

UNIVERSITY NAME

DOCTORAL THESIS

Conversations, Groupes et Communautés dans les Flots de Liens

Auteur :
Noé GAUMONT

Directeurs :
Clémence MAGNIEN
Matthieu LATAPY

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

in the

Research Group Name
Department or School Name

2 mai 2016

Table des matières

Acknowledgements	ii
1 Introduction	3
1.1 Communauté dans les graphes	3
1.2 Extension temporelle	3
2 Expected Nodes (COMPLENET)	5
3 Extensions temporelle des graphes	7
3.1 Ètat de l'art des formalismes et outils	7
3.2 L'approche flot de liens	7
3.2.1 Définition	7
3.2.2 Sous-flot induit	7
3.2.3 Degré et densité	7
3.3 Manipulation de flots	8
3.3.1 Représentation	8
Liste des notations	8
4 Étude d'une archive de courriels	11
4.1 Prétraitement sur le jeu de données	11
4.2 Caractéristiques élémentaires des discussions	12
4.3 Étude des discussions en tant que sous-flots	14
4.3.1 Application de la Δ -densité	14
4.3.2 Répartition temporelle et structurelle des discussions	17
4.3.3 Flot quotient	18
4.3.4 Conclusion	20
4.4 Détection de structures denses	20
4.4.1 Méthode de détection	20
4.4.2 Comparaison des partitions	21
4.4.3 Conclusion	24
5 Détection de groupes denses (SNAM)	25
6 Fonction de qualité	27
6.1 Définition	27
6.2 Générateur de flots de liens avec structure communautaire	27
7 Conclusion	29

Les axes des figures sont pour l'instant en anglais.

Tant que la bilbio n'est pas en version finale. Il n'est pas nécessaire de vérifier le formatage.

Chapitre 1

Introduction

1.1 Communauté dans les graphes

[3]

1.2 Extension temporelle

Chapitre 2

Expected Nodes (COMPLENET)

Chapitre 3

Extensions temporelle des graphes

3.1 Ètat de l'art des formalismes et outils

3.2 L'approche flot de liens

3.2.1 Définition

Un flot de liens est défini comme un triplet : $\mathcal{L} = (T, V, E)$, où $T = [\alpha, \omega]$ est un intervalle de temps, V un ensemble de n nœuds et $E \subseteq T \times T \times V \times V$ un ensemble de m liens. Les liens de E sont des quadruplets (b, e, u, v) , signifiant que la paire de nœuds (u, v) est connectée sur l'intervalle $[b, e] \subseteq [\alpha, \omega]$. Nous considérons un flot non orienté, i.e. $(b, e, u, v) = (b, e, v, u)$, et sans boucle, i.e. $u \neq v$.

Nous dénotons la durée du flot par $\bar{L} = \omega - \alpha$. $\beta_E = \min_{(b,e,u,v) \in E}(b)$ et $\psi_E = \max_{(b,e,u,v) \in E}(e)$ sont respectivement l'apparition du premier lien et la disparition du dernier lien.

Un flot de liens est simple si pour tout $(b, e, u, v) \in E$ et $(b', e', u, v) \in E$, $[b, e] \cap [b', e'] = \emptyset$. La simplification d'un flot de liens $\sigma = L$.

$V(E') = u_{(b,e,u,v) \text{ in } E'}$ sommets induits par un ensemble de liens.

$\xi(L, \Delta)$ ajout de Δ à chaque lien.

Que des flots avec durées ou bien delta densité ?

3.2.2 Sous-flot induit

Sous flot induit par un ensemble de lien $E' : L(E') = ([\beta_{E'}, \psi_{E'}, V(E'), E'])$.

Sous flot induit par un ensemble de paire nœuds $S \in V^2 : L(S). L(S) = ([\beta_{E'}, \psi_{E'}], V', E')$ avec $E' = \{(b, e, u, v) \in E, (u, v) \in S\}$. Par convention, on note $L(v) = L(\{v\}) \times V$.

Sous flot induit par un intervalle de temps $T' = [\alpha', \omega']$, $T' \subseteq T : L_{\alpha'.. \omega'}. L_{\alpha'.. \omega'} = ([\alpha', \omega'], V(E'), E')$ avec $E' = \{(b', e', u, v), \exists (b, e, u, v) \in E, b' = \max(b, \alpha'), e' = \min(e, \omega')\}$. Par convention, on note $L_{t..t} = L_t$

Il est aussi possible de combiner ces notions. Par exemple avec $V' \subset V, L_{\alpha'.. \omega'}(V'^2)$ est le sous flot correspondant au lien entre les nœuds de V' sur l'intervalle $[\alpha', \omega']$.

3.2.3 Degré et densité

Beaucoup de notions sont développées autour de cet objet. Degré temporelle d'un nœud u :

$$d_t(u) = |L_t(v)| = |\{(b, e, u, v) \in E, b \leq t \leq e\}| \quad (3.1)$$

Par extension pour un ensemble de sommets V' :

$$d_t(V') = |L_t(V'^2)| = |\{(b, e, u, v) \in E, u, v \in V', b \leq t \leq e\}| \quad (3.2)$$

Degré interne maintenant ?

$$d_t(E') = |\{(b, e, u, v) \in E', b \leq t \leq e\}| \quad (3.3)$$

Il est possible d'intégrer sur l'ensemble du temps

$$D_{\alpha..w}(v) = \int_{\alpha}^{\omega} d(v, t) dt = D(v) \quad (3.4)$$

Par convention, on notera le degré moyen de v : $d(v) = \langle d(v, t) \rangle = \frac{D(v)}{\omega - \alpha}$. Il en va de même pour les autres degrés.

C'est le cas de la densité :

$$\delta(L) = \frac{2 \sum_{l \in E}}{n(n-1)(\bar{L})} \quad (3.5)$$

3.3 Manipulation de flots

Existence de la lib

3.3.1 Représentation

Liste des notations

Symbol	description
L	Flot de liens
T	intervalle de temps
V	ensemble de nœuds
E	ensemble de liens : (b, e, u, v)
n	nombre de nœuds
$ L , E $	nombre de liens dans le flot
β_E	temps d'apparition du premier lien
ψ_E	temps de disparition du dernier lien
$\xi(L, \Delta)$	Flot de liens où chaque lien dure Δ
$L(V'^2)$	sous flot induits par les nœuds de V'
$L_{t..t'}$	sous flot induits par l'intervalle $[t, t']$
$d_t(v)$	degré de v à l'instant t
$d(v)$	degré moyen v sur T
$\delta(L)$	densité du flot
$\delta_\Delta(L)$	densité du flot ou chaque lien dure Δ

Chapitre 4

Étude d'une archive de courriels

L'étude de la structure des réseaux est un sujet qui est étudié depuis assez long-temps [REF](#). Ces études ont, dans un premier temps, permis de trouver comment caractériser une structure puis, dans un second temps, de proposer des méthodes de détections de ces structures. La littérature sur l'étude des flots de liens est encore récente et il n'existe que peu d'études [REF](#) sur les spécificités des structures dans les flots de liens.

Intro sur-
ement trop
général qui
sera bougée
ailleurs.

Nous nous intéressons ici à une archive de courriels publiquement disponibles¹. Cette archive contient l'ensemble des courriels échangés par différents utilisateurs pour résoudre un problème survenu lors de l'utilisation de Debian. Typiquement, une personne ayant un problème lors de l'installation envoie un courriel à la liste afin de demander de l'aide. Toute personne inscrite sur la liste reçoit ce courriel et peut y répondre ce qui donne lieu à une discussion visible par tous. Ces discussions ont déjà été étudiées dans le passé [2] mais cela a été fait en utilisant des méthodes statiques uniquement.

Or, ces données se représentent naturellement sous forme de flot de liens en associant chaque personne à un nœud et chaque courriel entre deux personnes à un lien dans le flot à l'instant où le courriel a été envoyé. L'avantage de ces données de communications est que nous connaissons la discussion (*thread*) dans laquelle a lieu chaque message. Une discussion est un ensemble de courriels dont tous les messages répondent à un message précédent de la discussion excepté pour le premier qui a initié la discussion et que nous appelons *racine*. Ainsi, nous étudions la structure des discussions dans le flot de liens représentant les courriels envoyés sur la liste.

