HEC MONTRÉAL

Détection de communautés dans un réseau de collaboration évolutif

Par

Laurie Cloutier

Sciences de la gestion
(Analytique d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

12/2016

A089W9.0911

Déclaration de l'étudiante, de l'étudiant
Éthique en recherche auprès des êtres humains

Registrariat

3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec), Canada H3T 2A7

HEC MONTRÉAL

Recherche ne nécessitant pas l'approbation du CER

Ce formulaire est requis pour les thèses, mémoires ou projets supervisés correspondant à une des deux situations suivantes :

1) un cas pédagogique;
2) une recherche menée auprès d'employés d'une organisation spécifique et qui servira exclusivement à des fins d'évaluation, de gestion ou d'amélioration de cette organisation.

Ou, la thèse, le mémoire ou le projet supervisé n'implique aucune des trois situations suivantes :

1) une collecte de données impliquant des sujets humains (par entrevue, groupe de discussion, questionnaire, observation ou toute autre méthode de collecte);
2) l'utilisation de données déjà collectées impliquant de l'information sur des sujets humains qui n'est pas accessible au public;
3) le couplage de plusieurs des données impliquant de l'information sur des sujets humains, que celle-ci soit publique ou non (le couplage est un recoupement de deux ensembles de données distincts qui permet de lier des données particulières entre elles).

Titre de la recherche : Détection de communautés dans un réseau de collaboration évolutif

Nom de l'étudiant : Laurie Cloutier

Signature : LaurieCloutier

Date : 07-12-2016

Nom du directeur : SYLVAIN PERRON

Signature : Sylvain Perron

Date : 7 décembre 2016

Veuillez remettre ce formulaire dûment complété et signé lors de votre dépôt initial

Pour toute question, veuillez vous adresser à cer@hec.ca

Imprimer

2016-03-10

Résumé

Les réseaux peuvent être fort utiles afin de refléter les interactions entre différents éléments et ainsi rendre compte des principes sous-jacents à leur comportement. Cependant, pour bien les comprendre, il est nécessaire de les représenter mathématiquement par un graphe, où chaque élément constitue un sommet et chaque relation est désignée par une arête. Les graphes permettent notamment l'étude du regroupement des sommets en communautés. Il s'agit alors de groupes formés à partir d'éléments du réseau qui sont fortement connectés ensemble, mais peu avec les autres. La détection de ceux-ci permet de mieux comprendre la dynamique entre les éléments, et ce, même pour des réseaux de très grandes tailles, d'où sa pertinence.

Ce document s'intéresse à expérimenter différents modèles de graphes représentant un réseau social évolutif, soit dont la composition change avec le temps, qui comporte des éléments de deux natures différentes. Ce dernier est formé plus spécifiquement comme un réseau de collaboration à partir d'une base de données réelles, dont aucune information contextuelle ou communauté n'est connue initialement. Divers paramètres seront déterminés pour chacun des modèles, de manière à sélectionner celui présentant les résultats les plus réalistes, en vue de leur analyse.

L'approche de résolution utilisée permet la détection de groupes de qualité, en plus de relever des tendances et phénomènes intéressants, quant au comportement des communautés et des individus, au fil du temps. Comme les données représentent des employés, les faits saillants soulevés grâce à la résolution du graphe ont de quoi intéresser leurs dirigeants, si ces derniers veulent bien comprendre la dynamique de leur équipe et ajuster leur gestion en conséquence. Certaines expérimentations restent toutefois à effectuer afin d'évaluer davantage la qualité de l'approche utilisée.

Mots-clés : *communauté, communauté dynamique, réseau évolutif, réseau social, réseau de collaboration, modélisation, analyse de communautés*

Abstract

Networks are known to be useful for describing the interactions between different elements and consequently to outline the underlying principles of their behavior. However, in order to understand them properly, it is necessary to represent them mathematically by a graph, where each element is represented by a vertex and each relation by an edge. Graphs also allow the study of vertices' grouping in communities. They are then groups (or clusters) formed by elements of the network which are highly connected together, but poorly linked with others. The community detection allows to better understand the dynamics between elements, even for networks of very large sizes, which demonstrates its relevance.

This thesis proposes to experiment different models of graphs representing an evolving social network, meaning its composition changes over time, which includes elements of two different kinds. This one is formed more specifically as a collaboration network from a real database, where any contextual information or community is initially known. Several parameters will be determined for each model in order to select the one with the most realistic results to analyze them.

The solution approach used allows the detection of quality groups and to reveal trends and interesting phenomena related to communities and individuals' behavior over time. As the data represents employees, the highlights found by our solution approach should interest managers, if they want to understand the dynamics of their team and then, adjust their management accordingly.

Keywords : *community, dynamic community, evolving network, social network, collaboration network, modeling, community analysis*

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

A^{3}CS	AdAptive Algorithm for Community Structure in dynamic networks
AFFECT	Adaptive Forgetting Factor for Evolutionary Clustering and Tracking
CNM	algorithme de Clauset, Newman et Moore
COPRA	Community Overlapping PRopagation Algorithm
CPM	Clique Percolation Method
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENGRAPH	DENsity based GRAPH clustering algorithm
DiDiC	Distributed Diffusive Clustering
DSBM	Dynamic Stochastic Block Model
DYN-LSNNIA	DYNamic Local Search Nondominated Neighbor Immune Algorithm
DYN-MOGA	DYNamic MultiObjective Genetic Algorithms
EO	Extremal Optimization
iLCD	intrinsic Longitudinal Community Detection
IM	Information Mutuelle
IMN	Information Mutuelle Normalisée
LPA	Label Propagation Algorithm
LPAcw	Label Propagation Algorithm with consensus weight
LPAm	Label Propagation Algorithm with modularity
LWEP	Local Weighted-Edge-based Pattern
MIEN	Modules Identification in Evolving Networks
MODEC	MOdeling and Detecting Evolutions of Communities
NEO-CDD	Nash Extremal Optimization for the Dynamic Community Detection problem
OSLOM	Order Statistics Local Optimization Method
PDEC	Particle-and-Density based Evolutionary Clustering
QCA	Quick Community Adaptation
WEO	Weighted Extremal Optimization

Avant-propos

Avec l'informatisation, les entreprises ont maintenant accès à un nombre croissant de données variées et ce, beaucoup plus facilement qu'auparavant. Dans ce contexte, il est naturel d'envisager une évolution des pratiques managériales afin de tirer avantage de ces nouveaux renseignements, disponibles en abondance dans l'organisation. Ce travail propose donc d'étudier les données d'une entreprise, dont l'exploitation pourrait permettre aux dirigeants d'envisager de nouvelles pratiques de gestion, plus adaptées à la réalité de leur organisation.

Les données sur lesquelles nous avons travaillé proviennent d'une grande organisation œuvrant dans le domaine des technologies de l'information. Elles concernent la collaboration de 1 045 employés à la conception d'un logiciel. Comme ces données représentent implicitement des relations entre les individus, soit par la réalisation de tâches communes, nous avons opté pour une modélisation des interactions par un réseau. Cet outil de modélisation mathématique semblait approprié, car il est basé sur des entités, appelées nœuds ou sommets, ainsi que les liens qui les unissent, nommés arêtes ou arcs. L'étude des réseaux est un domaine en pleine expansion. Nous avons donc utilisé des techniques qui en sont issues pour extraire, du réseau formé à partir des données fournies, l'information jugée pertinente du point de vue managérial.

L'approche de résolution proposée ouvre la porte à une méthodologie de gestion novatrice et prometteuse. Une fois la conception du réseau effectuée, les communautés qui s'y trouvaient ont été identifiées. Il s'agit, en fait, de groupes informels de collaboration, qui se caractérisent par une plus forte densité de liens entre leurs membres comparativement aux relations entretenues avec le reste du réseau. Comme les données recueillies couvrent plusieurs années, soit plus de cinq ans, il a été possible de constater l'évolution des communautés de travail au fil du temps, impliquant donc directement la gestion du personnel concerné.

La détection de communautés, dont la réalité sous-jacente nous était inconnue, s'est pratiquement effectuée à l'aveugle. Toutefois, des phénomènes intéressants quant aux comportements des employés ont été dégagés et pourraient être utiles aux dirigeants de l'organisation, afin de mieux comprendre la dynamique de leur équipe et ainsi, ajuster leur gestion en conséquence. Cette approche de résolution, qui n'a été que très peu testée malgré tout, s'annonce donc prometteuse.

En milieu d'affaires, un réseau s'apparente à la structure de l'entreprise, où divers employés interagissent ensemble. Ces relations entre collègues évoluent et peuvent changer avec le temps. Les modifications dans leur comportement se reflètent notamment sur leur regroupement avec les autres membres de l'organisation. Selon différents facteurs, dont leurs champs d'intérêt et leur personnalité [1], les employés interagissent davantage avec certains collègues qu'avec d'autres. Des regroupements sont ainsi formés et se distinguent les uns des autres. Il est alors question de communautés. Ces groupes sont dits informels, car ils sont créés par le libre choix des employés et non pas par la structure interne de l'entreprise, en vue de la réalisation de leurs tâches respectives [1]. L'affiliation à une communauté est alors changeante avec le temps. Selon leur niveau d'acception ou de divergence avec les idées partagées dans le groupe, les individus vont y rester ou s'en retirer. Leur identité peut également avoir un rôle à jouer quant à leur présence dans la communauté. Par exemple, les employés qui possèdent une identité plutôt traditionnelle pourraient se joindre à des groupes différents de ceux dont l'identité est basée sur le prestige ou encore, l'innovation [2].

Bien connaître son équipe est primordial pour les dirigeants d'entreprise [2]. C'est en connaissant bien la réalité humaine que le gestionnaire peut orienter ses actions pour augmenter la productivité de l'organisation. Les pratiques de direction sont influencées non seulement par le contexte organisationnel, dont les valeurs et objectifs de l'entreprise, et les particularités du dirigeant, mais aussi par les caractéristiques du personnel. Ainsi, bien connaître ces dernières est essentiel pour l'atteinte d'objectifs communs. D'autant plus, l'accès à ces informations sur les employés ne peut qu'augmenter la capacité d'influence du gestionnaire en plus de faciliter sa prise de décision.

Cependant, cet accès devient malheureusement plus difficile dès l'instant que l'entreprise grandit et que les gestionnaires ne connaissent pas personnellement chacun de leurs employés. Des techniques alternatives pour déceler automatiquement certaines réalités sont alors nécessaires. C'est une de ces techniques alternatives qui fait l'objet de ce mémoire.

Notre travail se concentre sur le comportement des 1 045 employés de l'entreprise concernée. L'analyse a été effectuée uniquement à partir de 1 607 632 commentaires émis par ces individus sur un forum pendant les cinq années et demie, soit 1 975 jours plus exactement, pendant lesquelles les données ont été recueillies. Ces commentaires sont répertoriés dans 324 860 conversations au total. La réalité sous-jacente à ces discussions nous est totalement inconnue, que ce soit par rapport au contexte de l'entreprise, à celui du forum ou encore, quant aux comportements des employés. De plus, le contenu des commentaires ne nous a pas été transmis. Les informations qui nous sont accessibles concernent l'auteur du commentaire, son bureau, son rôle dans la conversation, la date de son commentaire ainsi que la discussion à laquelle il a pris part. De cette manière, l'étude des comportements du personnel effectuée dans ce travail se base uniquement sur ces informations sans égard à la situation réelle de l'organisation.

D'abord, il est possible de connaître les grandes lignes du **comportement des employés** avec leurs collègues, que par une analyse statistique générale de leurs commentaires. Il s'agit en fait que d'une mise en contexte entourant la collaboration entre les employés de l'entreprise.

Par exemple, nous pouvons constater que les employés discutent de plus en plus ensemble par le biais du forum, plus les mois passent. Le nombre de discussions est à la hausse tout au long de l'horizon temporel étudié, excepté durant les six derniers mois où il chute. Des **relations** semblent donc s'établir entre les employés au fil du temps. Par ailleurs, ils échangent en plus grand nombre sur un nombre restreint de conversations durant les derniers mois. Malgré qu'ils joignent en moyenne 727 discussions au cours de leur participation, certains sont moins actifs que d'autres. En effet, alors qu'un employé a joint 10 conversations au total, un autre a pris part à 14 392 discussions. La **participation** est donc **variable** d'un employé à l'autre. Entre deux et trois individus prennent part à chaque conversation en moyenne. La durée de contribution aux échanges est également différente selon les membres de l'organisation. En moyenne, chacun a contribué aux discussions pendant près de trois ans. Il est à noter que certains employés, soit 118 d'entre eux, optent pour des pauses, c'est-à-dire que leur participation ne se réalise pas uniquement sur des périodes consécutives. Celles-ci, s'étalant sur six mois pour la moitié et près d'un an en moyenne, ont probablement été engendrées pour des raisons particulières qui nous restent inconnues sans informations supplémentaires de l'entreprise. Malgré que les individus étudiés proviennent de 46 bureaux différents, dont près de la moitié travaillent dans le même établissement, le

bureau d'appartenance n'influence pas nécessaire les regroupements du personnel au fil du temps.

La problématique au cœur de ce mémoire consiste à bien représenter les interactions entre les individus afin d'en dégager les différents groupes formés. Une fois ceux-ci trouvés, qu'on appellera communautés, il sera possible d'analyser le comportement des employés et des groupes qu'ils forment à travers le temps. Les principales tendances et divers faits saillants pourront être dégagés et permettront aux gestionnaires de mieux comprendre leur équipe. Ces phénomènes seront captés que par l'approche de résolution proposée dans ce travail sans aucune information additionnelle à celle fournie par l'entreprise, comme décrit précédemment.

Pour ce faire, les **relations entre les employés** sont représentées dans un réseau, simplifié à la figure 1, où chaque discussion et chaque individu sont illustrés. Chacune des conversations est reliée à ses participants ou, en d'autres mots, chaque employé est relié aux discussions auxquelles il a pris part. Pour permettre l'étude temporelle des comportements, tous les commentaires sont regroupés par période de six mois et ainsi, un même individu est représenté pour chaque semestre pendant lequel il a émis au moins un commentaire. Un employé non actif sur le forum pendant une période de six mois n'est donc pas représenté pour celle-ci. Cette approche permet de mieux cerner la durée réelle de participation des membres de l'entreprise aux discussions, en plus de montrer clairement les pauses prises par certains d'entre eux. La difficulté de cette approche est de déterminer à quel point le comportement d'un employé est cohérent à travers le temps, s'il a participé aux discussions de manière continue et si ce n'est pas le cas.

À partir de ce réseau, les interactions sont étudiées de manière à regrouper les employés discutant davantage ensemble qu'avec leurs autres collègues, signifiant qu'ils forment une communauté. L'objectif est de rassembler les membres de l'entreprise, de manière à ce que les commentaires émis à l'intérieur de leur groupe soient très nombreux et à l'inverse, plus rares avec les autres collègues. Il faut donc trouver des communautés bien distinctes les unes des autres. Le **niveau de cohérence dans le comportement** des employés à travers le temps a été établi à la vue de résultats obtenus en testant plusieurs valeurs et en choisissant celle permettant les résultats les plus réalistes.

La résolution a permis de déceler 29 groupes formés par les employés sur l'ensemble des cinq ans et demi de l'étude, dont chacun perdurait en moyenne pendant cinq ans. La composition de ces regroupements diffère entre eux et change au fil du temps. La

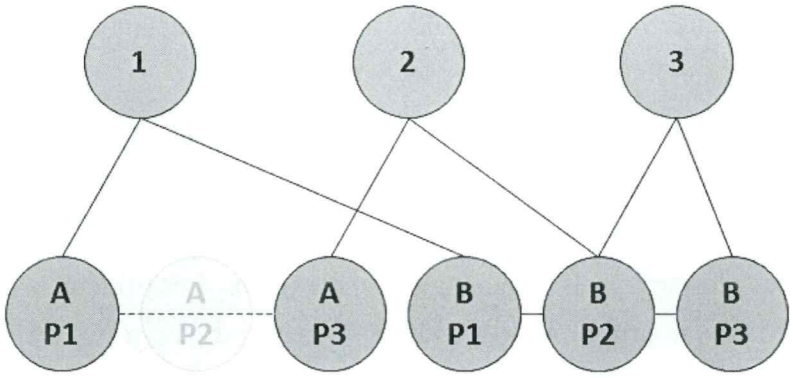


FIGURE 1 – Représentation des participations des employés aux discussions, selon la période à laquelle ils y ont pris part. Les sommets orangés représentent les conversations et les sommets bleus représentent les individus *A* et *B* aux périodes 1, 2 et 3, soit *P1, P2* et *P3* respectivement. L'individu *A* n'a contribué à aucune discussion pendant la deuxième période, il ne sera donc pas considéré à ce moment. Le lien entre un même individu correspond au niveau de cohérence dans son comportement au fil du temps.

tendance générale est à l'agrandissement de ces communautés plus le temps avance, ce qui est cohérent avec la participation grandissante des employés aux conversations. Cependant, certains groupes se distinguent davantage des autres, conformément à leur composition atypique, tout comme certains comportements se différencient des autres.

L'approche proposée dans ce travail a permis de montrer que les membres de l'entreprise développent véritablement des **affinités** avec certains de leurs collègues, plus qu'avec d'autres. En effet, avec le regroupement effectué, les employés participent, en moyenne, à quatre fois plus de discussions avec les membres de leur groupe qu'avec leurs autres collègues. Cela signifie donc que seulement 20% des conversations entretenues par un employé se font avec des individus qui ne font pas partie de sa communauté. La distinction entre les groupes formés est donc bien présente. Des relations sont ainsi créées et peuvent, sans aucun doute, influencer plusieurs aspects du comportement des individus au travail.

Il a aussi été constaté que les employés semblent plutôt **conservateurs** quant à leurs relations au fil du temps ainsi que leurs sujets de conversation. En fait, chaque individu joint en moyenne 1,13 communauté au cours de sa participation aux discussions. De plus, parmi les 118 employés arrêtant de prendre part à des conversations pendant six mois, 104, soit 88% d'entre eux, retrouvent leur même groupe de discussion à leur retour. Cette situation démontre donc que les participants modifient rarement leurs

champs d'intérêt ou leur comportement même pendant une pause.

Des relations changeantes, ou encore un intérêt variable pour les idées échangées peuvent toutefois expliquer pourquoi 15,98% des employés ont pris part à plus d'un groupe au cours des cinq ans et demi à l'étude. Le **manque de stabilité dans les relations interpersonnelles** de certains employés est un autre élément pouvant être remarqué par la résolution effectuée. Effectivement, les employés qui changent plus souvent de groupes sont ceux qui ont pris part à des discussions sur une plus longue période. Ils restent également moins longtemps dans chacune des communautés qu'ils joignent.

Les liens forts et profonds entretenus entre des individus très constants dans leurs relations et leurs sujets de discussion devraient intéresser les gestionnaires de l'organisation. Par exemple, nous décelons 41 employés qui restent dans la même communauté pendant les cinq ans et demi et dont, cinq et huit d'entre eux se retrouvent dans les mêmes communautés. Ces derniers entretiennent donc des relations profondes ensemble, ou encore s'intéressent réellement aux mêmes sujets, et ce, sur une longue période. L'étude de leur **influence** ainsi que le **rôle** qu'ils exercent dans leur groupe respectif pourrait s'avérer révélatrice sur leur comportement et celui de leur équipe.

Non seulement le comportement des individus peut différer avec le temps, mais la résolution a permis de constater que certaines périodes sont plus propices à ce phénomène. En effet, parmi les 203 changements de communautés relevés pour l'ensemble de l'horizon temporel étudié, une grande majorité d'entre eux se sont déroulés durant la troisième année. La figure 2 montre le nombre de changements de groupes survenus chaque semestre, en plus du nombre de communautés quittées par les employés changeant d'affiliation et le nombre de communautés dans lesquelles ils se sont ajoutés. Le nombre de groupes présents chaque période est également illustré.

Il y a donc des semestres où les employés semblent plus stables dans leurs relations alors qu'à d'autres, les **changements d'affinité** sont plus fréquents, comme c'est le cas durant les sixième et septième semestres. Il reste donc à savoir si un **évènement particulier relié à leur environnement**, ou au **contexte de l'entreprise**, aurait pu précipiter ces mouvements, ou encore si ces changements sont plutôt attribuables à l'**évolution des relations** en cause. Par ailleurs, parmi les changements relevés, seulement 7,39% d'entre eux sont explicables par le retour d'un employé de pause qui change de groupe à ce moment, en guise d'une modification de ses relations ou de ses champs d'intérêt pendant son temps d'arrêt. Aux semestres plus mouvementés, le nombre de commu-

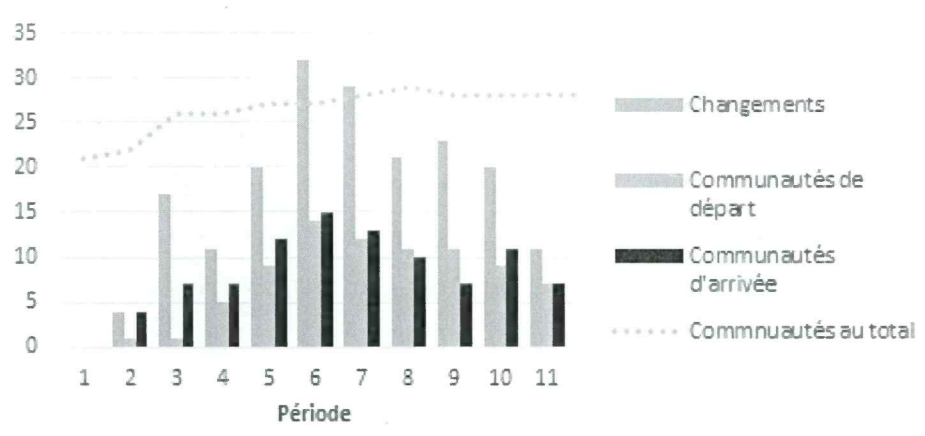


FIGURE 2 – Nombre d'employés qui changent de communauté chaque semestre, en plus du nombre de communautés dans lesquelles ils se trouvaient initialement ainsi que le nombre de communautés auxquelles ils se joignent au total. Les mouvements sont représentés au semestre où l'appartenance de l'individu est différente de celle de la période précédente. La courbe pointillée exprime le nombre total de communautés présentes chaque semestre.

nautés touchées par les changements constitue une proportion élevée de l'ensemble des groupes présents à ces moments. Il ne s'agit donc pas d'un phénomène isolé dans l'entreprise dans ces cas.

Les modifications dans le comportement des employés peuvent aller dans deux sens. L'approche proposée dans ce mémoire permet de détecter le **renforcement des relations**, comme aux huitième et neuvième semestres. Dans ces cas, il s'agit de participants qui ont initialement des liens faibles entre eux, mais ces relations se fortifient et les employés se retrouvent finalement ensemble dans de mêmes communautés. Cette situation se reflète par le nombre de groupes quittés lors des mouvements qui est supérieur au nombre de groupes rejoints. Par exemple, à la neuvième période, où la différence est la plus marquée, 13 des 23 employés changeant de communautés se retrouvent dans un même groupe, soit le plus grand à ce moment.

À l'inverse, par la résolution effectuée, il a aussi été possible d'identifier la **dégradation des relations** entre certains employés avec le temps, signifiant qu'ils entretenaient des liens serrés initialement pour ensuite se disperser chacun de leur côté ou presque. Cette situation se remarque par un nombre de communautés d'arrivée supérieur au nombre de communautés de départ. C'est le cas la majorité du temps, mais plus particulièrement

aux deuxième et troisième semestres. Les mouvements sont alors explicables uniquement par le départ d'employés d'un même groupe, soit celui ayant la plus grande taille jusque-là. Ces individus se divisent pour rejoindre respectivement quatre et sept communautés différentes au cours de ces périodes. Au troisième semestre, c'est près de 27% des individus de la communauté qui la quitte.

Il serait pertinent pour l'entreprise de chercher à comprendre la cause de ces départs, notamment afin de s'assurer que ce n'est pas un **conflit** important qui en est au cœur. En effet, les conflits peuvent être très dommageables autant pour les employés, notamment par l'ajout d'un stress, d'un mécontentement ou encore de l'adoption d'une attitude de fermeture, que pour l'organisation, en raison d'une diminution de la performance, d'un climat de travail malsain ou encore d'une détérioration des relations qui en découlent [3]. Un gestionnaire se doit donc de prévenir l'apparition de conflits dans l'entreprise en plus de résoudre rapidement ceux émergeant dans son équipe, et ce, au moindre coût pour l'organisation. En ayant un **meilleur suivi des relations** de leurs employés, les dirigeants sont mieux placés pour remplir adéquatement leurs responsabilités au sujet des conflits et éviter les conséquences néfastes de ceux-ci.

L'étude a permis de remarquer que l'**intérêt** de se joindre aux discussions avec les collègues est un phénomène constamment présent dans l'entreprise et plutôt répandu, constatant le nombre de groupes touchés par l'arrivée de nouveaux membres chaque semestre. En effet, environ une centaine d'individus se joignent aux conversations à chaque période, malgré une légère diminution des nouveaux participants au cours des derniers mois de l'horizon temporel étudié. La figure 3 illustre bien ce comportement.

Le regroupement de nouveaux participants ensemble à leur arrivée dans les discussions est une tendance répandue dans l'entreprise. Il est donc à se questionner si ces individus entretenaient une relation avant leur arrivée sur le forum. L'étude des **motivations** derrière ces ajouts ainsi que ces regroupements pourraient signifier aux gestionnaires l'intérêt réel des employés à se joindre aux différentes discussions du forum. L'**influence des membres** des groupes concernés est également matière à étude, tout comme la popularité des sujets qui y sont discutés. Ce phénomène est survenu notamment au cours des deuxième et troisième années de l'étude, où les employés se joignant aux conversations pour la première fois se sont retrouvés en grand nombre dans certains groupes. Cette situation a marqué significativement la taille des communautés *0*, *5*, *16* et *23*, dont l'évolution est représentée en Annexe à la figure B.6. Le groupe 0 a été le plus révélateur quant à cette situation, alors que 34 des 145 nouveaux participants à la qua-

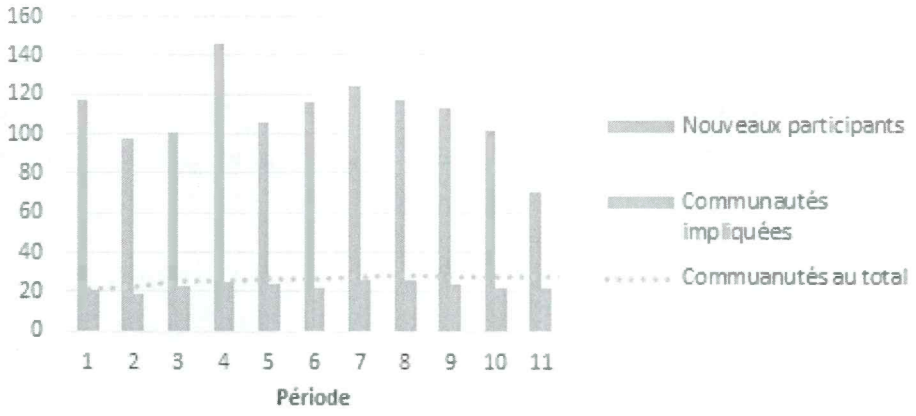


FIGURE 3 – Nombre de nouveaux participants aux discussions, en plus du nombre de groupe auxquels ils se joignent comparé au nombre total de communautés, et ce, à chaque période.

trième période s'y sont ajoutés, ce qui était très élevé comparativement à la croissance des autres communautés au même moment.

La résolution proposée dans ce travail a potentiellement permis d'identifier des **relations d'influence** entre les membres de l'entreprise. En effet, la participation de certains employés aux échanges semble être entrainée par celle de leurs collègues. Déjà, le regroupement de nouveaux participants ensemble a été abordé et est potentiellement explicable par l'influence de certains individus sur d'autres, à se joindre aux mêmes discussions. Également, les membres de l'organisation, prenant une pause d'au moins six mois au travers de leur contribution aux conversations, semblent **convaincre** de nouveaux participants à échanger avec eux à leur retour. Par exemple, deux groupes viennent à disparaitre au cours de l'horizon temporel étudié. Il s'agit de communautés qui se décomposent et ne compte plus qu'un seul membre, qui, lui, décide d'interrompre son apport aux conversations pendant un semestre. Cependant, à son retour, ce dernier retrouve son groupe, avec cette fois, uniquement de nouveaux participants. C'est ce phénomène qui laisse croire que l'individu, ayant pris une pause, a encouragé ses collègues pendant ce temps à expérimenter les échanges sur le forum avec lui, ce qui expliquerait leur regroupement ensemble à son retour. Cet employé agirait ainsi à titre de **leader**. Cette situation est également observable dans des groupes qui ne cessent d'exister. C'est le cas notamment pour la communauté *26* dont la taille augmente graduellement et qui est la plus élevée à la fin de l'horizon temporel. Toutefois,

une hausse significative du nombre d'employés échangeant dans ce groupe est remarquée à partir du cinquième semestre, où un employé revenant de pause retrouve son groupe d'échanges accompagné de 19 nouveaux participants. Ce phénomène est aussi observable de manière marquée avec la communauté *13*, dont le nombre de membres double pendant un seul semestre. À ce moment, trois employés reviennent d'une interruption de six mois de contribution aux conversations en plus de 26 de leurs collègues qui se joignent à eux pour leur première participation au forum. Il serait pertinent pour les gestionnaires d'investiguer davantage sur ce type de situations, car il semble que l'influence de certains employés pendant leur arrêt temporaire ne fait qu'**augmenter le niveau de participation** aux échanges à leur retour.

À l'inverse, ce phénomène pourrait aussi être explicable par le manque d'intérêt aux discussions de certains membres de l'organisation, qui quitteraient alors leur groupe et toutes conversations pendant au moins six mois. Leur retour aux échanges sur le forum pourrait alors être causé par l'influence de leurs collègues qui décident, eux aussi, de participer aux discussions, ce qui convainc ces employés à retrouver leur groupe et recommencer à prendre part aux échanges. Bien connaître les causes d'un tel phénomène ne pourrait qu'être bénéfique pour les gestionnaires qui connaitraient ainsi les véritables **rapports d'influence** qui existent au sein de leur personnel. Il serait alors possible pour eux de considérer ces informations lors de certaines situations et de l'exploiter à leur guise, au besoin.

L'approche proposée a également fait ressortir les arrêts de participation de certains employés aux discussions. En effet, malgré que des individus arrêtent de participer aux discussions chaque semestre, ils se font de plus en plus nombreux avec le temps, plus particulièrement à partir de la fin de la troisième année. La figure 4 présente justement le nombre d'employés quittant les discussions, en plus du nombre de communautés qui les regroupaient comparativement au nombre total de groupes existants chaque semestre. Les arrêts sont représentés à la première période où les employés concernés ne participent plus aux discussions.

Avec de tels résultats, il est à se questionner si une situation particulière se cache derrière cette **baisse de participation**. Cela pourrait autant toucher le **contexte de l'entreprise** elle-même, par exemple, que les **mesures incitatives** à échanger avec les collègues, l'**influence moindre** des membres des différents groupes, la présence de **conflits** entre les employés, le manque de **motivation** à se joindre aux discussions, ou encore le **manque d'intérêt** pour les sujets de conversation. La recherche et la connais-

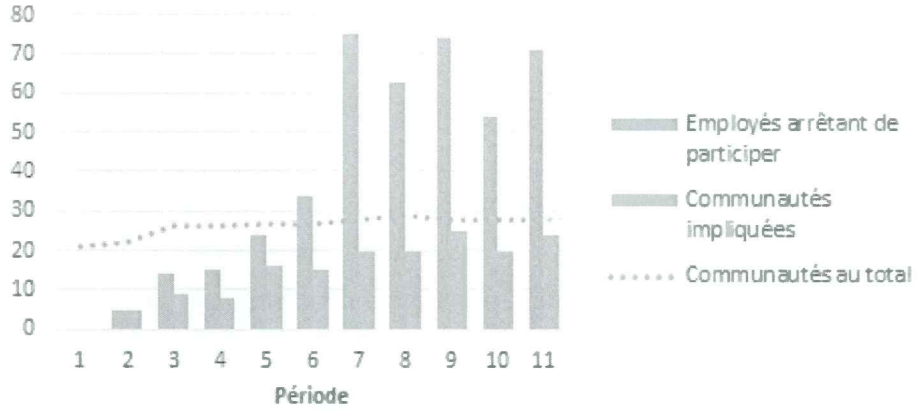


FIGURE 4 – Nombre d'employés arrêtant de participer, en plus du nombre de communautés desquelles ils partent comparé au nombre total de groupe, et ce, à chaque période.

sance des causes expliquant les départs plus nombreux seraient pertinentes pour les dirigeants qui veulent mieux comprendre le comportement des membres de leur équipe afin d'ajuster leur gestion en conséquence. Par ailleurs, les arrêts de participation constituent un phénomène répandu dans l'entreprise et non pas isolé à quelques groupes, ce qui augmente d'autant plus l'importance de leur démystification par les gestionnaires.

Deux communautés sont particulièrement impactées par ces départs. D'abord, la communauté *12*, qui est la plus populaire initialement, se décompose à partir de la deuxième année, et ce, jusqu'aux derniers mois de conversations relevés. Les premiers employés à quitter ce groupe se dirigent plutôt vers d'autres communautés alors que plus le temps passe, plus les départs se transforment en arrêt définitif de participation aux discussions. Alors que la plupart des autres groupes voient leur taille augmenter avec le temps, celui-ci va à l'encontre de cette tendance. Il serait donc pertinent pour les gestionnaires d'étudier ce qui a déclenché les départs d'employés de ce groupe et ce qui les a maintenus pendant près de trois ans. Les relations entretenues dans ce groupe pourraient être révélatrices du comportement de certains membres de l'organisation.

Puis, l'approche utilisée a également fait ressortir un phénomène de départs de masse constaté dans la communauté *0* du sixième au septième semestre. Effectivement, 37 employés ont quitté le groupe à ce moment et le forum du même coup, ce qui représente 49% des membres de cette communauté avant leur départ et $3,44\%$ de l'ensemble des employés participant aux conversations au cours des 1 975 jours relevés. Parmi eux,

neuf ont simplement pris une pause, alors que 28 ont quitté définitivement les discussions. Ces derniers s'ajoutent également à sept individus ayant quitté la communauté à la même période, pour se joindre à un autre regroupement. Malgré que le groupe *0* soit bien distinct des autres, que ses membres discutent peu avec leurs collègues des autres communautés et que ceux ayant quitté ne semblent pas avoir un **rôle central**, échangeant moins que la moyenne, les causes d'autant de départs simultanément sont plutôt intrigantes. Si un **conflit** en fait partie, il serait important que les gestionnaires s'y attardent avant que des conséquences négatives en découlent, ou encore pour limiter celles-ci si elles sont déjà en cours.

Avec autant de résultats révélateurs sur le comportement des employés de l'entreprise étudiée, il est possible de croire que l'approche de résolution utilisée dans ce travail peut réellement aider les gestionnaires à mieux comprendre la dynamique au sein de leur équipe. Plusieurs phénomènes peuvent être relevés et permettent de cibler les situations ou employés qui méritent une attention particulière en ce sens. Comme décrit plus tôt, avoir une connaissance approfondie de la réalité humaine de son entreprise ne peut qu'aider les dirigeants à orienter leurs actions, leur influence ainsi que leur prise de décision pour de meilleurs résultats et pour l'atteinte d'objectifs communs. C'est exactement ce que l'approche présentée dans ce travail permet à partir uniquement de l'étude de simples commentaires entre les employés, sans aucune connaissance de la situation réelle de l'entreprise.

Chapitre 1

Introduction

1.1 Contexte

De nombreuses situations réelles peuvent être représentées dans un réseau. En effet, dès qu'il existe des interactions ou encore des relations entre différents éléments, peu importe la nature de ceux-ci, il est possible de parler de réseaux. Ces derniers peuvent être étudiés de manière plus approfondie en les représentant par un graphe. Celui-ci constitue leur représentation mathématique, où des nœuds correspondent aux différents éléments et des arêtes relient ces derniers s'ils entretiennent une relation. Plusieurs aspects peuvent caractériser le graphe, par exemple, des liens de forces différentes, des éléments de plusieurs natures ou encore, une existence temporelle, c'est-à-dire qui s'étale dans le temps, au contraire d'une représentation du réseau à moment précis seulement. À partir des graphes, il est possible de détecter des communautés, c'est-à-dire des sous-ensembles d'éléments qui interagissent fortement ensemble et peu avec le reste du réseau. L'étude des communautés est un sujet populaire depuis les dernières années, et ce, dans plusieurs domaines. En effet, avec l'existence de différents types de réseaux, qu'ils soient notamment biologiques [9], sociaux [10] [11] ou technologiques [12], leur étude se prête bien et permet de comprendre différents phénomènes.

Initialement, les graphes servaient, notamment, à simplifier la représentation du réseau de manière à pouvoir y jeter un œil et comprendre rapidement les différents liens qui s'y trouvaient. Cependant, avec la popularité et la disponibilité grandissantes des ordinateurs, l'obtention de données est devenue un phénomène courant [13]. En effet, il est

maintenant possible de recueillir beaucoup plus de données qu'auparavant, et ce, beaucoup plus facilement. Dans ces contextes, le recours aux simples graphes est moins pertinent, car leur taille est multipliée, ce qui nuit à leur compréhension. Il est alors nécessaire de trouver des approches permettant de détecter les phénomènes cachés derrière ces données. C'est justement l'objectif principal de la détection de communautés dans les réseaux complexes, c'est-à-dire de très grande taille. Plusieurs approches ont été proposées au cours des dernières années, chacune d'elles ayant ses propres caractéristiques et étant adaptée à certains types de réseaux.

Lorsque les données recueillies s'étalent sur une certaine période de temps, plutôt que de représenter une situation à un moment précis, leur représentation peut se faire par un graphe dit évolutif. Les communautés qui s'y trouvent sont alors qualifiées de dynamiques. L'étude de l'évolution de ces communautés en plus du comportement des différents éléments qu'elles regroupent peut s'avérer des plus pertinentes. Il suffit cependant d'utiliser des approches de résolution permettant de retracer les évènements ayant lieu au fil du temps, que ce soit la naissance, la mort, la croissance ou la décroissance des communautés ou encore l'affiliation et la désaffiliation des éléments aux différents groupes formés. L'étude des réseaux sociaux constitue un bel exemple de ce type de phénomène. Des individus discutent ensemble de sujets variés au fil du temps et selon leurs interactions, ils peuvent être regroupés en communautés à différents moments. Il est ensuite intéressant de comprendre la formation de celles-ci et la manière dont les participants se comportent à travers le temps.

Comme les différentes approches de résolution pour la détection de communautés dans les réseaux sont plus récentes et moins nombreuses que celles pour les graphes statiques, soit à un moment précis, il semblait intéressant de s'y concentrer. Ainsi, ce mémoire se consacrera à la détection de communautés dans un graphe évolutif obtenu à partir d'une base de données réelles, en testant une approche de résolution, moins abordée dans la littérature jusqu'à maintenant.

1.2 Problématique

La problématique principale de ce mémoire consiste à détecter des communautés dans un réseau social et évolutif. Elle comporte deux principaux aspects, soit de trouver une approche de résolution adaptée au réseau à l'étude et capable de bien refléter sa situa-

tion, en plus d'être en mesure de déceler différents phénomènes marquant l'évolution des communautés trouvées ainsi que des membres qu'elles regroupent.

Le réseau étudié provient d'une base de données réelles qui nous a été fournie par une entreprise. Il correspond à un réseau de collaboration, soit une forme particulière de réseau social, où les employés échangent à propos de diverses possibilités pour le développement d'un logiciel. Cependant, excepté les données qui s'y trouvent et qui représentent les caractéristiques des différents commentaires émis sur un réseau social, aucune information additionnelle n'est connue sur tout ce qui entoure ces discussions. Ainsi, le modèle choisi pour représenter le réseau doit se rapprocher au maximum de ces données tout en étant le plus réaliste possible. La détermination de ce modèle constituera le défi majeur de ce mémoire, d'autant plus qu'il aura un impact sur la solution trouvée et analysée. Un algorithme devra ensuite être sélectionné, en fonction de sa capacité à traiter le type de réseau étudié, en vue de la détection de communautés dynamiques.

