



HAL
open science

Réseaux de Compétences : de l'Analyse des Réseaux Sociaux à l'Analyse Prédicative de Connaissances

Christophe Thovex

► **To cite this version:**

Christophe Thovex. Réseaux de Compétences : de l'Analyse des Réseaux Sociaux à l'Analyse Prédicative de Connaissances. Intelligence artificielle [cs.AI]. Université de Nantes, 2012. Français. NNT : . tel-00697798

HAL Id: tel-00697798

<https://theses.hal.science/tel-00697798>

Submitted on 16 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE NANTES

École Doctorale

SCIENCES ET TECHNOLOGIES DE L'INFORMATION
ET MATHÉMATIQUES (STIM, ED 503)

Année : 2012

Thèse de Doctorat de l'Université de Nantes

Spécialité : INFORMATIQUE

Présentée et soutenue publiquement par

Christophe Thovex

le 9 mars 2012

à l'Université de Nantes

Réseaux de Compétences : de l'Analyse des Réseaux Sociaux à l'Analyse Prédictive de Connaissances.

Jury

| | | | |
|-----------------|------------------------|---|---|
| Président : | ELIZABETH METAIS | - | Professeur, CNAM Paris (CEDRIC) |
| Rapporteurs : | CHANTAL REYNAUD | - | Professeur, Université Paris Sud 11 (LRI) |
| | BÉNÉDICTE LE GRAND | - | MdC HDR, Université Paris 6 (LIP6) |
| Examineurs : | ELIZABETH METAIS | - | Professeur, CNAM Paris (CEDRIC) |
| | GENOVEVA VARGAS-SOLAR | - | CR, Université de Grenoble (LIG, LAFMIA) |
| | PASCALE KUNTZ-COSPEREC | - | Professeur, Université de Nantes (LINA) |
| <i>Invité :</i> | GILLES GUERRIN | - | Directeur <i>OpenPortal Software S.A.</i> |

Directeur de Thèse : HENRI BRIAND - PR émérite, Université de Nantes
Laboratoire : Laboratoire d'Informatique Nantes Atlantique (UMR CNRS 6241)
Co-encadrant : FRANCKY TRICHET - MdC, Université de Nantes
Laboratoire : Laboratoire d'Informatique Nantes Atlantique (UMR CNRS 6241)
Composante de rattachement du directeur : POLYTECH NANTES

Remerciements

Merci au Professeur émérite HENRI BRIAND qui, le premier, m'a montré une voie que j'avais souvent longée avant de l'emprunter, maintenant devenue évidente à mon sens. Merci pour son soutien et son engagement, en faveur d'une première inscription à l'École Doctorale Sciences et Techniques de l'Informatique et Mathématiques (ED-STIM, 503) dans le cadre de la formation professionnelle continue.

Merci à FRANCKY TRICHET pour ses compétences pédagogiques, sa lucidité, sa franchise et son humanisme en permanence teinté de bonne humeur. Sa capacité de travail, d'imagination et de réflexion, son ouverture d'esprit, ont permis de courir le risque d'un sujet multidisciplinaire atypique, aussi improbable à l'origine, que créatif et satisfaisant à l'arrivée. Merci à sa capacité de discernement.

Merci à CHANTAL REYNAUD et à BÉNÉDICTE LE GRAND, ayant accepté le rôle de rapporteur, ainsi qu'à ELISABETH METAIS et à GENOVEVA VARGAS-SOLAR, ayant accepté de participer au Jury de thèse. Merci également à FABIEN GANDON, pour le suivi attentif et les conseils prodigués au cours des travaux de thèse.

Merci à PASCALE KUNTZ-COSPEREC, à PHILIPPE LERAY, aux chercheuses et chercheurs du Laboratoire d'Informatique de Nantes Atlantique et d'ailleurs rencontrés au détour de couloirs et de conférences, pour les échanges dynamiques et constructifs dont ils m'ont gratifié.

Merci à GILLES GUERRIN, à ALEX LANDREAU, à ARNAUD CLAVERIE, JÉRÔME ROCU, LAURENT VELOURS, CHRISTINE MIDOUX et aux membres de l'entreprise *OpenPortal Software* qui ont fourni le contexte expérimental, le savoir-faire et l'expertise nécessaires à l'application de nos travaux en contexte réel.

Merci à YIDE WANG, à JOSÉ MARTINEZ et au personnel de l'ED-STIM (ED 503), pour leur accueil et leur attention. Merci également à BERNARD MANGUY, aux responsables de l'Université de Nantes ainsi qu'au corps enseignant, pour leur effort de défense d'un système éducatif de qualité, fondement d'un progrès économique et social accessible et construit par tous.

Merci à SYLVIE LEROUX pour son soutien administratif sans faille et sa disponibilité constante.

Merci aux lecteurs de ces pages, à mes proches et à toutes celles et ceux qui me soutiennent. Merci enfin à RENÉ DESCARTES (1596-1650), par qui j'ai connu le doute cartésien et ses vertus, à FRIEDRICH WILHELM NIETZSCHE (1844-1900) et à MICHEL ONFRAY, philosophes militant pour *le gai savoir* et la connaissance scientifique « ouverte ». Ni immoraliste, ni hédoniste, pas même libre-penseur ou humaniste, il me semble être seulement un peu plus humain, grâce à la philosophie...

À bien y réfléchir, plutôt que d'innombrables congratulations, peut-être que d'emporter avec vous le souvenir d'une pensée vous ravirait plus longuement ?

Les paroles s'envolent, les écrits restent, les actes font l'histoire... et une fleur de lumière ensoleille le tout. J'y pense, j'y pense et puis ne l'oublie plus, c'est bien elle qu'il faut remercier !

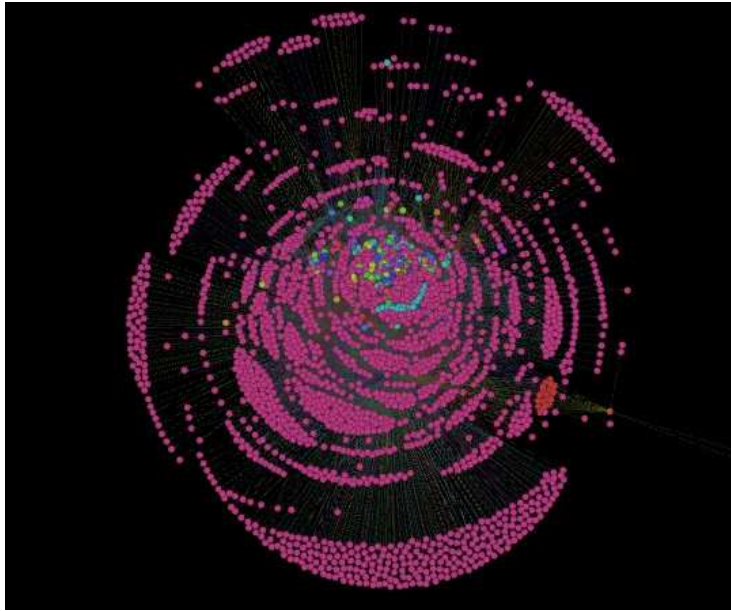


FIGURE 1 – Rose de connaissances.

Table des matières

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Problème adressé | 1 |
| 1.2 | Objectifs | 3 |
| 1.3 | Cadre applicatif | 4 |
| 1.4 | Contributions scientifiques | 5 |
| 1.5 | Structure du document | 8 |
| 2 | Analyse des réseaux sociaux : fouille de graphes et sociologie | 11 |
| 2.1 | Approche graphique en analyse des réseaux sociaux | 12 |
| 2.1.1 | Modèles statiques d'analyse des réseaux sociaux | 13 |
| 2.1.2 | Modèles dynamiques d'analyse des réseaux sociaux | 16 |
| 2.1.3 | Modèles sémantiques d'analyse des réseaux sociaux | 17 |
| 2.1.4 | Vers une analyse prédictive de connaissances | 20 |
| 2.2 | Psychologie sociale du développement | 20 |
| 2.2.1 | Structures organisationnelles | 21 |
| 2.2.2 | Rôle de l'organisation | 22 |
| 2.2.3 | Performance et climat social | 23 |
| 3 | Analyse sociographique et sémantique de réseau social | 27 |
| 3.1 | Fondements et objectifs | 28 |
| 3.2 | Réseau de compétences, structure sémantique de réseau social | 30 |
| 3.3 | Application contextuelle | 32 |
| 3.4 | Réseau global de compétences, structure simplifiée | 33 |
| 3.5 | Mesures statiques et sémantiques | 34 |
| 3.5.1 | C_p , facteur théorique adaptatif | 35 |
| 3.5.2 | Scénario d'usage type | 37 |
| 3.5.3 | Centralité d'intermédiarité sémantique - structure | 37 |
| 3.5.4 | Centralité de flux d'intermédiarité sémantique | 39 |
| 3.5.5 | Centralité de proximité de flux sémantique | 40 |
| 3.5.6 | Modularité de communauté sémantique | 42 |
| 3.5.7 | Centralité de communauté sémantique | 44 |
| 3.6 | Analyse sociophysique et prédictive des réseaux de compétences | 44 |
| 3.7 | Récapitulatif | 46 |
| 4 | Implémentation de réseaux de compétences | 47 |
| 4.1 | Recherche lexico-sémantique d'information dans le contenu endogène | 48 |
| 4.1.1 | Analyse lexicale et indexation du texte intégral | 50 |
| 4.1.2 | Construction de thésaurus et pondération sémantique | 50 |
| 4.1.3 | Développement et codage | 53 |
| 4.2 | Réseaux de compétences à la demande | 53 |

| | | |
|----------|---|-----------|
| 4.2.1 | Étape 1 : sélection et pondération des arcs de $aPcr$ | 55 |
| 4.2.2 | Étape 2 : sélection et valuation de l'intensité des nœuds | 56 |
| 4.2.3 | Propagation des valeurs sémantiques de $aPcr$ | 57 |
| 4.2.4 | Étape 9 : transformation des valeurs sémantiques de Cc_i | 60 |
| 4.2.5 | Étape 10 : équilibrage de pseudo-intensités au cœur d'un graphe de compétences | 61 |
| 4.3 | Valuation pseudo-électrique du graphe de compétences | 62 |
| 4.3.1 | Valuation pseudo-électriques d'arcs et d'extrémités à partir de connaissances-clés | 63 |
| 4.3.2 | Rétro-propagation de pseudo-valeurs électriques depuis Cc_i | 65 |
| 4.3.3 | Modélisation et vérification - solveur de circuit électrique | 72 |
| 4.4 | Un modèle de recommandations basé sur l'équilibre énergétique | 75 |
| 5 | Expérimentations, résultats et recommandations | 77 |
| 5.1 | Expérimentation du modèle prédictif | 78 |
| 5.1.1 | Données expérimentales | 78 |
| 5.1.2 | Résultats généraux sur l'échantillon étudié | 78 |
| 5.1.3 | Analyse d'une recommandation prédictive | 80 |
| 5.1.4 | Application aux graphes de compétences à la demande | 81 |
| 5.2 | Expérimentation de mesures sémantiques | 84 |
| 5.2.1 | Données expérimentales | 84 |
| 5.2.2 | Intermédiation de flux sémantique | 85 |
| 5.2.3 | Centralité de proximité de flux sémantique | 90 |
| 5.2.4 | Modularité de communauté sémantique | 93 |
| 5.2.5 | Mesure de stress et métrique de tension sémantique | 94 |
| 5.3 | Bilan d'expérimentations | 97 |
| 6 | Synthèse, perspectives et conclusion | 99 |
| 6.1 | Comparaison à des travaux similaires | 99 |
| 6.1.1 | Positionnement relatif à l'analyse statique | 99 |
| 6.1.2 | Positionnement relatif à l'analyse sémantique | 100 |
| 6.1.3 | Positionnement relatif à l'analyse dynamique | 102 |
| 6.2 | Points critiques et carences | 103 |
| 6.2.1 | Septicisme et légitimité en approches multidisciplinaires | 103 |
| 6.2.2 | Pertinence de l'indexation sémantique | 104 |
| 6.2.3 | Rigidité du processus de valuation de réseau de compétences | 105 |
| 6.2.4 | Lisibilité des grands réseaux | 105 |
| 6.3 | Perspectives | 106 |
| 6.3.1 | Vers un renforcement du modèle d'analyse sémantique | 106 |
| 6.3.2 | Vers un modèle prédictif basé sur l'électrodynamique | 107 |
| 6.4 | Conclusion | 108 |

| | |
|---|------------|
| A Annexes | 111 |
| A.1 Extrait annoté de documentation technique sur la classification SQL Server | 111 |
| A.1.1 TF.IDF avec CONTAINSTABLE et FREETEXTTABLE . . | 111 |
| A.1.2 Indice de JACCARD avec ISABOUT | 113 |
| A.2 Algorithme de propagation et de rétro-propagation | 115 |
| Bibliographie | 121 |

Introduction

Sommaire

| | | |
|------------|------------------------------------|----------|
| 1.1 | Problème adressé | 1 |
| 1.2 | Objectifs | 3 |
| 1.3 | Cadre applicatif | 4 |
| 1.4 | Contributions scientifiques | 5 |
| 1.5 | Structure du document | 8 |

Cette thèse présente les travaux de recherche effectués de 2009 à 2012 au Laboratoire d'Informatique de Nantes Atlantique (LINA, UMR CNRS 6241), par CHRISTOPHE THOVEX, Ingénieur de Recherche et Doctorant au sein de l'équipe *Connaissances et Décision* (CoD). Ces travaux s'inscrivent dans le cadre du projet SOCIOPRISE, retenu lors de l'appel à projets « Services innovants du Web » du Secrétariat d'État à la prospective et au développement de l'économie numérique, et mené en partenariat avec la société d'édition logicielle et d'ingénierie *OpenPortal S.A.* (<http://www.openportal.fr/>). Ils ont été financés par le Ministère de l'Économie, de l'Industrie et de l'Emploi (Fond Unique Interministériel - FUI) dans le cadre du plan de relance de l'économie nationale.

1.1 Problème adressé

« *The grand challenge is to boost the collective intelligence quotient of organizations and of society* »¹ [Engelbart 1968].

Le *capital humain* désigne les connaissances et compétences acquises par un individu, lui permettant de travailler et de produire des ressources, matérielles ou immatérielles, d'intérêt personnel et/ou collectif [Becker 1994]. Cette notion a été popularisée par G. BECKER, prix Nobel d'Économie en 1992.

Le *capital social* désigne l'interconnexion en réseau(x) de capitaux humains et individuels, formant un capital communautaire permettant de collaborer pour produire des ressources, matérielles ou immatérielles, d'intérêt personnel et/ou collectif. BURT résume le capital humain à « une qualité propre aux individus », et le capital social à « une qualité créée entre acteurs » [Burt 1995].

Nos tendances et besoins en communication appellent en permanence de nouvelles fonctions et applications sur les réseaux sociaux, comme en témoigne l'apparition permanente de nouveaux espaces d'échanges connectés sur le web. Ces espaces

1. D. ENGELBART a inventé la souris et un système précurseur de l'hypertexte.

numériques sont la cause, voire l'effet, de nouveaux modes de socialisation autour de pratiques communautaires, telles le partage instantané d'informations brèves (*i.e.*, Twitter) et le partage d'informations personnelles (*e.g.*, Facebook), de médias ou de signets (*e.g.*, YouTube, Diigo). Ces espaces d'échanges numériques, à l'avantage de nos espaces d'échanges réels, facilitent l'analyse statistique et l'application de métriques ou méthodes d'Analyse des Réseaux Sociaux (ARS, SNA - *Social Networks Analysis*).

En 1934, MORENO introduisait les notions de sociométrie et de réseau social, ainsi que l'association entre mathématiques et psychologie sociale [Moreno 1934]. Les métriques d'ARS basées sur les graphes sont apparues vers 1950, préfigurant la notion de *centralité* avec diverses études sociologiques et mathématiques sur le comportement de groupes [Bavelas 1948]. Des mesures sur les caractéristiques et comportements des communautés ont ensuite été développées, et certaines de ces mesures, applicables sur des graphes orientés ou non orientés, sont devenues des standards en ARS - *e.g.*, centralités d'*intermédiarité* [Freeman 1977, Newman 2005].

Les mesures standards d'ARS appliquent des méthodes statistiques exploratoires sur des structures graphiques - *i.e.*, des graphes sociaux - représentant divers liens identifiés au sein d'une population (*e.g.*, liens idéologiques). Elles permettent, par des principes arithmétiques, géodésiques et statistiques, d'identifier et de quantifier des amas, des communautés et des individus au sein d'un graphe social. En premier lieu, ce sont des mesures statiques, basées sur l'analyse instantanée, qui ne tiennent pas compte de l'évolution temporelle et de la dynamique des réseaux sociaux étudiés. Ces mesures ne constituent donc aucun modèle prédictif d'ARS. En second lieu, elles ne prennent en compte aucune autre sorte d'information relative à la population étudiée, que celle intrinsèquement représentée par la structure du graphe social et par les valeurs quantitatives attribuées aux éléments du graphe². La sémantique des résultats qu'elles produisent reste donc relativement pauvre, comparée à la richesse sémantique disponible dans l'environnement des réseaux sociaux actuels (*e.g.*, actions inhérentes aux médias sociaux, telles que la publication d'information textuelle et le suivi automatique des publications d'autrui).

Alors que les réseaux sociaux « publics » concernent la sphère personnelle et la vie privée, entreprises et institutions professionnelles leur accordent un intérêt croissant, notamment en vue d'apporter de nouvelles fonctions collaboratives favorisant la performance par de nouvelles formes d'organisation du travail et de gestion des ressources humaines. Le travail collaboratif évolue vers des outils Web « 2.0 », caractérisant une forme d'entreprise « 2.0 » sensible à la gestion du capital humain et social [Jean 2001]. Associé au développement de l'internet nomade et géolocalisé (4G, smartphones), ceci révèle une probable tendance à la spécialisation/localisation des réseaux sociaux, en voie d'adaptation aux usages de communautés professionnelles et territoriales.

Les institutions du monde professionnel se sont dotées de nombreux outils dé-

2. Les éléments d'un graphe sont nommés sommets et arêtes (graphe non orienté), ou nœuds et arcs (graphe orienté).

cisionnels visant à l'amélioration de la performance. Avec l'ouverture du monde entrepreneurial aux systèmes d'information collaboratifs, au web « 2.0 » et à un nouveau type d'organisation, dite *participative* dans une *société de la connaissance*, chacun pressent la nécessité et la possibilité de disposer d'outils décisionnels innovants, basés sur l'Analyse des Réseaux Sociaux (ARS).

Les mesures développées jusqu'à présent en ARS ne répondent que partiellement à cette perspective, car les limitations qu'elles doivent à l'approche mathématique ne permettent pas de produire des indicateurs aussi pertinents que souhaité par les professionnels des domaines intéressés - *e.g.*, Marketing, Gestion de Ressources Humaines (GRH). Ainsi, une agence de recrutement appliquant les mesures standards d'ARS pour trouver, sur les réseaux sociaux du web, des profils disposant de compétences spécifiques, n'a que peu de chances d'obtenir un résultat probant puisque ces mesures prennent en compte la structure du réseau, sans les connaissances sous-jacentes. Pour résorber ce type de carence fonctionnelle, il est nécessaire d'intégrer à ces mesures le traitement du *contenu endogène* aux réseaux sociaux - *i.e.*, les ressources présentes dans le contenu relatif à un réseau social. Dans le contexte évoqué, l'intégration de modèles d'indexation et de recherche *lexicale, syntaxique* et *sémantique* d'information textuelle, permet de dépasser les limitations d'origine mathématique, intrinsèques aux mesures standards d'ARS et à l'approche historique ayant permis leur développement. Parallèlement, les limitations dues à la nature statique de ces mesures posent un problème en terme d'analyse dynamique des réseaux sociaux, créant une carence de fonctions décisionnelles prédictives pour les entreprises et institutions exploitant l'analyse décisionnelle des réseaux sociaux, des réseaux du web aux réseaux sociaux professionnels internes.

En résumé, nous observons deux limitations intrinsèques aux mesures standards d'ARS, générant deux carences fonctionnelles dans le contexte de l'ARS appliquée au réseaux sociaux actuels :

1. la limitation mathématique génère une carence fonctionnelle en terme de *pertinence* des résultats obtenus, relativement à l'information disponible ;
2. la nature statique des mesures génère une carence de fonctions d'analyse comportementale, pour les entreprises et institutions exploitant l'analyse décisionnelle des réseaux sociaux.

1.2 Objectifs

En 2008, deux des fondateurs de Facebook déclaraient devoir quitter la société, pour concevoir un nouveau genre d'applications « qui deviendrait à votre vie professionnelle, ce que Facebook est à votre vie sociale »³.

Appliquées aux réseaux sociaux du web ou à des réseaux sociaux internes d'entreprises ou d'institutions, les mesures standards d'ARS présentent deux carences fonctionnelles, faute d'intégrer l'information disponible dans le contenu endogène du

3. Extrait de : « ... our products will become to your work life what Facebook.com is to your social life. » - <http://facereviews.com/2008/10/03/facebook-co-founder-moskovitz-to-leave>.

réseau social analysé et de produire une analyse dynamique. Les travaux développés dans le cadre de cette thèse répondent au double objectif de résorber ces carences.

Le premier objectif visé consiste à définir un facteur qualitatif, basé sur la sémantique du contenu endogène échangé entre individus constitutifs d'un réseau social. Ce facteur permet de qualifier les membres du réseau et les interactions qu'ils créent, en relation aux connaissances présentes dans le contenu endogène. Il est générique, en vue d'intégrer facilement la dimension sémantique aux principales mesures d'ARS. L'extension sémantique de certaines mesures de centralité standards d'ARS est présentée, grâce au facteur défini.

Le second objectif visé consiste à définir un modèle prédictif d'ARS qui développe une approche multidisciplinaire, croisant l'Ingénierie des Connaissances (IC), la Théorie des Graphes et l'une des 4 forces physiques fondamentales, la force électromagnétique⁴. Le modèle défini est applicable aux Réseaux Sociaux d'Entreprises et d'Institutions (RSEI). Les spécificités d'un RSEI sont :

1. graphe social de volume limité à 100 000 nœuds maximum,
2. membres du réseau social produisant et consommant du contenu professionnel - *i.e.*, contenu endogène restreint à un ou quelques domaines de connaissances connexes,
3. adoption de méthodes de travail collaboratif sur le contenu endogène (*e.g.*, partage de ressources).

Les travaux présentés introduisent de nouvelles synergies multidimensionnelles entre aspects statiques, sémantiques et dynamiques de l'analyse des RSEI.

1.3 Cadre applicatif

Les travaux relatés dans cette thèse ont principalement été réalisés dans le cadre du projet SOCIOPRISE, mené en partenariat avec l'éditeur de solutions logicielles de gestion du capital humain *OpenPortal*.

OpenPortal (<http://www.openportal.fr/>), anciennement « *web agency* » spécialisée en solutions de portails collaboratifs, propose des solutions « full web » en mode interne ou en SaaS, pour la gestion de la formation initiale, de la formation professionnelle continue, du plan de formation et des compétences. L'offre de services logiciels dédiés à la gestion du capital humain pour les entreprises et institutions (universités académiques et universités d'entreprise) couvre la gestion des entretiens individuels et la Gestion Prévisionnelle de l'Emploi et des Compétences (GPEC). Les solutions proposées, paramétrables et flexibles, s'adaptent à la plupart des contraintes métiers rencontrées. Couplées à l'annuaire d'entreprise interopérable *OpenPortal*, elles répondent ainsi aux exigences de 2500 utilisateurs et gestionnaires de ressources humaines, pour 110 clients répartis dans le monde (*e.g.*, Accenture, Areva, Thomson, Veolia, Vinci). *OpenPortal* S.A. est *Microsoft Gold Partner* et

4. La physique moderne définit deux forces nucléaires (faible et forte), la force électromagnétique et la gravité comme étant les 4 forces fondamentales de l'Univers.

développe ses produits logiciels sur un socle technologique propriétaire (.Net/SQL Server), à partir duquel elle interface, si nécessaire, les technologies Open Source (*e.g.*, Java, PHP).

Le projet SOCIOPRISE vise à définir les modèles théoriques et à développer les prototypes des futures solutions innovantes dédiées à la Gestion du Capital Social (GCS). Il ouvre une voie scientifique de convergence pour des modèles acquis en sciences exactes (statistiques, physique) et en sciences cognitives (informatique, linguistique, psychologie), visant au développement du bien-être social et économique des entreprises et institutions, par optimisation de l'adéquation entre équipes de travail, compétences, formation-éducation et organisation du travail. Ce projet, d'un budget global avoisinant 700 K€ pour une durée effective de 28 mois, a été retenu par la commission d'experts de la Direction Générale de la Compétitivité, de l'Industrie et des Services (DGCIS), dans le cadre de l'appel à projet « Services innovants du Web », lancé en 2009 par le Secrétariat d'État à la prospective et au développement de l'économie numérique. Il est issu d'un partenariat privé-public entre *OpenPortal* S.A. et le LINA (UMR CNRS 6241). *Socioprise* est financé à 60% par le Ministère de l'Économie, de l'Industrie et de l'Emploi, sur le Fonds Unique Interministériel (FUI) et dans le cadre du plan de relance de l'économie nationale. Il comporte trois volets basés sur l'Ingénierie des Connaissances (IC), dont le volet central correspond aux travaux présentés. Ce volet associe IC et ARS, pour définir et développer le prototype d'un modèle décisionnel innovant, dédié à la gestion du capital social et de la performance. Aucun produit du projet ne peut être directement commercialisé au terme du projet, mais les résultats obtenus autorisent la création par *OpenPortal* S.A à court et moyen terme, de valeur commerciale directement ou indirectement issue de ces résultats.

Afin d'éprouver sur un socle technologique totalement Open Source la faisabilité de certains modèles théoriques lui appartenant, le LINA a ponctuellement collaboré avec *DoYouBuzz*⁵, société innovante établie à Nantes et spécialisée dans le recrutement et la réputation en ligne. Cette collaboration a été réalisée dans le cadre d'un financement régional (Appel à projets *TIC et Innovation*). Les modèles éprouvés dans ce contexte sont issus d'une approche spécifique aux objectifs applicatifs de la start-up nantaise (*i.e.*, partage de CV en ligne et aide au recrutement), bien qu'également fondés sur le couplage IC-ARS. Ces modèles sortent du cadre de thèse. Ils sont signalés en complément d'information sur le contexte applicatif, dans le cas où cela puisse favoriser de futures perspectives de partenariats publics-privés ou privés-privés, pour la recherche et le développement d'applications innovantes dans le domaine de la gestion du capital social et des ressources humaines.

1.4 Contributions scientifiques

Les travaux d'ARS présentés dans cette thèse résultent de trois constats pluridisciplinaires. D'une part, l'ARS est une approche sociologique de la fouille de

5. <http://www.doyoubuzz.com/fr/>

graphes - étude de graphes sociaux. D'autre part, l'ARS peut intégrer deux dimensions : (1) une dimension statique analysant le graphe à l'instant t et (2) une dimension dynamique analysant le comportement du graphe, impliquant les instants T, T' et les états E, E' . Pour l'ARS dynamique, deux approches se distinguent : (1) les modèles dynamiques probabilistes, tels que les modèles prédictifs markoviens appliqués aux graphes sociaux [Pearson 2003], et (2) les modèles physiques statistiques et sociophysiques [Barrat 2006], [Galam 2008]. Le modèle électromagnétique prédictif prôné dans le cadre de nos travaux s'inscrit en parallèle à la sociophysique, s'en rapprochant par le fondement métaphysique et s'en différenciant par l'aspect non probabiliste.

Le troisième et dernier constat porte sur le fait que la sémantique du contenu partagé au sein des réseaux sociaux (*i.e.*, le *contenu endogène*), est restée marginalement intégrée aux modèles d'ARS, bien que l'on trouve les prémisses de son exploitation dans [Mika 2005]. De ce fait, un recruteur cherchant, à l'aide de métriques d'ARS appliquées sur un graphe social donné (*e.g.*, LinkedIn, Monster, Réseau interne d'entreprise), les *profils* pointus les plus recommandables dans leur communauté propre (*e.g.*, ingénieur turbines à gaz, doreur), se voit simplement recommander les individus les plus globalement connectés - *i.e.*, vedettes, personnages médiatiques, ou animateurs du réseau. Dans ce contexte, notre modèle permet de retrouver des communautés minoritaires, excentrées ou peu connectées, relatives à des compétences précises et/ou rares et d'identifier des individus clés au sein de communautés. Ce sont, typiquement, le spécialiste turbines à gaz ayant la meilleure expérience des principaux fabricants du secteur, ou le doreur en position centrale dans le réseau compagnonique de son corps de métiers. Ainsi, au sein d'une institution professionnelle, le modèle fournit des recommandations pour l'organisation agile des équipes en permettant d'identifier les ressources humaines éligibles sur un projet en terme de compétences, tout en minimisant le risque de pénaliser d'autres projets - *i.e.*, optimisation de la gestion du capital social.

L'ARS associe la sociologie, la statistique et la Théorie des Graphes. Un réseau social peut être formalisé à l'aide d'un graphe, orienté et valué ou non. Nos travaux intègrent une approche psychosociale des organisations et différencient trois formes d'ARS : (1) l'ARS *statique*, (2) l'ARS *sémantique* et (3) l'ARS *dynamique*. La figure 1.1 représente un exemple de graphe social d'entreprise, dans lequel différentes indications issues de l'ARS sont illustrées. Une *niche* (ou *clique*) est, par exemple, un groupe de sommets dans lequel chacun des sommets est directement connecté à tous les autres. Le concept de *trou structurel* a été défini par BURT comme l'absence de connexions horizontales entre branches distinctes d'une arborescence hiérarchique [Burt 1995]. Les autres notions présentées en figure 1.1 (*i.e.*, degré, centralité, intermédiarité, proximité), seront expliquées et approfondies dans la suite de cette thèse. L'illustration montre également que plusieurs réseaux peuvent coexister au sein d'une même organisation - *i.e.*, un réseau hiérarchique et un réseau collaboratif.

L'ARS *statique* étudie l'état E de graphes sociaux à un instant t . Elle est fondée sur des modèles et mesures de structures et de flux, permettant la classification de

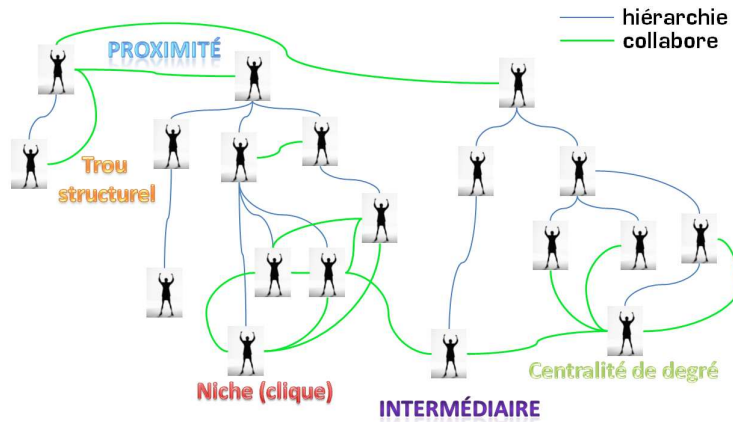


FIGURE 1.1 – Exemple de graphe social d'entreprise.

groupes d'individus ou de communautés, et la découverte de relations implicites entre individus au sein du graphe social par le calcul de *degré*, de *connectivité*, de *distance* ou de *flux*⁶.

L'ARS *sémantique* étudie les aspects conceptuels des graphes sociaux. Elle se réfère à l'Ingénierie des Connaissances (IC) et à l'ingénierie des ontologies, en corrélation aux sciences cognitives. Elle vise à produire des résultats pertinents en rapport aux connaissances qualifiant les communautés et individus d'un réseau social, et repose sur l'exploitation de la sémantique contenue dans les ressources - ou médias - diffusés au sein du réseau. Avec la croissance exponentielle des réseaux sociaux et de leurs flux d'information internes, l'ARS sémantique devient cruciale pour la découverte et la gestion de connaissances, des systèmes d'information collaboratifs d'entreprise aux grandes communautés du Web.

L'ARS *dynamique* étudie le comportement de graphes sociaux, impliquant les instants T et T' et les états E et E' . Les notions élémentaires d'analyse statique (degré, connectivité, distance, flux) sont associées à l'étude de comportements temporels, en vue de prévoir les évolutions de structures et de flux au sein des réseaux sociaux.

À notre connaissance, aucune approche actuelle ne vise à intégrer les trois formes d'analyse différenciées. Nos travaux répondent à ce manque et définissent un système multidimensionnel d'analyse statique, sémantique et dynamique de réseaux sociaux professionnels et institutionnels [Thovex 2010a].

Nos travaux développent plusieurs nouveaux concepts constituant une contribution pluridisciplinaire, de l'ARS vers l'analyse prédictive des réseaux de connaissances. En ARS statique, nous définissons 5 nouvelles mesures sémantiques de centralité et de *clustering* sémantique : (1) centralité d'intermédiation sémantique -

6. Le nombre d'arêtes sur un sommet donne son *degré*. Le nombre d'autres sommets accessibles depuis un sommet donne sa *connectivité*. La *distance* entre deux sommets est le nombre d'arêtes/arcs les séparant, ou le poids de ces arêtes/arcs. Un *flux* est donné en nombre d'unités circulant entre deux sommets.

structure, (2) centralité de flux d'intermédiation sémantique, (3) centralité de proximité sémantique - flux, (4) modularité de communauté sémantique, (5) centralité de communauté sémantique. La méthode généraliste définie pour générer ces centralités et modularité permet d'intégrer la dimension sémantique à d'autres mesures statiques d'ARS [Thovex 2010d, Thovex 2010e, Thovex 2011b].

En ARS dynamique, nous introduisons : (1) une métrique de *tension* d'un réseau social et sa notion paire, la *réactance*, utiles à l'évaluation du *stress* au sein du réseau, et (2) un modèle sémantique, prédictif et non probabiliste d'analyse dynamique des réseaux sociaux professionnels [Thovex 2010b, Thovex 2010c, Thovex 2011a, Thovex 2011c].

Nous définissons et expérimentons un modèle dynamique et prédictif, basé sur l'équilibre naturel des forces électrostatiques. Ce modèle s'inscrit en parallèle à la sociophysique, s'en rapprochant par le fondement métaphysique et s'en différenciant par l'aspect non probabiliste. Il définit une approche physique de la structure de graphe social qui développe et consolide la dimension électrique introduite en ARS dans [Brandes 2005] et [Newman 2005].

Au cours de la réalisation des travaux, il a été défini une structure de graphe adaptée aux objectifs du projet et à l'analyse statique, sémantique et dynamique des flux informationnels collaboratifs, porteurs des connaissances produites et consommées par les individus composant un réseau social professionnel. Cette structure est nommée *réseau de compétences*. Elle favorise l'application simultanée de standards de l'ARS, de l'ingénierie des connaissances et de l'électrodynamique [Thovex 2012a, Thovex 2012b].

Les contributions apportées, appliquées aux réseaux de compétences, introduisent un modèle multidimensionnel et pluridisciplinaire d'analyse des réseaux sociaux professionnels. Ce modèle fait converger la dimension sémantique avec la dimension électrodynamique, en vérifiant les lois d'OHM et de KIRCHHOFF. Il favorise l'émergence d'une méthode d'anticipation dynamique et prédictive de l'évolution des réseaux sociaux professionnels et de leurs compétences, à partir du contenu endogène.

Au sein d'une institution professionnelle, le modèle fournit des recommandations pour l'organisation agile des équipes en permettant d'identifier les ressources humaines éligibles sur un projet en terme de compétences, tout en minimisant le risque de pénaliser d'autres projets - *i.e.*, optimisation de la gestion du capital social.

1.5 Structure du document

Cinq chapitres composent la suite de ce document. Le chapitre 2 présente une synthèse des principes et méthodes rencontrés en ARS statique, sémantique et dynamique. Cet synthèse se justifie en terme applicatif par des éléments de psychologie sociale permettant l'introduction d'une forme d'analyse psychosociale des organisations.

Le chapitre 3 expose les contributions apportées par les hypothèses de travail

retenues. Un modèle théorique de type statique et sémantique, puis un modèle dynamique et sémantique y sont définis, ainsi qu'une structure de graphe hybride, tripartite et évolutive, nommée *réseau de compétences* et spécifiquement définie pour favoriser l'application des modèles théoriques présentés.

Le chapitre 4 décrit l'implémentation des *réseaux de compétences à la demande* et d'un *réseau global de compétences*. Ce chapitre introduit également un principe de propagation et rétro-propagation de valeurs d'*intensité sémantique* sur la structure de graphe de compétences, qui définit un processus transdisciplinaire intégrant le passage de la dimension sémantique à la dimension électrodynamique. La transition d'une dimension à l'autre autorise l'induction de valeurs pseudo-électriques, de *tension* et de *résistance*, qui conditionnent un modèle prédictif applicable aux réseaux de compétences.

Le chapitre 5 présente les résultats des expérimentations menées dans le contexte applicatif du projet SOCIOPRISE. Les résultats obtenus sont présentés et les graphes visualisés, pour l'application des contributions développées sur un réseau global de compétences et sur un réseau de compétences à la demande. Les expérimentations concernent l'application des mesures de centralités sémantiques, des métriques de tension et de *stress*, et de notre modèle prédictif basé sur l'équilibre naturel des forces électrostatiques. Les résultats sont évalués empiriquement, selon l'avis des experts et la mémoire collective d'entreprise, et soumis à discernement selon la connaissance du contexte que nous avons acquise.

Ces travaux rapprochent l'ARS statique et sémantique des réseaux de compétences, avec les propriétés électrodynamiques observées sur les réseaux électroniques. Ils fixent un modèle multidimensionnel et pluridisciplinaire d'ARS statique, sémantique et prédictif, basé sur l'IC et sur l'équilibre naturel des flux énergétiques. Ce modèle permet de produire des fonctions décisionnelles innovantes et des recommandations dédiées à la gestion du capital humain et social. Le chapitre 6 compare notre approche avec les travaux similaires et introduit une réflexion sur les perspectives scientifiques et socio-économiques de nos contributions. Il introduit également une approche psychosociale des principaux verrous organisationnels susceptibles de juguler ces perspectives.

Analyse des réseaux sociaux : fouille de graphes et sociologie

Sommaire

| | | |
|------------|--|-----------|
| 2.1 | Approche graphique en analyse des réseaux sociaux | 12 |
| 2.1.1 | Modèles statiques d'analyse des réseaux sociaux | 13 |
| 2.1.2 | Modèles dynamiques d'analyse des réseaux sociaux | 16 |
| 2.1.3 | Modèles sémantiques d'analyse des réseaux sociaux | 17 |
| 2.1.4 | Vers une analyse prédictive de connaissances | 20 |
| 2.2 | Psychologie sociale du développement | 20 |
| 2.2.1 | Structures organisationnelles | 21 |
| 2.2.2 | Rôle de l'organisation | 22 |
| 2.2.3 | Performance et climat social | 23 |

« Affronter le réel dans sa vérité et sa totalité ». C'est l'assertion donnée par S.R. STEINMETZ, ethnologue et sociologue (1862-1940), pour définir la *sociographie* - *i.e.*, étude descriptive des réalités et des faits sociaux [Steinmetz 1913]. Cette conceptualisation intuitive inspire la notion de *sociogramme* - ou représentation graphique des liens sociaux, géographiques, culturels ou économiques regroupant des individus -, une notion présentée avec le concept de *sociométrie* par J.L. MORENO [Moreno 1934]. La sociométrie préfigurait alors l'Analyse des Réseaux Sociaux (ARS).

La figure 2.1 présente un sociogramme utilisé par MORENO pour représenter la position d'un leader isolé (individu A). L'individu A est susceptible d'influencer 50 personnes, virtuellement prédisposées en sa faveur via l'individu B.

En définitive, l'ARS tend à identifier les groupes d'individus membres de réseaux et caractériser leurs positions et influences respectives, relations multiples et comportements (passé, présent, futur), dans des réseaux uniques ou superposés. L'ARS est une approche de la sociologie par la fouille de graphes. Ce chapitre offre un aperçu de ces domaines et de certaines de leurs spécificités pertinentes dans le cadre des travaux présentés. Il présente l'approche graphique, puis l'approche sociologique, correspondant aux objectifs du projet SOCIOPRISE.

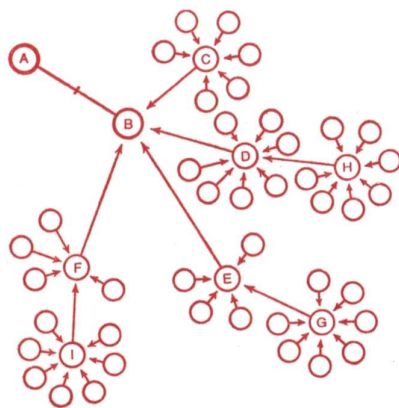


FIGURE 2.1 – Sociogramme de leader isolé [Moreno 1934].

2.1 Approche graphique en analyse des réseaux sociaux

Cette section résume les principes, méthodes et observations applicables en ARS, relativement à l’approche graphique développée pour SOCIOPRISE.

Un graphe non orienté connecte des sommets par des arêtes. Un graphe orienté connecte des nœuds par des arcs. Les éléments d’un graphe peuvent porter des étiquettes les qualifiant, et/ou des valeurs les quantifiant. La structure (*i.e.*, la topologie) du graphe et les flux y circulant, quantifiés par des valeurs et éventuellement qualifiés par des labels, caractérisent le graphe. Deux axes récurrents peuvent être différenciés en fouille de graphes : l’étude des structures (topologies) et l’étude des flux.

En ARS, l’étude des structures associe généralement mathématiques et sociologie. Elle peut conduire à l’édiction de principes stratégiques applicables en organisation du travail et gestion de ressources humaines, impliquant les réseaux sociaux professionnels et institutionnels. BURT considère par exemple les trous structurels (*i.e.*, l’absence de connexions horizontales entre branches distinctes d’une arborescence hiérarchique) comme facteur de progrès en entreprise, car en favorisant le pouvoir des dirigeants et l’application des procédures, ils contribuent au développement global de la société [Burt 1995]. Au contraire, pour LAZEGA, la diminution des trous structurels est un facteur de réactivité et de productivité, essentiel pour la performance [Lazega 2001].

L’étude des flux associe communément sciences physiques et mathématiques avec peu d’écho en sociologie, en dehors de modèles de propagation comme ceux de formation d’opinions cités dans [Barrat 2006] et [Galam 2008], d’études épidémiologiques [Zekri 2002], ou d’extrapolations récurrentes sur le paradoxe de *Milgram*. Le phénomène du petit monde (*small world*) a été découvert par S. MILGRAM suite à une expérience menée en 1967 à l’aide de courriers postaux. Le sociologue établit alors l’hypothèse que tout individu des États-Unis était joignable de n’importe quel autre via une moyenne de 6 intermédiaires, grâce au discernement [Travers 1969].

2.1.1 Modèles statiques d'analyse des réseaux sociaux

L'ARS statique étudie l'état E de graphes sociaux à un instant t , E étant défini par les structures et/ou les flux des graphes étudiés. Les prémisses de l'ARS sont apparues avec la notion de sociométrie, dans [Moreno 1934], puis dans [Freeman 1960]. Elles concernaient l'étude des relations et de la dominance dans les communautés. Ces notions se sont enrichies autour des mesures de *centralité* [Freeman 1977], caractérisant les propriétés de réseaux sociaux en terme de *pouvoir*, *prestige*, *proximité* ou *confiance* [Freeman 1989].

2.1.1.1 Modèles statiques d'analyse structurelle

Les mesures de centralité sont basées sur la comparaison du degré d'un sommet à ceux du graphe, voisins ou distants. Un sommet connecté (directement ou non) à un grand nombre de sommets du graphe porte un coefficient de *centralité de pouvoir* élevé. Un sommet connecté aux sommets à fort degré¹ du graphe social possède une *centralité de prestige* importante. Un sommet connecté à un grand nombre de sommets voisins ou proches possède une *centralité de proximité* élevée. Par induction, centralité de prestige et de proximité importantes pour un même sommet peuvent révéler un coefficient de *confiance* significatif².

Une mesure de *centralité d'intermédierité* quantifie l'importance d'un individu pour l'interconnexion des autres individus et communautés du graphe social. Cette mesure est basée sur la recherche de plus courtes chaînes, ou chemins géodésiques, dans [Freeman 1977]. En figure 2.2, le diamètre des sommets est déterminé par la centralité d'intermédierité de sommets [Freeman 1977] et les sommets violets sont les plus intermédiaires, en rapport à la topologie du graphe représenté.

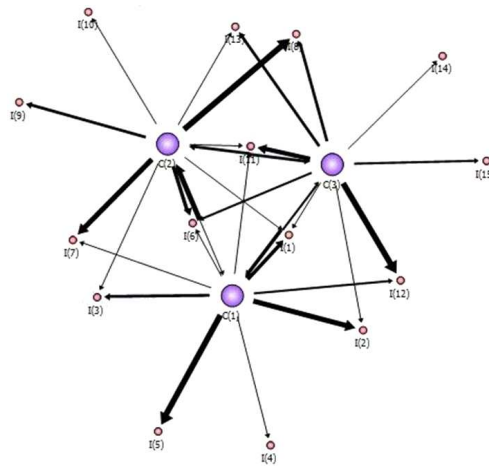


FIGURE 2.2 – Exemple de centralité d'intermédierité de sommet [Freeman 1977].

1. Des sommets portant beaucoup d'arêtes.
2. L'indice de confiance est également associé à un principe d'annotation de certains nœuds par des experts.

