Conversations, Groupes et Communautés dans les Flots de Liens

Noé Gaumont

8 avril 2016

Table des matières

Todo list

Chapitre 1

Introduction

1.1 Communauté dans les graphes

[2]

1.2 Extension temporelle

Chapitre 2

Modèle proche et outils pour les étudier

Chapitre 3

Étude de la structure d'un flot de liens

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.

Nous nous intéressons à une archive de courriels publiquement disponibles [1]. L'étude de fil de discussions a déjà étudié dans le passé [1] mais cela a été fait en utilisant des méthodes statiquement. Cette archives l'ensemble des courriels échangés par différent utilisateurs 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 cette demande et peut y répondre. Ces données se transposent facilement sous forme de flot de liens car une personne est un nœuds et chaque courriel entre deux personnes donne lieu à un lien dans le flot à l'instant où le courriel a été envoyé. Le message initiant la conversation est ignoré pour éviter les boucles. 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 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 cette lite de diffusion existe depuis 1994. L'aspect temporelle des discussion

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

est donc important.

3.1 Prétraitement sur le jeu de données

Bien que disponible, 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. 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 est nécessaire de vérifier leurs cohérences. Pour chaque courriel, sont connus d'envoie, l'auteur, le destinataire, le courriel auquel il répond et la discussion dans laquelle il apparait.

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[2]. Il s'agit de vérifications simple auquel il faut ajouter les vérification de la cohérence de la structure de discussions. Ainsi, une discussion est retirée du jeu de donnée si il manque la racine, un de ces messages a été retiré précédemment 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 courriels de début de discussions, ce sont 53753 courriels qui ont été filtrés soit environ 7%.

Parler de la transformation données vers formalisme flot

3.2 Caractéristiques basiques des discussions

Les caractéristiques les plus basiques des discussions sont le nombre de courriels, le nombre de personnes, le nombre de paires de personnes distincte en interaction direct et leur durée. Sur la figure 3.1 sont présentes les dis-

2. Ce cas est relativement rare.

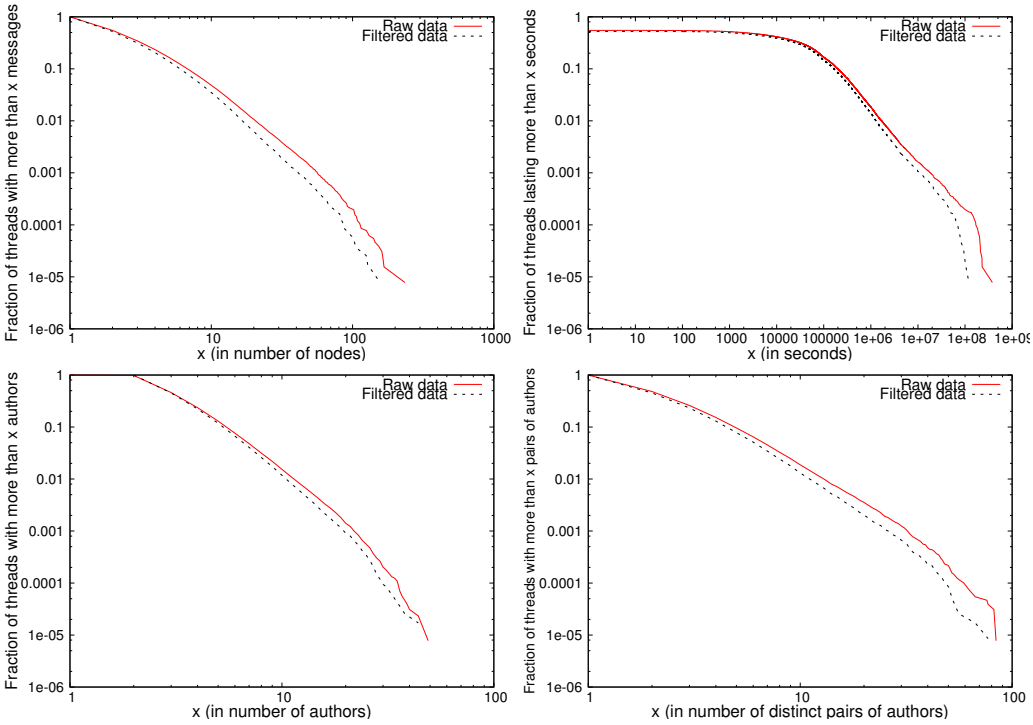


FIGURE 3.1 – Distribution cumulative inverse de différentes caractéristiques pour les données brutes (ligne pleine) et filtrées (ligne pointillé). En haut à gauche : nombre de courriels dans une discussion ; en haut à droite : durée d'une discussion ; en bas à gauche : nombre de personnes dans une discussion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

tributions cumulatives inverses de ces notions et on remarque qu'elles sont hétérogènes.

La distribution des durées des discussions, mets en avant que la majorité des discussions dure environ une journée (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 a été fixée à 2 ans. Enfin, on remarque que les données filtrées ne diffèrent pas qualitativement des données brutes.