Utiliser le formalisme de flot de liens est particulièrement intéressant car cette liste de diffusion existe depuis 1994. L'aspect temporel des discussions est donc important.

4.1 Prétraitement sur le jeu de données

Bien qu'accessible sur internet, ce jeu de données nécessite un ensemble de traitements avant de pouvoir exploiter les 724985 courriels que contenait l'archive en janvier 2015. Tout d'abords, les données ne sont pas sous la forme d'un flot de liens avec la structure des conversations. Les données sont accessibles via le site internet et ne sont pas structurées. Pour avoir ces informations sous la forme d'un flot de liens, un script d'extraction a été développé [URL](#). Lors de l'extraction, 2269 courriels n'ont pas pu être pris en compte car certaines informations étaient manquantes ou mal formées.

Une fois les informations récupérées, il faut les transformer en un flot de liens cohérent. Pour chaque message m , nous extrayons son auteur $a(m)$, l'instant $t(m)$

1. <https://lists.debian.org/debian-user/>

auquel le message a été posté², le message auquel il répond $p(m)$ trouvé via le champ IN-REPLY-TO, son destinataire $a(p(m))$ et la discussion $D(m)$ dans laquelle il apparaît. Comme les messages *racines* ne répondent à aucun autre, nous imposons $p(m) = m$. L'ensemble de liens du flot est donc $\{(t(m), a(m), a(p(m)))\}_m$. Nous ne prenons pas en compte la direction des liens.

Une fois le flot créé, il est encore nécessaire de vérifier sa cohérence. Un message peut être filtré pour différentes raisons : le courriel apparaît avant le message auquel il est censé répondre, le message auquel il répond n'est pas présent dans l'archive, l'auteur et le destinataire sont la même personne. Cette dernière condition permet notamment d'éviter la présence de boucles dans le flot. Cela concerne principalement les *racines*. Il s'agit de vérifications simples auquel il faut ajouter les vérifications sur la cohérence de la structure des discussions. Ainsi, une discussion est entièrement retirée du jeu de données s'il manque la *racine*, si un de ses messages a été retiré à l'étape précédente ou si la discussion a débuté trop récemment ou si elle dure trop longtemps. Les deux dernières conditions permettent d'éviter un biais envers les conversations incomplètes car trop longues ou trop récentes. La limite pour considérer une discussion trop récente ou trop longue a été fixé à 2 ans (6.3×10^7 s) en observant la distribution de durées des discussions, voir la figure 4.1.

Refaire les courbes qui sont avant la suppression à cause de la durée.

Une fois tout ces messages filtrés, nous obtenons un flot de liens avec 316 569 liens entre 34 648 personnes pendant presque 19 ans et 116 999 discussions. Mis à part les 237 664 messages de début de discussion, ce sont 168 482 courriels qui ont été filtrés soit environ 23%.

4.2 Caractéristiques élémentaires des discussions

Les caractéristiques les plus élémentaires des discussions sont le nombre de courriels, le nombres de personnes, le nombre de paires de personnes distinctes en interaction directes et leur durée. Dans la figure 4.1, sont présentées les distributions cumulatives inverses de ces quantités et on remarque qu'elles sont toutes hétérogènes. On remarque que les données filtrées ne diffèrent pas qualitativement des données brutes.

La distribution des durées des discussions montre que la majorité des discussions dure environ une journée ou moins (100000 secondes équivaut à moins de 28 heures). Par ailleurs, on remarque qu'il n'existe que quelques discussions durant plus d'un an. C'est pourquoi, la limite sur la durée des discussions a été fixée à 2 ans.

Ces premières observations sont nécessaires mais pas suffisantes pour comprendre les caractéristiques d'une discussion. Nous avons également étudié la corrélation entre ces différentes notions et une partie d'entre elles sont présentées dans la figure 4.2.

La corrélation entre la durée et le nombre de courriels, dans la figure 4.2 partie gauche, met en évidence que plus une discussion est grande en nombre de courriels plus elle dure longtemps, ce qui est attendu. Par contre, on observe que les petites discussions ont des durées très variables. Dans la partie droite de la figure 4.2 présentant la corrélation entre le nombre de courriels et d'auteurs, on observe un autre fait attendu [2] qui est qu'une discussion est constituée, en général, de plus de

2. Cet instant est convertit en *timestamp* en tenant compte des fuseaux horaires.

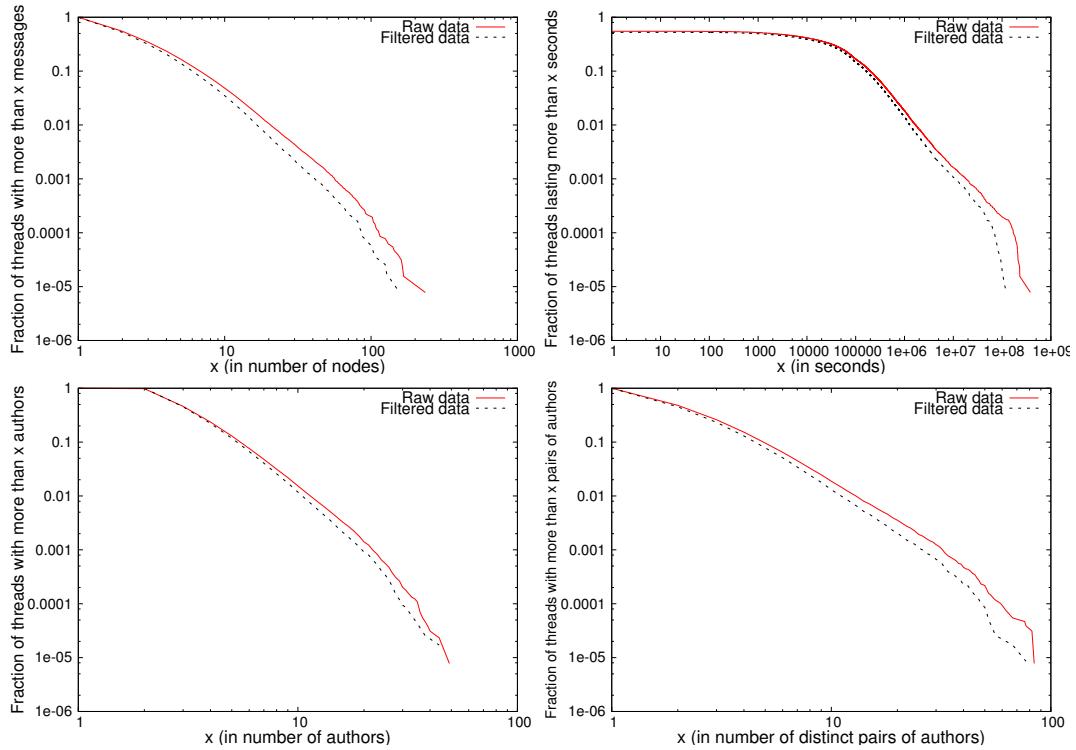


FIGURE 4.1 – Distribution cumulative inverse de différentes caractéristiques pour les données brutes (ligne pleine) et filtrées (ligne pointillée). En haut à gauche : nombre de courriels dans une discussion ; en haut à droite : durée d'une discussion ; en bas à gauche : nombre de personnes dans une discussion ; en bas à droite : nombre de paires d'auteurs distinct dans une .