Lorsque les arêtes ou les arcs d'un graphe portent des valeurs, le chemin géodésique ne correspond pas systématiquement au flux minimal, ou maximal, d'information circulant entre les sommets ou nœuds du graphe. Pour cette raison, la définition de centralité d'intermédiarité est basée sur la recherche de flux minimal/maximal dans [Newman 2005], où l'auteur introduit une analogie entre l'amplification des flux au sein d'un graphe et la résistance des circuits électriques. En figure 2.3, le diamètre des sommets est déterminé par la centralité d'intermédiarité de flux [Newman 2005] et certains sommets rouges sont aussi intermédiaires que les sommets violets, relativement aux flux circulant au sein du graphe représenté.

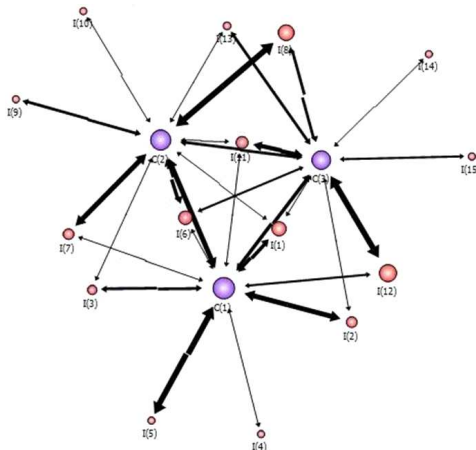


FIGURE 2.3 – Exemple de centralité d'intermédiarité de flux [Newman 2005].

Nous formalisons initialement la centralité d'intermédiarité (*betweenness centrality*) comme suit :

$$\forall i \neq u \neq j, \sigma(i, j) > 0, I_u = \sum_{(i,j)} \frac{\sigma(i, u, j)}{\sigma(i, j)} \quad (2.1)$$

où $\sigma(i, j)$ est le nombre de chemins géodésiques entre chaque paire de sommets (i, j) ; $\sigma(i, u, j)$ est le nombre de chemins géodésiques entre les paires (i, j) passant par u . Dans [Shimbel 1953], $\sigma(i, u, j)$ est également nommée *centralité de stress*.

Les propriétés structurelles sont définies pour les principaux types de graphes sociaux et fournissent des éléments d'ARS statique. Dans les graphes aléatoires [Erdős 1959], le degré des n sommets du graphe est déterminé par une probabilité $p(n)$ avec $p \mapsto [0; 1]$. Avec les graphes pseudo-aléatoires, le degré des n sommets est distribué suivant une loi de distribution uniforme (*e.g.*, loi de Gauss) par laquelle un graphe $G(V, E)$ possède une densité d'arcs $p = |E| \div \binom{|V|}{2}$. Pour les graphes libres d'échelle [Barabasi 1999], les nœuds les plus connectés accroissent leur degré de connexion suivant une loi de puissance (« *rich get richer* »). En déterminant des comportements caractéristiques à chaque type de réseaux et sous-graphes, ces propriétés structurelles statiques apportent aussi des éléments d'analyse dynamique des graphes sociaux.

2.1.1.2 Modèles statiques d'analyse des flux

Les travaux sur les arcs en Théorie des Graphes (*e.g.*, recherche de flux minimal/maximal), sont applicables à l'analyse statique de flux dans les réseaux sociaux, parfois avec des résultats intéressants. C'est le cas de l'étude du *petit monde* (*small world*) dans laquelle V. LATORA et M. MARCHIORI ont introduit la notion d'efficacité, définie comme une mesure de communication, inversement proportionnelle au plus court chemin valué entre deux sommets i et j [Latora 2001]. En rapport à l'hypothèse du petit monde, les travaux de LESKOVEC and HORVITZ sur les grands graphes sociaux (*i.e.*, MSN - 179 millions de sommets) ont actualisé l'hypothèse des six degrés de séparation³. La longueur maximum de plus courte chaîne trouvée par TRAVERS et MILGRAM était 14 et la longueur moyenne de plus courte chaîne était 5.2, pour 100% des 64 chaînes complètes [Travers 1969]. Au sein du réseau social planétaire expérimenté dans [Leskovec 2008], moins de 48% des sommets furent atteints en 6 sauts et des chemins géodésiques de longueur égale 29 ont été identifiés. En conséquence, au sein d'un réseau planétaire et malgré la réalité de l'effet *petit monde*, il serait erroné de croire en la capacité de chacun à joindre quiconque via 6 intermédiaires⁴.

Certains modèles physiques sont traités à l'aide de graphes pour faciliter la compréhension et la découverte de principes théoriques. Dans le domaine de l'électricité, les *loi des nœuds* et *loi des mailles* de KIRCHHOFF en sont une bonne illustration. La figure 2.4 montre un exemple d'application de la loi des nœuds, dans laquelle $i_1 + i_4 = i_2 + i_3$.

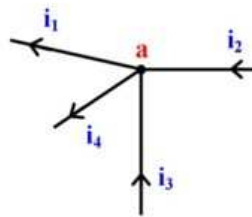


FIGURE 2.4 – Illustration de la loi des nœuds de KIRCHHOFF.

Les travaux de [Thomassen 1990] sur la résistance et les courants des réseaux infinis, en démontrant l'unité et la continuité des flux dans les grands graphes, préfiguraient également les travaux plus récents, inspirés par la métaphore électronique dans [Newman 2004, Newman 2005] et [Brandes 2005] pour identifier des structures communautaires au sein de graphes sociaux [Newman 2006]. Typiquement, à partir de ces modèles multidisciplinaires, évaluer le *stress* au travail reviendrait à estimer

3. Dans [Leskovec 2008], suivant une courbe longue-traîne, la distribution atteint 78% des sommets en 7 sauts et 90% des sommets en 7,8 sauts, dans un échantillon de 1000 sommets tirés aléatoirement du jeu de données.

4. Le calcul de $\log(n)$ sur les jeux de données expérimentés donne une longueur moyenne de plus courtes chaînes égale à 19 pour le réseau planétaire (179 Millions de sommets), 6.9 pour son échantillonnage (1000 sommets) et 4 pour le réseau de MILGRAM (55 sommets).

une tension individuelle basée sur l'intensité du travail, et comparée à une valeur moyenne ou à un rapport de charge sur capacité.

Par le manque de prise en compte du contenu endogène et des comportements évolutifs inhérents aux réseaux sociaux, l'approche statique de l'ARS semble insuffisante à elle seule, pour générer des indicateurs participant à la réduction du risque social⁵ [Blouin 1997], et à l'optimisation de la performance d'équipes de travail par adaptation du capital humain communautaire - *i.e.*, le capital social.

2.1.2 Modèles dynamiques d'analyse des réseaux sociaux

L'ARS dynamique étudie le comportement de graphes sociaux, impliquant les instants T, T' et les états E, E' . Les notions présentes en analyse statique (*i.e.*, degré, connectivité, distances, centralités, flux) y sont employées à l'étude de comportements temporels et de modèles prédictifs sur les structures et flux.

2.1.2.1 Modèles dynamiques d'analyse structurelle

Les propriétés dynamiques inhérentes aux types de graphes évoqués dans la section 2.1.1 offrent un point de passage de l'analyse statique à l'analyse dynamique de graphes sociaux [Krivelevich 2002]. Les modèles prédictifs markoviens, fonctionnant sur les probabilités de transition sont généralement plus efficaces pour l'analyse dynamique des flux que pour anticiper l'évolution des structures graphiques [Pearson 2003]. Les travaux récemment présentés dans [Chen 2011], montrent toutefois une utilisation possible du contenu endogène, pour qualifier des résultats issus de modèles cachés de MARKOV lexicaux (L-HMM) et de chaînes à conditions de champs aléatoires (CRF). GOODREAU propose une alternative applicable aux réseaux sociaux d'amitié, en associant de multiples distributions (degrés, voisinage, propriétés géodésiques, triangles) pour déduire des règles explicites d'identification de structures à partir de statistiques topologiques [Goodreau 2007]. L'effet des structures de graphes sur la dynamique de réseau permet d'induire des règles d'association [Benchettara 2009] et de prédire des comportements évolutifs [Goodreau 2007] suffisants pour en extraire des connaissances quasi-conceptuelles, dans un domaine défini [Ruiz 2007].

Les phénomènes de propagation en réseau sont étudiés d'un point de vue structurel dans divers domaines, tels que la médecine, l'environnement (*e.g.*, effets de pollution et de contaminations) ou l'analyse du Web social - *e.g.*, degré de robustesse [Albert 2000]. Comme en ARS statique, certains modèles physiques (*e.g.*, thermodynamique) génèrent également des résultats intéressants, tels l'utilisation du recuit simulé⁶ pour la découverte d'isomorphismes de graphes [Burel 1993]. Nous nous appuyons sur ces précédents pour oser une approche multidisciplinaire de l'ARS dynamique.

5. Le risque social désigne le *facteur de risque présent dans la structure sociale pouvant causer une maladie ou un traumatisme*.

6. Le recuit simulé, ou *simulated annealing*, est un principe de libération des contraintes de déformation d'un corps solide par échauffement localisé.

2.1.2.2 Modèles dynamiques d'analyse des flux

L'intérêt de modèles physiques, pour l'analyse statique de flux et l'analyse dynamique des réseaux sociaux, a été évoqué par certains travaux en électrodynamique et en thermodynamique - *i.e.*, Lois de KIRCHHOFF, [Thomassen 1990], [Burel 1993].

Encore peu rapprochés de l'ARS, les modèles de percolation (traversée, filtration) sont incontournables en matière d'analyse dynamique de flux dans les graphes [Langlands 1992]. Au sens commun, la percolation désigne le phénomène de traversée d'un solide ou d'un fluide (liquide ou gazeux). Les principes percolateurs s'accompagnent généralement de propagation par capillarité, porosité ou perméabilité [Koppal 2008]. La capillarité est un phénomène lié à la force de cohésion moléculaire d'un fluide (tension de surface) qui provoque son attraction à la surface d'un solide (*e.g.*, encre et buvard). La porosité désigne la capacité d'une matière à recevoir des fluides (*e.g.*, éponge) et la perméabilité, sa capacité à laisser passer les fluides ou modifier les champs magnétiques (perméabilité magnétique). Ces termes, exprimant des propriétés physiques, désignent également des phénomènes de flux observés sur les graphes [Langlands 1992] [Koppal 2008] et présumés présents au sein des réseaux sociaux. A ce titre, ils peuvent s'appliquer à l'analyse dynamique de flux d'information dans les réseaux sociaux, pour la découverte, par exemple, de *réseaux d'influence*⁷ [Jonasson 1999] ou pour l'étude de la propagation d'épidémies dans un réseau de petit monde et du comportement des amas enfouis dans le graphe social [Zekri 2002]. Certains modèles sociaux de formation d'opinion (*e.g.*, vote ou mouvements de mode) sont également fondés sur des principes physiques et s'inspirent, notamment, des travaux de ISING sur l'électron et le magnétisme [Galam 2008].

Parmi les domaines rencontrés en relation aux graphes et réseaux, l'électrostatique - *e.g.*, OHM, KIRCHHOFF - et la découverte des phénomènes magnétiques et électrodynamiques - *e.g.*, AMPÈRE, ISING - marquent une continuité historique en matière d'analyse dynamique de structures et de flux. Le rapprochement métaphysique entre ARS et force électromagnétique (*i.e.*, l'une des 4 forces de l'Univers, en physique moderne) se décline dans les deux sens, lorsque les physiciens recourent aux mesures de centralités pour prévenir le risque de défaillances dans les réseaux de puissance électriques [Wang 2010]. L'analyse dynamique de réseaux sociaux reste encore peu développée, et dans la continuité historique, nous développons une hypothèse audacieuse basée sur [Newman 2005, Brandes 2005] et [Galam 2008].

2.1.3 Modèles sémantiques d'analyse des réseaux sociaux

An ontology is an explicit specification of a conceptualization [Gruber 1995].

Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée qui représente les concepts, objets et autres entités supposés exister sur une aire d'intérêts, avec leurs relations. L'ARS sémantique étudie les aspects conceptuels

7. Un réseau d'influence est représenté par un sous-graphe où circulent les flux majoritaires d'un réseau d'information.

de graphes sociaux. Elle est fondée sur l'ingénierie des connaissances et en particulier sur l'ingénierie des ontologies, conjuguée aux principes d'ARS [Gruber 2008].

L'intégration d'ontologie dans un système lexical d'indexation et de recherche d'information, permet de trouver des ressources textuelles dont le signifié s'inscrit dans le domaine conceptuel où se situent les mots clés caractérisant une recherche d'information. Avec le mot clé « analyse », un système lexical de recherche d'information permet d'identifier les ressources indexées contenant les termes « analyse, analyses, analysé, analyser », voire « analytique ». Les ressources indexées dont ces termes sont absents, mais contenant des termes dont le signifié est proche ou similaire, tels que « observation, examen, étude, vérification », voire « raisonnement », sont ignorées.

Il existe divers niveaux de richesse ontologique, de la classification hiérarchique de termes au sein d'un thésaurus (*e.g.*, oiseau/passereau/{ moineau, rossignol, pinson }) à la définition d'ontologies lourdes incluant des systèmes d'axiomes de domaine et de règles - *e.g.*, *axiome* : un oiseau est un ovipare, *règle* : un oiseau vole, sauf si c'est une Autruche ou un Kagou ou un Kiwi ou un Dronte de Maurice (ou Dodo). Il existe également des mesures permettant de pondérer les concepts et relations sémantiques représentées au sein d'une ontologie, en fonction de leur importance lexicale et conceptuelle [Aimé 2009]. Les mesures de similarité et de proximité conceptuelles définies dans [Aimé 2011], peuvent produire des résultats en adéquation avec le domaine de connaissances représenté et avec la conceptualisation que s'en font les experts.

A l'heure actuelle, peu de travaux significatifs ont été publiés dans le domaine de l'ARS sémantique, alors que l'étude de [Limpens 2008] montre un besoin évident pour les applications du Web social. L'ARS sémantique favorise la création d'applications d'entreprises innovantes, d'applications mobiles ou de systèmes d'information géographiques intégrant la dimension conceptuelle aux réseaux sociaux créateurs et/ou consommateurs d'information partagée.

J. JUNG AND J. EUZENAT [Jung 2007] commentent la description d'une vue tridimensionnelle sur l'analyse sémantique de réseaux sociaux, rapprochant graphes sociaux, annotations (*tags*) et ontologies. La proposition superpose et fait coïncider les trois dimensions pour construire des ontologies « consensuelles » dont les annotations sont associées au graphe social. L'évaluation réciproque entre membres d'un réseau social donne un exemple d'annotation où l'interaction humaine avec le système produit une valuation des sommets du graphe social sur laquelle un *degré de confiance* (*cf.* section 2.1.1) relativement fiable et précis peut être calculé⁸ [Andrade 2007]. De récents travaux sur l'ARS sémantique ouvrent la voie de l'analyse sémantique statistique du web social et visent à rendre opérationnelles les grandes lignes de l'ARS en les intégrant aux ontologies et langages du web sémantique [Erétéo 2011]. La figure 2.5 résume le résultat obtenu grâce au modèle d'ARS défini dans [Erétéo 2011]. Dans la figure, les sommets rouges, bleus ou verts

8. On parle de *réseau de faveurs* quand la structure du graphe social dépend des évaluations entre pairs.

représentent respectivement des annotations, des ingénieurs et des superviseurs académiques.

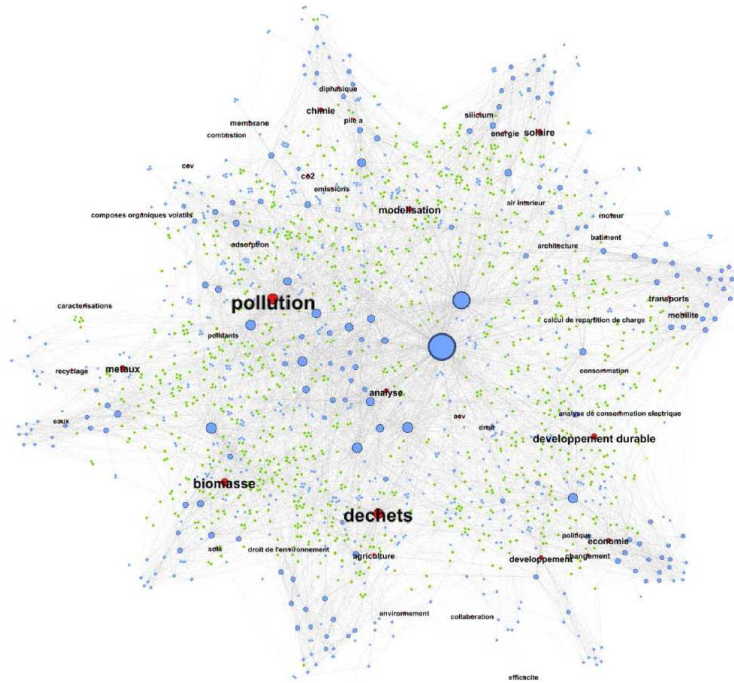


FIGURE 2.5 – Exemple de réseau social intégrant le contenu endogène.

L'analyse linguistique est un élément de l'ingénierie des connaissances et de la recherche d'information. En analyse linguistique, le taux de fréquence TF est défini par le rapport du nombre d'occurrences d'un terme t au sein d'un document d , sur le nombre de termes que contient d . La fréquence inverse de documents IDF est définie par le rapport du nombre $|d|$ des documents constituant un corpus, sur le nombre $|dt|$ de documents contenant un terme t . De nombreuses études sur le raffinement de TF et IDF , généralement combinés ($TF.IDF$), se trouvent dans la littérature scientifique. Les poids obtenus en recherche d'information à l'aide d'une mesure de type $TF.IDF$ indiquent la prégnance des termes recherchés dans le corpus. Notre modèle d'analyse linguistique est basé sur la définition publiée dans [Robertson 1976]. Il intègre l'indice de similarité de JACCARD [Jaccard 1901], basé sur les travaux publiés dans [Rogers 1960]. L'indice de similarité de JACCARD $J(A, B)$ définit la similarité de deux termes au sein d'un corpus, en fonction de l'ensemble A de documents contenant le terme t_1 et de l'ensemble B de documents contenant le terme t_2 . $J(A, B)$ établit le rapport du nombre de documents contenant simultanément les deux termes (*i.e.*, $|A \cap B|$) sur le nombre de documents contenant un terme ou l'autre terme (*i.e.*, $|A \cup B|$).

Le modèle présenté dans [Delalonde 2007], plus centré sur l'ingénierie des connaissances que sur l'ARS, est comparable au socle sémantique du modèle dé-

veloppé dans nos travaux. Les deux modèles recourent à des mesures standards d'analyse linguistique, de type *TF.IDF*. L'idée d'une ARS sémantique s'appuyant sur l'analyse linguistique semble désormais se généraliser, comme le montrent les travaux récents d'analyse de comportements en réseau à partir de blogs et de contenu textuel inhérent aux réseaux sociaux [Macskassy 2011], [Awadallah 2011].

Notre approche, en accord avec la vision de GRUBER, tend à intégrer les principes et langages du web sémantique aux réseaux sociaux, pour le développement de systèmes d'intelligence et de connaissances collectives, des grandes communautés du Web aux réseaux sociaux d'entreprises - *i.e.*, « *Collective Intelligence* », « *Collective Knowledge Systems* » [Gruber 2008]. Nous développons notre modèle en l'appliquant spécifiquement aux réseaux sociaux d'entreprise, dans le contexte du projet SOCIOPRISE.

2.1.4 Vers une analyse prédictive de connaissances

De l'introduction de la sociométrie, dans [Moreno 1934], aux avancées récentes en fouille du web sémantique de [Erétéo 2011], l'ARS a été un sujet de recherche multidisciplinaire, auquel des sociologues, statisticiens, informaticiens, physiciens et même biologistes, ont participé. Les nouvelles technologies, Internet et de nouveaux modes de vie sociale récemment apparus, montrent l'importance des réseaux sociaux et de leur analyse. La pluralité des origines et objectifs, qui a fait de l'ARS un domaine scientifique riche et complexe, a aussi rendu l'ARS difficile à gérer en tant que domaine unique de recherche. Alors que l'ARS statique est plutôt du ressort de la sociologie et des mathématiques, l'ARS dynamique est plutôt corrélée aux mathématiques et à la physique. L'informatique est de plus en plus impliquée en ARS statique ou dynamique, en accord avec son développement historique, et se trouve plus spécifiquement engagée en ARS sémantique, alors que le web sémantique implique également des aspects linguistiques et de psychologie cognitive.

Les aspects multidisciplinaires et historiques de l'ARS rendent difficile le développement et l'application de modèles impliquant les trois différentes approches, statique, dynamique et sémantique. Cependant, croiser ces trois approches peut produire des résultats pertinents et significativement efficaces. Nos travaux visent à prendre pour base la sociologie et l'analyse statique, et à intégrer des principes de physique et de sociophysique en profitant d'acquis en ingénierie des connaissances, afin de développer un modèle prédictif de gestion des connaissances par l'analyse des réseaux sociaux.

2.2 Psychologie sociale du développement

Cette section présente un aperçu des principes, méthodes et observations dans le domaine sociologique, correspondant aux objectifs d'optimisation de la performance, de réduction du risque psychosocial et d'amélioration du climat social, visés par SOCIOPRISE.

2.2.1 Structures organisationnelles

Du début de l'ère industrielle à nos jours, l'organisation moderne du travail repose sur deux formes de regroupement au sein des structures organisationnelles : le *regroupement par fonction* et le *regroupement par marché* [Louche 2007]. Si le contexte des structures organisationnelles est déterminant (*i.e.*, environnement, technologie, taille), ces deux formes de regroupement s'y retrouvent systématiquement - *e.g.*, organisation bureaucratique *vs.* organisation missionnaire.

Les entreprises et institutions adoptent généralement un modèle de classification empirique, lorsqu'elles définissent l'organisation du travail qui leur est propre. Ces approches contingentes s'inscrivent dans des principes connus de structuration organisationnelle, entre les configurations standardisées du modèle de MINTZBERG [Mintzberg 1982], favorables à la production de masse, et le flou adaptatif, propice à l'innovation, de structures organiques telles que l'*adhocratie* [Louche 2007]. Les approches contingentes évoquées ci-dessus, basées sur l'analyse statistique de relations causales, occultent en grande partie le rôle du personnel. En identifiant les abus d'interprétation des relations de causalité, CROZIER AND FRIEDBERG critiquent et dénoncent les limites de l'analyse contingente [Crozier 1977].

C'est à partir d'un courant d'approches cognitivistes⁹ que le savoir social organisé a été conçu en tant que reflet de la personnalité des individus le constituant. Il semble que ce soit STAW qui, le premier, ait eu l'idée d'inverser les relations causales caractéristiques de l'analyse contingente. Dans [Staw 1975], l'auteur relate une expérience réalisée sur le terrain, dans laquelle la performance connue d'un groupe d'employés est une variable indépendante, positive ou négative, et la structure du groupe est une variable fixe. L'expérience est construite de telle façon que ses résultats soient irréfutables. Ces résultats montrent que l'organisation du groupe est liée au niveau de performance indiqué après la réalisation d'une tâche ou d'une mission. L'auteur pose alors deux hypothèses : (1) la structure organisationnelle est conséquence, et non cause, de performance, et (2) les membres d'une organisation se font une théorie personnelle des corrélations entre caractéristiques organisationnelles et performance. Cette théorie, nommé *Théorie Implicite des Organisations* (TIO), est communément acceptée en psychologie sociale. La TIO propre à chaque responsable le mène à choisir la structure organisationnelle des groupes de travail, selon l'idée qu'il s'en fait. La TIO est également définie par le contexte présent et passé. Ceci pose la question du rôle de l'organisation, abordée dans la section 2.2.2.

L'*organisation liquide* des groupes de travail semble actuellement être la plus récente évolution de structure organisationnelle en termes de flexibilité, de gestion des compétences et d'optimisation du ratio charge/capacité [Demailly 2008]. Elle se définit par un mode de répartition des ressources basé sur l'exploitation maximale des compétences individuelles au sein des groupes de travail. L'individu y gagne en diversité d'actions, en entretien et développement de compétences, en reconnaissance relative aux compétences possédées, et en employabilité. L'organisation

9. Un courant cognitiviste d'analyse des structures organisationnelles s'est développé aux Etats-Unis après 1970.

y trouve une performance accrue, une réduction potentielle des charges de personnel et du *turn-over*, moyennant le développement du modèle de gestion du capital humain et social requis.

En résumé, la TIO constitue un verrou structurel, auquel les modèles de gestion du capital humain et social sont actuellement soumis. Ce verrou peut être levé par une approche inverse des causalités, en accord aux travaux de de Staw [Staw 1975] et de CROZIER AND FRIEDBERG [Crozier 1977], et par une structure organisationnelle basée sur l'exploitation maximale des compétences, telle que l'organisation liquide des groupes de travail [Demailly 2008].

2.2.2 Rôle de l'organisation

L'organisation joue un rôle primordial dans (1) la relation que lui confère l'individu, et dans (2) les processus collectifs.

La relation individu/organisation correspond à des rôles variables et évoluant en phases successives au cours du temps. MORELAND AND LEVINE identifient 3 rôles et 4 phases, qui définissent un principe de socialisation organisationnelle [Moreland 1982]. L'individu en phase de socialisation organisationnelle passe par les rôles de non-membres, de quasi-membres (individu en cours d'intégration dans l'organisation) et de membre à part entière. Les phases de socialisation correspondant respectivement à ces rôles sont l'investigation, la socialisation, le maintien et le souvenir, dernière phase dans laquelle l'individu est redevenu non-membre. Si ce modèle général semble plausible, il n'est pas certain que tous les entrants dans une organisation le vérifient. Cependant, la synthèse de FISHER [Fisher 1986] sur les diverses étapes de socialisation connues, produit un résultat ne démentant pas foncièrement le principe de MORELAND AND LEVINE.

En se basant sur ce fondement, il semble possible d'envisager des principes génériques facilitant l'accueil des entrants. Ces *tactiques de socialisation* seront d'autant plus utiles dans une organisation liquide des groupes de travail, où l'accueil des entrants est une micro-étape souvent répétée au sein de petits groupes. Elles peuvent, par exemple, observer les six étapes définies par VAN MAANEN AND SCHEIN [Van Maanen 1979]. Dans le cadre d'une structure d'organisation liquide, motivée par une approche inverse des causalités, les tactiques de socialisation sont un facteur de performance et de progrès social déterminant.

À notre connaissance, les systèmes actuels de gestion du capital humain et social ne tiennent pas compte de l'existence de modèles tels que ceux identifiés et définis par VAN MAANEN AND SCHEIN [Van Maanen 1979], MORELAND AND LEVINE [Moreland 1982] et FISHER [Fisher 1986], pour la gestion du relationnel individu/organisation.

La socialisation organisationnelle joue un rôle essentiel dans les processus collectifs, où elle représente un facteur de cohésion agissant sur le rendement des équipes et sur le climat social. Ces aspects sont étudiés dans la section 2.2.3.

2.2.3 Performance et climat social

Au sein de groupes de travail, performance et climat social représentent deux variables interdépendantes, soumises au contexte organisationnel et déterminantes dans l'exécution de processus collectifs, ou collaboratifs.

LOUCHE distingue deux modèles théoriques d'orientation des processus collectifs : le *modèle de la cohésion sociale* et le *modèle de l'identification sociale* [Louche 2007]. Le premier de ces modèles concerne principalement les relations et phénomènes organisationnels intragroupes, alors que le second modèle concerne plutôt les relations et phénomènes intergroupes. Les deux modèles prennent largement en compte la dimension affective individuelle dans l'organisation des processus collectifs.

Cohésion sociale et rendement

Le modèle de cohésion sociale est profondément ancré sur la dimension affective de l'individu au sein de groupes de travail exécutant des processus collectifs. L'attraction du groupe ou la crainte du rejet sont des facteurs d'intégration agissant sur le rendement individuel et collectif. Le travail de synthèse réalisé par VILLENEUVE ET LETARTE [Villeneuve 1995] à partir de 38 études montre, d'une part, que le rendement est en corrélation avec la cohésion, et d'autre part, que cette corrélation dépend du contexte institutionnel - *i.e.*, équipe sportive, armée, entreprise. L'étude de plusieurs exemples par TANNENBAUM montre que le développement du lien social au sein de groupes de travail favorise également l'adaptation au poste [Tannenbaum 1967].

Identification sociale et conflits

Les auteurs de [Tajfel 1971] montrent que des conflits intergroupes peuvent naître alors qu'aucun conflit d'intérêt n'existe entre les groupes. Dans [Louche 2007], le modèle de l'identification sociale permet d'expliquer ces conflits intergroupes, constatés dans les structures organisationnelles à groupement par fonction et n'apparaissant pas dans les structures à groupement par marché. Parce que l'individu pense pouvoir plus facilement trouver une identité sociale positive au sein de son groupe qu'au sein de groupes étrangers, l'identification sociale des membres de l'*in-group* tend vers le positif et celle des membres de l'*out-group* vers le négatif, entraînant un phénomène de discrimination intergroupe. Lorsque des individus de groupes différents partagent les mêmes tâches (groupement par fonction), l'identification sociale engendre des conflits entre individus par simple connaissance de l'appartenance au groupe étranger, même en l'absence de conflit d'intérêt. Lorsque les groupes travaillent sur des marchés différents (groupement par marché), l'identification sociale n'a plus d'influence sur les conflits puisque les tâches ne sont pas partagées entre les groupes.

En résumé, les modèles d'orientation de processus collectifs de LOUCHE encouragent (1) la cohésion interne des groupes de travail, et (2) la formation des groupes autour des tâches de travail - et/ou réciproquement.

Stress et performance

Pour l'année 2000, le coût du stress d'origine professionnelle en France a été évalué par l'Institut National de Recherche et de Sécurité (INRS), à 830 millions d'Euros [Van de Leemput 2005]. Le stress s'accompagne d'une détérioration de la relation et de l'engagement au travail. Il est l'objet de nombreuses études et peut être causé par divers facteurs, telles que l'environnement physique, le contenu du travail, l'organisation du travail, la relation individu/organisation, ou la latitude de décision. L'adoption du terme *stress* par le sens commun, fait que le savoir naïf élaboré à ce sujet biaise généralement les études faites dans les entreprises pour l'appréhender [Louche 2007].

Parmi les approches connues, nous considérons la *conception transactionnelle* du stress comme celle s'adaptant le mieux à notre modèle. Cette approche définit le stress comme une *relation particulière entre la personne et l'environnement, relation qui est évaluée par l'individu comme excédant ses ressources et menaçant son bien-être* - trad. de [Folkman 1984]. Elle considère le stress comme une réaction individuelle aux perturbations, conditionnée par les ressources dont l'individu dispose pour faire face à la situation (*coping*). Cette définition rappelle immanquablement celle de la réactance psychologique, caractérisant *un état de motivation négatif faisant suite à une menace (supposée réelle) d'une restriction de la liberté individuelle qui se traduit par une résistance à l'influence* [Brehm 1966].

Les méthodes de gestion du stress présentées dans [Folkman 1984] admettent des stratégies centrées sur le salarié et sur le soutien social. RASCLE et BRUCHON-SCHWEITZER montrent que la formation professionnelle, ainsi que la thérapie cognitive et comportementale permettent au salarié de développer les moyens nécessaires au *coping* [Rasclé 2004], [Rasclé 2006]. Le développement personnel et la psychologie du développement sont étudiés dans la dernière section.

Le travail collaboratif, une conséquence psychosociale

À notre connaissance, les modèles sur lesquels les solutions actuelles de gestion du capital humain et social sont basées, n'intègrent aucun des fondements théoriques que nous introduisons. Ils n'implémentent aucune fonction automatique d'analyse ou de gestion de la cohésion sociale, de l'identité sociale, ou du stress. Ces manques constituent un verrou, en termes de gestion du rendement, gestion des conflits et gestion de la performance¹⁰.

En accord avec [Staw 1975], [Crozier 1977] et [Louche 2007], la maximisation de la performance et l'amélioration du climat social ont pour conséquence une organisation liquide au sein de laquelle les groupes de travail se forment autour des compétences relatives aux tâches réalisées. La création et le maintien d'une telle organisation est favorisé par l'observation des modèles de socialisation organisationnelle et de gestion du stress définis dans [Tannenbaum 1967], [Folkman 1984], [Villeneuve 1995], [Rasclé 2004] et [Rasclé 2006].

10. La performance est le rapport de la valeur ajoutée sur l'effectif, tandis que le rendement représente la valeur ajoutée par l'individu.

Le *regroupement par fonction* et le *regroupement par marché*, sont les deux formes majeures de regroupement au sein de structures organisationnelles, du début de l'ère industrielle à nos jours [Louche 2007]. Les éléments de psychologie sociale présentés et les verrous identifiés à partir de ces éléments, conduisent à la définition d'une troisième forme de regroupement au sein des structures organisationnelles : le *regroupement par compétences*.

Cette forme innovante de regroupement ne s'oppose nullement au regroupement par fonction ou au regroupement par marché. Elle est simplement fondée sur une approche inverse des causalités dans laquelle l'individu, le capital humain et le capital social sont les causes de l'organisation et non ses conséquences, en accord aux travaux de STAW [Staw 1975] et de CROZIER AND FRIEDBERG [Crozier 1977]. Le regroupement par compétences est bien adapté à l'organisation liquide des groupes de travail. Son application aux structures organisationnelles est facilitée par les principes de socialisation organisationnelle de VAN MAANEN AND SCHEIN [Van Maanen 1979], MORELAND AND LEVINE [Moreland 1982] et FISHER [Fisher 1986].

Le modèle de regroupement par compétences en organisation liquide que nous introduisons est générateur de performance et de progrès social, en accord aux travaux de DEMAILLY [Demailly 2008]. Il autorise une nouvelle forme d'analyse non contingente et basée sur une approche réseaux. Cette forme d'analyse est nommée *Analyse Psychosociale des Réseaux de Compétences*.

Analyse sociographique et sémantique de réseau social

Sommaire

| | | |
|------------|---|-----------|
| 3.1 | Fondements et objectifs | 28 |
| 3.2 | Réseau de compétences, structure sémantique de réseau social | 30 |
| 3.3 | Application contextuelle | 32 |
| 3.4 | Réseau global de compétences, structure simplifiée | 33 |
| 3.5 | Mesures statiques et sémantiques | 34 |
| 3.5.1 | C_p , facteur théorique adaptatif | 35 |
| 3.5.2 | Scénario d'usage type | 37 |
| 3.5.3 | Centralité d'intermédiarité sémantique - structure | 37 |
| 3.5.4 | Centralité de flux d'intermédiarité sémantique | 39 |
| 3.5.5 | Centralité de proximité de flux sémantique | 40 |
| 3.5.6 | Modularité de communauté sémantique | 42 |
| 3.5.7 | Centralité de communauté sémantique | 44 |
| 3.6 | Analyse sociophysique et prédictive des réseaux de compétences | 44 |
| 3.7 | Récapitulatif | 46 |

Le cadre applicatif de nos travaux de thèse n'autorise pas le développement, en tant que tel, du modèle d'Analyse Psychosociale des Réseaux de Compétences précédemment introduit. De ce fait, les contributions scientifiques présentées dans ce chapitre sont centrées sur une approche graphique de l'ARS (*cf.* section 2.1), et conservent l'approche psychosociale (*cf.* section 2.2) en arrière-plan, sans la développer explicitement.

Le modèle théorique défini est multidisciplinaire, associant Ingénierie des Connaissances (IC), fouille de graphes et modèles physiques. Ils est également multidimensionnel en définissant, d'une part, un ensemble de mesures d'analyse statique et sémantique des réseaux sociaux professionnels et institutionnels, et d'autre part, une méthode prédictive bidimensionnelle, basée sur une structure sémantique de graphe social empruntant à l'IC et aux lois de l'électrodynamique. Cette méthode s'inscrivant en parallèle à la sociophysique de S. GALAM [Galambos 2008], reprend et développe l'analogie entre réseaux sociaux et réseaux de courants électriques, précédemment introduite dans [Brandes 2005] et [Newman 2005].

Les hypothèses multidisciplinaires et multidimensionnelles nécessitent généralement d'en justifier les fondements et objectifs théoriques, à défaut de pouvoir en démontrer mathématiquement la justesse dans bien des cas. Les fondements et objectifs de notre hypothèse sont exposés en section 3.1.

Une structure graphique particulière s'est rapidement imposée au cours de la conception du modèle présenté. Cette structure spécifique et évolutive a spontanément émergé de la superposition des contraintes fonctionnelles et technologiques, imposées par nos objectifs. Dans le cadre de *Socioprise*, elle se nomme *graphe de compétences* et s'accompagne d'une variante nommée *graphe global de compétences*. Elle est définie en section 3.2 et la section 3.3 en présente un exemple d'application contextuelle.

Nous présentons un principe générique permettant de transformer les mesures statiques de l'ARS en mesures sémantiques. Ce principe est applicable aux graphes sociaux et aux graphes de compétences, sans restriction par rapport aux mesures statiques d'ARS. Il est défini en section 3.5 et intégré à 5 mesures statiques (*cf.* sections 3.5.3 à 3.5.7).

3.1 Fondements et objectifs

Tandis que la sociométrie définit un réseau social à partir de liens affectifs, déduits de l'interrogation explicite de ses membres, la sociophysique utilise des méthodes et concepts issus de la physique pour décrire des comportements politiques et sociaux. La sociophysique de GALAM analyse les connaissances existantes à l'aide de méthodes empruntées à la physique statistique. Parce qu'elle compare le comportement de gens au comportement de particules et d'amas atomiques, elle n'est pas reconnue en tant que science exacte. Appliqués à la simulation de résultats de vote, les modèles sociophysiques sont cependant capables de prévoir un vote réel [Galam 2008].

Lorsqu'il intègre le concept d'inertie dans sa règle de majorité locale par un phénomène statistique, GALAM n'introduit pas pour autant $m.v^2$ (de $i = m.v^2$) dans le formalisme qu'il présente. Il conceptualise pourtant un phénomène de conservation du mouvement, constaté sur l'évolution statistique du modèle étudié et comparable au phénomène physique d'inertie. Dans ce cas, la similarité sémantique du concept d'inertie existant dans deux domaines distincts de connaissances (*i.e.*, statistique et physique) est suffisamment forte pour en autoriser le partage entre ces domaines.

A l'inverse, l'entropie¹ définie par SHANNON dans le modèle fondamental de la Théorie de l'Information [Shannon 1948], est formalisée par :

$$H = - \sum_{i=1}^n p_i \log p_i \quad (3.1)$$

En thermodynamique et à partir de la formule de BOLTZMANN, $S = k \ln \Omega$ où Ω est le nombre de micro-états distincts, l'entropie peut se formaliser en valeur

1. L'entropie exprime généralement le degré de désordre d'un système.

macroscopique par :

$$H = -K \sum_{i=1}^n p_i \ln p_i \quad (3.2)$$

Dans ce cas, sans entrer dans un niveau de détail sortant du cadre de cette thèse, le rapprochement entre la Théorie de l'Information et la thermodynamique semble évident, grâce à la similarité des formalismes mathématiques.

Nous établissons alors un constat : selon les cas, la correspondance de modèles conceptuels et de phénomènes issus de domaines distincts, peut ou ne peut pas, être démontrée de façon formelle. Le fait de savoir alors, si nous établissons ce constat par similarité conceptuelle ou par démonstration formelle constitue une méta-connaissance du constat établi, selon le raisonnement introduit dans [Pitrat 1990].

Nous ne pouvons ni démontrer, ni réfuter formellement la correspondance sémantique de termes interdisciplinaires tels que *tension* et *réactance*, alors qu'aux origines du langage, ces termes ont été définis par une même forme symbolique pour le vocabulaire commun ou scientifique - *i.e.*, tension nerveuse, tension/réactance électrique, réactance psychologique et sociale. Fait troublant, l'étymologie des termes *tension* et *réactance*, de racine latine, montre qu'ils ont été définis bien avant la reconnaissance du rôle de l'électro-chimie dans le corps humain (*e.g.*, systèmes nerveux et neuronaux) ou dans les rapports sociétaux - *e.g.* phéromones.

Lorsque, dans le contexte de nos travaux, nous avons voulu découvrir un modèle d'ARS permettant d'étudier les phénomènes de *tension* nerveuse (stress), de *réactance* et de *burn-out* au sein des réseaux sociaux professionnels, c'est probablement d'une réflexion intuitive, prônée par J. PITRAT pour la recherche de méta-connaissances, que l'évidence d'analogie entre les flux d'information, les réseaux de connaissances et les flux électriques est apparue, à l'instar des modèles sociophysiques de GALAM [Galam 2008].

L'objectif de notre approche est double. Il s'agit d'une part, de définir un modèle sémantique d'ARS capable de s'intégrer aux principaux modèles classiques de ce domaine, tels que les mesures de centralité [Freeman 1977] [Brandes 2005] et [Newman 2006]. D'autre part, il s'agit de proposer un ensemble innovant de fonctions décisionnelles basées sur l'ARS sémantique. Ces fonctions, statiques ou dynamiques, autorisent la production d'outils innovants de gestion du capital humain et social, en produisant des indicateurs visuels et des recommandations prédictives pour l'optimisation de la performance et la réduction du risque psychosocial.

Le pragmatisme tendant à reconnaître comme vrai ce qui fonctionne réellement, nous proposons de construire à la demande, pour un ensemble de mots-clés saisis, une structure de graphe synthétisant le RSEI étudié et le réseau sémantique induit par ses données endogènes. Nous nommons cette structure *réseau de compétences*.

3.2 Réseau de compétences, structure sémantique de réseau social

Un réseau de compétences est représenté par un graphe dont les nœuds correspondent aux individus et aux connaissances représentées par les termes du contenu endogène indexé. Une connaissance est nommée *Connaissance clé* lorsque le terme qui la représente sert de mot clé pour la recherche du réseau de compétences, spécifique à un ou plusieurs mots clés. Les termes dont la sémantique est associée à une connaissance clé dénotent les connaissances nommées *Connaissances relatives* à cette connaissance clé. Les associations sémantiques entre connaissances sont définies au sein d'une ontologie dont la richesse dépend du cadre applicatif choisi (*i.e.* thésaurus, ontologie légère, ontologie avec axiomes de domaines et règles). La nature des associations sémantiques entre connaissances dépend de la richesse ontologique voulue, de la relation commune de synonymie aux relations d'hyperonymie-hyponymie (*i.e.* généralisation-spécialisation), ou d'holonymie et meronymie (*i.e.*, tout et partie composante). Le nombre de niveaux de subsomption (*i.e.*, le nombre de niveaux hiérarchiques entre deux connaissances) définit la profondeur de l'ontologie.

Notre système produit, à la demande, le graphe social professionnel des individus couplés aux connaissances relatives et aux connaissances clés recherchées - *i.e.*, un *graphe de compétences*. Les communautés et individus remarquables du réseau de compétences obtenu sont ensuite identifiés à l'aide de métriques d'ARS classiques et de notre modèle dynamique prédictif.

Un graphe de compétences comporte trois types de nœuds et trois types d'arcs. Les nœuds peuvent être de type :

1. Cc , représentant un terme dénotant une connaissance clé,
2. Cr , représentant un terme dénotant une connaissance relative,
3. Pcr , représentant un individu en relation avec un ou plusieurs connaissances clés et/ou connaissances relatives,

Les arcs d'un graphe de compétences sont de type :

1. aCc , connectant deux nœuds Cc ,
2. aCr , connectant un nœud Cr à un nœud Cc , ou nœud Cr à un nœud Cr lorsque il existe plus d'un niveau de subsomption dans l'ontologie définie,
3. $aPcr$, connectant un nœud Pcr à un nœud Cc ou Cr

La figure 3.1 montre un exemple de structure de graphe de compétences. Dans l'exemple illustré par la figure, le graphe de compétences est tripartite car il n'existe qu'un seul niveau hiérarchique entre les nœuds Cc et les nœuds Cr . Ceci implique un thésaurus, ou une ontologie, restreint à un seul niveau de subsomption. Le diamètre et le nombre de parties partitionnant la structure de graphe de compétences, dépendent donc de la profondeur du thésaurus, ou de l'ontologie choisie pour représenter le domaine de connaissances relatif au contenu endogène.

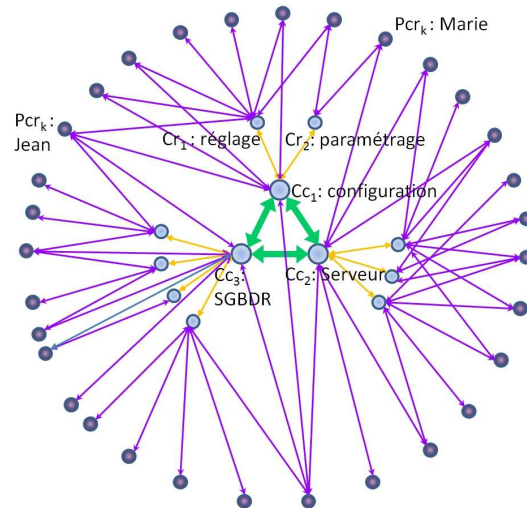


FIGURE 3.1 – Graphe de compétences : exemple pour 3 mots-clés, 9 connaissances relatives et 31 individus.

L'ensemble des arcs $acc(Cc_i, Cc_j)$ formant un circuit unique est nommé aCc . Les connaissances relatives sont connectées en étoile autour des sommets de Cc (connaissances clés) par l'ensemble d'arcs $aCr : \forall acr(Cr_j, Cc_i) \in aCr, Cr_j \in Cr, Cc_i \in Cc$.

