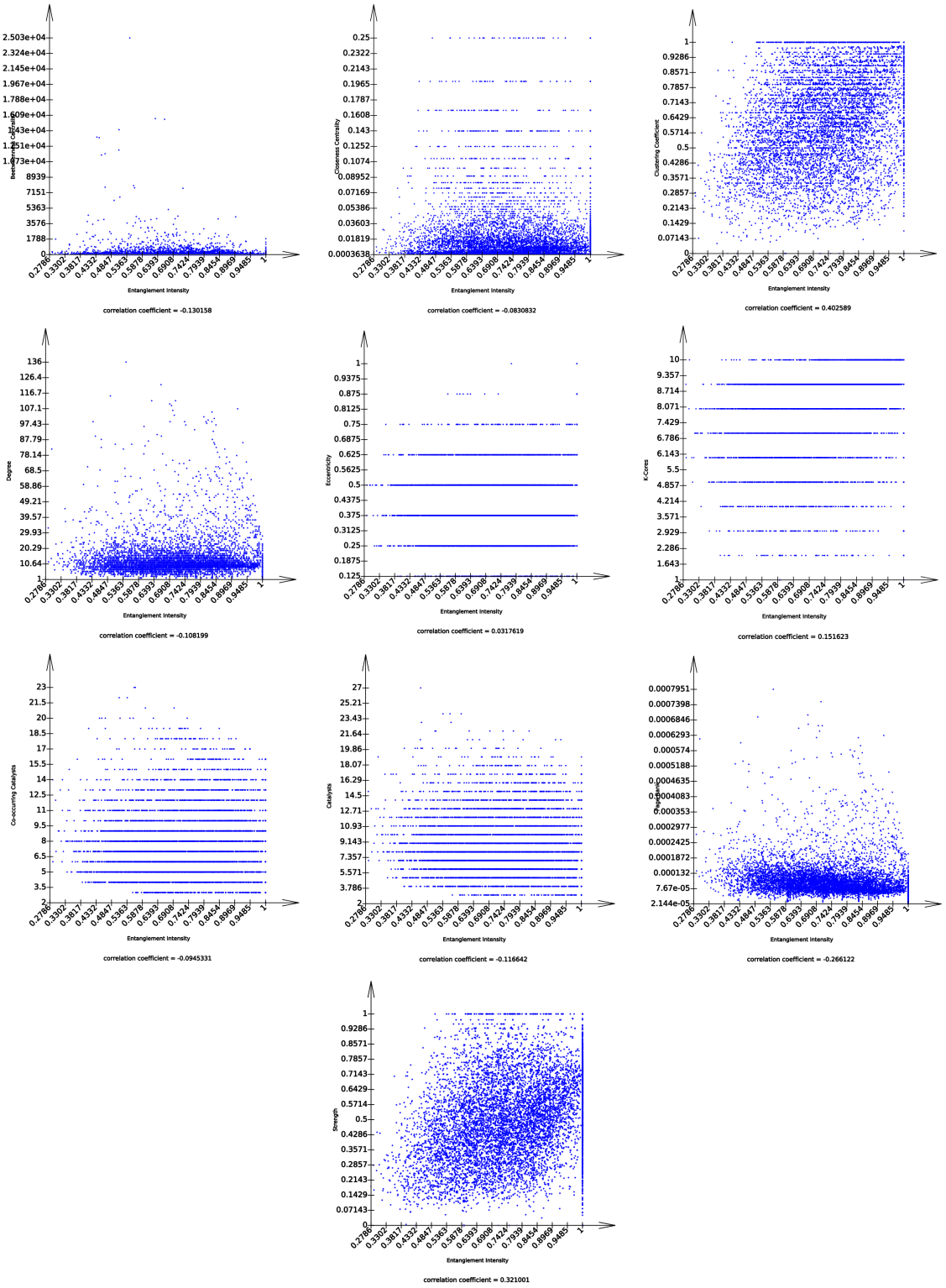


Figure A.1: 2D Plots with the entanglement intensity on the x-axis and another graph measure on the y-axis. Following the order: betweenness centrality, closeness centrality, clustering coefficient, degree, eccentricity, k-cores, multiplexity (equals the number co-occurrent catalysts), number of occurrent catalysts, PageRank, and strength