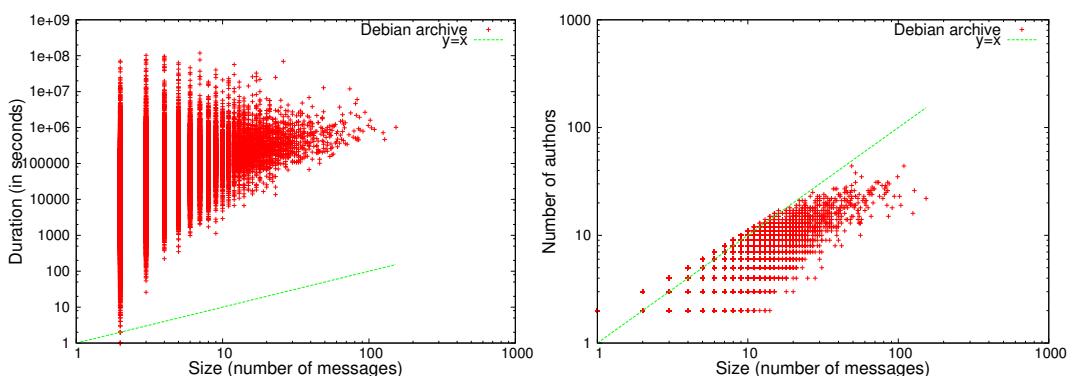


FIGURE 4.2 – Gauche : Corrélation entre le nombre de courriels et la durée d'une discussion. Droite : Corrélation entre le nombre de courriels et le nombre d'auteurs dans une discussion.

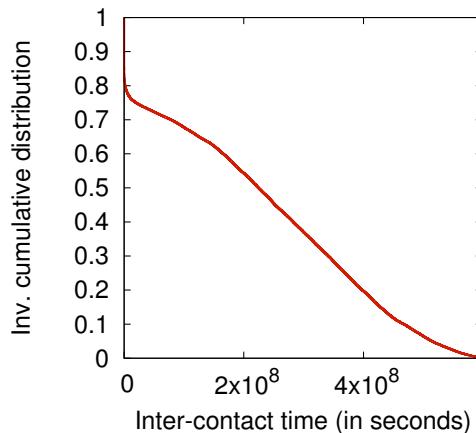


FIGURE 4.3 – Distribution des temps inter-contacts dans le fil de discussions.

messages que de participants. Ainsi lors d'une discussion, c'est un petit nombre de personnes qui échangent potentiellement beaucoup de messages.

Enfin, il est intéressant d'observer la dynamique des échanges entre deux personnes. Soit $\tau(u, v) = (t_{i+1} - t_i)_{i=0..k+1}$ la séquence des temps inter-contacts des k liens entre les nœuds u et v , où t_0 est le temps entre α et le premier lien et t_{k+1} est le temps entre le dernier lien et Ω . Il s'agit du temps écoulé avant que deux personnes se contactent à nouveau, indépendamment peu importe la conversation. Dans la figure 4.3 est représentée la distribution cumulative inverse du temps inter-contacts. 21% des temps inter-contacts sont inférieurs à 30 jours (2.6×10^6 s). Ce chiffre bien que relativement faible est tout de même important car il s'agit de discussions ouvertes où tout le monde peut participer. En particulier, une personne peut envoyer une demande d'aide à un moment donné et ne plus jamais échanger avec les mêmes personnes. Or, on observe que 21% des contacts sont renouvelés en moins de 30 jours. La participation est donc relativement élevée.

Il faudrait faire un inset

4.3 Étude des discussions en tant que sous-flots

4.3.1 Application de la Δ -densité

Jusqu'à maintenant aucune notion intrinsèquement liée aux flots de liens n'a été utilisée pour caractériser les discussions. Le but est d'évaluer si cette structure de flot peut se rapprocher d'une structure communautaire. Comme dit précédemment, les communautés sont souvent définies comme étant des structures devant être densément connectées. C'est pourquoi nous nous attachons à étudier la densité des discussions.

Comme ces données se modélisent par un flot de liens où les liens n'ont pas de durée, nous étudions la Δ -densité pour différentes valeurs de Δ entre 1 seconde et 20 ans. Tout d'abord dans la figure 4.4 est représentée la Δ -densité globale du le flot. En couvrant un spectre aussi large de Δ , on observe que la Δ -densité est croissante avec Δ mais surtout on observe bien la convergence de Δ -densité vers 3.139×10^{-4} , la densité du graphe agrégé, lorsque Δ est proche de $\omega - \alpha$.

Cependant, la Δ -densité du flot n'apporte que peu d'informations en elle-même. Elle est surtout utile pour comparer les valeurs de Δ -densité des sous-flots que sont les discussions. Ainsi dans la figure 4.5, est présentée la distribution cumulative inverse de la Δ -densité des discussions pour différentes valeurs de Δ . On remarque

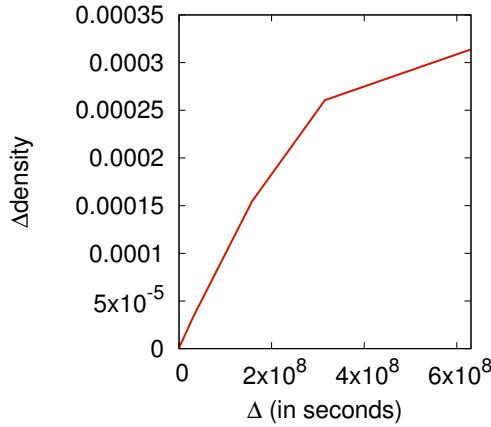


FIGURE 4.4 – Évolution de la Δ -densité du flot de liens pour Δ de 1 seconde à 20 ans. **Ajout bar horizontal de la densité statique dans le graphe agrégé.**

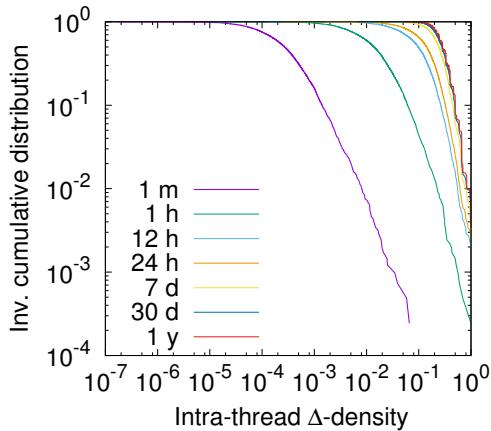


FIGURE 4.5 – Distribution cumulative inverse de la Δ -densité des discussions pour différentes valeurs de Δ s.

que les différentes valeurs de Δ ne semblent pas influencer qualitativement la distribution de Δ -densité. Cette courbe met surtout en évidence que les discussions sont des structures beaucoup plus denses que le flot. En effet, la densité médiane des discussions est, selon la valeur de Δ , entre 2.69×10^{-4} et 0.28 alors que le flot a une Δ -densité variant entre 1.05×10^{-10} et 3.42×10^{-5} . La Δ -densité des discussions est donc en moyenne 10^5 fois plus élevée que celle du flot. Bien que notable, ce fait est attendu notamment car le flot dure beaucoup plus longtemps et concerne beaucoup plus de noeuds que les discussions.

Afin d'aller plus loin dans l'étude de cette structure, il faut revenir à une définition plus précise de ce qu'est une bonne communauté. En soit, une valeur de densité n'est pas suffisante pour définir une structure communautaire. En effet, une discussion ayant une densité de 0.8 peut ne pas être une communauté tandis qu'une autre ayant une densité proche de zéro peut être une communauté. Il faut définir un point de comparaison pour effectivement affirmer qu'une structure est particulièrement dense. La prise en compte de la densité globale est un début mais n'est pas suffisante.

Une autre définition d'une communauté est qu'elle devrait être plus densément connectée à l'intérieur qu'avec les autres communautés adjacentes. Pour un graphe

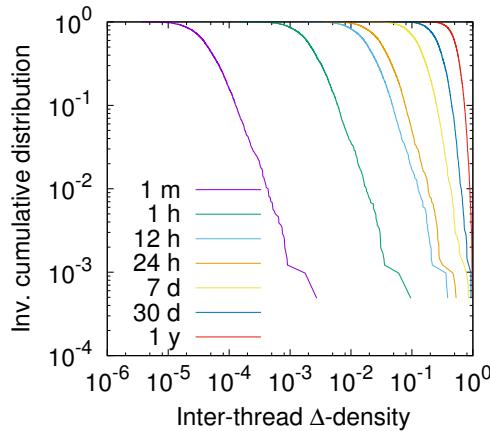


FIGURE 4.6 – Distribution cumulative inverse de la Δ -densité inter discussions pour différentes valeurs de Δ s.