Le deuxième élément de la problématique consiste à extraire de la solution des phénomènes réels intéressants, qui restent cachés autrement, en raison de l'ampleur de la situation réelle. La connaissance de ces phénomènes pourrait s'avérer très pertinente pour l'entreprise concernée qui comprendrait alors davantage le comportement de ses employés et pourrait ainsi adapter sa gestion en conséquence.

> C'est en connaissant bien la réalité humaine que le gestionnaire peut orienter ses actions pour augmenter la productivité de son organisation. Ainsi, être informé sur les caractéristiques de son personnel ne peut qu'augmenter sa capacité d'influence, en plus de faciliter sa prise de décision vers l'atteinte d'objectifs communs [2].

1.3 Plan du mémoire

Le prochain chapitre présentera la définition des différents concepts de base liés au contexte de ce mémoire ainsi que les notations qui seront utilisées dans le document. Il permettra d'assurer et faciliter la compréhension des propos qui suivront.

Le troisième chapitre se penche plutôt sur la littérature en faisant un état de tout ce qui a déjà été réalisé en lien avec la détection de communautés dynamiques dans les réseaux

évolutifs. La notion des réseaux est abordée pour commencer. Ensuite, le phénomène des communautés est expliqué plus en détail et est suivi d'un survol des différentes approches de résolution possibles pour les graphes statiques ainsi qu'évolutifs.

Le quatrième chapitre se concentra sur le contenu et les principales caractéristiques du jeu de données à la base du réseau étudié. La question de recherche sera également décrite précisément.

Le cinquième chapitre décrira la méthodologie qui sera utilisée pour répondre à la problématique de ce mémoire. La modélisation du réseau en vue de la détection des communautés qu'il regroupe sera présentée, suivie de l'approche de résolution qui sera adoptée pour identifier les communautés dynamiques. L'impact des différents paramètres et modèles sur les partitions obtenues sera évalué par la suite.

Le sixième chapitre exposera les analyses effectuées sur la partition sélectionnée à cet effet. Le choix de celle-ci sera d'abord justifié, puis l'étude détaillée des différents phénomènes marquant l'évolution des communautés et des individus sera présentée. Les particularités des liens de cette partition seront également évaluées.

Le dernier chapitre résumera enfin tout le travail effectué et ouvrira la porte sur de futurs travaux possibles.

Ces différentes étapes permettront de bien établir la situation réelle quant au comportement des employés dans l'entreprise, dans l'objectif de bien la comprendre et d'ajuster la gestion de l'organisation en conséquence.

Chapitre 2

Définitions et notations

2.1 Définitions

Afin de faciliter la lecture, nous donnerons ici quelques définitions utiles.

Un **réseau** est un diagramme comprenant un ensemble d'éléments, représentés par des points, dont certains sont reliés ensemble, par le biais d'une ligne [14]. Il peut être de plusieurs types, mais ce mémoire se concentrera principalement sur les réseaux sociaux, plus particulièrement, les réseaux de collaboration.

Porter [15] définit un **réseau social**, par un ensemble d'agents ou individus, dont il existe des liens entre eux. Ces liaisons représentent des interactions sociales ou des relations entre les agents du réseau et peuvent rendre compte des principes sous-jacents à leur comportement.

Un **réseau de collaboration** est une forme de réseau social où les différents individus sont amenés à travailler ensemble sur des projets communs [16]. Les interactions entre eux sont alors indirectes et se font plutôt par l'intermédiaire du projet qu'ils partagent.

Un **réseau** est dit **évolutif**, lorsque les éléments et les relations qui s'y trouvent changent au fil du temps.

Un **graphe** constitue une représentation mathématique d'un réseau. Il permet de modéliser ce dernier en désignant chaque membre par un nœud, aussi appelé sommet, et en reliant ceux-ci par des arêtes, ou liens, signifiant une relation qui les unit [17]. Les termes de réseaux et graphes ont leur propre définition. Toutefois, ils seront utilisés sans

distinction dans ce travail, étant donné qu'ils sont étroitement reliés et que l'utilisation de l'un ou de l'autre ne nuit pas à la compréhension.

Un **graphe pondéré** se distingue par l'intensité, représentée sous forme de poids, caractérisant les différents liens entre les sommets.

Un **graphe biparti** est caractérisé par la présence d'un ensemble de sommets, pouvant être divisé en deux classes distinctes. Les liens se font alors seulement d'une classe à l'autre et non entre les éléments d'une même classe.

Une **communauté** se présente souvent comme un groupe, dont les nœuds qui s'y trouvent sont densément connectés les uns aux autres, mais plus faiblement reliés d'un groupe à l'autre [18]. Il s'agit donc d'une forte concentration d'arêtes à l'intérieur même d'une communauté et d'une faible concentration de liens entre les communautés. Les sommets à l'intérieur d'une communauté partagent habituellement des propriétés communes ou possèdent un rôle similaire dans le réseau, ce qui explique leur appartenance à un même groupe. Fortunato [19] a défini le principe de communautés qui a été grandement utilisé par la suite. Il n'existe toutefois pas de définition exacte applicable dans tous les contextes.

Dans un réseau évolutif, les **communautés** qui existent au cours de plus d'une période sont qualifiées de **dynamiques**.

Une **partition** représente l'ensemble des communautés trouvées dans un graphe.

Un **instantané** désigne l'état du réseau à un moment précis ou encore regroupe les interactions ayant eu lieu dans le réseau au cours d'une période déterminée.

Un **lien intra-communauté** relie deux sommets appartenant à la même communauté. À l'inverse, un **lien inter-communautés** fait le lien entre deux nœuds se trouvant dans des communautés différentes.

Dans le milieu des affaires, un réseau s'apparente à la structure de l'entreprise, où divers employés interagissent ensemble. Ces relations entre collègues évoluent et peuvent changer avec le temps, représentant un réseau évolutif. Une communauté est formée par le regroupement de certains employés ensemble, qui interagissent fortement entre eux et peu avec les autres membres de l'organisation. La formation de ces groupes, dit informels, peut être influencée par plusieurs facteurs, dont notamment, les champs d'intérêts et la personnalité des individus qui se rassemblent [1]. Ces aspects nous seront toutefois inconnus dans l'étude qui sera réalisée.

2.2 Notations

Cette section présente les différentes notations qui seront utilisées tout au long du document, de manière à en assurer une cohésion.

Nous noterons :

- (u,v) une arête reliant les sommets u et v ;
- n le nombre de sommets du graphe ;
- m le nombre d'arêtes du graphe ;
- $G = G(V,E,W)$ un graphe statique pondéré et non orienté composé d'un ensemble de sommets V, d'un ensemble d'arêtes E et d'une matrice W, de taille $n \times n$ comportant les poids de chaque arête, par exemple, $W(u,v) = w_{uv}$ correspond au poids sur l'arête (u,v) de E. Si aucune arête ne relie deux sommets, le poids correspondant dans la matrice est nul. Si le réseau est non pondéré, tous les poids des arêtes sont à un et le graphe peut être exprimé par $G = G(V,E)$;
- $N(u)$ l'ensemble des voisins du sommet u de V ;
- A la matrice d'adjacence, de taille $n \times n$, où

$$A(u,v) = a_{uv} = \begin{cases} 1 & \text{si } (u,v) \text{ fait partie de } E, \\ 0 & \text{sinon.} \end{cases} ;$$

- d_u le degré du sommet u de V, correspond au nombre de sommets qui lui sont adjacents ;

— s_u la force d'un sommet u de V, correspond à la somme des poids sur les arêtes ayant u à une extrémité

$$s_u = \sum_{v \in V} a_{uv} w_{uv};$$

— $\mathscr{C} = \{C_1, C_2, ..., C_k\}$ la partition comprenant l'ensemble des sous-graphes ou communautés C_i du graphe G, où $C_i = G(V_{C_i}, E_{C_i}, W_{C_i})$;

— $N_{\mathscr{C}}$ le nombre de communautés du graphe ;

— n_{C_i} le nombre de sommets de la communauté C_i ;

— $m_{C_i}^{int}$ le nombre d'arêtes ayant leurs deux extrémités dans la communauté C_i ;

— $m_{C_i}^{ext}$ le nombre d'arêtes ayant une seule extrémité dans la communauté C_i ;

— Com_u un vecteur d'étiquettes représentant l'affiliation du sommet u à une ou plusieurs communautés ;

— $d_{C_i}^{int}(u)$ le degré interne du sommet u dans la communauté C_i, correspond au nombre de sommets de la communauté C_i qui lui sont adjacents ;

— $d_{C_i}^{ext}(u)$ le degré externe du sommet u dans la communauté C_i, correspond au nombre de sommets à l'extérieur de la communauté C_i qui lui sont adjacents.

Dans le contexte de graphes évolutifs, nous aurons aussi :

— $\mathscr{G} = \{G^{(0)}, G^{(1)}, ..., G^{(t_{max})}\}$ un graphe évolutif composé d'un ensemble d'instantanés, où $G^{(t)} = G(V^{(t)}, E^{(t)}, W^{(t)})$ représente l'état du graphe au temps t ;

— $\mathscr{C}^{(t)} = \{C_1^{(t)}, C_2^{(t)}, ..., C_k^{(t)}\}$ la partition comprenant l'ensemble des sous-graphes ou communautés $C_i^{(t)}$ de l'instantané $G^{(t)}$.

Les notations utilisées pour les graphes statiques et leurs communautés sont aussi applicables au contexte des graphes évolutifs, en y ajoutant seulement un indice t, correspondant au temps t étudié, soit l'instantané $G^{(t)}$.

Chapitre 3

Revue de littérature

L'étude des communautés dans les réseaux est un sujet récent qui a intéressé plusieurs spécialistes de différents domaines. Par les travaux de Girvan et Newman, le problème initial de partition dans un graphe a rejoint la physique statique ainsi que les mathématiques. Les problèmes de communautés dans les réseaux en sont donc émergés et ont été appliqués à plusieurs contextes et réalités. Notamment, la présence de communautés dans différents types de réseaux, dont notamment les réseaux biologiques [9], sociaux [10] [11] et technologiques [12], a été analysée sous plusieurs angles.

La taille des différents graphes à l'étude a grandement augmenté au cours des années avec la venue et la grande disponibilité des ordinateurs [13]. Les réseaux de communication permettant l'analyse de données ont pris de l'expansion, donnant lieu à des graphes à plus grande échelle qu'auparavant. De nouvelles méthodes pour l'analyse de réseaux, considérés complexes par leur grande taille, pouvant atteindre les millions ou milliards de sommets, ont alors vu le jour.

La représentation fidèle de la réalité par le biais des graphes a nécessité l'ajout de certaines caractéristiques à ceux-ci, multipliant ainsi les variantes étudiées. C'est le cas notamment des poids ajoutés aux différents liens, en guise de l'intensité des relations qu'ils traduisent. Il en est de même des graphes bipartis où les divers sommets se divisent en deux sous-ensembles différents. Les liens se font alors entre les sommets des différents sous-ensembles et non à l'intérieur d'un même sous-ensemble. [4]. La composante temporelle, qui permet de caractériser l'évolution des communautés dites dynamiques, a aussi marqué l'étude des réseaux et leur composition.

La détection des communautés dans les réseaux a plusieurs applications pratiques [18], d'où l'importance grandissante accordée récemment à ce type de problèmes. Pour les réseaux sociaux, par exemple, cela permet de comprendre les différents regroupements d'individus selon leurs intérêts ou toute autre caractéristique. Les pages abordant des sujets similaires sur le web peuvent également être regroupées pour former des communautés. Bref, la détection de communautés dans les réseaux permet de mieux comprendre ceux-ci et de les gérer plus facilement.

Le chapitre qui suit présente un survol des différents concepts, en lien avec la problématique de ce mémoire, qui ont déjà été abordés dans la littérature. La première section traite des réseaux, à la fois par rapport à leur structure, leur modélisation et leur étude. La section qui suit se concentre plutôt sur la détection des communautés dans les réseaux. Elle souligne notamment l'analyse des communautés et l'évaluation de leur qualité, en plus des différentes approches de résolution adaptées aux contextes statique et évolutif.

3.1 Les réseaux

Comme présenté au chapitre précédent, un réseau comprend un ensemble d'éléments qui interagissent ensemble. Toutefois, sa structure peut varier. Par exemple, il peut être qualifié de pondéré. Des poids sont alors affectés à ses liens afin de représenter l'intensité des différentes relations entretenues [20].

Un réseau peut également être caractérisé par une structure bipartie. Plusieurs définitions existent à ce point de vue, mais c'est celle reliée au réseau de collaboration qui sera présentée ici en raison du contexte de ce travail. Dans ce cas, un réseau biparti se définit comme un ensemble d'acteurs qui travaillent sur un ensemble de projets communs. La contribution des individus à chaque projet est représentée par un lien [16]. Les éléments du réseau se divisent donc en deux sous-ensembles distincts et les liens se font d'un sous-ensemble à l'autre seulement, c'est-à-dire entre les acteurs et les projets.

Un réseau évolutif se définit autant comme un réseau traditionnel, un réseau pondéré ou un réseau biparti auquel une information temporelle est ajoutée. Ainsi, peu importe les caractéristiques de sa structure, sa composition change à travers le temps, que ce soit par les éléments, les liens ou encore, les poids qui le composent [21]. Les individus et les relations, dans le contexte des réseaux sociaux, peuvent apparaitre à un certain

moment et disparaitre à un autre [22].

Les graphes permettent de modéliser plusieurs types de réseaux constitués dans la réalité [20]. Dans les graphes les plus simples, le lien entre les sommets est de type binaire, c'est-à-dire qu'il existe ou non. Deux sommets reliés par une arête sont qualifiés d'adjacents ou de voisins. Le degré d'un nœud représente le nombre de sommets auxquels il est adjacent.

Les graphes pondérés ont une troisième composante qui s'ajoute aux sommets et aux liens, soit les poids assignés aux arêtes. Selon le même principe que décrit précédemment, ceux-ci permettent de refléter l'intensité de chacune des relations dans le graphe. Dans ce contexte, le type de liens entre les nœuds n'est plus binaire, mais plutôt pondéré. Les poids sont alors présentés dans une matrice correspondant à tous les liens possibles entre les sommets du graphe [23]. Lorsque deux nœuds ne sont pas connectés ensemble, le poids attribué est 0. Cette matrice est normalement symétrique en raison de liens non orientés entre les sommets du graphe [24].

Dans ce type de graphe, les sommets ne sont pas distingués par leur degré, mais plutôt par leur force, qui se compose à la fois de leur degré et du poids de chacune de leurs arêtes [25]. En d'autres mots, il s'agit de la somme des poids de chacun des liens que comporte un sommet donné.

Selon la structure du réseau, les poids attribués aux arêtes touchant un nœud déterminé peuvent être d'un même ordre de grandeur [26]. Dans ce cas, chaque poids établi est identique et correspond à la force du sommet étudié divisé par son degré. Dans le cas inverse, les poids caractérisant certains liens du nœud en question dominent sur ceux de ses autres arêtes. Il existe alors une mesure de disparité pour évaluer la différence de grandeur entre les poids attribués aux liens d'un sommet donné. La disparité moyenne d'un nœud u est représentée par $Y(u)$ et s'énonce comme suit

$$Y(u) = \sum_{v \in V} \left[\frac{w_{uv}}{s_u} \right]^2 . \tag{3.1}$$

Si les arêtes du sommet donné ont des poids semblables, la valeur de $Y(u)$ se rapproche de $1/d_u$, d_u étant le degré du sommet u. Dans ce cas, $Y(u)$ est implicitement dépendant du degré du sommet. À l'inverse, si le poids d'une arête domine ceux des autres liens, la valeur de $Y(u)$ se rapproche d'un. Cette situation reflète l'indépendance de la disparité

moyenne du nœud par rapport à son degré [27].

Pour ce qui est des réseaux bipartis, le graphe correspondant est aussi composé d'un ensemble de nœuds, pouvant être séparé en deux sous-ensembles distincts, ainsi que d'arêtes entre les sommets interagissant ensemble [4]. Les nœuds d'un même sous-ensemble ne peuvent être reliés entre eux [28], ce qui signifie que les liens se font uniquement entre les sommets de différents sous-ensembles. Les deux types de nœuds ont des appellations différentes pour les distinguer, soit les nœuds du haut et les nœuds du bas [29]. La définition de graphe biparti peut être généralisée aux cas avec plus de deux sous-ensembles [19] alors que les liens se font, encore une fois, d'un sous-ensemble à une autre seulement.

Il arrive que pour l'étude des réseaux et leurs communautés, les graphes bipartis soient transformés en graphe standard. Dans ce cas, deux nœuds du bas, connectés au même sommet du haut, sont reliés directement ensemble dans la nouvelle version du graphe [4]. À ce moment, les nœuds du haut ne sont plus représentés. La figure 3.1 représente cette conversion. Cette modification dans la modélisation du réseau peut cependant engendrer une perte d'information quant à ses éléments [29].

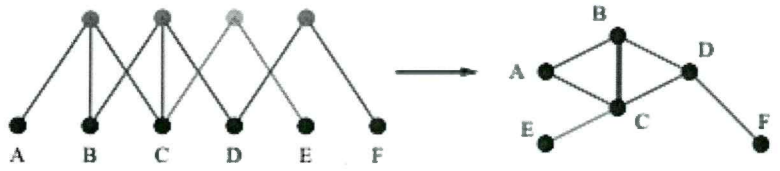


FIGURE 3.1 – Transformation d'un graphe biparti en un graphe traditionnel avec des sommets d'une seule classe [4].

Les réseaux évolutifs sont composés de nœuds et d'arêtes qui, plutôt que d'être considérés stables comme dans les graphes statiques, peuvent changer au fil du temps [30]. La modélisation de réseaux évolutifs se fait souvent par la décomposition de leur évolution en différents graphes statiques, nommés instantanés [31]. Chacun d'eux est composé, exactement comme les graphes statiques, soit des nœuds reliés entre eux par des arêtes. Ces instantanés peuvent être de deux types. Il existe ceux tirés d'un réseau original à un instant précis. Dans ce cas, les liens entre les sommets du réseau sont continus dans le temps, c'est-à-dire qu'ils représentent une relation durable. Les instantanés sont alors représentatifs de l'état du réseau à l'instant même où ils sont collectés. À l'inverse, les

liens entre les sommets peuvent constituer des interactions ponctuelles, c'est-à-dire non durables dans le temps. Dans ce contexte, les instantanés regroupent toutes les interactions entre les éléments du réseau, ayant eu lieu durant un certain intervalle de temps. Ces interactions sont alors représentées dans le même graphe statique correspondant à la période analysée. Les instantanés sont habituellement tirés du réseau original à des intervalles réguliers [32].

Les réseaux sociaux habituels sont notamment caractérisés par certains principes, dont la connectivité et la centralité [13]. La connectivité cherche à comprendre la présence ou non de liens entre les individus ainsi que la nature des différentes relations dans le réseau. La centralité se définit plutôt comme la présence d'un individu plus fortement lié aux autres membres du réseau [33]. Cet acteur a alors une position dite centrale et est reconnu avoir de l'influence sur le groupe. Étant relié à plusieurs autres membres, donc ayant un degré élevé, l'information qu'il transmet, qu'elle soit vraie ou fausse, va rapidement être diffusée dans le réseau. Dans le cas des graphes pondérés, la détermination de la centralité d'un individu considère plutôt la force du sommet, soit le nombre de liens qu'il possède en plus des poids attribués à ceux-ci [25]. Ainsi, les nœuds ayant une force plus élevée sont jugés des plus importants dans le réseau. Une mesure de centralité a été proposée par Holme et al. [30] pour le contexte des réseaux évolutifs. Elle cherche à déterminer à quelle vitesse l'information provenant d'un sommet atteint les autres nœuds du graphe, relativement à l'aspect temporel caractérisant ce dernier.

La présence de communautés à l'intérieur du réseau constitue un autre champ d'étude possible et dont la popularité a pris de l'ampleur au cours des dernières années. Les communautés se présentent souvent comme des groupes, dont les membres qui s'y trouvent interagissent grandement ensemble, mais moins avec les autres individus du réseau [18]. Ce principe sera détaillé plus en détail à la section suivante.

Les réseaux peuvent être étudiés sous trois principaux aspects, soit la centralité, les communautés et la propagation d'information. Ce travail se concentrera principalement sur l'analyse des communautés présentes dans les réseaux.

3.2 Détection de communautés

Comme présenté précédemment, une communauté se distingue par une forte concentration d'arêtes entre ses sommets et une plus faible entre ceux-ci et ceux des autres

communautés. Cette densité caractérisant les liens entre les sommets est représentée par la force de ces derniers, dans le cas des graphes pondérés. Ainsi, une communauté se présente comme un groupe de nœuds qui entretiennent des liens forts entre eux et plus faibles avec le reste du réseau [15]. Cette définition permet de refléter l'influence des poids sur la formation des communautés [34]. Selon Newman [23], ce sont surtout les poids élevés, en guise d'une relation étroite entre deux sommets ou une similarité entre ceux-ci, qui fournissent les informations les plus utiles à propos des communautés.

Des mesures ont été mises en place pour caractériser la densité à l'intérieur d'une communauté, $\delta_{int}(C_i)$, ainsi que celle à l'extérieur, soit entre les communautés, $\delta_{ext}(C_i)$, et s'énoncent comme suit

$$\delta_{int}(C_i) = \frac{m_{C_i}^{int}}{n_{C_i}(n_{C_i}-1)/2} \tag{3.2}$$

$$\delta_{ext}(C_i) = \frac{m_{C_i}^{ext}}{n_{C_i}(n-n_{C_i})} \ . \tag{3.3}$$

L'objectif, relativement à la définition d'une communauté, est d'obtenir une densité, dite intra-communauté, la plus élevée possible et une densité inter-communautés la plus faible possible [19].

Les communautés peuvent être étudiées selon deux approches [19]. Elles peuvent être définies localement, c'est-à-dire étudiées indépendamment du reste du réseau. Elles sont alors considérées comme un sous-groupe autonome. Le principe de clique est souvent relié à cette approche et représente un sous-ensemble de sommets tous connectés les uns aux autres. Les communautés peuvent également être définies globalement, soit en considérant l'ensemble du réseau. Dans ce cas, la structure d'un graphe en communautés est évaluée en comparaison avec un graphe aléatoire. C'est cette deuxième approche qui sera utilisée dans ce mémoire.

Le recouvrement de communautés constitue un phénomène observable dans certains réseaux [35]. Par exemple, dans le cas de communautés avec recouvrement, un même nœud peut appartenir à plus d'un groupe à la fois. À l'inverse, quand un sommet ne peut se trouver dans plus d'une communauté au même moment, les communautés sont dites sans recouvrement.

Les réseaux peuvent également être caractérisés par une certaine hiérarchie [36]. En ce sens, dans une communauté tirée du graphe, il existerait une communauté plus petite, dans laquelle il peut se trouver une autre communauté plus faible encore et ainsi de suite. En d'autres mots, des sous-groupes de plusieurs niveaux sont alors présents, imbriqués les uns dans les autres. Le dendrogramme est, notamment, un outil servant à visualiser et analyser les communautés faisant partie de telles hiérarchies. Les plus petites communautés sont d'abord trouvées et à chaque étape, elles se regroupent deux à deux pour former une communauté à un niveau supérieur [37].

Hopcroft et al. [38] ont été parmi les premiers à suggérer l'étude de communautés dans les réseaux évolutifs. Dans ce contexte, ces dernières sont dites dynamiques puisqu'elles se modifient en fonction des changements qui surviennent dans la composition du réseau. La définition des communautés de Fortunato [19], présentée précédemment, est applicable également dans le contexte temporel. S'il s'agit plutôt d'un réseau pondéré, la définition de communautés respective à ce type de réseau est applicable à celles dynamiques. Toutefois, peu importe la forme du réseau évolutif, une particularité s'ajoute à la définition des communautés dynamiques. En effet, celles-ci doivent être présentes dans le réseau à plusieurs occasions à l'intérieur de l'horizon temporel étudié [32]. Une communauté dynamique regroupe donc un ensemble de communautés, chacune relevée à un certain moment dans le temps. Chacune d'elles agit à titre d'observation, à un instant précis, de la communauté dynamique. Cette dernière peut ainsi être représentée par une ligne du temps composée des communautés détectées à chaque moment de l'analyse.

Les communautés dynamiques peuvent notamment être catégorisées de communautés naturelles. Il s'agit en fait de communautés qui ne sont que légèrement affectées lors de perturbations mineures du graphe [38]. Ces perturbations consistent à éliminer une certaine partie des nœuds du réseau ainsi que leurs arêtes respectives.

Asur et al. [22] ont défini divers évènements concernant l'évolution des individus au sein des communautés dynamiques influençant, du même coup, l'évolution de celles-ci.

- L'**apparition** représente l'arrivée d'un individu dans le réseau.
- La **disparition** caractérise le départ d'un individu du réseau.
- L'**affiliation** survient lorsqu'un individu se joint à une communauté.
- La **désaffiliation** consiste au départ d'un individu d'une communauté.

Les mêmes auteurs [22] ainsi que Palla et al. [39] ont présenté divers évènements qui peuvent survenir dans l'évolution des communautés dynamiques.

— La **continuation** signifie que tous les nœuds composant une communauté à un certain moment t y sont encore au temps $t+1$. Cet évènement indique qu'il n'y a aucun ajout ou retrait de nœuds à la communauté entre les deux moments observés. Les liens entre les sommets peuvent toutefois être différents.

— La **naissance** caractérise une communauté qui regroupe, au temps $t+1$, des sommets n'étant aucunement regroupés les uns aux autres au temps t.

— La **mort** d'une communauté survient lorsque tous les nœuds qui la composaient au temps t s'y détachent au temps $t+1$.

— La **fusion** représente le regroupement de deux communautés distinctes au temps t, ou une certaine proportion de leurs sommets, en une seule et même communauté au temps $t+1$.

— La **division** apparait lorsqu'une seule et même communauté au temps t, ou une certaine proportion de ses sommets, se divise en deux communautés distinctes au temps $t+1$.

— La **croissance** consiste à l'ajout de sommets au temps $t+1$ à une communauté existante au temps t. La taille de cette dernière augmente.

— La **contraction** signale le retrait de sommets au temps $t+1$ d'une communauté existante au temps t. La taille de cette dernière diminue.

— La **résurgence**, proposée par Cazabet [31], réfère à une communauté qui meurt à un certain moment puis nait à nouveau quelque temps plus tard. Il s'agit alors de communautés saisonnières, périodiques ou encore dont l'importance varie au fil du temps.

Dans un autre ordre d'idées, vérifier la qualité des communautés trouvées dans un réseau lorsqu'elles ne sont pas connues dans la réalité peut représenter un défi. En ce sens, certaines mesures ont été établies afin d'évaluer quantitativement leur qualité.

Radicchi [40] propose justement une telle mesure. Selon la définition de l'auteur, une **communauté** est caractérisée de **forte** si chaque nœud u qui s'y trouve a davantage de liens à l'intérieur de la communauté C_i qu'avec l'extérieur, soit

$$d_{C_i}^{int}(u) > d_{C_i}^{ext}(u). \tag{3.4}$$

À l'inverse, une **communauté** est qualifiée de **faible** si la somme des degrés à l'intérieur de celle-ci est supérieure à la somme des degrés dans le reste du réseau. Ce critère s'obtient par

$$\sum_{u \in C_i} d_{C_i}^{int}(u) > \sum_{u \in C_i} d_{C_i}^{ext}(u). \tag{3.5}$$

Cette mesure peut être étendue au contexte des réseaux pondérés [19]. Dans ce cas, une **communauté** est dite **forte,** si la force reliée aux arêtes intra-communauté des nœuds qu'elle regroupe est plus élevée que celle référant aux arêtes inter-communautés de ces sommets. À l'opposé, une **communauté** est jugée **faible** si la somme des forces de ses nœuds est supérieure à la force totale des autres sommets du réseau.

Cafieri et al. [41] ont établi le **ratio des arêtes** qui permet de comparer le nombre de liens à l'intérieur d'une communauté avec le nombre d'arêtes dont une seule des extrémités est dans la communauté. Il se calcule comme suit

$$r(C_i) = \frac{\sum_{u \in C_i} d_{C_i}^{int}(u)}{\sum_{u \in C_i} d_{C_i}^{ext}(u)}. \tag{3.6}$$

Pour une bonne qualité, un ratio élevé est souhaité.

La **conductance** d'une communauté constitue un autre critère d'évaluation de la qualité de celle-ci [42]. Il s'agit de la mesure des liens externes d'une communauté relativement à sa densité interne. Elle s'obtient de la manière suivante

$$\Phi(C_i) = \frac{m_{C_i}^{ext}}{min\{2m_{C_i}^{int} + m_{C_i}^{ext}, 2(m - m_{C_i}^{int}) - m_{C_i}^{ext}\}}. \tag{3.7}$$

Quant à eux, Shi et al. [43] ont développé la **coupe normalisée** qui évalue la différence entre une communauté et le reste du réseau, par l'expression suivante

$$\Phi_N(C_i) = \frac{m_{C_i}^{ext}}{2m_{C_i}^{int} + m_{C_i}^{ext}} + \frac{m_{C_i}^{ext}}{2(m - m_{C_i}^{int}) - m_{C_i}^{ext}}. \tag{3.8}$$

L'**expansion** est une autre mesure permettant de caractériser la qualité d'une communauté [31]. Elle représente le nombre de liens moyen qu'un nœud entretient avec des sommets à l'extérieur de son groupe. Elle se calcule ainsi

$$exp(C_i) = \frac{m_{C_i}^{ext}}{n_{C_i}}. \tag{3.9}$$

Le critère de **modularité** présenté par Girvan et Newman [44] est l'un des plus utilisés pour évaluer la qualité des communautés. Il compare la densité d'un sous-graphe représentant une communauté avec celle d'un sous-graphe où les arêtes sont distribuées aléatoirement en respect du degré des sommets. Il s'obtient comme suit

$$mod(C_i) = \frac{m_{C_i}^{int}}{m} - \left(\frac{\sum\limits_{u \in C_i} d_u}{2m} \right)^2. \tag{3.10}$$

L'hypothèse à la base de ce critère consiste au fait qu'un graphe aléatoire ne devrait pas avoir une structure en communautés. Ainsi, plus la densité de la partition analysée s'éloigne de celle d'un graphe aléatoire, plus les communautés détectées sont qualifiées de bonne qualité. La modularité totale d'une partition est représentée par la somme des modularités de chacune des communautés, décrite par

$$mod(\mathscr{C}) = \sum_{C \in \mathscr{C}} \left[\frac{m_C^{int}}{m} - \left(\frac{\sum\limits_{u \in C} d_u}{2m} \right)^2 \right]. \tag{3.11}$$

Elle peut également s'obtenir par la formule suivante, qui est équivalente, soit

$$mod(\mathscr{C}) = \frac{1}{2m} \sum_{u,v \in V} [A_{uv} - P_{uv}] \, \delta \left(Com_u, Com_v \right) \tag{3.12}$$

où P_{uv} représente la probabilité qu'il y ait une arête entre les nœuds u et v dans le graphe aléatoire, obtenu par $P_{uv} = \frac{d_u d_v}{2m}$, et $\delta(Com_u, C_v)$, soit le delta de Kronecker, vaut 1 si les deux nœuds sont dans la même communauté, 0 sinon. L'équation (3.12) peut être

réécrite ainsi

$$mod(\mathscr{C}) = \frac{1}{2m} \sum_{u,v \in V} \left[A_{uv} - \frac{d_u d_v}{2m} \right] \delta\left(Com_u, Com_v\right). \tag{3.13}$$

Ce critère a été généralisé par Newman afin de l'adapter aux réseaux pondérés [23]. Il permet de comparer la distribution des poids au sein d'une communauté avec celle présente dans un sous-graphe aléatoire obtenu en respectant la force des sommets. L'auteur utilise une transformation du réseau pondéré en multigraphes, où les liens sont dupliqués en fonction de la valeur du poids qui les caractérise. Il y applique ensuite une formule de modularité très semblable à celle présentée précédemment. L'équation adaptée est représentée ainsi [45]

$$mod(\mathscr{C}) = \frac{1}{2w} \sum_{u,v \in V} \left[w_{uv} - \frac{s_u s_v}{2w} \right] \delta\left(Com_u, Com_v\right). \tag{3.14}$$

Dans ce contexte, les valeurs de la matrice A_{uv} sont substituées par les poids entre les nœuds u et v et les degrés sont remplacés par les forces respectives. La valeur de m représentant le nombre d'arêtes du réseau est remplacé par w correspondant à la somme des poids du réseau, obtenue par $w = \frac{1}{2} \sum_{u,v \in V} w_{uv}$. Les significations des autres variables demeurent les mêmes. La même équation peut aussi être présentée, de façon équivalente, comme la somme des modularités de chacune des communautés du réseau [19], soit

$$mod(\mathscr{C}) = \sum_{C \in \mathscr{C}} \left[\frac{w_c}{w} - \left(\frac{\sum_{u \in C} s_c}{2w} \right)^2 \right] \tag{3.15}$$

où w_c représente la somme des poids sur les arêtes intérieures de la communauté et s_c la somme des forces des sommets à l'intérieur de la communauté.

Le critère de **modularité** peut tout aussi bien être étendu au contexte des réseaux évolutifs [7]. Dans ce cas, il est calculé à chaque instantané tiré du réseau original et a la même signification que dans le cas des réseaux statiques. L'application de ce critère ajusté nécessite toutefois que les nœuds de chaque instantané soient reliés avec eux-mêmes d'une période à l'autre. Dans ce contexte, la modularité d'une partition $\mathscr{C}$ comprenant un ensemble d'instantanés I, se calcule de la manière suivante

$$mod(\mathscr{C}) = \frac{1}{2w} \sum_{u,v \in Vr,t \in I} \left[\left(A_{uvt} - \gamma_t \frac{s_{ut} s_{vt}}{2m_t} \right) \delta_{rt} + \delta_{uv} B_{vrt} \right] \delta(Com_{ut} Com_{vr}) \quad (3.16)$$

où w est la somme des poids du réseau, γ_t est la résolution de l'instantané t, δ_{rt} vaut 1 si les instantanés r et t sont équivalents et 0 sinon, δ_{uv} vaut 1 si les nœuds u et v sont équivalents et 0 sinon, B_{vrt} vaut 1 si le nœud v de l'instantané r est relié avec lui-même dans l'instantané t et finalement, $\delta(Com_{ut} Com_{vr})$ vaut 1 si les nœuds u de l'instantané t et v de l'instantané r sont dans la même communauté.

Chakrabarti et al. [46] ont développé une autre mesure de la qualité des communautés détectées dans un réseau évolutif. Ce critère permet de considérer deux aspects simultanément, soit la **qualité statique** (Q_{stat}) et la **qualité séquentielle** de la structure modulaire (Q_{seq}). La qualité statique permet de s'assurer que les communautés formées à tout moment sont le reflet des données réelles du réseau. La qualité séquentielle mesure, pour sa part, la similarité temporelle. En d'autres mots, elle cherche à ce que l'état d'un nœud ou d'une communauté soit le plus semblable possible d'une période à l'autre. Un paramètre α permet de distribuer une valeur d'un entre ces deux mesures. La formule s'énonce donc comme suit

$$Q = \alpha Q_{stat} + (1 - \alpha) Q_{seq} \ . \quad (3.17)$$

La qualité d'une communauté représente à quel point le regroupement de certains employés se distingue des autres, par la quantité d'échanges réalisés entre eux comparativement aux discussions engendrées avec les autres membres de l'organisation.

La comparaison entre deux partitions est également possible et permet de mesurer la ressemblance entre leurs communautés respectives. Il existe notamment la mesure de l'**Information Mutuelle** (IM) [47][48] qui permet d'établir l'information commune partagée par deux partitions et l'**Information Mutuelle Normalisée** (IMN) [49][50] qui constitue une mesure de similarité entre ces dernières. L'indice de **Jaccard** [51]

permet d'établir la proportion du nombre de paires de sommets classés dans la même communauté dans deux partitions différentes, alors que l'indice de **Rand** [52] est similaire, mais plus complet. Ce dernier n'est toutefois utilisable que pour les partitions dont la solution est connue. Comme l'approche de résolution qui sera utilisée dans ce mémoire évite l'appariement de communautés, ces mesures ne seront pas utilisées et donc, pas détaillées davantage.

3.2.1 Méthodes de détection de communautés dans les réseaux statiques

Une multitude d'approches existent afin de détecter des communautés dans les réseaux. Fortunato en a fait une revue très détaillée en 2010, sans toutefois considérer la totalité d'entre elles [19]. Chaque méthode comporte ses particularités et est applicable pour un certain type de graphe. Les approches originales concernent les graphes statiques, c'est-à-dire stables dans le temps. Elles peuvent, dans certains cas, permettre les communautés avec recouvrement et dans d'autres cas, les communautés hiérarchiques. Les approches hiérarchiques, soit agglomératives [37] ou divisives [18][44], sont parmi les premières à avoir été étudiées. Il existe aussi les approches basées sur la modularité comprenant les algorithmes goutons [37][53], le recuit simulé [54], l'optimisation extrémale [55] et l'optimisation spectrale [56]. Des méthodes à base de cliques [35] ont également été développées en plus d'approches référant à la propagation d'étiquettes [57]. Les algorithmes spectraux [58] et l'inférence statistique [48], dont les modèles génératifs [59], constituent différents domaines de résolution. Des algorithmes dynamiques dont notamment, les marches aléatoires [60] et les modèles de spin [61] sont d'autres approches considérées pour la détection de communautés dans les réseaux.

Pour ne pas dépasser le cadre de ce travail, seulement quelques-unes de ces méthodes seront abordées. Il s'agit en fait des plus pertinentes à la problématique qui sera traitée. Les méthodes de détection de communautés pour les réseaux statiques seront d'abord présentées dans cette section afin de mieux comprendre leur adaptation aux graphes évolutifs à la section suivante.

Le concept de propagation d'étiquettes a été utilisé par Raghavan et al. [57] pour le développement d'un algorithme permettant la détection de communautés dans les ré-

seaux. Selon cet algorithme, **LPA** *(Label Propagation Algorithm)* ou **RAK** en référence aux auteurs, chaque nœud du graphe est associé à une étiquette, représentant la communauté à laquelle le sommet appartient et pouvant changer au cours de la résolution. Initialement, chacun des nœuds possède une étiquette unique. À chaque itération, l'étiquette d'un sommet est mise à jour en prenant la valeur de celle qui revient le plus souvent parmi ses voisins, comme l'énonce le critère de mise à jour suivant

$$l_v^{nouv} = argmax_l \left(\sum_{u \in V} A_{uv} \delta(l_u, l) \right) \tag{3.18}$$

où l_v^{nouv} représente la valeur de la nouvelle étiquette du nœud v, l_u l'étiquette actuelle du sommet u et $\delta(l_u, l)$ vaut 1 si l'étiquette l_u vaut l et 0 sinon. Si plusieurs étiquettes reviennent à la même fréquence parmi les sommets adjacents au nœud analysé, une d'entre elles est sélectionnée aléatoirement et devient la nouvelle étiquette du sommet. Ce procédé est répété jusqu'à ce que chaque nœud possède l'étiquette la plus fréquente parmi ses voisins. Les communautés sont alors représentées par les sommets ayant les mêmes étiquettes. Cet algorithme permet la résolution de graphe de grandes tailles assez rapidement. Cependant, la solution trouvée est influencée par l'ordre dans lequel les nœuds sont analysés. Ainsi, plusieurs solutions différentes peuvent être trouvées pour un même graphe [62].

Une variante à **LPA** a été proposée par Gregory [63] sous le nom de **COPRA** *(Community Overlapping PRopagation Algorithm)*. Initialement, une étiquette unique est attribuée à chacun des sommets. À chaque itération, le vecteur d'un nœud est construit en enregistrant les étiquettes de tous ses voisins en plus d'un coefficient caractérisant la force de chacun de ses liens. Les étiquettes enregistrées, ayant un coefficient respectif sous un seuil déterminé, sont éliminées du vecteur. Si tous les coefficients sont sous le seuil, l'étiquette reliée au coefficient le plus élevé est conservée seulement. L'algorithme s'arrête après un nombre donné d'itérations. Selon les expérimentations réalisées par l'auteur, l'algorithme **COPRA** semble bien réussir à détecter les communautés dans les réseaux de toute taille et possède une grande vitesse d'exécution. Malgré tout, les solutions obtenues sont influencées, tout comme le **LPA**, par l'ordre d'analyse des sommets du réseau. Le choix aléatoire entre plusieurs étiquettes de même fréquence parmi les voisins d'un nœud a toutefois moins d'impact sur la solution, étant donné que plusieurs étiquettes sont gardées en mémoire pour un même nœud, évitant ainsi les

choix arbitraires.