Une personne est associée aux connaissances clés et/ou relatives présentes dans le contenu endogène avec lequel elle interagit. Ces interactions sont classifiées selon deux types de rôles, *production* (P) ou *consommation* (C). Une paire (individu, ressource) prend le rôle production P, chaque fois que l'individu écrit de l'information dans la ressource (*e.g.*, document, annotation), en relation à une étape de workflow ou à un processus métier. L'arc orienté $aPcr(Pcr, Cr|Cc)$ correspondant représente alors ce rôle.

Une paire (individu, ressource) prend le rôle consommation C, à chaque fois que l'individu est censé lire de l'information depuis la ressource, en relation à une étape de workflow ou à un processus métier. L'arc orienté $aPcr(Cr|Cc, Cc)$ correspondant représente alors ce rôle.

Dans l'exemple présenté (figure 3.1), *Marie* est associée aux compétences en *paramétrage serveur*. Son niveau de compétences dans ce domaine de connaissances peut être estimé en fonction de la quantité d'information qu'elle produit et./ou consomme, respectivement aux nœuds. *Marie* et *Jean* peuvent également posséder des compétences dans d'autres domaines de connaissances, que ceux gravitant autour des connaissances clés dénotées par les termes « SGBDR, configuration, serveur », illustrés en figure 3.1. Afin de situer globalement un individu par rapport aux connaissances présentes au sein d'un corpus, nous avons défini une variante simplifiée de la structure de graphe de compétences, nommée graphe global de compétences (*cf.* section 3.4).

Avec l'ensemble des connaissances $C : C = Cc \cup Cr$, nous formalisons l'association entre C et Pcr par un ensemble d'arcs $aPcr : \forall aPcr(Pcr_k, C_j) \in aPcr, Pcr_k \in Pcr, C_j \in C$.

Nous définissons alors $GC(N, A) : N = Cc \cup Cr \cup Pcr, A = aCc \cup aCr \cup aPcr$, le graphe de compétences défini par l'union des sommets de Cc, Cr et Pcr et l'union des arcs de aCc, aCr et $aPcr$. Un graphe de compétences $GC(N, A)$ est :

- 1 - orienté, valué et connexe (*i.e.*, tout nœud $n \in N$ est connecté par au moins un arc $a \in A$),
- 2 - bidirectionnel, car pour tout arc $a \in A, a(n_i, n_j) \Leftrightarrow a(n_j, n_i)$,
- 3 - asymétrique, car chaque arc $a \in A$ de la paire $\{a(n_i, n_j); a(n_j, n_i)\}$ porte une valeur propre.

La figure 3.2 montre une structure de graphe de compétences basée sur une ontologie à trois niveaux de subsomption.

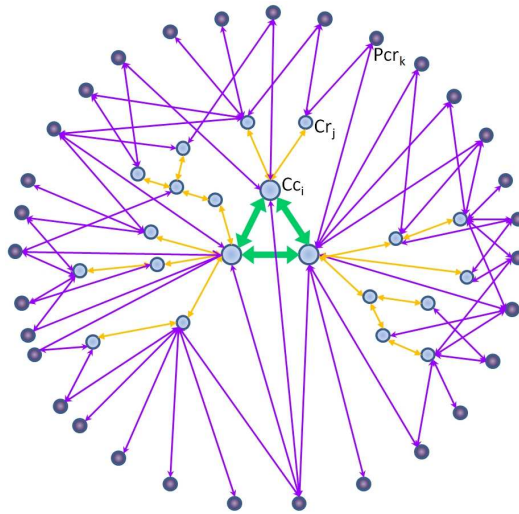


FIGURE 3.2 – Exemple de structure basée sur trois niveaux de subsomption.

3.3 Application contextuelle

Dans le contexte du projet SOCIOPRISE, considérons un exemple naïf illustré par la figure 3.1 avec les mots-clés *configuration* (Cc_1), *serveur* (Cc_2) et *SGBDR* (Cc_3)². Les connaissances sont conceptualisées par une spécification formelle et explicite, comme une ontologie ou un thésaurus. Dans le domaine de connaissances où s'inscrit le terme *configuration*, il existe deux termes associés : *réglage* et *paramétrage* - Cr_1 et Cr_2 sur la figure. *Serveur* porte trois termes associés, par exemple *JBoss*, *Debian* et *Windows 2003*. *Base de données* porte quatre termes associés, par exemple *Base de données*, *MySQL*, *Oracle* et *SQL Server*.

2. Le graphe de compétences réellement obtenu pour ces mots clés comprend environ 1600 arcs et ne serait pas adéquat pour une telle illustration.

Un individu $Pcr_k..Pcr_{k'}$ est présent dans le graphe recherché dès lors qu'il est en relation avec au moins un enregistrement de la base de données contenant un ou plusieurs des termes $Cc_{i..i'}$ et/ou $Cr_{j..j'}$. En fonction de la nature de la relation impliquant un enregistrement et un individu, l'un des deux types de rôle lui est assigné - *i.e.*, producteur (P) *vs.* consommateur (C). Les rôles P et C définissent l'orientation des arcs connectant un individu aux connaissances. De plus, un individu peut, pour une même connaissance, être à la fois producteur et consommateur, dans une période de temps donnée. Enfin, pour chaque triplet (*terme, individu, rôle*), une valeur de poids sémantique est calculée, en fonction du nombre de ressources impliquées et du poids sémantique du terme dans ces ressources. Ce poids sémantique, quantifiant la relation individu-connaissance en fonction du rôle (P, C), est calculé à l'aide de métriques classiques d'analyse linguistique [Robertson 1976], [Rogers 1960]. Ces métriques et le processus d'analyse linguistique et d'indexation sémantique sont présentées en section 4.1.

3.4 Réseau global de compétences, structure simplifiée

En vue de disposer de valeurs de référence pour la comparaison temporelle des résultats fournis par les mesures statiques et sémantiques d'ARS, nous définissons une variante de la structure de réseau de compétences permettant de réunir l'ensemble des données disponibles au sein d'un grand graphe unique.

Le réseau global de compétences est constitué à partir de l'indexation sémantique du contenu endogène (1), de la relation individu/contenu disponible dans les données sources (2), et du poids sémantique (*e.g.*, TFIDF) des termes du contenu endogène indexé, pour la relation individu/contenu (3). Un réseau global de compétences peut être comparé à un réseau de compétences dont la structure ne comporte aucune différenciation entre connaissances clés et connaissances relatives, à la différence de la structure de réseau de compétences. C'est un graphe biparti, représentant la relation pondérée entre tous les individus associés au contenu endogène et tous les termes présents dans l'index linguistique du contenu endogène. La figure 3.3 montre une structure de graphe global de compétences, à comparer aux figures 3.1 et 3.2.

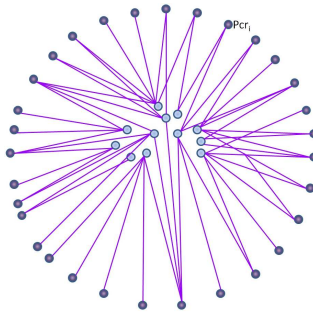


FIGURE 3.3 – Exemple de structure de graphe global de compétences pour comparaison.

dans la dimension sémantique et par les relations ontologiques de ce terme, le courriel associé à l'individu u est également associé aux ressources documentaires comportant des termes ou annotations similaires ou proches (*i.e.*, tuteur). Les individus associés à ces ressources documentaires sont alors associés à l'individu u , grâce au domaine de connaissance partagé. Les trois individus pointés par les associations in-

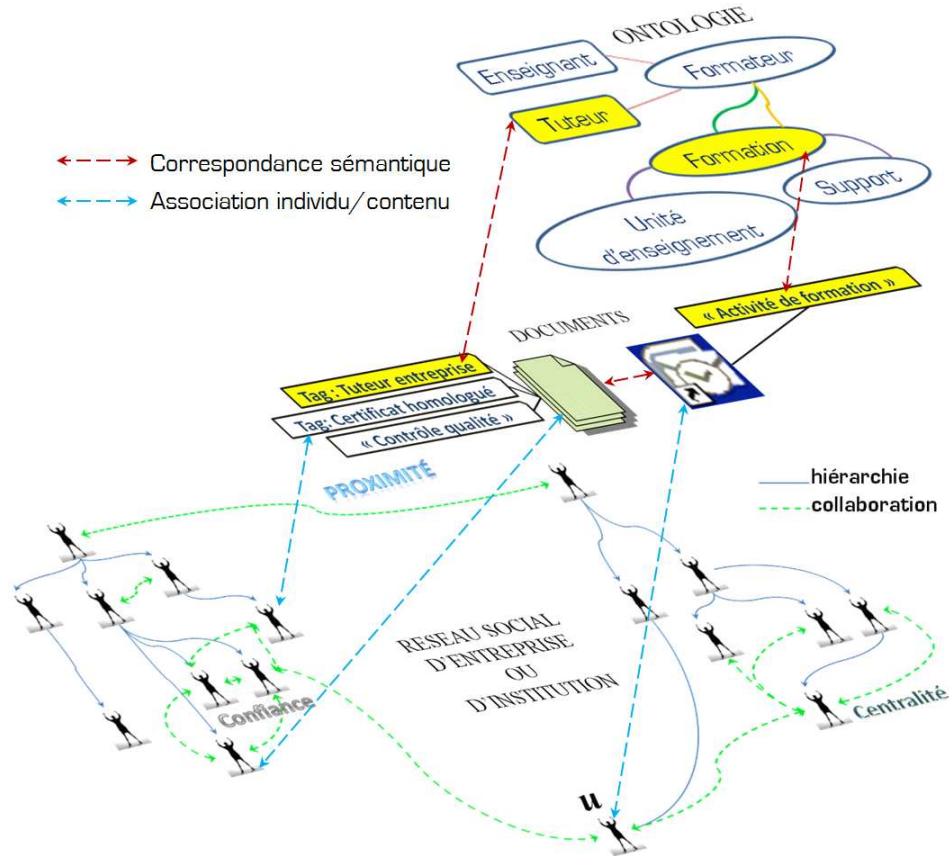


FIGURE 3.5 – Modèle multidimensionnel d'analyse statique des RSEI.

dividu/contenu, dans la figure 3.5, deviennent ainsi des nœuds d'un même graphe de compétences. De plus, les autres expressions indexant les ressources documentaires associées dans la figure (*i.e.*, certificat homologué, contrôle qualité), peuvent être utilisées pour la découverte de connaissances endogènes, par indexation sémantique et traitement automatique du langage naturel.

3.5.1 C_p , facteur théorique adaptatif

Dans le cadre du projet SOCIOPRISE, nous introduisons *Pregsem*, une fonction théorique, couvrant les différentes algèbres identifiées pour le développement de prototypes industriels utilisant des mesures sémantiques et statiques d'ARS. Elle permet de combiner des métriques de prégnance linguistique ou similarité sémantique

(e.g., TFIDF, Jaccard), avec des opérateurs algébriques simples tels que somme ou moyenne, afin de fournir un outil générique pour la valuation sémantique de graphes sociaux. *Pregsem* concourt à la définition d'un facteur sémantique, nommé C_p . Ce facteur autorise la réutilisation de mesures standards d'ARS, qui ne tiennent pas intrinsèquement compte du contenu endogène, et la définition de nouvelles mesures sémantiques d'ARS, dont les résultats dépendent du partage des connaissances présentes au sein du contenu endogène et relatif aux réseaux sociaux analysés. Nous définissons ainsi un moyen générique de transformer des mesures standards d'ARS en mesures sémantiques d'ARS. C_p se compose de deux fonctions, $f1$ et $f2$.

La fonction $f1$ est formalisée comme suit :

$$f(n, C,) \rightarrow R(C', Sim_{C'n}, Prox_{C'n}) \quad (3.3)$$

Elle produit la relation $R(C', Sim_{C'n}, Sim_{C'n}, Preg_{C'n}, Preg_{C'n})$ entre un nœud n et les connaissances endogènes C' , correspondant à l'ensemble de mots-clés C , avec $C \subset C'$ et $C' \setminus C$ représentant les connaissances définies dans le thésaurus, ou dans une ontologie comme illustré en figure 3.5. $Sim_{C'n}, Sim_{C'n}, Preg_{C'n}, Preg_{C'n}$ sont les valeurs de similarité et prégnance sémantiques, produites ou consommées par le nœud n , pour un élément de C' .

La fonction $f2$ est formalisée par :

$$Pregsem(f(n, C), Opc, Opa) \quad (3.4)$$

Elle produit une combinaison agrégée des valeurs de similarité et/ou prégnance sémantique résultant de $f(n, C)$. Le type de combinaison dépend de la valeur de l'opérateur *Opc* (i.e., $Sim_{C'n}$ et/ou $Prox_{C'n}$). Le type d'agrégation est défini par l'opérateur arithmétique *Opa* (i.e., somme, moyenne, etc.).

(3) Le facteur C_p est finalement défini comme suit :

$$C_p = Pregsem(f(n, C), Opc, Opa) \quad (3.5)$$

A titre d'exemple, nous utilisons $C_p = Pregsem(f(n, C), Sim, Somme)$, pour obtenir la relation sémantique entre n et le mot clé C , incluant la somme des valeurs de similarité sémantique, produites ou consommées, concernant n , pour les ressources indexées par C et ses connaissances relatives Cr (définies dans le thésaurus/ontologie). C_p retourne alors l'ensemble des arcs valués entre n , C et Cr . Pour chaque arc, la somme des valeurs de similarité, produites ou consommées par n , est issue de l'intégralité du contenu endogène. $C_p = Pregsem(f(n, C), Preg, Avg)$ retournerait une moyenne des valeurs de prégnance, au lieu d'une somme des valeurs de similarité pour chaque arc. $C_p = Pregsem(f(n, C), Preg_and_Sim, Avg)$ retournerait une moyenne des valeurs de prégnance et de similarité (produites ou consommées), pour chaque arc. Enfin, $C_p = Pregsem(f(n, C), Pregp_minus_Pregc, Sum)$ fournirait une somme de la différence entre valeurs de prégnance produites et consommées, pour chaque arc.

Le facteur C_p étant facilement intégrable à la plupart des mesures standards d'ARS, il constitue une méthode générique d'extension des mesures statiques d'ARS

en mesures sémantiques. Dans le cadre de nos travaux, nous avons identifié diverses métriques statiques pertinentes pour l'analyse sémantique des graphes sociaux en utilisant notre structure de graphe de compétences. Pour 5 de ces métriques, nous appliquons leurs fonctions respectives dans le cadre du projet SOCIOPRISE.

3.5.2 Scénario d'usage type

Afin de faciliter la compréhension du modèle, nous avons simulé le scénario d'usage suivant en s'appuyant sur les données réelles étudiées dans le cadre de SOCIOPRISE :

Le cycle de vie d'un projet de développement logiciel implique des connaissances connues, identifiées en tant que Connaissances clés. Ces Connaissances clés sont en relation, au sein d'un référentiel sémantique (*e.g.*, thésaurus, ontologie), avec des compétences techniques ou métiers, nommées Connaissances relatives.

(a) Les Connaissances clés sont identifiées par les expressions *DotNet 3.5*, *Visual Studio for Web Development*, *SQL Server 2008 R2*, et *gestion de la formation professionnelle continue*.

(b) Pour *DotNet 3.5*, les Connaissances relatives indexées sont *Framework .Net 3.5* et *ASP.NET*.

(c) Pour *Visual Studio for Web Development*, les Connaissances relatives sont *VSWD* et *Visual Studio pour Application Mobile*.

(d) Pour *SQL Server 2008 R2*, les Connaissances relatives sont *MS SQL Server*, *ADO.NET*, *Bases de données* et *Oracle*.

(e) Enfin, pour *gestion de la formation professionnelle continue*, les Connaissances relatives sont *FPC*, *formation continue* et *management de la formation*.

Le management et les équipes de production travaillent avec le graphe de ces compétences, pour la prise de décision et l'analyse décisionnelle basées sur le contenu et l'annuaire d'entreprise disponibles.

3.5.3 Centralité d'intermédiarité sémantique - structure

La définition de centralité d'intermédiarité de [Freeman 1979] est basée sur le nombre de chemins géodésiques entre toutes les paires de sommets (i, j) , passant par un sommet u . Dans un graphe de compétences $GC(N, A)$, avec N ensemble de nœuds et A ensemble d'arcs, ou dans un graphe quelconque $G(V, E)$ (avec V ensemble de sommets et E , ensemble des arêtes du graphe), nous étendons l'équation (2.1) par l'équation (3.6).

$$I_{C_p}(u) = \sum_{i,j} \frac{\sigma(i, u, j)}{\sigma(i, j)} \times C_p \quad (3.6)$$

La figure 3.6 résume les résultats produits par la centralité d'intermédiarité basée sur les chemins géodésiques.

Dans l'équation (3.6), la valeur d'intermédiarité est soumise au facteur C_p . La valeur de C_p est déterminée en fonction du contenu endogène et des arguments

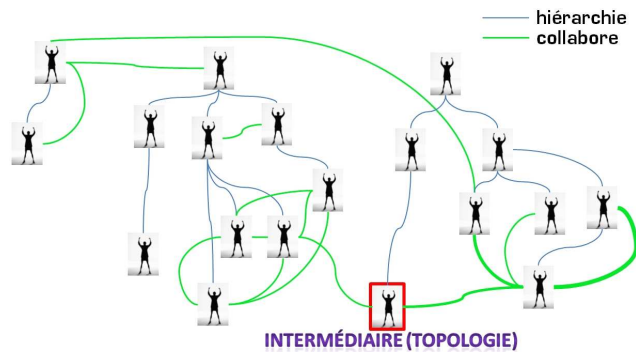


FIGURE 3.6 – Illustration de l’intermédiation de structure (topologique).

fournis à *Pregsem*, identiques pour chaque u . Elle est généralement élevée et peut être réduite sur un intervalle $[0; 1]$.

La valeur d’intermédiation sémantique d’un sommet u augmente en fonction de son intermédiation géodésique au sein du graphe, et de son implication dans la production et/ou la consommation de contenu endogène relatif aux connaissances clés fournies à *Pregsem*. Dans le contexte d’un RSEI, la mesure est donc susceptible d’indiquer :

1. une représentation ordonnée des spécialistes et/ou polyvalents d’une ou plusieurs compétences désignées ;
2. la reconnaissance de personnes centrales pour le partage d’information ou la collaboration relative à une ou plusieurs compétences ;
3. une représentation du réseau des leaders possibles pour un projet défini - Chefs de Projet (CP), Comités ;
4. l’identification de compétences manquantes dans une équipe projet - *i.e.*, lorsque $C_p = 0$.

Le classement produit dépendant surtout de la topologie du graphe, il peut être pertinent d’employer C_p avec une mesure d’intermédiation basée sur les flux (*e.g.*, mesure basée sur le flux électrique) telle que définie dans [Brandes 2005].

Exemple :

Pour une compétence donnée et l’une de ses compétences relatives - *Visual Studio for Web Development* - l’utilisateur identifie la communauté active de cette compétence, dans laquelle les polyvalents (*i.e.*, les individus de poids fort appartenant aux communautés *DotNet 3.5* et *SQL Server 2008 R2* et/ou *gestion de la formation professionnelle continue*), sont repérés par une interconnexion forte entre communautés de compétences distinctes, à l’opposé des spécialistes (*i.e.*, les individus de poids fort de la communauté *DotNet 3.5*). Pour l’ensemble des compétences données, les personnes fortement intermédiaires peuvent constituer un réseau de leaders dans l’équipe projet, représenter le projet (Comité) ou le diriger (CP), selon les compétences qu’ils possèdent - *e.g.*, compétences techniques, métiers, organisationnelles.

Les compétences manquantes sont révélées par les graphes de compétences sans individus connectés aux connaissances clés et/ou à leurs connaissances relatives, ou par les sommets portant une intermédiarité sémantique égale à 0, dans un graphe social quelconque.

3.5.4 Centralité de flux d'intermédiarité sémantique

La définition de centralité de flux d'intermédiarité de [Brandes 2005] est proche de la centralité d'intermédiarité de [Freeman 1979]. Elle est basée sur le flux minimal, au lieu du chemin géodésique, entre toutes les paires de sommets (i, j) , et ne s'appliquent donc que sur un graphe valué³. Lorsque les arêtes ou les arcs d'un graphe portent des valeurs différentes, le chemin géodésique ne correspond pas systématiquement au flux minimal, ou maximal, d'information circulant entre les sommets ou nœuds du graphe. Dans [Brandes 2005], l'auteur pallie à ce problème en définissant une centralité de flux d'intermédiarité, basée sur une analogie entre chaque arête/arcs d'un graphe et une résistance électrique. Cette analogie est également développée dans [Newman 2005]. Nous formalisons la centralité de flux d'intermédiarité sémantique de la même façon que dans l'équation (3.6), comme suit :

$$I_{C_p}(u) = \sum_{ij} \frac{\sigma(i, u, j)}{\sigma(i, j)} \times C_p \quad (3.7)$$

La figure 3.7 résume les résultats produits par la centralité d'intermédiarité de flux.

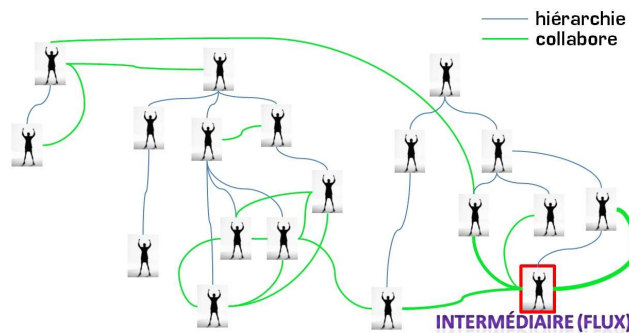


FIGURE 3.7 – Illustration de l'intermédiarité flux.

Dans l'équation (3.7), $\sigma(i, j)$ est la somme du flux minimal circulant entre chaque paire de sommets (i, j) , pour chaque paire de sommets du graphe analysé. $\sigma(i, u, j)$ est la somme de flux minimal circulant entre chaque paire de sommets (i, j) passant par u . Le mesure peut être implémentée à l'aide d'algorithmes de recherche de flux minimal ou de marche aléatoire, tels que définis dans [Brandes 2001], ou [Brandes 2005] et [Newman 2005].

3. La centralité d'intermédiarité de flux produit des résultats similaires à ceux de la centralité d'intermédiarité lorsque les arêtes d'un graphe portent une valeur égale à 1.

Comparée à la centralité d'intermédiarité sémantique (section 3.5.3) la centralité d'intermédiarité de flux sémantique n'est pas aussi dépendante de la topologie du graphe analysé. Dans le contexte d'un graphe valué quelconque, les différentes indications produites par les deux versions de centralité restent comparables, même si le classement des sommets qu'elles produisent est sensiblement différent, en fonction de la valeur des arcs du graphe et/ou des algorithmes de parcours employés⁴ - *e.g.*, flux minimal *vs.*, marche aléatoire. La centralité d'intermédiarité de flux sémantique permet d'exploiter les diverses combinaisons d'arguments de C_p - *e.g.*, recherche de flux minimal basée sur la valeur produite moins la valeur consommée entre paires de nœuds.

Comparée à la centralité d'intermédiarité sémantique, la centralité d'intermédiarité de flux sémantique produit, par définition, des indications plus pertinentes concernant :

1. la représentation ordonnée des spécialistes et/ou polyvalents d'une ou plusieurs compétences désignées ;
2. la reconnaissance de personnes centrales pour le partage d'information ou la collaboration relative à une ou plusieurs compétences ;
3. la représentation du réseau des leaders possibles pour un projet défini - Chefs de Projet (CP), Comités ;
4. l'identification de compétences manquantes dans une équipe projet.

Exemple :

Pour une compétence donnée et l'une de ses compétences relatives - *Visual Studio for Web Development* - l'utilisateur identifie la communauté active de cette compétence, dans laquelle les polyvalents (*i.e.*, les individus de poids fort appartenant aux communautés *DotNet 3.5* et *SQL Server 2008 R2* et/ou *gestion de la formation professionnelle continue*), sont repérés par une interconnexion forte entre communautés de compétences distinctes, à l'opposé des spécialistes (*i.e.*, les individus de poids fort de la communauté *DotNet 3.5*). Pour l'ensemble des compétences données, les personnes fortement intermédiaires peuvent constituer un réseau de leaders dans l'équipe projet, représenter le projet (Comité) ou le diriger (CP), selon les compétences qu'ils possèdent - *e.g.*, compétences techniques, métiers, organisationnelles. Les compétences manquantes sont révélées par les graphes de compétences sans individus connectés aux connaissances clés et/ou à leurs connaissances relatives, ou par les sommets portant une intermédiarité sémantique égale à 0, dans un graphe social quelconque.

3.5.5 Centralité de proximité de flux sémantique

La centralité de proximité définie dans [Beauchamp Murray 1965] et basée sur les chemins géodésiques, avec $d(p_u, p_i)$, distance entre deux sommets et n , nombre

4. la valeur de classement dépend de la valeur de flux minimal entre les paires de sommets du graphe étudié, puis de la topologie du graphe.

de sommets dans un graphe $G(V, E)$. Elle est formalisée comme suit.

$$C_{C_p}(u) = \frac{n - 1}{\sum_{i=1}^n d(p_u, p_i)} \quad (3.8)$$

Divers travaux ont introduit plusieurs définitions de la centralité de proximité, celle de [Brandes 2005] étant basée sur les flux de courants au lieu des plus courts chemins géodésiques, et comparant la distance graphique à la résistance électrique. Nous introduisons une nouvelle centralité de proximité de flux sémantique, basée sur [Beauchamp Murray 1965] et [Brandes 2005]. Cette mesure sémantique d'ARS est formalisée par l'équation (3.9), avec $a(p_u, p_i)$, poids des arcs d'un plus court chemin (p_u, p_i) - positif ou négatif. Nous employons $a(p_i, p_u)$, pour la centralité entrante et $a(p_u, p_i)$, pour la centralité sortante (graphes orientés).

$$C_{C_p}^f(u) = \frac{n - 1}{\sum_{i=1}^n a(p_u, p_i)} \times C_p \quad (3.9)$$

La figure 3.8 résume la différence de résultats produits par la centralité de proximité entre deux nœuds (encadrés en rouge dans l'illustration), selon qu'elle est basée sur la topologie du graphe ou sur les flux.

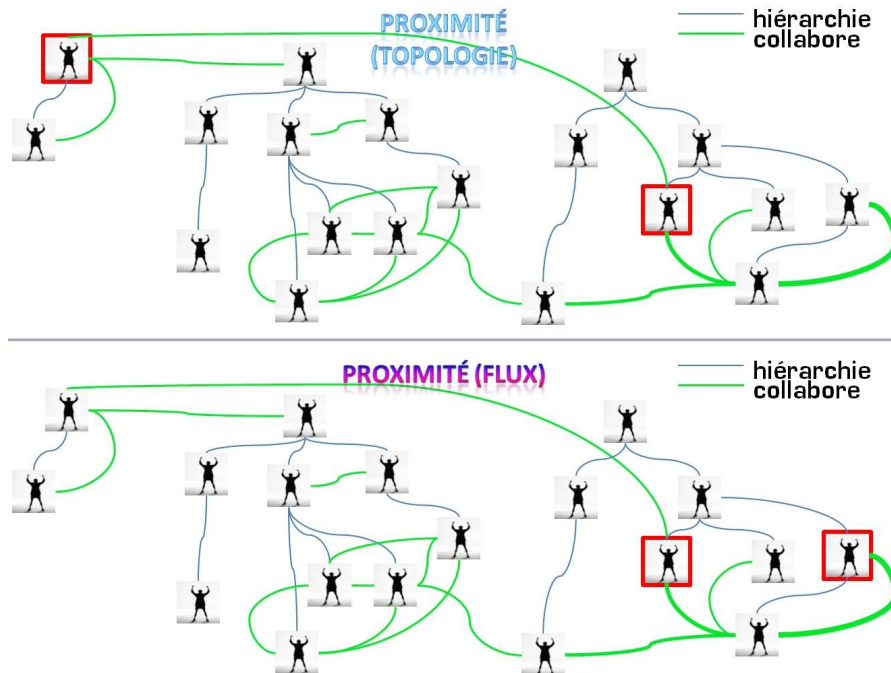


FIGURE 3.8 – Comparaison de centralités de proximité - structure *vs.* flux.

La mesure est basée sur le plus court chemin (u, i) , flux minimal représenté par les arcs pondérés d'extrémités u et i . La centralité de proximité de flux sémantique de u dépend de la topologie du graphe, de la pondération sémantique des arcs, et de l'implication de u dans la production et/ou consommation de

contenu endogène. Appliquée sur un réseau de compétences ou un réseau global de compétences, la valeur d'un nœud u augmente en fonction (1) du poids sémantique de u , et (2) de l'implication de u dans la production et/ou consommation de contenu endogène, relativement aux arguments passés à *Pregsem* - *i.e.*, $Pregsem(f(n, C), Pregp_minus_Pregc, Sum) * Pregp(n)$, pour C_p calculant la différence de valeur *TF.IDF* produite moins consommé entre deux nœuds, multipliée par la valeur *TF.IDF* produite de u . Dans le contexte d'un RSEI, la mesure fournit un ensemble intéressant d'indications, telles que :

1. représentation ordonnée des spécialistes et polyvalents d'une ou plusieurs compétences, en fonction du niveau de possession global ou respectif de chaque compétence ;
2. identification d'un réseau de leadership classé pour un projet connu (CP, Comité directeur) ;
3. représentation hiérarchique des compétences relatives à une communauté, identification de compétences en déficit ou excédentaires au sein d'un équipe projet ;
4. identification des consommateurs, producteurs et diffuseurs d'information relativement à une ou plusieurs compétences dans le graphe social.

Exemple :

Pour une compétence donnée ou l'une de ses compétences relatives (*e.g.*, *DotNet 3.5*), la métrique identifie les personnes possédant la compétence et les classe par niveau d'expertise. Avec un ensemble de compétences désignées, il est donc possible de visualiser l'adéquation d'une communauté avec l'expertise demandée pour un projet ou un plan stratégique, d'anticiper les besoins en formation comme d'identifier les talents, et de prévenir le risque de dégradation du niveau de performance ou de sur-qualité.

3.5.6 Modularité de communauté sémantique

En se basant sur [Clauset 2004], [Wakita 2007] ou [Blondel 2008], pour intégrer notre facteur C_p , la mesure pourrait être formalisée de différentes manières, à l'aide d'algorithmes de *clustering* tels que CNM (*i.e.*, Clauset, Newman, Moore). L'algorithme CNM infère les structures de communautés de la topologie d'un réseau [Clauset 2004]. Il se caractérise par une exécution rapide, mais semble limité lorsque le graphe analysé dépasse 500 000 nœuds, selon la profondeur du dendrogramme produit en sortie. Dans [Wakita 2007], les auteurs introduisent des raffinements de CNM, basés sur une approche équilibrée des calculs de fusion, grâce à un ratio de consolidation qui en réduit fortement la complexité organisationnelle. La méthode optimisée de détection de communautés définie dans [Blondel 2008], s'exécute également sur un graphe issu du web de 118 millions de nœuds et plus d'un milliard d'arêtes.

Notre structure hybride de graphe de compétences est constituée d'un réseau sémantique, le cœur de la structure, et d'un réseau social connecté à ce cœur. Les struc-

3.5.7 Centralité de communauté sémantique

Pour appliquer notre modèle sur un graphe social, nous devons intégrer le facteur C_p à la magnitude $|X_k|$, selon [Newman 2006]. Comme pour la modularité de communauté sémantique, nous employons la structure hybride de graphe de compétences pour obtenir une mesure pertinente de centralité de communauté sémantique, par simple application de la mesure standard sur la structure hybride. La mesure produit des indications sensiblement différentes de celles obtenues avec la modularité de communauté sémantique. Elle facilite l'identification de compétences minoritaires et majoritaires et la représentation ordonnées des spécialistes/polyvalents relatifs aux compétences minoritaires et majoritaires. La visualisation de résultats produits par la centralité de communauté sémantique est proche de celle présentée en figure 3.9.

Exemple :

L'utilisateur visualise les compétences les plus fortes/faibles et les communautés associées, au sein du réseau de compétences relatif aux mots-clés saisis - *i.e.*, *DotNet 3.5*, *Visual Studio for Web Development*, *SQL Server 2008 R2* et *gestion de la formation professionnelle continue*. Dans chaque communauté, les spécialistes d'une connaissance clé et de ses connaissances relatives (*i.e.*, un champ de connaissances) sont mis en avant, et inversement pour les polyvalents qui sont connectés à plusieurs champs de connaissances. Spécialistes et polyvalents sont ordonnés par niveaux de possession de leurs compétences respectives.

3.6 Analyse sociophysique et prédictive des réseaux de compétences

L'expérience d'E. BRANLY sur la radio-conduction (1890) a démontré que des flux peuvent circuler entre des points dépourvus de connections spécifiques. Sur le principe de la radio-conduction, nous émettons l'hypothèse que dans les graphes sociaux, il existe des flux d'information significatifs et invisibles entre des individus reliés ou non entre eux par les arêtes du graphe social. Nous étudions cette hypothèse en comparant les arcs du graphe de compétences à des conducteurs transportant des flux électriques, dans l'objectif de représenter les flux invisibles au sein du graphe et de quantifier la *réactance* conditionnant l'évolution de la structure du réseau social.

Dans le contexte défini pour le projet SOCIOPRISE, un intervenant consomme et produit de l'information en fonction de ses rôles durant le cycle de vie d'un workflow. Notre méthode consiste à établir un parallèle entre des valeurs électriques - *i.e.*, tension, résistance, intensité, puissance - et les valeurs du graphe de compétences - *i.e.*, des ratios sémantiques calculés à partir des poids attribués au contenu par le système lexico-sémantique d'indexation et de recherche d'information (*cf.* section 4.1).

Dans le contexte de SOCIOPRISE et au sein d'un graphe de compétences, nous considérons que la tension au travail par rapport à un terme augmente avec (1) l'intensité d'utilisation de ce terme, et (2) la rareté de ce terme au sein du corpus

d'entreprise. L'intensité électrique I exprime le débit de charge au sein d'un corps plus ou moins conducteur. Nous assimilons la sémantique du taux de fréquence TF à celle de l'intensité I et introduisons une métrique d'intensité sémantique Is .

La rareté d'un terme est classiquement donnée en linguistique par sa fréquence inverse IDF . La tension au travail augmentant avec l'intensité sémantique $Is \equiv TF$ et avec la rareté d'un terme $IDF(terme)$ au sein du corpus d'entreprise, nous introduisons l'équivalence $TF.IDF \equiv Is.R : R \equiv IDF$, définissant une métrique de résistance sémantique R . Puisque selon la loi d'OHM, $U = I.R$ (*i.e.*, Tension = Intensité.Résistance), en rapprochant la sémantique de $TF.IDF$ et celle de $I.R$, nous définissons une équivalence sémantique entre tension au travail et tension électrique, au sein des graphes de compétences. Cette équivalence introduit une métrique de tension sémantique Ts , et se formalise comme suit :

$$TF.IDF \equiv Is.R \equiv I.R = U \equiv Ts \quad (3.10)$$

Dans notre modèle, les flux d'information concernant un individu sont définis en tant que poids sémantiques, grâce aux métriques standards d'analyse linguistiques intégrées (*i.e.*, TF-IDF, *Jaccard*). Un graphe de compétences est intégralement valué par des poids sémantiques, de production ou consommation pour chaque élément du graphe. Les valeurs sémantiques $TF.IDF$ représentent des valeurs pseudo-électriques de tension. Nous utilisons également les valeurs temporelles (dates, heures, timestamps) liées aux informations et aux événements enregistrés dans la base de données pour calculer les valeurs pseudo-électriques de puissance, résistance, tension et intensité moyenne de chaque nœud et arc du graphe, en appliquant les lois d'OHM :

$U = R.I$ et $P = R.I^2$ U représente la tension électrique dépendante de la résistance R et de l'intensité I , et P représente la puissance délivrée par un sommet, de puissance maximum admissible notée $Pmax$, avec $Umax = \sqrt{R.Pmax}$ et $Imax = \sqrt{Pmax/R}$.

Notre modèle développe une analogie plus complète entre les flux des réseaux et les courants, que ceux précédemment définis dans [Thomassen 1990], [Brandes 2005] ou [Newman 2005]. L'originalité de nos travaux consiste à introduire une métrique de *tension/stress* du réseau social en relation aux notions d'*intensité* de flux traversant et de *résistance*, en appliquant les lois d'OHM sur le graphe de compétences. En introduisant un processus de valuation sémantique, le modèle permet d'estimer des valeurs de *tension/stress* pour chaque individu du réseau, un indicateur utile et doté d'un faible biais comparé aux habituels questionnaires d'enquêtes.

Finalement, l'étude de l'analogie que nous développons pourra être poursuivie en recourant à l'effet JOULE pour estimer l'*échauffement* des composants du réseau social d'entreprise et pour prévenir les risques de dégradation des performances, d'instabilité, de *burn-out* et de troubles psychosociaux. Par exemple, la *stress* pourra être comparé à une surcharge relative aux personnes travaillant hors de leur domaine de connaissance, alors que la *réactance* pourra apparaître avec la surcharge de personnes expertes dans le domaine de connaissance ou elles sont sollicitées.

3.7 Récapitulatif

Les contributions définies sont volontairement centrées sur une approche graphique de l'ARS (*cf.* section 2.1) qui place provisoirement l'approche psychosociale (*cf.* section 2.2) en arrière-plan.

Nous avons défini un principe générique d'ARS sémantique, rapprochant des méthodes d'IC telles que l'analyse linguistique et l'ingénierie des ontologies, et l'ARS statique. Ce principe est adaptable aux standards de l'ARS statique et a été intégré à cinq exemples de centralités et modularité.

Le développement d'un modèle théorique d'ARS sémantique a conduit à la définition d'une structure graphique sémantique et évolutive, le graphe de compétences. Les réseaux de compétences rendent possible le développement et l'application d'une analogie plus complète entre les flux des réseaux et les courants électriques, que celles précédemment définies dans [Thomassen 1990], [Brandes 2005] ou [Newman 2005]. Cette analogie permet l'introduction d'une métrique de *tension/stress*, basée sur la quantification sémantique du contenu endogène échangé au sein d'un réseau social professionnel.

Il résulte de ces contributions un modèle multidisciplinaire et bidimensionnel d'ARS dédié à l'optimisation de la performance et à la réduction du risque psychosocial dans les groupes de travail. Son implémentation en vue d'expérimentation est présentée au chapitre suivant - Chapitre 4.

Implémentation de réseaux de compétences

Sommaire

| | |
|--|-----------|
| 4.1 Recherche lexico-sémantique d'information dans le contenu endogène | 48 |
| 4.1.1 Analyse lexicale et indexation du texte intégral | 50 |
| 4.1.2 Construction de thésaurus et pondération sémantique | 50 |
| 4.1.3 Développement et codage | 53 |
| 4.2 Réseaux de compétences à la demande | 53 |
| 4.2.1 Étape 1 : sélection et pondération des arcs de $aPcr$ | 55 |
| 4.2.2 Étape 2 : sélection et valuation de l'intensité des nœuds | 56 |
| 4.2.3 Propagation des valeurs sémantiques de $aPcr$ | 57 |
| 4.2.4 Étape 9 : transformation des valeurs sémantiques de Cc_i | 60 |
| 4.2.5 Étape 10 : équilibrage de pseudo-intensités au cœur d'un graphe de compétences | 61 |
| 4.3 Valuation pseudo-électrique du graphe de compétences | 62 |
| 4.3.1 Valuation pseudo-électriques d'arcs et d'extrémités à partir de connaissances-clés | 63 |
| 4.3.2 Rétro-propagation de pseudo-valeurs électriques depuis Cc_1 | 65 |
| 4.3.3 Modélisation et vérification - solveur de circuit électrique | 72 |
| 4.4 Un modèle de recommandations basé sur l'équilibre énergétique | 75 |

Le modèle multidisciplinaire et bidimensionnel d'ARS défini précédemment a été implémenté en collaboration avec *OpenPortal Software*, puis expérimenté à partir de données réelles, dans le cadre du projet SOCIOPRISE. Ce chapitre décrit le contexte et le processus de construction de réseaux de compétences à la demande et de calcul de recommandations basées sur l'équilibre naturel énergétique.

La section 4.1 présente l'architecture, les données sources utilisées, et les méthodes d'analyse linguistique, d'indexation et de recherche sémantique d'information servant de socle pour l'implémentation de réseaux de compétences. La section 4.2 présente le sous-processus de construction de réseaux de compétences à la demande. La section 4.3 définit et présente l'implémentation du procédé de valuation dynamique qui permet de corrélérer la dimension sémantique et la dimension pseudo-électrique, en adéquation au sein d'un graphe de compétences. Enfin, la section 4.4

présente le modèle de recommandations implémenté, basé sur l'équilibre énergétique naturel.

4.1 Recherche lexico-sémantique d'information dans le contenu endogène

Notre processus d'identification et d'analyse de réseaux de compétences à partir de mots-clés repose sur un système d'indexation lexico-sémantique du contenu d'entreprise associé aux individus - *i.e.*, les ressources produites et consommées par les membres de l'entreprise utilisateurs d'un environnement logiciel collaboratif, constituant le contenu endogène.

Les corpus d'entreprises étant généralement conséquents, nous rejetons l'hypothèse d'une indexation sémantique fondée sur des ontologies métiers construites par les experts du domaine. Cette approche chronophage, nécessitant une adaptation permanente au contexte, ne peut correspondre aux objectifs industriels et commerciaux visés à terme dans le cadre du projet *Socioprise*. Le processus adopté intègre les standards industriels courants et permet les mises-à-jour automatiques planifiées. L'architecture générale mise en œuvre agrège un analyseur lexical, un thésaurus et des services multilingues d'indexation et de recherche en texte intégral - *cf.* figure 4.1.

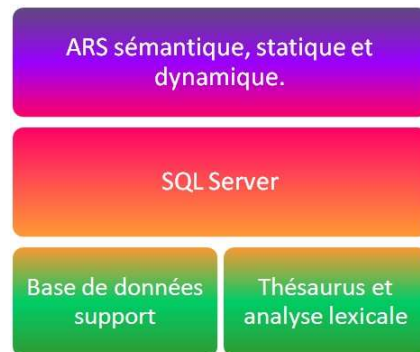


FIGURE 4.1 – Schéma simplifié de l'architecture système utilisée.

Le corpus expérimenté provient d'une base de données réelles, propre au projet SOCIOPRISE. Cette base de données stocke les enregistrements d'une application de *workflows*. Il s'agit de données représentatives, utilisées dans le cadre d'un processus métier réel. Nous disposons de divers états de la base de données, en fonction de dates données. Un état collecté en 2010 comporte environ 250 000 étapes commentées, concernant 33 000 entrées - *i.e.*, des actions initiales d'entrée dans le workflow. Ces entrées proviennent de 130 organisations, représentées par 540 personnes. Elles sont traitées par 80 acteurs, répartis en trois catégories et se partageant les tâches en fonction de leurs compétences personnelles. Chaque entrée comporte diverses

méta-données caractérisant le processus (*e.g.*, dates, actes). Chaque étape dans l'historique d'une entrée comporte des commentaires (texte), un état et diverses autres méta-données - non communiquées (confidentielles).

La figure 4.2 présente un aperçu simplifié du schéma relationnel des données étudiées, dans lequel les entrées et leurs étapes sont enregistrées dans les tables nommées *Message* et *MessageHistorique*. Les organisations, les personnes et les acteurs du processus métier sont enregistrés dans les tables *Client* et *Personne*. Les états des étapes du processus (table *Etat*), ainsi que les relations et les attributs *Contact*, *Développeur*, *Client*, *ExpertMetier* de la table *Message*, permettent la définition de rôles *Producteur* et *Consommateur*.

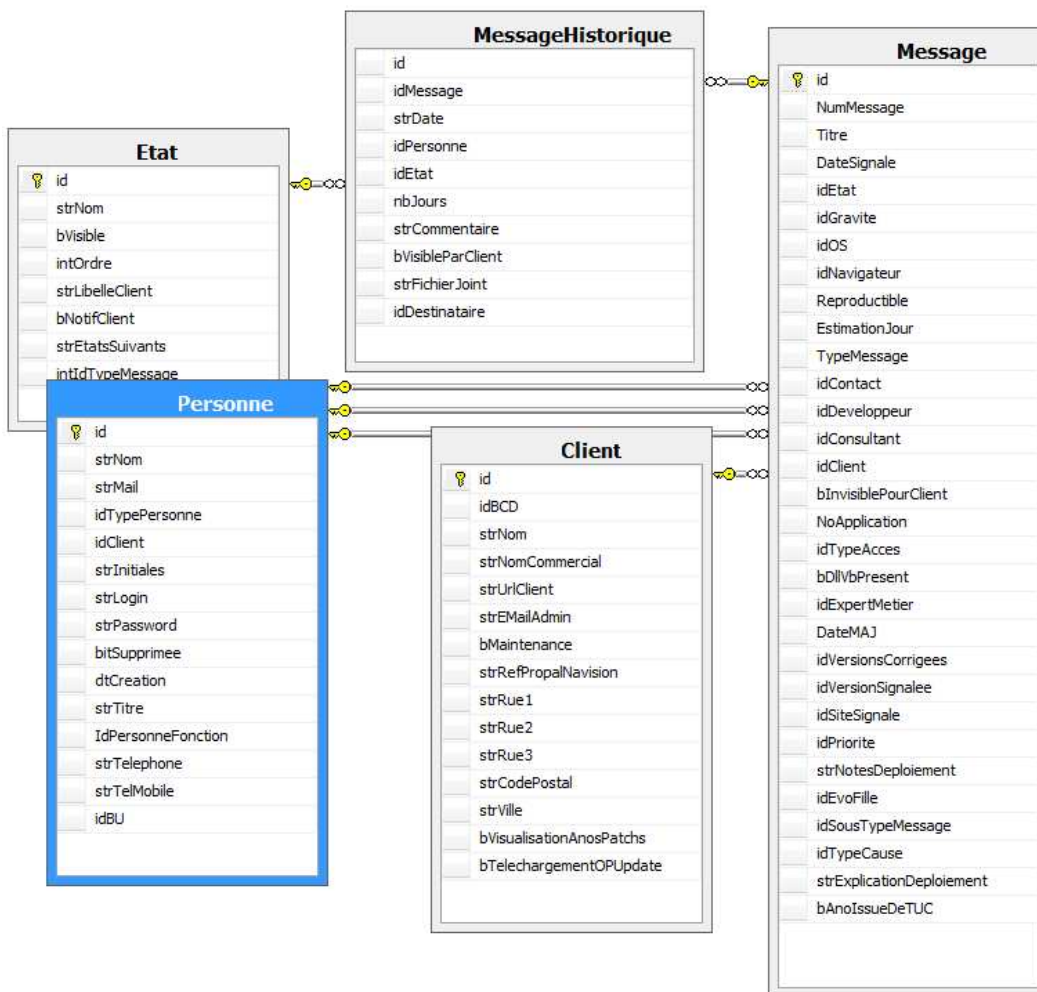


FIGURE 4.2 – Schéma conceptuel simplifié de la base de données étudiée.