$G = (V, E)$ et une communauté C_i de la partition $C = \{C_j\}_{j_1..k}$ de V en k communautés, cela se traduit par le calcul de la densité entre les communautés, $\delta^{inter}(C_i)$:

$$\delta^{inter}(C_i) = \frac{1}{|C| - 1} \sum_{j, i \neq j} \frac{|\{(u, v) \in E \text{ t.q. } u \in C_i \text{ and } v \in C_j\}|}{|C_i| \cdot |C_j|}. \quad (4.1)$$

Il s'agit tout simplement de la probabilité qu'un lien existe entre les noeuds des deux communautés. Encore une fois, cette notion n'a pas de sens direct dans le formalisme de flot de liens et il est nécessaire de l'adapter. Pour ce faire, nous définissons la Δ -densité inter discussions entre deux discussions D_i et D_j : $\delta^{inter}_\Delta(D_i, D_j)$. Soit $L' = \xi(L, \Delta)$, $V' = V(D_i) \cup V(D_j)$, $t'_\beta = t_\beta(D_i \cup D_j)$ et $t'_\psi = t_\psi(D_i \cup D_j)$. La définition de $\delta^{inter}_\Delta(D_i, D_j)$ est la suivante : $\delta^{inter}_\Delta(D_i, D_j) = \delta(L'_{t'_\beta..t'_\psi}(V'))$. Il s'agit donc de la Δ -densité du flot de liens induit par l'union des noeuds sur l'union de l'intervalle . Afin d'obtenir la Δ -densité inter discussions entre D_i et tout les autres discussions, nous utilisons la moyenne des densité inter discussion entre D_i et les autres discussions, soit :

$$\delta^{inter}_\Delta(D_i) = \frac{1}{|C| - 1} \sum_{j, i \neq j} \delta^{inter}_\Delta(D_i, D_j). \quad (4.2)$$

La distribution cumulative inverse de la Δ -densité inter discussions est présentée dans la figure 4.6 pour différentes valeurs de Δ . Bien que similaire, le comportement de la Δ -densité inter discussions diffère qualitativement de celui de la Δ -densité. La Δ -densité inter discussions croît également en fonction de Δ mais il y a toujours une différence notable entre $\Delta = 1 \text{ mois}$ et $\Delta = 1 \text{ an}$ ce qui n'est pas le cas pour la Δ -densité. Cette différence est normale car lors du calcul de Δ -densité le nombre de liens considérés est fixe peut importe Δ alors qu'il croît avec Δ lors du calcul de Δ -densité inter discussions. Un autre facteur est aussi la duré considérée, $t'_\psi - t'_\beta$, qui est plus longue que la durée des discussions.

Afin de comparer plus aisément Δ -densité et Δ -densité inter discussions, la corrélation entre ces deux mesures est présentée dans la figure 4.7 pour différentes valeurs de Δ . On remarque que les discussions sont effectivement plus denses intérieurement qu'avec les autres discussions. La différence est de plusieurs ordres de grandeur lorsque Δ est petit et elle diminue lorsque Δ croît. Pour $\Delta = 20 \text{ ans}$ dans la figure 4.7d, la différence n'est plus visible car à cette échelle de temps, l'ancrage

Faire un dessin de cette situation ?

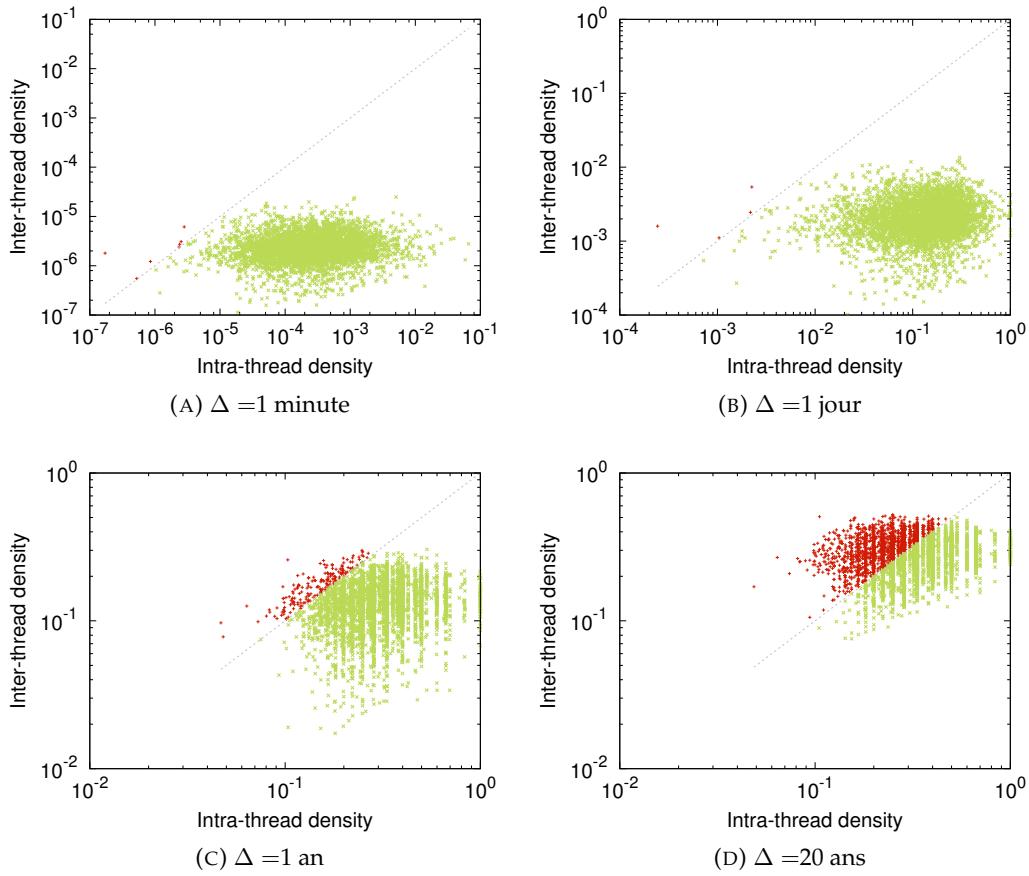


FIGURE 4.7 – Corrélations entre Δ -densité et Δ -densité inter discussions pour différentes valeurs de Δ . Une discussion est en vert (resp. rouge) si elle a une Δ -densité plus (resp. moins) élevée que sa Δ -densité inter discussions.

temporel des discussions n'est plus décisif. On remarque tout de même que pour $\Delta = 1$ an, la différence reste notable.

4.3.2 Répartition temporelle et structurelle des discussions

Nous avons étudié la densité des discussions et entre les discussions mais il est également intéressant d'observer comment ces discussions sont réparties topologiquement et temporellement. Pour étudier la répartition des discussions dans le temps, nous construisons un graphe d'intervalle $\text{REF}_X = (V_X, E_X)$ représentant le chevauchement temporel. Chaque discussion du flot devient un noeud de V_X et le lien (i, j) existe dans E_X si les discussions D_i et D_j correspondantes ont eu lieu au même instant, *i.e.* $[\alpha_i, \omega_i] \cap [\alpha_j, \omega_j] \neq \emptyset$. De manière similaire, nous définissons le graphe de chevauchement topologique $Y = (V_Y, E_Y)$. Les noeuds de ce graphe représentent encore une fois les discussions du flot et un lien existe entre deux discussions si au moins une personne a participé aux deux, *i.e.* $V(D_i) \cap V(D_j) \neq \emptyset$.

Ces deux graphes sont constitués de 116 999 noeuds et d'environ 2 millions de liens pour le graphe de chevauchement temporel et d'environ 63 millions de liens pour le graphe de chevauchement topologique. Par construction, ces graphes contiennent beaucoup d'informations sur les relations entre les discussions.

