Conversations, Groupes et Communautés dans les Flots de Liens

Noé Gaumont

19 avril 2016

Table des matières

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

[2]

1.2 Extension temporelle

Chapitre 2

Extensions temporelle des graphes

2.1 Ètat de l'art des formalismes et outils

2.2 L'approche flot de liens

2.2.1 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 $\sigma = L$.

$V(E') = u_{(b,e,u,v)\ inE'}$ 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 ?

2.2.2 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 $[\alpha', \omega']$.

2.2.3 Degré et densité

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

$$d_t(u) = |L_t(v)| = |\{(b, e, u, v) \in E,\ b \leq t \leq e\}| \tag{2.1}$$

Par extension pour un ensemble de sommets V' :

$$d_t(V') = |L_t(V'^2)| = |\{(b, e, u, v) \in E,\ u, v \in V',\ b \leq t \leq e\}| \tag{2.2}$$

Degré interne maintenant ?

$$d_t(E') = |\{(b, e, u, v) \in E',\ b \leq t \leq e\}| \tag{2.3}$$

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

$$D_{\alpha..\omega}(v) = \int_{\alpha}^{\omega} d(v, t)dt = D(v) \tag{2.4}$$

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 :

$$\delta(L) = \frac{2 \sum_{l \in E}}{n(n-1)(\bar{L})} \tag{2.5}$$

Symbole	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
$L(V'^2)$	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

2.3 Manipulation de flots

Existence de la lib

2.3.1 Représentation

Chapitre 3

Étude de la structure d'un flot de liens réel

L'étude de la structure des réseaux est un sujet qui est étudié depuis assez longtemps 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 surement trop général qui sera bougée ailleurs.

Nous nous intéressons ici à une archive de courriels publiquement disponibles [1]. Cette archive contient l'ensemble des courriels échangés par différent 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. Toutes personnes inscrites 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 [1] mais cela a été fait en utilisant des méthodes statique uniquement.

Hors, ces données se transposent facilement sous forme de flot de liens en transformant une personne en un nœuds et chaque courriel entre deux personnes en 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 dans laquelle a lieu chaque message. Une discussion est un ensemble de courriels dont tout les messages répondent à un message précédant 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 liens représentant les courriels envoyés sur la liste.

Utiliser le formalisme de flot de liens est particulièrement intéressant car

1. `https://lists.debian.org/debian-user/`

cette lite de diffusion existe depuis 1994. L'aspect temporelle des discussions est donc important.

3.1 Prétraitement sur le jeu de données

Bien que accessible sur internet, ce jeu de données nécessite un ensemble de traitement 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 flots 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. Pour chaque message m, nous extrayons son auteur $a(m)$, l'instant $t(m)$ auquel le message a été posté [2], le message auquel il réponds $p(m)$ trouvé via le champ In-Reply-To, son destinataire $a(p(m))$ et la discussion $D(m)$ dans laquelle il apparait. Comme les messages *racines* ne répondent à aucun autre, nous imposons $p(m) = a(m)$. L'ensemble de liens du flot est donc $\{t(m), a(m), a(p(m))\}_m$.

Une fois le flot créé, il est encore nécessaire de vérifier sa cohérence. Un message peut être filtré pour différentes raisons : le courriel apparait avant le message auquel il est censé répondre, le message auquel il réponds n'est pas présent dans l'archive, l'auteur et le destinataire sont la même personne [3]. Cette dernière condition permet notamment d'éviter la présence de boucle dans le flot. Il s'agit de vérifications simple auquel il faut ajouter les vérification de la cohérence de la structure des discussions. Ainsi, une discussion est entièrement retirée du jeu de données si il manque la *racine*,si un de ces messages a été retiré à l'étape précédante 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 en observant la distribution de durées des discussions, voir la figure 3.1.

Une fois tout ces messages filtrés, nous obtenons un flot de liens avec 554233 liens entre 34648 personnes pendant presque 19 ans et 116999 discussions. Mis à part les messages de début de discussions, se sont 53753 courriels qui ont été filtrés soit environ 7%.