Le même auteur a généralisé cet algorithme au cas des réseaux pondérés. Il s'applique sensiblement de la même manière, à une exception près. Le coefficient enregistré dans le vecteur est, dans ce contexte, multiplié par le poids sur l'arête étudiée. Le déploiement de l'algorithme reste entièrement le même pour la suite. Les tests réalisés par l'auteur démontrent qu'il s'illustre mieux que d'autres approches populaires pour les réseaux pondérés en plus de permettre la détection de communautés avec recouvrement. Il est aussi très rapide lors de la résolution.

LabelRank [64] constitue un autre algorithme développé à partir de la propagation d'étiquettes. Sensiblement comme le **COPRA**, des vecteurs sont assignés à chaque nœud et permettent d'enregistrer les étiquettes de leurs sommets adjacents en y ajoutant toutefois la probabilité d'appartenir à une certaine communauté. Les étiquettes subissent d'abord la phase de propagation, où une probabilité de distribution leur est assignée. L'inflation qui suit contracte la propagation en augmentant les hautes probabilités et en diminuant les plus faibles. Une coupure est ensuite réalisée de manière à éliminer, dans chacun des vecteurs, les étiquettes ayant une probabilité sous un certain seuil déterminé. La dernière opération, la mise à jour conditionnelle, consiste à ajuster l'étiquette d'un nœud lorsqu'elle diffère significativement de celles de ses voisins. L'algorithme s'arrête lorsque le premier des deux évènements suivants survient, soit il n'y a aucun changement d'étiquettes pendant une itération complète ou un nombre d'itérations prédéfini est atteint. Cet algorithme permet de stabiliser les solutions obtenues avec le **LPA** en plus d'en améliorer la performance.

L'algorithme **LPAcw** *(Label Propagation Algorithm with consensus weight)* a été présenté par Lou et al. [65]. Il vise à réduire la diversité des solutions trouvées par le **LPA** pour un même réseau, en réalisant un consensus sur celles-ci. **LPAcw** semble mieux fonctionner, selon les expérimentations des auteurs, que l'algorithme de propagation d'étiquettes original, et ce, tout en étant plus stable.

Une variante de l'algorithme **LPA** a été développée par Barber et Clark [66]. Il s'agit du **LPAm** *(Label Propagation Algorithm with modularity)*, soit la propagation d'étiquettes visant la maximisation du critère de modularité. Initialement, chaque sommet possède une étiquette unique. Une à la suite de l'autre, elles sont mises à jour en prenant la valeur de l'étiquette du nœud adjacent qui permet la plus grande augmentation de la modularité du graphe. Ce processus se fait selon la règle suivante

$$l_v^{nouv} = argmax_l \left(\sum_{u \in V} (A_{uv} - P_{uv}) \delta(l_u, l) \right) \tag{3.19}$$

où P_{uv} a été ajouté et s'obtient par $\frac{d_u d_v}{2m}$, jusqu'à ce qu'aucune amélioration ne soit possible. Cette variante est reconnue pour être aussi rapide que le **LPA** et a l'avantage additionnel de maximiser la modularité. Cependant, l'ordre de traitement des sommets a encore une influence sur la solution trouvée et l'algorithme peut rester coincé dans une solution locale. De plus, les communautés trouvées sont souvent d'un même ordre de grandeur quant au degré qu'elles possèdent.

Liu et Murata [67] ont alors proposé le **LPAm+** qui permet de sortir des optimums locaux trouvés par le **LPAm** afin de maximiser davantage la modularité du graphe. Partant de la solution obtenue par le **LPAm**, l'algorithme suggère de fusionner des communautés si cela permet un gain de modularité. Cependant, avant la fusion de deux communautés, il vérifie qu'il n'est pas plus avantageux de regrouper chacune d'elle avec une autre communauté du réseau quant au gain de modularité qu'il en résulte. La propagation d'étiquettes est appliquée à nouveau par la suite afin d'améliorer la modularité davantage. La fusion de communautés est alors analysée une autre fois et ainsi de suite jusqu'à ce qu'il n'y ait plus aucun gain de modularité possible. La partition trouvée a donc moins de chances d'être optimale que localement.

L'algorithme de **Louvain** développé par Blondel et al. [53] constitue une approche gloutonne, qui se base sur la maximisation du critère de modularité. Initialement, chaque nœud constitue lui-même une communauté. L'algorithme traite chacun des sommets en cherchant à le regrouper avec un de ses voisins de manière à obtenir un gain de modularité. Ainsi, deux sommets adjacents sont regroupés ensemble si, une fois fusionnés, leur modularité est supérieure à la somme de leur modularité individuelle. Si plusieurs nœuds adjacents permettent un gain de modularité, celui apportant le plus grand gain est sélectionné pour le regroupement. Le même procédé est répété avec le nouveau sommet ainsi formé, nommé supernoeud, dont les paramètres ont été ajustés en fonction des paramètres individuels des nœuds qui le composent. Dans le cas où aucun gain de modularité ne serait possible, l'algorithme passe au sommet suivant. Il s'arrête quand tous les nœuds ont été traités, en guise d'une modularité totale du graphe ne pouvant être améliorée davantage. La partition trouvée est un optimum local, car la liste des sommets n'est parcourue qu'une seule fois et le choix du nœud

de départ peut influencer les résultats qui suivent. Malgré tout, il a été démontré que cet algorithme s'illustre très bien et très rapidement sur des graphes non dirigés et non pondérés comparativement à d'autres méthodes [68]. Il permet aussi la résolution de graphes ayant jusqu'à 10^9 nœuds tout en étant encore assez rapide considérant leur taille. Des communautés hiérarchiques peuvent être trouvées, alors que les communautés avec recouvrement ne sont pas décelées. L'algorithme de **Louvain** est également applicable aux graphes pondérés. Dans ce contexte, son déploiement est exactement le même à la différence du critère de modularité utilisé qui est celui considérant les poids, comme présenté précédemment. Ces avantages en font l'un des algorithmes les plus utilisés pour la détection des communautés dans les réseaux.

Clauset et al. [69] suggèrent **CNM** (pour Clauset, Newman et Moore), un algorithme hiérarchique agglomératif. Initialement, chaque sommet du graphe est considéré comme une communauté distincte. À chaque itération, deux communautés sont fusionnées de manière à obtenir un gain de modularité optimal, jusqu'à ce qu'une seule communauté comprenne tous les sommets du graphe. Toutes les étapes de fusion sont reflétées dans un dendrogramme, duquel on extrait la valeur optimale de la modularité et la structure modulaire qui y correspond. Cette approche est reconnue pour son efficience par l'utilisation de structure de données ainsi que sa rapidité, même sur de larges graphes.

Duch et Arenas [70] proposent un algorithme basé sur l'optimisation extrémale, **EO** *(Extremal Optimization)* [55]. Comme cette heuristique de recherche locale permet l'obtention de bons résultats, en plus d'une rapidité d'exécution, Duch et Arenas ont cherché à l'utiliser pour la détection de communautés dans les réseaux. Pour ce faire, leur algorithme consiste à optimiser localement la disposition des nœuds dans les communautés, de manière à maximiser leur contribution à la modularité totale du graphe qui en découle. Initialement, tous les sommets sont regroupés aléatoirement en deux communautés de taille semblable. À chaque itération, le nœud ayant la plus faible contribution à une communauté y est retiré et déplacé dans l'autre communauté, de manière à augmenter la modularité du graphe. Les contributions sont alors mises à jour et le processus recommence, jusqu'à ce qu'aucune amélioration de la modularité ne soit possible. Les liens entre les deux communautés sont alors supprimés et l'algorithme est appliqué de façon récursive dans chacune d'elles. Cette boucle est répétée jusqu'à ce que la modularité ne puisse augmenter davantage. Les tests réalisés avec cet algorithme ont démontré qu'il permettait de bons résultats, et ce, lorsqu'appliqué à des réseaux de différentes tailles.

L'algorithme **EO** peut également être généralisé aux réseaux pondérés. Fan et al. [45] ont proposé **WEO** *(Weighted Extremal Optimization)* qui permet d'optimiser la modularité en tenant compte des poids. Pour le reste, la résolution se fait exactement de la même façon que dans les réseaux statiques. L'expérimentation de **WEO** a révélé qu'il fonctionnait bien et permettait mieux, que d'autres méthodes existantes, de comprendre la structure des réseaux pondérés réels.

L'algorithme **CPM** *(Clique Percolation Method)* présenté par Palla et al. [35] est l'un des plus connus pour la détection de communautés dans les réseaux. Cette méthode se base sur la formation de k-cliques dans le réseau, c'est-à-dire le regroupement des nœuds en cliques de tailles prédéfinies k. Une clique étant un ensemble de sommets tous reliés ensemble deux à deux. Une fois toutes les k-cliques possibles identifiées, une recherche locale est effectuée afin de déterminer celles dont seulement un nœud diffère dans leur composition pour ensuite les disposer dans la même communauté. L'algorithme s'arrête lorsque toutes les k-cliques ont été analysées relativement aux autres quant à leur regroupement en communautés. Ce procédé permet l'appartenance d'un sommet à plus d'une communauté ainsi que la présence de nœuds non regroupés. Malgré tout, cet algorithme se déploie moins bien dans les réseaux ayant une structure trop dense ou trop peu dense en raison de l'approche par cliques qui s'y adapte moins.

Cet algorithme est généralisable au contexte des réseaux pondérés [35]. Il possède alors deux paramètres, soit k la taille des k-cliques et w^*, le seuil au-dessus duquel les poids caractérisant les liens entre deux sommets doivent s'élever pour que les arêtes correspondantes soient considérées. Pour tenir compte de tous les liens du réseau, ce seuil peut simplement être établi à 0. Toutes les arêtes ayant un poids respectif sous le seuil w^* sont retirées du réseau, alors que les autres y sont conservés. Une fois tous les liens étudiés, l'algorithme est appliqué de manière identique à ce qui est fait avec un graphe standard. En fait, la valeur des poids sert seulement au tri des arêtes et n'a plus aucune importance lors de la résolution. La difficulté de cet algorithme concerne essentiellement la détermination initiale des deux paramètres, car les valeurs de ceux-ci influencent la partition finale trouvée.

OSLOM *(Order Statistics Local Optimization Method)* constitue une approche probabiliste présentée par Lancichinetti et al. [71]. Cette méthode consiste à optimiser localement une communauté de manière à ce qu'elle diverge significativement, au sens statistique, d'un modèle obtenu aléatoirement, comme présenté avec le critère de modularité. Le graphe initial peut être une partition trouvée par une autre approche ou

encore, un réseau non analysé. Dans ce deuxième cas, l'algorithme regroupe aléatoirement les sommets en communautés. Pour chaque communauté de la partition de départ, **OSLOM** vérifie la probabilité, par rapport au modèle nul, que les nœuds qui n'en font pas partie, mais dont un des voisins s'y trouve, y appartiennent aussi. Si la probabilité est suffisamment élevée, soit considérée statistiquement significative, ils y sont ajoutés. À l'inverse, les sommets dont la probabilité d'appartenir à ce groupe est non-significative, y sont retirés. La structure interne des communautés ainsi trouvées est analysée afin d'évaluer la fusion de certaines d'entre elles. Le choix de celles à étudier se fait aléatoirement en sélectionnant un sommet du réseau et en s'intéressant à la communauté à laquelle il appartient. Cette approche permet de considérer des nœuds non affiliés. La procédure est effectuée à plusieurs reprises, en raison de son caractère aléatoire, de manière à obtenir plusieurs solutions dont la distribution des sommets est statistiquement significative. L'application de cet algorithme à des réseaux de différentes tailles et de différents types a permis l'obtention d'excellents résultats. **OSLOM** a justement l'avantage d'être applicable à plusieurs types de réseaux, comme les réseaux comportant des communautés hiérarchiques ou avec recouvrement notamment.

Cet algorithme est aussi convenable aux réseaux pondérés [71]. Pour ce faire, une variable est ajoutée, soit la probabilité, pour une arête, d'être caractérisée par un certain poids relativement à un modèle nul, où la force des sommets ainsi que la distribution générale des poids sont connues. Cette nouvelle probabilité est combinée avec celle des réseaux statiques afin d'en former une seule. Pour le reste, **OSLOM** se déploie de la même manière en utilisant plutôt cette probabilité comme critère de distribution statistiquement significative des sommets en communautés. La performance de cet algorithme ajusté est très bonne comparativement à d'autres méthodes existantes pour les réseaux pondérés et plus spécifiquement lorsque la taille du réseau analysé est grande.

Pons et al. [72] proposent **WalkTrap**, un algorithme, basé sur le concept de marche aléatoire, qui permet de détecter les communautés dans les réseaux statiques. L'hypothèse des auteurs est qu'un marcheur qui se promène sur les arêtes d'un graphe devrait rester plus longtemps dans une communauté en raison du nombre élevé de liens qui s'y trouvent et du peu d' arêtes existantes pour en sortir. Une distance est établie entre chacun des sommets du graphe en fonction de la longueur de marche qui les sépare. Un algorithme agglomératif hiérarchique est ensuite utilisé afin de discerner les communautés en analysant la distance entre chaque sommet et en regroupant ceux qui se trouvent près.

Pour sa part, Newman propose la transformation d'un réseau pondéré en multigraphes non pondérés [23]. Ainsi, chaque arête est dupliquée selon son poids. Dans ce contexte, la détection de communautés peut se faire de la même manière que pour un graphe standard. N'importe laquelle des approches de résolution propres aux réseaux statiques, qu'elle soit adaptée aux graphes pondérés ou non, peut alors être utilisée.

> Le comportement des employés peut être étudié à un moment précis, comme c'est le cas avec les approches proprosées ci-haut. Cependant, ce qui rend la détection des communautés d'autant plus intéressante pour les gestinnaires, est l'étude de ces comportements au travers du temps par l'utilisation de réseaux évolutifs et des approches qui suivent.

3.2.2 Détection de communautés dans les réseaux évolutifs

Certains auteurs, dont Cazabet [31] et Aynaud et al. [73], ont présenté un large éventail des méthodes de détection de communautés dans les réseaux évolutifs développées avant 2013 alors que Correc [74] en a fait l'étude jusqu'en 2015. Le chapitre qui suit présente quatre grandes classes d'approches de résolution pour les graphes évolutifs. Pour chacune d'elle, les différentes approches connues à ce jour et pertinentes pour la problématique de ce mémoire sont détaillées.

L'approche initiale de détection de communautés dans les réseaux évolutifs consiste à résoudre chaque instantané indépendamment des autres et à apparier les communautés trouvées dans chacun d'eux par la suite. Ainsi, l'évolution des communautés peut être étudiée, une fois l'appariement réalisé. On parle alors d'approches indépendantes sur des instantanés successifs.

Hopcroft et al. [38] ont été parmi les premiers à proposer la détection de communautés dans les graphes évolutifs. Ces auteurs ont étudié deux instantanés successifs d'un réseau, qu'ils ont résolus séparément par un algorithme agglomératif hiérarchique. Ce dernier consiste à regrouper des nœuds ou des communautés qui sont similaires en fonction de la distance qui les sépare. L'ordre de traitement de ceux-ci influence toutefois la solution obtenue. Ainsi, pour minimiser cette instabilité, les auteurs ont développé une mesure d'appariement des communautés. Une perturbation de l'ordre de 5% du réseau est alors appliquée sur la partition obtenue d'un instantané. Les communautés de la partition initiale $\mathscr{C}$ et celles de la partition perturbée $\mathscr{C}'$ sont ensuite comparées, par la

mesure d'appariement, afin de vérifier leur ressemblance ou autrement dit, la constance de la structure modulaire malgré une perturbation. La mesure utilisée se définit ainsi

$$match(\mathscr{C},\mathscr{C}') = min\left(\frac{|\mathscr{C}\cap\mathscr{C}'|}{|\mathscr{C}|}, \frac{|\mathscr{C}\cap\mathscr{C}'|}{|\mathscr{C}'|}\right). \tag{3.20}$$

Plus la valeur obtenue est élevée, plus les communautés ont une composition et une taille semblables. Dans ce cas, elles sont qualifiées de communautés naturelles. La même mesure d'appariement est ensuite appliquée sur les communautés naturelles d'un instantané à l'autre de manière à retracer leur évolution.

L'algorithme **CPM** [35] a été ajusté par Palla et al. [39] de manière à pouvoir l'appliquer aux réseaux évolutifs. La détection des communautés se fait d'abord sur les différents instantanés du réseau à l'aide de **CPM**. Chaque instantané est regroupé avec celui de la période suivante pour former une paire, dont les nœuds et liens sont tous représentés dans un nouveau graphe. L'algorithme est alors appliqué à nouveau sur ce graphe. Cette procédure permet de faire correspondre les communautés des différentes périodes ensemble puisqu'une même communauté présente dans deux instantanés successifs en forme qu'une seule dans le réseau où ils sont regroupés. Il peut arriver que plus d'une communauté de chacun des deux instantanés se confondent dans le nouveau graphe. Elles sont alors appariées selon leur taux de chevauchement. Cette approche permet de retracer l'évolution des communautés en établissant leur présence d'un instantané à l'autre.

L'appariement des communautés ainsi que l'étude de leur évolution nécessitent un post-traitement lorsque l'approche de résolution utilisée est indépendante. Certains travaux ont donc été réalisés en ce sens.

Spiliopoulou et al. [8] proposent **MONIC**, un cadre méthodologique permettant de suivre l'évolution de clusters de données à travers le temps. Ce cadre permet de retracer les évènements dits externes d'un cluster, concernant sa relation avec les autres clusters, ainsi que les évènements internes, reliés à sa composition et sa forme. Les évènements externes sont les plus pertinents pour la suite de ce travail et sont présentés dans le tableau 3.1.

Greene et al. [32] présentent une autre stratégie pour comprendre l'évolution des communautés dynamiques. Leur approche consiste d'abord à résoudre séparément les ins-

Évènement	Notation	Indicateur
Le cluster naît	$\emptyset \rightarrow C_i^{(t+1)}$	$\forall j, C_i^{(t+1)} \neq match(C_j^{(t)})$
Le cluster survit	$C_i^{(t)} \rightarrow C_i^{(t+1)}$	$C_i^{(t+1)} = match(C_i^{(t)}) \wedge \forall C_j^{(t)} \neq C_i^{(t)}, C_i^{(t+1)} \neq C_j^{(t)}$
Le cluster est fusionné	$C_i^{(t)} \subseteq C_i^{(t+1)}$	$C_i^{(t+1)} = match(C_i^{(t)}) \wedge \exists C_j^{(t)} \neq C_i^{(t)}, Ci^{(t+1)} = match(C_j^{(t)})$
Le cluster est divisé	$C_i^{(t)} \rightarrow \{C_{i1}^{(t+1)}, ..., C_{ip}^{(t+1)}\}$	$\forall k \in 1..p, C_{ik}^{(t+1)} \cap C_i^{(t)}$ est assez large et $\cup_{k=1}^{p} C_{ik}^{(t+1)} \cap C_i^{(t)}$ est assez large
Le cluster disparaît	$C_i^{(t)} \rightarrow \emptyset$	Dans tous les autres cas

TABLEAU 3.1 – Évènements externes possibles pour les clusters, comme présenté par Spiliopoulou [8], où *match*() est une fonction d'appariement et signifie que les compositions de deux communautés sont identiques.

tantanés du réseau à l'aide d'un algorithme adapté aux graphes statiques. Ces instantanés sont ensuite comparés chronologiquement de manière à reconnaitre une même communauté d'une période à l'autre, en utilisant celles trouvées précédemment comme références. L'appariement de communautés à travers le temps forme une communauté dynamique, comme il a été défini au début de cette section. La plus récente observation d'une telle communauté est qualifiée de *fronts*. Les communautés trouvées dans l'instantané suivant (C_{ti}) sont comparées à ces *fronts* (F_j) afin de les joindre à la même communauté dynamique si elles sont suffisamment similaires. L'indice de Jaccard présenté précédemment est utilisé à cet effet et se présente comme suit

$$J(C_{ti}, F_j) = \frac{|C_{ti} \cap F_j|}{|C_{ti} \cup F_j|}. \tag{3.21}$$

Si la similarité entre les deux communautés de la paire excède un certain seuil prédéfini entre 0 et 1, elles sont couplées et assignées à la même communauté dynamique. Les *fronts* sont mis à jour après l'analyse de chaque instantané, et ce, jusqu'au dernier. La solution obtenue permet de tracer l'évolution des différentes communautés dynamiques

du réseau évolutif.

Takaffoli et al. [75] proposent aussi un cadre pour retracer l'évolution des communautés dynamiques. Un algorithme d'appariement est d'abord appliqué pour faire correspondre les communautés trouvées dans chacun des instantanés. Celles qui sont semblables sont alors regroupées pour former une *méta communauté*. Des évènements sont ensuite identifiés pour expliquer les différences entre les communautés d'une même *méta communauté*.

Chen et al. [76] suggèrent une approche de résolution qui se concentre plutôt sur des représentants du graphe et des communautés. Un nœud qui se trouve dans un instantané, mais aussi dans les instantanés des périodes précédente et suivante, est considéré comme un représentant du réseau. Dans ce contexte, les auteurs s'intéressent uniquement aux communautés de ces sommets. Ces dernières sont définies comme les cliques maximales du réseau. Un nœud représentatif d'une communauté est celui qui appartient au plus petit nombre d'autres communautés. Contrairement aux autres approches, celle-ci s'intéresse à la dynamique des communautés d'une période à l'autre plutôt qu'à leur stabilité. Ainsi, il est possible de retracer évolution des communautés en se concentrant sur les sommets représentant ces dernières.

Wang et al. [77] suggèrent **CommTracker**, qui se réfère également à des nœuds centraux. Ceux-ci sont toutefois déterminés en fonction de leur degré, ou de leur force dans le cas d'arêtes pondérées. Ces sommets sont utilisés en raison de leur stabilité dans le temps, ce qui permet une résolution plus précise et efficace. L'évolution des communautés est représentée par le comportement de ces nœuds d'un instantané à l'autre. Par exemple, deux sommets centraux, dans des communautés distinctes initialement, qui se retrouvent dans la même communauté à la période suivante, signifient une fusion des communautés initiales. L'enjeu de cette approche réside dans la représentativité adéquate du réseau original par les nœuds centraux.

Les approches informées sur des instantanés successifs forment la deuxième classe d'approches pour la résolution de réseaux évolutifs. Elles traitent les instantanés séparément également. Cependant, la partition trouvée pour un instantané donné est considérée lors de la résolution du suivant, dans le but de faciliter l'appariement des communautés. Ainsi, lorsque ces dernières sont plus difficiles à identifier, la référence au passé peut s'avérer utile.

L'algorithme de **Louvain** a été adapté par Aynaud et al. [78] au contexte temporel. Les auteurs l'initialisent à partir de la partition trouvée dans l'instantané précédent, mais dont certains nœuds ont été retirés. Cette initialisation, qui remplace celle où chaque nœud forme sa propre communauté, constitue le seul changement dans l'application de l'algorithme pour les graphes évolutifs.

Wang et al. [79] s'appuient sur le principe des *fronts* de Greene et al. [32] ainsi que des nœuds centraux de Wang et al. [77]. Les auteurs détectent les communautés du réseau à chaque instantané à partir de l'algorithme de **Louvain**. À chaque période, l'algorithme est initialisé à partir des nœuds centraux trouvés à la période précédente. Le processus d'appariement des communautés, d'un instantané à l'autre, se fait alors à partir de ces sommets, qui forment les *fronts* des communautés dynamiques.

DiDiC *(Distributed Diffusive Clustering)* est présenté par Gehweiler et al. [80]. Il s'agit d'une approche basée sur la distribution diffuse et qui optimise des mesures de qualité des partitions, dont la modularité. Initialement, chacun des nœuds est distribué aléatoirement dans une communauté, dont le nombre doit être connu. Aux périodes suivantes, l'algorithme est initialisé à partir de la partition trouvée à l'instantané précédent.

Takafolli et al. [81] proposent un algorithme basé sur l'optimisation de la mesure *L-metric*, qui s'applique localement sur un réseau pour ainsi accélérer la résolution. Cette mesure suppose qu'une communauté a peu de liens entre les sommets à sa frontière et la partie non considérée du réseau alors qu'elle en a davantage avec les communautés près d'elle. Pour chaque instantané étudié, les composantes connexes des communautés sont extraites et servent de point de départ pour la résolution de l'instantané suivant.

Kim et al. [82] suggèrent une approche multiobjectif pour la détection de communautés dynamiques. La résolution vise à maximiser à la fois le nombre de sommets regroupés en communautés et le nombre de sommets non affiliés. Plusieurs approches sont étudiées, dont des approches hybrides, afin d'obtenir des solutions Pareto optimales. Une mesure de similitude des éléments à l'intérieur d'une même communauté est utilisée pour évaluer et comparer la qualité des partitions trouvées. La résolution d'un instantané se base sur la solution obtenue dans celui qui le précède.

OSLOM est également applicable aux graphes évolutifs [71]. L'algorithme est alors déployé normalement sur le premier instantané. Comme il permet la résolution à partir d'une partition initiale, la partition trouvée à la première période sert de condition initiale pour la résolution de l'instantané suivant. Le même principe s'applique ainsi d'un

instantané à l'autre. L'évolution des différentes communautés peut alors facilement être retracée. Si deux instantanés successifs sont très différents en raison d'un évènement majeur survenu entre les périodes correspondantes, l'algorithme s'applique bien malgré tout et les deux partitions trouvées sont simplement non corrélées.

Wang et al. [83] présentent **LWEP** *(Local Weighted-Edge-based Pattern)* conçu pour les graphes évolutifs pondérés et localement homogènes. Par localement homogène, les auteurs insinuent des relations similaires entre des sommets dans une même région du graphe. L'application de **LWEP** enregistre des statistiques pour chaque nœud du réseau, en fonction des relations passées et présentes, et résout chacun des instantanés à partir de ces statistiques. Un appariement des communautés est réalisé par la suite grâce à l'indice de Jaccard.

NEO-CDD *(Nash Extremal Optimization for the Dynamic Community Detection problem)* constitue une approche de Lung et al [84], basée sur la théorie des jeux. Dans ce contexte, chaque nœud représente un joueur qui cherche à maximiser son profit en se joignant à la communauté de son choix. Le profit est estimé par la différence entre la qualité de la communauté avec et sans sa présence. Un algorithme d'optimisation extrémale est appliqué de manière à atteindre l'équilibre de Nash où tous les joueurs ne peuvent améliorer leur profit davantage. L'algorithme conserve la meilleure partition trouvée en mémoire et travaille sur un autre graphe pour la résolution. Quand un changement survient dans le réseau, c'est-à-dire que le profit d'un sommet est modifié, la partition avant ce changement est conservée en mémoire alors que le graphe mis à jour est réinitialisé puis résout.

Dinh et al. [85] suggèrent **MIEN** *(Modules Identification in Evolving Networks)*. Cette approche vise à accélérer la résolution en assurant le suivi des communautés par une représentation compacte des partitions à chaque période. Dans cette représentation, les communautés sont remplacées par des nœuds. Les liens entre ceux-ci sont pondérés en fonction du poids total des liens inter-communautés correspondants dans la partition originale. Les nouveaux sommets à chaque période sont ajoutés à cette représentation, sans être affiliés initialement. Un algorithme adapté aux graphes statiques pondérés est ensuite appliqué sur la représentation compacte pour chaque instantané à l'étude.

Riedy et al. [86] ont développé un algorithme agglomératif parallèle. Ce dernier vise à traiter les larges graphes, en mettant à jour les communautés d'une période à l'autre plutôt qu'en résolvant le graphe en entier sur chaque instantané. Une représentation sim-

plifiée du graphe est utilisée et développée de la même manière qu'avec **MIEN**. L'algorithme commence par retirer les nœuds touchés par des modifications au temps t de la partition obtenue au temps $t-1$. Il ne sera appliqué que sur ceux-ci. Les trois étapes principales sont le calcul de la modularité des sommets, leur couplage pour maximiser la modularité du graphe ainsi que la contraction, qui regroupe les nœuds couplés ensemble. Elles sont effectuées en boucle jusqu'à ce qu'il soit impossible d'améliorer la modularité davantage.

A^{3}CS *(AdAptive Algorithm for Community Structure in dynamic networks)* constitue une autre avancée de Dinh et al. [87]. Ce cadre méthodologique garantit la rapidité de résolution, par la maximisation de la modularité ainsi que l'exploitation de la propriété de la distribution des degrés par la loi de puissance. Pour accélérer la détection de communautés, les nœuds du graphe se voient identifiés par une étiquette, soit de leader, de suiveur ou d'indépendant. La formation des communautés se fait selon deux principes : tous les suiveurs d'un leader sont dans la même communauté que ce dernier et tous les indépendants sont non-affiliés. La résolution d'un instantané s'effectue à partir de la partition trouvée à la période précédente mise à jour avec les changements survenus depuis.

Afin de diminuer l'instabilité de certains algorithmes, Lancichinetti et Fortunato [88] ont développé une matrice de consensus qui compare les partitions de plusieurs itérations d'un même algorithme. Son but est d'obtenir une partition moyenne, c'est-à-dire la plus similaire à l'ensemble de celles trouvées. Chaque arête est étudiée et une valeur est déterminée en calculant son nombre d'apparitions dans une même communauté divisé par le nombre de partitions comparées. Si le résultat obtenu est inférieur à un certain seuil prédéterminé, l'arête est retirée, à moins que le graphe ne soit plus connexe après cette opération. La matrice est alors mise à jour, ainsi de suite jusqu'à ce qu'une partition unique soit atteinte et ne puisse être modifiée davantage. L'algorithme de résolution utilisé peut varier, pourvu qu'il soit adéquat pour les graphes pondérés. Des sous-ensembles d'instantanés consécutifs sont créés et la matrice de consensus leur est appliquée. L'indice de Jaccard est utilisé, par la suite, pour apparier les communautés.

Chakrabarti et al. [46] ont développé, non pas une méthode de résolution, mais plutôt un cadre évolutif. Ce dernier cherche à assurer la cohérence des communautés détectées à chaque période, en considérant à la fois la qualité statique et séquentielle des communautés, comme présentées précédemment. Les auteurs font donc plutôt référence au concept d'*Evolutionary clustering*. Ils ont expérimenté leur cadre évolutif avec la

détection de communautés dynamiques en utilisant deux approches différentes, soit l'algorithme des k-moyennes et un algorithme agglomératif hiérarchique.

Chi et al. [89] ont adapté légèrement ce cadre évolutif. En effet, ils ont ajusté la qualité séquentielle de deux manières différentes. Dans un cas, elle mesure la cohérence entre la partition courante et les données précédentes afin de conserver la qualité des communautés. Dans l'autre cas, elle mesure la ressemblance entre la partition courante et la précédente afin de préserver l'appartenance des sommets aux communautés. Un algorithme de classification spectrale est utilisé avec cette approche. Il vise à optimiser la qualité de la structure modulaire, en considérant à la fois, la qualité statique et la qualité séquentielle, selon l'une des deux propositions précédentes.

Lin et al. [90][91] utilisent un cadre similaire à celui de Chakrabarti et al. [46]. Ils proposent **FacetNet** qui, par un modèle génératif probabiliste, permet d'étudier à la fois la composition des communautés dans le réseau à une certaine période et leur évolution. Pour ce faire, cette approche maximise localement la cohérence de la solution avec les données à la base du réseau ainsi que l'historique concernant la composition des communautés aux périodes précédentes. **FacetNet** a l'avantage de détecter des communautés avec recouvrement. Cependant, plusieurs contraintes restreignent son utilisation, dont notamment la nécessité de connaitre le nombre de communautés dans le réseau initialement. Cette approche est donc moins pertinente pour l'étude des réseaux réels où cette information est inconnue. De plus, elle ne permet pas la naissance ni la mort des communautés au fil du temps. **FacetNet** s'adapte également difficilement aux réseaux de grandes tailles, en raison du fort volume d'itérations nécessaires à l'algorithme pour converger.

Kim et al. [92] proposent une méthode de densité et particule de détection **PDEC** *(Particle-and-Density based Evolutionary Clustering)*. Celle-ci permet de remédier aux limites de **FacetNet** en permettant la détection d'un nombre variable de communautés, qui peuvent naitre ou mourir à tout moment au cours de l'horizon temporel analysé. Le réseau est alors considéré comme un ensemble de particules, chacune contenant de l'information sur l'évolution des données ou de la structure modulaire. Dans ce contexte, les communautés représentent un sous-ensemble de particules densément connectées les unes aux autres. La méthode de classification présentée s'appuie sur le principe de densité, lui-même basé sur **DBSCAN** *(Density-Based Spatial Clustering of Applications with Noise)* [93]. Elle permet la détection de communautés de qualité, localement, en utilisant un cadre évolutif semblable à celui de Chakrabarti et al. [46], en plus

du critère de modularité. Un lissage temporel est ensuite réalisé sur les communautés trouvées à chaque période de manière à les faire correspondre et à pouvoir identifier différents évènements marquant leur évolution.

Takafolli et al. [94] se servent de **MODEC** *(MOdeling and Detecting Evolutions of Communities)*. Il s'agit d'un cadre évolutif qui se concentre plutôt sur les évènements qui caractérisent la modification des communautés à travers le temps, afin d'expliquer l'évolution de celles-ci.

Un cadre évolutif est suggéré par Xu et al. [95] et considère à la fois les données passées et présentes pour la détection de communautés dynamiques. Le poids assigné à la considération de chaque type de données est optimisé selon le critère de l'erreur moyenne au carré. L'algorithme spectral normalisé de Yu et Shi [96] est celui utilisé pour détecter les communautés, dont le nombre doit être spécifié initialement.

Xu et al. [97] présentent également **AFFECT** *(Adaptive Forgetting Factor for Evolutionary Clustering and Tracking)*. Cette approche permet l'application de leur cadre évolutif initial [95] aux algorithmes hiérarchiques et des k-moyennes, en plus de ceux spectraux comme il a déjà été présenté.

DYN-MOGA *(DYNamic MultiObjective Genetic Algorithms)* proposé par Folino et al. [98] aborde le problème de détection de communautés dynamiques comme un problème d'optimisation multiobjectif. D'un côté, il faut maximiser la qualité des instantanés, quant à la bonne représentation des données dans les communautés. De l'autre, il faut minimiser la différence entre la composition d'une même communauté d'une période à l'autre. Le premier objectif utilise le concept de pointage d'une communauté, alors que le deuxième se réfère à l'information mutuelle normalisée. L'optimisation simultanée de ces objectifs se fait par un algorithme génétique. Le même problème peut également être résolu par **DYN-LSNNIA** *(DYNamic Local Search Nondominated Neighbor Immune Algorithm)*, de Gong et al. [99], qui combine un algorithme immunitaire et une recherche locale en permettant des résultats plus précis.

La troisième classe d'approches regroupe les approches incrémentales sur des modifications successives. Celles-ci visent à mettre à jour les partitions trouvées lorsqu'un changement survient dans la structure du graphe, notamment l'ajout ou le retrait de sommets ou d'arêtes. Elles n'utilisent donc pas les instantanés comme les approches

précédentes.

Inspirés par le concept de coupe minimale d'un arbre développé et utilisé par Gomory et Hu [100] ainsi que Flake et al. [101], Görke et al. [5] suggèrent de considérer les modifications au graphe progressivement, comme présenté à la figure 3.2. De cette manière, la résolution du graphe s'effectue de façon continue et permet d'assurer une continuité dans les communautés trouvées.

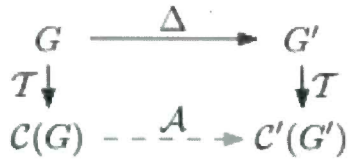


FIGURE 3.2 – Mise à jour des communautés $C(G)$ par l'approche $\mathscr{A}$ à chaque modification Δ comparativement à la résolution complète du graphe modifié G' à partir d'instantanés $\mathscr{T}$ [5][6].

Görke et al. [6] ont développé, selon le même principe de mise à jour progressive des communautés, **dGlobal** et **dLocal**, soit l'application d'algorithmes gloutons localement pour maximiser la modularité du graphe lors des mises à jour. **TDLocal** constitue une autre version de **dLocal** étudiée par les auteurs [102] et considère à la fois la maximisation de la modularité et la minimisation de l'indice de Rand présenté plus tôt. Un algorithme adapté aux graphes statiques peut être appliqué sur la fonction optimisant ces deux critères. Cette approche établit un compromis entre la qualité séquentielle du cadre de Chakrabarti et al. [46] et la qualité statique des communautés trouvées.

Basée sur l'algorithme **CNM**, Bansal et al. [103] utilisent une approche agglomérative hiérarchique où deux communautés sont fusionnées si cela permet un gain de modularité. Chaque modification du graphe est considérée individuellement, c'est-à-dire chaque ajout ou retrait d'arête. Lors d'un tel évènement, il suffit de remonter le dendrogramme de la résolution effectuée plus tôt jusqu'à atteindre les sommets concernés, puis d'appliquer l'algorithme agglomératif à partir de ce point. Cette approche est réputée être significativement plus rapide que la résolution complète du réseau sur plusieurs périodes.

Nguyen et al. [104] suggèrent **QCA** *(Quick Community Adaptation)* permettant de détecter les communautés dans les réseaux évolutifs, tout en retraçant leur évolution. Dans l'optique de diminuer le temps de résolution, cet algorithme adaptatif, basé sur le critère

de modularité, considère aussi chaque changement apporté dans le graphe, individuellement. Initialement, la partition optimale est trouvée par l'algorithme de **Louvain**. Quatre évènements peuvent survenir dans le réseau soit : l'ajout d'un nœud (sans arête ou avec une ou plusieurs arêtes le reliant), le retrait d'un nœud et ses liens respectifs, l'ajout d'une arête entre deux sommets existants ainsi que le retrait d'une arête. Pour chacun de ces évènements, **QCA** recherche localement la manière optimale de modifier le réseau et ses communautés pour en tenir compte de manière à maximiser la modularité. Cet algorithme a permis l'obtention de structure modulaire de grande qualité, et ce, dans un temps raisonnable. Une adaptation a aussi été développée pour la détection de communautés avec recouvrement. Malgré tout, plus l'horizon temporel analysé est grand, plus l'algorithme a tendance à former plusieurs communautés de petite taille en voulant maximiser la modularité localement. Le nombre de communautés est alors grandissant.

La maximisation de la modularité a aussi été adoptée par Shang et al. [105]. Ces auteurs utilisent l'algorithme de **Louvain** pour la résolution du graphe initial puis mettent à jour la partition à chaque changement d'arêtes, en fonction de son type, de manière à maximiser la modularité ou du moins, à la réduire le moins possible. Ces modifications peuvent concerner des arêtes de quatre types soit : intra-communauté, inter-communautés, mi-nouvelles, c'est-à-dire dont l'une des extrémités vient d'être ajoutée, et nouvelles, soit lorsque les deux extrémités viennent de s'insérer.