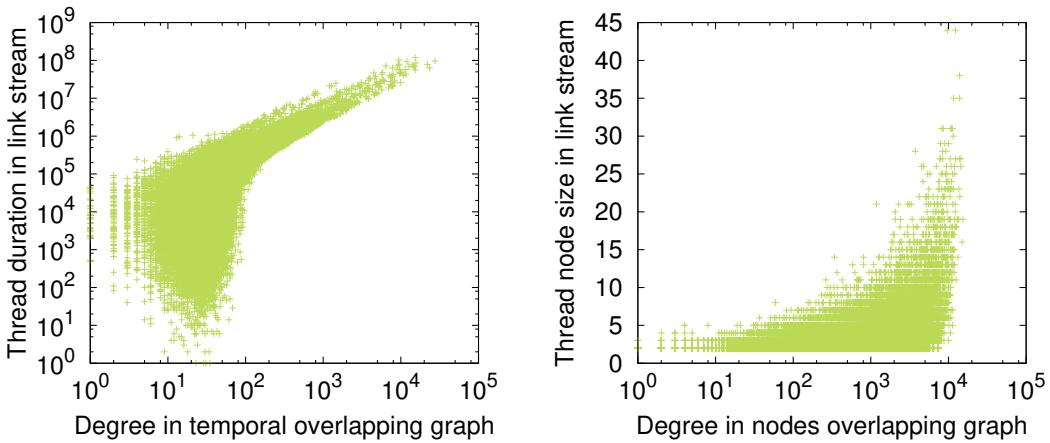


FIGURE 4.8 – Gauche : Corrélation entre le degré des discussions dans le graphe de chevauchement temporel et leur durée. Droite : Corrélation entre le degré des discussions dans le graphe de chevauchement topologique et leur nombre de participants.

Dans la figure 4.8(gauche), est représentée la corrélation entre le degré d'une discussion dans le graphe de chevauchement temporel X et sa durée. Il y a une corrélation évidente entre ces deux notions lorsque les discussions ont une durée supérieur à 10^5 secondes. Plus une discussion dure longtemps, plus elle a de chance d'avoir lieu en même temps que beaucoup d'autres discussions. On observe également que, même pour les discussions durant moins d'un jour (8.6×10^4 s), il peut y avoir jusqu'à une centaine d'autres discussions actives sur la même période.

La figure 4.8(droite) présente la corrélation entre le degré d'une discussion dans le graphe de chevauchement topologique Y et son nombre de participants. La corrélation est moins nette mais il y a tout de même une tendance. Par contre, on remarque de manière frappante que même une petite discussions peut partager des nœuds avec les énormément d'autres discussions.

4.3.3 Flot quotient

Le graphe quotient^{REF} est une autre notion clef pour étudier les relations entre les communautés d'un graphe $G = (V, E)$. Soit une partition $C = \{C_i\}_{1..k}$ des nœuds de G en k communautés, chaque communauté est représentée dans le graphe quotient \bar{G} par un nœuds dans V . Il y a un lien entre deux communauté C_i et C_j dans E si il existe au moins un lien entre un nœuds de C_i et un nœuds de C_j . Voir une illustration sur la figure 4.9. Il est possible d'ajouter un poids sur les liens de \bar{G} égale au nombre de liens reliant les communautés. Le graphe quotient permet de facilement étudier, dans un graphe, les relations entre les communautés.

Nous étendons ici cette notion de graphe quotient aux flots de liens. Nous définissons le flot quotient, $Q = (T_Q, V_Q, E_Q)$, induit par une partition $P = \{P_i\}_{1..k}$ en k sous-flots de la manière suivante. Chaque sous-fLOT P_i est représenté par un nœud dans V_Q . Il existe un lien (t, P_i, P_j) dans E_Q si il existe $(t_1, u, v) \in P_i$, $(t_2, u, v') \in P_j$ et $(t_3, u, v'') \in P_i$ avec $t_1 \leq t_2 \leq t_3$. En d'autre termes, il y a un lien dans E_Q si un nœud u a un lien dans P_j qui apparaît entre deux autres de ses liens du groupe P_i .

Le flot quotient induit par les discussions dans le jeu de données contient 12 281 269 liens impliquant 68 524 discussions différentes. Comme le jeu de données contient 116 999 discussions, il y a donc 48 475 discussions sans lien et qui ne seront pas prises en compte par la suite. Ce nombre de discussions non-reliées est

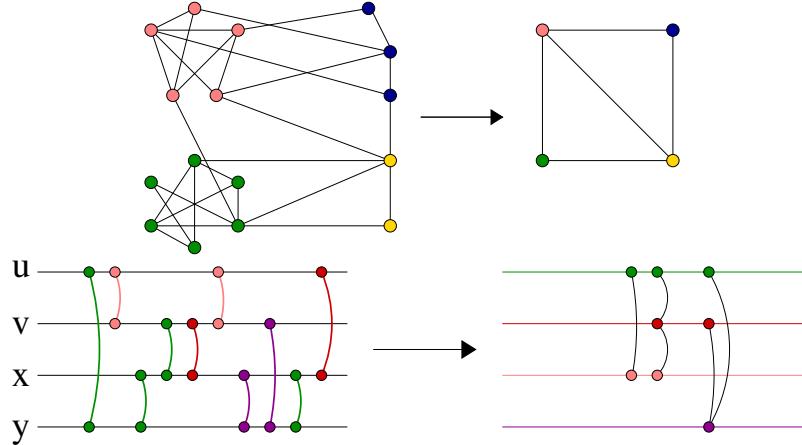


FIGURE 4.9 – Haut : Exemple de graphe ayant une structure communautaire et son graphe quotient associé. Bas : Exemple d'un flot de lien avec une structure ainsi que son flot quotient associé.

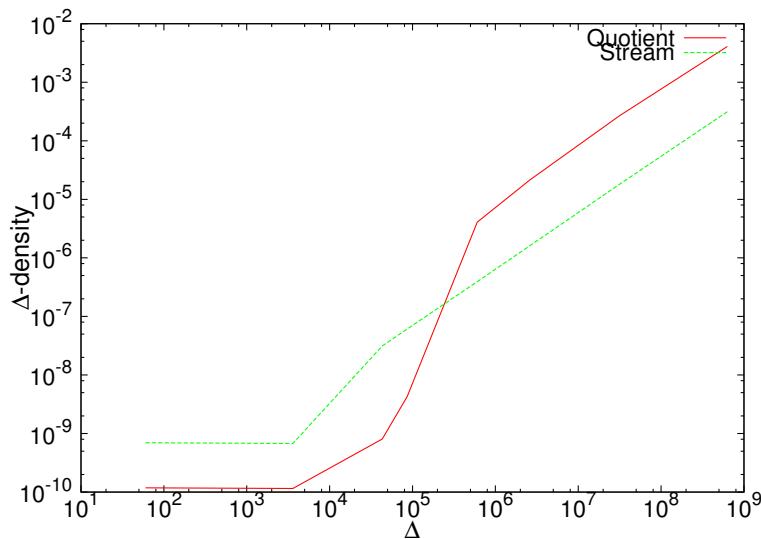


FIGURE 4.10 – Δ -densité du flot de liens et du flot de liens quotient en fonction de Δ pour $\Delta = 1mn, 1h, 12h, 1j, 7j, 30j, 1\text{ an}$ et 20 ans .

élevé comparé à ce qui est obtenu dans un graphe. En effet dans un graphe, un nœud de degré 0 correspond à une communauté qui est une composante connexe (ou un union de composantes connexes). En ajoutant l'information temporelle, les discussions sont séparées par le temps dans flot. C'est pourquoi un grand nombre de discussions n'ont pas de liens dans le flot quotient. Ce phénomène est d'autant plus vrai pour les petites discussions.

Il faut aussi noter qu'il y a environ 20 fois plus de liens dans le flot quotient que dans le flot initial. Cela est normal car un lien dans le flot peut donner lieu à plusieurs liens dans le flot quotient. Ce cas est visible dans la figure 4.9. Le lien (x, y) du groupe violet du flot à gauche donne lieu au lien (*violet*, *rouge*) et au lien (*violet*, *vert*) dans le flot quotient à droite.