Ces premières observations sont nécessaires mais pas suffisante pour comprendre les caractéristiques d'une discussion. Nous avons également étudié la 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, figure de 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 également

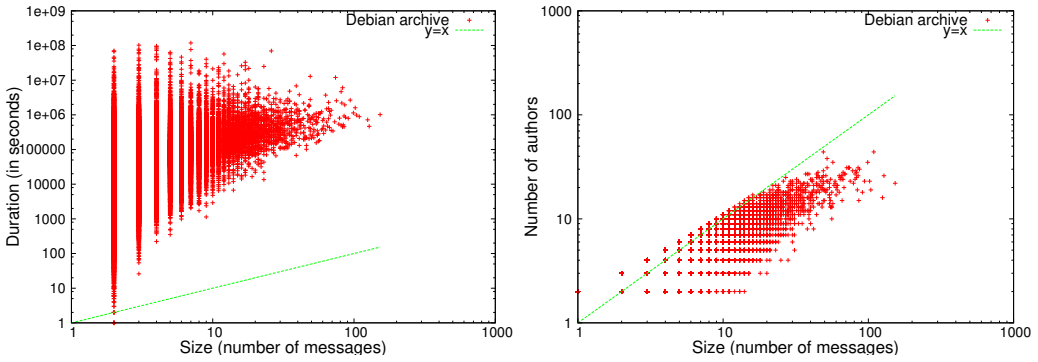


FIGURE 3.2 – Gauche : Corrélations entre le nombre de courriels et la durée d'une discussion. Droite : Corrélation entre le nombre de courriels et le nombre d'auteurs dans une discussion.

que les petites discussions ont des durées très variables. Sur la figure de droite présentant la corrélation entre le nombre de courriels et d'auteurs, on observe également un fait attendu [1] qui est qu'il y a plus de courriels que d'auteurs. Ainsi lors d'une discussion, c'est un petit nombre de personnes qui partagent beaucoup de messages.

$t_0 = \alpha$ et $t_k + 1 = \omega$

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 d'une pair de nœuds u et v dans V. 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 d'un fil de discussions ouvert où tout le monde peut participer. En particuliers, une personne peut envoyer une demande d'aide à un moment donner et ne plus jamais participer.

3.3 Étude des discussions en tant que sous-flots

3.3.1 Applications de la densité aux discussions

Jusqu'à maintenant aucune notion intrinsèquement liées aux flot de liens n'a été utilisée pour caractériser les discussion et leur répartition dans le flot de liens. Le but est d'évaluer si cette structure de flot peut se rapprocher d'une structure communautaire. Comme dit précédemment, les communautés sont

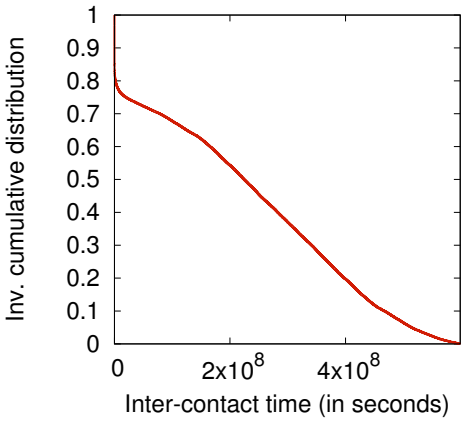


FIGURE 3.3 – Distribution des temps inter-contacts dans le fil de discussions.

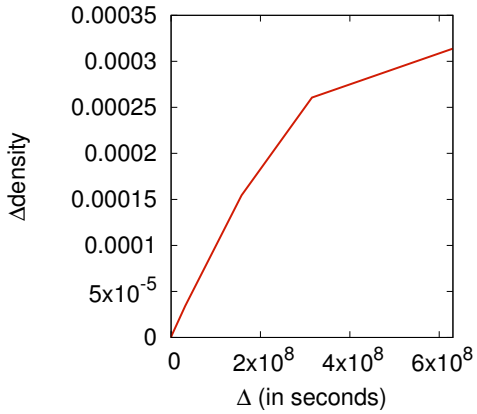


FIGURE 3.4 – Évolution de la Δ-densité du flot de liens pour Δ de 1 seconde à 20 ans.

souvent définies comme étant des structure 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. Tout d'abord sur la figure 3.4 est représenté la Δ-densité globale de tout le flot de liens. 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 la densité du graphe agrégé lorsque Δ est proche de $\omega - \alpha$.

Ajout bar horizontal de la densité statique.