Cazabet et al. [106][107][31] suggèrent, quant à eux, l'algorithme **iLCD** *(intrinsic Longitudinal Community Detection)*, permettant la détection de communautés avec recouvrement. Essentiellement, lorsqu'une arête inter-communautés est ajoutée (u,v), l'algorithme vérifie si le nœud v devrait être ajouté à la communauté du sommet u. Lorsqu'il s'agit du retrait d'une arête intra-communauté (u,v), l'algorithme évalue si des nœuds devraient quitter la communauté ou même si celle-ci devrait se diviser. Après chaque modification, la fusion de communautés se chevauchant est étudiée. Les décisions à chaque niveau se prennent en fonction de certaines métriques, à savoir la **représentativité** d'un nœud u par rapport à sa communauté C_i,

$$Rep(u,C_i) = \frac{d^{int}_{C_i}(u)}{d(u)}; \tag{3.22}$$

la **cohésion intrinsèque** de la communauté C_i, en guise de mesure de sa qualité,

$$CI(C_i) = \sum_{u \in C_i} Rep(u, C_i); \tag{3.23}$$

et la **force d'appartenance** de u à la communauté C_i

$$FA(u, C_i) = \sum_{v \in N(u)} Rep(v, C_i). \tag{3.24}$$

Xie et al. [108] ont adapté l'algorithme de propagation d'étiquettes **LabelRank** [64] pour la détection de communautés dans les réseaux évolutifs. **LabelRankT**, l'algorithme présenté par les auteurs, procède presqueidentiquement à la version originale. La seule différence concerne la mise à jour des sommets. Comme les vecteurs respectifs à chaque nœud comportent des informations sur la structure locale du réseau, ils permettent la transmission et la considération de ces informations d'un instantané à l'autre. Dans ce contexte, seuls les sommets qui ont subi une modification entre deux instantanés successifs voient leur vecteur mis à jour, suivi de leur étiquette, selon les quatre mêmes étapes que dans les graphes statiques. Parmi ces modifications possibles, les auteurs considèrent l'ajout ou le retrait d'un lien reliant un nœud existant ainsi que l'ajout ou le retrait d'un sommet du réseau.

L'algorithme **CPM** a été généralisé aux graphes évolutifs par Duan et al. [109]. L'ajout et le retrait d'arêtes sont alors considérés comme un flux constant de changement. Dans ce contexte, toutes les cliques maximales du graphe initial, de taille minimale k, sont identifiées. Elles sont chacune représentées par un sommet dans un graphe de cliques H et reliées entres elles si elles ont au moins $k-1$ nœuds en commun. Une recherche en profondeur identifie les composantes connexes de H de manière à obtenir une forêt, dont les sommets reliés à un même arbre forment une communauté. Le graphe de cliques H est mis à jour à chaque ajout ou retrait d'arêtes de manière à obtenir une rapidité de résolution.

Falkowski [110] présente **DENGRAPH** *(DENsity based GRAPH clustering algorithm)*, une approche de détection de communautés dans les réseaux évolutifs basée sur l'algorithme **DBSCAN** [93]. **DENGRAPH** cherche à identifier des sous-ensembles denses de nœuds inclus dans un certain rayon ε, de taille minimale η. Les sommets du graphe

sont analysés aléatoirement de manière à classer le sous-ensemble auquel ils appartiennent dans une communauté. Cette classification se fait en fonction de la densité caractérisant les liens entre le sous-ensemble du nœud étudié et les communautés existantes. Si le sommet analysé n'appartient à aucun sous-ensemble respectant les paramètres établis, il est laissé seul. Une mesure de distance caractérise également les liens entre chacun des nœuds du réseau. À chaque classification, les distances sont mises à jour et des changements dans la composition des communautés peuvent en découler. Par exemple, des sommets peuvent ne plus respecter le rayon de leur sous-ensemble initial ou à l'inverse s'ajouter dans un autre sous-ensemble. L'évolution des communautés s'étudie donc à partir de ces changements. Une variante a également été développée pour détecter les communautés avec recouvrement.

Ning et al. [111] proposent de trouver une partition initiale par un algorithme spectral, puis de mettre à jour le système de valeurs propres et vecteurs propres à chaque modification du réseau. Cette approche permet de suivre exactement l'évolution des données.

Les approches qui suivent cherchent plutôt à étudier simultanément tous les instantanés du graphe de manière à faire correspondre les communautés à travers le temps. Aucun appariement des communautés de chaque période n'est donc nécessaire une fois la partition trouvée. Il est alors question d'approches simultanées sur tous les instantanés et c'est celles-ci qui forment la dernière classe d'approches de résolution pour les réseaux évolutifs.

Aynaud et al. [112] utilisent une version modifiée de l'algorithme de **Louvain**, qu'ils appliquent sur tous les instantanés simultanément ou un sous-ensemble de ceux-ci. De cette manière, ils cherchent à optimiser la modularité moyenne sur tout l'horizon temporel plutôt qu'à un instant précis. Ainsi, la modularité du graphe s'obtient par la somme de celles à chaque période. Cette approche vise à obtenir des communautés cohérentes sur l'intervalle de temps analysé. Toutefois, les expérimentations des auteurs ont démontré que l'ordre des instantanés n'est pas respecté lors de la résolution, ce qui empêche de bien comprendre les liens de causalité. Également, les sommets d'un même individu ne sont pas différenciés d'un instantané à l'autre, ce qui engendre parfois le regroupement de deux membres dans la même communauté, alors qu'ils n'ont jamais été présents en même temps.

Une autre alternative de résolution est proposée par Gauvin et al. [113], qui cherchent simultanément à identifier les communautés dans un graphe évolutif en plus de retracer

leur évolution. Cette approche s'appuie sur des techniques de factorisation non négatives, sachant qu'elles permettent la détection de communautés avec recouvrement dans les graphes statiques. Une fois la résolution effectuée simultanément sur tous les instantanés, trois facteurs sont obtenus, deux représentant la structure modulaire du réseau, qui sont égaux lorsque le réseau est non dirigé, et un décrivant l'évolution de chaque communauté. Bien que cette approche permette les communautés avec recouvrement, seules les tendances globales, et non les temporaires, peuvent être dégagées de la partition pour expliquer l'évolution des communautés.

Mucha et al. [7] ajustent la représentation du réseau par les instantanés avant de résoudre ceux-ci. Pour ce faire, chaque sommet est relié avec lui-même, d'un instantané à l'autre, successifs ou non, comme présenté à la figure 3.3. L'algorithme de **Louvain** est alors appliqué sur ce modèle longitudinal par l'optimisation du critère de modularité ajusté aux graphes évolutifs. Les auteurs ont remarqué que cette approche permet de comprendre la composition et la dynamique de réseaux réels qu'il aurait été difficile, voire impossible, de cerner avec une utilisation indépendante des instantanés. Les résultats de leurs tests ont notamment permis de distinguer des évènements marquants ayant influencé la structure modulaire du réseau étudié, en plus de déceler certaines tendances quant au comportement des individus le composant.

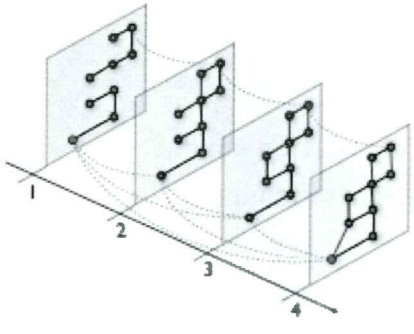


FIGURE 3.3 – Ajout d'arêtes reliant un même sommet d'un instantané à l'autre [7].

De manière semblable, Ben Jdidia et al. [114] suggèrent la considération simultanée de tous les nœuds du réseau évolutif. Dans ce cas, les sommets sont reliés avec eux-mêmes d'un instantané à l'autre, consécutif ou non, mais aussi, avec des nœuds des périodes consécutives avec qui ils partagent au moins un voisin commun. Une version ajustée de l'algorithme **WalkTrap** pour les graphes évolutifs est utilisée sur l'ensemble des instantanés pour la résolution. Les auteurs affirment que cette approche permet de

détecter des communautés réalistes ainsi que de suivre leur évolution. Ils ajoutent que les résultats obtenus peuvent expliquer d'autres phénomènes. Ils reprochent toutefois à leur modèle de ne pas être en mesure de détecter les communautés imbriquées les unes dans les autres ni d'étudier l'influence des membres au sein de leur groupe.

Tantipathananandh et al. [115] proposent un cadre permettant également de détecter les communautés à travers le temps et non pas sur chacun des instantanés séparément. Chaque nœud est relié avec lui-même d'un instantané à l'autre, consécutif ou non, et avec les sommets représentant le groupe dans lequel il se trouve à chaque période. Ces groupes, en guise de communautés, comprennent un ensemble de nœuds connectés les uns aux autres, mais aucunement reliés aux sommets des autres groupes. Ces groupes sont représentés par des nœuds dans les instantanés. Selon les auteurs, un sommet ne devrait pas changer d'affiliation souvent et devrait interagir avec sa communauté initiale la plupart du temps. Ainsi, ils ont développé une fonction de coûts composée de trois éléments relatifs aux nœuds, soit le coût de changer de communauté, celui de ne pas se trouver dans sa communauté d'origine à une certaine période ainsi que celui représentant le nombre de communautés différentes jointes au cours de l'horizon temporel étudié. Cette fonction de coût nécessite toutefois la connaissance des partitions à chaque période. Elle est ensuite optimisée par une heuristique sur l'ensemble des instantanés. La partition obtenue n'est donc pas nécessairement optimale. Malgré tout, cette approche permet de mieux comprendre le mouvement des individus au fil du temps en plus d'identifier les communautés et leur évolution. Tantipathananandh et al. [116] ont développé, quelques années plus tard, un ajustement à leur approche en permettant la détection de communautés dynamiques sans la connaissance au préalable des partitions à chaque période. Toutefois, cette variante reste encore non testée sur de larges graphes s'étalant sur un long horizon temporel.

Mitra et al. [117] introduisent un nouveau modèle de graphe où il n'existe pas d'instantanés et où toutes les interactions réalisées sur plusieurs périodes se trouvent dans le même modèle. Dans ce cas, un même sommet est dupliqué dans le graphe en fonction de son apparition à chacune des périodes analysées. Ce modèle est adapté aux réseaux où les relations entre les individus représentent des réponses à des évènements antérieurs. Ainsi, un lien peut exister entre un sommet d'une certaine période et un nœud de toute autre période selon le moment où la réponse a eu lieu. Cette approche permet la détection de communautés dont la composition peut changer entièrement à travers le temps. Les auteurs utilisent l'algorithme de **Louvain** pour la résolution de leur graphe

longitudinal. Cette approche permet l'étude de l'évolution complète du réseau en représentant fidèlement l'information temporelle originale. De plus, des évènements précis quant au comportement des individus et des communautés peuvent être retracés. Cependant, la densité du graphe influence la partition trouvée et un graphe peu dense entraine la formation d'un ensemble de petites communautés peu significatives.

Yang et al. [118] présentent le **DSBM** *(Dynamic Stochastic Block Model)*, soit une variante du *Stochastic Block Model* [119] dans le but de l'adapter aux graphes évolutifs. Avec cette approche, les sommets sont également dupliqués selon leur apparition dans le temps. Le cadre proposé, basé sur l'inférence bayésienne, est un modèle probabiliste unifié visant à identifier des communautés dynamiques et leur évolution. Il considère à la fois les observations passées et futures prévues lors de la résolution. Les expérimentations effectuées par les auteurs ont permis de distinguer avec précision la structure modulaire du graphe ainsi que son évolution, en plus de cerner des phénomènes sous-jacents aux données. Cependant, cette approche est adéquate seulement pour les réseaux de petite taille ainsi que plutôt denses et ne permet pas la fusion ou la division de communautés.

Chapitre 4

Problématique

Comme en témoigne la revue de littérature présentée au chapitre précédent, il existe une multitude d'approches et d'algorithmes de résolution pour la détection de communautés dans les réseaux. Ce mémoire se concentrera toutefois davantage sur une seule approche, qui sera adaptée au contexte de la base de données réelles qui nous a été fournie et pouvant être représentée par un réseau évolutif. La détection des communautés dans un tel réseau permettra à la fois d'évaluer l'approche de résolution utilisée, en plus de comprendre la dynamique qui existe entre les individus étudiés et les groupes qu'ils forment. Ce chapitre pose les bases au travail qui sera effectué dans ce mémoire.

La première section décrit la base de données qui sera analysée alors que la seconde aborde la question de recherche, liée à ces données, qui est adressée dans ce mémoire.

4.1 Description du jeu de données

4.1.1 Contenu du jeu de données

Le réseau qui sera étudié provient d'une base de données réelles qui nous a été fournie par une entreprise dans le domaine des technologies de l'information. Les données représentent des discussions entre des employés sur un forum au sujet de mises à jour possibles pour un logiciel. Chacune des 1 607 632 données recueillies a cinq composantes soit : *individu, mise à jour, rôle, bureau*, et *date*.

Plus précisément, voici le détail de chacune de ces composantes :

— *individu* : série de caractères qui identifie anonymement l'individu parmi les 1045 employés qui ont participé à au moins une discussion ;
— *mise à jour* : code numérique qui identifie la mise à jour discutée parmi les 324 860 qui ont été proposées et rassemblées ;
— *rôle* : rôle joué par l'employé dans une discussion parmi sept possibles selon son moment d'apparition et son pouvoir de décision. On y retrouve notamment l'initiateur qui propose une mise à jour, l'intervenant qui la commente, le réviseur qui la critique ainsi que l'approbateur qui permet sa mise en œuvre. Les sept rôles ne sont pas tous joués dans chaque mise à jour ;
— *bureau* : série de caractères pour identifier auquel des 46 bureaux participants est rattaché l'employé à la fin de la collecte des données ;
— *date* : date à laquelle le commentaire a été écrit par l'employé au cours des cinq ans et demi ou plutôt 1975 jours pendant lesquels les entrées ont été enregistrées.

Aucune information additionnelle à celles-ci ne nous a été transmise par l'entreprise, que ce soit par rapport au contexte de cette dernière, à celui du forum ou encore, quant aux comportements des employés. Ainsi, la résolution qui sera effectuée dans ce mémoire se basera uniquement sur les données obtenues sans égard à la situation réelle de la compagnie.

C'est seulement avec ces informations que l'approche proposée dans ce travail cherchera à déceler le comportement des employés au fil du temps. Ainsi, la résolution s'effectuera pratiquement à l'aveugle et visera à faire ressortir des tendances et phénomènes révélateurs, pour les gestinnaires, concernant la dynamique de leur équipe.

4.1.2 Analyse descriptive du jeu de données

Avant toute chose, il est pertinent de constater les principales caractéristiques et tendances des données fournies. Les analyses qui suivent font parfois référence à un horizon temporel. Il s'agit en fait de l'ensemble des 1975 jours pendant lesquels les commentaires des employés ont été relevés. Aux fins de l'analyse et des étapes qui suivent,

cette durée a été divisée en semestres. Les raisons de ce choix seront expliquées à la section 5.3.

Certaines analyses préliminaires ont permis de remarquer que le nombre d'employés impliqués dans les mises à jour est croissant sur l'horizon temporel tout comme le nombre de mises à jour discutées à chaque période. Cependant, lors du dernier semestre, une diminution de ces deux éléments est notée. Les figures 4.1 et 4.2 illustrent ces tendances. On y remarque notamment que le nombre d'employés semble croitre de manière constante comparativement au nombre de mises à jour qui lui, augmente de façon plus irrégulière. Effectivement, il s'élève légèrement entre certaines périodes, comme entre le quatrième et le cinquième semestre, alors qu'il croît significativement de la neuvième à la dixième période.

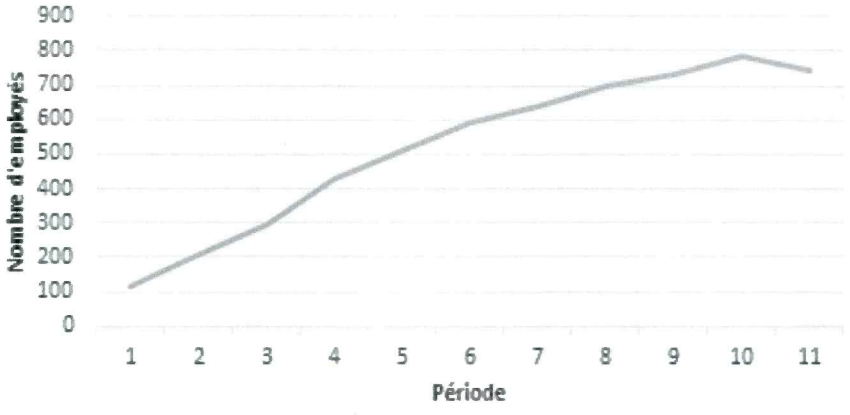


FIGURE 4.1 – Évolution du nombre d'employés impliqués dans les mises à jour au cours de l'horizon temporel étudié.

Ces deux évolutions sont à leur minimum initialement et atteignent leur maximum à la neuvième période. Le dernier semestre ne comporte que cinq mois de données plutôt que six comme les autres, étant donné les entrées qui nous ont été fournies et qui ne permettent pas de le compléter. Il a donc fallu évaluer si la diminution de la participation des employés et des mises à jour à cette période s'expliquait par ce manque de données. Une règle de proportionnalité a été établie selon le nombre d'individus et de mises à jour au cours des cinq premiers mois du semestre, pour estimer l'ampleur de leur présence au sixième mois. Avec cet ajustement, le nombre d'employés participant aux mises à jour lors de la dernière période s'élève à 893, ce qui est supérieur au niveau de participation des périodes précédentes, en guise d'une continuité de la hausse observée

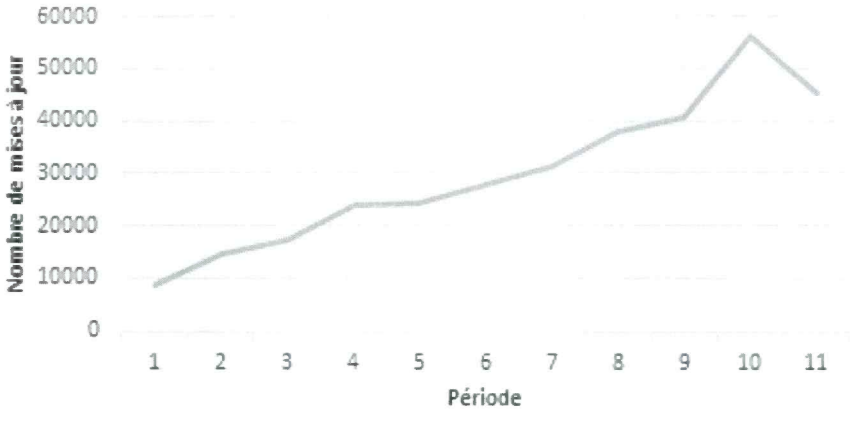


FIGURE 4.2 – Évolution du nombre de mises à jour discutées au cours de l'horizon temporel étudié.

plus le temps avance. Quant aux mises à jour, même avec un ajustement, elles sont présentes en moins grand nombre qu'au semestre précédent. Cette situation reflèterait donc une baisse réelle des mises à jour discutées lors de la dernière période. Malgré tout, aucune conclusion certaine ne peut être tirée sur cet aspect en raison de l'absence des données réelles pour ce dernier mois.

La figure 4.3 représente le nombre d'employés travaillant à chacun des 46 bureaux participant au forum. Ces derniers sont numérotés sur le graphique, car leur nom ne nous a pas été divulgué par la base de données. Nous pouvons remarquer que près de la moitié des participants proviennent du même emplacement. Les autres bureaux comptent plutôt entre 1 et 62 individus chacun. Malgré tout, le bureau d'appartenance des employés n'est pas censé déterminer ou encore influencer la formation des communautés.

En ce qui concerne le nombre de bureaux impliqués dans les mises à jour, la figure 4.4 présente son évolution. En fait, il est croissant jusqu'à la cinquième période, pour ensuite décroitre légèrement puis se stabiliser à partir du neuvième semestre. Cette information est cohérente avec le nombre d'employés participant aux discussions étant donné l'augmentation simultanée des deux courbes correspondantes aux premières périodes. Il est possible d'en déduire qu'initialement, les individus actifs, soit ceux qui prenaient part aux mises à jour, étaient répertoriés dans la moitié des bureaux à l'étude. Toutefois, ceux qui se sont joints aux discussions par la suite venaient de différents emplacements, jusqu'à inclure des employés rattachés à la quasi-totalité des bureaux

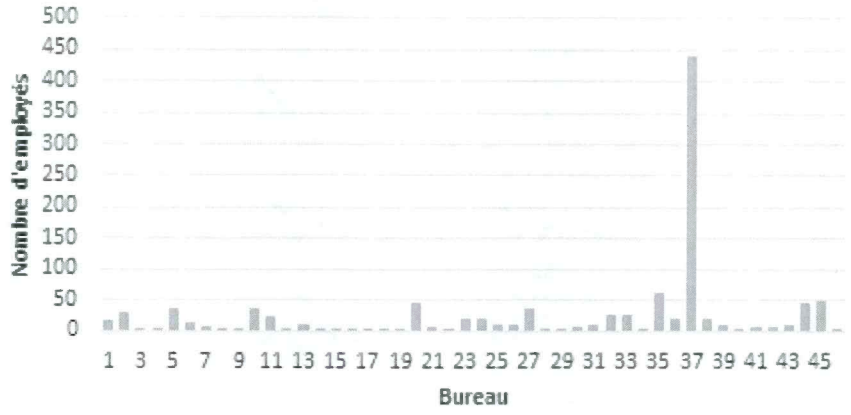


FIGURE 4.3 – Nombre d'employés impliqués dans les mises à jour et travaillant à chacun des 46 bureaux.

au cours d'une même période, soit 41 des 46, à la cinquième période.

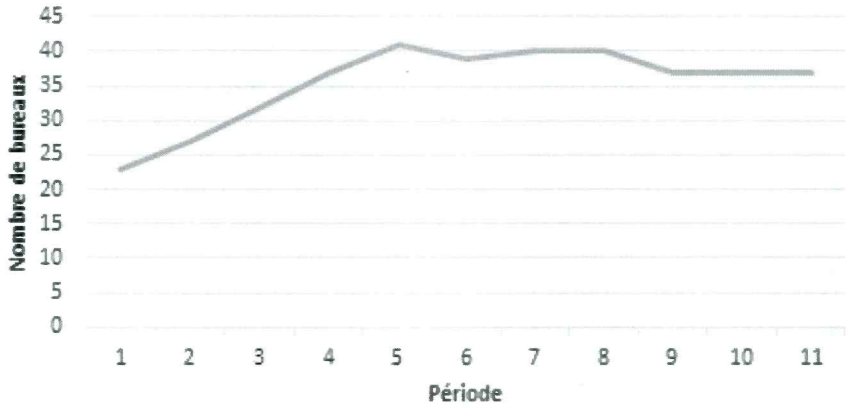


FIGURE 4.4 – Évolution du nombre de bureaux, dont au moins un employé est impliqué dans les mises à jour, au cours de l'horizon temporel étudié.

Les rôles des employés dans les discussions ont également été fournis. Certains rôles sont exercés plus souvent que d'autres, mais dans l'ensemble chacun d'eux suit la même tendance que le nombre de mises à jour à chaque période. La figure 4.5 illustre le nombre de commentaires émis selon le rôle de l'auteur pour chacun des semestres.

En considérant l'ensemble de l'horizon temporel étudié, nous remarquons que près de la moitié des employés ont discuté de 1 à 250 mises à jour, comme présenté à la

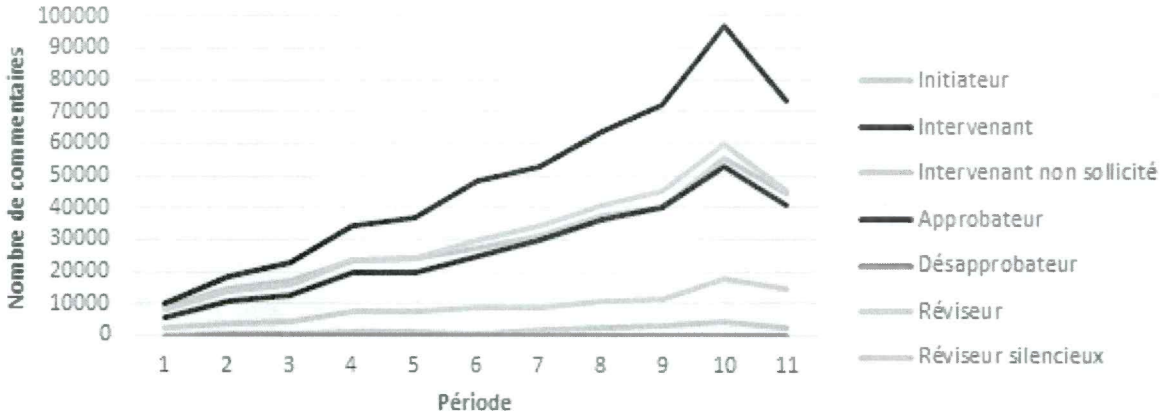


FIGURE 4.5 – Évolution du nombre de commentaires selon le rôle de l'employé qui l'a émis, au cours de l'horizon temporel étudié.

figure 4.6. Plus de 73% des participants se retrouvent dans les trois premiers intervalles, c'est-à-dire qu'ils se sont joints à, au plus, 750 discussions. En moyenne, les employés ont pris part à 727 mises à jour au total des périodes. Le moins actif a participé à 10 discussions au cours de l'horizon temporel alors que le plus actif a contribué à 14 392 mises à jour, s'éloignant significativement de la moyenne.

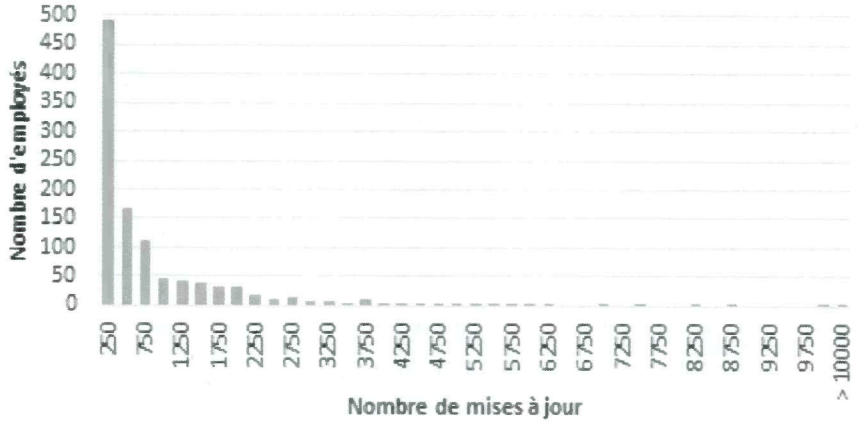


FIGURE 4.6 – Distribution des employés selon le nombre de mises à jour auxquelles ils ont participé durant l'ensemble de l'horizon temporel étudié. Les nombres de mises à jour présentés constituent la fin de l'intervalle commençant à zéro pour le premier et à la fin de l'intervalle précédent additionné d'un pour les suivants.

Plus spécifiquement, il est possible de regarder le nombre moyen de mises à jour auxquelles les individus ont participé chaque semestre, représenté dans la figure 4.7. Celui-ci s'élève à 109 au minimum à la cinquième période et à 172 au maximum au dixième semestre. Les variations dans la participation moyenne des employés proviennent du fait que le nombre de personnes actives sur le forum augmente au fil du temps jusqu'à la dixième période, alors que le nombre de mises à jour n'augmente pas aussi rapidement, et ce, jusqu'au sixième semestre inclusivement. Les discussions des individus sont donc concentrées sur un nombre restreint de mises à jour plus cette situation s'aggrave, d'où la diminution de la participation moyenne. Cependant, à partir de la sixième période, le nombre de nouveaux employés actifs ralentit alors que le nombre de mises à jour discutées augmente, d'où la croissance représentée dans la figure du sixième au dixième semestre. À la dernière période, ces deux éléments diminuent, mais pas à la même vitesse encore une fois, le nombre de mises à jour chutant davantage, d'où la diminution marquée.

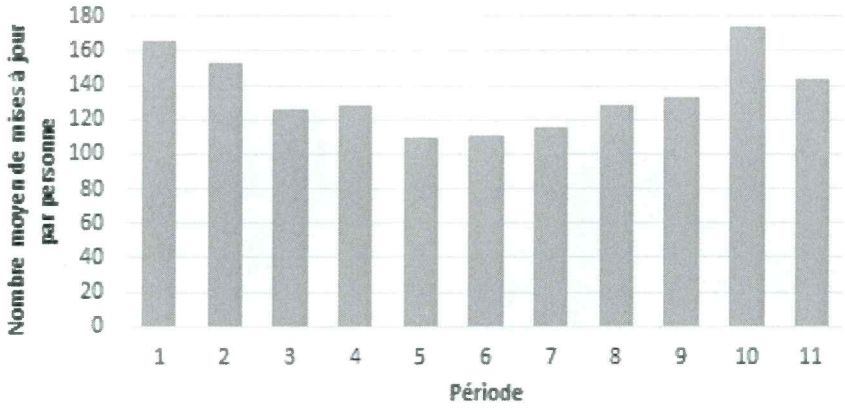


FIGURE 4.7 – Nombre moyen de mises à jour discutées par personne active au cours de l'horizon temporel étudié.

Le nombre d'employés par mise à jour peut également être analysé sur tout l'horizon temporel. La figure 4.8 présente la distribution des mises à jour selon leur nombre de participants. Plus de la moitié des discussions regroupent deux employés alors qu'il y en a près de 14% qui n'ont eu qu'un seul contributeur. Évidemment, la mise à jour la moins populaire a attiré un seul employé alors que la plus discutée en a rejoint 23. En moyenne, pour tout l'horizon temporel, une mise à jour regroupe 2,34 employés.

La participation moyenne des individus s'étale entre cinq et six périodes. Certains, soit

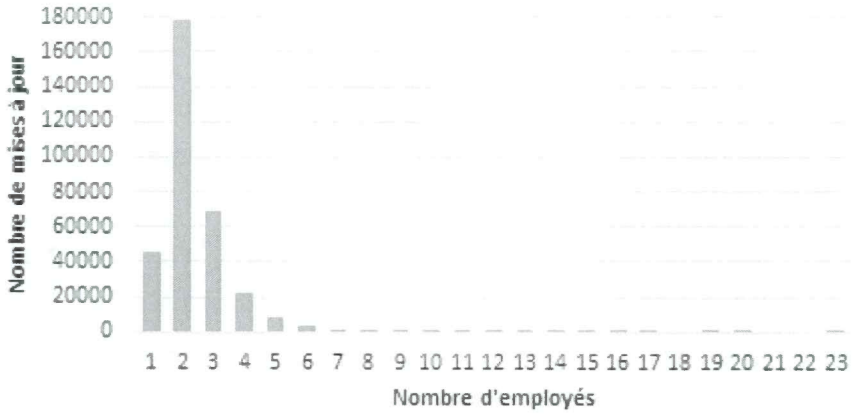


FIGURE 4.8 – Distribution des mises à jour selon le nombre d'employés qui y ont participé au cours de l'horizon temporel étudié.

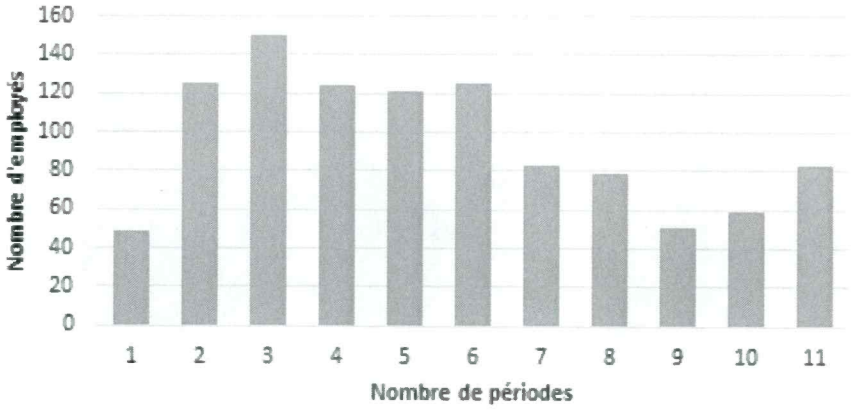


FIGURE 4.9 – Distribution des employés selon le nombre de périodes pendant lesquelles ils ont participé aux mises à jour sur l'ensemble de l'horizon temporel.

50 d'entre eux, se joignent aux discussions pendant un seul semestre alors que plus de 80 employés contribuent aux mises à jour à toutes les périodes. La figure 4.9 illustre la distribution des individus selon la durée de leur contribution sur le forum. Nous remarquons que 118 employés prennent des pauses, c'est-à-dire que leur participation ne se fait pas uniquement sur des périodes consécutives. Dans ces cas, l'arrêt des activités s'étale en moyenne sur deux périodes entre le début et la fin de leur participation. Malgré tout, pour la moitié des employés concernés, la pause ne dure qu'un seul semestre, comme présenté à la figure 4.10.

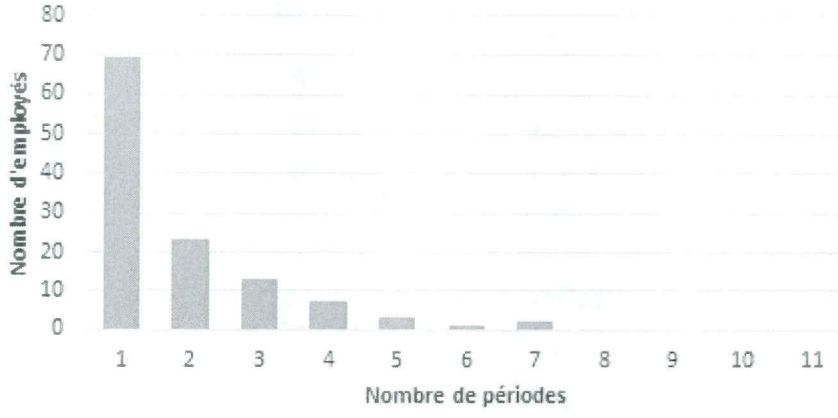


FIGURE 4.10 – Distribution des employés ne participant pas aux mises à jour de manière continue selon le nombre de périodes de pause qu'ils ont pris entre le début et la fin de leur contribution.

À chaque période, des individus se joignent pour la première fois aux discussions, comme le démontre la figure B.1 en Annexe. Le quatrième semestre est marqué par la plus grande arrivée de nouveaux participants sur le forum, soit plus de 140 employés, alors qu'à la dernière période moins de 20 individus contribuent pour la première fois aux discussions, soit le plus faible ajout relevé. Quant à la dernière intervention, elle se passe, pour la majorité des participants, soit 71% d'entre eux, à la fin de l'horizon temporel. Malgré tout, quelques employés se retirent dès la fin de la deuxième période, mais ce nombre est moindre. En fait, les retraits se multiplient plus le temps avance, ce qui est bien représenté sur la figure B.2 en Annexe.

D'autres analyses ont été réalisées, concernant notamment le nombre de commentaires effectués par période ainsi que le nombre de commentaires par personne au total de l'horizon temporel, en plus de celui à chaque période. Comme ces résultats suivent sensiblement les mêmes tendances que les analyses correspondantes avec les mises à jour, les figures ont été placées en Annexe. Il s'agit respectivement des figures B.3, B.4 et B.5. Elles ne seront pas commentées davantage n'étant pas directement représentées dans les différents modèles qui seront établis plus loin.

Il est donc possible de constater que des relations semblent s'établir entre les employés au fil du temps, en raison d'une participation accrue aux discussions. La contribution de chaque individu aux conversations semble toutefois variable.

4.2 Question de recherche

À partir du jeu de données, il est possible de former un réseau en y répertoriant les différentes interactions des employés. Celui-ci décrira donc la participation des individus aux mises à jour au fil du temps.

La problématique de ce mémoire consiste à développer un modèle adéquat pour représenter ce réseau ainsi que permettre la détection des communautés qui s'y trouvent. En se basant sur la partition trouvée, les comportements des employés ainsi que des communautés, au cours de l'horizon temporel, pourront être analysés plus en détail. Les principales tendances ou encore, divers faits saillants pourront en être dégagés. Comme aucune information nous a été transmise quant à la réalité derrière les données, il semble intéressant de les analyser et de chercher à cerner des phénomènes réels, captés que par la résolution du réseau.

Aucune communauté n'est connue initialement dans la compagnie. Les bureaux ne devraient pas, non plus, décrire des communautés, sachant que les employés sont encouragés à participer aux discussions initiées par tous les individus de l'entreprise et non pas particulièrement par ceux de leur établissement ou des autres rapprochés.

Ainsi, le choix du modèle utilisé sera très important pour la suite, car il influencera la détection des communautés et les résultats qui en découleront. Une attention particulière devra y être accordée.

Ainsi, la résolution s'effectuera pratiquement à l'aveugle, à l'aide uniquement des informations sur les commentaires qui nous ont été transmises. Elle visera à dégager les grandes tendances du comportement des employés au fil du temps, tout en permettant d'apercevoir certains événements importants survenus au cours de l'horizon temporel étudié.

Chapitre 5

Méthodologie

Ce chapitre présente le travail effectué en vue de répondre à la question de recherche adressée plus tôt. La modélisation du réseau évolutif formé à partir du jeu de données est d'abord présentée. L'approche de résolution pour la détection des communautés dynamiques est justifiée à la section qui suit. L'évaluation de l'impact de différents poids testés dans le graphe en plus des divers modèles étudiés est réalisée à la dernière section.

5.1 Modélisation

Une modélisation particulière du réseau évolutif sera étudiée dans ce mémoire. Cela permettra d'évaluer la pertinence et la performance d'une résolution avec celle-ci, qui n'est pas abordée telle quelle dans la littérature jusqu'à maintenant.

Correc [74] a déjà étudié la base de données utilisée pour ce mémoire. Son approche visait surtout à tester divers algorithmes de la littérature, pour la détection de communautés dans le réseau évolutif, modélisé avec différents instantanés. Dans le cas présent, l'aspect temporel sera plutôt représenté de manière longitudinale. En effet, les instantanés seront résolus simultanément, de manière semblable à ce que Mucha et al. [7] et quelques autres auteurs, présentés à la section 3.2.2, ont déjà proposé. Chaque instantané représentera les participations des individus aux mises à jour à chacune des périodes étudiées, tout comme ce que Correc avait fait. Cependant, un ajout sera réalisé à cette modélisation avec des instantanés et consistera à créer des liens dits tem-

porels, qui relieront les individus avec eux-mêmes d'une période à l'autre. Ces liens permettront de faire correspondre directement les communautés à travers le temps, par la résolution simultanée de tous les instantanés. Cette approche permettra, sans aucun doute, de faciliter l'étude de l'évolution des communautés, en n'ayant pas à apparier les communautés de chaque période, une fois la partition trouvée.

Le graphe utilisé sera composé de données hétérogènes, c'est-à-dire qu'il y a aura deux sous-ensembles de sommets, soit les individus et les mises à jour, qui seront reliés ensemble par des arêtes. Il ne sera toutefois pas qualifié de biparti, car pour ce faire, il faudrait des liens uniquement entre des sommets de types différents. Ce serait l'équivalent de résoudre chaque instantané séparément, chacun d'eux comportant des liens entre les employés et les mises à jour auxquelles ils ont participé. Cependant, comme des liens temporels seront ajoutés, reliant un même individu d'une période à l'autre, donc deux nœuds du même ensemble, la définition de graphe biparti ne sera plus respectée. Le graphe sera alors plutôt qualifié contenir des données hétérogènes seulement. La figure 5.1 représente bien cette différence, en version très simplifiée. On retrouve dans celle-ci des nœuds de deux types, soit les bleus en guise des employés et les orangés pour les mises à jour. L'exemple du haut présente quatre instantanés qui sont quatre graphes bipartis. En ajoutant les liens temporels entre les différentes représentations d'un même individu, soit entre les sommets bleus, dans l'exemple du bas, la structure du graphe biparti est brisée.

Plusieurs approches seront testées dans le but d'évaluer leur influence respective sur les partitions trouvées. Pour chacune d'elles, quatre questions se posent. D'abord, il faut déterminer la durée des périodes étudiées, tout comme il en est question avec l'utilisation indépendante d'instantanés. Il est important de choisir une durée suffisamment longue pour que des communautés significatives aient le temps de se créer. Elle ne doit cependant pas être trop grande non plus, afin de permettre d'étudier adéquatement l'évolution des communautés dynamiques. Deux variantes seront étudiées, soit la séparation de l'horizon temporel en semestres, ainsi qu'en trimestres. Celle qui semble la plus adéquate sera celle retenue pour l'analyse des résultats.

En raison de l'ajout des arêtes temporelles pour la modélisation, un deuxième élément doit être déterminé, soit la représentation des individus. D'un côté, ils peuvent être représentés à chaque période, qu'ils aient participé ou non au cours de celle-ci, alors que de l'autre, ils peuvent apparaitre, dans le graphe, qu'aux périodes où ils ont pris part à au moins une mise à jour.