La figure 4.10 présente la Δ -densité du flot de liens initial et du flot quotient pour différentes valeurs de Δ . Le flot quotient est plus Δ -dense que le flot initial pour des valeurs de Δ importantes. Ce résultat est comparable à ce qui est obtenu dans un graphe.

4.3.4 Conclusion

Nous avons utilisé le modèle de flot de liens pour étudier une archive de courriels provenant du projet Debian. Grâce au modèle de flot de liens, nous avons étudié des notions clefs pour mieux comprendre la répartition temporelle et topologique des discussions. Nous avons étudié la notion de Δ -densité sur les discussions en elles mêmes. Puis, nous avons étudié les relations entre les discussions avec la Δ -densité inter discussions, les projections en graphe de chevauchement temporel ou topologique et le flot quotient.

Cette étude repose en grande partie sur la notion de Δ -densité qui nécessite un paramètre fixé arbitrairement. Nous avons à chaque fois testé un ensemble de valeurs de Δ variant d'une seconde jusque parfois 20 ans et, lors de ces tests, aucune valeur Δ caractéristique n'a pu être identifiée. Il semble donc que la Δ -densité soit relativement robuste vis-à-vis de Δ dans ce contexte.

Nous avons tout d'abord observé que les discussions forment une structure plus dense que le flot de liens. De manière encore plus forte, nous avons constaté, grâce à la Δ -densité inter discussion, que les discussions sont plus denses en interne qu'en externe. C'est une caractéristique importante des communautés que l'on trouve dans les graphes mais qui n'avait pas été observée dans un contexte temporel. À partir de ces observations, nous avons également observé les relations entre les discussions. Via le graphe de chevauchement temporel, nous avons validé le fait que différentes discussions ont lieu en même temps et que par conséquent une agrégation temporelle entraînerait une perte d'information. De même via le graphe de chevauchement topologique, on remarque que la structure est très recouvrante sur les nœuds, rendant ainsi l'utilisation de partitions statiques de nœuds difficilement envisageable pour décrire les discussions.

On a pas étudié de mesure spéciale graphe sur X ou Y.

Faire évaluation des threads comme pertinent.

4.4 Détection de structures denses

À partir du constat que les discussions forment une structure particulière, il est naturel d'essayer de les retrouver automatiquement. Pour y parvenir, il faut un moyen capable de trouver des sous-flots-denses dans le flots. C'est à dire une méthode capturant des groupes de liens qui soient proche temporellement et topologiquement. Il serait tentant de directement d'optimiser la densité dans le flot mais ce n'est pas envisageable car un groupe constitué d'un unique lien a une densité de 1. Il faut donc trouver une autre méthode. C'est pourquoi nous avons construit une autre projection du flot en un graphe statique afin d'y appliquer une méthode de détection de communautés. Le problème est alors de réussir à créer une transformation de telle sorte que les informations temporelles et topologiques ne soient pas complètement détruites.

4.4.1 Méthode de détection

Dans la transformation que nous appliquons, nous créons un graphe non-orienté et non-pondéré $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Chaque lien du flot est représenté par un nœud. Deux liens (b, e, u, v) et (b', e', u', v') sont connectés dans le graphe s'ils partagent un nœud, i.e. $\{u, v\} \cap \{u', v'\} \neq \emptyset$, et si les intervalles s'intersectent, i.e. $[b, e] \cap [b', e'] \neq \emptyset$, voir figure 4.11. Ainsi, un lien dans le graphe représente une connexion structurelle et

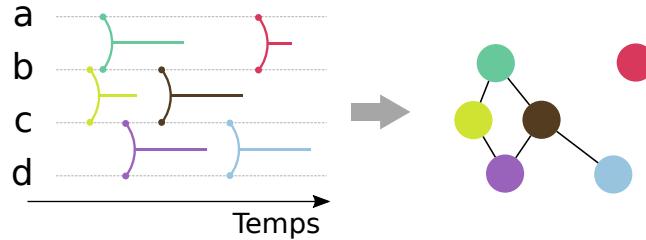


FIGURE 4.11 – Transformation d'un flot de liens avec 4 noeuds (a-d) et 6 liens à gauche en un graphe à droite à 6 noeuds. La couleur d'un noeuds dans le graphe indique le lien du flot qu'il représente.

temporelle entre deux liens du flot de liens. Les groupes denses dans le graphe représentent donc des groupes de liens connectés temporellement et topologiquement dans le flot.

Cette définition n'est valide que pour un flot de liens avec durée. Or, les courriels échangés n'ont pas de durées. Lors du calcul de la densité dans la section 4.3.1, nous avions ajouté une durée arbitraire Δ . Ici, il n'est pas très pertinent d'appliquer la même logique. En effet, si on utilise un Δ faible, alors il n'y aura que très peu de liens dans \mathcal{E} et les noeuds représentant les liens d'une discussions ne seront pas forcément connexes. Il paraît illusoire d'espérer retrouver les discussions dans \mathcal{G} si elles ne sont même pas connexes. C'est pourquoi nous adoptons une autre manière d'ajouter une durée sur les liens.

Pour chaque message m , nous connaissons $p(m)$, le message auquel il répond dans la discussion. Nous définissons un autre flot, \mathfrak{L} , où les liens représentant les messages sont de la forme $(t(p(m)), t(m), a(m), a(p(m)))$. Ainsi, deux messages, m_1 et m_2 , se succédant dans une discussion sont par définition reliés topologiquement car $a(m_1) = a(p(m_2))$. Ces deux messages sont aussi reliés temporellement car nous avons la relation suivante :

$$\begin{aligned} [t(p(m_1)), t(m_1)] \cap [t(p(m_2)), t(m_2)] &= \\ [t(p(m_1)), t(m_1)] \cap [t(m_1), t(m_2)] &= [t(m_1)] \neq \emptyset. \end{aligned}$$

Par construction, une discussion est donc représentée dans \mathcal{G} par un ensemble connexe de noeuds. Un fois \mathcal{G} construit, on peut appliquer un algorithme de détection de communautés.

Avec cette construction, \mathcal{G} contient plus d'1 millions de liens pour les 116 999 discussions présentes. Sur ce graphe, nous avons appliqué l'algorithme de Louvain [1] qui optimise la modularité. D'autres algorithmes peuvent également être appliqués s'ils capturent des groupes de noeuds disjoints et qu'ils passent à l'échelle. Les groupes trouvés par Louvain sont des communautés dans \mathcal{G} . Par conséquent, ils sont censé être densément connectés dans \mathcal{G} . Comme un lien de \mathcal{G} correspond à une connexion temporelle et topologique dans le flot, on peut espérer qu'ils correspondent à des groupes denses dans le flot.

4.4.2 Comparaison des partitions

Avant de comparer la structure des discussions, D , et la partition, \mathfrak{D} , trouvée par la méthode de Louvain sur \mathcal{G} , il est nécessaire de décrire cette dernière. Dans la figure 4.12, les distributions cumulatives inverses du nombre de liens, du nombre de noeuds et de leur durée sont présentées pour les groupes de \mathfrak{D} . Pour rappel, les mêmes données sont représentées pour les discussions. On remarque tout de suite