4.1.1 Analyse lexicale et indexation du texte intégral

La première étape du processus d'implémentation consiste à générer un index du corpus expérimental en langue française. Cet index fortement bruité comporte environ 145 000 clés uniques pour 290 000 enregistrements indexés. Les éléments indexés sont majoritairement des titres et des commentaires. L'index comporte les principales inflexions grammaticales de la langue française, telles que genre, nombre ou conjugaison - *e.g.*, service, services, trouver, trouvées, trouverait, etc.

Le système d'indexation utilisé comporte un analyseur syntaxique capable d'éliminer certains mots vides (*i.e.*, articles et pronoms de la langue française). Afin de réduire le bruit persistant (*e.g.*, prépositions et termes absents des dictionnaires), nous peuplons une liste de mots-bruits par référencement lexical - *i.e.*, par comparaison aux lexiques de référence. Suite à l'examen des lexiques français en libre distribution, nous avons retenu deux ressources : *Lexique 3*, une base de données lexicale libre de l'Université Descartes - Paris [Ferrand 2001] et *Morphalou*, un lexique ouvert du Centre National de Ressources Textuelles et Lexicales. *Lexique 3* comporte environ 135 000 mots simples ou composés (termes), basés sur 55 000 lemmes (syllabes). *Morphalou* ne comporte pas de noms propres, onomatopées, anagrammes et voisins orthographiques, mais référence 554 000 termes. Après une manipulation conservant certaines structures lexicales pertinentes pour les experts métiers (*i.e.*, jargon, acronymes courants et abréviations métiers), nous obtenons une liste de 107 000 mots bruits et un index réduit à 26 000 clés uniques, soit un gain de facteur 6 par rapport à l'index initial (145 000 clés uniques).

Au terme du référencement lexical, quelques faux-positifs apparaissent dans la liste de mots bruits. Ce sont typiquement (1) des francisations (*e.g.*, *framework*, *export*), (2) des noms propres ou dénominations (*e.g.*, *Perrot*, *google*), (3) des acronymes et abréviations (*e.g.*, *DG*, *compil*), ou (4) des voisins orthographiques (*e.g.* *fonctionnalité*). Il est possible d'intégrer les faux-positifs dans un lexique de référence, par expertise ou par traitement automatique du langage naturel, afin d'améliorer la pertinence de l'indexation automatique. Dans ce cadre, l'usage de lexiques spécialisés métiers améliorera l'implémentation du modèle en contextes métiers particuliers (*e.g.*, métiers du droit, industries spécialisées).

4.1.2 Construction de thésaurus et pondération sémantique

Les méthodes d'apprentissage non supervisé du langage s'appuient sur des lexiques sémantiques (*e.g.*, wikipedia, wordnet, dbpedia) et/ou des algorithmes de traitement automatique du langage - *e.g.*, loi de Zipf, Conditional Random Fields - dédiés à l'apprentissage de la morphologie des langues [Delalonde 2007, Lavallée 2010, Chen 2011]. L'indexation sémantique du contenu endogène nécessite la construction automatique d'un thésaurus, mais ces méthodes complexes sont difficilement compatibles avec les contraintes du projet SOCIOPRISE, dans lequel le traitement automatique du langage n'est pas une priorité.

Afin de favoriser l'application industrielle du modèle, nous avons intégré un thé-

saurus construit à partir de l'index épuré et d'un dictionnaire de synonymes propriétaire - plus de 44 000 entrées associées à 458 000 termes. Notre automatisations produit environ 45 000 associations - des synonymes ou hyponymes/hyperonymes prépondérants - à partir de 19 500 entrées. Ce thésaurus ne comporte qu'un seul niveau hiérarchique, ne représentant pas systématiquement une subsomption sémantique. Chaque terme est enrichi de sa catégorie grammaticale parmi les choix possibles *nom*, *verbe*, *adverbe*, *adjectif* ou *divers*. Nous obtenons une moyenne de 5 termes associés par terme indexé trouvant au moins une correspondance sémantique, pour un thésaurus comportant environ 34 000 entrées uniques¹ - *e.g.*, *aberration* : absurdité, astigmatisme, erreur, folie, irisation.

Sur des termes génériques employés dans le contexte spécifique au projet, la construction automatique de thésaurus définie dans le cadre de SOCIOPRISE révèle les limites d'une implémentation simplifiée. Le terme *serveur*, désignant une machine informatique dans le contenu endogène, est par exemple associé au terme *servante*, association uniquement acceptable pour un registre de langue générique. L'intégration de dictionnaires de synonymes spécialisés ou d'ontologies, pour un coût d'implémentation alourdi, permettrait d'éradiquer ce type d'imprécisions.

4.1.2.1 Prégnance linguistique et similarité sémantique

L'index épuré et le thésaurus sont utilisés lors d'une recherche en texte intégral. Chaque entrée du corpus retournée par une interrogation comporte un coefficient propre - son poids - calculé en fonction des occurrences de mots-clés dans le texte de l'entrée et de l'ensemble des résultats retournés.

Deux algorithmes classiques sont employés pour calculer ce poids, sous forme d'un coefficient de *prégnance linguistique*, de type BM25 [Robertson 1976] - *i.e.* un poids de type *TF.IDF* -, et d'un indice de *similarité sémantique* [Jaccard 1901, Rogers 1960] - *i.e.*, un *indice de JACCARD*. Dans le contexte de SOCIOPRISE, les équations utilisées sont définies dans [Zaragoza 2004] et intégrées au SGBDR gratuit utilisé.

Dans l'implémentation présentée, la commande CONTAINSTABLE produit un classement de type *TF.IDF* (*i.e.*, variable $Rank \equiv IDF.TF$), défini et formalisé par :

$$ContainsRank = \min \left(1000, \log_2 \left(2 + \frac{Idoc}{Tdoc} \right) * 16 * \frac{Docc}{Dtrm} \right) \quad (4.1)$$

Idoc est le nombre de documents indexés (enregistrements pour une propriété dans l'index), *Tdoc* est le nombre de documents contenant un terme T_i , *Docc* est le nombre d'occurrences de T_i dans un document D_j et *Dtrm* le nombre de termes dans D_j . Dans cette formule, $Docc/Dtrm$ correspond à la définition de *TF* dans [Salton 1986] et $CR_{IDF} = \log_2 \left(2 + \frac{Idoc}{Tdoc} \right)$ est une variante simple de l'*IDF* de [Sparck Jones 1972]. La valeur par défaut 1000 est retenue si $IDF.TF > 1000$.

Pour la commande FREETEXTTABLE, le classement est calculé à partir de la formule BM25 [Robertson 1976]. BM25 représente l'état de l'art en terme de raf-

1. Certaines entrées de l'index n'ont aucune association dans le thésaurus.

TABLE 4.1 – Comparaison du calcul de prégnance sémantique.

| Test | Mots-clés | Nb. résultats | Ecart-type | Test Z |
|------------------|-------------------------------|---------------|------------|--------|
| A (TFIDF simple) | configuration, serveur, | 1346 | 51,7 | 0,5 |
| A (BM25) | configuration, serveur, | 1359 | 35,3 | 0,5 |
| B (TFIDF simple) | configuration, serveur, poste | 2031 | 66 | 0,5 |
| B (BM25) | configuration, serveur, poste | 2041 | 28,8 | 0,5 |
| C (TFIDF simple) | validation, message, erreur | 534 | 55,8 | 0,5 |
| C (BM25) | validation, message, erreur | 539 | 33,6 | 0,5 |

finement de $TF.IDF$. La partie IDF implémentée dans `FREETEXTTABLE` est extrapolée de `BM25` et intègre la prise en compte d’annotations utilisateur (de 0 à 5 par document), relatives à la pertinence du document D_j par rapport au terme T_i . Elle se formalise comme suit :

$$FTT_{IDF} = \log_{10} \left(\frac{(AD_{trm} + 0,5) * (I_{doc} - A_{doc} + AD_{trm} + 0,5)}{(A_{doc} - AD_{trm} + 0,5) * (T_{doc} - AD_{trm} + 0,5)} \right) \quad (4.2)$$

A_{doc} , AD_{trm} , sont respectivement le nombre de documents annotés et de documents annotés contenant le terme T_i . En l’absence d’annotations, FTT_{IDF} se formalise comme suit :

$$FTT_{IDF} = \log_{10} \left(\frac{I_{doc} + 0,5}{T_{doc} + 0,5} \right) \quad (4.3)$$

Connaissant la valeur de la variable $Rank$, avec $Rank \equiv Ts \equiv U$, et la définition de IDF pour les commandes `CONTAINS` et `FREETEXT` (*i.e.*, CR_{IDF} et FTT_{IDF}), nous définissons les pseudo-valeurs électriques d’intensité I et de résistance R . Conformément à l’équation 3.10 (*cf.* section 3.6), nous formalisons I et R comme suit, pour l’une ou l’autre des commandes :

$$I = \frac{U}{R} \equiv \frac{Rank}{R_{IDF} | FTT_{IDF}}, \quad (4.4)$$

$$R = \frac{U}{I} \equiv R_{IDF} | FTT_{IDF} = Rank / \frac{Rank}{R_{IDF} | FTT_{IDF}} \quad (4.5)$$

Le détail des définitions implémentées est indiqué et commenté en annexe, section A.1.

Nous avons évalué les différences de classement des résultats obtenus en fonction de deux modes de calcul de poids (*i.e.*, TF-IDF simple, BM25), sur trois recherches A, B et C. Les résultats de cette évaluation sont présentés dans le tableau 4.1.2.1. La valeur de test Z égale à 0,5 indique que la moyenne arithmétique des échantillons - *i.e.* les valeurs $Rank$ calculées - est stable, au centre de la distribution des échantillons.

Les courbes moyennes, préfixées par M, du graphique correspondant au tableau de comparaison du calcul de prégnance sémantique, présenté en figure 4.3, montrent

la séparation plus importante du classement produit par le *TFIDF* simplifié (commande *ContainsTable*), comparé au classement *TF.IDF* BM25 (commande *FreeTextTable*).

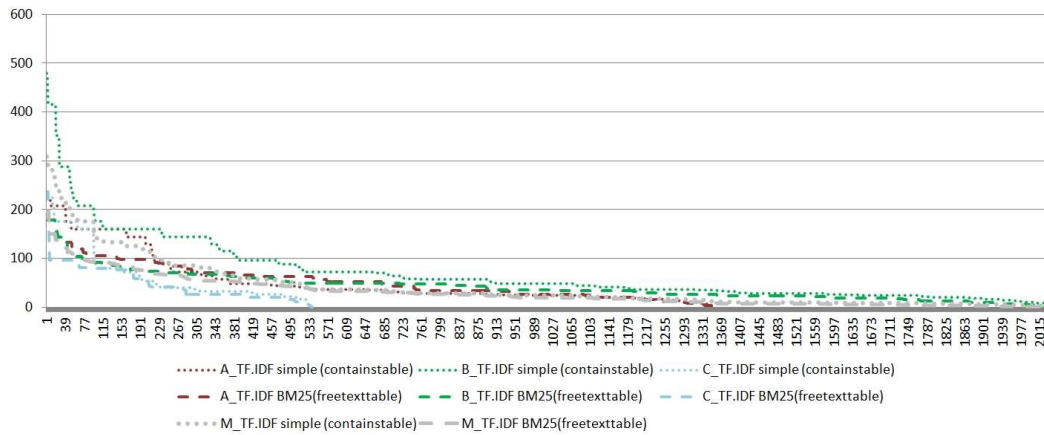


FIGURE 4.3 – Distribution de valeurs de prégnance linguistique.

4.1.3 Développement et codage

Les fonctions transactionnelles intégrées à la plateforme logicielle permettent d’implémenter les définitions des valeurs respectives de Ts , Is , R dans une interrogation SQL avec `CONTAINS``TABLE`, comme dans l’exemple présenté en figure 4.4.

Le modèle implémenté définit un index lexical et un thésaurus basés sur le corpus. Dans ce contexte, les valeurs de prégnance linguistique et de similarité sémantique fournies par les commandes `CONTAINS``TABLE`, `FREETEXT``TABLE`, et éventuellement `ISABOUT`, correspondent à des valeurs sémantiques de tension, d’intensité et de résistance pour tout arc connectant un individu à un terme indexé du contenu endogène, au sein d’un réseau de compétences. Le modèle implémenté utilise cette propriété pour définir un procédé automatique de construction de réseaux de compétences à la demande, en fonction de mots clés. Ce procédé est présenté dans la section 4.2.

4.2 Réseaux de compétences à la demande

Soit le graphe de compétences $GC(N, A)$, défini en section 3.2. Pour rappel, ce graphe est :

1. orienté, valué et connexe (*i.e.*, tout nœud $n \in N$ est connecté par au moins un arc $a \in A$),
2. bidirectionnel, car pour tout arc $a \in A$, $a(n_i, n_j) \leftrightarrow a(n_j, n_i)$


```

SELECT 'configuration' as motcle, id_mess, id_hist, idContact as idpers,
Contact_mess as Personne, RANK as poids_U,
(log(2+((SELECT COUNT(DISTINCT document_id) Idoc
FROM
sys.dm_fts_index_keywords_by_document(DB_ID('support'),OBJECT_ID('FT_Chaine_index')))) / (SELECT sum(document_count) as TDoc
FROM
sys.dm_fts_index_keywords(DB_ID('support'),OBJECT_ID('FT_Chaine_index'))
WHERE display_term='configuration'
GROUP BY display_term))) / LOG(2)) as R,
(RANK/(log(2+((SELECT COUNT(DISTINCT document_id) Idoc
FROM
sys.dm_fts_index_keywords_by_document(DB_ID('support'),OBJECT_ID('FT_Chaine_index')))) / (SELECT sum(document_count) as TDoc
FROM
sys.dm_fts_index_keywords(DB_ID('support'),OBJECT_ID('FT_Chaine_index'))
WHERE display_term='configuration'
GROUP BY display_term))) / LOG(2))) as I
FROM dbo.FT_Chaine_index AS FTCI
INNER JOIN CONTAINSTABLE(dbo.FT_Chaine_index, *, 'FORMSOF(INFLECTIONAL,
configuration) OR configuration NEAR serveur OR configuration NEAR poste',
LANGUAGE N'French') AS RESFT
ON FTCI.id = RESFT.[KEY]

```

| motcle | id_mess | id_hist | idpers | Personne | poids | R | I |
|---------------|---------|---------|--------|----------|-------|------------------|------------------|
| configuration | 19982 | 466204 | 927 | D | 120 | 7.17990909001493 | 16.7133035384645 |
| configuration | 19929 | 515541 | 907 | V | 288 | 7.17990909001493 | 40.1119284923148 |
| configuration | 20004 | 466406 | 1044 | P | 57 | 7.17990909001493 | 7.93881918077064 |
| configuration | 20061 | 467169 | 160 | N | 144 | 7.17990909001493 | 20.0559642461574 |
| configuration | 20148 | 469070 | 754 | L | 72 | 7.17990909001493 | 10.0279821230787 |
| configuration | 20148 | 469209 | 754 | L | 48 | 7.17990909001493 | 6.6853214153858 |
| configuration | 20674 | 473320 | 555 | A | 41 | 7.17990909001493 | 5.71037870897537 |
| configuration | 20224 | 468249 | 1278 | P | 48 | 7.17990909001493 | 6.6853214153858 |
| configuration | 20248 | 470446 | 891 | M | 48 | 7.17990909001493 | 6.6853214153858 |
| configuration | 20278 | 471469 | 268 | F | 96 | 7.17990909001493 | 13.3706428307716 |
| configuration | 20319 | 469322 | 1127 | E | 123 | 7.17990909001493 | 17.1311361269261 |
| configuration | 20340 | 469067 | 9 | C | 88 | 7.17990909001493 | 12.256422594874 |
| configuration | 20377 | 469329 | 1127 | E | 41 | 7.17990909001493 | 5.71037870897537 |

FIGURE 4.4 – Exemple de liste d’arcs obtenue par interrogation SQL.

3. asymétrique, car chaque arc $a \in A$ de la paire $[a(n_i, n_j); a(n_j, n_i)]$ porte une valeur propre.

Nous construisons $GC(N, A)$ par une approche *bottom-up*, à partir de mots clés saisis. Cette approche détermine un sous-processus basé sur la répartition naturelle des intensités au sein d’un réseau électrique. Elle débute par la sélection et la pondération sémantique des relations entre individus et connaissances indexées, et se poursuit par l’application d’une heuristique de propagation bidimensionnelle, cohérente dans la dimension sémantique, comme dans la dimension pseudo-électrique du modèle.

Les premières étapes du sous-processus permettent la construction à la demande d’une structure de réseau de compétences, en rapport à des mots-clés. Elles produisent une valuation partielle du graphe de compétences et se décomposent comme suit :

- **Etape 1.** sélection et pondération par valeurs sémantiques, des arcs $apcr(n_i, n_j) \in aPcr$ dont l’une des extrémités $n \in Pcr$ (cf. section 3.2) ;

- **Étape 2.** sélection et pondération par valeurs sémantiques des nœuds $n \in N$, de $GC(N, A)$.

La valuation totale du graphe de compétences est ensuite obtenue par propagation des valeurs sémantiques de $aPcr$, durant les étapes suivantes :

- **Étape 3.** valuation des arcs de aCr ,
- **Étape 4.** valuation finale de l'intensité sémantique de Cc_i ;
- **Étape 5.** valuation de la tension de Cr_j et Pcr_k ;
- **Étape 6.** adaptation bidimensionnelle de résistance $Rapcr(Cr_j, Pcr_k)$;
- **Étape 7.** valuation de la tension sémantique de Pcr_k ;
- **Étape 8.** valuation d'arcs de aCc ;

Enfin, deux étapes supplémentaires permettent de justifier la transformation des intensités sémantiques des nœuds de Cc en pseudo-valeurs électriques, puis de les équilibrer selon les lois de *Kirchhoff* :

- **Étape 9.** transformation des valeurs sémantiques de Cc_i ;
- **Étape 10.** équilibrage de pseudo-intensités au cœur d'un graphe de compétences.

4.2.1 Étape 1 : sélection et pondération des arcs de $aPcr$

À partir du modèle théorique, une implémentation programmatique appelant les transactions SQL intégrées à la plateforme logicielle est développée. Cette implémentation donnant $aPcr$ valué, permet d'évaluer les poids produits ou consommés de chaque individu, relativement aux connaissances clés et connaissances relatives - cf. figure 4.5.

| idpers | Personne | terme_source | terme_associe | tpoidscons | tpoidsprod |
|--------|----------|--------------|---------------|------------|------------|
| 301 | 1057 | T | validation | NULL | 34 |
| 302 | 1079 | T | erreur | faute | 44 |
| 303 | 1092 | Il | erreur | faute | 25 |
| 304 | 1096 | B | erreur | faute | 29 |
| 305 | 1099 | B | message | pensee | 54 |
| 306 | 1100 | F | erreur | faute | 176 |
| 307 | 1111 | L | erreur | confusion | 48 |
| 308 | 1127 | E | erreur | confusion | 96 |
| 309 | 1134 | Jr | erreur | faute | 160 |
| 310 | 1136 | P | validation | NULL | 34 |
| 311 | 1147 | R | erreur | confusion | 32 |
| 312 | 1154 | P. | erreur | confusion | 38 |
| 313 | 1154 | P | erreur | faute | 60 |

FIGURE 4.5 – Exemple de poids produit/consommé.

Les données de la relation définissant les arcs $aPcr$ couvrent Pcr, Cr, Cc , définis en section 3.2. Il est possible d'en extraire l'ensemble des nœuds $n \in N$, de $GC(N, A)$.

| idpers | Personne | terme_source | terme_associe | tpoidscons | tRcons | tlcons | tpoidsprod | tRprod | tlprod |
|--------|----------|---------------|---------------|------------|------------------|------------------|------------|------------------|------------------|
| 1837 | R | configuration | NULL | 57 | 7,17990909001493 | 7,93881918077064 | 57 | 7,17990909001493 | 7,93881918077064 |
| 1868 | IL | configuration | NULL | 32 | 7,17990909001493 | 4,45688094359053 | 32 | 7,17990909001493 | 4,45688094359053 |
| 1903 | V | configuration | NULL | 57 | 7,17990909001493 | 7,93881918077064 | 57 | 7,17990909001493 | 7,93881918077064 |
| 4 | O | configuration | figure | 451 | 7,32309232119819 | 61,5860049578356 | NULL | NULL | NULL |
| 7 | E | configuration | figure | NULL | NULL | NULL | 122 | 7,41821298506141 | 16,4460093348196 |
| 9 | CI | configuration | figure | 581 | 7,17990909001493 | 80,9202446320656 | 632 | 7,24543238147427 | 87,2273684612599 |
| 10 | R | configuration | figure | 72 | 7,17990909001493 | 10,0279821230787 | NULL | NULL | NULL |
| 18 | Y | configuration | figure | 306 | 7,32054787015144 | 41,8001501291555 | 602 | 7,29065780141916 | 82,5714244718519 |

FIGURE 4.6 – Extrait de résultats avec Ts , Is et R produit/consommé.

L'intensité Is est calculée selon les définitions formalisées en section 4.1.2.1.

4.2.2 Étape 2 : sélection et valuation de l'intensité des nœuds

Nous définissons deux valeurs d'intensité sémantique pour tout $n \in N$, une valeur de production φ_n et une valeur de consommation γ_n . Ces valeurs sont obtenues, pour chaque nœud, en fonction de la somme des valeurs d'intensité I portées par les arcs entrants ou sortants. Nous les formalisons comme suit.

1. Pour Pcr , $\forall n \in Pcr$:

$$\varphi_{n_k} = \sum \varphi(I_{apcr}(n_k, n)) : n \in Cr|Cc \quad (4.6)$$

$$\gamma_{n_k} = \sum \gamma(I_{apcr}(n_k, n)) : n \in Cr|Cc \quad (4.7)$$

Les individus Pcr_k forment la couche périphérique du graphe et sont considérés comme composants actifs, générateurs et consommateurs de la connaissance représentée dans les couches centrales (Cr, Cc) du graphe. La valeur sémantique d'intensité de production φ_n ou de consommation γ_n , d'un individu k est la somme des intensités sémantiques des arcs de production ou de consommation $Ia(k, C) : C \in Cr|Cc$, de cet individu vers les connaissances clés et connaissances relatives.

1. Pour Cr , $\forall n \in Cr$:

$$\varphi_{n_j} = \sum \varphi(I_{apcr}(n_j, n_k)) : n_k \in Pcr, \quad (4.8)$$

$$\gamma_{n_j} = \sum \gamma(I_{apcr}(n_j, n_k)) : n_k \in Pcr \quad (4.9)$$

Pour une connaissance relative Cr_j , la valeur d'intensité sémantique produite ou consommée $\varphi_n|\gamma_n$ est la somme des valeurs d'intensité sémantique des arcs de production des individus de Pcr vers cette connaissance relative.

1. Pour Cc , $\forall n \in Cc$:

$$\varphi_n = \sum \varphi(I_{acr}(n, n_j)) : n_j \in Cr + \sum \varphi(I_{apcr}(n_j, n_k)) : n_k \in Pcr \quad (4.10)$$

$$\gamma_n = \sum \gamma(I_{acr}(n, n_j)) : n_j \in Cr + \sum \gamma(I_{apcr}(n_j, n_k)) : n_k \in Pcr \quad (4.11)$$

La valeur d'intensité sémantique produite ou consommée $\varphi_n|\gamma_n$ d'une connaissance clé Cc_i est la somme des valeurs d'intensité sémantiques des arcs $Ia(k, C)$ de production ou de consommation connectant les individus de Pcr aux connaissances clés, éléments de Cc_i .

4.2.3 Propagation des valeurs sémantiques de $aPcr$

A ce stade du processus de valuation du graphe de compétences $GC(N, A)$, tous les arcs de $aPcr$ et tous les nœuds du graphe de compétences portent une valeur d'intensité sémantique. Du fait que les arcs de aCr ne sont pas encore valués (intensité, résistance et tension), seule la valeur la valeur IDF des nœuds de Cc est fixée et les arcs de aCc ne portent aucune valeur².

Nous complétons la valuation du graphe en définissant les valeurs d'intensité et de tension des arcs de aCr , calculées au prorata inverse des valeurs de résistance sémantique relatives aux nœuds de Cc (*i.e.*, l' IDF de ces nœuds), ce qui permet de finaliser le calcul des valeurs d'intensité produites ou consommées des nœuds de Cc . Connaissant l'intensité sémantique des nœuds de Cc , et connaissant la résistance et l'intensité sémantique des arcs $apcr(Cc_i, Pcr_k) : Cc_i \in Cc, Pcr_k \in Pcr$ et des arcs $acr(Cc_i, Cr_j) : Cc_i \in Cc, Cr_j \in Cr$, le processus de valuation du graphe peut déterminer les valeurs de tension sémantique, produites ou consommées, des nœuds de Cr et des nœuds de Pcr .

Les valeurs d'intensité des nœuds de Cc , cohérentes dans la dimension sémantique et observant spontanément la *loi des mailles* de KIRCHHOFF (loi des nœuds), favorisent le passage de la dimension sémantique à la dimension électrodynamique, au sein de $GC(N, A)$. Elles sont ensuite équilibrées selon *loi des nœuds* de KIRCHHOFF, afin de produire une recommandation prédictive basée sur l'équilibre naturel des flux électroniques. Cette dernière étape achève le processus de valuation complète de $GC(N, A)$, par propagation de valeurs d'intensité, de résistance et de tension sémantiques, basées sur $TF.IDF$.

4.2.3.1 Étape 3 : valuation des arcs de aCr

Un arc $acr(Cc_i, Cr_j) : Cc_i \in Cc, Cr_j \in Cr$ porte des valeurs d'intensité et de tension sémantique, produites ou consommées, calculées selon les lois d'OHM appliquées aux groupements de résistances en parallèle. Sa valeur de résistance sémantique (IDF) est celle du nœud $Cc_i \in Cc$. Pour $n = |acr(Cc_i, Cr_j)|$ où j est constant, nous définissons et formalisons Rt , résistance totale du groupement de $nacr(Cc_i, Cr_j)$, ainsi que les valeurs sémantiques, produites ou consommées, d'intensité Is et de tension Ts .

$$Rt = \frac{1}{\sum_{k=1}^n 1/Racr_k(Cc_i, Cr_j)} \quad (4.12)$$

2. Ce modèle de propagation des valeurs d'intensité sémantique aux nœuds du graphe de compétences ne s'applique pas de la même façon aux valeurs de tension sémantiques ($TF.IDF$).

$$Is(acr(Cc_i, Cr_j)) = Rt/Rs(acr(Cc_i, Cr_j)), \quad (4.13)$$

$$Ts(acr(Cc_i, Cr_j)) = Is(Cc_i) * Rt \quad (4.14)$$

4.2.3.2 Étape 4 : valuation finale de l'intensité de Cc_i

En pratique dans le modèle, un nœud $Cr_j \in Cr$ distribue les intensités sémantiques, produites ou consommées, des nœuds de Pcr qui lui sont connectés, aux nœuds de Cc auxquels il est connecté, comme illustré par la figure 4.7 - cf. équations 4.10 et 4.11. Cette distribution est calculée au prorata inverse des résistances $Racr_k(Cc_i, Cr_j)$. Connaissant maintenant l'intensité sémantique des arcs

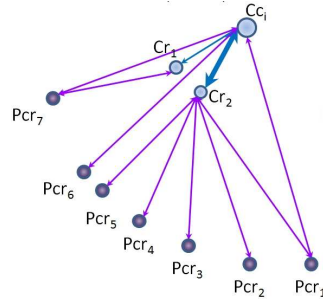


FIGURE 4.7 – Distribution tripartite des flux dans un graphe de compétences.

$apcr(Cc_i, Pcr_k) : Cc_i \in Cc, Pcr_k \in Pcr$ et des arcs $acr(Cc_i, Cr_j) : Cc_i \in Cc, Cr_j \in Cr$, l'intensité sémantique des nœuds de Cc peut être calculée, selon les équations 4.10 et 4.11.

4.2.3.3 Étape 5 : valuation de la tension sémantique de Cr_j

Dans notre modèle, un nœud $Cr_j \in Cr$ est comparé à un collecteur, tandis qu'un nœud $Cc_i \in Cc$ est comparé à un composant accumulateur et transformateur d'intensités sémantiques, produites ou consommées.

Connaissant $Isacr(Cc_i, Cr_j)$, valeur d'intensité sémantique d'un arc $acr(Cc_i, Cr_j)$ et connaissant la valeur de résistance sémantique $Racr(Cc_i, Cr_j) \equiv IDF(Cc_i)$, nous définissons la valeur pseudo-électrique de tension d'un nœud Cr_j en fonction des règles de groupement de résistances dans les circuits électriques.

Nous comparons l'ensemble des arcs $acr(Cc_i, Cr_j)$ à un groupement de résistances parallèle afin de calculer la valeur de tension sémantique $Ts(Cr_j)$. Pour un nœud $Cr_j : n = |apcr(Cc_i, Cr_j)|$, le groupement observe les règles suivantes :

$$Rt(Cr_j) = \frac{1}{\sum_{k=1}^n 1/Racr_k(Cc_i, Cr_j)} \quad (4.15)$$

$$It(Cr_j) = \sum_{k=1}^n Isacr_k(Cc_i, Cr_j), Ts(Cr_j) = Rt(Cr_j) . It(Cr_j) \quad (4.16)$$

4.2.3.4 Étape 6 : adaptation bidimensionnelle de résistance $R_{apcr}(Cr_j, Pcr_k)$

Dans le modèle développé, un nœud $Cr_j \in Cr$ prend pour valeur $IDF(Cr_j)$ la valeur $IDF(Cc_i)$ du nœud $Cc_i \in Cc$, auquel il est connecté. Un arc $apcr(Cr_j, Pcr_k) : Pcr_k \in Cr$ hérite implicitement de la valeur $IDF(Cr_j)$ - i.e., $R_{apcr}(Cr_j, Pcr_k) = Rt(Cr_j) = IDF(Cc_i)$.

Lorsque Cr_j , est connecté à plusieurs nœuds Cc_i , nous recourons à l'heuristique de calcul de résistance totale définie par la dimension pseudo-électrique et formalisée dans l'équation (Equation 4.15), s'appliquant à $Rt(Cr_j)$. La valeur initiale de résistance sémantique $R_{apcr}(Cr_j, Pcr_k)$ prend alors une valeur $Rt(Cr_j) \neq IDF(Cc_i)$. Ce cas est illustré dans la figure 4.8, dans laquelle $R_{apcr}(Cr_1, Pcr_k)$ passe de $IDF(Cc_i) = 7,18$, à $Rt(Cr_j) = 4,1$ (de gauche à droite).

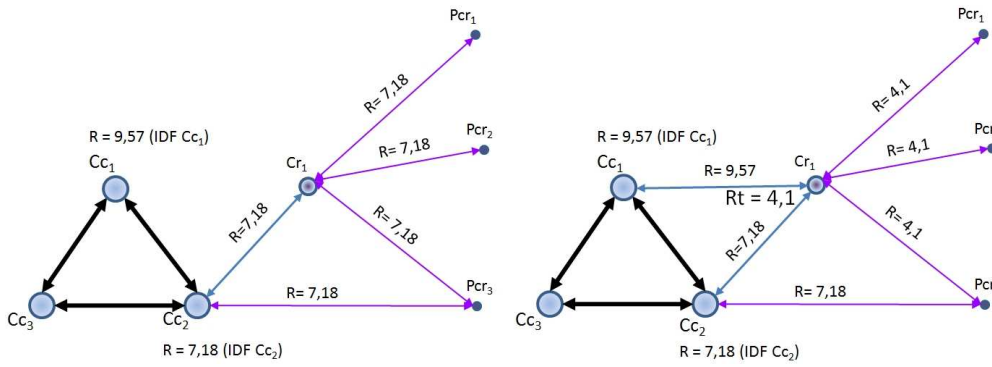


FIGURE 4.8 – Adaptation bidimensionnelle de $R_{apcr}(Cr_j, Pcr_k)$.

4.2.3.5 Étape 7 : valuation de la tension sémantique de Pcr_k

Dans notre modèle théorique, un nœud $Pcr_k \in Pcr$ est comparé à un producteur ou un consommateur d'intensités sémantiques (intensités produites ou consommées).

Nous avons défini, selon l'équivalence $R \equiv IDF$, la valeur initiale de résistance sémantique $R_{apcr}(Cc_i, Pcr_k) = IDF(Cc_i)$, d'un arc $apcr(Cc_i, Pcr_k)$. Nous avons ensuite défini la valeur $R_{apcr}(Cr_j, Pcr_k) = Rt(Cr_j)$, d'un arc $apcr(Cr_j, Pcr_k)$. Comme pour un nœud Cr_j connecté à plusieurs nœuds Cc_i , pour un nœud Pcr_k connecté à plusieurs Cc_i , nous recourons à l'heuristique de calcul de résistance totale définie par la dimension pseudo-électrique et formalisée dans l'équation (Eq. para), s'appliquant alors à $Rt(Pcr_k)$.

Connaissant $Rt(Pcr_k)$, connaissant $Isapcr(Cc_i, Pcr_k)$, valeur d'intensité sémantique d'un arc $apcr(Cc_i, Pcr_k)$, et connaissant $Isapcr'(Cr_j, Pcr_k)$, valeur d'un arc $apcr'(Cr_j, Pcr_k)$, nous recourons à nouveau à la dimension pseudo-électrique pour définir la valeur de tension sémantique $Ts(Pcr_k)$. Nous comparons l'ensemble des arcs $apcr(C, Pcr_k) = apcr(Cc_i, Pcr_k) \cup apcr'(Cr_j, Pcr_k)$ à un groupement de

résistances parallèle. Pour un nœud $Cr_j \in Cr : n = |apcr(C, Pcr_k)|$, nous définissons et formalisons la définition de $Ts(Pcr_k)$, à partir de $Rt(Pcr_k)$ et de l'intensité sémantique totale $Ist(Pcr_k)$:

$$Ist(Pcr_k) = \sum_{k=1}^n Isapcr_k(C, Pcr_k), Ts(Pcr_k) = Rt(Pcr_k) \cdot Ist(Pcr_k)$$

4.2.3.6 Étape 8 : valuation d'arcs de aCc

Dans la structure de graphe de compétences, les nœuds $Cc_i \in Cc$ et les paires d'arcs $\{acc(Cc_i, Cc_{i'}), acc(Cc_{i'}, Cc_i)\} \in aCc$ constituent un circuit unique.

Soit $m = |(Cc)|$.

Si $m > 2$, le circuit unique comporte $2m \{acc(Cc_i, Cc_{i'}), acc(Cc_{i'}, Cc_i)\}$. Si $m = 2$, le circuit comporte seulement $m \{acc(Cc_i, Cc_{i'}), acc(Cc_{i'}, Cc_i)\}$.

Un arc $acc(Cc_i, Cc_{i'})$ a pour valeur la différence de flux $DdF(acc(Cc_i, Cc_{i'}))$ des tensions sémantiques ($TF.IDF$) produites $\varphi(Cc_i)$, ou consommées $\gamma(Cc_i)$, par ses extrémités respectives, réparties sur le nombre d'arcs qu'il porte, telle que :

$$m = 1 \leftrightarrow \nexists DdF(acc(Cc_i, Cc_{i'})), \varphi(Cc_i) = \gamma(Cc_i) = (\varphi(Cc_i) + \gamma(Cc_i))/2$$

$$m = 2 \leftrightarrow DdF(acc(Cc_i, Cc_{i'})) = \varphi(Cc_i) - \gamma(Cc_{i'})$$

$$m > 2 \leftrightarrow DdF(acc(Cc_i, Cc_{i'})) = \varphi(Cc_i/2) - \gamma(Cc_{i'}/2)$$

4.2.4 Étape 9 : transformation des valeurs sémantiques de Cc_i

Pour trois points A, B, C formant une maille électrique : $U_{AB} + U_{BC} + U_{CA} = 0$.
- Loi des mailles de KIRCHHOFF, dans laquelle u est la tension, ou différence de potentiel, entre deux points.

Afin de vérifier la validité mathématique de notre modèle lors du passage de la dimension sémantique à la dimension électrodynamique, nous avons calculé les flux sémantiques orientés provenant des arcs $aPcr$ (*i.e.*, des arcs connectant les individus aux connaissances) et circulant au cœur du graphe, entre les sommets de Cc (Connaissances clés). Nous avons constaté que la répartition des tensions sémantiques $Ts \equiv TF.IDF$ observe scrupuleusement la loi des mailles de KIRCHHOFF, ce qui nous a conduit à approfondir l'hypothèse.

La figure 4.9 présente un exemple de valeurs sémantiques calculées sur un graphe de compétences et convergeant en son centre.

Ces valeurs définissent des flux sémantiques, centralisés par les nœuds Cc_i . Dans notre modèle, nous assimilons ces flux à des tensions électriques et vérifions que le circuit $GC(Cc, aCc)$, où aCc est l'ensemble des arcs connectant les sommets Cc entre eux, obéit bien à la loi des mailles, en calculant les Différences de Flux (DdF) circulant sur les arcs de aCc ³.

3. Afin de respecter le cadre théorique dans lequel la loi des nœuds se vérifie, nous considérons que les éléments de aCc et de Cc sont des dipôles non rayonnants - *i.e.*, neutres en électrostatique et en électromagnétisme. Afin de respecter la cohérence du modèle réel avec la métaphore électrique, nous considérons que les nœuds de Cc sont à la terre - *i.e.*, différence de potentiel ou tension nulles.

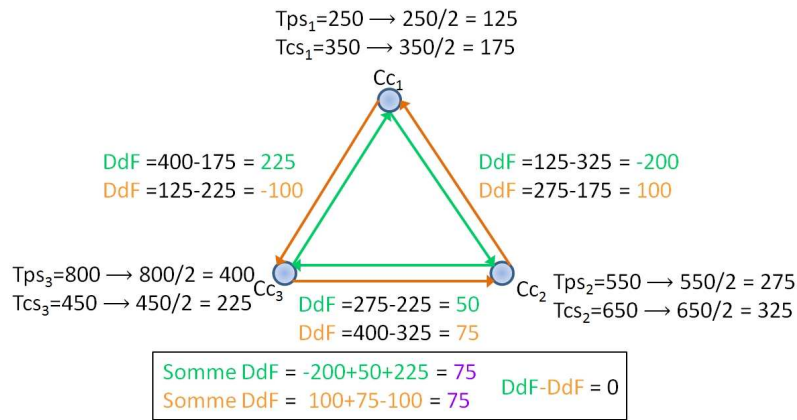


FIGURE 4.9 – Vérification de la loi des mailles de KIRCHHOFF sur le circuit unique $GC(Cc, aCc)$.

En accord avec les règles de distribution de flux de courant électrique dans un circuit, les flux sont divisés en fonction du nombre d'arcs connectés à un nœud (*i.e.*, son degré), avec une résistance identique de chaque arc. Dans la figure 4.9, où $m = |Cc| = 3$, la valeur produite P ou consommée C d'un nœud est divisée par $m - 1 = 2$, pour chaque nœud. Puis, les valeurs P sortantes sont comparées aux valeurs C entrantes, afin de calculer la DdF de chaque paire de nœuds, dans les deux orientations possibles. Considérant le flux de tension vert, opposé au flux orange dans la figure 4.9, la différence des DdF totales est égale à 0 ($75 - 75$) et respecte donc la loi des mailles de KIRCHHOFF. Par précaution, nous avons également vérifié la démonstration avec $|aCc| = 4|5$, ce qui rend envisageable, au sein d'un graphe de compétences, la transformation d'une valeur sémantique en valeur pseudo-électrique.

4.2.5 Étape 10 : équilibrage de pseudo-intensités au cœur d'un graphe de compétences

La somme des intensités entrantes est égale à la somme des intensités sortantes, dans un point quelconque du circuit - Loi des nœuds de KIRCHHOFF.

Lors de la valuation sémantique du graphe de compétences, nous avons pagé les poids sémantiques de l'extérieur vers le centre du graphe. Puis, nous avons converti les valeurs convergeant au centre du graphe en valeurs pseudo-électriques d'intensité, et vérifié l'adéquation de la métaphore avec la loi des nœuds de KIRCHHOFF. Par ses travaux sur les transistors lasers, M. FENG a fait apparaître que les lois de KIRCHHOFF ne s'appliquaient pas seulement aux flux d'électrons, mais également aux flux photoniques [Then 2010]. Les travaux de FENG ont permis la construction d'un composant opto-électronique, basé sur une extension des lois de KIRCHHOFF et l'ayant conduit à affirmer que les lois des courants de KIRCHHOFF devraient être redéfinies en tant que *lois des courants et des énergies* de KIRCHHOFF

. Nous définissons alors un principe d'équilibrage des flux dans le circuit unique $GC(Cc, aCc)$, à partir de la loi des nœuds, dans l'objectif de produire un modèle de recommandations basé sur l'équilibre énergétique naturel.

Nous présentons les intensités sémantiques en figure 4.10, comme dans l'exemple illustré en figure 4.9.

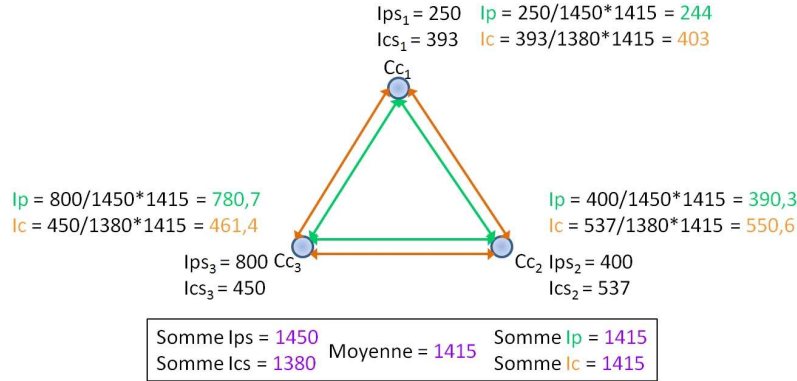


FIGURE 4.10 – Equilibrage des intensités au sein du circuit unique $GC(Cc, aCc)$, basé sur la loi des nœuds de KIRCHHOFF.

Le modèle produit $Ip = Ic$ et le circuit unique obéit à la loi des nœuds de KIRCHHOFF. La figure 4.10, illustre comment est appliquée la loi des nœuds au sein de $GC(Cc, aCc)$, en déduisant des valeurs moyennes Ip et Ic , produites ou consommées, à partir de la moyenne des sommes d'intensités sémantiques (TF) Ips et Ics .

Ce modèle définit des valeurs pseudo-électriques d'intensité qui observent l'équilibre naturel des flux électroniques correspondant à la loi des nœuds de KIRCHHOFF, au sein de $GC(Cc, aCc)$. Nous utilisons ces valeurs pour produire un système prédictif de recommandations et d'aide à la décision, visant à l'optimisation de la performance et à la prévention du risque psychosocial dans les entreprises et institutions. Ce système implémente un sous-processus de valuation dynamique de graphe de compétences, défini dans la section 4.3.

4.3 Valuation pseudo-électrique du graphe de compétences

Nous avons introduit l'équivalence $TF.IDF \equiv Is.R \equiv I.R = U \equiv Ts$ (cf. équation 3.10). En appliquant les lois de KIRCHHOFF au circuit unique $GC(Cc, aCc)$, nous avons comparé les valeurs de tension et d'intensité sémantique du graphe de compétences à des pseudo-valeurs électriques de tension et d'intensité. Les valeurs de tension sémantique $TF.IDF \equiv Ts$, sont transformées en pseudo-valeurs électriques de tension $U \equiv Ts$, par correspondance à la loi des mailles (cf. figure 4.9). Les valeurs d'intensité sémantique $Is \equiv TF$, sont initialement équilibrées et trans-

formées en pseudo-valeurs électriques d'intensité $I \equiv Is$, par application de la loi des nœuds (cf. figure 4.10). Sur ces pseudo-valeurs électriques s'appliquent les lois d'OHM et de KIRCHHOFF, avec $TF \equiv I$ et $IDF \equiv R$.