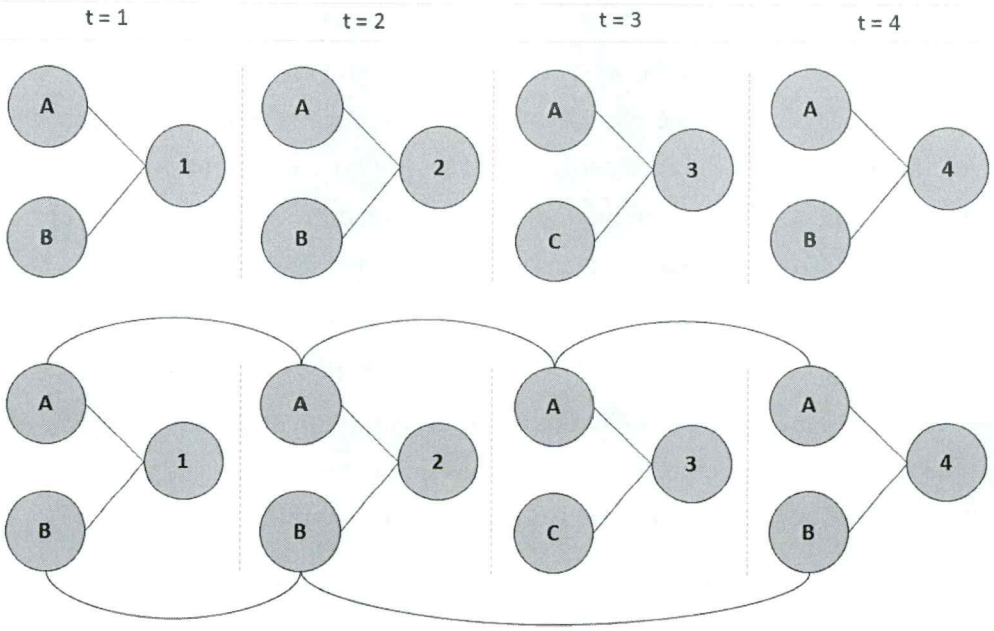


FIGURE 5.1 – Représentation de deux graphes évolutifs sur différents instantanés $G^{(t)}$, $t = 1,2,3,4$. Tous deux ont des données hétérogènes, soit des individus, en bleus, A, B, C, et des mises à jour, en orangé, 1, 2, 3, 4, mais seulement celui du haut est biparti.

Un troisième élément s'ajoute aussi aux paramètres à établir et concerne la mise en place des liens reliant un individu avec lui-même d'une période à l'autre. Ces arêtes peuvent notamment être ajoutées au graphe pour relier tous les sommets d'un même individu, en ordre chronologique d'apparition. Une autre possibilité est de relier les apparitions d'un employé que si elles sont réalisées à des périodes consécutives. Dans ce cas, il faut toutefois que les sommets des individus, aux semestres où ils sont inactifs, ne soient pas représentés dans le graphe. Par exemple, si un employé contribue aux discussions de mises à jour à la quatrième et la sixième période, en étant inactif et donc non représenté entre temps, ses sommets distancés ne seraient pas reliés.

Le quatrième et dernier élément à déterminer est le niveau des poids qui seront attribués aux différentes arêtes du graphe. Des poids pour les liens entre les individus et les mises à jour auxquelles ils ont participé devront être établis en plus d'autres poids pour caractériser les arêtes entre un employé et lui-même d'une période à l'autre. Initialement, des poids d'un seront affectés à tous les liens entre un individu et une mise à jour. Ces poids agiront à titre de référence en restant à un pour toutes les résolutions effectuées dans le cadre de ce mémoire. Ce niveau signifie la présence de l'employé

dans la discussion correspondante. Pour ce qui est des poids sur les liens temporels, ils varieront d'une résolution à l'autre, de manière à pouvoir évaluer leur impact sur la solution obtenue avec une telle modélisation. À chacune des résolutions, tous les liens entre les individus et eux-mêmes seront affectés de poids identiques. Les poids testés varieront entre 0,1 et 1000. Le maintien des poids d'un entre les individus et les mises à jour pour toutes les résolutions empêchera d'influencer les solutions obtenues. De cette façon, la variation entre les partitions sera totalement attribuable au changement des poids sur les arêtes temporelles, ce qui permettra d'en étudier le réel impact.

Trois modèles seront établis pour la détection de communautés dans le réseau étudié. Ils reflètent tous les différentes caractéristiques de ce dernier, c'est-à-dire les données hétérogènes, la pondération ainsi que l'information temporelle.

Dans le premier d'entre eux, chaque individu se verra dupliqué selon le nombre de périodes à l'étude. Chaque sommet ainsi créé sera relié avec son correspondant de la période suivante. La figure 5.2 illustre un exemple de cette représentation. Ce sont donc les nœuds des employés de chacune des périodes qui seront reliés aux mises à jour selon la date de participation à ces dernières. En fait, cette duplication revient à différencier un même individu selon l'instantané dans lequel il se trouve. Comme tous les instantanés seront résolus simultanément, cette identification temporelle permettra de distinguer les participations de chaque période et d'ainsi retracer l'évolution de l'employé une fois la partition trouvée. La figure 5.3 montre un exemple simple de ces liens. Un employé n'ayant participé à aucune discussion à une certaine période, par exemple *A P2*, sera représenté dans le graphe, relié avec ses homologues des périodes immédiatement avant et après, sans aucun lien vers une mise à jour. Les poids sur les différentes arêtes seront établis comme il a été présenté précédemment.

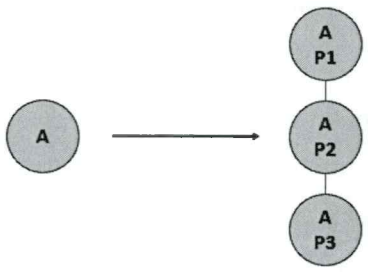


FIGURE 5.2 – Duplication d'un individu sur trois périodes, représentées par *P1, P2* et *P3*, en reliant tous les sommets de périodes consécutives.

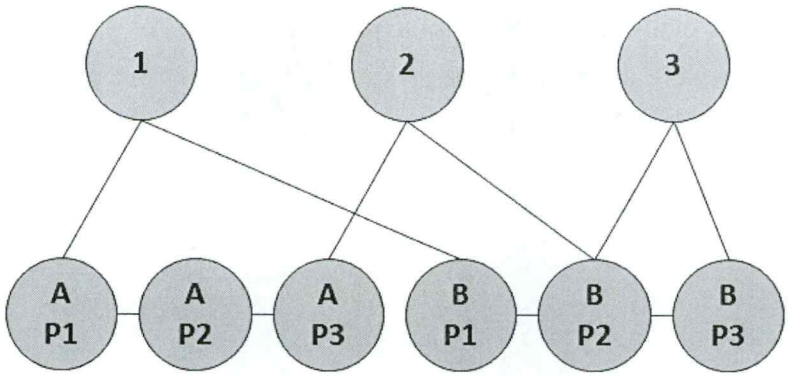


FIGURE 5.3 – Liaisons entre les individus et les mises à jour auxquelles ils ont participé selon la période à laquelle ils y ont pris part. Les sommets orangés représentent les mises à jour et les sommets bleus représentent les individus *A* et *B* aux périodes 1, 2 et 3, soit *P1, P2* et *P3* respectivement.

Les mises à jour ne seront pas dupliquées dans le temps, car elles ne sont discutées en moyenne pendant qu'une seule période. Pour celles dont ce n'est pas le cas, des employés de périodes différentes y seront reliés tout simplement. Il pourrait aussi arriver qu'un même employé y participe au cours de deux périodes. Ces deux situations sont illustrées respectivement, à la figure 5.3, avec les mises à jour 2 et 3.

Dans le deuxième modèle, chaque individu sera également dupliqué dans le temps, mais seulement pour les périodes pendant lesquelles il aura participé à au moins une mise à jour. Tous les sommets d'un même employé seront ensuite reliés avec celui de la période d'activité suivante, que celle-ci soit immédiatement après ou non. La figure 5.4 constitue un exemple simple de ce modèle. Dans ce cas, le nœud *A P2* ne sera plus présent et les nœuds qui le précède et le suive seront reliés directement ensemble. Les poids qui caractériseront la force des relations seront attribués comme décrit précédemment à une exception près. Lorsque l'arête relira les sommets d'un même individu, de deux périodes non consécutives, le poids établi sera le même qu'ailleurs, mais divisé par le nombre de périodes séparant les deux nœuds. Par exemple, dans le cas des sommets *A P1* et *A P3*, si le poids posé sur les arêtes entre les employés est d'un normalement, le poids sur le lien entre ces nœuds serait de $1/2$. Il s'agit en fait de la division du poids d'un par la différence de deux périodes entre les participations de l'individu *A*. Comme le comportement d'un employé est susceptible de changer, lorsque ce dernier ne participe à aucune mise à jour pendant au moins une période, il semble adéquat de réduire la force du lien entre le sommet qui précède son départ et celui à son retour. Le poids

réduit pourrait être établi de différentes façons, mais une seule a été retenue pour de ne pas dépasser le cadre de ce mémoire. Sa simplicité et sa correspondance de manière proportionnelle avec l'arrêt effectué la rendaient adéquate au contexte de l'étude.

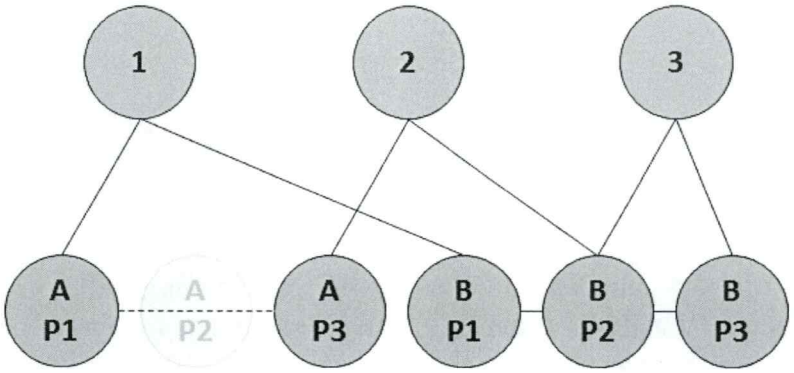


FIGURE 5.4 – Représentation des individus aux périodes pendant lesquelles ils ont pris part à des mises à jour seulement. Le sommet en gris, qui était présent dans le modèle précédent, ne fait plus partie du graphe, car l'individu *A* n'a contribué à aucune discussion pendant la deuxième période. Une arête est créée entre les nœuds qui le précède et le suive.

Le troisième modèle est semblable au deuxième, à une exception près. Seuls les sommets représentant les périodes pendant lesquelles l'employé a été actif sur le forum sont représentés dans le graphe. Cependant, les liens entre un individu et lui-même d'une période à l'autre sont fixés seulement si les nœuds concernés sont de périodes consécutives. Si ce n'est pas le cas, aucune arête ne les relira. La figure 5.5 présente bien ce modèle. Ainsi, plutôt que d'établir un lien entre les sommets *A P1* et *A P3* avec un poids ajusté, aucune liaison n'est réalisée. C'est donc l'équivalent d'une arête, comme il était fait précédemment, mais avec un poids nul. Cette variante permet de dissocier complètement un employé de lui-même lorsqu'il ne participe pas aux mises à jour de façon continue. La résolution permettra donc de vérifier si son comportement diffère à son retour, ce qui entrainerait un changement de communautés. Pour ce qui est des autres liens établis dans le graphe, ils seront caractérisés par des poids comme il a été décrit initialement.

Ce mémoire se consacrera donc à l'étude d'un réseau social, modélisé par un graphe évolutif à la fois pondéré et avec des données hétérogènes, de manière à représenter le plus fidèlement possible la réalité derrière les données qui nous ont été transmises. Comme les employés interagissent ensemble par l'intermédiaire des mises à jour et

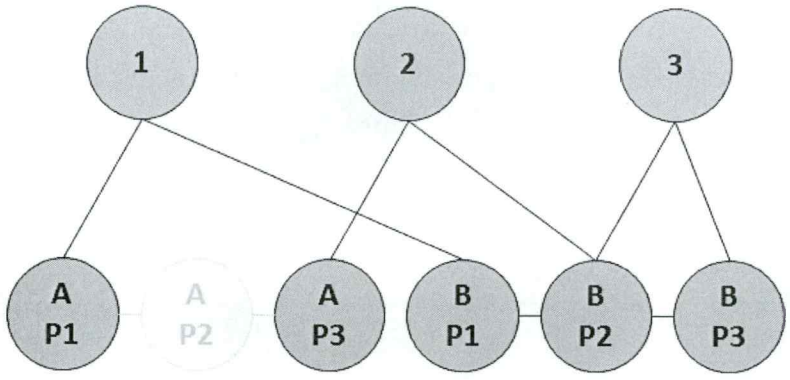


FIGURE 5.5 – Représentation des individus aux périodes pendant lesquelles ils ont pris part à des mises à jour seulement. Le sommet en gris ne fait plus partie du graphe, n'étant relié à aucune mise à jour, et ses arêtes ne sont pas remplacées.

non pas directement entre eux, il sera question d'une forme particulière de réseau social, nommée réseau de collaboration. L'approche de résolution choisie devra permettre l'analyse d'une partition obtenue à partir d'un tel graphe.

Chaque discussion et chaque employé sont représentés dans un réseau. Chacun des individus est relié aux conversations auxquelles il a pris part selon la période de six mois pendant laquelle son commentaire a été émis. En fait, pour l'étude temporelle des comportements, tous les commentaires sont regroupés par semestre et ainsi, un même individu est représenté à chacun des semestres, ou encore aux semestres où il a participé seulement, selon les modèles. La difficulté réside dans le niveau de cohérence du comportement des employés au fil du temps, à établir.

Certaines caractéristiques peuvent être dégagées à partir des modèles établis en vue de la résolution. Le tableau 5.1 fait notamment référence au nombre de sommets et d'arêtes du graphe en plus du degré moyen des nœuds selon les différents modèles utilisés. Comme présentée précédemment, le premier modèle consiste à la duplication des employés sur toutes les périodes en reliant chaque sommet créé avec celui de la période suivante. Étant donné que l'étude se fait sur 11 périodes, nombre obtenu par la division de l'horizon temporel en semestres, chaque individu est présent à 11 répétitions dans le graphe. Le deuxième modèle concerne la duplication des individus pour les périodes pendant lesquels ils ont participé à au moins une mise à jour. Dans celui-ci, chaque

nœud créé est relié avec celui de la période de participation suivante, en utilisant un poids ajusté, s'il ne s'agit pas de la période immédiatement après. Le troisième et dernier modèle représente ce même type de duplication sans toutefois relier les sommets d'un même individu qui ne sont pas de périodes consécutives.

Types de sommets	|V|	|E|	$\sum_u \frac{d_u}{N}$
Modèle 1			
Individus	11 495	10 450	68,44
Mises à jour	324 860	765 791	2,36
Tous sommets confondus	336 355	776 241	4,61
Modèle 2			
Individus	5 750	4 647	134,80
Mises à jour	324 860	765 791	2,36
Tous sommets confondus	330 610	770 438	4,66
Modèle 3			
Individus	5 750	4 514	134,75
Mises à jour	324 860	765 791	2,36
Tous sommets confondus	330 610	770 305	4,63

TABLEAU 5.1 – Nombre de sommets et d'arêtes pour chaque type de nœuds, soit individus ou mises à jour, ainsi qu'au total pour chacun des modèles. Le degré moyen est aussi présenté dans tous les cas. Le nombre d'arêtes associé aux individus ne représente en fait que les liens temporels, c'est-à-dire avec eux-mêmes. Le nombre d'arêtes des mises à jour réfère au nombre de liens entre les deux types de sommets.

Étant donné le grand nombre de mises à jour comparativement au nombre d'employés dans le réseau, le degré moyen de chacun de ces sommets diffère grandement. On peut en déduire qu'avec le premier modèle, un individu participe en moyenne à 66 mises à jour par période, soit son degré moyen de 68 diminué de deux étant donné les liens avec lui-même aux périodes immédiatement avant et après. En raison de la présence de sommets représentant des employés ne participant à aucune mise à jour, le degré moyen pour l'ensemble de ce type de nœuds est grandement affecté à la baisse. Effectivement, avec les autres modèles, il passe plutôt à près de 135. Cela signifie qu'en moyenne, un

individu participe à environ 133 mises à jour par période, ce qui est cohérent avec les observations relevées par l'analyse descriptive du jeu de données. Cette participation moyenne est obtenue en diminuant de deux le degré moyen de 135 pour ne pas considérer les liens temporels qui ne représentent pas des contributions au forum. Le degré moyen des sommets constituant les mises à jour est le même pour tous les modèles, ce qui est normal étant donné qu'aucun changement n'est effectué quant aux arêtes les reliant. Il ne diffère que très peu du nombre moyen de participants aux mises à jour présenté à la section 4.1.2. Le léger écart provient du fait qu'un individu, participant à une même discussion à deux périodes différentes, est comptabilisé deux fois pour le calcul du degré moyen contrairement à une seule fois pour tout l'horizon temporel dans les analyses réalisées plus tôt. Cette situation survient à 1 852 reprises, d'où la différence. Il s'agit cependant de 1 852 interactions en moins sur 765 791 liens au total des semestres, ce qui est plutôt négligeable.

5.2 Détection de communautés dynamiques

Une fois les modèles établis, il est possible de les résoudre, selon les différents paramètres à tester, de manière à en dégager les communautés qui s'y trouvent. Pour ce faire, certains choix doivent être effectués quant à la méthode de résolution à utiliser.

5.2.1 Choix du critère d'évaluation

Le choix du critère de modularité, adapté aux graphes pondérés, pour l'évaluation de la qualité des communautés a d'abord été réalisé. La modularité constitue, en fait, le critère le plus utilisé et reconnu dans le domaine. Malgré certaines critiques à son égard et les défauts qu'il possède, c'est celui qui semble le plus fiable. Il permet également de mesurer la qualité des communautés sans devoir se référer à une partition de référence, représentant la solution cherchée. Cette particularité le rend donc adéquat dans un contexte comme celui de ce mémoire, où aucune communauté n'est connue à l'avance. De plus, ce critère est généralisable aux réseaux pondérés, ce qui le rend d'autant plus pertinent. Comme la problématique se concentre principalement sur la modélisation du réseau, le critère choisi ne constitue pas un élément central de l'approche de résolution utilisée. Le choix standard de la modularité semble donc justifié.

5.2.2 Choix de l'algorithme de résolution

Une multitude d'algorithmes existe afin de détecter des communautés dans un réseau. Cependant, ils ne sont pas tous adaptés à tous les types de graphes. Dans le cas présent, il faut s'assurer d'en sélectionner un qui considère à la fois l'aspect temporel, la pondération sur les liens ainsi que les données hétérogènes.

La détection des communautés, dans le réseau évolutif décrit précédemment, se fera par l'algorithme de **Louvain**, en raison de ses propriétés qui s'agencent aux caractéristiques du graphe établi. Il permet à la fois la résolution de graphes pondérés, par l'optimisation de la modularité considérant les poids, et comportant des données hétérogènes, en plus de proposer une solution juste rapidement. La résolution de tous les instantanés simultanément sera également possible avec cet algorithme. Comme le critère choisi pour l'évaluation de la qualité des communautés est la modularité, d'autres approches de résolution auraient évidemment pu être choisies. Il existe notamment le **LPAm**, mais celui-ci ne considère pas les poids attribués sur les liens et n'est pas aussi robuste que l'algorithme de **Louvain**. Ce dernier constitue donc un choix standard et adapté au type de réseau étudié, ce qui en justifie l'utilisation.

Cet algorithme a été implanté dans un programme maison, permettant la résolution de grands graphes ayant jusqu'à quelques millions de nœuds. Cette propriété est pertinente étant donné l'ampleur du réseau à étudier qui possède au plus 336 355 sommets et près de 800 000 arêtes. Comme décrit à la section 3.2.1, l'algorithme de **Louvain** est l'un des plus utilisés dans le domaine en raison de son efficacité et sa rapidité même avec les graphes de grande taille. Il semble donc bien adapté au contexte de la présente étude.

Son utilisation lors de la détection des communautés dynamiques a permis d'atteindre un haut niveau de modularité, comme présenté dans le tableau A.1 en Annexe. L'algorithme exécutait plusieurs résolutions simultanément en ne présentant toutefois que la meilleure d'entre elles en fonction de la modularité la plus élevée trouvée. C'est cette solution qui est inscrite dans le tableau A.1. Dans tous les cas, le critère est suffisamment élevé pour permettre une analyse pertinente des communautés.

5.3 Évaluation de l'impact des poids et des modèles

Les différents modèles testés, en plus des divers poids attribués lors des résolutions, ont permis l'obtention d'une multitude de partitions. Cependant, analyser chacune d'elle en détail dépasserait le cadre de ce mémoire. Ainsi, l'impact des poids établis sur les liens entre les individus et eux-mêmes sera analysé de façon générale dans cette section, en plus de l'influence des modèles. Le nombre de communautés trouvées, l'appartenance des employés, la durée de vie des communautés ainsi que leur taille constitueront les éléments de base aux comparaisons. Le chapitre suivant sera réservé à l'analyse en détail d'une partition considérée des plus adéquates et représentatives parmi celles trouvées.

Avant tout, il est important de mentionner que le nombre de périodes utilisé dans les résolutions qui suivent, peu importe le modèle, est de 11. Il a été obtenu par la division de l'horizon temporel en semestres. Cette durée est suffisamment longue pour que des communautés aient le temps de se former. Elle permet également un nombre acceptable de périodes pour l'analyse des mouvements des employés à travers le temps, en plus de l'évolution des communautés. La variante des trimestres a été écartée, car il s'agit d'une période trop courte pour permettre la formation de communautés significatives. Cette décision a été prise à la vue de partitions obtenues avec ce paramètre et dont les individus étaient très volatiles à travers le temps, comparativement aux résultats avec les semestres.

5.3.1 Nombre de communautés

Le nombre de communautés trouvées pour chaque modèle et chacun des poids testés se trouve au tableau A.1 et est illustré à la figure 5.6. Il est possible de remarquer que les poids attribués influencent le nombre de communautés détectées et ce peu importe le modèle choisi.

Le premier modèle, qui inclut la présence de tous les individus à toutes les périodes, qu'ils aient participé ou non pendant celles-ci, a généré quelques partitions ressortant du lot. En effet, le nombre de communautés trouvées, dans les graphes avec des poids inférieurs à deux, était plutôt élevé comparativement aux autres modèles utilisant des poids semblables, en s'élevant entre 91 et 121 communautés. En étudiant le nombre de

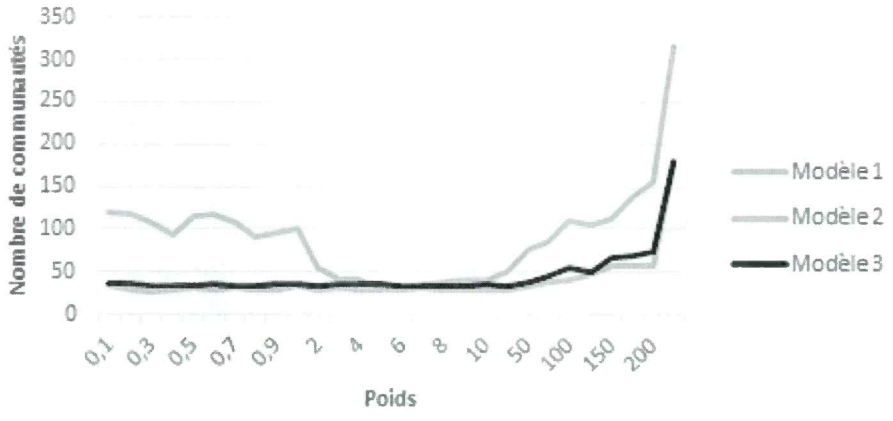


FIGURE 5.6 – Nombre de communautés trouvées avec chaque modèle et chacun des poids testés.

communautés à chaque période ainsi que la taille de celles-ci, l'origine de ces résultats a été trouvée. Comme des sommets d'individus sont représentés sans être reliés à des mises à jour, avec des poids faibles, ils sont peu connectés au reste du réseau et se retrouvent seuls dans leur communauté. Dans le pire cas, soit avec des poids de 0,1, 88 sommets sont dans cette situation, ce qui influence largement le nombre de communautés trouvées. Ces nœuds, seuls pendant les premiers semestres, se joignent toutefois à d'autres communautés existantes à partir de la huitième période, ce qui diminue significativement le nombre de communautés durant les derniers semestres. La cause de ce changement à ce moment reste toutefois inexplicable. Des données supplémentaires seraient nécessaires pour approfondir la compréhension de cette situation.

Pour ce qui est des deuxième et troisième modèles, le nombre de communautés trouvées est plutôt semblable pour les poids allant de 0,1 à 50 et de 0,1 à 25 respectivement. Cela signifierait donc qu'avec des poids faibles, les liens entre les mises à jour et les employés influencent davantage les partitions trouvées que les arêtes temporelles. Ainsi, peu importe les poids établis entre les individus, le nombre de communautés détectées est semblable. Cependant, dans les deux cas, à partir des poids de 25, ce nombre est croissant, plus la pondération établie augmente. À ce moment, ce sont alors les liens entre les participants qui semblent avoir le plus d'influence sur la formation des communautés. Dans ce contexte, celles-ci sont de plus en plus nombreuses et regroupent tous les sommets de seulement quelques individus, ce qui est peu pertinent. Il sera donc important de choisir un poids qui n'est ni trop faible, ni trop élevé, de manière à

considérer le plus équitablement possible les deux types de liens.

5.3.2 Appartenance des individus

Un autre élément qui est influencé par la variation des paramètres est la durée d'appartenance des individus aux communautés. La figure 5.7 présente le nombre moyen de communautés d'appartenance des employés sur l'ensemble de l'horizon temporel, selon les différents modèles et poids testés. La figure 5.8 montre plutôt le nombre total de changements de communautés réalisés au fil des périodes selon les paramètres établis.

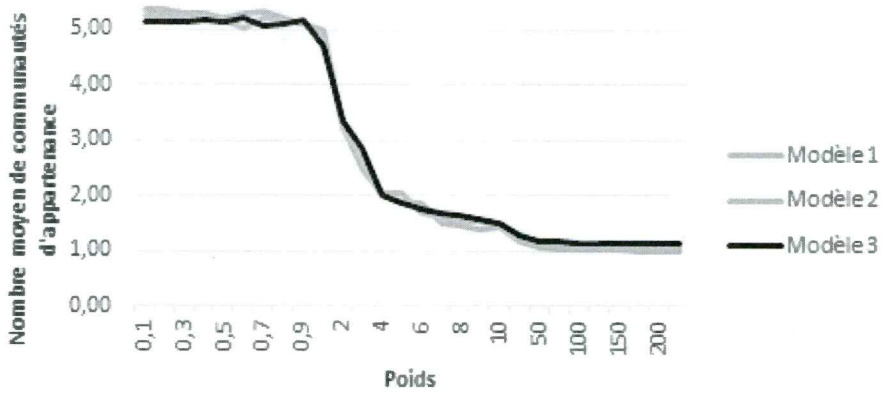


FIGURE 5.7 – Nombre moyen de communautés que les employés ont rejoint sur l'ensemble de l'horizon temporel avec chaque modèle et chacun des poids testés.

Plus les poids sont élevés, plus les employés sont stables, c'est-à-dire qu'ils restent longtemps dans une même communauté. Avec des poids faibles, les individus changent souvent de groupes et ce, peu importe le modèle choisi. En moyenne, avec de telles pondérations, chaque participant se joint à plus de cinq communautés différentes sur l'ensemble de l'horizon temporel, pour un total de plus de 4 000 changements. Avec des poids plus élevés, les employés restent très souvent dans la même communauté d'une période à l'autre, ce qui réduit le nombre total de changements qui passe sous la centaine avec les deux premiers modèles. Le troisième d'entre eux, en éliminant certains liens entre les individus, engendre plus de changements, pour un total plutôt sous les 200 avec des poids élevés. En raison de la présence de liens tous pareillement pondérés entre les sommets d'un même employé, le premier modèle favorise davantage la stabilité des participants. Ainsi, il en vient à ce que tous les individus restent dans

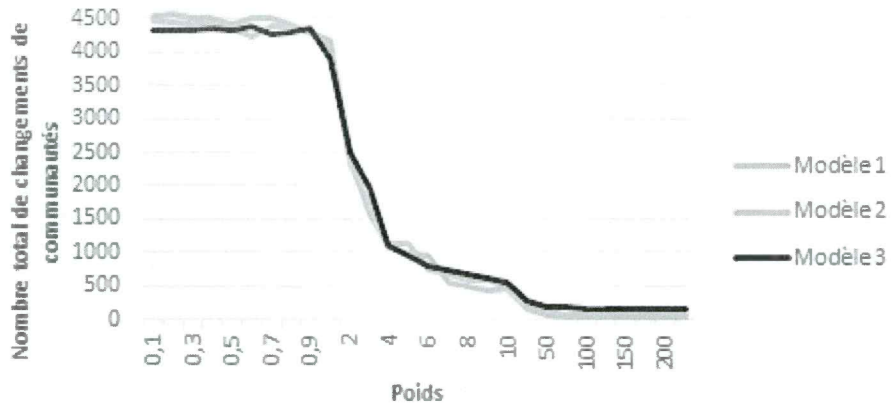


FIGURE 5.8 – Nombre total de changements de communautés réalisés sur l'ensemble de l'horizon temporel avec chaque modèle et chacun des poids testés.

la même communauté sur tout l'horizon temporel alors qu'avec des poids identiques, les autres modèles entrainent encore des changements. Il sera donc important de choisir une partition qui engendre suffisamment de mouvements, tout en étant respectable, pour permettre une analyse adéquate et réaliste du comportement des employés.

5.3.3 Durée de vie des communautés

Relativement aux analyses précédentes, la durée de vie des communautés se trouve également affectée par le modèle utilisé et le niveau des poids attribués aux arêtes temporelles. La durée de vie moyenne des communautés, en termes du nombre de périodes d'existence, est présentée à la figure 5.9 selon les différents paramètres testés.

En général, plus les poids sont élevés, plus les communautés perdurent à travers le temps. Cela s'explique par le fait, qu'avec des poids faibles, leur composition est majoritairement basée sur les participations communes aux mises à jour. Comme ces dernières se déroulent habituellement pendant qu'un seul semestre, les communautés ont plus de difficulté à perdurer. À l'inverse, avec des poids élevés, la formation des communautés se base principalement sur les liens temporels, donc elles regroupent plusieurs sommets de quelques individus. Son existence suit alors la durée de participation des employés qu'elle regroupe.

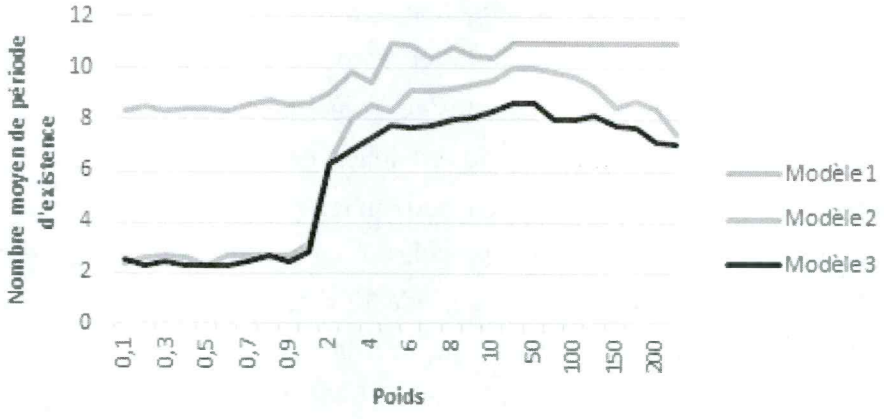


FIGURE 5.9 – Durée de vie moyenne des communautés, en termes du nombre de périodes d'existence, pour chaque modèle et chacun des poids testés.

Le premier modèle favorise une durée de vie plus grande des communautés, en raison de la représentation d'un même participant à tous les semestres, qu'il ait été actif ou non. Dans ce contexte, l'utilisation de poids élevés engendre la formation de nombreuses communautés qui durent sur tout l'horizon temporel, car elles regroupent tous les sommets d'un même individu. Avec des poids de 1000, c'est le cas de la totalité des communautés détectées. Cependant, cette situation est moins pertinente et réaliste, car à certaines périodes, la communauté existe, mais regroupe uniquement des employés qui ne participent pas.

Les deux autres modèles permettent des résultats similaires quant à la durée de vie des communautés. Dans ces cas, elles survivent moins longtemps, en raison des participations limitées des employés, qui sont mieux représentées dans le graphe. Le dernier modèle, par le retrait de certains liens lorsqu'une pause survient, engendre plus de communautés de plus courte durée que les partitions obtenues avec le deuxième modèle. Au moment du choix de la partition idéale pour l'analyse, il sera important de s'assurer que la durée de vie des communautés représente autant que possible une situation réaliste.

5.3.4 Taille des communautés

La taille des communautés est également influencée par les poids attribués avec chacun des modèles testés. Effectivement, avec des poids faibles, il existe, en général,

quelques petites communautés et toujours une significativement plus grande, qui prend de l'ampleur d'une période à l'autre. Celle-ci regroupe la majorité des individus actifs initialement et s'agrandit chaque semestre en accueillant, en plus, une grande partie des nouveaux participants. Selon les poids attribués, cette communauté peut regrouper jusqu'à plus de 600 employés à la fin de l'horizon temporel. Sa croissance se remarque cependant seulement avec les derniers modèles. Pour ce qui est du premier, comme le nombre de sommets est constant d'une période à l'autre, dès le début, une très grande communauté est formée et elle perdure jusqu'à la fin en regroupant toujours autant d'employés. Avec des poids plus élevés, les différentes communautés sont plutôt de tailles semblables et sous la centaine en termes du nombre de participants qu'elles regroupent.

En lien avec les analyses précédentes, nous remarquons que le nombre de communautés peut être similaire avec différents poids, mais les caractéristiques de celles-ci diffèrent tout de même. Avec des poids très faibles, il n'y a que quelques communautés à chaque période, dont une très grande, et celles-ci durent que quelques semestres, voire un seul. Avec des poids légèrement supérieurs, plus de communautés sont trouvées à chaque période. Elles sont de tailles semblables, soit moins de 100 employés en général, voire moins de 50 pour la majorité, et leur durée de vie est plus grande.

Chapitre 6

Analyse des résultats

L'analyse des résultats permet de comprendre la composition et l'évolution des communautés en plus du comportement des employés. Ainsi, la partition semblant la plus réaliste sera sélectionnée. Elle sera ensuite étudiée quant aux caractéristiques de ses communautés et aux mouvements des employés, de manière à dégager les tendances générales du réseau évolutif formé à partir du jeu de données. Une évaluation des divers types de liens sera réalisée pour terminer.

6.1 Choix de la partition à analyser

La partition choisie pour l'analyse détaillée provient du deuxième modèle. Avec celui-ci, un individu est représenté aux périodes pendant lesquelles il a participé à des mises à jour seulement et tous ses sommets sont reliés, qu'il ait pris une pause ou non. Le premier modèle, par la présence de sommets reliés à aucune mise à jour, était moins adéquat pour refléter le contexte réel traduit par la base de données. De plus, comme il a été présenté précédemment, ces sommets influencent les résultats alors qu'ils sont peu pertinents. Le troisième modèle, soit celui représentant les individus aux périodes où ils ont participé à des mises à jour et qui relie les sommets de périodes consécutives seulement, semblait aussi légèrement moins approprié. L'absence de liens entre les sommets d'un même individu à deux périodes non successives se rapproche moins de la réalité que le modèle choisi. En effet, il semble préférable d'établir une arête entre de tels nœuds avec un poids ajusté que ne pas en mettre du tout, en guise d'un individu

complètement différent.

La pondération choisie pour ce modèle est 25. Cette décision a été prise au regard des différentes analyses effectuées au chapitre précédent. D'abord, il semble y avoir une coupure à partir de ce poids dans le nombre de communautés détectées. Avec des poids inférieurs, la résolution se base davantage sur les participations communes aux mises à jour alors qu'avec des poids supérieurs, ce sont les liens temporels qui ont le plus d'influence. Cette partition a donc l'avantage de considérer plus équitablement les deux types de liens lors de la détection, en se situant entre ces deux tendances. L'appartenance des individus, la durée de vie des communautés ainsi que la taille de celles-ci semblent également réalistes et seront analysées davantage dans les sections qui suivent.

> Le niveau de cohérence dans le comportement des employés à travers le temps a été établi à la vue des résultats obtenus en testant plusieurs valeurs et en sélectionnant celle permettant la solution la plus réaliste.

Les différents mouvements réalisés à chaque période dans cette partition sont répertoriés des tableaux A.2 à A.11. Les plus marquants d'entre eux seront abordés dans ce chapitre selon les différents concepts à l'étude.

6.2 Caractéristiques des communautés trouvées

Les communautés seront étudiées sous différents aspects, à savoir, leur existence, en termes temporels, ainsi que leur taille.

La résolution du graphe évolutif avec les paramètres choisis a permis de détecter 29 communautés. Le nombre de communautés à chaque période est toutefois variable, comme l'illustre la figure 6.1. Initialement, 21 communautés sont formées, puis de nouvelles s'ajoutent aux périodes qui suivent, jusqu'à atteindre un maximum de 29 communautés, toutes présentes au huitième semestre.

La figure 6.2 montre plus clairement les communautés en action à chaque période. Cette représentation permet de voir explicitement l'arrivée et le départ de chacune des communautés. On y remarque également que deux d'entre elles, soit la *7* et la *17*, sont présentes de façon discontinue sur l'horizon temporel. Il est alors possible de parler

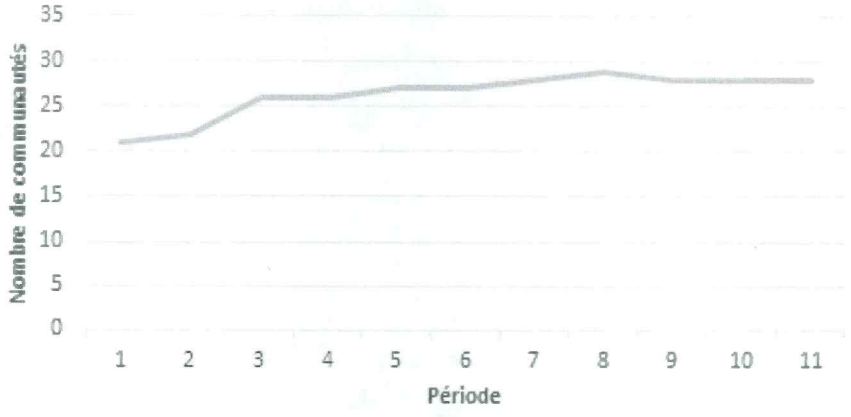


FIGURE 6.1 – Nombre de communautés présentes à chaque période de l'horizon temporel.

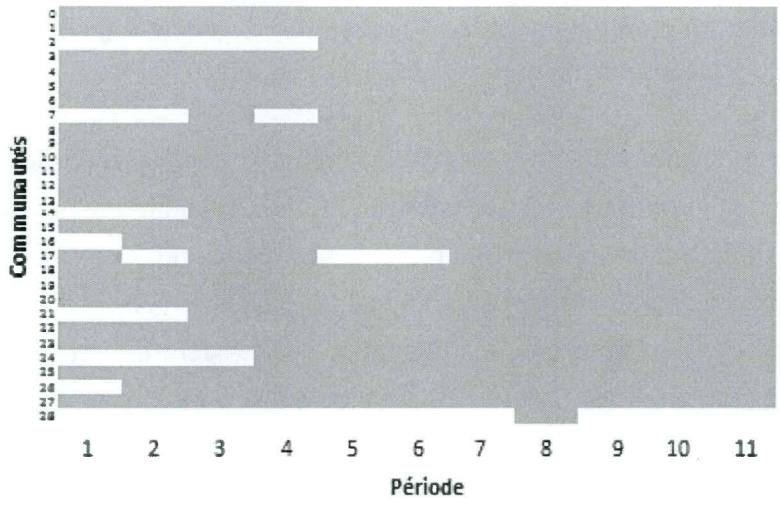


FIGURE 6.2 – Périodes d'existence de chacune des communautés détectées. Les cases bleues représentent la présence de la communauté, soit lorsqu'elle regroupe au moins un employé, alors que les cases blanches signifient son absence, ce qui équivaut à une taille nulle.

de résurgence de celles-ci. La communauté dans le bas de la figure, la *28*, pour sa part, regroupe des employés pendant une seule période. Excepté cette communauté qui disparait après le huitième semestre, toutes les autres perdurent jusqu'à la fin de l'horizon temporel. La figure 6.3 regroupe le nombre de communautés qui naissent et qui meurent à chaque période. La résurgence des communautés *7* et *17* est représentée

par une mort lorsqu'elles disparaissent et une nouvelle naissance à leur retour.