2. Cet instant est convertit en *timestamp* en tenant compte des fuseaux horaires.
3. Cela concerne principalement les *racines*

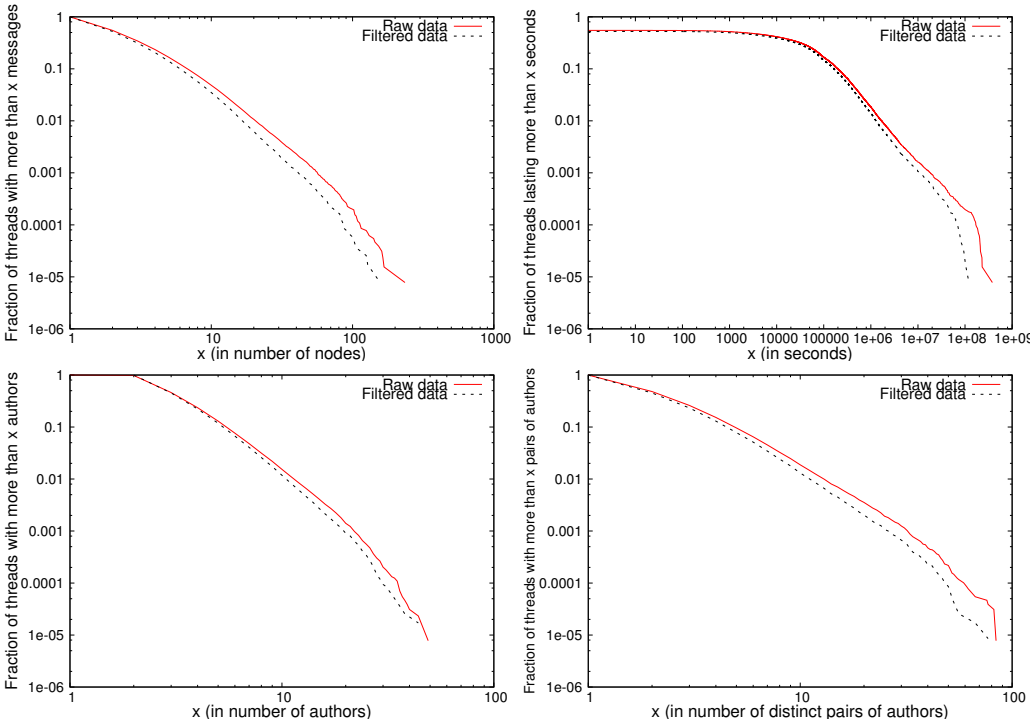


FIGURE 3.1 – Distribution cumulative inverse de différentes caractéristiques pour les données brutes (ligne pleine) et filtrées (ligne pointillé). En haut à gauche : nombre de courriels dans une discussion ; en haut à droite : durée d'une discussion ; en bas à gauche : nombre de personnes dans une discussion ; en bas à droite : nombre de paires d'auteurs distinct dans une .

3.2 Caractéristiques basiques des discussions

Les caractéristiques les plus basiques des discussions sont les nombres de courriels, les nombres de personnes, les nombres de paires de personnes distincte en interaction direct et leurs durées. Sur la figure 3.1, sont présentées les distributions cumulatives inverses de ces notions 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 équivaux 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 suffisante pour comprendre les caractéristiques d'une discussion. Nous avons également étudié la

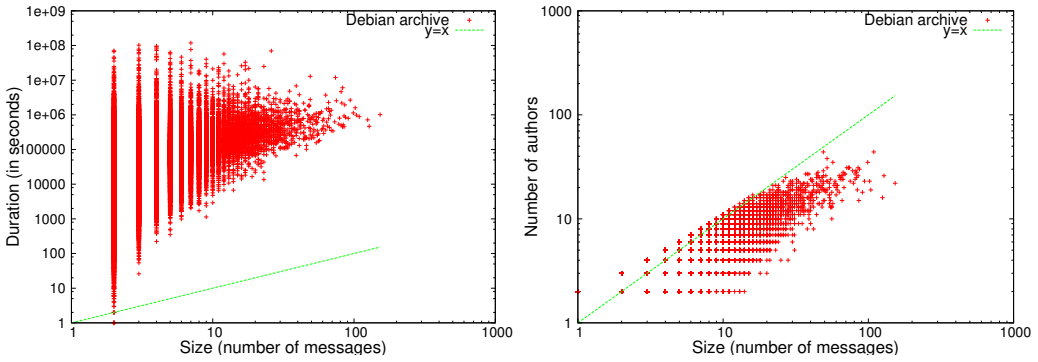


FIGURE 3.2 – Gauche : Corrélations entre le nombre de courriels et la durée d'une discussion. Droite : Corrélation entre le nombre de courriels et le nombre d'auteurs dans une discussion.

corrélation de ces différentes notions et une partie d'entre elles sont présentées sur la figure 3.2.