Nous définissons un procédé de rétro-propagation des pseudo-valeurs électriques, symétrique et cohérent avec celui défini pour la propagation des valeurs sémantiques. Nous appliquons ce procédé sur tous les éléments du graphe de compétences, à partir des pseudo-valeurs électriques d'intensité $I(Cc_i)$ des nœuds de Cc , en vue de confronter les résultats obtenus à l'avis des experts et au modèle réel d'activité représenté par le corpus expérimental.

A ce stade, nous avons défini :

1. un principe de propagation de valeurs sémantiques, des arcs bidirectionnels $apcr(Cc_i|Cr_j, Pcr_k)$ à tous les autres éléments d'un graphe de compétences $GC(N, A)$;
2. un modèle de distribution des flux issu de lois physiques, simulant des pseudo-valeurs électriques d'intensité, résistance et tension par observation des lois d'OHM et de KIRCHHOFF et formalisant le comportement des flux sémantiques dans les réseaux de compétences ;
3. un principe d'équilibrage de valeurs sémantiques accumulées ou restituées au sein de $GC(Cc, aCc)$, piloté par pseudo-valeurs électriques de tension et d'intensité équivalentes.

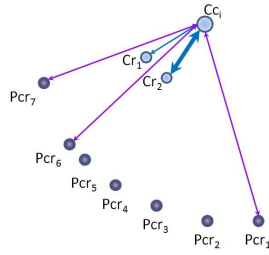
L'objectif est d'étendre le principe d'équilibrage (3) à l'ensemble du graphe de compétences $GC(N, A)$, afin d'en évaluer la pertinence en comparant, pour tout élément d'un réseau de compétences, les valeurs sémantiques de tension et d'intensité obtenues par équilibrage de flux électroniques, aux valeurs sémantiques initiales, figurant la notion de *stress* et de performance au sein d'un réseau de compétences. Les recommandations que représentent les valeurs équilibrées, une fois interprétées et confrontées à la connaissance des experts, seront-elles pertinentes pour l'optimisation de la performance et la prévention du risque psychosocial, au sein de réseaux sociaux professionnels ?

Afin de disposer de pseudo-valeurs électriques d'intensité et de tension pour tout élément d'un réseau de compétences $GC(N, A)$, nous définissons un principe de rétro-propagation des pseudo-valeurs électriques équilibrées de $GC(Cc, aCc)$ vers la périphérie de $GC(N, A)$.

4.3.1 Valuation pseudo-électriques d'arcs et d'extrémités à partir de connaissances-clés

Les nœuds du circuit unique $GC(Cc, aCc)$ portent des pseudo-valeurs d'intensité, produite ou consommée, calculées selon l'équilibre naturel des flux électroniques - lois de KIRCHHOFF - au sein de $GC(Cc, aCc)$.

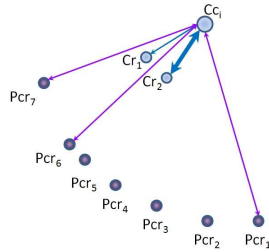
Soit les arcs $acr(Cc_i, Cr_j) : Cc_i \in Cc, Cr_j \in Cr$, les arcs $apcr(Cc_i, Pcr_k) : Cc_i \in Cc, Pcr_k \in Pcr$ et $aCp(Cc_i, C_j|k) = acr(Cc_i, Cr_j) \cup apcr(Cc_i, Pcr_k)$, illustré en figure 4.11

FIGURE 4.11 – Exemple de cinq arcs $acp(Cc_i, C_j)$.

L'union définie par $aCp(Cc_i, C_j|_k)$ est vue comme un ensemble de résistances montées en parallèle, pour tout nœud Cc_i . Nous en connaissons les règles de calcul de tension et d'intensité, formalisée par le système d'équations (Equation 4.15). Notons que lorsque les n résistances sont toutes de valeur identique, alors :

$$Rt = \frac{R}{n}, I(R_k) = It/n$$

Soit un arc $apcr(Cr_j, Pcr_k) : Cr_j \in Cr, Pcr_k \in Cr$, comme illustré en figure 4.12.

FIGURE 4.12 – Exemple de 6 arcs $apcr(Cr_j, Pcr_k)$.

Dans la figure 4.12, l'ensemble des arcs $apcr(Cr_2, Pcr_{1...5})$ est également assimilé à un groupement de résistances parallèle.

Considérant un circuit formé successivement d'un arc $acr(Cc_i, Cr_j)$ et d'un arc $apcr(Cr_j, Pcr_k) : Cr_j \in Cr, Pcr_k \in Pcr$, ce circuit est assimilé à deux résistances (les nœuds Cr_j, Pcr_k) montées en série, illustré dans la figure 4.13.

En électronique, dans un montage de résistances en série :

1. l'intensité est constante, $It = I_1 = I_2 = I_n$;
2. la résistance totale du circuit $Rt = R_1 + R_2 + \dots + R_n$;
3. la tension varie au prorata de la résistance, $Ur_1 = Ut * (R_1/Rt) = \sum_{k=1}^n Ur_k * (R_1/Rt)$;
4. la puissance varie selon la tension au prorata de la résistance, $Pr_1 = Ur_1.It$ et $Pt = Ut.It$.

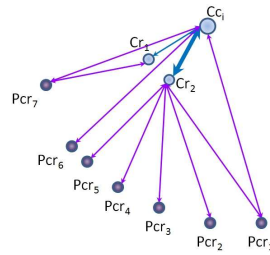


FIGURE 4.13 – Exemple de sous-graphe de compétences assimilé à un montage série-parallèle de résistances.

Dans un graphe de compétences $GC(N, A)$, un nœud $Cr_j \in Cr$ n'existe que s'il existe au moins un arc $apcr(Cr_j, Pcr_k)$, et il n'existe aucun arc connectant directement deux nœuds Cr_j et $Cr_{j'}$. Pour tout nœud $Cc_i \in Cc$, il existe donc deux cas de chemins possibles, soit un chemin $\langle Cc_i, Pcr_k \rangle$, soit un chemin $\langle Cc_i, Cr_j, Pcr_k \rangle$. Un graphe de compétences combine généralement ces deux cas. Nous définissons le procédé de calcul des valeurs pseudo-électriques d'intensité et de tension sur les arcs et les nœuds intégrant ces deux cas de chemins et leur combinaison, puis en vérifions les résultats à l'aide d'un solveur de circuit électronique, sur un échantillon simulé de $GC(N, A)$, dont la structure, illustrée en figure 4.14, a pour racine le nœud Cc_2 . Les valeurs initiales d'intensité sont celles déjà présentées en figure 4.10.

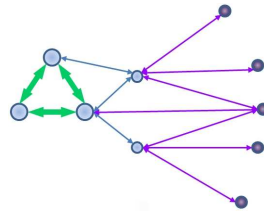


FIGURE 4.14 – Echantillon de test.

Les résultats obtenus par propagation des valeurs sémantiques sur l'échantillon de test sont illustrés en détail par la figure 4.15.

4.3.2 Rétro-propagation de pseudo-valeurs électriques depuis Cc_i

La rétro-propagation des valeurs pseudo-électriques d'intensité et de tension à partir d'un nœud $Cc_i \in Cc$ s'effectue en une seule passe, combinant les deux cas de chemins identifiés. Elle représente un modèle prédictif visant à produire des recommandations pour l'équilibre des compétences, produites ou consommées, entre individus au sein d'un réseau de compétences.

Le modèle théorique défini, basé sur la distribution de flux électroniques au sein de groupements de résistances en parallèle et en série, est implémenté pour

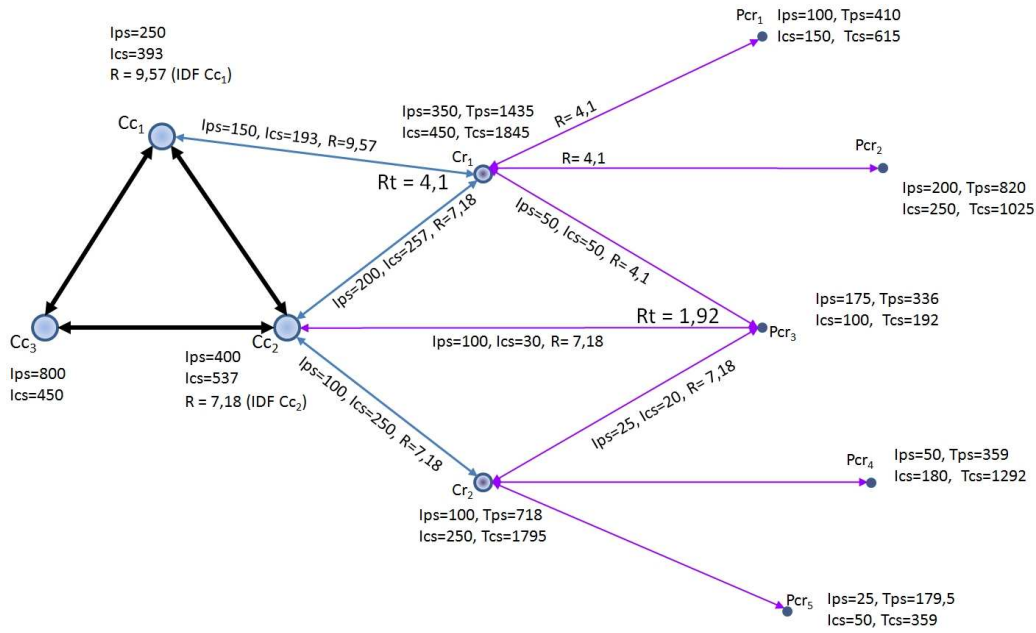


FIGURE 4.15 – Application test du modèle de propagation de flux sémantiques.

comparaison dans un simulateur/solveur de circuits électriques⁴.

La proportion des pseudo-valeurs d'intensités I entre les arcs $apcr(Cc_i, Pcr_k)$ d'un groupement en parallèle d'arcs $\langle Cc_i, Pcr_k \rangle$, est fonction de la proportion des pseudo-valeurs de résistance des arcs. La valeur de résistance sémantique R , produite ou consommée, présente sur tout arc d'un graphe de compétences, équivaut à sa pseudo-valeur électrique de résistance. Le processus de rétro-propagation de pseudo-valeurs électriques que nous définissons est comparable à celui de propagation des valeurs sémantiques, au sein d'un graphe de compétences. Cette similarité permet de maintenir la cohérence entre la dimension sémantique et la dimension électrique, tout en produisant une répartition de flux différente, d'une dimension à l'autre.

Nous étudions l'application des lois électriques sur deux cas de chemins identifiés, $\langle Cc_i, Pcr_k \rangle$ et $\langle Cc_i, Cr_j, Pcr_k \rangle$, entre un nœud Cc_i et les nœuds Pcr_k , et leur combinaison au sein d'un graphe de compétences.

4.3.2.1 Cas 1 : groupement de chemins $\langle Cc_i, Pcr_k \rangle$

Soit $n = |apcr(Cc_i, Pcr_k)|$, avec i constant. La pseudo-valeur électrique d'intensité $I(Cc_i)$ équilibrée est connue. Pour le groupement $\langle Cc_i, Pcr_k \rangle$, nous définissons et formalisons un groupement de résistances en parallèle comme suit :

$$Rt \langle Cc_i, Pcr_k \rangle = \frac{1}{\sum_{k=1}^n \frac{1}{Rapcr(Cc_i, Pcr_k)}}$$

4. Le solveur utilisé est *Solvelec* - <http://www.physicsbox.com/indexsolveelec2en.html>.

$$I_t \langle Cc_i, Pcr_k \rangle = I(Cc_i),$$

$$I_{apcr}(Cc_i, Pcr_k) = I(Cc_i) * \frac{Rt}{Rapcr(Cc_i, Pcr_k)}$$

Le groupement $\langle Cc_i, Pcr_k \rangle$ est vu comme une résistance unique, de pseudo-valeur électrique Rt et nous formalisons sa tension, constante sur chacun de ses arcs, comme suit :

$$U_t \langle Cc_i, Pcr_k \rangle = I_t \langle Cc_i, Pcr_k \rangle * Rt \langle Cc_i, Pcr_k \rangle = I(Cc_i) * Rt \langle Cc_i, Pcr_k \rangle$$

Dans l'exemple illustré en figure 4.16, nous obtenons :

$$I_{apcr}(Cc_1, Pcr_2) = I(Cc_1) * \frac{1 / \left(\frac{1}{Rapcr(Cc_1, Pcr_2)} + \frac{1}{Rapcr(Cc_1, Pcr_3)} + \frac{1}{Rapcr(Cc_1, Pcr_4)} \right)}{Rapcr(Cc_1, Pcr_2)},$$

et $U_{apcr}(Cc_1, Pcr_k) =$

$$I(Cc_1) * 1 / \left(\frac{1}{Rapcr(Cc_1, Pcr_2)} + \frac{1}{Rapcr(Cc_1, Pcr_3)} + \frac{1}{Rapcr(Cc_1, Pcr_4)} \right)$$

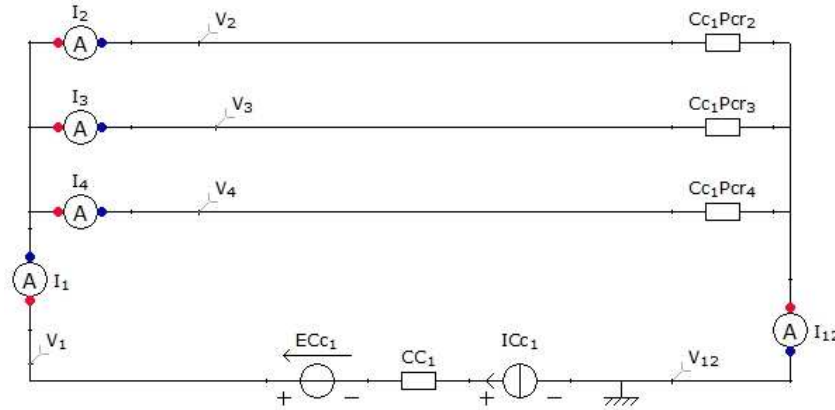


FIGURE 4.16 – Simulation électrique de groupement $\langle Cc_i, Pcr_k \rangle$.

4.3.2.2 Une limitation de l'ambivalence entre dimension sémantique et électrique

L'épreuve, à l'aide d'un solveur, du modèle théorique implémenté sous forme de circuit électrique, permet de discerner une limitation présente au passage de la dimension sémantique à la dimension électrique. Cette limitation apparait quand il existe plusieurs sources d'intensité Cc_i , directement connectées à un nœud $Cc_j|Pcr_k$.

Un circuit électrique forme par définition une boucle, plus ou moins complexe. Nous considérons donc, qu'il existe un bus commun, sous-jacent à la structure de graphe de compétences, acheminant les flux circulant entre le centre et les extrémités du graphe vers une terre commune, représentée dans la figure 4.16 et la figure 4.17.

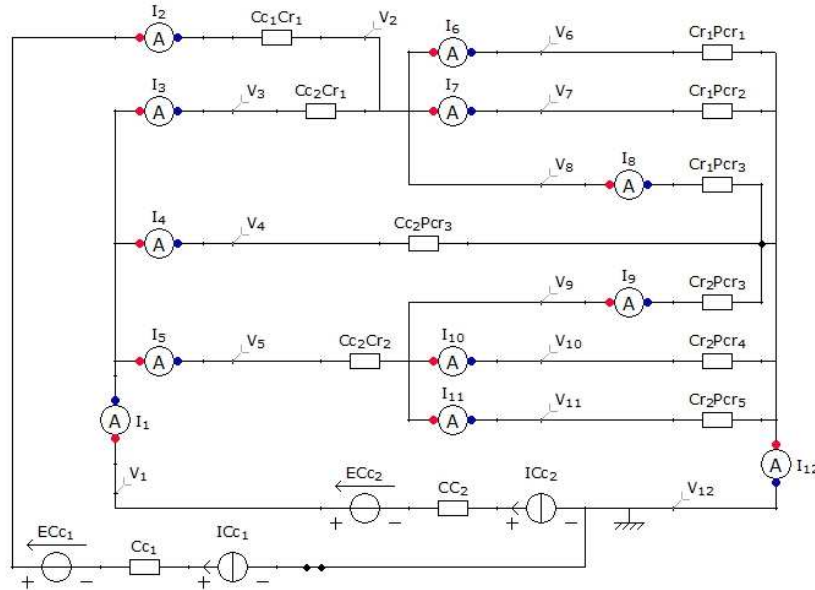


FIGURE 4.17 – Représentation du graphe de compétences échantillon sous forme de circuit électrique.

La figure 4.17 représente le graphe de compétences échantillon sous forme de circuit électrique, tel qu'implémenté et soumis au solveur logiciel utilisé pour l'expérimentation.

Lorsque plusieurs sources d'intensité dépendantes Cc_i sont connectées à un même nœud Cr_j et/ou Pcr_k , comme illustré en figure 4.17, des interactions se produisent entre les sources et se répercutent sur les composants du circuit. Ces interactions causées par la dépendance des sources, alourdissent fortement le calcul de rétro-propagation, par récurrence. Afin de conserver au mieux l'ambivalence définie entre la dimension sémantique et la dimension électrique, sans perdre la capacité d'exécution du modèle prédictif développé, le processus de rétro-propagation isole partiellement les sources communes. Pour les sources Cc_1, Cc_2 - figure 4.17 - la rétro-propagation des intensités, produites ou consommées de Cc_2 , intègre celles de Cc_1 sans les réinjecter sur Cc_2 , et la résistance totale du circuit strictement alimenté par Cc_1 , est isolée de celle du circuit strictement alimenté par Cc_2 . La distribution des flux en est sensiblement affectée, sans que la cohérence générale et l'ambivalence du modèle ne soit remise en cause.

4.3.2.3 Cas 2 : groupement de chemins $\langle Cc_i, Cr_j, Pcr_k \rangle$

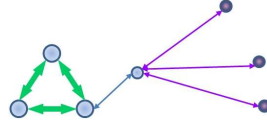
Tout groupement unitaire de chemins $\langle Cc_i, Cr_j, Pcr_k \rangle$ de distance $d = 2$, dans lequel i, j sont constants, forme un pseudo-groupement série-parallèle de résistances où $n = |apcr(Cr_j, Pcr_k)|$.

Définition des résistances

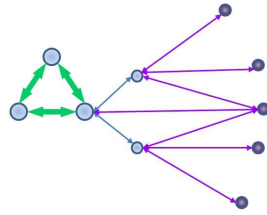
Comme avec les groupements d'arcs parallèles $\langle Cc_i, Pcr_k \rangle$, pour les groupements $\langle Cc_i, Cr_j \rangle$ et $\langle Cr_j, Pcr_k \rangle$, nous définissons $Rt \langle Cc_i, Cr_j \rangle$, $Rt \langle Cr_j, Pcr_k \rangle$, résistances totales définies selon les règles de groupement de résistances en parallèle. Dans un groupement unitaire $\langle Cc_i, Cr_j, Pcr_k \rangle$ où i, j sont constants, tel qu'illustré par la figure 4.18, la pseudo-valeur de résistance totale est définie comme un groupement de résistances en série-parallèle. La résistance totale du groupement se formalise comme suit :

$$Ru \langle Cc_i, Cr_j, Pcr_k \rangle = Racr(Cc_i, Cr_j) + Rt \langle Cr_j, Pcr_k \rangle$$

$$Ru \langle Cc_i, Cr_j, Pcr_k \rangle = Racr(Cc_i, Cr_j) + \frac{1}{\sum_{k=1}^n \frac{1}{Racpr(Cr_j, Pcr_k)}} \quad (4.17)$$

FIGURE 4.18 – Groupement unitaire $\langle Cc_i, Cr_j, Pcr_k \rangle$.

Pour un nœud Cc_i , nous définissons un groupement parallèle $\langle Cc_i, Cr_j \rangle$ de groupements unitaires série-parallèles $\langle Cc_i, Cr_j, Pcr_k \rangle$, correspondant à un groupement $\langle Cc_i, Cr_j, Pcr_k \rangle$ dans lequel i est constant, illustré par la figure 4.19.

FIGURE 4.19 – Groupement $\langle Cc_i, Cr_j, Pcr_k \rangle$.

Avec $l = |acr(Cc_i, Cr_j)|$ et $n = |apcr(Cr_j, Pcr_k)|$, nombre d'arcs $apcr(Cr_j, Pcr_k)$ pour un nœud Cr_j , la pseudo-valeur de résistance totale du groupement se formalise comme suit :

$$Rt \langle Cc_i, Cr_j, Pcr_k \rangle = \frac{1}{\sum_{j=1}^l \frac{1}{Ru \langle Cc_i, Cr_j, Pcr_k \rangle}}$$

$$Rt \langle Cc_i, Cr_j, Pcr_k \rangle = \frac{1}{\sum_{j=1}^l 1 / \left(Racr (Cc_i, Cr_j) + \frac{1}{\sum_{k=1}^n \frac{1}{Rapcr (Cr_j, Pcr_k)}} \right)} \quad (4.18)$$

Définition des intensités

Pour un groupement parallèle $\langle Cc_i, Cr_j, Pcr_k \rangle$ de groupements unitaires série-parallèles (figure 4.19), la pseudo-valeur d'intensité électrique $Iacr (Cc_i, Cr_j)$ est calculée au prorata inverse des pseudo-valeurs unitaires de résistances $Racr (Cc_i, Cr_j)$ et en fonction de $I (Cc_i)$, intensité électrique totale du groupement partiel $\langle Cc_i, Cr_j \rangle$. $I \langle Cc_i, Cr_j \rangle$ se formalise comme suit :

$$Iacr (Cc_i, Cr_j) = I (Cc_i) * \frac{Rt \langle Cc_i, Cr_j, Pcr_k \rangle}{(Racr (Cc_i, Cr_j) + Rt \langle Cr_j, Pcr_k \rangle)}$$

$$Iacr (Cc_i, Cr_j) = I (Cc_i) * \frac{1 / \sum_{j=1}^l 1 / \left(Racr (Cc_i, Cr_j) + \frac{1}{\sum_{m=1}^n \frac{1}{Racr (Cr_j, Pcr_k)}} \right)}{\left(Racr (Cc_i, Cr_j) + 1 / \sum_{m=1}^n \frac{1}{Racr (Cr_j, Pcr_k)} \right)} \quad (4.19)$$

Nous définissons l'intensité, produite ou consommée, d'un nœud Cr_j :

$$I (Cr_j) = \sum Iacr (Cc_i, Cr_j) \quad (4.20)$$

Connaissant $I (Cr_j)$, nous définissons $I \langle Cr_j, Pcr_k \rangle$, formalisée par :

$$Iapcr (Cr_j, Pcr_k) = I (Cr_j) * \frac{Rt \langle Cr_j, Pcr_k \rangle}{Rapcr (Cr_j, Pcr_k)} \quad (4.21)$$

Définition des tensions

Pour un groupement parallèle $\langle Cc_i, Cr_j, Pcr_k \rangle$ de groupements unitaires série-parallèles, la tension pseudo-électrique U est constante dans les arcs $\langle Cc_i, Cr_j \rangle$ ou $\langle Cr_j, Pcr_k \rangle$, où $U = Ut$. Elle se calcule en fonction de la résistance totale $Rt \langle Cc_i, Cr_j, Pcr_k \rangle$ et de l'intensité $I (Cc_i)$. Nous formalisons $U \langle Cc_i, Cr_j \rangle$ comme suit :

$$Uacr (Cc_i, Cr_j) = I (Cc_i) * Rt \langle Cc_i, Cr_j, Pcr_k \rangle$$

$$Uacr (Cc_i, Cr_j) = I (Cc_i) * \frac{1}{\sum_{j=1}^l 1 / \left(Racr (Cc_i, Cr_j) + 1 / \sum_{k=1}^n \frac{1}{Rapcr (Cr_j, Pcr_k)} \right)} \quad (4.22)$$

Enfin, $Uapcr (Cr_j, Pcr_k)$ se définit et formalise par :

$$Uapcr (Cr_j, Pcr_k) = I (Cr_j) * Rt \langle Cr_j, Pcr_k \rangle$$

$$Uapcr (Cr_j, Pcr_k) = \sum Iacr (Cc_i, Cr_j) * \frac{1}{\sum_{k=1}^n \frac{1}{Rapcr (Cr_j, Pcr_k)}} \quad (4.23)$$

4.3.2.4 Cas 3 : groupement de chemins $\langle Cc_i, Pcr_k \rangle$ et $\langle Cc_i, Cr_j, Pcr_k \rangle$

Les deux cas de chemins, $\langle Cc_i, Pcr_k \rangle$ et $\langle Cc_i, Cr_j, Pcr_k \rangle$, coexistent généralement dans un graphe de compétences. Nous formalisons donc la définition des pseudo-valeurs d'intensité et de tension dans ce contexte mixte.

Pour un nœud Cc_i , nous considérons le groupement des chemins $\langle Cc_i, Pcr_k \rangle$ et des chemins $\langle Cc_i, Cr_j, Pcr_k \rangle$ comme un groupement de résistances parallèle. Connaissant $Rt \langle Cc_i, Pcr_k \rangle$ [E.2] et $Rt \langle Cc_i, Cr_j, Pcr_k \rangle$ [Eq.8], nous définissons et formalisons la résistance totale des chemins $\langle Cc_i, Pcr_k \rangle$ et $\langle Cc_i, Cr_j, Pcr_k \rangle$ comme suit :

$$Rt \langle Cc_i, Pcr_k | Cc_i, Cr_j, Pcr_k \rangle = \frac{1}{1/Rt \langle Cc_i, Pcr_k \rangle + 1/Rt \langle Cc_i, Cr_j, Pcr_k \rangle} \quad (4.24)$$

Pour les deux cas de chemins combinés, les pseudo-valeurs d'intensité sont redéfinies en fonction de $Rt \langle Cc_i, Pcr_k | Cc_i, Cr_j, Pcr_k \rangle$. Nous formalisons ainsi :

$$I_{apcr} (Cc_i, Pcr_k) = I (Cc_i) * \frac{1/(1/Rt \langle Cc_i, Pcr_k \rangle + 1/Rt \langle Cc_i, Cr_j, Pcr_k \rangle)}{Rapcr (Cc_i, Pcr_k)}$$

$$I_{apcr} (Cc_i, Cr_j) = I (Cc_i) * \frac{Rt \langle Cc_i, Pcr_k | Cc_i, Cr_j, Pcr_k \rangle}{(Rapcr (Cc_i, Cr_j) + Rt \langle Cr_j, Pcr_k \rangle)}$$

$$I_{apcr} (Cc_i, Cr_j) = I (Cc_i) * \frac{1/(1/Rt \langle Cc_i, Pcr_k \rangle + 1/Rt \langle Cc_i, Cr_j, Pcr_k \rangle)}{\left(Rapcr (Cc_i, Cr_j) + 1/\sum_{m=1}^n \frac{1}{Rapcr (Cr_j, Pcr_k)} \right)}$$

$$I_{apcr} (Cr_j, Pcr_k) = I (Cr_j) * \frac{Rt \langle Cr_j, Pcr_k \rangle}{Rapcr (Cr_j, Pcr_k)}$$

$$I_{apcr} (Cr_j, Pcr_k) = \sum I_{acr} (Cc_i, Cr_j) * \frac{Rt \langle Cr_j, Pcr_k \rangle}{Rapcr (Cr_j, Pcr_k)} \quad (4.25)$$

Pour les deux cas de chemins combinés, la tension des arcs $\langle Cc_i, Pcr_k \rangle$ ou $\langle Cc_i, Cr_j \rangle$ est redéfinie et formalisée par :

$$U \langle Cc_i, Pcr_k | Cr_j \rangle = I (Cc_i) * \frac{1}{1/Rt \langle Cc_i, Pcr_k \rangle + 1/Rt \langle Cc_i, Cr_j, Pcr_k \rangle} \quad (4.26)$$

Enfin, $U \langle Cr_j, Pcr_k \rangle$ se redéfinit et formalise par :

$$U \langle Cr_j, Pcr_k \rangle = I (Cr_j) * Rt \langle Cr_j, Pcr_k \rangle$$

$$U \langle Cr_j, Pcr_k \rangle = \sum I_{acr} (Cc_i, Cr_j) * \frac{1}{\sum_{k=1}^n \frac{1}{Rapcr (Cr_j, Pcr_k)}} \quad (4.27)$$

Pour le sous-graphe arborescent $GCCI (N, A) : \forall a (n, n'), a (n', n'') n = Cc \wedge n \in Cr | Pcr \wedge n'' \in Pcr$, une arborescence dont la racine est strictement le nœud Cc_i , nous avons maintenant défini les équations permettant la rétro-propagation des pseudo-valeurs électriques d'intensité et de tension, à partir des pseudo-valeurs d'intensité électriques et équilibrées. Ces équations, en respectent les lois d'OHM et les règles de groupement de résistances en série, en parallèle et en série-parallèle, permettent la valuation des arcs de $GCCI (N, A)$ à partir de la pseudo-valeur équilibrée d'intensité $I (Cc_i)$.

4.3.3 Modélisation et vérification - solveur de circuit électrique

Pour valider nos définitions initiales, nous avons représenté le graphe de compétences échantillon dans un solveur de circuit électrique - cf. figure 4.17. Le processus de rétro-propagation isolant partiellement les sources d'intensité $Cc_{i+1} \dots Cc_n$ relatives à un nœud Cc_i (cf. section 4.3.2.2), nous utilisons le circuit électrique représenté en figure 4.20, pour comparer les résultats produits par notre modèle théorique à ceux du solveur.

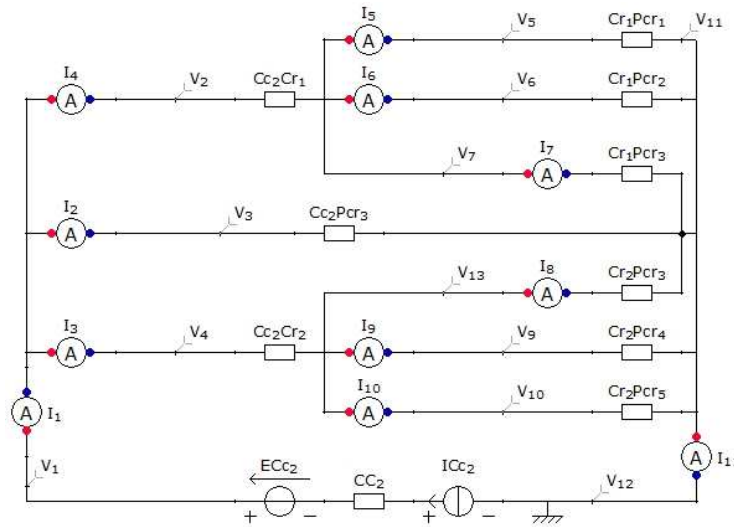


FIGURE 4.20 – Circuit électrique d'expérimentation du processus de rétro-propagation.

Nous avons comparé les valeurs d'arcs produites par le solveur, à celles obtenues par nos définitions, illustrées en figure 4.21 et figure 4.22. Les valeurs obtenues par notre modèle théorique correspondent intégralement à celles générées par le solveur.

Pseudo-valeurs des nœuds

Un nœud Cr_j est distributeur, et porte les pseudo-valeurs d'intensité des arcs $acr(Cc_i, Cr_j)$ dont il est extrémité. Un nœud Pcr_k est producteur/consommateur, et porte les pseudo-valeurs cumulées des arcs $apcr(Cc_i, Pcr_k)$ et $apcr(Cr_j, Pcr_k)$ dont il est extrémité. Les pseudo-valeurs électriques d'intensité et de tension d'un nœud Cc_i (accumulateur et transformateur) sont connues. Nous observons le modèle électronique pour définir le calcul des pseudo-valeurs cumulées d'un nœud $Cr_j|Pcr_k$, et vérifions nos définitions à l'aide du solveur.

Une pseudo-valeur, produite ou consommée, d'intensité de Pcr_k est la somme des pseudo-valeurs d'intensité des arcs dont il est extrémité (groupement de résistances parallèle). Une pseudo-valeur, produite ou consommée, de tension $U(Pcr_k)$, est le produit de l'intensité totale des arcs auxquels il est connecté par la résistance totale

| Arc | Résistance | Inverse resistance | | Arcs | I(arc) - A | U(arc) - V |
|---------|------------|--------------------|-------------|--------------------------|-------------|------------|
| Cc2Cr1 | 7,18 | 0,139275766 | | Cc2-Cr1 | 126,5933143 | 1081,9509 |
| Cc2Cr2 | 7,18 | 0,139275766 | | Cc2-Cr2 | 113,017151 | 1081,9509 |
| Cc2Pcr3 | 7,18 | 0,139275766 | | Cc2-Pcr3 | 150,6895347 | 1081,9509 |
| Cr1Pcr1 | 4,1 | 0,243902439 | | Cr1-Pcr1 | 42,19777142 | 173,01086 |
| Cr1Pcr2 | 4,1 | 0,243902439 | | Cr1-Pcr2 | 42,19777142 | 173,01086 |
| Cr1Pcr3 | 4,1 | 0,243902439 | | Cr1-Pcr3 | 42,19777142 | 173,01086 |
| Cr2Pcr3 | 7,18 | 0,139275766 | | Cr2-Pcr3 | 37,67238368 | 270,48771 |
| Cr2Pcr4 | 7,18 | 0,139275766 | | Cr2-Pcr4 | 37,67238368 | 270,48771 |
| Cr2Pcr5 | 7,18 | 0,139275766 | | Cr2-Pcr5 | 37,67238368 | 270,48771 |
| | | | | Inverse Rt Cr+Pcr | | |
| I(Cc2) | 390,3 | Rt des Pcr de Cr1 | 1,366666667 | | | |
| U(Cc2) | 390,3 | Rt des Pcr de Cr2 | 2,393333333 | | | |
| | | Rt Cc2-Cr1-Pcr | 8,546666667 | 0,11700468 | | |
| | | Rt Cc2-Cr2-Pcr | 9,573333333 | 0,104456825 | | |
| | | Rt Cc2-Pcr3 | 7,18 | 0,139275766 | | |
| | | RT Cc2 | 2,772100587 | 0,360737271 | | |

FIGURE 4.21 – Valeurs d’arcs obtenues par le modèle théorique sur le graphe de compétences échantillon.

de ces arcs. Nous formalisons ainsi :

$$n = |\text{apcr}(Cc_i, Pcr_k)|, n' = |\text{apcr}(Cr_j, Pcr_k)|$$

$$I(Pcr_k) = \sum_{i=1}^n I_{\text{apcr}}(Cc_i, Pcr_k) + \sum_{j=1}^{n'} I_{\text{apcr}}(Cr_j, Pcr_k) \quad (4.28)$$

$$U(Pcr_k) = \frac{1}{\sum_{i=1}^n \frac{1}{R_{\text{apcr}}(Cc_i, Pcr_k)} + \sum_{j=1}^{n'} \frac{1}{R_{\text{apcr}}(Cr_j, Pcr_k)}} * I(Pcr_k) \quad (4.29)$$

Les valeurs calculées obéissent aux lois d’OHM ($U = R.I$). La figure 4.22 présente les résultats obtenus à l’aide du solveur, sur l’échantillon de graphe de compétences étudié, dont les valeurs relatives à Cc_2 ont été isolées - *i.e.*, l’intensité de Cc_2 est uniquement dépendante de Cc_2 . Le circuit électrique correspondant est celui illustré en figure 4.20, rappelé en bas à gauche. Les valeurs correspondent bien à celles obtenues par le modèle algébrique, présentées en figure 4.21.

Conformément au modèle théorique défini, les pseudo-valeurs électriques de Cr_1 sont dépendantes des intensités, produites ou consommées, accumulées et transformées par les nœuds Cc_2 et Cc_1 . Les figure 4.20 et 4.22 illustrent la cohérence du modèle de rétro-propagation défini et du modèle électrique, sans tenir compte de cette dépendance. Les valeurs sémantiques de tension et d’intensité produites ou consommées Tps, Tcs, Ips, Ics , et les valeurs pseudo-électriques Up, Uc, Ip, Ic obtenues sur l’échantillon de graphe de compétences, sont illustrées en figure 4.23.

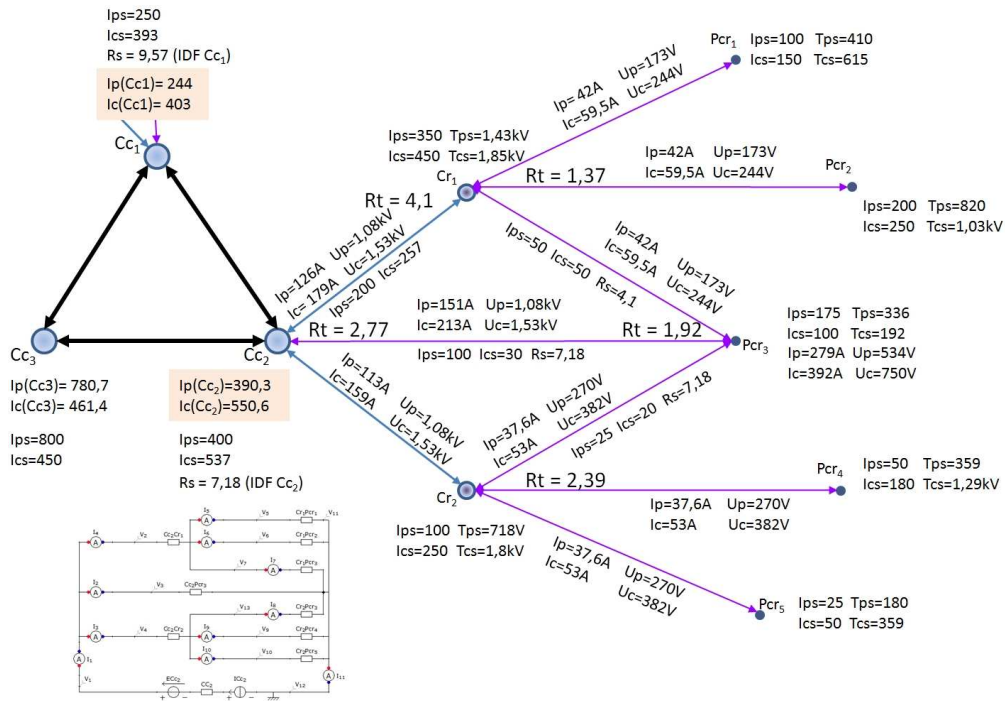


FIGURE 4.22 – Comparaison de valeurs sémantiques et pseudo-valeurs électriques équilibrées.

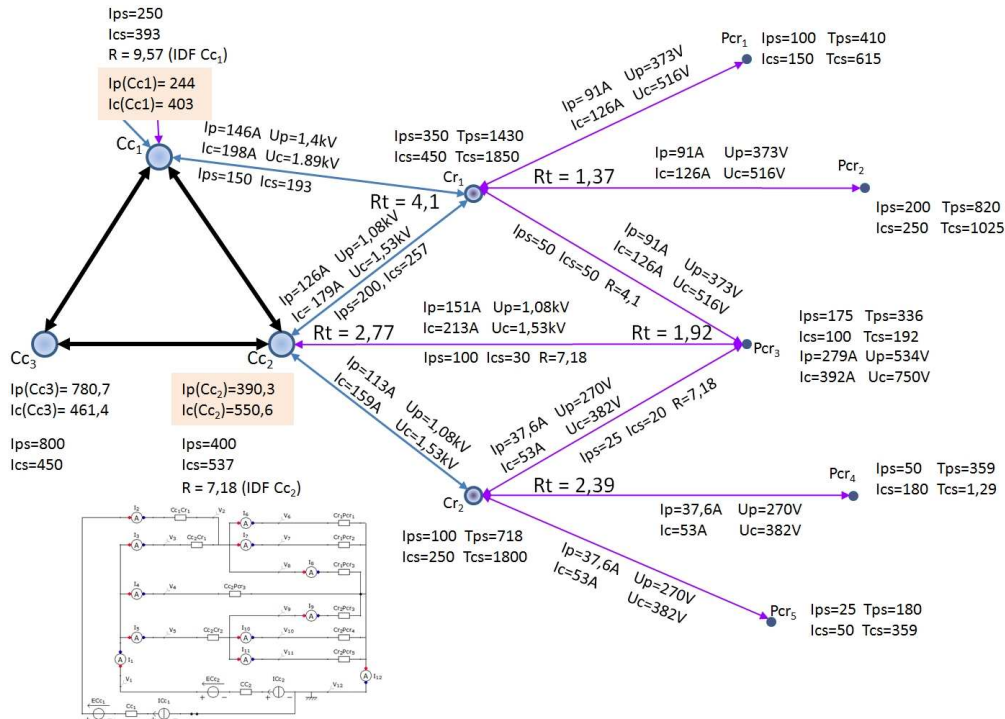


FIGURE 4.23 – Rétro-propagation de pseudo-valeurs électriques sur l'échantillon étudié.

La figure 4.23 montre comment les valeurs d'intensité, produites ou consommées, accumulées et transformées par Cc_1 modifient les pseudo-valeurs d'intensité et de tension des nœuds $Cr_1, Pcr_1, Pcr_2, Pcr_3$ et des arcs les connectant directement, dans notre modèle. Dans la figure 4.23, les valeurs Rt montrent la résistance totale du groupe de résistances parallèles. Ces valeurs sont calculées et utilisées en accord aux règles de distribution électronique dans les circuits séries/parallèle, durant les processus de propagation et de rétro-propagation.

4.4 Un modèle de recommandations basé sur l'équilibre énergétique

Pour tout graphe de compétences construit à la demande, nous avons défini :

1. un processus de propagation de valeurs sémantiques basé sur $(TF.IDF)$ et sur les lois d'OHM ;
2. un principe de transformation de valeurs de flux sémantiques en valeurs pseudo-électriques, basé sur la loi des mailles de KIRCHHOFF ;
3. un principe d'équilibrage de valeurs pseudo-électriques, basée sur la loi des nœuds de Kirchhoff ;
4. un processus de rétro-propagation de valeurs pseudo-électriques équilibrée, basé sur les lois d'OHM.

L'algorithme implémentant ces 4 éléments est communiqué en annexe ⁵ - cf section A.2.

Les valeurs pseudo-électriques d'intensité équilibrées au sein du circuit unique $GC(Cc, aCc)$, bien que dépendantes des valeurs sémantiques, propagent un équilibre des flux obéissant au modèle physique, au sein du graphe de compétences. Comparées aux valeurs sémantiques initiales des éléments du réseau de compétences, ces valeurs d'intensité peuvent ainsi représenter un système de recommandations, basé sur l'équilibre naturel des flux énergétiques.

Dans l'illustration (cf. figure 4.23), les résistances initiales ne sont pas modifiées, contrairement à ce qui se passe dans la simulation électrique avec le circuit correspondant, illustré en bas à gauche de la figure. Cette imperfection du modèle théorique de rétro-propagation est la conséquence de la limitation que nous lui avons imposée en section 4.3.2.2. Notre processus de rétro-propagation isole les sources dépendantes pour éviter une complexité récursive inutile ; hormis cela, il respecte scrupuleusement les règles de distribution électronique des circuits séries/parallèles. Cela représente une évolution significative de l'analogie entre flux électroniques et ARS, depuis les travaux de [Brandes 2005] et [Newman 2005].

5. Cet algorithme est écrit dans un pseudo-langage procédural, pour favoriser sa transcription rapide par un développeur habitué à la programmation de procédures stockées - e.g., PL-SQL Oracle.

Expérimentations, résultats et recommandations

Sommaire

| | | |
|------------|---|-----------|
| 5.1 | Expérimentation du modèle prédictif | 78 |
| 5.1.1 | Données expérimentales | 78 |
| 5.1.2 | Résultats généraux sur l'échantillon étudié | 78 |
| 5.1.3 | Analyse d'une recommandation prédictive | 80 |
| 5.1.4 | Application aux graphes de compétences à la demande | 81 |
| 5.2 | Expérimentation de mesures sémantiques | 84 |
| 5.2.1 | Données expérimentales | 84 |
| 5.2.2 | Intermédiarité de flux sémantique | 85 |
| 5.2.3 | Centralité de proximité de flux sémantique | 90 |
| 5.2.4 | Modularité de communauté sémantique | 93 |
| 5.2.5 | Mesure de stress et métrique de tension sémantique | 94 |
| 5.3 | Bilan d'expérimentations | 97 |

Le développement du modèle de recommandations basé sur l'équilibre énergétique, présenté en section 4.4, a nécessité d'alterner plusieurs fois entre la phase d'implémentation et la phase d'expérimentation qui, à chaque itération, ont respectivement progressé par ajustement mutuel, de façon agile. La suite de ce chapitre met tout d'abord en avant les résultats générés à l'aide de ce modèle, sur l'échantillon utilisé en phase d'implémentation et sur un exemple de réseau de compétences à la demande, issus de données réelles.

La section 5.1 présente l'interprétation, confrontée au point de vue de quelques experts du domaine de connaissances étudié¹, d'une recommandation prédictive basée sur l'équilibre naturel énergétique et sur la métrique de tension/stress introduite par notre modèle d'analyse sociographique et sémantique de réseaux sociaux (*cf.* Chapitre 3).

La section présente les résultats de l'expérimentation de nos mesures statiques et sémantiques d'ARS. Nous évaluons la validité du facteur C_p (*cf.* section 3.5.1) avec 2 des mesures définies en section 3.5. Nous évaluons ensuite la modularité de communauté définie dans [Blondel 2008] (*cf.* section 3.5.6). Ces mesures sont

1. Les experts métiers consultés au sein de l'entreprise occupent les rôles de Responsable Support *Testing*, Architecte Fonctionnel, et Directeur Produit et Innovation.

appliquées sur un réseau global de compétences et sur un exemple de réseau de compétences construit à la demande, tous deux issus de données réelles.