La Δ-densité du flot liens, cependant, n'apporte que peu d'informations

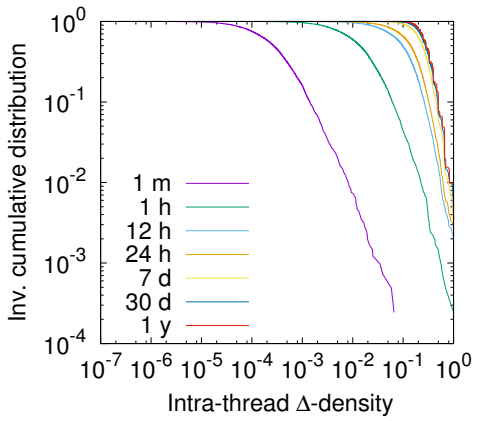


FIGURE 3.5 – Distribution cumulative inverse de la Δ-densité des discussions pour différentes valeurs de Δs.

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 valeur de Δ, entre 2.69×10^{-4} et 0.28 alors que le flot a une densité 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 une preuve pour définir une structure comme 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.

Faire SNAM sur les threads

Une autre définition d'une communauté est qu'une communauté 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, $\delta^{inter}(C_i)$:

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

où $\delta(G_{ij})$ est la densité du graphe constitué des nœuds des communautés i et j et liens de G entre ces deux communautés. Il s'agit tout simplement de la probabilité qu'il y 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 lien et il est nécessaire de l'adapter. Pour ce faire, nous définissons le sous-flot inter-discussion entre les discussion D_i et D_j de la manière suivante : $L_{ij} = (T_{ij}, V_{ij}, E_{ij})$, avec $T_{ij} = [\min(\alpha_i, \alpha_j), \max(\omega_i, \omega_j)]$, $V_{ij} = V_i \cup V_j$ et $E_{ij} = \{(t,u,v) : t \in T_{ij}, u, v \in V_{ij}, (t,u,v) \in E \setminus D_i \cup D_j\}$. La Δ-densité inter discussions entre D_i et D_j est alors la Δ-densité du sous-flot L_{ij}. 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 :

C'est pas vraiment ça ...

Faire un dessin de cette situation ?

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

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 qualitatif de la Δ-densité inter discussions diffère de la Δ-densité des discussions. La Δ-densité inter discussions croit également en fonction de Δ mais il y a toujours une différence notable entre $\Delta = 1$ *mois* et $\Delta = 1$ *année* ce qui n'est pas le cas pour la Δ-densité. Cette différence est normal 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.

Peut etre aussi parce que L_{ij} dure plus longtemps

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 dense intérieurement qu'avec les autres discussions. La différence est de plusieurs ordres de grandeur lorsque Δ est petit et elle diminue lorsque Δ croit. Pour $\Delta = 20$ *ans* sur la figure 3.7d, la différence n'est plus visible. A cette échelle de temps, l'ancrage temporel des discussions n'est plus décisif. On remarque que même pour $\Delta = 1$ *an*, la différence est notable.

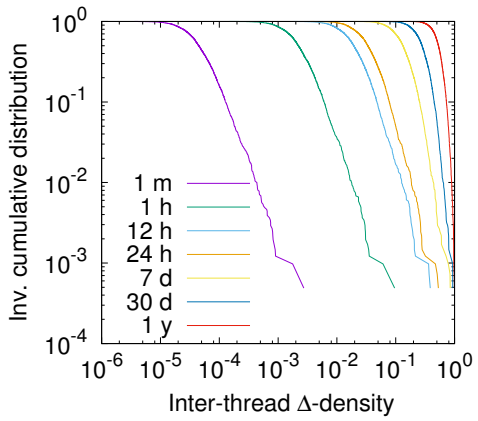
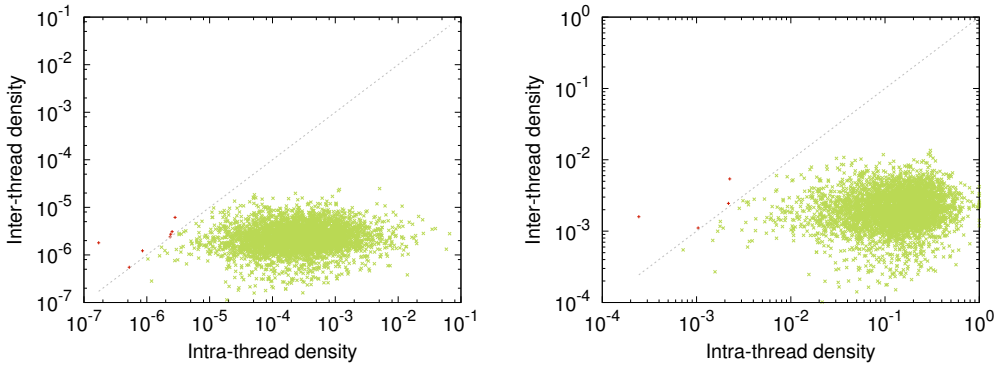


FIGURE 3.6 – Distribution cumulative inverse de la Δ-densité inter discussions pour différentes valeurs de Δs.

3.3.2 Graphe et flot quotient des discussions

3.4 Détection automatique des discussions ?

Nope



(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 *Delta*. Une discussion est en vert (resp. rouge) si elle a une Δ-densité plus (resp. moins) élevée que sa Δ-densité inter 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