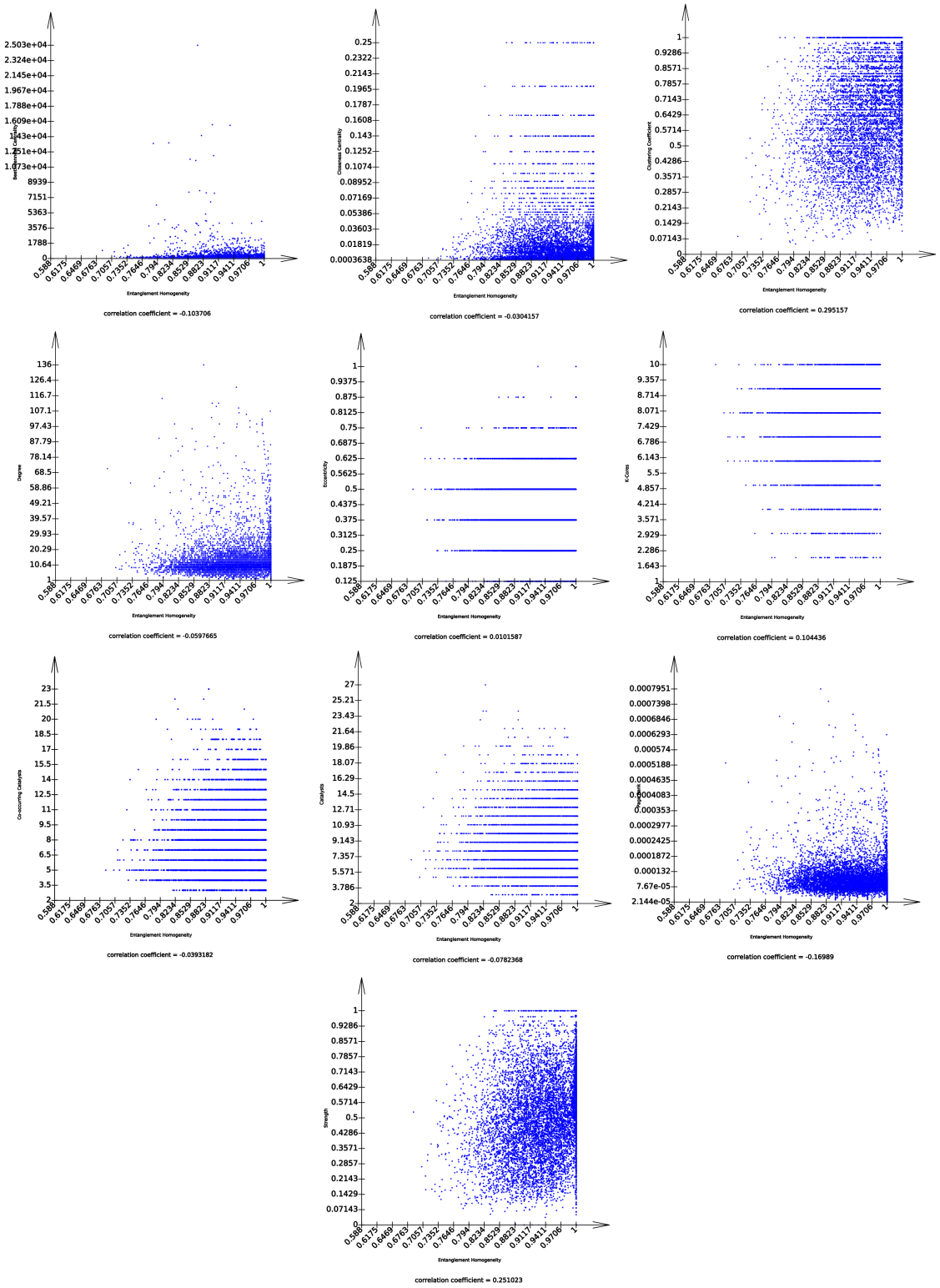


Figure A.2: 2D Plots with the entanglement homogeneity on the x-axis and another graph measure on the y-axis. Following the order: betweenness centrality, closeness centrality, clustering coefficient, degree, eccentricity, k-cores, multiplexity (equals the number co-occurrent catalysts), number of occurrent catalysts, PageRank, and strength

CORRELATION OF MEASURES IN A DOCUMENT NETWORK

B

Inspired by the work we started to study correlation between entanglement measures and other known network measures (presented in Appendix A) in our whole network of documents. We were able to extend the study to every pair of the measures giving rise to this correlation table. This table only aims at giving a reference point of the different correlation between traditional measures in our specific case of document networks.

	Betweenness centrality	***Closeness centrality***	***Clustering coefficient***	***Degree***	***Eccentricity***	***K*-cores**	***Multiplexity***	***Catalysts***	***PageRank***	***Strength***
Btw. centrality		-0.093	-0.278	0.519	0.106	0.110	0.241	0.221	0.486	-0.214
Cls. centrality	-0.093		0.254	-0.184	-0.596	-0.384	-0.186	-0.109	0.047	0.474
Cluster. coef.	-0.278	0.254		-0.451	-0.228	0.017	-0.178	-0.156	-0.451	0.800
Degree	0.519	-0.184	-0.451		-0.028	0.461	0.427	0.341	0.938	-0.123
Eccentricity	0.106	-0.596	-0.228	-0.028		0.022	-0.001	-0.019	0.147	-0.480
***K*-cores**	0.110	-0.384	0.017	0.461	0.022		0.486	0.296	0.291	0.173
Multiplexity	0.241	-0.186	-0.178	0.427	-0.001	0.486		0.881	0.328	-0.064
Catalysts	0.221	-0.109	-0.156	0.341	-0.019	0.296	0.881		0.277	-0.071
PageRank	-0.486	0.047	-0.451	0.938	0.147	0.291	0.328	0.277		-0.072
Strength	-0.214	0.474	0.800	-0.123	-0.480	0.173	-0.064	-0.071	-0.072	

Table B.1: Comparison of Pearson's correlation coefficient between the entanglement intensity and homogeneity and nine other measures.