La corrélation entre la durée et le nombre de courriels, sur la figure 3.2 partie gauche, mets 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. Sur la partie droite de la figure 3.2 présentant la corrélation entre le nombre de courriels et d'auteurs, on observe un autre fait attendu [1] qui est qu'une discussion est constitué de plus de messages que de participants. Ainsi lors d'une discussion, c'est un petit nombre de personnes qui partagent 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 de 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. Sur la figure 3.3 est représentée la distribution cumulative inverse du temps inter-contacts. On remarque que 21% des temps inter-contacts est inférieurs à 30 jours. 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 particuliers, une personne peut envoyer une demande d'aide à un moment donner et ne plus jamais échanger avec les même personnes. Or, on observe que 21% des contacts sont renouvelés en moins de 30 jours. La participation est donc quand relativement élevée.

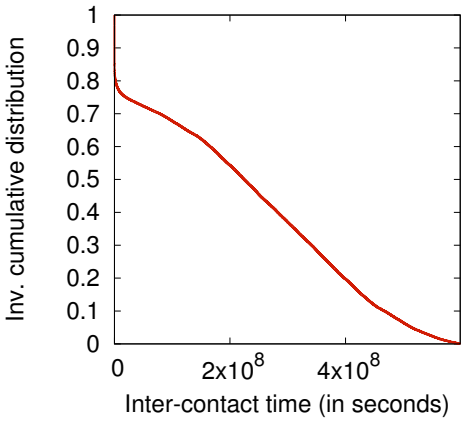


FIGURE 3.3 – Distribution des temps inter-contacts dans le fil de discussions.

3.3 Étude des discussions en tant que sous-flots

3.3.1 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 ou chaque lien n'a pas de durée, nous étudions la Δ-densité pour différentes valeurs de Δ entre 1 seconde et 20 ans. Tout d'abord sur la figure 3.4 est représenté 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 sur la figure 3.5, est présentée la distribution cumulative inverse de la Δ-densité des discussions pour différentes valeurs de Δ. On remarque que les différentes valeurs de Δ ne semblent pas influencer qualitativement la distribution de Δ-densité. Cette courbe mets 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

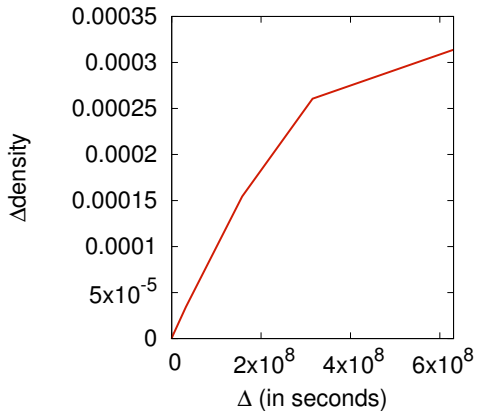


FIGURE 3.4 – Évolution de la Δ-densité du flot de liens pour Δ de 1 seconde à 20 ans. Ajout bar horizontal de la densité statique dans le graphe agrégé.

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 élevé que celle du flot. Bien que notable, ce fait est attendu notamment car le flot dure beaucoup plus longtemps et concerne beaucoup plus de nœuds 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 tant dis 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 connecté à l'intérieur qu'avec les autres communautés adjacentes. Pour un graphe $G = (V, E)$ 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, $\delta^{inter}(C_i)$:

$$\delta^{inter}(C_i) = \frac{1}{|C| - 1} \sum_{j, i \neq j} \frac{|\{(u, v) \in E \text{ t.q. } u \in C_i \text{ and } v \in C_j\}|}{|C_i| \cdot |C_j|}. \tag{3.1}$$

Il s'agit tout simplement de la probabilité qu'un lien existe entre les nœuds 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