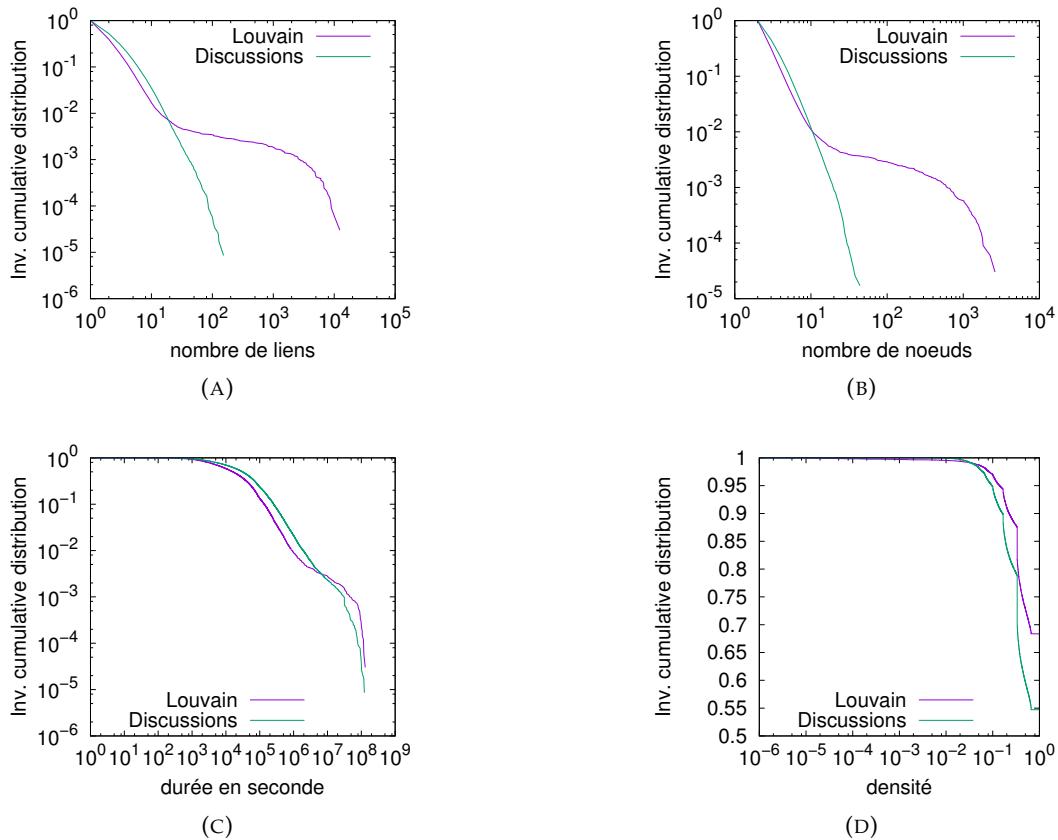


FIGURE 4.12 – Distribution cumulative inverses du nombre de liens (a), du nombre de noeuds (b), de la durée (c) et de la densité (d) pour les groupes trouvés par Louvain et les discussions.

que \mathfrak{D} contient des groupes beaucoup plus gros en nombre de noeuds et de liens alors qu'ils ont les mêmes durées.

Ces deux structures sont donc très différentes mais cela pourrait être dû à l'algorithme de Louvain qui n'est pas adapté pour trouver des groupes denses. C'est pourquoi, nous avons également observé la densité des groupes de D et \mathfrak{D} dans le flot \mathcal{L} . Le résultat est visible dans la figure 4.12d. Comme les liens de \mathcal{L} ont une durée, il est possible d'utiliser directement la densité au lieu de la Δ -densité utilisée précédemment. On remarque que les groupes de \mathfrak{D} , bien que plus gros, sont plus denses que les groupes de D . Cependant la distribution cumulative inverse cache les effets de la taille sur la densité. Or à taille égale (entre 2 et 160) liens, on remarque que les groupes trouvés par Louvain sont légèrement plus dense en médian, 0.34 contre 0.33, et également plus denses en moyenne ,0.46 contre 0.38.

En revanche, les plus gros groupes ($|\mathfrak{D}_j| > 160$) trouvés par Louvain ont une densité plus faible mais c'est attendu à cause de leur taille.

Si les groupes de \mathfrak{D} sont plus denses, c'est peut être car ils regroupent plusieurs discussion de D dans un groupe. Pour comparer deux partitions, l'indice de Jaccard est classiquement utilisé pour calculer la *précision* et le *rappel* qui sont définis de la manière suivante :

$$\text{précision}(D_i) = \max_j \frac{|\mathfrak{D}_j \cap D_i|}{|\mathfrak{D}_j|}, \quad \text{rappel}(D_i) = \max_j \frac{|\mathfrak{D}_j \cap D_i|}{|D_i|}.$$

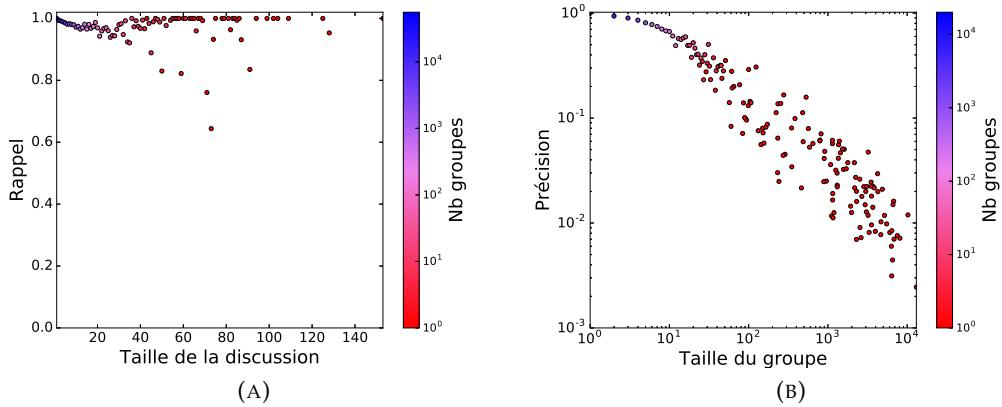


FIGURE 4.13 – Précision (B) et rappel (A) entre les discussions et les groupes trouvés par Louvain en fonction de la taille. Chaque point représente la moyenne pour les groupes ayant la même taille. La couleur du point indique le nombre de groupes ayant la même taille.

Dans la figure 4.13, est présenté la *précision* et le *rappel* des discussions en fonctions de leurs tailles. Chaque point représente la moyenne du *rappel* et de *précision* pour les groupes d'une taille donné. On voit qu'il y a un important rappel et ceux même pour les grandes discussions, ce qui veut dire qu'en générale une discussion D_i est totalement incluse dans un groupe \mathfrak{D}_j . En revanche, la précision est très faible car un groupe \mathfrak{D}_j contient plusieurs discussions, ce qui est cohérent avec la taille très importante des groupes de \mathfrak{D}_j .

Il semble donc que la partition \mathfrak{D} soit proche de D mais que ces groupes soient plus gros. Pour circonvenir à ce problème, nous appliquons de manière récursive l'algorithme de Louvain sur chaque graphe induit par un groupe \mathfrak{D}_j . Ce processus permet de subdiviser de manière récursive chaque groupe \mathfrak{D}_j . Par construction, un groupe trouvé au niveau h , $\mathfrak{D}_j(h)$, est donc inclus dans un groupe trouvé au niveau $h - 1$, c'est à dire $\mathfrak{D}_j(h) \subset \mathfrak{D}_j(h - 1)$. Le niveau 0 est la première partition trouvée par l'algorithme de Louvain sur le \mathbb{G} .