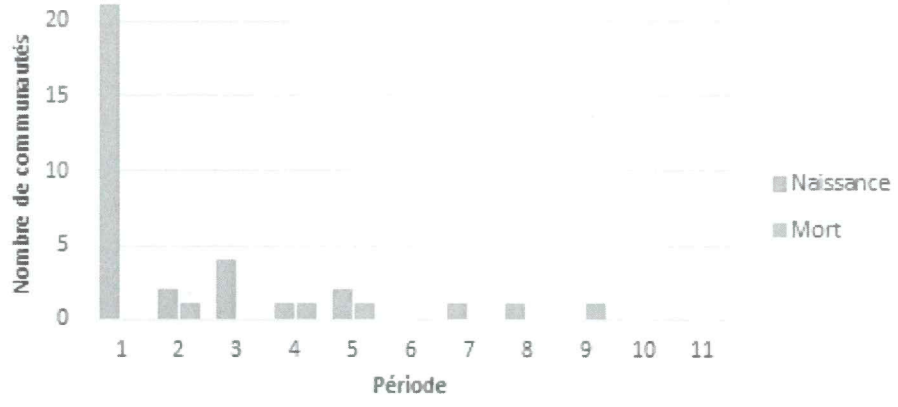


FIGURE 6.3 – Distribution des communautés selon la période où elles naissent et celle où elles meurent. La mort est constatée à la période où la communauté n'est plus présente.

La durée de vie des communautés est assez semblable pour la majorité d'entre elles, comme illustrée à la figure 6.4. Effectivement, 20 des 29 communautés sont présentes pendant tout l'horizon temporel. Il s'agit de toutes celles qui ont été formées dès la première période, à l'exception de la communauté *17*, en raison de son absence pendant quelques semestres. En moyenne, les communautés existent durant dix périodes. Le phénomène de continuation est donc très répandu dans ce réseau.

Bien que la majorité des communautés existent au cours de tous les semestres ou presque, leur taille diffère et varie au fil du temps. La figure 6.5 présente justement l'évolution de la taille de chacune des communautés au cours de l'horizon temporel.

Cette figure, quoique bien remplie, illustre des faits intéressants. Certaines courbes se démarquent des autres, notamment celles représentant les communautés *0, 12, 13, 20* et *26*. Elles seront analysées plus en détail à la section 6.3.5. On constate aussi que les communautés *7* et *17* touchent l'axe horizontal au moment où elles meurent avant de resurgir par la suite. Malgré tout, de manière générale, il semble y avoir une tendance à la hausse du nombre d'individus par communauté plus le temps avance. Celles de grande taille se multiplient d'une période à l'autre. Le regroupement des communautés selon leur taille chaque semestre est présenté à la figure 6.6.

FIGURE 6.4 – Distribution des communautés selon leur nombre de périodes d'existence sur l'ensemble de l'horizon temporel.

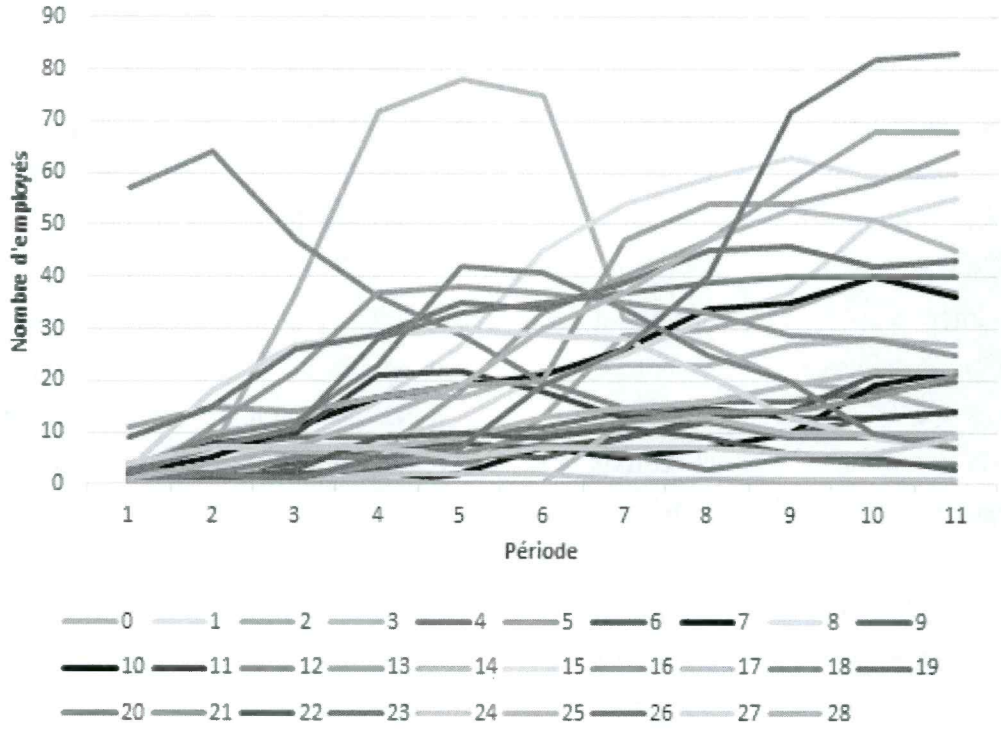


FIGURE 6.5 – Évolution de la taille des communautés, en termes du nombre d'employés qu'elles regroupent, sur l'ensemble de l'horizon temporel.

Plus le temps s'écoule, plus le nombre de participants augmente, jusqu'à la dixième période, alors que le nombre de communautés est plutôt semblable d'un semestre à

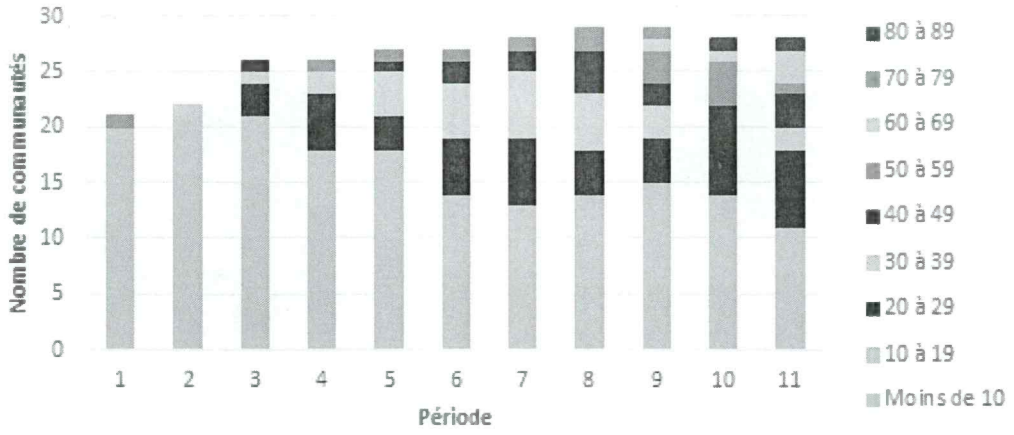


FIGURE 6.6 – Distribution des communautés selon leur taille, en termes du nombre d'employés qu'elles regroupent, chaque semestre.

l'autre. Il est donc normal que la taille des communautés soit croissante avec le temps. Il est possible de remarquer, en effet, que le nombre de petites communautés, soit celles regroupant moins de dix employés, est moindre plus le temps avance. À l'inverse, les communautés avec 20 participants ou plus, voire plus de 40, sont de plus en plus nombreuses. Toutefois, la communauté représentée en vert à la deuxième période, signifiant qu'elle rassemble entre 60 et 69 individus, semble se diviser au troisième semestre. À ce moment, le plus grand groupe est composé, au plus, de 49 employés, ce qui est bien inférieur. En faisant correspondre cette situation avec la figure 6.5, il est possible de déduire qu'il s'agit de la communauté *12*. Une situation semblable survient également avec la communauté *0* entre le sixième et le septième semestre. Celle-ci, désignée en orange jusqu'à la sixième période, dans la figure 6.6, en guise d'un regroupement entre 70 et 79 participants, semble être décomposée à la période suivante. En effet, la plus grande communauté à ce moment est d'au plus 59 individus. Au dernier semestre, la taille moyenne des communautés diminue, en raison, notamment, du nombre de participations en baisse au cours de celui-ci.

Ces analyses permettent de constater que le comportement des communautés est différent, à la fois d'une communauté à l'autre et d'un semestre à l'autre. Malgré tout, quelques tendances sont partagées par la majorité d'entre elles, notamment concernant leur durée de vie ainsi que le moment de leur naissance et leur mort. Une étude plus approfondie de la composition des communautés, en termes d'employés, sera réalisée à la section qui suit.

Des groupes distincts semblent véritablement exister dans cette entreprise, signifiant que les employés développement davanatge d'affinités avec certains de leurs collègues qu'avec d'autres. La composition de ces communautés diffère entre elles et changent au fil du temps. Certaines tendances générales sont observales, tout comme certains groupes se distinguent davantage des autres. Une étude détaillée de ceux-ci sera réalisée à la section suivante.

6.3 Caractéristiques des individus

Il est intéressant d'étudier le comportement des employés afin de mieux comprendre leur regroupement en communautés. En ce sens, l'appartenance des employés aux communautés sera analysée, en plus des changements de communautés, des nouvelles participations, des arrêts de participation ainsi que des mouvements marquants.

6.3.1 Appartenance des employés

Dans la partition choisie, les individus se joignent en moyenne à 1,18 communauté sur l'ensemble de l'horizon temporel. La figure 6.7 présente justement la distribution des participants selon le nombre de communautés qu'ils ont rejoint au total des périodes.

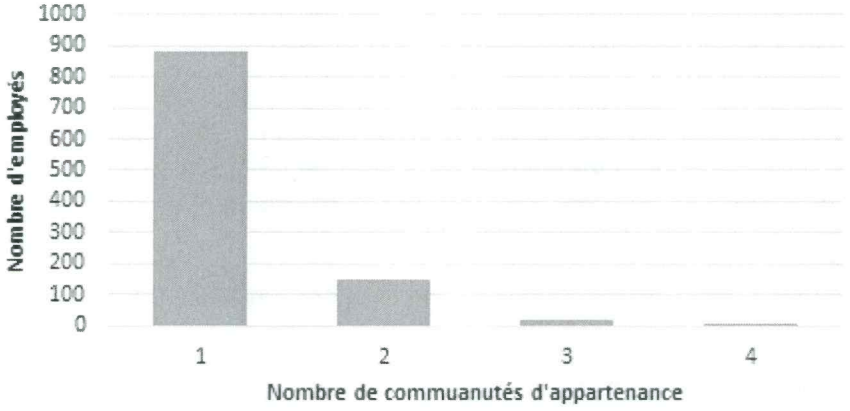


FIGURE 6.7 – Distribution des employés selon le nombre de communautés qu'ils ont rejoint sur l'ensemble de l'horizon temporel.

Il est possible de constater que la très grande majorité des individus, soit 84,02% d'entre eux, restent dans la même communauté tout au long de leur participation. Les employés les plus actifs rejoignent, quant à eux, quatre communautés différentes au cours de leur participation au forum et se présentent au nombre de deux dans cette situation. La figure 6.8 permet de mettre en relation le nombre de communautés d'appartenance avec le nombre de périodes d'activité des employés concernés. Cette figure permet de constater que les individus les plus stables sont ceux qui, en moyenne, ont participé pendant le moins de semestres. En fait, le nombre moyen de communautés jointes par chaque employé augmente plus sa participation s'est étalée dans le temps. Cependant, ces deux éléments ne croient pas de façon proportionnelle. Cela engendre donc une durée moindre d'appartenance à une même communauté, plus le participant est actif longtemps.

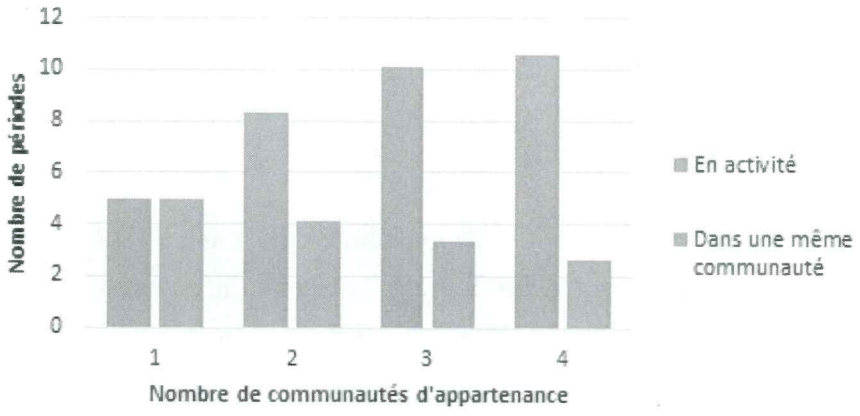


FIGURE 6.8 – Nombre moyen de semestres d'activités des employés et nombre moyen de périodes qu'ils ont passé dans une même communauté, en fonction du nombre de communautés qu'ils ont rejoint sur l'ensemble de l'horizon temporel.

En se concentrant sur les individus les plus stables, soit ceux qui sont restés dans la même communauté tout au long de leur participation, il est possible de remarquer que certains d'entre eux ont participé durant une longue période, voire durant tous les semestres de l'horizon temporel. La figure 6.9 présente la distribution des employés stables en fonction de la durée de leur participation.

Cette figure permet de constater que près de 40 employés ont participé à des discussions pendant les 11 semestres à l'étude et sont restés dans la même communauté tout ce temps. Cette situation traduit donc d'un réel intérêt de ces individus envers les mises à

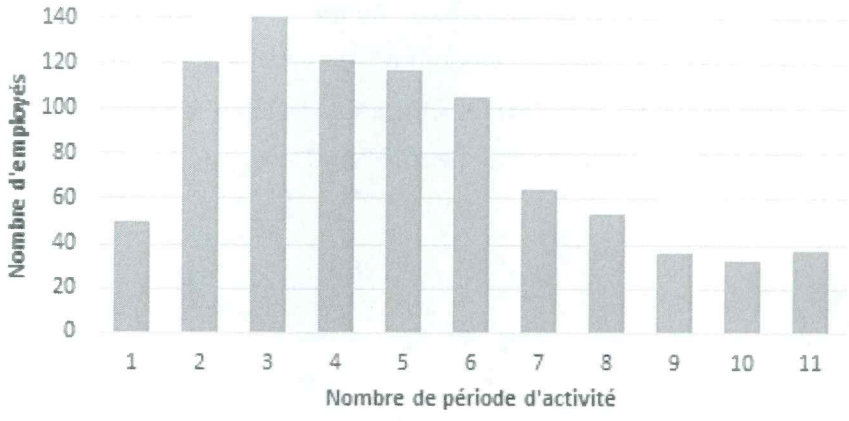


FIGURE 6.9 – Distribution des employés stables sur tout l'horizon temporel en fonction du nombre de semestres pendant lesquels ils ont participé aux mises à jour.

jour auxquelles ils ont contribué, ou le domaine de celles-ci ainsi que pour les relations créées avec les autres membres de leur regroupement. Ces employés sont dispersés dans 16 communautés différentes. Toutefois, les communautés *12* et *13* en regroupent respectivement cinq et huit chacune. Ces derniers passent donc toutes les périodes ensemble dans la même communauté, ce qui signifie qu'ils entretiennent des liens forts et durables. Une situation semblable survient également pour les 36 et 32 participants qui ont été actifs pendant respectivement neuf et dix périodes.

> Les employés semblent donc plutôt conservateurs quant à leurs relations au fil du temps ainsi qu'à leurs sujets de conversation. Certains semblent entretenir des liens profonds et ce, pendant cinq ans et demi. L'étude de leur influence ainsi que le rôle qu'ils exercent dans leur groupe respectif pourrait s'avérer révélatrice sur leur comportement et celui de leur équipe.

6.3.2 Changements de communautés

En revenant à la figure 6.5, il est possible de chercher à comprendre davantage ce qui s'est passé d'une période à l'autre quant aux mouvements des individus et leurs répercussions sur la formation et la composition des différentes communautés. D'abord, il est intéressant de poser un regard sur les mouvements qui ont eu lieu au fil du temps. La figure 6.10 présente justement le nombre de changements de communautés surve-

nus à chaque période, en plus du nombre de communautés quittées par les employés changeant d'affiliation et le nombre de communautés dans lesquelles ils se sont ajoutés. Le nombre de groupes présents chaque semestre est également illustré. Au total, 203 changements de communautés ont été effectués sur l'ensemble de l'horizon temporel.

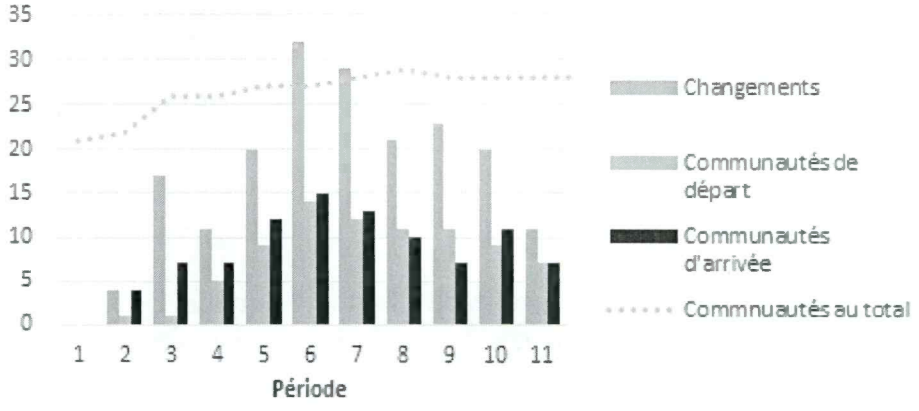


FIGURE 6.10 – Nombre d'employés qui changent de communauté à chaque période, en plus du nombre de communautés dans lesquelles ils se trouvaient initialement ainsi que le nombre de communautés auxquelles ils se joignent au total. Les mouvements sont représentés au semestre où l'appartenance de l'individu est différente de celle de la période précédente. La courbe pointillée exprime le nombre total de communautés présentes chaque semestre.

Ce graphique démontre clairement que les mouvements sont plus nombreux aux sixième et septième périodes, tout en s'élevant à 20 et plus du cinquième au dixième semestre. Toutefois, les participants sont plus nombreux durant ces périodes également, ce qui peut expliquer les changements plus fréquents. Lorsque le nombre de communautés d'arrivée est supérieur au nombre de communautés de départ des employés qui bougent, cela signifie que ces derniers entretenaient des liens plus serrés initialement pour ensuite se disperser chacun de leur côté ou presque. C'est le cas la majorité du temps, mais plus particulièrement aux deuxième et troisième semestres. Les mouvements sont alors explicables uniquement par le départ d'employés de la communauté *12*, soit celle ayant la plus grande taille jusque-là, comme illustré à la figure 6.5. Ces individus se divisent pour rejoindre respectivement quatre et sept communautés différentes au cours de ces périodes. Au troisième semestre, c'est près de 27% des individus de la communauté *12* qui la quitte. Les raisons de ces mouvements demeurent inconnues, en raison du peu d'informations dont nous disposons, mais leur étude pourrait

s'avérer intéressante pour l'entreprise. Aux huitième et neuvième périodes, c'est plutôt l'inverse qui se produit. En effet, le nombre de communautés quittées par les employés changeant d'affiliation est supérieur au nombre de communautés auxquelles ils se greffent. Dans ces cas, il s'agit de participants qui ont initialement des liens faibles entre eux, mais ces liens se fortifient et les employés se retrouvent finalement ensemble dans de mêmes communautés. Pour ce qui est du neuvième semestre, où la différence est marquée davantage, 13 des 23 employés changeant de communautés se rejoignent dans la communauté *26*, ce qui explique, en partie, l'ampleur que prend celle-ci à ce moment. En général, le nombre de communautés de départ et d'arrivée suit la même tendance que le nombre de changements, à l'exception de la troisième période, pour les raisons déjà mentionnées. Aux semestres plus mouvementés, le nombre de communautés touchées par les changements constitue une proportion élevée de l'ensemble des communautés présentes à ces moments. Il ne s'agit donc pas d'un phénomène isolé dans le réseau dans ces cas.

Il est possible de remarquer des périodes où les changements d'affinité sont plus fréquents. La cause de ceux-ci devrait intéresser les gestionnaires qui veulent mieux comprendre leur équipe. Certaines situations relatent d'un renforcement des relations entre certains employés qui quittent différents groupes pour se retrouver dans une même communauté ensemble. À l'inverse, une dégradation des relations est aussi observable à certains moments, où des employés quittent leur groupe pour se disperser dans des conversations différentes. Dans ce cas, les dirigeants devraient s'assurer qu'un conflit n'est pas en cause, car celui-ci pourrait engendrer des conséquences néfastes autant pour les employés que pour l'organisation [1].

6.3.3 Nouvelles participations

La figure 6.3 présentée plus tôt peut également être analysée plus en détail. Comme une grande partie des communautés naissantes se créent à partir de nouveaux participants, il est intéressant d'examiner ces arrivées avec la figure 6.11. Afin d'avoir une compréhension plus exacte de la situation, la figure 6.12 distingue les nouveaux participants qui reviennent de pause de ceux qui font leur première apparition sur le forum.

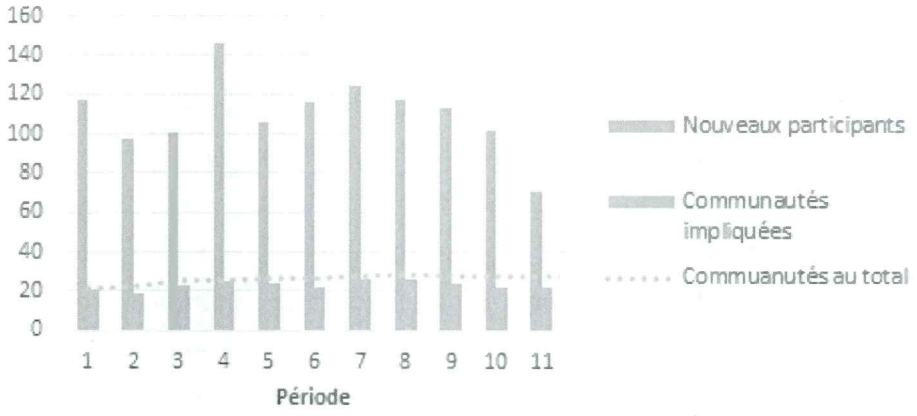


FIGURE 6.11 – Nombre de nouveaux participants, en plus du nombre de communautés auxquelles ils se joignent comparé au nombre total de communautés, et ce, à chaque période.

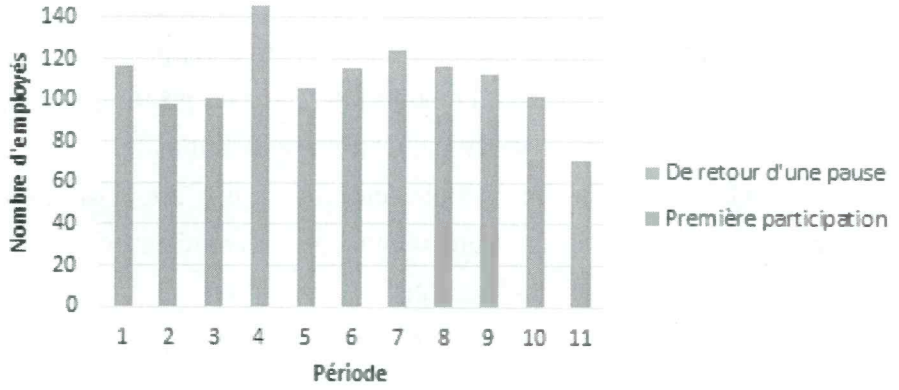


FIGURE 6.12 – Distribution des nouveaux participants à chaque période, selon s'ils reviennent de pause ou s'ils participent pour la première fois à une mise à jour.

À la lumière de ces graphiques, il est maintenant possible d'analyser les différents mouvements des individus. Nous remarquons notamment qu'une centaine de nouveaux participants se joignent aux discussions sur les mises à jour à chaque période. Au quatrième semestre, 145 employés s'ajoutent aux conversations, dont 34 se retrouvent dans la même communauté, soit la *0*, qui est celle ayant la plus grande taille à ce moment. Seulement cinq individus reviennent de pause à cette période, donc près de 97% des nouveaux participants en sont à leur première contribution. Au dernier semestre, les

arrivées sont moindres, d'où, notamment, la diminution des employés actifs au cours de celui-ci. De plus, la très grande majorité de celles-ci consistent en des individus revenant de pauses, plutôt que des employés qui participent pour la première fois aux mises à jour. Une grande partie des communautés accueillent de nouveaux participants, soit entre 19 et 25 selon les périodes, alors qu'il existe au maximum 29 communautés simultanément. Le rapprochement entre les bandes orangées et la courbe pointillée grise sur la figure 6.11 démontre ce phénomène. Les nouveaux participants sont donc répartis dans une grande proportion des communautés à leur arrivée. La figure 6.12 permet de constater que plus le temps avance, plus les nouveaux participants sont, en fait, des employés qui ont déjà contribué aux mises à jour, mais qui ont pris une pause et qui reviennent se joindre aux discussions. À partir du septième semestre, le nombre d'individus participant pour la toute première fois aux mises à jour est décroissant, et ce, jusqu'à la fin de l'horizon temporel. Il serait donc intéressant, pour l'entreprise, de déterminer si ces ajouts, qui se font plus rares durant ces périodes, sont attribuables à un manque d'intérêt des employés envers les discussions ou bien au peu d'individus qui n'ont encore jamais contribué aux mises à jour. Il est toutefois normal que les participants revenant de pauses se fassent plus nombreux d'un semestre à l'autre, en raison du temps nécessaire pour qu'ils participent une première fois aux discussions puis prennent ensuite une pause d'au moins une période.

La naissance de nouvelles communautés s'explique souvent par le rassemblement de nouveaux participants. C'est le cas pour la totalité d'entre elles, à l'exception de deux, soit une à la cinquième période, qui regroupe à la fois un nouveau participant et un ayant déjà contribué aux mises à jour et une autre au huitième semestre, qui se compose uniquement d'un individu ayant changé de communauté. En général, les nouvelles communautés rassemblent moins de dix employés. Il est intéressant de remarquer que les communautés *7* et *17*, qui resurgissent aux périodes cinq et sept respectivement, se composent uniquement de nouveaux participants. Alors que la première d'entre elles n'en compte que deux, dont un revenant de pause, l'autre en dénombre 12, dont un seul a déjà contribué aux mises à jour. Les deux participants de retour aux discussions rejoignent la même communauté qu'avant leur départ. Avec des informations supplémentaires sur les individus concernés et leur comportement dans l'entreprise, il serait pertinent de vérifier si les participants qui prennent des pauses encouragent des collègues à se joindre aux discussions pour une première fois pendant ce temps. À l'inverse, il pourrait aussi s'agir d'un individu manquant d'intérêt pour les discussions qui

décide de prendre une pause puis qui revient sur le forum en raison de ses collègues qui commencent à participer. Ces situations pourraient expliquer le regroupement de nouveaux participants avec un individu ayant pris une pause, au retour de ce dernier sur le forum.

> Le désir de se joindre aux discussions semble constant et plutôt répandu dans l'entreprise. L'étude des motivations des employés à participer aux échanges avec leurs collègues pourrait permettre aux gestionnaire de mieux comprendre les relations entretenues dans leur équipe. Par exemple, des relations d'influence pourraient exister et être à l'origine de tels comportements. Il est tout à l'avantage des dirigeants de connaître ces rapports et les leaders de leur équipe, s'ils veulent, par exemple, en tirer avantage dans leur gestion et lorsqu'il est temps de persuader leur équipe pour une situation quelconque.

6.3.4 Arrêts de participation

À l'inverse des communautés naissantes qui sont habituellement formées de nouveaux participants, les communautés qui meurent résultent souvent de l'arrêt de participation des employés qu'elles regroupaient. La figure 6.13 présente justement le nombre d'employés quittant les discussions, en plus du nombre de communautés qui les regroupaient comparativement au nombre total de communautés existantes à chaque période.

Les arrêts sont représentés à la première période où les employés concernés ne participent plus aux mises à jour. La figure 6.14 différencie les départs d'individus qui sont temporaires, représentant une pause, de ceux permanents, soit pour le reste de l'horizon temporel.

Bien que des participants s'ajoutent aux discussions au fil du temps, d'autres les quittent. Ce phénomène, présenté à la figure 6.13, est davantage marqué à partir de la septième période. Il faut rappeler que la participation moyenne des employés aux mises à jour se situe entre cinq et six semestres. Ainsi, les départs constatés à partir de cette période représentent l'arrêt de participation des individus qui ont commencé tôt à contribuer aux mises à jour. Malgré tout, des départs sont observés à chaque période, excepté la première évidemment. Les arrêts survenus au deuxième semestre signifient que les employés ont participé à la première période, sans continuer à la suivante. Aux cinquième et sixième semestres, les individus qui quittent les discussions proviennent de

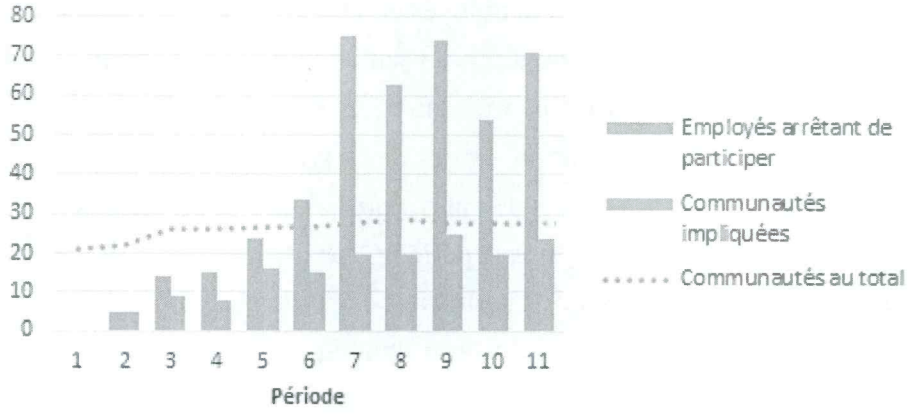


FIGURE 6.13 – Nombre d'employés arrêtant de participer, en plus du nombre de communautés desquelles ils partent comparé au nombre total de communautés, et ce, à chaque période.

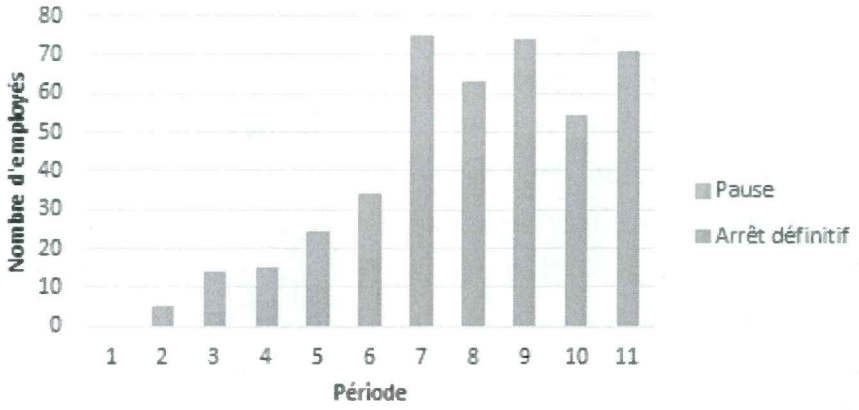


FIGURE 6.14 – Distribution des employés arrêtant de participer à chaque période, selon s'ils prennent une pause ou s'ils arrêtent définitivement leur contribution aux mises à jour.

16 et 15 communautés différentes respectivement. Ce nombre s'élève aux alentours de 20 par la suite. Cette situation permet de constater que les arrêts de participation ne sont pas un phénomène isolé dans quelques communautés, mais plutôt une tendance populaire dans tout le réseau. Le rapprochement entre les bandes orangées et la courbe pointillée grise dans la figure 6.13 dénote justement de la grande proportion de communautés concernées par ces départs. Parmi les arrêts de participation, certains

sont temporaires, c'est-à-dire que les employés concernés prennent une pause pendant au moins un semestre puis contribuent à nouveau aux mises à jour par la suite. Pour d'autres, il s'agit plutôt d'un arrêt permanent pour le restant de l'horizon temporel. La figure 6.14 présente la distribution de ces deux types d'arrêts. Il est possible de remarquer qu'initialement, le nombre de pauses prises et celui de départs définitifs effectués étaient plutôt équivalents à chaque période. À partir du septième semestre, cette tendance s'atténue, alors que les arrêts permanents sont de plus en plus fréquents comparativement à ceux temporaires, qui finissent même par être inexistants. Étant donné qu'aux premières périodes il y a moins de participants, il est normal que les départs soient plus rares. Aucune pause n'est représentée au dernier semestre, car les données pour la période suivante sont inconnues.

Comme présenté à la section 4.1.2, 118 individus ne participent pas aux discussions de façon continue sur l'ensemble de l'horizon temporel. Cependant, 128 pauses sont constatées au total des périodes, ce qui signifie que dix employés prennent une pause à deux reprises au cours de leur participation aux mises à jour. Parmi ces 128 arrêts, seulement 15 engendrent un changement de communautés des participants à leur retour. Ces 15 mouvements sont effectués par 14 employés différents, c'est-à-dire que l'un d'entre eux change d'affiliation à deux reprises. Ce phénomène démontre donc que les participants modifient rarement leurs intérêts ou comportements lorsqu'ils prennent une pause. Ainsi, il y a de grandes chances, soit plus de 88% selon cette partition, qu'à leur retour sur le forum, ils rejoignent la même communauté qu'avant leur départ. Cela signifie que ces individus ont un attachement particulier soit pour les liens qu'ils entretiennent avec les autres membres de leur groupe ou bien, pour le sujet des discussions ou le domaine des mises à jour auxquelles ils contribuent. Un parallèle peut également être réalisé avec le total de 203 changements survenant au cours de l'horizon temporel étudié. En effet, il est possible de remarquer que seulement 15 d'entre eux sont attribuables à des individus revenant de pause, ce qui ne représente que 7,39%. Ainsi, d'autres causes se cachent derrière la majorité des mouvements effectués par les employés, d'une communauté à l'autre, et seule l'entreprise pourrait investiguer plus loin à ce point de vue.

La mort des communautés est explicable en grande partie par le départ des individus qui les composaient. Comme présenté à la figure 6.2, la presque totalité des communautés perdure jusqu'au dernier semestre. Celles dont ce n'est pas le cas, soit les communautés *7, 17* et *28*, sont celles qui représentent les morts reportées à la figure 6.3. Par exemple,

les morts recensées aux semestres deux et cinq concernent le départ en pause du seul employé composant la communauté *17* à ces moments. Selon le même principe, le seul participant dans la communauté *7* a arrêté temporairement à la quatrième période, d'où la dissolution de celle-ci durant ce temps. Puis, la communauté éteinte au neuvième semestre, soit la *28*, a été formée à la période précédente par un individu venant d'un autre groupe, qui a mis fin à sa participation définitivement par la suite.

> Les gestionnaires devraient revoir le contexte de l'entreprise elle-même, les incitatifs à échanger avec les collègues, la présence de conflits dans leur équipe, la motivation et l'intérêt des employés pour les discussions ainsi que l'influence des membres de l'organisation pour la participation aux conversations, afin de mieux comprendre ce qui se cache derrière les arrêts de participation. Par ailleurs, ce phénomène est plutôt répandu dans l'entreprise et non pas isolé dans seulement quelques groupes, ce qui augmente d'autant plus l'importance de sa compréhension par les dirigeants.

6.3.5 Mouvements marquants

Certains phénomènes marquants reliés au comportement des employés influencent grandement la composition des communautés et méritent une attention particulière. À partir de la figure 6.5, nous pouvons relever quelques communautés dont la courbe se distingue de la tendance générale du graphique, ce qui signifie que certains évènements particuliers ont affecté leur existence. Une étude plus approfondie de ces évènements sera présentée ci-dessous. Les courbes représentant la taille des communautés qui seront analysées plus en détail sont présentées distinctement à la figure B.6 en Annexe.

La communauté *0* est sans aucun doute celle dont la taille varie le plus, au cours de l'horizon temporel, et ce, autant à la hausse qu'à la baisse. De la deuxième à la quatrième période, un nombre élevé d'employés se joignent à la communauté. Il s'agit en fait de 30 et 34 nouveaux participants qui s'ajoutent respectivement aux troisième et quatrième semestres ainsi qu'un individu ayant changé de communautés chaque fois. Le regroupement d'autant de nouveaux participants est plutôt élevé comparativement aux ajouts réalisés dans les autres communautés au même moment. Il serait intéressant pour l'entreprise de vérifier si ces individus entretenaient déjà certaines relations ensemble avant de joindre le forum. Malgré tout, avec les données fournies, il a été

possible de vérifier le bureau d'appartenance de chacun d'eux. À la troisième période, les nouveaux participants proviennent de six établissements différents alors qu'au semestre suivant, douze bureaux les regroupent. Dans les deux cas, une majorité travaille au même emplacement, mais ce dernier regroupe également une grande partie de l'ensemble des participants, comme présenté à la figure 4.3. Ainsi, aucune conclusion ne peut être tirée de cette situation. La communauté prend ensuite de l'ampleur jusqu'à la cinquième période, mais de manière beaucoup moins significative qu'auparavant. Elle atteint alors sa taille maximale de 78 employés pour ensuite en perdre quelques-uns au semestre qui suit. Puis, de la sixième à la septième période, un nombre important de participants quittent la communauté. En effet, sept individus se dispersant vers six communautés différentes, en plus de 37 départs relevés, dont 28 employés arrêtant définitivement leur participation ainsi que neuf prenant une pause. Parmi ces derniers, cinq reviennent dans la même communauté à la huitième période, un supplémentaire au semestre suivant puis un autre encore à la dernière période. Pour ce qui des deux autres ayant pris une pause, ils prennent part à des communautés différentes à leur retour. Cette vague massive d'arrêts de participation a de quoi retenir l'attention de l'entreprise. En raison du manque de données à notre disposition pouvant aider à expliquer ce phénomène, il serait intéressant pour la compagnie de chercher à comprendre ce qui a pu se passer à ce moment, pour qu'autant d'employés, représentant 3,44% du nombre total de participants et 49% de leur communauté avant leur départ, arrêtent de contribuer aux mises en jour en même temps.

Selon les analyses réalisées, il est possible de remarquer que ces individus participaient moins que la moyenne en termes du nombre de discussions auxquelles ils prenaient part, exception faite de la cinquième période pour les employés qui ont quitté temporairement seulement. Ce phénomène est illustré clairement à la figure 6.15.

Ces employés se sont joints au forum entre le troisième et le cinquième semestre. Cela signifie donc que les liens créés entre eux sont de durées différentes selon les cas. Les premiers individus à avoir quitté les discussions avaient déjà diminué leur contribution aux mises à jour au cours de la sixième période. Au contraire de l'hypothèse envisagée, soit que ces employés étaient des leaders et avaient engendré le départ des autres membres de leur communauté, il semble plutôt que les premiers à avoir cessé de participer ne détenaient pas un rôle central dans leur groupe.