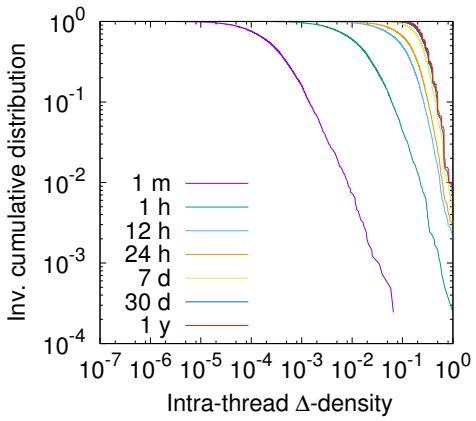


FIGURE 3.5 – Distribution cumulative inverse de la Δ-densité des discussions pour différentes valeurs de Δs.

faire, nous définissons la Δ-densité inter discussions entre deux discussions D_i et D_j : $\delta_\Delta^{inter}(D_i, D_j)$. 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 $\delta_\Delta^{inter}(D_i, D_j)$ est la suivante : $\delta_\Delta^{inter}(D_i, D_j) = d(L'_{t'_\beta..t'_\psi}(V'))$. Il s'agit donc de la Δ-densité du flot de liens induit par l'union des nœuds 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 :

Faire un dessin de cette situation ?

$$\delta_\Delta^{inter}(D_i) = \frac{1}{|C| - 1} \sum_{j, i \neq j} \delta_\Delta^{inter}(D_i, D_j). \tag{3.2}$$

La distribution cumulative inverse de la Δ-densité inter discussions est sur la figure 3.6 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 croit également en fonction de Δ mais il y a toujours une différence notable entre $\Delta = 1$ *mois* et $\Delta = 1$ *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 import Δ alors qu'il croit 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 sur la figure 3.7 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

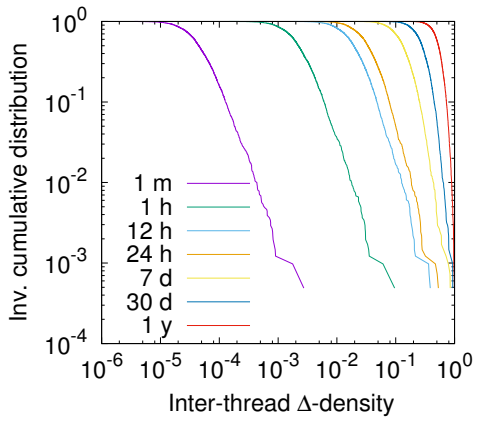


FIGURE 3.6 – Distribution cumulative inverse de la Δ-densité inter discussions pour différentes valeurs de Δs.

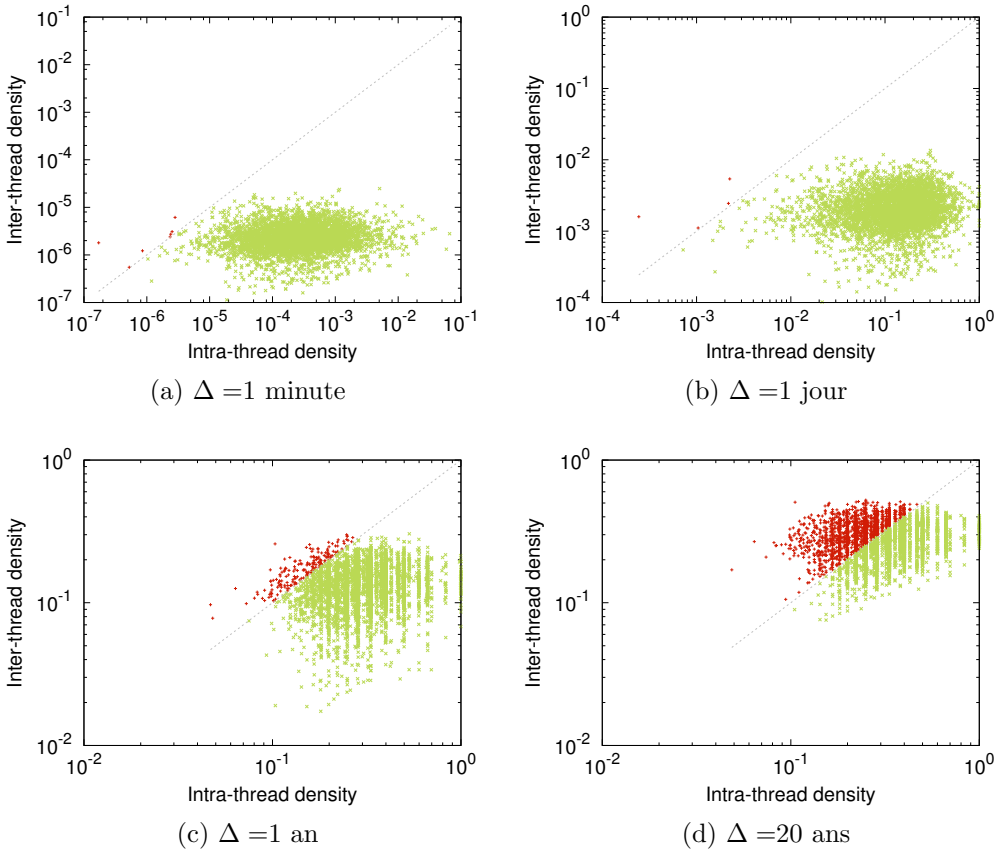
croit. Pour $\Delta = 20$ *ans* sur la figure 3.7d, la différence n'est plus visible car à cette échelle de temps, l'ancrage temporel des discussions n'est plus décisif. On remarque tout de même que pour $\Delta = 1$ *an*, la différence reste notable.