La mesure de stress de [Shimbel 1953] équivaut au numérateur de la centralité d'intermédiarité de [Freeman 1977] - cf. équation 2.1, section 2.1.1. Nous comparons la notion de *stress*, définie dans [Shimbel 1953], avec notre métrique de tension sémantique, en comparant la définition de [Shimbel 1953] au numérateur de notre centralité d'intermédiarité de flux sémantique - cf. section 3.5.4.

La section 5.2 présente les expérimentations suivantes :

1. centralité d'intermédiarité de flux sémantique - cf. section 3.5.4 ;
2. centralité de proximité sémantique - cf. section 3.5.5 ;
3. modularité de communauté sémantique [Blondel 2008] - cf. section 3.5.6 ;
4. comparaison de la notion de stress, définie dans [Shimbel 1953] avec notre métrique de tension sémantique.

5.1 Expérimentation du modèle prédictif

Nous comparons les valeurs sémantiques, produites ou consommées, *Ips*, *Ics* et *Tps* avec les pseudo-valeurs électriques équilibrées, sur les éléments de l'échantillon étudié durant la phase d'implémentation - cf. figure 4.23 et sur un exemple de réseau de compétences. Les résultats sont ensuite interprétés en termes experts, sous forme de recommandations induites par le modèle.

5.1.1 Données expérimentales

Le modèle d'analyse sociographique et sémantique de réseaux sociaux que nous avons défini (cf. Chapitre 3) a été implémenté sur un échantillon de données simulé, puis sur un réseau de compétences, relativement au contexte réel étudié.

L'échantillon de données simulé représente de façon synthétique, une structure de réseau de compétences hypothétique extraite manuellement des données étudiées dans le cadre du projet SOCIOPRISE. Le réseau de compétences expérimental est construit automatiquement en fonction des mots clés *validation*, *message*, *erreur*, saisis dans le prototype logiciel développé à l'aide de l'implémentation présentée en sections 4.1 et 4.2. Il est extrait des données réelles étudiées dans le cadre de SOCIOPRISE, représentant l'activité d'environ 600 personnes engagées sur un processus métier collaboratif - cf. figure 4.2.

5.1.2 Résultats généraux sur l'échantillon étudié

La figure 5.1 agrège les valeurs sémantiques et pseudo-électriques, produites ou consommées, résultant de l'expérimentation. En titres de colonnes, *Sum* or *S* et *Dev* ou *D* signifient *somme* et *écart-type*. Les quatre dernières colonnes agrègent les diverses valeurs produites ou consommées, définies dans les sections précédentes.

| Nœud | Ips | Ics | Ip | Ic | Tps | Tcs | Up | Uc | Sum Is | Sum I | Sum T | Sum U | Dev Is | Dev I | Dev Tp | Dev U | S I / S Is | S U / S T | Dev I / Dev Is | Dev U / Dev Ts |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|------------|------------|-------------|------------|------------|------------|----------------|----------------|
| Cc1 | 250 | 393 | 244 | 403 | | | | | 643 | 647 | | | 101 | 112 | | | 101% | | 111% | |
| Cc2 | 400 | 537 | 390 | 550 | | | | | 937 | 940 | | | 97 | 113 | | | 100% | | 117% | |
| Cc3 | 800 | 450 | 781 | 461 | | | | | 1250 | 1242 | | | 247 | 226 | | | 99% | | 91% | |
| Cr1 | 350 | 450 | 272 | 377 | 1430 | 1850 | 373 | 516 | 800 | 649 | 3280 | 889 | 71 | 74 | 297 | 102 | 81% | 27% | 105% | 34,25% |
| Cr2 | 100 | 250 | 113 | 160 | 718 | 1800 | 270 | 382 | 350 | 273 | 2518 | 652 | 106 | 33 | 765 | 79 | 78% | 26% | 31% | 10,38% |
| Pcr1 | 100 | 150 | 91 | 126 | 410 | 615 | 373 | 516 | 250 | 217 | 1025 | 889 | 35 | 25 | 145 | 101 | 87% | 87% | 70% | 69,76% |
| Pcr2 | 200 | 250 | 91 | 126 | 820 | 1003 | 373 | 516 | 450 | 217 | 1823 | 889 | 35 | 25 | 129 | 101 | 48% | 49% | 70% | 78,14% |
| Pcr3 | 175 | 100 | 279 | 392 | 336 | 192 | 534 | 750 | 275 | 671 | 528 | 1284 | 53 | 80 | 102 | 153 | 244% | 243% | 151% | 150,00% |
| Pcr4 | 50 | 180 | 38 | 53 | 359 | 1292 | 270 | 382 | 230 | 91 | 1651 | 652 | 92 | 11 | 660 | 79 | 40% | 39% | 12% | 12,00% |
| Pcr5 | 25 | 50 | 38 | 53 | 180 | 359 | 270 | 382 | 75 | 91 | 539 | 652 | 18 | 11 | 127 | 79 | 121% | 121% | 60% | 62,57% |
| Total | 2450 | 2810 | 2337 | 2701 | 4253 | 7111 | 2463 | 3445 | 5260 | 5038 | 11364 | 5908 | 255 | 257 | 2021 | 695 | 96% | 52% | 101% | 34,37% |

FIGURE 5.1 – Extrait de résultats issus d'expérimentation sur un échantillon de graphe de compétences.

Le modèle pseudo-électrique génère une légère perte de l'intensité globale sans modification de son écart type, sur l'échantillon étudié, perte en partie causée par les erreurs d'arrondi. Il produit une forte diminution de la tension globale (52%) et de son écart type (34%). Les résultats montrent que la répartition naturelle des flux électriques, pour l'échantillon étudié, limite la tension globale comme la dispersion des tensions, et augmente les disparités d'intensités sans modifier l'intensité globale.

Résultats sur les nœuds Cc_i

Pour les nœuds Cc_i , la somme des valeurs Ips, Ics et leur distribution entre les nœuds sont peu affectées. L'écart-type des valeurs Ips, Ics est lissé, passant, respectivement aux nœuds $\{Cc_1, Cc_2, Cc_3\}$, de $t\{101; 97; 247\}$ à $\{112; 113; 226\}$. Ces résultats montrent que l'application du modèle sur les nœuds accumulateurs Cci , produit des pseudo-valeurs électriques transformées, favorisant l'équilibrage entre valeurs produites et consommées sur l'ensemble de l'échantillon étudié.

Résultats sur les nœuds Cr_j

Les résultats obtenus montrent une diminution sensible des intensités circulant par les nœuds Cr_j (81% et 78%). Cette diminution s'accompagne d'une forte baisse de tension sur ces nœuds (27% et 26%), pour l'échantillon étudié. La topologie d'un graphe de compétences et l'application du modèle électrique tendent à redistribuer les intensités et tensions associées aux connaissances relatives, vers les connaissances clés.

Résultats sur les nœuds Pcr_k

Les nœuds Pcr_k sont les plus fortement affectés par l'application du processus de rétro-propagation de pseudo-valeurs électriques. Les variations d'intensités, produites et consommées, allant de 40% à 244%, sont corrélées aux variations de tension. Cette corrélation reflète le modèle électrique ($U = R.I$) et s'étend aux variations d'écart-type obtenues, s'étalant de 12% à 151%.

En moyenne, les nœuds Pcr_k uniquement connectés au nœud Cr_1 voient leurs valeurs d'intensité et de tension réduite à 68%. Ceux connectés au nœud Cr_2 voient leurs valeurs d'intensité et de tension réduite à 91%, leurs valeurs produites passant à 106% ($(Is + Ts)/(I/T)$) et leurs valeurs consommées passant à 68%. Le nœud

Pcr_3 , connecté aux nœuds Cc_2, Cr_1, Cr_2 , voit ses valeurs produites passer à 159% et ses valeurs consommées à 391%, soit une augmentation moyenne de 275%.

L'application du modèle aux nœuds Pcr_k produit un recentrage important de l'activité sur les individus directement connectés aux connaissances clés, ou connectés aux connaissances clés et à des connaissances relatives. L'équilibre entre connaissances produites et consommées est également affecté, un nœud Pcr_k initialement de tendance *producteur* pouvant devenir *fortement consommateur*, ou inversement (cf. figures 4.23 et 5.1).

5.1.3 Analyse d'une recommandation prédictive

L'observation des résultats produits sur l'échantillon de graphes de compétences étudié montre que notre modèle produit une recommandation qui peut aider à réduire le stress au travail, à intensité d'activité constante. Ceci peut favoriser la performance et l'amélioration du climat social.

Pour un ensemble de connaissances nommées, le principe expérimenté redistribue la charge sur les individus possédant les connaissances clés et lisse la charge des groupes d'individus possédant les connaissances relatives.

Deux groupes d'individus possédant des connaissances relatives se voient recommander des comportements différents, l'un étant enclin à diminuer uniformément et faiblement son activité sur un domaine de connaissance, l'autre étant invité à consommer significativement moins de connaissances, tout en produisant sensiblement plus de connaissances dans le domaine concerné. Pour ces derniers, nous comprenons cette recommandation comme une nécessité d'augmentation de compétences dans le domaine concerné, révélant un besoin de formation académique, ou de tutorat par les individus possédant déjà les compétences.

Sur l'échantillon étudié, l'individu le plus compétent et connecté - Pcr_3 - voit ses valeurs produites significativement augmentées (1,6) et ses valeurs consommées très fortement multipliées (3,6). L'augmentation conjointe de ces valeurs équilibre les diminutions recommandées aux individus moins compétents, uniquement connectés aux connaissances relatives. L'augmentation des valeurs produites semble être une recommandation pertinente pour l'optimisation de la performance, relativement à l'expertise visible du domaine de connaissance concerné que l'individu possède. Une seconde indication le concernant apparaît à l'observation des résultats. Pour cet expert, la très forte multiplication des valeurs consommées inverse la tendance initiale, entre valeurs produites et consommées. Produisant initialement 1,7 fois plus qu'il ne consommait, la recommandation lui indique de consommer désormais 1,4 fois plus qu'il ne produira, en plus de l'augmentation globale de charge (intensité et tension) lui incombant sur le domaine de connaissance concerné. Cette inversion peut indiquer une évolution du rôle de Pcr_3 dans le réseau de compétences étudié, enclin à devenir le principal consommateur de connaissances dans ce réseau. Le nouveau rôle de Pcr_3 pourrait s'assimiler à une activité d'instruction du flux de connaissances entrant, en vue de sa redistribution aux autres membres du réseau, sous le contrôle et la gestion de l'individu expert.

Cette analyse des résultats et l'interprétation qui en ressort semblent pertinentes, d'après les experts du domaine consultés - *i.e.*, Responsable Support *Testing*, Architecte Fonctionnel, Directeur Produit et Innovation. L'organisation actuelle des équipes partageant les données étudiées, correspond à un processus de distribution par niveau de complexité (*i.e.*, niveau 1/niveau 2/niveau 3). Dans ce processus, la distribution des données entrantes n'est que partiellement orientée par les compétences spécifiques de chacun. En phase initiale du processus, les étapes correspondent à une distribution pseudo-aléatoire des tâches au sein du groupe de travail, basée sur la disponibilité (variable) et sur le placement de chacun dans l'organigramme hiérarchique (constant, à court terme). Une entrée dans le processus n'est orientée vers l'expert compétent, qu'après avoir été traitée sans succès par d'autres membres du groupe de travail ; elle est alors orientée par le dernier membre qui l'a examinée, selon ses compétences propres et selon la perception interne qu'il possède de l'organisation - *i.e.*, une perception conditionnant la TIO de chacun, *cf.* section 2.2.1. La supervision du processus est donc floue durant les étapes initiales, puis relativement incertaine sur les données entrantes accédant aux étapes de dernier niveau.

Le modèle prédictif expérimenté implique et permet une organisation différente du groupe de travail. Dans cette organisation, le processus de distribution des données entrantes est centré sur les compétences. En phase initiale, les étapes correspondent alors à une distribution des tâches déterministe et probabiliste, et non pseudo-aléatoire. La supervision des étapes est également conditionnée de bout en bout et *a priori* par le degré d'expertise relatif au domaine de connaissances concerné, et non par la TIO et le niveau de complexité identifié *a posteriori*. L'analyse sommaire menée avec l'Architecte Fonctionnel a conclu à l'intérêt potentiel d'une telle organisation, et son application expérimentale au sein d'un groupe témoin visant à comparer la performance et le climat social avec l'organisation actuelle, a été évoquée.

5.1.4 Application aux graphes de compétences à la demande

Nous avons appliqué notre algorithme de propagation et rétro-propagation de valeurs (*cf.* annexes, section A.2) sur un réseau de compétences, construit à la demande à partir des mots-clés *validation*, *message*, *erreur*. La figure 5.2 présente l'effet du modèle prédictif de recommandations en 4 illustrations du réseau de compétences étudié, à chaque fois disposé dans l'espace par le même algorithme (*i.e.*, une spatialisation *Parallel Force Atlas*). Un gradient de couleur représente l'amplitude des valeurs, du bleu pour le minimum au rouge pour le maximum.

En figure 5.2, les 2 graphes de gauche sont colorisés et dimensionnés selon la différence d'*intensité sémantique produite moins consommée*, en haut, et d'*intensité équilibrée produite moins consommée*, en bas - *i.e.*, $I_s = I_{ps} - I_{cs}$ et $I = I_p - I_c$. Les 2 graphes de droite sont colorisés et dimensionnés selon la différence de *tension*

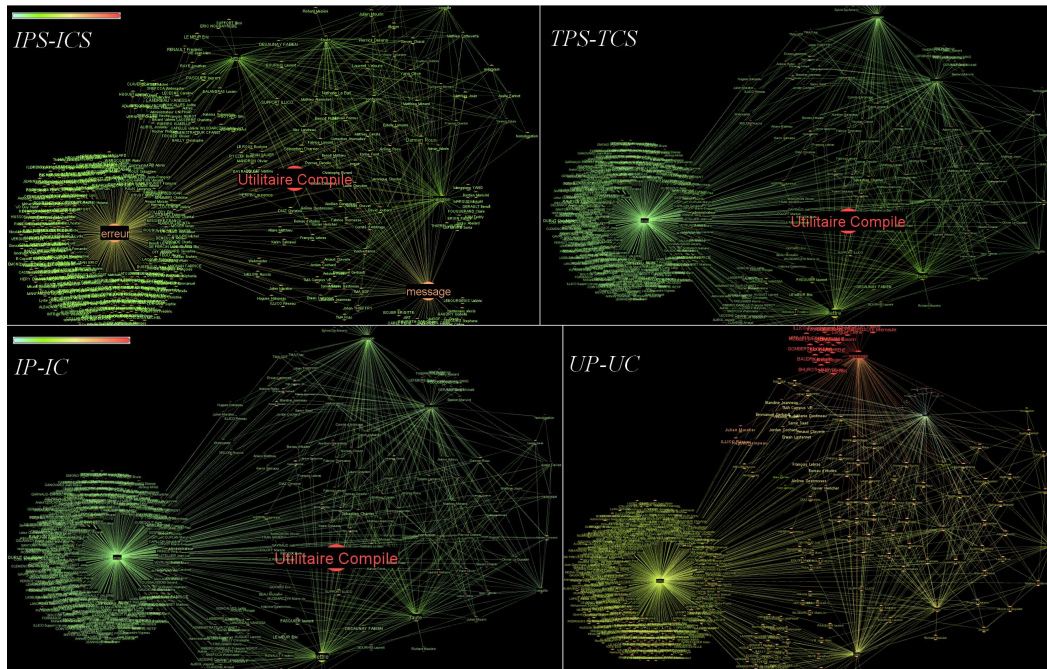


FIGURE 5.2 – Recommandation prédictive sur un réseau de compétences.

sémantique produite moins consommée, en haut et de *tension équilibrée produite moins consommée*, en bas - *i.e.*, $Ts = Tps - Tcs$ et $U = Up - Uc$. Les 2 graphes du haut de la figure 5.2 sont donc colorisés et dimensionnés en fonction des valeurs sémantiques et ceux du bas de la figure, en fonction des valeurs recommandées selon l'équilibre énergétique naturel - *i.e.*, la distribution naturelle des flux électroniques dans les groupements de résistances en série et/ou en parallèle².

La figure 5.3 présente la distribution des valeurs initiales au sein du réseau de compétences expérimental, pour les 2 graphes du haut de la figure 5.2.

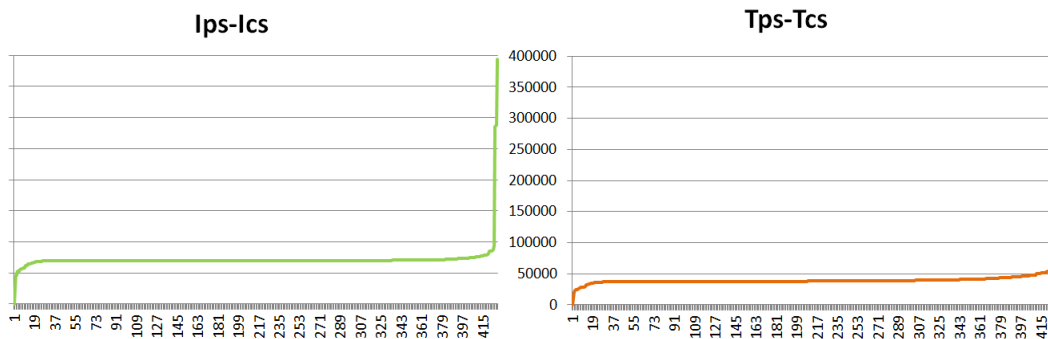


FIGURE 5.3 – Distribution de valeurs initiales.

2. Les valeurs calculées ont été normalisées dans \mathbb{R}_+ pour être chargées dans le logiciel de visualisation utilisé - *i.e.*, *Gephi*, <http://gephi.org>.

La figure 5.4 présente la distribution des valeurs équilibrées. En visualisant les 4 réseaux illustrés en figures 5.2, 5.3 et 5.4, nous constatons que :

1. les intensités sont plus fortes que les tensions, pour les valeurs initiales et les valeurs équilibrées - *i.e.*, couleurs plus chaudes illustrées en figure 5.2 ;
2. les valeurs initiales sont plus discriminantes que les valeurs équilibrées - *i.e.*, population moins importante des valeurs moyennes initiales, *cf.* figure 5.3 et température des couleurs en figure 5.2 ;
3. les valeurs de tension équilibrées ne présentent plus aucun *point chaud* (*cf.* nœud « Utilitaire Compile ») et sont fortement recentrées autour de la moyenne - *cf.* couleur et dimension des nœuds du graphe $Up - Uc$, en figure 5.2 et courbe illustrée au centre de la figure 5.3 ;
4. les maximas sont plus faibles pour les valeurs équilibrées que pour les valeurs initiales, environ 5 fois moindres pour les valeurs d'intensité et 60 fois moindres pour les valeurs de tension - *cf.* échelles verticales et courbes des figures 5.3 et 5.4 ;
5. les valeurs moyennes sont plus faibles pour les valeurs équilibrées que pour les valeurs initiales, environ 2,5 fois moindres pour les valeurs d'intensité et 14 fois moindres pour les valeurs de tension - *cf.* échelles verticales et courbes des figures 5.3 et 5.4.

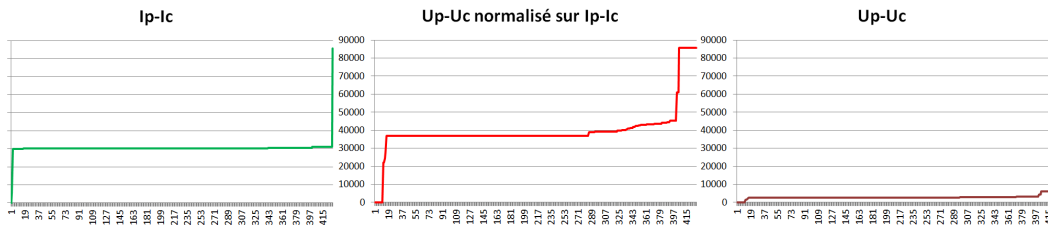


FIGURE 5.4 – Distribution de valeurs équilibrées.

Ces observations s'accordent avec l'analyse d'une recommandation prédictive effectuée sur un échantillon de réseau de compétences - *cf.* section 5.1.3. Elles confirment qu'à valeurs d'intensité constantes, notre modèle prédictif de recommandations régule et diminue les valeurs de tension au sein des réseaux de compétences. Dans l'hypothèse métaphorique introduite en section 3.1, puis définie et formalisée en section 3.6, les recommandations issues de notre modèle prédictif sont susceptibles de favoriser la performance des groupes de travail et l'amélioration du climat social.

En étudiant plus précisément les graphes illustrés en figure 5.2, nous avons commencé à identifier de nombreux autres aspects intéressants des résultats produits par notre modèle prédictif, basé sur l'équilibre énergétique naturel - *e.g.*, termes et

communauté en tension exhibés en haut du graphe $Up - Uc$ de la figure 5.2. Les contraintes du projet SOCIOPRISE ne permettent pas d’explorer en profondeur tous les aspects prometteurs, théoriques et applicatifs, décelés grâce à cette première approche. Pour cela, le travail effectué devra être poursuivi ultérieurement, dans un cadre différent.

Actuellement, le modèle multidisciplinaire et bidimensionnel d’analyse prédictive des réseaux de compétences que nous avons défini, produit (1) un principe de réduction de la tension au travail (*stress*), (2) des indications précises pour la gestion de la formation (plan prévisionnel) et (3) des recommandations pour la gestion du capital humain et social - *e.g.*, Gestion Prévisionnelle des Emplois et Carrières (GPEC). La validité de ces résultats semble indubitable dans le contexte du projet SOCIOPRISE. Sur un échantillon simple, simulé aléatoirement et sans prédispositions particulières, l’étude des résultats obtenus révèle de nombreuses indications et recommandations, semblant propices à l’amélioration de la performance et du climat social. Ces indications sont générées de façon automatique sur les réseaux de compétences à la demande. Le modèle que nous avons défini devrait toutefois être appliqué à l’échelle sur diverses données réelles et si possible, dans plusieurs contextes métiers, pour autoriser le passage à l’échelle industrielle.

5.2 Expérimentation de mesures sémantiques

Nous présentons les expériences et résultats obtenus par application de mesures de centralité sémantiques, définies selon la méthode générique présentée en section 3.5, sur un réseau global de compétences (section 3.4) et sur un réseau de compétences construit à la demande (section 3.2). Ces réseaux sont extraits des données réelles étudiées à l’aide du développement réalisé dans le cadre de SOCIOPRISE³. Les mesures expérimentées sont choisies de façon à évaluer la pertinence du facteur sémantique C_p (*cf.* section 3.5) sur des centralités de chemin géodésique et/ou de flux. Ce sont (1) une mesure d’intermédiarité de flux sémantique, (2) une mesure de proximité sémantique, et (3) modularité de communauté sémantique appliquée à la structure de réseau global de compétences et de réseau de compétences à la demande. Nous comparons finalement à la mesure de stress de [Shimbel 1953], une mesure de stress sémantique utilisant notre métrique de tension.

5.2.1 Données expérimentales

Les quatre mesures expérimentées sont appliquées sur les deux structures sémantiques de réseaux sociaux que nous avons définies. Le réseau global de compétences introduit en section 3.4 est calculé sur des données récentes (2011). C’est un graphe filtré dans lequel la valeur de tension sémantique produite ou consommée des éléments (nœud ou arc) est supérieure ou égale à 10 000. Il représente 2275 nœuds

3. Le développement logiciel est principalement réalisé par *OpenPortal Software*, en concertation permanente avec le LINA.

connectés par 28 500 arcs mutuels⁴. Sans filtrage, le réseau global de compétences extrait de la base de données étudiée se compose de 14 300 nœuds connectés par 1 620 000 arcs mutuels (*i.e.*, 810 000 arcs sortants ou entrants). Parmi ces 14 300 nœuds, 650 représentent des individus⁵ et 13 600 représentent les termes indexés⁶.

Le réseau de compétences à la demande utilisé dans l'expérimentation s'appuie sur les mêmes données que le réseau global. Il est construit par interrogation automatique, à l'aide du prototype développé avec *OpenPortal Software*, et respecte la structure définie en section 3.2. Il se compose de 428 nœuds, connectés par 1491 arcs mutuels. Parmi ces 428 nœuds, 416 représentent les individus, 3 représentent les connaissances clés (*i.e.*, *validation*, *message*, *erreur*) et 9 représentent les connaissances relatives⁷. Celles-ci sont plus ou moins pertinentes dans le réseau de compétences étudié, en rapport à l'aspect générique du référentiel employé durant la construction automatique du thésaurus intégré - *cf.* section 4.1.2.

5.2.2 Intermédiarité de flux sémantique

Appliquée aux graphes expérimentaux, l'intermédiarité de flux sémantique définie dans l'équation (3.7) (section 3.5.4) produit les résultats présentés ci-dessous.

5.2.2.1 Application au graphe global de compétences

La mesure est calculée à partir de la tension produite, puis à partir de la différence entre tension produite et consommée. Avec les valeurs produites, la longue traîne illustrée par la figure 5.5 montre des valeurs faibles pour la population étudiée (< 0.2) puisque la mesure peut générer un résultat normalisé dans l'intervalle $[0; 1]$. La dispersion est concentrée sur 9% de la population.

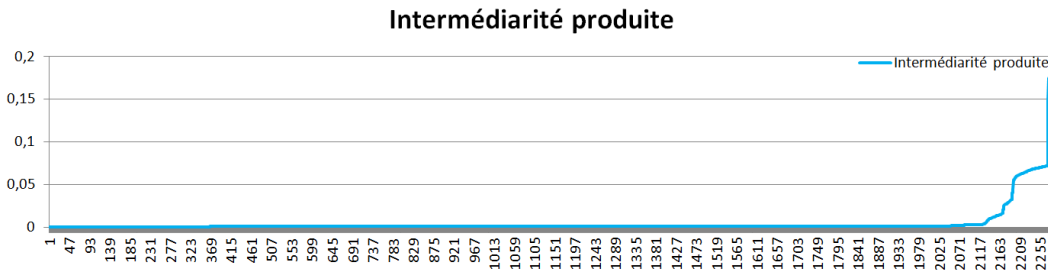


FIGURE 5.5 – Intermédiarité sémantique : distribution des valeurs de *tension produite*.

4. Les nœuds rendus isolés par le filtrage ont été supprimés.

5. Environ 10 individus fictifs ont été ajoutés par les experts aux individus réels, afin de répondre à des besoins particuliers du processus métiers - *e.g.*, « Utilitaire Compile ».

6. Il n'y a que des connaissances clés - *i.e.*, pas de connaissances relatives définies dans un thésaurus ou une ontologie - dans une structure de graphe global de compétences (*cf.* section 3.4).

7. Les connaissances relatives sont dénotées par les termes *confusion*, *coquille*, *faute*, *homologation*, *illusion*, *impair*, *inadvertance*, *lettre* et *quiproquo*.

L'utilisation de valeurs de tension produite moins consommée fait apparaître des valeurs négatives (déficit) et une dispersion concentrée sur une frange encore plus limitée (1%) de la population, illustrées en figure 5.6. Le calcul de la différence entre tension produite et consommée rend la mesure plus sélective et en améliore la pertinence en introduisant des valeurs négatives.

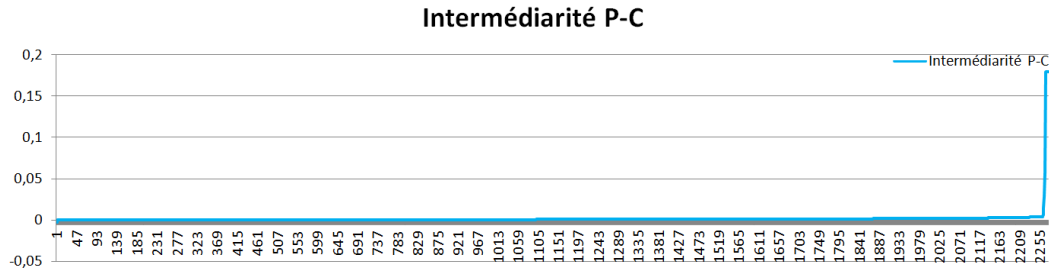


FIGURE 5.6 – Intermédialité sémantique : distribution des valeurs de *tension produite moins consommée*.

Le graphe présenté en figure 5.7 fait apparaître la topologie se profilant à l’affichage de 10% des nœuds du graphe. Les nœuds du réseau global de compétences sont dimensionnés et colorisés en fonction du coefficient de centralité d’intermédialité sémantique calculé sur la tension produite. Leur taille est proportionnelle à leur centralité et leur couleur passe du bleu au rouge, en passant par le vert et le jaune, entre les valeurs de centralité minimum ou maximum obtenues. Cette pre-

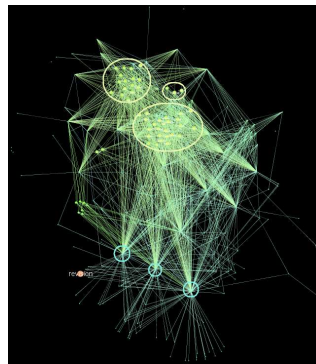


FIGURE 5.7 – Intermédialité sémantique : topologie du graphe global à 10% de nœuds affichés.

mière visualisation fait apparaître la prégnance du terme *révision* (en rouge), et la spatialisation utilisée⁸ fait apparaître des hubs⁹ (cercles bleus) et des amas d’autorités¹⁰ (cercles jaunes).

8. Le placement est de type attraction-répulsion - <http://gephi.org/plugins/parallel-force-atlas>.

9. Un hub est un nœud de degré élevé par rapport au degré moyen.

10. Une autorité est un nœud de poids fort par rapport au poids moyen.

Les hubs identifiés correspondent à des individus et les autorités correspondent aux termes prégnants et centraux du corpus, en terme de centralité d'intermédiarité - *i.e.*, des termes fortement impliqués dans les flux interconnectant les individus.

La figure 5.8 montre la topologie à l'affichage de 50% des nœuds du graphe. Cette visualisation fait apparaître deux nouveaux termes prégnants (*migration* et sa déclinaison plurielle, en rouge) et deux nouveaux amas de termes excentrés (cercles verts) présentant à la fois un poids et un degré significatifs.

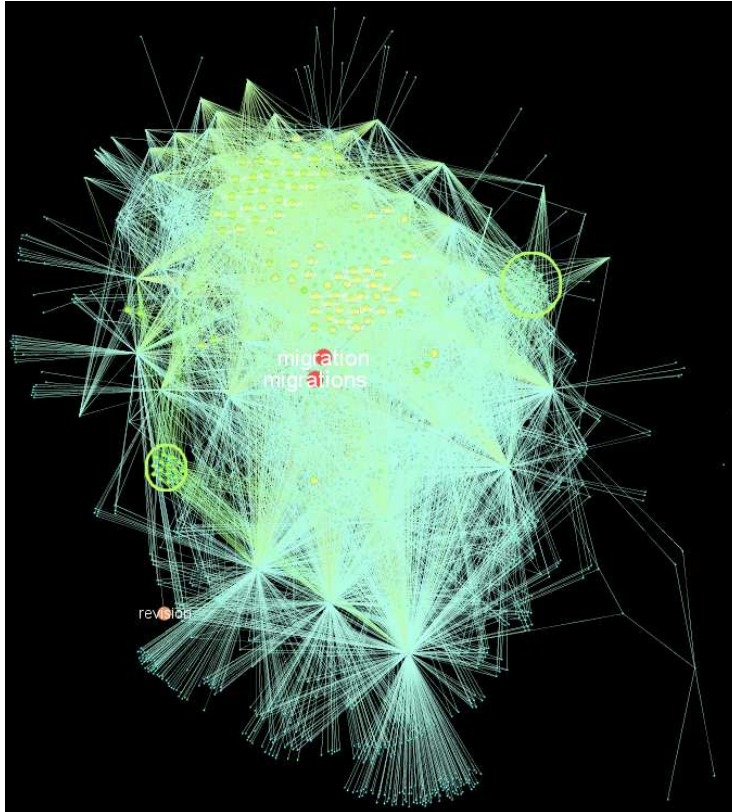


FIGURE 5.8 – Intermédiarité sémantique : topologie du graphe global à 50% de nœuds affichés.

La figure 5.9 présente une visualisation de 100% des nœuds du graphe variant de gauche à droite dans le paramétrage de la spatialisation. A gauche dans la figure, les nœuds sont plus ramassés, à droite les nœuds sont plus épars. La figure 5.9 facilite la visualisation des nœuds en ne présentant pas les arcs. Ces variantes de visualisation facilitent l'identification de particularités, détaillées ci-après. Un zoom sur les amas d'autorités illustrés en figure 5.9, permet d'identifier les connaissances en tension, relativement à la production de contenu endogène dans le corpus étudié. Dans la figure 5.10, ces connaissances sont indiquées par des nuages de termes récurrents - *i.e.*, des *amas d'autorités*. Le zoom permet d'identifier, par ordre décroissant de prégnance, les connaissances *migration*, *révision*, *devoir*, *faire*, *correction*, *sien*,

mien. Ces connaissances résument la sémantique des termes constituant les amas d'autorités. Elles semblent correspondre au point de vue expert de l'entreprise. Les indications obtenues confortent ainsi la vision intuitive et non vérifiée des experts, sur l'activité globale des équipes. L'apparition des termes *mien* et *sien*, parmi les connaissances prépondérantes apporte une information supplémentaire inattendue. La sémantique de ces termes induit des hyperonymes tels que *propriété* ou *appartenance* et dans le contexte, peut se trouver proche de *responsabilité*. L'antonymie des deux termes peut éventuellement suggérer la présence de conflits. Une approche par la psychologie sociale du champ sémantique impliquant ces connaissances devrait permettre d'identifier des leviers d'optimisation de la performance et du climat social au sein des équipes.

Enfin, la figure 5.11 présente une communauté d'experts plus ou moins spécialistes ou polyvalents. Les experts spécialistes sont les hubs excentrés de la masse principale, tandis que les experts polyvalents sont les hubs se situant près du centre de la masse graphique représentée.

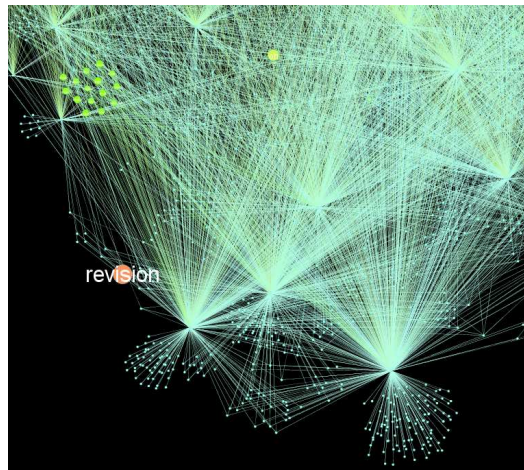


FIGURE 5.11 – Intermédierité sémantique : communauté d'experts, spécialistes ou polyvalents.

5.2.2.2 Application au graphe de compétences à la demande

Les nœuds intermédiaires sont ceux qui sont sur le plus grand nombre de plus faibles flux, parmi les plus faibles flux entre toutes les paires de nœuds du réseau étudié. L'application de notre mesure d'intermédierité sémantique sur le réseau de compétences expérimental (*cf.* section 5.2.1) a produit les résultats présentés en figure 5.12.

En figure 5.12, le terme *erreur*, dénotant une connaissance clé, est peu intermédiaire en terme de tension produite Tps , à gauche dans l'illustration. Il devient le principal intermédiaire lorsque que la centralité est calculée sur la différence

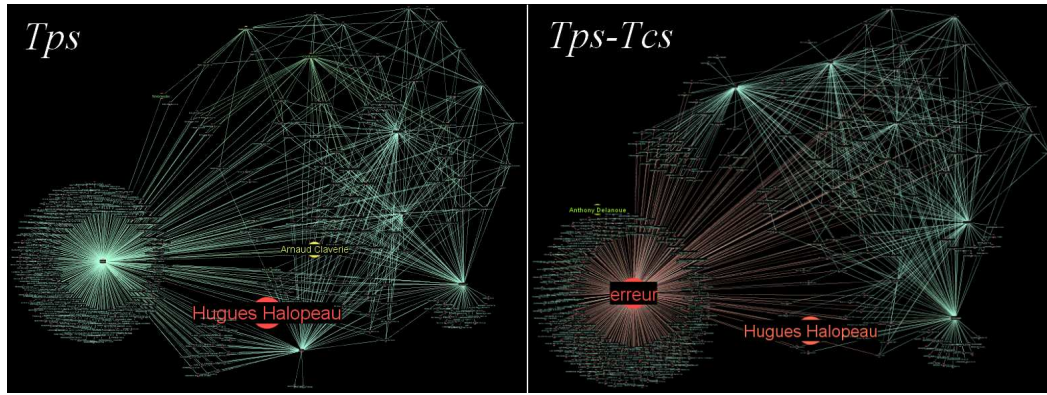


FIGURE 5.12 – Intermédiarité sémantique : tension produite *vs.* différence entre tension produite et consommée.

des valeurs de tension produites et consommées - *i.e.*, $Tps - Tcs$. Les 3 individus exhibés dans les deux illustrations de la figure 5.12 sont des personnes ayant ponctuellement produit de faibles quantité d'information impliquant plusieurs des connaissances représentées dans le graphe. Ce sont en pratique, un consultant de l'entreprise employé de longue date, apparaissant dans le réseau de tension produite Tps , et deux stagiaires, ayant travaillé quelques semaines dans l'entreprise.

La comparaison présentée en figure 5.12 est intéressante car elle montre que la centralité d'intermédiarité de flux sémantique est capable d'identifier (1) des polyvalents non experts et (2) des connaissances en tension, pour lesquelles la production d'information correspond quantitativement à la demande et réquisitionne un nombre significatif de personnes.

Conformément à notre définition, l'ensemble de ces résultats démontre que la centralité d'intermédiarité de flux sémantique est susceptible de produire des indications pertinentes concernant :

1. la représentation ordonnée des spécialistes et/ou polyvalents d'une ou plusieurs compétences désignées - *e.g.*, figure 5.8 ;
2. la reconnaissance de personnes centrales pour le partage d'information ou la collaboration relative à une ou plusieurs compétences - *e.g.*, figures 5.9 et 5.8 ;
3. la représentation du réseau des leaders possibles pour un projet défini - Chefs de Projet (CP), Comités - *e.g.*, figures 5.11 et 5.12 ;
4. l'identification de compétences manquantes dans une équipe projet - *e.g.*, figure 5.12.

5.2.3 Centralité de proximité de flux sémantique

Dans le contexte d'un RSEI, la centralité de proximité de flux sémantique (*cf.* section 3.5.5) est susceptible de fournir des indications telles que :

1. représentation ordonnée des spécialistes et polyvalents d'une ou plusieurs compétences, en fonction du niveau de possession global ou respectif de chaque compétence ;
2. identification d'un réseau de leadership classé pour un projet connu (CP, Comité directeur) ;
3. représentation hiérarchique des compétences relatives à une communauté, identification de compétences en déficit ou excédentaires au sein d'un équipe projet ;
4. identification des consommateurs, producteurs et diffuseurs d'information relativement à une ou plusieurs compétences dans le graphe social.

5.2.3.1 Application au graphe global de compétences

Appliquée à notre réseau global expérimental, ces indications sont présentées en s'appuyant sur la figure 5.13. À gauche de la figure, nous avons appliqué une centralité de proximité de [Beauchamp Murray 1965]. À droite, nous avons appliqué notre mesure en employant la différence de tension produite et consommée. L'algorithme de placement est différent pour les deux graphes illustrés.

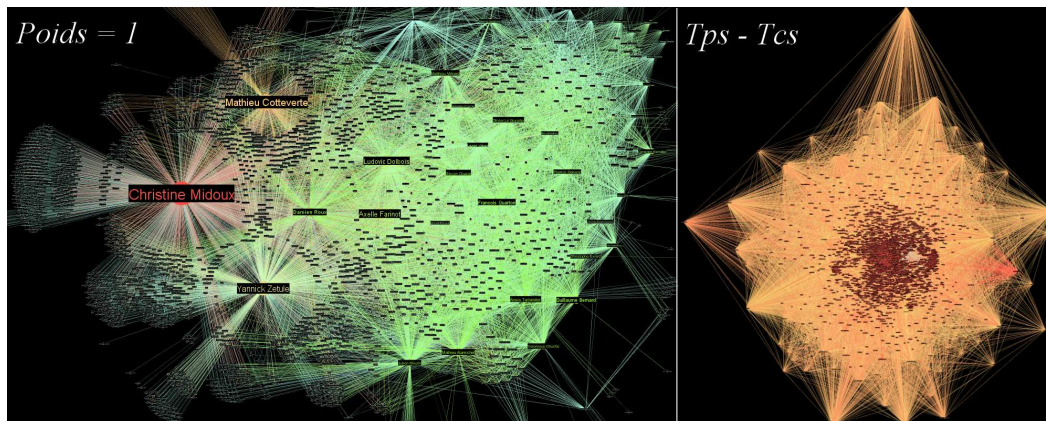


FIGURE 5.13 – Proximité sémantique : différence de tension produite moins consommée.

À l'analyse de la figure 5.13, nous constatons que sur un graphe global de compétences, la centralité de proximité de [Beauchamp Murray 1965] présente des similitudes avec notre centralité d'intermédierité de flux - *cf.* figure 5.8. L'illustration de droite présente les connaissances indexées du contenu endogène au centre et les individus en périphérie du graphe. La distance entre un individu et les connaissances auxquelles il est connecté dépend de la différence de tension produite et consommée. La couleur des nœuds et des arcs dépend de la centralité de proximité de flux sémantique.

La position de individus et la navigation interactive dans le réseau $Tps - Tcs$ de la figure 5.13, permet d'indiquer :

1. la représentation ordonnée des spécialistes et polyvalents d'une ou plusieurs compétences, en fonction du niveau de possession global ou respectif de chaque compétence ;
2. le réseau de leadership classé pour un projet connu (CP, Comité directeur) ;
3. les consommateurs, producteurs et diffuseurs d'information relativement à une ou plusieurs compétences dans le graphe social.

Les couleurs et la disposition des connaissances, relativement au centre du réseau $Tps - Tcs$ et aux personnes représentées, permettent de visualiser le classement des compétences relatives aux communautés et d'identifier des compétences en déficit ou excédentaires, au sein d'une équipe projet.

5.2.3.2 Application au graphe de compétences à la demande

Du fait de la faible profondeur hiérarchique du thésaurus implémenté dans notre modèle d'IC - cf. section 4.1.2, notre réseau de compétence expérimental a un diamètre égal à 4 - *i.e.*, la longueur du plus long chemin géodésique. Les centralités de proximité étant basées sur la distance entre les paires de nœuds, leur application au réseau de compétences expérimental présente un intérêt limité, illustré par le réseau affiché à 50% de la figure 5.14.

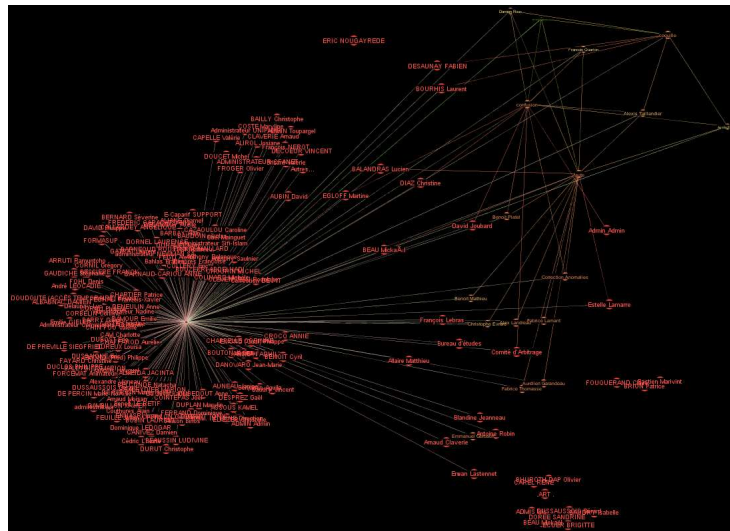


FIGURE 5.14 – Proximité sémantique : différence de tension produite moins consommée.

5.2.4 Modularité de communauté sémantique

La mesure de modularité sémantique que nous avons définie fonctionne sur notre structure de réseau de compétences - cf. section 3.5.6.

5.2.4.1 Application au graphe global de compétences

Rien n'interdit d'appliquer une modularité de communauté à la structure de réseau global de compétences (cf. section 3.4), mais la sémantique du contenu endogène n'ayant plus qu'une influence mineure sur la topologie du graphe étudié dans un tel cas, les résultats ne peuvent être comparés à celle d'une modularité de communauté sémantique.

5.2.4.2 Application au graphe de compétences à la demande

Nous appliquons la modularité de communauté définie dans [Blondel 2008] au réseau de compétences expérimental. La figure 5.15 présente le résultat obtenu.