Les mises à jour présentes dans la communauté *0* ont également été étudiées pour vérifier si elles pouvaient éclaircir le mouvement des employés qui s'y trouvent. La

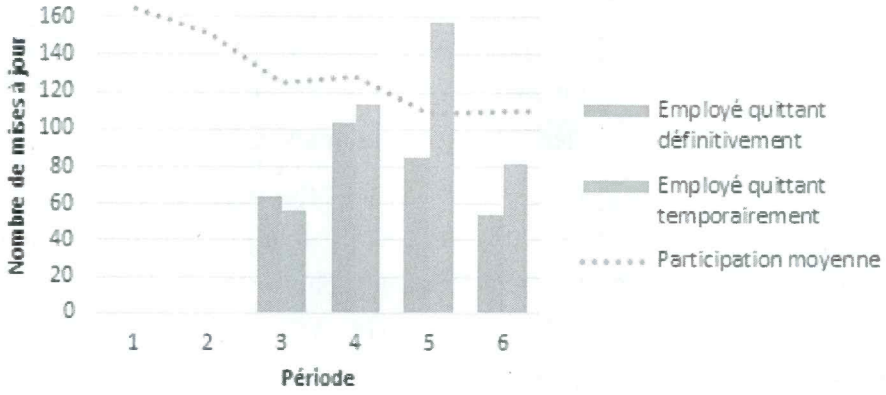


FIGURE 6.15 – Nombre de mises à jour auxquelles les employés, qui ont quitté la communauté *0* au septième semestre, de façon permanente ou temporaire, ont participé en moyenne à chaque période avant leur départ. La courbe en gris représente la moyenne de participation des autres employés du réseau.

tendance de la courbe représentant le nombre de mises à jour dans la communauté à chaque période est plutôt semblable à celle obtenue avec le nombre d'individus aux mêmes moments. Cependant, le nombre de mises à jour commence à chuter avant le départ des nombreux participants. Malgré tout, la diminution est la plus marquée du sixième au septième semestre, soit au même moment que la vague massive de départs des employés. En fait, les mises à jour sont de plus en plus nombreuses dans la communauté plus le temps passe et ce, jusqu'à la cinquième période où 4 814 s'y retrouvent. Par la suite, leur nombre chute significativement pour n'atteindre que 180 dès le septième semestre et moins de 50 au cours des trois dernières périodes. Au total, 9,18% des arêtes liées aux mises à jour de ce regroupement à tout moment vont vers l'extérieur de la communauté. Le nombre de liens se dirigeant vers des individus d'autres communautés varie à l'inverse du nombre total de liens touchant à la communauté au même moment. En effet, lorsque le nombre d'arêtes est faible, une grande partie d'entre elles sont inter-communautés, soit près de 50%. Dans la situation contraire, où un grand nombre de liens touchent à la communauté *0*, notamment de la troisième à la sixième période, entre 6% et 13% seulement de ceux-ci se dirigent hors de la communauté. Bref, même avec les mises à jour, cette communauté semble significative et rien de particulier ne peut expliquer la vague massive de départs au sixième semestre.

La mesure du ratio des arêtes ainsi que celle de l'expansion peuvent également être appliquées à la communauté *0* pour vérifier sa qualité. Le ratio des arêtes s'élève à 4,61 pour l'ensemble de la communauté, ce qui signifie qu'il existe près de cinq arêtes intra-communauté pour chaque lien inter-communautés qu'elle comprend. Ce ratio est acceptable. L'expansion atteint pour sa part 0,47, ce qui implique que chaque sommet de la communauté *0* entretient moins d'un lien en moyenne avec un nœud n'appartenant pas au même groupe. Cette mesure peut être analysée séparément pour chaque type de sommets. En fait, elle s'élève à 0,22 en ne considérant que les nœuds de mises à jour et à 10,00 en ne tenant compte que de ceux d'employés. Il est normal que le nombre d'arêtes inter-communautés liées aux individus soit supérieur, en raison de leur degré moyen qui est plus élevé également. Malgré tout, l'expansion générale de la communauté est faible, ce qui laisse croire à une bonne qualité.

En raison des résultats observés, il était quand même important de vérifier si le regroupement de ces 37 individus dans la même communauté avait été engendré en raison de la méthode de résolution retenue, qui aurait considéré leur départ simultané lors de la formation de leur communauté. Pour ce faire, tous les sommets et liens du réseau ont été éliminés après la sixième période, soit à partir du moment où les nombreux arrêts de participation sont survenus. Ce graphe modifié a été résolu selon la même approche que précédemment. La partition trouvée fournit un regroupement semblable de ces employés. En fait, 34 d'entre eux se retrouvent dans la même communauté au sixième semestre alors que les trois autres sont répartis dans deux communautés différentes. Cette situation prouve donc que le départ au même moment de ces 37 employés ne jouait pas une influence majeure dans la formation des communautés lors de la résolution initiale.

Un autre impact de l'algorithme utilisé, en raison de son approche gloutonne, peut être de regrouper quelques communautés ensemble afin d'en former une plus grande, alors que cette dernière est moins représentative de la réalité. Pour vérifier que la communauté *0* n'ait pas été formée ainsi, sa composition a été testée. Pour ce faire, seulement les liens et sommets appartenant à cette communauté, sur l'ensemble de l'horizon temporel, ont été conservés, en prenant soin d'éliminer les liens inter-communautés touchant à cette dernière. La même approche de résolution a encore une fois été appliquée sur cet autre graphe modifié. À partir de la partition trouvée, il était possible de vérifier si les 37 employés, quittant la communauté *0* en même temps, se divisaient en plusieurs communautés distinctes, en guise de biais dans leur regroupement initial. Les résultats obtenus démontrent bien que leur appartenance à une même communauté

dans la partition originale est justifiée. En effet, à la sixième période, soit avant leur retrait du forum, quatre communautés sont présentes et l'une d'elle comprend 33 des 37 employés étudiés. Ainsi, la communauté initiale *0* ne semble pas avoir été formée en raison de l'approche gloutonne de l'algorithme utilisée, mais plutôt, car ces participants qui quittent les discussions simultanément constituent véritablement un tout.

À ce point, la recherche de la cause de cette vague massive de départs simultanés dépasse le cadre de ce mémoire et nécessiterait des informations supplémentaires de la part de l'entreprise.

Mis à part la communauté *0*, qui se distingue grandement par sa composition au fil du temps, quelques autres communautés méritent également une attention particulière. Celles-ci, toutes représentées à la figure B.6 en Annexe, sont impactées, à un moment ou à un autre, par un mouvement important de participants. En y allant dans l'ordre de la figure, la communauté *5* subit une augmentation marquée du nombre d'employés qui la joignent du quatrième au sixième semestre. En effet, à la cinquième période, onze nouveaux participants s'ajoutent à la communauté en plus d'un employé venant d'un autre groupe. Au semestre suivant, c'est dix nouveaux participants qui y adhèrent, en plus de cinq individus venant tous de la même communauté. Il est à se questionner si ces nouveaux participants étaient déjà en relation avant leur arrivée sur le forum.

La communauté *12* a plutôt un comportement à l'inverse de la tendance générale. En effet, initialement, il s'agit de la plus grande communauté, elle regroupe déjà 64 employés au deuxième semestre. Cela représente alors plus de 30% du nombre total de participants à ce moment. Cependant, cette communauté commence à se décomposer dès la période suivante, et ce, jusqu'à la fin de l'horizon temporel. En effet, 17 employés la quittent au troisième semestre pour se disperser dans sept communautés différentes, en plus de trois autres qui arrêtent de participer définitivement au même moment. À la période suivante, six individus de la communauté se dirigent vers quatre autres communautés distinctes alors que six autres arrêtent de participer, dont deux temporairement seulement. Au cinquième semestre, encore six changements sont constatés vers trois communautés différentes, en plus de deux départs définitifs. Bref, le nombre d'employés qui quittent la communauté *12* diminue avec le temps, mais celle-ci continue tout de même à décroitre jusqu'au dernier semestre. Il serait intéressant pour l'entreprise de vérifier ce qui a enclenché les nombreux départs à partir de la deuxième période et ce qui les a maintenus jusqu'à la toute fin. Il semble réellement se passer quelque chose dans cette communauté pour que les liens se dégradent alors que l'inverse se produit

dans la majorité des autres groupes.

La communauté *13*, pour sa part, connait une croissance lente au cours des premiers semestres, mais sa taille augmente significativement à la septième période. À ce moment, 26 nouveaux participants se joignent à elle, en plus de trois employés qui y étaient déjà plus tôt, mais qui avaient pris une pause et y reviennent. Cette situation amène à se questionner à ce qui a déjà été énoncé plus tôt, soit à savoir si les individus qui prennent des pauses créent des liens avec des collègues pendant ce temps et les encouragent à participer aux mises à jour avec eux à leur retour, ou encore s'ils recommencent à participer en raison de l'arrivée de leurs collègues sur le forum. Il serait pertinent pour l'entreprise d'investiguer à ce point de vue, car si ce comportement constitue une tendance générale des employés qui s'arrêtent temporairement, il permet réellement d'augmenter le niveau de participation aux discussions.

La taille de la communauté *16* varie généralement peu, excepté à la quatrième période où elle croît rapidement. À ce moment, 15 nouveaux participants viennent s'y ajouter. Lorsque de nombreux employés faisant leur première intervention au même moment dans les discussions se retrouvent dans la même communauté, il est intéressant de se demander si ces derniers entretenaient déjà certaines relations avant leur arrivée. L'entreprise est la mieux placée pour répondre à cette question, étant donné qu'aucune information ne nous a été fournie quant à la situation réelle derrière les données.

La composition de la communauté *20* varie de manière semblable à celle de la communauté *0*, mais de façon moins marquée. Le nombre de participants qu'elle regroupe augmente au cours des premiers semestres jusqu'à la cinquième période. À ce moment, 15 nouveaux participants s'ajoutent, en plus de quatre employés changeant de communautés et d'un autre y appartenant plus tôt et qui revient de pause. Tout comme la communauté *0*, le nombre d'individus qui s'y trouvent commence à chuter après le sixième semestre. Les départs se font toutefois moins nombreux à chaque période. La communauté se désagrège donc plus lentement. Il serait toutefois intéressant de vérifier si la tendance observée dans l'évolution de la composition de ces deux communautés, soit la *0* et la *20*, s'explique pareillement dans les deux cas.

La taille de la communauté *23* croît d'une période à l'autre tout au long de l'horizon temporel. Cependant, une augmentation plus importante est remarquée au quatrième semestre. En effet, à ce moment, 16 nouveaux participants se joignent à ceux qu'elle regroupe déjà. Cette situation va donc dans le même sens des questionnements posés

avec les communautés *5* et *16*, à savoir si des employés peuvent développer des liens avant de rejoindre le forum.

La communauté *26* est celle qui termine le dernier semestre avec la plus grande taille en termes du nombre d'employés qu'elle comprend. Alors qu'initialement elle ne comportait que très peu de participants, sa croissance s'est amorcée plus significativement à partir de la cinquième période. Une arrivée massive d'individus est toutefois remarquée au neuvième semestre, alors que 19 nouveaux participants s'y ajoutent, dont un seul revient de pause.

Bref, toutes ces communautés évoluent de manière particulière et leur étude, en lien avec la situation réelle, pourrait véritablement être révélatrice et bénéfique pour l'entreprise.

Tous ces mouvements marquants méritent une attention particulière de la part des gestionnaires, si ces derniers veulent démystifier le comportement des membres de leur organisation. Ils représentent en fait, les situations et employés qui semblent avoir marqués davantage la dynamique de leur équipe. Des vagues de départs massives aux arrivées nombreuses dans un même groupe simultanément, en passant par l'influence des employés en pause ou de l'influence qu'ils subissent, aux risques de conflits et aux comportements variables de certains employés, tous ces événements marquent la composition des équipes actuelles de l'entreprise ainsi que le comportement de leurs membres, d'où l'importance de bien les comprendre afin d'ajuster la gestion en conséquence.

6.4 Caractéristiques des liens

Comme présenté à la section 3.2, il faut idéalement le plus de liens possible à l'intérieur d'une même communauté alors qu'une plus faible concentration est souhaitée d'une communauté à l'autre. Sur cette base, nous pouvons donc étudier les liens de la partition trouvée.

Le graphe étudié comporte au total 795 791 arêtes entre les individus et les mises à jour. Dans la partition trouvée, 153 954 d'entre elles, représentant 20, 1% du total, lient des sommets de communautés différentes, elles sont donc inter-communautés. En généra-

lisant le ratio des arêtes présenté à la section 3.2, à l'ensemble du graphe, il est possible de trouver qu'il s'élève à environ 4,17. Cela signifie donc qu'en moyenne, les sommets de l'ensemble des communautés entretiennent environ quatre liens intra-communauté pour chacune de leurs arêtes inter-communautés.

Afin de mieux comprendre pourquoi les nœuds reliés ensemble ne se trouvent pas toujours dans la même communauté, l'analyse a été poussée pour les principaux concernés. Elle a révélé que 983 participants sur le total des 1045 à l'étude ont au moins un lien inter-communautés avec une mise à jour. C'est donc dire qu'ils ne se retrouvent pas dans la même communauté d'au moins une mise à jour à laquelle ils ont contribué. En fait, c'est le cas pour 4 918 des sommets représentant les employés à chaque période, soit plus de 85% du nombre total de ce type de nœuds. Chacun d'eux est relié en moyenne à 151 mises à jour, dont 31, soit 20,53%, se trouvent à l'extérieur de leur communauté. De plus, ces individus sont plus actifs en moyenne que l'ensemble du réseau en participant à environ 151 discussions par période comparativement à la moyenne générale trouvée précédemment, qui s'élève à 133. Malgré tout, avec autant de liens, il serait difficile que toutes les mises à jour soient regroupées dans la même communauté que chacun des nœuds y participant, tout en s'assurant que les communautés trouvées soient significatives.

Dans le même ordre d'idées, 110 060 mises à jour, soit près de 34% de toutes celles analysées, entretiennent au moins un lien inter-communautés. En les étudiant davantage nous avons soulevé que celles-ci étaient discutées en moyenne par 2,84 employés, soit légèrement plus que la moyenne générale de 2,34. De plus, environ 1,39 lien, soit 48,78% de leurs liens au total, se rendent dans une autre communauté. C'est donc dire que près de la moitié des participants à ces mises à jour se trouvent dans une autre communauté que celles-ci. La figure 6.16 représente bien l'évolution du nombre d'arêtes total reliant les mises à jour ainsi que de la proportion d'entre elles qui passent d'une communauté à l'autre.

Ce graphique montre que le nombre de liens inter-communautés suit une tendance semblable au nombre de liens total touchant aux mises à jour, tout en étant moins nombreux. La figure 6.17 illustre plus clairement la proportion des arêtes touchant aux mises à jour qui sont reliées à un individu d'une autre communauté. Cette proportion s'élève aux alentours de 20% du deuxième au dernier semestre, en variant entre 17,19% et 22,67%. Elle est cependant plus faible à la première période en atteignant près de 10% à ce moment. Comme le nombre de participants est moindre initialement et que le nombre de

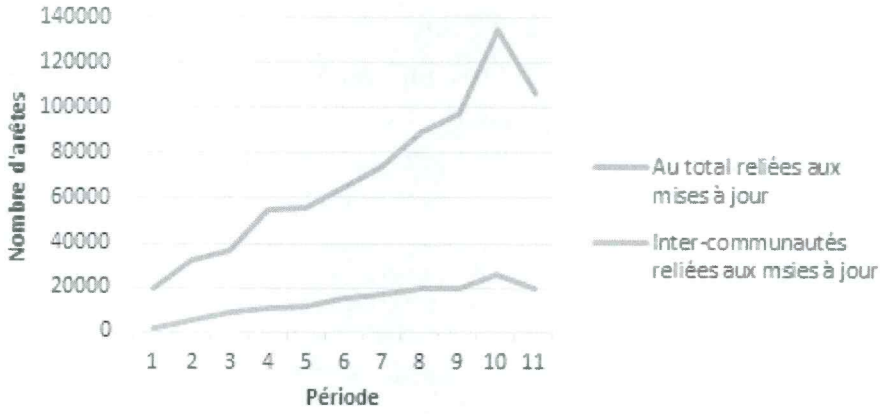


FIGURE 6.16 – Nombre total d'arêtes reliées aux mises à jour à chaque période, en plus du nombre total de liens inter-communautés touchant aux mises à jour aux mêmes moments.

communautés est relativement grand, comparativement au même rapport aux autres périodes, il semble normal que la proportion des liens qui sont inter-communautés soit moindre. Bref, les liens inter-communautés, qui constituent heureusement qu'une mince partie de l'ensemble des arêtes du réseau, semblent plutôt réalistes, car elles relient des employés plus actifs que la moyenne à des discussions plus populaires que la moyenne.

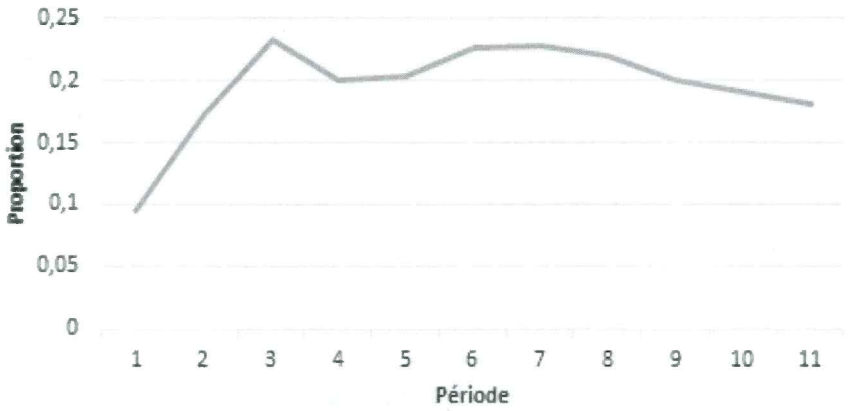


FIGURE 6.17 – Proportion des arêtes reliées aux mises à jour qui sont inter-communautés à chaque période.

Des affinités sont réellement présentes entre certains employés, ce qui expliquent leurs échanges quatre fois plus nombreux entre eux qu'avec les autres collègues. Les groupes formés se distinctent bien les uns des autres.

Chapitre 7

Conclusion et avenues de recherche

L'objectif de ce mémoire était de trouver une approche de résolution permettant la détection de communautés dans un réseau pondéré et évolutif comportant des données hétérogènes. La partition trouvée devait également permettre l'analyse des différents évènements marquant l'évolution des communautés et les mouvements des employés. Pour y arriver, trois modèles différents ont été testés en fonction de certains paramètres établis, permettant chacun la détection de communautés de qualité. L'un d'entre eux s'est démarqué par les résultats qu'il engendrait, au niveau du nombre de communautés trouvées, de la taille et la durée de vie de celles-ci, ainsi que de l'appartenance des individus. Il semblait le plus réaliste en raison de la représentation des individus aux périodes où ils ont participé seulement, en plus de limiter les changements au retour de pause des employés, en associant leur présence sur le forum avant et après l'arrêt. Une pondération a ensuite été sélectionnée, selon l'importance accordée aux deux types de liens dans son graphe, de manière à analyser en détail la partition qu'elle engendrait. Aucune attente n'était envisagée quant aux résultats en raison de la méconnaissance de la situation réelle sous-jacente aux données. Cependant, l'analyse s'est avérée très intéressante et a permis de soulever certaines tendances et quelques phénomènes importants quant à la participation des employés et à leur mouvement d'une communauté à l'autre au fil du temps. Des informations additionnelles de la part de l'entreprise seraient toutefois nécessaires afin de mieux comprendre et justifier certains des faits saillants relevés. Les expérimentations effectuées ont donc permis de répondre convenablement à la problématique posée initialement, en modélisant adéquatement le jeu de données, de manière à identifier des communautés dynamiques qui s'y trouvaient et à

pouvoir les analyser par la suite.

Il est important de se rappeler qu'il s'agit d'une nouvelle approche pour le jeu de données étudié. Effectivement, celui-ci avait été analysé, antérieurement, quant à la performance de différents algorithmes de la littérature, en utilisant une approche indépendante sur les instantanés. La modélisation proposée dans ce mémoire a permis de relever les mêmes phénomènes et tendances que ce qui avait été obtenu avec cette approche. Cependant, elle a également permis de constater des phénomènes additionnels, comme la vague massive de départs observée à la sixième période, qui n'avait pas été relevée dans les études précédentes.

Le modèle utilisé présente également plusieurs autres avantages. En plus de définir des communautés à travers le temps, il permet une plus grande stabilité de celles-ci, comparativement à une résolution indépendante sur les instantanés. Il permet aussi d'apparier automatiquement les communautés de différentes périodes par la résolution simultanée de tous les instantanés. Il accélère ainsi le processus de détection en évitant le post-traitement une fois les communautés trouvées. De plus, cette approche s'est avérée robuste au niveau de son extension. Effectivement, si une période est ajoutée et que le graphe est résolu à nouveau, les partitions trouvées initialement ne devraient pas changer significativement.

Les différents évènements externes des communautés, proposés par Spiliopoulou [8] et présentés à la section 3.1, peuvent également être identifiés en utilisant l'approche de ce mémoire. La survie des communautés est donnée directement par la résolution, alors que leur naissance et leur mort peuvent facilement être obtenues. La fusion et la division peuvent aussi être étudiées, mais requièrent une analyse plus poussée. En raison du modèle utilisé, la fusion se distingue par une communauté qui disparait et dont la majorité des sommets se retrouvent dans une autre communauté par la suite. À l'inverse, la division se caractérise par une communauté, dont une certaine proportion des nœuds la quittent pour en former une nouvelle. Tous les évènements sont donc repérables, bien que certains nécessitent un peu plus de travail.

Une des limites de cette approche concerne toutefois la taille du graphe qui est créé. Comme les sommets sont dupliqués pour chacune des périodes où les individus ont été actifs, leur nombre total ainsi que le nombre de liens dans le graphe sont multipliés. Dans le cas de la base de données étudiée, seulement 1045 nœuds devaient être dupliqués, et cela sur onze périodes, ce qui était plutôt raisonnable. Cependant, avec un jeu

de données de plus grande taille, ou ayant davantage de sommets à dupliquer, l'impact sur la taille aurait pu être plus important en nuisant, du même coup, à la résolution.

Malgré tout, cette approche de résolution pour les réseaux évolutifs ouvre la porte à de nombreuses questions managériales. La détection de communautés dans le réseau créé à partir de la base de données, dont la réalité sous-jacente nous était inconnue, s'est pratiquement effectuée à l'aveugle. Toutefois, des phénomènes intéressants quant aux comportements des individus ont été dégagés et pourraient être utiles aux dirigeants de l'entreprise, afin de mieux comprendre la dynamique de leur équipe et ainsi, ajuster leur gestion en conséquence. Bref, cette approche de résolution, qui n'a été que très peu testée malgré tout, s'annonce plutôt prometteuse.

Évidemment, plusieurs avenues différentes pourraient être envisagées afin d'expérimenter davantage l'approche de résolution utilisée dans ce mémoire. En effet, plusieurs décisions ont été prises au cours du processus et ont sans doute influencé les partitions trouvées. Ainsi, l'impact de décisions différentes pourrait être intéressant à évaluer. Il est d'abord question de la détermination des poids sur les divers liens du réseau. Il pourrait être pertinent de fixer les poids entre les employés et les mises à jour en fonction du nombre de commentaires effectués dans la même discussion. Ainsi, des individus contribuant à plusieurs reprises à une même mise à jour auraient peut-être plus de chance de se retrouver ensemble dans une communauté que deux autres ayant participé qu'une seule fois chacun. Les poids attribués reflèteraient donc davantage cette situation. Également, des poids différents pourraient être établis sur les liens entre les employés et eux-mêmes, plutôt que de les fixer tous au même niveau, avant l'ajustement pour la représentation des pauses. La pondération pourrait, par exemple, considérer le nombre de mises à jour auxquelles l'individu a participé à chaque période. Ainsi, un poids élevé serait établi sur le lien reliant un employé très actif au cours de deux périodes consécutives alors qu'un poids plus faible serait attribué pour un individu peu actif. Une règle de détermination de ces poids en fonction des degrés des sommets des participants pourrait donc être développée et testée. De cette manière, l'équilibre entre les poids attribués à chacun des types de liens serait visé.

Un autre aspect pouvant être étudié davantage consiste à l'algorithme de résolution. Pour ne pas dépasser le cadre de ce mémoire, les choix standards de la modularité et l'algorithme de **Louvain** ont été effectués. Cependant, il serait intéressant d'expérimenter la résolution avec d'autres algorithmes, qui seraient tout aussi adaptés au type du réseau étudié. Comme l'algorithme choisi ne renvoie pas toujours la même partition

pour un même graphe, les résultats pourraient être comparés à ceux obtenus autrement.

De plus, les modèles développés ont été testés sur une seule base de données. Leur application dans un autre contexte, avec des données différentes, mais comportant les mêmes caractéristiques quant à l'aspect temporel et hétérogène du réseau qu'elles permettent de former, serait des plus pertinentes. Il serait alors possible de vérifier si elles se prêtent à différentes tailles de réseaux et à divers domaines tout en permettant des résultats aussi intéressants que dans ce mémoire.

Bref, l'approche de résolution utilisée dans ce mémoire ouvre la porte sur de nombreuses autres expérimentations. Elle sera possiblement bien utile pour des gestionnaires de différents domaines qui souhaitent mieux comprendre la dynamique de leur équipe.

Annexe A

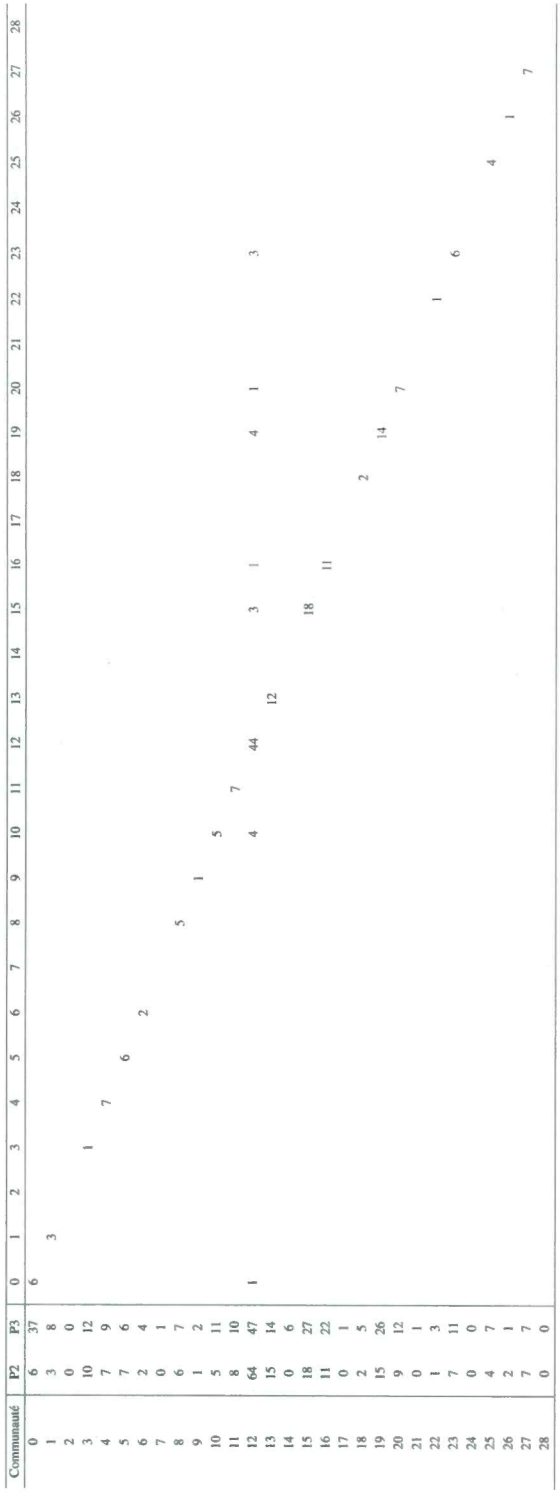
Tableaux

	Modèle 1		Modèle 2		Modèle 3	
Poids	$N_{\mathscr{C}}$	$mod(\mathscr{C})$	$N_{\mathscr{C}}$	$mod(\mathscr{C})$	$N_{\mathscr{C}}$	$mod(\mathscr{C})$
0,1	121	0,9000	34	0,9009	35	0,8988
0,2	117	0,9035	30	0,8993	37	0,8945
0,3	107	0,9024	27	0,8983	33	0,8948
0,4	94	0,8998	30	0,9000	34	0,8982
0,5	116	0,8984	31	0,8963	34	0,8933
0,6	117	0,8950	30	0,8951	36	0,8947
0,7	109	0,9016	32	0,9000	34	0,8931
0,8	91	0,8977	30	0,8940	33	0,8943
0,9	95	0,8961	30	0,8992	35	0,8954
1	101	0,8922	33	0,8907	36	0,8862
2	56	0,8415	28	0,8495	34	0,8422
3	42	0,8260	32	0,8145	37	0,8218
4	41	0,8038	29	0,8041	35	0,7907
5	30	0,7959	30	0,7915	36	0,7842
6	31	0,7902	29	0,7827	34	0,7728
7	35	0,7821	32	0,7822	34	0,7728
8	38	0,7821	29	0,7793	33	0,7773
9	41	0,7807	28	0,7793	34	0,7741
10	42	0,7823	30	0,7731	36	0,7674
25	50	0,7977	29	0,7685	34	0,7660
50	78	0,8258	33	0,7842	39	0,7779
75	87	0,8482	39	0,7957	46	0,7927
100	110	0,8642	40	0,8115	55	0,8075
125	106	0,8780	47	0,8244	50	0,8231
150	112	0,8892	57	0,8350	67	0,8335
175	139	0,8986	58	0,8427	70	0,8419
200	155	0,9049	57	0,8501	74	0,8508
1000	317	0,9679	176	0,9384	181	0,9383

TABLEAU A.1 – Nombre de communautés et modularité obtenues pour la meilleure résolution de l'algorithme de **Louvain** pour chaque modèle avec différents poids testés.

Communauté	P1	P2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	2	6	1																												
1	1	3		1																											
2	0	0																													
3	3	10				3																									
4	3	7					3																								
5	4	7						4																							
6	1	2							1																						
7	0	0																													
8	2	6									2																				
9	1	1										1																			
10	2	5											2																		
11	3	8												3																	
12	57	64												1	52							1				1					
13	11	15														1												1			
14	0	0																													
15	3	18																3													
16	0	11																													
17	1	0																													
18	3	2																			2										
19	9	15																				9									
20	3	9																					3								
21	0	0																													
22	1	1																							1						
23	2	7																								2					
24	0	0																													
25	1	4																										1			
26	0	2																													
27	4	7																												4	
28	0	0																													

TABLEAU A.2 – Mouvements des employés d'une communauté à l'autre du premier au deuxième semestre.

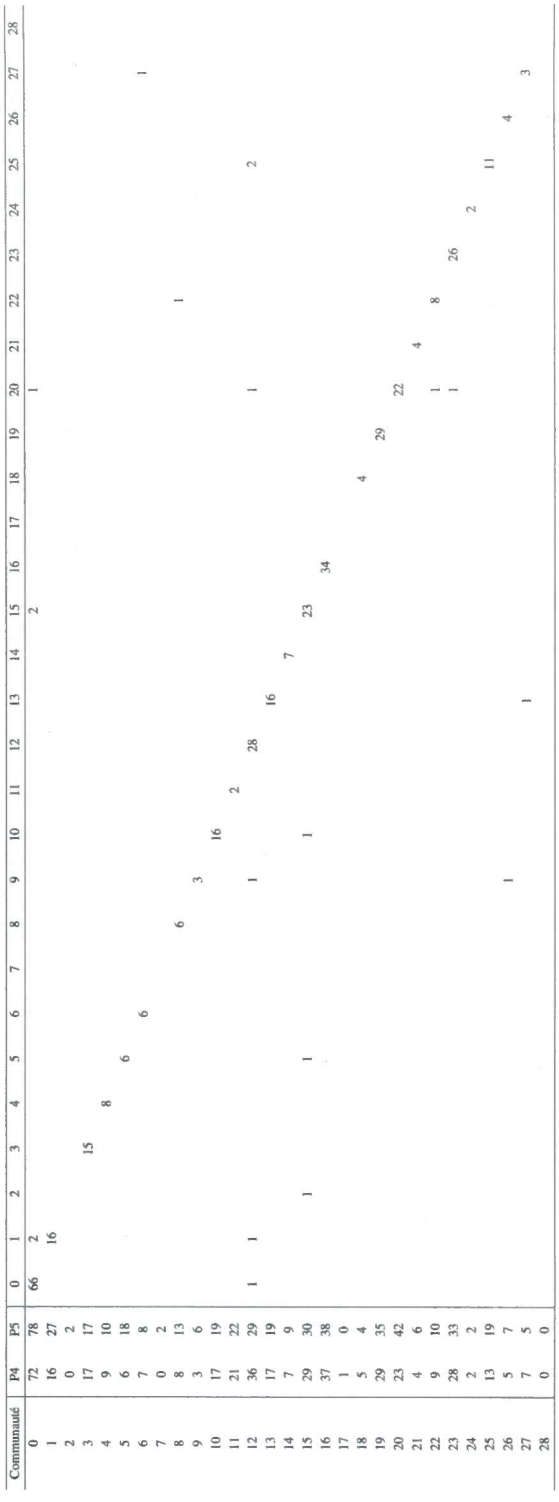


Communauté	P2	P3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	6	37	6																												
1	3	8		3																											
2	0	0																													
3	10	12				1																									
4	7	9					7																								
5	7	6						6																							
6	2	4							2																						
7	0	1																													
8	6	7									5																				
9	1	2										1																			
10	5	11											5																		
11	8	10												7																	
12	64	47	1										4		44			3	1			4	1			3					
13	15	14														12															
14	0	6																													
15	18	27																18													
16	11	22																	11												
17	0	1																													
18	2	5																			2										
19	15	26																				14									
20	9	12																					7								
21	0	1																													
22	1	3																							1						
23	7	11																								6					
24	0	0																													
25	4	7																										4			
26	2	1																											1		
27	7	7																												7	
28	0	0																													

TABLEAU A.3 – Mouvements des employés d'une communauté à l'autre du deuxième au troisième semestre.

Communauté	P3	P4	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	37	72	37																												
1	8	16		8																											
2	0	0																													
3	12	17				11																									
4	9	9					8																								
5	6	6						5																							
6	4	7							4																						
7	1	0																													
8	7	8									6																				
9	2	3										2																			
10	11	17											11																		
11	10	21												1																	
12	47	36													35																
13	14	17														14						1				1					
14	6	7															6														
15	27	29	1															24							1						
16	22	37																	22												
17	1	1																		1											
18	5	5																			4										
19	26	29																				22									
20	12	23																					12								
21	1	4																						1							
22	3	9																							3						
23	11	28																								11					
24	0	2																													
25	7	13																										7			
26	1	5																											1		
27	7	7								1																				6	
28	0	0																													

TABLEAU A.4 – Mouvements des employés d'une communauté à l'autre du troisième au quatrième semestre.

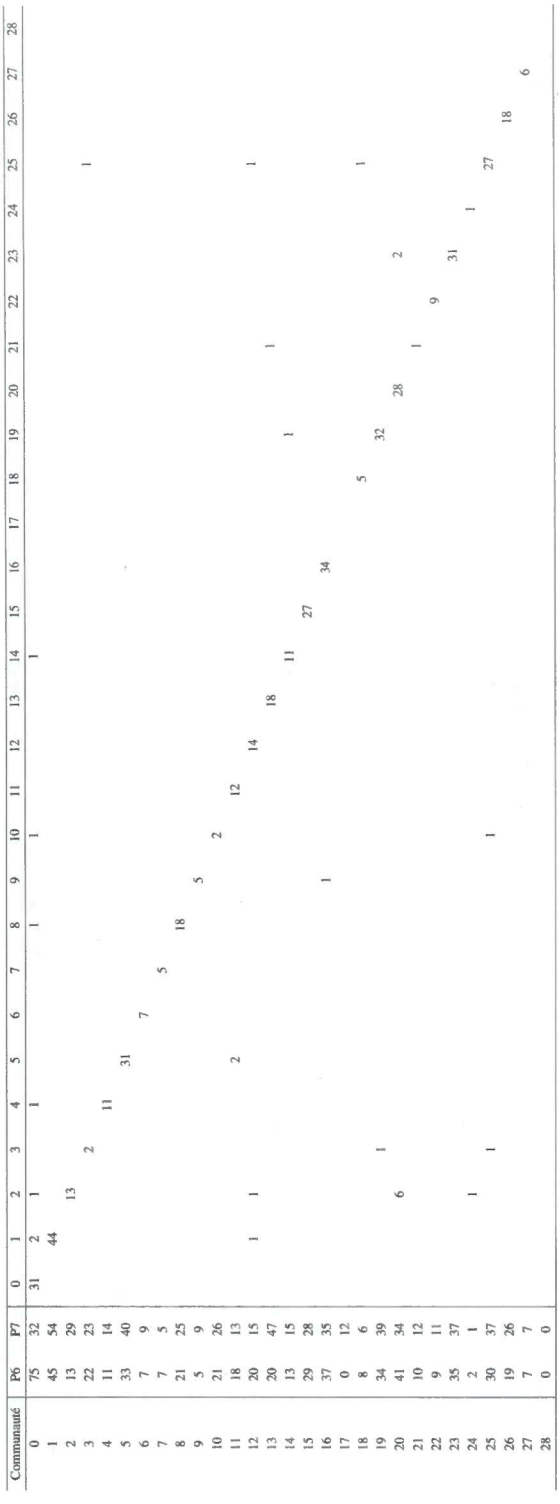


Communauté	P4	P5	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	72	78	66	2														2					1								
1	16	27		16																											
2	0	2																													
3	17	17				15																									
4	9	10					8																								
5	6	18						6																							
6	7	8							6																						
7	0	2																													
8	8	13									6														1						
9	3	6										3																			
10	17	19											16																		
11	21	22												2																	
12	36	29	1	1								1			28								1					2			
13	17	19														16															
14	7	9															7														
15	29	30			1			1					1					23													
16	37	38																	34												
17	1	0																			4										
18	5	4																				29									
19	29	35																					22								
20	23	42																						4							
21	4	6																					1		8						
22	9	10																					1			26					
23	28	33																									2				
24	2	2																										11			
25	13	19										1																	4		
26	5	7														1														3	
27	7	5																													
28	0	0																													

TABLEAU A.5 – Mouvements des employés d'une communauté à l'autre du quatrième au cinquième semestre.

Communauté	P5	P6	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	78	75	74	1																											
1	27	45		26																											
2	2	13			1																										
3	17	22				16																									
4	10	11					1																								
5	18	33						18																							
6	8	7							5																						
7	2	7								2																					
8	13	21	1								1																				
9	6	5										5																			
10	19	21											18																1		
11	22	18												15			1														
12	29	20		2				5	1						2																
13	19	20														18															
14	9	13															8														
15	30	29																26				1									
16	38	37											1						32												
17	0	0																													
18	4	8																		4											
19	35	34			1						2			1								28									
20	42	41			1																	1	34			1		1			
21	6	10															1							5							
22	10	9																							9						
23	33	35									1										1					28		1			
24	2	2																									2				
25	19	30				1																						18			
26	7	19																											7		
27	5	7																												4	
28	0	0																													

TABLEAU A.6 – Mouvements des employés d'une communauté à l'autre du cinquième au sixième semestre.