3.3.2 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 REF$X = (V_X, E_X)$ représentant le chevauchement temporel. Chaque discussion du flot devient un nœud 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 nœuds de se 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 116999 nœuds et de environ 2 millions de liens pour le graphe de chevauchement temporel et de 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.

Sur la figure 3.8(gauche), est représentée la corrélation entre le degré d'une discussion dans le graphe de chevauchement temporel X et sa durée.



(a) Δ =1 minute (b) Δ =1 jour

(c) Δ =1 an (d) Δ =20 ans

FIGURE 3.7 – 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.

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 jours (8.6×10^4 s), il y a jusqu'à une centaine d'autres discussions actives sur la même période.

La figure 3.8(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 partagent énormément de nœuds avec les autres discussions.

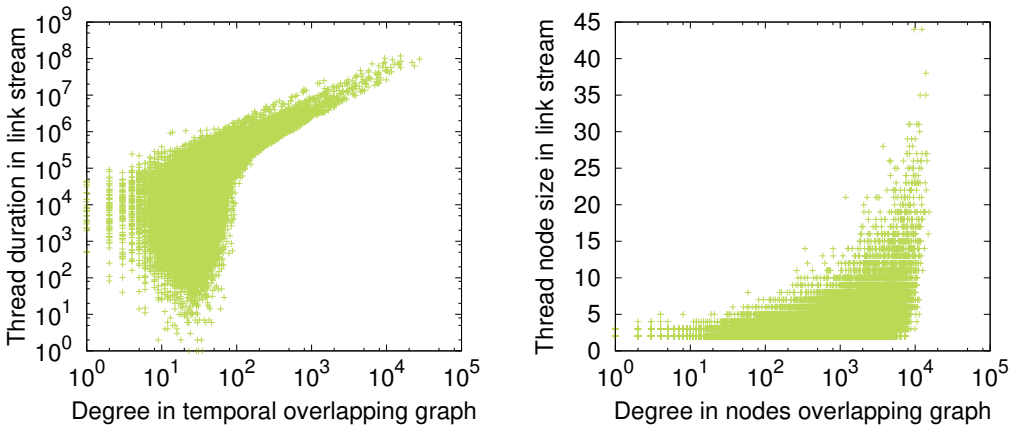


FIGURE 3.8 – Gauche : Corrélation entre le degré des discussions dans le graphe de chevauchement temporel et leur durée. Droite : Corrélation entre le degré des discussions dans le graphe de chevauchement topologique et leur nombre de participants.

3.3.3 Flot quotient

Le graphe quotientREF 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 3.9. 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 apparait entre deux autres de ses liens du groupe P_i.