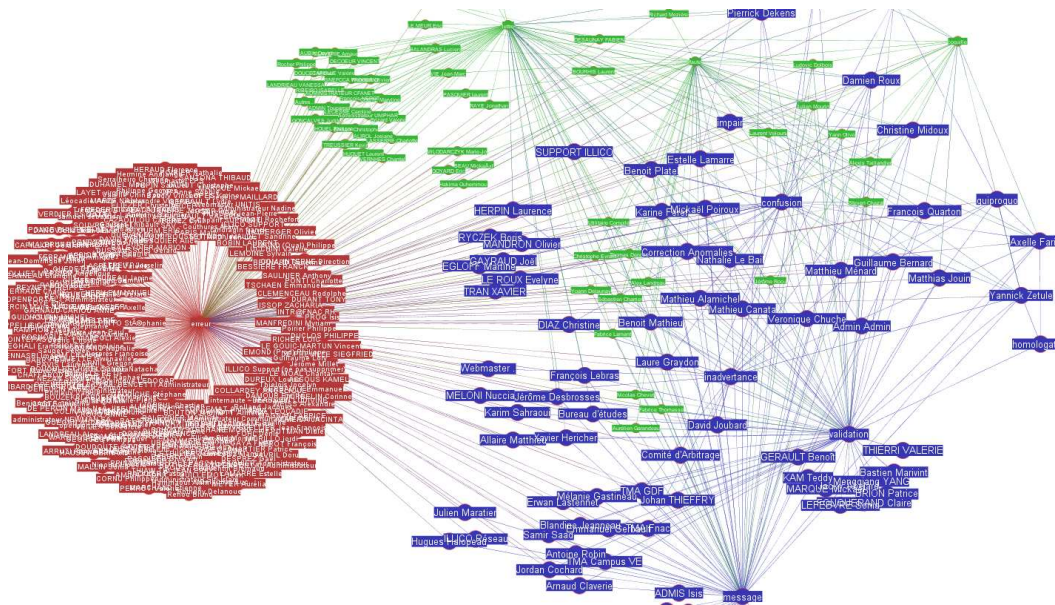


FIGURE 5.15 – Modularité de communauté sémantique appliquée à un réseau de compétences.

En figure 5.15, la dimension et la couleur des nœuds sont déterminées selon les trois communautés définies par le modularité de communauté au sein du réseau de compétences étudié. La disposition du réseau dépend également de ces communautés. L'observation du réseau montre une communauté majoritaire, interconnectée autour des connaissances dénotées par les termes *validation* et *message*, en bleu

dans l'illustration. La communauté de second rang, en rouge, est interconnectée autour des connaissances dénotées par le terme *erreur*, ce qui peut sembler surprenant si l'on considère le degré du nœud le représentant. Enfin, la dernière communauté est interconnectée autour des connaissances relatives dénotées par des termes tels que *faute*, *coquille* ou *lettre*, indépendamment des termes dénotant les connaissances clés. L'observation détaillée de ces résultats fait également apparaître la connectivité de termes *a priori* dissociés dans le thésaurus implémenté, tels que *message*, *validation*, *inadvertance*. Cette remarque pourrait éventuellement s'avérer perspicace, dans l'hypothèse d'un modèle d'auto-apprentissage sémantique contextuel, basé sur l'ARS sémantique.

L'observation de la figure 5.15 montre que la modularité de communauté sémantique permet d'identifier les compétences les plus en commun, les communautés possédant ces compétences et, ce qui apporte une vision pertinente du capital social, comment les connaissances spécifiques sont connectées et fédèrent les individus.

5.2.5 Mesure de stress et métrique de tension sémantique

Nous comparons la notion de *stress*, définie dans [Shimbel 1953], avec notre métrique de tension sémantique, en comparant la définition de [Shimbel 1953] au numérateur de notre centralité d'intermédiarité de flux sémantique - cf. section 3.5.4. L'évaluation se fait sur le réseau global de compétences expérimental.

5.2.5.1 Centralité de stress

La figure 5.16 présente l'application de la centralité de stress de [Shimbel 1953] au réseau global de compétences expérimental.

À l'observation de la figure 5.16, nous discernons un vaste ensemble de termes dénotant des connaissances éparses, ordonnées dans un classement peu distinctif au sein du graphe représenté. Des termes récurrents, tels que les inflexions des verbes *devoir* et *faire* sont visiblement mis en avant, indiquant une tendance générale à la nécessité d'action du groupe de travail analysé. Cette indication semble si naturellement évidente dans le contexte, qu'elle ne nécessite pas même l'avis des experts pour être vérifiée. Elle peut cependant manquer de pertinence, après réflexion.

5.2.5.2 Stress relatif à la métrique de tension consommée

La figure 5.17 présente l'application du numérateur de notre centralité d'intermédiarité de flux sémantique, basée sur les valeurs de *tension consommée* du réseau global de compétences expérimental. Dans le processus métier représenté par le contenu de la base de données endogène, la tension consommée est en général issue de la production d'information par les clients, et par les membres de l'entreprise lorsqu'ils se trouvent destinataires de l'information, au cours du processus métier. À l'observation de la figure 5.17, nous constatons que le nombre de termes mis en avant est nettement plus faible que précédemment, avec la mesure statique de stress de [Shimbel 1953]. Le classement des termes dénotant les connaissances exhibées

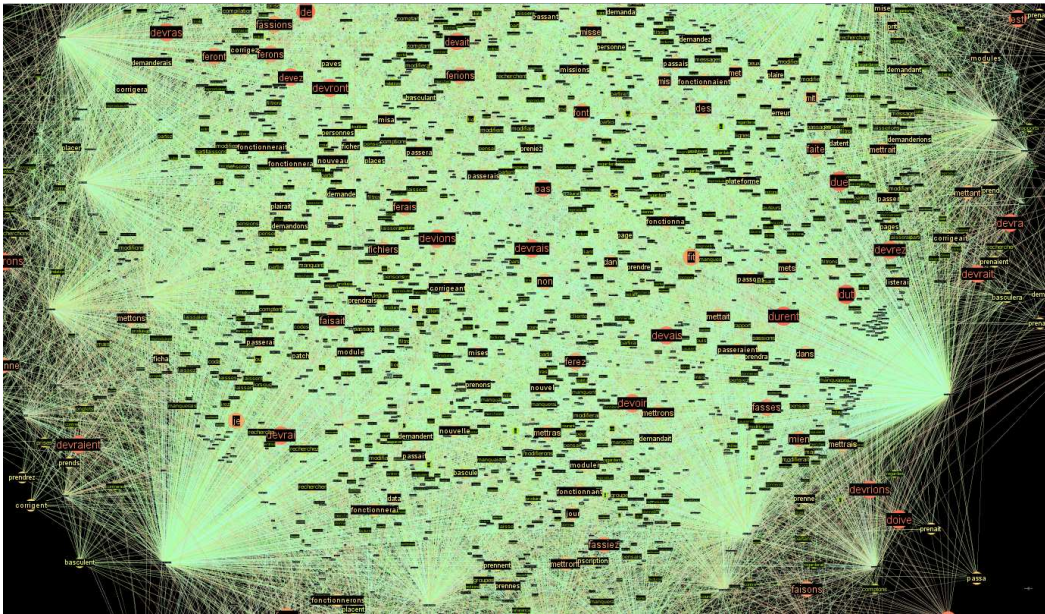


FIGURE 5.16 – Mesure statique de stress appliquée au réseau global de compétences expérimental.

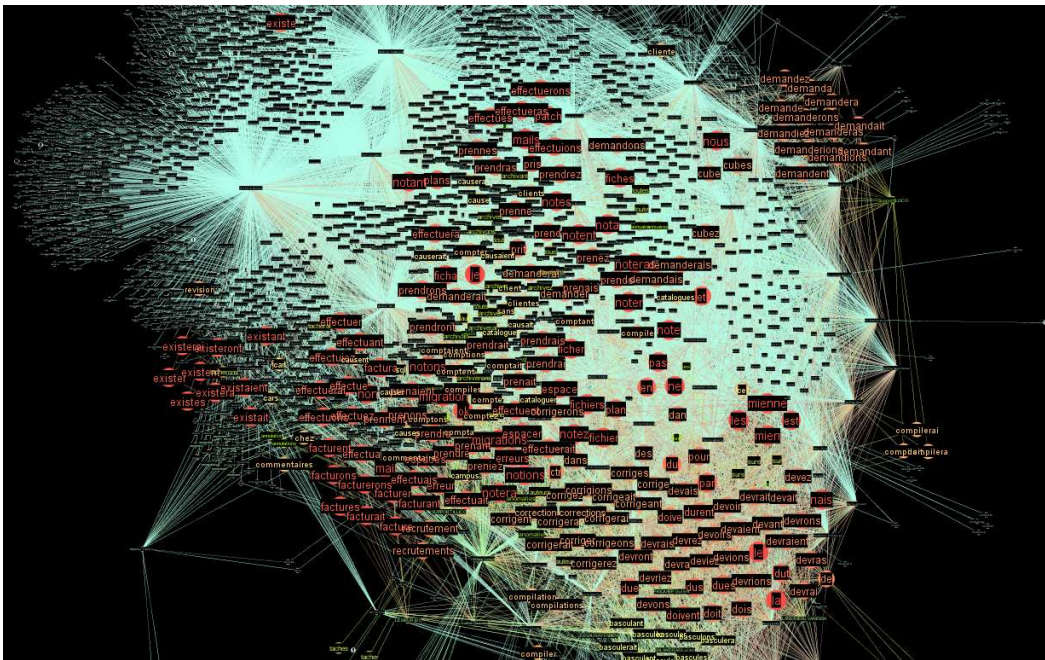


FIGURE 5.17 – Mesure sémantique du stress révélé par la tension consommée au sein du réseau global de compétences.

de *bascule* est majoritairement représentée par les inflexions du terme éponyme, alors que le terme *migration* et son pluriel, précédemment exhibés en figure 5.9 pour dénoter la connaissance centrale intermédiaire au sein du groupe de travail, sont relégués au second plan¹¹. La notion de *bascule*, révélée par la différence entre tension produite et consommée au sein du graphe global de compétences étudié, est susceptible de représenter la principale source de stress du groupe de travail en interne - *i.e.*, une source de stress n'étant pas directement issue de demandes explicites.

5.3 Bilan d'expérimentations

Nous avons expérimenté (1) un modèle prédictif de recommandations basé sur l'équilibre énergétique naturel, et (2) 4 mesures sémantiques, issues de nos contributions. En conclusion des résultats obtenus, nous avons d'abord pu valider empiriquement le modèle multidisciplinaire et multidimensionnel défini et implémenté¹², basé sur le rapprochement métaphorique de l'IC, de la force électromagnétique et de l'ARS. Nous avons ainsi démontré pragmatiquement le bien-fondé d'une hypothèse audacieuse.

Nous avons ensuite vérifié la correspondance applicative entre les mesures de centralités sémantiques définies en section 3.5, et les résultats obtenus par l'application de ces mesures à des réseaux de compétences construits automatiquement à partir de données réelles. À l'issue des vérifications, l'expérimentation a confirmé la perspicacité de notre analyse théorique.

Nous avons enfin profité de la disponibilité des données expérimentales pour comparer la mesure statique de stress de [Shimbel 1953], avec une mesure sémantique du stress révélé par les valeurs de tension au sein d'un réseau global de compétences. L'aspect informel et improvisé de cette dernière série d'expériences, opposé aux résultats observés, semble avoir confirmé la flexibilité et l'accessibilité de notre approche, et des structures graphiques que nous avons définies.

En conclusion, le bilan de ces expérimentations, conforté par l'avis favorable des experts, semble favorable à l'ouverture de nouvelles perspectives dans le cadre de nos travaux. Ces perspectives sont présentées dans le prochain et dernier chapitre - *cf.* chapitre 6.

11. En figure 5.9, la centralité d'intermédiarité sémantique est calculée en fonction des valeurs de tension produite.

12. La validation a été obtenue pour un échantillon simulé et pour un réseau de compétences construit automatiquement, à partir de données réelles.

Synthèse, perspectives et conclusion

Sommaire

| | |
|---|------------|
| 6.1 Comparaison à des travaux similaires | 99 |
| 6.1.1 Positionnement relatif à l'analyse statique | 99 |
| 6.1.2 Positionnement relatif à l'analyse sémantique | 100 |
| 6.1.3 Positionnement relatif à l'analyse dynamique | 102 |
| 6.2 Points critiques et carences | 103 |
| 6.2.1 Septicisme et légitimité en approches multidisciplinaires | 103 |
| 6.2.2 Pertinence de l'indexation sémantique | 104 |
| 6.2.3 Rigidité du processus de valuation de réseau de compétences | 105 |
| 6.2.4 Lisibilité des grands réseaux | 105 |
| 6.3 Perspectives | 106 |
| 6.3.1 Vers un renforcement du modèle d'analyse sémantique | 106 |
| 6.3.2 Vers un modèle prédictif basé sur l'électrodynamique | 107 |
| 6.4 Conclusion | 108 |

Ce chapitre final synthétise le bilan des travaux présentés. La section 6.1 établit le bilan de nos travaux en les comparant aux travaux équivalents introduits dans l'état de l'art. La section 6.2 formule une analyse critique de nos contributions.

6.1 Comparaison à des travaux similaires

Nous avons défini un modèle dans lequel l'ARS est abordée sous trois angles distincts : statique, sémantique et dynamique. Les contributions développées dans ce modèle sont foncièrement multidisciplinaires. Ce sont les aspects sémantiques et dynamiques de l'ARS qui les caractérisent, l'aspect statique bénéficiant déjà de nombreux acquis.

6.1.1 Positionnement relatif à l'analyse statique

Dans le contexte de *Socioprise*, notre principale contribution à l'ARS statique est constituée par, (1) le transfert de compétences scientifiques, (2) l'implémentation de modèles scientifiques par une société industrialisant ses solutions commerciales et (3) le développement des mesures standards d'ARS identifiées. D'un point de

vue scientifique, nos contributions à l'ARS statique ne sont pas originales, car elles ne sont que partiellement basées sur des mesures standards de centralité et de modularité, préalablement définies dans la littérature [Moreno 1934, Shimbel 1953, Freeman 1960, Freeman 1977, Newman 2005, Brandes 2005, Blondel 2008].

Alors que les réseaux sociaux intègrent désormais le monde des entreprises et des institutions, au cours du projet *Socioprise*, notre partenaire et nous avons été surpris du manque flagrant d'information sur l'existence et les possibilités offertes par l'ARS statique, au sein des communautés professionnelles concernées par les réseaux sociaux. Les usages professionnels des réseaux sociaux sont naissants, les volumes de données disponibles sont croissants, les besoins en indicateurs que peuvent déjà satisfaire partiellement les modèles d'ARS statiques sont évidents, mais sauf exception, les communautés d'experts n'ont pas encore saisi les opportunités qu'offre l'ARS. De par sa relative simplicité, l'ARS statique favorise l'accès à une forme innovante d'analyse décisionnelle en adéquation avec les réseaux sociaux, pour les analystes, gestionnaires et décideurs de nombreux secteurs d'activités.

Nos travaux dans le cadre du projet SOCIOPRISE ont contribué à pérenniser l'ARS statique, en la démocratisant auprès d'une société fournissant à de nombreux clients, PME et grands comptes, des solutions logicielles de gestion de la formation et du développement du capital humain. Si nos contributions à l'ARS statique sont clairement négligeables du point de vue strictement scientifique, elles peuvent toutefois s'avérer significatives en terme de valorisation, directe ou indirecte, des produits de la recherche au sein de communautés professionnelles diverses.

6.1.2 Positionnement relatif à l'analyse sémantique

L'une des constituantes de la multidisciplinarité de notre modèle d'ARS est l'Ingénierie des Connaissances (IC), sur laquelle nous l'avons fondé. Notre système d'indexation lexico-sémantique se compose, (1) d'un analyseur syntaxique évitant les mots vides, (2) d'une liste automatique de mots-bruits spécifiques au corpus étudié, et (3) d'un thésaurus automatique constitué de synonymes et d'hyponymes et/ou hyperonymes. Ce thésaurus, rudimentaire mais à construction automatique et spécifique au contenu endogène¹, est susceptible d'être remplacé par une ontologie. La construction automatique d'ontologie reste actuellement un défi difficile à relever [Wang 2004, Boonchom 2010] et l'éventualité d'une ontologie construite par les experts du domaine, généralement rédhitoire et incompatible avec les objectifs commerciaux de par son coût, est exclue dans le contexte de SOCIOPRISE.

Le modèle présenté dans [Delalonde 2007], plus centré sur l'IC que sur l'ARS, est comparable au socle sémantique développé dans nos travaux. Les deux modèles recourent à des mesures standards d'analyse linguistique de type *TF.IDF* (et/ou indice de *Jaccard*) [Jaccard 1901, Sparck Jones 1972, Salton 1986]. En comparaison des mesures de similarité et de proximité sémantique définies dans [Aimé 2011], il est clair que les mesures de prégnance linguistique [Robertson 1976] et de similarité sémantique [Rogers 1960] que nous utilisons peuvent sembler limitées. Ces mesures

1. Le modèle implémenté est également capable de traiter un corpus multilingue.

présentent l'avantage de pouvoir se passer d'ontologie. Elles répondent à un pré-requis et autorisent l'implémentation du facteur générique C_p que nous définissons, au cœur de notre contribution à l'ARS sémantique - cf. section 3.5.1.

Intégré à 3 mesures statiques d'ARS (cf. sections 3.5.3, 3.5.4 et 3.5.5), C_p a permis de définir 3 nouvelles mesures sémantiques de centralité. En ce sens, notre modèle ne peut être confondu avec celui présenté dans [Erétéo 2011] (cf. figure 2.5, section 2.1.3), expérimentant sur des réseaux sociaux préalablement pourvus d'annotations sémantiques², l'implémentation de mesures statiques de centralité à l'aide d'outils du Web sémantique³ - e.g., centralité d'intermediarité [Freeman 1977]. Actuellement, le modèle expérimenté dans [Erétéo 2011] représente l'état de l'art en ARS sémantique, à notre connaissance. Il présente l'avantage d'être applicable aux réseaux sociaux du Web. Notre modèle n'exploite pas toute la richesse des ressources du Web sémantique (e.g., thésaurus ou ontologie tels que *DBPedia*, *WordNet*, ou *OpenCalais*). Entièrement automatique, il présente cependant l'avantage de pouvoir fonctionner sans solliciter les experts du domaine de connaissances étudié. Il est également capable d'indexer le contenu textuel accessible sur le Web.

Nous avons défini et implémenté deux structures graphiques spécifiques, respectivement nommées *réseau de compétences* et *réseau global de compétences*. La topologie d'un réseau global de compétences est comparable à la topologie définie dans [Erétéo 2011]. La structure de réseau de compétences associe le réseau sémantique des connaissances présentes dans le contenu endogène (i.e., le corpus créé et partagé par les membres du réseau social étudié) au réseau social des individus associés à ces connaissances. Cette structure est spécifique à notre modèle multidisciplinaire et nous n'en connaissons aucune équivalence. Elle permet la détection de communautés de compétences - cf. sections 3.5.6 et 3.5.7.

Les éléments d'IC intégrés à notre modèle multidisciplinaire d'ARS forment une composante essentielle de son implémentation. Notre contribution à l'ARS sémantique se caractérise par la volonté de construire une passerelle scientifique entre l'ARS et l'IC, et de rendre cette passerelle accessible aux PME et grands comptes dont ce n'est pas le cœur de métier. Nous avons focalisé nos travaux sur la complexité à définir et intégrer les composantes scientifiques optimales, maximisant l'efficacité et l'accessibilité d'un modèle multidisciplinaire et multidimensionnel, en imposant des contraintes techniques ne nécessitant pas d'investissement lourd durant les phases de commercialisation, de personnalisation et d'exploitation, en terme de compétences scientifiques et techniques. Le modèle implémenté a ainsi permis le développement d'un nouveau prototype d'ARS sémantique, évolutif et favorable à l'industrialisation de solutions décisionnelles innovantes, couplant IC et ARS.

2. Les annotations sémantiques du modèle implémenté sont définies par les experts et superviseurs du domaine de connaissances relatif au corpus étudié.

3. Les outils du Web sémantique sont des langages et micro-formats tels que RDF, SKOW, OWL ou SPARQL, qui permettent de décrire, d'indexer et de rechercher la sémantique des ressources textuelles du Web.

6.1.3 Positionnement relatif à l'analyse dynamique

Parmi les domaines scientifiques associés aux graphes et réseaux, l'électrostatique - *e.g.*, OHM, KIRCHHOFF - et la découverte des phénomènes magnétiques et électrodynamiques - *e.g.*, AMPÈRE, ISING - marquent une continuité historique en matière d'analyse dynamique de structures et de flux graphiques. Le rapprochement métaphysique entre ARS et force électromagnétique (*i.e.*, l'une des 4 forces de l'Univers, en Physique moderne) se décline dans les deux sens, lorsque les physiciens recourent aux mesures de centralités pour prévenir le risque de défaillances dans les réseaux de puissance électriques [Wang 2010]. L'analyse dynamique de réseaux sociaux reste encore peu développée, et dans la continuité historique, nous avons développé une hypothèse audacieuse, basée sur [Newman 2005, Brandes 2005] et [Galam 2008].

Les modèles multidisciplinaires associant force électromagnétique et ARS produisent souvent des résultats pertinents [Newman 2005, Brandes 2005, Wang 2010]. Si l'on considère la profondeur de l'adéquation qu'ils définissent entre les deux disciplines, ces modèles sont pourtant incomplets.

[Newman 2005] et [Brandes 2005] ont introduit la notion de *résistance* et de *tension* électrique dans l'ARS. La multidisciplinarité de leurs approches respectives a permis de définir de nouvelles mesures de centralité, basées sur les flux plutôt que sur les chemins géodésiques. Dans [Wang 2010], les mesures standards de centralité (*i.e.*, des mesures d'ARS statique) sont associées à la notion d'*impédance* et de *puissance*, au sein de graphes représentant des réseaux électriques. Tous ces modèles s'appuient sur une adéquation partielle entre le modèle physique et le modèle graphique, car aucun ne prend en compte l'intégralité des lois définissant le comportement des flux électroniques dans les circuits électriques. Vouloir pallier totalement à ce manque est utopique, mais puisque les modèles de [Newman 2005, Brandes 2005, Wang 2010] ont déjà produit des résultats pertinents en l'état, il est logique de les faire évoluer en perfectionnant l'adéquation multidisciplinaire qu'ils ont introduit. Les travaux de S. GALAM, précisant avec rigueur l'adéquation entre physique statistique et sociologie statistique, s'inscrivent également dans cette approche [Galam 2008].

Le modèle d'ARS dynamique que nous avons expérimenté implémente une définition précise de l'adéquation multidisciplinaire introduite par [Newman 2005] et [Brandes 2005], en ARS statique. Notre contribution à l'ARS dynamique est basée sur l'introduction des lois d'OHM et de KIRCHHOFF dans la dimension graphique et dans la dimension sémantique de notre modèle. Elle est définie pour être implémentée sur notre structure de réseau de compétences, et respecte les règles de distribution des flux électroniques dans les groupements en série et/ou en parallèle de résistances. Notre modèle d'ARS dynamique représente ainsi une évolution significative de l'adéquation multidisciplinaire entre la force électromagnétique et l'ARS, depuis les travaux de [Brandes 2005] et [Newman 2005]. À notre connaissance, il n'existe aucune équivalence théorique ou applicative à notre modèle prédictif d'ARS dynamique, basé sur l'équilibre énergétique naturel.

6.2 Points critiques et carences

L'analyse critique de nos contributions étudie leurs points faibles et carences principales, du point de vue théorique de l'approche développée, puis du point de vue applicatif du modèle implémenté.

6.2.1 Septicisme et légitimité en approches multidisciplinaires

Lorsque, dans le contexte de nos travaux, nous avons voulu découvrir un modèle d'ARS permettant d'étudier l'optimisation quantitative et qualitative des flux de travail, les phénomènes de *tension* nerveuse (stress), de *réactance* et de *burn-out* au sein des réseaux sociaux professionnels, c'est probablement d'une réflexion intuitive, que l'évidence d'analogie entre les flux d'information, les réseaux de connaissances et les flux électriques est apparue. Nous ne pouvons ni démontrer, ni réfuter formellement l'origine étymologique et sémantique commune de termes interdisciplinaires tels que *tension* et *réactance*, définis par une même forme symbolique pour le vocabulaire commun et scientifique, et ce dans plusieurs langues - *i.e.*, tension nerveuse, tension/réactance électrique, réactance psychologique. Les fondements de l'Univers, de la pensée et du langage sont encore peu connus, et si des formes de vie civilisées étaient apparues ailleurs dans l'Univers quelques dizaines de milliers d'années avant que cela ne se produise sur Terre (*i.e.*, le temps d'une respiration à l'échelle chronologique de l'Univers), nous pourrions éventuellement les voir, mais devrions certainement évoluer encore autant de milliers d'années avant d'approcher la compréhension de leurs connaissances actuelles. En ce sens, nos travaux peuvent inspirer un scepticisme légitime, comme la centralité d'intermédiation [Newman 2005] a pu le faire avant de devenir une référence incontournable de l'ARS.

Cependant, le modèle d'ARS statique, sémantique et dynamique que nous avons défini et expérimenté autorise la production d'outils innovants de gestion du capital humain et social, en produisant des indicateurs visuels et des recommandations prédictives pour l'optimisation de la performance et la réduction du risque psychosocial. Le pragmatisme tendant à reconnaître comme vrai ce qui fonctionne réellement, d'un point de vue pragmatique, nos contributions démontrent empiriquement l'*unité d'échelle* (*i.e.*, une unité d'espace et de temps) entre IC, ARS et force électromagnétique que la multidisciplinarité de notre modèle exhibe, à l'instar de l'approche intrinsèque aux travaux de [Newman 2005, Brandes 2005, Galam 2008, Wang 2010].

En règle générale, plusieurs aspects critiques freinent le développement de contributions multidisciplinaires. D'une part, il n'est pas toujours évident d'imaginer et d'assembler des informations issues de disciplines éloignées pour produire une unité d'échelle, à partir d'impressions souvent aussi difficiles à appréhender intuitivement que la « fractalité » du monde objectif de B. MANDELBROT (1924-2010) et les *attracteurs étranges* de E.N. LORENZ (1917-2008) [Mandelbrot 1982, Lorenz 1969] - cf. figure 6.1, attracteur de Lorenz extrait de *Nature journal*, 2000/08/31.

D'autre part, il est rare d'être en mesure de réunir au sein d'un même projet, les ressources nécessaires et suffisantes pour que les contributions propres à chaque

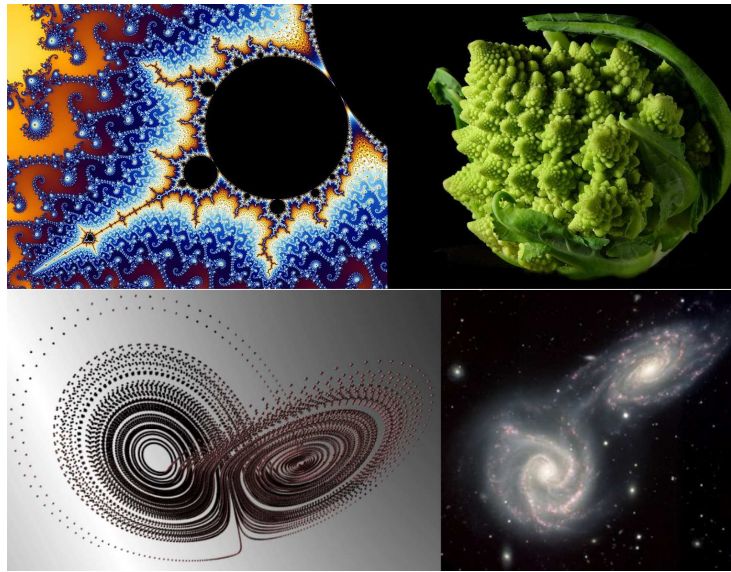


FIGURE 6.1 – Fractales et attracteurs étranges dans la Nature.

discipline soient significatives, prises séparément, sans que le projet ne deviennent trop complexe pour être appréhendé dans son ensemble.

En pratique, l'implémentation et l'expérimentation du modèle développé ont permis de révéler des points critiques et carences précises, sur les aspects suivants de nos contributions :

1. pertinence de l'indexation sémantique ;
2. rigidité du processus de valuation de réseau de compétences ;
3. difficulté de visualisation du réseau global de compétences.

6.2.2 Pertinence de l'indexation sémantique

Le prototype expérimenté implémente un thésaurus à construction automatique, basé sur des dictionnaires génériques de langue française. Ce choix garantit une adaptation relativement aisée au traitement de corpus multilingues, l'environnement logiciel disposant d'analyseurs lexicaux pour environ 50 langues. Les résultats obtenus dans le contexte du projet SOCIOPRISE ont révélé les limites de référentiels lexicaux génériques, pour la recherche sémantique d'information dans les corpus professionnels.

Lors d'une recherche sur le terme « serveur », les experts du domaine informatique souhaitent pouvoir identifier les ressources indexées comportant des termes dont le signifié est proche ou similaire, tels que :

1. des acronymes - *e.g.*, IIS (pour *Internet Information Server*) ;
2. des voisins orthographiques ou linguistiques - *e.g.*, *Server* ;
3. des hyponymes du domaine - *e.g.*, *Apache*, *Tomcat*, *Oracle*.

Le prototype permet l'ajout manuel de termes dans les référentiels lexicaux (*cf.* section 4.1.1). Un terme ajouté par les experts est conservé et chaque mise à jour de l'index de recherche le prend alors en compte, autorisant l'identification des termes voulus lors de la recherche sémantique d'information. Le solution que représente cette méthode implique de recourir aux experts de chaque domaine de connaissances concerné, dans le contexte de réseaux sociaux professionnels.

Nous avons également relevé des aberrations sémantiques contextuelles pour le terme « serveur », telles que : « serveuse, servante, barman ». La solution de contournement de ces problèmes serait donc de remplacer les dictionnaires génériques par des dictionnaires spécialisés. Pour certains secteurs et selon les langues à indexer, ces référentiels existent et peuvent rapidement être intégrés au prototype (*e.g.*, droit, médecine); pour d'autres, il peut s'avérer obligatoire de définir préalablement les référentiels nécessaires.

6.2.3 Rigidité du processus de valuation de réseau de compétences

Le processus de valuation multidimensionnelle des réseaux de compétences à la demande (*cf.* sections 4.2 et 4.3) a été spécifiquement défini pour des réseaux basés sur un thésaurus à un seul niveau de hiérarchie sémantique (i.e. synonymie et/ou subsomption unique). L'évolution du prototype vers un modèle d'IC implémentant un thésaurus, ou une ontologie comportant plusieurs niveaux hiérarchiques, impliquerait le changement de profondeur des réseaux de compétences à la demande, puisque ceux-ci représentent partiellement le référentiel sémantique implémenté.

Pour permettre l'évolution du modèle d'IC, il est donc indispensable de redéfinir le modèle théorique de valuation des réseaux de compétences à la demande de façon dynamique, afin de pouvoir l'implémenter sur des graphes de profondeur variable. Une telle évolution est réalisable, moyennant un travail raisonnable de *refactoring*.

6.2.4 Lisibilité des grands réseaux

Lors des expérimentations, la compréhension visuelle du graphe global de compétences que nous avons étudié s'est révélée difficile, à cause de la dimension du graphe et de la quantité d'information représentée - *cf.* section 5.2. Cette difficulté acceptable dans un cadre expérimental, constitue un frein important à l'appropriation du modèle par des non-spécialistes, malgré sa pertinence. Nous avons identifié ce problème avec les experts *OpenPortal Software* et envisagé conjointement les solutions suivantes :

1. représentation partielle par filtrage ;
2. visualisation et manipulation tridimensionnelles ;
3. factorisation de graphes et calcul d'arbre couvrant.

La représentation partielle par filtrage consiste simplement à filtrer la quantité d'éléments d'un réseau global de compétences apparaissant durant sa visualisation, à partir de seuils variables, applicables sur les valeurs de pondération et/ou sur la cardinalité des relations relatives aux éléments du graphe global de compétences -

e.g., valeurs de tension, d'intensité ou de résistance, centralités sémantiques, degré des nœuds. La colorisation du graphe en fonction de ses propriétés en favorise également la compréhension. De telles fonctions sont généralement implémentées dans les applications et bibliothèques logicielles dédiées à la visualisation de graphes.

La visualisation en trois dimensions permet une meilleure perception visuelle des réseaux globaux de compétences. Elle autorise la navigation interactive en immersion au sein des grands graphes de terrain [Greffard 2012]. La 3D stéréoscopique, solution innovante se généralisant dans les systèmes de visualisation (*e.g.*, projecteurs, écrans plats), est devenue accessible aux PME et industries via des plateformes interactives ouvertes au grand public - *e.g.*, *Kinect/XBox 360*.

Enfin, la factorisation de graphes et le calcul d'arbre couvrant pourra faciliter la compréhension d'un réseau global de compétences, en produisant des sous-graphes basés sur le graphe complet et représentant les informations essentielles du graphe complet selon les critères de simplification choisis par l'utilisateur. Divers algorithmes et techniques de factorisation de graphes et de calculs d'arbres couvrants ont été définis et publiés dans la littérature [Hong Tuan 1995, Propp 1998].

6.3 Perspectives

Nous plaçons deux thèmes prioritaires en perspective du modèle multidisciplinaire et multidimensionnel que nous avons défini : (1) renforcer notre modèle d'IC, afin d'améliorer la pertinence des résultats de recherche d'information sémantique produits par le prototype implémenté, dans les divers contextes professionnels et linguistique visés pour la diffusion de solutions logicielles commerciales ; (2) perfectionner encore l'adéquation entre force électromagnétique et ARS, en introduisant la Loi d'AMPÈRE dans notre modèle, en vue de définir un modèle prédictif électrodynamique, basé sur l'induction électromagnétique.

La section 6.3.1 présente les grandes lignes des hypothèses de travail que nous souhaitons développer à moyen terme, en vue de renforcer le modèle d'IC que nous avons implémenté. L'hypothèse présentée en section 6.3.2 introduit un modèle de structure graphique prédictive, basée sur l'application de lois électrodynamiques sur les réseaux de compétences. Ce modèle n'a pas été encore développé. Il reste une perspective à implémenter et éprouver, dans l'objectif de produire un prototype prédictif de l'évolution topologique des réseaux de compétences, basé sur l'induction électromagnétique.

6.3.1 Vers un renforcement du modèle d'analyse sémantique

Dans le contexte de SOCIOPRISE, le développement de méthodes d'IC basées sur les ontologies est restreint par les contraintes du projet. Cependant, nous avons initialement envisagé l'usage d'ontologies et de langages du web sémantique (*e.g.*, RDF, OWL), au lieu d'un simple thésaurus automatique. Les travaux de [Zhuhadar 2011] captent les mots-clés les plus fréquemment recherchés par les utilisateurs pour guider la construction d'ontologies. L'index des ressources textuelles est plongé dans

l'ontologie construite, permettant une approche visuelle du réseau sémantique intrinsèque au contenu. Le modèle constitue une amélioration pertinente de notre vision initiale, en ajoutant une fonction cartographique des connaissances à notre première approche. Dans [Bhattacharyya 2011], la similarité des termes du contenu décrivant un individu est évaluée en s'appuyant sur *WordNet* pour construire le graphe hiérarchique, nommée forêt conceptuelle, des connaissances présentes dans le contenu endogène d'un réseau social. Cette méthode peut également enrichir notre modèle d'IC, bien qu'il faille s'attendre à rencontrer des imprécisions dues à l'emploi d'un lexique générique, telles que déjà constatées (sections 4.1.1 et 4.1.2). Afin de renforcer notre modèle d'IC, il peut être opportun de combiner notre approche aux modèles définis dans [Zhuhadar 2011] et [Bhattacharyya 2011].

6.3.2 Vers un modèle prédictif basé sur l'électrodynamique

A l'aide du modèle physique sous-jacent, nous introduisons la loi d'AMPÈRE définissant l'induction électromagnétique, définie dans [Thovex 2010e] et basée sur [Gardiol 1987], accompagnée de seuils limitant l'oscillation constante. Le modèle obtenu permet d'anticiper l'apparition et la disparition d'arcs du graphe de compétences, pour un intervalle de temps $[t, t']$. La figure 6.2 montre un effet possible du modèle sur une partie d'un graphe de compétences.

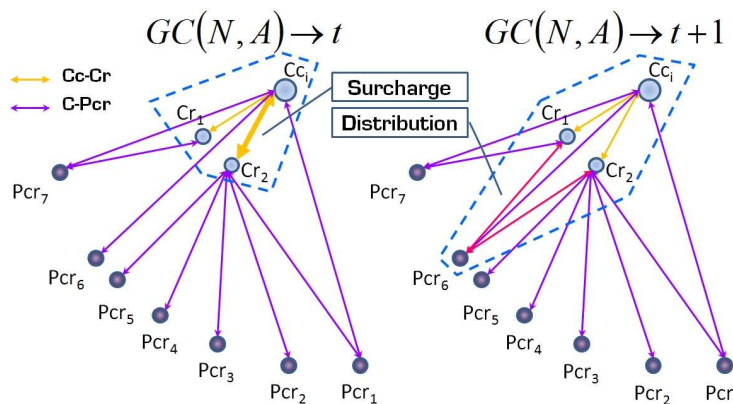


FIGURE 6.2 – Evolution prédictive d'un sous-graphe de compétences de t à $t + 1$.

Dans $GC(N, A) \rightarrow t$ (cf. figure 6.2), du fait de flux électrodynamiques forts entre $[Pcr_1 \dots Pcr_5]$ et Cr_2 , en fonction d'un seuil haut, (Cr_2, Cc_1) est surchargé - pour les flux orientés de $[Pcr_1 \dots Pcr_5]$ vers Cr_2 et Cc_1 . Il en émane une force électromagnétique forte - *i.e.*, une attraction ou répulsion. Par rapport à Cc_1 , Cr_2 est proche de Cr_1 et Pcr_5 est proche de Pcr_6 . Ainsi, dans $GC'(N, A) \rightarrow t + 1$, la surcharge se propage à Cr_1 et Pcr_6 , faisant apparaître les arcs (Pcr_6, Cr_1) et (Pcr_6, Cr_2) , en redistribuant la puissance dans (Cr_2, Cc_1) , (Cr_1, Cc_1) et $[Pcr_1 \dots Pcr_5]$.

La structure du graphe de compétences a évolué, en laissant apparaître Pcr_6 comme un nouvel individu polyvalent relativement au domaine de connaissance de

Cc_1 , alors qu'il était auparavant spécialiste de Cc_1 . De plus, puisque la structure et les flux du réseau ont évolué, les mesures statiques et sémantiques retourneront des résultats différents, produisant de nouvelles recommandations.

La profondeur des changements peut être ajustée avec les seuils haut/bas et un *rayon de proximité*. Le processus peut également être exécuté en cascade, pour une séquence d'intervalles de temps, afin d'obtenir une visualisation continue de l'évolution du réseau de compétences.

6.4 Conclusion

Le cœur de notre contribution à l'ARS se trouve dans le modèle transdisciplinaire et multidimensionnel, dynamique et non probabiliste que nous avons défini pour l'optimisation de la performance et du bien-être professionnel. À notre connaissance, nous expérimentons le premier modèle unifiant les mesures classiques de l'analyse des réseaux sociaux et l'ingénierie des connaissances avec l'électrodynamique, définissant de nouvelles mesures sémantiques d'ARS et métriques sociophysiques, et proposant un modèle d'ARS prédictif non probabiliste. Ce modèle repose sur une structure propre de *graphe de compétences*.

Nos travaux visent à définir un modèle à la fois statique, dynamique et sémantique d'Analyse de Réseaux Sociaux d'Entreprises et d'Institutions (ARSEI). Ils fondent une forme innovante d'outils d'aide à la décision, pour la gestion du capital humain et social dans les entreprises et institutions. Leur originalité réside, d'une part, dans l'intégration au sein d'un même modèle des aspects statiques, dynamiques et sémantiques de l'ARSEI et, d'autre part, dans la définition de contributions fondées sur des apports multidisciplinaires. Ces contributions introduisent (1) un modèle générique d'extension sémantique pour l'analyse statique des réseaux sociaux, (2) 5 nouvelles mesures sémantiques de centralité et de clustering sémantique, (3) une structure graphique et sémantique supportant l'analyse statique et dynamique, (4) des métriques sémantiques et sociophysiques pour l'évaluation de la *tension* et de la *réactance*, et (5) un modèle sémantique prédictif d'analyse socio-dynamique des réseaux sociaux professionnels et institutionnels, basé sur l'équilibre naturel énergétique de la force électromagnétique.

D'un point de vue applicatif, le partenariat avec *OpenPortal Software* se poursuit au terme du projet SOCIOPRISE. D'un point de vue théorique, en vue de renforcer l'ingénierie des connaissances implémentée dans le modèle développé, nous souhaitons tester le principe de visualisation défini dans [Zhuhadar 2011] et les fonctions de génération de forêt et de similarité sémantique, définies dans [Bhattacharyya 2011]. Afin de favoriser l'appropriation de nos résultats par les non-spécialistes, nous prévoyons d'étudier les algorithmes et techniques de factorisation de graphes et de calculs d'arbres couvrants, facilitant la visualisation de réseau global de compétences [Hong Tuan 1995, Propp 1998]. En vue d'améliorer l'étude prédictive des réseaux sociaux, la métaphore développée sera complétée en recourant à l'effet JOULE pour estimer l'*échauffement* des composants du réseau social d'entreprise et pour pré-

venir les risques de dégradation des performances, d'instabilité, de *burn-out* et de troubles psychosociaux. Notre contribution multidisciplinaire couplant ainsi électrodynamique et thermodynamique, nous espérons introduire de nouvelles heuristiques et contributions dédiées à l'ARS dynamique.

Nous réfléchissons à certains principes de percolation de réseaux présentés dans [Barrat 2006], en vue d'introduire dans notre approche multidisciplinaire de l'ARS une dimension mécanique, destinée à l'identification de connaissances enfouies et latentes au sein de graphes de compétences géolocalisés. Appliqué au Web social, le modèle pressenti vise à anticiper les tendances populaires et économiques au sein d'un périmètre géosocial. Enfin, d'un point de vue plus fondamental, nous nous intéressons à la métaconnaissance de l'*unité d'échelle multidisciplinaire*, de la philosophie des sciences cognitives à la percolation des modèles mathématiques, physiques, biologiques et psychologiques dans l'ARS.

La perspective théorique de l'approche présentée est de formaliser « un modèle sémantique pluridisciplinaire et multidimensionnel d'analyse statique et dynamique des réseaux sociaux professionnels élargis » - *i.e.*, un modèle complexe applicable à un bassin d'entreprises et institutions. La perspective applicative est d'assister le développement de la performance territoriale par l'optimisation des forces vives, tout en favorisant la réduction du risque psychosocial et le bien-être professionnel.

Annexes

Sommaire

| | | |
|------------|---|------------|
| A.1 | Extrait annoté de documentation technique sur la classification SQL Server | 111 |
| A.1.1 | TF.IDF avec CONTAINSTABLE et FREETEXTTABLE . . . | 111 |
| A.1.2 | Indice de JACCARD avec ISABOUT | 113 |
| A.2 | Algorithme de propagation et de rétro-propagation | 115 |

A.1 Extrait annoté de documentation technique sur la classification SQL Server

Les notes de l'auteur livrées ci-dessous s'appuient sur la documentation technique SQL Server 2005 et SQL Server 2008 R2 - <http://msdn.microsoft.com/en-us/library/ms142524.aspx>.

A.1.1 TF.IDF avec CONTAINSTABLE et FREETEXTTABLE

CONTAINSTABLE

Extrait de documentation technique.

StatisticalWeight = $\text{Log}_2((2 + \text{IndexedRowCount}) / \text{KeyRowCount})$

StatisticalWeight = $\text{Log}_2((2 + \text{nb documents indexés}) / \text{nb documents contenant le terme})$

IDF [Sparck Jones 1972] : $\log(\text{nombre total de documents dans le corpus} / \text{nombre de documents où le terme apparaît})$

TF [Salton 1986] : $\text{Nombre d'occurrences d'un terme dans un doc} / \text{nombre termes du doc}$

Rank = $\min(\text{MaxQueryRank}, \text{HitCount} * 16 * \text{StatisticalWeight} / \text{MaxOccurrence})$

Rank = $\min(1000, 16 * \text{Nombre d'occurrences du terme dans un doc} * \text{Log}_2((2 + \text{nb documents indexés}) / \text{nb documents contenant le terme}) / \text{Nb de termes ds document})$

Rank = min(1000, Log2((2 + nb documents indexés) / nb documents contenant le terme)*16*Nombre d'occurrences du terme dans un doc / Nb de termes ds document)

Rank = min(1000, 16*IDF*TF)

Conclusion : Rank est une forme simple de TF-IDF

FREETEXTABLE

Extrait de documentation technique et de [Robertson 1976]

Rank = $\Sigma[\text{Terms in Query}] w \left(\left(\frac{1 + tf}{K + tf} \right) * \left(\frac{k3 + 1}{k3 + qtf} \right) \right)$

Where :

w is the Robertson-Sparck Jones weight. In simplified form, w is defined as :

$w = \log_{10} \left(\left(\frac{r + 0.5}{N - R + r + 0.5} \right) / \left(\frac{R - r + 0.5}{n - r + 0.5} \right) \right)$

N is the number of indexed rows for the property being queried.

n is the number of rows containing the word.

K is $(k1 * (1 - b) + (b * dl / avdl))$.

dl is the property length, in word occurrences.

avdl is the average length of the property being queried, in word occurrences.

k1, b, and k3 are the constants 1.2, 0.75, and 8.0, respectively.

tf is the frequency of the word in the queried property in a specific row.

qtf is the frequency of the term in the query.

R est l'ensemble des documents annotés pertinents (si tags utilisateurs) [Zaragoza 2004]. r est l'ensemble des documents annotés pertinents (si tags utilisateurs) contenant le terme calculé [Zaragoza 2004].

Si un terme est un mot clé ou une inflection, son rank est calculé. Si c'est un synonyme il prend le rank du mot clé.