Communauté	P6	P7	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	75	32	31	2	1		1				1		1				1														
1	45	54		44																											
2	13	29			13																										
3	22	23				2																						1			
4	11	14					11																								
5	33	40						31																							
6	7	9							7																						
7	7	5								5																					
8	21	25									18																				
9	5	9										5																			
10	21	26											2																		
11	18	13						2						12																	
12	20	15		1											14																
13	20	47			1											18															
14	13	15															11														
15	29	28																27													
16	37	35										1							34												
17	0	12																			5										
18	8	6																				32						1			
19	34	39				1																	28								
20	41	34			6																			1		2					
21	10	12																							9						
22	9	11																								31					
23	35	37																									1				
24	2	1			1																							27			
25	30	37				1							1																18		
26	19	26																												6	
27	7	7																													
28	0	0																													

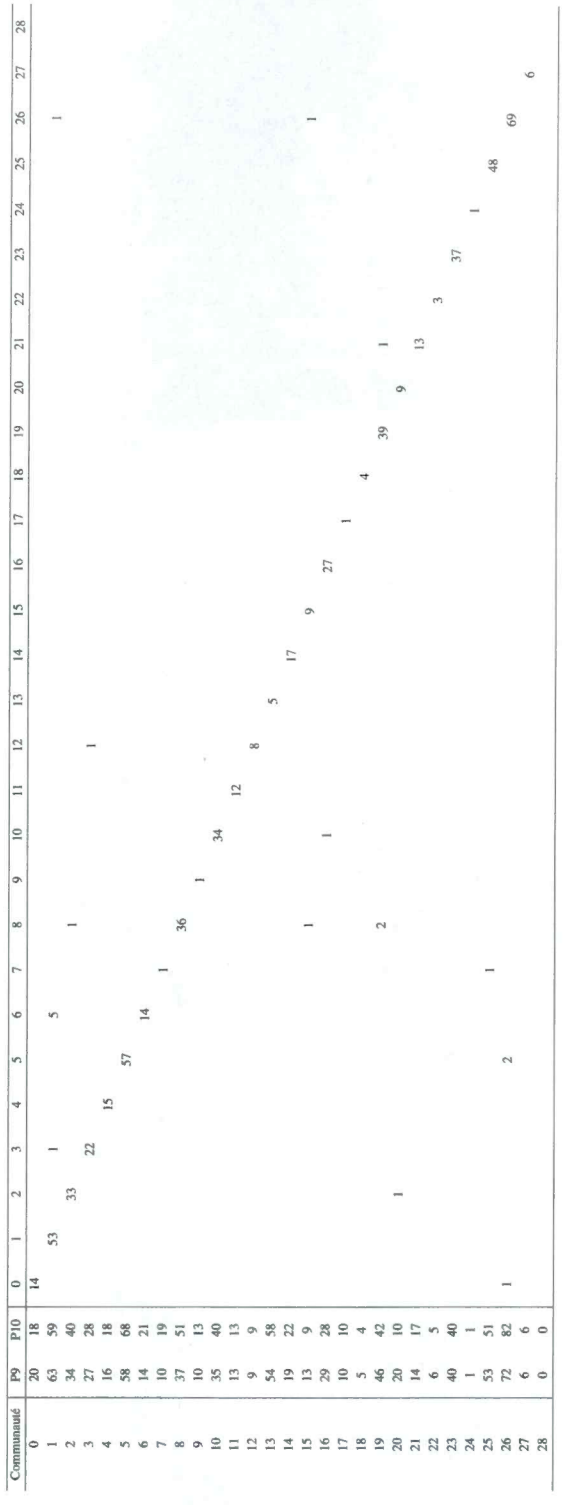
TABLEAU A.7 – Mouvements des employés d'une communauté à l'autre du sixième au septième semestre.

Communauté	P7	P8	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	32	27	22	2													1					1						1			
1	54	59		49																											
2	29	30			27																										
3	23	23				2																									
4	14	16					13																								
5	40	47						38					2																		
6	9	13							8																						
7	5	7								5																					
8	25	32									23																				
9	9	12										9																			
10	26	34										1	24																		
11	13	15											1	11								1									
12	15	12											1		11																
13	47	54														44															
14	15	16															14							1							
15	28	21									2							19													
16	35	33																	27												
17	12	13																		12											
18	6	3																			3										
19	39	45		1																		36									
20	34	25									1												22					1			
21	12	14																						12							
22	11	9																							8						
23	37	39																								34			1		
24	1	1																									1				
25	37	47																										34			
26	26	40																											26		
27	7	7																												5	
28	0	1																													

TABLEAU A.8 – Mouvements des employés d'une communauté à l'autre du septième au huitième semestre.

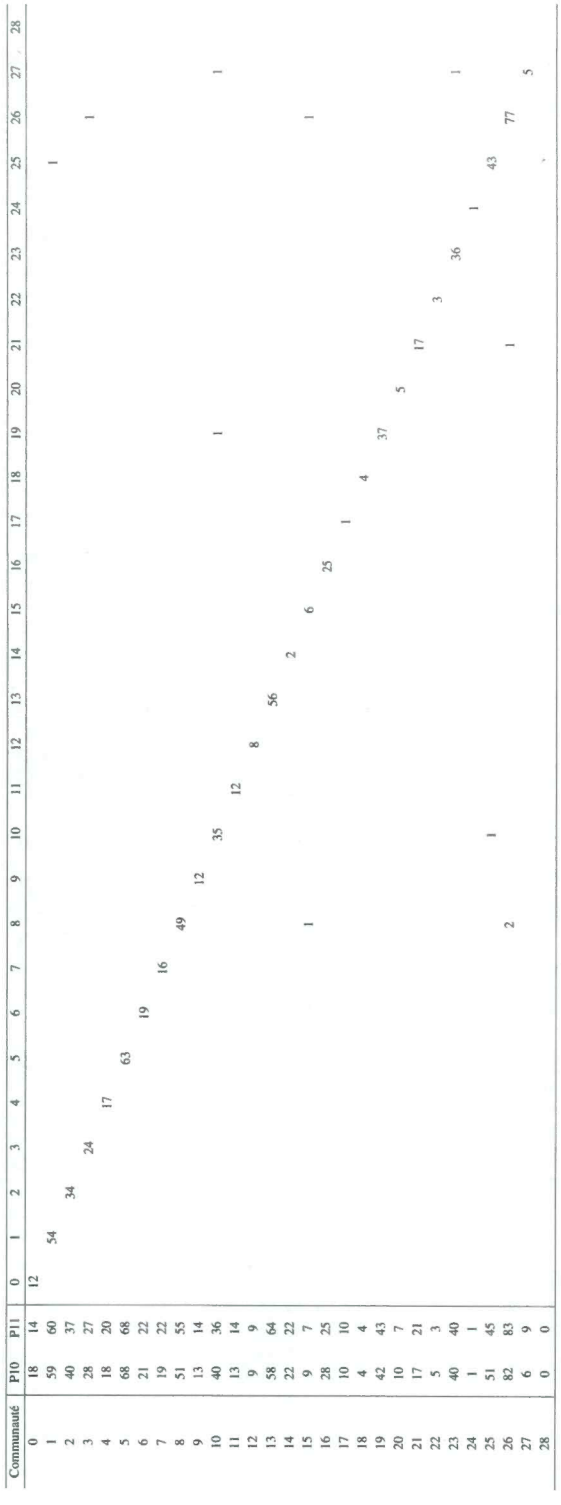
Communauté	P8	P9	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	27	20	16	1																									1		
1	59	63		55																									1		
2	30	34			29																										
3	23	27		2		19																									
4	16	16					14																								
5	47	58						44																							
6	13	14							13																						
7	7	10								6																					
8	32	37									3																				
9	12	10			1							1																			
10	34	35											33																		
11	15	13												13																	
12	12	9				1									9																
13	54	54														47															
14	16	19															15														
15	21	13									1							12													
16	33	29																	27										6		
17	13	10																		9											
18	3	5																			3										
19	45	46																				35		2					1		
20	25	20																					19						1		
21	14	14														2								9					2		
22	9	6																							5				1		
23	39	40																								36	1				
24	1	1																										44			
25	47	53																											4		
26	40	72																												5	
27	7	6																													
28	1	0																													

TABLEAU A.9 – Mouvements des employés d'une communauté à l'autre du huitième au neuvième semestre.



Communauté	P9	P10	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	20	18	14																												
1	63	59		53		1			5		1				1														1		
2	34	40			33																										
3	27	28				22																									
4	16	18					15																								
5	58	68						57																							
6	14	21							14																						
7	10	19								1																					
8	37	51									36																				
9	10	13										1																			
10	35	40											34																		
11	13	13												12																	
12	9	9													8																
13	54	58														5															
14	19	22															17														
15	13	9									1							9													
16	29	28											1						27										1		
17	10	10																		1											
18	5	4																			4										
19	46	42									2											39									
20	20	10			1																		9	1							
21	14	17																						13							
22	6	5																							3						
23	40	40																								37					
24	1	1																									1				
25	53	51								1																		48			
26	72	82		1				2																					69		
27	6	6																												6	
28	0	0																													

TABLEAU A.10 – Mouvements des employés d'une communauté à l'autre du neuvième au dixième semestre.



Communauté	P10	P11	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	18	14	12																												
1	59	60		54																								1			
2	40	37			34																										
3	28	27				24																									
4	18	20					17																						1		
5	68	68						63																							
6	21	22							19																						
7	19	22								16																					
8	51	55									49																				
9	13	14										12																			
10	40	36											35									1								1	
11	13	14												12																	
12	9	9													8																
13	58	64														56															
14	22	22															2														
15	9	7									1							6												1	
16	28	25																	25												
17	10	10																		1											
18	4	4																			4										
19	42	43																				37									
20	10	7																					5								
21	17	21																						17							
22	5	3																							3						
23	40	40																								36					
24	1	1																									1				
25	51	45											1															43			
26	82	83									2													1					77		
27	6	9																												5	
28	0	0																													

TABLEAU A.11 – Mouvements des employés d'une communauté à l'autre du dixième au onzième semestre.

Annexe B

Figures

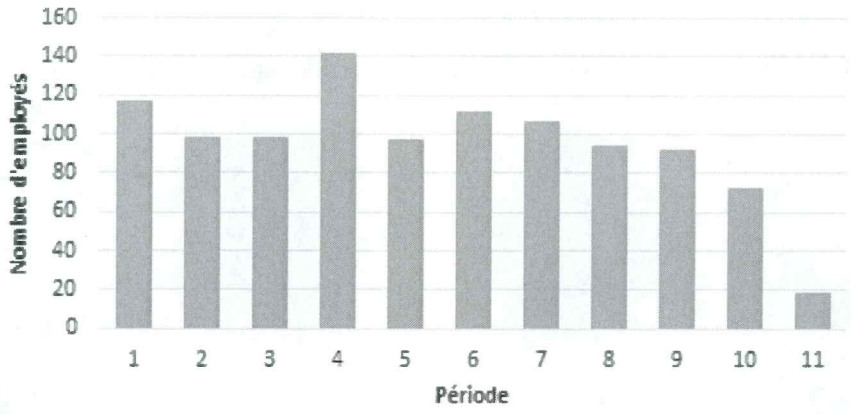


FIGURE B.1 – Distribution des employés selon la période de leur première intervention sur le forum.

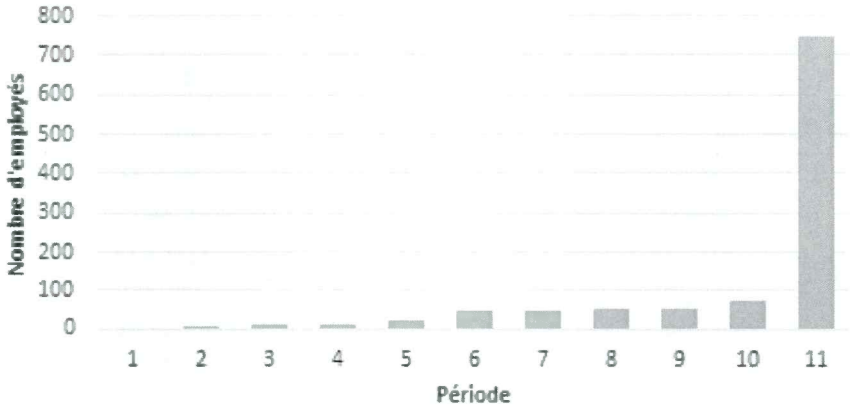


FIGURE B.2 – Distribution des employés selon la période de leur dernière intervention sur le forum.

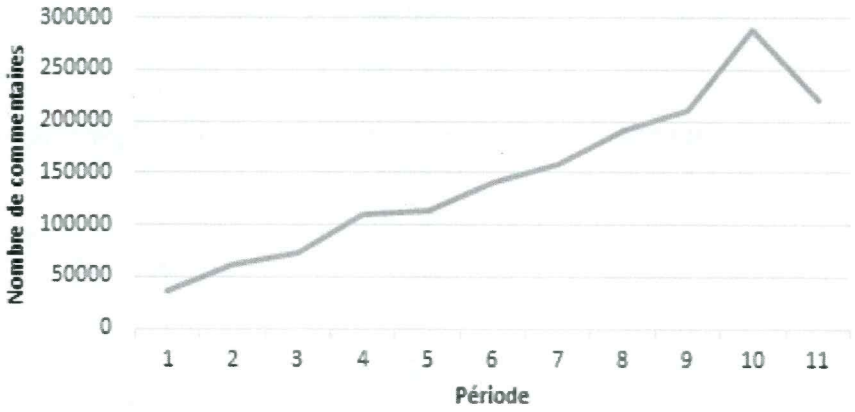


FIGURE B.3 – Évolution du nombre de commentaires émis sur l'ensemble de l'horizon temporel.

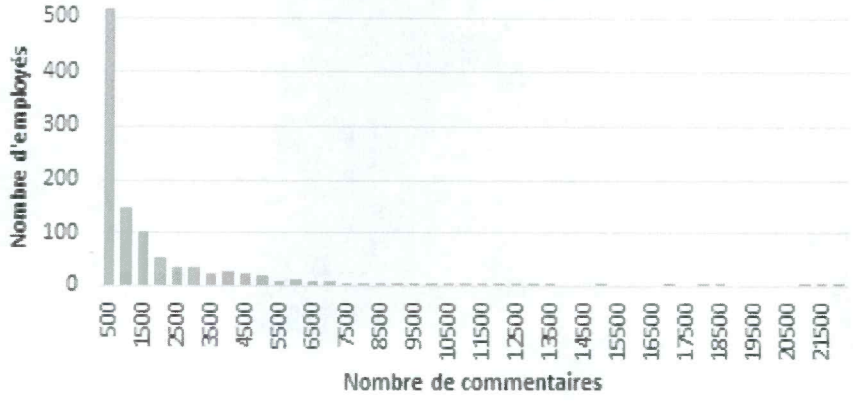


FIGURE B.4 – Distribution des employés selon le nombre de commentaires effectués durant l'ensemble de l'horizon temporel. Le nombre de commentaires affichés constitue la fin de l'intervalle, commençant à 0 pour le premier et à la fin de l'intervalle précédent additionné d'un pour les suivants.

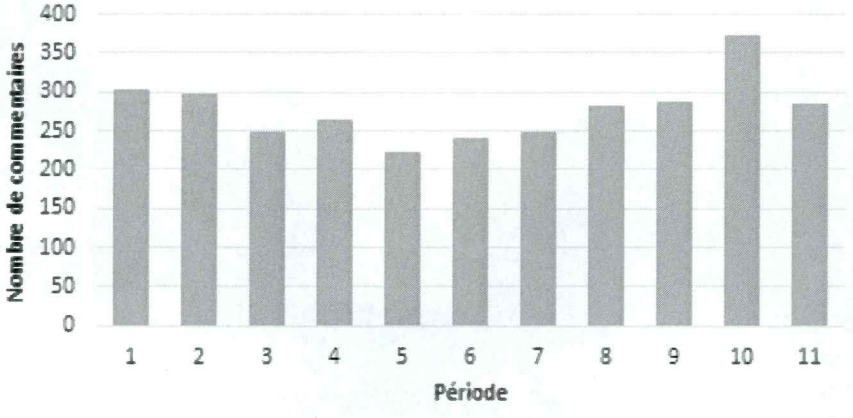


FIGURE B.5 – Nombre moyen de commentaires par personne active à chaque période.

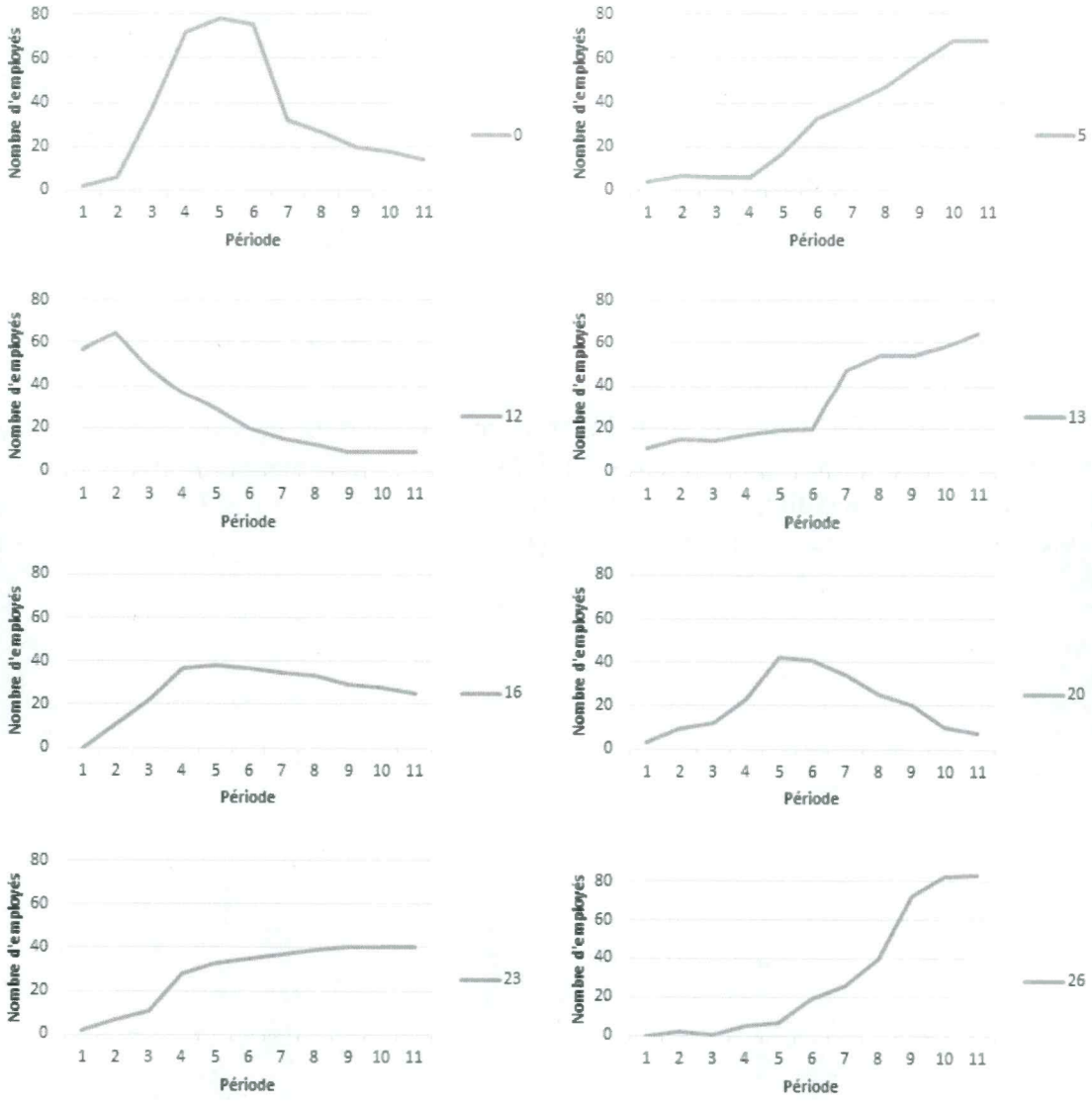


FIGURE B.6 – Nombre d'employés regroupés respectivement dans les communautés *0, 5, 12, 13, 16, 20, 23* et *26*, à chaque période.

Bibliographie

[1] C. LANOUE, « Fondements psychologiques et organisation ». HEC Montréal, 2013.

[2] S. BOUTHILLIER, « Management ». HEC Montréal, 2012.

[3] M. COSSETTE, « Gestion des ressources humaines ». HEC Montréal, 2013.

[4] J. GUILLAUME et M. LATAPY, « Bipartite structure of all complex networks », *Information Processing Letters*, vol. 90, no. 5, p. 215–221, 2004.

[5] R. GÖRKE, T. HARTMANN et D. WAGNER, « Dynamic graph clustering using minimum-cut trees », *Algorithms and Data Structures*, vol. 5664, p. 339–350, 2009.

[6] R. GÖRKE, P. MAILLARD, C. STAUDT et D. WAGNER, « Modularity-driven clustering of dynamic graphs », *in Proceedings of the 9th international conference on Experimental Algorithms*, p. 436–448, Springer, 2010.

[7] P. MUCHA, T. RICHARDSON, K. MACON, M. PORTER et J. ONNELA, « Community structure in time-dependent, multiscale, and multiplex networks », *Science*, vol. 328, no. 5980, p. 876–878, 2010.

[8] M. SPILIOPOULOU, I. NTOUTSI, Y. THEODORIDIS et R. SCHULT, « Monic : modeling and monitoring cluster transitions », *in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 706–711, ACM, 2006.

[9] B. JUNKER et F. SCHREIBER, *Analysis of biological network*. 2008.

[10] S. WASSERMAN et K. FAUST, « Social network analysis : methods and applications », *Acta Sociologica*, vol. 37, no. 4, p. 467–482, 1994.

[11] J. SCOTT, *Social network analysis*. 1 éd., 1991.

[12] M. FALOUTSOS, P. FALOUTSOS et C. FALOUTSOS, « On power-law relationships of the internet topology », *Computer Communication Review*, vol. 9, no. 4, p. 251–262, 1999.

[13] M. NEWMAN, « The structure and function of complex networks », *Siam Review*, vol. 45, no. 2, p. 167–256, 2003.

[14] J. BONDY et U. MURTY, *Graph theory with applications*. 1976.

[15] M. PORTER, J. ONNELA et P. MUCHA, « Communities in network », *Notices of the AMS*, vol. 56, no. 9, p. 1082–1097, 2009.

[16] R. GUIMERA, M. SALES-PARDO et L. AMARAL, « Module identification in bipartite and directed networks », *Physical Review E*, vol. 76, no. 3, p. 036102, 2007.

[17] P. TURÁN, « On the theory of graphs », *in Colloquium Mathematicum 3*, p. 19–30, icm, 1954.

[18] M. GIRVAN et M.E.J.NEWMAN, « Community structure in social and biological networks », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, p. 7821–7826, 2002.

[19] S. FORTUNATO, « Community detection in graphs », *Physics Reports*, vol. 486, no. 3, p. 75–174, 2010.

[20] S. BOCCALETTI, V. LATORA, Y. MORENO, M. CHAVEZ et D. HWANG, « Complex networks : Structure and dynamics », *Physics Reports - Review Section of Physics Letters*, vol. 424, no. 4, p. 175–308, 2006.

[21] C. BILGIN et B. YENER, « Dynamic network evolution : Models, clustering, anomaly detection », *IEEE Networks*, 2006.

[22] S. ASUR, S. PARTHASARATHY et D. UCAR, « An event-based framework for characterizing the evolutionary behavior of interaction graphs », *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 4, p. 16, 2009.

[23] M. NEWMAN et M. GIRVAN, « Analysis of weighted networks », *Physical Review E*, vol. 70, no. 5, p. 056131, 2004.

[24] A. BARRAT, M. BARTHELEMY et A. VESPIGNANI, « Weighted evolving networks : Coupling topology and weight dynamics », *Physical Review Letters*, vol. 92, no. 22, p. 228701, 2004.

[25] A. BARRAT, M. BARTHELEMY, R. PASTOR-SATORRAS et A. VESPIGNANI, « The architecture of complex weighted networks », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, p. 3747–3752, 2004.

[26] M. BARTHELEMY, A. BARRAT, R. PASTOR-SATORRAS et A. VESPIGNANI, « Characterization and modeling of weighted networks », *Physica A - Statistical Mechanics and its Applications*, vol. 346, no. 1, p. 34–43, 2005.

[27] E. ALMAAS, B. KOVACS, T. VISCEK, Z. OLTVAI et A.-L. BARABÁSI, « Global organization of metabolic fluxes in the bacterium escherichia coli », *Nature*, vol. 427, no. 6977, p. 839–843, 2004.

[28] M. BARBER, « Modularity and community detection in bipartite networks », *Physical Review E*, vol. 76, no. 6, p. 066102, 2007.

[29] P. ZHANG, J. WANG, X. LI, Z. DI et Y. FAN, « The clustering coefficient and community structure of bipartite networks », *Physica A : Statistical Mechanics and its Applications*, vol. 387, no. 27, p. 6869–6875, 2008.

[30] P. HOLME et J. SARAMAKI, « Temporal networks », *Physics Reports - Review Section of Physics Letters*, vol. 519, no. 3, p. 97–125, 2012.

[31] R. CAZABET, *Détection de communautés dynamiques dans des réseaux temporels*. Thèse doctorat, Université Paul Sabatier- Toulouse III, 2013.

[32] D. GREENE, D. DOYLE et P. CUNNINGHAM, « Tracking the evolution of communities in dynamic social networks », *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, p. 176–183, 2010.

[33] L. FREEMAN, « A set of measures of centrality based upon betweenness », *Sociometry*, vol. 40, no. 1, p. 35–41, 1977.

[34] J. KUMPULA, J. ONNELA, J. SARAMAKI, K. KASKI et J. KERTESZ, « Emergence of communities in weighted networks », *Physical Review Letters*, vol. 99, no. 22, p. 228701, 2007.

[35] G. PALLA, I. DERENYI, I. FARKAS et T. VICSEK, « Uncovering the overlapping community structure of complex networks in nature and society », *Nature*, vol. 435, no. 7043, p. 814–818, 2005.

[36] H. SIMON, « The architecture of complexity », *Proceedings of the American Philosophical Society*, vol. 106, no. 6, p. 467–482, 1962.

[37] M. NEWMAN, « Fast algorithm for detecting community structure in networks », *Physical Review E*, vol. 69, no. 6, p. 066133, 2004.

[38] J. HOPCROFT, O. KHAN, B. KULIS et B. SELMAN, « Tracking evolving communities in large linked networks », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, p. 5249–5253, 2004.

[39] G. PALLA, A.-L. BARABÁSI et T. VICSEK, « Quantifying social group evolution », *Nature*, vol. 446, no. 7136, p. 664–667, 2010.

[40] F. RADICCHI, C. CASTELLANO, F. CECCONI, V. LORETO et D. PARISI, « Defining and identifying communities in networks », *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, p. 2658–2663, 2004.

[41] S. CAFIERI, P. HANSEN et L. LIBERTI, « Edge ratio and community structure in networks », *Physical review E*, vol. 81, no. 2, p. 026105, 2010.

[42] B. BOLLOBÁS, *Modern graph theory*, vol. 184. 1998.

[43] J. SHI et J. MALIK, « Normalized cuts and image segmentation », *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, p. 888–905, 2000.

[44] M. NEWMAN et M. GIRVAN, « Finding and evaluating community structure in networks », *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

[45] Y. FAN, M. LI, P. ZHANG, J. WU et Z. DI, « Accuracy and precision of methods for community identification in weighted networks », *Physica A - Statistical Mechanics and its Applications*, vol. 377, no. 1, p. 363–372, 2007.

[46] D. CHAKRABARTI, R. KUMAR et A. TOMKINS, « Evolutionary clustering », *in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 554–560, ACM, 2006.

[47] T. COVER et J. THOMAS, *Elements of information theory*. 1991.

[48] D. MACKAY, *Information Theory, Inference, and Learning Algorithm*. 2003.

[49] A. FRED et A. JAIN, « Robust data clustering », *in Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, p. 128–133, IEEE, 2003.

[50] L. DANON, J. DIAZ-GUILERA, J. DUCH et A. ARENAS, « Comparing community structure identification », *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005.

[51] A. JAIN et R. DUBES, *Algorithms for clustering data*. 1988.

[52] W. RAND, « Objective criteria for the evaluation of clustering methods », *Journal of the American Statistical Association*, vol. 66, no. 336, p. 846–850, 1971.

[53] V. BLONDEL, J. GUILLAUME, R. LAMBIOTTE et E. LEFEBVRE, « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics - Theory and Experiment*, vol. 2008, no. 10, 2008.

[54] S. KIRKPATRICK, C. GELATT et M. VECCHI, « Optimization by simulated annealing », *Science*, vol. 220, no. 4598, p. 671–680, 1983.

[55] S. BOETTCHER et A. PERCUS, « Optimization with extremal dynamics », *Physical Rewiew Letters*, vol. 86, no. 23, p. 5211–5214, 2001.

[56] M.E.J.NEWMAN, « Finding community structure in networks using the eigenvectors of matrices », *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.

[57] U. RAGHAVAN, R. ALBERT et S. KUMARA, « Near linear time algorithm to detect community structures in large-scale networks », *Physical Review E*, vol. 76, no. 3, p. 036106, 2007.

[58] L. DONETTI et M. M. NOZ, « Detecting network communities : a new systematic and efficient algorithm », *Journal of Statistical Mechanics - Theory and Experiment*, p. P10012, 2004.

[59] M. HANDCOCK, A. RAFTERY et J. TANTRUM, « Model based clustering for social networks », *Journal of the Royal Statistical Society Series A - Statistics in Society*, vol. 170, no. 2, p. 301–322, 2007.

[60] B. HUGHES, *Random Walks and Random Environments : Random Walks*. 1 éd., 1995.

[61] F. WU, « The potts model », *Reviews of Modern Physics*, vol. 54, no. 1, p. 235–268, 1982.

[62] I. LEUNG, P. HUI, P. LIO et J. CROWCROFT, « Towards real-time community detection in large networks », *Physical Review E*, vol. 79, no. 6, p. 066107, 2009.

[63] S. GREGORY, « Finding overlapping communities in networks by label propagation », *New Journal of Physics*, vol. 12, p. 103018, 2010.

[64] J. XIE et B. SZYMANSKI, « Labelrank : A stabilized label propagation algorithm for community detection in networks », *in Network Science Workshop (NSW), 2013 IEEE 2nd*, p. 138–143, IEEE, 2013.

[65] H. LOU, S. LI et Y. ZHAO, « Detecting community structure using label propagation with weighted coherent neighborhood propinquity », *Physica A - Statistical Mechanics and its Applications*, vol. 392, no. 14, p. 3095–3105, 2013.

[66] M. BARBER et J. CLARK, « Detecting network communities by propagating labels under constraints », *Physical Review E*, vol. 80, no. 2, p. 026129, 2009.

[67] X. LIU et T. MURATA, « Advanced modularity-specialized label propagation algorithm for detecting communities in networks », *Physica A*, vol. 389, no. 7, p. 1493–1500, 2010.

[68] A. LANCICHINETTI et S. FORTUNATO, « Community detection algorithms : A comparative analysis », *Physical Review E*, vol. 80, no. 5, p. 056117, 2009.

[69] A. CLAUSET, M. NEWMAN et C. MOORE, « Finding community structure in very large networks », *Physical Review E*, vol. 70, no. 6, p. 066111, 2004.

[70] J. DUCH et A. ARENAS, « Community detection in complex networks using extremal optimization », *Physical Review E*, vol. 72, no. 2, p. 027104, 2005.

[71] A. LANCICHINETTI, F. RADICCHI, J. RAMASCO et S. S. FORTUNATO, « Finding statistically significant communities in networks », *Plos One*, vol. 6, no. 4, p. e18961, 2011.

[72] P. PONS et M. LATAPY, « Computing communities in large networks using random walks », *in Computer and Information Sciences - ISCIS 2005*, p. 284–293, ISCIS, 2005.

[73] T. AYNAUDA, J.-L. GUILLAUMEA, Q. WANGB et E. FLEURY, « Communities in evolving networks : Definitions, detection, and analysis techniques », *Dynamics On and Of Complex Networks*, vol. 2, p. 159–200, 2011.

[74] A. CORREC, « Comparaison d'algorithmes de détection de communautés dynamiques dans les réseaux évolutifs », Mém. D.E.A., HEC Montréal, 2015.

[75] M. TAKAFFOLI, F. SANGI, J. FAGNAN et O. ZAIANE, « Community evolution mining in dynamic social networks », *Procedia - Social and Behavioral Sciences*, vol. 22, p. 49–58, 2011.

[76] Z. CHEN, K. A. WILSON, Y. JIN, W. HENDRIX et N. F. SAMATOVA, « Detecting and tracking community dynamics in evolutionary networks », *in 2010 IEEE International Conference on Data Mining Workshops*, p. 318–327, IEEE, 2010.

[77] Y. WANG, B. WU et N. DU, « Community evolution of social network : feature, algorithm and model », *Physics and Society*, 2008. arXiv :0804.4356.

[78] T. AYNAUD et J.-L. GUILLAUME, « Static community detection algorithms for evolving networks », *in Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, p. 508–514, IEEE, 2010.

[79] Q. WANG et E. LEURY, « Mining time-dependent communities », *LAWDN - Latin-American Workshop on Dynamic Networks*, 2010.

[80] J. GEHWEILER et H. MEYERHENKE, « A distributed diffusive heuristic for clustering a virtual p2p supercomputer », *in Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on*, p. 1–8, IEEE, 2010.

[81] M. TAKAFFOLI, R. RABBANY et O. ZAIANE, « Incremental local community identification in dynamic social networks », *in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, p. 90–94, ACM, 2013.

[82] K. KIM, R. MCKAY et B.-R. MOON, « Multiobjective evolutionary algorithms for dynamic social network clustering », *in Proceedings of the 12th annual conference on Genetic and evolutionary computation*, p. 1179–1186, ACM, 2010.

[83] C.-D. WANG, J.-H. LAI et S. PHILIP, « Dynamic community detection in weighted graph streams », *Proc. of SDM*, p. 151–161, 2013.

[84] R. LUNG, C. CHIRA et A. ANDREICA, « Game theory and extremal optimization for community detection in complex dynamic networks », *Plos One*, vol. 9, no. 2, p. e86891, 2014.

[85] T. DINH, I. SHIN, N. THAI et M. THAI, « A general approach for modules identification in evolving networks », *Dynamics of Information Systems*, p. 83–100, 2010.

[86] J. RIEDY et D. BADER, « Multithreaded community monitoring for massive streaming graph data », *in 2013 IEEE 27th International Symposium on Parallel & Distributed Processing Workshops and PhD Forum*, p. 1646–1655, IEEE, 2013.

[87] T. DINH et N. N. M. THAI, « An adaptive approximation algorithm for community detection in dynamic scale-free networks », *in INFOCOM, 2013 Proceedings IEEE*, p. 55–59, IEEE, 2013.

[88] A. LANCICHINETTI et S. FORTUNATO, « Consensus clustering in complex networks », *Scientific Reports*, vol. 2, 2012.

[89] Y. CHI, X. SONG, D. ZHOU, K. HINO et B. TSENG, « Evolutionary spectral clustering by incorporating temporal smoothness », *in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 153–162, ACM, 2007.

[90] Y.-R. LIN, Y. CHI, S. ZHU, H. SUNDARAM et B. TSENG, « Facetnet : a framework for analyzing communities and their evolutions in dynamic networks », *in Proceedings of the 17th international conference on World Wide Web*, p. 685–694, ACM, 2008.

[91] Y. LIN, Y. CHI, S. ZHU, H. SUNDARAM et B. TSENG, « Analyzing communities and their evolutions in dynamic social networks », *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 2, p. 8, 2009.

[92] M. KIM et J. HAN, « A particle-and-density based evolutionary clustering method for dynamic networks », *Proceedings of the VLDB Endowment*, vol. 2, no. 1, p. 622–633, 2009.

[93] M. ESTER, H.-P. KRIEGEL, J. SANDER et X. XU, « A density-based algorithm for discovering clusters in large spatial databases with noise », *KKD*, vol. 96, p. 226–231, 1996.

[94] M. TAKAFFOLI, F. SANGI, J. FAGNAN et O. ZAIANE, « Modec-modeling and detecting evolutions of communities », *in Fifth International AAAI Conference on Weblogs and Social Media*, ACM, 2011.

[95] K. XU, M. KLIGER et A. HERO, « Tracking communities in dynamic social networks », *Social Computing, Behavioral-Cultural Modeling and Prediction*, vol. 6589, p. 219–226, 2011.

[96] S. YU et J. SHI, « Multiclass spectral clustering », *in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, p. 313–319, IEEE, 2003.

[97] K. XU, M. KLIGER et A. HERO, « Adaptive evolutionary clustering », *Data Mining and Knowledge Discovery*, vol. 28, no. 2, p. 304–336, 2014.

[98] F. FOLINO et C. PIZZUTI, « A multiobjective and evolutionary clustering method for dynamic networks », *in Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, p. 256–263, IEEE, 2010.

[99] M.-G. GONG, L.-J. ZHANG, J.-J. MA et L.-C. JIAO, « Adaptive evolutionary clustering », *Journal of Computer Science and Technology*, vol. 27, no. 3, p. 455–467, 2012.

[100] R. GOMORY et T. HU, « Multi-terminal network flows », *Journal of the Society for Industrial and Applied Mathematics*, vol. 9, no. 4, p. 551–570, 1961.

[101] G. FLAKE, R. TARJAN et K. TSIOUTSIOULIKLIS, « Graph clustering and minimum cut trees », *Internet Math*, vol. 1, no. 4, p. 385–408, 2003.

[102] R. GÖRKE, P. MAILLARD, A. SCHUMM, C. STAUDT et D. WAGNER, « Dynamic graph clustering combining modularity and smoothness », *Journal of Experimental Algorithmics (JEA)*, vol. 18, p. 1–5, 2013.

[103] S. BANSAL, S. BHOWMICK et P. PAYMAL, « Fast community detection for dynamic complex networks », *in Complex Networks*, p. 196–207, Springer, 2011.

[104] N. NGUYEN, T. DINH, Y. SHEN et M. THAI, « Dynamic social community detection and its applications », *Plos One*, vol. 9, no. 4, p. e91431, 2014.

[105] J. SHANG, L. LIU, F. XIE, Z. CHEN, J. MIAO, X. FANG et C. WU, « A real-time detecting algorithm for tracking community structure of dynamic networks », *in 6th SNA-KDD Workshop*, vol. 12, SIGKDD, 2012.

[106] R. CAZABET, F. AMBLARD et C. HANACHI, « Detection of overlapping communities in dynamical social networks », *in Social Computing (SocialCom), 2010 IEEE Second International Conference on*, p. 309–314, IEEE, 2010.

[107] R. CAZABET et F. AMBLARD, « Simulate to detect : a multi-agent system for community detection », *in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, vol. 2, p. 402–408, IEEE, 2011.

[108] M. C. e. B. K. S. J. XIE, « Labelrankt : Incremental community detection in dynamic networks via label propagation », *Proceedings of the Workshop on Dynamic Networks Management and Mining*, p. 25–32, 2013.

[109] D. DUAN, Y. LI, R. LI et Z. LU, « Incremental k-clique clustering in dynamic social networks », *Artificial Intelligence Review*, vol. 38, no. 2, p. 129–147, 2012.

[110] T. FALKOWSKI, *Community analysis in dynamic social networks*. Thèse doctorat, Sierke, 2009.

[111] H. NING, W. XU, Y. CHI, Y. GONG et T. HUANG, « Incremental spectral clustering by efficiently updating the eigen-system », *Pattern Recognition*, vol. 43, no. 1, p. 113–127, 2010.

[112] T. AYNAUD et J. L. GUILLAUME, « Long range community detection », *in LAWDN - LatinAmerican Workshop on Dynamic Networks,*, ISCIS, 2010.

[113] L. GAUVIN, A. PANISSON et C. CATTUTO, « Detecting the community structure and activity patterns of temporal networks : a non-negative tensor factorization approach », *Plos One*, vol. 9, no. 1, p. e86028, 2014.

[114] M. B. JDIDIA, C. ROBARDET et E. FLEURY, « Communities detection and analysis of their dynamics in collaborative networks », *in 2nd International Conference on Digital Information Management (ICDIM)*, p. 44–54, IEEE, 2007.

[115] C. TANTIPATHANANANDH, T. BERGER-WOLF, et D. KEMPE, « A framework for community identification in dynamic social networks », *in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 717–726, ACM, 2007.

[116] C. TANTIPATHANANANDH et T. BERGER-WOLF, « Finding communities in dynamic social networks », *in 2011 IEEE 11th International Conference on Data Mining*, p. 1236–1241, IEEE, 2011.

[117] B. MITRA, L. TABOURIER et C. ROTH, « Intrinsically dynamic network communities », *Computer Networks*, vol. 56, no. 3, p. 1041–1053, 2012.

[118] T. YANG, Y. CHI, S. ZHU, Y. GONG et R. JIN., « Detecting communities and their evolutions in dynamic social networks-a bayesian approach », *Machine Learning*, vol. 82, no. 2, p. 157–189, 2011.

[119] P. HOLLAND, K. LASKEY et S. LEINHARDT, « Stochastic blockmodels : first steps », *Social Networks*, vol. 5, no. 2, p. 109–137, 1983.