Soit $D_i \in D$ et $\mathfrak{D}_{\tilde{j}}(h)$ avec $h \in \mathbf{N}$, le groupe trouvé par la méthode de Louvain au niveau h qui soit le plus proche de D_i au niveau h , c'est-à-dire $|\mathfrak{D}_{\tilde{j}}(h) \cap D_i| = \max_j |\mathfrak{D}_j(h) \cap D_i|$. Avec ces définitions, on observe la relation suivante : $\mathfrak{D}_{\tilde{j}}(h) \cap D_i \subseteq \mathfrak{D}_{\tilde{j}}(h-1) \cap D_i$. La définition de *rappel* de l'équation 4.4.2 n'est donc pas adapté pour les niveaux inférieurs et nous l'adaptons de la manière suivante :

$$rappel(D_i, h) = \max_j \frac{|\mathfrak{D}_j(h) \cap D_i|}{|D_i \cap \mathfrak{D}_{\tilde{j}}(h-1)|}. \quad (4.3)$$

Ainsi, le *rappel* au niveau h prends en compte le maximum d'élément qu'il est possible de trouver à ce niveau. La définition de *précision* ne pose quant à elle pas de problème. Dans la figure 4.14, sont représentés le *rappel* adapté et la *précision* pour le premier et deuxième niveau de récursion de la même manière que pour la figure 4.13. On remarque que, dès le premier niveau, le rappel baisse et que ce phénomène s'amplifie fortement au niveau suivant. Cela implique que les discussions ne sont plus incluses dans un groupe mais au contraire réparties dans plusieurs. La précision quant à elle augmente légèrement mais cela est dû à la baisse de la taille des groupes trouvés. Il semble donc qu'il ne soit pas possible avec cette

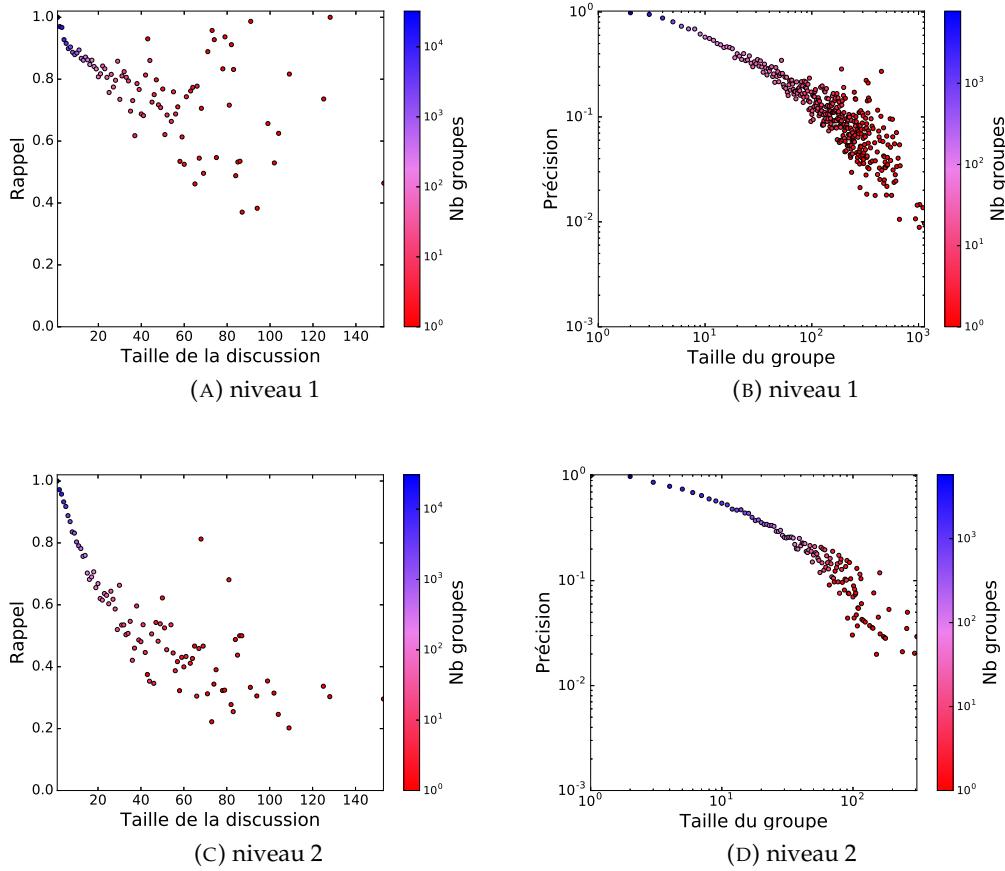


FIGURE 4.14 – Précision (B,D,F) et rappel (A,B,C) entre les discussions et les groupes trouvés par Louvain à différent niveaux récursifs. Chaque point représente la moyenne pour les groupes ayant la même taille. La couleur du point indique le nombre de groupes ayant la même taille.

approche de retrouver automatiquement les discussions.

4.4.3 Conclusion

Nous avons avec cette méthode mis en évidence des groupes denses. Les groupes trouvés sont plus gros et plus denses que la structure des discussions. Cependant, ces observations ne remettent pas en cause les conclusions faites dans la section 4.3 pour plusieurs raisons. Tout d'abord, les flots de lien étudiés ne sont pas exactement les-mêmes ($L \neq \mathcal{L}$). Ce changement de flot est nécessaire pour le fonctionnement de la méthode de détection. Ensuite, les deux structures ne sont pas complètement différentes car les groupes trouvés semblent en fait agréger plusieurs discussions. Malheureusement, nous n'avons pas réussi avec notre méthode à isoler chaque discussion malgré notre approche récursive. Pourtant, il semble que la structure trouvée ai du sens au vu des valeurs de densité des groupes.

Chapitre 5

Détection de groupes denses (SNAM)

Chapitre 6

Fonction de qualité

6.1 Définition

**6.2 Générateur de flots de liens avec structure communau-
taire**

Chapitre 7

Conclusion

Bibliographie

- [1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2008(10) :P10008, oct 2008.
- [2] Remi Dorat, Matthieu Latapy, Bernard Conein, and Nicolas Auray. Multi-level analysis of an interaction network between individuals in a mailing-list. *Ann Telecommun*, 62 :325–349, 2007.
- [3] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5) :75–174, feb 2010.

Table des figures

4.1	Distribution cumulative inverse de différentes caractéristiques pour les données brutes (ligne pleine) et filtrées (ligne pointillée). En haut à gauche : nombre de courriels dans une discussion ; en haut à droite : durée d'une discussion ; en bas à gauche : nombre de personnes dans une discussion ; en bas à droite : nombre de paires d'auteurs distinct dans une	13
4.2	Gauche : Corrélations entre le nombre de courriels et la durée d'une discussion. Droite : Corrélation entre le nombre de courriels et le nombre d'auteurs dans une discussion.	13
4.3	Distribution des temps inter-contacts dans le fil de discussions.	14
4.4	Évolution de la Δ -densité du flot de liens pour Δ de 1 seconde à 20 ans. <i>Ajout bar horizontal de la densité statique dans le graphe agrégé.</i>	15
4.5	Distribution cumulative inverse de la Δ -densité des discussions pour différentes valeurs de Δ s.	15
4.6	Distribution cumulative inverse de la Δ -densité inter discussions pour différentes valeurs de Δ s.	16
4.7	Corrélations entre Δ -densité et Δ -densité inter discussions pour différentes valeurs de Δ . Une discussion est en vert (resp. rouge) si elle a une Δ -densité plus (resp. moins) élevée que sa Δ -densité inter discussions.	17
4.8	Gauche : Corrélation entre le degré des discussions dans le graphe de chevauchement temporel et leur durée. Droite : Corrélation entre le degré des discussions dans le graphe de chevauchement topologique et leur nombre de participants.	18
4.9	Haut : Exemple de graphe ayant une structure communautaire et son graphe quotient associé. Bas : Exemple d'un flot de lien avec une structure ainsi que son flot quotient associé.	19
4.10	Δ -densité du flot de liens et du flot de liens quotient en fonction de Δ pour $\Delta = 1mn, 1h, 12h, 1j, 7j, 30j, 1\text{ an}$ et 20 ans	19
4.11	Transformation d'un flot de liens avec 4 nœuds (a-d) et 6 liens à gauche en un graphe à droite à 6 nœuds. La couleur d'un nœuds dans le graphe indique le lien du flot qu'il représente.	21
4.12	Distribution cumulative inverses du nombre de liens (a), du nombre de nœuds (b), de la durée (c) et de la densité (d) pour les groupes trouvés par Louvain et les discussions.	22
4.13	Précision (B) et rappel (A) entre les discussions et les groupes trouvés par Louvain en fonction de la taille. Chaque point représente la moyenne pour les groupes ayant la même taille. La couleur du point indique le nombre de groupes ayant la même taille.	23
4.14	Précision (B,D,F) et rappel (A,B,C) entre les discussions et les groupes trouvés par Louvain à différent niveaux récursifs. Chaque point représente la moyenne pour les groupes ayant la même taille. La couleur du point indique le nombre de groupes ayant la même taille.	24