$w = \log_{10} \left(\left(\frac{r + 0.5}{N - R + r + 0.5} \right) / \left(\frac{R - r + 0.5}{n - r + 0.5} \right) \right)$

$w = \log_{10} \left(\left(\frac{\text{nb documents annotés contenant terme} + 0.5}{\text{nb documents indexés} - \text{nb documents annotés} + \text{nb documents annotés contenant terme} + 0.5} \right) / \left(\frac{\text{nb doc retournés contenant terme} - \text{nb documents annotés contenant terme} + 0.5}{\text{nb documents annotés} - \text{nb documents annotés contenant terme} + 0.5} \right) \right)$

Conclusion partielle : w est un IDF raffiné. Dans un système sans annotation :

$w = \log_{10} \left(\left(\frac{0.5 * (\text{nb documents indexés} + 0.5)}{\text{nb documents indexés} + 0.5} \right) / \left(\frac{0.5 * (\text{nb doc retournés contenant terme} + 0.5)}{\text{nb doc retournés contenant terme} + 0.5} \right) \right)$

Extrait de documentation technique :

K is $(k1 * ((1 - b) + (b * dl / avdl)))$.

$K = (1.2 * (0.25 + (0.75 * \text{nb termes ds doc retourné} / \text{moyenne nb termes ds docs indexés})))$.

$\text{Rank} = \Sigma[\text{Terms in Query}] w(((1 + 1) \text{tf}) / (K + \text{tf})) * ((k3 + 1) \text{qtf} / (k3 + \text{qtf}))$

$\text{Rank} = \Sigma[\text{Terms in Query}] w((2 \text{tf}) / (K + \text{tf})) * (9 \text{qtf} / (8 + \text{qtf}))$

$\text{Rank} = \Sigma[\text{Terms in Query}] w * ((2 \text{nb occ terme ds un doc/nombre termes du doc}) / ((1.2 * (0.25 + (0.75 * \text{nb termes ds doc retourné} / \text{moyenne nb termes ds docs indexés}))) + (\text{nb occ terme ds un doc/nombre termes du doc})) * ((9 \text{nb occ terme corpus/nb termes corpus}) / (8 + (\text{nb occ terme corpus/nb termes corpus})))$

$\text{Rank} = \Sigma[\text{Terms in Query}] \log_{10}((\text{nb documents indexés} + 0.5) / (\text{nb doc retournés contenant terme} + 0.5)) * ((2 \text{nb occ terme ds un doc/nombre termes du doc}) / ((1.2 * (0.25 + (0.75 * \text{nb termes ds doc retourné} / \text{moyenne nb termes ds docs indexés}))) + (\text{nb occ terme ds un doc/nombre termes du doc})) * ((9 \text{nb occ terme corpus/nb termes corpus}) / (8 + (\text{nb occ terme corpus/nb termes corpus})))$

A.1.2 Indice de Jaccard avec ISABOUT

Extrait de documentation technique :

The ranking is computed for each term in the query and then combined as described below.

ContainsRank = same formula used for CONTAINSTABLE ranking of a single term (above).

Weight = the weight specified in the query for each term. Default weight is 1.

WeightedSum = $\Sigma[\text{key}=1 \text{ to } n] \text{ContainsRankKey} * \text{WeightKey}$

$\text{Rank} = (\text{MaxQueryRank} * \text{WeightedSum}) / ((\Sigma[\text{key}=1 \text{ to } n] \text{ContainsRankKey}^2) + (\Sigma[\text{key}=1 \text{ to } n] \text{WeightKey}^2) - (\text{WeightedSum}))$

$\text{ContainsRank} = \min(1000, 16 * \text{IDF} * \text{TF})$

$\text{WeightedSum} = \Sigma[\text{key}=1 \text{ to } n] \min(1000, 16 * \text{IDF} * \text{TF}) * \text{WeightKey}$

$\text{Rank} = (\text{MaxQueryRank} * \text{WeightedSum}) / ((\Sigma[\text{key}=1 \text{ to } n] \text{ContainsRankKey}^2) + (\Sigma[\text{key}=1 \text{ to } n] \text{WeightKey}^2) - (\text{WeightedSum}))$

$\text{Rank} = (1000 * (\Sigma[\text{key}=1 \text{ to } n] \min(1000, \text{Log}_2((2 + \text{nb documents indexés}) / \text{nb documents contenant le terme}) * 16 * \text{Nombre d'occurrences du terme dans un doc} / \text{Nb de termes ds document}) * \text{WeightKey})) / ((\Sigma[\text{key}=1 \text{ to } n] \min(1000, \text{Log}_2((2 + \text{nb documents indexés}) / \text{nb documents contenant le terme}) * 16 * \text{Nombre d'occurrences du terme dans un doc} / \text{Nb de termes ds document})^2) + (\Sigma[\text{key}=1 \text{ to } n] \text{WeightKey}^2) - (\Sigma[\text{key}=1 \text{ to } n] \min(1000, \text{Log}_2((2 + \text{nb documents indexés}) / \text{nb documents contenant le terme}) * 16 * \text{Nombre d'occurrences du terme dans un doc} / \text{Nb de termes ds document})))$

/ nb documents contenant le terme)*16*Nombre d'occurrences du terme dans un doc / Nb de termes ds document)* WeightKey))

$$\text{Rank} = (1000 * (\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)* \text{WeightKey})) / ((\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)^2) + (\sum[\text{key}=1 \text{ to } n] \text{WeightKey}^2) - (\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)* \text{WeightKey}))$$

L'indice de similarité de JACCARD est défini par $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ [Jaccard 1901].

ROGERS et TANIMOTO étendent cette définition, et formalisent la similarité $T(X, Y) = \frac{\sum_i (A_i \wedge B_i)}{\sum_i (A_i \vee B_i)}$ [Rogers 1960],

Lorsque les données sont binaires, $\sum_i (A_i \wedge B_i) = A.B$ et $\sum_i (A_i) = |A^2|$. Ces deux indices correspondent à :

$$S(A, B) = \frac{A.B}{|A^2| + |B^2| - A.B}$$

Rank est un raffinement de l'indice de JACCARD sur des données binaires, considérant

$$A = \sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)$$

$$B = \text{WeightKey}$$

$$|A^2| = n = |B^2|$$

Conclusion

$$\text{Rank} = (1000 * (\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)* \text{WeightKey})) / ((\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)^2) + (\sum[\text{key}=1 \text{ to } n] \text{WeightKey}^2) - (\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)* \text{WeightKey}))$$

Avec WeigthKey=1 (valeur par défaut) :

$$\text{Rank} = (1000 * \sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)) /$$

$$(\sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF)^2 + \sum[\text{key}=1 \text{ to } n] 1^2 - \sum[\text{key}=1 \text{ to } n] \min(1000, 16*\text{IDF}*TF))$$

$$S(A, B) = \frac{1000 A}{|A^2| + |B^2| - A} = \frac{1000 A}{2n - A}$$

A.2 Algorithme de propagation et de rétro-propagation

Pour tout graphe de compétences construit à la demande, nous avons défini :

1. un processus de propagation de valeurs sémantiques basé sur $(TF.IDF)$ et sur les lois d'OHM ;
2. un principe de transformation de valeurs de flux sémantiques en valeurs pseudo-électriques, basé sur la loi des mailles de KIRCHHOFF ;
3. un principe d'équilibrage de valeurs pseudo-électriques, basée sur la loi des nœuds de Kirchhoff ;
4. un processus de rétro-propagation de valeurs pseudo-électriques équilibrée, basé sur les lois d'OHM.

Nous présentons le processus complet implémentant ces 4 éléments sous la forme d'un algorithme écrit en pseudo-code. A l'initialisation de l'algorithme, les nœuds et les arcs du graphe $GC(N, A)$ sont connus, les valeurs initiales Is, Us, R sont connues pour les arcs de $aPcr$ et les valeurs $R(Cc_i)$ sont connues.

```

/* PROPAGATION DE VALEURS SEMANTIQUES */
/* Mise à jour des  $apcr(Cr_j, Pcr_k), R$  en fonction des  $acr(Cc_i, Cr_j), R$ 
intensités et tensions sémantiques des nœuds  $Cr_j$  et intensités sémantiques des arcs  $acr(Cc_i, Cr_j)$  */
For each  $Cr_j$ 
sumRCc = 0, sumRPcr = 0, lps = 0, lcs = 0
For each  $acr(Cc_i, Cr_j)$ 
sumRCc = sumRCc + 1/acr(Cr_j, Cc_i).R // somme des inverses de résistances
End For
 $Cr_j, RtCc = 1/\text{sumRCc}$  //résistance totale du groupement  $\langle Cr_j, Cc_i \rangle$ 
For each  $apcr(Cr_j, Pcr_k)$ 
 $apcr(Cr_j, Pcr_k).R = Cr_j, RtCc$  //  $apcr(Cr_j, Pcr_k).R$  dépend de la résistance du groupe  $\langle Cr_j, Cc_i \rangle$ 
 $apcr(Cr_j, Pcr_k).Tps = apcr(Cr_j, Pcr_k) * R$  // ajuste T selon R  $\langle Cr_j, Cc_i \rangle$ 
 $apcr(Cr_j, Pcr_k).Tcs = apcr(Cr_j, Pcr_k) * R$  // ajuste T selon R  $\langle Cr_j, Cc_i \rangle$ 
sumRPcr = sumRPcr + 1/apcr(Cr_j, Pcr_k).R // somme inverses de résistances pour retro-propagation
lps = lps + apcr(Cr_j, Pcr_k).Ips // somme partielle des intensités produites
lcs = lcs + apcr(Cr_j, Pcr_k).Ics // somme partielle des intensités consommées
End For
 $Cr_j, RtPcr = 1/\text{sumRPcr}$  //résistance totale du groupement  $\langle Cr_j, Pcr_k \rangle$  pour retro-propagation
 $Cr_j, Ips = lps$  // intensité produite de  $Cr_j$ 
 $Cr_j, Ics = lcs$  // intensité consommée de  $Cr_j$ 
 $Cr_j, Tps = Cr_j, RtCc * Cr_j, Ips$  // tension produite de  $Cr_j$ 
 $Cr_j, Tcs = Cr_j, RtCc * Cr_j, Ics$  // tension consommée de  $Cr_j$ 
/* Intensités sémantiques des arcs  $acr(Cc_i, Cr_j)$ 
For each  $acr(Cc_i, Cr_j)$  // dans le groupement  $\langle Cr_j, Cc_i \rangle, Ir = It * (Rt/R)$ 
 $acr(Cc_i, Cr_j).Ips = Cr_j, Ips * (Cr_j, RtCc / Cc_i, R)$ 
 $acr(Cc_i, Cr_j).Ics = Cr_j, Ics * (Cr_j, RtCc / Cc_i, R)$ 
End For
End For
/* Résistance totale groupement  $\langle Pcr_k, Cc_i | Cr_j \rangle$ ,
intensités et tensions sémantiques des nœuds  $Pcr_k$  */
For each  $Pcr_k$ 
sumR = 0, lps = 0, lcs = 0
For each  $apcr(Pcr_k, Cc_i)$ 
sumR = sumR + 1/apcr(Pcr_k, Cc_i).R // somme partielle des inverses de résistances
lps = lps + apcr(Pcr_k, Cc_i).Ips // somme partielle des intensités produites
lcs = lcs + apcr(Pcr_k, Cc_i).Ics // somme partielle des intensités consommées
End For
For each  $apcr(Pcr_k, Cr_j)$ 
sumR = sumR + 1/apcr(Cr_j, Pcr_k).R // somme partielle des inverses de résistances
lps = lps + apcr(Pcr_k, Cr_j).Ips // somme partielle des intensités produites
lcs = lcs + apcr(Pcr_k, Cr_j).Ics // somme partielle des intensités consommées
End For
 $Pcr_k, Rt = 1/\text{sumR}$  // Résistance totale de groupement  $\langle Pcr_k, Cc_i | Cr_j \rangle$ 
 $Pcr_k, Ips = lps$  // intensité produite
 $Pcr_k, Ics = lcs$  // intensité consommée
 $Pcr_k, Tps = Pcr_k, Rt * Pcr_k, Ips$  // tension produite
 $Pcr_k, Tcs = Pcr_k, Rt * Pcr_k, Ics$  // tension consommée
End For

```

```

/* Résistance totale de groupement  $(Cc_i, Pcr_k | Cr_j)$ ,
   intensités sémantiques des nœuds  $Cc_i$ ,
   équilibrage des intensités de Cc */
Tips = 0, Tics = 0
For each  $Cc_i$ 
  sumR = 0, Ips = 0, Ics = 0
  For each  $apcr(Cc_i, Pcr_k)$ 
    sumR = sumR + 1/apcr( $Pcr_k, Cc_i$ ).R // somme partielle des inverses de résistances
    Ips = Ips + apcr( $Pcr_k, Cc_i$ ).Ips // somme partielle des intensités produites
    Ics = Ics + apcr( $Pcr_k, Cc_i$ ).Ics // somme partielle des intensités consommées
  End For
  For each  $apcr(Cc_i, Cr_j)$ 
    sumR = sumR + 1/[apcr( $Cr_j, Pcr_k$ ).R +  $Cr_j.RtPcr$ ] // somme partielle série-parallèle
    Ips = Ips + apcr( $Pcr_k, Cr_j$ ).Ips // somme partielle des intensités produites
    Ics = Ics + apcr( $Pcr_k, Cr_j$ ).Ics // somme partielle des intensités consommées
  End For
   $Cc_i.Rt = 1/sumR$  // Résistance totale de groupement  $(Pcr_k, Cc_i | Cr_j)$  pour rétro-p.
   $Cc_i.Ips = Ips$  // intensité produite
   $Cc_i.Ics = Ics$  // intensité consommée
  Tips = Tips + Ips // somme des intensités produites
  Tics = Tics + Ics // somme des intensités consommées
End For
Ims = (Tips + Tics)/2 // moyenne des intensités produites et consommées de Cc
For each  $Cc_i$  // Equilibrage des intensités de Cc
   $Cc_i.Ip = Cc_i.Ips / Tips * Ims$ 
   $Cc_i.Ic = Cc_i.Ics / Tics * Ims$ 
End For

/* RETRO-PROPAGATION DE PSEUDO-VALEURS ELECTRIQUES */
/* Intensités et tensions des arcs  $apcr(Cc_i, Cr_j)$  et des nœuds  $Cr_j$  */
For each  $acr(Cc_i, Cr_j)$  // dans le groupement  $(Cc_i, Cr_j, Pcr_k)$ ,  $l(Cc_i, Cr_j) = It * (Rt/R(Cc_i, Cr_j, Pcr_k))$ 
   $apcr(Cc_i, Cr_j).Ip = Cc_i.Ip * [Cc_i.Rt / (acr(Cc_i, Cr_j).R +  $Cr_j.RtPcr$ )]$  // intensité produite
   $apcr(Cc_i, Cr_j).Ic = Cc_i.Ic * [Cc_i.Rt / (acr(Cc_i, Cr_j).R +  $Cr_j.RtPcr$ )]$  // intensité consommée
   $apcr(Cc_i, Cr_j).Up = Cc_i.Ip * Cc_i.Rt$  // tension produite
   $apcr(Cc_i, Cr_j).Uc = Cc_i.Ic * Cc_i.Rt$  // tension consommée
End For
/* Intensités et tensions des nœuds  $Cr_j$  */
For each  $Cr_j$ 
  IpCr = 0, IcCr = 0,
  For each  $acr(Cc_i, Cr_j)$ 
    IpCr = IpCr + apcr( $Cc_i, Cr_j$ ).Ip // intensité produite  $Cr_j$ 
    IcCr = IcCr + apcr( $Cc_i, Cr_j$ ).Ic // intensité consommée  $Cr_j$ 
  End for
   $Cr_j.Ip = IpCr$  // total intensité produite  $Cr_j$ 
   $Cr_j.Ic = IcCr$  // total intensité consommée  $Cr_j$ 
   $Cr_j.Up = Cr_j.RtPcr * Cr_j.Ip$  // tension produite  $Cr_j$ 
   $Cr_j.Uc = Cr_j.RtPcr * Cr_j.Ic$  // tension consommée  $Cr_j$ 
End For

```

```

/* Intensités et tensions des arcs  $apcr(Cc_i, Pcr_k)$  */
For each  $apcr(Cc_i, Pcr_k)$  // dans le groupement  $(Cc_i, Pcr_k), l(Cc_i, Pcr_k) = It * [Rt / R(Cc_i, Cr_j, Pcr_k)]$ 
   $apcr(Cc_i, Pcr_k).Ip = Cc_i.Ip * [Cc_i.Rt / apcr(Cc_i, Pcr_k).R]$  // intensité produite
   $apcr(Cc_i, Pcr_k).Ic = Cc_i.Ic * [Cc_i.Rt / apcr(Cc_i, Pcr_k).R]$  // intensité consommée
   $apcr(Cc_i, Pcr_k).Up = Cc_i.Ip * Cc_i.Rt$  // tension produite
   $apcr(Cc_i, Pcr_k).Uc = Cc_i.Ic * Cc_i.Rt$  // tension consommée
End For

/* Intensités et tensions des arcs  $apcr(Cr_j, Pcr_k)$  */
For each  $apcr(Cr_j, Pcr_k)$  // dans le groupement  $(Cr_j, Pcr_k), l(Cr_j, Pcr_k) = It * [Rt / R(Cr_j, Pcr_k)]$ 
   $apcr(Cr_j, Pcr_k).Ip = Cr_j.Ip * [Cr_j.Rt / apcr(Cr_j, Pcr_k).R]$  // intensité produite
   $apcr(Cr_j, Pcr_k).Ic = Cr_j.Ic * [Cr_j.Rt / apcr(Cr_j, Pcr_k).R]$  // intensité consommée
   $apcr(Cr_j, Pcr_k).Up = Cr_j.Ip * Cr_j.Rt$  // tension produite
   $apcr(Cr_j, Pcr_k).Uc = Cr_j.Ic * Cr_j.Rt$  // tension consommée
End For

/* Intensités et tensions des nœuds  $Pcr_k$  */
For each  $Pcr_k$ 
   $IpPcr = 0, IcPcr = 0$ 
  For each  $apcr(Cc_i, Pcr_k)$ 
     $IpPcr = IpPcr + apcr(Cc_i, Pcr_k).Ip$  // intensité produite  $Pcr_k$ 
     $IcPcr = IcPcr + apcr(Cc_i, Pcr_k).Ic$  // intensité consommée  $Pcr_k$ 
  End for
  For each  $apcr(Cr_j, Pcr_k)$ 
     $IpPcr = IpPcr + apcr(Cr_j, Pcr_k).Ip$  // intensité produite  $Pcr_k$ 
     $IcPcr = IcPcr + apcr(Cr_j, Pcr_k).Ic$  // intensité consommée  $Pcr_k$ 
  End for
   $Pcr_k.Ip = IpPcr$  // total intensité produite  $Pcr_k$ 
   $Pcr_k.Ic = IcPcr$  // total intensité consommée  $Pcr_k$ 
   $Pcr_k.Up = Pcr_k.Rt * Pcr_k.Ip$  // tension produite  $Pcr_k$ 
   $Pcr_k.Uc = Pcr_k.Rt * Pcr_k.Ic$  // tension consommée  $Pcr_k$ 
End For
/* FIN */

```

Résumé : En 1977, FREEMAN formalisait les premières mesures génériques d'Analyse de Réseaux Sociaux (ARS). Puis, les réseaux sociaux du Web « 2.0 » sont devenus planétaires (*e.g.*, FaceBook, MSN). Cette thèse définit un modèle sémantique, non probabiliste et prédictif, pour l'analyse décisionnelle de réseaux sociaux professionnels et institutionnels. Ce modèle, en parallèle à la sociophysique de GALAM, intègre des méthodes de traitement sémantique du langage naturel et d'ingénierie des connaissances, des mesures de sociologie statistique et des lois électrodynamiques, appliquées à l'optimisation de la performance économique et du climat social. Il a été développé et expérimenté dans le cadre du projet SOCIOPRISE, financé par le Secrétariat d'État à la prospective et au développement de l'économie numérique.

Mots clés : analyse, réseaux, sociaux, statique, dynamique, sémantique, ontologie, sociophysique.

Skills Networks : From Social Networks Analysis to Predictive Knowledge Analysis.

Abstract : In 1977, FREEMAN formalised generic measures of Social Networks Analysis (SNA). Then, the Web “2.0” social networks have become global networks (*e.g.*, FaceBook, MSN). This thesis defines a semantic model, non probabilist and predictive, for the decisional analysis of professional and institutional social networks. The presented multidisciplinary model, in parallel to the GALAM sociophysics, integrates some semantic methods of natural language processing and knowledge engineering, some measures of statistic sociology and some electrodynamic laws, applied to the economic performance and social climate optimisation. It has been developed and experimented in line with the SOCIOPRISE project, funded by the French State Secretariat for the prospective and development of the digital economy.

Keywords : social, networks, analysis, static, dynamic, semantic, ontology, sociophysics.

Bibliographie

- [Aimé 2011] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *SemioSem et Proxima : mesures sémiotiques de similarité et de proximité conceptuelles*. In Actes des 22èmes Journées francophones d'Ingénierie des Connaissances, IC 2011, Chambéry, France, pages 539–554, 2011. (Cité en pages 18 et 100.)
- [Aimé 2009] X. Aimé, F. Furst, P. Kuntz et F. Trichet. *Gradients de prototypicalité conceptuelle et lexicale : une contribution à la pragmatization des ontologies de domaine*. Revue des Nouvelles Technologies de l'Information (RNTI) - Extraction et Gestion des Connaissances (EGC'08), vol. 11-1, pages 127–132, 2009. (Cité en page 18.)
- [Albert 2000] R. Albert, H. Jeong et A-L. Barabasi. *The Internet's Achille's Heel : Error and attack tolerance of complex networks*. Nature, vol. 406, pages 200–210, 2000. (Cité en page 16.)
- [Andrade 2007] N Andrade, F. Brasileiro, W. Cirne et M. Mowbray. *Automatic grid assembly by promoting collaboration in peer-to-peer grids*. Journal of Parallel and Distributed Computing, vol. 67, no. 8, pages 957–966, 2007. (Cité en page 18.)
- [Awadallah 2011] A.M.H. Awadallah. *Identifying Roles in Social Networks using Linguistic Analysis*. PhD thesis, University of Michigan, 2011. (Cité en page 20.)
- [Barabasi 1999] A-L. Barabasi et R. Albert. *Emergence of Scaling in Random Networks*. Science Magazine, vol. Vol. 286, no. no. 5439, pages 509–512, 1999. (Cité en page 14.)
- [Barrat 2006] A. Barrat, M. Barthelemy et A. Vespignani. *Réseaux complexes et physique statistique*. Images de la Physique, pages 47–53, 2006. (Cité en pages 6, 12 et 109.)
- [Bavelas 1948] A. Bavelas. *A mathematical model for Group Structure*. Human Organization, vol. 7, pages 16–30, 1948. (Cité en page 2.)
- [Beauchamp Murray 1965] A. Beauchamp Murray. *An improved index of centrality*. Behavioral Science, vol. 10, no. 2, pages 161–163, 1965. (Cité en pages 40, 41 et 91.)
- [Becker 1994] G.S. Becker. *Human capital : A theoretical and empirical analysis, with special reference to education*. University of Chicago Press, 3rd édition, 1994. (Cité en page 1.)
- [Benchettara 2009] N. Benchettara, R. Kanawati et C. Rouveirol. *Calcul de recommandations par prédiction de liens dans un graphe biparti*. In Actes de l'Atelier AGS : Apprentissage de Graphes pour les Systèmes complexes, 2009. (Cité en page 16.)

- [Bhattacharyya 2011] P. Bhattacharyya, A. Garg et S.F. Wu. *Analysis of user keyword similarity in online social networks*. Social Network Analysis and Mining Journal (SNAM), vol. 3, pages 143–158, 2011. Journal by Springer. (Cité en pages 107 et 108.)
- [Blondel 2008] V. D. Blondel, J.L. Guillaume, R. Lambiotte et E. Lefebvre. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics : Theory and Experiment, page 10008, 2008. (Cité en pages 42, 77, 78, 93 et 100.)
- [Blouin 1997] M. Blouin et C. Bergeron. Dictionnaire de la réadaptation, tome2 : termes d'intervention et d'aides techniques. Québec : Les Publications du Québec, 1997. (Cité en page 16.)
- [Boonchom 2010] V. Boonchom et N. Soonthornphisaj. *Legal Ontology Construction Using ATOB Algorithm*. In BIS (Workshops), volume 57 of *Lecture Notes in Business Information Processing*, pages 268–279. Springer, 2010. (Cité en page 100.)
- [Brandes 2001] U. Brandes. *A Faster Algorithm for Betweenness Centrality*. Journal of Mathematical Sociology, vol. 25, pages 163–177, 2001. (Cité en page 39.)
- [Brandes 2005] U. Brandes et D. Fleischer. *Centrality measures based on current flow*. In 22nd Symp. Theoretical Aspects of Computer Science (STACS 05), numéro 3404 de LNCS, pages 533–544. Springer Verlag, 2005. (Cité en pages 8, 15, 17, 27, 29, 38, 39, 41, 45, 46, 75, 100, 102 et 103.)
- [Brehm 1966] J.W. Brehm. A theory of psychological reactance. New York Academic Press, 1966. (Cité en page 24.)
- [Burel 1993] G. Burel et J-Y. Catros. *Mise en correspondance de graphes par recuit simulé. Application au recalage d'images*. Revue technique - Thomson-CSF ISSN 0035-4279 CODEN RTTCBG, vol. 25, no. 1, pages 213–237, 1993. (Cité en pages 16 et 17.)
- [Burt 1995] R.S. Burt. *Le capital social, les trous structuraux et l'entrepreneur*. Revue Française de Sociologie, vol. 36, no. 4, pages 599–628, 1995. (Cité en pages 1, 6 et 12.)
- [Chen 2011] L. Chen et L. Qi. *Social opinion mining for supporting buyer's complex decision making : exploratory user study and algorithm comparison*. Social Network Analysis and Mining Journal (SNAM), vol. 1, no. 4, pages 301–320, 2011. Journal by Springer. (Cité en pages 16 et 50.)
- [Clauset 2004] A. Clauset, M.E.J. Newman et C. Moore. *Finding community structure in very large networks*. Physical Review E, vol. 70, page 6, 2004. (Cité en page 42.)
- [Crozier 1977] M. Crozier et E. Friedberg. L'acteur et le système : les contraintes de l'action collective. Points. Politique. Éditions du Seuil, 1977. (Cité en pages 21, 22, 24 et 25.)

- [Delalonde 2007] C. Delalonde et E. Soulier. *Recherche et échange de connaissances dans des équipes distribuées*. In Actes des 18es Journées Francophones d'Ingénierie des Connaissances, Grenoble, France (IC 2007), 2007. (Cité en pages 19, 50 et 100.)
- [Demailly 2008] L. Demailly. *Politiques de la relation : approche sociologique des métiers et activités*. Presses Universitaires Septentrion, 2008. (Cité en pages 21, 22 et 25.)
- [Engelbart 1968] Douglas C. Engelbart et William K. English. *A research center for augmenting human intellect*. In AFIPS '68 (Fall, part I) : Proceedings of the December 9-11, 1968, fall joint computer conference, part I, pages 395–410, New York, NY, USA, 1968. ACM. (Cité en page 1.)
- [Erdős 1959] P. Erdős et A. Rényi. *On random graphs*. *Publicationes Mathematicae*, vol. 6, pages 290–297, 1959. (Cité en page 14.)
- [Erétéo 2011] G. Erétéo. *Analyse sémantique des réseaux sociaux*. PhD thesis, Université de Nice Sophia-Antipolis, Laboratoire d'Informatique, Signaux, et Systèmes de Sophia-Antipolis (L3S, UMR6070 CNRS), 2011. (Cité en pages 18, 20 et 101.)
- [Ferrand 2001] L. Ferrand, R. Matos, B. New et C. Pallier. *Une base de données lexicales du français contemporain sur internet : LEXIQUE*. *L'Année Psychologique*, vol. 101, pages 447–462, 2001. (Cité en page 50.)
- [Fisher 1986] C. D. Fisher. *Organizational socialization : an integrative review*. *Research in personnel and Human Resource Management*, vol. 4, pages 101–145, 1986. (Cité en pages 22 et 25.)
- [Folkman 1984] S. Folkman et R. S. Lazarus. *Stress, appraisal, and coping*. Springer Publishing Company, 1984. (Cité en page 24.)
- [Freeman 1960] L.C. Freeman, W. Bloomberg, S.P. Koff, M.H. Sunshine et T.J. Fararo. *Local community leadership*. Syracuse, 1960. (Cité en pages 13 et 100.)
- [Freeman 1977] L.C. Freeman. *A set of measures of centrality based on betweenness*. *Sociometry*, vol. 40, pages 35–41, 1977. (Cité en pages 2, 13, 29, 34, 78, 100 et 101.)
- [Freeman 1979] L.C. Freeman. *Centrality in social networks conceptual clarification*. *Social Networks*, vol. 1, no. 3, pages 215–239, 1979. (Cité en pages 37 et 39.)
- [Freeman 1989] L.C. Freeman, D.R. White et A.K. Romney. *Research methods in social network analysis*. George Mason University Press, 1989. (Cité en page 13.)
- [Galam 2008] S. Galam. *Sociophysics : a review of Galam models*. *International Journal of Modern Physics*, vol. C 19, 3, pages 409–440, 2008. (Cité en pages 6, 12, 17, 27, 28, 29, 102 et 103.)
- [Gardiol 1987] F Gardiol. *Electromagnétisme*. Dunod, 1987. (Cité en page 107.)

- [Goodreau 2007] S. M. Goodreau. *Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network*. PubMedCentral, 2007. (Cité en page 16.)
- [Greffard 2012] N. Greffard, F. Picarougne et P. Kuntz. *Visual Community Detection : An Evaluation of 2D, 3D Perspective and 3D Stereoscopic Displays*. In Graph Drawing, volume 7034/2012 of *Lecture Notes in Computer Science*, pages 215–225. Springer, 2012. (Cité en page 106.)
- [Gruber 1995] T.R. Gruber. *Toward principles for the design of ontologies used for knowledge sharing*. *International Journal of Human Computer Studies*, vol. 43, no. 5/6, pages 907–928, 1995. (Cité en page 17.)
- [Gruber 2008] R. T. Gruber. *Collective knowledge systems : Where the Social Web meets the Semantic Web*. *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 6, no. 1, pages 4–13, February 2008. (Cité en pages 18 et 20.)
- [Hong Tuan 1995] R. Hong Tuan. *Factorization of Spanning Trees on Feynman Graphs*. Rapport technique LPTHE Orsay 92/59, Cornell University Library, 1995. (Cité en pages 106 et 108.)
- [Jaccard 1901] P. Jaccard. *Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines*. *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pages 241–272, 1901. (Cité en pages 19, 51, 100 et 114.)
- [Jean 2001] G. Jean et D. Rallet. *Capital humain et capital social*. *Nouveaux Regards*, vol. 14, no. 14, 2001. (Cité en page 2.)
- [Jonasson 1999] J. Jonasson et J.E. Steiff. *Amenability and Phase Transition in the Ising Model*. *Journal of Theoretical Probability*, vol. 12, no. 2, pages 549–559, 1999. (Cité en page 17.)
- [Jung 2007] J. Jung et J. Euzenat. *Towards Semantic Social Networks*. In *ESWC 07 : Proceedings of the 4th European Semantic Web Conference*, pages 267–280, Berlin, Heidelberg, 2007. Springer-Verlag. (Cité en page 18.)
- [Koppal 2008] A. Koppal. *The Ising Model and Percolation on Graphs*, April 2008. <http://www1.cs.columbia.edu/coms6998/Notes/lecture22.pdf>. (Cité en page 17.)
- [Krivelevich 2002] M. Krivelevich et B. Sudakov. *Sparse Pseudo-Random Graphs Are Hamiltonian*. *J. Graph Theory*, vol. 42, pages 17–33, 2002. (Cité en page 16.)
- [Langlands 1992] R.P. Langlands, C. Pichet, P. Pouliot et Y. Saint Aubin. *On the universality of crossing probabilities in two-dimensional percolation*. *Journal of statistical physics*, vol. 67, pages 553–574, 1992. (Cité en page 17.)
- [Latora 2001] V. Latora et M. Marchiori. *Efficient behavior of small-world networks*. *Physical Review Letters*, vol. 87, no. 19, page 198701, 2001. (Cité en page 15.)

- [Lavallée 2010] J.F. Lavallée et P. Langlais. *Apprentissage non supervisé de la morphologie d'une langue par généralisation de relations analogiques*. In TALN 2010, Montréal, Québec, Canada, jul 2010. (Cité en page 50.)
- [Lazega 2001] E. Lazega. The collegial phenomenon : The social mechanisms of cooperation among peers in a corporate law partnership. Oxford, 2001. (Cité en page 12.)
- [Leskovec 2008] J. Leskovec et E. Horvitz. *Planetary-Scale Views on a Large Instant-Messaging Network*. In WWW 2008, April 21-25, 2008, Beijing, China, 2008. (Cité en page 15.)
- [Limpens 2008] F. Limpens, F. Gandon et M. Buffa. *Bridging Ontologies and Folksonomies to Leverage Knowledge Sharing on the Social Web : a Brief Survey*. In 1st International Workshop on Social Software Engineering and Applications (SoSEA), L'Aquila (IT), 15th September, 2008, 2008. (Cité en page 18.)
- [Lorenz 1969] E.N. Lorenz. *Deterministic Nonperiodic Flow*. Journal of the Atmospheric Sciences, vol. 20, no. 2, pages 130–141, 1969. (Cité en page 103.)
- [Louche 2007] C. Louche. Psychologie sociale des organisations - 2ème édition. Armand Colin, 2007. (Cité en pages 21, 23, 24 et 25.)
- [Macskassy 2011] S.A. Macskassy. *Contextual linking behavior of bloggers : leveraging text mining to enable topic-based analysis*. Social Network Analysis and Mining Journal (SNAM), vol. 4, pages 355–375, 2011. Journal by Springer. (Cité en page 20.)
- [Mandelbrot 1982] B. Mandelbrot. The fractal geometry of nature. W H Freeman & Co, 1982. (Cité en page 103.)
- [Mika 2005] P. Mika. *Ontologies Are Us : A unified model of social networks and semantics*. In International Semantic Web Conference, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. International Semantic Web Conference 2005, Springer, November 2005. (Cité en page 6.)
- [Mintzberg 1982] H. Mintzberg et P. Romelaer. Structure et dynamique des organisations. Editions d'organisation, 1982. (Cité en page 21.)
- [Moreland 1982] R. Moreland et J. Levine. *Socialization in small groups : temporal changes in individual-group relations*. Advances in Experimental Social Psychology, vol. 15, pages 137–192, 1982. (Cité en pages 22 et 25.)
- [Moreno 1934] J. Moreno. *Who shall survive?* - (trad. fr) Fondements de la sociométrie . PUF, 1934. (Cité en pages 2, 11, 12, 13, 20 et 100.)
- [Newman 2004] M.E.J. Newman. *Detecting community structure in networks*. The European Physical Journal B - Condensed Matter and Complex Systems, vol. 38, no. 2, pages 321–330, 2004. (Cité en page 15.)
- [Newman 2005] M.E.J. Newman. *A measure of betweenness centrality based on random walks*. Social Networks, vol. 27, no. 1, pages 39–54, 2005. (Cité en pages 2, 8, 14, 15, 17, 27, 39, 45, 46, 75, 100, 102 et 103.)

- [Newman 2006] M.E.J. Newman. *Finding community structure in networks using the eigenvectors of matrices*. Physical Review E, vol. 74, no. 3, page 19, sep 2006. (Cité en pages 15, 29 et 44.)
- [Pearson 2003] M. Pearson et P. West. *Drifting smoke rings : social network analysis and Markov processes in a longitudinal study of friendship groups and risk taking*. Connections : bulletin of the International Network for Social Network Analysis, vol. 25, no. 2, pages 59–76, 2003. (Cité en pages 6 et 16.)
- [Pitrat 1990] J. Pitrat. Métaconnaissance : Futur de l'intelligence artificielle. Hermès, 1990. (Cité en page 29.)
- [Propp 1998] J.G. Propp et D.B. Wilson. *How to Get a Perfectly Random Sample from a Generic Markov Chain and Generate a Random Spanning Tree of a Directed Graph*. Journal of Algorithms, vol. 27, pages 170–217, 1998. (Cité en pages 106 et 108.)
- [Rasclé 2004] N. Rasclé et M. Bruchon-Schweitzer. La psychologie appliquée à l'analyse des problèmes sociaux, chapitre Le stress au travail, pages 65–82. PUF, 2004. (Cité en page 24.)
- [Rasclé 2006] N. Rasclé. Psychologie appliquée à la gestion des ressources humaines, chapitre Comment gérer le stress au travail, pages 299–318. Editions d'organisation, 2ème édition, 2006. (Cité en page 24.)
- [Robertson 1976] S E. Robertson et K. Sparck Jones. *Relevance weighting of search terms*. Journal of the American Society for Information Science, vol. 27, no. 3, pages 129–146, May-June 1976. (Cité en pages 19, 33, 51, 100 et 112.)
- [Rogers 1960] D.J. Rogers et T.T. Tanimoto. *A Computer Program for Classifying Plants*. Science, vol. 132, pages 1115–1118, 1960. (Cité en pages 19, 33, 51, 100 et 114.)
- [Ruiz 2007] I.L. Ruiz, M.U. Cuadrado et M. A. Gomez-Nieto. *New Graph Similarity Measurements based on Isomorphic and Nonisomorphic Data Fusion and their Use in the Prediction of the Pharmacological Behavior of Drugs*. International Journal of Biological and Medical Sciences, vol. 2, pages 125–129, 2007. (Cité en page 16.)
- [Salton 1986] G. Salton et M.J. MacGill. Introduction to modern information retrieval, chapitre 6 - Retrieval Refinements, pages 201–215. McGraw-Hill Book Company, 1986. (Cité en pages 51, 100 et 111.)
- [Shannon 1948] C.E. Shannon. *A mathematical theory of communication*. Bell System Technical Journal, vol. 27, pages 379–453 and 623–656, July and October 1948. (Cité en page 28.)
- [Shimbel 1953] A. Shimbel. *Structural parameters of communication networks*. Bulletin of Mathematical Biophysics, vol. 15, pages 501–507, 1953. Stress rate. (Cité en pages 14, 78, 84, 94, 96, 97 et 100.)
- [Sparck Jones 1972] K. Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, vol. 28, pages 11–21, 1972. (Cité en pages 51, 100 et 111.)

- [Staw 1975] B. Staw. *Attributions of the "cause" of performance : a general alternative of interpretation of cross-sectional research on organizations*. *Organizational Behavior and Human Performance*, vol. 13, pages 414–432, 1975. (Cité en pages 21, 22, 24 et 25.)
- [Steinmetz 1913] S.R. Steinmetz. Die stellung der soziographie in der reihe der geisteswissenschaften. *Archiv für Rechts - und Wirtschaftsphilosophie*, 1913. (Cité en page 11.)
- [Tajfel 1971] H. Tajfel, M. Billig, R. Bundy et C. Flament. *Social categorization and intergroup behavior*. *European Journal of Social Psychology*, vol. 1, pages 149–178, 1971. (Cité en page 23.)
- [Tannenbaum 1967] A. S. Tannenbaum. *Psychologie sociale de l'organisation industrielle*. Editions Hommes et Techniques, 1967. (Cité en pages 23 et 24.)
- [Then 2010] H.W. Then, M. Feng et N. Holonyak. *Microwave circuit model of the three-port transistor laser*. *Journal of Applied Physics*, vol. 107, no. 9, page 094509, 2010. (Cité en page 61.)
- [Thomassen 1990] C. Thomassen. *Resistances and currents in infinite electrical networks*. *J. Comb. Theory Ser. B*, vol. 49, no. 1, pages 87–102, 1990. (Cité en pages 15, 17, 45 et 46.)
- [Thovex 2010a] C. Thovex et F. Trichet. *Analyse statique et sémantique de réseaux sociaux d'entreprises et institutions : vers un modèle multidimensionnel convergent*. In Actes de l'atelier « Web Social », 10ième Journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010), Hammamet, Tunisie, pages 517–528, 2010. (Cité en page 7.)
- [Thovex 2010b] C. Thovex et F. Trichet. *Dynamic and Semantic Social Networks Analysis : a New Model Based on a Multidisciplinary Approach*. In Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering (SEKE 2010), San Francisco Bay, USA, pages 548–552, 2010. (Cité en page 8.)
- [Thovex 2010c] C. Thovex et F. Trichet. *A Multidisciplinary Model of Dynamic and Semantic Social Networks Analysis for Institutions*. In Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2010, Odense, Denmark, pages 397–399. Institute of Electrical and Electronics Engineers Computer Society (IEEE-CS), 2010. (Cité en page 8.)
- [Thovex 2010d] C. Thovex et F. Trichet. *A semantic social network analysis model for knowledge sharing recommendations*. Journée thématique : Fouille de grands graphes JFGG 2010 (MARAMI 2010), octobre 2010. (Cité en page 8.)
- [Thovex 2010e] C. Thovex et F. Trichet. *Static, Dynamic and Semantic Dimensions : Towards a Multidisciplinary Approach of Social Networks Analysis*. In Proceedings of the 4th international conference on Knowledge Science, engineering and management, KSEM'10, Belfast, Northern Ireland, volume

- 6291 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 567–572. Springer Verlag / Heidelberg, 2010. (Cité en pages 8 et 107.)
- [Thovex 2011a] C. Thovex et F. Trichet. *A semantic and multidisciplinary model for professional and social networks analysis*. In Proceedings of The International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, Kaohsiung, Taiwan, pages 45–52, 2011. (Cité en page 8.)
- [Thovex 2011b] C. Thovex et F. Trichet. *Static and Semantic Social Networks Analysis : Towards a Multidimensional Convergent Model*. In Proceedings of The First International Conference on Advanced Collaborative Networks, Systems and Applications, COLLA 2011, Luxembourg, pages 58–63, 2011. (Cité en page 8.)
- [Thovex 2011c] C. Thovex et F. Trichet. *Un Modèle Décisionnel, Sémantique et Pluridisciplinaire d'Analyse des Réseaux Sociaux Professionnels*. In Actes des 22èmes Journées francophones d'Ingénierie des Connaissances, IC 2011, Chambéry, France, 2011. (Cité en page 8.)
- [Thovex 2012a] C. Thovex et F. Trichet. Book chapter in : *Dynamic Analysis for Social Network (à paraître)*, chapitre Static, Dynamic and Semantic Dimensions : From Social Networks Analysis to Predictive Knowledge Networks. iConcept Press Ltd, 2012. (Cité en page 8.)
- [Thovex 2012b] C. Thovex et F. Trichet. *Semantic Social Networks Analysis : Towards a Sociophysical Knowledge Analysis (à paraître)*. *Social Network Analysis and Mining Journal (SNAM)*, vol. 2, no. 1, 2012. Journal by Springer. (Cité en page 8.)
- [Travers 1969] J. Travers et S. Milgram. *An Experimental Study of the Small World Problem*. *Sociometry*, vol. 32, no. 4, pages 425–443, 1969. (Cité en pages 12 et 15.)
- [Van de Leemput 2005] C. Van de Leemput. *Stress et mal-être au travail*. Dunod, 2005. (Cité en page 24.)
- [Van Maanen 1979] J. Van Maanen et E. Schein. *Towards a theory of organizational socialization*. *Research in Organizational Behavior*, vol. 1, pages 209–264, 1979. (Cité en pages 22 et 25.)
- [Villeneuve 1995] M. Villeneuve et H. Letarte. *La cohésion est-elle un prédicteur valable de la performance des équipes de travail ? Une méta-analyse*. *Psychologie du Travail et des Organisations*, vol. 1, pages 86–103, 1995. (Cité en pages 23 et 24.)
- [Wakita 2007] K. Wakita et T. Tsurumi. *Finding community structure in mega-scale social networks*. In Proceedings of the 16th international conference on World Wide Web, WWW '07, pages 1275–1276, New York, NY, USA, 2007. ACM. (Cité en page 42.)
- [Wang 2004] Peng. Wang, Xu Baowen, Lu Jiangjiang, Li Yanhui et Kang Dazhou. *Theory and semi-automatic generation of bridge ontology in multi-ontologies*

- environment*. In On the Move to Meaningful Internet Systems 2004 : OTM 2004 Workshops, volume 3292 of *Lecture Notes in Computer Science*. Springer, 2004. (Cité en page 100.)
- [Wang 2010] Z. Wang, A. Scaglione et R.J. Thomas. *Electrical centrality measures for electric power grid vulnerability analysis*. In IEEE, editeur, Proceedings of the 49th IEEE Conference on Decision and Control, CDC 2010, December 15-17, 2010, Atlanta, Georgia, USA, pages 5792–5797, 2010. (Cité en pages 17, 102 et 103.)
- [Zaragoza 2004] H. Zaragoza, N. Craswell, M. Taylor, S. Saria et S E. Robertson. *Microsoft Cambridge at TREC-13 : Web and HARD tracks*. In Proceedings of Text REtrieval Conference, TREC 2004, 2004. (Cité en pages 51 et 112.)
- [Zekri 2002] N. Zekri et J-P. Clerc. *Etude statistique et dynamique de la propagation d'épidémies dans un réseau de petit monde*. Comptes Rendus Physique, vol. 3, no. 6, pages 741–747, 2002. (Cité en pages 12 et 17.)
- [Zhuhadar 2011] L. Zhuhadar, O. Nasraoui, R. Wyatt et R. Yang. *Visual knowledge representation of conceptual semantic networks*. Social Network Analysis and Mining Journal (SNAM), vol. 3, pages 219–299, 2011. Journal by Springer. (Cité en pages 106, 107 et 108.)