Le flot quotient induit par les discussions dans le jeu de données contient 12281269 liens impliquant 68524 discussions différentes. Comme notre jeu de données contient 116999 discussions, il y a donc 48475 discussions étant reliées à aucune autres et qui ne seront pas pris en compte par la suite. Ce nombre de discussions non-reliées est élevé comparé à ce qui est obtenu dans un graphe. En effet dans un graphe, un nœuds non-relié corresponds à

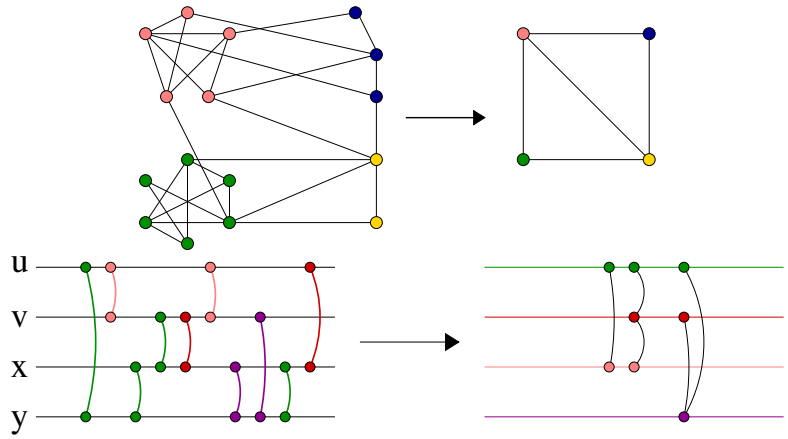


FIGURE 3.9 – Haut : Exemple de graphe ayant une structure communautaire et son graphe quotient associé. Bas : Exemple d'un flot de lien avec une structure ainsi que son flot quotient associé.

une communauté qui soit une composante connexe. En ajoutant l'information temporelle, les discussions sont séparées par le temps dans flot. C'est pourquoi un grand nombre de discussion ne sont pas reliées dans le flot quotient. Ce phénomène est d'autant plus présent 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 sur la figure 3.9. 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 3.10 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 importante. Ce résultat est comparable a ce qui est obtenu dans un graphe.

3.3.4 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 temporel et topologique des discussions. Nous avons étudié la notion de Δ-densité sur les discussions en elle même. 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

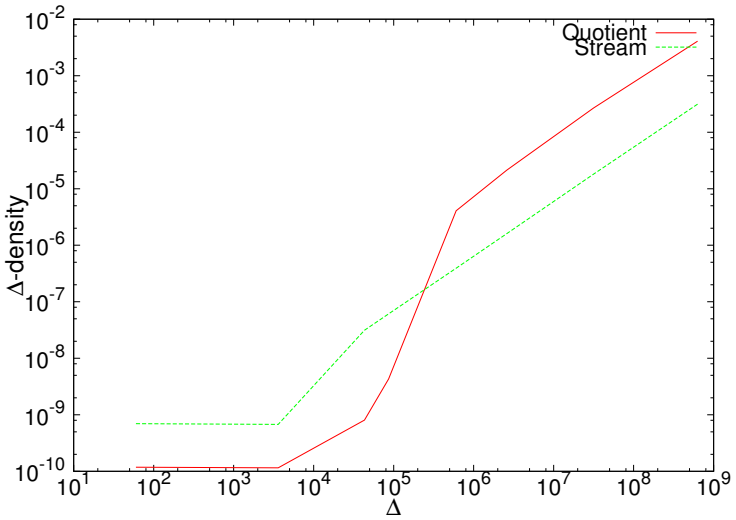


FIGURE 3.10 – Δ-densité du flot de liens et du flot de liens quotien en fonction de Δ pour $\Delta = 1mn, 1h, 12h, 1j, 7j, 30j, 1\ an$ et 20 ans.

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 valeurs 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 dense 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 temporel entrainerait 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.

3.4 Détection automatique des discussions ?

Chapitre 4

Détection de groupes denses (SNAM)

Chapitre 5

Expected Nodes (COMPLENET)

Chapitre 6

Fonction de qualité

6.1 Définition

6.2 Générateur de flots de liens avec structure communautaire

Chapitre 7

Conclusion

Bibliographie

[1] 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.

[2] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5) :75–174, feb 2010.

Table des figures